

**Comprehensive Compendium:**  
Calculus II, Calculus III, Newtonian Mechanics, Probability and Statistics, Linear algebra

**Nathan Warner**



**Northern Illinois  
University**

Computer Science  
Northern Illinois University  
August 28, 2023  
United States

## Contents

<b>1</b>	<b>Calculus II</b>	<b>5</b>
1.1	Chapter 1 Definitions and Theorems . . . . .	5
1.2	Chapter 2 Definitions and Theorems . . . . .	6
1.3	Chapter 3 Definitions and Theorems . . . . .	9
1.4	Chapter 5 Definitions and Theorems . . . . .	14
1.5	Chapter 6 Definitions and Theorems . . . . .	22
1.6	Chapter 6 Problems to Remember . . . . .	29
<b>2</b>	<b>Fundemental Physics: Classical Mechanics</b>	<b>32</b>
2.1	Chapter 1: Units and Measurement . . . . .	32
2.1.1	Key Terms . . . . .	32
2.1.2	Defintions and Theorems (and important things) . . . . .	34
2.1.3	Fundemental Tables and figures . . . . .	38
2.1.4	Memorize conversions . . . . .	40
2.1.5	Problems to remember . . . . .	41
2.2	Chapter 2: Vectors . . . . .	42
2.2.1	Vocabulary . . . . .	42
2.2.2	Definitions and theorems (and important things) . . . . .	43
2.2.3	Problems to remember . . . . .	51
2.3	Chapter 3: Motion along a straight line . . . . .	52
2.3.1	Definitions and theorems . . . . .	52
2.4	Chapter 4: Motion in two and three dimensions . . . . .	56
2.4.1	Definitions and theorems . . . . .	56
2.5	Chapter 5: Newton's laws of motion . . . . .	63
2.5.1	Definitions and Theorems . . . . .	63
2.6	Chapter 6: Applications of Newton's Laws . . . . .	69

2.6.1	Definitions and theorems . . . . .	69
2.7	Chapter 7: Work and Kinetic energy . . . . .	74
2.7.1	Definitions and theorems . . . . .	74
2.8	Chapter 8: Potential Energy and Conservation of Energy . . . . .	78
2.8.1	Definitions and theorems . . . . .	78
2.9	Chapter 13: Gravitation . . . . .	82
2.9.1	Definitions and theorems . . . . .	82
2.10	Chapter 9: Linear momentum and collisions . . . . .	88
2.10.1	Definitions and Theorems . . . . .	88
2.11	Chapter 10: Fixed-axis rotation . . . . .	94
2.11.1	Definitions and theorems . . . . .	94
2.12	Chapter 11: Angular Momentum . . . . .	98
2.12.1	Definitions and Theorems . . . . .	98
2.13	Chapter 12: Static equilibrium and elasticity . . . . .	100
2.13.1	Definitions and Theorems . . . . .	100
<b>3</b>	<b>Newtonian Mechanics Formulas Simplified</b>	101
3.1	General . . . . .	101
3.2	Chapter 3 & 4: Kinematic equations . . . . .	102
3.3	Chapter 5 & 6: Newtons laws of motion (forces) . . . . .	103
3.4	Chapter 7 & 8: Work and energy . . . . .	104
3.5	Chapter 13: Gravitation . . . . .	105
3.6	Chapter 9: Momentum, Impulse, collisions, center of mass, average force, rocket equation . . . . .	106
3.7	Circular Motion . . . . .	107
3.8	Chapter 10 & 11: Fixed axis rotation and angular momentum . . . . .	108
<b>4</b>	<b>Calculus III</b>	111
4.1	Chapter 1: Parametric equations and polar coordinates . . . . .	111
4.1.1	Definitions and Theorems . . . . .	111
4.1.2	Problems to remember . . . . .	118
4.2	Chapter 2: Vectors in Space . . . . .	120
4.2.1	Definitions and Theorems . . . . .	120
4.3	Conic sections and quadric surfaces . . . . .	131

4.3.1	Conic sections: Parabola, Ellipse, and Hyperbola . . . . .	131
4.3.2	Quadric Surfaces . . . . .	136
4.4	Chapter 3: Vector-Valued Functions . . . . .	139
4.4.1	Definitions and Theorems . . . . .	139
4.5	Chapter 4: Differentiation of Functions of Several Variables . . . . .	146
4.5.1	Definitions and Theorems . . . . .	146
4.6	Chapter 5: Multiple integration . . . . .	160
4.6.1	Definitions and theorems . . . . .	160
4.7	Chapter 6: Vector calculus . . . . .	172
4.7.1	Definitions and theorems . . . . .	172
<b>5</b>	<b>Probability and Statistics</b>	<b>178</b>
5.1	Chapter 1: Overview and Descriptive Statistics . . . . .	178
5.1.1	Definitions and Theorems . . . . .	178
5.1.2	Frequency Distribution . . . . .	182
5.1.3	Steam and Leaf Displays . . . . .	183
5.1.4	Dotplots . . . . .	184
5.1.5	Histogram for discrete data . . . . .	185
5.1.6	Histogram for continuous data: Equal class widths . . . . .	186
5.1.7	Histogram for Continuous Data: Unequal Class Widths . . . . .	187
5.1.8	Boxplots . . . . .	188
5.2	Chapter 2: Probability . . . . .	189
5.2.1	Definitions and Theorems . . . . .	189
5.2.2	Examples in Probability Theory . . . . .	195
5.2.3	Examples in counting . . . . .	197
5.3	Chapter 3: Discrete Random Variables and Probability Distributions . . . . .	198
5.3.1	Definitions and Theorems . . . . .	198
5.3.2	Discrete rv distribution problems . . . . .	209
5.4	Chapter 4: Continuous Random variables and Probability Distributions . . . . .	211
5.4.1	Definitions and Theorems . . . . .	211
5.5	Chapter 5: Joint Probability Distributions and Random Samples . . . . .	219
5.5.1	Definitions and Theroems . . . . .	219
5.6	Chapter 7: Statistical Intervals Based on a Single Sample . . . . .	229

5.6.1	Definitions and Theorems . . . . .	229
5.7	Chapter 8: Tests of Hypotheses Based on a Single Sample . . . . .	235
5.7.1	Definitions and Theorems . . . . .	235
5.8	Chapter 9: Inferences Based on Two Samples . . . . .	243
5.8.1	Definitions and Theorems . . . . .	243
<b>6</b>	<b>Linear Algebra</b>	<b>247</b>
6.1	By The Professor . . . . .	247
6.1.1	Prelude . . . . .	247
6.1.2	Intro to linear systems and matrices . . . . .	253
6.1.3	Properties of linearity, more on linear systems and matrices . . . . .	256
6.1.4	Algebra of matrices, matrix operations . . . . .	265
6.1.5	Composition, rotations . . . . .	272
6.1.6	Linear transformations, surjective, injective, bijective. Invertibility, and basic uses of determinants . . . . .	276
6.1.7	The inverse of a square matrix, computation of determinants . . . . .	291
6.1.8	Eigenvectors, Eigenvalues . . . . .	300
6.1.9	Basis, change of basis, diagonalization . . . . .	302
6.1.10	Vector spaces, Abstract Vector Spaces, Subspaces . . . . .	308
6.1.11	Kernel, Image, Orthogonality, and the Rank Nullity theorem . . . . .	324
6.1.12	More on dimensions and the rank nullity theorem . . . . .	337
6.1.13	Linear algebra with complex numbers . . . . .	341
6.1.14	Proofs . . . . .	342
6.2	By The Book . . . . .	346
6.3	More on calculus with linear algebra . . . . .	347
6.4	Just Formulas and Theorems . . . . .	348
<b>7</b>	<b>Combinatorics</b>	<b>349</b>
7.1	Introduction . . . . .	349
7.2	Induction and recurrence relations . . . . .	351

## Calculus II

### 1.1 Chapter 1 Definitions and Theorems

- **Mean Value Theorem For Integrals:** If  $f(x)$  is continuous over an interval  $[a,b]$ , then there is at least one point  $c \in [a, b]$  such that

$$f(c) = \frac{1}{b-a} \int f(x) dx.$$

- **Integrals resulting in inverse trig functions:**

1.

$$\int \frac{dx}{\sqrt{a^2 - x^2}} = \sin^{-1} \frac{x}{|a|} + C.$$

2.

$$\int \frac{dx}{a^2 + x^2} = \frac{1}{a} \tan^{-1} \frac{x}{a} + C.$$

3.

$$\int \frac{dx}{x\sqrt{x^2 - a^2}} = \frac{1}{|a|} \sec^{-1} \frac{|x|}{a} + C.$$

## 1.2 Chapter 2 Definitions and Theorems

- **Area between two curves, integrating on the x-axis:**

$$A = \int_a^b [f(x) - g(x)] dx \quad (1)$$

Where  $f(x) \geq g(x)$

$$A = \int_a^b [g(x) - f(x)] dx.$$

for  $g(x) \geq f(x)$

- **Area between two curves, integrating on the y-axis:**

$$A = \int_c^d [u(y) - v(y)] dy \quad (2)$$

- **Areas of compound regions:**

$$\int_a^b |f(x) - g(x)| dx.$$

- **Area of complex regions:**

$$\int_a^b f(x) dx + \int_b^c g(x) dx.$$

- **Slicing Method:**

$$V(s) = \sum_{i=1}^n A(x_i^*) \Delta x = \int_a^b A(x) dx.$$

- **Disk Method along the x-axis:**

$$V = \int_a^b \pi[f(x)]^2 dx \quad (3)$$

- **Disk Method along the y-axis:**

$$V = \int_c^d \pi[g(y)]^2 dy \quad (4)$$

- **Washer Method along the x-axis:**

$$V = \int_a^b \pi[(f(x))^2 - (g(x))^2] dx \quad (5)$$

- **Washer Method along the y-axis:**

$$V = \int_c^d \pi[(u(y))^2 - (v(y))^2] dy \quad (6)$$

- Radius if revolved around other line (Washer Method):

$$\begin{aligned} If: & \quad x = -k \\ Then: & \quad r = Function + k. \end{aligned}$$

$$\begin{aligned} If: & \quad x = k \\ Then: & \quad r = k - Function. \end{aligned}$$

- Method of Cylindrical Shells (x-axis):

$$V = \int_a^b 2\pi x f(x) dx \quad (7)$$

- Method of Cylindrical Shells (y-axis):

$$V = \int_c^d 2\pi y g(y) dy \quad (8)$$

- Region revolved around other line (method of cylindrical shells)::

$$\begin{aligned} If: & \quad x = -k \\ Then: & \quad V = \int_a^b 2\pi(x + k)(f(x)) dx. \end{aligned}$$

$$\begin{aligned} If: & \quad x = k \\ Then: & \quad V = \int_a^b 2\pi(k - x)(f(x)) dx. \end{aligned}$$

- A Region of Revolution Bounded by the Graphs of Two Functions (method cylindrical shells):

$$V = \int_a^b 2\pi x [f(x) - g(x)] dx.$$

- Arc Length of a Function of x:

$$\text{Arc Length} = \int_a^b \sqrt{1 + [f'(x)]^2} dx \quad (9)$$

- Arc Length of a Function of y:

$$\text{Arc Length} = \int_c^d \sqrt{1 + [g'(y)]^2} dy \quad (10)$$

- Surface Area of a Function of x (Around x):

$$\text{Surface Area} = \int_a^b 2\pi f(x) \sqrt{1 + [f'(x)]^2} dx \quad (11)$$

- Surface Area of a Function of x (Around y):

$$\text{Surface Area} = \int_a^b 2\pi x \sqrt{1 + [f'(x)]^2} dx \quad (12)$$

$$\text{Or: } \int_a^b 2\pi u(y) \sqrt{1 + (u'(y))^2} dy \quad (13)$$

- **Natural logarithm function:**

$$\ln x = \int_1^x \frac{1}{t} dt \quad (14)$$

- **Exponential function:**

$$y = e^x, \quad \ln y = \ln(e^x) = x \quad (15)$$

- **Logarithm Differentiation:**

$$f'(x) = f(x) \cdot \frac{d}{dx} \ln(f'(x)).$$

**Note:** Use properties of logs before you differentiate whats inside the logarithm

### 1.3 Chapter 3 Definitions and Theorems

- Integration by parts formula:

$$\int u \, dv = uv - \int v \, du.$$

- Integration by parts for definite integral:

$$\int_a^b u \, dv = uv \Big|_a^b - \int_a^b v \, du$$

- To integrate products involving  $\sin(ax)$ ,  $\sin(bx)$ ,  $\cos(ax)$ , and  $\cos(bx)$ , use the substitutions::

– Sine Products:

$$\sin(ax) \sin(bx) = \frac{1}{2} \cos((a-b)x) - \frac{1}{2} \cos((a+b)x)$$

– Sine and Cosine Products:

$$\sin(ax) \cos(bx) = \frac{1}{2} \sin((a-b)x) + \frac{1}{2} \sin((a+b)x)$$

– Cosine Products:

$$\cos(ax) \cos(bx) = \frac{1}{2} \cos((a-b)x) + \frac{1}{2} \cos((a+b)x)$$

– Power Reduction Formula (sine):

$$\begin{aligned} \int \sin^n x \, dx &= -\frac{1}{n} \sin^{n-1} x \cos x + \frac{n-1}{n} \int \sin^{n-2} x \, dx \\ \int_0^{\frac{\pi}{2}} \sin^n x \, dx &= \frac{n-1}{n} \int_0^{\frac{\pi}{2}} \sin^{n-2} x \, dx. \end{aligned}$$

– Power Reduction Formula (cosine):

$$\begin{aligned} \int \cos^n x \, dx &= \frac{1}{n} \cos^{n-1} x \sin x + \frac{n-1}{n} \int \cos^{n-2} x \, dx \\ \int_0^{\frac{\pi}{2}} \cos^n x \, dx &= \frac{n-1}{n} \int_0^{\frac{\pi}{2}} \cos^{n-2} x \, dx. \end{aligned}$$

– Power Reduction Formula (secant):

$$\begin{aligned} \int \sec^n x \, dx &= \frac{1}{n-1} \sec^{n-1} x \sin x + \frac{n-2}{n-1} \int \sec^{n-2} x \, dx \\ \int \sec^n x \, dx &= \frac{1}{n-1} \sec^{n-2} x \tan x + \frac{n-2}{n-1} \int \sec^{n-2} x \, dx \end{aligned}$$

– Power Reduction Formula (tangent):

$$\int \tan^n x \, dx = \frac{1}{n-1} \tan^{n-1} x - \int \tan^{n-2} x \, dx$$

- Trigonometric Substitution:

- $\sqrt{a^2 - x^2}$  use  $x = a \sin \theta$  with domain restriction  $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$
- $\sqrt{a^2 + x^2}$  use  $x = a \tan \theta$  with domain restriction  $(-\frac{\pi}{2}, \frac{\pi}{2})$
- $\sqrt{x^2 - a^2}$  use  $x = a \sec \theta$  with domain restriction  $\left[0, \frac{\pi}{2}\right) \cup \left[\pi, \frac{3\pi}{2}\right)$

- **Steps for fraction decomposition:**

1. Ensure  $\deg(Q) < \deg(P)$ , if not, long divide
2. Factor denominator
3. Split up fraction into factors
4. Multiply through to clear denominator
5. Group terms and equalize
6. Solve for constants
7. Plug constants into split up fraction
8. Compute integral

- **Solving for constants:** Either:

- Plug in values (often the roots)
- Equalize

- **Cases for partial fractions:**

- Non repeated linear factors
- Repeated linear factors
- Nonfactorable quadratic factors

- **Midpoint rule:**

$$M_n = \sum_{i=1}^n f(m_i) \Delta x.$$

- **Absolute error:**

$$err = \left| \text{Actual} - \text{Estimated} \right|.$$

- **Relative error:**

$$err = \left| \frac{\text{Actual} - \text{Estimated}}{\text{Actual}} \right| \cdot 100\%.$$

- **Error upper bound for midpoint rule:**

$$E_M \leq \frac{M(b-a)^3}{24n^2}$$

Where  $M$  is the maximum value of the second derivative

- **Trapezoidal rule:**

$$T_n \frac{1}{2} \Delta x (f(x_0) + 2f(x_1) + 2f(x_2) + \cdots + 2f(x_{n-1}) + f(x_n))$$

- **Error upper bound for trapezoidal rule:**

$$E_T \leq \frac{M(b-a)^3}{12n^2}$$

Where  $M$  is the maximum value of the second derivative

- **Simpson's rule:**

$$S_n = \frac{\Delta x}{3} (f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + 2f(x_4) + 4f(x_5) + \cdots + 2f(x_{n-2}) + 4f(x_{n-1}) + f(x_n))$$

- **Error upper bound for Simpson's rule:**

$$E_S \leq \frac{M(b-a)^5}{180n^4}$$

Where  $M$  is the maximum value of the fourth derivative

- **Finding  $n$  with error bound functions:**

1. Find  $f''(x)$
2. Find maximum values of  $f''(x)$  in the interval
3. Plug into error bound function
4. Set value  $\leq$  desired accuracy (ex: 0.01)
5. Solve:
6. If we were to truncate, we would use the ceil function  $\lceil n \rceil$  DO NOT FLOOR

- **Improper integrals (Infinite interval):**

$$\begin{aligned} - \int_a^{+\infty} f(x) dx &= \lim_{t \rightarrow +\infty} \int_a^t f(x) dx \\ - \int_{-\infty}^b f(x) dx &= \lim_{t \rightarrow -\infty} \int_t^b f(x) dx \\ - \int_{-\infty}^{+\infty} f(x) dx &= \int_{-\infty}^0 f(x) dx + \int_0^{+\infty} f(x) dx \end{aligned}$$

- **Improper integral (discontinuous):**

- Let  $f(x)$  be continuous on  $[a, b]$ , then;

$$\int_a^b f(x) dx = \lim_{t \rightarrow b^-} \int_a^t f(x) dx .$$

- Let  $f(x)$  be continuous on  $(a, b]$ , then;

$$\int_a^b f(x) dx = \lim_{t \rightarrow b^+} \int_t^b f(x) dx .$$

In each case, if the limit exists, then the improper integral is said to converge. If the limit does not exist, then the improper integral is said to diverge.

- Let  $f(x)$  be continuous on  $[a, b]$  except at a point  $c \in (a, b)$ , then;

$$\int_a^b f(x) dx = \int_a^c f(x) dx + \int_c^b f(x) dx .$$

If either integral diverges, then  $\int_a^b f(x) dx$  diverges

- **Comparison theorem:** Let  $f(x)$  and  $g(x)$  be continuous over  $[a, +\infty)$ . Assume that  $0 \leq f(x) \leq g(x)$  for  $x \geq a$ .
  - If  $\int_a^{+\infty} f(x) dx = \lim_{t \rightarrow +\infty} \int_a^t f(x) dx = +\infty$ ,  
then  $\int_a^{+\infty} g(x) dx = \lim_{t \rightarrow +\infty} \int_a^t g(x) dx = +\infty$ .
  - If  $\int_a^{+\infty} g(x) dx = \lim_{t \rightarrow +\infty} \int_a^t g(x) dx = L$ , where  $L$  is a real number,  
then  $\int_a^{+\infty} f(x) dx = \lim_{t \rightarrow +\infty} \int_a^t f(x) dx = M$  for some real number  $M \leq L$ .

- **P-integrals:**

$$\begin{aligned}
 - \int_0^{+\infty} \frac{1}{x^p} dx &= \begin{cases} \frac{1}{p-1} & \text{if } p > 1 \\ +\infty & \text{if } p \leq 1 \end{cases} \\
 - \int_0^1 \frac{1}{x^p} dx &= \begin{cases} \frac{1}{1-p} & \text{if } p < 1 \\ +\infty & \text{if } p \geq 1 \end{cases} \\
 - \int_a^{+\infty} \frac{1}{x^p} dx &= \begin{cases} \frac{a^{1-p}}{p-1} & \text{if } p > 1 \\ +\infty & \text{if } p \leq 1 \end{cases} \\
 - \int_0^a \frac{1}{x^p} dx &= \begin{cases} \frac{a^{1-p}}{1-p} & \text{if } p < 1 \\ +\infty & \text{if } p \geq 1 \end{cases}
 \end{aligned}$$

- **Bypass L'Hospital's Rule:**

$$\ln(\ln(x)), \ln(x), \dots, x^{\frac{1}{100}}, x^{\frac{1}{3}}, \sqrt{x}, 1, x^2, x^3, \dots, e^x, e^{2x}, e^{3x}, \dots, e^{x^2}, \dots, e^{e^x}.$$

Essentially what it means is things on the right grow faster than things on the left.  
Thus, if we have say:

$$\lim_{x \rightarrow \infty} \frac{x^2}{e^{2x}}.$$

We can be sure that it is zero. Because this is  $x^2 \cdot e^{-2x}$ . If we take  $\lim_{x \rightarrow \infty} x^2 e^{-2x}$ , we get  $\infty \cdot 0$ . As we see by the sequence  $e^{-2x}$  overrules  $x^2$  and we can say the limit is zero.

- **Consideration for Limits:** Let  $f : A \rightarrow B$  be a function defined by  $x \mapsto f(x)$ . If a point  $c$  lies outside the domain  $A$ , then the expression  $\lim_{x \rightarrow c} f(x)$  is not meaningful, and we classify this limit as undefined. For instance, the function arcsine has a domain of  $[-1, 1]$ . Therefore, limits like  $\lim_{x \rightarrow a} \sin^{-1} x$  where  $a \notin [-1, 1]$  are undefined.

- **Why does:**

$$\lim_{x \rightarrow 2} \tan^{-1} \frac{1}{x-2}.$$

$$\begin{aligned}
 &= \lim_{x \rightarrow 2^-} \tan^{-1} \frac{1}{x-2} && = \lim_{x \rightarrow 2^+} \tan^{-1} \frac{1}{x-2} \\
 &= \lim_{x \rightarrow -\infty} \tan^{-1} x && = \lim_{x \rightarrow +\infty} \tan^{-1} x \\
 &= -\pi/2. && = \frac{\pi}{2}.
 \end{aligned}$$

## 1.4 Chapter 5 Definitions and Theorems

- Sequence notation:

$$\{a_n\}_{n=1}^{\infty}, \text{ or simply } \{a_n\}.$$

- Sequence notation (ordered list):

$$a_1, a_2, a_3, \dots, a_n, \dots$$

- Arithmetic Sequence Difference:

$$d = a_n - a_{n-1}.$$

- Arithmetic sequence (common difference between subsequent terms) general form:

$$\text{Index starting at 0 : } a_n = a + nd$$

$$\text{Index starting at 1 : } a_n = a + (n - 1)d$$

- Arithmetic sequence (common difference between subsequent terms) recursive form:

$$a_n = a_{n-1} + d.$$

- Sum of arithmetic sequence:

$$S_n = \frac{n}{2} [a + a_n]$$

$$S_n = \frac{n}{2} [2a + (n - 1)d].$$

- Geometric sequence form common ratio:

$$r = \frac{a_n}{a_{n-1}}.$$

- Geometric sequence general form:

$$a_n = ar^n \text{ (Index starting at 0)}$$

$$a_n = a^{n+1} \text{ (index starting at 0 and a=r)}$$

$$a_n = ar^{n-1} \text{ (Index starting at 1)}$$

$$a_n = a^n \text{ (index starting at 1 and a=r).}$$

- Geometric sequence recursive form:

$$a_n = ra_{n-1}.$$

- Sum of geometric sequence (finite terms):

$$S_n = \frac{a(1 - r^n)}{1 - r} \quad r \neq 1.$$

- **Convergence / Divergence:** If

$$\lim_{n \rightarrow +\infty} a_n = L.$$

We say that the sequence converges, else it diverges

- **Formal definition of limit of sequence:**

$$\lim_{n \rightarrow +\infty} a_n = L \iff \forall \varepsilon > 0, \exists N \in \mathbb{Z} \mid |a_n - L| < \varepsilon, \text{ if } n \geq N.$$

Then we can say

$$\lim_{n \rightarrow +\infty} a_n = L \text{ or } a_n \rightarrow L.$$

- **Limit of a sequence defined by a function:** Consider a sequence  $\{a_n\}$  : such that  $a_n = f(n)$  for all  $n \geq 1$ . If there exists a real number  $L$  such that

$$\lim_{x \rightarrow \infty} f(x) = L,$$

then  $\{a_n\}$  converges and

$$\lim_{n \rightarrow \infty} a_n = L.$$

- **Algebraic limit laws:** Given sequences  $\{a_n\}$  and  $\{b_n\}$  and any real number  $c$ , if there exist constants  $A$  and  $B$  such that  $\lim_{n \rightarrow \infty} a_n = A$  and  $\lim_{n \rightarrow \infty} b_n = B$ , then

- $\lim_{n \rightarrow \infty} c = c$
- $\lim_{n \rightarrow \infty} ca_n = c \lim_{n \rightarrow \infty} a_n = cA$
- $\lim_{n \rightarrow \infty} (a_n \pm b_n) = \lim_{n \rightarrow \infty} a_n \pm \lim_{n \rightarrow \infty} b_n = A \pm B$
- $\lim_{n \rightarrow \infty} (a_n \cdot b_n) = (\lim_{n \rightarrow \infty} a_n) \cdot (\lim_{n \rightarrow \infty} b_n) = A \cdot B$
- $\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = \frac{\lim_{n \rightarrow \infty} a_n}{\lim_{n \rightarrow \infty} b_n} = \frac{A}{B}$ , provided  $B \neq 0$  and each  $b_n \neq 0$ .

- **Continuous Functions Defined on Convergent Sequences:** Consider a sequence  $\{a_n\}$  and suppose there exists a real number  $L$  such that the sequence  $\{a_n\}$  converges to  $L$ . Suppose  $f$  is a continuous function at  $L$ . Then there exists an integer  $N$  such that  $f$  is defined at all values  $a_n$  for  $n \geq N$ , and the sequence  $\{f(a_n)\}$  converges to  $f(L)$ .

- **Squeeze Theorem for Sequences:** Consider sequences  $\{a_n\}$ ,  $\{b_n\}$ , and  $\{c_n\}$  :: Suppose there exists an integer  $N$  such that

$$a_n \leq b_n \leq c_n \text{ for all } n \geq N.$$

If there exists a real number  $L$  such that

$$\lim_{n \rightarrow \infty} a_n = L = \lim_{n \rightarrow \infty} c_n,$$

then  $\{b_n\}$  converges and  $\lim_{n \rightarrow \infty} b_n = L$

- **Bounded above:** A sequence  $\{a_n\}$  : is bounded above if there exists a real number  $M$  such that

$$a_n \leq M$$

for all positive integers  $n$ .

- **Bounded below:** A sequence  $\{a_n\}$  is bounded below if there exists a real number  $M$  such that

$$M \leq a_n$$

for all positive integers  $n$ .

- **Bounded:** A sequence  $\{a_n\}$  is a bounded sequence if it is bounded above and bounded below.
- **Unbounded:** If a sequence is not bounded, it is an unbounded sequence.
- **If a sequence  $\{a_n\}$  converges, then it is bounded.:**
- **Increasing sequence:** A sequence  $\{a_n\}$  : is increasing for all  $n \geq n_0$  if

$$a_n \leq a_{n+1} \text{ for all } n \geq n_0.$$

- **Decreasing sequence:** A sequence  $\{a_n\}$  : is decreasing for all  $n \geq n_0$  if

$$a_n \geq a_{n+1} \text{ for all } n \geq n_0.$$

- **monotone sequence:** for all  $n \geq n_0$  if it is increasing for all  $n \geq n_0$  or decreasing for all  $n \geq n_0$
- **Monotone Convergence Theorem:** If  $\{a_n\}$  is a bounded sequence and there exists a positive integer  $n_0$  such that  $\{a_n\}$  is monotone for all  $n \geq n_0$ , then  $\{a_n\}$  : converges.
- **Infinite Series form::**

$$\sum_{n=1}^{\infty} a_n = a_1 + a_2 + a_3 + \dots$$

- **Partial sum ( $k^{th}$  partial sum):**

$$S_k = \sum_{n=1}^k a_n = a_1 + a_2 + a_3 + \dots + a_k.$$

- **Convergence of infinity series notation:**

For a series, say...

$$\sum_{n=1}^{\infty} a_n .$$

its convergence is determined by the limit of its sequence of partial sums. Specifically, if

$$\lim_{n \rightarrow +\infty} S_n = S \rightarrow \sum_{n=1}^{\infty} a_n = S.$$

- **Harmonic series:**

$$\sum_{n=1}^{\infty} \frac{1}{n} = \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \dots$$

Which diverges to  $+\infty$

- **Algebraic Properties of Convergent Series:** Let  $\sum_{n=1}^{\infty} a_n$  and  $\sum_{n=1}^{\infty} b_n$  be convergent series. Then the following algebraic properties hold:

1. The series  $\sum_{n=1}^{\infty} (a_n + b_n)$  converges and

$$\sum_{n=1}^{\infty} (a_n + b_n) = \sum_{n=1}^{\infty} a_n + \sum_{n=1}^{\infty} b_n. \quad (\text{Sum Rule}).$$

2. The series  $\sum_{n=1}^{\infty} (a_n - b_n)$  converges and

$$\sum_{n=1}^{\infty} (a_n - b_n) = \sum_{n=1}^{\infty} a_n - \sum_{n=1}^{\infty} b_n. \quad (\text{Difference Rule}).$$

3. For any real number  $c$ , the series  $\sum_{n=1}^{\infty} ca_n$  converges and

$$\sum_{n=1}^{\infty} ca_n = c \sum_{n=1}^{\infty} a_n. \quad (\text{Constant Multiple Rule}).$$

- **Geometric series convergence or divergence:** :

$$\sum_{n=1}^{\infty} ar^{n-1} = \begin{cases} \frac{a}{1-r} & \text{if } |r| < 1 \\ \text{diverges} & \text{if } |r| \geq 1 \end{cases}.$$

- **Divergence test:** In the context of sequences, if  $\lim_{n \rightarrow \infty} a_n = c \neq 0$  or the limit does not exist, then the series  $\sum_{n=1}^{\infty} a_n$  is said to diverge. The converse is not true.

Because:

$$\lim_{k \rightarrow \infty} a_k = \lim_{k \rightarrow \infty} (S_k - S_{k-1}) = \lim_{k \rightarrow \infty} S_k - \lim_{k \rightarrow \infty} S_{k-1} = S - S = 0..$$

- **Integral Test Prelude:** for any integer  $k$ , the  $k$ th partial sum  $S_k$  satisfies

$$S_k = a_1 + a_2 + a_3 + \cdots + a_k < a_1 + \int_1^k f(x) dx < a_1 + \int_1^{\infty} f(x) dx..$$

and

$$S_k = a_1 + a_2 + a_3 + \cdots + a_k > \int_1^{k+1} f(x) dx..$$

- **Integral test:** Suppose  $\sum_{n=1}^{\infty} a_n$  is a series with positive terms  $a_n$ . Suppose there exists a function  $f$  and a positive integer  $N$  such that the following three conditions are satisfied:

1.  $f$  positive, continuous, and decreasing on  $[N, \infty)$

2.  $f(n) = a_n$  for all integers  $n \geq N$ ,  $N \in \mathbb{Z}^+$

Then the series  $\sum_{n=1}^{\infty} a_n$  and the improper integral  $\int_N^{\infty} f(x) dx$  either both converge or both diverge..

- **P-series:**  $\forall p \in \mathbb{R}$ , the series

$$\sum_{n=1}^{\infty} \frac{1}{n^P}.$$

Is called a **p-series**. Furthermore,

$$\sum_{n=1}^{\infty} \frac{1}{n^p} \begin{cases} \text{converges if } p > 1 \\ \text{diverges if } p \leq 1. \end{cases}.$$

- **P-series extended:**

$$\sum_{n=2}^{\infty} \frac{1}{n \ln(n)^p} \begin{cases} \text{converges if } p > 1 \\ \text{diverges if } p \leq 1. \end{cases}.$$

- **Remainder estimate for the integral test:** Suppose  $\sum_{n=1}^{\infty} a_n$  is a convergent series with positive terms. Suppose there exists a function  $f$  and a positive integer  $M$  satisfying the following three conditions:

1.  $f$  is positive, decreasing, and continuous on  $[M, \infty)$
2.  $f(n) = a_n$  for all integers  $n \geq M$ .

Let  $S_N$  be the  $N$ th partial sum of  $\sum_{n=1}^{\infty} a_n$ . For all positive integers  $N$ ,

$$S_N + \int_{N+1}^{\infty} f(x) dx < \sum_{n=1}^{\infty} a_n < S_N + \int_N^{\infty} f(x) dx.$$

In other words, the remainder  $R_N = \sum_{n=1}^{\infty} a_n - S_N = \sum_{n=N+1}^{\infty} a_n$  satisfies the following estimate:

$$\int_{N+1}^{\infty} f(x) dx < R_N < \int_N^{\infty} f(x) dx.$$

This is known as the remainder estimate

To find a value of  $N$  such that we are within a desired margin of error, Since we know  $R_N < \int_N^{\infty} f(x) dx$ . Simply compute the improper integral and set the result  $<$  the desired error to solve for  $N$

- **Find  $a_n$  given the expression for the partial sum:**

$$a_n = S_n - S_{n-1}.$$

- **telescoping series:** Telescoping series are a type of series where each term cancels out a part of another term, leaving only a few terms that do not cancel. When you sum the series, most of the terms collapse or "telescope," which simplifies the calculation of the sum. Here are some key points and generalizations you can note about telescoping series:

- Partial Fraction Decomposition
- Cancellation Pattern: In a telescoping series, look for a pattern where a term in one fraction will cancel out with a term in another fraction.
- Write out Terms
- What is left is  $S_n$ , thus the sum of the series is the  $\lim_{n \rightarrow \infty} S_n$

Try:

$$\sum_{n=2}^{\infty} \frac{1}{n^2 - 1}.$$

Hint, its not only the first and last terms cancel, we also have a  $\frac{1}{n}$ , when  $a_{n-1}$ : Answer is  $\frac{3}{4}$

- **Comparison test for series:**

1. Suppose there exists an integer  $N$  such that  $0 \leq a_n \leq b_n$  for all  $n \geq N$ . If  $\sum_{n=1}^{\infty} b_n$  converges, then  $\sum_{n=1}^{\infty} a_n$  converges.
2. Suppose there exists an integer  $N$  such that  $a_n \geq b_n \geq 0$  for all  $n \geq N$ . If  $\sum_{n=1}^{\infty} b_n$  diverges, then  $\sum_{n=1}^{\infty} a_n$  diverges.

- **Limit Comparison Test:** Let  $a_n, b_n \geq 0$  for all  $n \geq 1$ .

- If  $\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = L \neq 0$ , then  $\sum_{n=1}^{\infty} a_n$  and  $\sum_{n=1}^{\infty} b_n$  both converge or both diverge.
- If  $\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = 0$  and  $\sum_{n=1}^{\infty} b_n$  converges, then  $\sum_{n=1}^{\infty} a_n$  converges.
- If  $\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = \infty$  and  $\sum_{n=1}^{\infty} b_n$  diverges, then  $\sum_{n=1}^{\infty} a_n$  diverges.

**Note:** Note that if  $\frac{a_n}{b_n} \rightarrow 0$  and  $\sum_{n=1}^{\infty} b_n$  diverges, the limit comparison test gives no information. Similarly, if  $\frac{a_n}{b_n} \rightarrow \infty$  and  $\sum_{n=1}^{\infty} b_n$  converges, the test also provides no information.

Consider the series

$$\sum_{n=1}^{\infty} \frac{n^4 + 6}{n^5 + 4}.$$

To find our  $b_n$  we can only focus on the leading coefficients. Thus:

$$b_n = \frac{n^4}{n^5} = \frac{1}{n}.$$

So our test...

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{a_n}{b_n} &= \frac{\frac{n^4 + 6}{n^5 + 4}}{\frac{1}{n}} \\ &= \lim_{n \rightarrow \infty} \frac{n(n^4 + 6)}{n^5 + 4} \\ &= \lim_{n \rightarrow \infty} \frac{n^5 + 6n}{n^5 + 4} \\ &= 1. \end{aligned}$$

Since  $\lim_{n \rightarrow \infty} \frac{a_n}{b_n} \neq 0 \vee +\infty$ . And  $\frac{1}{n}$  diverges, we can conclude that  $a_n$  will also diverge.

- **Determine which series (or function) is greater:**

- **Subtraction:** Given two functions  $f(x) = \frac{1}{x}$  and  $g(x) = \frac{x^4 + 6}{x^5 + 4}$  :, we want to compare them by considering the function  $h(x) = f(x) - g(x)$ :

$$h(x) = f(x) - g(x) = \frac{1}{x} - \frac{x^4 + 6}{x^5 + 4}$$

To compare these directly, it would be helpful to have a common denominator:

$$h(x) = \frac{x^4 + 4 - (x^4 + 6)}{x(x^5 + 4)} = \frac{-2}{x(x^5 + 4)}$$

Now, we can see that the sign of  $h(x)$  depends on the sign of  $x$  because the denominator  $x(x^5 + 4)$  is always positive for  $x \neq 0$ . So:

- \* For  $x > 0$ ,  $h(x) < 0$ , which means  $f(x) < g(x)$ .
- \* For  $x < 0$ ,  $h(x) > 0$ , which means  $f(x) > g(x)$ .

- **Alternating Series:** Any series whose terms alternate between positive and negative values is called an alternating series. An alternating series can be written in the form

$$\sum_{n=1}^{\infty} (-1)^{n+1} b_n = b_1 - b_2 + b_3 - b_4 + \dots$$

or

$$\sum_{n=1}^{\infty} (-1)^n b_n = -b_1 + b_2 - b_3 + b_4 - \dots$$

Where  $b_n > 0$  for all positive integers  $n$ .

- **alternating series test (Leibniz criterion):** An alternating series of the form

$$\sum_{n=1}^{\infty} (-1)^{n+1} b_n \quad \text{or} \quad \sum_{n=1}^{\infty} (-1)^n b_n$$

converges if

- $0 < b_{n+1} \leq b_n \forall n \geq 1$
- $\lim_{n \rightarrow \infty} b_n = 0$ .

**Note:** We remark that this theorem is true more generally as long as there exists some integer  $N$  such that  $0 < b_{n+1} \leq b_n$  for all  $n \geq N$ .

**Additional note:** The AST allows us to consider just the positive terms to check for these two conditions because if a series of decreasing positive terms that approach zero is alternated in sign, the alternating series will converge. This is a special property of alternating series that does not generally hold for non-alternating series.

- **Show decreasing (For the AST):** Consider the series

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^2}.$$

So you see we have  $b_n = \frac{1}{n^2}$ . For the AST, we must show that this is decreasing. If  $b_{n+1} = \frac{1}{(n+1)^2}$ . Then we see

$$\frac{1}{(n+1)^2} < \frac{1}{n^2}.$$

Thus it is decreasing for  $n \geq 1$  ( $b_{n+1} < b_n$ ) ■

- **Remainders in alternating series:** Consider an alternating series of the form

$$\sum_{n=1}^{\infty} (-1)^{n+1} b_n \quad \text{or} \quad \sum_{n=1}^{\infty} (-1)^n b_n,$$

that satisfies the hypotheses of the alternating series test. Let  $S$  denote the sum of the series and  $S_N$  denote the  $N$ -th partial sum. For any integer  $N \geq 1$ , the remainder  $R_N = S - S_N$  satisfies

$$|R_N| \leq b_{N+1}.$$

This tells us that if we stop at the  $N^{th}$  term, the error we are making is at most the size of the next term

- **Absolute and conditional convergence:**

- A series  $\sum_{n=1}^{\infty} a_n$  exhibits absolute convergence if  $\sum_{n=1}^{\infty} |a_n|$  converges.
- A series  $\sum_{n=1}^{\infty} a_n$  exhibits conditional convergence if  $\sum_{n=1}^{\infty} a_n$  converges but  $\sum_{n=1}^{\infty} |a_n|$  diverges.
- If  $\sum_{n=1}^{\infty} |a_n|$  converges then  $\sum_{n=1}^{\infty} a_n$  converges

**Note:** if  $|a_n|$  diverges, we cannot have absolute convergence, thus we must examine to see if normal  $a_n$  converges, in which case we would have conditional convergence

**Big Note:** If a series not strictly decreasing, we can still check for absolute/conditional convergence. Take  $\sum_{n=1}^{\infty} \frac{\sin n}{3^n + 4}$  for example.

- **Ratio test:** Let  $\sum_{n=1}^{\infty} a_n$  be a series with nonzero terms. Let

$$\rho = \lim_{n \rightarrow \infty} \left| \frac{a_{n+1}}{a_n} \right| ..$$

Then:

- If  $0 \leq \rho < 1$ , then  $\sum_{n=1}^{\infty} a_n$  converges absolutely.
- If  $\rho > 1$  or  $\rho = \infty$ , then  $\sum_{n=1}^{\infty} a_n$  diverges.
- If  $\rho = 1$ , the test does not provide any information.

**Note:** The ratio test is useful for series whose terms involve factorials

- **Root test:** Consider the series  $\sum_{n=1}^{\infty} a_n$ . Let

$$\rho = \lim_{n \rightarrow \infty} \sqrt[n]{|a_n|} ..$$

- If  $0 \leq \rho < 1$ , then  $\sum_{n=1}^{\infty} a_n$  converges absolutely.
- If  $\rho > 1$  or  $\rho = \infty$ , then  $\sum_{n=1}^{\infty} a_n$  diverges.
- If  $\rho = 1$ , the test does not provide any information.

**Note:** The root test is useful for series whose terms involve exponentials

- **Which tests require positive terms:**

- **Integral Test:** This test applies to series where the terms come from a function that is positive, continuous, and decreasing on a certain interval. The convergence or divergence of the series is determined by the convergence or divergence of the corresponding improper integral of the function.
- **Remainder estimate for the integral test:**
- **Comparison Test:** This test compares the terms of a series to those of another series with known convergence behavior. It requires that the terms of both series be positive or non-negative.
- **Limit Comparison Test:** Similar to the Comparison Test, this test involves comparing the terms of two series by taking the limit of the ratio of their terms. It requires that the terms of both series be positive.
- **alternating series,  $b_n$ :** must have only positive terms

## 1.5 Chapter 6 Definitions and Theorems

- Euler definition for  $e$ :

$$\begin{aligned} e^a &= \lim_{n \rightarrow \infty} \left(1 + \frac{a}{n}\right)^n \\ \frac{1}{e^a} &= \lim_{n \rightarrow \infty} \left(1 + \frac{-a}{n}\right)^n \\ \frac{1}{e^a} &= \lim_{n \rightarrow \infty} \left(\frac{n}{n+a}\right)^n. \end{aligned}$$

- Other definition for  $e$ :

$$\begin{aligned} e &= \sum_{n=0}^{\infty} \frac{1}{n!} \\ e - 1 &= \sum_{n=1}^{\infty} \frac{1}{n!} \\ e^x &= \sum_{n=0}^{\infty} \frac{x^n}{n!}. \end{aligned}$$

- Power series: A series of the form

$$\sum_{n=0}^{\infty} c_n x^n = c_0 + c_1 x + c_2 x^2 + \dots.$$

is a power series centered at  $x = 0$ .

A series of the form

$$\sum_{n=0}^{\infty} c_n (x-a)^n = c_0 + c_1(x-a) + c_2(x-a)^2 + \dots.$$

is a power series centered at  $x = a$ .

- Convergence of a Power Series:

Consider the power series  $\sum_{n=0}^{\infty} c_n (x-a)^n$ . The series satisfies exactly one of the following properties:

- (i) The series converges at  $x = a$  and diverges for all  $x \neq a$ .
- (ii) The series converges for all real numbers  $x$ .
- (iii) There exists a real number  $R > 0$  such that the series converges if  $|x-a| < R$  and diverges if  $|x-a| > R$ . At the values  $x$  where  $|x-a| = R$ , the series may converge or diverge.

- A power series always converges at its center:

- Radius of convergence: Consider the power series  $\sum_{n=0}^{\infty} c_n (x-a)^n$ . The set of real numbers  $x$  where the series converges is the interval of convergence. If there exists a real number  $R > 0$  such that the series converges for  $|x-a| < R$  and diverges for  $|x-a| > R$ , then  $R$  is the radius of convergence. If the series converges only at  $x = a$ , we say the radius of convergence is  $R = 0$ . If the series converges for all real numbers  $x$ , we say the radius of convergence is  $R = \infty$

- **Finding interval of convergence and radius of convergence:**
  - Fact: power series is always convergent on its center
  - Use ratio test (values of  $\rho$ )
  - Use  $\rho < 1$  to find Radius of convergence
  - Test end points of interval by plugging into original series and seeing whether the series is convergent or divergent
- **If  $\rho = 0$ , the power series converges for all  $x$ :**
- **If  $\rho = \infty$ :** the series diverges for all  $x \neq a$
- **Combining Power Series:** Suppose that the two power series  $\sum_{n=0}^{\infty} c_n x^n$  and  $\sum_{n=0}^{\infty} d_n x^n$  converge to the functions  $f$  and  $g$ , respectively, on a common interval  $I$ .
  - (i) The power series  $\sum_{n=0}^{\infty} (c_n x^n \pm d_n x^n)$  converges to  $f \pm g$  on  $I$ .
  - (ii) For any integer  $m \geq 0$  and any real number  $b$ , the power series  $\sum_{n=0}^{\infty} b x^m c_n x^n$  converges to  $b x^m f(x)$  on  $I$ .

Eg: If we know  $\sum_{n=0}^{\infty} a_n x^n$  has  $I = (-1, 1)$ . Then

$$\begin{aligned} & \sum_{n=0}^{\infty} a_n 3^n x^n \\ &= \sum_{n=0}^{\infty} a_n (3x)^n \\ & I = (-3, 3). \end{aligned}$$

- (iii) For any integer  $m \geq 0$  and any real number  $b$ , the series  $\sum_{n=0}^{\infty} c_n (bx^m)^n$  converges to  $f(bx^m)$  for all  $x$  such that  $bx^m$  is in  $I$ .

- **For part I, II, and III, the interval of the combined series is the smaller interval:**
- **Cauchy product (Multiplying power series):** Suppose that the power series  $\sum_{n=0}^{\infty} c_n x^n$  and  $\sum_{n=0}^{\infty} d_n x^n$  converge to  $f$  and  $g$ , respectively, on a common interval  $I$ . Let

$$\begin{aligned} e_n &= c_0 d_n + c_1 d_{n-1} + c_2 d_{n-2} + \cdots + c_{n-1} d_1 + c_n d_0 \\ &= \sum_{k=0}^n c_k d_{n-k} \end{aligned}$$

Then

$$\left( \sum_{n=0}^{\infty} c_n x^n \right) \left( \sum_{n=0}^{\infty} d_n x^n \right) = \sum_{n=0}^{\infty} e_n x^n$$

and

$$\sum_{n=0}^{\infty} e_n x^n \text{ converges to } f(x) \cdot g(x) \text{ on } I.$$

The series  $\sum_{n=0}^{\infty} e_n x^n$  is known as the Cauchy product of the series  $\sum_{n=0}^{\infty} c_n x^n$  and  $\sum_{n=0}^{\infty} d_n x^n$ .

- **Sterling's Approximation:**

$$n! \approx \sqrt{2\pi n} \left( \frac{n}{e} \right)^n.$$

- **Gamma function (extension of the factorial function):**

$$\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} dt$$

Thus,  $n! = \Gamma(n+1)$ .

- **Cool definition for  $e^x$ :**

$$\begin{aligned} f'(x) &= rf(x) \\ \implies f(x) &= ce^{rx}. \end{aligned}$$

- **Term-by-Term Differentiation and Integration for Power Series.:** Suppose that the power series  $\sum_{n=0}^\infty c_n(x-a)^n$  converges on the interval  $(a-R, a+R)$  for some  $R > 0$ . Let  $f$  be the function defined by the series

$$f(x) = \sum_{n=0}^\infty c_n(x-a)^n = c_0 + c_1(x-a) + c_2(x-a)^2 + c_3(x-a)^3 + \dots$$

for  $|x-a| < R$ . Then  $f$  is differentiable on the interval  $(a-R, a+R)$  and we can find  $f'$  by differentiating the series term-by-term:

$$f'(x) = \sum_{n=1}^\infty nc_n(x-a)^{n-1} = c_1 + 2c_2(x-a) + 3c_3(x-a)^2 + \dots$$

for  $|x-a| < R$ . Also, to find  $\int f(x) dx$ , we can integrate the series term-by-term. The resulting series converges on  $(a-R, a+R)$ , and we have

$$\int f(x) dx = C + \sum_{n=0}^\infty \frac{c_n(x-a)^{n+1}}{n+1} = C + c_0(x-a) + \frac{c_1(x-a)^2}{2} + \frac{c_2(x-a)^3}{3} + \dots$$

for  $|x-a| < R$ .

**NOTE!** when a power series is differentiated or integrated term-by-term, it says nothing about what happens at the endpoints.

- **Uniqueness of Power Series:** Let  $\sum_{n=0}^\infty c_n(x-a)^n$  and  $\sum_{n=0}^\infty d_n(x-a)^n$  be two convergent power series such that

$$\sum_{n=0}^\infty c_n(x-a)^n = \sum_{n=0}^\infty d_n(x-a)^n$$

for all  $x$  in an open interval containing  $a$ . Then  $c_n = d_n$  for all  $n \geq 0$ .

- **When finding the Cauchy product of two power series, we include the zero term when finding the new power series general term. For integrating and differentiating, we do not:**
- **Taylor and Maclaurin series:** If  $f$  has derivatives of all orders at  $x = a$ , then the Taylor series for the function  $f$  at  $a$  is

$$\sum_{n=0}^\infty \frac{f^{(n)}(a)}{n!}(x-a)^n = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \dots + \frac{f^{(n)}(a)}{n!}(x-a)^n + \dots \quad (16)$$

The Taylor series for  $f$  at 0 is known as the Maclaurin series for  $f$ .

- **Uniqueness of Taylor series:** If a function  $f$  has a power series at  $a$  that converges to  $f$  on some open interval containing  $a$ , then that power series is the Taylor series for  $f$  at  $a$ .
- **Taylor-Macluarin Polynomials:** If  $f$  has  $n$  derivatives at  $x = a$ , then the  $n$ th Taylor polynomial for  $f$  at  $a$  is

$$p_n(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n. \quad (17)$$

The  $n$ th Taylor polynomial for  $f$  at 0 is known as the  $n$ th Maclaurin polynomial for  $f$ .

- **Taylor's Theorem with Remainder:** Let  $f$  be a function that can be differentiated  $n+1$  times on an interval  $I$  containing the real number  $a$ . Let  $p_n$  be the  $n$ th Taylor polynomial of  $f$  at  $a$  and let

$$R_n(x) = f(x) - p_n(x)$$

be the  $n$ th remainder. Then for each  $x$  in the interval  $I$ , there exists a real number  $c$  between  $a$  and  $x$  such that

$$R_n(x) = \frac{f^{(n+1)}(c)}{(n+1)!}(x-a)^{n+1}.$$

If there exists a real number  $M$  such that  $|f^{(n+1)}(x)| \leq M$  for all  $x \in I$ , then

$$|R_n(x)| \leq \frac{M}{(n+1)!}|x-a|^{n+1}$$

$$\forall x \in I$$

- **Maclaurin Series/Polynomials for sine:** The Taylor series for the sine function is

$$\sin(x) = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots \quad \text{For } x \in \mathbb{R}.$$

Where  $p_n$  obeys

$$\begin{aligned} p_{2m+1} &= p_{2m+2} \\ &= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots + \frac{(-1)^m x^{2m+1}}{(2m+1)!}. \end{aligned}$$

**Note:** When discussing specific polynomials, say  $P_5$  for example, we arnt talking about the first 5 terms in the series above, we are talking about the polynomial up to degree 5. Thus it would have 3 terms

- **Maclaurin Series/Polynomials for cosine:** Similar to the sine function, the Maclaurin series for the cosine function is

$$\cos(x) = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n)!} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \dots \quad \text{For } x \in \mathbb{R}.$$

Where  $p_n$  obeys

$$\begin{aligned} p_{2m} &= p_{2m+1} \\ &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \dots + (-1)^n \frac{x^{2m}}{(2m)!}. \end{aligned}$$

- **Maclaurin Series/Polynomials for  $e^x$ :** We find the Maclaurin series for the exponential function to be

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots \quad \text{For } x \in \mathbb{R}.$$

**Note:** this definition is described above but now we have a way of showing its truthiness

- **Convergence of Taylor Series:** Suppose that  $f$  has derivatives of all orders on an interval  $I$  containing  $a$ . Then the Taylor series

$$\sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!} (x-a)^n$$

converges to  $f(x)$  for all  $x$  in  $I$  if and only if

$$\lim_{n \rightarrow \infty} R_n(x) = 0$$

for all  $x$  in  $I$ .

**Note:** With this theorem, we can prove that a Taylor series for  $f$  at  $a$  converges to  $f$  if we can prove that the remainder  $R_n(x) \rightarrow 0$ . To prove that  $R_n(x) \rightarrow 0$ , we typically use the bound

$$|R_n(x)| \leq \frac{M}{(n+1)!} |x-a|^{n+1}$$

from Taylor's theorem with remainder.

- **Using taylor series to find limits:** Consider the limit  $\lim_{x \rightarrow 0^+} \frac{\cos \sqrt{x} - 1}{2x}$ . We know we have a problem if we attempt to use the **direct substitution property**. Thus, we can substitute  $\cos(\sqrt{x})$  for its **Maclaurin series** and see what happens. We know the maclaurin series for  $\cos(x)$  : is

$$\begin{aligned} \cos x &\sim \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n)!} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \dots \\ \Rightarrow \cos \sqrt{x} &\sim \sum_{n=0}^{\infty} (-1)^n \frac{(\sqrt{x})^{2n}}{(2n)!} = 1 - \frac{(\sqrt{x})^2}{2!} + \frac{(\sqrt{x})^4}{4!} - \frac{(\sqrt{x})^6}{6!} + \dots \\ &= \sum_{n=0}^{\infty} (-1)^n \frac{x^n}{(2n)!} = 1 - \frac{x}{2!} + \frac{x^2}{4!} - \frac{x^3}{6!} + \dots \end{aligned}$$

So we have

$$\begin{aligned} &\lim_{x \rightarrow 0^+} \frac{\left(1 - \frac{x}{2!} + \frac{x^2}{4!} - \frac{x^3}{6!} + \dots\right) - 1}{2x} \\ &= \lim_{x \rightarrow 0^+} \frac{\left(-\frac{x}{2!} + \frac{x^2}{4!} - \frac{x^3}{6!} + \dots\right)}{2x} \\ &= \lim_{x \rightarrow 0^+} \left(-\frac{x}{2!} + \frac{x^2}{4!} - \frac{x^3}{6!} + \dots\right) \cdot \frac{1}{2x} \\ &= \lim_{x \rightarrow 0^+} -\frac{1}{4} \\ &= -\frac{1}{4}. \end{aligned}$$

- **Multiplying a known Taylor series by some other function:** Consider  $f(x) = x \cos x$ . Since we know that the taylor series for  $\cos(x) = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n)!}$ , which converges  $\forall x \in \mathbb{R}$ . We can easily just multiply this by  $x$  to get the taylor series for  $f(x) = x \cos(x)$ . Since the Taylor series for  $\cos(x)$  : converges for all real  $x$ , multiplying it by  $x$  won't affect its convergence properties. The resulting series will still converge for all  $x$ .

**Note:** The product of the Taylor series and the function will be valid only where both the series converges and the function is well-defined. Probably analyze the convergence of the product series.

- **Multiplying a known Taylor series by some other function where convergence is affected:** Consider the example above. Although this time suppose we multiply  $\cos(x)$  by  $\frac{1}{x}$  instead of  $x$ . We know the resulting taylor series must not be convergent at  $x = 0$  because  $\frac{1}{x}$  is not defined at zero. It is important to understand that we will not get this conclusion from the ratio test alone. The Ratio Test alone does not account for points where the series or its terms are not defined.

**TLDR:** Be mindful about the domain of the function you are multiplying the Taylor series by. Do not only rely on the ratio test to find points of convergence.

- **Analytic function:**

- An analytic function is infinitely differentiable within its domain.
- An analytic function can be represented by a convergent power series (like a Taylor series) around any point in its domain.
- The power series representing an analytic function not only exists but also converges to the function within a certain radius around the point of expansion

- **Maclaurin series for  $\frac{1}{1-x}$ :**

$$\frac{1}{1-x} \sim \sum_{n=0}^{\infty} x^n \quad \text{for } |x| < 1.$$

- **Maclaurin series for  $\ln(1+x)$ :**

$$\ln(1+x) \sim \sum_{n=1}^{\infty} (-1)^{n+1} \frac{x^n}{n} \quad \text{for } |x| < 1.$$

- **Maclaurin series for  $\tan^{-1} x$ :**

$$\tan^{-1} x \sim \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{2n+1} \quad \text{for } |x| \leq 1.$$

- **Binomial expansion for  $(1+x)^r$  for  $r \in \mathbb{Z}^+$ :**

$$(1+x)^r = \sum_{n=0}^r \binom{r}{n} x^n \quad r \in \mathbb{Z}^+.$$

- **Binomial expansion for  $(1+x)^r$  for  $r \in \mathbb{R}$ :**

$$(1+x)^r = \sum_{n=0}^{\infty} \binom{r}{n} x^n = 1 + rx + \frac{r(r-1)}{2!} x^2 + \dots + \frac{r(r-1)\cdots(r-n+1)}{n!} x^n + \dots$$

Where

$$\binom{r}{n} = \frac{f^{(n)}(0)}{n!} = \frac{r(r-1)(r-2)\cdots(r-n+1)}{n!}.$$

**Note:** When  $n = 0$ ,  $\binom{r}{0} = 1$ , when  $n = 1$ ,  $\binom{r}{1} = r$ , when  $n = 2$ ,  $\binom{r}{2} = \frac{r(r-1)}{2!}$ , etc...

- **The binomial theorem:** For any real number  $r$ , the Maclaurin series for  $f(x) = (1+x)^r$  is the binomial series. It converges to  $f$  for  $|x| < 1$ , and we write

$$(1+x)^r = \sum_{n=0}^{\infty} \binom{r}{n} x^n = 1 + rx + \frac{r(r-1)}{2!} x^2 + \cdots + \frac{r(r-1)\cdots(r-n+1)}{n!} x^n + \cdots$$

for  $|x| < 1$ .

## 1.6 Chapter 6 Problems to Remember

- **Problem to remember (Properties of power series):** Evaluate the infinite series by identifying it as the value of an integral of a geometric series.

$$\sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1}.$$

**Remark.** If we can find which geometric power series's integral (with some bounds) gives us the given series, we can then integrate the function representation to get the value of the original series. Consider the geometric power series

$$\frac{1}{1+x^2} = \frac{1}{1-(-x^2)} = \sum_{n=0}^{\infty} (-x^2)^n = \sum_{n=0}^{\infty} (-1)^n x^{2n}.$$

Suppose we then integrate the power series

$$\begin{aligned} & \int \sum_{n=0}^{\infty} (-1)^n x^{2n} dx \\ &= \sum_{n=0}^{\infty} (-1)^n \int x^{2n} dx \\ &= \frac{1}{2n+1} x^{2n+1}. \end{aligned}$$

Now we must deduce for which bounds will the FTC give us the original series  $\sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1}$ . We come to the conclusion

$$\sum_{n=0}^{\infty} (-1)^n \int_0^1 x^{2n} dx = \sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1}.$$

This implies we can integrate the function representation of the geometric power series we just integrated to get the value of the infinite series  $\sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1}$ . Thus,

$$\int_0^1 \frac{1}{1+x^2} dx = \frac{\pi}{4}.$$

- **Problem to remember (Properties of power series):** Find the power series for  $f(x) = \ln x$  : centered at  $x = 9$  by using term-by-term integration or differentiation.

*Solution.* The goal is to find a function that resembles one we know (sum of geometric series  $\frac{a}{1-x}$ ) such that if we integrate or differentiate we can get  $\ln x$ . Since we know the integral of  $\frac{1}{x}$  is  $\ln x$ , and we can easily manipulate  $\frac{1}{x}$  to be in the form  $\frac{a}{1-x}$ , we choose  $\frac{1}{x}$  to be the function to examine. Thus,

$$\frac{1}{x} = \frac{1}{9+x-9} = \frac{1}{9-(-(x-9))} = \frac{1/9}{1-\left(\frac{-(x-9)}{9}\right)}$$

$$\text{If } f(x) = \frac{a}{1-x} \sim \sum_{n=0}^{\infty} a(x^n) = a + ax + ax^2 + ax^3 + \dots$$

$$\Rightarrow f(x) = \frac{1/9}{1-\left(\frac{-(x-9)}{9}\right)} \sim \sum_{n=0}^{\infty} \frac{1}{9} \left(\frac{-(x-9)}{9}\right)^n = \sum_{n=0}^{\infty} \frac{(-1)^n (x-9)^n}{9^{n+1}}.$$

Then we can throw in some integrals

$$\begin{aligned} \int f(x) dx &= \int \frac{1/9}{1 - \left(\frac{-(x-9)}{9}\right)} dx = \int \sum_{n=0}^{\infty} \frac{(-1)^n (x-9)^n}{9^{n+1}} dx \\ \ln x &= \sum_{n=0}^{\infty} \frac{(-1)^n}{9^{n+1}} \int (x-9)^n dx \\ \ln x &= \sum_{n=0}^{\infty} \frac{(-1)^n (x-9)^{n+1}}{(n+1)9^{n+1}} + C. \end{aligned}$$

If we let  $x = 9$

$$\begin{aligned} \ln 9 &= \sum_{n=0}^{\infty} \frac{(-1)^n (9-9)^{n+1}}{(n+1)9^{n+1}} + C \\ \ln 9 &= C. \end{aligned}$$

Thus, we have

$$f(x) = \ln 9 + \sum_{n=0}^{\infty} \frac{(-1)^n (9-9)^{n+1}}{(n+1)9^{n+1}}$$

**Note:** the "+C" is initially omitted from  $\ln(x)$  because we're considering a specific antiderivative. When you integrate the power series, you include "+C" to account for the general form of the antiderivative. The value of  $C$  is then determined using a specific condition to match the specific antiderivative you're interested in.

- **Problem to remember:** Say we want to find the power series for  $7x \ln 1+x$ . We can first find the power series for  $\ln(1+x)$ :

$$\frac{d}{dx} \ln(1+x) = \frac{1}{1+x} = \frac{1}{1-(-x)}.$$

We know the power series for  $\frac{1}{1-(-x)}$  is

$$\begin{aligned} &\sum_{n=0}^{\infty} (-1)^n x^n \\ \implies &\int \frac{1}{1+x} = \int \sum_{n=0}^{\infty} (-1)^n x^n \\ \ln(1+x) &= \sum_{n=0}^{\infty} (-1)^n \frac{x^{n+1}}{n+1} + C \\ \ln(1+x) &= \sum_{n=0}^{\infty} (-1)^n \frac{x^{n+1}}{n+1}. \end{aligned}$$

Now that we have found the power series for  $\ln(1+x)$ , to find the power series for  $7x \ln(1+x)$ ...

$$\begin{aligned} &7x \sum_{n=0}^{\infty} (-1)^n \frac{x^{n+1}}{n+1} \\ &\sum_{n=0}^{\infty} 7x \left( (-1)^n \frac{x^{n+1}}{n+1} \right) \\ &= \sum_{n=0}^{\infty} (-1)^n \frac{7x^{n+2}}{n+1}. \end{aligned}$$

- **Problem to remember (Cumbersome taylor polynomial):** Suppose we have some function  $f$ , and we would like to find the Taylor polynomial up to degree 3. Say  $f(x) = e^{2x} \cos(x)$ . We could find each derivative up to degree 3, however, given that we know the taylor series for both  $e^{2x}$  and  $\cos(x)$  :

$$e^{2x} = \sum_{n=0}^{\infty} \frac{(2x)^n}{n!} = 1 + 2x + \frac{4x^2}{2!} + \frac{8x^3}{3!} + \dots$$

$$\cos(x) = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n)!} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \dots .$$

We can find  $P_3(x)$  by multiplying these Taylor series. Thus,

$$(1 + 2x + \frac{4x^2}{2!} + \frac{8x^3}{3!} + \dots)(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \dots).$$

And we can find  $P_3$  to be

$$P_3 = 1 + 2x + \frac{3}{2}x^2 + \frac{1}{3}x^3.$$

- **Problem to Remember (Using known Taylor series to find sum of series):**  
Consider the series  $\sum_{n=0}^{\infty} (-1)^n \frac{(\frac{1}{25})^{n-3}}{2n+1}$  :

We notice this resembles the Taylor series for the arctangent function  $\tan^{-1} x = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{2n+1}$  for  $|x| \leq 1$ . Thus, we manipulate the series to better conform to the Taylor series for  $\tan^{-1} x$ .

$$\begin{aligned} & \sum_{n=0}^{\infty} (-1)^n \frac{(\frac{1}{25})^{n-3}}{2n+1} \\ &= \sum_{n=0}^{\infty} (-1)^n \frac{(\frac{1}{5})^{2n-6}}{2n+1} \\ &= \sum_{n=0}^{\infty} (-1)^n \frac{(\frac{1}{5})^{2n} 5^6}{2n+1} \\ &= \sum_{n=0}^{\infty} (-1)^n \frac{(\frac{1}{5})^{2n} 5^6 (\frac{1}{5}) 5}{2n+1} \\ &= \sum_{n=0}^{\infty} (-1)^n \frac{(\frac{1}{5})^{2n+1} 5^7}{2n+1} \end{aligned}$$

Thus the sum will be

$$5^7 \tan^{-1} \frac{1}{5}.$$

# Fundamental Physics: Classical Mechanics

## 2.1 Chapter 1: Units and Measurement

### 2.1.1 Key Terms

- **Physics**, which comes from the Greek *phύσις*: meaning “nature,” is concerned with describing the interactions of energy, matter, space, and time to uncover the fundamental mechanisms that underlie every phenomenon.
- **model**: is a representation of something that is often too difficult (or impossible) to display directly. Although
- **theory**: is a testable explanation for patterns in nature supported by scientific evidence and verified multiple times by various groups of researchers.
- **law**: uses concise language to describe a generalized pattern in nature supported by scientific evidence and repeated experiments. Often, a law can be expressed in the form of a single mathematical equation.
- **metric system**:
- **centimeter–gram–second (cgs)**: system.
- **English units**: (also known as the customary or imperial system). English units were historically used in nations once ruled by the British Empire and are still widely used in the United States.
- **the foot–pound–second (fps)**: system.
- **base quantities**: for that system
- **base units**:
- **derived unit**:
- **conversion factor**: is a ratio that expresses how many of one unit are equal to another unit.
- **dimension**: of any physical quantity expresses its dependence on the base quantities as a product of symbols (or powers of symbols) representing the base quantities.
- **dimensionless**: (or sometimes “of dimension 1,” because anything raised to the zero power is one).
- **pure numbers**:
- **dimensionally consistent**:
  - Every term in an expression must have the same dimensions; it does not make sense to add or subtract quantities of differing dimension (think of the old saying: “You can’t add apples and oranges”). In particular, the expressions on each side of the equality in an equation must have the same dimensions.
  - The arguments of any of the standard mathematical functions such as trigonometric functions (such as sine and cosine), logarithms, or exponential functions that appear in the equation must be dimensionless. These functions require pure numbers as inputs and give pure numbers as outputs.

- **Estimation:** means using prior experience and sound physical reasoning to arrive at a rough idea of a quantity's value.
- **Accuracy:** is how close a measurement is to the accepted reference value for that measurement.
- **precision of measurements:** refers to how close the agreement is between repeated independent measurements (which are repeated under the same conditions).
- **discrepancy:** from the accepted reference value.
- **Discrepancy:** (or “measurement error”) is the difference between the measured value and a given standard or expected value.
- Another method of expressing uncertainty is as a percent of the measured value.

### 2.1.2 Definitions and Theorems (and important things)

- Order of Magnitude (Method one):

$$m = \log_{10} l, \quad l \in \mathbb{R}.$$

For example

$$\log 450 \approx 3.$$

Thus the order of magnitude is  $10^3$

**Note:** Always round s.t  $m \in \mathbb{R}$

- Order of Magnitude (Method two): We begin by writing the number in scientific notation. For example, suppose we want to find the order of magnitude for 800, we first write

$$8 \cdot 10^2.$$

We then check to see if the first factor is greater or less than  $\sqrt{10} \approx 3$

- If it is less than  $\sqrt{10} \approx 3$ , we round it down to one
- If it is greater than  $\sqrt{10} \approx 3$ , we round it up to ten

Since 8 is greater than  $\sqrt{10} \approx 3$ , our number becomes  $10 \cdot 10^2 = 10^{2+1} = 10^3$ . Thus, we have order  $10^3$

- second:
- meter:
- kilogram:
- Newton:
- $10^3$  (Megagram) is also called a *metric ton*, abbreviated *t*:
- Average speed:

$$\frac{\text{distance}}{\text{time}} = \frac{d}{t}.$$

- Dimension notation: Square brackets

$$[r] = L.$$

- Checking for dimensional consistency: check that each term in a given equation has the same dimensions as the other terms in that equation and that the arguments of any standard mathematical functions are dimensionless.

- Dimension for velocity:

$$[v] = LT^{-1}.$$

- Dimension for acceleration:

$$[a] = LT^{-2}.$$

- **Dimensions for derivatives:**

$$\left[ \frac{dv}{dt} \right] = \frac{[v]}{[t]}.$$

- **Dimensions for integrals:**

$$\left[ \int v dt \right] = [v] \cdot [t]..$$

- **Volume:**

$$V = AD.$$

Where  $A$  is the area and  $D$  is the depth

- **Mass:**

$$M = \rho V.$$

Where  $\rho$  is the density and  $V$  is the volume

- **Density (Volume density):**

$$\rho = \frac{M}{V}.$$

Where  $M$  is the mass and  $V$  is the volume

- **Density (Area density):**

$$\rho = \frac{M}{A}.$$

Where  $M$  is the mass and  $A$  is the surface area

- **Geometric mean of bounds:**

$$(order_1 \times order_2)^{0.5}.$$

- **Significant figures:**

– All non-zero numbers are Significant.

$$563 \rightarrow 3.$$

– Zeros are significant if they reside between two significant figures.

$$5002 \rightarrow 4.$$

– Leading zeros are never significant.

$$0.0056 \rightarrow 2.$$

– Trailing zeros without a decimal point are not significant

$$500 \rightarrow 1.$$

– Trailing zeros with a decimal point are significant. The decimal point indicates that the zeros are measured and are significant.

$$500.0 \rightarrow 4.$$

- **Percent uncertainty:** If a measurement  $A$  is expressed with uncertainty  $\delta A$ , the percent uncertainty is defined as

$$\text{Percent uncertainty} = \frac{\delta A}{A} \times 100\%.$$

where  $A$  is the average, and  $\delta A$  is the margin of error

**Note:** The value of the PU will take the place of the margin of error, so something like  $5.1 \text{ lbs} \pm 0.3 \text{ lbs}$  will become  $5.1 \pm 6\%$

- We can find the margin of error ( $\delta A$ ) by taking half of the range:
- If the measurements going into the calculation have small uncertainties (a few percent or less), then the method of adding percents can be used for multiplication or division. This method states the percent uncertainty in a quantity calculated by multiplication or division is the sum of the percent uncertainties in the items used to make the calculation. For example, if a floor has a length of  $4.00 \text{ m}$  and a width of  $3.00 \text{ m}$ , with uncertainties of 2% and 1%, respectively, then the area of the floor is  $12.0 \text{ m}^2$  and has an uncertainty of 3%. (Expressed as an area, this is  $0.36 \text{ m}^2$  [ $12.0 \text{ m}^2 \times 0.03$ ], which we round to  $0.4 \text{ m}^2$  since the area of the floor is given to a tenth of a square meter.):
- When combining measurements with different degrees of precision with the mathematical operations of addition, subtraction, multiplication, or division, then the number of significant digits in the final answer can be no greater than the number of significant digits in the least-precise measured value. There are two different rules, one for multiplication and division and the other for addition and subtraction. There are two different rules:
  - For multiplication and division:, the result should have the same number of significant figures as the quantity with the least number of significant figures entering into the calculation
  - For addition and subtraction:, the answer can contain no more decimal places than the least-precise measurement.
- If a quantity increases  $n\%$ , that is the same as saying that it is multiplied by a factor of:

$$1 + \left( \frac{n}{100} \right).$$

- If a quantity decreases  $n\%$ , that is the same as saying that it is multiplied by a factor of:

$$1 - \left( \frac{n}{100} \right).$$

- **Proportional notation:** Suppose  $A$  is proportional to  $B$ , then we say

$$A \propto B.$$

This means if  $B$  increases by some factor, then  $A$  must increase by the same factor.

In other words, the ratio of two values of  $B$  is equal to the ratio of the corresponding values of  $A$ :

$$\frac{B_2}{B_1} = \frac{A_2}{A_1}.$$

Ex: Given the circumference formula for a circle

$$c = 2\pi r.$$

We can say

$$C \propto r.$$

If the radius doubles, the circumference also doubles

- **Numbers that are exact (defined) have an infinite number of significant figures:** because they are not measurements with any uncertainty, but rather are defined values or counts of discrete objects. For example, considering a conversion of 93.4 beats/min to beats/hour, we would compute

$$\frac{93.4 \text{ beats}}{1 \text{ m}} \cdot \frac{60 \text{ m}}{1 \text{ h}} \\ = 5604 \text{ b/h.}$$

Since the number 60 is an exact number (it is a defined conversion factor), we should report the answer with three significant figures (since 93.4 has three). Thus, we round to 5600. To explicitly express 5600 with three sig figs, we write as  $5.60 \cdot 10^3$

- **Explicitly express sig figs with scientific notation:**

$$5600 \rightarrow 2 \\ 5.60 \cdot 10^3 \rightarrow 3.$$

- **Forms of scientific notation:**

$$10^{10} = 1.0 \cdot 10^{10} = 1.0e + 10.$$

- **Decimal places in addition and subtraction:** consider the example

$$501.258313 + 54.5235 + 350.257 = 906.038813.$$

We would report the value as

$$906.039.$$

Because the least precise measurement has three decimal points

- **you can err on the side of including extra ones:**
- **Weight is the measure of the force exerted on an object due to gravity. It is calculated as the mass of the object multiplied by the acceleration due to gravity.:**
- **Newton:**

$$N = kg \cdot \frac{m}{s^2}.$$

### 2.1.3 Fundamental Tables and figures

- Known ranges of length, mass, and time:

Length in Meters (m)	Masses in Kilograms (kg)	Time in Seconds (s)
$10^{-15} \text{ m} = \text{diameter of proton}$	$10^{-30} \text{ kg} = \text{mass of electron}$	$10^{-22} \text{ s} = \text{mean lifetime of very unstable nucleus}$
$10^{-14} \text{ m} = \text{diameter of large nucleus}$	$10^{-27} \text{ kg} = \text{mass of proton}$	$10^{-17} \text{ s} = \text{time for single floating-point operation in a supercomputer}$
$10^{-10} \text{ m} = \text{diameter of hydrogen atom}$	$10^{-15} \text{ kg} = \text{mass of bacterium}$	$10^{-15} \text{ s} = \text{time for one oscillation of visible light}$
$10^{-7} \text{ m} = \text{diameter of typical virus}$	$10^{-5} \text{ kg} = \text{mass of mosquito}$	$10^{-13} \text{ s} = \text{time for one vibration of an atom in a solid}$
$10^{-2} \text{ m} = \text{pinky fingernail width}$	$10^{-2} \text{ kg} = \text{mass of hummingbird}$	$10^{-3} \text{ s} = \text{duration of a nerve impulse}$
$10^0 \text{ m} = \text{height of 4 year old child}$ 	$10^0 \text{ kg} = \text{mass of liter of water}$ 	$10^0 \text{ s} = \text{time for one heartbeat}$ 
$10^2 \text{ m} = \text{length of football field}$	$10^2 \text{ kg} = \text{mass of person}$	$10^5 \text{ s} = \text{one day}$
$10^7 \text{ m} = \text{diameter of Earth}$	$10^{19} \text{ kg} = \text{mass of atmosphere}$	$10^7 \text{ s} = \text{one year}$
$10^{13} \text{ m} = \text{diameter of solar system}$	$10^{22} \text{ kg} = \text{mass of Moon}$	$10^9 \text{ s} = \text{human lifetime}$
$10^{16} \text{ m} = \text{distance light travels in a year (one light-year)}$	$10^{25} \text{ kg} = \text{mass of Earth}$	$10^{11} \text{ s} = \text{recorded human history}$
$10^{21} \text{ m} = \text{Milky Way diameter}$	$10^{30} \text{ kg} = \text{mass of Sun}$	$10^{17} \text{ s} = \text{age of Earth}$
$10^{26} \text{ m} = \text{distance to edge of observable universe}$	$10^{53} \text{ kg} = \text{upper limit on mass of known universe}$	$10^{18} \text{ s} = \text{age of the universe}$

- SI Units: Base and Derived Units:

ISQ Base Quantity	SI Base Unit
Length	meter (m)
Mass	kilogram (kg)
Time	second (s)
Electrical current	ampere (A)
Thermodynamic temperature	kelvin (K)
Amount of substance	mole (mol)
Luminous intensity	candela (cd)

- **Metric Prefixes:**

Prefix	Symbol	Meaning	Prefix	Symbol	Meaning
yotta-	Y	$10^{24}$	yocto-	y	$10^{-24}$
zetta-	Z	$10^{21}$	zepto-	z	$10^{-21}$
exa-	E	$10^{18}$	atto-	a	$10^{-18}$
peta-	P	$10^{15}$	femto-	f	$10^{-15}$
tera-	T	$10^{12}$	pico-	p	$10^{-12}$
giga-	G	$10^9$	nano-	n	$10^{-9}$
mega-	M	$10^6$	micro-	$\mu$	$10^{-6}$
kilo-	k	$10^3$	milli-	m	$10^{-3}$
hecto-	h	$10^2$	centi-	c	$10^{-2}$
deka-	da	$10^1$	deci-	d	$10^{-1}$

- **Base Quantity & Symbol for Dimension:**

Base Quantity	Symbol for Dimension
Length	L
Mass	M
Time	T
Current	I
Thermodynamic temperature	$\Theta$
Amount of substance	N
Luminous intensity	J

**2.1.4 Memorize conversions**

- **Newton to Pounds:**

$$1N = 0.225lbs.$$

- **Pounds to newtons:**

$$1lb = 4.448 N.$$

- **Length/Distance :**

- 1 inch (in) = 2.54 centimeters (cm)
- 1 centimeter (cm) = 0.393701 inches (in)
- 1 foot (ft) = 0.3048 meters (m)
- 1 meter (m) = 3.28 feet (ft)
- 1 mile (mi) = 1.6 kilometers (km)
- 1 kilometer (km) = 0.621371 miles (mi)

- **Weight to mass or mass to weight:**

- 1 pound = 0.453592 kilograms.
- 1 kilogram = 2.2 pounds.

### 2.1.5 Problems to remember

- **Restating mass:** Restate the mass  $1.93 \times 10^{13}$  : using a metric prefix such that the resulting numerical value is bigger than one but less than 1000.

First, we must restate in terms of grams. Since  $1\text{kg} = 10^3\text{g}$ , we write

$$\begin{aligned} 1.93 \times 10^{13} \times 10^3\text{g} \\ 1.93 \times 10^{16}\text{g}. \end{aligned}$$

Since  $1\text{Pg} = 10^{15}\text{g}$ , we can write

$$1.93 \times 10^1\text{Pg}.$$

Since  $16 - 15 = 1$

- **Unit conversion:** The distance from the university to home is 10 mi and it usually takes 20 min to drive this distance. Calculate the average speed in meters per second (m/s). (Note: Average speed is distance traveled divided by time of travel.)

**Note:** There are 1609 meters in 1 mile

First, we can compute the average speed with the units given ( $\frac{\text{miles}}{\text{minute}}$ )

$$\text{Average Speed} = \frac{\text{miles}}{\text{minute}} = \frac{10}{20} = 0.5 \text{ mi/min.}$$

Now we simply convert to m/s

$$\begin{aligned} \frac{0.5 \text{ mi}}{1 \text{ min}} \times \frac{1 \text{ min}}{60 \text{ sec}} \times \frac{1609 \text{ m}}{1 \text{ mi}} \\ \approx 13 \text{ m/s.} \end{aligned}$$

- **Unit conversion:** The density of iron is  $7.86\text{g}/\text{cm}^3$  under standard conditions. Convert this to  $\text{kg}/\text{m}^3$  :

$$\begin{aligned} \frac{7.86 \text{ g}}{1 \text{ cm}^3} \times \left( \frac{100 \text{ cm}}{1 \text{ m}} \right)^3 \times \frac{1 \text{ kg}}{1000 \text{ g}} \\ = \frac{7.86(100^3)(1 \text{ kg})}{1000(1 \text{ m})} \\ = 7.86 \cdot 10^3 \text{ kg/m}^3. \end{aligned}$$

- **Proportional:** I found that if I drive my car 110 miles, I use 4 gallons of gas. If I assume that the relationship between gas guzzled and distance driven is linearly proportional, how many gallons of gas do I use if I drive 275 miles?

To answer this lets find the linear equation

$$4 \text{ gal} = k(100 \text{ mi}).$$

Where  $k$  is some arbitrary factor, Since we know the relationship is proportional, we can write

$$\begin{aligned} \frac{X}{4\text{gal}} &= \frac{k(275 \text{ mi})}{k(110 \text{ mi})} \\ X &= 10 \text{ gal.} \end{aligned}$$

## 2.2 Chapter 2: Vectors

### 2.2.1 Vocabulary

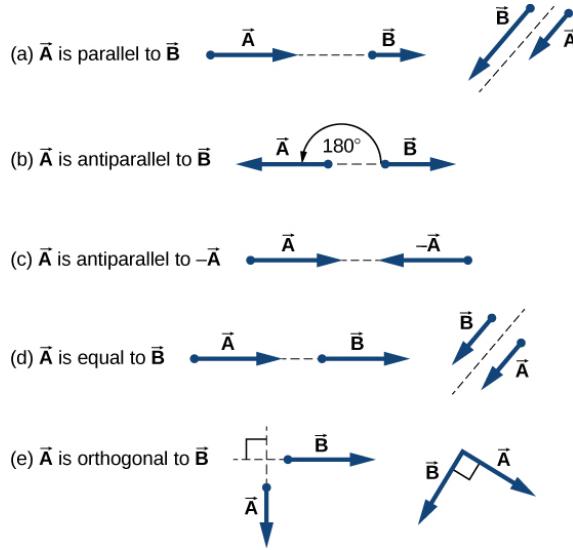
- **scalar quantities:**
- **vector quantities:** Examples of vector quantities include displacement, velocity, position, force, and torque.

### 2.2.2 Definitions and theorems (and important things)

- **direction:**
- **Algebraic Operations with vectors:**
  - **add or subtract:** two vectors
  - **multiply:** a vector by a scalar or by another vector
  - **cannot:** divide by a vector. The operation of division by a vector is not defined.
- **Vector Notation:** We denote a vector with a bold face letter with an arrow above it. For example

$$\vec{V}.$$

- **Displacement:** General term used to describe change in position.
- **The magnitude of a vector is the length of the arrow used to represent it:**
- **Vector relations:**



- **Scalars:** When a vector  $\vec{A}$  is multiplied by a positive scalar  $\alpha$ , the result is a new vector  $\vec{B}$  that is parallel to  $\vec{A}$ ::

$$\vec{B} = \alpha \vec{A}.$$

The magnitude of this new vector  $\vec{B}$  is

$$B = |\alpha|A.$$

Where  $B$  is the magnitude of  $\vec{B}$  and  $A$  is the magnitude of  $\vec{A}$

- **Antiparallel vectors:** If they are antiparallel (Parallel, but in different directions), we write

$$\vec{A} = -\vec{B}.$$

- **Resultant vectors**

$$\vec{R} = \vec{A} + \vec{B}.$$

- **Vector Laws :**

- **Commutitive law:**  $\vec{A} + \vec{B} = \vec{B} + \vec{A}$ :
- **Assosiative law:**  $(\vec{A} + \vec{B}) + \vec{C} = \vec{A} + (\vec{B} + \vec{C})$ :
- **Distributive law:**  $\alpha_1 \vec{A} + \alpha_2 \vec{A} = (\alpha_1 + \alpha_2) \vec{A}$ :

- **Vector addition:**

$$\vec{A} + \vec{B}.$$

When two vectors are parallel, we can simply sum their magnitudes. However, if the vectors lie in different directions, the approach for vector addition involves finding their  $x$  and  $y$  components, summing them and then finding the magnitude.

- **Vector Subtraction:**

$$\vec{A} + (-\vec{B}).$$

When two vectors are aligned but point in exactly opposite directions, you can subtract their magnitudes (assuming you define one direction as positive and the other as negative) to find the net effect. This is a specific case of adding magnitudes where the direction is implicitly considered through subtraction.

- **Unit Vector notation:** A unit vector in a normed vector space is a vector of length 1. We declare unit vectors with a hat instead of an arrow, consider the following example

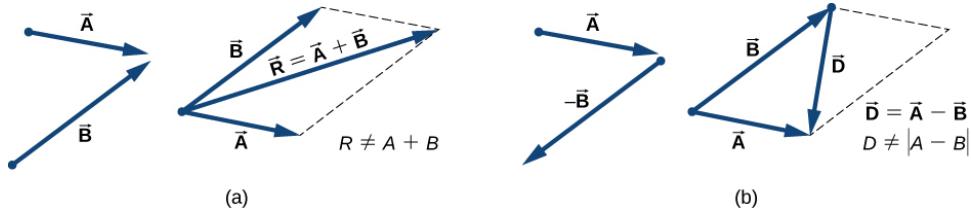
$$\hat{\mathbf{u}}.$$

We usually denote the unit vector along the positive x-axis  $\hat{i}$ , the unit vector along the positive y-axis  $\hat{j}$ , and the unit vector along the positive z-axis  $\hat{k}$

- **Unit vector example:** For example, instead of saying vector  $\vec{D}_{AB}$  has a magnitude of 6.0 km and a direction of northeast, we can introduce a unit vector  $\hat{u}$  that points to the northeast and say succinctly that  $\vec{D}_{AB} = (6.0 \text{ km})\hat{u}$ . Then the southwesterly direction is simply given by the unit vector  $-\hat{u}$ . In this way, the displacement of 6.0 km : in the southwesterly direction is expressed by the vector

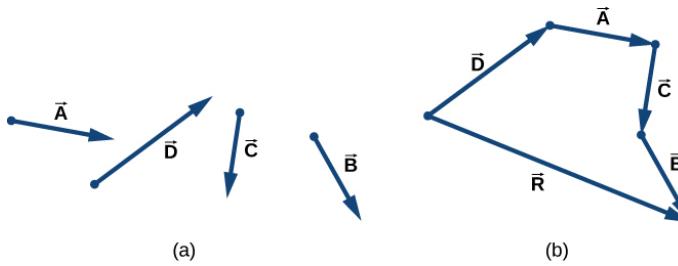
$$\vec{D}_{BA} = (-6.0 \text{ km})\hat{u}..$$

- **Parallelogram rule for resultant or difference vectors in two dimensions.:**



It follows from the parallelogram rule that neither the magnitude of the resultant vector nor the magnitude of the difference vector can be expressed as a simple sum or difference of magnitudes  $A$  and  $B$ , because the length of a diagonal cannot be expressed as a simple sum of side lengths.

- **tail-to-head geometric construction:**



- **Vector  $x$  and  $y$  components:** The  $x$  component can be denoted,  $\vec{A}_x$  the  $y$  component can be denoted  $\vec{A}_y$ . Thus, the vector can be represented as

$$\vec{A} = \vec{A}_x + \vec{A}_y.$$

- **Unit vectors of the axes:** It is customary to denote the positive direction on the  $x$ -axis by the unit vector  $\hat{i}$  and the positive direction on the  $y$ -axis by the unit vector  $\hat{j}$ . Unit vectors of the axes,  $\hat{i}$  and  $\hat{j}$ , define two orthogonal directions in the plane. The  $x$ - and  $y$ -components of a vector can now be written in terms of the unit vectors of the axes:

$$\begin{cases} \vec{A}_x = A_x \hat{i} \\ \vec{A}_y = A_y \hat{j} \end{cases} \quad (18)$$

- **Component form of a vector:**

$$\vec{A} = A_x \hat{i} + A_y \hat{j}.$$

- **Finding components given initial and terminal points:** If we know the coordinates  $b(x_b, y_b)$  of the origin point of a vector (where  $b$  stands for "beginning") and the coordinates  $e(x_e, y_e)$  of the end point of a vector (where  $e$  stands for "end"), we can obtain the scalar components of a vector simply by subtracting the origin point coordinates from the end point coordinates:

$$\begin{aligned} A_x &= x_e - x_b \\ A_y &= y_e - y_b. \end{aligned}$$

- **Magnitude  $A$  of a vector with components  $A_x$  and  $A_y$ :**

$$\begin{aligned} A^2 &= A_x^2 + A_y^2 \\ A &= \sqrt{A_x^2 + A_y^2}. \end{aligned}$$

This equation works even if the scalar components of a vector are negative.

- **Finding theta for a vector (used for direction angles):**

$$\tan \theta = \frac{A_y}{A_x}.$$

- **Direction angles for vectors in the first:**

- **Direction angles for vectors in the second quadrant:**

$$\theta_A = \theta.$$

- Direction angles for vectors in the second quadrant:

$$\theta_A = 180 - \theta.$$

- Direction angles for vectors in the second quadrant:

$$\theta_A = 180 + \theta.$$

- Direction angles for vectors in the fourth quadrant:

$$\theta_A = 360 - \theta.$$

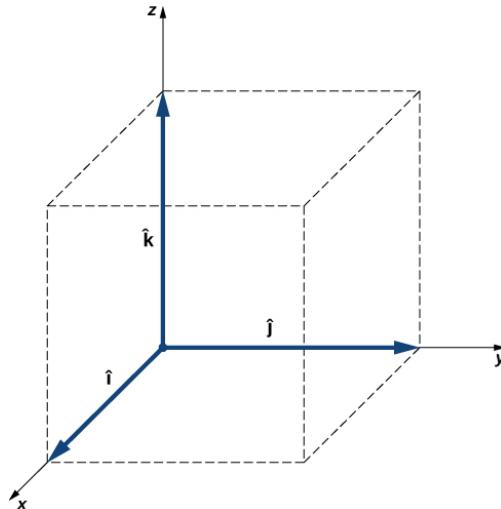
- Finding  $A_x$  and  $A_y$  when the magnitude and direction angle are known:

$$\begin{cases} A_x = A \cos \theta_A \\ A_y = A \sin \theta_A \end{cases} \quad (19)$$

- Polar form:

$$\begin{aligned} x &= r \cos \varphi \\ y &= r \sin \varphi. \end{aligned}$$

- Three dimensional plane:



- $z$  component of a vector:

$$\vec{A}_z = A_z \hat{k}.$$

Where  $A_z$  is given by

$$z_e - z_b.$$

- **Vector in three dimensions:** A vector in three-dimensional space is the vector sum of its three vector components.

$$\vec{A} = A_x \hat{i} + A_y \hat{j} + A_z \hat{k}..$$

- **Magnitude of a vector in three dimensions:**

$$A = \sqrt{A_x^2 + A_y^2 + A_z^2}.$$

- **Null vector:** Denoted by

$$\vec{0}.$$

Has all components 0. Thus,

$$\vec{0} = 0\hat{i} + 0\hat{j} + 0\hat{k}.$$

Thus, it has no direction and no length

- **equal vectors:** if and only if their difference is the null vector: Hence, we can write  $\vec{A} = \vec{B}$  if and only if the corresponding components of vectors  $\vec{A}$  and  $\vec{B}$  are equal:

$$\vec{A} = \vec{B} \Leftrightarrow \begin{cases} A_x = B_x \\ A_y = B_y \\ A_z = B_z \end{cases}$$

- **Components of a resultant vector:**

$$\begin{cases} R_x = A_x + B_x \\ R_y = A_y + B_y \\ R_z = A_z + B_z \end{cases} \quad (20)$$

- **components of a resultant of many vectors:** if we are to sum up  $N$  vectors  $\vec{F}_1, \vec{F}_2, \vec{F}_3, \dots, \vec{F}_N$ , where each vector is  $\vec{F}_k = F_{kx}\hat{i} + F_{ky}\hat{j} + F_{kz}\hat{k}$ , the resultant vector  $\vec{F}_R$  is

$$\begin{cases} F_{R_x} = \sum_{k=1}^N F_{kx} = F_{1x} + F_{2x} + F_{3x} + \dots + F_{Nx} \\ F_{R_y} = \sum_{k=1}^N F_{ky} = F_{1y} + F_{2y} + F_{3y} + \dots + F_{Ny} \\ F_{R_z} = \sum_{k=1}^N F_{kz} = F_{1z} + F_{2z} + F_{3z} + \dots + F_{Nz} \end{cases} \quad (21)$$

With the component form

$$\vec{F}_R = F_{R_x}\hat{i} + F_{R_y}\hat{j} + F_{R_z}\hat{k}.$$

- **Finding unit vector (Direction) of some vector:** Suppose we have some vector  $\vec{V}$ . Then

$$\hat{V} = \frac{\vec{V}}{V}.$$

- **Dot product:**

$$\vec{A} \cdot \vec{B} = AB \cos \varphi.$$

Where  $\varphi$  is the angle between the vectors

- **Dot product of two parallel vectors:**

$$\vec{A} \cdot \vec{B} = AB \cos 0^\circ = AB.$$

- Dot product of two anti-parallel vectors:

$$\vec{A} \cdot \vec{B} = AB \cos 180^\circ = -AB.$$

- Dot product of two orthogonal vectors:

$$\vec{A} \cdot \vec{B} = AB \cos 90^\circ = 0.$$

- Dot product of a vector with itself:

$$\vec{A} \cdot \vec{A} = AA \cos 0 = A^2.$$

- **Dot products of unit vectors:** Scalar products of the unit vector of an axis with other unit vectors of axes always vanish (equals 0) because these unit vectors are orthogonal:

- **Dot product of the same unit vector:** The dot product of the same unit vector is 1

$$\hat{i} \cdot \hat{i} = i^2 = 1.$$

- Using dot product to find scalar x-component:

$$\vec{A} \cdot \hat{i} = \|\vec{A}\| \|\hat{i}\| \cos \theta_A = A \cos \theta_A = A_x.$$

- Using dot product to find scalar y-component:

$$\vec{A} \cdot \hat{j} = \|\vec{A}\| \|\hat{j}\| \cos (90^\circ - \theta_A) = A \sin \theta_A = A_y.$$

- **Trig complementary angles:** For a right angle triangle, the sine of the complementary angle is the cosine of the angle. And vice versa

$$\sin 90^\circ - \theta = \cos \theta$$

$$\cos 90^\circ - \theta = \sin \theta.$$

- Dot product second method of computation:

$$\vec{A} \cdot \vec{B} = A_x B_x + A_y B_y + A_z B_z.$$

- **Equation for  $\cos(\varphi)$ :**

$$\cos(\varphi) = \frac{\vec{A} \cdot \vec{B}}{AB}.$$

- **Si unit for work:** is the joule (J), where

$$1 \text{ J} = 1 \text{ N} \cdot \text{m}.$$

- **The Work of a Force:** When force  $\vec{F}$  pulls on an object and when it causes its displacement  $\vec{D}$ , we say the force performs work. The amount of work the force does is the scalar product  $\vec{F} \cdot \vec{D}$ .

- **Cross Product (Vector Product):** The vector product of two vectors  $\vec{A}$  and  $\vec{B}$  is denoted by  $\vec{A} \times \vec{B}$  and is often referred to as a cross product. The vector product is a vector that has its direction perpendicular to both vectors  $\vec{A}$  and  $\vec{B}$ . In other words, vector  $\vec{A} \times \vec{B}$  is perpendicular to the plane that contains vectors  $\vec{A}$  and  $\vec{B}$ . The magnitude of the vector product is defined as

$$\|\vec{A} \times \vec{B}\| = AB \sin \varphi, \quad (22)$$

where angle  $\varphi$ , between the two vectors, is measured from vector  $\vec{A}$  (first vector in the product) to vector  $\vec{B}$  (second vector in the product), as indicated in Figure 2.29, and is between  $0^\circ$  and  $180^\circ$ .

- **antiparallel:** ( $\varphi = 180^\circ$ ) because  $\sin 0^\circ = \sin 180^\circ = 0$ .
- **anti-commutative:**

$$\vec{A} \times \vec{B} = -\vec{B} \times \vec{A}.$$

- **Torque:** Denoted with the greek letter "tau" is the vector product of the distance between the pivot to force with the force:

$$\vec{\tau} = \vec{R} \times \vec{F}.$$

- **the cross product has the following distributive property::**

$$\vec{A} \times (\vec{B} + \vec{C}) = \vec{A} \times \vec{B} + \vec{A} \times \vec{C}.$$

- **Cross product of the same unit vectors:** The cross product of the same unit vectors is 0
- **Cross product between unit vectors:**

$$\begin{cases} \hat{i} \times \hat{j} = +\hat{k} \\ \hat{j} \times \hat{i} = -\hat{k} \\ \hat{j} \times \hat{k} = +\hat{i} \\ \hat{k} \times \hat{j} = -\hat{i} \\ \hat{k} \times \hat{i} = +\hat{j} \\ \hat{i} \times \hat{k} = -\hat{j} \end{cases} \quad (23)$$

**Notice:** The cross product of two different unit vectors is always a third unit vector.

- **Computation of the cross product:**

$$\vec{C} = \vec{A} \times \vec{B} = (A_y B_z - A_z B_y) \hat{i} + (A_z B_x - A_x B_z) \hat{j} + (A_x B_y - A_y B_x) \hat{k}.$$

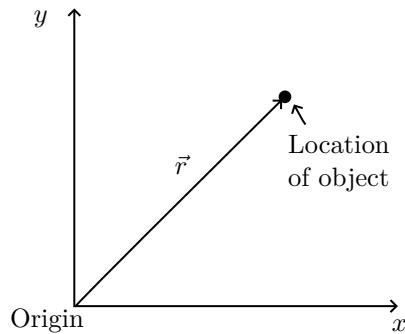
- **vector quantity:** of speed plus the direction. Velocity is

$$\vec{v} = \frac{\text{Displacement}}{\text{Time}} = \frac{\|\vec{d}\|}{t}.$$

Where  $\|\vec{d}\|$  is the magnitude of the displacement vector, and  $t$  is the total elapsed time.

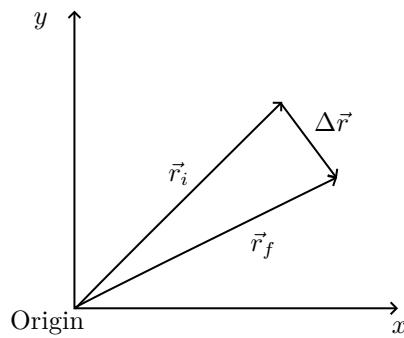
- **position vector:** Denoted

$$\vec{r}.$$



- **Displacement:** is defined as the change in the position vector. The final position vector minus the initial position vector

$$\Delta\vec{r} = \vec{r}_f - \vec{r}_i.$$



### 2.2.3 Problems to remember

- **Unit vectors:** A long measuring stick rests against a wall in a physics laboratory with its 200-cm end at the floor. A ladybug lands on the 100-cm mark and crawls randomly along the stick. It first walks 15 cm toward the floor, then it walks 56 cm toward the wall, then it walks 3 cm toward the floor again. Then, after a brief stop, it continues for 25 cm toward the floor and then, again, it crawls up 19 cm toward the wall before coming to a complete rest (Figure 2.8). Find the vector of its total displacement and its final resting position on the stick.

If we choose the direction along the stick toward the floor as the direction of unit vector  $\hat{u}$ , then the direction toward the floor is  $+ \hat{u}$  and the direction toward the wall is  $- \hat{u}$ . The ladybug makes a total of five displacements:

$$\begin{aligned}\vec{D}_1 &= (15 \text{ cm})(+ \hat{u}), \quad \vec{D}_2 \\ &= (56 \text{ cm})(- \hat{u}), \quad \vec{D}_3 \\ &= (3 \text{ cm})(+ \hat{u}), \quad \vec{D}_4 \\ &= (25 \text{ cm})(+ \hat{u}), \quad \text{and} \quad \vec{D}_5 \\ &= (19 \text{ cm})(- \hat{u})..\end{aligned}$$

The total displacement  $\vec{D}$  is the resultant of all its displacement vectors.

The resultant of all the displacement vectors is

$$\begin{aligned}\vec{D} &= \vec{D}_1 + \vec{D}_2 + \vec{D}_3 + \vec{D}_4 + \vec{D}_5 \\ &= (15 \text{ cm})(+ \hat{u}) + (56 \text{ cm})(- \hat{u}) + (3 \text{ cm})(+ \hat{u}) + (25 \text{ cm})(+ \hat{u}) + (19 \text{ cm})(- \hat{u}) \\ &= (15 - 56 + 3 + 25 - 19) \text{ cm } \hat{u} = -32 \text{ cm } \hat{u}.\end{aligned}$$

In this calculation, we use the distributive law. The result reads that the total displacement vector points away from the 100-cm mark (initial landing site) toward the end of the meter stick that touches the wall. The end that touches the wall is marked 0 cm, so the final position of the ladybug is at the  $(100 - 32)$  cm = 68 cm mark.

## 2.3 Chapter 3: Motion along a straight line

### 2.3.1 Definitions and theorems

- **Kinematics:** is a subfield of physics, developed in classical mechanics, that describes the motion of points, bodies, and systems of bodies without considering the forces that cause them to move
- **Displacement:** Displacement  $\Delta x$  is the change in position of an object:

$$\Delta x = x_f - x_0,$$

where  $\Delta x$  is displacement,  $x_f$  is the final position, and  $x_0$  is the initial position.

- We define total displacement  $\Delta x_{\text{Total}}$ , as the sum of the individual displacements, and express this mathematically with the equation

$$\Delta x_{\text{Total}} = \sum \Delta x_i.$$

- **the distance traveled:** is the sum of the magnitudes of the individual displacements:

$$x_{\text{total}} = \sum_{i=1}^n |\Delta x_i|.$$

- **average velocity.** If  $x_1$  and  $x_2$  are the positions of an object at times  $t_1$  and  $t_2$  respectively, then

$$\begin{aligned} \text{Average Velocity} &= \bar{v} = \frac{\text{Displacement between two points}}{\text{Time needed to make the displacement}} \\ \bar{v} &= \frac{\Delta x}{\Delta t} = \frac{x_2 - x_1}{t_2 - t_1}. \end{aligned}$$

This vector quantity is simply the total displacement between two points divided by the time taken to travel between them. The time taken to travel between two points is called the **elapsed time**  $\Delta t$

- **instantaneous velocity:** of an object is the limit of the average velocity as the elapsed time approaches zero, or the derivative of  $x$  with respect to  $t$ :

$$\begin{aligned} v(t) &= \lim_{\Delta t \rightarrow 0} \frac{f(t + \Delta t) - f(t)}{\Delta t} \\ &= \frac{d}{dt} x(t) = \frac{d\vec{r}}{dt}.. \end{aligned}$$

- **average speed:** by finding the total distance traveled divided by the elapsed time:

$$\text{Average speed} = \bar{s} = \frac{\text{Total distance}}{\text{Elapsed time}}..$$

- **instantaneous speed:** from the magnitude of the instantaneous velocity:

$$\text{instantaneous speed} = |v(t)|.$$

- **Calculating Instantaneous Velocity:** When calculating instantaneous velocity, we need to specify the explicit form of the position function  $x(t)$ . If each term in the  $x(t)$  equation has the form of  $At^n$  where  $A$  is a constant and  $n$  is an integer, this can be differentiated using the power rule to be:

$$\frac{d(At^n)}{dt} = Ant^{n-1}.$$

- **Average acceleration:** is the rate at which velocity changes:

$$\bar{a} = \frac{\Delta v}{\Delta t} = \frac{v_f - v_0}{t_f - t_0},$$

where  $\bar{a}$  is average acceleration,  $v$  is velocity, and  $t$  is time. (The bar over the  $a$  means average acceleration.)

**Note:** acceleration occurs when velocity changes in magnitude (an increase or decrease in speed) or in direction, or both.

- **Acceleration as a vector:** Acceleration is a vector in the same direction as the change in velocity,  $\Delta v$ . Since velocity is a vector, it can change in magnitude or in direction, or both. Acceleration is, therefore, a change in speed or direction, or both.

Keep in mind that although acceleration is in the direction of the change in velocity, it is not always in the direction of motion. When an object slows down, its acceleration is opposite to the direction of its motion. Although this is commonly referred to as deceleration

- **Distance over constant acceleration:**

$$d = \frac{1}{2}at^2.$$

Where  $a$  is the acceleration, and  $t$  is the time

- **instantaneous acceleration:**

$$a(t) = \frac{d}{dt}v(t).$$

- **Derivative of velocity function:** Suppose we have some function

$$v(t) = (20m/s)t - (10m/s^2)t^2.$$

When we find the acceleration function, we are taking the derivative of the velocity function. Thus, we are finding the change in velocity with respect to time. Consequently, our terms become

$$a(t) = 20m/s^2 - (10m/s^3)t.$$

**Notice:** How we are dividing by an additional unit of time, thus the exponents for our seconds increases by one.

- **Simplified notation:** If we take initial time to be zero, and final quantitys without subscript, then we have

$$\Delta t = t$$

$$\Delta x = x - x_0$$

$$\Delta v = v - v_0$$

- **Assumption of constant acceleration:** This assumption allows us to avoid using calculus to find instantaneous acceleration. Since acceleration is constant, the average and instantaneous accelerations are equal—that is,

$$\bar{a} = a = \text{constant}.$$

Thus, we can use the symbol  $a$  for acceleration at all times.

- **Final Position function:**

$$x = x_0 + \bar{v}t.$$

- **Average velocity under constant acceleration:**

$$\bar{v} = \frac{v_0 + v}{2} \quad (\text{Constant } a).$$

This reflects the fact that when acceleration is constant,  $\bar{v}$  is just the simple average of the initial and final velocities.

- **Final Velocity function under constant acceleration:**

$$v = v_0 + at \quad (\text{Constant } a).$$

- **Equation for final position under constant acceleration:**

$$x = x_0 + v_0 t + \frac{1}{2}at^2 \quad (\text{Constant } a).$$

When initial position and velocity are both zero, we have

$$x = \frac{1}{2}at^2 \quad (\text{Constant } a).$$

- **relationships seen in final position under constant acceleration equation:**

- Displacement depends on the square of the elapsed time when acceleration is not zero.
- If acceleration is zero, then initial velocity equals average velocity  $v_0 = \bar{v}$ , and  $x = x_0 + v_0 t + \frac{1}{2}at^2$  becomes  $x = x_0 + v_0 t$ .

- **Final velocity equation (no time required):**

$$v^2 = v_0^2 + 2a(x - x_0) \quad (\text{Constant } a).$$

- **additional insights into the general relationships among physical quantities::**

- The final velocity depends on how large the acceleration is and the distance over which it acts.
- For a fixed acceleration, a car that is going twice as fast doesn't simply stop in twice the distance. It takes much farther to stop. (This is why we have reduced speed zones near schools.)

- **acceleration in terms of velocities and displacement:**

$$a = \frac{v^2 - v_0^2}{2(x - x_0)}.$$

- **two-body pursuit problems:**
  - Find equations of motion for both bodies (with the same parameter)
  - eliminate the parameter
  - plug in knowns to solve for the unknown
- **constant acceleration, independent of their mass.:**
- **free fall:**
- **acceleration due to gravity:**

$$g = 9.8 \text{ m/s}^2.$$

**Note:** If we define the upward direction as positive, then  $a = g = -9.8\text{m/s}^2$ , and if we define the downward direction as positive, then  $a = g = 9.8\text{m/s}^2$

- **vertical displacement:** We denote vertical displacement with the symbol  $y$
- **Kinematic equations for objects in free fall:** We assume here that acceleration equals  $-g$  (with the positive direction upward).

$$\begin{aligned} v &= v_0 - gt \\ y &= y_0 + v_0 t - \frac{1}{2}gt^2 \\ v^2 &= v_0^2 - 2g(y - y_0). \end{aligned}$$

- **Change in velocity:**

$$\Delta v = a\Delta t.$$

If you don't care about position this relates the acceleration, intervals of time and changes in velocity

- **Change in position:**

$$\Delta x = \frac{1}{2}(v_f - v_i)\Delta t.$$

## 2.4 Chapter 4: Motion in two and three dimensions

### 2.4.1 Definitions and theorems

- **Location of a particle in space:**

$$\begin{aligned}x &= x(t) \\y &= y(t) \\z &= z(t).\end{aligned}$$

Where  $x, y, z$  are functions of time ( $t$ ).

- **Position vector in space:**

$$\vec{r}(t) = x(t)\hat{i} + y(t)\hat{j} + z(t)\hat{k}.$$

- **Displacement vector in space:** Suppose we have a particle. At time  $t_1$  the particle is located at  $P_1$  with position vector  $\vec{r}(t_1)$ . At some later time  $t_2$ , the particle is located at  $P_2$  with position vector  $\vec{r}(t_2)$ . The displacement vector  $\Delta\vec{r}$  is found by subtracting  $\vec{r}(t_1)$  from  $\vec{r}(t_2)$ :

$$\Delta\vec{r} = \vec{r}(t_2) - \vec{r}(t_1).$$

Thus, the displacement vector  $\Delta\vec{r} = \vec{r}(t_2) - \vec{r}(t_1)$  is the vector from  $P_1$  to  $P_2$

- **Instantaneous velocity vector in two and three dimensions:** We can do the same operation in two and three dimensions, but we use vectors. The instantaneous velocity vector is now

$$\begin{aligned}\vec{v} &= \lim_{\Delta t \rightarrow 0} \frac{\vec{r}(t - \Delta t) - \vec{r}(t)}{\Delta t} \\&= \frac{d\vec{r}}{dt}.\end{aligned}$$

- **Velocity in component form:**

$$\vec{v}(t) = v_x\hat{i} + v_y\hat{j} + v_z\hat{k}.$$

Where

$$v_x(t) = \frac{dx(t)}{dt}, \quad v_y(t) = \frac{dy(t)}{dt}, \quad v_z(t) = \frac{dz(t)}{dt}.$$

- **Average velocity in two and three dimensions:** If only the average velocity is of concern, we have the vector equivalent of the one-dimensional average velocity for two and three dimensions

$$\vec{v}_{\text{avg}} = \frac{\vec{r}(t_2) - \vec{r}(t_1)}{t_2 - t_1}.$$

- **The Independence of Perpendicular Motions:** When we look at the three-dimensional equations for position and velocity written in unit vector notation, we see the components of these equations are separate and unique functions of time that do not depend on one another. Motion along the x direction has no part of its motion along the y and z directions, and similarly for the other two coordinate axes. Thus, the motion of an object in two or three dimensions can be divided into separate, independent motions along the perpendicular axes of the coordinate system in which the motion takes place.

- **Independence of motion:** In the kinematic description of motion, we are able to treat the horizontal and vertical components of motion separately. In many cases, motion in the horizontal direction does not affect motion in the vertical direction, and vice versa.
- **Instantaneous Acceleration:** This acceleration vector is the instantaneous acceleration and it can be obtained from the derivative with respect to time of the velocity function. The only difference in two or three dimensions is that these are now vector quantities. Taking the derivative with respect to time  $\vec{v} : (t)$ . We find

$$\vec{a}(t) = \frac{d\vec{v}(t)}{dt}.$$

- The acceleration in terms of components is:

$$\vec{a}(t) = \frac{dv_x(t)}{dt}\hat{i} + \frac{dv_y(t)}{dt}\hat{j} + \frac{dv_z(t)}{dt}\hat{k}.$$

- acceleration in terms of the second derivative of the position function::

$$\vec{a}(t) = \frac{d^2x(t)}{dt^2}\hat{i} + \frac{d^2y(t)}{dt^2}\hat{j} + \frac{d^2z(t)}{dt^2}\hat{k}.$$

- **Equations for position and velocity in the two and three dimensions (8):** For simplicity I will only list the ones for the x-direction. However, keep in mind that these same equations hold for  $y$  and  $z$

- Position with initial position and average velocity::

$$x(t) = x_0 + (v_x)_{\text{avg}} t.$$

- Final velocity with initial velocity, acceleration, and time:

$$v_x(t) = v_{0x} + a_x t.$$

- Position:

$$x(t) = x_0 + v_{0x}t + \frac{1}{2}a_x t^2.$$

- Velocity with initial velocity, acceleration, final and initial position:

$$v_x^2(t) = v_{0x}^2 + 2a_x(x - x_0).$$

- These equations can be substituted into the equations for the position and velocity vectors in component form, and velocity vector in component form without the z-component to obtain the position vector and velocity vector as a function of time in two dimensions::

$$\begin{aligned}\vec{r}(t) &= x(t)\hat{i} + y(t)\hat{j} \\ \vec{v}(t) &= v_x(t)\hat{i} + v_y(t)\hat{j}.\end{aligned}$$

- trajectory:

- **Acceleration in projectile motion:** Defining the positive direction to be upward, the components of acceleration are then very simple:

$$\begin{aligned}a_y &= -9.8 \text{ m/s}^2 \\ a_x &= 0.\end{aligned}$$

Because gravity is vertical,  $a_x = 0$ . If  $a_x = 0$ , this means the initial velocity in the  $x$  direction is equal to the final velocity in the  $x$  direction, or  $v_x = v_{0x}$

- **Kinematic equations for motion in a uniform gravitational field:**

- Horizontal motion

$$v_{0x} = v_x, \quad x = x_0 + v_x t.$$

- **Average velocity (vertical):**

- Basic average velocity

$$V_{y,\text{avg}} = \frac{y - y_0}{\Delta t}.$$

- **Average velocity in special conditions:**

$$V_{y,\text{avg}} = \frac{v_i + v_f}{2}.$$

**Note:** gives the same result for average velocity in scenarios of uniform acceleration, such as projectile motion or free fall, under specific conditions. This formula calculates the average of the initial and final velocities, assuming that the acceleration is constant throughout the motion. It works perfectly for the vertical component of projectile motion when considering the ascent or descent separately, or any motion where the start and end points are symmetrical in terms of velocity but in opposite directions (e.g., going up and coming back down to the same height).

- Vertical motion

$$\begin{aligned} y &= y_0 + \frac{1}{2}(v_{0y} + v_y)t \\ v_y &= v_{0y} + gt \\ y &= y_0 + v_{0y}t + \frac{1}{2}gt^2 \\ v_y^2 &= v_{0y}^2 + 2g(y - y_0). \end{aligned}$$

- **Projectile motion max height:**

$$h = \frac{v_{0y}^2}{2g}.$$

This equation defines the maximum height of a projectile above its launch position and it depends only on the vertical component of the initial velocity.

- **Parabolic trajectory:** If the motion is parabolic, we should use the equation with  $t^2$  : to solve for  $t$  and vertical displacement
- **Time of flight:** We can solve for the time of flight of a projectile that is both launched and impacts on a flat horizontal surface by performing some manipulations of the kinematic equations. We note the position and displacement in y must be zero at launch and at impact on an even surface. Thus, We set the displacement in y equal to zero and find

$$T_{\text{tof}} = \frac{2(v_0 \sin \theta_0)}{g}.$$

**Note:** This is the time of flight for a projectile both launched and impacting on a flat horizontal surface. This equation does not apply when the projectile lands at a different elevation than it was launched,

- **Trajectory:** The trajectory of a projectile can be found by eliminating the time variable  $t$  from the kinematic equations for arbitrary  $t$  and solving for  $y(x)$ . We take  $x_0 = y_0 := 0$  so the projectile is launched from the origin.

$$y = (\tan \theta_0)x - \left[ \frac{g}{2(v_0 \cos \theta_0)^2} \right] x^2.$$

**Note:** This trajectory equation is of the form  $y = ax + bx^2$  which is an equation of a parabola with coefficients

$$a = \tan \theta_0, \quad b = -\frac{g}{2(v_0 \cos \theta_0)^2}.$$

- **Range:** From the trajectory equation we can also find the range, or the horizontal distance traveled by the projectile.

$$R = \frac{v_0^2 \sin(2\theta_0)}{g}.$$

- **time of flight, trajectory, and range:** are derived with  $g$  as negative, thus we do not make  $g$  negative when plugging into the equation.
- **Centripetal Acceleration:** a particle moving in a circle at a constant speed has an acceleration with magnitude

$$a_c = \frac{v^2}{r}.$$

**Note:** The direction of the acceleration vector is toward the center of the circle

- **Position vector for a particle executing circular motion:** As the particle moves on the circle, its position vector sweeps out the angle  $\theta$  with the x-axis. Vector  $\vec{r}(t)$  making an angle  $\theta$  with the x-axis is shown with its components along the x- and y-axes. The magnitude of the position vector is  $A = |\vec{r}(t)|$  and is also the radius of the circle, so that in terms of its components,

$$\vec{r}(t) = A \cos(\omega t) \hat{i} + A \sin(\omega t) \hat{j}.$$

- **Average speed for circular motion:** If  $T$  is the item it takes to go around the circle once, and the distance around a circle is  $2\pi r$ , then

$$\begin{aligned} v &= \frac{2\pi r}{T} \\ \implies T &= \frac{2\pi r}{v}. \end{aligned}$$

- **Amplitude ( $A$ ):** In the context of uniform circle motion, we define the amplitude  $A$  as the radius of the circle. Thus,

$$A = r.$$

- **Angular frequency:** In the previous equation,  $\omega$  is a constant called the angular frequency of the particle.

The angular frequency has units of radians (rad) per second and is simply the number of radians of angular measure through which the particle passes per second. The angle  $\theta$  that the position vector has at any particular time is  $\omega t$

- **Angular frequency computation:** If  $T$  is the period of motion, or the time to complete one revolution ( $2\pi$  rad), then

$$\begin{aligned}\omega &= \frac{2\pi}{T} \\ \implies T &= \frac{2\pi}{\omega}.\end{aligned}$$

- **Example: Finding  $\omega$ :** A flywheel is rotating at 21 rev/s. What is the total angle, in radians, through which a point on the flywheel rotates in 37 s?

To find  $\omega$ , we must first deduce  $T$ , if we complete 21 revolutions in one second, then the time to complete one revolution is  $\frac{1}{21}s$ . Thus, we have  $\omega$  as

$$\frac{2\pi}{\frac{1}{21}} = 41\pi.$$

Which gives us

$$\begin{aligned}\theta &= \omega t \\ &= 41\pi(37s).\end{aligned}$$

- **Finding  $\theta$  with  $\omega$ :**

$$\theta = \omega t.$$

- **Velocity for circular motion:**

$$\vec{v}(t) = \frac{d\vec{r}(t)}{dt} = -A\omega \sin(\omega t)\hat{i} + A\omega \cos(\omega t)\hat{j}.$$

- **Acceleration for circular motion:**

$$\vec{a}(t) = \frac{d\vec{v}(t)}{dt} = -A\omega^2 \cos(\omega t)\hat{i} - A\omega^2 \sin(\omega t)\hat{j}.$$

From this equation we see that the acceleration vector has magnitude  $A\omega^2$  and is directed opposite the position vector, toward the origin, because  $\vec{a}(t) = -\omega^2\vec{r}(t)$ .

- **Nonuniform Circular Motion Tangential Acceleration:** Circular motion does not have to be at a constant speed. A particle can travel in a circle and speed up or slow down, showing an acceleration in the direction of the motion.

In uniform circular motion, the particle executing circular motion has a constant speed and the circle is at a fixed radius. If the speed of the particle is changing as well, then we introduce an additional acceleration in the direction tangential to the circle. Such accelerations occur at a point on a top that is changing its spin rate, or any accelerating rotor. If the speed of the particle is changing, then it has a **tangential acceleration** that is the time rate of change of the magnitude of the velocity

$$a_T = \left| \frac{d\vec{v}}{dt} \right|.$$

- **Tangential acceleration total acceleration:** The direction of tangential acceleration is tangent to the circle whereas the direction of centripetal acceleration is radially inward toward the center of the circle. Thus, a particle in circular motion with a tangential acceleration has a total acceleration that is the vector sum of the centripetal and tangential accelerations

$$\vec{a} = \vec{a}_c + \vec{a}_T.$$

Thus total acceleration can be found by finding the magnitude of this vector

$$\|\vec{a}\| = \sqrt{\vec{a}_c^2 + \vec{a}_T^2}.$$

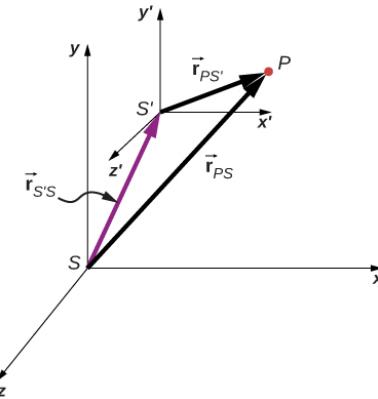
With

$$\tan \theta = \frac{y}{x}.$$

- **Reference frames:** When we say an object has a certain velocity, we must state it has a velocity with respect to a given reference frame. In most examples we have examined so far, this reference frame has been Earth.
- **Relative Motion in One Dimension Example:** Consider an example of a person sitting in a train moving east. If we choose east as the positive direction and Earth as the reference frame, then we can write the velocity of the train with respect to the Earth as  $\vec{v}_{TE} = 10 \text{ m/s } \hat{i}$  east, where the subscripts TE refer to train and Earth. Let's now say the person gets up out of her seat and walks toward the back of the train at  $2 \text{ m/s}$ . This tells us she has a velocity relative to the reference frame of the train. Since the person is walking west, in the negative direction, we write her velocity with respect to the train as  $\vec{v}_{PT} = -2 \text{ m/s } \hat{i}$ . We can add the two velocity vectors to find the velocity of the person with respect to Earth. This relative velocity is written as

$$\vec{v}_{PE} = \vec{v}_{PT} + \vec{v}_{TE}.$$

- **Relative Velocity in Two Dimensions Example:** Consider a particle  $P$  and reference frames  $S$  and  $S'$ . The position of the origin of  $S'$  as measured in  $S$  is  $\vec{r}_{S'S}$ , the position of  $P$  as measured in  $S'$  is  $\vec{r}_{PS'}$ , and the position of  $P$  as measured in  $S$  is  $\vec{r}_{PS}$ .



From this, we see

$$\vec{r}_{PS} = \vec{r}_{PS'} + \vec{r}_{S'S}.$$

- **Relative velocities (Still using previous example):** The relative velocities are the time derivatives of the position vectors. Therefore,

$$\vec{v}_{PS} = \vec{v}_{PS'} + \vec{v}_{S'S}.$$

So we see the velocity of a particle relative to  $S$  is equal to its velocity relative to  $S'$  plus the velocity of  $S'$  relative to  $S$

- **Relative accelerations (Still using same example):** We can also see how the accelerations are related as observed in two reference frames by differentiating

$$\vec{a}_{PS} = \vec{a}_{PS'} + \vec{a}_{S'S}.$$

We see that if the velocity of  $S'$  relative to  $S$  is a constant, then  $\vec{a}_{S'S} = 0$  and

$$\vec{a}_{PS} = \vec{a}_{PS'}.$$

This says the acceleration of a particle is the same as measured by two observers moving at a constant velocity relative to each other.

- **Percent difference equation:** If  $A$  is the experimental value, and  $B$  is the actual value, then the percent difference is given by

$$\text{Percent difference} = \frac{|A - B|}{A} \cdot 100\%.$$

## 2.5 Chapter 5: Newton's laws of motion

### 2.5.1 Definitions and Theorems

- **Dynamics:** is the study of how forces affect the motion of objects and systems.
- **constraints of Newtonian mechanics:** Newton's laws produce a good description of motion only when the objects are moving at speeds much less than the speed of light and when those objects are larger than the size of most molecules (about  $10^{-9}$  : m in diameter).
- **Intuitive definition of force:** A push or a pull—is a good place to start.

**Note:** We know that a push or a pull has both magnitude and direction (therefore, it is a vector quantity), so we can define force as the push or pull on an object with a specific magnitude and direction. Force can be represented by vectors or expressed as a multiple of a standard force.

- **Standard force:** A quantitative definition of force can be based on some standard force, just as distance is measured in units relative to a standard length. One possibility is to stretch a spring a certain fixed distance, and use the force it exerts to pull itself back to its relaxed shape—called a restoring force—as a standard. The
- **Contact forces:** Contact forces are due to direct physical contact between objects.
- **Field forces:** Field forces, however, act without the necessity of physical contact between objects. They depend on the presence of a “field” in the region of space surrounding the body under consideration.
- **Field:** You can think of a field as a property of space that is detectable by the forces it exerts.
- **The newton:** 1 N is the force needed to accelerate an object with a mass of 1 kg at a rate of  $1m/s^2$ . Hence,  $1N=1kg \cdot m/s^2$  :
- **Net external force:** The resultant of forces is call the net external force  $\vec{F}_{\text{net}}$  : and is found by taking the vector sum of all external forces acting on an object or system

$$\vec{F}_{\text{net}} = \sum \vec{F} = \vec{F}_1 + \vec{F}_2 + \dots$$

- **Newton's first law of motion:** A body at rest remains at rest or, if in motion, remains in motion at constant velocity unless acted on by a net external force.

**Note:** Newton's first law says that there must be a cause for any change in velocity (a change in either magnitude or direction) to occur. This cause is a net external force,

- **object would not slow down if friction were eliminated.:**
- **mass:** is a measure of the amount of matter in something.
- **Gravitation:** is the attraction of one mass to another

**Note:** such as the attraction between yourself and Earth that holds your feet to the floor. The magnitude of this attraction is your weight, and it is a force.

- **Inertia:** is the ability of an object to resist changes in its motion—in other words, to resist acceleration.

**Note:** Newton's first law is often called the **law of inertia**. The inertia of an object is measured by its mass.

- **Inertial reference frame:** In principle, we can make the net force on a body zero. If its velocity relative to a given frame is constant, then that frame is said to be inertial.

So by definition, an inertial reference frame is a reference frame in which Newton's first law is valid. Newton's first law applies to objects with constant velocity. From this fact, we can infer the following statement.

A reference frame moving at constant velocity relative to an inertial frame is also inertial. A reference frame accelerating relative to an inertial frame is not inertial.

- **Newton's first law in vector form::**

$$\vec{v} = \text{constant when } \vec{F}_{\text{net}} = \vec{0}_N.$$

- **static equilibrium:** Static systems do not change over time. Static equilibrium involves objects at rest
- **Dynamic equilibrium:** Dynamic systems change over time. Dynamic equilibrium involves objects in motion without acceleration,
- **a net force of zero means that an object is either at rest or moving with constant velocity,:;**
- **Acceleration is proportional to the net external force:.** That is,

$$a \propto \sum \vec{F}.$$

- **It also seems reasonable that acceleration should be inversely proportional to the mass of the system.:**

$$a \propto \frac{1}{m}.$$

In other words, the larger the mass (the inertia), the smaller the acceleration produced by a given force.

- **Experiments have shown that acceleration is exactly inversely proportional to mass, just as it is directly proportional to net external force.:**
- **Newton's second law of motion:** The acceleration of a system is directly proportional to and in the same direction as the net external force acting on the system and is inversely proportional to its mass. In equation form, Newton's second law is

$$\vec{a} = \frac{\vec{F}_{\text{net}}}{m},$$

where  $\vec{a}$  is the acceleration,  $\vec{F}_{\text{net}}$  is the net force, and  $m$  is the mass. This is often written in the more familiar form

$$\vec{F}_{\text{net}} = \sum \vec{F} = m\vec{a},$$

but the first equation gives more insight into what Newton's second law means. When only the magnitude of force and acceleration are considered, this equation can be written in the simpler scalar form:

$$F_{\text{net}} = ma.$$

- **Component Form of Newton's Second Law:** The equations of motion in three dimensions can be represented as:

$$\sum F_x = ma_x \quad (24)$$

$$\sum F_y = ma_y \quad (25)$$

$$\sum F_z = ma_z \quad (26)$$

- **A body's mass is a measure of its inertia:**
- **Newton's Second Law and Momentum:** Newton actually stated his second law in terms of momentum: "The instantaneous rate at which a body's momentum changes is equal to the net force acting on the body." ("Instantaneous rate" implies that the derivative is involved.) This can be given by the vector equation

$$\vec{F}_{\text{net}} = \frac{d\vec{p}}{dt}.$$

Momentum was described by Newton as "quantity of motion," a way of combining both the velocity of an object and its mass.

- **Momentum:** For now, it is sufficient to define momentum  $\vec{p}$  as the product of the mass of the object  $m$  and its velocity  $\vec{v}$ :

$$\vec{p} = m\vec{v}.$$

- **Weight:** If air resistance is negligible, the net force on a falling object is the gravitational force, commonly called its weight  $\vec{w}$  : , or its force due to gravity acting on an object of mass  $m$

Weight can be denoted as a vector because it has a direction; down is, by definition, the direction of gravity, and hence, weight is a downward force. The magnitude of weight is denoted as  $w$ .

The gravitational force on a mass is its weight. We can write this in vector form, where  $\vec{w}$  is weight and  $m$  is mass, as

$$\vec{w} = mg\vec{g}.$$

In scalar form, we can write

$$w = mg.$$

- **When the net external force on an object is its weight, we say that it is in free fall:**
- **Newton's third law of motion:** Whenever one body exerts a force on a second body, the first body experiences a force that is equal in magnitude and opposite in direction to the force that it exerts. Mathematically, if a body  $A$  exerts a force  $\vec{F}$  on body  $B$ , then  $B$  simultaneously exerts a force  $-\vec{F}$  on  $A$ , or in vector equation form,

$$\vec{F}_{AB} = -\vec{F}_{BA}$$

Newton's third law represents a certain symmetry in nature: Forces always occur in pairs, and one body cannot exert a force on another without experiencing a force itself.

- **a normal force** and here is given by the symbol  $\vec{N}$  :

$$\vec{N} = -mg\vec{g}.$$

**Note:** The word normal means perpendicular to a surface.

For objects resting on horizontal surfaces, this becomes the scalar form

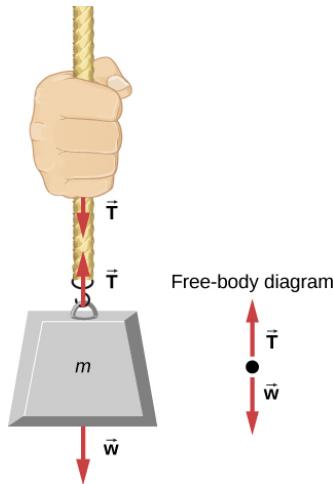
$$N = mg.$$

**Note:** The normal force can be less than the object's weight if the object is on an incline.

- **Normal force is perpendicular to the surface:**
- **tension:** is a force along the length of a medium; in particular, it is a pulling force that acts along a stretched flexible connector, such as a rope or cable.

Any flexible connector, such as a string, rope, chain, wire, or cable, can only exert a pull parallel to its length; thus, a force carried by a flexible connector is a tension with a direction parallel to the connector.

Consider the following figure



If the 5.00-kg mass in the figure is stationary, then its acceleration is zero and the net force is zero. The only external forces acting on the mass are its weight and the tension supplied by the rope. Thus,

$$\begin{aligned} F_{\text{net}} &= T - w = 0 \\ \implies T &= w = mg. \end{aligned}$$

With this, (neglecting the mass of the rope), we see that the tension would be

$$\begin{aligned} T &= mg = (5.00\text{kg})(9.8\text{m/s}^2) \\ &= 49\text{N}. \end{aligned}$$

**Observation:** If we cut the rope and insert a spring, the spring would extend a length corresponding to a force of 49.0 N, providing a direct observation and measure of the tension force in the rope.

- **When to use normal force:**

- **Contact Between Surfaces:** Use normal force in scenarios where two surfaces are in contact, such as an object resting on a table, a book lying on a shelf, or a box pushed against a wall.

- **Supporting Weight:** If an object is supported by a surface (preventing it from falling due to gravity), there's a normal force exerted by the surface on the object, equal and opposite to the component of the object's weight perpendicular to the surface.
- **Inclined Planes:** On an inclined plane, normal force acts perpendicular to the surface, supporting the component of the object's weight perpendicular to the plane. It's crucial for calculating the net force acting along the plane.

- **When to not use normal force:**

- **No Direct Contact:** In scenarios where there's no direct physical contact between surfaces (e.g., objects in free fall, satellites in orbit, or a mass hanging from a string), normal force does not apply.
- **Tension-Dominated Problems:** In problems primarily involving tension (such as a tightrope walker, a hanging pendulum, or objects connected by ropes over pulleys), the focus is on tension forces rather than normal forces.

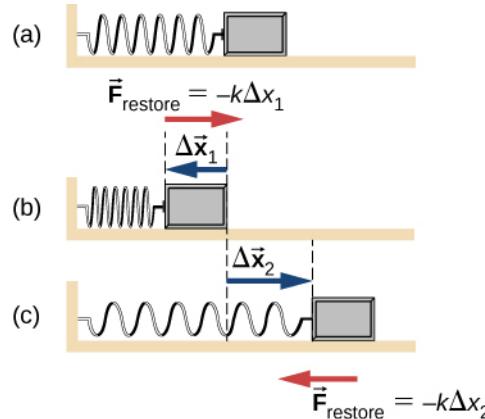
- **Tension T created when a perpendicular force ( $F_{\perp}$ ) is exerted at the middle of a flexible connector:**

$$T = \frac{F_{\perp}}{2 \sin(\theta)}.$$

The angle between the horizontal and the bent connector is represented by  $\theta$ . In this case,  $T$  becomes large as  $\theta$  approaches zero. Even

- **Hooke's law:** A spring is a special medium with a specific atomic structure that has the ability to restore its shape, if deformed. To restore its shape, a spring exerts a restoring force that is proportional to and in the opposite direction in which it is stretched or compressed. This is the statement of a law known as Hooke's law, which has the mathematical form

$$\vec{F} = -k\vec{x}.$$



A spring exerts its force proportional to a displacement, whether it is compressed or stretched.

- The spring is in a relaxed position and exerts no force on the block.
- The spring is compressed by displacement  $\Delta\vec{x}_1$  of the object and exerts restoring force  $-k\Delta\vec{x}_1$ .

- (c) The spring is stretched by displacement  $\vec{\Delta x}_2$  of the object and exerts restoring force  $-k\vec{\Delta x}_2$ .

- **Force on an inclined plane practice:** What force (in N) must be applied to a 250.0 kg crate on a frictionless plane inclined at  $30^\circ$  to cause an acceleration of  $7.6 \text{ m/s}^2$  up the plane? (Enter the magnitude.)

First, let's draw a diagram

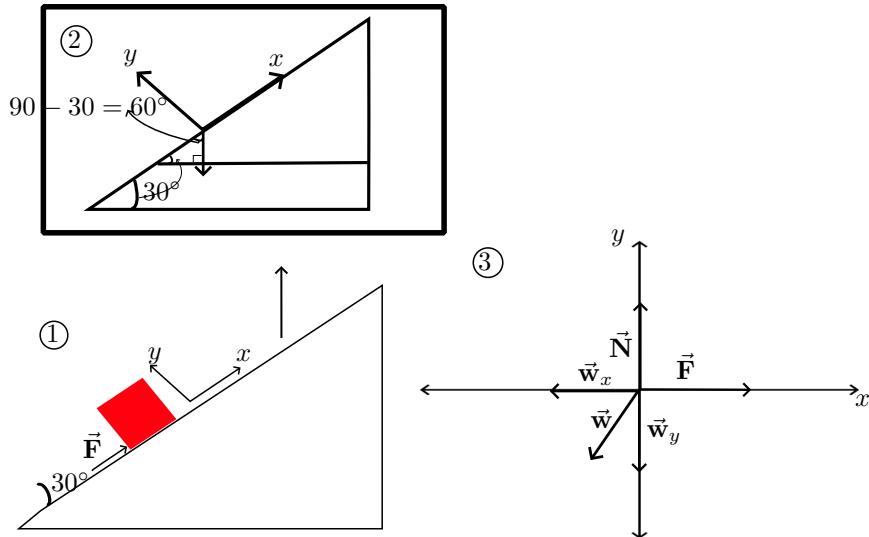


Figure 3 shows how we set up our coordinate axis, and figure 2 shows how we find the angle between the weight vector and the negative x-axis

To find the Force vector  $\vec{F}$ , we need to first find the net force in the horizontal direction. Since there is seemingly no vertical acceleration, we deduce that the net force must be zero. We find

$$W_x = \vec{W} \cos(60^\circ) = mg \cos(60^\circ)$$

Thus the net force in the horizontal direction is given by

$$\begin{aligned} F_{\text{net},x} &= \vec{F} - W_x = ma \\ &\implies \vec{F} - wg \cos(60^\circ) = ma \\ &\implies \vec{F} = ma + mg \cos(60^\circ) \\ &\therefore \vec{F} = (250)(7.6) + (250)(9.8) \cos(60^\circ) \\ &= 3125. \end{aligned}$$

- **Average frictional force when stopping in car:**

$$|\vec{F}_s| = m|(\vec{a})_{\text{avg}}|.$$

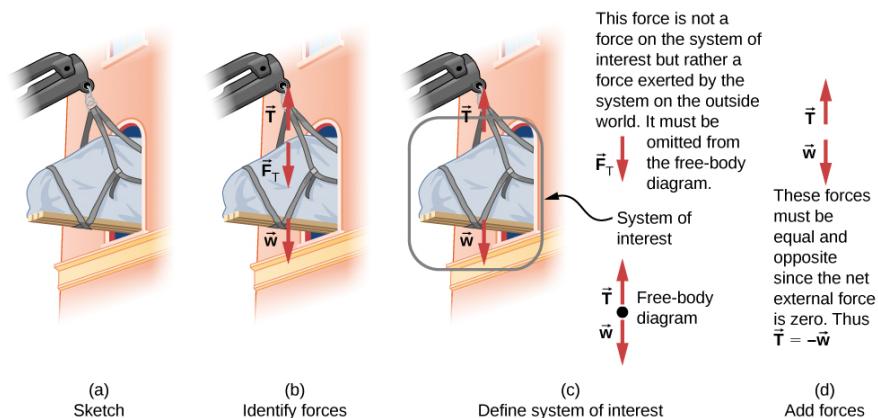
## 2.6 Chapter 6: Applications of Newton's Laws

### 2.6.1 Definitions and theorems

- **Problem solving / Omitting forces :**

- Identify the physical principles involved by listing the givens and the quantities to be calculated.
- Sketch the situation, using arrows to represent all forces.
- Determine the system of interest. The result is a free-body diagram that is essential to solving the problem.
- Apply Newton's second law to solve the problem. If necessary, apply appropriate kinematic equations from the chapter on motion along a straight line.
- Check the solution to see whether it is reasonable.

Suppose we want to lift a grand piano into a second-story apartment. Consider the sketch



In this case we want to find the tension of the rope, so the system of interest is the piano. Thus, the force that the piano exerts on the rope is a force by the system and not of use. We only care about external forces for whatever system we are analyzing

- **Weight and normal force are external:**

- **When tensions are equal:**

- **Single Rope with No Mass and no friction:** In a system where a massless rope passes over a frictionless pulley connecting two masses, the tension throughout the rope is constant if the pulley is ideal (massless and frictionless) and the rope is massless. This is because the rope cannot exert any force by itself, and there's no mechanism (like friction or pulley mass) to change the tension in different parts of the rope.
- **Static Equilibrium:** In a scenario where the system is in static equilibrium (not moving), and if we ignore the mass of the rope or the friction in the pulley, the tension on either side of the pulley must be equal to maintain equilibrium. This is because, in equilibrium, the sum of forces in any direction must be zero, and thus the forces (tensions) pulling on either side of the pulley must balance out.

- **When Tensions are Not Equal:**

- **Rope with Mass::**

- \* If the rope has mass, the tension varies along the length of the rope. The tension is higher closer to where the heavier load is applied because it has to support more weight (including the weight of the rope itself).

- **Multiple Pulleys or Complex Systems::**

- \* In systems involving multiple pulleys or complex arrangements, tension can vary significantly throughout the system. Each segment of the rope can have different tensions due to varying angles of pull, different masses being lifted, or the mechanical advantage created by the pulleys.

- **Varying Angle Measures::**

- \* When forces are applied at different angles, especially in systems involving pulleys or objects being pulled at angles, the tension in the rope or cable can differ based on how the components of the forces contribute to the tension.

- **Acceleration::** In systems where the masses are accelerating, the tension in the rope can differ if the rope passes over a pulley that changes the direction of the tension force. If one mass is accelerating downward, it might cause the rope on its side to have more tension compared to the other side, especially if the masses or forces acting on them are not symmetrical.

- **Non-Ideal Conditions::** When the rope has mass or the pulley has friction, the tension can vary along the rope. For example, the part of the rope supporting a heavier mass will have more tension compared to the side with a lighter mass. Similarly, friction in the pulley can cause a difference in tension on either side because it requires additional force to overcome the friction, leading to higher tension on one side.

- **Friction:** is a force that opposes relative motion between systems in contact.

- **Sliding friction:** is parallel to the contact surfaces between systems and is always in a direction that opposes motion or attempted motion of the systems relative to each other.

- **Static friction:** If two systems are in contact and stationary relative to one another, then the friction between them is called static friction.

- **Kinetic friction:** If two systems are in contact and moving relative to one another, then the friction between them is called kinetic friction.

- **Magnitude of static friction:** The magnitude of static friction  $f_s$  is

$$f_s \leq \mu_s N \quad (27)$$

where  $\mu_s$  is the coefficient of static friction and  $N$  is the magnitude of the normal force.

- Static friction is a responsive force that increases to be equal and opposite to whatever force is exerted, up to its maximum limit. Once the applied force exceeds  $f_s(\max)$ , the object moves. Thus,

$$f_s(\max) = \mu_s N.$$

- **Magnitude of kinetic friction:** The magnitude of kinetic friction  $f_k$  is given by

$$f_k = \mu_k N \quad (28)$$

where  $\mu_k$  is the coefficient of kinetic friction.

- the coefficients of kinetic friction are less than their static counterparts.:
- Coefficient of friction is a unitless quantity with a magnitude usually between 0 and 1.0. The actual value depends on the two surfaces that are in contact.:
- neither formula is accurate for lubricated surfaces or for two surfaces sliding across each other at high speeds.:
- **Friction on an inclined plane:** The basic physics is the same. We usually generalize the sloping surface and call it an inclined plane but then pretend that the surface is flat.
- **Angular velocity (radial acceleration),** is the vector quantity  $\omega$ , as we know the term angular frequency refers to the scalar quantity  $\omega = \frac{2\pi}{T}$  :, where  $T$  is the time it takes to complete one revolution, and  $\omega$  is given in rads/sec. This acceleration acts along the radius of the curved path and is thus also referred to as a radial acceleration.
- **Centripetal acceleration in terms of angular velocity:**

$$a_c = r\omega^2.$$

- Any net force causing uniform circular motion is called a centripetal force.: The direction of a centripetal force is toward the center of curvature,
- The direction of a centripetal force is toward the center of curvature, the same as the direction of centripetal acceleration.:
- Equations for centripetal force:

$$\begin{aligned} F_c &= ma_c \\ F_c &= m \frac{v^2}{r} \\ F_c &= mr\omega^2. \end{aligned}$$

- Centripetal force  $\vec{F}_c$  is always perpendicular to the path and points to the center of curvature,:
- Friction in regards to cars and tires:

- **Tires Rolling Without Slipping:** When a car is moving normally, and its tires are rolling without slipping, the friction between the tires and the road is static friction. Despite the car moving, the point of the tire that contacts the road is momentarily at rest relative to the road surface. This static friction is what allows the car to start moving from rest, stop, and turn without skidding. It's also the force that propels the car forward; the engine exerts a torque on the wheels, and it's the static frictional force at the contact patch of the tire that moves the car forward.
- **Tires Slipping or Skidding:** When the tires are slipping or skidding on the road surface, the friction between the tires and the road is kinetic (sliding) friction. This occurs when the force applied (for example, through braking or accelerating too hard) exceeds the maximum static frictional force that can be developed between the tire and the road. Kinetic friction also acts when a car is drifting, where the tires are intentionally made to slip sideways.

**Note:** The key to understanding why static friction is involved with a moving car lies in the behavior of the tires in contact with the road. When a car moves and its tires are rolling without slipping, the point of contact between each tire and the road does not slide across the road surface. Instead, it momentarily "sticks" to the road before lifting off again as the tire rotates. This means that, at any given instant, the part of the tire touching the road is stationary relative to the road surface. The force preventing the tire from slipping and allowing it to roll is provided by static friction.

- **Finding theta for ideal banking (frictionless curve):**

$$\theta = \tan^{-1} \left( \frac{v^2}{rg} \right).$$

- **Drag force always opposes the motion of an object.:**
- **the magnitude of the drag force  $F_D$  is proportional to the square of the speed of the object.:**

$$F_D \propto v^2.$$

When taking into account other factors, this relationship becomes

$$F_D = \frac{1}{2} C \rho A v^2.$$

where  $C$  is the drag coefficient,  $A$  is the area of the object facing the fluid, and  $\rho$  is the density of the fluid.

This equation can also be written in a more generalized fashion as

$$F_D = bv^n.$$

Where  $b$  is a constant equivalent to  $0.5C\rho A$

**Note:** The value of the drag coefficient  $C$  is determined empirically,

- **What is terminal velocity:** Terminal velocity is the constant speed that a freely falling object eventually reaches when the resistance of the medium through which it is falling prevents further acceleration.
- **Terminal Velocity:** Consider a skydiver falling through air under the influence of gravity. Once air resistance becomes equal to the force of gravity, the net force will be zero and there will be no acceleration. At this point, the person's velocity remains constant and we say that the person has reached his terminal velocity. We have

$$mg = \frac{1}{2} C \rho A v^2$$

$$v_T = \sqrt{\frac{2mg}{\rho CA}}.$$

- **The density of air is approximately:**

$$\rho = 1.21 \text{ kg/m}^3.$$

- **Stoke's Law:** For a spherical object falling in a medium, the drag force is

$$F_s = 6\pi r \eta v.$$

where  $r$  is the radius of the object,  $\eta$  is the viscosity of the fluid, and  $v$  is the object's velocity.

- A car approaches the top of a hill that is shaped like a vertical circle with a radius of 55.0m. What is the fastest speed that the car can go over the hill without losing contact with the ground?. In this case, we have  $F_y = mg - N = ma_c \therefore$  But once the car loses contact with the road (goes past its max velocity to stay on the road), normal force goes to zero because the car would no longer be in contact with the road. Thus, the maximum velocity would be

$$\begin{aligned} F_y &= mg = a_c \\ \implies mg &= m \frac{v^2}{r} \\ \implies v &= \sqrt{g \cdot r}. \end{aligned}$$

## 2.7 Chapter 7: Work and Kinetic energy

### 2.7.1 Definitions and theorems

- **work:** is done on an object when energy is transferred to the object. In other words, work is done when a force acts on something that undergoes a displacement from one position to another. Forces can vary as a function of position, and displacements can be along various paths between two points.
  - **Increment of work  $dW$ :** We first define the increment of work  $dW$  done by a force  $\vec{F}$  acting through an infinitesimal displacement  $d\vec{r}$  : as the dot product of these two vectors:
- $$dW = \vec{F} \cdot d\vec{r} = |\vec{F}| |\vec{r}| \cos \theta. \quad (29)$$
- **Work done by a force:** We add up the contributions for infinitesimal displacements, along a path between two positions, to get the total work.

The work done by a force is the integral of the force with respect to displacement along the path of the displacement:

$$W_{AB} = \int_{\text{path AB}} \vec{F} \cdot d\vec{r}.$$

- **SI unit for work:** The units of work are units of force multiplied by units of length, which in the SI system is newtons times meters,  $N \cdot m$ . This combination is called a joule, abbreviated  $J$
- **American unit for work:** States, the unit of force is the pound (lb) and the unit of distance is the foot (ft), so the unit of work is the foot-pound (ft-lb).
- **Work done by constant forces and contact forces:** The simplest work to evaluate is that done by a force that is constant in magnitude and direction. In this case, we can factor out the force; the remaining integral is just the total displacement, which only depends on the end points A and B, but not on the path between them:

$$\begin{aligned} W_{AB} &= \vec{F} \cdot \int_A^B d\vec{r} \\ &= \vec{F} \cdot (\vec{r}_B - \vec{r}_A) \\ &= |\vec{F}| |\vec{r}_B - \vec{r}_A| \cos(\theta). \end{aligned}$$

That is, work is just  $\vec{F} \cdot \vec{d} = Fd \cos(\theta)$

- **Work of the normal force.** If the object being displaced never leaves the surface, then the displacement  $d\vec{r}$  : is tangent to the surface. Since these two vectors are orthogonal, their dot product is zero. Thus, we have

$$dW_N = \vec{N} \cdot d\vec{r} = \vec{0}.$$

**Note:** If the displacement  $d\vec{r}$  did have a relative component perpendicular to the surface, the object would either leave the surface or break through it, and there would no longer be any normal contact force.

- **"Tangent to the surface":** When the displacement is tangent to the surface, it means the movement of the object is along the surface

**Note:** This is the same as saying the displacement is parallel to the surface.

- **Work done by kinetic friction:** If the magnitude of  $\vec{f}_k$  is constant (as it would be if all the other forces on the object were constant), then the work done by friction is

$$W_{\text{fr}} = \int_A^B \vec{f}_k \cdot d\vec{r} = -f_k \int_A^B |d\vec{r}| = -f_k |l_{AB}|,$$

where  $|l_{AB}|$  is the path length on the surface.

**Note:** kinetic friction  $\vec{f}_k$  is opposite to  $d\vec{r}$  relative to the surface, so the work done by kinetic friction is negative.

- **Work done by static friction:** The force of static friction does no work in the reference frame between two surfaces because there is never displacement between the surfaces.

As an external force, static friction can do work. Static friction can keep someone from sliding off a sled when the sled is moving and perform positive work on the person.

- **the work done against a force is the negative of the work done by the force.:**
- **Work done by gravity:** The work done by a constant force of gravity on an object depends only on the object's weight and the difference in height through which the object is displaced.

$$W_{\text{grav},AB} = -mg\hat{\mathbf{j}} \cdot (\vec{r}_B - \vec{r}_A) = -mg(y_B - y_A).$$

**Note:** Gravity does negative work on an object that moves upward ( $y_B > y_A$ ), or, in other words, you must do positive work against gravity to lift an object upward. Alternately, gravity does positive work on an object that moves downward ( $y_B < y_A$ ), or you do negative work against gravity to "lift" an object downward, controlling its descent so it doesn't drop to the ground. ("Lift" is used as opposed to "drop".)

- **Work Done by Forces that Vary:** In general, forces may vary in magnitude and direction at points in space, and paths between two points may be curved. The infinitesimal work done by a variable force can be expressed in terms of the components of the force and the displacement along the path,

$$dW = F_x dx + F_y dy + F_z dz.$$

- **Springs, Hooke's law, and work (perfectly elastic spring):** The force exerted by a perfectly elastic spring is governed by Hooke's law  $\vec{F} = -k\Delta\vec{x}$ , where  $k$  is the spring constant and  $\Delta\vec{x} = \vec{x} - \vec{x}_{\text{eq}}$  is the displacement from the spring's equilibrium position. This equilibrium position aligns with the spring's unstretched position in the absence of other forces or when they are neutralized. For work calculation by the spring force, the x-axis is aligned along the spring's length, with the origin at  $x_{\text{eq}} = 0$ . Thus, positive  $x$  denotes stretching and negative  $x$  compression. The work done by the spring force as  $x$  varies from  $x_A$  to  $x_B$  is calculated under this framework.

The work done by the spring from  $x_A$  to  $x_B$  is given by:

$$\begin{aligned} W_{\text{spring},AB} &= \int_A^B F_x dx = -k \int_A^B x dx \\ &= -\frac{k}{2} [x^2]_A^B = -\frac{1}{2}k(x_B^2 - x_A^2). \end{aligned}$$

**Note:** Notice that  $W_{AB}$  Depends only on the starting and ending points,  $A$  and  $B$ , and is independent of the actual path between them, as long as it starts at  $A$  and ends at  $B$ . That is, the actual path could involve going back and forth before ending.

- **Kinetic energy definition:** The kinetic energy of an object is the form of energy that it possesses due to its motion
- **Kinetic energy of a particle with mass  $m$ :**

$$K = \frac{1}{2}mv^2.$$

- **Kinetic energy for a system of particles:**

$$k = \sum \frac{1}{2}mv^2.$$

- **Kinetic energy in terms of a particles momentum (single particle):** that just as we can express Newton's second law in terms of either the rate of change of momentum or mass times the rate of change of velocity, so the kinetic energy of a particle can be expressed in terms of its mass and momentum  $\vec{p} = m\vec{v}$ , instead of its mass and velocity. Since  $v = \frac{p}{m}$ , we see that

$$k = \frac{1}{2} \frac{p^2}{m} = \frac{p^2}{2m}.$$

- **Units of kinetic energy:** The units of kinetic energy are also the units of force times distance, which are the units of work, or joules.
- **Work-energy theorem:** The net work done on a particle equals the change in the particle's kinetic energy:

$$W_{\text{net}} = K_B - K_A. \quad (30)$$

- **Average power:** we first define average power as the work done during a time interval, divided by the interval,

$$\bar{P} = \frac{\Delta W}{\Delta t}.$$

- **Instantaneous power (or just plain power):** Power is defined as the rate of doing work, or the limit of the average power for time intervals approaching zero,

$$P = \frac{dW}{dt}.$$

- **Work in constant power:** If the power is constant over a time interval, the average power for that interval equals the instantaneous power, and the work done by the agent supplying the power is

$$W = P\Delta t.$$

- **Work in varying power:** If the power during an interval varies with time, then the work done is the time integral of the power,

$$W = \int P dt.$$

- **Unit for power:** We can also define power as the rate of transfer of energy. Work and energy are measured in units of joules, so power is measured in units of joules per second, which has been given the SI name watts,

$$1J/s = 1W.$$

- **Horsepower:**

$$1hp = 746W.$$

- **Power in terms of forces and velocity:**

$$\begin{aligned} P &= \frac{dW}{dt} = \frac{\vec{F} \cdot d\vec{r}}{dt} = \vec{F} \cdot \left( \frac{d\vec{r}}{dt} \right) \\ &= \vec{F} \cdot \vec{v}. \end{aligned}$$

- **Finding angle when given grade:** Suppose we want to find the angle associated with a 15% grade. This means for every unit increase in the x-direction, we move up 15% or 0.15.



Thus, we see

$$\begin{aligned} \tan(\theta) &= \frac{0.15}{1} \\ \theta &= \tan^{-1}(0.15) \\ &\approx 8.53^\circ. \end{aligned}$$

## 2.8 Chapter 8: Potential Energy and Conservation of Energy

### 2.8.1 Definitions and theorems

- **Potential energy definition:** Potential energy is the energy an object posses by virtue of its position in a field
- **Potential energy difference:** We define the difference of potential energy from point A to point B as the negative of the work done:

$$\Delta U_{AB} = U_B - U_A = -W_{AB}.$$

- **Potential energy:** we need to define potential energy at a given position in such a way as to state standard values of potential energy on their own, rather than potential energy differences. We do this by rewriting the potential energy function in terms of an arbitrary constant,

$$\Delta U = U(\vec{r}) - U(\vec{r}_0).$$

**Note:** the lowest height in a problem is usually defined as zero potential energy, or if an object is in space, the farthest point away from the system is often defined as zero potential energy. Then, the potential energy, with respect to zero at  $\vec{r}_0$ , is just  $U(\vec{r})$

- **Change in kinetic energy with no friction or air resistance:** As long as there is no friction or air resistance, the change in kinetic energy of the football equals negative of the change in gravitational potential energy of the football. This can be generalized to any potential energy:

$$\Delta K_{AB} = -\Delta U_{AB}.$$

- **Gravitational potential energy function:**

$$U(y) = mgy + C.$$

**Note:** You can choose the value of the constant. However, for solving most problems, the most convenient constant to choose is zero for when  $y = 0$ , which is the lowest vertical position in the problem.

- **Elastic potential energy:**

$$U(x) = \frac{1}{2}kx^2 + C.$$

**Note:** If the spring force is the only force acting, it is simplest to take the zero of potential energy at  $x = 0$ , when the spring is at its unstretched length. Then, the constant is zero. (Other choices may be more convenient if other forces are acting.)

- **Conservative Force:** Conservative force, in physics, any force, such as the gravitational force between Earth and another mass, whose work is determined only by the final displacement of the object acted upon. The total work done by a conservative force is independent of the path resulting in a given displacement and is equal to zero when the path is a closed loop

- **Non-Conservative Force:** Non-conservative forces are dissipative forces such as friction or air resistance. These forces take energy away from the system as the system progresses, energy that you can't get back. These forces are path dependent; therefore it matters where the object starts and stops.
- **Work done by a conservative force:** The work done by a conservative force is independent of the path; in other words, the work done by a conservative force is the same for any path connecting two points:

$$W_{AB,\text{path-1}} = \int_A^B \text{on path-1} \vec{F}_{\text{cons}} \cdot d\vec{r} = W_{AB,\text{path-2}} = \int_A^B \text{on path-2} \vec{F}_{\text{cons}} \cdot d\vec{r}.$$

- **Work done by a non-conservative force:** The work done by a non-conservative force depends on the path taken.
- **When is a force conservative:** A force is conservative if the work it does around any closed path is zero:

$$W_{\text{closed path}} = \oint \vec{F}_{\text{cons}} \cdot d\vec{r} = 0.$$

- **Proving whether or not a force is conservative:** One answer is that the work done is independent of path if the infinitesimal work  $\vec{F} \cdot d\vec{r}$  is an exact differential, the way the infinitesimal net work was equal to the exact differential of the kinetic energy,  $dW_{\text{net}} = m\vec{v} \cdot d\vec{v} = d(\frac{1}{2}mv^2)$ ,

There are mathematical conditions that you can use to test whether the infinitesimal work done by a force is an exact differential, and the force is conservative. These conditions only involve differentiation and are thus relatively easy to apply. In two dimensions, the condition for  $\vec{F} \cdot d\vec{r} = F_x dx + F_y dy$  to be an exact differential is

$$\frac{dF_x}{dy} = \frac{dF_y}{dx}.$$

- **Non-conservative forces don't have potential energy:** we note that non-conservative forces do not have potential energy associated with them because the energy is lost to the system and can't be turned into useful work later. So there is always a conservative force associated with every potential energy.
- **The infinitesimal increment of potential energy:** force. The infinitesimal increment of potential energy is the dot product of the force and the infinitesimal displacement,

$$\begin{aligned} dU &= -\vec{F} \cdot d\vec{\ell} = -F_\ell d\ell \\ \implies F_\ell &= -\frac{dU}{d\ell}. \end{aligned}$$

Here, we chose to represent the displacement in an arbitrary direction by  $d\vec{\ell}$  so as not to be restricted to any particular coordinate direction.

In words, the component of a conservative force, in a particular direction, equals the negative of the derivative of the corresponding potential energy, with respect to a displacement in that direction.

- **Force with derivative of potential energy: One dimension:**

$$\vec{F} = f_x \hat{i} = -\frac{dU}{dx} \hat{i}.$$

- **Force with derivative of potential energy: Two dimensions:**

$$\vec{F} = f_x \hat{i} + f_y \hat{j} = -\frac{dU}{dx} \hat{i} + -\frac{dU}{dy} \hat{j}.$$

- **Mechanical energy:** mechanical energy is simply all the energy that an object has because of its motion and its position.

$$E = K + U.$$

Where  $E$  denotes the mechanical energy

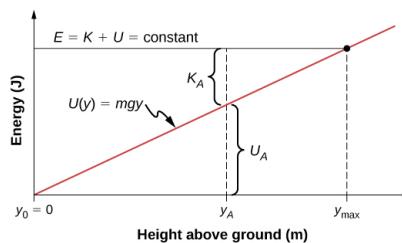
- **Conservation of Energy:** The mechanical energy  $E$  of a particle stays constant unless forces outside the system or non-conservative forces do work on it, in which case, the change in the mechanical energy is equal to the work done by the non-conservative forces:

$$W_{nc,AB} = \Delta(K + U)_{AB} = \Delta E_{AB}.$$

- **Position  $x$  as a function of time  $t$  with potential energy:**

$$t = \int_{x_0}^x \frac{dx}{\sqrt{2[E - U(x)]/m}}.$$

- **Potential energy diagram example:**



The line at energy  $E$  represents the constant mechanical energy of the object, whereas the kinetic and potential energies,  $K_A$  and  $U_A$ , are indicated at a particular height  $y_A$ . You can see how the total energy is divided between kinetic and potential energy as the object's height changes.

- **Max height of a particle:** Using the fact that kinetic energy must be positive, we can write  $K = E - U \geq 0 \implies U \leq E$ . If we use the gravitational potential energy reference point of zero at  $y_0$  :, we can rewrite the gravitational potential energy  $U$  as  $mgy$ . Solving for  $y$  results in

$$y \leq \frac{E}{mg} = y_{\max}.$$

- **Max speed of a particle:** At ground level  $y_0$  :, the potential energy is zero, and the kinetic energy and the speed are maximum.

$$U_0 = 0 = E - K_0$$

$$E = K_0 = \frac{1}{2}mv_0^2$$

$$v_0 = \pm \sqrt{\frac{2E}{m}}.$$

**Note:** The maximum speed  $\pm v_0$  gives the initial velocity necessary to reach  $y_{\max}$ , the maximum height, and  $-v_0$  represents the final velocity, after falling from  $y_{\max}$ .

- **Finding range of  $x$ :** Using the inequality  $0 \leq K \leq E$ , we can find the allowable range for  $x$  values. Considering the turning points, where all of the energy is potential, we have  $K = 0$  and  $U = E$ . For a elastic spring, we have

$$\begin{aligned}\frac{1}{2}kx^2 &= E \\ \implies x &= \pm\sqrt{\frac{2E}{k}}.\end{aligned}$$

- **Finding range of  $x$ : Example:** Suppose we have the function  $U(x) = 2(x^4 - x^2)$ , and  $E = -\frac{1}{4}$  ∵. To find the allowable range for  $x$  values, we use the condition

$$\begin{aligned}U + K &= E \implies K = E - U \geq 0 \\ \implies -\frac{1}{4} - 2(x^4 - x^2) &\geq 0 \\ \implies 2(x^4 - x^2) &\leq -\frac{1}{4} \\ \implies 2\left(\left(x^4 - \frac{1}{2}\right)^2 - \frac{1}{4}\right) &\leq -\frac{1}{4} \\ \implies \left(x^4 - \frac{1}{2}\right)^2 &\leq \frac{-\frac{1}{4} + \frac{1}{2}}{2} \\ \implies \left(x^4 - \frac{1}{2}\right)^2 &\leq \frac{1}{8} \\ \implies -\frac{1}{\sqrt{8}} + \frac{1}{2} \leq x^2 &\leq \frac{1}{\sqrt{8}} + \frac{1}{2}.\end{aligned}$$

This represents two allowed regions,  $x_p \leq x \leq x_R$  and  $-x_R \leq x \leq -x_p$ , where  $x_p = 0.38$  and  $x_R = 0.92$  (in meters).

- **Manipulating inequalities: Squared terms:** For  $a$  non-negative, we have

$$\begin{aligned}x^2 &\leq a \\ \implies -\sqrt{a} \leq x &\leq \sqrt{a}.\end{aligned}$$

We also have

$$\begin{aligned}a &\leq x^2 \leq b \quad (\text{Suppose}) \\ \implies \sqrt{a} &\leq x \leq \sqrt{b} \quad \text{and} \quad -\sqrt{b} \leq x \leq -\sqrt{a}.\end{aligned}$$

For  $a, b$  non-negative. The implication imposed by this is that  $x$  can be within either ranges.

- **Finding equilibrium points:** To find the equilibrium points we just need to find relative extrema. Relative maxima is deemed unstable while relative minima is deemed stable

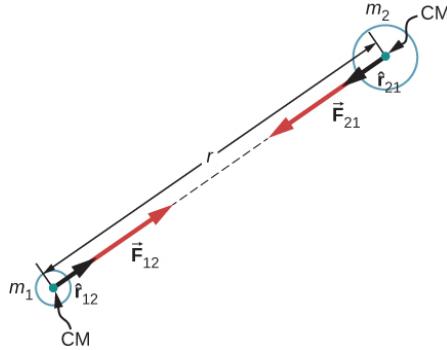
## 2.9 Chapter 13: Gravitation

### 2.9.1 Definitions and theorems

- **Newton's Law of Gravitation:** Newton's law of gravitation can be expressed as

$$\vec{F}_{12} = G \frac{m_1 m_2}{r^2} \hat{r}_{12} \quad (31)$$

where  $\vec{F}_{12}$  is the force on object 1 exerted by object 2 and  $\hat{r}_{12}$  is a unit vector that points from object 1 toward object 2.



**Note:** Notice how we take  $r$  to be the distance between the center of mass for both objects.

- **Universal gravitational constant:** The constant  $G$  is the previous equation is called the *universal gravitational constant*: and cavendish determined it to be

$$G = 6.67 \times 10^{-11} N \cdot \frac{m^2}{kg^2}.$$

- **the law of gravitation applies to spherically symmetrical objects:**, where the mass of each body acts as if it were at the center of the body.
- **Weight with NLOG:** We now know that this force is the gravitational force between the object and Earth. If we substitute  $mg$  for the magnitude of  $\vec{F}_{12}$  in Newton's law of universal gravitation,  $m$  for  $m_1$ , and  $M_E$  for  $m_2$ , we obtain the scalar equation

$$mg = G \frac{m M_E}{r^2} \quad (32)$$

where  $r$  is the distance between the centers of mass of the object and Earth.

**Note:** The average radius of Earth is about 6370 km. Hence, for objects within a few kilometers of Earth's surface, we can take  $r = R_E$

- **Calculating  $g$ :**, dividing the previous equation by  $m$ , we get

$$g = G \frac{M_E}{r^2} \quad (33)$$

- **Radius of the earth**  $R_E$ :

$$R_E = 6.37 \times 10^6 \text{ m.}$$

- **Mass of the earth:**

$$M_E = 5.95 \times 10^{24} \text{ kg.}$$

- **$g$  on the moon's surface:**

$$g = 1.6 \text{ m/s}^2.$$

- **The gravitational field caused by mass  $M$ :**

$$\vec{g} = G \frac{M}{r^2} \hat{r}.$$

We identify this vector field  $\vec{g}$  as the *gravitational field caused by mass  $M$*

- **Weight at the equator with Earth's spin:**

$$\sum F = F_s - mg = ma_c.$$

Where  $a_c = -\frac{v^2}{r}$

- **Results Away from the Equator:** At any other latitude  $\lambda$ , the situation is more complicated. The centripetal acceleration is directed toward point P in the figure, and the radius becomes  $r = R_E \cos \lambda$ . The vector sum of the weight and  $\vec{F}_s$  must point toward point P, hence  $\vec{F}_s$  no longer points away from the center of Earth. (The difference is small and exaggerated in the figure.)



- **Gravity Away from the Surface:** Earlier we stated that the law of gravitation applies to spherically symmetrical objects, where the mass of each body acts as if it were at the center of the body.  $(\vec{F}_{12} = G \frac{m_1 m_2}{r^2} \hat{r}_{12})$  and  $(g = G \frac{M_E}{r^2})$ .

both equations are valid only for values of  $r \geq R_E$ . For  $r < R_E$  these equations are not valid. However, we can determine  $g$  for these cases using a principle that comes from Gauss's law,

A consequence of Gauss's law, applied to gravitation, is that only the mass *within*  $r$  contributes to the gravitational force. Also, that mass, just as before, can be considered to be located at the center. The gravitational effect of the mass *outside*  $r$  has zero net effect.

For a spherical planet with constant density, the mass within  $r$  is the density times the volume within  $r$ . This mass can be considered located at the center. Replacing  $M_E$  with only the mass within  $r$ ,  $M = \rho \times (\text{volume of a sphere})$ , and  $R_E$  with  $r$ , we get

$$g = \frac{GM_E}{R_E^2} = G\rho \left( \frac{4}{3}\pi r^3 \right) \frac{1}{r^2} = \frac{4}{3}G\rho\pi r.$$

**Note:** The value of  $g$ , and hence your weight, decreases linearly as you descend down a hole to the center of the spherical planet. At the center, you are weightless, as the mass of the planet pulls equally in all directions.

Actually, Earth's density is not constant, nor is Earth solid throughout.

- **Gravitational Potential Energy beyond Earth:** We return to the definition of work and potential energy to derive an expression that is correct over larger distances.

$$U = -\frac{GM_E m}{r}.$$

- **Conservation of Energy with Gravitation:**

$$\begin{aligned} E &= K_1 + U_1 = K_2 + U_2 \\ \implies &= \frac{1}{2}mv_1^2 - \frac{GMm}{r_1} = \frac{1}{2}mv_2^2 - \frac{GMm}{r_2}. \end{aligned}$$

**Note:** Note that we use  $M$ , rather than  $M_E$ , as a reminder that we are not restricted to problems involving Earth. However, we still assume that  $m \ll M$ . (For problems in which this is not true, we need to include the kinetic energy of both masses and use conservation of momentum to relate the velocities to each other. But the principle remains the same.)

- **Escape velocity:** *Escape velocity* is often defined to be the **minimum initial velocity** of an object that is required to escape the surface of a planet (or any large body like a moon) and never return. As usual, we assume no energy lost to an atmosphere, should there be any.

Consider the case where an object is launched from the surface of a planet with an initial velocity directed away from the planet. With the minimum velocity needed to escape, the object would just come to rest infinitely far away, that is, the object gives up the last of its kinetic energy just as it reaches infinity, where the force of gravity becomes zero. Since  $U \rightarrow 0$  as  $r \rightarrow \infty$

Thus, we find the escape velocity from the surface of an astronomical body of mass  $M$  and radius  $R$  by setting the total energy equal to zero. At the surface of the body, the object is located at  $r_1 = R$  and it has escape velocity  $v_1 = v_{\text{esc}}$ . It reaches  $r_2 = \infty$  with velocity  $v_2 = 0$ . Substituting the equation above and solving for  $v_{\text{esc}}$ , we find

$$v_{\text{esc}} = \sqrt{\frac{2GM}{R}}.$$

**Note:** Notice that  $m$  has canceled out of the equation. The escape velocity is the same for all objects, regardless of mass. Also, we are not restricted to the surface of the planet;  $R$  can be any starting point beyond the surface of the planet.

- **Not gravitationally Bound:** As noted earlier, we see that  $U \rightarrow 0$  as  $r \rightarrow \infty$ . If the total energy is zero, then as  $m$  reaches a value of  $r$  that approaches infinity,  $U$  becomes zero and so must the kinetic energy. Hence,  $m$  comes to rest infinitely far away from  $M$ . It has "just escaped"  $M$ . If the total energy is positive, then kinetic energy remains at  $r = \infty$  and certainly  $m$  does not return. When the total energy is zero or greater, then we say that  $m$  is not gravitationally bound to  $M$ .
- **Gravitationally Bonud:** On the other hand, if the total energy is negative, then the kinetic energy must reach zero at some finite value of  $r$ , where  $U$  is negative and equal to the total energy. The object can never exceed this finite distance from  $M$ , since to do so would require the kinetic energy to become negative, which is not possible. We say  $m$  is gravitationally bound to  $M$ .
- **Speed of orbit:**

$$v_{\text{orbit}} = \sqrt{\frac{GM_E}{r}}.$$

- **Period of a circular orbit:**

$$T = 2\pi \sqrt{\frac{r^3}{GM_E}}.$$

- **Total energy for a circular orbit:**

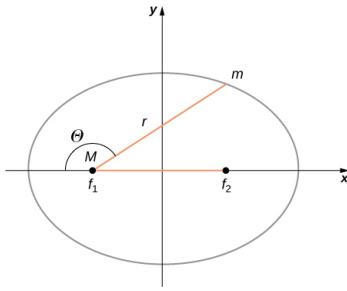
$$E = K + U = -\frac{GM_E m}{2r}.$$

For circular orbits, the magnitude of the kinetic energy is exactly one-half the magnitude of the potential energy.

- **Keplars first law:** Kepler's first law states that every planet moves along an ellipse, with the Sun located at a focus of the ellipse.
- **Form of all conics:** There are four different conic sections, all given by the equation

$$\frac{a}{r} = 1 + e \cos(\theta).$$

The variables  $r$  and  $\theta$  are shown in the figure below. In the case of an ellipse. The constants  $a$  and  $e$  are determined by the total energy and angular momentum of the satellite at a given point. The constant  $e$  is called the eccentricity. The values of  $a$  and  $e$  determine which of the four conic sections represents the path of the satellite.



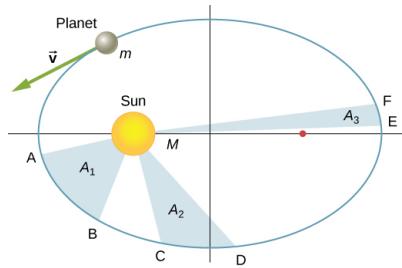
Every path taken by  $m$  is one of the four conic sections: a circle or an ellipse for bound or closed orbits, or a parabola or hyperbola for unbounded or open orbits.

- **Total energy for an elliptical orbit:**

$$E = -\frac{GmM_s}{2a}.$$

Where  $M_s$  is the mass of the sun and  $a$  is the semi-major axis

- **Keplar's second law:** Kepler's second law states that a planet sweeps out equal areas in equal times, that is, the area divided by time, called the areal velocity, is constant.



The time it takes a planet to move from position  $A$  to  $B$ , sweeping out area  $A_1$ , is exactly the time taken to move from position  $C$  to  $D$ , sweeping area  $A_2$ , and to move from  $E$  to  $F$ , sweeping out area  $A_3$ . These areas are the same:  $A_1 = A_2 = A_3$ .

- **semi-major axis:**

$$a = \frac{1}{2}(\text{aphelion} + \text{perihelion}).$$

- **Keplar's third law:** Kepler's third law states that the square of the period is proportional to the cube of the semi-major axis of the orbit.

$$T^2 = \frac{4\pi^2}{GM}a^3.$$

- **Schwarzschild radius:** For any mass  $M$ , if that mass were compressed to the extent that its radius becomes less than the Schwarzschild radius, then the mass will collapse to a singularity, and anything that passes inside that radius cannot escape. Once inside  $R_S$ , the arrow of time takes all things to the singularity. (In a broad mathematical sense, a singularity is where the value of a function goes to infinity. In this case, it is a point in space of zero volume with a finite mass. Hence, the mass density and gravitational energy become infinite.) The Schwarzschild radius is given by

$$R_S = \frac{2GM}{c^2}.$$

**Note:** The Schwarzschild radius is also called the event horizon of a black hole.

## 2.10 Chapter 9: Linear momentum and collisions

### 2.10.1 Definitions and Theorems

- **Momentum:** is a quantity of motion and is defined as the vector

$$\vec{p} = m\vec{v}.$$

So we see the momentum  $p$  of an object is the product of its mass and its velocity:

**Note:** kinetic energy. It is perhaps most useful when determining whether an object's motion is difficult or easy to change over a short time interval.

- **Impulse:** The product of a force and a time interval (over which that force acts) is called impulse, and is given the symbol  $\vec{J}$ . Formally we write

Let  $\vec{F}(t)$  be the force applied to an object over some differential time interval  $dt$ . The resulting impulse on the object is defined as

$$d\vec{J} = \vec{F}(t)dt.$$

The total impulse over the interval  $t_f - t_i$  is

$$\vec{J} = \int_{t_1}^{t_2} d\vec{J} \quad \text{or} \quad \int_{t_1}^{t_2} \vec{F}(t) dt.$$

- **Impulse with the mean value theorem:** To calculate the impulse using Equation 9.3, we need to know the force function  $F(t)$ , which we often don't. However, a result from calculus is useful here: Recall that the average value of a function over some interval is calculated by

$$f_{\text{ave}}(x) = \frac{1}{\Delta x} \int_{x_i}^{x_f} f(x) dx$$

where  $\Delta x = x_f - x_i$ . Applying this to the time-dependent force function, we obtain

$$\vec{F}_{\text{ave}} = \frac{1}{\Delta t} \int_{t_i}^{t_f} \vec{F}(t) dt.$$

Therefore, from the equation above

$$\vec{J} = \vec{F}_{\text{ave}} \Delta t.$$

**Note:** The idea here is that you can calculate the impulse on the object even if you don't know the details of the force as a function of time; you only need the average force. In fact, though, the process is usually reversed: You determine the impulse (by measurement or calculation) and then calculate the average force that caused that impulse.

- **Impulse again:** To calculate the impulse, a useful result follows from writing the force in Equation 9.3 as  $\vec{F}(t) = m\vec{a}(t)$ :

$$\begin{aligned} \vec{J} &= \int_{t_i}^{t_f} \vec{F}(t) dt = m \int_{t_i}^{t_f} \vec{a}(t) dt = m[\vec{v}(t_f) - \vec{v}_i]. \\ &\therefore \vec{J} = m\Delta\vec{v}. \end{aligned}$$

- **impulse-momentum theorem:** Because  $m\vec{v}$  is the momentum of a system,  $m\Delta\vec{v}$  is the change of momentum  $\Delta\vec{p}$ . This gives us the following relation,

An impulse applied to a system changes the system's momentum, and that change of momentum is exactly equal to the impulse that was applied:

$$\vec{J} = \Delta \vec{p}.$$

- **Two crucial concepts in the impulse-momentum theorem:**

- Impulse is a vector quantity; an impulse of, say,  $-(10 \text{ N} \cdot \text{s})\hat{i}$  is very different from an impulse of  $+(10 \text{ N} \cdot \text{s})\hat{i}$ ; they cause completely opposite changes of momentum.
- An impulse does not cause momentum; rather, it causes a change in the momentum of an object. Thus, you must subtract the initial momentum from the final momentum, and—since momentum is also a vector quantity—you must take careful account of the signs of the momentum vectors.

- **Problem solving: Impulse momentum theorem:**

- **Average force with momentum:**

$$\vec{F}_{\text{ave}} = \frac{\Delta \vec{p}}{\Delta t}.$$

- **Force with rate of change in momentum (Newton's second law in terms of momentum):**

$$\vec{F} = \frac{d\vec{p}}{dt}.$$

This says that the rate of change of the system's momentum (implying that momentum is a function of time) is exactly equal to the net applied force (also, in general, a function of time). This is, in fact, Newton's second law, written in terms of momentum rather than acceleration.

- **Newton's third law with momentum:**

$$m_1 \frac{d\vec{v}_1}{dt} = -m_2 \frac{d\vec{v}_2}{dt}.$$

- **Newton's third law with momentum if mass remains constant:** We can then pull the masses inside the derivatives (see equation above)

$$\frac{d}{dt}(m_1 \vec{v}_1) = -\frac{d}{dt}(m_2 \vec{v}_2).$$

and thus

$$\frac{d\vec{p}_1}{dt} = -\frac{d\vec{p}_2}{dt}.$$

This says that the rate at which momentum changes is the same for both objects. The masses are different, and the changes of velocity are different, but the rate of change of the product of  $m$  and  $\vec{v}$  are the same.

- **Newton's third law with momentum if mass remains constant (additive inverse version):**

$$\frac{d\vec{p}_1}{dt} + \frac{d\vec{p}_2}{dt} = 0.$$

This says that during the interaction, although object 1's momentum changes, and object 2's momentum also changes, these two changes cancel each other out, so that the total change of momentum of the two objects together is zero.

Since the total combined momentum of the two objects together never changes, then we could write

$$\begin{aligned}\frac{d}{dt}(\vec{p}_1 + \vec{p}_2) &= 0 \\ \Rightarrow \vec{p}_1 + \vec{p}_2 &= \text{constant.}\end{aligned}$$

- **Conservation laws:** If the value of a physical quantity is constant in time, we say that the quantity is conserved.
- **Requirements for Momentum Conservation:** There is a complication, however. A system must meet two requirements for its momentum to be conserved:
  - The mass of the system must remain constant during the interaction. As the objects interact (apply forces on each other), they may transfer mass from one to another; but any mass one object gains is balanced by the loss of that mass from another. The total mass of the system of objects, therefore, remains unchanged as time passes:

$$\left[ \frac{dm}{dt} \right]_{\text{system}} = 0.$$

- The net external force on the system must be zero. As the objects collide, or explode, and move around, they exert forces on each other. However, all of these forces are internal to the system, and thus each of these internal forces is balanced by another internal force that is equal in magnitude and opposite in sign. As a result, the change in momentum caused by each internal force is cancelled by another momentum change that is equal in magnitude and opposite in direction. Therefore, internal forces cannot change the total momentum of a system because the changes sum to zero. However, if there is some external force that acts on all of the objects (gravity, for example, or friction), then this force changes the momentum of the system as a whole; that is to say, the momentum of the system is changed by the external force. Thus, for the momentum of the system to be conserved, we must have

$$\vec{F}_{\text{ext}} = \vec{0}.$$

- **Closed system:** A system of objects that meets these two requirements is said to be a closed system (also called an isolated system). Thus, the more compact way to express this is shown below.
- **Law of conservation of momentum:** The total momentum of a closed system is conserved:

$$\sum_{j=1}^N \vec{P}_j = \text{constant.}$$

This statement is called the **Law of Conservation of Momentum**

**Note:** Note that there absolutely can be external forces acting on the system; but for the system's momentum to remain constant, these external forces have to cancel, so that the net external force is zero. Billiard balls on a table all have a weight force acting on them, but the weights are balanced (canceled) by the normal forces, so there is no net force.

- **The meaning of a "system":** A **system** (mechanical) is the collection of objects in whose motion (kinematics and dynamics) you are interested. If you are analyzing the bounce of a ball on the ground, you are probably only interested in the motion of the ball, and not of Earth; thus, the ball is your system. If you are analyzing a car crash, the two cars together compose your system
- **Problem solving: Conservation of momentum:** Using conservation of momentum requires four basic steps. The first step is crucial:
  1. Identify a closed system (total mass is constant, no net external force acts on the system).
  2. Write down an expression representing the total momentum of the system before the "event" (explosion or collision).
  3. Write down an expression representing the total momentum of the system after the "event."
  4. Set these two expressions equal to each other, and solve this equation for the desired quantity.
- **Explosions:** if the object is initially motionless, then the system (which is just the object) has no momentum and no kinetic energy. After the explosion, the net momentum of all the pieces of the object must sum to zero (since the momentum of this closed system cannot change). However, the system will have a great deal of kinetic energy after the explosion, although it had none before. Thus, we see that, although the momentum of the system is conserved in an explosion, the kinetic energy of the system most definitely is not; it increases. This interaction—one object becoming many, with an increase of kinetic energy of the system—is called an **explosion**.

**Note:** Where does the energy come from? Does conservation of energy still hold? Yes; some form of potential energy is converted to kinetic energy. In the case of gunpowder burning and pushing out a bullet, chemical potential energy is converted to kinetic energy of the bullet, and of the recoiling gun. For a bow and arrow, it is elastic potential energy in the bowstring.

- **Inelastic collisions:** two or more objects collide with each other and stick together, thus (after the collision) forming one single composite object. The total mass of this composite object is the sum of the masses of the original objects, and the new single object moves with a velocity dictated by the conservation of momentum. However, it turns out again that, although the total momentum of the system of objects remains constant, the kinetic energy doesn't; but this time, the kinetic energy decreases. This type of collision is called inelastic.
- **Perfectly inelastic:** Any collision where the objects stick together will result in the maximum loss of kinetic energy (i.e.,  $K_f$  will be a minimum). Such a collision is called **perfectly inelastic**.
- **Elastic collisions:** The extreme case on the other end is if two or more objects approach each other, collide, and bounce off each other, moving away from each other at the same relative speed at which they approached each other. In this case, the total kinetic energy of the system is conserved. Such an interaction is called elastic.

- **Classifying collision types:**

- If  $0 < K_f < K_i$ , the collision is inelastic.
- If  $K_f$  is the lowest energy, or the energy lost by both objects is the most, the collision is perfectly inelastic (objects stick together).
- If  $K_f = K_i$ , the collision is elastic.

- **Problem solving: Collisions:**

1. Define a closed system.
2. Write down the expression for conservation of momentum.
3. If kinetic energy is conserved, write down the expression for conservation of kinetic energy; if not, write down the expression for the change of kinetic energy.
4. You now have two equations in two unknowns, which you solve by standard methods.

- **Momentum in two dimensions:**

$$p_{f,x} = p_{1,i,x} + p_{2,i,x}$$

$$p_{f,y} = p_{1,i,y} + p_{2,i,y}.$$

- **Force of a system of particles:** Note that the change of momentum of a system is entirely due the external forces, the sum of the chaneg in momentum of internal forces is zero

$$\vec{\mathbf{F}}_{\text{ext}} = \sum_{j=1}^N \frac{d\vec{\mathbf{p}}_j}{dt} .$$

We also know

$$\vec{\mathbf{P}}_{\text{cm}} = \sum_{j=1}^N \vec{\mathbf{p}}_j .$$

Thus,

$$\vec{\mathbf{F}} = \frac{d\vec{\mathbf{p}}_{cm}}{dt}.$$

- **Center of mass:**

$$\vec{\mathbf{r}}_{cm} = \frac{1}{M} \sum_{j=1}^N m_j \vec{\mathbf{r}}_j .$$

- **Instantaneous velocity at center of mass:**

$$\begin{aligned} \vec{\mathbf{v}}_{\text{cm}} &= \frac{d}{dt} \left( \frac{1}{M} \sum_{j=1}^N m_j \vec{\mathbf{r}}_j \right) \\ &= \frac{1}{M} \sum_{j=1}^N m_j \vec{\mathbf{v}}_j . \end{aligned}$$

- **Center of Mass of Continuous Objects:** If the object in question has its mass distributed uniformly in space, rather than as a collection of discrete particles, then  $m_j \rightarrow dm$ , and the summation becomes an integral:

$$\vec{r}_{\text{cm}} = \frac{1}{M} \int \vec{r} dm.$$

- **Conservation of momentum regarding center of mass:**

$$M\vec{v}_{\text{cm,f}} = M\vec{v}_{\text{cm,i}}.$$

Which implies, for a closed system with constant  $M$ , the velocities are equal

- **Rocket equation:**

$$\Delta V = u \ln \left( \frac{m_0}{m} \right).$$

Where  $\Delta v$  is the change in velocity of the rocket,  $u$  is the effective exhaust velocity (which is the average velocity at which the exhaust gases are ejected from the rocket as seen from the rocket frame),  $m_0$  is the initial mass, and  $m$  is the final mass after burning some amount of fuel.

- **Rocket in a Gravitational Field:**

$$\Delta V = u \ln \left( \frac{m_0}{m} \right) - g\Delta t.$$

- **Final velocity of each object in elastic collision:**

$$v_{1,f} = \frac{v_{1,i}(m_1 - m_2) + 2m_2 v_{2,i}}{m_1 + m_2}$$

$$v_{2,f} = \frac{v_{2,i}(m_2 - m_1) + 2m_1 v_{1,i}}{m_1 + m_2}$$

## 2.11 Chapter 10: Fixed-axis rotation

### 2.11.1 Definitions and theorems

- Arc length of a circle:

$$s = r\theta$$

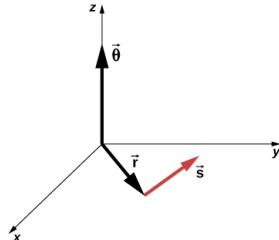
$$\implies \theta = \frac{s}{r}.$$



- Arc length with vectors:

$$\vec{s} = \vec{\theta} \times \vec{r}.$$

That is, the arc length is the cross product of the angle vector and the position vector,



- **Magnitude of angular velocity:** The magnitude of the angular velocity, denoted by  $\omega$ , is the time rate of change of the angle  $\theta$  as the particle moves in its circular path. The instantaneous angular velocity is defined as the limit in which  $\Delta t \rightarrow 0$  in the average angular velocity  $\bar{\omega} = \frac{\Delta\theta}{\Delta t}$ :

$$\omega = \lim_{\Delta t \rightarrow 0} \frac{\Delta\theta}{\Delta t} = \frac{d\theta}{dt}.$$

Units of angular velocity are  $\text{rad s}^{-1}$

- **Tangential speed**  $v_t$ :

$$v_t = r\omega.$$

- **Tangential speed with vectors:**

$$\vec{v} = \vec{\omega} \times \vec{r}.$$

- **Angular acceleration  $\alpha$ :**

$$\alpha = \lim_{\Delta t \rightarrow 0} \frac{\Delta\omega}{\Delta t} = \frac{d\omega}{dt} = \frac{d^2\theta}{dt^2},$$

The units are  $\text{rad s}^{-2}$

- **Tangential acceleration with cross product:**

$$\vec{a} = \vec{\alpha} \times \vec{r}.$$

- **Tangential acceleration as scalar:**

$$a = \alpha r.$$

- **Average angular velocity under constant acceleration:**

$$\bar{\omega} = \frac{\omega_0 + \omega_f}{2}.$$

- **Rotation with constant angular acceleration:**

- Final  $\theta$

$$\theta_f = \theta_0 + \bar{\omega}t.$$

- Final  $\omega$

$$\omega_f = \omega + \alpha t.$$

- Final  $\theta$

$$\theta_f = \theta_0 + \omega_0 t + \frac{1}{2}\alpha t^2.$$

- Final  $\omega$

$$\omega_f^2 = \omega_0^2 + 2\alpha(\Delta\theta).$$

- **Rotational kinetic energy:**

$$K = \frac{1}{2}mv_t^2 = \frac{1}{2}m(\omega r)^2.$$

- **Rotational kinetic energy for a rigid body:**

$$K = \sum_j \frac{1}{2}m_j v_j^2 = \sum_j \frac{1}{2}m_j(r_j\omega_j)^2.$$

But since  $\omega_j = \omega \forall$  masses, we have

$$K = \frac{1}{2} \left( \sum_j m_j r_j^2 \right) \omega^2.$$

- **Moment of inertia:**

$$I = \sum_j m_j r_j^2.$$

With units  $\text{kg} \cdot \text{m}^2$

- Rotational kinetic energy using MOI notation:

$$K = \frac{1}{2} I \omega^2.$$

- Moment of inertia with infinitesimally small piece of mass:

$$I = \int r^2 dm.$$

- Linear mass density  $\lambda$ : Which is the mass per unit length

$$\lambda = \frac{m}{L} \implies m = \lambda L.$$

Provided the mass density of the object is uniform

- Mass differential with  $\lambda$ :

$$dm = d(\lambda L) = \lambda(dL).$$

- Parallel-axis theorem:

Let  $m$  be the mass of an object and let  $d$  be the distance from an axis through the object's center of mass to a new axis. Then we have

$$I_{\text{parallel-axis}} = I_{\text{center of mass}} + md^2.$$

- Surface mass density (uniform): Which is the mass per unit surface area.

$$\begin{aligned} \sigma &= \frac{m}{A}, \quad \text{or} \quad \sigma A = m \\ \implies dm &= \sigma(dA). \end{aligned}$$

- Moment of inertia for compound objects:

$$I_{\text{total}} = \sum_i I_i.$$

- Moment of inertia for uniform thin rod with axis through center of mass:

$$I = \frac{1}{12} ML^2.$$

- Moment of inertia for uniform thin rod with axis at one of the ends:

$$I = \frac{1}{3} ML^2.$$

- Moment of inertia for a uniform disk with axis at center:

$$I = \frac{1}{2} m R^2.$$

- Moment of inertia for a solid cylinder:

$$I = \frac{1}{2} m R^2.$$

- Moment of inertia for a solid sphere:

$$I = \frac{2}{5} m R^2.$$

- Torque: A force that tends to cause rotation.

When a force  $\vec{F}$  is applied to a point  $P$  whose position is  $\vec{r}$  relative to  $O$  (Figure 10.32), the torque  $\vec{\tau}$  around  $O$  is

$$\vec{\tau} = \vec{r} \times \vec{F}.$$

With magnitude

$$\|\vec{\tau}\| = rF \sin(\theta).$$

- **Lever arm:** The quantity  $r \sin(\theta)$  from the equation above is denoted  $r_{\perp}$  and is defined to be the perpendicular distance from  $O$  to the line determined by the vector  $\vec{F}$ . Thus we can write the magnitude of the torque as

$$\|\vec{\tau}\| = r_{\perp} F.$$

- **Net torque:**

$$\vec{\tau}_{\text{net}} = \sum_i \vec{\tau}_i.$$

- **Newton's second law for rotation:** If more than one torque acts on a rigid body about a fixed axis, then the sum of the torques equals the moment of inertia times the angular acceleration:

$$\tau_{\text{net}} = \sum_i \tau_i = I\alpha.$$

- **rotational work:** The total work done on a rigid body is the sum of the torques integrated over the angle through which the body rotates

$$W = \int \sum \vec{\tau} \cdot d\vec{\theta}.$$

- **Rotational work for constant  $\tau$ :**

$$W = \tau\theta.$$

Provided theta starts at zero

- **Incremental rotational work:**

$$dW = \left( \sum_i \tau_i \right) d\theta.$$

- **Work-energy theorem for rotation:** The work-energy theorem for a rigid body rotating around a fixed axis is

$$W_{AB} = K_B - K_A$$

where

$$K = \frac{1}{2}I\omega^2$$

and the rotational work done by a net force rotating a body from point A to point B is

$$W_{AB} = \int_{\theta_A}^{\theta_B} \left( \sum_i \tau_i \right) d\theta.$$

- **Work energy theorem for a constant  $\tau$ :**

$$W_{AB} = \tau(\theta_B - \theta_A) = \frac{1}{2}I(\omega_B^2 - \omega_A^2).$$

- **Power for rotational motion:**

$$P = \tau\omega.$$

## 2.12 Chapter 11: Angular Momentum

### 2.12.1 Definitions and Theorems

- Velocity of pure rolling motion at center of mass (Without slipping):

$$V_{cm} = R\omega.$$

- Linear acceleration of a rolling object:

$$A_{cm} = R\alpha.$$

- Distance traveled:

$$d_{cm} = R\theta.$$

- Acceleration of a rolling object without slipping down an incline plane:

$$a_{cm} = \frac{mg \sin(\theta)}{m + \left(\frac{I_{cm}}{r^2}\right)}.$$

- Conservation of Mechanical Energy in Rolling Motion:

$$E_T = \frac{1}{2}mv_{cm}^2 + \frac{1}{2}I_{cm}\omega^2 + mgh.$$

In the absence of any nonconservative forces that would take energy out of the system in the form of heat, the total energy of a rolling object without slipping is conserved and is constant throughout the motion.

Energy is conserved for an object that is not slipping (has static friction)

- **Angular momentum of a particle:** The angular momentum  $\vec{l}$  of a particle is defined as the cross-product of  $\vec{r}$  and  $\vec{p}$ , and is perpendicular to the plane containing  $\vec{r}$  and  $\vec{p}$ :

$$\begin{aligned} \vec{l} &= \vec{r} \times \vec{p}. \\ \implies l &= rp \sin(\theta). \end{aligned}$$

Where  $\theta$  is the angle between  $\vec{r}$  and  $\vec{p}$ . The units of angular momentum are  $kg \cdot \frac{m^2}{s}$

- **Torque on a particle :**

$$\begin{aligned} \frac{d\vec{l}}{dt} &= \sum \vec{\tau} \\ \implies \frac{d\vec{l}}{dt} &= \tau_{net}. \end{aligned}$$

- **Angular momentum of a rigid body rotating:** The angular momentum along the axis of rotation of a rigid body rotating with angular velocity  $\omega$  about the axis is

$$L = I\omega.$$

- **Law of conservation of angular momentum:** The angular momentum of a system of particles around a point in a fixed inertial reference frame is conserved if there is no net external torque around that point:

$$\frac{d\vec{L}}{dt} = 0.$$

Or

$$\vec{L} = \vec{I}_1 + \vec{I}_2 + \dots + \vec{I}_n = \text{const.}$$

This law is analogous to linear momentum being conserved when the external force on a system is zero.

- **Precession is a change in the orientation of the rotational axis of a rotating body.:**
- **Precession angular velocity:**

$$\omega_P = \frac{rMg}{I\omega}.$$

## 2.13 Chapter 12: Static equilibrium and elasticity

### 2.13.1 Definitions and Theorems

- **equilibrium:** when both its linear and angular acceleration are zero relative to an inertial frame of reference. This means that a body in equilibrium can be moving, but if so, its linear and angular velocities must be constant.
- **static equilibrium:** when it is at rest in our selected frame of reference
- **First equilibrium condition:** The first equilibrium condition for the static equilibrium of a rigid body expresses translational equilibrium:

$$\sum_k \vec{F}_k = \vec{0}.$$

- **Second equilibrium condition:** The second equilibrium condition for the static equilibrium of a rigid body expresses rotational equilibrium:

$$\sum_k \vec{\tau}_k = \vec{0}.$$

- **net gravitational torque:** occurs on an object.

Gravitational torque is the torque caused by weight. This gravitational torque may rotate the object if there is no support present to balance it. The magnitude of the gravitational torque depends on how far away from the pivot the CM is located.

## Newtonian Mechanics Formulas Simplified

### 3.1 General

- **Displacement:**

$$\Delta x = x_f - x_i.$$

- **Velocity:**

$$\bar{v} = \frac{\text{Displacement}}{\Delta t}$$

$$v = \frac{dx}{dt}.$$

- **Speed:**

$$\text{Average speed} = \frac{\text{Distance traveled}}{\Delta t}$$

$$\text{Speed} = |v(t)|.$$

- **Acceleration:**

$$\bar{a} = \frac{\Delta v}{\Delta t}$$

$$a = \frac{dv}{dt}.$$

### 3.2 Chapter 3 & 4: Kinematic equations

- Position:

$$x = x_0 + \bar{v}t$$

$$x = x_0 + v_0 t + \frac{1}{2}at^2.$$

- Velocity:

$$v = v_0 + at$$

$$v^2 = v_0^2 + 2a\Delta x.$$

- Average velocity under constant acceleration:

$$\bar{v} = \frac{v_0 + v}{2}.$$

- Time of flight:

$$\frac{2(v_0 \sin(\theta_0))}{g}.$$

- Range:

$$R = \frac{v_0^2 \sin(2\theta_0)}{g}.$$

### 3.3 Chapter 5 & 6: Newton's laws of motion (forces)

- Newton's first law:

$$\vec{v} = \text{const when } \vec{F}_{net} = \vec{0}.$$

- Newton's Second law:

$$\vec{F} = m\vec{a}.$$

- Newton's third law:

$$\vec{F}_{AB} = -\vec{F}_{BA}.$$

- Weight:

$$\vec{w} = m\vec{g}.$$

- Normal force:

$$\vec{N} = -m\vec{g}.$$

- Hooke's law:

$$\vec{F}_{restore} = -k\vec{x}.$$

- Static friction:

$$\begin{aligned} f_s &\leq \mu_s N \\ \max(f_s) &= \mu_s N. \end{aligned}$$

- Kinetic friction:

$$f_k = \mu_k N.$$

- Centripetal force:

$$\begin{aligned} F_c &= ma_c \\ F_c &= m \frac{v^2}{r} \\ F_c &= mr\omega^2. \end{aligned}$$

- Ideal banking:

$$\theta = \tan^{-1} \left( \frac{v^2}{rg} \right).$$

- Drag force:

$$F_D = \frac{1}{2} \rho C A v^2.$$

where  $C$  is the drag coefficient,  $A$  is the area of the object facing the fluid, and  $\rho$  is the density of the fluid.

- Terminal velocity:

$$v_T = \sqrt{\frac{2mg}{\rho CA}}.$$

- Density of air:

$$\rho = 1.21 \frac{kg}{m^3}.$$

### 3.4 Chapter 7 & 8: Work and energy

- Work:

$$W = \vec{F} \cdot d\vec{r} = Fr \cos(\theta).$$

- Work of friction:

$$W_{fk} = -f_k |\ell|.$$

- Work of gravity:

$$-mgh.$$

- Work of a spring:

$$W = -\frac{1}{2}k\ell^2.$$

- Kinetic energy:

$$K = \frac{1}{2}mv^2.$$

- Net work (work energy theorem):

$$W_{net} = \Delta k.$$

- Power:

$$\begin{aligned} P &= \frac{\Delta w}{\Delta t} \\ P &= \vec{F} \cdot \vec{v}. \end{aligned}$$

- Potential energy:

$$\begin{aligned} U &= -W \\ \implies \Delta U &= -W_{AB}. \end{aligned}$$

- Mechanical energy:

$$E = K + U.$$

- Forces with potential energy:

$$\vec{F} = F_x \hat{i} + F_y \hat{j} = -\frac{dU}{dx} \hat{i} - \frac{dU}{dy} \hat{j}.$$

### 3.5 Chapter 13: Gravitation

- Newton's law of universal gravitation:

$$\vec{F}_{12} = \frac{GMm}{R^2} \hat{r}.$$

- Universal gravitation constant:

$$G = -6.67 \times 10^{-11} N \frac{m^2}{kg^2}.$$

- Weight:

$$mg = \frac{GMm}{R^2}.$$

- Gravitational field:

$$\vec{g} = \frac{GM}{R^2} \hat{r}.$$

- Gravitational potential energy:

$$U = -\frac{GMm}{r}.$$

- Escape velocity:

$$v_{esc} = \sqrt{\frac{2GM}{r}}.$$

- Orbit velocity:

$$v_{orb} = \sqrt{\frac{GM}{r}}.$$

- Period of orbit:

$$T = 2\pi \sqrt{\frac{r^3}{GM}}$$

$$T^2 = \frac{4\pi^2}{GM} a^3.$$

- Orbital energy:

$$E = -\frac{GMm}{2r}$$

### 3.6 Chapter 9: Momentum, Impulse, collisions, center of mass, average force, rocket equation

- Momentum:

$$\vec{p} = m\vec{v}.$$

- Impulse:

$$\vec{J} = \Delta \vec{p} = m\Delta \vec{v}.$$

- Average Force:

$$\vec{F}_{ave} = \frac{\vec{J}}{\delta t} = \frac{\Delta \vec{p}}{\delta t}.$$

- Center of mass:

$$\vec{r}_{cm} = \frac{1}{M} \sum_j m_j \vec{r}_j.$$

- Rocket equation:

$$\Delta v = v_{exh} \ln \left( \frac{m_0}{m_f} \right).$$

- Final velocitys in elastic collisions:

$$v_{1f} = \frac{v_{1i}(m_1 - m_2) + 2m_2 v_{2i}}{m_1 + m_2}$$

$$v_{2f} = \frac{v_{2i}(m_2 - m_1) + 2m_1 v_{1i}}{m_1 + m_2}.$$

- Types of collisions:

- Inelastic (Stick):

$$p_i = p_f \implies mv_1 + mv_2 = (m_1 + m_2)v_f$$

$$k_i \neq k_f.$$

- Elastic (Don't stick):

$$p_i = p_f$$

$$k_i = k_f.$$

### 3.7 Circular Motion

- **Centripetal acceleration:**

$$a_c = \frac{v^2}{r}$$

$$a_c = r\omega^2.$$

- **Period:**

$$T = \frac{2\pi}{\omega}.$$

Where  $T$  is the time it takes to complete one rotation

- **Frequency:**

$$f = \frac{1}{T}.$$

Where  $f$  is the number of rotations per unit time

- **Translational velocity in circular motion:**

$$v = \frac{2\pi r}{T}$$

$$v = 2\pi f$$

$$v = r\omega.$$

- **Angular velocity (angular frequency):**

$$\omega = \frac{2\pi}{T}.$$

Units are  $\frac{\text{rad}}{\text{sec}}$

- **Amplitude of circular motion:**

$$A = r.$$

- **Radians traversed:**

$$\theta = \omega t.$$

- Circular motion position, velocity, acceleration vectors

$$\vec{r}(t) = A \cos(\omega t) \hat{i} + A \sin(\omega t) \hat{j}$$

$$\vec{v}(t) = -A\omega \sin(\omega t) \hat{i} + A\omega \cos(\omega t) \hat{j}$$

$$\vec{a}(t) = -A\omega^2 \cos(\omega t) \hat{i} - A\omega^2 \sin(\omega t) \hat{j}.$$

### 3.8 Chapter 10 & 11: Fixed axis rotation and angular momentum

- **Vectors:**

- **Angular velocity ( $\omega$ ):**
- **Tangential (translational) velocity ( $v$ ):**
- **Angular acceleration ( $\alpha$ ):**
- **Tangential acceleration ( $a$ ):**
- **Torque ( $\tau$ ):**
- **Linear momentum ( $p$ ):**
- **Angular momentum ( $L$ ):**

- **Directions of the vectors:**

- **Angular Velocity ( $\omega$ ):** Directed along the axis of rotation, following the right-hand rule.
- **Tangential (Translational) Velocity ( $v$ ):** Directed tangent to the circle at the point of the object.
- **Angular Acceleration ( $\alpha$ ):** Directed along the axis of rotation, according to the right-hand rule, representing the rate of change of angular velocity.
- **Tangential Acceleration ( $a$ ):** Points in the direction of the change of the tangential velocity, either in the direction of  $v$  (if speeding up) or opposite to it (if slowing down).
- **Torque ( $\tau$ ):** Follows the right-hand rule, directed along the axis of the rotation it produces.
- **Linear Momentum ( $p$ ):** Directed along the tangential velocity vector, as it is mass times velocity.
- **Angular Momentum ( $L$ ):** Directed along the axis of rotation, determined by the right-hand rule, a product of moment of inertia and angular velocity.

- **Tangential acceleration:**

$$a = r\alpha.$$

- **Rotational kinematics:**

$$\begin{aligned}\omega_f &= \omega_0 + \alpha t \\ \omega_f^2 &= \omega_0^2 + 2\alpha\Delta\theta \\ \theta_f &= \theta_0 + \bar{\omega}t \\ \theta_f &= \theta_0 + \omega_0 t + \frac{1}{2}\alpha t^2.\end{aligned}$$

- **Moment of inertia: General:**

$$I = mr^2.$$

- **Moment of inertia: Uniform thin rod, axis through center of mass:**

$$I = \frac{1}{12}mL^2.$$

- **Moment of inertia: Uniform thin rod, axis at one of the ends:**

$$I = \frac{1}{3}mL^2.$$

- **Moment of inertia:** Uniform thin disk, axis through center:

$$I = \frac{1}{2}mr^2.$$

- **Moment of inertia:** Solid cylinder, axis through center:

$$I = \frac{1}{2}mr^2.$$

- **Moment of inertia:** Solid sphere, axis through center:

$$I = \frac{2}{5}mr^2.$$

- **Moment of inertia:** Thin circular hoop:

$$I = mr^2.$$

- **Parallel-axis theorem:**

$$I_{\text{Parallel-axis}} = I_{cm} + md^2.$$

- **Rotational energy:**

$$\begin{aligned} K &= \frac{1}{2} \left( \sum_j m_j r_j^2 \right) \omega^2 \quad \text{omega const} \\ K &= \frac{1}{2} I \omega^2. \end{aligned}$$

- **Torque:**

$$\begin{aligned} \vec{\tau} &= \vec{r} \times \vec{F} = rf \sin(\theta) \\ \vec{\tau}_{net} &= I\alpha. \end{aligned}$$

- **Rotational work:**

$$W = \tau\theta$$

$$W_{net} = \tau\Delta\theta = \frac{1}{2} I \Delta \omega^2.$$

- **Rotational power:**

$$P = \tau\omega.$$

- **Velocity center mass:**

$$v_{cm} = r\omega.$$

- **Acceleration center mass:**

$$a_{cm} = r\alpha.$$

- **Distance:**

$$d = r\theta.$$

- Acceleration of a rolling object without slipping down an incline plane:

$$a_{cm} = \frac{mg \sin(\theta)}{m + \left(\frac{I_{cm}}{r^2}\right)}.$$

- Kinetic energy in rolling motion:

$$K = \frac{1}{2}mv^2 + \frac{1}{2}I\omega^2.$$

- Mechanical energy in rolling motion:

$$E = \frac{1}{2}mv^2 + \frac{1}{2}I\omega^2 + mgh.$$

- Angular momentum:

$$\vec{L} = \vec{r} \times \vec{p} = rp \sin(\theta)$$

$$L = I\omega.$$

# Calculus III

## 4.1 Chapter 1: Parametric equations and polar coordinates

### 4.1.1 Definitions and Theorems

- **Parametric equations, parameter:** If  $x$  and  $y$  are continuous functions of  $t$  on an interval  $I$ , then the equations

$$x = x(t) \quad \text{and} \quad y = y(t)$$

are called **parametric equations** and  $t$  is called the **parameter**.

- **parametric curve:** or plane curve, and is denoted by  $C$ .
- **Eliminating the parameter:** This allows us to rewrite the two equations as a single equation relating the variables  $x$  and  $y$ . Then we can apply any previous knowledge of equations of curves in the plane to identify the curve.
  - Solve one of the equations for  $t$
  - Plug the equation for  $t$  into the equation still in terms of  $t$
  - Sketch the curve on the interval  $I$
- **Domain consideration when graphing by eliminating the parameter:** Suppose we have the equations o

$$\begin{aligned} x &= 2t^2 \\ y &= t^4 + 1. \end{aligned}$$

If we solve  $x$  for  $t$ , we get

$$t = \pm \sqrt{\frac{1}{2}x}.$$

We see that the domain of  $t$  is restricted to  $t \geq 0$ , thus, the graph of this equation will only have the positive side.

- **Parameterizing a curve:** is when we start with the equation of a curve and determine a pair of parametric equations for that curve.
  - To find the first Parameterization, Define  $x(t) = t$  and  $y(t)$  as the function given, with  $t$  instead of  $x$
  - Verify there is no restriction on the domain of the original graph. Thus there is no restriction on the values of  $t$
  - To find the second parameterization, choose some function for  $x(t)$ . Ensure that the domain is the set of all real numbers
  - Plug  $x(t)$  into the original function and solve to get  $y(t)$
- **Parametric equations for a cycloid:**

$$x(t) = a(t - \sin t), \quad y(t) = a(1 - \cos t).$$

- The general parametric equations for a hypocycloid are::

$$x(t) = (a - b) \cos t + b \cos \left( \frac{a - b}{b} \right) t$$

$$y(t) = (a - b) \sin t + b \sin \left( \frac{a - b}{b} \right) t$$

- **Derivatives of Parametric Equations:** Consider the plane curve defined by the parametric equations  $x = x(t)$  and  $y = y(t)$ . Suppose that  $x'(t)$  and  $y'(t)$  exist, and assume that  $x'(t) \neq 0$ . Then the derivative  $\frac{dy}{dx}$  is given by

$$\frac{dy}{dx} = \frac{\frac{dy}{dt}}{\frac{dx}{dt}} = \frac{y'(t)}{x'(t)}.$$

- **Second order derivatives of parametric functions:**

$$\frac{d^2y}{dx^2} = \frac{d}{dx} \left( \frac{dy}{dx} \right) = \frac{\left( \frac{d}{dt} \right) \left( \frac{dy}{dx} \right)}{\frac{dx}{dt}}.$$

- **Integral involving parametric equations:** Consider the non-self-intersecting plane curve defined by the parametric equations

$$x = x(t), \quad y = y(t), \quad a \leq t \leq b$$

and assume that  $x(t)$  is differentiable. The area under this curve is given by

$$A = \int_a^b y(t)x'(t) dt.$$

- **Arc Length of a Parametric Curve:** Consider the plane curve defined by the parametric equations

$$x = x(t), \quad y = y(t), \quad t_1 \leq t \leq t_2$$

and assume that  $x(t)$  and  $y(t)$  are differentiable functions of  $t$ . Then the arc length of this curve is given by

$$s = \int_{t_1}^{t_2} \sqrt{\left( \frac{dx}{dt} \right)^2 + \left( \frac{dy}{dt} \right)^2} dt.$$

Or simply

$$s = \int_a^b \sqrt{1 + \left( \frac{dy}{dx} \right)^2} dx.$$

- **Surface area for a parametric curve:** The analogous formula for a parametrically defined curve is

$$S = 2\pi \int_a^b y(t) \sqrt{(x'(t))^2 + (y'(t))^2} dt$$

provided that  $y(t)$  is not negative on  $[a, b]$ .

- **Parametric equations for a circle (with  $(h, k) = (0, 0)$ ):**

$$x(t) = r \cos(t)$$

$$y(t) = r \sin(t).$$

For  $0 \leq t \leq 2\pi$

- Parametric equations for the upper half of a semi-circle (with  $(h, k) = (0, 0)$ ):

$$\begin{aligned}x(t) &= r \cos(t) \\y(t) &= r \sin(t).\end{aligned}$$

For  $0 \leq t \leq \pi$

- Parametric equations for the lower half of a semi-circle (with  $(h, k) = (0, 0)$ ):

$$\begin{aligned}x(t) &= r \cos(t) \\y(t) &= r \sin(t).\end{aligned}$$

For  $\pi \leq t \leq 2\pi$

- Note: A graph has a horizontal tangent line when the derivative equals zero.:
- Note: A graph has a vertical tangent line when the derivative does not exist:
- Converting Points between Coordinate Systems: Given a point  $P$  in the plane with Cartesian coordinates  $(x, y)$  and polar coordinates  $(r, \theta)$ , the following conversion formulas hold true:

$$x = r \cos \theta \quad \text{and} \quad y = r \sin \theta.$$

$$r^2 = x^2 + y^2 \quad \text{and} \quad \tan \theta = \frac{y}{x}.$$

These formulas can be used to convert from rectangular to polar or from polar to rectangular coordinates.

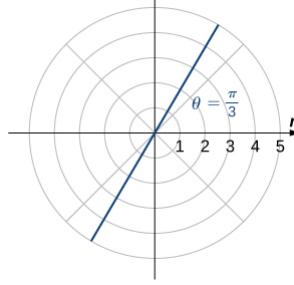
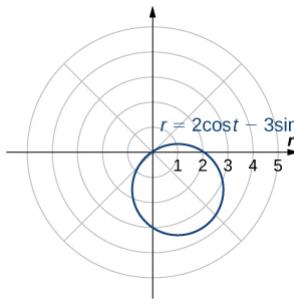
- Sometimes finding  $\theta$  with the formula is not possible. Consider the point  $P(0, 3)$ , using the formula  $\tan \theta = \frac{3}{0}$  is undefined. Instead, we graph the point  $(0, 3)$  and observe the angle between the  $x$  and  $y$  axis is  $\frac{\pi}{2}$ :
- The polar representation of a point is not unique.: Every point in the plane has an infinite number of representations in polar coordinates. However, each point in the plane has only one representation in the rectangular coordinate system.
- polar axis:.
- pole:, or origin, of the coordinate system, and corresponds to  $r = 0$ .
- If it is negative:, move along the ray that is opposite the terminal ray of the given angle.
- Polar equation: has the form

$$r = f(\theta).$$

- Plotting a Curve in Polar Coordinates:

1. Create a table with two columns. The first column is for  $\theta$ , and the second column is for  $r$ .
2. Create a list of values for  $\theta$ .
3. Calculate the corresponding  $r$  values for each  $\theta$ .
4. Plot each ordered pair  $(r, \theta)$  on the coordinate axes.
5. Connect the points and look for a pattern.

- Types of polar curves:

Name	Equation	Example
Line passing through the pole with slope $\tan K$	$\theta = K$	
Circle	$r = a \cos \theta + b \sin \theta$	
Spiral	$r = a + b\theta$	

Name	Equation	Example
Cardioid	$r = a(1 + \cos\theta)$ $r = a(1 - \cos\theta)$ $r = a(1 + \sin\theta)$ $r = a(1 - \sin\theta)$	
Limacon	$r = a\cos\theta + b$ $r = a\sin\theta + b$	
Rose	$r = a\cos(b\theta)$ $r = a\sin(b\theta)$	

- If the coefficient of  $\theta$  is odd, then the number of petals equals the coefficient.:

- **Graphing a polar curve.:**

- Make graph in rectangular system (use  $\theta$  as horizontal axis and  $r$  as vertical)
- Translate over to polar
- The circles represent values of  $r$

- **Area of a Region Bounded by a Polar Curve:** Suppose  $f$  is continuous and nonnegative on the interval  $\alpha \leq \theta \leq \beta$  with  $0 < \beta - \alpha \leq 2\pi$ . The area of the region bounded by the graph of  $r = f(\theta)$  between the radial lines  $\theta = \alpha$  and  $\theta = \beta$  is

$$A = \frac{1}{2} \int_{\alpha}^{\beta} [f(\theta)]^2 d\theta = \frac{1}{2} \int_{\alpha}^{\beta} r^2 d\theta.$$

- **Area between two polar curves:**

$$A = \frac{1}{2} \int_{\alpha}^{\beta} f(\theta)^2 - g(\theta)^2 d\theta.$$

Where  $f(\theta) \geq g(\theta) \forall \alpha \leq \theta \leq \beta$

- **Bounds of integration for area outside some curve and inside some curve:** We find the bounds of integration the same way we found them for regularo functions, we find the points of intersection by setting the two functions equal to each other
- **Arc Length of a Curve Defined by a Polar Function:** Let  $f$  be a function whose derivative is continuous on an interval  $\alpha \leq \theta \leq \beta$ . The length of the graph of  $r = f(\theta)$  from  $\theta = \alpha$  to  $\theta = \beta$  is

$$\begin{aligned} L &= \int_{\alpha}^{\beta} \sqrt{[f(\theta)]^2 + [f'(\theta)]^2} d\theta \\ &= \int_{\alpha}^{\beta} \sqrt{r^2 + \left(\frac{dr}{d\theta}\right)^2} d\theta.. \end{aligned}$$

- **Absolute value bars:** Consider the integral

$$20 \int_0^{2\pi} \sqrt{\cos^2\left(\frac{\theta}{2}\right)} d\theta.$$

If we cancel out the sqrt and square, we must include absolute value bars. This is because  $\cos\left(\frac{\theta}{2}\right)$  can be negative on the interval  $[0, 2\pi]$ . The integral becomes

$$20 \int_0^{2\pi} \left| \cos\left(\frac{\theta}{2}\right) \right| d\theta.$$

To solve this, we use symmetry to change the bounds. Because the graph of the polar curve  $(10 + 10 \cos \theta)$  is symmetric, we can integrate instead over the range  $0, \pi$ , and multiply the result by 2.

- **Derivative of polar equaotion:** Given some function  $r = f(\theta)$ . We can find the derivative with

$$\frac{dy}{dx} = \frac{\frac{dy}{d\theta}}{\frac{dx}{d\theta}} = \frac{\frac{dr}{d\theta} \sin \theta + r \cos \theta}{\frac{dr}{d\theta} \cos \theta - r \sin \theta}.$$

- **Find equation of tangent line given  $r = f(\theta)$  and point  $\theta$ :**

- Find  $r$  with given  $\theta$ .
- Plug  $r$  into  $x = r \cos \theta$ ,  $y = r \sin \theta$  to get point  $P(x, y)$
- Find  $\frac{dy}{dx}$  of  $r = f(\theta)$
- Plug in  $\theta$  to get slope  $m$
- Use point slope form to find equation

### 4.1.2 Problems to remember

- **Eliminating the parameter:** Sometimes it is necessary to be a bit creative in eliminating the parameter. The parametric equations for this example are

$$x(t) = 4 \cos t, \quad y(t) = 3 \sin t.$$

Solving either equation for  $t$  directly is not advisable because sine and cosine are not one-to-one functions. However, dividing the first equation by 4 and the second equation by 3 (and suppressing the  $t$ ) gives us

$$\cos t = \frac{x}{4}, \quad \sin t = \frac{y}{3}.$$

Now use the Pythagorean identity  $\cos^2 t + \sin^2 t = 1$  and replace the expressions for  $\sin t$  and  $\cos t$  with the equivalent expressions in terms of  $x$  and  $y$ . This gives

$$\begin{aligned} \left(\frac{x}{4}\right)^2 + \left(\frac{y}{3}\right)^2 &= 1 \\ \frac{x^2}{16} + \frac{y^2}{9} &= 1. \end{aligned}$$

This is the equation of a horizontal ellipse centered at the origin, with semimajor axis 4 and semiminor axis 3 as shown in the following graph.

- **Convert from cartesian to polar:** Consider the point  $P(1, 1)$ . First we find  $r$  and  $\theta$

$$\begin{aligned} r^2 &= x^2 + y^2 \\ r &= \sqrt{1^2 + 1^2} \\ r &= \sqrt{2} \\ \tan \theta &= \frac{y}{x} = \frac{1}{1} \\ \theta &= \tan^{-1} 1 \\ \theta &= \frac{\pi}{4}. \end{aligned}$$

Thus we have the point  $(\sqrt{2}, \frac{\pi}{4})$

- **Graphing a polar equation:** Graph the curve defined by the function  $r = 4 \sin \theta$ . Identify the curve and rewrite the equation in rectangular coordinates.

Because the function is a multiple of a sine function, it is periodic with period  $2\pi$ , so use values for  $\theta$  between 0 and  $2\pi$

$\theta$	$r = 4 \sin \theta$	$\theta$	$r = 4 \sin \theta$
0	0	$\pi$	0
$\frac{\pi}{6}$	2	$\frac{7\pi}{6}$	-2
$\frac{\pi}{4}$	$2\sqrt{2} \approx 2.8$	$\frac{5\pi}{4}$	$-2\sqrt{2} \approx -2.8$
$\frac{\pi}{3}$	$2\sqrt{3} \approx 3.4$	$\frac{4\pi}{3}$	$-2\sqrt{3} \approx -3.4$
$\frac{\pi}{2}$	4	$\frac{3\pi}{2}$	-4
$\frac{2\pi}{3}$	$2\sqrt{3} \approx 3.4$	$\frac{5\pi}{3}$	$-2\sqrt{3} \approx -3.4$
$\frac{3\pi}{4}$	$2\sqrt{2} \approx 2.8$	$\frac{7\pi}{6}$	$-2\sqrt{2} \approx -2.8$
$\frac{5\pi}{6}$	2	$\frac{11\pi}{6}$	-2
$2\pi$	0		

Plotting these points gives the graph of a circle. The equation  $r = 4 \sin \theta$  can be converted into rectangular coordinates by first multiplying both sides by  $r$ . This gives the equation  $r^2 = 4r \sin \theta$ . Next use the facts that  $r^2 = x^2 + y^2$  and  $y = r \sin \theta$ . This gives  $x^2 + y^2 = 4y$ . To put this equation into standard form, subtract  $4y$  from both sides of the equation and complete the square:

$$x^2 + (y^2 - 4y) = x^2 + (y^2 - 4y + 4) = x^2 + (y - 2)^2 = 0 + 4.$$

This is the equation of a circle with radius 2 and center  $(0, 2)$  in the rectangular coordinate system.

- **Finding the Arc Length of a Polar Curve:**

$$\text{Find the arc length of the polar curve } r = 2 + 2 \cos \theta.$$

When  $\theta = 0$ ,  $r = 2 + 2 \cos(0) = 4$ . Furthermore, as  $\theta$  goes from 0 to  $2\pi$ , the cardioid is traced out exactly once. Therefore, these are the limits of integration. Using  $f(\theta) = 2 + 2 \cos(\theta)$ ,  $\alpha = 0$ , and  $\beta = 2\pi$ . Thus we have

$$\begin{aligned} L &= \int_{\alpha}^{\beta} \sqrt{[f(\theta)]^2 + [f'(\theta)]^2} d\theta \\ &= \int_0^{2\pi} \sqrt{[2 + 2 \cos(\theta)]^2 + [-2 \sin(\theta)]^2} d\theta \\ &= \int_0^{2\pi} \sqrt{4 + 8 \cos(\theta) + 4 \cos^2(\theta) + 4 \sin^2(\theta)} d\theta \\ &= \int_0^{2\pi} \sqrt{4 + 8 \cos(\theta) + 4(\cos^2(\theta) + \sin^2(\theta))} d\theta \\ &= \int_0^{2\pi} \sqrt{8 + 8 \cos(\theta)} d\theta \\ &= 2 \int_0^{2\pi} \sqrt{2 + 2 \cos(\theta)} d\theta.. \end{aligned}$$

Next, using the identity  $\cos(2\alpha) = 2 \cos^2(\alpha) - 1$ , add 1 to both sides and multiply by 2. This gives  $2 + 2 \cos(2\alpha) = 4 \cos^2(\alpha)$ . Substituting  $\alpha = \frac{\theta}{2}$  gives  $2 + 2 \cos(\theta) = 4 \cos^2\left(\frac{\theta}{2}\right)$ , so the integral becomes

$$\begin{aligned} L &= 2 \int_0^{2\pi} \sqrt{2 + 2 \cos(\theta)} d\theta \\ &= 2 \int_0^{2\pi} \sqrt{4 \cos^2\left(\frac{\theta}{2}\right)} d\theta \\ &= 2 \int_0^{2\pi} 2 \left| \cos\left(\frac{\theta}{2}\right) \right| d\theta.. \end{aligned}$$

The absolute value is necessary because the cosine is negative for some values in its domain. To resolve this issue, change the limits from 0 to  $\pi$  and double the answer. This strategy works because cosine is positive between 0 and  $\frac{\pi}{2}$ . Thus,

$$\begin{aligned} L &= 4 \int_0^{2\pi} \left| \cos\left(\frac{\theta}{2}\right) \right| d\theta \\ &= 8 \int_0^{\pi} \cos\left(\frac{\theta}{2}\right) d\theta \\ &= 8 \left[ 2 \sin\left(\frac{\theta}{2}\right) \right]_0^{\pi} \\ &= 16.. \end{aligned}$$

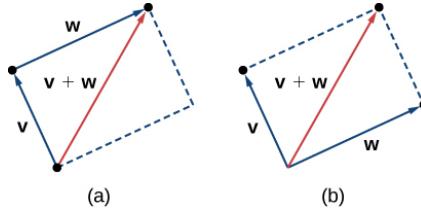
## 4.2 Chapter 2: Vectors in Space

### 4.2.1 Definitions and Theorems

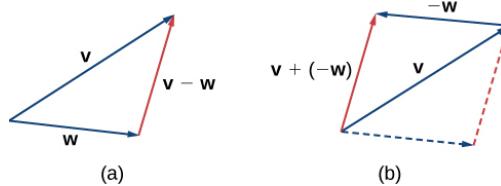
- **terminal point:** of the vector.
- The length of the line segment represents its magnitude. We use the notation  $\|\vec{v}\|$  to denote the magnitude of the vector  $\vec{v}$ .
- **zero vector**, denoted  $\vec{0}$ :
- **equivalent vectors.** We treat equivalent vectors as equal, even if they have different initial points. Thus, if  $\vec{v}$  and  $\vec{w}$  are equivalent, we write

$$\vec{v} = \vec{w}.$$

- **Scalar Multiplication:** The product  $k\vec{v}$  of a vector  $\vec{v}$  and a scalar  $k$  is a vector with a magnitude that is  $|k|$  times the magnitude of  $\vec{v}$ , and with a direction that is the same as the direction of  $\vec{v}$  if  $k > 0$ , and opposite the direction of  $\vec{v}$  if  $k < 0$ . This is called scalar multiplication. If  $k = 0$  or  $\vec{v} = \vec{0}$ , then  $k\vec{v} = \vec{0}$ .
- **Vector Addition:** The sum of two vectors  $\vec{v}$  and  $\vec{w}$  can be constructed graphically by placing the initial point of  $\vec{w}$  at the terminal point of  $\vec{v}$ . Then, the vector sum,  $\vec{v} + \vec{w}$ , is the vector with an initial point that coincides with the initial point of  $\vec{v}$  and has a terminal point that coincides with the terminal point of  $\vec{w}$ . This operation is known as vector addition.



- **Vector difference:** We define  $\vec{v} - \vec{w}$  as  $\vec{v} + (-\vec{w}) = \vec{v} + (-1)\vec{w}$ . The vector  $\vec{v} - \vec{w}$  is called the vector difference. Graphically, the vector  $\vec{v} - \vec{w}$  is depicted by drawing a vector from the terminal point of  $\vec{w}$  to the terminal point of  $\vec{v}$ .



- **Triangle inequality:** the length of any one side is less than the sum of the lengths of the remaining sides. So we have

$$\|\vec{v} + \vec{w}\| \leq \|\vec{v}\| + \|\vec{w}\|.$$

- **standard-position:** vector.

- **Component form of a vector:** The vector with initial point  $(0, 0)$  and terminal point  $(x, y)$  can be written in component form as

$$\vec{v} = \langle x, y \rangle.$$

The scalars  $x$  and  $y$  are called the components of  $\mathbf{v}$ .

- :
- **Component form of a vector not in standard position:** Let  $\vec{v}$  be a vector with initial point  $(x_i, y_i)$  and terminal point  $(x_t, y_t)$ . Then we can express  $\vec{v}$  in component form as  $\vec{v} = \langle x_t - x_i, y_t - y_i \rangle$ .
- **Magnitude of vector:** If a vector is given by components  $\langle x, y \rangle$ . Then this is the vector with initial point at the origin  $(0, 0)$ , and terminal point at  $(x, y)$ . We find the magnitude of the vector with

$$\|\vec{v}\| = \sqrt{x^2 + y^2}.$$

If the vector is not in standard position, and we have initial point  $(x_1, y_1)$  and terminal point  $(x_2, y_2)$ , then we find the magnitude with

$$\|\vec{v}\| = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$

- **Scalar multiplication, and vector addition (component form):** Let  $\mathbf{v} = \langle x_1, y_1 \rangle$  and  $\mathbf{w} = \langle x_2, y_2 \rangle$  be vectors, and let  $k$  be a scalar.

Scalar multiplication:  $k\mathbf{v} = \langle kx_1, ky_1 \rangle$

Vector addition:  $\mathbf{v} + \mathbf{w} = \langle x_1, y_1 \rangle + \langle x_2, y_2 \rangle = \langle x_1 + x_2, y_1 + y_2 \rangle$ .

- **Properties of Vector Operations:** Let  $\mathbf{u}$ ,  $\mathbf{v}$ , and  $\mathbf{w}$  be vectors in a plane. Let  $r$  and  $s$  be scalars.

1.  $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$  (Commutative property)
2.  $(\mathbf{u} + \mathbf{v}) + \mathbf{w} = \mathbf{u} + (\mathbf{v} + \mathbf{w})$  (Associative property)
3.  $\mathbf{u} + \mathbf{0} = \mathbf{u}$  (Additive identity property)
4.  $\mathbf{u} + (-\mathbf{u}) = \mathbf{0}$  (Additive inverse property)
5.  $r(s\mathbf{u}) = (rs)\mathbf{u}$  (Associativity of scalar multiplication)
6.  $(r+s)\mathbf{u} = r\mathbf{u} + s\mathbf{u}$  (Distributive property)
7.  $r(\mathbf{u} + \mathbf{v}) = r\mathbf{u} + r\mathbf{v}$  (Distributive property)
8.  $1\mathbf{u} = \mathbf{u}$ ,  $0\mathbf{u} = \mathbf{0}$  (Identity and zero properties)

**Note** CAAA<sup>-1</sup>(ASM)D<sup>2</sup>(I&Z)

- **Finding components of a vector given the magnitude and the angle  $\theta$ :**

$$x = \|\vec{v}\| \cos \theta$$

$$y = \|\vec{v}\| \sin \theta.$$

- **Unit vector:** A unit vector is a vector with magnitude 1. For any nonzero vector  $\vec{v}$ , we can use scalar multiplication to find a unit vector  $\vec{u}$  that has the same direction as  $\vec{v}$ . To do this, we multiply the vector by the reciprocal of its magnitude:

$$\vec{u} = \frac{1}{\|\vec{v}\|} \vec{v}.$$

- **xy-plane:**

$$\{(x, y, 0) : x, y \in \mathbb{R}\}.$$

Described by the equation  $z = 0$

- **xz-plane:**

$$\{(x, 0, z) : x, z \in \mathbb{R}\}.$$

Described by the equation  $y = 0$

- **yz-plane:**

$$\{(0, y, z) : y, z \in \mathbb{R}\}.$$

Described by the equation  $x = 0$

- **octants.** The octants fill  $\mathbb{R}^3$  the same way that quadrants fill  $\mathbb{R}^2$  ;,
- **Distance formula for three-dimensional space:** The distance  $d$  between points  $(x_1, y_1, z_1)$  and  $(x_2, y_2, z_2)$  is given by the formula

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}.$$

- **Equations of planes parallel to coordinate planes:**

1. The plane in space that is parallel to the xy-plane and contains point  $(a, b, c)$  can be represented by the equation  $z = c$ .
  2. The plane in space that is parallel to the xz-plane and contains point  $(a, b, c)$  can be represented by the equation  $y = b$ .
  3. The plane in space that is parallel to the yz-plane and contains point  $(a, b, c)$  can be represented by the equation  $x = a$
- **sphere** is the set of all points in space equidistant from a fixed point, the center of the sphere, just as the set of all points in a plane that are equidistant from the center represents a circle. In a sphere, as in a circle, the distance from the center to a point on the sphere is called the *radius*:-.
  - **Standard equation of a sphere:** The sphere with center  $(a, b, c)$  and radius  $r$  can be represented by the equation

$$(x - a)^2 + (y - b)^2 + (z - c)^2 = r^2.$$

This equation is known as the **standard equation of a sphere**.

- **Properties of vectors in space:** Let  $\vec{v} = \langle x_1, y_1, z_1 \rangle$  and  $\vec{w} = \langle x_2, y_2, z_2 \rangle$  be vectors, and let  $k$  be a scalar.

- Scalar multiplication:

$$k\vec{v} = \langle kx_1, ky_1, kz_1 \rangle.$$

- Vector addition:

$$\vec{v} + \vec{w} = \langle x_1, y_1, z_1 \rangle + \langle x_2, y_2, z_2 \rangle = \langle x_1 + x_2, y_1 + y_2, z_1 + z_2 \rangle.$$

- Vector subtraction:

$$\vec{v} - \vec{w} = \langle x_1, y_1, z_1 \rangle - \langle x_2, y_2, z_2 \rangle = \langle x_1 - x_2, y_1 - y_2, z_1 - z_2 \rangle.$$

- Vector magnitude:

$$\|\vec{v}\| = \sqrt{x_1^2 + y_1^2 + z_1^2}.$$

- Unit vector in the direction of  $\vec{v}$ :

$$\begin{aligned}\frac{1}{\|\vec{v}\|}\vec{v} &= \frac{1}{\|\vec{v}\|}\langle x_1, y_1, z_1 \rangle \\ &= \left\langle \frac{x_1}{\|\vec{v}\|}, \frac{y_1}{\|\vec{v}\|}, \frac{z_1}{\|\vec{v}\|} \right\rangle, \text{ if } \vec{v} \neq 0.\end{aligned}$$

- **The dot product:** The dot product of vectors  $\vec{u} = \langle u_1, u_2, u_3 \rangle$  and  $\vec{v} = \langle v_1, v_2, v_3 \rangle$  is given by the sum of the products of the components

$$\vec{u} \cdot \vec{v} = u_1 v_1 + u_2 v_2 + u_3 v_3.$$

The dot product **does not** return a new vector, the result is a **scalar**

- **Properties of the dot product:** Let  $\vec{u}$ ,  $\vec{v}$ , and  $\vec{w}$  be vectors, and let  $c$  be a scalar.

1. Commutative property:  $\vec{u} \cdot \vec{v} = \vec{v} \cdot \vec{u}$
2. Distributive property:  $\vec{u} \cdot (\vec{v} + \vec{w}) = \vec{u} \cdot \vec{v} + \vec{u} \cdot \vec{w}$
3. Associative property of scalar multiplication:  $(c\vec{u}) \cdot \vec{v} = (c\vec{u}) \cdot \vec{v} = \vec{u} \cdot (c\vec{v})$
4. Property of magnitude:  $\vec{v} \cdot \vec{v} = \|\vec{v}\|^2$

- **Evaluating a dot product:** The dot product of two vectors is the product of the magnitude of each vector and the cosine of the angle between them:

$$\vec{u} \cdot \vec{v} = \|\vec{u}\| \cdot \|\vec{v}\| \cdot \cos \theta.$$

- **Find the measure of the angle between two nonzero vectors:**

$$\cos \theta = \frac{\vec{u} \cdot \vec{v}}{\|\vec{u}\| \|\vec{v}\|}.$$

**Note:** We are considering  $0 \leq \theta \leq \pi$

- **Vector Projection:** The vector projection of  $\mathbf{v}$  onto  $\mathbf{u}$  has the same initial point as  $\mathbf{u}$  and  $\mathbf{v}$  and the same direction as  $\mathbf{u}$ , and represents the component of  $\mathbf{v}$  that acts in the direction of  $\mathbf{u}$ :

$$\text{proj}_{\vec{u}} \vec{v} = \frac{\vec{v} \cdot \vec{u}}{\|\vec{u}\|^2} \vec{u}.$$

We say "The vector projection of  $\vec{v}$  onto  $\vec{u}$ "

- **Scalar projection notation:** This is the length of the vector projection and is denoted

$$\|\text{proj}_{\vec{u}} \vec{v}\| = \text{comp}_{\vec{u}} \vec{v} = \frac{\vec{u} \cdot \vec{v}}{\|\vec{u}\|}.$$

- **Decompose some vector  $\vec{v}$  into orthogonal components such that one of the component vectors has the same direction as  $\vec{u}$ :**

- First, we compute  $\vec{p} = \text{proj}_{\vec{u}} \vec{v}$
- Then, we define  $\vec{q} = \vec{v} - \vec{p}$
- Check that  $\vec{q}$  and  $\vec{p}$  are orthogonal by finding  $\vec{q} \cdot \vec{p}$

- **Work:** When a constant force is applied to an object so the object moves in a straight line from point  $P$  to point  $Q$ , the work  $W$  done by the force  $\mathbf{F}$ , acting at an angle  $\theta$  from the line of motion, is given by

$$W = \vec{F} \cdot \overrightarrow{PQ} = \|\vec{F}\| \|\overrightarrow{PQ}\| \cos \theta.$$

- **Two vectors are orthogonal if:**

$$\vec{u} \cdot \vec{v} = 0.$$

- **Two vectors are parallel if:**

$$\exists \alpha \text{ s.t } \alpha \vec{u} = \vec{v}.$$

- **Scalar projection components of a vector:**

$$\vec{v} = \langle \text{comp}_i \vec{v}, \text{comp}_j \vec{v}, \text{comp}_k \vec{v} \rangle.$$

- **Find direction angles for some vector:** Suppose we have some vector  $\vec{v}$  : to find the direction angles, we use the formula

$$\cos \theta = \frac{\vec{v} \cdot \vec{u}}{\|\vec{v}\| \|\vec{u}\|}.$$

With the unit vectors  $\hat{i}, \hat{j}, \hat{k}$ . This will give angles  $\alpha, \beta, \gamma$

- **The Cross Product:** Let  $\mathbf{u} = \langle u_1, u_2, u_3 \rangle$  and  $\mathbf{v} = \langle v_1, v_2, v_3 \rangle$ . Then, the cross product  $\mathbf{u} \times \mathbf{v}$  is vector

$$\begin{aligned} \mathbf{u} \times \mathbf{v} &= (u_2 v_3 - u_3 v_2) \mathbf{i} - (u_1 v_3 - u_3 v_1) \mathbf{j} + (u_1 v_2 - u_2 v_1) \mathbf{k} \\ &= \langle u_2 v_3 - u_3 v_2, -(u_1 v_3 - u_3 v_1), u_1 v_2 - u_2 v_1 \rangle. \end{aligned}$$

**Note:** The cross product only works in  $\mathbb{R}^3$ , additionally, we measure the angle between  $\vec{u}$  and  $\vec{v}$  in  $\vec{u} \times \vec{v}$  from  $\vec{u}$  to  $\vec{v}$

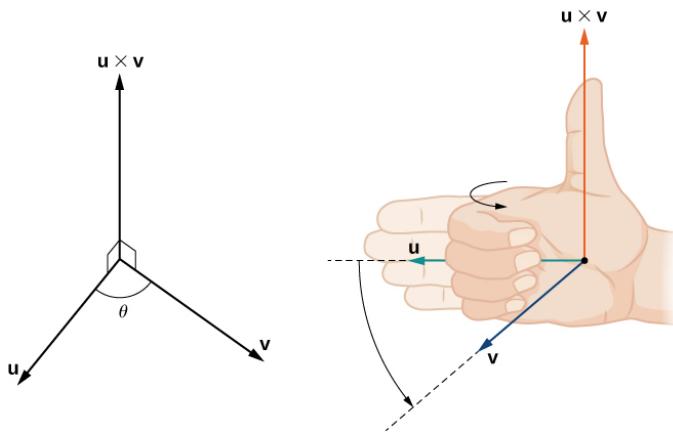
- **Cross product using matrix and discriminant,** suppose we have vectors  $\vec{u}$  und  $\vec{v}$  :. Then we can express them in matrix form as

$$\vec{u} \times \vec{v} = \begin{bmatrix} \hat{i} & \hat{j} & \hat{k} \\ u_x & u_y & u_z \\ v_x & v_y & v_z \end{bmatrix}.$$

Then we can find the discriminant of this matrix to compute the cross product

$$\vec{u} \times \vec{v} = (u_y v_z - u_z v_y) \hat{i} - (u_x v_z - u_z v_x) \hat{k} + (u_x v_y - u_y v_x) \hat{j}.$$

- **Right hand rule for cross product:** The direction of  $\mathbf{u} \times \mathbf{v}$  is given by the right-hand rule. If we hold the right hand out with the fingers pointing in the direction of  $\mathbf{u}$ , then curl the fingers toward vector  $\mathbf{v}$ , the thumb points in the direction of the cross product, as shown.



- **Properties of the Cross Product:** Let  $\mathbf{u}$ ,  $\mathbf{v}$ , and  $\mathbf{w}$  be vectors in space, and let  $c$  be a scalar.

1. Anticommutative property:  $\mathbf{u} \times \mathbf{v} = -(\mathbf{v} \times \mathbf{u})$
2. Distributive property:  $\mathbf{u} \times (\mathbf{v} + \mathbf{w}) = \mathbf{u} \times \mathbf{v} + \mathbf{u} \times \mathbf{w}$
3. Multiplication by a constant:  $c(\mathbf{u} \times \mathbf{v}) = (c\mathbf{u}) \times \mathbf{v} = \mathbf{u} \times (c\mathbf{v})$
4. Cross product of the zero vector:  $\mathbf{u} \times \mathbf{0} = \mathbf{0} \times \mathbf{u} = \mathbf{0}$
5. Cross product of a vector with itself:  $\mathbf{v} \times \mathbf{v} = \mathbf{0}$
6. Scalar triple product:  $\mathbf{u} \cdot (\mathbf{v} \times \mathbf{w}) = (\mathbf{u} \times \mathbf{v}) \cdot \mathbf{w}$

**Note:** (AC)D(MC)(ZV)(IT)(TP)

- **Magnitude of the Cross Product:** Let  $\mathbf{u}$  and  $\mathbf{v}$  be vectors, and let  $\theta$  be the angle between them. Then,  $\|\mathbf{u} \times \mathbf{v}\| = \|\mathbf{u}\| \cdot \|\mathbf{v}\| \cdot \sin \theta$ .

- **Applications of the cross product:**

- Finding a vector orthogonal to two given vectors
- Computing areas of triangles and parallelograms
- Determining the volume of the three-dimensional geometric shape made of parallelograms known as a parallelepiped:w

- **Area of a Parallelogram:** If we locate vectors  $\mathbf{u}$  and  $\mathbf{v}$  such that they form adjacent sides of a parallelogram, then the area of the parallelogram is given by  $\|\mathbf{u} \times \mathbf{v}\|$ .

- **Triple Scalar Product:**

The triple scalar product of vectors  $\mathbf{u}$ ,  $\mathbf{v}$ , and  $\mathbf{w}$  is  $\mathbf{u} \cdot (\mathbf{v} \times \mathbf{w})$ .

The triple scalar product is the determinant of the  $3 \times 3$  matrix formed by the components of the vectors

- **triple scalar product identities:**

- (a)  $\mathbf{u} \cdot (\mathbf{v} \times \mathbf{w}) = -\mathbf{u} \cdot (\mathbf{w} \times \mathbf{v})$
- (b)  $\mathbf{u} \cdot (\mathbf{v} \times \mathbf{w}) = \mathbf{v} \cdot (\mathbf{w} \times \mathbf{u}) = \mathbf{w} \cdot (\mathbf{u} \times \mathbf{v})$

- **parallelepiped:** Let  $\mathbf{u}$  and  $\mathbf{v}$  be two vectors in standard position. If  $\mathbf{u}$  and  $\mathbf{v}$  are not scalar multiples of each other, then these vectors form adjacent sides of a parallelogram.

Now suppose we add a third vector  $\mathbf{w}$  that does not lie in the same plane as  $\mathbf{u}$  and  $\mathbf{v}$  but still shares the same initial point. Then these vectors form three edges of a parallelepiped, a three-dimensional prism with six faces that are each parallelograms,

- **Volume of a Parallelepiped:** The volume of a parallelepiped with adjacent edges given by the vectors  $\mathbf{u}$ ,  $\mathbf{u}$ , and  $\mathbf{w}$  is the absolute value of the triple scalar product:

$$V = \left| \mathbf{u} \cdot (\mathbf{v} \times \mathbf{w}) \right|.$$

- **Torque:** Measures the tendency of a force to produce rotation about an axis of rotation. Let  $\mathbf{r}$  be a vector with an initial point located on the axis of rotation and with a terminal point located at the point where the force is applied, and let vector  $\mathbf{F}$  represent the force. Then torque is equal to the cross product of  $\mathbf{r}$  and  $\mathbf{F}$ :

$$\tau = \mathbf{r} \times \mathbf{F}.$$

- **Choosing  $\alpha$  to make parallel vectors equal:** Suppose we have two vectors  $\mathbf{u}$  and  $\mathbf{v}$ , and  $\exists \alpha \in \mathbb{R}$  s.t  $\alpha\mathbf{v} = \mathbf{u}$ . Then

$$\alpha = \frac{\|\mathbf{u}\|}{\|\mathbf{v}\|}.$$

Or if they are **anti-parallel**

$$\alpha = -\frac{\|\mathbf{u}\|}{\|\mathbf{v}\|}.$$

- **The zero vector is considered to be parallel to all vectors:**

- **vector equation of a line:**

$$\mathbf{r} = \mathbf{r}_0 + t\mathbf{v}.$$

Where  $\mathbf{v}$  is the direction vector (vector parallel to the line),  $t$  is some scalar, and  $\mathbf{r}$ ,  $\mathbf{r}_0$  are position vectors

- **Parametric and Symmetric Equations of a Line:** A line  $L$  parallel to vector  $\mathbf{v} = \langle a, b, c \rangle$  and passing through point  $P(x_0, y_0, z_0)$  can be described by the following parametric equations:

$$x = x_0 + ta, \quad y = y_0 + tb, \quad \text{and} \quad z = z_0 + tc.$$

If the constants  $a$ ,  $b$ , and  $c$  are all nonzero, then  $L$  can be described by the symmetric equation of the line:

$$\frac{x - x_0}{a} = \frac{y - y_0}{b} = \frac{z - z_0}{c}.$$

**Note:** The parametric equations of a line are not unique. Using a different parallel vector or a different point on the line leads to a different, equivalent representation. Each set of parametric equations leads to a related set of symmetric equations, so it follows that a symmetric equation of a line is not unique either.

- **Vector equation of a line reworked:** Suppose we have some line, with points  $P(x_0, y_0, z_0)$ ,  $Q(x_1, y_1, z_1)$ . Where  $\mathbf{p} = \langle x_0, y_0, z_0 \rangle$  and  $\mathbf{q} = \langle x_1, y_1, z_1 \rangle$  are the corresponding position vectors. Suppose we also have  $\mathbf{r} := \langle x, y, z \rangle$ . Then our vector equation for a line becomes

$$\mathbf{r} = \mathbf{p} + t(\vec{PQ}).$$

By properties of vectors, we get the vector equation of a line passing through points  $P$  and  $Q$  to be

$$\mathbf{r} = (1-t)\mathbf{p} + t\mathbf{q}.$$

- **Equation of a line segment between two points  $P$  and  $Q$ :** Using the result from the previous item, we find the vector equation of the line segment between  $P$  and  $Q$  is

$$\mathbf{r} = (1-t)\mathbf{p} + t\mathbf{q}, \quad 0 \leq t \leq 1.$$

Because when  $t = 0$ ,  $\mathbf{r} = \mathbf{p}$ . When  $t = 1$ ,  $\mathbf{r} = \mathbf{q}$

- **parametric equations for this line segment:**

$$\begin{cases} x = x_0 + t(x_1 - x_0) \\ y = y_0 + t(y_1 - y_0) \\ z = z_0 + t(z_1 - z_0) \end{cases} \quad (34)$$

For  $0 \leq t \leq 1$

- **Distance from a Point to a Line:** Let  $L$  be a line in space passing through point  $P$  with direction vector  $\mathbf{v}$ . If  $M$  is any point not on  $L$ , then the distance from  $M$  to  $L$  is

$$d = \frac{\|\overrightarrow{PM} \times \mathbf{v}\|}{\|\mathbf{v}\|}$$

- **skew:** lines
- **Classifying lines in space:**

		Lines Share A Common Point?	
		Yes	No
Direction Vectors Are Parallel?	Yes	Equal	Parallel but not equal
	No	Intersecting	Skew

- **Testing for parallel, intersecting, or skew:** Suppose we have two sets of parametric equations for two lines.

- First we test for parallel lines, We find their direction vectors, if their direction vectors are parallel, the lines are parallel.

- If they are not parallel, we can test for intersection. This is best explained with an example

Suppose we have the equations

$$\begin{aligned} L_1 : \quad x &= 3 + 2t & y &= 4 - t & z &= 1 + 3t \\ L_2 : \quad x &= 1 + 4s & y &= 3 - 2s & z &= 4 + 5s. \end{aligned}$$

First, we equate each set of equations

$$\begin{aligned} 3 + 2t &= 1 + 4s \\ 4 - t &= 3 - 2s \\ 1 + 3t &= 4 + 5s. \end{aligned}$$

Next, we solve the first equation for one of the variables, in this case we choose to solve for  $t$

$$t = 2s - 1.$$

Now, we want to plug this into the second equation. This gives

$$\begin{aligned} 4 - (2s - 1) &= 3 - 2s \\ 2 &= 0. \end{aligned}$$

Since we this statement is not true, we know the lines must not be intersecting. In this case, we know they must be skew

However, suppose we got some value for  $s$ , such as  $s = 1$ , we can then plug this value into the equation we got for  $t$  ( $t = 2s - 1$ ) to get a value for  $t$ , after we have a value for  $t$ , we can plug both  $s$  and  $t$  into the third equation and evaluate the truthiness of its outcome. If the outcome is true, we have a valid solution for the system of equations.

**Note:** If the lines are known to be skew, then we compute the dot product of their direction angles to check for orthogonality.

- **a plane is determined by three.:**
- **there is exactly one plane containing both lines.:**
- **determined by a line and any point that does not lie on the line.:**
- We say that  $\mathbf{n}$  is a normal vector, or perpendicular to the plane.
- **Vector equation of a plane:** Given a point  $P$  and vector  $\mathbf{n}$ , the set of all points  $Q$  satisfying the equation  $\mathbf{n} \cdot \overrightarrow{PQ} = 0$  forms a plane. The equation

$$\mathbf{n} \cdot \overrightarrow{PQ} = 0$$

is known as the vector equation of a plane.

- **Scalar equation of a plane:** The scalar equation of a plane containing point  $P = (x_0, y_0, z_0)$  with normal vector  $\mathbf{n} = \langle a, b, c \rangle$  is

$$a(x - x_0) + b(y - y_0) + c(z - z_0) = 0.$$

- **General form of the equation of a plane:** This equation (the one above) can be expressed as  $ax + by + cz + d = 0$ , where  $d = -ax_0 - by_0 - cz_0$ . This form of the equation is sometimes called the general form of the equation of a plane.
- any three points that do not all lie on the same line determine a plane. Given three such points, we can find an equation for the plane containing these points.
- **The Distance between a Plane and a Point:** Suppose a plane with normal vector  $\mathbf{n}$  passes through point  $Q$ . The distance  $d$  from the plane to a point  $P$  not in the plane is given by

$$d = \|\text{proj}_{\mathbf{n}} \vec{QP}\| = \left| \text{comp}_{\mathbf{n}} \vec{QP} \right| = \frac{|\vec{QP} \cdot \mathbf{n}|}{\|\mathbf{n}\|}.$$

- **The two distinct planes are parallel or they intersect. When two planes are parallel, their normal vectors are parallel. When two planes intersect, the intersection is a line:**
- **Finding line of intersection for two planes:** Suppose we are given the parametric equations for two planes
  - Check if their normals are parallel, if not, we know the planes must intersect
  - Find a common point that satisfies both equations (perhaps the origin)
  - Eliminate one of the variables (perhaps by adding them)
  - After elimination we should have one variable equal to one of the others, we then plug this into one of the equations to get two variables in terms of the third.
  - Define this third variable in terms of  $t$  to get  $x, y, z$  in terms of  $t$ , these equations will be the parametric equations for the line of intersection
- **find the angle formed by the intersection of two planes:** We can use normal vectors to calculate the angle between the two planes.

We can find the measure of the angle  $\theta$  between two intersecting planes by first finding the cosine of the angle, using the following equation:

$$\cos \theta = \frac{|\mathbf{n}_1 \cdot \mathbf{n}_2|}{\|\mathbf{n}_1\| \|\mathbf{n}_2\|}.$$

We can then use the angle to determine whether two planes are parallel or orthogonal or if they intersect at some other angle.

- **Finding the distance between two parallel planes:** Let  $P(x_0, y_0, z_0)$  be a point. The distance from  $P$  to plane  $ax + by + cz + k = 0$  is given by

$$d = \frac{|ax_0 + by_0 + cz_0 + k|}{\sqrt{a^2 + b^2 + c^2}}.$$

- **Find the intersection of a plane with a line:**

- First, we get the line in parametric form, with  $t$  is the parameter.
- When then plug our parametric equations into the corresponding plane equation variables
- Solve for  $t$ ,
- Plug  $t$  into parametric equation to get point  $(x, y, z)$

## 4.3 Conic sections and quadric surfaces

### 4.3.1 Conic sections: Parabola, Ellipse, and Hyperbola

- **Conic sections, The parabola:** A parabola is the set of all points in a plane equidistant from a fixed point  $F$  (the focus) and a fixed line  $\ell$  (the directrix) that lie in the plane.
- **Parabola that opens up:** Given a parabola opening upward with vertex located at  $(h, k)$  and focus located at  $(h, k + p)$ , where  $p$  is a constant, the equation for the parabola is given by

$$y = \frac{1}{4p}(x - h)^2 + k$$

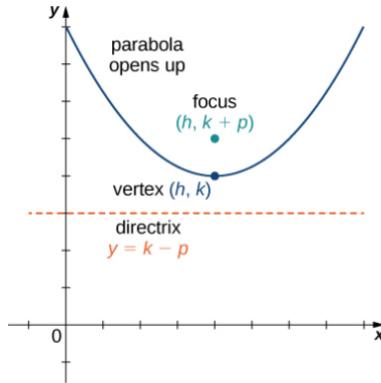
Vertex:  $(h, k)$

Focus:  $(h, k + p)$

Directrix  $y = k - p$

AOS  $x = h$ .

This is the **standard form** of a parabola.



- **Parabola that opens down:**

$$y = -\frac{1}{4p}(x - h)^2 + k$$

Vertex:  $(h, k)$

Focus:  $(h, k - p)$

Directrix  $y = k + p$

AOS  $x = h$ .



- Parabola opens right:

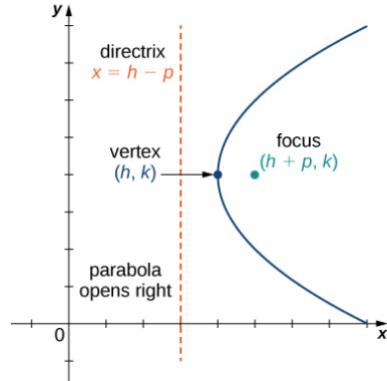
$$x = \frac{1}{4p}(y - k)^2 + h$$

Vertex:  $(h, k)$

Focus:  $(h + p, k)$

Directrix  $x = h - p$

AOS  $y = k$ .



- Parabola opens left:

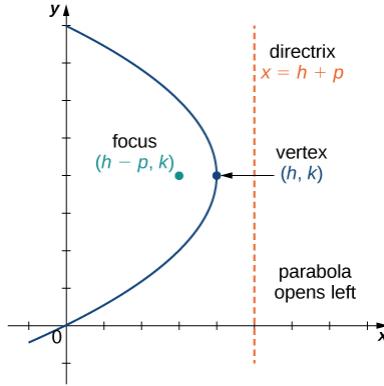
$$x = -\frac{1}{4p}(y - k)^2 + h$$

Vertex:  $(h, k)$

Focus:  $(h - p, k)$

Directrix  $x = h + p$

AOS  $y = k$ .



- **General form of a parabola:** The equation of a parabola can be written in the general form, though in this form the values of  $h$ ,  $k$ , and  $p$  are not immediately recognizable. The general form of a parabola is written as

$$ax^2 + bx + cy + d = 0 \quad \text{or} \quad ay^2 + bx + cy + d = 0.$$

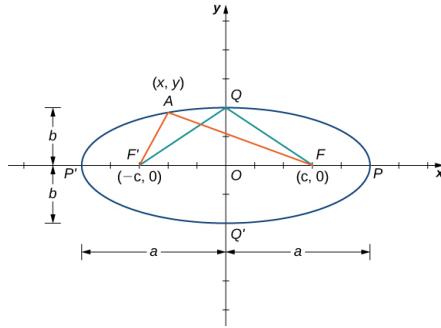
**Note:** The first equation represents a parabola that opens either up or down. The second equation represents a parabola that opens either to the left or to the right. To put the equation into standard form, use the method of completing the square.

- **Finding  $h$  and  $k$  for a parabola:**  $h$  and  $k$  for a parabola are given by

$$h = -\frac{2b}{a}$$

$$k = f(h).$$

- **Conic sections, The ellipse:** An ellipse is the set of all points for which the sum of their distances from two fixed points (the foci) is constant.



**Foci:** There are two foci

**Directrices:** There are two directrices (plural of directrix)

**Major axis:** An ellipse has two axes, one short and one long. The major axis is the longer of the two and has length  $2a$

**Minor axis:** The shorter axis has length  $2b$

**Finding  $c$  with  $a$  and  $b$ ,** to find  $c$ , we use

$$c^2 = a^2 - b^2.$$

- **Ellipse with horizontal major axis:** This type of ellipse has the form

$$\frac{(x-h)^2}{a^2} + \frac{(y-k)^2}{b^2} = 1$$

Center  $(h, k)$

Foci:  $(h \pm c, k)$

Directrices:  $x = h \pm \frac{a^2}{c}$ .

- **Ellipse with vertical major axis::** This type of ellipse has the form

$$\frac{(x-h)^2}{b^2} + \frac{(y-k)^2}{a^2} = 1$$

Center:  $(h, k)$

Foci:  $(h, k \pm c)$

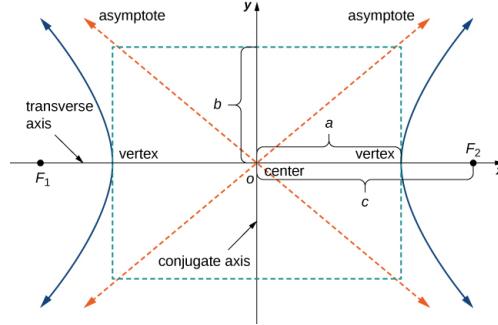
Directrices:  $y = k \pm \frac{a^2}{c}$ .

- **General form of an ellipse:** The equation of an ellipse is in general form if it is in the form

$$Ax^2 + By^2 + Cx + Dy + E = 0.$$

where  $A$  and  $B$  are either both positive or both negative. To convert the equation from general to standard form, use the method of completing the square.

- **Conic sections, the Hyperbola:** A hyperbola is the set of all points where the difference between their distances from two fixed points (the foci) is constant.



**Foci:** There are two foci

**Directrices:** There are two directrices (plural of directrix)

**Asymptotes:** There are two Asymptotes

**Transverse axis (major axis)** has length  $2a$

**Conjugate axis (minor axis)** has length  $2b$

**Finding  $c$  with  $a$  and  $b$ ,** to find  $c$ , we use

$$c^2 = a^2 + b^2.$$

- Hyperbola opening left and right (horizontal major axis):

$$\frac{(x-h)^2}{a^2} - \frac{(y-k)^2}{b^2} = 1$$

Center:  $(h, k)$

Foci:  $(h \pm c, k)$

$$\text{Asymptotes: } y = k \pm \frac{b}{a}(x - h)$$

$$\text{Directrices: } x = h \pm \frac{a^2}{c}.$$

- Hyperbola opening up and down (vertical major axis):

$$\frac{(y-k)^2}{a^2} - \frac{(x-h)^2}{b^2} = 1$$

Center:  $(h, k)$

Foci:  $(h, k \pm c)$

$$\text{Asymptotes: } y = k \pm \frac{a}{b}(x - h)$$

$$\text{Directrices: } y = k \pm \frac{a^2}{c}.$$

- **Hyperbola general form:** The equation of a hyperbola is in general form if it is in the form

$$Ax^2 + By^2 + Cx + Dy + E = 0.$$

where  $A$  and  $B$  have opposite signs. In order to convert the equation from general to standard form, use the method of completing the square.

- **Eccentricity and directrix of conic sections:** The eccentricity  $e$  of a conic section is defined to be the distance from any point on the conic section to its focus, divided by the perpendicular distance from that point to the nearest directrix. This value is constant for any conic section, and can define the conic section as well:

- If  $e = 1$ , the conic is a parabola.
- If  $e < 1$ , it is an ellipse.
- If  $e > 1$ , it is a hyperbola.

The eccentricity of a circle is zero. The directrix of a conic section is the line that, together with the point known as the focus, serves to define a conic section. Hyperbolas and noncircular ellipses have two foci and two associated directrices. Parabolas have one focus and one directrix.



- Eccentricity for parabola:

$$e = 1.$$

- Eccentricity for ellipses:

$$e = \frac{c}{a}.$$

- Eccentricity for hyperbolas:

$$e = \frac{c}{a}.$$

#### 4.3.2 Quadric Surfaces

- **rulings:**
- **traces:** of a surface are the cross-sections created when the surface intersects a plane parallel to one of the coordinate planes.
- **quadric surfaces:**
- **Quadric surfaces:** are the graphs of equations that can be expressed in the form

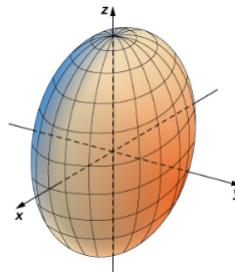
$$Ax^2 + By^2 + Cz^2 + Dxy + Exz + Fyz + Gx + Hy + Jz + K = 0..$$

**Note:** When a quadric surface intersects a coordinate plane, the trace is a conic section.

- **ellipsoid:** is a surface described by an equation of the form

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1.$$

**Note:** Set  $x = 0$  to see the trace of the ellipsoid in the  $yz$ -plane. To see the traces in the  $xy$ - and  $xz$ -planes, set  $z = 0$  and  $y = 0$ , respectively. Notice that, if  $a = b$ , the trace in the  $xy$ -plane is a circle. Similarly, if  $a = c$ , the trace in the  $xz$ -plane is a circle and, if  $b = c$ , then the trace in the  $yz$ -plane is a circle. A sphere, then, is an ellipsoid with  $a = b = c$ .



- **Sphere:** is an ellipsoid with

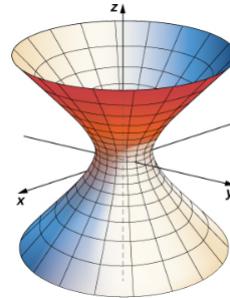
$$a = b = c.$$

**Traces:** All three traces are ellipses

- **two unique traces:**, these traces make the name of the surface.
- **Hyperboloid of one sheet:**

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1.$$

Two of the variables have positive coefficients and one has a negative coefficient. The axis of the surface corresponds to the variable with the negative coefficient

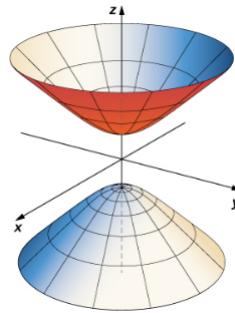


**Traces:** One ellipse and two hyperbololas

- **Hyperboloid of two sheets:**

$$\frac{z^2}{c^2} - \frac{x^2}{a^2} - \frac{y^2}{b^2} = 1.$$

Two of the variables have negative coefficients and one has a positive coefficient. The axis of the surface corresponds to the variable with the positive coefficient. The surface does not intersect the coordinate plane perpendicular to the axis

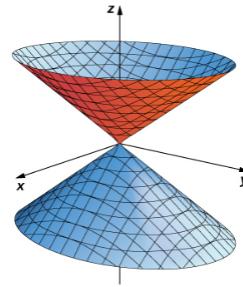


**Traces:** One ellipse (or the empty set (no trace)), and two hyperbololas

- **Elliptic cone:**

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 0.$$

The axis of the surface corresponds to the variable with a negative coefficient. The traces in the coordinate planes parallel to the axis are intersecting lines



**Traces:** One ellipse and two hyperbolas.

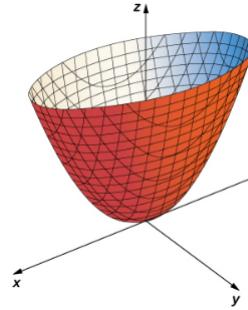
In the  $xz$ -plane, we have a pair of lines that intersect at the origin

In the  $yz$ -plane, we have a pair of lines that intersect at the origin

- **Elliptic paraboloid:**

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = z.$$

The axis of the surface corresponds to the linear variable

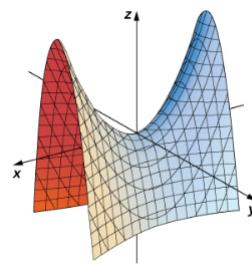


**Traces:** One ellipse and two parabolas

- **Hyperbolic paraboloid:**

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = z.$$

The axis of the surface corresponds to the linear variable



**Traces:** One hyperbola and two parabolas

## 4.4 Chapter 3: Vector-Valued Functions

### 4.4.1 Definitions and Theorems

- **vector-valued function:** is a function of the form

$$\mathbf{r}(t) = f(t)\mathbf{i} + g(t)\mathbf{j} \quad \text{or} \quad \mathbf{r}(t) = f(t)\mathbf{i} + g(t)\mathbf{j} + h(t)\mathbf{k},$$

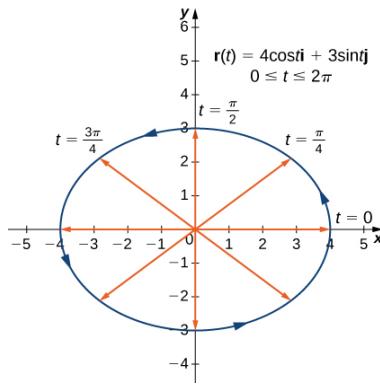
where the **component functions**  $f$ ,  $g$ , and  $h$ , are real-valued functions of the parameter  $t$ . Vector-valued functions are also written in the form

$$\mathbf{r}(t) = \langle f(t), g(t) \rangle \quad \text{or} \quad \mathbf{r}(t) = \langle f(t), g(t), h(t) \rangle.$$

In both cases, the first form of the function defines a two-dimensional vector-valued function; the second form describes a three-dimensional vector-valued function.

**Note:** The parameter  $t$  can lie between two real numbers:  $a \leq t \leq b$ . Another possibility is that the value of  $t$  might take on all real numbers. Last, the component functions themselves may have domain restrictions that enforce restrictions on the value of  $t$ . We often use  $t$  as a parameter because  $t$  can represent time.

- **Graphing vector-valued functions: Plane curve:** The graph of a vector-valued function of the form  $\mathbf{r}(t) = f(t)\mathbf{i} + g(t)\mathbf{j}$  consists of the set of all  $(t, \mathbf{r}(t))$ , and the path it traces is called a **plane curve**.
- **Graphing vector-valued functions: Space curve:** The graph of a vector-valued function of the form  $\mathbf{r}(t) = f(t)\mathbf{i} + g(t)\mathbf{j} + h(t)\mathbf{k}$  consists of the set of all  $(t, \mathbf{r}(t))$ , and the path it traces is called a **space curve**.
- **Vector Parameterization:** Any representation of a plane curve or space curve using a vector-valued function is called a **vector parameterization** of the curve.
- **Note on graphing vector-valued functions:** When graphing a vector-valued function, we typically graph the vectors in the domain of the function in standard position, because doing so guarantees the uniqueness of the graph.
- **How to graph vector-valued functions:** As with any graph, we start with a table of values. We then graph each of the vectors in the second column of the table in standard position and connect the terminal points of each vector to form a curve



- **similarity between vector-valued functions and parameterized curves.**: Given a vector-valued function  $\mathbf{r}(t) = f(t)\mathbf{i} + g(t)\mathbf{j}$ , we can define  $x = f(t)$  and  $y = g(t)$ . If a restriction exists on the values of  $t$  (for example,  $t$  is restricted to the interval  $[a, b]$  for some constants  $a < b$ ), then this restriction is enforced on the parameter. The graph of the parameterized function would then agree with the graph of the vector-valued function, except that the vector-valued graph would represent vectors rather than points. Since we can parameterize a curve defined by a function  $y = f(x)$ , it is also possible to represent an arbitrary plane curve by a vector-valued function.
- **Limits of a Vector-Valued Function (Rigorous):** A vector-valued function  $\mathbf{r}$  approaches the limit  $L$  as  $t$  approaches  $a$ , written

$$\lim_{t \rightarrow a} \mathbf{r}(t) = L,$$

provided

$$\lim_{t \rightarrow a} \|\mathbf{r}(t) - L\| = 0.$$

**Note:** This is a rigorous definition of the limit of a vector-valued function. In practice, we use the following theorem:

- **Limit of a Vector-Valued Function:** Let  $f$ ,  $g$ , and  $h$  be functions of  $t$ . Then the limit of the vector-valued function  $\mathbf{r}(t) = f(t)\mathbf{i} + g(t)\mathbf{j}$  as  $t$  approaches  $a$  is given by

$$\lim_{t \rightarrow a} \mathbf{r}(t) = \left[ \lim_{t \rightarrow a} f(t) \right] \mathbf{i} + \left[ \lim_{t \rightarrow a} g(t) \right] \mathbf{j},$$

provided the limits  $\lim_{t \rightarrow a} f(t)$  and  $\lim_{t \rightarrow a} g(t)$  exist. Similarly, the limit of the vector-valued function  $\mathbf{r}(t) = f(t)\mathbf{i} + g(t)\mathbf{j} + h(t)\mathbf{k}$  as  $t$  approaches  $a$  is given by

$$\lim_{t \rightarrow a} \mathbf{r}(t) = \left[ \lim_{t \rightarrow a} f(t) \right] \mathbf{i} + \left[ \lim_{t \rightarrow a} g(t) \right] \mathbf{j} + \left[ \lim_{t \rightarrow a} h(t) \right] \mathbf{k},$$

provided the limits  $\lim_{t \rightarrow a} f(t)$ ,  $\lim_{t \rightarrow a} g(t)$ , and  $\lim_{t \rightarrow a} h(t)$  exist.

- **Continuity of a vector-valued function:** Let  $f$ ,  $g$ , and  $h$  be functions of  $t$ . Then, the vector-valued function  $\mathbf{r}(t) = f(t)\mathbf{i} + g(t)\mathbf{j}$  is continuous at point  $t = a$  if the following three conditions hold:

1.  $\mathbf{r}(a)$  exists.
2.  $\lim_{t \rightarrow a} \mathbf{r}(t)$  exists.
3.  $\lim_{t \rightarrow a} \mathbf{r}(t) = \mathbf{r}(a)$ .

Similarly, the vector-valued function  $\mathbf{r}(t) = f(t)\mathbf{i} + g(t)\mathbf{j} + h(t)\mathbf{k}$  is continuous at point  $t = a$  if the following three conditions hold:

1.  $\mathbf{r}(a)$  exists.
2.  $\lim_{t \rightarrow a} \mathbf{r}(t)$  exists.
3.  $\lim_{t \rightarrow a} \mathbf{r}(t) = \mathbf{r}(a)$ .

- **Example problem:** Find the vector equation that represents the curve of intersection of the cylinder  $x^2 + y^2 := 9$  and the surface  $z = x + 3y$

To start, we can easily find  $x(t)$  and  $y(t)$ .

**Proposition.**

$$\begin{aligned}x(t) &= 3 \cos t \\y(t) &= 3 \sin t.\end{aligned}$$

We can verify this simply

$$\begin{aligned}\frac{1}{3}x &= \cos t \\ \frac{1}{3}y &= \sin t.\end{aligned}$$

If  $\cos^2 t + \sin^2 t = 1$ , Then

$$\begin{aligned}\frac{1}{9}x^2 + \frac{1}{9}y^2 &= 1 \\ x^2 + y^2 &= 9.\end{aligned}$$

To find the function for  $z(t)$ , we see that our surface is already solved for  $z$ , thus we simply replace  $x$  and  $y$  with what we have for  $x(t)$  and  $y(t)$

$$\implies z(t) = 3 \cos t + 9 \sin t.$$

- **The derivative of a vector-valued function  $r(t)$ :** The derivative of a vector-valued function  $\mathbf{r}(t)$  is

$$\mathbf{r}'(t) = \lim_{\Delta t \rightarrow 0} \frac{\mathbf{r}(t + \Delta t) - \mathbf{r}(t)}{\Delta t}, \quad (35)$$

provided the limit exists. If  $\mathbf{r}'(t)$  exists, then  $\mathbf{r}$  is differentiable at  $t$ . If  $\mathbf{r}'(t)$  exists for all  $t$  in an open interval  $(a, b)$ , then  $\mathbf{r}$  is differentiable over the interval  $(a, b)$ . For the function to be differentiable over the closed interval  $[a, b]$ , the following two limits must exist as well:

$$\mathbf{r}'(a) = \lim_{\Delta t \rightarrow 0^+} \frac{\mathbf{r}(a + \Delta t) - \mathbf{r}(a)}{\Delta t} \quad (36)$$

and

$$\mathbf{r}'(b) = \lim_{\Delta t \rightarrow 0^-} \frac{\mathbf{r}(b + \Delta t) - \mathbf{r}(b)}{\Delta t}. \quad (37)$$

- **Differentiation of Vector-Valued Functions:** Let  $f$ ,  $g$ , and  $h$  be differentiable functions of  $t$ .

- If  $\mathbf{r}(t) = f(t)\mathbf{i} + g(t)\mathbf{j}$ , then  $\mathbf{r}'(t) = f'(t)\mathbf{i} + g'(t)\mathbf{j}$ .
- If  $\mathbf{r}(t) = f(t)\mathbf{i} + g(t)\mathbf{j} + h(t)\mathbf{k}$ , then  $\mathbf{r}'(t) = f'(t)\mathbf{i} + g'(t)\mathbf{j} + h'(t)\mathbf{k}$ .

- **Properties of the Derivative of Vector-Valued Functions:** Let  $\mathbf{r}$  and  $\mathbf{u}$  be differentiable vector-valued functions of  $t$ , let  $f$  be a differentiable real-valued function of  $t$ , and let  $c$  be a scalar.

1.  $\frac{d}{dt}[\mathbf{cr}(t)] = \mathbf{cr}'(t)$ , Scalar multiple
2.  $\frac{d}{dt}[\mathbf{r}(t) \pm \mathbf{u}(t)] = \mathbf{r}'(t) \pm \mathbf{u}'(t)$ , Sum and difference
3.  $\frac{d}{dt}[f(t)\mathbf{u}(t)] = f'(t)\mathbf{u}(t) + f(t)\mathbf{u}'(t)$ , Scalar product
4.  $\frac{d}{dt}[\mathbf{r}(t) \cdot \mathbf{u}(t)] = \mathbf{r}'(t) \cdot \mathbf{u}(t) + \mathbf{r}(t) \cdot \mathbf{u}'(t)$ , Dot product
5.  $\frac{d}{dt}[\mathbf{r}(t) \times \mathbf{u}(t)] = \mathbf{r}'(t) \times \mathbf{u}(t) + \mathbf{r}(t) \times \mathbf{u}'(t)$ , Cross product
6.  $\frac{d}{dt}[\mathbf{r}(f(t))] = \mathbf{r}'(f(t)) \cdot f'(t)$ , Chain rule
7. If  $\mathbf{r}(t) \cdot \mathbf{r}(t) = c$ , then  $\mathbf{r}(t) \cdot \mathbf{r}'(t) = 0$ .

- **principal unit tangent vector:** Let  $C$  be a curve defined by a vector-valued function  $\mathbf{r}$ , and assume that  $\mathbf{r}'(t)$  exists when  $t = t_0$ . A tangent vector  $\mathbf{v}$  at  $t = t_0$  is any vector such that, when the tail of the vector is placed at point  $\mathbf{r}(t_0)$  on the graph, vector  $\mathbf{v}$  is tangent to curve  $C$ . Vector  $\mathbf{r}'(t_0)$  is an example of a tangent vector at point  $t = t_0$ . Furthermore, assume that  $\mathbf{r}'(t) \neq 0$ . The principal unit tangent vector at  $t$  is defined to be

$$\mathbf{T}(t) = \frac{\mathbf{r}'(t)}{\|\mathbf{r}'(t)\|}, \quad (38)$$

provided  $\|\mathbf{r}'(t)\| \neq 0$ .

- **the derivative provides a tangent vector to the curve represented by the function.:**
- **The Fundamental Theorem of Calculus applies to vector-valued functions as well.:**
- **Integrals of Vector-Valued Functions:** Let  $f$ ,  $g$ , and  $h$  be integrable real-valued functions over the closed interval  $[a, b]$ . The indefinite integral of a vector-valued function  $\mathbf{r}(t) = f(t)\mathbf{i} + g(t)\mathbf{j} + h(t)\mathbf{k}$  is

$$\int [f(t)\mathbf{i} + g(t)\mathbf{j} + h(t)\mathbf{k}] dt = \left[ \int f(t) dt \right] \mathbf{i} + \left[ \int g(t) dt \right] \mathbf{j} + \left[ \int h(t) dt \right] \mathbf{k}. \quad (39)$$

The definite integral of the vector-valued function is

$$\int_a^b [f(t)\mathbf{i} + g(t)\mathbf{j} + h(t)\mathbf{k}] dt = \left[ \int_a^b f(t) dt \right] \mathbf{i} + \left[ \int_a^b g(t) dt \right] \mathbf{j} + \left[ \int_a^b h(t) dt \right] \mathbf{k}. \quad (40)$$

- **Integration constant for the integral of a vector-valued function:**

$$\begin{aligned} \int [f(t)\mathbf{i} + g(t)\mathbf{j}] dt &= \left[ \int f(t) dt \right] \mathbf{i} + \left[ \int g(t) dt \right] \mathbf{j} \\ &= (F(t) + C_1)\mathbf{i} + (G(t) + C_2)\mathbf{j} \\ &= F(t)\mathbf{i} + G(t)\mathbf{j} + C_1\mathbf{i} + C_2\mathbf{j} \\ &= F(t)\mathbf{i} + G(t)\mathbf{j} + \mathbf{C}. \end{aligned}$$

where  $\mathbf{C} = C_1\mathbf{i} + C_2\mathbf{j}$ . Therefore, the integration constant becomes a constant vector.

- **Arc Length for Vector Functions:**

1. **Plane curve:** Given a smooth curve  $C$  defined by the function  $\mathbf{r}(t) = f(t)\mathbf{i} + g(t)\mathbf{j}$ , where  $t$  lies within the interval  $[a, b]$ , the arc length of  $C$  over the interval is

$$s = \int_a^b \sqrt{[f'(t)]^2 + [g'(t)]^2} dt = \int_a^b \|\mathbf{r}'(t)\| dt. \quad (41)$$

2. **Space curve:** Given a smooth curve  $C$  defined by the function  $\mathbf{r}(t) = f(t)\mathbf{i} + g(t)\mathbf{j} + h(t)\mathbf{k}$ , where  $t$  lies within the interval  $[a, b]$ , the arc length of  $C$  over the interval is

$$s = \int_a^b \sqrt{[f'(t)]^2 + [g'(t)]^2 + [h'(t)]^2} dt = \int_a^b \|\mathbf{r}'(t)\| dt. \quad (42)$$

**Note:** Note that the formulas are defined for smooth curves: curves where the vector-valued function  $r(t)$  is differentiable with a non-zero derivative. The smoothness condition guarantees that the curve has no cusps (or corners) that could make the formula problematic.

- **Arc-length function definition:** If a vector-valued function represents the position of a particle in space as a function of time, then the arc-length function measures how far that particle travels as a function of time.
- **Arc-length function:** Let  $\mathbf{r}(t)$  describe a smooth curve for  $t \geq a$ . Then the arc-length function is given by

$$s(t) = \int_a^t \|\mathbf{r}'(u)\| du \quad (43)$$

Furthermore,  $\frac{ds}{dt} = \|\mathbf{r}'(t)\| > 0$ . If  $\|\mathbf{r}'(t)\| = 1$  for all  $t \geq a$ , then the parameter  $t$  represents the arc length from the starting point at  $t = a$ .

**Note:** If a vector-valued function represents the position of a particle in space as a function of time, then the arc-length function measures how far that particle travels as a function of time.

Since  $s(t)$  measures distance traveled as a function of time,  $s'(t)$  measures the speed of the particle at any given time.

- **Arc-length parametrization:**

1. Find  $s(t)$
2. Solve  $s(t)$  for  $t$
3. Plug expression for  $t$  into  $\mathbf{r}(t)$  to get  $\mathbf{r}(s)$

The vector-valued function is now written in terms of the parameter  $s$ . Since the variable  $s$  represents the arc length, we call this an arc-length parameterization of the original function  $r(t)$ . One advantage of finding the arc-length parameterization is that the distance traveled along the curve starting from  $s = 0$  is now equal to the parameter  $s$ .

**Eg:** Suppose we have the function  $r(t) = 4 \cos t \hat{\mathbf{i}} + 4 \sin t \hat{\mathbf{j}}$  for  $t \geq 0$ . First, we find  $s(t)$

$$\begin{aligned} r'(u) &= -4 \sin(u) \hat{\mathbf{i}} + 4 \cos(u) \hat{\mathbf{k}} \\ \implies s(t) &= \int_0^t \sqrt{16 \sin^2 u + 16 \cos^2 u} du \\ &= \int_0^t 4 du = 4u \Big|_0^t = 4t. \end{aligned}$$

Now we solve for  $t$

$$s = 4t \implies t = \frac{1}{4}s.$$

Pluggin back into  $\vec{r}(t)$  we get

$$\vec{r}(s) = 4 \cos\left(\frac{1}{4}s\right) \hat{\mathbf{i}} + 4 \sin\left(\frac{1}{4}s\right) \hat{\mathbf{j}}.$$

This is the arc-length parameterization of  $\mathbf{r}(t)$ . Since the original restriction on  $t$  was given by  $t \geq 0$ , the restriction on  $s$  becomes  $\frac{s}{4} \geq 0$ , or  $s \geq 0$ .

- **Curvature:** The concept of curvature provides a way to measure how sharply a smooth curve turns. Let  $C$  be a smooth curve in the plane or in space given by  $\mathbf{r}(s)$ , where  $s$  is the arc-length parameter. The curvature  $\kappa$  at  $s$  is

$$\kappa = \left\| \frac{d\mathbf{T}}{ds} \right\| = \|\mathbf{T}'(s)\|. \quad (44)$$

- **Alternate formulas for curvature:** If  $C$  is a smooth curve given by  $\mathbf{r}(t)$ , then the curvature  $\kappa$  of  $C$  at  $t$  is given by

$$\kappa = \frac{\|\mathbf{T}'(t)\|}{\|\mathbf{r}'(t)\|}. \quad (45)$$

If  $C$  is a three-dimensional curve, then the curvature can be given by the formula

$$\kappa = \frac{\|\mathbf{r}'(t) \times \mathbf{r}''(t)\|}{\|\mathbf{r}'(t)\|^3}. \quad (46)$$

If  $C$  is the graph of a function  $y = f(x)$  and both  $y'$  and  $y''$  exist, then the curvature  $\kappa$  at point  $(x, y)$  is given by

$$\kappa = \frac{|y''|}{[1 + (y')^2]^{3/2}}. \quad (47)$$

- **The curvature of a circle:** is given by

$$\frac{1}{\text{radius}}.$$

- **Principal unit normal vector:** Let  $C$  be a three-dimensional smooth curve represented by  $\mathbf{r}$  over an open interval  $I$ . If  $\mathbf{T}'(t) \neq 0$ , then the principal unit normal vector at  $t$  is defined to be

$$\mathbf{N}(t) = \frac{\mathbf{T}'(t)}{\|\mathbf{T}'(t)\|}. \quad (48)$$

- **Principal unit binormal vector:** The binormal vector at  $t$  is defined as

$$\mathbf{B}(t) = \mathbf{T}(t) \times \mathbf{N}(t), \quad (49)$$

where  $\mathbf{T}(t)$  is the unit tangent vector.

**Note:** the binormal vector will already be of magnitude one

- We can only find a binormal vector for a space curve, not a two dimensional curve (plane curve):
- **Normal plane:** The unit normal vector and the binormal vector form a plane that is perpendicular to the curve at any point on the curve, called the normal plane
- **Frenet frame of reference (also called the TNB frame):**
- **osculating plane of  $C$  at any point  $P$  on the curve.:**

## 4.5 Chapter 4: Differentiation of Functions of Several Variables

### 4.5.1 Definitions and Theorems

- **Function of two variables:** A function of two variables  $z = f(x, y)$  maps each ordered pair  $(x, y)$  in a subset  $D$  of the real plane  $\mathbb{R}^2$  to a unique real number  $z$ . The set  $D$  is called the domain of the function. The range of  $f$  is the set of all real numbers  $z$  that has at least one ordered pair  $(x, y) \in D$  such that  $f(x, y) = z$  as shown in the following figure.
- **Surface:** The graph of a function  $z = f(x, y)$  of two variables is called a surface.
- **level curve of a function of two variables:** for the value  $c$  is defined to be the set of points satisfying the equation  $f(x, y) = c$
- **contour map.:**
- **vertical traces:** are graphed in the  $xz$ - or  $yz$ -planes.
- **vertical trace:** of the function can be either the set of points that solves the equation  $f(a, y) = z$  for a given constant  $x = a$  or  $f(x, b) = z$  for a given constant  $y = b$ .
- **Limit laws for functions of two variables:** Let  $f(x, y)$  and  $g(x, y)$  be defined for all  $(x, y) \neq (a, b)$  in a neighborhood around  $(a, b)$ , and assume the neighborhood is contained completely inside the domain of  $f$ . Assume that  $L$  and  $M$  are real numbers such that  $\lim_{(x,y) \rightarrow (a,b)} f(x, y) = L$  and  $\lim_{(x,y) \rightarrow (a,b)} g(x, y) = M$ , and let  $c$  be a constant. Then each of the following statements holds:

**Constant Law:**

$$\lim_{(x,y) \rightarrow (a,b)} c = c \quad (4.2)$$

**Identity Laws:**

$$\lim_{(x,y) \rightarrow (a,b)} x = a \quad (4.3)$$

$$\lim_{(x,y) \rightarrow (a,b)} y = b \quad (4.4)$$

**Sum Law:**

$$\lim_{(x,y) \rightarrow (a,b)} (f(x, y) + g(x, y)) = L + M \quad (4.5)$$

**Difference Law:**

$$\lim_{(x,y) \rightarrow (a,b)} (f(x, y) - g(x, y)) = L - M \quad (4.6)$$

**Constant Multiple Law:**

$$\lim_{(x,y) \rightarrow (a,b)} (cf(x, y)) = cL \quad (4.7)$$

**Product Law:**

$$\lim_{(x,y) \rightarrow (a,b)} (f(x, y)g(x, y)) = LM \quad (4.8)$$

**Quotient Law:**

$$\lim_{(x,y) \rightarrow (a,b)} \frac{f(x, y)}{g(x, y)} = \frac{L}{M} \quad \text{for } M \neq 0 \quad (4.9)$$

**Power Law:**

$$\lim_{(x,y) \rightarrow (a,b)} (f(x, y))^n = L^n \quad (4.10)$$

for any positive integer  $n$ . **Root Law:**

$$\lim_{(x,y) \rightarrow (a,b)} \sqrt[n]{f(x,y)} = \sqrt[n]{L} \quad (4.11)$$

for all  $L$  if  $n$  is odd and positive, and for  $L \geq 0$  if  $n$  is even and positive provided that  $f(x,y) \geq 0$  for all  $(x,y) \neq (a,b)$  in neighborhood of  $(a,b)$ .

- **Method to show that a limit does not exist:** To show that a limit exists can be quite challenging. To show that the limit exists is to show that it is the same along each and every path (infinitely many). To show that it does not exist, however, we only need to show that the limit differs for at least two unique paths

**Example:** Show that the following limit does not exist

$$\lim_{(x,y) \rightarrow (0,0)} \frac{x^2 - y^2}{2x^2 + y^2}.$$

We see that evaluating this limit directly yields an undefined result. Thus, we shall show that the limit differs among paths taken. Suppose we imagine our point at  $(0,0)$ , if we choose to go along the path  $x = 0$ , our limit becomes

$$\lim_{(0,y) \rightarrow (0,0)} \frac{-y^2}{y^2} = -1.$$

Likewise, lets go along the line  $y = 0$ , we get

$$\lim_{(x,0) \rightarrow (0,0)} \frac{x^2}{2x^2} = \frac{1}{2}.$$

Since  $-1 \neq \frac{1}{2}$ . We assert the limit does not exist at point  $(0,0)$  and hence move forward

- **L'Hospital's Rule in limits of functions of two or more variables:** It is sadly the case that we cannot use L'Hospital's Rule for limits of two variables, we can however use it when showing the limits are different among different paths. If we choose our path such that one variable vanishes, then the function is now of one variable and we are free to use L'Hospital's Rule to evaluate it.

**Example:**

$$\lim_{(x,y) \rightarrow (0,0)} \frac{3xy}{3x^2 + y^2}.$$

If we choose our path of interest to be along  $y = x$ , we get

$$\lim_{(x,x) \rightarrow (0,0)} \frac{3x^2}{4x^2} = \lim_{x \rightarrow 0} \frac{3x^2}{4x^2}.$$

We are now free to use L'Hospital's Rule, although it is clearly not necessary in this case.

- **Using polar coordinates to evaluate limits:**

$$\lim_{(x,y) \rightarrow (0,0)} \frac{xy}{\sqrt{x^2 + y^2}}.$$

We swap  $x$  and  $y$  to polar form

$$\begin{aligned} & \lim_{r \rightarrow 0} \frac{r \cos(\theta)r \sin(\theta)}{\sqrt{r^2 \cos^2(\theta) + r^2 \sin^2(\theta)}} \\ &= \lim_{r \rightarrow 0} r \cos(\theta) \sin(\theta) \\ &= 0. \end{aligned}$$

- **Squeeze theorem for multivariable limits:** Suppose we have

$$\lim_{(x,y) \rightarrow (0,0)} \frac{5x^2y}{x^2 + y^2}.$$

We want to consider the positive function, thus we examine

$$\begin{aligned} & \lim_{(x,y) \rightarrow (0,0)} \left| \frac{5x^2y}{x^2 + y^2} \right| \\ &= \lim_{(x,y) \rightarrow (0,0)} \frac{5x^2|y|}{x^2 + y^2}. \end{aligned}$$

We remark that

$$\lim_{(x,y) \rightarrow (a,b)} \left| f(x,y) \right| = \left| \lim_{(x,y) \rightarrow (a,b)} f(x,y) \right|.$$

We notice

$$0 \leq \frac{5x^2}{x^2 + y^2} \leq 5.$$

From here, we multiply all sides by  $|y|$ , note that this is why we wanted to consider the absolute value of the function, so we can peacefully manipulate the inequality

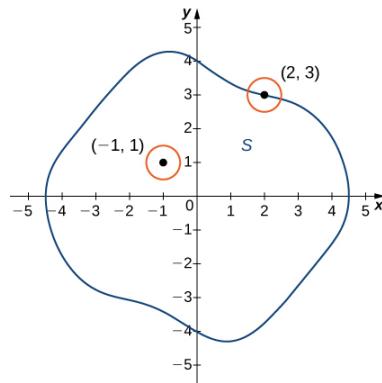
$$0 \leq \frac{5x^2|y|}{x^2 + y^2} \leq 5|y|.$$

Now, we take limits and notice the squeeze

$$\lim_{(x,y) \rightarrow (0,0)} 0 \leq \lim_{(x,y) \rightarrow (0,0)} \frac{5x^2|y|}{x^2 + y^2} \leq \lim_{(x,y) \rightarrow (0,0)} 5|y|.$$

Thus, the limit of the middle function must be zero.

- **Limits of functions of three variables:** Similar to the functions of two variables, we can choose a path along an axis by having two of the variables be zero. Furthermore, we can go along a curve that's parametric. Thus, we set  $x, y$  and  $z$  to functions of a parameter  $t$ . This allows us to travel along a curve  $C$ . The best way to approach this is to choose the functions of  $t$  such that the degrees in both the numerator and denominator match up.
- **Interior Points and Boundary Points:** Let  $S$  be a subset of  $\mathbb{R}^2$ 
  - interior point:** of  $S$  if there is a  $\delta$ -disk centered around  $P_0$  contained completely in  $S$ .
  - boundary point:** of  $S$  if every  $\delta$ -disk centered around  $P_0$  contains points both inside and outside  $S$ .
  - open set:** if every point of  $S$  is an interior point.
  - closed set:** if it contains all its boundary points.
  - connected set:** if it cannot be represented as the union of two or more disjoint, nonempty open subsets
  - region:** if it is open, connected, and nonempty.



- **Continuity of a function of two variables:** A function  $f(x, y)$  is continuous at a point  $(a, b)$  in its domain if the following conditions are satisfied:

1.  $f(a, b)$  exists.
2.  $\lim_{(x,y) \rightarrow (a,b)} f(x, y)$  exists.
3.  $\lim_{(x,y) \rightarrow (a,b)} f(x, y) = f(a, b)$ .

- **The Sum of Continuous Functions Is Continuous:**

- **The Product of Continuous Functions Is Continuous:**

- **The Composition of Continuous Functions Is Continuous:**

- **Sketching graphs for domain and continuity:**

- Identify domain restrictions
- If domain restriction is an inequality, change inequality to equality and solve for  $y$
- Graph function of  $y$ , if inequality is strict, curve should be dotted. If non-strict, solid
- Test points outside/inside (or above/below for line) with original inequality (from domain restriction), shade regions that yield true

**Example:** Graph the set of points of continuity for the following function:

$$\ln(x + 6y).$$

We see

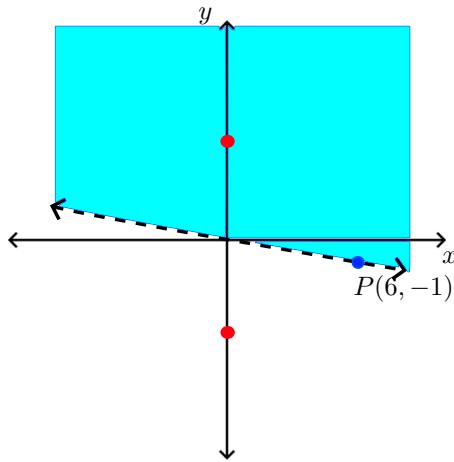
$$\begin{aligned} x + 6y &> 0 \quad (\text{strict}) \\ \implies y &= -\frac{1}{6}x. \end{aligned}$$

Graph:

The shaded region is found with test points  $T_1(0, -3)$  and  $T_2(0, 3)$  (red points).  
We see

$$\begin{aligned} 0 + 6(-3) &\not> 0 \\ 0 + 6(3) &> 0. \end{aligned}$$

- **Constant of proportionality:**



- Direct proportionality

$$y = kx.$$

Where  $k$  is the constant of proportionality

The constant of proportionality can be determined if you know the values of the two variables at a specific point. For direct proportionality, if you know a pair of values  $(x_0, y_0)$ , you can find  $k$  by rearranging the formula:

$$k = \frac{y}{x}.$$

- Inversely proportional

$$y = \frac{k}{x}.$$

For inverse proportionality, given a pair of values  $(x_0, y_0)$ ,  $k$  can be found as

$$k = xy.$$

We use  $\propto$  notation without the constant, but when we replace it with equality, we need to introduce the constant to maintain equality

- **Limit definition of a partial derivative:** Let  $f(x, y)$  be a function of two variables. Then the partial derivative of  $f$  with respect to  $x$ , written as  $\frac{\partial f}{\partial x}$ , or  $f_x$ , is defined as

$$\frac{\partial f}{\partial x} = \lim_{h \rightarrow 0} \frac{f(x + h, y) - f(x, y)}{h}.$$

The partial derivative of  $f$  with respect to  $y$ , written as  $\frac{\partial f}{\partial y}$ , or  $f_y$ , is defined as

$$\frac{\partial f}{\partial y} = \lim_{k \rightarrow 0} \frac{f(x, y + k) - f(x, y)}{k}.$$

**Note:** The logic remains the same for functions of three variables

- **Fast computation of partial derivatives:** To compute partial derivatives without using the limit definition, we let the variable that we aren't interested in be a constant, this allows us to compute the derivative as if it were ordinary.

- **Interpretation of partial derivatives:** The partial derivative in the x-direction is the slope for slight movement in the x-direction, etc
- **higher-order partial derivatives:** There are four second-order partial derivatives for any function (provided they all exist):

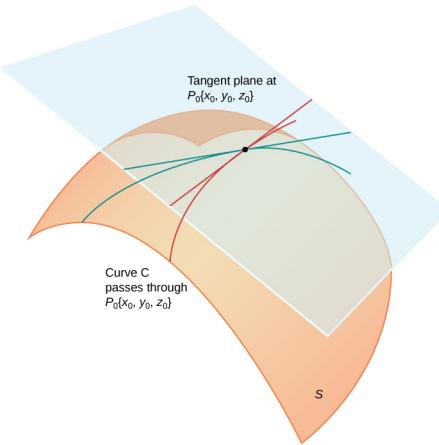
$$\begin{aligned}\frac{\partial^2 f}{\partial x^2} &= \frac{\partial}{\partial x} \left[ \frac{\partial f}{\partial x} \right], \\ \frac{\partial^2 f}{\partial x \partial y} &= \frac{\partial}{\partial x} \left[ \frac{\partial f}{\partial y} \right], \\ \frac{\partial^2 f}{\partial y \partial x} &= \frac{\partial}{\partial y} \left[ \frac{\partial f}{\partial x} \right], \\ \frac{\partial^2 f}{\partial y^2} &= \frac{\partial}{\partial y} \left[ \frac{\partial f}{\partial y} \right]..\end{aligned}$$

**Note:** in this notation, the order in which we take derivatives goes from right to left (of the denominator)

- **Higher-order partial derivatives alternate notation:** An alternative notation for each is  $f_{xx}$ ,  $f_{yx}$ ,  $f_{xy}$ , and  $f_{yy}$ , respectively. Higher-order partial derivatives calculated with respect to different variables, such as  $f_{xy}$  and  $f_{yx}$ , are commonly called mixed partial derivatives.

**Note:** with this notation, the order is from left to right

- **Tangent plane:** Let  $P_0 = (x_0, y_0, z_0)$  be a point on a surface  $S$ , and let  $C$  be any curve passing through  $P_0$  and lying entirely in  $S$ . If the tangent lines to all such curves  $C$  at  $P_0$  lie in the same plane, then this plane is called the tangent plane to  $S$  at  $P_0$ .



- **For a tangent plane to a surface to exist at a point on that surface, it is sufficient for the function that defines the surface to be differentiable at that point,:**
- **Equation of a tangent plane:** Let  $S$  be a surface defined by a differentiable function  $z = f(x, y)$ , and let  $P_0 = (x_0, y_0)$  be a point in the domain of  $f$ . Then, the equation of the tangent plane to  $S$  at  $P_0$  is given by

$$z = f(x_0, y_0) + f_x(x_0, y_0)(x - x_0) + f_y(x_0, y_0)(y - y_0).$$

- **Linear approximation with tangent planes:** Given a function  $z = f(x, y)$  with continuous partial derivatives that exist at the point  $(x_0, y_0)$ , the linear approximation of  $f$  at the point  $(x_0, y_0)$  is given by the equation

$$L(x, y) = f(x_0, y_0) + f_x(x_0, y_0)(x - x_0) + f_y(x_0, y_0)(y - y_0)$$

- **a surface is considered to be smooth at point  $P$  if a tangent plane to the surface exists at that point.:**
- **For a tangent plane to exist at the point  $(x_0, y_0)$  the partial derivatives must therefore exist at that point. However, this is not a sufficient condition for smoothness,:**
- **Differentiability (book definition):** A function  $f(x, y)$  is differentiable at a point  $P(x_0, y_0)$  if, for all points  $(x, y)$  in a  $\delta$  disk around  $P$ , we can write

$$f(x, y) = f(x_0, y_0) + f_x(x_0, y_0)(x - x_0) + f_y(x_0, y_0)(y - y_0) + E(x, y) \quad (50)$$

where the error term  $E$  satisfies

$$\lim_{(x,y) \rightarrow (x_0,y_0)} \frac{E(x, y)}{\sqrt{(x - x_0)^2 + (y - y_0)^2}} = 0.$$

- **Differentiability (Professor's definition):** The geometric concept of having a tangent plane at  $(x_0, y_0)$  is equivalent to the approximation  $f(x, y) \approx L(x, y)$  being sufficiently good. When this happens, we call  $f(x, y)$  differentiable at  $(x_0, y_0)$ . Formally,

$$\lim_{(x,y) \rightarrow (x_0,y_0)} E(x, y) = 0.$$

Where

$$E(x, y) = \frac{f(x, y) - L(x, y)}{|(x, y) - (x_0, y_0)|}.$$

- **Differentiability criteria:**  $f(x, y)$  is differentiable at  $(x_0, y_0)$  iff

1.  $f(x, y) = L(x) + E(x, y) \cdot |(x, y) - (x_0, y_0)|$
2.  $\lim_{(x,y) \rightarrow (x_0,y_0)} E(x, y) = 0$

Differentiability for functions of several variables (e.g.,  $f(x, y)$ ) extends the concept from single-variable calculus. A function  $f(x, y)$  is differentiable at a point  $(x_0, y_0)$  if it can be well approximated by a linear function (the tangent plane) near that point.

- **Show that a function is differentiable:** Show that

$$f(x, y) = 2x^2 - 4y.$$

Is differentiable at the point  $(2, -3)$

First, we find  $f(x, y)$ ,  $f_x(x, y)$ , and  $f_y(x, y)$ . We find

$$\begin{aligned} f(2, -3) &= 20 \\ f_x(2, -3) &= 8 \\ f_y(2, -3) &= -4. \end{aligned}$$

This implies

$$\begin{aligned} f(x, y) &= L(x) + E(x, y) \\ \implies E(x, y) &= f(x, y) - L(x, y) \\ \therefore E(x, y) &= 2x^2 - 8x + 8. \end{aligned}$$

Now we need to show that  $\lim_{(x,y) \rightarrow (x_0, y_0)} \frac{E(x, y)}{\sqrt{(x-x_0)^2 + (y-y_0)^2}} = 0$ . Thus,

$$\begin{aligned} &\lim_{(x,y) \rightarrow (2, -3)} \frac{2x^2 - 8x + 8}{\sqrt{(x-2)^2 + (y+3)^2}} \\ &= \lim_{(x,y) \rightarrow (2, -3)} \frac{2(x-2)^2}{\sqrt{(x-2)^2 + (y+3)^2}} \\ &\leqslant \lim_{(x,y) \rightarrow (2, -3)} \frac{2((x-2)^2 + (y+3)^2)}{\sqrt{(x-2)^2 + (y+3)^2}} \\ &= \lim_{(x,y) \rightarrow (2, -3)} 2\sqrt{(x-2)^2 + (y+3)^2} \\ &= 0. \end{aligned}$$

Since  $E(x, y) \geq 0$  for any value of  $x$  or  $y$ , the original limit must be equal to zero. Therefore,  $f(x, y) = 2x^2 - 4y$  is differentiable at point  $(2, -3)$

- **Differentiability Implies Continuity:** Let  $z = f(x, y)$  be a function of two variables with  $(x_0, y_0)$  in the domain of  $f$ . If  $f(x, y)$  is differentiable at  $(x_0, y_0)$ , then  $f(x, y)$  is continuous at  $(x_0, y_0)$ .
- **Continuity of First Partials Implies Differentiability:** Let  $z = f(x, y)$  be a function of two variables with  $(x_0, y_0)$  in the domain of  $f$ . If  $f(x, y)$ ,  $f_x(x, y)$ , and  $f_y(x, y)$  all exist in a neighborhood of  $(x_0, y_0)$  and are continuous at  $(x_0, y_0)$ , then  $f(x, y)$  is differentiable there.
- **Total differential:** Let  $z = f(x, y)$  be a function of two variables with  $(x_0, y_0)$  in the domain of  $f$ , and let  $\Delta x$  and  $\Delta y$  be chosen so that  $(x_0 + \Delta x, y_0 + \Delta y)$  is also in the domain of  $f$ . If  $f$  is differentiable at the point  $(x_0, y_0)$ , then the differentials  $dx$  and  $dy$  are defined as

$$dx = \Delta x \quad \text{and} \quad dy = \Delta y.$$

The differential  $dz$ , also called the total differential of  $z = f(x, y)$  at  $(x_0, y_0)$ , is defined as

$$dz = f_x(x_0, y_0)dx + f_y(x_0, y_0)dy.$$

- $\Delta z$ :

$$\Delta z = f(x + \Delta x, y + \Delta y) - f(x, y).$$

We use  $dz$  to approximate  $\Delta z$ , so

$$\Delta z \approx dz = f_x(x_0, y_0)dx + f_y(x_0, y_0)dy.$$

With can further approximate with

$$\begin{aligned} f(x + \Delta x, y + \Delta y) &= f(x, y) + \Delta z \\ &\approx f(x, y) + f_x(x_0, y_0)\Delta x + f_y(x_0, y_0)\Delta y. \end{aligned}$$

- **Chain Rule for One Independent Variable:** Suppose that  $x = g(t)$  and  $y = h(t)$  are differentiable functions of  $t$  and  $z = f(x, y)$  is a differentiable function of  $x$  and  $y$ . Then  $z = f(x(t), y(t))$  is a differentiable function of  $t$  and

$$\frac{dz}{dt} = \frac{\partial z}{\partial x} \cdot \frac{dx}{dt} + \frac{\partial z}{\partial y} \cdot \frac{dy}{dt},$$

where the ordinary derivatives are evaluated at  $t$  and the partial derivatives are evaluated at  $(x, y)$ .

- **Chain Rule for Two Independent Variables:** Suppose  $x = g(u, v)$  and  $y = h(u, v)$  are differentiable functions of  $u$  and  $v$ , and  $z = f(x, y)$  is a differentiable function of  $x$  and  $y$ . Then,  $z = f(g(u, v), h(u, v))$  is a differentiable function of  $u$  and  $v$ , and

$$\frac{\partial z}{\partial u} = \frac{\partial z}{\partial x} \frac{\partial x}{\partial u} + \frac{\partial z}{\partial y} \frac{\partial y}{\partial u}$$

and

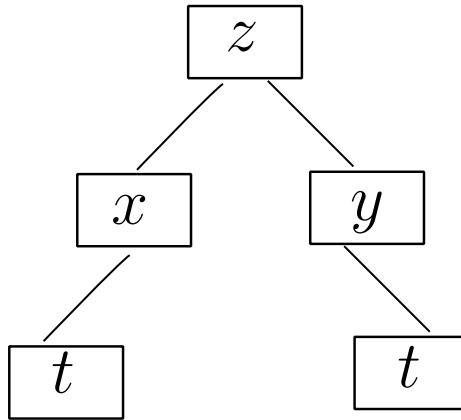
$$\frac{\partial z}{\partial v} = \frac{\partial z}{\partial x} \frac{\partial x}{\partial v} + \frac{\partial z}{\partial y} \frac{\partial y}{\partial v}.$$

- **Generalized Chain Rule:** Let  $w = f(x_1, x_2, \dots, x_m)$  be a differentiable function of  $m$  independent variables, and for each  $i \in \{1, \dots, m\}$ , let  $x_i = x_i(t_1, t_2, \dots, t_n)$  be a differentiable function of  $n$  independent variables. Then

$$\frac{\partial w}{\partial t_j} = \frac{\partial w}{\partial x_1} \frac{\partial x_1}{\partial t_j} + \frac{\partial w}{\partial x_2} \frac{\partial x_2}{\partial t_j} + \dots + \frac{\partial w}{\partial x_m} \frac{\partial x_m}{\partial t_j}$$

for any  $j \in \{1, 2, \dots, n\}$ .

- **Chain rule by a tree diagram.** Suppose we have some function  $z = f(x, y)$ , where  $x = g(t)$  and  $y = h(t)$ , so  $z$  has two variables  $x$  and  $y$ , both of which depend on  $t$ . To find the formula for the derivative  $\frac{dz}{dt}$  :, we can create a tree diagram



The way we use the diagram is simple. We want to find  $\frac{dz}{dt}$ , which you notice is an ordinary derivative, because the leaf nodes are all the same variable. First, we start with the left side, we take derivatives all the way down, multiplying, and then the same for the right side. We sum the two sides. Thus this becomes

$$\frac{dz}{dt} = \frac{\delta z}{\delta x} \cdot \frac{dx}{dt} + \frac{\delta z}{\delta y} \cdot \frac{dy}{dt}.$$

- **Changing variables after computing derivative:** After we compute the derivative, we need to change all instances of  $x$  and  $y$  by using the equation that these variables are equal to. This way our final answer is in terms of the leaf node variables (see the tree diagram above)
- **Implicit Differentiation of a Function of Two or More Variables:** Suppose the function  $z = f(x, y)$  defines  $y$  implicitly as a function  $y = g(x)$  of  $x$  via the equation  $f(x, y) = 0$ . Then

$$\frac{dy}{dx} = -\frac{\partial f / \partial x}{\partial f / \partial y}$$

provided  $\partial f / \partial y(x, y) \neq 0$ .

If the equation  $f(x, y, z) = 0$  defines  $z$  implicitly as a differentiable function of  $x$  and  $y$ , then

$$\frac{\partial z}{\partial x} = -\frac{\partial f / \partial x}{\partial f / \partial z} \quad \text{and} \quad \frac{\partial z}{\partial y} = -\frac{\partial f / \partial y}{\partial f / \partial z}$$

as long as  $\partial f / \partial z(x, y, z) \neq 0$ .

- **Tangent plane for a surface defined implicitly:**

$$F_x(x_0, y_0, z_0)(x - x_0) + F_y(x_0, y_0, z_0)(y - y_0) + F_z(x_0, y_0, z_0)(z - z_0) = 0.$$

- **Direction Derivatives limit definition:** Suppose  $z = f(x, y)$  is a function of two variables with a domain of  $D$ . Let  $(a, b) \in D$  and define  $\mathbf{u} = \cos \theta \mathbf{i} + \sin \theta \mathbf{j}$ . Then the directional derivative of  $f$  in the direction of  $\mathbf{u}$  is given by

$$D_{\mathbf{u}}f(a, b) = \lim_{h \rightarrow 0} \frac{f(a + h \cos \theta, b + h \sin \theta) - f(a, b)}{h}, \quad (51)$$

provided the limit exists.

- **Directional derivatives with partial derivatives:** Let  $z = f(x, y)$  be a function of two variables  $x$  and  $y$ , and assume that  $f_x$  and  $f_y$  exist and  $f(x, y)$  is differentiable everywhere. Then the directional derivative of  $f$  in the direction of  $\mathbf{u} = \cos \theta \mathbf{i} + \sin \theta \mathbf{j}$  is given by

$$D_{\mathbf{u}}f(x, y) = f_x(x, y) \cos \theta + f_y(x, y) \sin \theta.$$

- **Gradient:** Let  $z = f(x, y)$  be a function of  $x$  and  $y$  such that  $f_x$  and  $f_y$  exist. The vector  $\nabla f(x, y)$  is called the gradient of  $f$  and is defined as

$$\nabla f(x, y) = f_x(x, y) \mathbf{i} + f_y(x, y) \mathbf{j}. \quad (52)$$

The vector  $\nabla f(x, y)$  is also written as “grad  $f$ .”

- **Divide by the norm:** If the vector that is given for the direction of the derivative is not a unit vector, then it is only necessary to divide by the norm of the vector.

- **Directional derivative with the gradient:**

$$D_u f(x, y) = \nabla f(x, y) \cdot \mathbf{u}.$$

- **Properties of the Gradient:** Suppose the function  $z = f(x, y)$  is differentiable at  $(x_0, y_0)$  (Figure 4.41).

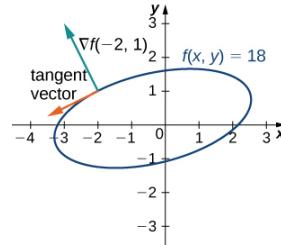
1. If  $\nabla f(x_0, y_0) = 0$ , then  $D_{\mathbf{u}} f(x_0, y_0) = 0$  for any unit vector  $\mathbf{u}$ .
2. If  $\nabla f(x_0, y_0) \neq 0$ , then  $D_{\mathbf{u}} f(x_0, y_0)$  is maximized when  $\mathbf{u}$  points in the same direction as  $\nabla f(x_0, y_0)$ . The maximum value of  $D_{\mathbf{u}} f(x_0, y_0)$  is  $\|\nabla f(x_0, y_0)\|$ .
3. If  $\nabla f(x_0, y_0) \neq 0$ , then  $D_{\mathbf{u}} f(x_0, y_0)$  is minimized when  $\mathbf{u}$  points in the opposite direction from  $\nabla f(x_0, y_0)$ . The minimum value of  $D_{\mathbf{u}} f(x_0, y_0)$  is  $-\|\nabla f(x_0, y_0)\|$ .

- **Gradient Is Normal to the Level Curve:** Suppose the function  $z = f(x, y)$  has continuous first-order partial derivatives in an open disk centered at a point  $(x_0, y_0)$ . If  $\nabla f(x_0, y_0) \neq 0$ , then  $\nabla f(x_0, y_0)$  is normal to the level curve of  $f$  at  $(x_0, y_0)$ .

- **Finding a tangent vector with normal vector:** Suppose we use the theorem above to find the vector tangent to a level curve at some point, we can then find a tangent vector by reversing the components and multiplying either one by negative one.

Example:

$$\begin{aligned}\nabla f(-2, 1) &= 9\hat{\mathbf{i}} + 18\hat{\mathbf{j}} \quad (\text{would be normal to some level curve}) \\ -18\hat{\mathbf{i}} - 9\hat{\mathbf{j}} &\quad \text{Tangent vector at } (-2, 1).\end{aligned}$$



- **Gradient in three variables:** Let  $w = f(x, y, z)$  be a function of three variables such that  $f_x$ ,  $f_y$ , and  $f_z$  exist. The vector  $\nabla f(x, y, z)$  is called the gradient of  $f$  and is defined as

$$\nabla f(x, y, z) = f_x(x, y, z)\mathbf{i} + f_y(x, y, z)\mathbf{j} + f_z(x, y, z)\mathbf{k}. \quad (53)$$

$\nabla f(x, y, z)$  can also be written as  $\text{grad}f(x, y, z)$ .

- **Directional derivative for functions of three variables (limit definition):** Suppose  $w = f(x, y, z)$  is a function of three variables with a domain of  $D$ . Let  $(x_0, y_0, z_0) \in D$  and let  $\mathbf{u} = \cos \alpha \mathbf{i} + \cos \beta \mathbf{j} + \cos \gamma \mathbf{k}$  be a unit vector. Then, the directional derivative of  $f$  in the direction of  $\mathbf{u}$  is given by

$$D_{\mathbf{u}}f(x_0, y_0, z_0) = \lim_{t \rightarrow 0} \frac{f(x_0 + t \cos \alpha, y_0 + t \cos \beta, z_0 + t \cos \gamma) - f(x_0, y_0, z_0)}{t}, \quad (54)$$

provided the limit exists.

**Note:** The components of the unit vector are called the **directional cosines**

- **Directional derivative for functions of three variables (p.d definition):** Let  $f(x, y, z)$  be a differentiable function of three variables and let  $\mathbf{u} = \cos \alpha \mathbf{i} + \cos \beta \mathbf{j} + \cos \gamma \mathbf{k}$  be a unit vector. Then, the directional derivative of  $f$  in the direction of  $\mathbf{u}$  is given by

$$\begin{aligned} D_{\mathbf{u}}f(x, y, z) &= \nabla f(x, y, z) \cdot \mathbf{u} \\ &= f_x(x, y, z) \cos \alpha + f_y(x, y, z) \cos \beta + f_z(x, y, z) \cos \gamma. \end{aligned}$$

- **Critical points:** Let  $z = f(x, y)$  be a function of two variables that is defined on an open set containing the point  $(x_0, y_0)$ . The point  $(x_0, y_0)$  is called a critical point of a function of two variables  $f$  if one of the two following conditions holds:

1.  $f_x(x_0, y_0) = f_y(x_0, y_0) = 0$
2. Either  $f_x(x_0, y_0)$  or  $f_y(x_0, y_0)$  does not exist.

- **Local max / Absolute max (global max):** Let  $z = f(x, y)$  be a function of two variables that is defined and continuous on an open set containing the point  $(x_0, y_0)$ . Then  $f$  has a local maximum at  $(x_0, y_0)$  if

$$f(x_0, y_0) \geq f(x, y)$$

for all points  $(x, y)$  within some disk centered at  $(x_0, y_0)$ . The number  $f(x_0, y_0)$  is called a local maximum value. If the preceding inequality holds for every point  $(x, y)$  in the domain of  $f$ , then  $f$  has a global maximum (also called an absolute maximum) at  $(x_0, y_0)$ .

- **Local min / Absolute min (global min):** The function  $f$  has a local minimum at  $(x_0, y_0)$  if

$$f(x_0, y_0) \leq f(x, y)$$

for all points  $(x, y)$  within some disk centered at  $(x_0, y_0)$ . The number  $f(x_0, y_0)$  is called a local minimum value. If the preceding inequality holds for every point  $(x, y)$  in the domain of  $f$ , then  $f$  has a global minimum (also called an absolute minimum) at  $(x_0, y_0)$ .

- **Local Extremum:** If  $f(x_0, y_0)$  is either a local maximum or local minimum value, then it is called a local extremum.
- **Fermat's Theorem for Functions of Two Variables:** Let  $z = f(x, y)$  be a function of two variables that is defined and continuous on an open set containing the point  $(x_0, y_0)$ . Suppose  $f_x$  and  $f_y$  each exists at  $(x_0, y_0)$ . If  $f$  has a local extremum at  $(x_0, y_0)$ , then  $(x_0, y_0)$  is a critical point of  $f$ .
- **Saddle point:** Given the function  $z = f(x, y)$ , the point  $(x_0, y_0, f(x_0, y_0))$  is a saddle point if both  $f_x(x_0, y_0) = 0$  and  $f_y(x_0, y_0) = 0$ , but  $f$  does not have a local extremum at  $(x_0, y_0)$ .

- **Second derivative test:** Let  $z = f(x, y)$  be a function of two variables for which the first- and second-order partial derivatives are continuous on some disk containing the point  $(x_0, y_0)$ . Suppose  $f_x(x_0, y_0) = 0$  and  $f_y(x_0, y_0) = 0$ . Define the quantity

$$D = f_{xx}(x_0, y_0)f_{yy}(x_0, y_0) - (f_{xy}(x_0, y_0))^2.$$

- I. If  $D > 0$  and  $f_{xx}(x_0, y_0) > 0$ , then  $f$  has a local minimum at  $(x_0, y_0)$ .
- II. If  $D > 0$  and  $f_{xx}(x_0, y_0) < 0$ , then  $f$  has a local maximum at  $(x_0, y_0)$ .
- III. If  $D < 0$ , then  $f$  has a saddle point at  $(x_0, y_0)$ .
- IV. If  $D = 0$ , then the test is inconclusive.



- **Extreme Value Theorem:** A continuous function  $f(x, y)$  on a closed and bounded set  $D$  in the plane attains an absolute maximum value at some point of  $D$  and an absolute minimum value at some point of  $D$ .
- **Finding extreme values:** Assume  $z = f(x, y)$  is a differentiable function of two variables defined on a closed, bounded set  $D$ . Then  $f$  will attain the absolute maximum value and the absolute minimum value, which are, respectively, the largest and smallest values found among the following:
  1. The values of  $f$  at the critical points of  $f$  in  $D$ .
  2. The values of  $f$  on the boundary of  $D$ .
- **Method of Lagrange Multipliers: One Constraint:** Let  $f$  and  $g$  be functions of two variables with continuous partial derivatives at every point of some open set containing the smooth curve  $g(x, y) = 0$ . Suppose that  $f$ , when restricted to points on the curve  $g(x, y) = 0$ , has a local extremum at the point  $(x_0, y_0)$  and that  $\nabla g(x_0, y_0) \neq 0$ . Then there is a number  $\lambda$  called a Lagrange multiplier, for which

$$\nabla f(x_0, y_0) = \lambda \nabla g(x_0, y_0).$$

- **Problem-Solving Strategy: Steps for Using Lagrange Multipliers:** Follow these steps to solve the optimization problem using Lagrange multipliers:
  1. Determine the objective function  $f(x, y)$  and the constraint function  $g(x, y)$ . Does the optimization problem involve maximizing or minimizing the objective function?
  2. Set up a system of equations using the following template:

$$\begin{aligned} \nabla f(x_0, y_0) &= \lambda \nabla g(x_0, y_0) \\ g(x_0, y_0) &= 0. \end{aligned}$$

3. Solve for  $x_0$  and  $y_0$ .

4. The largest of the values of  $f$  at the solutions found in step 3 maximizes  $f$ ; the smallest of those values minimizes  $f$ .

- **Problems with Two Constraints:** The method of Lagrange multipliers can be applied to problems with more than one constraint. In this case, the optimization function,  $w$ , is a function of three variables:

$$w = f(x, y, z)$$

and it is subject to two constraints:

$$g(x, y, z) = 0 \quad \text{and} \quad h(x, y, z) = 0.$$

There are two Lagrange multipliers,  $\lambda_1$  and  $\lambda_2$ , and the system of equations becomes

$$\begin{aligned}\nabla f(x_0, y_0, z_0) &= \lambda_1 \nabla g(x_0, y_0, z_0) + \lambda_2 \nabla h(x_0, y_0, z_0) \\ g(x_0, y_0, z_0) &= 0 \\ h(x_0, y_0, z_0) &= 0.\end{aligned}$$

- **Using gradient to find tangent plane to level surface:** Suppose we have some function  $z = f(x, y)$ , we rearrange the function such that it becomes  $F(x, y, z) = 0$ . We then use the fact that  $\nabla F(x_0, y_0, z_0)$  is normal to the surface  $F(x, y, z)$  at  $P_0$ :

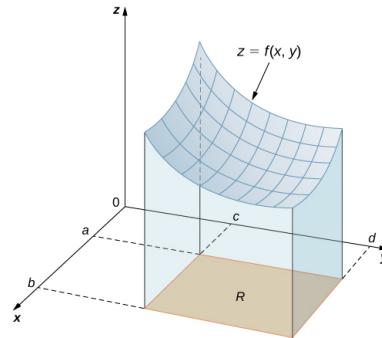
## 4.6 Chapter 5: Multiple integration

### 4.6.1 Definitions and theorems

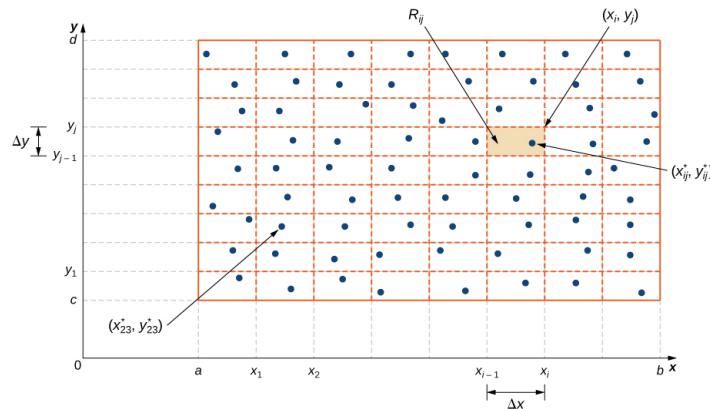
- **Double integral intro 1: Rectangular region  $R$  and solid  $S$ :** Consider a continuous function  $f(x, y) \geq 0$  of two variables defined on the closed rectangle  $R$ :

$$R = [a, b] \times [c, d] = \{(x, y) \in \mathbb{R}^2 \mid a \leq x \leq b, c \leq y \leq d\}$$

The graph of  $f$  represents a surface above the  $xy$ -plane with equation  $z = f(x, y)$  where  $z$  is the height of the surface at the point  $(x, y)$ . Let  $S$  be the solid that lies above  $R$  and under the graph of  $f$ . The base of the solid is the rectangle  $R$  in the  $xy$ -plane. We want to find the volume  $V$  of the solid  $S$ .



- **Double integral intro 2: Divisions of  $R$ :** We divide the region  $R$  into small rectangles  $R_{ij}$ , each with area  $\Delta A$  and with sides  $\Delta x$  and  $\Delta y$ . We do this by dividing the interval  $[a, b]$  into  $m$  subintervals and dividing the interval  $[c, d]$  into  $n$  subintervals. Hence,  $\Delta x = \frac{b-a}{m}$ ,  $\Delta y = \frac{d-c}{n}$ , and  $\Delta A = \Delta x \Delta y$ .



- **Double integral intro 3: volume of the subregions:** The volume of a thin rectangular box above  $R_{ij}$  is  $f(x_{ij}^*, y_{ij}^*) \Delta A$ , where  $(x_{ij}^*, y_{ij}^*)$  is an arbitrary sample point in each  $R_{ij}$  as shown in the following figure.



- **Double integral intro 4: Double Riemann sum:** Using the same idea for all the subrectangles, we obtain an approximate volume of the solid  $S$  as

$$V \approx \sum_{i=1}^m \sum_{j=1}^n f(x_{ij}^*, y_{ij}^*) \Delta A.$$

This sum is known as a double Riemann sum and can be used to approximate the value of the volume of the solid. Here, the double sum means that for each subrectangle, we evaluate the function at the chosen point, multiply by the area of each rectangle, and then add all the results.

As we have seen in the single-variable case, we obtain a better approximation to the actual volume if  $m$  and  $n$  become larger.

$$V = \lim_{m,n \rightarrow \infty} \sum_{i=1}^m \sum_{j=1}^n f(x_{ij}^*, y_{ij}^*) \Delta A$$

or

$$V = \lim_{\Delta x, \Delta y \rightarrow 0} \sum_{i=1}^m \sum_{j=1}^n f(x_{ij}^*, y_{ij}^*) \Delta A.$$

Note that the sum approaches a limit in either case, and the limit is the volume of the solid with the base  $R$ . Now we are ready to define the double integral.

- **The double integral over a rectangular region:** The double integral of the function  $f(x, y)$  over the rectangular region  $R$  in the  $xy$ -plane is defined as

$$\iint_R f(x, y) dA = \lim_{m,n \rightarrow \infty} \sum_{i=1}^m \sum_{j=1}^n f(x_i^*, y_j^*) \Delta A \quad (55)$$

**Note:** If  $f(x, y) \geq 0$ , then the volume  $V$  of the solid  $S$ , which lies above  $R$  in the  $xy$ -plane and under the graph of  $f$ , is the double integral of the function  $f(x, y)$  over the rectangle  $R$ . If the function is ever negative, then the double integral can be considered a “signed” volume in a manner similar to the way we defined net signed area in The Definite Integral.

**Example:** Suppose we have the surface defined by  $z = f(x, y) = 3x^2 + y$  with the region  $[0, 2] \times [0, 2]$

$$\iint_R 3x^2 + y dA = \lim_{m,n \rightarrow \infty} \sum_{i=1}^m \sum_{j=1}^n [(x_i^*)^2 + y_j^*] \Delta A.$$

- **Double integral exists and function is integrable:** The double integral of the function  $z = f(x, y)$  exists provided that the function  $f$  is not too discontinuous. If the function is bounded and continuous over  $R$  except on a finite number of smooth curves, then the double integral exists and we say that  $f$  is integrable over  $R$ .
- **Double integral with  $dx$  and  $dy$ :** Since  $\Delta A = \Delta x \Delta y = \Delta y \Delta x$ , we can express  $dA$  as  $dx dy$  or  $dy dx$ . This means that, when we are using rectangular coordinates, the double integral over a region  $R$  denoted by  $\iint_R f(x, y) dA$  can be written as  $\iint_R f(x, y) dx dy$  or  $\iint_R f(x, y) dy dx$ .
- **Properties of double integrals:** Assume that the functions  $f(x, y)$  and  $g(x, y)$  are integrable over the rectangular region  $R$ ;  $S$  and  $T$  are subregions of  $R$ ; and assume that  $m$  and  $M$  are real numbers.

I. The sum  $f(x, y) + g(x, y)$  is integrable and

$$\iint_R [f(x, y) + g(x, y)] dA = \iint_R f(x, y) dA + \iint_R g(x, y) dA.$$

II. If  $c$  is a constant, then  $cf(x, y)$  is integrable and

$$\iint_R cf(x, y) dA = c \iint_R f(x, y) dA.$$

III. If  $R = S \cup T$  and  $S \cap T = \emptyset$  except an overlap on the boundaries, then

$$\iint_R f(x, y) dA = \iint_S f(x, y) dA + \iint_T f(x, y) dA.$$

IV. If  $f(x, y) \geq g(x, y)$  for  $(x, y)$  in  $R$ , then

$$\iint_R f(x, y) dA \geq \iint_R g(x, y) dA.$$

V. If  $m \leq f(x, y) \leq M$ , then

$$m \times A(R) \leq \iint_R f(x, y) dA \leq M \times A(R).$$

VI. In the case where  $f(x, y)$  can be factored as a product of a function  $g(x)$  of  $x$  only and a function  $h(y)$  of  $y$  only, then over the region  $R = \{(x, y) | a \leq x \leq b, c \leq y \leq d\}$ , the double integral can be written as

$$\iint_R f(x, y) dA = \left( \int_a^b g(x) dx \right) \left( \int_c^d h(y) dy \right).$$

**Example:** Evaluate the integral  $\iint_R ye^x \cos(x) dA$  over the region  $R = \{(x, y) | 0 \leq x \leq \frac{\pi}{2}, 0 \leq y \leq 1\}$ .

$$\begin{aligned} & \left( \int_0^{\frac{\pi}{2}} \cos(x) dx \right) \left( \int_0^1 e^y dy \right) \\ &= e - 1. \end{aligned}$$

- **Illustrating property V:** Over the region  $R = \{(x, y) \mid 1 \leq x \leq 3, 1 \leq y \leq 2\}$ , we have  $2 \leq x^2 + y^2 \leq 13$ . Find a lower and an upper bound for the integral  $\iint_R (x^2 + y^2) dA$ . For a lower bound, integrate the constant function 2 over the region  $R$ . For an upper bound, integrate the constant function 13 over the region  $R$ .

$$\begin{aligned} \int_1^2 \int_1^3 2 dx dy &= \int_1^2 [2x]_1^3 dy = \int_1^2 2(2) dy = 4y \Big|_1^2 = 4(2 - 1) = 4 \\ \int_1^2 \int_1^3 13 dx dy &= \int_1^2 [13x]_1^3 dy = \int_1^2 13(2) dy = 26y \Big|_1^2 = 26(2 - 1) = 26.. \end{aligned}$$

Hence, we obtain  $4 \leq \iint_R (x^2 + y^2) dA \leq 26$ .

- **Iterated integrals:** Assume  $a, b, c$ , and  $d$  are real numbers. We define an iterated integral for a function  $f(x, y)$  over the rectangular region  $R = [a, b] \times [c, d]$  as

(a)

$$\int_a^b \int_c^d f(x, y) dy dx = \int_a^b \left[ \int_c^d f(x, y) dy \right] dx.$$

(b)

$$\int_c^d \int_a^b f(x, y) dx dy = \int_c^d \left[ \int_a^b f(x, y) dx \right] dy.$$

**Note:** The notation

$$\int_a^b \left[ \int_c^d f(x, y) dy \right] dx$$

means that we integrate  $f(x, y)$  with respect to  $y$  while holding  $x$  constant. Similarly, the notation

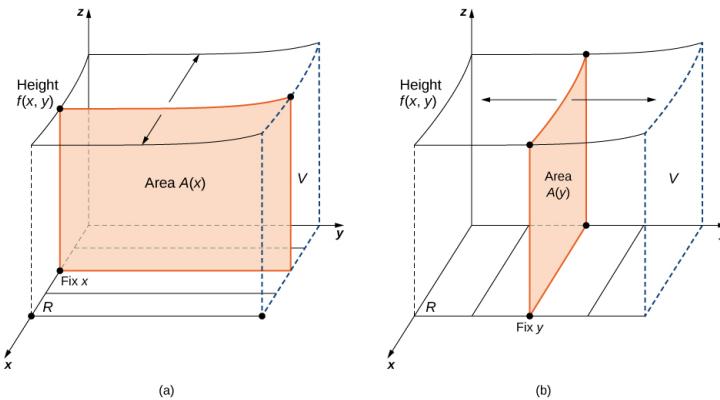
$$\int_c^d \left[ \int_a^b f(x, y) dx \right] dy$$

means that we integrate  $f(x, y)$  with respect to  $x$  while holding  $y$  constant.

- **Fubini's Theorem:** Suppose that  $f(x, y)$  is a function of two variables that is continuous over a rectangular region  $R = \{(x, y) \in \mathbb{R}^2 \mid a \leq x \leq b, c \leq y \leq d\}$ . Then we see from the figure below that the double integral of  $f$  over the region equals an iterated integral,

$$\iint_R f(x, y) dA = \iint_R f(x, y) dx dy = \int_a^b \left( \int_c^d f(x, y) dy \right) dx = \int_c^d \left( \int_a^b f(x, y) dx \right) dy.$$

More generally, Fubini's theorem is true if  $f$  is bounded on  $R$  and  $f$  is discontinuous only on a finite number of continuous curves. In other words,  $f$  has to be integrable over  $R$ .



- **Area of a region  $R$ :** The area of the region  $R$  is given by

$$A(R) = \iint_R 1 \, dA.$$

- **Recall:** Average value of a function of one variable: the average value of a function of one variable on an interval  $[a, b]$  is given by

$$\bar{f} = \frac{1}{b-a} \int_a^b f(x) \, dx.$$

- **Average value of a function of two variables:** The average value of a function of two variables over a region  $R$  is given by

$$\bar{f} = \frac{1}{A(R)} \iint_R f(x, y) dA.$$

- **Non rectangular region  $D$ :** Since  $D$  is bounded on the plane, there must exist a rectangular region  $R$  on the same plane that encloses the region  $D$ , that is, a rectangular region  $R$  exists such that  $D$  is a subset of  $R$  ( $D \subseteq R$ ).

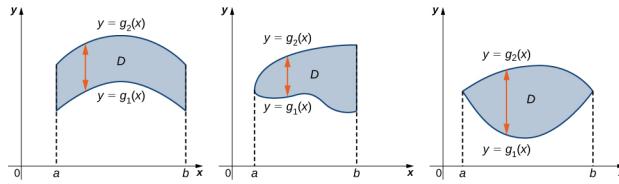
We extend the definition of the function to include all points on the rectangular region  $R$  and then use the concepts and tools from the preceding section. But how do we extend the definition of  $f$  to include all the points on  $R$ ? We do this by defining a new function  $g(x, y)$  on  $R$  as follows:

$$g(x, y) = \begin{cases} f(x, y) & \text{if } (x, y) \text{ is in } D \\ 0 & \text{if } (x, y) \text{ is in } R \text{ but not in } D \end{cases}$$

**Note:** we assume the boundary to be a piecewise smooth and continuous simple closed curve. We must be careful about  $g(x, y)$  and verify that  $g(x, y)$  is an integrable function over the rectangular region  $R$ . This happens as long as the region  $D$  is bounded by simple closed curves.

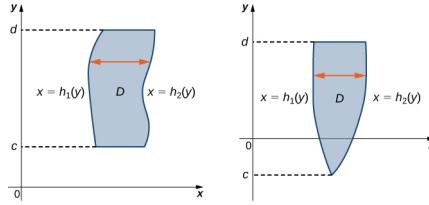
- **Types of planar bounded regions:** A region  $D$  in the  $(x, y)$ -plane is of **Type I** if it lies between two vertical lines and the graphs of two continuous functions  $g_1(x)$  and  $g_2(x)$ .

$$D = \{(x, y) \mid a \leq x \leq b, g_1(x) \leq y \leq g_2(x)\}.$$



A region  $D$  in the  $xy$ -plane is of **Type II** if it lies between two horizontal lines and the graphs of two continuous functions  $h_1(y)$  and  $h_2(y)$ .

$$D = \{(x, y) \mid c \leq y \leq d, h_1(y) \leq x \leq h_2(y)\}.$$



- **Double Integrals over Nonrectangular Regions:** Suppose  $g(x, y)$  is the extension to the rectangle  $R$  of the integrable function  $f(x, y)$  defined on the region  $D$ , where  $D$  is inside  $R$ . Then  $g(x, y)$  is integrable and we define the double integral of  $f(x, y)$  over  $D$  by

$$\iint_D f(x, y) dA = \iint_R g(x, y) dA.$$

**Note:** The equality works because the values of  $g(x, y)$  are 0 for any point  $(x, y)$  that lies outside  $D$ , and hence these points do not add anything to the integral. However, it is important that the rectangle  $R$  contains the region  $D$ .

- **Fubini's Theorem (Strong Form):** For a function  $f(x, y)$  that is continuous on a region  $D$  of Type I, we have

$$\iint_D f(x, y) dA = \iint_D f(x, y) dy dx = \int_a^b \left[ \int_{g_1(x)}^{g_2(x)} f(x, y) dy \right] dx. \quad (56)$$

Similarly, for a function  $f(x, y)$  that is continuous on a region  $D$  of Type II, we have

$$\iint_D f(x, y) dA = \iint_D f(x, y) dx dy = \int_c^d \left[ \int_{h_1(y)}^{h_2(y)} f(x, y) dx \right] dy. \quad (57)$$

- **Decomposing Regions into Smaller Regions:** Suppose the region  $D$  can be expressed as  $D = D_1 \cup D_2$  where  $D_1$  and  $D_2$  do not overlap except at their boundaries. Then

$$\iint_D f(x, y) dA = \iint_{D_1} f(x, y) dA + \iint_{D_2} f(x, y) dA.$$

- **We can always change our region from type 1 to type 2 to make our iterated integral easier to solve:**

- **Area of a plane bounded by a region  $D$ :** The area of a plane-bounded region  $D$  is defined as the double integral  $\iint_D 1 dA$ .

- **Average value of a function over a general region:** If  $f(x, y)$  is integrable over a plane-bounded region  $D$  with positive area  $A(D)$ , then the average value of the function is

$$f_{\text{ave}} = \frac{1}{A(D)} \iint_D f(x, y) dA.$$

Note that the area is  $A(D) = \iint_D 1 dA$ .

- **Fubini's Theorem for Improper Integrals:** If  $D$  is a bounded rectangle or simple region in the plane defined by  $\{(x, y) : a \leq x \leq b, g(x) \leq y \leq h(x)\}$  and also by  $\{(x, y) : c \leq y \leq d, j(y) \leq x \leq k(y)\}$ , and  $f$  is a nonnegative function on  $D$  with finitely many discontinuities in the interior of  $D$ , then

$$\iint_D f dA = \int_{x=a}^{x=b} \int_{y=g(x)}^{y=h(x)} f(x, y) dy dx = \int_{y=c}^{y=d} \int_{x=j(y)}^{x=k(y)} f(x, y) dx dy.$$

**Note:** It is very important to note that we required that the function be nonnegative on  $D$  for the theorem to work. We consider only the case where the function has finitely many discontinuities inside  $D$ .

- **Improper Integrals on an Unbounded Region:** If  $R$  is an unbounded rectangle such as  $R = \{(x, y) : a \leq x < \infty, c \leq y < \infty\}$ , then when the limit exists, we have

$$\begin{aligned} \iint_R f(x, y) dA &= \lim_{(b,d) \rightarrow (\infty, \infty)} \int_a^b \left( \int_c^d f(x, y) dy \right) dx \\ &= \lim_{(b,d) \rightarrow (\infty, \infty)} \int_c^d \left( \int_a^b f(x, y) dx \right) dy. \end{aligned}$$

$$\lim_{(b,d) \rightarrow (\infty, \infty)} \frac{1}{4} (1 - e^{-b^2})(1 - e^{-d^2}).$$

- **Joint density function:** Consider a pair of continuous random variables  $X$  and  $Y$ , such as the birthdays of two people or the number of sunny and rainy days in a month. The joint density function  $f$  of  $X$  and  $Y$  satisfies the probability that  $(X, Y)$  lies in a certain region  $D$ :

$$P((X, Y) \in D) = \iint_D f(x, y) dA.$$

Since the probabilities can never be negative and must lie between 0 and 1, the joint density function satisfies the following inequality and equation:

$$f(x, y) \geq 0 \quad \text{and} \quad \iint_{\mathbb{R}^2} f(x, y) dA = 1.$$

- **Independent random variable classification:** The variables  $X$  and  $Y$  are said to be independent random variables if their joint density function is the product of their individual density functions:

$$f(x, y) = f_1(x)f_2(y).$$

- **Probability theory expected values:** In probability theory, we denote the expected values  $E(X)$  and  $E(Y)$ , respectively, as the most likely outcomes of the events. The expected values  $E(X)$  and  $E(Y)$  are given by

$$E(X) = \iint_S xf(x, y) dA \quad \text{and} \quad E(Y) = \iint_S yf(x, y) dA,$$

where  $S$  is the sample space of the random variables  $X$  and  $Y$ .

- **Double integral in polar coordinates:** The double integral of the function  $f(r, \theta)$  over the polar rectangular region  $R$  in the  $r\theta$ -plane is defined as

$$\begin{aligned} & \iint_R f(r, \theta) dA \\ &= \lim_{m,n \rightarrow \infty} \sum_{i=1}^m \sum_{j=1}^n f(r_{ij}^*, \theta_{ij}^*) \Delta A \\ &= \lim_{m,n \rightarrow \infty} \sum_{i=1}^m \sum_{j=1}^n f(r_{ij}^*, \theta_{ij}^*) r_{ij}^* \Delta r \Delta \theta. \end{aligned}$$

- **Iterated integral for polar regions:**

$$\begin{aligned} \iint_R f(r, \theta) dA &= \iint_R f(r, \theta) r dr d\theta \\ &= \int_{\theta=\alpha}^{\theta=\beta} \int_{r=a}^{r=b} f(r, \theta) r dr d\theta. \end{aligned}$$

- **Integral when changinig from  $xy$  to  $r\theta$ :**

$$\iint_R f(x, y) dA = \iint_R f(r \cos(\theta), r \sin(\theta)) r dr d\theta.$$

**NOTICE:** The extra  $r$  at the end of the integrand

- **Double Integrals over General Polar Regions:** If  $f(r, \theta)$  is continuous on a general polar region  $D$  as described above, then the double integral over  $D$  can be expressed as:

$$\iint_D f(r, \theta) r dr d\theta = \int_{\theta=\alpha}^{\theta=\beta} \int_{r=h_1(\theta)}^{r=h_2(\theta)} f(r, \theta) r dr d\theta$$

- **Area of a polar region:**

$$A = \int_{\alpha}^{\beta} \int_{h_1(\theta)}^{h_2(\theta)} r dr d\theta.$$

- **Equation of an arbitrary cone with radius  $a$  and height  $h$ :**

$$z = h - \frac{h}{a} \sqrt{x^2 + y^2}.$$

- **Entire  $xy$  plane region to polar region:** Suppose we have the region  $\mathbb{R}^2$ , ie the entire  $xy$ -plane. This can be seen as

$$0 \leq \theta \leq 2\pi, \quad 0 \leq r < \infty.$$

- **Triple integral:** The triple integral of a function  $f(x, y, z)$  over a rectangular box  $B$  is defined as

$$\lim_{l,m,n \rightarrow \infty} \sum_{i=1}^l \sum_{j=1}^m \sum_{k=1}^n f(x_{ijk}^*, y_{ijk}^*, z_{ijk}^*) \Delta x \Delta y \Delta z = \iiint_B f(x, y, z) dV$$

if this limit exists.

**Note:** When the triple integral exists on  $B$ , the function  $f(x, y, z)$  is said to be integrable on  $B$ . Also, the triple integral exists if  $f(x, y, z)$  is continuous on  $B$ . Therefore, we will use continuous functions for our examples. However, continuity is sufficient but not necessary; in other words,  $f$  is bounded on  $B$  and continuous except possibly on the boundary of  $B$ .

The sample point  $(x_{ijk}^*, y_{ijk}^*, z_{ijk}^*)$  can be any point in the rectangular sub-box  $B_{ijk}$  and all the properties of a double integral apply to a triple integral. **Fubini's Theorem for Triple Integrals** If  $f(x, y, z)$  is continuous on a rectangular box  $B = [a, b] \times [c, d] \times [e, f]$ , then

$$\iiint_B f(x, y, z) dV = \int_e^f \int_c^d \int_a^b f(x, y, z) dx dy dz.$$

This integral is also equal to any of the other five possible orderings for the iterated triple integral.

- **Triple integral over general regions:** The triple integral of a continuous function  $f(x, y, z)$  over a general three-dimensional region

$$E = \{(x, y, z) \mid (x, y) \in D, u_1(x, y) \leq z \leq u_2(x, y)\}$$

in  $\mathbb{R}^3$ , where  $D$  is the projection of  $E$  onto the  $xy$ -plane, is

$$\iiint_E f(x, y, z) dV = \iint_D \left[ \int_{u_1(x, y)}^{u_2(x, y)} f(x, y, z) dz \right] dA.$$

- **Volume of a general region  $E$  with triple integrals:** To find the volume of a general region  $E$ , we use

$$\iiint_E dV.$$

- **Average Value of a Function of Three Variables:** If  $f(x, y, z)$  is integrable over a solid bounded region  $E$  with positive volume  $V(E)$ , then the average value of the function is

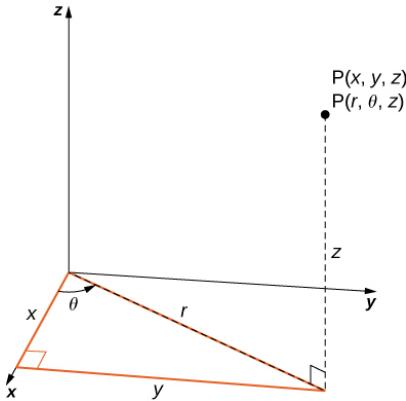
$$f_{\text{ave}} = \frac{1}{V(E)} \iiint_E f(x, y, z) dV.$$

Note that the volume is  $V(E) = \iiint_E 1 dV$ .

- **Equation of a plane with intercepts  $(a, 0, 0)$ ,  $(0, b, 0)$ ,  $(0, 0, c)$ :**

$$\frac{x}{a} + \frac{y}{b} + \frac{z}{c} = 1.$$

- **Cylindrical coordinates:** In three-dimensional space  $\mathbb{R}^3$ , a point with rectangular coordinates  $(x, y, z)$  can be identified with cylindrical coordinates  $(r, \theta, z)$  and vice versa. We can use these same conversion relationships, adding  $z$  as the vertical distance to the point from the  $xy$ -plane as shown in the following figure.



- **Fubini's Theorem in Cylindrical Coordinates:** Suppose that  $g(x, y, z)$  is continuous on a portion of a circular cylinder  $B$ , which when described in cylindrical coordinates looks like  $B = \{(r, \theta, z) \mid a \leq r \leq b, \alpha \leq \theta \leq \beta, c \leq z \leq d\}$ .

Then  $g(x, y, z) = g(r \cos \theta, r \sin \theta, z) = f(r, \theta, z)$  and

$$\iiint_B g(x, y, z) dV = \int_c^d \int_{\alpha}^{\beta} \int_a^b f(r, \theta, z) r dr d\theta dz.$$

**Note:** The iterated integral may be replaced equivalently by any one of the other five iterated integrals obtained by integrating with respect to the three variables in other orders.

- **Cylindrical region is a general solid:** If the cylindrical region over which we have to integrate is a general solid, we look at the projections onto the coordinate planes. Hence the triple integral of a continuous function  $f(r, \theta, z)$  over a general solid region  $E = \{(r, \theta, z) \mid (r, \theta) \in D, u_1(r, \theta) \leq z \leq u_2(r, \theta)\}$  in  $\mathbb{R}^3$ , where  $D$  is the projection of  $E$  onto the  $r\theta$ -plane, is

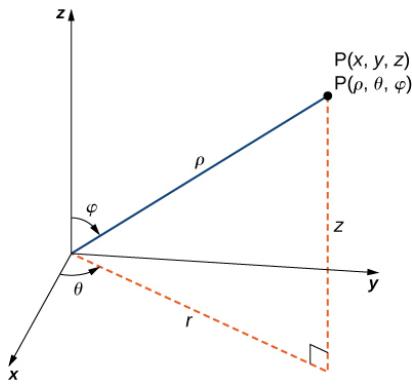
$$\iiint_E f(r, \theta, z) r dr d\theta dz = \iint_D \left[ \int_{u_1(r, \theta)}^{u_2(r, \theta)} f(r, \theta, z) dz \right] r dr d\theta.$$

In particular, if  $D = \{(r, \theta) \mid g_1(\theta) \leq r \leq g_2(\theta), \alpha \leq \theta \leq \beta\}$ , then we have

$$\iiint_E f(r, \theta, z) r dr d\theta dz = \int_{\theta=\alpha}^{\theta=\beta} \int_{r=g_1(\theta)}^{r=g_2(\theta)} \int_{z=u_1(r, \theta)}^{z=u_2(r, \theta)} f(r, \theta, z) r dz dr d\theta.$$

Similar formulas exist for projections onto the other coordinate planes. We can use polar coordinates in those planes if necessary.

- **Spherical coordinates:** In three-dimensional space  $\mathbb{R}^3$  in the spherical coordinate system, we specify a point  $P$  by its distance  $\rho$  from the origin, the polar angle  $\theta$  from the positive  $x$ -axis (same as in the cylindrical coordinate system), and the angle  $\phi$  from the positive  $z$ -axis and the line  $OP$ . Note that  $\rho \geq 0$  and  $0 \leq \phi \leq \pi$ .



**Note:** Spherical coordinates are useful for triple integrals over regions that are symmetric with respect to the origin.

- **Rectangular to spherical:**

$$x = \rho \sin(\varphi) \cos(\theta)$$

$$y = \rho \sin(\varphi) \sin(\theta)$$

$$z = \rho \cos(\varphi)$$

$$\rho^2 = x^2 + y^2 + z^2$$

$$\tan(\theta) = \frac{y}{x}$$

$$\varphi = \cos^{-1} \left( \frac{z}{\sqrt{x^2 + y^2 + z^2}} \right).$$

- **Spherical to cylindrical:**

$$r = \rho \sin(\varphi)$$

$$\theta = \theta$$

$$z = \rho \cos(\varphi).$$

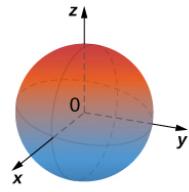
- **Cylindrical to spherical:**

$$\rho = \sqrt{r^2 + z^2}$$

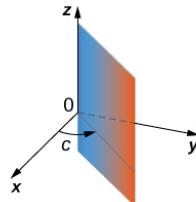
$$\theta = \theta$$

$$\varphi = \cos^{-1} \left( \frac{z}{\sqrt{r^2 + z^2}} \right).$$

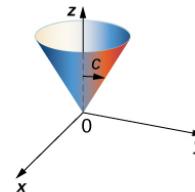
- **solid regions that are convenient to express in spherical coordinates.:**



Sphere  $\rho = c$  (constant)

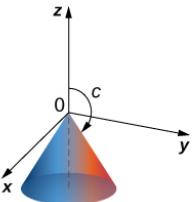


Half plane  $\theta = c$  (constant)



$$0 < c < \frac{\pi}{2}$$

Half cone  $\varphi = c$  (constant)



$$\frac{\pi}{2} < c < \pi$$

- **Fubini's Theorem for Spherical Coordinates:** If  $f(\rho, \theta, \phi)$  is continuous on a spherical solid box  $B = [a, b] \times [\alpha, \beta] \times [\gamma, \psi]$ , then

$$\iiint_B f(\rho, \theta, \phi) \rho^2 \sin \phi \, d\rho \, d\phi \, d\theta = \int_{\theta=\alpha}^{\theta=\beta} \int_{\phi=\gamma}^{\phi=\psi} \int_{\rho=a}^{\rho=b} f(\rho, \theta, \phi) \rho^2 \sin \phi \, d\rho \, d\phi \, d\theta.$$

This iterated integral may be replaced by other iterated integrals by integrating with respect to the three variables in other orders.

- **Spherical region is a general solid:**
- **Jacobian:** The Jacobian of the  $C^1$  transformation  $T(u, v) = (g(u, v), h(u, v))$  is denoted by  $J(u, v)$  and is defined by the  $2 \times 2$  determinant

$$J(u, v) = \left| \frac{\partial(x, y)}{\partial(u, v)} \right| = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \left( \frac{\partial x}{\partial u} \frac{\partial y}{\partial v} - \frac{\partial x}{\partial v} \frac{\partial y}{\partial u} \right).$$

- **one-to-one:** transformation if no two points map to the same image point.

## 4.7 Chapter 6: Vector calculus

### 4.7.1 Definitions and theorems

- **Vector field:** A vector field is a map of vectors.

A vector field  $\mathbf{F}$  in  $\mathbb{R}^2$  is an assignment of a two-dimensional vector  $\mathbf{F}(x, y)$  to each point  $(x, y)$  of a subset  $D$  of  $\mathbb{R}^2$ . The subset  $D$  is the domain of the vector field.

A vector field  $\mathbf{F}$  in  $\mathbb{R}^3$  is an assignment of a three-dimensional vector  $\mathbf{F}(x, y, z)$  to each point  $(x, y, z)$  of a subset  $D$  of  $\mathbb{R}^3$ . The subset  $D$  is the domain of the vector field.

- **Representing a vector field:** A vector field in  $\mathbb{R}^2$  can be represented in either of two equivalent ways. The first way is to use a vector with components that are two-variable functions:

$$\mathbf{F}(x, y) = \langle P(x, y), Q(x, y) \rangle.$$

The second way is to use the standard unit vectors:

$$\mathbf{F}(x, y) = P(x, y)\mathbf{i} + Q(x, y)\mathbf{j}.$$

A vector field is said to be continuous if its component functions are continuous.

- **Graphing a vector field:** Representing it visually by sketching it is more complex because the domain of a vector field is in  $\mathbb{R}^2$ , as is the range. Therefore, the “graph” of a vector field in  $\mathbb{R}^2$  lives in four-dimensional space. Since we cannot represent four-dimensional space visually, we instead draw vector fields in  $\mathbb{R}^2$  in a plane itself. To do this, draw the vector associated with a given point at the point in a plane. For example, suppose the vector associated with point  $(4, -1)$  is  $\langle 3, 1 \rangle$ . Then, we would draw vector  $\langle 3, 1 \rangle$  at point  $(4, -1)$ .

We should plot enough vectors to see the general shape, but not so many that the sketch becomes a jumbled mess. If we were to plot the image vector at each point in the region, it would fill the region completely and is useless. Instead, we can choose points at the intersections of grid lines and plot a sample of several vectors from each quadrant of a rectangular coordinate system in  $\mathbb{R}^2$ .

- **Radial fields:** Radial fields model certain gravitational fields and energy source fields,

In a **radial field**, all vectors either point directly toward or directly away from the origin. Furthermore, the magnitude of any vector depends only on its distance from the origin. In a radial field, the vector located at point  $(x, y)$  is perpendicular to the circle centered at the origin that contains point  $(x, y)$ , and all other vectors on this circle have the same magnitude.

- **Rotational fields:** Rotational fields model the movement of a fluid in a vortex.

In contrast to radial fields, in a rotational field, the vector at point  $(x, y)$  is tangent (not perpendicular) to a circle with radius  $r = \sqrt{x^2 + y^2}$ . In a standard rotational field, all vectors point either in a clockwise direction or in a counterclockwise direction, and the magnitude of a vector depends only on its distance from the origin.

- **Unit vector field:** A vector field  $\vec{\mathbf{F}}$  is a unit vector field if the magnitude of each vector in the field is one

- **Gradient field:** A vector field  $\mathbf{F}$  in  $\mathbb{R}^2$  or in  $\mathbb{R}^3$  is a gradient field if there exists a scalar function  $f$  such that  $\nabla f = \mathbf{F}$ .
- **Conservative vector field (Gradient field) existence:** A vector field  $\mathbf{F}$  is a conservative vector field, or a gradient field if there exists a scalar function  $f$  such that  $\nabla f = \mathbf{F}$ .
- **Potential functions:**  $f$  is called a potential function for  $\mathbf{F} \iff \nabla f = \mathbf{F}$
- **Uniqueness of Potential Functions:** Let  $\mathbf{F}$  be a conservative vector field on an open and connected domain, and let  $f$  and  $g$  be functions such that  $\nabla f = \mathbf{F}$  and  $\nabla g = \mathbf{F}$ . Then, there is a constant  $C$  such that  $f = g + C$ .
- **The Cross-Partial Property of Conservative Vector Fields:** Let  $\mathbf{F}$  be a vector field in two or three dimensions such that the component functions of  $\mathbf{F}$  have continuous first-order partial derivatives on the domain of  $\mathbf{F}$ .

If  $\mathbf{F}(x, y) = \langle P(x, y), Q(x, y) \rangle$  is a conservative vector field in  $\mathbb{R}^2$ , then  $\frac{\partial P}{\partial y} = \frac{\partial Q}{\partial x}$ .

If  $\mathbf{F}(x, y, z) = \langle P(x, y, z), Q(x, y, z), R(x, y, z) \rangle$  is a conservative vector field in  $\mathbb{R}^3$ , then

$$\frac{\partial P}{\partial y} = \frac{\partial Q}{\partial x}, \quad \frac{\partial Q}{\partial z} = \frac{\partial R}{\partial y}, \quad \text{and} \quad \frac{\partial R}{\partial x} = \frac{\partial P}{\partial z}.$$

- **Evaluating a Scalar Line Integral:** Let  $f$  be a continuous function with a domain that includes the smooth curve  $C$  with parameterization  $\mathbf{r}(t)$ ,  $a \leq t \leq b$ . Then

$$\int_C f \, ds = \int_a^b f(\mathbf{r}(t)) \|\mathbf{r}'(t)\| \, dt.$$

- **Scalar Line Integral Calculation:** Let  $f$  be a continuous function with a domain that includes the smooth curve  $C$  with parameterization  $\mathbf{r}(t) = \langle x(t), y(t), z(t) \rangle$ ,  $a \leq t \leq b$ . Then

$$\int_C f(x, y, z) \, ds = \int_a^b f(\mathbf{r}(t)) \sqrt{(x'(t))^2 + (y'(t))^2 + (z'(t))^2} \, dt.$$

Similarly,

$$\int_C f(x, y) \, ds = \int_a^b f(\mathbf{r}(t)) \sqrt{(x'(t))^2 + (y'(t))^2} \, dt.$$

if  $C$  is a planar curve and  $f$  is a function of two variables.

- **Independence of parameterization:** integral. Scalar line integrals are independent of parameterization, as long as the curve is traversed exactly once by the parameterization.
- **Arc length of a curve with line integral:**

$$S = \int_C ds.$$

- **Vector line integral:** The vector line integral of vector field  $\mathbf{F}$  along oriented smooth curve  $C$  is

$$\int_C \mathbf{F} \cdot \mathbf{T} \, ds = \lim_{n \rightarrow \infty} \sum_{i=1}^n \mathbf{F}(P_i^*) \cdot \mathbf{T}(P_i^*) \Delta s_i$$

if that limit exists.

Which we often write as

$$\begin{aligned} \int_C \mathbf{F} \cdot \mathbf{r}'(t) dt \\ = \int_C \mathbf{F} \cdot d\mathbf{r}. \end{aligned}$$

Where  $d\mathbf{r}$  denotes the differential

$$d\mathbf{r} = \langle x'(t), y'(t), z'(t) \rangle dt.$$

- **Another way to write vector line integral:** Let  $\mathbf{F}(x, y) = \langle P(x, y), Q(x, y) \rangle$  be a two-dimensional vector field. The integral  $\int_C \mathbf{F} \cdot \mathbf{T} ds$  is sometimes written as

$$\int_C P dx + Q dy.$$

- **Orientation reversal:**

$$\int_{-C} \mathbf{F} \cdot d\mathbf{r} = - \int_C \mathbf{F} \cdot d\mathbf{r}.$$

That is, reversing the orientation of a curve changes the sign of a line integral.

- **Flux:** The flux of  $\mathbf{F}$  across  $C$  is the line integral

$$\int_C \mathbf{F} \cdot \frac{\mathbf{n}(t)}{\|\mathbf{n}\|(t)} ds.$$

If  $\mathbf{F}$  is a velocity field of a fluid and  $C$  is a curve that represents a membrane, then the *flux* of  $\mathbf{F}$  across  $C$  is the quantity of fluid flowing across  $C$  per unit time, or the *rate of flow*.

- **Alternate way to write flux:**

$$\int_C -Q dx + P dy.$$

- **Circulation of a vector field:** The line integral of vector field  $\mathbf{F}$  along an oriented closed curve is called the circulation of  $\mathbf{F}$  along  $C$ . Circulation line integrals have their own notation:

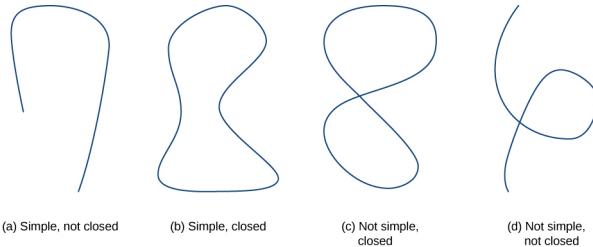
$$\oint_C \mathbf{F} \cdot \mathbf{T} ds.$$

The circle on the integral symbol denotes that  $C$  is “circular” in that it has no endpoints.

the value of the circulation  $\oint \mathbf{v} \cdot \mathbf{T} ds$  measures the tendency of the fluid to move in the direction of  $C$ .

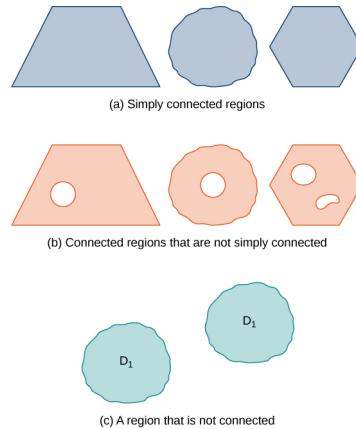
- **closed curve** if there is a parameterization  $\mathbf{r}(t)$ ,  $a \leq t \leq b$  of  $C$  such that the parameterization traverses the curve exactly once and  $\mathbf{r}(a) = \mathbf{r}(b)$ .
- **simple curve** if  $C$  does not cross itself. That is,  $C$  is simple if there exists a parameterization  $\mathbf{r}(t)$ ,  $a \leq t \leq b$  of  $C$  such that  $\mathbf{r} : (a, b)$  is one-to-one over  $(a, b)$ .

It is possible for  $\mathbf{r}(a) = \mathbf{r}(b)$ , meaning that the simple curve is also closed.



- **connected region:** if, for any two points  $P_1$  and  $P_2$ , there is a path from  $P_1$  to  $P_2$  with a trace contained entirely inside  $D$ .
- **simply connected region:** if  $D$  is connected and for any simple closed curve  $C$  that lies inside  $D$ , the curve  $C$  can be shrunk continuously to a point while staying entirely inside  $D$ .

All simply connected regions are connected, but not all connected regions are simply connected



- **The fundamental theorem for line integrals:** Let  $C$  be a piecewise smooth curve with parameterization  $\mathbf{r}(t), a \leq t \leq b$ .

Let  $f$  be a function of two or three variables with first-order partial derivatives that exist and are continuous on  $C$ . Then,

$$\int_C \nabla f \cdot d\mathbf{r} = f(\mathbf{r}(b)) - f(\mathbf{r}(a)).$$

- **Path independent:** Let  $\mathbf{F}$  be a vector field with domain  $D$ . The vector field  $\mathbf{F}$  is *independent of path* (or path independent) if

$$\int_{C_1} \mathbf{F} \cdot d\mathbf{r} = \int_{C_2} \mathbf{F} \cdot d\mathbf{r}$$

for any paths  $C_1$  and  $C_2$  in  $D$  with the same initial and terminal points.

- **Path Independence of Conservative Fields:** If  $\mathbf{F}$  is a conservative vector field, then  $\mathbf{F}$  is independent of path.
- **The Path Independence Test for Conservative Fields:** If  $\mathbf{F}$  is a continuous vector field that is independent of path and the domain  $D$  of  $\mathbf{F}$  is open and connected, then  $\mathbf{F}$  is conservative.
- **Problem-Solving Strategy: Finding a Potential Function for a Conservative Vector Field:**
  1. Integrate  $P$  with respect to  $x$ . This results in a function of the form  $g(x, y) + h(y)$ , where  $h(y)$  is unknown.
  2. Take the partial derivative of  $g(x, y) + h(y)$  with respect to  $y$ , which results in the function  $g_y(x, y) + h'(y)$ .
  3. Use the equation  $g_y(x, y) + h'(y) = Q(x, y)$  to find  $h'(y)$ .
  4. Integrate  $h'(y)$  to find  $h(y)$ .
  5. Any function of the form  $f(x, y) = g(x, y) + h(y) + C$ , where  $C$  is a constant, is a potential function for  $\mathbf{F}$ .

- **The Cross-Partial Test for Conservative Fields:** If  $\mathbf{F} = \langle P, Q, R \rangle$  is a vector field on an open, simply connected region  $D$  and the conditions

$$\frac{\partial P}{\partial y} = \frac{\partial Q}{\partial x}, \quad \frac{\partial P}{\partial z} = \frac{\partial R}{\partial x}, \quad \text{and} \quad \frac{\partial Q}{\partial z} = \frac{\partial R}{\partial y}$$

hold throughout  $D$ , then  $\mathbf{F}$  is conservative.

- **Cross-Partial Property of Conservative Fields:** Let  $\mathbf{F} = \langle P, Q, R \rangle$  be a vector field on an open, simply connected region  $D$ . Then  $\mathbf{F}$  is conservative if and only if the following conditions are satisfied throughout  $D$ :

$$\frac{\partial P}{\partial y} = \frac{\partial Q}{\partial x}, \quad \frac{\partial P}{\partial z} = \frac{\partial R}{\partial x}, \quad \text{and} \quad \frac{\partial Q}{\partial z} = \frac{\partial R}{\partial y}.$$

- **Green's Theorem, Circulation Form:** Let  $D$  be an open, simply connected region with a boundary curve  $C$  that is a piecewise smooth, simple closed curve oriented counterclockwise. Let  $\mathbf{F} = \langle P, Q \rangle$  be a vector field with component functions that have continuous partial derivatives on  $D$ . Then,

$$\oint_C \mathbf{F} \cdot d\mathbf{r} = \oint_C P dx + Q dy = \iint_D \left( \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA.$$

**Note:** Greens theorem can only be used for a two-dimensional vector field

- **Green's Theorem, Flux Form:** Let  $D$  be an open, simply connected region with a boundary curve  $C$  that is a piecewise smooth, simple closed curve that is oriented counterclockwise. Let  $\mathbf{F} = \langle P, Q \rangle$  be a vector field with component functions that have continuous partial derivatives on an open region containing  $D$ . Then,

$$\oint_C \mathbf{F} \cdot \mathbf{N} ds = \iint_D (P_x + Q_y) dA.$$

- **source-free field:** is a field with a flux that is zero along any closed curve
- **Source-free field  $\mathbf{F} = \langle P, Q \rangle$  on a simply connected domain:**

1. The flux  $\oint_C \mathbf{F} \cdot \mathbf{N} ds$  across any closed curve  $C$  is zero.
2. If  $C_1$  and  $C_2$  are curves in the domain of  $\mathbf{F}$  with the same starting points and endpoints, then  $\int_{C_1} \mathbf{F} \cdot \mathbf{N} ds = \int_{C_2} \mathbf{F} \cdot \mathbf{N} ds$ . In other words, flux is independent of path.
3. There is a stream function  $g(x, y)$  for  $\mathbf{F}$ . A stream function for  $\mathbf{F} = \langle P, Q \rangle$  is a function  $g$  such that  $P = g_y$  and  $Q = -g_x$ . Geometrically,  $\mathbf{F}(a, b)$  is tangential to the level curve of  $g$  at  $(a, b)$ . Since the gradient of  $g$  is perpendicular to the level curve of  $g$  at  $(a, b)$ , stream function  $g$  has the property  $\mathbf{F}(a, b) \cdot \nabla g(a, b) = 0$  for any point  $(a, b)$  in the domain of  $g$ . (Stream functions play the same role for source-free fields that potential functions play for conservative fields.)
4.  $P_x + Q_y = 0$

# Probability and Statistics

## 5.1 Chapter 1: Overview and Descriptive Statistics

### 5.1.1 Definitions and Theorems

- **Data:** Collection of facts
- **Population:** Encompasses the complete set of data
- **census:** When desired information is available for all objects in the population
- **sample:** Subset of the population
- **variable:** any characteristic whose value may change from one object to another in the population
- **univariate:** A univariate data set consists of observations on a single variable.
- **bivariate:** We have bivariate data when observations are made on each of two variables
- **Multivariate:** Multivariate data arises when observations are made on more than one variable
- **descriptive statistics:** Descriptive statistics refers to a branch of statistics that involves summarizing, organizing, and presenting data meaningfully and concisely
- **inferential statistics:** Techniques for generalizing from a sample to a population are gathered within the branch of our discipline called inferential statistics.
- **conceptual or hypothetical population:** populations that are conceptual or theoretical and may not exist in reality
- **discrete numerical variable:** A numerical variable is discrete if its set of possible values either is finite or else can be listed in an infinite sequence
- **continuous numerical variable:** A numerical variable is discrete if its set of possible values either is finite or else can be listed in an infinite sequence
- **frequency of a value:** The frequency of any particular  $x$  value is the number of times that value occurs in the data set
- **relative frequency of a value:** The relative frequency of a value is the fraction or proportion of times the value occurs. In theory, the relative frequencies should sum to 1, but in practice the sum may differ slightly from 1 because of rounding

$$\text{relative frequency of a value} = \frac{\text{number of times the value occurs}}{\text{number of observations in the data set}}.$$

- **frequency distribution:** A frequency distribution is a tabulation of the frequencies and/or relative frequencies.
- **class intervals or classes:** ranges of values that data is divided into

- Rule to determine number of classes:

$$\text{Number of classes} \approx \sqrt{\text{Number of observations}}.$$

- Continuous histogram unequal class width rectangle height:

$$\text{Rectangle height} = \frac{\text{Relative frequency of the class}}{\text{Class width}}.$$

Why Use This Formula?

- **Equal Area Representation:** When class intervals have different widths, simply using the frequency as the height would result in misleading representations. For instance, a wider class interval might appear more significant just because it is wider, not necessarily because it has a higher frequency.
- **Density Representation:** The formula normalizes the frequency by the class width, which gives a measure of frequency density. This normalization ensures that the area of the rectangle (height  $\times$  width) represents the true relative frequency of the data in that interval.

The resulting rectangle heights are usually called **densities**, and the vertical scale is the **density scale**. This prescription will also work when class widths are equal.

- **Density histogram property:** A density histogram does have one interesting property. Multiplying both sides of the formula for density by the class width gives

$$\text{Relative frequency} = (\text{class width})(\text{density}) = (\text{rectangle width})(\text{rectangle height}) = \text{rectangle area}.$$

That is, the area of each rectangle is the relative frequency of the corresponding class. Furthermore, since the sum of relative frequencies should be 1, the total area of all rectangles in a density histogram is 1. It is always possible to draw a histogram so that the area equals the relative frequency (this is true also for a histogram of discrete data)—just use the density scale

- **Histogram shapes:**

- **unimodal:** histogram is one that rises to a single peak and then declines
- **bimodal:** histogram has two different peaks.
- **multimodal:**
- **symmetric:** if the left half is a mirror image of the right half
- **negatively skewed (skewed left):** if the stretching is to the left



- **Sample Mean  $\bar{x}$ :** The sample mean  $\bar{x}$  : of observations is given by

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

The numerator of  $\bar{x}$  can be written more informally as  $\sum_{i=1}^n x_i$ , where the summation is over all sample observations.

- **Population mean:** Denote by  $\mu$
- **Sample Median  $\tilde{x}$  (Sorted set):**

$$\tilde{x} = \begin{cases} \left(\frac{n+1}{2}\right)^{\text{th}} \text{ ordered element} & \text{if } n = 2k + 1, k \in \mathbb{Z} \\ (\text{ave}(\frac{n}{2}, \frac{n+2}{2}))^{\text{th}} \text{ ordered element} & \text{if } n = 2k, k \in \mathbb{Z} \end{cases}.$$

- **Population median:** Denoted by  $\tilde{\mu}$  :
- **trimmed mean** is a compromise between  $\bar{x}$  and  $\tilde{x}$  :. A 10% trimmed mean, for example, would be computed by eliminating the smallest 10% and the largest 10% of the sample and then averaging what remains.
- **trimmed mean** is denoted by  $\bar{x}_{tr(\gamma)}$ , for a 10% trimmed mean we have  $\bar{x}_{tr(10)}$  :. And we can say the trimmed percentage is given by  $100\alpha\%$
- **sample proportion:** The sample proportion, denoted by  $\hat{p}$  :, is the ratio of members of a sample that have a particular characteristic to the total members in the sample. It is an estimate of the population proportion. It is calculated as:

$$\hat{p} = \frac{x}{n}.$$

where  $x$  is the number of members in the sample with the characteristic of interest, and  $n$  is the total number of members in the sample.

- **Population proportion:** The population proportion, denoted by  $p$ , is the ratio of members of a population that have a particular characteristic to the total members in the population. It is calculated as:

$$p = \frac{X}{N}.$$

where  $x$  is the number of members in the sample with the characteristic of interest, and  $n$  is the total number of members in the sample.

- **point estimate:** is a single value that serves as an estimate of a population parameter
- **dispersion:**, refers to the extent to which data points in a statistical distribution or data set differ from the mean or from each other. It provides a measure of how spread out the values are in a data set.
- **Range:** The simplest measure of variability in a sample is the range which is the difference between the largest and smallest sample values
- **deviations from the mean:** the deviations from the mean are obtained by subtracting  $\bar{x}$  from each of the  $n$  sample observations. A deviation will be positive if the observation is larger than the mean (to the right of the mean on the measurement axis) and negative if the observation is smaller than the mean
- **The sum of the deviations is always zero:**

$$\sum(x_i - \bar{x}) = 0.$$

180

- **sample variance**, denoted by  $s^2$ , is given by

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{S_{xx}}{n-1}$$

The **sample standard deviation**, denoted by  $s$ , is the (positive) square root of the variance:

$$s = \sqrt{s^2}$$

- **Sample variance alternate formula**:

$$\begin{aligned} S_{xx} &= \sum x_i^2 - \frac{1}{n} \left( \sum x_i \right)^2 \\ \implies s^2 &= \frac{n \sum x_i^2 - (\sum x_i)^2}{n^2 - n}. \end{aligned}$$

- **Population variance**, denote by  $\sigma^2$  :, is given by

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}.$$

The **population standard deviation**, denoted by  $\sigma$ , is given by

$$\sigma = \sqrt{\sigma^2}.$$

- **Boxplot definitions**:

Order the  $n$  observations from smallest to largest and separate the smallest half from the largest half; the median is included in both halves if  $n$  is odd. Then the **lower fourth** is the median of the smallest half and the **upper fourth** is the median of the largest half. A measure of spread that is resistant to outliers is the **fourth spread**  $f_s$ , given by

$$f_s = \text{upper fourth} - \text{lower fourth}.$$

Any observation farther than  $1.5f_s$  from the closest fourth is an outlier. An outlier is extreme if it is more than  $3f_s$  from the nearest fourth, and it is mild otherwise

- **Effects of Linear Transformations on Mean, Median, sd, and Variance**: Assume that  $Y$  is a linear transformation of  $X$ . Then,

$$Y = bX + A$$

What is the relationship between the mean, median, standard deviation, and variance of  $X$  and the mean, median, standard deviation, and variance of  $Y$ ?

- **Mean of  $Y$** :

$$\text{Mean of } Y = b(\text{Mean of } X) + A$$

- **Median of  $Y$** :

$$\text{Median of } Y = b(\text{Median of } X) + A$$

- **Standard Deviation of  $Y$** :

$$\text{sd of } Y = b(\text{sd of } X)$$

- **Variance of  $Y$** :

$$\text{Variance of } Y = b^2(\text{Variance of } X)$$

### 5.1.2 Frequency Distribution

A frequency distribution is a tabulation of data containing frequencies and or relative frequencies. For example,

**Table 1.1 Frequency Distribution for Hits in Nine-Inning Games**

Hits/Game	Number of Games	Relative Frequency	Hits/Game	Number of Games	Relative Frequency
0	20	.0010	14	569	.0294
1	72	.0037	15	393	.0203
2	209	.0108	16	253	.0131
3	527	.0272	17	171	.0088
4	1048	.0541	18	97	.0050
5	1457	.0752	19	53	.0027
6	1988	.1026	20	31	.0016
7	2256	.1164	21	19	.0010
8	2403	.1240	22	13	.0007
9	2256	.1164	23	5	.0003
10	1967	.1015	24	1	.0001
11	1509	.0779	25	0	.0000
12	1230	.0635	26	1	.0001
13	834	.0430	27	1	.0001
				19,383	1.0005

### 5.1.3 Stem and Leaf Displays

Consider a numerical data set  $x_1, x_2, x_3, \dots, x_n$  for which each  $x_i$  consists of at least two digits. A quick way to obtain an informative visual representation of the data set is to construct a stem-and-leaf display.

#### Constructing a Stem-and-Leaf Display

1. Select one or more leading digits for the stem values. The trailing digits become the leaves.
2. List possible stem values in a vertical column.
3. Record the leaf for each observation beside the corresponding stem value.
4. Indicate the units for stems and leaves someplace in the display.

Suppose we have a set containing the weights of 20 people

101, 114, 117, 121, 124, 128, 129, 130,  
 137, 139, 139, 142, 144, 148, 149, 149,  
 153, 157, 163, 175

Since all the weights start with 1, we will use the first two digits as the stems. Our stem and leaf display may look something like

Stem	leaf
10	1
11	4,7
12	1,4,8,9
13	0,7,9,9
14	2,4,8,9,9
15	3,7
16	3
17	5

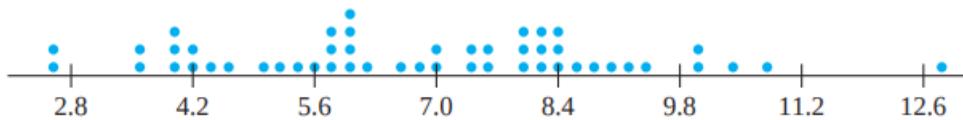
### 5.1.4 Dotplots

A dotplot is an attractive summary of numerical data when the data set is reasonably small or there are relatively few distinct data values. Each observation is represented by a dot above the corresponding location on a horizontal measurement scale. When a value occurs more than once, there is a dot for each occurrence, and these dots are stacked vertically. As with a stem-and-leaf display, a dotplot gives information about location, spread, extremes, and gaps.

Suppose we have the data

```
10.8 6.9 8.0 8.8 7.3 3.6 4.1 6.0 4.4 8.3  
8.1 8.0 5.9 5.9 7.6 8.9 8.5 8.1 4.2 5.7  
4.0 6.7 5.8 9.9 5.6 5.8 9.3 6.2 2.5 4.5  
12.8 3.5 10.0 9.1 5.0 8.1 5.3 3.9 4.0 8.0  
7.4 7.5 8.4 8.3 2.6 5.1 6.0 7.0 6.5 10.3
```

Then our dotplot would look something like



### 5.1.5 Histogram for discrete data

Some numerical data is obtained by counting to determine the value of a variable (the number of traffic citations a person received during the last year, the number of customers arriving for service during a particular period), whereas other data is obtained by taking measurements (weight of an individual, reaction time to a particular stimulus). The prescription for drawing a histogram is generally different for these two cases

#### Constructing a histogram for discrete data

First, determine the frequency and relative frequency of each  $x$  value. Then mark possible  $x$  values on a horizontal scale. Above each value, draw a rectangle whose height is the relative frequency (or alternatively, the frequency) of that value.

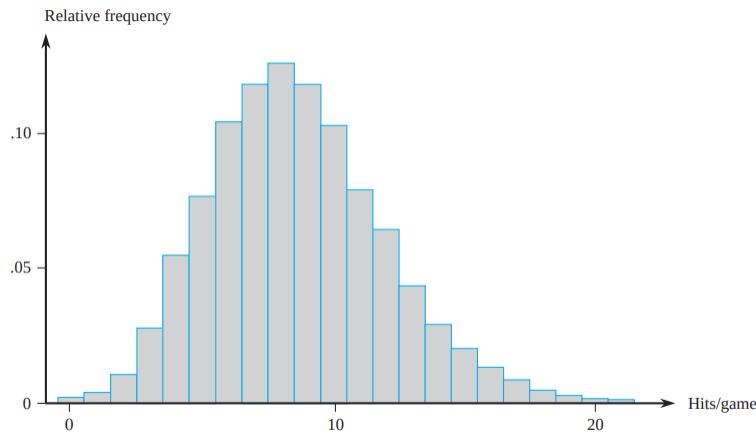
This construction ensures that the area of each rectangle is proportional to the relative frequency of the value. Thus if the relative frequencies of  $x = 1$  and  $x = 5$  are .35 and .07, respectively, then the area of the rectangle above 1 is five times the area of the rectangle above 5.

Let's refer back to the following frequency distribution

**Table 1.1 Frequency Distribution for Hits in Nine-Inning Games**

Hits/Game	Number of Games	Relative Frequency	Hits/Game	Number of Games	Relative Frequency
0	20	.0010	14	569	.0294
1	72	.0037	15	393	.0203
2	209	.0108	16	253	.0131
3	527	.0272	17	171	.0088
4	1048	.0541	18	97	.0050
5	1457	.0752	19	53	.0027
6	1988	.1026	20	31	.0016
7	2256	.1164	21	19	.0010
8	2403	.1240	22	13	.0007
9	2256	.1164	23	5	.0003
10	1967	.1015	24	1	.0001
11	1509	.0779	25	0	.0000
12	1230	.0635	26	1	.0001
13	834	.0430	27	1	.0001
				19,383	1.0005

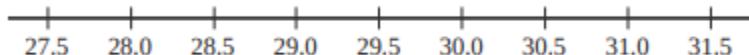
We can then construct a histogram for the relative frequencies



**Figure 1.7 Histogram of number of hits per nine-inning game**

### 5.1.6 Histogram for continuous data: Equal class widths

Constructing a histogram for continuous data (measurements) entails subdividing the measurement axis into a suitable number of class intervals or classes, such that each observation is contained in exactly one class. Suppose, for example, that we have 50 observations on  $x =$  fuel efficiency of an automobile (mpg), the smallest of which is 27.8 and the largest of which is 31.4. Then we could use the class boundaries 27.5, 28.0, 28.5, . . . , and 31.5 as shown here:



One potential difficulty is that occasionally an observation lies on a class boundary so therefore does not fall in exactly one interval

One way to deal with this problem is to use boundaries like 27.55, 28.05, . . . , 31.55. Adding a hundredths digit to the class boundaries prevents observations from falling on the resulting boundaries. Another approach is to use the classes 27.5 -<28.0, 28.0-<28.5,...,31.0-<31.5. Then 29.0 falls in the class 29.0-<29.5 rather than in the class 28.5-<29.0. In other words, with this convention, an observation on a boundary is placed in the interval to the right of the boundary.

#### Constructing a Histogram for Continuous Data: Equal Class Widths

Determine the frequency and relative frequency for each class. Mark the class boundaries on a horizontal measurement axis. Above each class interval, draw a rectangle whose height is the corresponding relative frequency (or frequency)

Consider the following data

Class	1-<3	3-<5	5-<7	7-<9	9-<11	11-<13	13-<15	15-<17	17-<19
Frequency	1	1	11	21	25	17	9	4	1
Relative frequency	.011	.011	.122	.233	.278	.189	.100	.044	.011

Then the histogram would look something like



Figure 1.8 Histogram of the energy consumption data from Example 1.10

### 5.1.7 Histogram for Continuous Data: Unequal Class Widths

Equal-width classes may not be a sensible choice if there are some regions of the measurement scale that have a high concentration of data values and other parts where data is quite sparse

Using a small number of equal-width classes results in almost all observations falling in just one or two of the classes. If a large number of equal-width classes are used, many classes will have zero frequency. A sound choice is to use a few wider intervals near extreme observations and narrower intervals in the region of high concentration.

After determining frequencies and relative frequencies, calculate the height of each rectangle using the formula

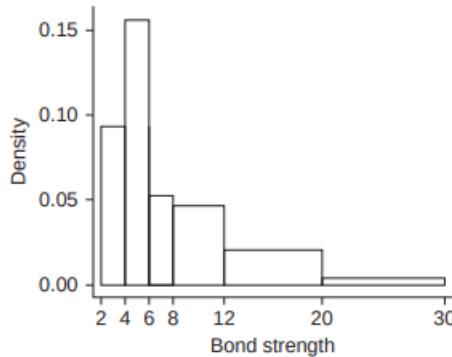
$$\text{Rectangle height} = \frac{\text{Relative frequency of the class}}{\text{Class width}}.$$

The resulting rectangle heights are usually called **densities**, and the vertical scale is the **density scale**. This prescription will also work when class widths are equal.

Consider the following data and frequency distribution

11.5	12.1	9.9	9.3	7.8	6.2	6.6	7.0	13.4	17.1	9.3	5.6
5.7	5.4	5.2	5.1	4.9	10.7	15.2	8.5	4.2	4.0	3.9	3.8
3.6	3.4	20.6	25.5	13.8	12.6	13.1	8.9	8.2	10.7	14.2	7.6
5.2	5.5	5.1	5.0	5.2	4.8	4.1	3.8	3.7	3.6	3.6	3.6
<i>Class</i>	2-<4	4-<6	6-<8	8-<12	12-<20	20-<30					
<i>Frequency</i>	9	15	5	9	8	2					
<i>Relative frequency</i>	.1875	.3125	.1042	.1875	.1667	.0417					
<i>Density</i>	.094	.156	.052	.047	.021	.004					

Then the histogram would look something like



When class widths are unequal, not using a density scale will give a picture with distorted areas. For equal-class widths, the divisor is the same in each density calculation, and the extra arithmetic simply results in a rescaling of the vertical axis

### 5.1.8 Boxplots

Stem-and-leaf displays and histograms convey rather general impressions about a data set, whereas a single summary such as the mean or standard deviation focuses on just one aspect of the data. In recent years, a pictorial summary called a boxplot has been used successfully to describe several of a data set's most prominent features. These features include (1) center, (2) spread, (3) the extent and nature of any departure from symmetry, and (4) identification of "outliers," observations that lie unusually far from the main body of the data. Because even a single outlier can drastically affect the values of  $\bar{x}$  and  $s$ , a boxplot is based on measures that are "resistant" to the presence of a few outliers—the median and a measure of variability called the fourth spread.

Order the  $n$  observations from smallest to largest and separate the smallest half from the largest half; the median is included in both halves if  $n$  is odd. Then the **lower fourth** is the median of the smallest half and the **upper fourth** is the median of the largest half. A measure of spread that is resistant to outliers is the **fourth spread**  $f_s$ , given by

$$f_s = \text{upper fourth} - \text{lower fourth}.$$

The simplest boxplot is based on the following five-number summary:

smallest  $x_i$     lower fourth    median    upper fourth    largest  $x_i$

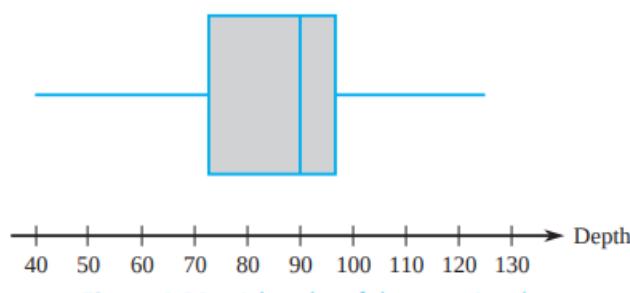
First, draw a horizontal measurement scale. Then place a rectangle above this axis; the left edge of the rectangle is at the lower fourth, and the right edge is at the upper fourth (so box width =  $f_s$ ). Place a vertical line segment or some other symbol inside the rectangle at the location of the median; the position of the median symbol relative to the two edges conveys information about skewness in the middle 50% of the data. Finally, draw "whiskers" out from either end of the rectangle to the smallest and largest observations. A boxplot with a vertical orientation can also be drawn by making obvious modifications in the construction process.

$40 \ 52 \ 55 \ 60 \ 70 \ 75 \ 85 \ 85 \ 90 \ 90 \ 92 \ 94 \ 94 \ 95 \ 98 \ 100 \ 115 \ 125 \ 125$

The five-number summary is as follows:

smallest  $x_i = 40$     lower fourth = 72.5     $\bar{x} = 90$     upper fourth = 96.5  
largest  $x_i = 125$

Figure 1.20 shows the resulting boxplot. The right edge of the box is much closer to the median than is the left edge, indicating a very substantial skew in the middle half of the data. The box width ( $f_s$ ) is also reasonably large relative to the range of the data (distance between the tips of the whiskers).



**Figure 1.20** A boxplot of the corrosion data

## 5.2 Chapter 2: Probability

### 5.2.1 Definitions and Theorems

- **An experiment:** is any activity or process whose outcome is subject to uncertainty
- **sample space** of an experiment, denoted by  $\mathcal{S} :$ , is the set of all possible outcomes of that experiment.
- **event** is any collection (subset) of outcomes contained in the sample space  $\mathcal{S} :$
- **compound:** if it consists of more than one outcome.
- **Some relations from set theory:**
  1. **complement:** of an event  $A$ , denoted by  $A^c$ , is the set of all outcomes in the sample space  $S$  that are not contained in  $A$ .
  2. **union:** of two events  $A$  and  $B$ , denoted by  $A \cup B$  and read “ $A$  or  $B$ ,” is the event consisting of all outcomes that are either in  $A$  or in  $B$  or in both events (so that the union includes outcomes for which both  $A$  and  $B$  occur as well as outcomes for which exactly one occurs)—that is, all outcomes in at least one of the events.
  3. **intersection:** of two events  $A$  and  $B$ , denoted by  $A \cap B$  and read “ $A$  and  $B$ ,” is the event consisting of all outcomes that are in both  $A$  and  $B$ .
  4. **disjoint:** events.

- **Other way to denote disjoint sets:** Given two sets  $A$  and  $B$ ,

$$\forall x(x \in A \implies x \notin B).$$

- **Probability:** Given an experiment and a sample space  $\mathcal{S} :$ , the objective of probability is to assign to each event  $A$  a number  $P(A)$ , called the probability of the event  $A$ , which will give a precise measure of the chance that  $A$  will occur.
- **Axioms of probability:** To ensure that the probability assignments will be consistent with our intuitive notions of probability, all assignments should satisfy the following axioms (basic properties) of probability. For any event  $A$ ,  $P(A) \geq 0$ .

$$P(\mathcal{S}) = 1.$$

If  $A_1, A_2, A_3, \dots$  is an infinite collection of disjoint events, then

$$P(A_1 \cup A_2 \cup A_3 \cup \dots) = \sum_{i=1}^{\infty} P(A_i)$$

$P(\emptyset) = 0$  where  $\emptyset$  is the **null event**

- **Limiting (long-run) relative frequency for some event  $A$ :** as  $n$  gets arbitrarily large, approaches a limiting value referred to as the limiting (or long-run) relative frequency of the event  $A$ . The objective interpretation of probability identifies this limiting relative frequency with  $P(A)$
- **Interpretation of probability (example):** if  $B$  is the event that an appliance of a particular type will need service while under warranty, then  $P(B) = 0.1$  is interpreted to mean that in the long run 10% of such appliances will need warranty service. This doesn't mean that exactly 1 out of 10 will need service, or that exactly 10 out of 100 will need service, because 10 and 100 are not the long run.
- **Probability Properties:**
  - For any event  $A$ ,  $P(A) + P(A^c) = 1$ , from which  $P(A) = 1 - P(A^c)$ .

**Note:** This proposition is surprisingly useful because there are many situations in which  $P(A^C)$  is more easily obtained by direct methods than is  $P(A)$ .

- For any event  $A$ ,  $P(A) \leq 1$

- **Disjoint vs independent events:**

- **disjoint:** or mutually exclusive if they cannot both occur at the same time. This means that the intersection of the two events is the empty set:

$$A \cap B = \emptyset.$$

- **independent:** if the occurrence of one event does not affect the probability of the occurrence of the other event

- **Addition and Multiplication rules for independent and disjoint events:**

For two disjoint events  $A$  and  $B$

- Probability of  $A$  or  $B$ :

$$P(A \cup B) = P(A) + P(B)$$

Since disjoint events cannot occur at the same time, there is no overlap between them.

- Probability of  $A$  and  $B$ :

$$P(A \cap B) = 0$$

Because disjoint events cannot both occur at the same time, their intersection is the empty set.

For two independent events  $A$  and  $B$

- Probability of  $A$  or  $B$ :

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

This formula accounts for the fact that there may be an overlap between  $A$  and  $B$ , which is subtracted to avoid double-counting.

**Note:** This formula is also used for dependent events. In other words, this is the formula to use if the events are not disjoint.

- Probability of  $A$  and  $B$ :

$$P(A \cap B) = P(A) \cdot P(B)$$

Since the events are independent, the probability of both occurring is the product of their individual probabilities.

For any three events  $A$ ,  $B$ , and  $C$

- Probability of  $A$  or  $B$  or  $C$

$$\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) \\ &\quad - P(B \cap C) + P(A \cap B \cap C). \end{aligned}$$

This can be verified by examining a Venn diagram of  $A \cup B \cup C$ . When  $P(A)$ ,  $P(B)$ , and  $P(C)$  are added, certain intersections are counted twice, so they must be subtracted out, but this results in  $P(A \cap B \cap C)$  being subtracted once too often.

- **Splitting event into union of disjoint events:**

$$\begin{aligned} A &= (A \cap B) \cup (A \cap B^C) \\ B &= (A \cap B) \cup (B \cap A^C). \end{aligned}$$

With probability,

$$\begin{aligned} P(A) &= P((A \cap B) \cup (A \cap B^C)) \\ &= P(A \cap B) + P(A \cap B^C). \end{aligned}$$

Since the events are disjoint, we can simply add them.

- **Determining Probabilities Systematically:** Consider a sample space that is either finite or "countably infinite" (the latter means that outcomes can be listed in an infinite sequence, so there is a first outcome, a second outcome, a third outcome, and so on—for example, the battery testing scenario of Example 2.12). Let  $E_1, E_2, E_3, \dots$  denote the corresponding simple events, each consisting of a single outcome. A sensible strategy for probability computation is to first determine each simple event probability, with the requirement that  $\sum P(E_i) = 1$ . Then the probability of any compound event  $A$  is computed by adding together the  $P(E_i)$ 's for all  $E_i$ 's in  $A$ :

$$P(A) = \sum_{\text{all } E'_i \text{ s in } A} P(E_i)$$

- **Equally Likely Outcomes:** In many experiments consisting of  $N$  outcomes, it is reasonable to assign equal probabilities to all  $N$  simple events. These include such obvious examples as tossing a fair coin or fair die once or twice (or any fixed number of times), or selecting one or several cards from a well-shuffled deck of 52. With  $p = P(E_i)$  for every  $i$ ,

$$1 = \sum_{i=1}^N P(E_i) = \sum_{i=1}^N p = p \cdot N \quad \text{so } p = \frac{1}{N}$$

That is, if there are  $N$  equally likely outcomes, the probability for each is  $1/N$ .

Now consider an event  $A$ , with  $N(A)$  denoting the number of outcomes contained in  $A$ . Then

$$P(A) = \sum_{E_i \text{ in } A} P(E_i) = \sum_{E_i \text{ in } A} \frac{1}{N} = \frac{N(A)}{N}$$

Thus when outcomes are equally likely, computing probabilities reduces to counting: determine both the number of outcomes  $N(A)$  in  $A$  and the number of outcomes  $N$  in  $S$ , and form their ratio.

- **The Product Rule for Ordered Pairs:** If the first element or object of an ordered pair can be selected in  $n_1$  ways, and for each of these  $n_1$  ways the second element of the pair can be selected in  $n_2$  ways, then the number of pairs is  $n_1 n_2$ .
- **The product rule for k-tuples:** Suppose a set consists of ordered collections of  $k$  elements ( $k$ -tuples) and that there are  $n_1$  possible choices for the first element; for each choice of the first element, there are  $n_2$  possible choices of the second element; ...; for each possible choice of the first  $k - 1$  elements, there are  $n_k$  choices of the  $k$ -th element. Then there are  $n_1 \times n_2 \times \cdots \times n_k$  possible  $k$ -tuples.

- **Permutations and Combinations:** An ordered subset is called a **permutation**. The number of permutations of size  $k$  that can be formed from the  $n$  individuals or objects in a group will be denoted by  $P_{k,n}$ . An unordered subset is called a **combination**. One way to denote the number of combinations is  $C_{k,n}$ , but we shall instead use notation that is quite common in probability books:  $\binom{n}{k}$ , read “ $n$  choose  $k$ ”.

- **Permutation Formulas:**

- **Without repetition:**

$$P(n, k) = \frac{n!}{(n - k)!}.$$

- **With repetition:**

$$n^k.$$

- **Combination formulas:**

- **Without repetition:**

$$C(n, k) = \binom{n}{k} = \frac{P_{k,n}}{k!} = \frac{n!}{k!(n - k)!}.$$

- **With repetition:**

$$C(n + k - 1, k) = \binom{n + k - 1}{k} = \frac{(n + k - 1)!}{k!(n - 1)!}.$$

- **Conditional Probability:** For any two events  $A$  and  $B$ , the conditional probability of  $A$  given that  $B$  has occurred is defined by

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

- **Complements with condition probability:** For two events  $A$  and  $B$

$$\begin{aligned} P(B | A) + P(B^C | A) &= 1 \\ P(B | A^C) + P(B^C | A^C) &= 1. \end{aligned}$$

The first one states that given  $A$  has occurred, either  $B$  occurs or it does not. Ie the sum of the probabilities is one. For the second one we are stating that either  $B$  occurs given  $A^C$  does not occur, or neither occur.

- **When is conditional probability most enlightening:**

- For dependent events, conditional probability provides insights into how the occurrence of one event affects the likelihood of another event.
- If events were independent, then the probability of one event occurring does not affect the probability of the other event occurring. Consider  $P(A) = 0.6$  and  $P(B) = 0.4$ . If we discern these events to be independent, then  $P(A \cap B) = P(A)P(B) = 0.24$ . We can then see

$$P(A | B) = \frac{P(A \cap B)}{P(B)} = \frac{0.24}{0.4} = 0.6$$

In other words, the probability of  $A$  occurring given  $B$  has already occurred is just the probability that  $A$  will occur. Hence, they are independent.

- Now suppose we are given that  $P(A \cap B) = 0.3$ . Since the multiplication rule for independent events does not hold in this case, we conclude that these events must be dependent. In this case,  $P(A | B)$  and  $P(B | A)$  would be different from the individual probabilities  $P(A)$  and  $P(B)$ . For example,

$$P(A | B) = \frac{P(A \cap B)}{P(B)} = \frac{0.3}{0.4} = 0.75$$

and

$$P(B | A) = \frac{P(A \cap B)}{P(A)} = \frac{0.3}{0.6} = 0.5$$

This demonstrates that the probability of one event occurring affects the probability that the other will occur, indicating dependence.

In summary:

- For independent events,  $P(A | B) = P(A)$  and  $P(B | A) = P(B)$ . The events do not influence each other, and the intersection is simply  $P(A \cap B) = P(A)P(B)$ .
- For dependent events,  $P(A | B) \neq P(A)$  and  $P(B | A) \neq P(B)$ . The occurrence of one event influences the probability of the other, and we use conditional probability to capture this relationship.

- **Multiplication rule for dependent events:**

$$P(A \cap B) = P(A) \times P(B | A).$$

or equivalently,

$$P(A \cap B) = P(B) \times P(A | B).$$

This rule is important because it is often the case that  $P(A \cap B)$  is desired, whereas both  $P(B)$  and  $P(A|B)$  can be specified from the problem description.

This rule can also be extended, for example

$$\begin{aligned} P(A_1 \cap A_2 \cap A_3) &= P(A_1 \cap A_2) \cdot P(A_3 | A_1 \cap A_2) \\ &= P(A_1) \cdot P(A_2 | A_1) \cdot P(A_3 | A_1 \cap A_2). \end{aligned}$$

where  $A_1$  occurs first, followed by  $A_2$ , and finally  $A_3$ .

- **Complex conditional probabilities:**

$$P(A | B \cup C) = \frac{P(A \cap (B \cup C))}{P(B \cup C)}.$$

Etc..

- **Prior probability:**, in Bayesian statistics, is the probability of an event before new data is collected
- **posterior probability:**, in Bayesian statistics, is the revised or updated probability of an event occurring after taking into consideration new information
- **The Law of Total Probability:** Let  $A_1, \dots, A_k$  be mutually exclusive and exhaustive events. Then for any other event  $B$ ,

$$P(B) = P(B|A_1)P(A_1) + \dots + P(B|A_k)P(A_k)$$

$$= \sum_{i=1}^k P(B|A_i)P(A_i)$$

- **Exhaustive Events:** A set of events  $A_1, A_2, \dots, A_k$  is exhaustive if at least one of the events must occur, i.e.,

$$P(A_1 \cup A_2 \cup \dots \cup A_k) = 1$$

- **Bayes' Theorem:** Let  $A_1, A_2, \dots, A_k$  be a collection of  $k$  mutually exclusive and exhaustive events with prior probabilities  $P(A_i)$  ( $i = 1, \dots, k$ ). Then for any other event  $B$  for which  $P(B) > 0$ , the posterior probability of  $A_j$  given that  $B$  has occurred is

$$P(A_j|B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^k P(B|A_i) \cdot P(A_i)} \quad j = 1, \dots, k$$

- **independent:** if  $P(A | B) = P(A)$  and are dependent otherwise.

**Note:** The definition of independence might seem “unsymmetric” because we do not also demand that  $P(B|A) = P(B)$ . However, using the definition of conditional probability and the multiplication rule,

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

The right-hand side of Equation is  $P(B)$  if and only if  $P(A|B) = P(A)$  (independence), so the equality in the definition implies the other equality (and vice versa). It is also straightforward to show that if  $A$  and  $B$  are independent, then so are the following pairs of events: (1)  $A'$  and  $B$ , (2)  $A$  and  $B'$ , and (3)  $A'$  and  $B'$ .

- **independent:** iff

$$P(A \cap B) = P(A) \cdot P(B).$$

- Events  $A_1, \dots, A_n$  are **mutually independent** if for every  $k$  ( $k = 2, 3, \dots, n$ ) and every subset of indices  $i_1, i_2, \dots, i_k$ ,

$$P(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_k}) = P(A_{i_1}) \cdot P(A_{i_2}) \cdots P(A_{i_k})$$

### 5.2.2 Examples in Probability Theory

- In a certain residential suburb, 60% of all households get Internet service from the local cable company, 80% get television service from that company, and 50% get both services from that company. If a household is randomly selected, what is the probability that it gets at least one of these two services from the company, and what is the probability that it gets exactly one of these services from the company? With  $A = \{\text{gets Internet service}\}$  and  $B = \{\text{gets TV service}\}$ , the given information implies that  $P(A) = .6$ ,  $P(B) = .8$ , and  $P(A \cap B) = .5$ . The foregoing proposition now yields  $P$  (subscribes to at least one of the two services)

$$= P(A \cup B) = P(A) + P(B) - P(A \cap B) = .6 + .8 - .5 = .9$$

The event that a household subscribes only to tv service can be written as  $A' \cap B$  [(not Internet) and TV]. Now Figure 2.4 implies that

$$.9 = P(A \cup B) = P(A) + P(A' \cap B) = .6 + P(A' \cap B)$$

from which  $P(A' \cap B) = .3$ . Similarly,  $P(A \cap B') = P(A \cup B) - P(B) = .1$ . This is all illustrated in Figure 2.5, from which we see that

$$P(\text{exactly one}) = P(A \cap B') + P(A' \cap B) = .1 + .3 = .4$$

- During off-peak hours a commuter train has five cars. Suppose a commuter is twice as likely to select the middle car (#3) as to select either adjacent car (#2 or #4), and is twice as likely to select either adjacent car as to select either end car (#1 or #5). Let  $p_i = P(\text{car } i \text{ is selected}) = P(E_i)$ . Then we have  $p_3 = 2p_2 = 2p_4$  and  $p_2 = 2p_1 = 2p_5 = p_4$ . This gives

$$1 = \sum P(E_i) = p_1 + 2p_1 + 4p_1 + 2p_1 + p_1 = 10p_1$$

implying  $p_1 = p_5 = .1$ ,  $p_2 = p_4 = .2$ ,  $p_3 = .4$ . The probability that one of the three middle cars is selected (a compound event) is then  $p_2 + p_3 + p_4 = .8$ .

- Conditional Probability:** Complex components are assembled in a plant that uses two different assembly lines,  $A$  and  $A'$ . Line  $A$  uses older equipment than  $A'$ , so it is somewhat slower and less reliable. Suppose on a given day line  $A$  has assembled 8 components, of which 2 have been identified as defective ( $B$ ) and 6 as nondefective ( $B'$ ), whereas  $A'$  has produced 1 defective and 9 nondefective components. This information is summarized in the accompanying table.

	$B$	$B'$
Line A	2	6
Line A'	1	9

Unaware of this information, the sales manager randomly selects 1 of these 18 components for a demonstration. Prior to the demonstration

$$P(\text{line A component selected}) = P(A) = \frac{N(A)}{N} = \frac{8}{18} = .44$$

However, if the chosen component turns out to be defective, then the event  $B$  has occurred, so the component must have been 1 of the 3 in the  $B$  column of the table. Since these 3 components are equally likely among themselves after  $B$  has occurred,

$$P(A|B) = \frac{2}{3} = \frac{2/18}{3/18} = \frac{P(A \cap B)}{P(B)}$$

**Why must we use conditional probability here?:** Here we can't use the multiplication rule for independent events because the events in this case are actually **dependent**. The events are dependent because the probability of a component being defective ( $B$ ) is different depending on whether it came from line  $A$  or line  $A'$ . Knowing that a component is defective changes the likelihood of it being from line  $A$  compared to line  $A'$

If events were independent:

- The defective rate would be the same for both lines.
- Knowing a component is defective would not provide any additional information about which line it came from.

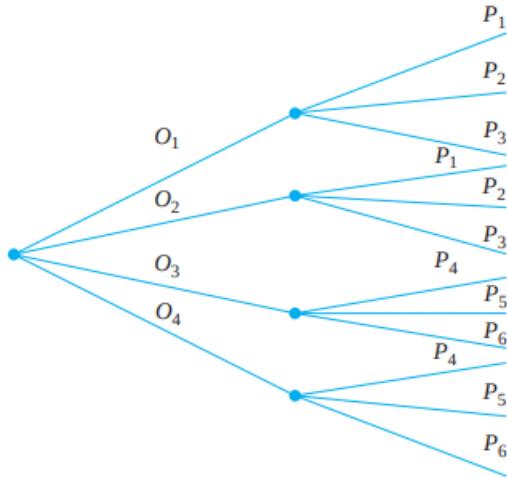
If a component is defective, it is more likely to be from line  $A$  due to the higher defective rate

Simply put, the probability of a component being defective is conditional on the line it came from, making the events  $A$  and  $B$  dependent.

### 5.2.3 Examples in counting

- A family has just moved to a new city and requires the services of both an obstetrician and a pediatrician. There are two easily accessible medical clinics, each having two obstetricians and three pediatricians. The family will obtain maximum health insurance benefits by joining a clinic and selecting both doctors from that clinic. In how many ways can this be done? Denote the obstetricians by  $O_1, O_2, O_3$ , and  $O_4$  and the pediatricians by  $P_1, \dots, P_6$ . Then we wish the number of pairs  $(O_i, P_j)$  for which  $O_i$  and  $P_j$  are associated with the same clinic. Because there are four obstetricians,  $n_1 = 4$ , and for each there are three choices of pediatrician, so  $n_2 = 3$ . Applying the product rule gives  $N = n_1 n_2 = 12$  possible choices.

**Note:** In many counting and probability problems, a configuration called a **tree diagram** can be used to represent pictorially all the possibilities



## 5.3 Chapter 3: Discrete Random Variables and Probability Distributions

### 5.3.1 Definitions and Theorems

- **random variable (rv):** is any rule that associates a number with each outcome in  $\mathcal{S}$ . In mathematical language, a random variable is a function whose domain is the sample space and whose range is the set of real numbers.

**Note:** Random variables are customarily denoted by uppercase letters, such as  $X$  and  $Y$ , near the end of our alphabet. In contrast to our previous use of a lowercase letter, such as  $x$ , to denote a variable, we will now use lowercase letters to represent some particular value of the corresponding random variable. The notation means that  $x$  is the value associated with the outcome  $s$  by the rv  $X$ .

- **Bernoulli random variable:** Any random variable whose only possible values are 0 and 1 is called a Bernoulli random variable
- **discrete random variable:** is an rv whose possible values either constitute a finite set or else can be listed in an infinite sequence in which there is a first element, a second element, and so on (“countably” infinite)
- **continuous:** if both of the following apply:
  1. Its set of possible values consists either of all numbers in a single interval on the number line (possibly infinite in extent, e.g., from  $-\infty$  to  $\infty$ ) or all numbers in a disjoint union of such intervals (e.g.,  $[0, 10] \cup [20, 30]$ ).
  2. No possible value of the variable has positive probability, that is,  $P(X = c) = 0$  for any possible value  $c$ .

**Note:** Although any interval on the number line contains an infinite number of numbers, it can be shown that there is no way to create an infinite listing of all these values—there are just too many of them. The second condition describing a continuous random variable is perhaps counterintuitive, since it would seem to imply a total probability of zero for all possible values. But we shall see in Chapter 4 that intervals of values have positive probability; the probability of an interval will decrease to zero as the width of the interval shrinks to zero.

- A **probability distribution for a discrete random variable** is a function that assigns probabilities to each possible value of the variable, such that the sum of all assigned probabilities equals 1. It is denoted

$$P(x).$$

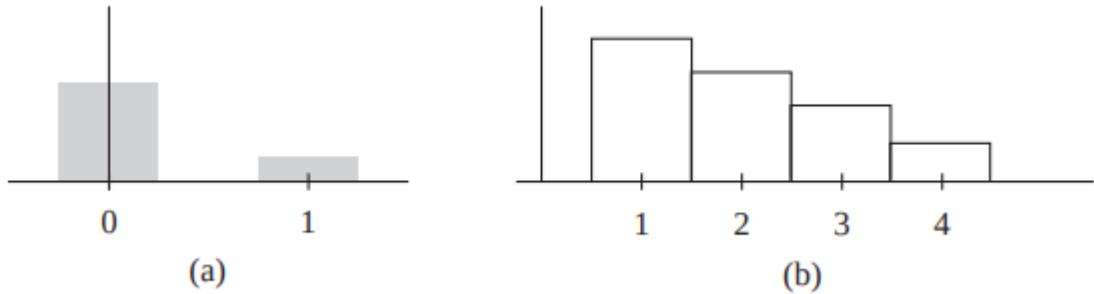
- **probability mass function:** (pmf) of a discrete rv is defined for every number  $x$  by

$$p(x) = P(X = x) = P(\text{all } s \in \mathcal{S} : X(s) = x).$$

- **pmf conditions:**

1.  $p(x) \geq 0$
2.  $\sum p(x) = 1$

- **probability histogram:** is similar to histograms discussed in Chapter 1. Above each  $y$  with  $p(y) > 0$ , construct a rectangle centered at  $y$ . The height of each rectangle is proportional to  $p(y)$ , and the base is the same for all rectangles. When possible values are equally spaced, the base is frequently chosen as the distance between successive  $y$  values (though it could be smaller). Figure 3.4 shows two probability histograms.



- **Bernoulli pmf:** The pmf of any Bernoulli rv can be expressed in the form  $p(1) = \alpha$  and  $p(0) = 1 - \alpha$ , where  $0 < \alpha < 1$ . Because the pmf depends on the particular value of  $\alpha$ , we often write  $p(x; \alpha)$  rather than just  $p(x)$ :

$$p(x; \alpha) = \begin{cases} 1 - \alpha & \text{if } x = 0 \\ \alpha & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases} \quad (3.1)$$

Then each choice of  $\alpha$  in Expression (3.1) yields a different pmf.

- Suppose  $p(x)$  depends on a quantity that can be assigned any one of a number of possible values, with each different value determining a different probability distribution. Such a quantity is called a **parameter** of the distribution. The collection of all probability distributions for different values of the parameter is called a **family** of probability distributions.

**Note:** The quantity  $\alpha$  in Expression (3.1) is a parameter. Each different number  $\alpha$  between 0 and 1 determines a different member of the Bernoulli family of distributions.

- **geometric distribution:** is a probability distribution that models the number of trials needed to achieve the first success in a sequence of independent Bernoulli trials, where each trial has the same probability of success. In other words, it describes the probability of having a certain number of failures before the first success.

There are two common variants of the geometric distribution, depending on the specific definition used:

1. **The Number of Trials Until the First Success:** This version includes the successful trial. If  $X$  is a geometric random variable, then  $X$  represents the number of trials until the first success, including the success.

– The probability mass function (pmf) for this version is:

$$P(X = k) = (1 - p)^{k-1} p \quad \text{for } k = 1, 2, 3, \dots$$

where  $p$  is the probability of success on each trial.

2. **The Number of Failures Before the First Success:** This version does not include the successful trial. If  $Y$  is a geometric random variable, then  $Y$  represents the number of failures before the first success.

– The pmf for this version is:

$$P(Y = k) = (1 - p)^k p \quad \text{for } k = 0, 1, 2, \dots$$

where  $p$  is the probability of success on each trial.

- **Mean and variance of geometric distributions:**

- **Mean (expected value):** The mean of the geometric distribution is  $\frac{1}{p}$  : for the number of trials until the first success, and  $\frac{1-p}{p}$  for the number of failures before the first success.
- **Variance:**  $\frac{1-p}{p^2}$  : for both definitions

- **cumulative distribution function (cdf):**  $F(x)$  of a discrete rv variable  $X$  with pmf  $p(x)$  is defined for every number  $x$  by

$$F(x) = P(X \leq x) = \sum_{y:y \leq x} p(y) \quad (3.3)$$

For any number  $x$ ,  $F(x)$  is the probability that the observed value of  $X$  will be at most  $x$ .

- **Recall: Geometric series:**

An infinite geometric series takes the form

$$\sum_{n=0}^{\infty} ar^n = \begin{cases} \frac{a}{1-r} & \text{if } |r| < 1 \\ \infty & \text{otherwise} \end{cases}.$$

This is the standard form, the other form is

$$\sum_{n=1}^{\infty} ar^{n-1} = \begin{cases} \frac{a}{1-r} & \text{if } |r| < 1 \\ \infty & \text{otherwise} \end{cases}.$$

The partial sum of a geometric series is

$$S_{n+1} = \sum_{k=0}^n ar^k = \frac{a(1 - r^{n+1})}{1 - r}, \quad r \neq 1.$$

The other option is

$$S_n = \sum_{k=1}^n ar^{k-1} = \sum_{k=0}^{n-1} ar^k = \frac{a(1 - r^n)}{1 - r}, \quad r \neq 1.$$

**Note:** When the index starts at zero, we say we are summing the first  $n + 1$  terms instead of the first  $n$  terms.

Summing  $m$  to  $N$  terms, we get

$$\sum_{k=m}^{m+N-1} ar^k = ar^m \sum_{k=0}^{N-1} r^k = ar^m \left( \frac{1 - r^N}{1 - r} \right), \quad r \neq 1.$$

- **pmf from cdf:** For any two numbers  $a$  and  $b$  with  $a \leq b$ ,

$$P(a \leq X \leq b) = F(b) - F(a-)$$

where “ $a-$ ” represents the largest possible  $X$  value that is strictly less than  $a$ . In particular, if the only possible values are integers and if  $a$  and  $b$  are integers, then

$$P(a \leq X \leq b) = P(X = a \text{ or } a + 1 \text{ or } \dots \text{ or } b) = F(b) - F(a - 1)$$

Taking  $a = b$  yields  $P(X = a) = F(a) - F(a - 1)$  in this case.

**Note:** The reason for subtracting  $F(a-)$  rather than  $F(a)$  is that we want to include  $P(X = a)$ ;  $F(b) - F(a)$  gives  $P(a < X \leq b)$ . This proposition will be used extensively when computing binomial and Poisson probabilities

- Let  $X$  be a discrete rv with set of possible values  $D$  and pmf  $p(x)$ . The **expected value** or **mean value** of  $X$ , denoted by  $E(X)$  or  $\mu_X$  or just  $\mu$ , is

$$E(X) = \mu_X = \sum_{x \in D} x \cdot p(x)$$

- Expected value from function:** If the rv  $X$  has a set of possible values  $D$  and pmf  $p(x)$ , then the expected value of any function  $h(X)$ , denoted by  $E[h(X)]$  or  $\mu_{h(X)}$ , is computed by

$$E[h(X)] = \sum_{x \in D} h(x) \cdot p(x)$$

**Note:** That is,  $E[h(X)]$  is computed in the same way that  $E(X)$  itself is, except that  $h(x)$  is substituted in place of  $x$

- Rules of expected value:** The  $h(X)$  function of interest is quite frequently a linear function  $aX + b$ . In this case,  $E[h(X)]$  is easily computed from  $E(X)$ .

$$E(aX + b) = a \cdot E(X) + b$$

(Or, using alternative notation,  $\mu_{aX+b} = a \cdot \mu_X + b$ )

We also have

$$E(aX + bY) = aE(X) + bE(Y).$$

- Variance of discrete rv  $X$ :** Let  $X$  have pmf  $p(x)$  and expected value  $\mu$ . Then the variance of  $X$ , denoted by  $V(X)$  or  $\sigma_X^2$ , or just  $\sigma^2$ , is

$$V(X) = \sum_D (x - \mu)^2 \cdot p(x) = \mathbb{E}[(X - \mu)^2]$$

The number of arithmetic operations necessary to compute  $\sigma^2$  can be reduced by using an alternative formula

$$V(X) = \sigma^2 = \left[ \sum_D x^2 \cdot p(x) \right] - \mu_x^2 = E(X^2) - [E(X)]^2.$$

- Standard Deviation of discrete rv  $X$ :** The standard deviation (SD) of  $X$  is

$$\sigma_X = \sqrt{\sigma_X^2}$$

- Variance/Standard deviation rules (discrete rv):**

$$V(aX + b) = \sigma_{aX+b}^2 = a^2 \cdot \sigma_X^2 \quad \text{and} \quad \sigma_{aX+b} = |a| \cdot \sigma_X$$

In particular,

$$\sigma_{aX} = |a| \cdot \sigma_X, \quad \sigma_{X+b} = \sigma_X$$

- Binomial experiment/conditions:**

1. **trials:** , where  $n$  is fixed in advance of the experiment.
2. Each trial can result in one of the same two possible outcomes (dichotomous trials), which we generically denote by success (S) and failure (F).
3. The trials are independent, so that the outcome on any particular trial does not influence the outcome on any other trial.
4. The probability of success  $P(S)$  is constant from trial to trial; we denote this probability by  $p$

An experiment for which Conditions 1–4 are satisfied is called a **binomial experiment**

**Note:** Many experiments involve a sequence of independent trials for which there are more than two possible outcomes on any one trial. A binomial experiment can then be created by dividing the possible outcomes into two groups

- **Dichotomous:** If something's dichotomous, it's divided into two distinct parts
- **Rule:** Consider sampling without replacement from a dichotomous population of size  $N$ . If the sample size (number of trials)  $n$  is at most 5% of the population size, the experiment can be analyzed as though it were exactly a binomial experiment.

**Example:** A certain state has 500,000 licensed drivers, of whom 400,000 are insured. A sample of 10 drivers is chosen without replacement. The  $i$ th trial is labeled  $S$  if the  $i$ th driver chosen is insured. Although this situation would seem identical to that of Example 3.29, the important difference is that the size of the population being sampled is very large relative to the sample size. In this case

$$P(S \text{ on } 2 | S \text{ on } 1) = \frac{399,999}{499,999} = 0.80000$$

and

$$P(S \text{ on } 10 | S \text{ on first } 9) = \frac{399,991}{499,991} = 0.799996 \approx 0.80000$$

These calculations suggest that although the trials are not exactly independent, the conditional probabilities differ so slightly from one another that for practical purposes the trials can be regarded as independent with constant  $P(S) = 0.8$ . Thus, to a very good approximation, the experiment is binomial with  $n = 10$  and  $p = 0.8$ .

- **binomial random variable:**  $X$  associated with a binomial experiment consisting of  $n$  trials is defined as

$x$  = the number of S's among the  $n$  trials

**Note:** We will often write  $X \sim Bin(n, p)$  to indicate that  $X$  is a binomial rv based on  $n$  trials with success probability  $p$ .

- **Binomial pmf notation:** Because the pmf of a binomial rv  $X$  depends on the two parameters  $n$  and  $p$ , we denote the pmf by  $b(x; n, p)$ .
- **Binomial probability distribution:**

$$b(x; n, p) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{for } x = 0, 1, 2, \dots, n \\ 0 & \text{otherwise} \end{cases}$$

- **Cdf of binomial experiment:** For  $X \sim \text{Bin}(n, p)$ , the cdf will be denoted by

$$B(x; n, p) = P(X \leq x) = \sum_{y=0}^x b(y; n, p) \quad x = 0, 1, \dots, n$$

- **Using Binomial Tables:** Even for a relatively small value of  $n$ , the computation of binomial probabilities can be tedious. Appendix Table A.1 tabulates the cdf  $F(x) = P(X \leq x)$  for  $n = 5, 10, 15, 20, 25$  in combination with selected values of  $p$ . Various other probabilities can then be calculated using the proposition on cdf's from Section 3.2. A table entry of 0 signifies only that the probability is 0 to three significant digits since all table entries are actually positive.

**Table A.1 Cumulative Binomial Probabilities**  
a.  $n = 5$

		$p$															
		0.01	0.05	0.10	0.20	0.25	0.30	0.40	0.50	0.60	0.70	0.75	0.80	0.90	0.95	0.99	
x	0	.951	.774	.590	.328	.237	.168	.078	.031	.010	.002	.001	.000	.000	.000	.000	
	1	.999	.977	.919	.737	.633	.528	.337	.188	.087	.031	.016	.007	.000	.000	.000	
	2	1.000	.999	.991	.942	.896	.837	.683	.500	.317	.163	.104	.058	.009	.001	.000	
	3	1.000	1.000	1.000	.993	.984	.969	.913	.812	.663	.472	.367	.263	.181	.081	.023	.001
	4	1.000	1.000	1.000	1.000	.999	.998	.990	.969	.922	.832	.763	.672	.510	.226	.049	

b.  $n = 10$

		$p$														
		0.01	0.05	0.10	0.20	0.25	0.30	0.40	0.50	0.60	0.70	0.75	0.80	0.90	0.95	0.99
x	0	.904	.599	.349	.107	.056	.028	.006	.001	.000	.000	.000	.000	.000	.000	.000
	1	.996	.914	.736	.376	.244	.149	.046	.011	.002	.000	.000	.000	.000	.000	.000
	2	1.000	.988	.930	.678	.526	.383	.167	.055	.012	.002	.000	.000	.000	.000	.000
	3	1.000	.999	.987	.879	.776	.650	.382	.172	.055	.011	.004	.001	.000	.000	.000
	4	1.000	1.000	.998	.967	.922	.850	.633	.377	.166	.047	.020	.006	.000	.000	.000
	5	1.000	1.000	1.000	.994	.980	.953	.834	.623	.367	.150	.078	.033	.002	.000	.000
	6	1.000	1.000	1.000	.999	.996	.989	.945	.828	.618	.350	.224	.121	.013	.001	.000
	7	1.000	1.000	1.000	1.000	.998	.988	.945	.833	.617	.474	.322	.207	.12	.004	.000
	8	1.000	1.000	1.000	1.000	1.000	.998	.989	.954	.851	.756	.624	.464	.264	.086	.004
	9	1.000	1.000	1.000	1.000	1.000	1.000	.999	.994	.972	.944	.893	.851	.401	.096	

c.  $n = 15$

		$p$														
		0.01	0.05	0.10	0.20	0.25	0.30	0.40	0.50	0.60	0.70	0.75	0.80	0.90	0.95	0.99
x	0	.860	.463	.206	.035	.013	.005	.000	.000	.000	.000	.000	.000	.000	.000	.000
	1	.990	.829	.549	.167	.080	.035	.005	.000	.000	.000	.000	.000	.000	.000	.000
	2	1.000	.964	.816	.398	.236	.127	.027	.004	.000	.000	.000	.000	.000	.000	.000
	3	1.000	.995	.944	.648	.461	.297	.091	.018	.002	.000	.000	.000	.000	.000	.000
	4	1.000	.999	.987	.836	.686	.515	.217	.059	.009	.001	.000	.000	.000	.000	.000
	5	1.000	1.000	.998	.939	.852	.722	.403	.151	.034	.004	.001	.000	.000	.000	.000
	6	1.000	1.000	1.000	.982	.943	.869	.610	.304	.095	.015	.004	.001	.000	.000	.000
	7	1.000	1.000	1.000	.996	.983	.950	.787	.500	.213	.050	.017	.004	.000	.000	.000
	8	1.000	1.000	1.000	.999	.996	.985	.905	.696	.390	.131	.057	.018	.000	.000	.000
	9	1.000	1.000	1.000	.999	.996	.996	.966	.849	.597	.278	.148	.061	.002	.000	.000
x	10	1.000	1.000	1.000	1.000	.999	.991	.941	.783	.485	.314	.164	.013	.001	.000	.000
	11	1.000	1.000	1.000	1.000	1.000	.998	.982	.909	.703	.539	.352	.056	.005	.000	.000
	12	1.000	1.000	1.000	1.000	1.000	1.000	.996	.973	.873	.764	.602	.184	.036	.000	.000
	13	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.995	.965	.920	.833	.451	.171	.010	.000
	14	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.995	.987	.965	.794	.537	.140	.000

(continued)

**Table A.1 Cumulative Binomial Probabilities (cont.)**d.  $n = 20$ 

$$B(x; n, p) = \sum_{y=0}^x b(y; n, p)$$

	<i>p</i>														
	0.01	0.05	0.10	0.20	0.25	0.30	0.40	0.50	0.60	0.70	0.75	0.80	0.90	0.95	0.99
0	.818	.358	.122	.012	.003	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000
1	.983	.736	.392	.069	.024	.008	.001	.000	.000	.000	.000	.000	.000	.000	.000
2	.999	.925	.677	.206	.091	.035	.004	.000	.000	.000	.000	.000	.000	.000	.000
3	1.000	.984	.867	.411	.225	.107	.016	.001	.000	.000	.000	.000	.000	.000	.000
4	1.000	.997	.957	.630	.415	.238	.051	.006	.000	.000	.000	.000	.000	.000	.000
5	1.000	1.000	.989	.804	.617	.416	.126	.021	.002	.000	.000	.000	.000	.000	.000
6	1.000	1.000	.998	.913	.786	.608	.250	.058	.006	.000	.000	.000	.000	.000	.000
7	1.000	1.000	1.000	.968	.898	.772	.416	.132	.021	.001	.000	.000	.000	.000	.000
8	1.000	1.000	1.000	.990	.959	.887	.596	.252	.057	.005	.001	.000	.000	.000	.000
9	1.000	1.000	1.000	.997	.986	.952	.755	.412	.128	.017	.004	.001	.000	.000	.000
x	10	1.000	1.000	1.000	.999	.996	.983	.872	.588	.245	.048	.014	.003	.000	.000
11	1.000	1.000	1.000	1.000	.999	.995	.943	.748	.404	.113	.041	.010	.000	.000	.000
12	1.000	1.000	1.000	1.000	1.000	.999	.979	.868	.584	.228	.102	.032	.000	.000	.000
13	1.000	1.000	1.000	1.000	1.000	1.000	.994	.942	.750	.392	.214	.087	.002	.000	.000
14	1.000	1.000	1.000	1.000	1.000	1.000	.998	.979	.874	.584	.383	.196	.011	.000	.000
15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.994	.949	.762	.585	.370	.043	.003	.000
16	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.999	.984	.893	.775	.589	.133	.016	.000
17	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.996	.965	.909	.794	.323	.075	.001
18	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.999	.992	.976	.931	.608	.264	.017
19	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.999	.997	.988	.878	.642	.182

(continued)

$$A'_1 \cap A'_2 \cap A'_3 = (A_1 \cup A_2 \cup A_3)'.$$

- **Binomial distribution mean and variance:** If  $X \sim \text{Bin}(n, p)$ , then  $E(X) = np$ ,  $V(X) = np(1 - p) = npq$ , and  $\sigma_X = \sqrt{npq}$  (where  $q = 1 - p$ ).
- **Bernoulli distribution:** The Bernoulli distribution is a special case of the binomial distribution where a single trial is conducted (so  $n$  would be 1 for such a binomial distribution)
- **The Hypergeometric Distribution:** The assumptions leading to the hypergeometric distribution are as follows:
  1. The population or set to be sampled consists of  $N$  individuals, objects, or elements (a finite population).
  2. Each individual can be characterized as a success (S) or a failure (F), and there are  $M$  successes in the population.
  3. A sample of  $n$  individuals is selected without replacement in such a way that each subset of size  $n$  is equally likely to be chosen
- **Hypergeometric distribution notation:** The random variable of interest is  $X =$  the number of S's in the sample. The probability distribution of  $X$  depends on the parameters  $n$ ,  $M$ , and  $N$ , so we wish to obtain  $P(X = x) = h(x; n, M, N)$ .
- **Hypegeometric distribution range of  $X$  values:** In general, if the sample size  $n$  is smaller than the number of successes in the population ( $M$ ), then the largest possible  $X$  value is  $n$ . However, if  $M < n$  (e.g., a sample size of 25 and only 15 successes in the population), then  $X$  can be at most  $M$ . Similarly, whenever the number of population failures ( $N - M$ ) exceeds the sample size, the smallest possible  $X$  value is 0 (since all sampled individuals might then be failures). However, if  $N - M < n$ , the smallest possible  $X$  value is  $n - (N - M)$ . Thus, the possible values of  $X$  satisfy the restriction

$$\max(0, n - (N - M)) \leq x \leq \min(n, M).$$

- **Hypegeometric distribution probabilities:** If  $X$  is the number of S's in a completely random sample of size  $n$  drawn from a population consisting of  $M$  S's and  $(N - M)$  F's, then the probability distribution of  $X$ , called the **hypergeometric distribution**, is given by

$$P(X = x) = h(x; n, M, N) = \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}}$$

for  $x$ , an integer, satisfying  $\max(0, n - N + M) \leq x \leq \min(n, M)$ .

- **The mean and variance of the hypergeometric:** The mean and variance of the hypergeometric rv  $X$  having pmf  $h(x; n, M, N)$  are

$$E(X) = n \cdot \frac{M}{N} \quad V(X) = \left( \frac{N-n}{N-1} \right) \cdot n \cdot \frac{M}{N} \cdot \left( 1 - \frac{M}{N} \right)$$

The ratio  $M/N$  is the proportion of S's in the population. If we replace  $M/N$  by  $p$  in  $E(X)$  and  $V(X)$ , we get

$$E(X) = np \quad V(X) = \left( \frac{N-n}{N-1} \right) \cdot np(1-p)$$

Expression shows that the means of the binomial and hypergeometric rv's are equal, whereas the variances of the two rv's differ by the factor  $\left( \frac{N-n}{N-1} \right)$ , often called the *finite population correction factor*. This factor is less than 1, so the hypergeometric variable has smaller variance than does the binomial rv. The correction factor can be written  $\left( \frac{1-\frac{n}{N}}{1-\frac{1}{N}} \right)$ , which is approximately 1 when  $n$  is small relative to  $N$ .

- **$N$  not known:** Suppose the population size  $N$  is not actually known, so the value  $x$  is observed and we wish to estimate  $N$ . It is reasonable to equate the observed sample proportion of S's,  $x/n$ , with the population proportion,  $M/N$ , giving the estimate

$$\hat{N} = \frac{M \cdot n}{x}$$

If  $M = 100$ ,  $n = 40$ , and  $x = 16$ , then  $\hat{N} = 250$ .

Our general rule of thumb in Section 3.4 stated that if sampling was without replacement but  $n/N$  was at most .05, then the binomial distribution could be used to compute approximate probabilities involving the number of S's in the sample. A more precise statement is as follows: Let the population size,  $N$ , and number of population S's,  $M$ , get large with the ratio  $M/N$  approaching  $p$ . Then  $h(x; n, M, N)$  approaches  $b(x; n, p)$ ; so for  $n/N$  small, the two are approximately equal provided that  $p$  is not too near either 0 or 1. This is the rationale for the rule.

- **The Negative Binomial Distribution Conditions:** The negative binomial rv and distribution are based on an experiment satisfying the following conditions:

1. The experiment consists of a sequence of independent trials.
2. Each trial can result in either a success (S) or a failure (F).
3. The probability of success is constant from trial to trial, so  $P(\text{S on trial } i = p)$  for  $i = 1, 2, 3, \dots$
4. The experiment continues (trials are performed) until a total of  $r$  successes have been observed, where  $r$  is a specified positive integer

- **Negative Binomial Random Variable:**

- **Standard:**  $X =$  the number of trials needed to achieve  $r$  successes
- **Alternate:** The random variable of interest is  $X =$  the number of failures that precede the  $r$ th success;  $X$  is called a *negative binomial random variable*: because, in contrast to the binomial rv, the number of successes is fixed and the number of trials is random.

- **Negative Binomial pmf:**

- **Standard:** The pmf of the negative binomial random variable  $X$  with parameters  $r$  (number of successes) and  $p$  (probability of success) is given by:

$$P(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r} \quad \text{for } x = r, r+1, r+2, \dots$$

where  $y$  represents the total number of trials needed to achieve  $r$  successes.

- **Alternate:** The pmf of the negative binomial rv  $X$  with parameters  $r =$  number of  $S$ 's and  $p = P(S)$  is

$$nb(x; r, p) = \binom{x+r-1}{r-1} p^r (1-p)^x \quad x = 0, 1, 2, \dots$$

Where  $x$  values are the number of failures before the  $r^{th}$  success

- **Negative binomial expected value and variance:**

- **Standard:**

$$E(x) = \frac{r}{p} \quad \text{and} \quad V(X) = \frac{r(1-p)}{p^2}.$$

- **Alternate:** If  $X$  is a negative binomial rv with pmf  $nb(x; r, p)$ , then

$$E(X) = \frac{r(1-p)}{p} \quad \text{and} \quad V(X) = \frac{r(1-p)}{p^2}$$

- **Poisson Distribution:** A discrete random variable  $X$  is said to have a *Poisson distribution* with parameter  $\mu$  ( $\mu > 0$ ) if the pmf of  $X$  is

$$p(x; \mu) = \frac{e^{-\mu} \cdot \mu^x}{x!} \quad \text{for } x = 0, 1, 2, 3, \dots$$

**Note:** It is no accident that we are using the symbol  $\mu$  for the Poisson parameter; we shall see shortly that  $\mu$  is in fact the expected value of  $X$ . The letter  $e$  in the pmf represents the base of the natural logarithm system; its numerical value is approximately 2.71828. In contrast to the binomial and hypergeometric distributions, the Poisson distribution spreads probability over all non-negative integers, an infinite number of possibilities.

It is not obvious by inspection that  $p(x; \mu)$  specifies a legitimate pmf, let alone that this distribution is useful. First of all,  $p(x; \mu) > 0$  for every possible  $x$  value because of the requirement that  $\mu > 0$ . The fact that  $\sum p(x; \mu) = 1$  is a consequence of the Maclaurin series expansion of  $e^\mu$  (check your calculus book for this result):

$$e^\mu = 1 + \mu + \frac{\mu^2}{2!} + \frac{\mu^3}{3!} + \dots = \sum_{x=0}^{\infty} \frac{\mu^x}{x!} \quad (3.18)$$

If the two extreme terms in (3.18) are multiplied by  $e^{-\mu}$  and then this quantity is moved inside the summation on the far right, the result is

$$1 = \sum_{x=0}^{\infty} \frac{e^{-\mu} \cdot \mu^x}{x!}$$

- **Rationale for the poisson distribution:** The rationale for using the Poisson distribution in many situations is provided by the following proposition.

Suppose that in the binomial pmf  $b(x; n, p)$ , we let  $n \rightarrow \infty$  and  $p \rightarrow 0$  in such a way that  $np$  approaches a value  $\mu > 0$ . Then  $b(x; n, p) \rightarrow p(x; \mu)$ .

**Note:** According to this proposition, *in any binomial experiment in which  $n$  is large and  $p$  is small,  $b(x; n, p) \approx p(x; \mu)$* , where  $\mu = np$ . As a rule of thumb, this approximation can safely be applied if  $n > 50$  and  $np < 5$ .

- **Example: Using the poisson distribution to estimate binomial probabilities:**

If a publisher of nontechnical books takes great pains to ensure that its books are free of typographical errors, so that the probability of any given page containing at least one such error is .005 and errors are independent from page to page, what is the probability that one of its 400-page novels will contain exactly one page with errors? At most three pages with errors?

With  $S$  denoting a page containing at least one error and  $F$  an error-free page, the number  $X$  of pages containing at least one error is a binomial rv with  $n = 400$  and  $p = .005$ , so  $np = 2$ . We wish

$$P(X = 1) = b(1; 400, .005) \approx p(1; 2) = \frac{e^{-2}(2)^1}{1!} = .270671$$

The binomial value is  $b(1; 400, .005) = .270669$ , so the approximation is very good.

Similarly,

$$\begin{aligned} P(X \leq 3) &\approx \sum_{x=0}^3 p(x; 2) = \sum_{x=0}^3 \frac{e^{-2}(2)^x}{x!} \\ &= .135335 + .270671 + .270671 + .180447 = .8571 \end{aligned}$$

and this again is quite close to the binomial value  $P(X \leq 3) = .8576$ .

- **Mean and variance of poisson distribution:** If  $X$  has a Poisson distribution with parameter  $\mu$ , then  $E(X) = V(X) = \mu$ .
- **The Poisson Process introduction:** A very important application of the Poisson distribution arises in connection with the occurrence of events of some type over time. Events of interest might be visits to a particular website, pulses of some sort recorded by a counter, email messages sent to a particular address, accidents in an industrial facility, or cosmic ray showers observed by astronomers at a particular observatory. We make the following assumptions about the way in which the events of interest occur:

1. There exists a parameter  $\alpha > 0$  such that for any short time interval of length  $\Delta t$ , the probability that exactly one event occurs is  $\alpha \cdot \Delta t + o(\Delta t)$

2. The probability of more than one event occurring during  $\Delta t$  is  $o(\Delta t)$  [which, along with Assumption 1, implies that the probability of no events during  $\Delta t$  is  $1 - \alpha \cdot \Delta t - o(\Delta t)$ ].
3. The number of events occurring during the time interval  $\Delta t$  is independent of the number that occur prior to this time interval.

Informally, Assumption 1 says that for a short interval of time, the probability of a single event occurring is approximately proportional to the length of the time interval, where  $\alpha$  is the constant of proportionality. Now let  $P_k(t)$  denote the probability that  $k$  events will be observed during any particular time interval of length  $t$ .

- **The Poisson Process:**

$$P_k(t) = e^{-\alpha t} \cdot \frac{(\alpha t)^k}{k!}$$

so that the number of events during a time interval of length  $t$  is a Poisson rv with parameter  $\mu = \alpha t$ . The expected number of events during any such time interval is then  $\alpha t$ , so the expected number during a unit interval of time is  $\alpha$ .

The occurrence of events over time as described is called a Poisson process; the parameter  $\alpha$  specifies the *rate* for the process.

- **Little-o notation:**  $o(\Delta t)$  is the "little-o notation," which is used in mathematics to describe the limiting behavior of a function when the argument tends toward a particular value. Specifically,  $o(\Delta t)$  denotes a function that grows significantly slower than  $\Delta t$  as  $\Delta t$  approaches 0.

Formally,  $f(\Delta t) = o(\Delta t)$  as  $\Delta t \rightarrow 0$  if and only if  $\frac{f(\Delta t)}{\Delta t} \rightarrow 0$  as  $\Delta t \rightarrow 0$ .

- **Poisson process example:**

Suppose pulses arrive at a counter at an average rate of six per minute, so that  $\alpha = 6$ . To find the probability that in a .5-min interval at least one pulse is received, note that the number of pulses in such an interval has a Poisson distribution with parameter  $\alpha t = 6 \cdot 0.5 = 3$  (.5 min is used because  $\alpha$  is expressed as a rate per minute). Then with  $X$  = the number of pulses received in the 30-sec interval,

$$P(1 \leq X) = 1 - P(X = 0) = 1 - \frac{e^{-3}(3)^0}{0!} = 1 - \frac{e^{-3} \cdot 1}{1} = 1 - e^{-3} = 1 - 0.0498 \approx 0.950$$

### 5.3.2 Discrete rv distribution problems

- Consider whether the next person buying a computer at a certain electronics store buys a laptop or a desktop model. Let

$$X = \begin{cases} 1 & \text{if the customer purchases a desktop computer} \\ 0 & \text{if the customer purchases a laptop computer} \end{cases}$$

If 20% of all purchasers during that week select a desktop, the pmf for  $X$  is

$$\begin{aligned} p(0) &= P(X = 0) = P(\text{next customer purchases a laptop model}) = 0.8 \\ p(1) &= P(X = 1) = P(\text{next customer purchases a desktop model}) = 0.2 \\ p(x) &= P(X = x) = 0 \text{ for } x \neq 0 \text{ or } 1 \end{aligned}$$

An equivalent description is

$$p(x) = \begin{cases} 0.8 & \text{if } x = 0 \\ 0.2 & \text{if } x = 1 \\ 0 & \text{if } x \neq 0 \text{ or } 1 \end{cases}$$

The figure below is a picture of this pmf, called a line graph.  $X$  is, of course, a Bernoulli rv and  $p(x)$  is a Bernoulli pmf.



- Starting at a fixed time, we observe the gender of each newborn child at a certain hospital until a boy (B) is born. Let  $p = P(B)$ , assume that successive births are independent, and define the rv  $X$  by  $x = \text{number of births observed}$ . Then

$$\begin{aligned} p(1) &= P(X = 1) = P(B) = p \\ p(2) &= P(X = 2) = P(GB) = P(G) \cdot P(B) = (1 - p)p \end{aligned}$$

and

$$p(3) = P(X = 3) = P(GGB) = P(G) \cdot P(G) \cdot P(B) = (1 - p)^2 p$$

Continuing in this way, a general formula emerges:

$$p(x) = \begin{cases} (1 - p)^{x-1} p & \text{if } x = 1, 2, 3, \dots \\ 0 & \text{otherwise} \end{cases}$$

The parameter  $p$  can assume any value between 0 and 1. Expression (3.2) describes the family of *geometric distributions*. In the gender example,  $p = .51$  might be appropriate, but if we were looking for the first child with Rh-positive blood, then we might have  $p = .85$ .

- **cdf problem:** A store carries flash drives with either 1 GB, 2 GB, 4 GB, 8 GB, or 16 GB of memory. The accompanying table gives the distribution of  $Y$  = the amount of memory in a purchased drive:

$y$	1	2	4	8	16
$p(y)$	.05	.10	.35	.40	.10

Let's first determine  $F(y)$  for each of the five possible values of  $Y$ :

$$\begin{aligned} F(1) &= P(Y \leq 1) = P(Y = 1) = p(1) = .05 \\ F(2) &= P(Y \leq 2) = P(Y = 1 \text{ or } 2) = p(1) + p(2) = .15 \\ F(4) &= P(Y \leq 4) = P(Y = 1 \text{ or } 2 \text{ or } 4) = p(1) + p(2) + p(4) = .50 \\ F(8) &= P(Y \leq 8) = P(Y = 1 \text{ or } 2 \text{ or } 4 \text{ or } 8) = p(1) + p(2) + p(4) + p(8) = .90 \\ F(16) &= P(Y \leq 16) = 1 \end{aligned}$$

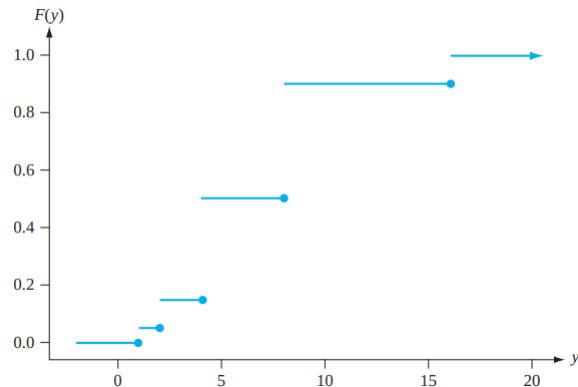
Now for any other number  $y$ ,  $F(y)$  will equal the value of  $F$  at the closest possible value of  $Y$  to the left of  $y$ . For example,

$$\begin{aligned} F(2.7) &= P(Y \leq 2.7) = P(Y \leq 2) = F(2) = .15 \\ F(7.999) &= P(Y \leq 7.999) = P(Y \leq 4) = F(4) = .50 \end{aligned}$$

If  $y$  is less than 1,  $F(y) = 0$  [e.g.  $F(.58) = 0$ ], and if  $y$  is at least 16,  $F(y) = 1$  [e.g.  $F(25) = 1$ ]. The cdf is thus

$$F(y) = \begin{cases} 0 & y < 1 \\ .05 & 1 \leq y < 2 \\ .15 & 2 \leq y < 4 \\ .50 & 4 \leq y < 8 \\ .90 & 8 \leq y < 16 \\ 1 & 16 \leq y \end{cases}$$

A graph of this cdf is shown in the figure below



**Note:** For  $X$  a discrete rv, the graph of  $F(x)$  will have a jump at every possible value of  $X$  and will be flat between possible values. Such a graph is called a **step function**.

## 5.4 Chapter 4: Continuous Random variables and Probability Distributions

### 5.4.1 Definitions and Theorems

- **probability density function:** (pdf) of  $X$  is a function  $f(x)$  such that for any two numbers  $a$  and  $b$  with  $a \leq b$ ,

$$P(a \leq X \leq b) = \int_a^b f(x) dx$$

That is, the probability that  $X$  takes on a value in the interval  $[a, b]$  is the area above this interval and under the graph of the density function, as illustrated in Figure 4.2. The graph of  $f(x)$  is often referred to as the **density curve**.

- **pdf conditions:** For  $f(x)$  to be a legitimate pdf, it must satisfy the following two conditions:
  1.  $f(x) \geq 0$  for all  $x$
  2.  $\int_{-\infty}^{\infty} f(x) dx = \text{area under the entire graph of } f(x) = 1$
- **Uniform Distribution:** A continuous rv  $X$  is said to have a **uniform distribution** on the interval  $[A, B]$  if the pdf of  $X$  is

$$f(x; A, B) = \begin{cases} \frac{1}{B-A} & A \leq x \leq B \\ 0 & \text{otherwise} \end{cases}$$

**Note:** The expected value for a uniform distribution  $E(X)$  is

$$\frac{a+b}{2}.$$

- **cumulative distribution function:**  $F(x)$  for a continuous rv  $X$  is defined for every number  $x$  by

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(y) dy$$

For each  $x$ ,  $F(x)$  is the area under the density curve to the left of  $x$ . This is illustrated in Figure 4.5, where  $F(x)$  increases smoothly as  $x$  increases.

- **Using F(x) to Compute Probabilities:** Let  $X$  be a continuous rv with pdf  $f(x)$  and cdf  $F(x)$ . Then for any number  $a$ ,

$$P(X > a) = 1 - F(a)$$

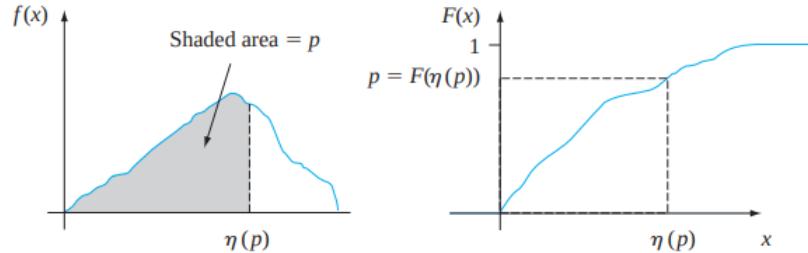
and for any two numbers  $a$  and  $b$  with  $a < b$ ,

$$P(a \leq X \leq b) = F(b) - F(a)$$

- **Obtaining f(x) from F(X):** If  $X$  is a continuous rv with pdf  $f(x)$  and cdf  $F(x)$ , then at every  $x$  at which the derivative  $F'(x)$  exists,  $F'(x) = f(x)$ .
- **Percentiles of a continuous distribution:** Let  $p$  be a number between 0 and 1. The  $(100p)$ th percentile of the distribution of a continuous rv  $X$ , denoted by  $\eta(p)$ , is defined by

$$p = F(\eta(p)) = \int_{-\infty}^{\eta(p)} f(y) dy$$

According to this expression,  $\eta(p)$  is that value on the measurement axis such that  $100p\%$  of the area under the graph of  $f(x)$  lies to the left of  $\eta(p)$  and  $100(1-p)\%$  lies to the right. Thus  $\eta(.75)$ , the 75th percentile, is such that the area under the graph of  $f(x)$  to the left of  $\eta(.75)$  is .75. Figure 4.10 illustrates the definition.



**Figure 4.10** The  $(100p)$ th percentile of a continuous distribution

- **Percentile example:** The distribution of the amount of gravel (in tons) sold by a particular construction supply company in a given week is a continuous rv  $X$  with pdf

$$f(x) = \begin{cases} \frac{3}{2}(1-x^2) & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

The cdf of sales for any  $x$  between 0 and 1 is

$$F(x) = \int_0^x \frac{3}{2}(1-y^2) dy = \frac{3}{2} \left( y - \frac{y^3}{3} \right) \Big|_0^x = \frac{3}{2} \left( x - \frac{x^3}{3} \right) = \frac{3}{2} \left( x - \frac{x^3}{3} \right)$$

The graphs of both  $f(x)$  and  $F(x)$  appear in Figure 4.11. The  $(100p)$ th percentile of this distribution satisfies the equation

$$p = F(\eta(p)) = \frac{3}{2} \left[ \eta(p) - \frac{(\eta(p))^3}{3} \right]$$

that is,

$$(\eta(p))^3 - 3\eta(p) + 2p = 0$$

For the 50th percentile,  $p = .5$ , and the equation to be solved is  $\eta^3 - 3\eta + 1 = 0$ ; the solution is  $\eta = \eta(.5) = .347$ . If the distribution remains the same from week to week, then in the long run 50% of all weeks will result in sales of less than .347 ton and 50% in more than .347 ton.

- **median** of a continuous distribution, denoted by  $\tilde{\mu}$ , is the 50th percentile, so  $\tilde{\mu}$  satisfies  $.5 = F(\tilde{\mu})$ . That is, half the area under the density curve is to the left of  $\tilde{\mu}$  and half is to the right of  $\tilde{\mu}$ .
- **Symmetric pdf:** A continuous distribution whose pdf is symmetric—the graph of the pdf to the left of some point is a mirror image of the graph to the right of that point—has median equal  $\tilde{\mu}$  to the point of symmetry, since half the area under the curve lies to either side of this point

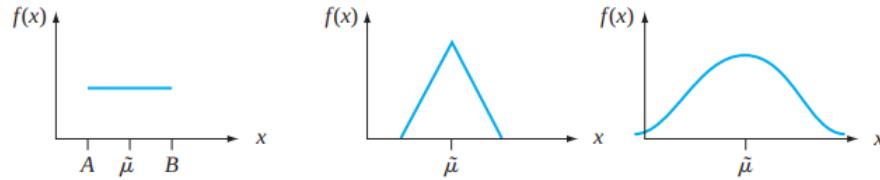


Figure 4.12 Medians of symmetric distributions

- **pdf Expected value:** The **expected** or **mean value** of a continuous rv  $X$  with pdf  $f(x)$  is

$$\mu_X = E(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

If  $X$  is a continuous rv with pdf  $f(x)$  and  $h(X)$  is any function of  $X$ , then

$$E[h(X)] = \mu_{h(X)} = \int_{-\infty}^{\infty} h(x) \cdot f(x) dx$$

- **pdf Variance, Standard Dev:** The **variance** of a continuous random variable  $X$  with pdf  $f(x)$  and mean value  $\mu$  is

$$\sigma_X^2 = V(X) = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x) dx = E[(X - \mu)^2]$$

The **standard deviation** (SD) of  $X$  is  $\sigma_X = \sqrt{V(X)}$ .

- **pdf Variance Shortcut:**

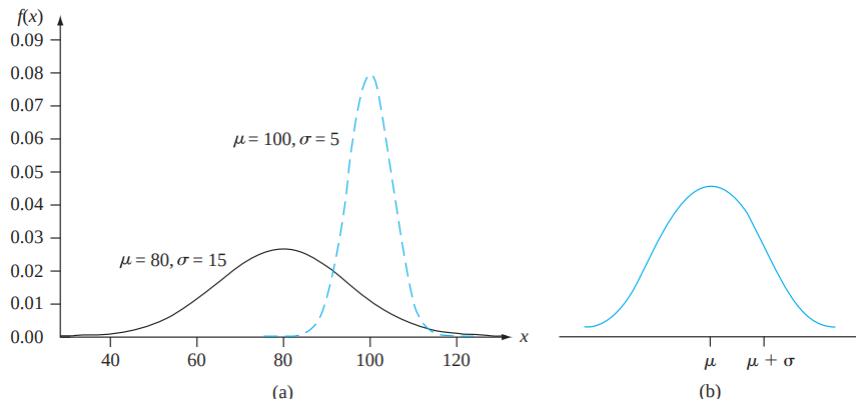
$$V(X) = E(X^2) - [E(X)]^2.$$

- **Normal (Guassian) distribution:** A continuous rv  $X$  is said to have a **normal distribution** with parameters  $\mu$  and  $\sigma$  (or  $\mu$  and  $\sigma^2$ ), where  $-\infty < \mu < \infty$  and  $0 < \sigma$ , if the pdf of  $X$  is

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad -\infty < x < \infty$$

The statement that  $X$  is normally distributed with parameters  $\mu$  and  $\sigma^2$  is often abbreviated  $X \sim N(\mu, \sigma^2)$

Clearly  $f(x; \mu, \sigma) \geq 0$ , but a somewhat complicated calculus argument must be used to verify that  $\int_{-\infty}^{\infty} f(x; \mu, \sigma) dx = 1$ . It can be shown that  $E(X) = \mu$  and  $V(X) = \sigma^2$ , so the parameters are the mean and the standard deviation of  $X$ . Figure 4.13 presents graphs of  $f(x; \mu, \sigma)$  for several different  $(\mu, \sigma)$  pairs. Each density curve is symmetric about  $\mu$  and bell-shaped, so the center of the bell (point of symmetry) is both the mean of the distribution and the median. The value of  $\sigma$  is the distance from  $\mu$  to the inflection points of the curve (the points at which the curve changes from turning downward to turning upward). Large values of  $\sigma$  yield graphs that are quite spread out about  $\mu$ , whereas small values of  $\sigma$  yield graphs with a high peak above  $\mu$  and most of the area under the graph quite close to  $\mu$ . Thus a large  $\sigma$  implies that a value of  $X$  far from  $\mu$  may well be observed, whereas such a value is quite unlikely when  $\sigma$  is small.

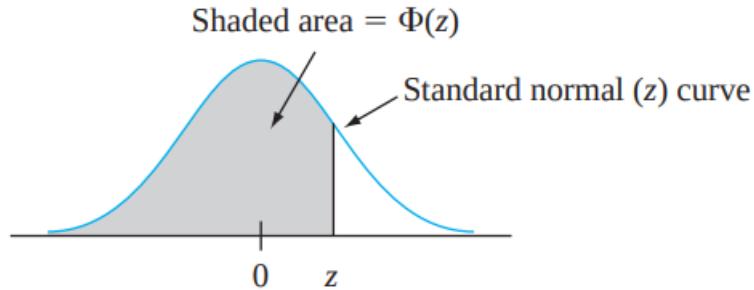


- **The Standard Normal Distribution:** The normal distribution with parameter values  $\mu = 0$  and  $\sigma = 1$  is called the *standard normal distribution*. A random variable having a standard normal distribution is called a *standard normal random variable* and will be denoted by  $Z$ . The pdf of  $Z$  is

$$f(z; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \quad -\infty < z < \infty$$

The graph of  $f(z; 0, 1)$  is called the *standard normal* (or  $z$ ) curve. Its inflection points are at 1 and  $-1$ . The cdf of  $Z$  is  $P(Z \leq z) = \int_{-\infty}^z f(y; 0, 1) dy$ , which we will denote by  $\Phi(z)$ .

The standard normal distribution almost never serves as a model for a naturally arising population. Instead, it is a reference distribution from which information about other normal distributions can be obtained. Appendix Table A.3 gives  $\Phi(z) = P(Z \leq z)$ , the area under the standard normal density curve to the left of  $z$ , for  $z = -3.49, -3.48, \dots, 3.48, 3.49$ . Figure 4.14 illustrates the type of cumulative area (probability) tabulated in Table A.3. From this table, various other probabilities involving  $Z$  can be calculated.



<i>z</i>	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0017	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0038
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0352	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0722	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3482
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
-0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

(continued)

Table A.3 Standard Normal Curve Areas (cont.)

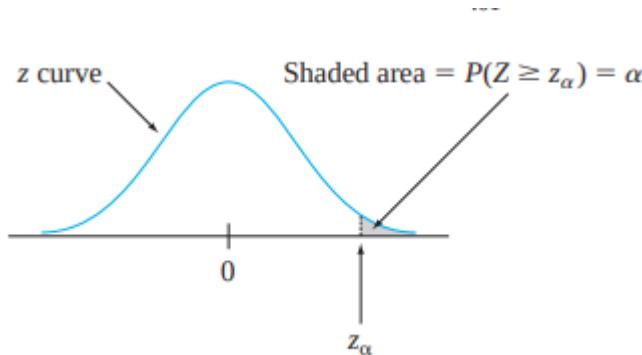
$\Phi(z) = P(Z \leq z)$

<i>z</i>	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8109	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9278	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

- **Percentiles of the Standard Normal Distribution:** For any  $p$  between 0 and 1, Appendix Table A.3 can be used to obtain the  $(100p)$ th percentile of the standard normal distribution.

- **$z_\alpha$  Notation for  $z$  Critical Values:**  $z_\alpha$  will denote the value on the  $z$  axis for which  $\alpha$  of the area under the  $z$  curve lies to the right of  $z_\alpha$ . (See Figure 4.19.)

For example  $z_{.01}$ , captures upper-tail area .10, and captures upper-tail area .01.



Since  $\alpha$  of the area under the  $z$  curve lies to the right of  $z_\alpha$ ,  $1 - \alpha$  of the area lies to its left. Thus  $z_\alpha$  is the  $100(1 - \alpha)$ th percentile of the standard normal distribution. By symmetry the area under the standard normal curve to the left of  $-z_\alpha$  is also  $\alpha$ . The  $z_\alpha$ 's are usually referred to as  $z$  critical values. Table 4.1 lists the most useful  $z$  percentiles and  $z_\alpha$  values.

**Table 4.1 Standard Normal Percentiles and Critical Values**

Percentile	90	95	97.5	99	99.5	99.9	99.95
$\alpha$ (tail area)	.1	.05	.025	.01	.005	.001	.0005
$z_\alpha = 100(1 - \alpha)$ th percentile	1.28	1.645	1.96	2.33	2.58	3.08	3.27

- **Nonstandard Normal Distributions:** When  $X \sim N(\mu, \sigma^2)$ , probabilities involving  $X$  are computed by “standardizing.” The *standardized variable* is  $(X - \mu)/\sigma$ . Subtracting  $\mu$  shifts the mean from  $\mu$  to zero, and then dividing by  $\sigma$  scales the variable so that the standard deviation is 1 rather than  $\sigma$ .

If  $X$  has a normal distribution with mean  $\mu$  and standard deviation  $\sigma$ , then

$$Z = \frac{X - \mu}{\sigma}$$

has a standard normal distribution. Thus

$$\begin{aligned} P(a \leq X \leq b) &= P\left(\frac{a - \mu}{\sigma} \leq Z \leq \frac{b - \mu}{\sigma}\right) \\ &= \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right) \\ P(X \leq a) &= \Phi\left(\frac{a - \mu}{\sigma}\right) \quad P(X \geq b) = 1 - \Phi\left(\frac{b - \mu}{\sigma}\right) \end{aligned}$$

- **Percentiles of an Arbitrary Normal Distribution:** The  $(100p)^{\text{th}}$  percentile of a normal distribution with mean  $\mu$  and standard deviation  $\sigma$  is easily related to the  $(100p)^{\text{th}}$  percentile of the standard normal distribution.

$$(100p)^{\text{th}} \text{ percentile for normal } (\mu, \sigma) = \mu + [(100p)^{\text{th}} \text{ for standard normal}] \cdot \sigma$$

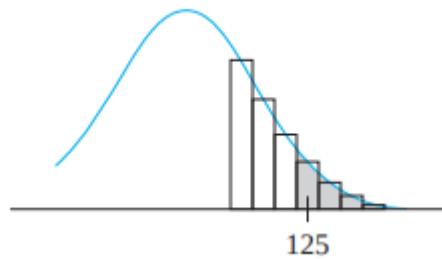
Another way of saying this is that if  $z$  is the desired percentile for the standard normal distribution, then the desired percentile for the normal  $(\mu, \sigma)$  distribution is  $z$  standard deviations from  $\mu$ .

- **The Normal Distribution and Discrete Populations:** The normal distribution is often used as an approximation to the distribution of values in a discrete population. In such situations, extra care should be taken to ensure that probabilities are computed in an accurate manner.

For example:

IQ in a particular population (as measured by a standard test) is known to be approximately normally distributed with  $\mu = 100$  and  $\sigma = 15$ . What is the probability that a randomly selected individual has an IQ of at least 125? Letting  $X$  = the IQ of a randomly chosen person, we wish  $P(X \geq 125)$ . The temptation here is to standardize  $X \geq 125$  as in previous examples. However, the IQ population distribution is actually discrete, since IQs are integer-valued. So the normal curve is an approximation to a discrete probability histogram, as pictured in Figure 4.24.

The rectangles of the histogram are centered at integers, so IQs of at least 125 correspond to rectangles beginning at 124.5, as shaded in Figure 4.24. Thus we really want the area under the approximating normal curve to the right of 124.5. Standardizing this value gives  $P(Z \geq 1.63) = 0.0516$ , whereas standardizing 125 results in  $P(Z \geq 1.67) = 0.0475$ . The difference is not great, but the answer 0.0516 is more accurate. Similarly,  $P(X = 125)$  would be approximated by the area between 124.5 and 125.5, since the area under the normal curve above the single value 125 is zero.



The correction for discreteness of the underlying distribution in Example 4.19 is often called a **continuity correction**. It is useful in the following application of the normal distribution to the computation of binomial probabilities

- **Approximating the Binomial Distribution:** Let  $X$  be a binomial rv based on  $n$  trials with success probability  $p$ . Then if the binomial probability histogram is not too skewed,  $X$  has approximately a normal distribution with  $\mu = np$  and  $\sigma = \sqrt{npq}$ . In particular, for  $x$  = a possible value of  $X$ ,

$$P(X \leq x) = B(x, n, p) \approx (\text{area under the normal curve to the left of } x + 0.5)$$

$$= \Phi \left( \frac{x + 0.5 - np}{\sqrt{npq}} \right)$$

$$\begin{aligned} P(a \leq X \leq B) &= P \left( \frac{a - 0.5 - np}{\sqrt{npq}} \leq Z \leq \frac{b + 0.5 - np}{\sqrt{npq}} \right) \\ &= \Phi \left( \frac{b + 0.5 - np}{\sqrt{npq}} \right) - \Phi \left( \frac{a - 0.5 - np}{\sqrt{npq}} \right). \end{aligned}$$

- **Approximating the binomial distribution conditions:** In practice, the approximation is adequate provided that both  $np \geq 10$  and  $nq \geq 10$ , since there is then enough symmetry in the underlying binomial distribution.

## 5.5 Chapter 5: Joint Probability Distributions and Random Samples

### 5.5.1 Definitions and Theroems

- **Two Discrete Random Variables:** **Joint probability mass function:** The probability mass function (pmf) of a single discrete rv  $X$  specifies how much probability mass is placed on each possible  $X$  value. The joint pmf of two discrete rv's  $X$  and  $Y$  describes how much probability mass is placed on each possible pair of values  $(x, y)$ .

Let  $X$  and  $Y$  be two discrete rv's defined on the sample space  $\mathcal{S}$  of an experiment. The joint probability mass function  $p(x, y)$  is defined for each pair of numbers  $(x, y)$  by

$$p(x, y) = P(X = x \text{ and } Y = y)$$

It must be the case that  $p(x, y) \geq 0$  and

$$\sum_x \sum_y p(x, y) = 1.$$

Now let  $A$  be any set consisting of pairs of  $(x, y)$  values (e.g.,  $A = \{(x, y) : x + y = 5\}$  or  $\{(x, y) : \max(x, y) \leq 3\}$ ). Then the probability  $P((X, Y) \in A)$  is obtained by summing the joint pmf over pairs in  $A$ :

$$P((X, Y) \in A) = \sum_{(x, y) \in A} p(x, y).$$

- **Marginal probability mass function:** The *marginal probability mass function* of  $X$ , denoted by  $p_X(x)$ , is given by

$$p_X(x) = \sum_{y: p(x, y) > 0} p(x, y) \quad \text{for each possible value } x$$

Similarly, the *marginal probability mass function* of  $Y$  is

$$p_Y(y) = \sum_{x: p(x, y) > 0} p(x, y) \quad \text{for each possible value } y.$$

- **Example: Joint pmf and Marginal pmf:** A large insurance agency services a number of customers who have purchased both a homeowner's policy and an automobile policy from the agency. For each type of policy, a deductible amount must be specified. For an automobile policy, the choices are 100 and 250, whereas for a homeowner's policy, the choices are \$0, \$100, and \$200. Suppose an individual with both types of policy is selected at random from the agency's files. Let  $X$  = the deductible amount on the auto policy and  $Y$  = the deductible amount on the homeowner's policy. Possible  $(X, Y)$  pairs are then  $(100, 0)$ ,  $(100, 100)$ ,  $(100, 200)$ ,  $(250, 0)$ ,  $(250, 100)$ , and  $(250, 200)$ ; the joint pmf specifies the probability associated with each one of these pairs, with any other pair having probability zero. Suppose the joint pmf is given in the accompanying *joint probability table*:

$p(x, y)$	$y = 0$	$y = 100$	$y = 200$
$x = 100$	0.20	0.10	0.20
$x = 250$	0.05	0.15	0.30

Then  $p(100, 100) = P(X = 100 \text{ and } Y = 100) = P(\$100 \text{ deductible on both policies}) = .10$ . The probability  $P(Y \geq 100)$  is computed by summing probabilities of all  $(x, y)$  pairs for which  $y \geq 100$ :

$$P(Y \geq 100) = p(100, 100) + p(250, 100) + p(100, 200) + p(250, 200) = .75$$

Once the joint pmf of the two variables  $X$  and  $Y$  is available, it is in principle straightforward to obtain the distribution of just one of these variables. As an example, let  $X$  and  $Y$  be the number of statistics and mathematics courses, respectively, currently being taken by a randomly selected statistics major. Suppose that we wish the distribution of  $X$ , and that when  $X = 2$ , the only possible values of  $Y$  are 0, 1, and 2. Then

$$p_X(2) = P(X = 2) = P(\{(X, Y) = (2, 0) \text{ or } (2, 1) \text{ or } (2, 2)\}) = p(2, 0) + p(2, 1) + p(2, 2)$$

That is, the joint pmf is summed over all pairs of the form  $(2, y)$ . More generally, for any possible value of  $X$ , the probability  $p_X(x)$  results from holding  $x$  fixed and summing the joint pmf  $p(x, y)$  over all  $y$  for which the pair  $(x, y)$  has positive probability mass. The same strategy applies to obtaining the distribution of  $Y$  by itself.

The possible  $X$  values are  $x = 100$  and  $x = 250$ , so computing row totals in the joint probability table yields

$$p_X(100) = p(100, 0) + p(100, 100) + p(100, 200) = .50$$

and

$$p_X(250) = p(250, 0) + p(250, 100) + p(250, 200) = .50$$

The marginal pmf of  $X$  is then

$$p_X(x) = \begin{cases} .5 & x = 100, 250 \\ 0 & \text{otherwise} \end{cases}$$

Similarly, the marginal pmf of  $Y$  is obtained from column totals as

$$p_Y(y) = \begin{cases} .25 & y = 0, 100 \\ .50 & y = 200 \\ 0 & \text{otherwise} \end{cases}$$

so  $P(Y \geq 100) = p_Y(100) + p_Y(200) = .75$  as before.

- **Two Continuous Random Variables: Joint pdf:** Let  $X$  and  $Y$  be continuous rv's. A *joint probability density function*  $f(x, y)$  for these two variables is a function satisfying  $f(x, y) \geq 0$  and

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$$

Then for any two-dimensional set  $A$

$$P((X, Y) \in A) = \iint_A f(x, y) dx dy$$

In particular, if  $A$  is the two-dimensional rectangle  $\{(x, y) : a \leq x \leq b, c \leq y \leq d\}$ , then

$$P((X, Y) \in A) = P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f(x, y) dy dx$$

- **Marginal pdf:** The *marginal probability density functions* of  $X$  and  $Y$ , denoted by  $f_X(x)$  and  $f_Y(y)$ , respectively, are given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy \quad \text{for } -\infty < x < \infty$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx \quad \text{for } -\infty < y < \infty$$

- **Independent Random Variables:** Two random variables  $X$  and  $Y$  are said to be *independent* if for every pair of  $x$  and  $y$  values

$$p(x, y) = p_X(x) \cdot p_Y(y) \quad \text{when } X \text{ and } Y \text{ are discrete}$$

or

$$f(x, y) = f_X(x) \cdot f_Y(y) \quad \text{when } X \text{ and } Y \text{ are continuous}$$

If this is not satisfied for all  $(x, y)$ , then  $X$  and  $Y$  are said to be *dependent*.

The definition says that two variables are independent if their joint pmf or pdf is the product of the two marginal pmf's or pdf's. Intuitively, independence says that knowing the value of one of the variables does not provide additional information about what the value of the other variable might be.

- **More Than Two Random Variables:** If  $X_1, X_2, \dots, X_n$  are all discrete random variables, the joint pmf of the variables is the function

$$p(x_1, x_2, \dots, x_n) = P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$$

If the variables are continuous, the joint pdf of  $X_1, \dots, X_n$  is the function  $f(x_1, x_2, \dots, x_n)$  such that for any  $n$  intervals  $[a_1, b_1], \dots, [a_n, b_n]$ ,

$$P(a_1 \leq X_1 \leq b_1, \dots, a_n \leq X_n \leq b_n) = \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} f(x_1, \dots, x_n) dx_n \cdots dx_1$$

- **Multinomial experiment, distribution:** In a binomial experiment, each trial could result in one of only two possible outcomes. Consider now an experiment consisting of  $n$  independent and identical trials, in which each trial can result in any one of  $r$  possible outcomes. Let  $p_i = P(\text{outcome } i \text{ on any particular trial})$ , and define random variables by  $X_i = \text{number of trials resulting in outcome } i$  ( $i = 1, \dots, r$ ). Such an experiment is called a *multinomial experiment*, and the joint pmf of  $X_1, \dots, X_r$  is called the *multinomial distribution*. By using a counting argument analogous to the one used in deriving the binomial distribution, the joint pmf of  $X_1, \dots, X_r$  can be shown to be

$$p(x_1, \dots, x_r) = \begin{cases} \frac{n!}{(x_1!)(x_2!)\cdots(x_r!)} p_1^{x_1} \cdots p_r^{x_r} & x_i = 0, 1, 2, \dots, \text{ with } x_1 + \cdots + x_r = n \\ 0 & \text{otherwise} \end{cases}$$

The case  $r = 2$  gives the binomial distribution, with  $X_1 = \text{number of successes}$  and  $X_2 = n - X_1 = \text{number of failures}$ .

- **Multinomial Example:** If the allele of each of ten independently obtained pea sections is determined and  $p_1 = P(AA)$ ,  $p_2 = P(Aa)$ ,  $p_3 = P(aa)$ ,  $X_1 = \text{number of AAs}$ ,  $X_2 = \text{number of Aas}$ , and  $X_3 = \text{number of aa's}$ , then the multinomial pmf for these  $X_i$ 's is

$$p(x_1, x_2, x_3) = \frac{10!}{(x_1!)(x_2!)(x_3!)} p_1^{x_1} p_2^{x_2} p_3^{x_3}, \quad x_i = 0, 1, \dots \quad \text{and} \quad x_1 + x_2 + x_3 = 10$$

With  $p_1 = p_3 = .25$ ,  $p_2 = .5$ ,

$$P(X_1 = 2, X_2 = 5, X_3 = 3) = p(2, 5, 3) = \frac{10!}{2!5!3!} (.25)^2 (.5)^5 (.25)^3 = .0769$$

- The random variables  $X_1, X_2, \dots, X_n$  are said to be *independent* if for every subset  $X_{i_1}, X_{i_2}, \dots, X_{i_k}$  of the variables (each pair, each triple, and so on), the joint pmf or pdf of the subset is equal to the product of the marginal pmf's or pdf's.

Thus if the variables are independent with  $n = 4$ , then the joint pmf or pdf of any two variables is the product of the two marginals, and similarly for any three variables and all four variables together. Most importantly, once we are told that  $n$  variables are independent, then the joint pmf or pdf is the product of the  $n$  marginals.

- **Conditional Distributions:** Let  $X$  and  $Y$  be two continuous rv's with joint pdf  $f(x, y)$  and marginal  $X$  pdf  $f_X(x)$ . Then for any  $X$  value  $x$  for which  $f_X(x) > 0$ , the *conditional probability density function of  $Y$  given that  $X = x$*  is

$$f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)} \quad -\infty < y < \infty$$

If  $X$  and  $Y$  are discrete, replacing pdf's by pmf's in this definition gives the *conditional probability mass function of  $Y$  when  $X = x$* .

- **Joint probability distribution: Expected values:** Let  $X$  and  $Y$  be jointly distributed rv's with pmf  $p(x, y)$  or pdf  $f(x, y)$  according to whether the variables are discrete or continuous. Then the expected value of a function  $h(X, Y)$ , denoted by  $E[h(X, Y)]$  or  $\mu_{h(X, Y)}$ , is given by

$$E[h(X, Y)] = \begin{cases} \sum_x \sum_y h(x, y) \cdot p(x, y) & \text{if } X \text{ and } Y \text{ are discrete} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y) \cdot f(x, y) dx dy & \text{if } X \text{ and } Y \text{ are continuous} \end{cases}$$

Generally,

$$\begin{aligned} E(X) &= \sum_x \sum_y x \cdot p(x, y) = \sum_x x \cdot p_X(x) \\ E(Y) &= \sum_x \sum_y y \cdot p(x, y) = \sum_y y \cdot p_Y(y) \\ E(XY) &= \sum_x \sum_y xy \cdot p(x, y). \end{aligned}$$

**Note:** If  $X$  and  $Y$  are independent, then  $p(x, y) = p(x)p(y)$ . Thus,

$$\begin{aligned} E(XY) &= \sum_x \sum_y xy \cdot p(x, y) = \sum_x \sum_y xy \cdot p(x)p(y) \\ &= \sum_x x \cdot p(x) \cdot \sum_y y \cdot p(y) \\ &= E(X) \cdot E(Y). \end{aligned}$$

A similar argument can be shown for continuous rv's  $X$  and  $Y$

This does **not** hold for dependent variables  $X$  and  $Y$ .

- **Covariance:** When two random variables  $X$  and  $Y$  are not independent, it is frequently of interest to assess how strongly they are related to one another

The covariance between two rv's  $X$  and  $Y$  is

$$\begin{aligned}\text{Cov}(X, Y) &= E[(X - \mu_X)(Y - \mu_Y)] \\ &= \begin{cases} \sum_x \sum_y (x - \mu_X)(y - \mu_Y)p(x, y) & \text{if } X \text{ and } Y \text{ are discrete} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y)f(x, y) dx dy & \text{if } X \text{ and } Y \text{ are continuous} \end{cases}\end{aligned}$$

- **What does covariance measure:** Covariance measures the degree to which two random variables,  $X$  and  $Y$ , change together. In this context:
  - **Positive Covariance:** Indicates that  $X$  (major failures) and  $Y$  (minor failures) tend to increase together.
  - **Negative Covariance:** Indicates that  $X$  and  $Y$  tend to change in opposite directions.
  - **Zero Covariance:** Indicates no linear relationship between  $X$  and  $Y$ .

- **Covariance shortcut:**

$$\text{Cov}(X, Y) = E(XY) - \mu_x \mu_y.$$

- **Correlation Coefficient:** The **correlation coefficient** of  $X$  and  $Y$ , denoted by  $\text{Corr}(X, Y)$ ,  $\rho_{X,Y}$ , or just  $\rho$ , is defined by

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

Ranges from  $-1 \leq \rho \leq 1$

- **What does  $\rho$  measure:**

- **Range:**  $-1 \leq \rho \leq 1$ 
  - \*  $\rho = 1$ : Perfect positive linear relationship.
  - \*  $\rho = -1$ : Perfect negative linear relationship.
  - \*  $\rho = 0$ : No linear relationship.
- **Sign:** Indicates the direction of the relationship.
  - \*  $\rho > 0$ : Positive correlation (as one variable increases, the other increases).
  - \*  $\rho < 0$ : Negative correlation (as one variable increases, the other decreases).
- **Magnitude:** Indicates the strength of the linear relationship.
  - \* 0.0 to  $\pm 0.2$ : Very weak to no correlation.
  - \*  $\pm 0.2$  to  $\pm 0.4$ : Weak correlation.
  - \*  $\pm 0.4$  to  $\pm 0.6$ : Moderate correlation.
  - \*  $\pm 0.6$  to  $\pm 0.8$ : Strong correlation.
  - \*  $\pm 0.8$  to  $\pm 1.0$ : Very strong correlation.

**Note:** Correlation does not imply causation.

- **Cov vs Corr:**

- **Covariance ( $\text{Cov}(X, Y)$ ):**
  - \* Measures the degree to which two random variables  $X$  and  $Y$  change together.
  - \* Formula:  $\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$
  - \* Interpretation:

- $\text{Cov}(X, Y) > 0$ :  $X$  and  $Y$  tend to increase together.
- $\text{Cov}(X, Y) < 0$ :  $X$  and  $Y$  tend to move in opposite directions.
- $\text{Cov}(X, Y) = 0$ :  $X$  and  $Y$  are linearly independent.

\* Units: Covariance has units that are the product of the units of  $X$  and  $Y$ .

– **Correlation ( $\rho$  or  $r$ ):**

\* Standardized measure of the linear relationship between two random variables.

\* Formula:  $\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$

\* Interpretation:

- $\rho = 1$ : Perfect positive linear relationship.
- $\rho = -1$ : Perfect negative linear relationship.
- $\rho = 0$ : No linear relationship.

\* Range:  $-1 \leq \rho \leq 1$

\* Units: Correlation is dimensionless (no units).

**Key Differences:**

- **Scale:** Covariance depends on the units of the variables, while correlation is dimensionless.
- **Interpretation:** Correlation provides a normalized measure of the strength and direction of a linear relationship, making it easier to interpret and compare across different datasets.

• **Correlation Properties:**

1. If  $a$  and  $c$  are either both positive or both negative,

$$\text{Corr}(aX + b, cY + d) = \text{Corr}(X, Y)$$

2. For any two random variables  $X$  and  $Y$ ,

$$-1 \leq \text{Corr}(X, Y) \leq 1$$

3. If  $X$  and  $Y$  are independent, then  $\rho = 0$ , but  $\rho = 0$  does not imply independence.
4.  $\rho = 1$  or  $-1$  iff  $Y = aX + b$  for some numbers  $a$  and  $b$  with  $a \neq 0$ .

• **uncorrelated:**

- **statistic:** is any quantity whose value can be calculated from sample data. Prior to obtaining data, there is uncertainty as to what value of any particular statistic will result. Therefore, a statistic is a random variable and will be denoted by an uppercase letter; a lowercase letter is used to represent the calculated or observed value of the statistic.
- **Sampling Distribution:** The probability distribution of a statistic is sometimes referred to as its **sampling distribution** to emphasize that it describes how the statistic varies in value across all samples that might be selected.
- **random sample:** of size  $n$  if:
  1. The  $X_i$ 's are independent rv's.
  2. Every  $X_i$  has the same probability distribution.

**Note:** Conditions 1 and 2 can be paraphrased by saying that the  $X_i$ 's are *independent and identically distributed (iid)*. If sampling is either with replacement or from an infinite (conceptual) population, Conditions 1 and 2 are satisfied exactly. These conditions will be approximately satisfied if sampling is without replacement, yet the sample size  $n$  is much smaller than the population size  $N$ . In practice, if  $n/N \leq 0.05$  (at most 5% of the population), the conditions are satisfied.

- **Simulation Experiments:** The second method of obtaining information about a statistic's sampling distribution is to perform a simulation experiment. This method is usually used when a derivation via probability rules is too difficult or complicated to be carried out. Such an experiment is virtually always done with the aid of a computer. The following characteristics of an experiment must be specified:

1. The statistic of interest ( $\bar{X}$ ,  $S$ , a particular trimmed mean, etc.)
2. The population distribution (normal with  $\mu = 100$  and  $\sigma = 15$ , uniform with lower limit  $A = 5$  and upper limit  $B = 10$ , etc.)
3. The sample size  $n$  (e.g.,  $n = 10$  or  $n = 50$ )
4. The number of replications  $k$  (number of samples to be obtained)

Then use appropriate software to obtain  $k$  different random samples, each of size  $n$ , from the designated population distribution. For each sample, calculate the value of the statistic and construct a histogram of the  $k$  values. This histogram gives the *approximate* sampling distribution of the statistic. The larger the value of  $k$ , the better the approximation will tend to be (the actual sampling distribution emerges as  $k \rightarrow \infty$ ). In practice,  $k = 500$  or  $1000$  is usually sufficient if the statistic is "fairly simple."

- **The Distribution of the Sample Mean:** Let  $X_1, X_2, \dots, X_n$  be a random sample from a distribution with mean value  $\mu$  and standard deviation  $\sigma$ . Then

1.  $\mathbb{E}(\bar{X}) = \mu_{\bar{X}} = \mu$
2.  $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$  and  $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

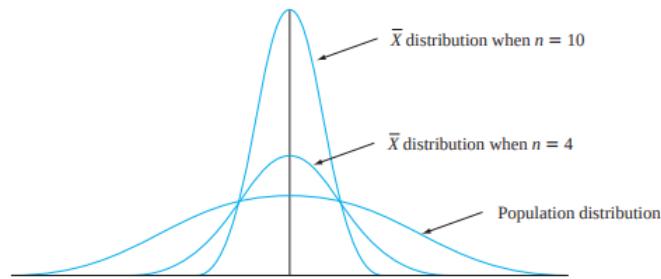
In addition, with  $T_o = X_1 + \dots + X_n$  (the sample total),  $\mathbb{E}(T_o) = n\mu$ ,

$$\text{Var}(T_o) = n\sigma^2, \text{ and } \sigma_{T_o} = \sqrt{n}\sigma.$$

According to Result 1, the sampling (i.e., probability) distribution of  $\bar{X}$  is centered precisely at the mean of the population from which the sample has been selected. Result 2 shows that the  $\bar{X}$  distribution becomes more concentrated about  $\mu$  as the sample size  $n$  increases. In marked contrast, the distribution of  $T_o$  becomes more spread out as  $n$  increases. Averaging moves probability in toward the middle, whereas totaling spreads probability out over a wider and wider range of values. **Note:** The standard deviation  $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$  is often called the *standard error of the mean*; it describes the magnitude of a typical or representative deviation of the sample mean from the population mean.

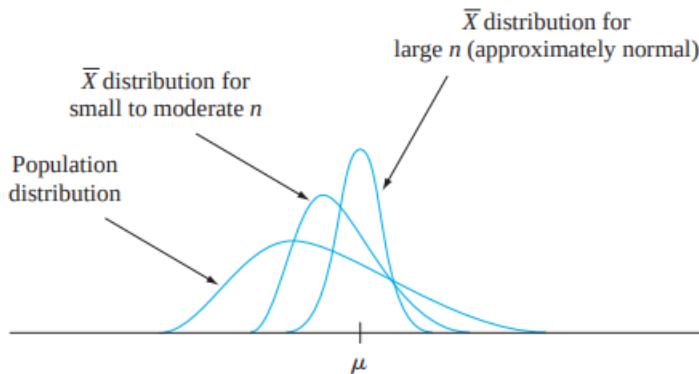
- **The Case of a Normal Population Distribution:** Let  $X_1, X_2, \dots, X_n$  be a random sample from a *normal* distribution with mean  $\mu$  and standard deviation  $\sigma$ . Then for any  $n$ ,  $\bar{X}$  is normally distributed (with mean  $\mu$  and standard deviation  $\sigma/\sqrt{n}$ ), as is  $T_o$  (with mean  $n\mu$  and standard deviation  $\sqrt{n}\sigma$ ).

We know everything there is to know about the  $\bar{X}$  and  $T_o$  distributions when the population distribution is normal. In particular, probabilities such as  $P(a \leq \bar{X} \leq b)$  and  $P(c \leq T_o \leq d)$  can be obtained simply by standardizing.



- **The Central Limit Theorem:** Let  $X_1, X_2, \dots, X_n$  be a random sample from a distribution with mean  $\mu$  and variance  $\sigma^2$ . Then if  $n$  is sufficiently large,  $\bar{X}$  has approximately a normal distribution with  $\mu_{\bar{X}} = \mu$  and  $\sigma_{\bar{X}}^2 = \sigma^2/n$ , and  $T_o$  also has approximately a normal distribution with  $\mu_{T_o} = n\mu$ ,  $\sigma_{T_o}^2 = n\sigma^2$ . The larger the value of  $n$ , the better the approximation.

According to the CLT, when  $n$  is large and we wish to calculate a probability such as  $P(a \leq \bar{X} \leq b)$ , we need only “pretend” that  $\bar{X}$  is normal, standardize it, and use the normal table. The resulting answer will be approximately correct. The exact answer could be obtained only by first finding the distribution of  $\bar{X}$ , so the CLT provides a truly impressive shortcut. The proof of the theorem involves much advanced mathematics.



- **Invoking the CLT: Condition:**
  - For most situations,  $n > 30$  is a good rule of thumb.
  - For populations that are approximately normal, smaller sample sizes may suffice.
  - For highly non-normal populations, larger sample sizes may be required.
- **Linear combination of the  $X_i$ 's:** Given a collection of  $n$  random variables  $X_1, \dots, X_n$  and  $n$  numerical constants  $a_1, \dots, a_n$ , the random variable

$$Y = a_1X_1 + \dots + a_nX_n = \sum_{i=1}^n a_iX_i$$

is called a *linear combination* of the  $X_i$ 's.

- **Linear combination mean and variance:** Notice that we are not requiring the  $X_i$ 's to be independent or identically distributed. All the  $X_i$ 's could have different distributions and therefore different mean values and variances. We first consider the expected value and variance of a linear combination.

Let  $X_1, X_2, \dots, X_n$  have mean values  $\mu_1, \mu_2, \dots, \mu_n$  respectively, and variances  $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$  respectively.

1. Whether or not the  $X_i$ 's are independent,

$$\begin{aligned} E(a_1X_1 + a_2X_2 + \dots + a_nX_n) &= a_1E(X_1) + a_2E(X_2) + \dots + a_nE(X_n) \\ &= a_1\mu_1 + \dots + a_n\mu_n \end{aligned} \tag{58}$$

2. If  $X_1, \dots, X_n$  are independent,

$$\begin{aligned} V(a_1X_1 + a_2X_2 + \dots + a_nX_n) &= a_1^2V(X_1) + a_2^2V(X_2) + \dots + a_n^2V(X_n) \\ &= a_1^2\sigma_1^2 + \dots + a_n^2\sigma_n^2 \end{aligned} \quad (59)$$

and

$$\sigma_{a_1X_1+\dots+a_nX_n} = \sqrt{a_1^2\sigma_1^2 + \dots + a_n^2\sigma_n^2} \quad (60)$$

3. For any  $X_1, \dots, X_n$ ,

$$V(a_1X_1 + \dots + a_nX_n) = \sum_{i=1}^n \sum_{j=1}^n a_i a_j \text{Cov}(X_i, X_j) \quad (61)$$

- **The Difference Between Two Random Variables:** An important special case of a linear combination results from taking  $n = 2$ ,  $a_1 = 1$ , and  $a_2 = -1$ :

$$Y = a_1X_1 + a_2X_2 = X_1 - X_2 \quad (62)$$

We then have the following corollary to the proposition.

$$E(X_1 - X_2) = E(X_1) - E(X_2) \quad \text{for any two rv's } X_1 \text{ and } X_2 \quad (63)$$

$$V(X_1 - X_2) = V(X_1) + V(X_2) \quad \text{if } X_1 \text{ and } X_2 \text{ are independent rv's} \quad (64)$$

- **The Case of Normal Random Variables:** If  $X_1, X_2, \dots, X_n$  are independent, normally distributed rv's (with possibly different means and/or variances), then any linear combination of the  $X_i$ 's also has a normal distribution. In particular, the difference  $X_1 - X_2$  between two independent, normally distributed variables is itself normally distributed.
- **Using the normal distribution to approximate the poisson distribution:** For large  $\mu$  ( $\mu \geq 30$ ), the poission distribution  $p(x; \mu)$  can be approximated by  $X \sim N(\mu, \sqrt{\mu})$

## 5.6 Chapter 7: Statistical Intervals Based on a Single Sample

### 5.6.1 Definitions and Theorems

- **95% Confidence Interval for  $\mu$ :** If, after observing  $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$ , we compute the observed sample mean  $\bar{x}$  and then substitute  $\bar{x}$  into (7.4) in place of  $\bar{X}$ , the resulting fixed interval is called a 95% confidence interval for  $\mu$ . This CI can be expressed either as

$$\left( \bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}} \right) \text{ is a 95% CI for } \mu$$

or as

$$\bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}} \text{ with 95% confidence}$$

A concise expression for the interval is  $\bar{x} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}}$ , where  $-$  gives the left endpoint (lower limit) and  $+$  gives the right endpoint (upper limit).

- **Interpreting a Confidence Level:** A 95% confidence interval for a population parameter (such as the mean,  $\mu$ ) is an interval estimate calculated from the sample data, within which we expect the true population parameter to lie 95% of the time. In other words, if we were to take many random samples from the population and compute a 95% confidence interval for each sample, about 95% of those intervals would contain the true population parameter. It provides a range of plausible values for the parameter, reflecting the uncertainty due to sampling variability.

A common incorrect way to interpret a 95% confidence interval is to say that there is a 95% probability that the true population parameter lies within the specific interval computed from a single sample. This interpretation is incorrect because the true population parameter is fixed and either lies within the interval or it does not; the 95% confidence level refers to the long-run frequency of the intervals containing the parameter if we repeated the sampling process many times.

- **Any confidence interval:** A  $100(1 - \alpha)\%$  confidence interval for the mean  $\mu$  of a normal population when the value of  $\sigma$  is known is given by

$$\left( \bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \right)$$

or, equivalently, by  $\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ .

- **error rate:** and  $100(1 - \alpha)\%$  is called the confidence coefficient
- **Confidence Level, Precision:** Why settle for a confidence level of 95% when a level of 99% is achievable? Because the price paid for the higher confidence level is a wider interval. Since the 95% interval extends  $1.96 \cdot \frac{\sigma}{\sqrt{n}}$  to each side of  $\bar{x}$ , the width of the interval is  $2(1.96) \cdot \frac{\sigma}{\sqrt{n}} = 3.92 \cdot \frac{\sigma}{\sqrt{n}}$ . Similarly, the width of the 99% interval is  $2(2.58) \cdot \frac{\sigma}{\sqrt{n}} = 5.16 \cdot \frac{\sigma}{\sqrt{n}}$ . That is, we have more confidence in the 99% interval precisely because it is wider. The higher the desired degree of confidence, the wider the resulting interval will be.

Thus it cannot be said unequivocally that a 99% interval is to be preferred to a 95% interval; the gain in reliability entails a loss in precision

An appealing strategy is to specify both the desired confidence level and interval width and then determine the necessary sample size.

- **Determine sample size needed for confidence interval with width  $w$ :** The sample size necessary for the CI to have a width  $w$  is

$$n = \left\lceil \left( 2z_{\alpha/2} \cdot \frac{\sigma}{w} \right)^2 \right\rceil.$$

The smaller the desired width  $w$ , the larger  $n$  must be. In addition,  $n$  is an increasing function of  $\sigma$  (more population variability necessitates a larger sample size) and of the confidence level  $100(1 - \alpha)$  (as  $\alpha$  decreases,  $z_{\alpha/2}$  increases).

The half-width  $1.96\sigma/\sqrt{n}$  of the 95% CI is sometimes called the *bound on the error of estimation* associated with a 95% confidence level. That is, with 95% confidence, the point estimate  $\bar{x}$  will be no farther than this from  $\mu$ . Before obtaining data, an investigator may wish to determine a sample size for which a particular value of the bound is achieved. For example, with  $\mu$  representing the average fuel efficiency (mpg) for all cars of a certain type, the objective of an investigation may be to estimate  $\mu$  to within 1 mpg with 95% confidence. More generally, if we wish to estimate  $\mu$  to within an amount  $B$  (the specified bound on the error of estimation) with  $100(1 - \alpha)\%$  confidence, the necessary sample size results from replacing  $2w$  by  $1/B$  in the formula in the preceding box.

Alternately, with  $E = \frac{w}{2}$

$$n = \left\lceil \left( \frac{Z_{\alpha/2} \cdot \sigma}{E} \right)^2 \right\rceil$$

- **Large sample CI for  $\mu$ :** If  $n$  is sufficiently large, the standardized variable

$$Z = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has approximately a standard normal distribution. This implies that

$$\bar{X} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

is a **large-sample confidence interval** for  $\mu$  with confidence level approximately  $100(1 - \alpha)\%$ . This formula is valid regardless of the shape of the population distribution.

The qualifier *approximately* is very important here.

- **Large sample CI sample size:** Unfortunately, the choice of sample size to yield a desired interval width is not as straightforward here as it was for the case of known  $\sigma$ . This is because the width of (7.8) is  $2z_{\alpha/2}s/\sqrt{n}$ . Since the value of  $s$  is not available before the data has been gathered, the width of the interval cannot be determined solely by the choice of  $n$ . The only option for an investigator who wishes to specify a desired width is to make an educated guess as to what the value of  $s$  might be. By being conservative and guessing a larger value of  $s$ , an  $n$  larger than necessary will be chosen. The investigator may be able to specify a reasonably accurate value of the population range (the difference between the largest and smallest values). Then if the population distribution is not too skewed, dividing the range by 4 gives a ballpark value of what  $s$  might be.

**Example:** The charge-to-tap time (min) for carbon steel in one type of open hearth furnace is to be determined for each heat in a sample of size  $n$ . If the investigator believes that almost all times in the distribution are between 320 and 440, what sample size would be appropriate for estimating the true average time to within 5 min. with a confidence level of 95%?

A reasonable value for  $s$  is  $(440 - 320)/4 = 30$ . Thus

$$n = \left[ \frac{(1.96)(30)}{5} \right]^2 = 138.3$$

Since the sample size must be an integer,  $n = 139$  should be used. Note that estimating to within 5 min. with the specified confidence level is equivalent to a CI width of 10 min.

- **Sample proportion  $\hat{p}$ : Definition:** The sample proportion  $\hat{p}$  (read as "p-hat") is the fraction of individuals in a sample from the population that have the characteristic of interest. For example, if we survey 100 people in the city and find that 60 of them own a car, the sample proportion  $\hat{p}$  is 0.60 or 60%.

**Formula:** The sample proportion  $\hat{p}$  is calculated as:

$$\hat{p} = \frac{x}{n}$$

where  $x$  is the number of individuals in the sample with the characteristic of interest, and  $n$  is the total number of individuals in the sample.

**Usage:** The sample proportion is a statistic, which is a value calculated from the sample data. It serves as an estimate of the population proportion  $p$ . The reliability of this estimate depends on the size of the sample and how well the sample represents the population.

- **$\mu_{\hat{p}}$  and  $\sigma_{\hat{p}}$ :**

$$\begin{aligned}\mu_{\hat{p}} &= p \\ \sigma_{\hat{p}} &= \sqrt{\frac{p(1-p)}{n}}.\end{aligned}$$

If we don't have  $p$ , we use  $\hat{p}$  as an unbiased estimator of  $p$ . Thus,

$$\begin{aligned}\mu_{\hat{p}} &= \hat{p} \\ \sigma_{\hat{p}} &= \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.\end{aligned}$$

- **A General Large-Sample Confidence Interval:** The large-sample intervals  $\bar{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}$  and  $\bar{x} \pm z_{\alpha/2} \cdot s/\sqrt{n}$  are special cases of a general large-sample CI for a parameter  $\theta$ . Suppose that  $\hat{\theta}$  is an estimator satisfying the following properties: (1) It has approximately a normal distribution; (2) it is (at least approximately) unbiased; and (3) an expression for  $\sigma_{\hat{\theta}}$ , the standard deviation of  $\hat{\theta}$ , is available. For example, in the case  $\theta = \mu$ ,  $\hat{\theta} = \bar{X}$  is an unbiased estimator whose distribution is approximately normal when  $n$  is large and  $\sigma_{\bar{X}} = \sigma/\sqrt{n}$ . Standardizing  $\hat{\theta}$  yields the rv  $Z = (\hat{\theta} - \theta)/\sigma_{\hat{\theta}}$ , which has approximately a standard normal distribution. This justifies the probability statement

$$P \left( -z_{\alpha/2} < \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} < z_{\alpha/2} \right) \approx 1 - \alpha \quad (7.9)$$

Suppose first that  $\sigma_{\hat{\theta}}$  does not involve any unknown parameters (e.g., known  $\sigma$  in the case  $\theta = \mu$ ). Then replacing each  $<$  in (7.9) by  $=$  results in  $\theta = \hat{\theta} \pm z_{\alpha/2} \cdot \sigma_{\hat{\theta}}$ , so the lower and upper confidence limits are  $\hat{\theta} - z_{\alpha/2} \cdot \sigma_{\hat{\theta}}$  and  $\hat{\theta} + z_{\alpha/2} \cdot \sigma_{\hat{\theta}}$ , respectively. Now suppose that  $\sigma_{\hat{\theta}}$  does not involve  $\theta$  but does involve at least one other unknown parameter. Let  $\hat{\sigma}_{\hat{\theta}}$  be the estimate of  $\sigma_{\hat{\theta}}$  obtained by using estimates in place of the unknown parameters (e.g.,  $s/\sqrt{n}$  estimates  $\sigma/\sqrt{n}$ ). Under general conditions (essentially that  $\hat{\sigma}_{\hat{\theta}}$  be close to  $\sigma_{\hat{\theta}}$  for most samples), a valid CI is  $\hat{\theta} \pm z_{\alpha/2} \cdot \hat{\sigma}_{\hat{\theta}}$ . The large-sample interval  $\bar{x} \pm z_{\alpha/2} \cdot s/\sqrt{n}$  is an example.

Finally, suppose that  $\sigma_{\hat{\theta}}$  does involve the unknown  $\theta$ . This is the case, for example, when  $\theta = p$ , a population proportion. Then  $(\hat{p} - p)/\sigma_{\hat{p}} = z_{\alpha/2}$  can be difficult to solve. An approximate solution can often be obtained by replacing  $\theta$  in  $\sigma_{\hat{\theta}}$  by its estimate  $\hat{\theta}$ . This results in an estimated standard deviation  $\hat{\sigma}_{\hat{p}}$ , and the corresponding interval is again  $\hat{p} \pm z_{\alpha/2} \cdot \hat{\sigma}_{\hat{p}}$ .

In words, this CI is a

point estimate of  $\theta \pm (z \text{ critical value})(\text{estimated standard error of the estimator})$

- **A Confidence Interval for a Population Proportion:** Let  $\tilde{p} = \frac{\hat{p} + \frac{z_{\alpha/2}^2}{2n}}{1 + \frac{z_{\alpha/2}^2}{n}}$ . Then a **confidence interval for a population proportion  $p$**  with confidence level approximately  $100(1 - \alpha)\%$  is

$$\tilde{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}/n + \frac{z_{\alpha/2}^2}{4n^2}}{1 + \frac{z_{\alpha/2}^2}{n}}}$$

where  $\hat{q} = 1 - \hat{p}$  and, as before, the  $-$  in (7.10) corresponds to the lower confidence limit and the  $+$  to the upper confidence limit.

This is often referred to as the **score CI for  $p$** .

For  $n$  large, the expression becomes approximately

$$\hat{p} \pm Z_{\alpha/2} \sqrt{\hat{p} \cdot \frac{\hat{q}}{n}}.$$

- **PP CI Approximation conditions**
  1.  $np \geq 10$
  2.  $n(1 - p) \geq 10$
- **Sample size  $n$  for PP CI:** The sample size needed for a population proportion confidence interval with width  $w$  is given by

$$n = \frac{2z^2 \hat{p}\hat{q} - z^2 w^2 \pm \sqrt{4z^4 \hat{p}\hat{q}(\hat{p}\hat{q} - w^2) + w^2 z^4}}{w^2}.$$

Neglecting the terms in the numerator involving  $w^2$  gives the approximate solution

$$n \approx \frac{4z^2 \hat{p}\hat{q}}{w^2}.$$

If  $\hat{p}$  is unknown using  $\hat{p} = 0.5$  gives the largest possible value of  $n$

**Note:** The term  $z$  is  $Z_{\frac{\alpha}{2}}$

- **Confidence bounds:** A large-sample upper confidence bound for  $\mu$  is

$$\mu < \bar{x} + z_\alpha \cdot \frac{s}{\sqrt{n}}$$

and a large-sample lower confidence bound for  $\mu$  is

$$\mu > \bar{x} - z_\alpha \cdot \frac{s}{\sqrt{n}}$$

A one-sided confidence bound for  $p$  results from replacing  $z_{\alpha/2}$  by  $z_\alpha$  and  $\pm$  by either  $+$  or  $-$  in the CI formula (7.10) for  $p$ . In all cases the confidence level is approximately  $100(1 - \alpha)\%$ .

- ***t* distribution assumption:**

**Assumption:** The population of interest is normal, so that  $X_1, \dots, X_n$  constitutes a random sample from a normal distribution with both  $\mu$  and  $\sigma$  unknown.

The key result underlying the interval in Section 7.2 was that for large  $n$ , the rv  $Z = (\bar{X} - \mu)/(S/\sqrt{n})$  has approximately a standard normal distribution. When  $n$  is small,  $S$  is no longer likely to be close to  $\sigma$ , so the variability in the distribution of  $Z$  arises from randomness in both the numerator and the denominator. This implies that the probability distribution of  $(\bar{X} - \mu)/(S/\sqrt{n})$  will be more spread out than the standard normal distribution. The result on which inferences are based introduces a new family of probability distributions called *t* distributions.

- ***t* distribution:** When  $\bar{X}$  is the mean of a random sample of size  $n$  from a normal distribution with mean  $\mu$ , the rv

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \tag{7.13}$$

has a probability distribution called a *t* distribution with  $n - 1$  degrees of freedom (df).

- ***t* distribution degrees of freedom:** Degrees of freedom (df) in a *t*-distribution refer to the number of independent values or quantities that can vary in an analysis without violating any given constraints. In the context of a *t*-distribution, the degrees of freedom are typically related to the sample size

When you estimate the population mean using a sample mean, you often don't know the population standard deviation and instead use the sample standard deviation. The number of degrees of freedom is usually the sample size minus one ( $n - 1$ ). This adjustment is necessary because one parameter (the sample mean) is used to estimate the population mean.

For a sample of size  $n$ , the degrees of freedom for the *t*-distribution would be  $n - 1$

- **Properties of *t* distributions:** Let  $t_\nu$  denote the *t* distribution with  $\nu$  df.

1. Each  $t_\nu$  curve is bell-shaped and centered at 0.
2. Each  $t_\nu$  curve is more spread out than the standard normal (z) curve.
3. As  $\nu$  increases, the spread of the corresponding  $t_\nu$  curve decreases.
4. As  $\nu \rightarrow \infty$ , the sequence of  $t_\nu$  curves approaches the standard normal curve (so the z curve is often called the *t* curve with  $df = \infty$ ).



**Figure 7.7**  $t_\nu$  and z curves

**Note:** The number of df for  $T$  in (7.13) is  $n - 1$  because, although  $S$  is based on the  $n$  deviations  $X_1 - \bar{X}, \dots, X_n - \bar{X}$ ,  $\sum(X_i - \bar{X}) = 0$  implies that only  $n - 1$  of these are “freely determined.” The number of df for a  $t$  variable is the number of freely determined deviations on which the estimated standard deviation in the denominator of  $T$  is based.

- **$t$  critical values:-** Let  $t_{\alpha,\nu}$  = the number on the measurement axis for which the area under the  $t$  curve with  $\nu$  df to the right of  $t_{\alpha,\nu}$  is  $\alpha$ ;  $t_{\alpha,\nu}$  is called a *t critical value*.
- **$t$  CI for  $\mu$ :** Let  $\bar{x}$  and  $s$  be the sample mean and sample standard deviation computed from the results of a random sample from a normal population with mean  $\mu$ . Then a  $100(1 - \alpha)\%$  confidence interval for  $\mu$  is

$$\left( \bar{x} - t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}} \right) \quad (7.15)$$

or, more compactly,  $\bar{x} \pm t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}$ .

An *upper confidence bound* for  $\mu$  is

$$\bar{x} + t_{\alpha,n-1} \cdot \frac{s}{\sqrt{n}}$$

and replacing + by - in this latter expression gives a *lower confidence bound* for  $\mu$ , both with confidence level  $100(1 - \alpha)\%$ .

- **$t$  Prediction interval:** A *prediction interval* (PI) for a single observation to be selected from a normal population distribution is

$$\bar{x} \pm t_{\alpha/2,n-1} \cdot s \sqrt{1 + \frac{1}{n}} \quad (7.16)$$

The *prediction level* is  $100(1 - \alpha)\%$ . A lower prediction bound results from replacing  $t_{\alpha/2}$  by  $t_\alpha$  and discarding the + part of (7.16); a similar modification gives an upper prediction bound.

## 5.7 Chapter 8: Tests of Hypotheses Based on a Single Sample

### 5.7.1 Definitions and Theorems

- **alternative hypothesis:**, denoted by  $H_a$ , is the assertion that is contradictory to  $H_0$ .

The null hypothesis will be rejected in favor of the alternative hypothesis only if sample evidence suggests that  $H_0$  is false. If the sample does not strongly contradict  $H_0$ , we will continue to believe in the plausibility of the null hypothesis.

- **Hypothesis-testing conclusions:** The two possible conclusions from a hypothesis-testing analysis are then *reject  $H_0$*  or *fail to reject:  $H_0$* .
- **Test of hypotheses:** A test of hypotheses is a method for using sample data to decide whether the null hypothesis should be rejected
- **Assertions, null value:** In our treatment of hypothesis testing,  $H_0$  will generally be stated as an equality claim. If  $\theta$  denotes the parameter of interest, the null hypothesis will have the form  $H_0 : \theta = \theta_0$ , where  $\theta_0$  is a specified number called the *null value* of the parameter (value claimed for  $\theta$  by the null hypothesis).

The alternative to the null hypothesis  $H_0 : \theta = \theta_0$  will look like one of the following three assertions:

1.  $H_a : \theta > \theta_0$  (in which case the implicit null hypothesis is  $\theta \leq \theta_0$ ),
2.  $H_a : \theta < \theta_0$  (in which case the implicit null hypothesis is  $\theta \geq \theta_0$ ), or
3.  $H_a : \theta \neq \theta_0$

For example, let  $\sigma$  denote the standard deviation of the distribution of inside diameters (inches) for a certain type of metal sleeve. If the decision was made to use the sleeve unless sample evidence conclusively demonstrated that  $\sigma > .001$ , the appropriate hypotheses would be  $H_0 : \sigma = .001$  versus  $H_a : \sigma > .001$ . The number  $\theta_0$  that appears in both  $H_0$  and  $H_a$  (separates the alternative from the null) is called the **null value**.

- **Test Procedures:** A test procedure is a rule, based on sample data, for deciding whether to reject  $H_0$ :

A test procedure is specified by the following:

1. A test statistic, a function of the sample data on which the decision (reject  $H_0$  or do not reject  $H_0$ ) is to be based
2. A rejection region, the set of all test statistic values for which  $H_0$  will be rejected

The null hypothesis will then be rejected if and only if the observed or computed test statistic value falls in the rejection region.

**Example:** Suppose a cigarette manufacturer claims that the average nicotine content  $\mu$  of brand B cigarettes is (at most) 1.5 mg. It would be unwise to reject the manufacturer's claim without strong contradictory evidence, so an appropriate problem formulation is to test  $H_0 : \mu = 1.5$  versus  $H_a : \mu > 1.5$ . Consider a decision rule based on analyzing a random sample of 32 cigarettes. Let  $\bar{X}$  denote the sample average nicotine content. If  $H_0$  is true,  $E(\bar{X}) = \mu = 1.5$ , whereas if  $H_0$  is false, we expect  $\bar{X}$  to exceed 1.5. Strong evidence against  $H_0$  is provided by a value  $\bar{x}$  that considerably exceeds 1.5. Thus we might use  $\bar{X}$  as a test statistic along with the rejection region  $\bar{x} \geq 1.6$ .

- **Errors in Hypothesis Testing:**

1. **type I error** consists of rejecting the null hypothesis  $H_0$  : when it is true.
2. **type II error** involves not rejecting  $H_0$  when  $H_0$  : is false.

**Note:** In the best of all possible worlds, test procedures for which neither type of error is possible could be developed. However, this ideal can be achieved only by basing a decision on an examination of the entire population. The difficulty with using a procedure based on sample data is that because of sampling variability, an unrepresentative sample may result, e.g., a value of  $\bar{X}$  that is far from  $\mu$  or a value of  $\hat{p}$  that differs considerably from  $p$ .

- **Error probabilities:** The choice of a particular rejection region cutoff value fixes the probabilities of type I and type II errors. These error probabilities are traditionally denoted by  $\alpha$  and  $\beta$ , respectively. Because  $H_0$  specifies a unique value of the parameter, there is a single value of  $\alpha$ . However, there is a different value of  $\beta$  for each value of the parameter consistent with  $H_a$ .

- **Tailed tests:**

1. **Upper-Tailed Test:** An upper-tailed test is used when the alternative hypothesis states that the parameter of interest is greater than the null hypothesis value. The rejection region is in the upper tail of the sampling distribution.

$$\begin{aligned} \text{Null hypothesis } (H_0) : \theta &\leq \theta_0 \quad \text{or} \quad \theta = \theta_0 \\ \text{Alternative hypothesis } (H_a) : \theta &> \theta_0. \end{aligned}$$

2. **Lower-Tailed Test:** A lower-tailed test is used when the alternative hypothesis states that the parameter of interest is less than the null hypothesis value. The rejection region is in the lower tail of the sampling distribution.

$$\begin{aligned} \text{Null hypothesis } (H_0) : \theta &\geq \theta_0 \quad \text{or} \quad \theta = \theta_0 \\ \text{Alternative hypothesis } (H_a) : \theta &< \theta_0. \end{aligned}$$

3. **Two-Tailed Test:** A two-tailed test is used when the alternative hypothesis states that the parameter of interest is not equal to the null hypothesis value. The rejection regions are in both tails of the sampling distribution.

$$\begin{aligned} \text{Null hypothesis } (H_0) : \theta &= \theta_0 \\ \text{Alternative hypothesis } (H_a) : \theta &\neq \theta_0. \end{aligned}$$

- Suppose an experiment and a sample size are fixed and a test statistic is chosen. Then decreasing the size of the rejection region to obtain a smaller value of  $\alpha$  results in a larger value of  $\beta$  for any particular parameter value consistent with  $H_a$ .

Thus, **Any attempt to decrease  $\alpha$  would in turn lead to an increase of  $\beta$**

This proposition says that once the test statistic and  $n$  are fixed, there is no rejection region that will simultaneously make both  $\alpha$  and all  $\beta$ 's small. A region must be chosen to effect a compromise between  $\alpha$  and  $\beta$ .

- **Significance level, level  $\alpha$  test:** Because of the suggested guidelines for specifying  $H_0$  and  $H_a$ , a type I error is usually more serious than a type II error (this can always be achieved by proper choice of the hypotheses). The approach adhered to by most statistical practitioners is then to specify the largest value of  $\alpha$  that can be tolerated

and find a rejection region having that value of  $\alpha$  rather than anything smaller. This makes  $\beta$  as small as possible subject to the bound on  $\alpha$ . The resulting value of  $\alpha$  is often referred to as the *significance level* of the test. Traditional levels of significance are .10, .05, and .01, though the level in any particular problem will depend on the seriousness of a type I error—the more serious this error, the smaller should be the significance level. The corresponding test procedure is called a **level  $\alpha$  test** (e.g., a level .05 test or a level .01 test). A test with significance level  $\alpha$  is one for which the type I error probability is controlled at the specified level.

**Example:** Again let  $\mu$  denote the true average nicotine content of brand B cigarettes. The objective is to test  $H_0 : \mu = 1.5$  versus  $H_a : \mu > 1.5$  based on a random sample  $X_1, X_2, \dots, X_{32}$  of nicotine content. Suppose the distribution of nicotine content is known to be normal with  $\sigma = .20$ . Then  $\bar{X}$  is normally distributed with mean value  $\mu_{\bar{X}} = \mu$  and standard deviation  $\sigma_{\bar{X}} = .20/\sqrt{32} = .0354$ .

Rather than use  $\bar{X}$  itself as the test statistic, let's standardize  $\bar{X}$ , assuming that  $H_0$  is true.

$$\text{Test statistic: } Z = \frac{\bar{X} - 1.5}{\sigma/\sqrt{n}} = \frac{\bar{X} - 1.5}{.0354}$$

$Z$  expresses the distance between  $\bar{X}$  and its expected value when  $H_0$  is true as some number of standard deviations. For example,  $z = 3$  results from an  $\bar{x}$  that is 3 standard deviations larger than we would have expected it to be were  $H_0$  true.

Rejecting  $H_0$  when  $\bar{x}$  "considerably" exceeds 1.5 is equivalent to rejecting  $H_0$  when  $z$  "considerably" exceeds 0. That is, the form of the rejection region is  $z \geq c$ . Let's now determine  $c$  so that  $\alpha = .05$ . When  $H_0$  is true,  $Z$  has a standard normal distribution. Thus

$$\alpha = P(\text{type I error}) = P(\text{rejecting } H_0 \text{ when } H_0 \text{ is true}) = P(Z \geq c \text{ when } Z \sim N(0, 1))$$

The value  $c$  must capture upper-tail area .05 under the  $z$  curve. Either from Section 4.3 or directly from Appendix Table A.3,  $c = z_{.05} = 1.645$ .

Notice that  $z \geq 1.645$  is equivalent to  $\bar{X} - 1.5 \geq (.0354)(1.645)$ , that is,  $\bar{x} \geq 1.56$ . Then  $\beta$  involves the probability that  $\bar{X} < 1.56$  and can be calculated for any  $\mu$  greater than 1.5.

- **Test about a population mean, Case I: Normal population with  $\sigma$  known:**

Null hypothesis:  $H_0 : \mu = \mu_0$

Test statistic value:  $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$

Alternative Hypothesis	Rejection Region for Level $\alpha$ Test
------------------------	--

$H_a : \mu > \mu_0$

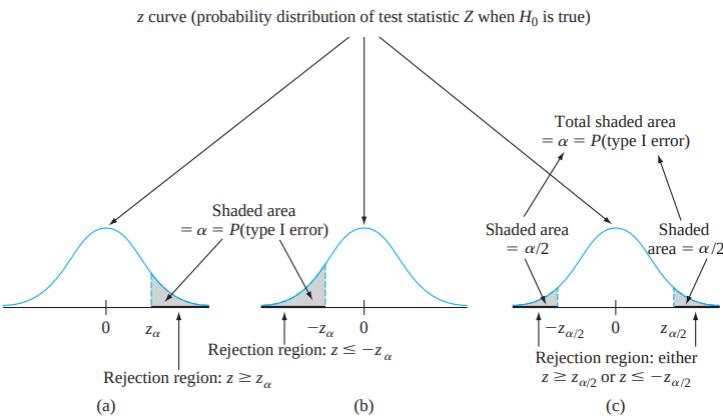
$z \geq z_\alpha$  (upper-tailed test)

$H_a : \mu < \mu_0$

$z \leq -z_\alpha$  (lower-tailed test)

$H_a : \mu \neq \mu_0$

either  $z \geq z_{\alpha/2}$  or  $z \leq -z_{\alpha/2}$  (two-tailed test)



- **Hypothesis test steps:** Use of the following sequence of steps is recommended when testing hypotheses about a parameter.

1. Identify the parameter of interest and describe it in the context of the problem situation.
2. Determine the null value and state the null hypothesis.
3. State the appropriate alternative hypothesis.
4. Give the formula for the computed value of the test statistic (substituting the null value and the known values of any other parameters, but not those of any samplebased quantities).
5. State the rejection region for the selected significance level  $\alpha$ .
6. Compute any necessary sample quantities, substitute into the formula for the test statistic value, and compute that value.
7. Decide whether  $H_0$  should be rejected, and state this conclusion in the problem context.

The formulation of hypotheses (Steps 2 and 3) should be done before examining the data.

**Example:** A manufacturer of sprinkler systems used for fire protection in office buildings claims that the true average system-activation temperature is 130°F. A sample of  $n = 9$  systems, when tested, yields a sample average activation temperature of 131.08°F. If the distribution of activation times is normal with standard deviation 1.5°F, does the data contradict the manufacturer's claim at significance level  $\alpha = .01$ ?

1. Parameter of interest:  $\mu$  = true average activation temperature.
2. Null hypothesis:  $H_0 : \mu = 130$  (null value =  $\mu_0 = 130$ ).
3. Alternative hypothesis:  $H_a : \mu \neq 130$  (a departure from the claimed value in either direction is of concern).
4. Test statistic value:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{x} - 130}{1.5/\sqrt{9}}$$

5. Rejection region: The form of  $H_a$  implies use of a two-tailed test with rejection region either  $z \geq z_{.005}$  or  $z \leq -z_{.005}$ . From Section 4.3 or Appendix Table A.3,  $z_{.005} = 2.58$ , so we reject  $H_0$  if either  $z \geq 2.58$  or  $z \leq -2.58$ .

6. Substituting  $n = 9$  and  $\bar{x} = 131.08$ :

$$z = \frac{131.08 - 130}{1.5/\sqrt{9}} = \frac{1.08}{0.5} = 2.16$$

That is, the observed sample mean is a bit more than 2 standard deviations above what would have been expected were  $H_0$  true.

7. The computed value  $z = 2.16$  does not fall in the rejection region ( $-2.58 < 2.16 < 2.58$ ), so  $H_0$  cannot be rejected at significance level .01. The data does not give strong support to the claim that the true average differs from the design value of 130.

**$\beta$  and Sample Size Determination for case I:** The  $z$  tests for case I are among the few in statistics for which there are simple formulas available for  $\beta$ , the probability of a type II error

**Alternative Hypothesis      Type II Error Probability  $\beta(\mu')$  for a Level  $\alpha$  Test**

$$\begin{aligned} H_a : \mu > \mu_0 & \quad \Phi\left(z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right) \\ H_a : \mu < \mu_0 & \quad 1 - \Phi\left(-z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right) \\ H_a : \mu \neq \mu_0 & \quad \Phi\left(z_{\alpha/2} + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right) - \Phi\left(-z_{\alpha/2} + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right) \end{aligned}$$

where  $\Phi(z)$  is the standard normal cdf.

The sample size  $n$  for which a level  $\alpha$  test also has  $\beta(\mu') = \beta$  at the alternative value  $\mu'$  is

$$n = \begin{cases} \left(\frac{\sigma(z_\alpha + z_\beta)}{\mu_0 - \mu'}\right)^2 & \text{for a one-tailed (upper or lower) test} \\ \left(\frac{\sigma(z_{\alpha/2} + z_\beta)}{\mu_0 - \mu'}\right)^2 & \text{for a two-tailed test (an approximate solution)} \end{cases}$$

- **Test about a population mean, Case II: Large-sample tests:** When the sample size is large, the  $z$  tests for case I are easily modified to yield valid test procedures without requiring either a normal population distribution or known  $\sigma$

When the sample size is large, the  $z$  tests for case I are easily modified to yield valid test procedures without requiring either a normal population distribution or known  $\sigma$ . The key result was used in Chapter 7 to justify large-sample confidence intervals: A large  $n$  implies that the standardized variable

$$Z = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has *approximately* a standard normal distribution. Substitution of the null value  $\mu_0$  in place of  $\mu$  yields the test statistic

$$Z = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

In this case the significance level is approximately (rather than exactly)  $\alpha$

- **Case II condition:** The rule of thumb to characterize a large sample size will be  $n > 40$

- Test about a population mean, Case III: A Normal Population Distribution:

### The One-Sample $t$ -Test

Null hypothesis:  $H_0 : \mu = \mu_0$

Test statistic value:  $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$

Alternative Hypothesis	Rejection Region for a Level $\alpha$ Test
$H_a : \mu > \mu_0$	$t \geq t_{\alpha, n-1}$ (upper-tailed)
$H_a : \mu < \mu_0$	$t \leq -t_{\alpha, n-1}$ (lower-tailed)
$H_a : \mu \neq \mu_0$	either $t \geq t_{\alpha/2, n-1}$ or $t \leq -t_{\alpha/2, n-1}$ (two-tailed)

- Tests about a population proportion: Large sample tests:

Null hypothesis:  $H_0 : p = p_0$

Test statistic value:  $z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$

Alternative Hypothesis	Rejection Region
$H_a : p > p_0$	$z \geq z_\alpha$ (upper-tailed)
$H_a : p < p_0$	$z \leq -z_\alpha$ (lower-tailed)
$H_a : p \neq p_0$	either $z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$ (two-tailed)

- Tests about a population proportion: Large sample test conditions: These test procedures are valid provided that  $np_0 \geq 10$  and  $n(1 - p_0) \geq 10$ .
- Tests about a population proportion: Large sample test  $\beta$  and sample size determination:

### Alternative Hypothesis

$$\begin{aligned} H_a : p > p_0 \quad \beta(p') &= \Phi \left( \frac{p_0 - p' + z_\alpha \sqrt{p_0(1-p_0)/n}}{\sqrt{p'(1-p')/n}} \right) \\ H_a : p < p_0 \quad \beta(p') &= 1 - \Phi \left( \frac{p_0 - p' - z_\alpha \sqrt{p_0(1-p_0)/n}}{\sqrt{p'(1-p')/n}} \right) \\ H_a : p \neq p_0 \quad \beta(p') &= \Phi \left( \frac{p_0 - p' + z_{\alpha/2} \sqrt{p_0(1-p_0)/n}}{\sqrt{p'(1-p')/n}} \right) - \Phi \left( \frac{p_0 - p' - z_{\alpha/2} \sqrt{p_0(1-p_0)/n}}{\sqrt{p'(1-p')/n}} \right) \end{aligned}$$

The sample size  $n$  for which the level  $\alpha$  test also satisfies  $\beta(p') = \beta$  is

$$n = \left( \frac{z_\alpha \sqrt{p_0(1-p_0)} + z_\beta \sqrt{p'(1-p')}}{p' - p_0} \right)^2 \quad \text{one-tailed test}$$

$$n = \left( \frac{z_{\alpha/2} \sqrt{p_0(1-p_0)} + z_{\beta} \sqrt{p'(1-p')}}{p' - p_0} \right)^2 \quad \text{two-tailed test (an approximate solution)}$$

- **Hypothesis testing: P-value method:** The **P-value** is the probability, calculated assuming that the null hypothesis is true, of obtaining a value of the test statistic at least as contradictory to  $H_0$  as the value calculated from the available sample.

This definition is quite a mouthful. Here are some key points:

- The P-value is a probability.
- This probability is calculated assuming that the null hypothesis is true.
- Beware: The P-value is not the probability that  $H_0$  is true, nor is it an error probability!
- To determine the P-value, we must first decide which values of the test statistic are at least as contradictory to  $H_0$  as the value obtained from our sample.

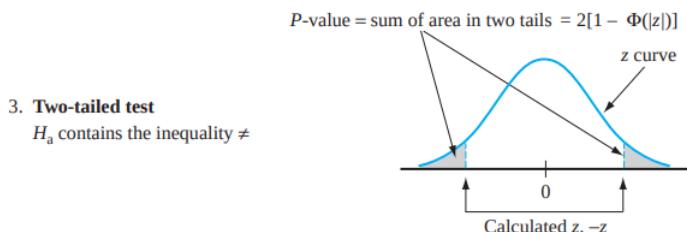
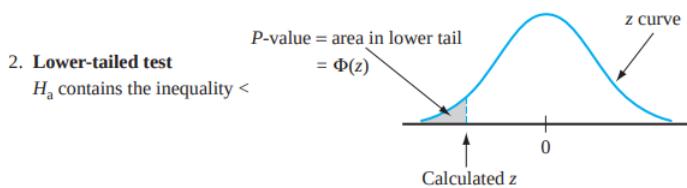
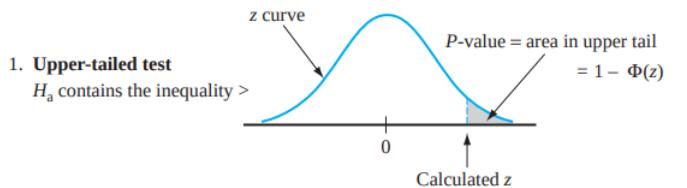
- **Decision rule based on the P-value:** Select a significance level  $\alpha$  (as before, the desired type I error probability). Then

Reject  $H_0$  if P-value  $\leq \alpha$   
Do not reject  $H_0$  if P-value  $> \alpha$ .

- **Observed significance level:** The P-value is the smallest significance level  $\alpha$  at which the null hypothesis can be rejected. Because of this, the P-value is alternatively referred to as the **observed significance level (OSL)** for the data.

- **P-Values for z Tests:**

$$\text{P-value : } P = \begin{cases} 1 - \Phi(z) & \text{for a right-tailed } z \text{ test} \\ \Phi(z) & \text{for a left-tailed } z \text{ test} \\ 2(1 - \Phi(|z|)) & \text{for a two-tailed } z \text{ test} \end{cases}$$



- **P-Values for t Tests:** Just as the P-value for a  $z$  test is a  $z$  curve area, the P-value for a  $t$  test will be a  $t$ -curve area. The following figure illustrates the three different cases. The number of df for the one-sample  $t$  test is  $n - 1$



## 5.8 Chapter 9: Inferences Based on Two Samples

### 5.8.1 Definitions and Theorems

- **Inference based on two samples,  $\mu_1 - \mu_2$ : Conditions (assumptions for CI also):**
  1.  $X_1, X_2, \dots, X_m$  is a random sample from a distribution with mean  $\mu_1$  and variance  $\sigma_1^2$ .
  2.  $Y_1, Y_2, \dots, Y_n$  is a random sample from a distribution with mean  $\mu_2$  and variance  $\sigma_2^2$ .
  3. The  $X$  and  $Y$  samples are independent of one another.
- **Inference based on two samples,  $\mu_1 - \mu_2$ : Expected value and standard deviation:** The expected value of  $\bar{X} - \bar{Y}$  is  $\mu_1 - \mu_2$ , so  $\bar{X} - \bar{Y}$  is an unbiased estimator of  $\mu_1 - \mu_2$ . The standard deviation of  $\bar{X} - \bar{Y}$  is

$$\sigma_{\bar{X} - \bar{Y}} = \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

- **Inference based on two samples,  $\mu_1 - \mu_2$ : Hypothesis Testing: Null hypothesis:  $H_0: \mu_1 - \mu_2 = \Delta_0$**

**Test statistic value:**

$$z = \frac{\bar{X} - \bar{Y} - \Delta_0}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}$$

### Alternative Hypothesis Region for Level $\alpha$ Test

$H_a: \mu_1 - \mu_2 > \Delta_0 \quad z \geq z_\alpha$  (upper-tailed)

$H_a: \mu_1 - \mu_2 < \Delta_0 \quad z \leq -z_\alpha$  (lower-tailed)

$H_a: \mu_1 - \mu_2 \neq \Delta_0 \quad \text{either } z \geq z_{\alpha/2} \text{ or } z \leq -z_{\alpha/2}$  (two-tailed)

Because these are  $z$  tests, a  $P$ -value is computed as it was for the  $z$  tests in Chapter 8 [e.g.,  $P$ -value =  $1 - \Phi(z)$  for an upper-tailed test].

- **Inference based on two samples,  $\mu_1 - \mu_2$ : Large-sample tests:** Use of the test statistic value

$$z = \frac{\bar{X} - \bar{Y} - \Delta_0}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}$$

along with the previously stated upper-, lower-, and two-tailed rejection regions based on  $z$  critical values gives large-sample tests whose significance levels are approximately  $\alpha$ . These tests are usually appropriate if both  $m > 40$  and  $n > 40$ . A  $P$ -value is computed exactly as it was for our earlier  $z$  tests.

- **Confidence intervals for  $\mu_1 - \mu_2$ :** Provided that  $m$  and  $n$  are both large, a CI for  $\mu_1 - \mu_2$  with a confidence level of approximately  $100(1 - \alpha)\%$  is

$$\bar{X} - \bar{Y} \pm z_{\alpha/2} \sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}$$

where  $-$  gives the lower limit and  $+$  the upper limit of the interval. An upper or a lower confidence bound can also be calculated by retaining the appropriate sign (+ or  $-$ ) and replacing  $z_{\alpha/2}$  by  $z_\alpha$ .

- **Two sample t-test and confidence intervals:** Assumptions: Both population distributions are normal, so that  $X_1, X_2, \dots, X_m$  is a random sample from a normal distribution and so is  $Y_1, \dots, Y_n$  (with the  $X$ 's and  $Y$ 's independent of one another). The plausibility of these assumptions can be judged by constructing a normal probability plot of the  $x_i$ 's and another of the  $y_j$ 's.
- **Two sample t-test: Test statistic and df computation:** When the population distributions are both normal, the standardized variable

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}$$

has approximately a  $t$  distribution with df  $\nu$  estimated from the data by

$$\nu = \frac{\left(\frac{s_1^2}{m} + \frac{s_2^2}{n}\right)^2}{\frac{\left(\frac{s_1^2}{m}\right)^2}{m-1} + \frac{\left(\frac{s_2^2}{n}\right)^2}{n-1}} = \frac{[(se_1)^2 + (se_2)^2]^2}{\frac{(se_1)^4}{m-1} + \frac{(se_2)^4}{n-1}}$$

where

$$se_1 = \frac{s_1}{\sqrt{m}}, \quad se_2 = \frac{s_2}{\sqrt{n}}$$

(round  $\nu$  down to the nearest integer).

- **Two sample t confidence interval ( $\mu_1 - \mu_2$ ):** The *two-sample t* confidence interval for  $\mu_1 - \mu_2$  with confidence level  $100(1 - \alpha)\%$  is then

$$\bar{X} - \bar{Y} \pm t_{\alpha/2, \nu} \sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}$$

A one-sided confidence bound can be calculated as described earlier.

- **Two sample t-test:** The *two-sample t test* for testing  $H_0 : \mu_1 - \mu_2 = \Delta_0$  is as follows:

Test statistic value:

$$t = \frac{\bar{X} - \bar{Y} - \Delta_0}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}$$

#### Alternative Hypothesis Region for Approximate Level $\alpha$ Test

$$H_a : \mu_1 - \mu_2 \geq \Delta_0 \quad t \geq t_{\alpha, \nu} \text{ (upper-tailed)}$$

$$H_a : \mu_1 - \mu_2 \leq \Delta_0 \quad t \leq -t_{\alpha, \nu} \text{ (lower-tailed)}$$

$$H_a : \mu_1 - \mu_2 \neq \Delta_0 \quad \text{either } t \geq t_{\alpha/2, \nu} \text{ or } t \leq -t_{\alpha/2, \nu} \text{ (two-tailed)}$$

A  $P$ -value can be computed as described in Section 8.4 for the one-sample  $t$  test.

- **Pooled t Procedures:** Alternatives to the two-sample  $t$  procedures just described result from assuming not only that the two population distributions are normal but also that they have equal variances. That is, the two population distribution curves are assumed normal with equal spreads, the only possible difference between them being where they are centered.

Let  $\sigma^2$  denote the common population variance. Then standardizing  $\bar{X} - \bar{Y}$  gives

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2}{m} + \frac{\sigma^2}{n}}} = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma^2 \left( \frac{1}{m} + \frac{1}{n} \right)}}$$

which has a standard normal distribution. Before this variable can be used as a basis for making inferences about  $\mu_1 - \mu_2$ , the common variance must be estimated from sample data. One estimator of  $\sigma^2$  is  $S_1^2$ , the variance of the  $m$  observations in the first sample, and another is  $S_2^2$ , the variance of the second sample. Intuitively, a better estimator than either individual sample variance results from combining the two sample variances. A first thought might be to use  $(S_1^2 + S_2^2)/2$ . However, if  $m > n$ , then the first sample contains more information about  $\sigma^2$  than does the second sample, and an analogous comment applies if  $m < n$ . The following weighted average of the two sample variances, called the *pooled* (i.e., combined) *estimator of  $\sigma^2$* , adjusts for any difference between the two sample sizes:

$$\begin{aligned} S_p^2 &= \frac{m-1}{m+n-2} \cdot S_1^2 + \frac{n-1}{m+n-2} \cdot S_2^2 \\ \implies S_p &= \sqrt{\frac{(m-1)s_1^2 + (n-1)s_2^2}{m+n-2}}. \end{aligned}$$

The first sample contributes  $m-1$  degrees of freedom to the estimate of  $\sigma^2$ , and the second sample contributes  $n-1$  df, for a total of  $m+n-2$  df. Statistical theory says that if  $S_p^2$  replaces  $\sigma^2$  in the expression for  $Z$ , the resulting standardized variable has a  $t$  distribution based on  $m+n-2$  df. In the same way that earlier standardized variables were used as a basis for deriving confidence intervals and test procedures, this  $t$  variable immediately leads to the pooled  $t$  CI for estimating  $\mu_1 - \mu_2$  and the pooled  $t$  test for testing hypotheses about a difference between means.

Thus, we have a test statistic

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{S_p^2 \left( \frac{1}{m} + \frac{1}{n} \right)}}.$$

And we can carry out the test as usual from here.

**Note:** The pooled standard deviation  $S_p$  can also be applied to the z-test

- **Pooled  $t$  CI:** The  $100(1 - \alpha)\%$  CI for  $\mu_1 - \mu_2$  is

$$(\bar{x} - \bar{y}) \pm t_{\frac{\alpha}{2}, \nu} \sqrt{S_p^2 \left( \frac{1}{m} + \frac{1}{n} \right)}.$$

With  $\nu = m + n - 2$  degrees of freedom

- **The paired  $t$  test (Dependent samples):**

**Null hypothesis:**  $H_0 : \mu_D = \Delta_0$

(where  $D = X - Y$  is the difference between the first and second observations within a pair, and  $\mu_D = \mu_1 - \mu_2$ )

**Test statistic value:**

$$t = \frac{\bar{d} - \Delta_0}{s_D / \sqrt{n}}$$

(where  $\bar{d}$  and  $s_D$  are the sample mean and standard deviation, respectively, of the  $d_i$ 's)

**Alternative Hypothesis**

$$H_a : \mu_D > \Delta_0$$

$$H_a : \mu_D < \Delta_0$$

$$H_a : \mu_D \neq \Delta_0$$

**Rejection Region for Level  $\alpha$  Test**

$$t \geq t_{\alpha, n-1}$$

$$t \leq -t_{\alpha, n-1}$$

$$\text{either } t \geq t_{\alpha/2, n-1} \text{ or } t \leq -t_{\alpha/2, n-1}$$

A  $P$ -value can be calculated as was done for earlier  $t$  tests.

- **Inference about a difference between two population proportions:** Let  $\hat{p}_1 = X/m$  and  $\hat{p}_2 = Y/n$ , where  $X \sim \text{Bin}(m, p_1)$  and  $Y \sim \text{Bin}(n, p_2)$  with  $X$  and  $Y$  independent variables. Then

$$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$$

so  $\hat{p}_1 - \hat{p}_2$  is an unbiased estimator of  $p_1 - p_2$ , and

$$V(\hat{p}_1 - \hat{p}_2) = \frac{p_1 q_1}{m} + \frac{p_2 q_2}{n} \quad (\text{where } q_i = 1 - p_i)$$

- **Large sample two population proportion test:**

**Null hypothesis:**  $H_0 : p_1 - p_2 = 0$

**Test statistic value (large samples):**

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p(1-p) \left( \frac{1}{m} + \frac{1}{n} \right)}}$$

Where  $p = \frac{x_1+x_2}{m+n}$  (Under assumption the null hypothesis is true, both  $p$ 's should be equal, so we use this value as our estimator). Note, we do not use this for confidence intervals.

**Alternative Hypothesis**

$$H_a : p_1 - p_2 > 0$$

$$H_a : p_1 - p_2 < 0$$

$$H_a : p_1 - p_2 \neq 0$$

**Rejection Region for Approximate Level  $\alpha$  Test**

$$z \geq z_\alpha$$

$$z \leq -z_\alpha$$

$$\text{either } z \geq z_{\alpha/2} \text{ or } z \leq -z_{\alpha/2}$$

A  $P$ -value is calculated in the same way as for previous  $z$  tests.

- **Large sample two population proportion test: Conditions:** The test can safely be used as long as  $m\hat{p}_1, m\hat{q}_1, n\hat{p}_2$ , and  $n\hat{q}_2$  are all at least 10.
- **$\hat{p}_1 - \hat{p}_2$  CI :** A CI for  $p_1 - p_2$  with confidence level approximately  $100(1 - \alpha)\%$  is

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{m} + \frac{\hat{p}_2 \hat{q}_2}{n}}$$

This interval can safely be used as long as  $m\hat{p}_1, m\hat{q}_1, n\hat{p}_2$ , and  $n\hat{q}_2$  are all at least 10.

# Linear Algebra

## 6.1 By The Professor

### 6.1.1 Prelude

- **Binomial theorem:** The binomial theorem provides a formula for expanding the power of a binomial expression  $(a + b)^n$  into a sum involving terms of the form  $\binom{n}{k}a^{n-k}b^k$ , where  $\binom{n}{k}$  is the binomial coefficient. The theorem is expressed as:

$$(a + b)^n = \sum_{k=0}^n \binom{n}{k} a^{n-k} b^k$$

This formula allows for the expansion of any binomial raised to a positive integer power.

- **Rational number:** A rational number is one that can be expressed as:

$$\frac{p}{q}, \quad p, q \in \mathbb{Z} \quad q \neq 0.$$

- **Irrational number:** An irrational number is a real number that cannot be expressed as a fraction of two integers, meaning it cannot be written in the form  $\frac{p}{q}$ , where  $p$  and  $q$  are integers and  $q \neq 0$ .

**General Definition:** A number  $x$  is irrational if it satisfies the following condition:

$$x \in \mathbb{R} \setminus \mathbb{Q}$$

This means that  $x$  belongs to the set of real numbers  $\mathbb{R}$  but does not belong to the set of rational numbers  $\mathbb{Q}$ .

**Decimal Representation:** An irrational number has a non-repeating, non-terminating decimal expansion.

- **Recall: Proper subset vs subset:**

**Subset:** A set  $A$  is a subset of a set  $B$  (denoted  $A \subseteq B$ ) if every element of  $A$  is also an element of  $B$ .

**Proper Subset:** A set  $A$  is a proper subset of a set  $B$  (denoted  $A \subset B$ ) if  $A \subseteq B$  and  $A \neq B$ , meaning  $A$  is contained within  $B$  but is not equal to  $B$ .

- **Number sets:** The hierarchy of the number sets is as follows

$$\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R} \subset \mathbb{C}.$$

With

$\mathbb{N}$  : The set of natural numbers (e.g., 1, 2, 3, ...)

$\mathbb{Z}$  : The set of integers (e.g., ..., -2, -1, 0, 1, 2, ...)

$\mathbb{Q}$  : The set of rational numbers (numbers that can be expressed as  $\frac{p}{q}$ ,

where  $p, q \in \mathbb{Z}$  and  $q \neq 0$ )

$\mathbb{R}$  : The set of real numbers (including both rational and irrational numbers)

$\mathbb{C}$  : The set of complex numbers (numbers of the form  $a + bi$ , where  $a, b \in \mathbb{R}$  and  $i^2 = -1$ ).

**Note:** The sets  $\mathbb{N}, \mathbb{Z}, \mathbb{Q}$  are described as *countably infinite*. Whereas  $\mathbb{R}, \mathbb{C}$  are *uncountably infinite*.

Furthermore, The set  $\overline{\mathbb{Q}}$  (the set of irrational numbers) fits within the set of real numbers  $\mathbb{R}$  but outside the set of rational numbers  $\mathbb{Q}$ . This is because  $\mathbb{R}$  is the union of  $\mathbb{Q}$  (the rationals) and  $\overline{\mathbb{Q}}$  (the irrationals).

$\overline{\mathbb{Q}}$  is a subset of  $\mathbb{R}$  but is disjoint from  $\mathbb{Q}$ . Therefore,  $\overline{\mathbb{Q}}$  lies "within"  $\mathbb{R}$ , just like  $\mathbb{Q}$ , but not within any of the earlier sets in the hierarchy.

- **Cardinality of sets:** For some set  $A$  The cardinality (size) of that set uses the following notation

$$|A| = n(A) = \text{the number of elements in the set.}$$

- **Countably infinite sets:** A set is countably infinite if there exists a way to list its elements in a sequence, where each element is matched with exactly one natural number. This means that despite being infinite, the elements can be "counted" one by one, just like we count natural numbers

The set of natural numbers, the set of integers, and the set of rational numbers are countably infinite.

- **Uncountably infinite sets:** Uncountably infinite refers to the size of a set that is so large that its elements cannot be put into a one-to-one correspondence with the natural numbers ( $\mathbb{N}$ ). In other words, there is no way to "list" all the elements of an uncountably infinite set in a sequence, even if the sequence were to go on forever.

This means that even if you tried to list all the elements in an infinite sequence, there would always be elements left out, making the set "larger" than countably infinite sets.

$\mathbb{R}$  and  $\mathbb{C}$  are uncountably infinite, Any interval of real numbers, such as  $[0,1]$  or  $(-\infty, \infty)$ , is also uncountably infinite.

- **Cardinality of countably infinite sets:** The cardinality of any countably infinite set is denoted by  $\aleph_0$  (aleph-null), which is the smallest type of infinity in set theory. For example,

$$\mathbb{N} : |\mathbb{N}| = \aleph_0$$

$$\mathbb{Z} : |\mathbb{Z}| = \aleph_0$$

$$\mathbb{Q} : |\mathbb{Q}| = \aleph_0$$

All these sets, despite having different structures, have the same "size" in terms of cardinality because they can each be matched one-to-one with the natural numbers.

- **Cardinality of uncountably infinite sets:** The cardinality of uncountably infinite sets is denoted by  $\mathfrak{c}$ , which stands for the "cardinality of the continuum." This term comes from the fact that  $\mathfrak{c}$  represents the number of points on a continuous line (i.e., the real number line).

The most well-known example of an uncountably infinite set is the set of real numbers  $\mathbb{R}$ , especially within any interval (e.g.,  $[0, 1]$ ).

In notation:

$$\begin{aligned} |\mathbb{R}| &= \mathfrak{c} \\ |\mathbb{C}| &= \mathfrak{c}. \end{aligned}$$

- **Connection between  $\aleph_0$  and  $\mathfrak{c}$ :** The cardinality of countably infinite sets ( $\aleph_0$ ) is strictly less than the cardinality of uncountably infinite sets ( $\mathfrak{c}$ ). This distinction highlights that uncountably infinite sets are much larger in terms of size, even though both are infinite.

In notation:

$$\aleph_0 < \mathfrak{c}.$$

$\mathfrak{c}$  is vastly larger than  $\aleph_0$ . The set of real numbers (and other uncountably infinite sets) contains infinitely more elements than any countably infinite set, despite both being infinite.

- **The continuum hypothesis:** The Continuum Hypothesis, which is independent of standard set theory (Zermelo-Fraenkel with the Axiom of Choice, ZFC), posits that there is no cardinality between  $\aleph_0$  and  $\mathfrak{c}$ . This suggests that  $\mathfrak{c}$  is the next "step up" in size after  $\aleph_0$ , with no intermediate infinite cardinalities.
- **Fundamental theorem of algebra:** The Fundamental Theorem of Algebra states that every non-constant polynomial equation with complex coefficients has at least one complex root

Any polynomial  $P(x)$  of degree  $n$ , where  $n \geq 1$  and the coefficients of the polynomial are complex numbers, can be factored into exactly  $n$  linear factors in the complex numbers.

Recall that all real numbers are by definition complex, real numbers are of the form  $a + bi$ , where  $a$  is real and  $b = 0$ .

Consider  $f(x) = x^2 - 1$ . Since this is a degree two polynomial, the theorem states that we will find exactly two complex roots, where  $f(x) = 0$

- **Fields:** A mathematical field is a set equipped with two operations—typically called addition and multiplication—that satisfy specific properties. Here's a basic outline:

**Set:** A field consists of a set of elements. Common examples include the set of real numbers  $\mathbb{R}$ , the set of complex numbers  $\mathbb{C}$ , and the set of rational numbers  $\mathbb{Q}$ .

**Addition:** The set is closed under addition. This means if you take any two elements from the set and add them, the result is also an element of the set. The field must have:

- **Commutativity:**  $a + b = b + a$
- **Associativity:**  $(a + b) + c = a + (b + c)$
- **Additive Identity:** There exists an element 0 in the field such that  $a + 0 = a$  for any element  $a$ .
- **Additive Inverse:** For every element  $a$ , there exists an element  $-a$  such that  $a + (-a) = 0$ .

**Multiplication:** The set is also closed under multiplication, and the field satisfies:

- **Commutativity:**  $a \times b = b \times a$
- **Associativity:**  $(a \times b) \times c = a \times (b \times c)$
- **Multiplicative Identity:** There exists an element 1 (different from 0) in the field such that  $a \times 1 = a$  for any element  $a$ .
- **Multiplicative Inverse:** For every element  $a \neq 0$ , there exists an element  $a^{-1}$  such that  $a \times a^{-1} = 1$ .

**Distributivity:** Multiplication is distributive over addition, meaning:

$$a \times (b + c) = (a \times b) + (a \times c)$$

Fields form the foundation for many areas of mathematics, particularly in algebra, analysis, and number theory.

**Notes: Additive Inverse:** For every element  $a$  in the field, there exists another element  $-a$  such that when  $a$  is added to  $-a$ , the result is the additive identity 0. This means that what we typically call "subtraction" is just adding the inverse of a number.

**No Explicit Subtraction:** Instead of saying "subtraction," we can describe subtraction as the addition of a positive number and its additive inverse. For example, instead of saying  $a - b$ , you could say  $a + (-b)$ , where  $-b$  is the additive inverse of  $b$ .

- **Scalar field:** A scalar field is a mathematical function that assigns a single scalar value (a real number) to every point in a space. The value of the scalar field can vary from point to point, but at each point in the space, the field is represented by just one number.
- **Vector field:** A vector field is a mathematical function that assigns a vector to every point in a space. Each vector has both magnitude and direction, and these can vary from point to point in the field.
- **Mapping notation:** Mapping notation is a concise and formal way of describing functions in mathematics, particularly how elements from one set (the domain) are associated with elements in another set (the codomain). It is used to specify the domain, codomain, and the rule of the function in a clear and structured manner.

The general form of mapping notation is:

$$f : A \rightarrow B.$$

- $f$  is the name of the function.

- $A$  is the domain (the set of inputs).
- $B$  is the codomain (the set of outputs).
- The arrow  $\rightarrow$  indicates the direction of the mapping from the domain to the codomain.

Consider the function  $f : \mathbb{R} \rightarrow \mathbb{R}$ , where  $f(x) = x^2$ :

$$f : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto x^2.$$

Consider the function  $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ , where  $g(x, y) = x + y$ , then

$$g : \mathbb{R}^2 \rightarrow \mathbb{R}, (x, y) \mapsto x + y.$$

- **Some known maps**

$$\begin{aligned} f : \mathbb{R} \rightarrow \mathbb{R}^2 &\implies \text{Parameterized curve} \\ f : \mathbb{R}^k \rightarrow \mathbb{R} &\implies \text{Scalar field} \\ f : \mathbb{R}^k \rightarrow \mathbb{R}^2 &\implies \text{Vector field.} \end{aligned}$$

- **Affine functions:** An affine function is a type of function defined as  $f(x) = mx + b$ , where  $m$  is the slope and  $b$  is the intercept. It is similar to a linear function but can include a non-zero intercept, allowing the graph to shift vertically. Affine functions preserve straight lines and parallelism but do not necessarily pass through the origin. In higher dimensions, affine transformations include translations and linear transformations.

Affine functions are closely related to linear functions, but they are not the same. A linear function is a special case of an affine function where the intercept  $b$  is zero, meaning it passes through the origin

- **Unit interval:** The unit interval refers to the set of real numbers between 0 and 1, inclusive. It is denoted as  $[0, 1]$
- **Basics of imaginary numbers:** The concept of an imaginary number is based on the unit  $i$ , defined by the property  $i^2 = -1$

Imaginary numbers can be added, subtracted, and multiplied like real numbers

$$\begin{aligned} 2i + 3i &= 5i \\ 2i \times 3i &= 6i^2 = 6(-1) = -6. \end{aligned}$$

Another property that we will soon see comes from the complex conjugate is  $\frac{1}{i} = -i$

- **basics of complex numbers:** A complex number is a number that has both a real part and an imaginary part. It's generally written in the form  $a + bi$ , where  $a$  and  $b$  are real numbers.  $a$  is the real part,  $b$  is the imaginary part.

We can add or subtract two complex numbers

$$\begin{aligned} (a + bi) + (c + di) &= (a + c) + i(b + d) \\ (a + bi) - (c + di) &= (a - c) + i(b - d). \end{aligned}$$

We can also multiply or divide two complex numbers

$$(a + bi)(c + di) = (ac - bd) + i(ad + bc)$$

To divide, we first need to discuss the complex conjugate. For a complex number  $c + di$ , the complex conjugate is  $c - di$ . Thus,

$$\begin{aligned}\frac{a + bi}{c + di} &= \frac{a + bi}{c + di} \cdot \frac{c - di}{c - di} \\ &= \left( \frac{ac + bd}{c^2 + d^2} \right) + i \left( \frac{bc - ad}{c^2 + d^2} \right).\end{aligned}$$

Using the complex conjugate,

$$\frac{1}{i} = \frac{1}{i} \cdot \frac{-i}{-i} = -i.$$

Complex and imaginary numbers adhere to most of the traditional rules of algebra. However, they introduce some unique aspects that differ from real numbers.

- **No natural ordering:** Real numbers have a natural ordering (e.g., you can say  $3 > 2$ ), but there is no way to define a similar ordering for complex numbers. You can't say, for example, that  $i > 1$  or  $i < 1$ , because imaginary and complex numbers don't fit into a one-dimensional ordered system.

This breaks the rule that any two numbers  $a$  and  $b$  can be compared as either  $a > b$ ,  $a < b$ , or  $a = b$ . Complex numbers lack this property because they exist in a two-dimensional plane.

### 6.1.2 Intro to linear systems and matrices

- **Linear equation:** The equation

$$a_1x_1 + a_2x_2 + \dots + a_nx_n = b. \quad (1)$$

Which expresses the real or complex quantity  $b$  in terms of the unknowns  $x_1, x_2, \dots, x_n$  and the real or complex constants  $a_1, a_2, \dots, a_n$ , is called a **linear equation**. In many applications we are given  $b$  and must find numbers  $x_1, x_2, \dots, x_n$  satisfying (1).

**Note:** A linear equation, by definition, involves addition, writing something like  $a_1x_1 \cdot a_2x_2$  the equation becomes **nonlinear**

- **Solution to a linear equation:** A **solution** to linear Equation (1) is a sequence of  $n$  numbers  $s_1, s_2, \dots, s_n$ , which has the property that (1) is satisfied when  $x_1 = s_1, x_2 = s_2, \dots, x_n = s_n$  are substituted in (1). Thus  $x_1 = 2, x_2 = 3$ , and  $x_3 = -4$  is a solution to the linear equation

$$6x_1 - 3x_2 + 4x_3 = -13,$$

because

$$6(2) - 3(3) + 4(-4) = -13.$$

- **System of  $m$  linear equations in  $n$  unknowns:** More generally, a **system of  $m$  linear equations in  $n$  unknowns**,  $x_1, x_2, \dots, x_n$ , or a **linear system**, is a set of  $m$  linear equations each in  $n$  unknowns. A linear

A linear system can conveniently be written as

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n &= b_m. \end{aligned} \quad (2)$$

Thus the  $i$ th equation is

$$a_{i1}x_1 + a_{i2}x_2 + \dots + a_{in}x_n = b_i.$$

In (2) the  $a_{ij}$  are known constants. Given values of  $b_1, b_2, \dots, b_m$ , we want to find values of  $x_1, x_2, \dots, x_n$  that will satisfy each equation in (2).

- **Linear system solution:** A **solution** to linear system (2) is a sequence of  $n$  numbers  $s_1, s_2, \dots, s_n$ , which has the property that each equation in (2) is satisfied when  $x_1 = s_1, x_2 = s_2, \dots, x_n = s_n$  are substituted.
- **Linear system Consistency:** If the linear system (2) has no solution, it is said to be **inconsistent**. if it has a solution, it is said to be **consistent**
- **Homogeneous system:** If  $b_1 = b_2 = \dots = b_m = 0$  then (2) is called a homogeneous system.
- **Trivial solution:** Note that if  $x_1 = x_2 = \dots = x_n = 0$  is always a solution, it is called the trivial solution to the homogeneous system.
- **Non trivial solution:** A solution to a homogeneous system in which not all of  $x_1, x_2, \dots, x_n$  are zero is called a nontrivial solution

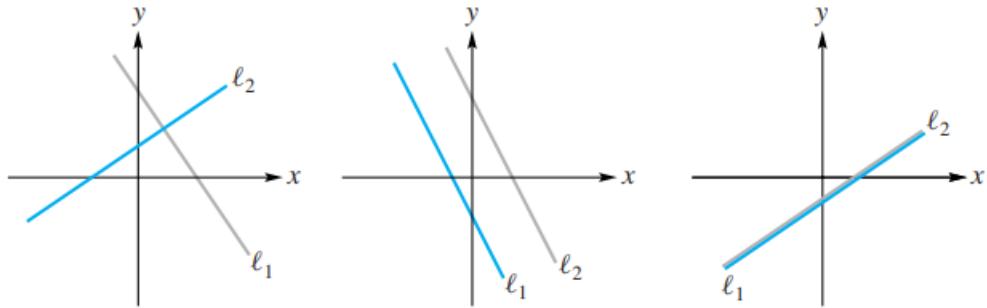
- **Equivalent systems:** Consider another system of  $r$  linear equations in  $n$  unknowns:

$$\begin{aligned} c_{11}x_1 + c_{12}x_2 + \cdots + c_{1n}x_n &= d_1 \\ c_{21}x_1 + c_{22}x_2 + \cdots + c_{2n}x_n &= d_2 \\ &\vdots \\ c_{r1}x_1 + c_{r2}x_2 + \cdots + c_{rn}x_n &= d_r. \end{aligned} \tag{3}$$

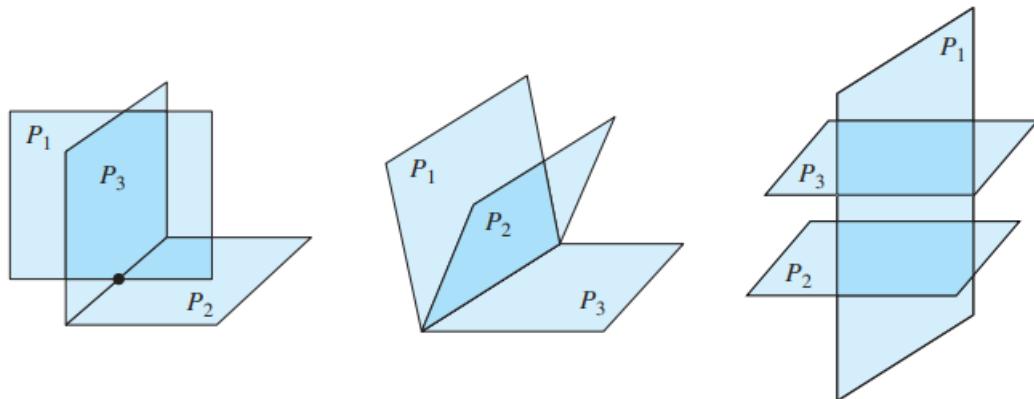
We say that (2) and (3) are **equivalent** if they both have exactly the same solutions.

- **Method of elimination (Gaussian elimination):** To find a solution to a linear system, we shall use a technique called the method of elimination; that is, we eliminate some variables by adding a multiple of one equation to another equation. Elimination merely amounts to the development of a new linear system that is equivalent to the original system, but is much simpler to solve
- **Possible solutions to a linear system of two unknowns:** The linear system can have a **unique solution**, **no solution**, or **infinitely many solutions**.

(a) A unique solution.      (b) No solution.      (c) Infinitely many solutions.



(a) A unique solution.      (b) Infinitely many solutions.      (c) No solution.



- If two planes in three-dimensional space are parallel but not coincident (not the same plane), there is a constant distance (space) between them, and they **do not intersect at any point**.
- **Gaussian elimination mechanics:** If we examine the method of elimination more closely, we find that it involves three manipulations that can be performed on a linear system to convert it into an equivalent system. These manipulations are as follows:

1. Interchange the  $i$ th and  $j$ th equations.
2. Multiply an equation by a nonzero constant.
3. Replace the  $i$ th equation by  $c$  times the  $j$ th equation plus the  $i$ th equation,  $i \neq j$ .  
That is, replace

$$a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n = b_i$$

by

$$(ca_{j1} + a_{i1})x_1 + (ca_{j2} + a_{i2})x_2 + \cdots + (ca_{jn} + a_{in})x_n = cb_j + b_i.$$

### 6.1.3 Properties of linearity, more on linear systems and matrices

- **Linear equations in mapping notation:** In mapping notation, a linear equation can be described as a function that maps a vector from one vector space to another. Let's consider a linear equation in the form:

$$a_1x_1 + a_2x_2 + \dots + a_nx_n = b.$$

This equation can be viewed as a mapping  $\mathbb{R}^n \rightarrow \mathbb{R}$  where  $\mathbb{R}^n$  is the space of all  $n$ -dimensional vectors of real numbers, and  $\mathbb{R}$  is the space of real numbers. The function  $f$  is defined as:

$$f(\vec{x}) = a_1x_1 + a_2x_2 + \dots + a_nx_n.$$

Where  $\vec{x} = \langle a_1x_1 + a_2x_2 + \dots + a_nx_n \rangle \in \mathbb{R}^n$

- **The properties of linear equations:** A function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  representing a linear equation is linear, meaning it satisfies the following properties for all vectors  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$  and all scalars  $c \in \mathbb{R}$ :

- **Additivity:**  $f(\mathbf{x} + \mathbf{y}) = f(\mathbf{x}) + f(\mathbf{y})$
- **Homogeneity of Degree 1:**  $f(c\mathbf{x}) = cf(\mathbf{x})$

It follows from this that  $f(c\mathbf{x})$ , when  $c = 0$  implies  $f(0\mathbf{x}) = 0f(\mathbf{x}) = 0$ . Thus, we add the property

- **Scale by zero:**  $f(0) = 0$

These properties define a linear function and imply that the graph of a linear equation is a straight line (in 2D) or a plane (in 3D).

- **A way to look at linear systems with matrices:** If we examine the method of elimination described in Section 1.1, we can make the following observation: Only the numbers in front of the unknowns  $x_1, x_2, \dots, x_n$  and the numbers  $b_1, b_2, \dots, b_m$  on the right side are being changed as we perform the steps in the method of elimination. Thus we might think of looking for a way of writing a linear system without having to carry along the unknowns. Matrices enable us to do this--that is, to write linear systems in a compact form that makes it easier to automate the elimination method by using computer software in order to obtain a fast and efficient procedure for finding solutions. The use of matrices, however, is not merely that of a convenient notation. We now develop operations on matrices and will work with matrices according to the rules they obey; this will enable us to solve systems of linear equations and to handle other computational problems in a fast and efficient manner. Of course, as any good definition should do, the notion of a matrix not only provides a new way of looking at old problems, but also gives rise to a great many new questions, some of which we study in this book.
- **Matrix definition:** An  $m \times n$  matrix  $A$  is a rectangular array of  $mn$  real or complex numbers arranged in  $m$  horizontal rows and  $n$  vertical columns:

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}$$

The  $i$ th row of  $A$  is:

$$\{a_{i1}, a_{i2}, \dots, a_{in}\} \quad (1 \leq i \leq m)$$

The  $j$ th column of  $A$  is:

$$\begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{bmatrix} \quad (1 \leq j \leq n)$$

- **Some matrix terminology:** We shall say that  $A$  is  $m$  by  $n$  (written as  $m \times n$ ). If  $m = n$ , we say that  $A$  is a *square matrix of order  $n$* , and that the numbers  $a_{11}, a_{22}, \dots, a_{nn}$  form the *main diagonal* of  $A$ . We refer to the number  $a_{ij}$ , which is in the  $i$ th row and  $j$ th column of  $A$ , as the  $i, j$ th element of  $A$ , or the  $(i, j)$  entry of  $A$ , and we often write (1) as:

$$A = [a_{ij}]$$

- **$n$ -vectors, vectors:** An  $n \times 1$  matrix is also called an  *$n$ -vector* and is denoted by lowercase boldface letters. When  $n$  is understood, we refer to  $n$ -vectors merely as *vectors*. Vectors are discussed at length in Section 4.1.

$$\vec{u} = \begin{bmatrix} 1 \\ 2 \\ -1 \\ 0 \end{bmatrix} \quad \text{is a 4-vector and} \quad \vec{v} = \begin{bmatrix} 1 \\ -1 \\ 3 \end{bmatrix} \quad \text{is a 3-vector.}$$

The  $n$ -vector all of whose entries are zero is denoted by  $\vec{0}$ .

**Note:** Observe that if  $A$  is an  $n \times n$  matrix, then the rows of  $A$  are  $1 \times n$  matrices and the columns of  $A$  are  $n \times 1$  matrices.

- **$\mathbb{R}^n$ , The set of real entered n-vectors:** The set of all  $n$ -vectors with real entries is denoted by  $\mathbb{R}^n$
- **$\mathbb{C}^n$ , The set of complex entered n-vectors:** The set of all  $n$ -vectors with complex entries is denoted by  $\mathbb{C}^n$
- **Expressing linear equations in matrix notation:** Consider a simple linear equation in two variables:

$$a_1x_1 + a_2x_2 = b$$

This can be represented in matrix form as:

$$\begin{bmatrix} a_1 & a_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = [b]$$

Here:

- $[a_1 \ a_2]$  is a row vector representing the coefficients of the variables  $x_1$  and  $x_2$ .
- $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$  is a column vector representing the variables.
- $[b]$  is the constant on the right-hand side.

- **Expressing linear systems in matrix notation:** Now, consider a system of linear equations:

$$a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1$$

$$a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2$$

⋮

$$a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m$$

This system can be expressed in matrix form as:

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

Where:

- $\mathbf{A}$  is the coefficient matrix:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}$$

- $\mathbf{x}$  is the column vector of variables:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

- $\mathbf{b}$  is the column vector of constants on the right-hand side:

$$\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

- **Some maps we know and their linear equations:**

- $f : \mathbb{R} \rightarrow \mathbb{R}^2$ : For this we take in a single input and expect a 2-vector output, thus

$$\begin{aligned} f : \mathbb{R} \rightarrow \mathbb{R}^2 &\implies x \rightarrow (a_1x, a_2x) \\ &\implies f(x) = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} x = \vec{a}x. \end{aligned}$$

- $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ : Since this function maps a vector in  $\mathbb{R}^2$  to a single real number, we have the (only) linear equation of the form

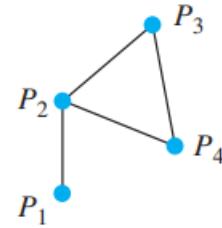
$$\begin{aligned} f : \mathbb{R}^2 \rightarrow \mathbb{R} &\implies (x_1, x_2) \rightarrow (a_1x_1 + a_2x_2) \\ &\implies f(\vec{x}) = f\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = a_1x_1 + a_2x_2 = [a_1 \ a_2] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= \vec{a}\vec{x}. \end{aligned}$$

- $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ : We have,

$$\begin{aligned} f : \mathbb{R}^2 \rightarrow \mathbb{R}^2 &\implies (x_1, x_2) \rightarrow (ax_1 + bx_2, cx_1 + dx_2) \\ &\implies f(\vec{x}) = f\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} ax_1 + bx_2 \\ cx_1 + dx_2 \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= A\vec{x}. \end{aligned}$$

- **Graph, nodes, vertices, edges:** By a **graph** we mean a set of points called **nodes** or **vertices**, some of which are connected by lines called **edges**. The nodes are usually labeled as  $P_1, P_2, \dots, P_k$  and for now we allow an edge to be traveled in either direction. One mathematical representation of a graph is constructed from a table. For example, the following table represents the graph shown:

	$P_1$	$P_2$	$P_3$	$P_4$
$P_1$	0	1	0	0
$P_2$	1	0	1	1
$P_3$	0	1	0	1
$P_4$	0	1	1	0



The  $(i, j)$  entry = 1 if there is an edge connecting vertex  $P_i$  to vertex  $P_j$ ; otherwise, the  $(i, j)$  entry = 0.

- **Incidence matrix:** The *incidence matrix*  $A$  is the  $k \times k$  matrix obtained by omitting the row and column labels from the preceding table. The incidence matrix for the corresponding graph is

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

- **Row-echelon form:** Row-echelon form is a specific kind of matrix form used in linear algebra, particularly in the process of solving systems of linear equations using Gaussian elimination

### Properties:

- **Leading Coefficient of 1:** In each non-zero row, the first non-zero number (from the left) is 1. This is called the leading coefficient or pivot.
- **Zeros Below the Leading Coefficient:** All entries directly below a leading coefficient (in the same column) are zeros
- **Staircase Pattern:** The leading coefficient of each row must be to the right of the leading coefficient of the row directly above it. This creates a "staircase" or "triangular" pattern when viewed across the rows.
- **Rows of All Zeros (if any):** If there are any rows that consist entirely of zeros, they are placed at the bottom of the matrix.

Consider the matrix

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

Here, the vertical bar separates the matrix  $A = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & -1 \\ 0 & 0 & 2 \end{bmatrix}$

from the vector  $b = \begin{bmatrix} 4 \\ 3 \\ 5 \end{bmatrix}$ .

This matrix is in row-echelon form because:

- The first non-zero entry in the first row is 1.
- The first non-zero entry in the second row (which is also 1) is to the right of the leading 1 in the first row.
- The first non-zero entry in the third row (2) is to the right of the leading 1 in the second row.
- There are zeros below each leading coefficient.
- The last row is all zeros and is at the bottom of the matrix.

- **Augmented matrix:** The matrix we saw above, with the vertical bar is called an **augmented matrix**. Formally, given a linear system  $A\vec{x} = \vec{b}$ , the augmented matrix is then  $[A|\vec{b}]$
- **Back substitution from row-echelon form:** Back substitution is a method used to solve a system of linear equations once the system's matrix is in row-echelon form. This method works by solving the equations starting from the bottom row and moving upward, substituting the values obtained into the equations above.

For example, given the system

$$\begin{cases} x + 2y + 3z = 4 \\ y - z = 3 \\ 2z = 5 \end{cases} = \left[ \begin{array}{ccc|c} 1 & 2 & 3 & 4 \\ 0 & 1 & -1 & 3 \\ 0 & 0 & 2 & 5 \end{array} \right].$$

Since this matrix is already in row-echelon form, we can just focus on the back substitution, starting at the bottom,

$$\begin{aligned} 2z = 5 &\implies z = \frac{5}{2} \\ y - z = 3 &\implies y - \frac{5}{2} = 3 \implies y = \frac{11}{2} \\ x + 2y + 3z = 4 &\implies x + 2\left(\frac{11}{2}\right) + 3\left(\frac{5}{2}\right) = 4 \implies x = -\frac{29}{2}. \end{aligned}$$

- **Reduced row-echelon form:** This form is a product of further reducing row-echelon form to get a matrix in the form

$$\left[ \begin{array}{ccc|c} 1 & 0 & 0 & b_1 \\ 0 & 1 & 0 & b_2 \\ 0 & 0 & 1 & b_3 \end{array} \right].$$

- **Gaussian elimination:** For the linear system  $A\vec{x} = \vec{b}$  form the augmented matrix  $[A | \vec{b}]$ . Compute the row echelon form of the augmented matrix; then the solution can be computed using back substitution
- **Gauss-jordan reduction:** For the linear system  $A\vec{x} = \vec{b}$  form the augmented matrix  $[A | \vec{b}]$ . Compute the reduced row echelon form of the augmented matrix; then the solution can be computed using back substitution
- **Row reduction: Getting a matrix into row-echelon form or reduced row-echelon form:** Given a system of linear equations, we can perform a couple operations to reduce, we have

- **Row Swapping:** Swapping rows
- **Row Multiplication:** Multiplying a row by some constant factor
- **Row addition/subtraction:** Multiplying one row by some constant factor and adding or subtracting it to a different row (not changing the row we scaled, only changing the row we add it to),

**Note:** These operations do not change the solution set.

- **Mechanical process for guassian elimination:**

1. Start at the first entry, we want this to be one, after we get this to be one we go down, getting the entries remaining entries in the column to be zero
2. Go to the next entry in the main diagonal, we also want this to be one, then we go down to get the remaining entries zero.
3. If we can't go down, we go up.

- **Inconsistent system of linear equations:** If while we are doing row reduction (guassian elimination) on a linear system, and one row becomes inconsistent, we assert the entire system is then inconsistent. Then solution set would then be the empty set  $\emptyset$

**Example:** Suppose we have the system

$$\begin{cases} x + 2y + 6z = 5 \\ -x + y - 2z = 3 \\ x - 4y - 2z = 1 \end{cases} .$$

The augmented matrix is then

$$\left[ \begin{array}{ccc|c} 1 & 2 & 6 & 5 \\ -1 & 1 & -2 & 3 \\ 1 & -4 & -2 & 1 \end{array} \right] .$$

After some row reduction, we get

$$\left[ \begin{array}{ccc|c} 1 & 2 & 6 & 5 \\ 0 & 1 & \frac{4}{3} & \frac{8}{3} \\ 0 & 0 & 0 & 12 \end{array} \right] .$$

We notice that this last row implies  $0 = 12$ . Thus we assert this row is inconsistent, which implies the entire system is inconsistent, and thus the solution set is

$$S = \emptyset.$$

- **Redundancy in linear systems:** Take another look at the last version of the augmented matrix in the above example, it may turn out that the last row gives the information  $0 = 0$ . In this case, there is no useful information to be obtained from the last row and we declare it redundant. All the information we need about the system is then obtained by the first and second rows.
- **Parametric Solution of a System of Linear Equations with Infinite Solutions: Example problem:**

Given the system

$$\begin{cases} 3x - 2y + 3z = 8 \\ x + 3y + 6z = -3 \\ 2x + 6y + 12z = -6 \end{cases}.$$

Which yields the augmented matrix

$$\left[ \begin{array}{ccc|c} 3 & -2 & 3 & 8 \\ 1 & 3 & 6 & -3 \\ 2 & 6 & 12 & -6 \end{array} \right].$$

After some row reduction we get

$$\left[ \begin{array}{ccc|c} 1 & 0 & \frac{21}{11} & \frac{18}{11} \\ 0 & 1 & \frac{15}{11} & -\frac{17}{11} \\ 0 & 0 & 0 & 0 \end{array} \right].$$

Which implies

$$\begin{aligned} x &= \frac{18}{11} - \frac{21}{11}z \\ y &= -\frac{17}{11} - \frac{15}{11}z. \end{aligned}$$

The solution set is uncountably infinite

Now, let  $\alpha$  be a real number, we have

$$\begin{aligned} \left( \frac{18}{11} - \frac{21}{11}\alpha \right) \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + \left( -\frac{17}{11} - \frac{15}{11}\alpha \right) \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + \alpha \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \\ = \begin{bmatrix} \frac{18}{11} \\ -\frac{17}{11} \\ 0 \end{bmatrix} + \alpha \begin{bmatrix} -\frac{21}{11} \\ -\frac{15}{11} \\ 1 \end{bmatrix}. \end{aligned}$$

Now, we have a vector form to the set of solutions in the form  $\vec{p} + t\vec{v}$

$$\vec{x} = \vec{p} + \alpha \vec{d}.$$

**Why introduce  $\alpha$ ?**: By introducing  $\alpha$ , we explicitly indicate that the solution set is parameterized by a single variable that can vary freely, independent of the original variables  $x, y$ , and  $z$ . This makes it clear that the solution set forms a line (or possibly a plane, etc., in higher dimensions) in the solution space, depending on the number of free parameters.

- **Does the solution set form a line, plane, hyperplane, or something else?**: The formation of the solution set depends on the number of free variables,

- **No free variables (one unique solution)**: Intersects at a point
- **One free variable (Uncountable solutions)**: Solution set is a line (1-dimensional subspace)
- **Two free variable (Uncountable solutions)**: Solution set forms a plane (2-dimensional subspace)
- **Three free variable (Uncountable solutions)**: Solution set is a three dimensional subspace (In  $\mathbb{R}^3$  it would be the whole space)
- **$k$  free variables**: Solution set is a  $k$ -dimensional subspace in  $\mathbb{R}^n$

**Note:** A  $k$ -dimensional subspace in  $\mathbb{R}^n$  means that the solution set spans a  $k$ -dimensional space within the  $n$ -dimensional ambient space  $\mathbb{R}^n$ .

- **The ambient space ( $\mathbb{R}^n$ ):** In mathematics,  $\mathbb{R}^n$  (the  $n$ -dimensional real coordinate space) can indeed be referred to as an ambient space. In this context, the ambient space refers to the larger space within which objects (such as vectors, points, curves, or surfaces) reside or are embedded.
- **More on linear maps:**  $L : \mathbb{R} \rightarrow \mathbb{R}^2$ : For this we have

$$L(x) = (a_1x, a_2x).$$

Which can be written as

$$L(x) = \begin{bmatrix} a \\ b \end{bmatrix}.$$

Or

$$\begin{aligned} a_1x &= a \\ a_2x &= b. \end{aligned}$$

From here, we can derive (using gaussian elimination)

$$\begin{bmatrix} 1 & \frac{a}{a_1} \\ 0 & b - \frac{a_2a}{a_1} \end{bmatrix}.$$

Assume  $b - \frac{a_2a}{a_1} = 0$  (Otherwise the system would be inconsistent), then

$$\begin{aligned} x &= \frac{a}{a_1} \\ \implies b - a_2x &= 0 \\ \implies x &= \frac{b}{a_2} \\ \implies \frac{a}{a_1} &= \frac{b}{a_2} \\ \implies \frac{a_2}{a_1} &= \frac{b}{a}. \end{aligned}$$

We can see that this is unlikely, and the system will probably be inconsistent

In practical situations, there is no reason to expect that the ratio  $\frac{a}{b}$  would match the ratio  $\frac{a_2}{a_1}$ , unless they were specially chosen or related in some way. The constants  $a_1$  and  $a_2$  come from the definition of the linear map, while  $a$  and  $b$  represent desired outcomes (or targets), which are often independent of the map's coefficients.

- **Determine if three planes intersect at a unique point:** For this, we find all three normal vectors  $\vec{n}_1, \vec{n}_2$ , and  $\vec{n}_3$ . Then we find the triple scalar product, that is

$$\vec{n}_1 \cdot (\vec{n}_2 \times \vec{n}_3).$$

If this value is non-zero, we have intersection at a unique point. If the value is zero, we either have no intersection, or intersection at a line.

- **Determine if three planes intersect at a line:** If the solution to the system all involve a free parameter, then we know all three planes intersect at a point

- **Determine if three planes have two planes that intersect at a line, and the third intersects at a point:** If the solution to the system has two variables that involve a free parameter, while one is constant.
- **Transpose of a matrix:** If  $A = [a_{ij}]$  is an  $m \times n$  matrix, then the transpose of  $A$ ,  $A^T = [a_{ij}^T]$ , is the  $n \times m$  matrix defined by  $a_{ij}^T = a_{ji}$ . Thus the transpose of  $A$  is obtained from  $A$  by interchanging the rows and columns of  $A$ .

**Example:** If  $A$  is a  $2 \times 3$  matrix defined by

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}.$$

Then the transpose of  $A$ ,  $A^T$  is

$$A^T = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}.$$

Which is a  $3 \times 2$  matrix

- **Recall:** A  $m \times n$  matrix has  $m$  rows and  $n$  columns
- **Some more matrix notation:** We can denote the space that a matrix lives by writing

$$\mathbb{R}_{m \times n}.$$

- **Matrix maps:** Suppose we have

$$L : \mathbb{R}_{1 \times 2} \rightarrow \mathbb{R}_{2 \times 1}.$$

Then

$$L([x \ y]) = \begin{bmatrix} ax + by \\ cx + dy \end{bmatrix}.$$

#### 6.1.4 Algebra of matrices, matrix operations

- **Dot product:** The **dot product**, or **inner product**, of the  $n$ -vectors in  $\mathbb{R}^n$

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

is defined as

$$\mathbf{a} \cdot \mathbf{b} = a_1b_1 + a_2b_2 + \cdots + a_nb_n = \sum_{i=1}^n a_i b_i.$$

**Note:** Only defined for vectors of the same size

- **Matrix multiplication (Matrix-vector product) (formal):** If  $A = [a_{ij}]$  is an  $m \times p$  matrix and  $B = [b_{ij}]$  is a  $p \times n$  matrix, then the **product** of  $A$  and  $B$ , denoted  $AB$ , is the  $m \times n$  matrix  $C = [c_{ij}]$ , defined by

$$\begin{aligned} c_{ij} &= a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{ip}b_{pj} \\ &= \sum_{k=1}^p a_{ik}b_{kj} \quad (1 \leq i \leq m, 1 \leq j \leq n). \end{aligned}$$

This says that the  $i, j$ th element in the product matrix is the dot product of the transpose of the  $i$ th row,  $\text{row}_i(A)$ —that is,  $(\text{row}_i(A))^T$ —and the  $j$ th column,  $\text{col}_j(B)$ , of  $B$ ;

$$\begin{aligned} \text{row}_i(A) &\left[ \begin{array}{cccc} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ip} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mp} \end{array} \right] \quad \text{col}_j(B) \\ &\left[ \begin{array}{ccccc} b_{11} & b_{12} & \cdots & b_{1j} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2j} & \cdots & b_{2n} \\ \vdots & \vdots & & \vdots & & \vdots \\ b_{p1} & b_{p2} & \cdots & b_{pj} & \cdots & b_{pn} \end{array} \right] \\ &= \left[ \begin{array}{ccccc} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & & \vdots \\ c_{m1} & c_{m2} & \cdots & c_{mn} \end{array} \right] \\ (\text{row}_i(A))^T \cdot \text{col}_j(B) &= \sum_{k=1}^p a_{ik}b_{kj} = c_{ij} \end{aligned}$$

Observe that the product of  $A$  and  $B$  is defined only when the number of rows of  $B$  is exactly the same as the number of columns of  $A$

- **More Matrix multiplication (Matrix-vector product):** Suppose we have two vectors  $M_{m \times n}$  and  $N_{n \times r}$ . And we want to perform the operation

$$M_{m \times n} N_{n \times r}.$$

As long as the number of columns in the first matrix and the number of rows in the second matrix are the same, then this operation will hold. The resulting matrix will then be

$$(MN)_{m \times r}.$$

**Note:** We would like this operation to correspond to composition.

Consider

$$\begin{bmatrix} a \\ b \end{bmatrix}_{2 \times 1} \begin{bmatrix} c & d & e \end{bmatrix}_{1 \times 3}.$$

Thus the resulting matrix will be  $2 \times 3$

$$= \begin{bmatrix} ac & ad & ae \\ bc & bd & be \end{bmatrix}.$$

Consider

$$\begin{bmatrix} c & d \end{bmatrix}_{1 \times 2} \begin{bmatrix} a \\ b \end{bmatrix}_{2 \times 1}.$$

This one is a bit strange, we get

$$\begin{bmatrix} ca + db \end{bmatrix}_{1 \times 1}.$$

- **Commutativity of matrix-vector product:** Multiplication of matrices requires much more care than their addition, since the algebraic properties of matrix multiplication differ from those satisfied by the real numbers. Part of the problem is due to the fact that  $AB$  is defined only when the number of columns of  $A$  is the same as the number of rows of  $B$ . Thus, if  $A$  is an  $m \times p$  matrix and  $B$  is a  $p \times n$  matrix, then  $AB$  is an  $m \times n$  matrix. What about  $BA$ ? Four different situations may occur:
  1.  $BA$  may not be defined; this will take place if  $n \neq m$ .
  2. If  $BA$  is defined, which means that  $m = n$ , then  $BA$  is  $p \times p$  while  $A$  is  $m \times n$ ; thus, if  $m \neq p$ ,  $AB$  and  $BA$  are of different sizes.
  3. If  $AB$  and  $BA$  are both of the same size, they may be equal.
  4. If  $AB$  and  $BA$  are both of the same size, they may be unequal.
- **Dot Product as matrix-vector product:** If  $\mathbf{u}$  and  $\mathbf{v}$  are  $n$ -vectors ( $n \times 1$  matrices), then it is easy to show by matrix multiplication that

$$\mathbf{u} \cdot \mathbf{v} = \mathbf{u}^T \mathbf{v}.$$

- **The Matrix-Vector Product Written in Terms of Columns:** Let  $A$  be a  $m \times n$  matrix, with

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}.$$

Let  $c$  be an  $n$ -vector, that is, an  $n \times 1$  matrix

$$\vec{\mathbf{c}} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix}.$$

Since  $A$  is  $m \times n$  and  $\vec{\mathbf{c}}$  is  $n \times 1$ , the matrix product  $A\vec{\mathbf{c}}$  is the  $m \times 1$  matrix

$$\begin{aligned}
A\vec{c} &= \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} (\text{row}_1(A))^T \cdot \mathbf{c} \\ (\text{row}_2(A))^T \cdot \mathbf{c} \\ \vdots \\ (\text{row}_m(A))^T \cdot \mathbf{c} \end{bmatrix} \\
&= \begin{bmatrix} a_{11}c_1 + a_{12}c_2 + \cdots + a_{1n}c_n \\ a_{21}c_1 + a_{22}c_2 + \cdots + a_{2n}c_n \\ \vdots \\ a_{m1}c_1 + a_{m2}c_2 + \cdots + a_{mn}c_n \end{bmatrix}.
\end{aligned}$$

This last expression can be written as

$$\begin{aligned}
&\begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{bmatrix} c_1 + \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{bmatrix} c_2 + \cdots + \begin{bmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{bmatrix} c_n \\
&= c_1 \text{col}_1(A) + c_2 \text{col}_2(A) + \cdots + c_n \text{col}_n(A).
\end{aligned}$$

Thus the product  $Ac$  of an  $m \times n$  matrix  $A$  and an  $n \times 1$  matrix  $c$  can be written as a linear combination of the columns of  $A$ , where the coefficients are the entries in the matrix  $c$ .

In our study of linear systems of equations we shall see that these systems can be expressed in terms of a matrix–vector product. This point of view provides us with an important way to think about solutions of linear systems.

If  $A$  is an  $m \times p$  matrix and  $B$  is a  $p \times n$  matrix, we can then conclude that the  $j$ th column of the product  $AB$  can be written as a linear combination of the columns of matrix  $A$ , where the coefficients are the entries in the  $j$ th column of matrix  $B$ :

$$\text{col}_j(AB) = A\text{col}_j(B) = b_{1j}\text{col}_1(A) + b_{2j}\text{col}_2(A) + \cdots + b_{pj}\text{col}_p(A).$$

- **Linear systems as a linear combination of the matrices**

Recall for a system of  $m$  equations in  $n$  unknowns, we can express that system in matrix form as  $A\vec{x} = \vec{b}$ , where  $A$  is the coefficient matrix, etc. If we join  $A$  and  $\vec{x}$ , we get

$$\begin{aligned}
A\vec{x} &= \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \end{bmatrix} \\
&= \begin{bmatrix} a_{11}x_1 \\ a_{21}x_1 \\ \vdots \\ a_{m1}x_1 \end{bmatrix} + \begin{bmatrix} a_{12}x_2 \\ a_{22}x_2 \\ \vdots \\ a_{m2}x_2 \end{bmatrix} + \cdots + \begin{bmatrix} a_{1n}x_n \\ a_{2n}x_n \\ \vdots \\ a_{mn}x_n \end{bmatrix} \\
&= x_1 \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{bmatrix} + x_2 \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{bmatrix} + \cdots + x_n \begin{bmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{bmatrix} \\
&= x_1 \text{col}_1(A) + x_2 \text{col}_2(A) + \cdots + x_n \text{col}_n(A).
\end{aligned}$$

Thus  $A\mathbf{x}$  is a linear combination of the columns of  $A$  with coefficients that are the entries of  $\mathbf{x}$ . It follows that the matrix form of a linear system,  $A\mathbf{x} = \mathbf{b}$ , can be expressed as

$$x_1 \text{col}_1(A) + x_2 \text{col}_2(A) + \cdots + x_n \text{col}_n(A) = \mathbf{b}. \quad (6)$$

Conversely, an equation of the form in (6) always describes a linear system in the standard system form

- **Note about consistent systems:**  $A\mathbf{x} = \mathbf{b}$  is consistent if and only if  $\mathbf{b}$  can be expressed as a linear combination of the columns of the matrix  $A$ .
- **Linear systems as a linear combination of the matrices**

Recall for a system of  $m$  equations in  $n$  unknowns, we can express that system in matrix form as  $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$ , where  $A$  is the coefficient matrix, etc. If we join  $A$  and  $\vec{\mathbf{x}}$ , we get

$$\begin{aligned} A\mathbf{x} &= \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \end{bmatrix} \\ &= \begin{bmatrix} a_{11}x_1 \\ a_{21}x_1 \\ \vdots \\ a_{m1}x_1 \end{bmatrix} + \begin{bmatrix} a_{12}x_2 \\ a_{22}x_2 \\ \vdots \\ a_{m2}x_2 \end{bmatrix} + \cdots + \begin{bmatrix} a_{1n}x_n \\ a_{2n}x_n \\ \vdots \\ a_{mn}x_n \end{bmatrix} \\ &= x_1 \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{bmatrix} + x_2 \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{bmatrix} + \cdots + x_n \begin{bmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{bmatrix} \\ &= x_1 \text{col}_1(A) + x_2 \text{col}_2(A) + \cdots + x_n \text{col}_n(A). \end{aligned}$$

Thus  $A\mathbf{x}$  is a linear combination of the columns of  $A$  with coefficients that are the entries of  $\mathbf{x}$ . It follows that the matrix form of a linear system,  $A\mathbf{x} = \mathbf{b}$ , can be expressed as

$$x_1 \text{col}_1(A) + x_2 \text{col}_2(A) + \cdots + x_n \text{col}_n(A) = \mathbf{b}. \quad (6)$$

Conversely, an equation of the form in (6) always describes a linear system in the standard system form

- **Note about consistent systems:**  $A\mathbf{x} = \mathbf{b}$  is consistent if and only if  $\mathbf{b}$  can be expressed as a linear combination of the columns of the matrix  $A$ .
- **Matrix maps (Evaluation):** To get a connection between products and evaluation we need to use column matrices in the domain and range. Ie

$$\mathbb{R}^m = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix}.$$

For example, suppose we have  $L : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ , with

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

Then

$$L(\vec{v}) = L\vec{v} = \begin{bmatrix} 1 & 0 \\ -2 & 3 \end{bmatrix} \vec{v}.$$

If  $\vec{v} = \begin{bmatrix} 4 \\ -1 \end{bmatrix}$ , then

$$L\left(\begin{bmatrix} 4 \\ -1 \end{bmatrix}\right) = \begin{bmatrix} 1 & 0 \\ -2 & 3 \end{bmatrix} \begin{bmatrix} 4 \\ -1 \end{bmatrix} = \begin{bmatrix} 4 \\ -11 \end{bmatrix}.$$

- **Matrix operations:** We next define a number of operations that will produce new matrices out of given matrices. When we are dealing with linear systems, for example, this will enable us to manipulate the matrices that arise and to avoid writing down systems over and over again. These operations and manipulations are also useful in other applications of matrices.

- **Matrix addition:** If  $A = [a_{ij}]$  and  $B = [b_{ij}]$  are both  $m \times n$  matrices, then the sum  $A + B$  is an  $m \times n$  matrix  $C = [c_{ij}]$  defined by  $c_{ij} = a_{ij} + b_{ij}$ ,  $i = 1, 2, \dots, m$ ;  $j = 1, 2, \dots, n$ . Thus, to obtain the sum of  $A$  and  $B$ , we merely add corresponding entries.

**Note:** It should be noted that the sum of the matrices  $A$  and  $B$  is defined only when  $A$  and  $B$  have the same number of rows and the same number of columns, that is, only when  $A$  and  $B$  are of the same size. We now make the convention that when  $A + B$  is written, both  $A$  and  $B$  are of the same size.

- **Scalar Multiplication:** If  $A = [a_{ij}]$  is an  $m \times n$  matrix and  $r$  is a real number, then the scalar multiple of  $A$  by  $r$ ,  $rA$ , is the  $m \times n$  matrix  $C = [c_{ij}]$ , where  $c_{ij} = ra_{ij}$ ,  $i = 1, 2, \dots, m$  and  $j = 1, 2, \dots, n$ ; that is, the matrix  $C$  is obtained by multiplying each entry of  $A$  by  $r$ .
- **Matrix difference:** If  $A$  and  $B$  are  $m \times n$  matrices, we write  $A + (-1)B$  as  $A - B$  and call this the difference between  $A$  and  $B$ .

- **The  $m \times n$  zero matrix:** The matrix  $0$  is called the  $m \times n$  **zero matrix**, where all entries are zeros.

#### • Algebraic properties of matrix operations

- **Scalar multiplication:** As we already saw, we have multiplication by a scalar

$$\begin{aligned} s \begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix} \\ = \begin{bmatrix} sa & sb & sc \\ sd & se & sf \end{bmatrix}. \end{aligned}$$

If  $s = 0$ , then

$$\begin{aligned} 0 \begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix} \\ = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \end{aligned}$$

**Note:** The result is the  $2 \times 3$  zero matrix

- **Laws of matrix addition:**

- \* **Addition with the zero matrix:**  $0 + A = A$
- \* **Commutative law for matrix addition:**  $A + B = B + A$
- \* **Associativity of matrix addition:**  $(A + B) + C = A + (B + C)$

- **Laws of matrix subtraction:**

- \*  $A - 0 = A$
- \*  $A - A = 0$
- \*  $B - A = (-1)(A - B)$

- **Matrix difference (subtraction):** We can give a definition to the subtraction operator by just defining it as using matrix addition and multiplication by a scalar  $A - B = A + (-1B)$
- **Note on matrix multiplication:** Matrix multiplication is general **not** commutative, it can be, but it isn't always. Also, in the real numbers, we know for

$$ab = 0.$$

Then either  $a$  is zero,  $b$  is zero, or they are both zero. This is not always the case with matrix multiplication, it is possible to multiply two non-zero matrices and get the zero matrix as a result.

- **Properties of matrix multiplication:**

1. If  $A$ ,  $B$ , and  $C$  are matrices of the appropriate sizes, then

$$A(BC) = (AB)C.$$

2. If  $A$ ,  $B$ , and  $C$  are matrices of the appropriate sizes, then

$$(A + B)C = AC + BC.$$

3. If  $A$ ,  $B$ , and  $C$  are matrices of the appropriate sizes, then

$$C(A + B) = CA + CB.$$

- **Properties of Scalar Multiplication:** If  $r$  and  $s$  are real numbers and  $A$  and  $B$  are matrices of the appropriate sizes, then

1.  $r(sA) = (rs)A$
2.  $(r + s)A = rA + sA$
3.  $r(A + B) = rA + rB$
4.  $A(rB) = r(AB) = (rA)B$

- **Properties of Transpose:** If  $r$  is a scalar and  $A$  and  $B$  are matrices of the appropriate sizes, then

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

- **Note on cancellation:** If  $a$ ,  $b$ , and  $c$  are real numbers for which  $ab = ac$  and  $a \neq 0$ , it follows that  $b = c$ . That is, we can cancel out the nonzero factor  $a$ . However, the cancellation law does not hold for matrices.
- **Differences between matrix multiplication and multiplication of real numbers:** We summarize some of the differences between matrix multiplication and the multiplication of real numbers as follows: For matrices  $A$ ,  $B$ , and  $C$  of the appropriate sizes,
  1.  $AB$  need not equal  $BA$ .
  2.  $AB$  may be the zero matrix with  $A \neq 0$  and  $B \neq 0$ .
  3.  $AB$  may equal  $AC$  with  $B \neq C$ .

### 6.1.5 Composition, rotations

- **Matrix Composition:** Recall for composition of functions we write  $f(g(x)) = (f \circ g)(x)$

Suppose we want

$$\begin{aligned} L : \mathbb{R}^2 &\rightarrow \mathbb{R}^2 \quad (\vec{v} \rightarrow L(\vec{v})) \\ K : \mathbb{R}^2 &\rightarrow \mathbb{R}^2 \quad (L(\vec{v}) \rightarrow K(L(\vec{v}))). \end{aligned}$$

Then we write

$$\begin{aligned} K(L(\vec{v})) &= (K \circ L)(\vec{v}) \\ &= (KL)\vec{v}. \end{aligned}$$

**Example:** Suppose

$$\begin{aligned} L &= \begin{bmatrix} a & b \\ c & d \end{bmatrix} \\ K &= \begin{bmatrix} e & f \\ g & h \end{bmatrix} \\ \vec{v} &= \begin{bmatrix} i \\ j \end{bmatrix}. \end{aligned}$$

Then  $L(\vec{v}) = \begin{bmatrix} ai + bj \\ ci + dj \end{bmatrix}$ ,  $K(L(\vec{v})) = \begin{bmatrix} e(ai + bj) + f(ci + dj) \\ g(ai + bj) + h(ci + dj) \end{bmatrix}$ . If  $KL = \begin{bmatrix} ea + fc & eb + fd \\ ga + hc & gb + hd \end{bmatrix}$ , then

$$\begin{aligned} (KL)\vec{v} &= \begin{bmatrix} ea + fc & eb + fd \\ ga + hc & gb + hd \end{bmatrix} \begin{bmatrix} i \\ j \end{bmatrix} \\ &= \begin{bmatrix} i(ea + fc) + j(eb + fd) \\ i(ga + hc) + j(gb + hd) \end{bmatrix} \\ &= \begin{bmatrix} iea + ifc + jeb + jfd \\ iga + ihc + jgb + jhd \end{bmatrix} \\ &= \begin{bmatrix} iea + jeb + fci + jfd \\ iga + jgb + ihc + jhd \end{bmatrix} \\ &= \begin{bmatrix} e(ai + bj) + f(ci + dj) \\ g(ai + bj) + h(ci + dj) \end{bmatrix}. \end{aligned}$$

Thus,  $K(L(\vec{v})) = (K \circ L)(\vec{v}) = (KL)\vec{v}$

- **2D Rotation matrix (Rotate a 2d vector by  $\theta$  radians):** For this we look at rotation of the basis vectors (unit vectors)  $\hat{i}$  and  $\hat{j}$ . To understand how any arbitrary vector rotates, it's enough to see how the standard basis vectors rotate

Rotating  $\hat{i} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$  by an angle  $\theta$ , we get  $\begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix}$ . Rotating  $\hat{j}$  we get  $\begin{bmatrix} -\sin(\theta) \\ \cos(\theta) \end{bmatrix}$

To describe how a vector transforms under this rotation, we use these rotated basis vectors. The vector  $\vec{v}$  is written as a linear combination of the basis vectors  $\vec{v} = x\hat{i} + y\hat{j}$

After rotation, the vector transforms as

$$\begin{aligned}\vec{v}' &= x \begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix} + y \begin{bmatrix} -\sin(\theta) \\ \cos(\theta) \end{bmatrix} \\ \implies \vec{v}' &= \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}.\end{aligned}$$

Thus we write the rotation matrix, the matrix formed by combining the rotated versions of the basis vectors into columns as

$$R(\theta) = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}.$$

- **Recall: Trig key formulas**

- Pythagorean Identities:

$$\begin{aligned}\sin^2 \theta + \cos^2 \theta &= 1 \\ \tan^2 \theta + 1 &= \sec^2 \theta \\ \cot^2 \theta + 1 &= \csc^2 \theta.\end{aligned}$$

- Sum and difference:

$$\begin{aligned}\cos(\alpha + \beta) &= \cos \alpha \cos \beta - \sin \alpha \sin \beta \\ \cos(\alpha - \beta) &= \cos \alpha \cos \beta + \sin \alpha \sin \beta \\ \sin(\alpha + \beta) &= \sin \alpha \cos \beta + \cos \alpha \sin \beta \\ \sin(\alpha - \beta) &= \sin \alpha \cos \beta - \cos \alpha \sin \beta \\ \tan(\alpha + \beta) &= \frac{\tan(\alpha) + \tan(\beta)}{1 - \tan(\alpha) \tan(\beta)} \\ \tan(\alpha - \beta) &= \frac{\tan(\alpha) - \tan(\beta)}{1 + \tan(\alpha) \tan(\beta)}.\end{aligned}$$

- Double Angle/Half Angle:

$$\begin{aligned}\sin(2\theta) &= 2 \sin \theta \cos \theta \\ \cos 2\theta &= 1 - 2 \sin^2 \theta \\ \tan 2\theta &= \frac{2 \tan \theta}{1 - \tan^2 \theta} \\ \sin \frac{\theta}{2} &= \pm \sqrt{\frac{1 - \cos \theta}{2}} \\ \cos \frac{\theta}{2} &= \pm \sqrt{\frac{1 + \cos \theta}{2}} \\ \tan \frac{\theta}{2} &= \pm \sqrt{\frac{1 - \cos \theta}{1 + \cos \theta}}.\end{aligned}$$

## – Product to Sum:

$$\begin{aligned}\sin a \sin b &= \frac{1}{2}[\cos(a - b) - \cos(a + b)] \\ \cos a \cos b &= \frac{1}{2}[\cos(a - b) + \cos(a + b)] \\ \sin a \cos b &= \frac{1}{2}[\sin(a + b) + \sin(a - b)] \\ \cos a \sin b &= \frac{1}{2}[\sin(a + b) - \sin(a - b)].\end{aligned}$$

## – Sum to Product:

$$\begin{aligned}\sin a + \sin b &= 2 \sin \frac{a+b}{2} \cos \frac{a-b}{2} \\ \sin a - \sin b &= 2 \sin \frac{a-b}{2} \cos \frac{a+b}{2} \\ \cos a + \cos b &= 2 \cos \frac{a+b}{2} \cos \frac{a-b}{2} \\ \cos a - \cos b &= -2 \sin \frac{a+b}{2} \sin \frac{a-b}{2}.\end{aligned}$$

- **Rotating a vector twice, by  $\theta$  and by  $\varphi$ :** Rotating twice is a composition of two rotations, ie  $R(\varphi) \cdot R(\theta) \cdot \vec{v}$

**Note:** When you apply a series of transformations to a vector, the order of matrix multiplication is right to left:

We have

$$\begin{aligned}&\begin{bmatrix} \cos(\varphi) & -\sin(\varphi) \\ \sin(\varphi) & \cos(\varphi) \end{bmatrix} \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \\ &= \begin{bmatrix} \cos(\varphi)\cos(\theta) - \sin(\varphi)\sin(\theta) & -\cos(\varphi)\sin(\theta) - \sin(\varphi)\cos(\theta) \\ \sin(\varphi)\cos(\theta) + \cos(\varphi)\sin(\theta) & -\sin(\varphi)\sin(\theta) + \cos(\varphi)\cos(\theta) \end{bmatrix}.\end{aligned}$$

Using sum and difference formulas, we get

$$\begin{bmatrix} \cos(\varphi + \theta) & -\sin(\varphi + \theta) \\ \sin(\varphi + \theta) & \cos(\varphi + \theta) \end{bmatrix}.$$

**Note:** Rotation of a vector by two angles is commutative, this means  $R(\varphi) \cdot R(\theta) = R(\theta) \cdot R(\varphi)$

- **3D rotation, but keeping one variable constant, ie rotating about one of the coordinate axis.**

All these cases below will require a  $3 \times 3$  matrix

– Rotation about the x-axis (rotation in the  $yz$ -plane):

$$R_x(\theta) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta) & -\sin(\theta) \\ 0 & \sin(\theta) & \cos(\theta) \end{bmatrix}.$$

– Rotation about the y-axis (rotation in the  $xz$ -plane)

$$R_y(\theta) = \begin{bmatrix} \cos(\theta) & 0 & -\sin(\theta) \\ 0 & 1 & 0 \\ \sin(\theta) & 0 & \cos(\theta) \end{bmatrix}.$$

- Rotation about the z-axis (rotation in the  $xy$ -plane)

$$R_z(\theta) \begin{bmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

**Notes on consecutive rotations:** Two consecutive rotations about different axes is **not** commutative, however if you rotate about the same axis it is.

### 6.1.6 Linear transformations, surjective, injective, bijective. Invertibility, and basic uses of determinants

- Review of onto (surjective), one-to-one (injective), and invertible (bijective):

– **Onto (surjective):** A function is onto if every element in the codomain has at least one preimage in the domain. What is required is that every element in the codomain must be the image of at least one element in the domain. In other words, for every element  $y$  in the codomain, there exists at least one  $x$  in the domain such that  $f(x) = y$

To determine if a function is onto, we check whether every element in the codomain has at least one preimage in the domain. In terms of graphs, For a function to be onto, the graph must "hit" or cover every point in the codomain. This means there should be no "gaps" in the values of the function that leave some elements of the codomain unmapped.

**Note:** For a function to be surjective, the codomain must equal the domain. If  $f : A \rightarrow B$ , then  $\text{Range}(f) = B$  must be true. (See notes on codomain vs range below)

The whole codomain must be used, the whole codomain must be used, multiple members of the domain may map to the same element in the codomain.

– **One-to-one (injective):** An injective function, also called a one-to-one function, is a type of function that preserves uniqueness. More formally, a function  $f : A \rightarrow B$  is injective if and only if different inputs in the domain map to different outputs in the codomain. In other words:

$$f(x_1) = f(x_2) \implies x_1 = x_2$$

This means no two distinct elements from the domain can be mapped to the same element in the codomain. If two elements from the domain have the same image, the function is not injective.

**Example:**  $f(x) = 2x$  is injective because if  $f(x_1) = f(x_2)$ , it implies  $2x_1 = 2x_2$ , and hence  $x_1 = x_2$ .

**Horizontal Line Test:** For functions that are graphically represented, you can check if a function is injective using the horizontal line test. If any horizontal line intersects the graph of the function at most one point, then the function is injective.

– **Invertible (bijective):** A function is bijective if it is both surjective and injective.

- **Reason onto status:** To reason about onto (also called surjective) functions without using a graph, we can use logical and algebraic methods

For an onto function, we typically need to show that for any arbitrary  $y$  in the codomain, we can find at least one  $x$  in the domain such that  $f(x) = y$ .

**Example:** Consider  $f(x) = 2x + 1$  from  $\mathbb{R} \rightarrow \mathbb{R}$ . To show this function is onto, we take an arbitrary element  $y \in \mathbb{R}$  and solve for  $x$

$$\begin{aligned} y &= 2x + 1 \\ x &= \frac{y - 1}{2}. \end{aligned}$$

Thus, for any  $y \in \mathbb{R}$ , there is an  $x = \frac{y-1}{2} \in \mathbb{R}$  such that  $f(x) = y$ . Therefore,  $f(x) = 2x + 1$  is onto because we can find an  $x$  for every  $y \in \mathbb{R}$ .

- **onto, one-to-one, invertible example:** Consider  $f(x) = ax$  from  $\mathbb{R} \rightarrow \mathbb{R}$ .

If  $a = 0$  then we have

$$f(x) = 0 \quad \forall x.$$

This means that no matter what real number  $x$  you choose, the output is always 0.

For the function  $f : \mathbb{R} \rightarrow \mathbb{R}$  to be onto, every real number in the codomain  $\mathbb{R}$  (the set of all real numbers) must have a corresponding input from the domain  $\mathbb{R}$  that maps to it. However, with  $f(x) = 0$  for all  $x$ , the range of the function is just  $\{0\}$ . This means The only output the function can produce is 0. There are many real numbers in the codomain  $R$ , such as 1, -2, etc that cannot be reached by any input  $x \in \mathbb{R}$

If  $A = 0$ , then it is onto

$$y = ax \implies x = \frac{1}{a}y.$$

We can now take a look at 1-1. Again, if  $a = 0$ , the function is not one-to-one. As we saw before when  $a = 0$ , we have

$$f(x) = 0 \quad \forall x.$$

This means it is not always the case that  $f(x_1) = f(x_2) \implies x_1 = x_2$  When  $a = 0$ , we have:

$$f(x_1) = 0 \text{ and } f(x_2) = 0 \text{ for any } x_1, x_2 \in \mathbb{R}.$$

Thus,  $f(x_1) = f(x_2) = 0$ , but  $x_1$  and  $x_2$  can be different. This violates the injectivity condition because the same output (0) can come from different inputs.

When  $a \neq 0$ ,  $f(x) = ax$  is 1-1. We have

$$f(x_1) = f(x_2) \implies ax_1 = ax_2 \implies x_1 = x_2.$$

Thus, when  $a \neq 0$ , the function is invertible, and we can easily find the inverse

$$\begin{aligned} f(x) &= ax = y \implies y = ax \\ x &= ay \implies y = \frac{1}{a}x. \end{aligned}$$

Thus,

$$f^{-1}(x) = \frac{1}{a}x.$$

Furthermore, we can say

$$\begin{aligned} f_a^{-1} &: \mathbb{R} \rightarrow \mathbb{R} \\ f_a^{-1}(x) &= \frac{1}{a}x \\ f_a^{-1} &= f_{\frac{1}{a}} \\ f_1^{-1} &= f_1. \end{aligned}$$

- **Codomain vs range:**

- **Codomain:** The set of all possible outputs that a function is allowed to map to, as defined when the function is created. It includes every value the function could theoretically output, whether or not it actually does. For  $f : A \rightarrow B$ ,  $B$  is the codomain.
- **Range (Image):** The set of all actual outputs that the function produces when applied to elements of the domain. The range is a subset of the codomain, containing only the values the function actually maps to.

- **Another example of onto, 1-1, etc..:** Let's take a look at  $f(x) = x^2$  from  $\mathbb{R}$  to  $\mathbb{R}$ .

- **Check surjective:**

$$\begin{aligned} f(x) = x^2 &\implies y = x^2 \\ &\implies x = \pm\sqrt{y}. \end{aligned}$$

Since  $y$  can't be negative (and we know negative numbers are in the codomain  $\mathbb{R}$ ), it is not onto. To fix this, we would need to redefine the map as  $f : \mathbb{R} \rightarrow [0, \infty)$

- **Check injective:**

$$\begin{aligned} f(x_1) = f(x_2) &\implies x_1 = x_2 \\ &\implies x_1^2 = x_2^2 \\ &\implies x_1 = \pm\sqrt{x_2^2} \\ &\implies x_1 = \pm|x_2| \\ &\implies x_1 = \pm x_2. \end{aligned}$$

To fix this, we would need to restrict the domain and not allow negative numbers, this would mean  $x_1 = x_2$ . Thus, to make this function both surjective and injective (and thus invertible), the map would need to be  $f : [0, \infty) \rightarrow [0, \infty)$ . And

$$f^{-1}(x) = \sqrt{x}.$$

- **Injective, surjective, but  $\mathbb{R} \rightarrow \mathbb{R}^2$ :** Let's define  $L : \mathbb{R} \rightarrow \mathbb{R}^2$ ,  $L(x) = (ax, bx)$ . Then choose  $(\alpha, \beta) \in \mathbb{R}^2$ . We then have

$$\begin{aligned} ax = \alpha &\implies x = \frac{\alpha}{a} \\ bx = \beta &\implies x = \frac{\beta}{b} \\ &\implies \frac{\alpha}{a} = \frac{\beta}{b}. \end{aligned}$$

which implies that not all pairs  $(\alpha, \beta) \in \mathbb{R}^2$  can be reached. Only pairs that satisfy this condition can be mapped by  $L(x)$ . Thus, not onto.

What about 1-1? Well

$$\begin{aligned} L(x_1) = L(x_2) &\implies (ax_1, bx_1) = (ax_2, bx_2) \\ &\implies ax_1 = ax_2 \\ &\quad bx_1 = bx_2. \end{aligned}$$

Thus injective as long as both are not 0

- **Injective, surjective in general with maps:** In general

- $\mathbb{R}^n \rightarrow \mathbb{R}^k$  where  $n < k$  can never be onto, maybe 1-1 though

This type of map can never be surjective (onto), because a lower-dimensional space cannot fill a higher-dimensional space. For example, a line ( $\mathbb{R}^1$ ) can never cover the whole plane ( $\mathbb{R}^2$ )

However, it can be injective (1-1), because you can map different points in  $\mathbb{R}^n$  to distinct points in  $\mathbb{R}^k$ . For example, a line can injectively map into a plane without overlapping itself.

When  $n < k$ , you are mapping from a lower-dimensional space (like a line) into a higher-dimensional space (like a plane). There is "extra room" in the higher-dimensional space, so it's possible to place distinct points from the lower-dimensional space in such a way that they don't overlap in the higher-dimensional space.

- $\mathbb{R}^n \rightarrow \mathbb{R}^k$  where  $n > k$  can never be 1-1, maybe onto though

This type of map can be surjective (onto), because you can potentially cover all of  $\mathbb{R}^k$  by projecting parts of  $\mathbb{R}^n$  onto  $\mathbb{R}^k$ . For example, projecting a 3D space onto a 2D plane can cover the entire plane.

However, this type of map can never be injective (1-1), because different points in  $\mathbb{R}^n$  must collapse to the same point in  $\mathbb{R}^k$  (there's "not enough space" to map all distinct points in  $\mathbb{R}^n$  uniquely to  $\mathbb{R}^k$ ). For example, you cannot injectively map 3D space ( $\mathbb{R}^3$ ) into a 2D plane ( $\mathbb{R}^2$ ) without some overlap.

- **Injective, surjective for  $\mathbb{R}^2 \rightarrow \mathbb{R}^2$**

$$\begin{aligned} L : \mathbb{R}^2 &\rightarrow \mathbb{R}^2 \\ L \left( \begin{bmatrix} x \\ y \end{bmatrix} \right) &= \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}. \end{aligned}$$

If  $a, c = 0$

$$\begin{aligned} \implies by &= e \implies y = \frac{e}{b} \\ \implies dy &= f \implies y = \frac{f}{d} \end{aligned}$$

This expression is not likely to hold. For this to be subjective we would need to be able to produce all vectors in  $\mathbb{R}^2$ , the condition above means that  $e$  and  $f$  cannot be chosen freely. It also doesn't depend on  $x$  at all, this is trouble some. How are we expected to generate all vectors in  $\mathbb{R}^2$  if the transformation ignores one of the inputs.  $y$  alone isn't enough to generate all possible vectors in  $\mathbb{R}^2$ . Since the transformation only depends on  $y$ , the output vectors can only vary based on that one value. But to cover all vectors in  $\mathbb{R}^2$ , you need both  $x$  and  $y$  to contribute to the output.

- **Matrix as a linear transformation:** A matrix in linear algebra represents a transformation that acts on vectors. For example, if you have a  $2 \times 2$  matrix  $A$  and a vector  $v$ , multiplying the matrix by the vector gives you a new vector:

$$Av = v'.$$

This transformation can stretch, shrink, rotate, or reflect the vector  $v$ . The matrix encodes all the information about how vectors (and shapes) are transformed.

The determinant of a matrix  $A$  gives a scalar value that tells you how the transformation scales areas or volumes in the space:

**In 2D:** The determinant of a  $2 \times 2$  matrix tells you how the transformation changes the area of shapes.

- If  $\det(A) = 2$ , it means the transformation doubles the area of any shape in the 2D plane.
- If  $\det(A) = 0.5$ , the transformation shrinks the area to half its original size.
- If  $\det(A) = 0$ , the transformation collapses the area to zero (e.g., maps everything onto a line).

While it's true that a matrix transforms vectors, those vectors can be thought of as forming shapes in space. The determinant tells us how those shapes — formed by sets of vectors — are transformed by the matrix

**In 3D:** The determinant of a  $3 \times 3$  matrix tells you how the transformation changes the volume of shapes in 3D space.

- A non-zero determinant means the transformation stretches or shrinks volume, while a zero determinant collapses the volume to a lower dimension.

A matrix contains rows (or columns) that represent how the basis vectors of the space are transformed:

In  $\mathbb{R}^2$ , the columns of a  $2 \times 2$  matrix tell you where the unit vectors  $\mathbf{i}$  and  $\mathbf{j}$  (representing the x- and y-axes) are mapped after the transformation.

For example, if the matrix is:

$$A = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix},$$

It means the transformation stretches both the x-axis and y-axis by a factor of 2, doubling the area. The determinant of this matrix is  $2 \times 2 = 4$ , indicating the area is scaled by a factor of 4.

- **surjective and bijective in terms of matrix transformations:**

If a matrix  $A$  represents a linear transformation from  $\mathbb{R}^n \rightarrow \mathbb{R}^m$ , the transformation is onto if for every vector  $b \in \mathbb{R}^m$  in the codomain, there exists at least one vector  $x \in \mathbb{R}^n$  such that  $Ax = b$ .

**Geometric Interpretation:** Surjectivity means the matrix transformation "covers" the entire codomain, hitting every possible point. In 2D, this would mean the entire plane is covered by the transformation.

A matrix  $A$  is injective if for every  $x_1$  and  $x_2$  in the domain, if  $Ax_1 = Ax_2$ , then  $x_1 = x_2$ . This means that no two distinct input vectors can be mapped to the same output vector.

**Geometric Interpretation:** Injectivity means the transformation doesn't collapse any part of the domain into a lower dimension, so no information is lost.

- **If dimensions are the same:** If  $\mathbb{R}^n \rightarrow \mathbb{R}^n$ , if its onto, its most likely 1-1.
- **Rank of a matrix:** The rank of a matrix refers to the number of linearly independent rows or columns in the matrix. Essentially, the rank tells you the "dimension" of the space spanned by the rows or columns.
  - **Row rank:** The number of linearly independent rows.
  - **Column rank:** The number of linearly independent columns.

For any matrix, the row rank is always equal to the column rank. This common number is simply called the rank of the matrix.

- **Full Rank:** A matrix is said to have full rank if its rank is as large as possible, meaning that all of its rows or columns are linearly independent.

The rank of an  $m \times n$  matrix is bounded by

$$\text{Rank}(A) \leq \min(m, n).$$

If a matrix has full rank, it means the matrix is capable of fully transforming vectors in the space without collapsing dimensions.

If the matrix has full rank, it is invertible. This is because the matrix does not collapse any part of the vector space it transforms.

If the matrix has

$$\det(A) \neq 0.$$

It has full rank

When a matrix has full rank, it essentially means that the system of equations it represents is well-behaved, and every input gets mapped to exactly one output without overlap

When a matrix has full rank, it has as many independent equations as there are variables (in a square matrix case). This means that each variable (input) has its own effect on the output, and no two variables end up pointing to the same place.

### Full rank:

- **One-to-one mapping:** Full rank means there is no redundancy in the equations. If you solve this system, no two variables will point to the same result, because the equations are independent of each other. Each variable contributes uniquely to the solution.

### Rank-deficient

- **Overlapping outputs:** Some inputs are getting sent to the same output, because the system of equations isn't fully independent. This means you might not be able to tell which input caused which output (no unique solution).
- **Multiple or no solutions:** When rank is less than full, you might get cases where there are infinitely many solutions (because some inputs can lead to the same output) or no solution at all, depending on how things are set up.
- **Rank-Deficient:** A matrix is rank-deficient if its rank is less than the maximum possible rank. In other words, some of the rows or columns are linearly dependent (one can be expressed as a combination of others).

When a matrix is rank-deficient, it means the transformation it represents collapses some part of the space into a lower dimension.

- In  $\mathbb{R}^2$ , a rank-1 matrix would map all points onto a line, losing one dimension. In  $\mathbb{R}^3$ , a rank-2 matrix would map all points onto a plane, collapsing one dimension of the space.

If the matrix has

$$\det(A) = 0.$$

It is rank deficient

- **Singular matrix:** A rank-deficient square matrix is singular, meaning it cannot be inverted. This happens because there is a loss of information in the transformation — it collapses some dimensions, making it impossible to "go back" to the original input.
- **Determine rank of matrix:** To determine the rank of a matrix, we need to make sure the  $m \times n$  matrix has  $m$  linearly independent rows. To do this, we can perform Gaussian elimination to get the augmented matrix in row echelon form. The rank of the matrix,  $\text{rank}(A)$  is the number of non zero rows.
- **What does the rank tell us about solutions:** If we have full rank, then given a target, there will be a unique solution. If we are rank-deficient, there may be no solution, or infinitely many solutions.
- **Linear dependence:** Linear dependence tells us that we will lose information because it implies that some vectors (or columns/rows of a matrix) do not add any new, unique directions to the space. These dependent vectors can be expressed as combinations of other vectors, meaning they don't span new dimensions, and as a result, the transformation collapses part of the input space into a lower-dimensional output space.
- **Geometric interpretation of a linear transformation.** Suppose

$$\begin{aligned} & \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \\ &= \begin{bmatrix} ax + by \\ cx + dy \end{bmatrix} \\ &= x \begin{bmatrix} a \\ c \end{bmatrix} + y \begin{bmatrix} b \\ d \end{bmatrix}. \end{aligned}$$

This shows that multiplying a vector by a matrix transforms the vector into a new vector that is a linear combination of the matrix's columns. Geometrically, this implies that the action of the matrix on a vector can be viewed as scaling and rotating the vector in the direction of the matrix's column vectors.

- **Geometric interpretation of any vector:** Examine the result

$$\begin{aligned} & x \begin{bmatrix} 1 \\ 0 \end{bmatrix} + y \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} x \\ y \end{bmatrix}. \end{aligned}$$

The fact that any vector

$$\begin{bmatrix} x \\ y \end{bmatrix}$$

can be expressed as

$$x \begin{bmatrix} 1 \\ 0 \end{bmatrix} + y \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

means that the standard basis vectors

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

span the entire 2D plane ( $\mathbb{R}^2$ ). This implies that any point or vector in the 2D space can be reached by scaling and adding these two vectors.

$$\text{Suppose } L = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

And

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} e \\ f \end{bmatrix}.$$

Then by the observation above and the properties of linearity, we can say

$$\begin{aligned} & L \left( \begin{bmatrix} x \\ y \end{bmatrix} \right) \\ & L \left( x \begin{bmatrix} 1 \\ 0 \end{bmatrix} + y \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right) \\ &= L \left( x \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right) + L \left( y \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right) \\ &= xL \left( \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right) + yL \left( \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right). \end{aligned}$$

And we also know

$$\begin{aligned} L \left( \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right) &= \begin{bmatrix} a \\ c \end{bmatrix} \\ L \left( \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right) &= \begin{bmatrix} b \\ d \end{bmatrix}. \end{aligned}$$

- **Span:** The span of a set of vectors is the collection of all possible linear combinations of those vectors

if a matrix has full rank, the span of its columns (or rows) is the entire codomain (or the entire vector space that the matrix maps to)

- **Intro to row space, column space, null space (kernel):**

- **Column space ( $C(A)$ ):** The set of all linear combinations of the columns of a matrix  $A$ . It represents the range of the matrix and consists of all possible outputs of the matrix transformation. It's a subspace of  $\mathbb{R}^m$  (for an  $m \times n$  matrix)
- **Row space ( $R(A)$ ):** The set of all linear combinations of the rows of a matrix  $A$ . It is the span of the rows of the matrix and forms a subspace of  $\mathbb{R}^n$
- **Kernel/Null space ( $\ker(A)/N(A)$ ):** The set of all vectors  $\mathbf{x}$  such that  $A\mathbf{x} = 0$ , where  $A$  is a matrix. It represents the solutions to the homogeneous system and is a subspace of  $\mathbb{R}^n$

Each of these spaces relates to the structure and solvability of linear systems and the transformation properties of the matrix.

- **Linear combination of vectors:** A linear combination of vectors is a sum of those vectors, each multiplied by a scalar. So, for a matrix  $A$ , if  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$  are its column vectors, any vector in the column space can be written as:

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \dots + c_n\mathbf{v}_n.$$

Where  $c_1, c_2, \dots, c_n$  are scalars

As you vary the scalars in the linear combination of the matrix's columns, you generate all possible vectors in the column space (also known as the range) of the matrix.

**Note:** The same applies for the row space

- **Row space vs column space vs null space:**

- The column space tells you what the matrix outputs.
- The row space tells you about the constraints or conditions that the solutions to the matrix system must satisfy.
- The null space tells you what the matrix "loses". If a vector  $\mathbf{x}$  is in the null space, it gets mapped to the zero vector, meaning it is "annihilated" by the matrix. Vectors in the null space represent dependencies between the columns of the matrix. If the matrix has non-trivial solutions to  $A\mathbf{x} = 0$ , it indicates that the columns are linearly dependent.

Geometrically, the null space represents all the directions in which the matrix compresses space to a lower dimension. For example, in  $\mathbb{R}^3$ , if the null space is a line, the matrix compresses all points along that line to the origin.

If the null space is non-trivial, it indicates the matrix transformation has lost some dimensions.

**Note:** A non-trivial null space refers to a null space that contains vectors other than just the zero vector.

- **More on the row space:** In a system of linear equations, the row space reveals various types of constraints, depending on the number of independent equations and how the rows of the matrix relate to one another.

- **Unique Solution (Full Rank, Independent Rows):** If the rows of the matrix are linearly independent and span the entire row space, the system has a unique solution. This means the constraints from the equations are sufficient to pin down exactly one solution.
- **No Solution (Inconsistent System):**
- **Infinite Solutions (Dependent Rows, Underdetermined System):** If some rows are linearly dependent, the system will have fewer constraints than unknowns, leading to infinitely many solutions. In this case, the system is underdetermined, meaning there aren't enough independent constraints to specify a unique solution, allowing multiple solutions (often forming a plane, line, or higher-dimensional space).
- **Zero Solutions (Trivial System):** In a homogeneous system, if the rows are independent but fewer than the number of variables, there is only the trivial solution. This means that the row space spans a subspace of dimension less than the total space, so the only solution is the zero vector.
- **Linear dependence in rows vs columns:** The rows of a matrix are linearly dependent if at least one row can be written as a linear combination of the other rows.

This means there is redundancy in the information that the rows provide.

If the rows of a matrix are linearly dependent, the row space has a lower dimension than the number of rows, meaning the system has fewer independent constraints than it might appear.

Consider the matrix

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

Here, the second row is  $2 \times$  the first row, and the third row is  $3 \times$  the first row. Therefore, the rows are linearly dependent because each row is a scalar multiple of the first row.

This means there's only one independent constraint in the row space. The row space is spanned by a single vector (the first row), even though the matrix has three rows. Geometrically, the row space collapses to a lower dimension (a line in 3D space).

If the rows are linearly dependent, the system of equations might be underdetermined, leading to infinite solutions or no solutions.

The columns of a matrix are linearly dependent if at least one column can be written as a linear combination of the other columns.

This implies that some columns do not contribute "new" information, and there is a loss of dimensionality in the column space.

If the columns are linearly dependent, the column space has a lower dimension than the number of columns, meaning the matrix cannot map onto all of  $\mathbb{R}^m$  (the output space).

Consider the matrix

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

Here, the second and third columns are linearly dependent on the first column (they are multiples of the same vector). In fact, each column is identical in this case, so all the columns are linearly dependent.

This means the column space of this matrix is spanned by a single vector, even though there are three columns.

Geometrically, the matrix can only map vectors to a line in  $\mathbb{R}^3$ , rather than a full plane or space.

### Summary:

- **Row dependence** affects the row space, which corresponds to the constraints on the system of equations. Linearly dependent rows mean some equations are redundant, reducing the number of independent constraints.
- **Column dependence** affects the column space, which corresponds to the set of possible outputs of the matrix. Linearly dependent columns mean the matrix has reduced rank, implying the system might not span the full space, and it could have a non-trivial null space (leading to infinite or no solutions).

**Recall:** To check if a linear map has full rank, it is sufficient to check whether all the columns of the matrix representing the linear map are linearly independent

**Important:** The dimension of both the row space and the column space of a matrix is equal to the number of linearly independent rows and linearly independent columns, respectively. This common dimension is called the rank of the matrix.

- **Revisiting rank:** A matrix has full rank if its rank (the dimension of the column space or row space) is equal to the smaller of the number of rows or columns.

For an  $m \times n$  matrix

- If  $m \leq n$  (more rows than columns), the matrix has full rank if its rank is  $m$  (the number of rows).
- If  $n \leq m$  (more columns than rows), the matrix has full rank if its rank is  $n$  (the number of columns).

For a matrix to have full rank, the following must hold:

- The column space must have the maximum possible dimension, meaning the columns must be linearly independent

If the columns are linearly independent, the matrix has full column rank, and the rank of the matrix is equal to the number of columns,  $n$

The rank of a matrix is defined as the dimension of the column space (or row space, as they are equal).

For an  $m \times n$  matrix, where  $m > n$  (more rows than columns), the rank can be at most  $n$ , the number of columns. In other words, the rank of the matrix is limited by the number of columns, not rows.

**Important:** When a matrix (or linear map) has full rank, it means the mapping does not "squash" or lose any dimensions.

- **Rank of the null space:** The null space of a matrix has dimension  $n - \text{rank}$ , where  $n$  is the number of columns in the matrix. This is a consequence of the Rank-Nullity Theorem (Not yet stated).
- **Nullity:** The nullity of a matrix is the dimension of the null space (i.e., the number of independent vectors that get mapped to the zero vector by the matrix).
- **Zero nullity:** If the nullity is zero, this means that the null space has dimension 0. This implies that the only vector in the null space is the zero vector itself,  $\mathbf{0}$ .
- **non-zero nullity:** We know that there are infinitely many vectors in the kernel (null space) of a matrix when the nullity (dimension of the null space) is greater than zero.
- **Check if linear map is injective or surjective (Formal):** To check whether a linear transformation is surjective or injective, we use specific properties of the matrix representing the transformation.

Let  $T : V \rightarrow W$  be a linear transformation, where  $V$  and  $W$  are vector spaces and  $A$  is the matrix representation of  $T$

- **Checking Injectivity (One-to-One):** A linear transformation  $T$  is injective (one-to-one) if:

$$T(\mathbf{v}_1) = T(\mathbf{v}_2) \implies \mathbf{v}_1 = \mathbf{v}_2.$$

Equivalently,  $T$  is injective if the only solution to  $T(\mathbf{v}) = \mathbf{0}$  (the null space or kernel of  $T$ ) is  $\mathbf{v} = \mathbf{0}$ .

Alternatively, a matrix  $A$  is injective if the rank of the matrix (the number of linearly independent columns) is equal to the number of columns of the matrix. This means the matrix has full column rank

Lastly, we can just check the determinant if the matrix is square.

Consider the linear transformation  $T : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ , represented by the matrix:

$$A = \begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix}.$$

This matrix defines the linear transformation  $T(\mathbf{v}) = A\mathbf{v}$ . We want to check whether this transformation is injective, meaning that:

$$T(\mathbf{v}_1) = T(\mathbf{v}_2) \implies \mathbf{v}_1 = \mathbf{v}_2.$$

In other words, if two vectors  $\mathbf{v}_1$  and  $\mathbf{v}_2$  are mapped to the same vector by  $T$ , they must be the same vector.

Assume  $T(\mathbf{v}_1) = T(\mathbf{v}_2)$ , which implies:

$$A\mathbf{v}_1 = A\mathbf{v}_2.$$

This can be rewritten as:

$$A(\mathbf{v}_1 - \mathbf{v}_2) = 0.$$

Let  $\mathbf{w} = \mathbf{v}_1 - \mathbf{v}_2$ , so the equation becomes:

$$A\mathbf{w} = 0.$$

This means that  $\mathbf{w}$  is in the kernel (null space) of  $A$ .

To check if the kernel of  $A$  contains any non-zero vectors, solve the equation:

$$\begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

This gives the system of equations:

$$2w_1 + w_2 = 0,$$

$$w_2 = 0.$$

From the second equation, we have  $w_2 = 0$ . Substituting this into the first equation gives:

$$2w_1 = 0 \implies w_1 = 0.$$

Thus,  $\mathbf{w} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ , meaning that the only solution is the zero vector.

Since  $\mathbf{w} = 0$ , we have  $\mathbf{v}_1 - \mathbf{v}_2 = 0$ , which implies  $\mathbf{v}_1 = \mathbf{v}_2$ .

Thus,  $T(\mathbf{v}_1) = T(\mathbf{v}_2) \implies \mathbf{v}_1 = \mathbf{v}_2$ , and therefore, the transformation  $T$  is injective. There are no non-zero vectors in the kernel of  $A$ , so the transformation does not collapse any vectors together.

- **Check surjectivity:** A linear transformation  $T$  is surjective (onto) if for every vector  $\mathbf{w} \in W$  (the codomain), there exists a vector  $\mathbf{v} \in V$  (the domain) such that:

$$T(\mathbf{v}) = \mathbf{w}.$$

This means that  $T$  "covers" the entire codomain  $W$ , or in other words, the image of  $T$  is the entire space  $W$ .

#### Steps to check surjectivity:

1. **Image:** The transformation  $T$  is surjective if the image (column space or range) of the matrix  $A$  spans the entire codomain  $W$ . This means

$$\text{im}(A) = W.$$

2. **Rank:** For surjectivity, the rank of the matrix must be equal to the dimension of the codomain. If  $A$  is an  $m \times n$  matrix, the matrix is surjective if its rank is equal to  $m$  (the number of rows).

3. **Determinant for square matrices:** If  $A$  is a square matrix,  $T$  is surjective if

$$\det(A) \neq 0,$$

because a non-zero determinant implies the matrix has full rank, covering the entire codomain.

- **Find a linear map  $L : \mathbb{R}^3 \rightarrow \mathbb{R}^2$  that is not onto:** We know that a map from a higher dimension to a lower dimension cannot be injective. However, it is possible for it to be surjective (onto). One way we would formulate such a matrix is make one that is rank deficient. Such as

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

For this map to be onto, it would need to be rank 2. However, we see it is only rank 1. Thus, not onto.

Alternately we could use the concept of onto. For the map to be onto every vector in the codomain would need to be hit. Thus, we design the map in such a way that it is impossible for all targets to be hit by the transformation.

We could simply ask what map gives

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} \mapsto \begin{bmatrix} x \\ 0 \end{bmatrix}.$$

The choice of mapping to  $\begin{bmatrix} x \\ 0 \end{bmatrix}$  is somewhat arbitrary, but it is a good choice for designing a map that is not onto. Thus, the matrix would be

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

- **Check if a map  $L : \mathbb{R}^3 \rightarrow \mathbb{R}^2$  is onto:** Suppose we had the map

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

This implies

$$\begin{aligned} \begin{bmatrix} 1 & 0 & 2 \\ 3 & -1 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} &= \begin{bmatrix} u \\ v \end{bmatrix} \\ \implies \begin{bmatrix} x + 2z \\ 3x - y + z \end{bmatrix} &= \begin{bmatrix} u \\ v \end{bmatrix}. \end{aligned}$$

Where  $\begin{bmatrix} u \\ v \end{bmatrix}$  is the target vector, and  $u, v \in \mathbb{R}$

If this map were onto, then choose any arbitrary  $u, v \in \mathbb{R}$ , we would be able to hit that target with some  $x, y, z \in \mathbb{R}$

If we let  $z = 0$ , then

$$\begin{aligned} x &= u \\ y &= 3u - v \\ z &= 0. \end{aligned}$$

Implies

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} u \\ 3u - v \\ 0 \end{bmatrix}.$$

Thus, no matter the choice of  $u, v$  (target), we can always find a preimage in the domain.

Also, we can show that its not injective, let  $z$  be free

$$\begin{aligned} x + 2z &= u \implies x = u - 2z \\ 3x - y + z &= v \implies 3(u - 2z) - y + z = v \\ \implies 3u - 6z - y + z &= v \\ \implies y &= 3u - v - 5z. \end{aligned}$$

Thus, for any arbitrary choose of the target vector,

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} u - 2z \\ 3u - v - 5z \\ z \end{bmatrix}.$$

Thus, there are infinitely many input vectors that map to the same output (target) vector.

- **Check if a map  $L : \mathbb{R}^3 \rightarrow \mathbb{R}^2$  is onto or one-to-one** Suppose we had the map

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

The rank of the matrix is 2, meaning the image of the matrix spans a 2-dimensional subspace in  $\mathbb{R}^2$

Since the codomain is also  $\mathbb{R}^2$ , and the rank is 2 (meaning the image is 2-dimensional and fills  $\mathbb{R}^2$ ), the transformation is onto.

A matrix is one-to-one (injective) if different input vectors map to different output vectors, which means there are no nonzero vectors in the kernel (the set of vectors that map to zero).

The matrix is  $2 \times 3$ , meaning it maps from a higher-dimensional space ( $\mathbb{R}^3$ ) to a lower-dimensional space ( $\mathbb{R}^2$ ).

Since the matrix has rank 2, the null space (or kernel) of the matrix has dimension  $3 - 2 = 1$ . This means there is a one-dimensional subspace of vectors in  $\mathbb{R}^3$  that are mapped to the zero vector in  $\mathbb{R}^2$ .

In other words, there are infinitely many vectors in  $\mathbb{R}^3$  (those that lie in the kernel) that are mapped to the same output (the zero vector), which shows that the matrix is not injective.

### 6.1.7 The inverse of a square matrix, computation of determinants

- **Matrix invertibility:** Suppose we have

$$\begin{aligned} ax + by &= e \\ cx + dy &= f. \end{aligned}$$

Then we have the augmented matrix

$$\left[ \begin{array}{cc|c} a & b & e \\ c & d & f \end{array} \right].$$

Solving this system (assuming a nice system, no troublesome dividing by zeros). We get

$$\left[ \begin{array}{cc|c} 1 & 0 & \frac{de-bf}{ad-bc} \\ 0 & 1 & \frac{af-ce}{ad-bc} \end{array} \right].$$

Which implies

$$\begin{aligned} x &= \frac{de - bf}{ad - bc} \\ y &= \frac{af - ce}{ad - bc}. \end{aligned}$$

As long as the determinant  $ad - bc \neq 0$ , we have unique solutions for all  $(e, f) \in \mathbb{R}^2$ . Thus, onto and one-to-one. If  $ad - bc = 0$  the matrix is not invertible, and the system may have no solution or infinitely many solutions.

- **Identity matrix:** An identity matrix is a square matrix in which all the elements of principal diagonals are one, and all other elements are zero. It is denoted by  $I_n$  or simply  $I$ .
- **Inverse matrix:** Recall for a function  $f : A \rightarrow B$  if  $f$  is invertible then it has an inverse  $f^{-1} : B \rightarrow A$ , and  $(f^{-1} \circ f)(a) = f^{-1}(f(a)) = a$ ,  $(f \circ f^{-1})(b) = f(f^{-1}(b)) = b$  (it goes both ways).

For a matrix we would like the same thing. In the section below, we saw that for a matrix  $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ ,

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

If we call the matrix on the left side  $B$ , and define  $A : \mathbb{R}^2 \rightarrow \mathbb{R}^2 \implies A : \vec{v} \rightarrow \vec{w}$ , and  $B : \mathbb{R}^2 \rightarrow \mathbb{R}^2$

Then we can show that  $B$  is actually the inverse of  $A$ ,  $A^{-1}$  if  $(B \circ A)(\vec{v}) = \vec{v}$ , and  $(A \circ B)(\vec{w}) = \vec{w}$

This exercise is trivial and we will just assert that it is true.

Recall that a matrix is only invertible if  $\det(A) = |A| \neq 0$

**Note:** Only square matrices are invertible

- **Determine invertibility for  $\mathbb{R}^2 \rightarrow \mathbb{R}^2$ :** The determinant of a matrix plays a key role in determining whether a matrix is invertible (has an inverse) because it gives us crucial information about the transformation the matrix represents

For a matrix  $A$ , the determinant essentially tells us how the matrix scales areas or volumes during a transformation:

For a  $2 \times 2$  matrix acting on vectors in  $\mathbb{R}^2$ , the determinant measures how the matrix scales the area of any region in 2D space.

If the determinant is zero, the matrix collapses the area (or volume in higher dimensions) to zero, meaning the transformation is singular — it squashes space in some way, making it impossible to reverse the transformation.

A zero determinant means that the matrix does not have full rank (it's rank-deficient), meaning its rows or columns are linearly dependent. Linear dependence implies that some rows or columns can be written as combinations of others, meaning the matrix cannot span the entire space it operates in. This makes it impossible to uniquely solve the system  $A\vec{x} = \vec{b}$  for  $x$ , which is a key requirement for invertibility

If the determinant of a matrix is non-zero, it means the matrix is invertible because the transformation is non-degenerate — it does not collapse space, so it can be reversed.

The matrix has full rank (its rows or columns are linearly independent), meaning it spans the entire space and can uniquely map input vectors to output vectors.

**Note:** A nonzero determinant indicates that the matrix has full rank, but this applies specifically to square matrices

The determinant is defined only for square matrices

- **Non-square matrix and the nature of its transformation:** while we can't use the determinant, we can still obtain valuable information about the nature of the linear transformation using other concepts. Mainly the technique described above where we get the matrix in row echelon form and count the non-zero rows.
- **More on the determinant:** We can think of the determinant as a map. For a  $2 \times 2$  square matrix,

$$(a, b, c, d) \mapsto ad - bc.$$

We see that this is **not** linear. Thus,

$$\det(sA) \neq s\det(A).$$

In fact, we find

$$\det(sA) = s^n \det(A).$$

for a  $n \times n$  square matrix, where  $n$  is the number of rows and columns.

- **Left and right inverse: Left Inverse:** A matrix  $B$  is a left inverse of  $A$  if  $BA = I$ , where  $I$  is the identity matrix. This means  $B$  "undoes"  $A$  when multiplied from the left.

**Right Inverse:** A matrix  $C$  is a right inverse of  $A$  if  $AC = I$ . This means  $C$  "undoes"  $A$  when multiplied from the right.

- (**Finite dimensions**) **One sided inverse theorem for matrices (inveribility criterion):** It states that for square matrices in finite dimensions, the existence of either a left inverse or a right inverse implies the existence of the other, and hence the matrix is invertible. Specifically, if a square matrix  $A$  has a left inverse  $B$  (such that  $BA = I$ ) or a right inverse  $C$  (such that  $AC = I$ ), then both inverses must be equal, making  $A$  invertible with the unique inverse  $A^{-1} = B = C$ .
- (**Finite dimensions**) **Inverse Uniqueness Theorem for matrices:** If  $A$  is an invertible  $n \times n$  matrix, then there is exactly one matrix  $B$  such that  $AB = BA = I$ . This unique matrix is called the inverse of  $A$  and is denoted by  $A^{-1}$ .
- **Scheme to get inverse of any square matrix (Guassian elimination):** Let's take the  $2 \times 2$  matrix. We know

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

We know that we need one such matrix such that

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x & t \\ y & z \end{bmatrix} = I.$$

For now, let's only look at the first column of the RHS matrix

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

We can easily solve this using guassian elimination, but how can we use these ideas to solve the whole system. Well, whatever we do during guassian elimination only affects the LHS constants, and despite what the RHS is, the steps on the left remain the same. Thus

$$\begin{array}{cc|cc} a & b & 1 & 0 \\ c & d & 0 & 1 \end{array}.$$

We can think of this as solving two systems at the same time, it would be like solving

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

and

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} t \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Separately. Thus, we will have the form

$$A \mid I .$$

Which yields

$$I \mid A^{-1} .$$

- **Minor of a matrix element  $M_{ij}$ :** The minor of an element  $a_{ij}$  is the determinant of the submatrix obtained by removing the  $i$ -th row and  $j$ -th column from  $A$ .
- **Principal minors:** The principal minors of a square matrix are the determinants of its leading submatrices. For an  $n \times n$  matrix, the  $k$ -th principal minor is the determinant of the  $k \times k$  upper-left submatrix, formed by selecting the first  $k$  rows and the first  $k$  columns of the original matrix.

For an  $n \times n$  matrix  $H$ , its principal minors are:

1. The first principal minor is the determinant of the top-left  $1 \times 1$  submatrix, which is just the first element of the matrix,  $H_{11}$ .

$$H_1 = \det(H_{11})$$

2. The second principal minor is the determinant of the top-left  $2 \times 2$  submatrix:

$$H_2 = \det \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix}$$

3. The third principal minor is the determinant of the top-left  $3 \times 3$  submatrix:

$$H_3 = \det \begin{pmatrix} H_{11} & H_{12} & H_{13} \\ H_{21} & H_{22} & H_{23} \\ H_{31} & H_{32} & H_{33} \end{pmatrix}$$

4. Continue this process until the  $n$ th principal minor, which is the determinant of the full matrix  $H$ .

- **Cofactor:** A cofactor of an element in a square matrix is a signed minor that is used in calculating the determinant of a matrix and in finding the matrix inverse.

The cofactor of the element  $a_{ij}$  is given by:

$$C_{ij} = (-1)^{i+j} M_{ij}.$$

The minor  $M_{ij}$  is simply the determinant of the  $(n - 1) \times (n - 1)$  matrix that remains after excluding the row and column of  $a_{ij}$ .

The sign factor alternates based on the position of the element  $a_{ij}$ . It follows a checkerboard pattern:

$$\begin{bmatrix} + & - & + & \dots \\ - & + & - & \dots \\ + & - & + & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

**Example:** For a  $3 \times 3$  matrix:

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix},$$

The cofactor of  $a_{11}$  is:

$$C_{11} = (-1)^{1+1} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} = \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix}.$$

The cofactor of  $a_{12}$  is:

$$C_{12} = (-1)^{1+2} \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} = - \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix}.$$

- **Determinant of any square matrix (laplace/cofactor expansion):** This method involves expanding the determinant along any row or column of the matrix. For an  $n \times n$  matrix  $A$ , the determinant  $\det(A)$  is defined recursively using smaller matrices called minors.

For a matrix  $A = [a_{ij}]$ :

$$\begin{aligned}\det(A) &= \sum_{j=1}^n (-1)^{i+j} a_{ij} \det(A_{ij}) \\ &= \sum_{j=1}^n a_{ij} C_{ij}.\end{aligned}$$

Or:

$$\begin{aligned}\det(A) &= \sum_{i=1}^n (-1)^{i+j} a_{ij} \det(A_{ij}) \\ &= \sum_{i=1}^n a_{ij} C_{ij}.\end{aligned}$$

where:

- $a_{ij}$  is the element in the  $i$ -th row and  $j$ -th column of  $A$ ,
- $(-1)^{i+j}$  is the sign factor (alternating signs in a checkerboard pattern),
- $A_{ij}$  is the  $(n - 1) \times (n - 1)$  submatrix obtained by removing the  $i$ -th row and  $j$ -th column from  $A$ .

You can perform the cofactor expansion along any row or column, often choosing the one with the most zeros to simplify calculations.

- **Determinant of the identity matrix:** The determinant of the identity matrix is one

$$\det(I) = 1.$$

- **Product of two determinants:** The product of two determinants  $\det(A)\det(B) = \det(AB)$ . This is particularly useful when we want to find the determinant of a product of two matrices, we can assert, for two matrices  $A$ , and  $B$

$$\det(AB) = \det(A)\det(B).$$

The verification for this is not stated here but is a simple exercise.

**Note:** This holds for any finite number of matrices, not just two

- **Determinant of the inverse matrix:** If we have some matrix  $A$ , which has an inverse  $A^{-1}$ , then

$$\begin{aligned}\det(AA^{-1}) &= \det(I) = 1 = \det(A)\det(A^{-1}) \\ \implies \det(A^{-1}) &= \frac{1}{\det(A)}.\end{aligned}$$

- **Another look at the matrix equation  $A\vec{x} = \vec{b}$ :** Suppose we have

$$\begin{cases} ax + by + cz = j \\ dx + ey + fz = k \\ gx + hy + iz = \ell \end{cases}.$$

Which can be written in matrix form as

$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} j \\ k \\ \ell \end{bmatrix}.$$

Or simply  $A\vec{x} = \vec{b}$ . Assume  $A$  is invertible and has an inverse  $A^{-1}$ . Then

$$A^{-1}(A\vec{x}) = A^{-1}\vec{b}.$$

Recall for a function  $f(x)$ , then

$$\begin{aligned} x &= y \\ \implies f(x) &= f(y). \end{aligned}$$

$A^{-1}$  is a linear map (function), so applying it to both sides we can expect the equality to hold. Furthermore

$$\begin{aligned} A^{-1}(A\vec{x}) &= A^{-1}\vec{b} \\ (A^{-1}A)\vec{x} &= A^{-1}\vec{b} \\ I\vec{x} &= A^{-1}\vec{b} \\ \therefore \vec{x} &= A^{-1}\vec{b}. \end{aligned}$$

Thus, if we know  $A$  is invertible and we know its inverse, we can solve for  $\vec{x}$  given any arbitrary  $\vec{b}$  using the inverse of  $A$

Moreover,

$$\vec{b} = \begin{bmatrix} j \\ k \\ \ell \end{bmatrix} = j \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + k \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + \ell \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

Then,

$$A^{-1}\vec{b} = A^{-1} \left( \begin{bmatrix} j \\ k \\ \ell \end{bmatrix} \right) = A^{-1} \left( j \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + k \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + \ell \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right).$$

Since we know  $A^{-1}$  is linear,

$$\begin{aligned} \vec{x} &= A^{-1}\vec{b} = A^{-1} \left( \begin{bmatrix} j \\ k \\ \ell \end{bmatrix} \right) = A^{-1} \left( j \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \right) + A^{-1} \left( k \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \right) + A^{-1} \left( \ell \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right) \\ &= jA^{-1} \left( \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \right) + kA^{-1} \left( \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \right) + \ell A^{-1} \left( \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right). \end{aligned}$$

**Example:** Suppose  $A = \begin{bmatrix} 1 & 2 & -2 \\ 1 & 3 & -1 \\ 0 & 1 & 2 \end{bmatrix}$ , then  $A^{-1} = \begin{bmatrix} 7 & -6 & 4 \\ -2 & 2 & -1 \\ 1 & -1 & 1 \end{bmatrix}$

let the output vector be  $\begin{bmatrix} 3 \\ 0 \\ -1 \end{bmatrix}$ . Then

$$\begin{aligned}\vec{x} &= \begin{bmatrix} 7 & -6 & 4 \\ -2 & 2 & -1 \\ 1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 0 \\ -1 \end{bmatrix} \\ &= \begin{bmatrix} 17 \\ -5 \\ 2 \end{bmatrix}.\end{aligned}$$

- **Matrix products and their inverse:** Suppose we have two matrices  $A$ , and  $B$ , we know we can find the product  $AB$ . It could happen that this product is invertible. We know

$$\det(AB) = \det(A)\det(B).$$

If  $AB$  is invertible, then  $\det(AB) \neq 0$ , which implies  $\det(A)$  and  $\det(B)$  cannot be zero, which implies they too must be invertible. Thus, if the product of matrices is invertible, then all factors must be invertible.

Furthermore, how could we find  $(AB)^{-1}$  intuition suggests it could be either

$$\begin{aligned}A^{-1}B^{-1} \\ \text{or } B^{-1}A^{-1}.\end{aligned}$$

If  $(AB)^{-1}(AB) = I$ , assume the second equation is the true one, then

$$\begin{aligned}(AB)^{-1}(AB) &= I \\ \implies (B^{-1}A^{-1})(AB) &= I \\ \implies B^{-1}(A^{-1}A)B &= I \\ \implies B^{-1}IB &= I \\ \implies B^{-1}B &= I \\ \implies I &= I.\end{aligned}$$

Thus, the second equation is verified, and we assert

$$(AB)^{-1} = B^{-1}A^{-1}.$$

**Note:** The first equation would not be verified because commutativity is not a guarantee like we have with associativity.

- **Upper and Lower triangles:** An  $n \times n$  matrix  $A = [a_{ij}]$  is called *upper triangular* if  $a_{ij} = 0$  for  $i > j$ .

It is called *lower triangular* if  $a_{ij} = 0$  for  $i < j$ . A diagonal matrix is both upper triangular and lower triangular.

The matrix

$$A = \begin{bmatrix} 1 & 3 & 3 \\ 0 & 3 & 5 \\ 0 & 0 & 2 \end{bmatrix}$$

is upper triangular, and

$$B = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 3 & 0 \\ 3 & 5 & 2 \end{bmatrix}$$

is lower triangular.

- **Determinant of upper and lower triangles:** Recall to compute a determinant, we use  $\det(A) = \sum_{i=1}^n a_{ij}(-1)^{i+j} \det(A_{ij}) = \sum_{j=1}^m a_{ij}(-1)^{i+j} \det(A_{ij})$ , where we can choose any arbitrary row or column to compute the determinant with, usually choosing the one with the most zeros to make things easy. For an upper triangle,

$$A = \begin{bmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{bmatrix}.$$

We can choose the first column to compute the determinant, since we have two zeros. Thus, the determinant expands to

$$\begin{aligned} \det(A) &= a(df - e(0)) - 0(bf - c(0)) + 0(be - cd) \\ &= adf. \end{aligned}$$

Thus, the determinant is the product of the main diagonal. The same logic applies to lower triangles.

- **Symmetric matrices:** A matrix  $A$  with real entries is called *symmetric* if  $A^T = A$ .

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 5 & 6 \end{bmatrix}$$

is a symmetric matrix.

- **Skew Symmetric matrices:** A matrix  $A$  with real entries is called *skew symmetric* if  $A^T = -A$ .

$$B = \begin{bmatrix} 0 & 2 & -3 \\ -2 & 0 & -4 \\ -3 & 4 & 0 \end{bmatrix}$$

is a skew symmetric matrix.

- **Determinant of the transpose:** We assert

$$\det(A^T) = \det(A).$$

For a  $2 \times 2$  matrix,

$$\begin{aligned} A &= \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad \det(A) = ad - bc \\ A^T &= \begin{pmatrix} a & c \\ b & d \end{pmatrix}, \quad \det(A^T) = ad - cb = ad - bc. \end{aligned}$$

Next, consider a  $3 \times 3$  matrix. Let  $A = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}$ , then  $A^T = \begin{pmatrix} a & d & g \\ b & e & h \\ c & f & i \end{pmatrix}$ . To find the determinants, we invoke the lapace expansion  $\det(A) = \sum_{j=1}^3 (-1)^{i+j} a_{ij} \det(A_{ij})$ , with the signs following  $S = \begin{pmatrix} + & - & + \\ - & + & - \\ + & - & + \end{pmatrix}$ . Notice that the transpose of the sign matrix is symmetric, in other words  $S^T = \begin{pmatrix} + & - & + \\ - & + & - \\ + & - & + \end{pmatrix}$ . To find the determinant of  $A$ , we expand over the first row

$$\det(A) = a(ei - fh) - b(di - fg) + c(dh - eg).$$

To find the determinant of the transpose, we notice that we can just expand over the first column, which is the first row in  $A$ , using the fact that the sign matrix does not change, the determinant will therefore be the same.

### 6.1.8 Eigenvectors, Eigenvalues

- **Revisiting square matrices, surjectivity, injectivity:** When you're working with square matrices the concepts of one-to-one (injectivity) and onto (surjectivity) are closely tied together:

For a square matrix  $A$ , if it is injective, it is automatically surjective, and vice versa. This is because the domain and codomain are the same size (both are  $\mathbb{R}^n$ ).

This means that for square matrices, you only need to check one condition (either injectivity or surjectivity) to determine if the matrix is invertible (which corresponds to having a non-zero determinant).

If a linear transformation is not injective, we can conclude the determinant must be zero.

If a linear transformation is not surjective, we can conclude the determinant must be zero.

- **Eigenvectors, Eigenvalues:** An **eigenvector** of a square matrix  $A$  is a non-zero vector  $\mathbf{v}$  such that when  $A$  acts on  $\mathbf{v}$ , the result is a scalar multiple of  $\mathbf{v}$ . Mathematically, this is written as:

$$A\mathbf{v} = \lambda\mathbf{v} \quad \mathbf{v} \neq \mathbf{0}$$

where  $\lambda$  is a scalar known as the eigenvalue corresponding to the eigenvector  $\mathbf{v}$ .

An **eigenvalue**  $\lambda$  is the scalar that represents the factor by which the eigenvector is scaled during the transformation. The eigenvalue corresponds to each eigenvector and provides information about the nature of the transformation (scaling, rotation, etc.).

Since the left side is matrix multiplication and the right side is vector multiplication by a scalar, we can rewrite the equation above as

$$\begin{aligned} A\mathbf{v} &= (\lambda I)\mathbf{v} \\ \implies A\mathbf{v} - \lambda I\mathbf{v} &= \mathbf{0} \\ \implies \mathbf{v}(A - \lambda I) &= \mathbf{0}. \end{aligned}$$

This is a homogeneous system. If the transformation map  $(A - \lambda I)$  is one-to-one and thus invertible, the only solution would be the trivial solution ( $\mathbf{v} = \mathbf{0}$ ). In order to have non-zero solutions for  $\mathbf{v}$  (eigenvectors), the system above would need to not be one-to-one (multiple solutions to the solution vector  $\mathbf{0}$ ), and thus

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

- **Characteristic equation, characteristic polynomial:** The characteristic equation of a square matrix  $A$  is the equation obtained by setting the determinant of  $A - \lambda I$  equal to zero:

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

where  $\lambda$  represents the eigenvalues of  $A$  and  $I$  is the identity matrix. Solving this equation gives the eigenvalues of  $A$ . The characteristic polynomial is the polynomial in  $\lambda$  obtained from the determinant  $\det(A - \lambda I)$ . It is typically expressed as:

$$p(\lambda) = \det(A - \lambda I),$$

and its roots are the eigenvalues of the matrix  $A$ .

- **Finding eigenvectors and eigenvalues:** To find the eigenvalues  $\lambda$ , we need to solve the characteristic equation:

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

This equation determines the values of  $\lambda$  for which the matrix  $A - \lambda I$  is singular (non-invertible), which leads to non-zero solutions for  $\mathbf{v}$ .

Once the eigenvalues  $\lambda$  are found, substitute each  $\lambda$  into the equation  $(A - \lambda I)\mathbf{v} = 0$  and solve for  $\mathbf{v}$ . These are the eigenvectors corresponding to each eigenvalue.

- **Characteristic polynomial** ( $2 \times 2$ ): The characteristic polynomial is

$$\det(A - \lambda I) = 0 \implies \lambda^2 - (a + d)\lambda + (ad - bc) = 0.$$

Recall the trace of a matrix  $A$  is  $\text{tr}(A) = a + d$ , or the sum of the diagonal. Thus,

$$\lambda^2 - \text{tr}(A)\lambda + \det(A) = 0.$$

Furthermore, we can use the quadratic formula to get an expression for  $\lambda$ .

$$\begin{aligned}\lambda &= \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \\ \implies \lambda &= \frac{a + d \pm \sqrt{(-(a + d))^2 - 4(ad - bc)}}{2}.\end{aligned}$$

Which, after some algebra, simplifies to

$$\lambda = \frac{a + d}{2} \pm \sqrt{\left(\frac{a - d}{2}\right)^2 + bc}.$$

Define  $D = \sqrt{\left(\frac{a+d}{2}\right)^2 + bc}$ . Then

$$\begin{cases} \text{No solution} & \text{if } D < 0 \\ \text{One solution} & \text{if } D = 0 \\ \text{Two solutions} & \text{if } D > 0 \end{cases}$$

### 6.1.9 Basis, change of basis, diagonalization

- **Basis:** A basis of a vector space is a set of linearly independent vectors that span the entire space. This means any vector in the space can be uniquely expressed as a linear combination of the basis vectors. A basis provides a reference framework for representing vectors in that space.

For an  $n$ -dimensional vector space, a basis will consist of exactly  $n$  vectors. The coordinates of a vector relative to a basis are the coefficients used in this linear combination.

In short, a basis defines the "building blocks" for all vectors in a vector space.

- **Standard basis (Implied basis):** From vector calculus, we know that  $\hat{i}, \hat{j}$  are the unit vectors that describe the 2 dimensional cartesian plane. Where

$$\begin{aligned}\hat{i} &= \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\ \hat{j} &= \begin{bmatrix} 0 \\ 1 \end{bmatrix}.\end{aligned}$$

Thus, the standard implied basis when working in  $\mathbb{R}^2$  is the matrix

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

And a vector  $\vec{v} = \begin{bmatrix} x \\ y \end{bmatrix}$  can be represented as scaling these basis vectors and then adding them. I.e

$$\begin{bmatrix} x \\ y \end{bmatrix} = x \begin{bmatrix} 1 \\ 0 \end{bmatrix} + y \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Thus, the standard basis for  $\mathbb{R}^n$  is the  $n \times n$  identity matrix  $I$

- **Basis notation:** When working with vectors, the choice of basis determines how we interpret where the vector sits. For the standard basis, the components of the vector are precisely where it will be. Thus, for the standard basis, we write

$$[\vec{v}].$$

If the basis were non-standard, we would need to specify it. We write

$$[\vec{v}]_B.$$

Where  $B$  is then defined as the matrix representing the basis.

- **Changing basis:** The goal of changing the basis is often to simplify computations or better understand the structure of the vector space or linear transformation. For example, in one basis, a matrix might be complex and difficult to work with, but in another basis, it might become diagonal or easier to handle.

**Example:** Suppose we have some vector in the standard basis

$$[\vec{v}] = \begin{bmatrix} 1 \\ 2 \end{bmatrix} = 1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + 2 \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Now suppose we define some new basis

$$B = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}.$$

To get the vector  $\vec{v} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$  in the new basis we compute  $B^{-1}\vec{v}$

Assuming  $B$  is invertible. Thus, this computation tells us what we need to scale the basis vectors before adding them to get the vector  $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$  in the standard basis.

Computing  $B\vec{v}$  tells us what the vector  $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$  in the new basis represents in the standard basis.

1.  $B^{-1}\vec{v}$ : Converts the vector  $\vec{v}$  from the standard basis to the new basis defined by  $B$ .
2.  $B\vec{v}$ : Converts the vector  $\vec{v}$  from the new basis back to the standard basis.

- **Change of basis explained:** We have

$$\begin{aligned} - B^{-1}v &= v_B \\ - Bv_B &= v \end{aligned}$$

Suppose we have a vector  $v = \begin{pmatrix} a \\ b \end{pmatrix}$  in the standard basis. Define a new basis  $b_1 = \begin{pmatrix} u \\ v \end{pmatrix}$ ,  $b_2 = \begin{pmatrix} x \\ y \end{pmatrix}$ , then  $B = \begin{pmatrix} u & x \\ v & y \end{pmatrix}$ . Where  $v$  under this basis becomes  $v_B = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$ . Then

$$\begin{aligned} \begin{pmatrix} a \\ b \end{pmatrix} &= \alpha \begin{pmatrix} u \\ v \end{pmatrix} + \beta \begin{pmatrix} x \\ y \end{pmatrix} \\ &= \begin{pmatrix} \alpha u + \beta x \\ \alpha v + \beta y \end{pmatrix} \\ &= \begin{pmatrix} u & x \\ v & y \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \\ \therefore v &= Bv_B. \end{aligned}$$

From this,

$$\begin{aligned} v &= Bv_B \\ B^{-1}v &= B^{-1}Bv_B \\ B^{-1}v &= Iv_B \\ \therefore B^{-1}v &= v_B. \end{aligned}$$

**Note:** The new basis must span the same space as the old basis for the change of basis to work correctly

- **Basis in maps:** Suppose we have some linear map

$$L : V \rightarrow W$$

where  $V$  and  $W$  are vector spaces. Suppose  $V$  has a basis  $\beta = \{b_1, \dots, b_n\}$ , and  $W$  has a basis  $\gamma = \{c_1, \dots, c_n\}$ . Then, the matrix representation of  $L$  with respect to these bases is denoted as

$$[L]_{\beta}^{\gamma}$$

Now, suppose we define new bases  $\beta'$  for  $V$  and  $\gamma'$  for  $W$ . We want to find the matrix representation of the linear map  $L$  in the new bases.

We have the following transformations:

$$V_{\beta'} \rightarrow W_{\gamma'} \quad \text{and} \quad V_{\beta} \rightarrow W_{\gamma}$$

To find the matrix of the map  $L : V_{\beta'} \rightarrow W_{\gamma'}$ , we need to relate the new basis vectors to the old basis vectors. Specifically, we perform the following steps:

- To change from the new basis  $\beta'$  to the old basis  $\beta$ , we multiply by the change-of-basis matrix  $B$ , so we have  $B[v]_{\beta'} = [v]_{\beta}$ .
- To change from the old basis  $\gamma$  to the new basis  $\gamma'$ , we multiply by the inverse of the change-of-basis matrix  $C^{-1}$ , so we have  $C^{-1}[w]_{\gamma} = [w]_{\gamma'}$ .

Thus, the matrix representation of  $L$  in the new bases  $\beta'$  and  $\gamma'$  is given by:

$$[L]_{\beta'}^{\gamma'} = C^{-1}[L]_{\beta}^{\gamma}B$$

- **Diagonalization and eigenbases:** We want to find some new basis  $B$  such that the linear map becomes diagonal. That is

$$B^{-1}LB = L_D.$$

We can achieve this via eigenvectors. By changing the basis to one formed by the eigenvectors, we simplify the linear map so that it acts independently on each direction (each eigenvector). In this new basis, the map scales each eigenvector by its corresponding eigenvalue, without mixing different directions. This independence is what makes the matrix diagonal, making the transformation much easier to understand and work with.

Consider a matrix  $A$  and a basis formed from its eigenvectors, say  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ . For simplicity, let's assume  $A$  has  $n$  linearly independent eigenvectors (which guarantees diagonalization).

We can express any vector  $\mathbf{x}$  in terms of this new basis as a linear combination of the eigenvectors

$$\mathbf{x} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n$$

When we apply the matrix  $A$  to  $\mathbf{x}$ , because each eigenvector  $\mathbf{v}_i$  satisfies  $A\mathbf{v}_i = \lambda_i\mathbf{v}_i$ , we get:

$$A\mathbf{x} = c_1\lambda_1\mathbf{v}_1 + c_2\lambda_2\mathbf{v}_2 + \cdots + c_n\lambda_n\mathbf{v}_n$$

This shows that  $A$  acts on each eigenvector individually by multiplying it by its corresponding eigenvalue. The action of  $A$  is now "separated" along each eigenvector direction.

To represent  $A$  in the new basis (the eigenvector basis), we express the transformation in matrix form with respect to this new basis. Let's call the matrix  $P$ , where the columns of  $P$  are the eigenvectors of  $A$

The matrix  $A$  in the original basis acts in a complicated way, but in the eigenvector basis, the transformation is simplified. Specifically, in the eigenvector basis, applying  $A$  scales each eigenvector by its corresponding eigenvalue. This means that, with respect to this basis, the matrix representing  $A$  becomes diagonal

$$\begin{aligned} P^{-1}AP &= D \\ \implies A &= PDP^{-1}. \end{aligned}$$

where  $D$  is a diagonal matrix with the eigenvalues  $\lambda_1, \lambda_2, \dots, \lambda_n$  on the diagonal:

$$D = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}$$

Consider a matrix  $\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$ . When this map acts on a vector  $\begin{pmatrix} x \\ y \end{pmatrix}$ ,

$$\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1x \\ 2y \end{pmatrix}.$$

Ie the vector is scaled by 1 in the  $x$  direction, and 2 in the  $y$  direction

- **Diagonalization example:** Suppose  $A = \begin{pmatrix} 4 & 1 \\ 2 & 3 \end{pmatrix}$ . Then

$$\begin{aligned} \lambda &= \frac{a+d}{2} \pm \sqrt{\left(\frac{a-d}{2}\right)^2 + bc} \\ \implies \lambda_1 &= 2 \\ \lambda_2 &= 5. \end{aligned}$$

For  $\lambda_1$ :

$$\begin{aligned} \begin{pmatrix} 4 & 1 \\ 2 & 3 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} &= 2 \begin{pmatrix} x \\ y \end{pmatrix} \\ \implies \begin{pmatrix} 4x+y \\ 2x+3y \end{pmatrix} &= \begin{pmatrix} 2x \\ 2y \end{pmatrix} \\ \implies 4x+y &= 2x \\ \implies y &= -2x. \end{aligned}$$

let  $x = 2$ , then  $v_1 = \begin{pmatrix} 2 \\ -4 \end{pmatrix}$

For  $\lambda_2$ :

$$\begin{aligned} 4x+y &= 5x \\ \implies y &= x. \end{aligned}$$

let  $x = 2$ , then  $v_2 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$ . Thus, one choice of basis is

$$B = \begin{pmatrix} 2 & 2 \\ -4 & 2 \end{pmatrix}.$$

Let's see what happens when we change the basis of  $A$  to  $B$

$$\begin{aligned} B^{-1}AB &= \frac{1}{12} \begin{pmatrix} 2 & -2 \\ 4 & 2 \end{pmatrix} \begin{pmatrix} 4 & 1 \\ 2 & 3 \end{pmatrix} \begin{pmatrix} 2 & 2 \\ -4 & 2 \end{pmatrix} \\ &= \begin{pmatrix} 2 & 0 \\ 0 & 5 \end{pmatrix} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} = D. \end{aligned}$$

- **Diagonalization formulas:** For a basis  $P$  composed of eigenvectors,

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

$$D = P^{-1}AP.$$

Where  $A$  is the original map, and  $D$  is the diagonalized map obtained by the second formula above. We saw in the previous example

$$D = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{pmatrix}.$$

- **Effect of changing basis on characteristic polynomial:** If changing the basis of  $A$  with a new basis  $B$  yields  $B^{-1}AB$ , then

$$\begin{aligned} \det(B^{-1}AB - \lambda I) &= 0 \\ \det(B^{-1}AB - \lambda B^{-1}B) &= 0 \\ \det(B^{-1}(AB - \lambda B)) &= 0 \\ \det(B^{-1}(A - \lambda I)B) &= 0 \\ \det(B^{-1}) \det(A - \lambda I) \det(B) &= 0 \\ \det(B^{-1}B) \det(A - \lambda I) &= 0 \\ \det(I) \det(A - \lambda I) &= 0 \\ \det(A - \lambda I) &= 0. \end{aligned}$$

Thus, no effect ■

- **Determinant after changing basis:** Suppose we have some matrix  $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ . If we introduce a new basis matrix  $B$ , then the map becomes  $B^{-1}AB$ . And the determinant is

$$\begin{aligned} \det(B^{-1}AB) &= \det(B^{-1}) \det(A) \det(B) \\ &= \det(B^{-1}B) \det(A) \\ &= \det(I) \det(A) \\ &= \det(A). \end{aligned}$$

Thus, the determinant remains the same. We say that the determinant is invariant under a change of basis.

- **Trace after changing basis:** First, we remark about a property of the trace. Specifically

$$\text{Tr}(XYZ) = \text{Tr}(YZX) = \text{Tr}(ZXY).$$

This property only holds for cyclic permutations, meaning you can rotate the matrices around but not arbitrarily rearrange them.

With this property in mind

$$\text{Tr}(B^{-1}AB) = \text{Tr}(ABB^{-1}) = \text{Tr}(A).$$

Thus, the trace is also invariant under change of basis.

- **Note about diagonalization and eigenbases:** Eigenbases (or eigenspaces) are typically defined only for linear maps of the form  $L : V \rightarrow V$ , where the map  $L$  acts on a vector space  $V$  and maps vectors within that same space.

If a linear map  $L$  is of the form  $L : V \rightarrow W$ , where  $W$  is a different vector space (or even a subspace of  $V$ ), the concept of eigenvectors and eigenvalues doesn't apply in the usual sense. The reason is that the output of  $L$  is not necessarily a scalar multiple of the input vector  $v$ , and it may not even belong to the same vector space.

For eigenvectors and eigenvalues to be meaningful, you need the transformation to act within a single vector space, ensuring that the transformed vector remains in the same space, allowing us to compare it directly to the original vector.

- **Diagonalization of a  $3 \times 3$  matrix:** Suppose we have the matrix

$$A = \begin{pmatrix} 0 & -6 & 4 \\ -1 & -1 & 2 \\ -2 & -8 & 7 \end{pmatrix}.$$

Then, if we have eigenvectors, we know

$$\begin{aligned} \det \begin{pmatrix} -\lambda & -6 & 4 \\ -1 & -1 - \lambda & 2 \\ -2 & -8 & 7 - \lambda \end{pmatrix} &= 0 \\ \implies -\lambda((-1 - \lambda)(7 - \lambda) - 2(-8)) \\ + 1(-6(7 - \lambda) - 4(-8)) - 2(-6(2) - 4(-8)) &= 0 \\ \implies -\lambda^3 + 6\lambda^2 - 11\lambda + 6 &= 0. \end{aligned}$$

We could try guessing a solution, say  $\lambda = 1$ ,  $-1 + 6 - 11 + 6 = 0 \implies 0 = 0$ . Thus,  $(\lambda - 1)$  is a factor. We use this to reverse engineer the remaining quadratic. We get

$$(\lambda - 1)(-\lambda^2 + 5\lambda - 6).$$

Implies the roots are 1, 2, 3. Thus, we have three distinct eigenvalues, which means we must have three linearly independent eigenvectors. We can then use these eigenvalues to get the vectors that serve as an eigenbases, then use

$$D = B^{-1}AB.$$

To find the diagonalized matrix.

### 6.1.10 Vector spaces, Abstract Vector Spaces, Subspaces

- **Vector Space:** A vector space is a set of vectors that satisfy

1. Space has a zero vector
2. Closed under addition
3. Closed under multiplication by a scalar

We also need the space to have the algebraic axioms.

- **Commutativity of addition:**  $u + v = v + u$
- **Associativity of addition:**  $(u + v) + w = u + (v + w)$
- **Additive identity:** There exists a vector  $0 \in V$  such that  $v + 0 = v$  for all  $v \in V$ .
- **Additive inverse:** For every  $v \in V$ , there exists a vector  $-v \in V$  such that  $v + (-v) = 0$ .
- **Associativity of scalar multiplication:**  $c(dv) = (cd)v$
- **Distributivity of scalar multiplication over vector addition:**  $c(u + v) = cu + cv$
- **Distributivity of scalar multiplication over scalar addition:**  $(c + d)v = cv + dv$
- **Scalar identity:**  $1v = v$ , where 1 is the multiplicative identity in the field.

If we have these properties and algebraic axioms, we have a valid vector space.

- **Abstract vector space:** An abstract vector space is a generalization of this concept, where the elements (vectors) may not have a concrete geometric form, such as functions, polynomials, or matrices, but still follow the same axioms

For example, A set of matrices can be defined as an abstract vector space

- **Basis of a vector space, span of the basis:** The basis of a vector space is a choice of  $n$  vectors  $b_1, \dots, b_n$  such that

$$\mathbf{v} = s_1b_1 + \dots + s_nb_n.$$

If we are able to generate all vectors in the space by simple scaling the vectors by some constant and adding them, the basis  $b_1, \dots, b_n$  are said to **span** the vector space.

**Example:**

$$\mathbf{v} = (x, y) = x(1, 0) + y(0, 1).$$

Where the basis in this case is  $b_1 = (1, 0)$ ,  $b_2 = (0, 1)$ ,  $s_1 = x$ , and  $s_2 = y$

- **Number of basis vectors for the space:** In a simple vector space  $\mathbb{R}^n$ , the number of basis vectors is  $n$
- **Basis vectors are not unique:** Basis vectors are not unique for a vector space because a vector space can have multiple sets of linearly independent vectors that span the same space. As long as the vectors satisfy the conditions of being linearly independent and spanning the entire space, different sets of vectors can serve as bases for the same vector space.

- **Basis example:** Suppose in  $\mathbb{R}^2$  we have the basis vectors  $(1, 1), (1, -1)$ , then

$$(x, y) = s_1(1, 1) + s_2(1, -1).$$

Then

$$\begin{aligned} x &= s_1 + s_2 \\ y &= s_1 - s_2. \end{aligned}$$

Solving this linear system yields

$$\begin{aligned} s_1 &= \frac{x+y}{2} \\ s_2 &= \frac{x-y}{2}. \end{aligned}$$

This expresses  $s_1$  and  $s_2$  in terms of the original coordinates  $(x, y)$ , showing how the coordinates in the new basis relate to the standard basis in  $\mathbb{R}^2$ .

- **Non-valid basis:** If the basis were non-valid, meaning the vectors are linearly dependent, the set would no longer qualify as a basis because the vectors would fail to span the space or be linearly independent.
  - **Linear Dependence:** If the two vectors in the supposed "basis" are linearly dependent (for example, if one vector is a scalar multiple of the other), the set no longer forms a valid basis because the vectors do not cover the full space. This would result in an inability to represent all points in  $\mathbb{R}^2$  using the given vectors.
  - **No Unique Solutions:** In the example above, solving for  $s_1$  and  $s_2$  involves solving a linear system of equations. If the basis is invalid due to linear dependence, this system becomes underdetermined, meaning there are infinitely many solutions or no solutions at all, rather than unique solutions for  $s_1$  and  $s_2$ .
  - **Inability to Express All Vectors:** A valid basis allows every vector in  $\mathbb{R}^2$  to be uniquely represented as a linear combination of the basis vectors. In the case of a non-valid basis, some vectors in  $\mathbb{R}^2$  cannot be expressed at all by the given vectors, or they could be represented by more than one combination, violating the uniqueness property of a basis.
- **Redundant basis vectors:** It may be the case that some basis vectors are redundant, for example if we choose (or are given) three basis vectors for  $\mathbb{R}^2$ .
- **Throwing out redundant basis vectors:** We must discard the redundant basis vectors. Suppose in  $\mathbb{R}^2$  we have the basis  $b_1 = (1, 1), b_2 = (1, -1), b_3 = (-1, 1)$ . Then

$$(x, y) = s_1(1, 1) + s_2(1, -1) + s_3(-1, 1).$$

Since we have three basis vectors in  $\mathbb{R}^2$ , we believe one should be redundant. Assume its  $(-1, 1)$ , then

$$(-1, 1) = s_1(1, 1) + s_2(1, -1).$$

Setting up this equation allows us to check for linear dependence.

$$\begin{aligned} (-1, 1) &= s_1(1, 1) + s_2(1, -1) \\ \implies -1 &= s_1 + s_2 \\ \implies 1 &= s_1 - s_2. \end{aligned}$$

Solving this system gives

$$\begin{aligned} s_1 &= 0 \\ s_2 &= -1. \end{aligned}$$

Thus,

$$(-1, 1) = 0(1, 1) + (-1)(1, -1).$$

Which implies the basis vector  $(-1, 1)$  is actually a linear combination of the first two (and linearly dependent), and we don't actually need it. The first two gives us all the information we need to uniquely express all vectors in the space.

- **The number of linearly independent vectors per space:** In  $\mathbb{R}^n$  the maximum number of linearly independent vectors we can have is  $n$ . For example, in  $\mathbb{R}^2$ , the maximum number of linearly independent vectors we can have is 2. This is why we need exactly  $n$  vectors to form a basis in  $\mathbb{R}^n$ , and having more than  $n$  will also result in the case of allowing us to find and throw out the linearly dependent ones.

In other words, There are only  $n$  linearly independent vectors in  $\mathbb{R}^n$  because the dimension of  $\mathbb{R}^n$  is  $n$ , which means that the space has exactly  $n$  independent directions, or degrees of freedom

- **Definition of dimension:** The dimension of a vector space is the number of vectors in a basis for that space. A basis is a set of linearly independent vectors that spans the entire space. In  $\mathbb{R}^n$ , any valid basis must have exactly  $n$  vectors, because it takes  $n$  vectors to fully describe the space.

A set of vectors is linearly independent if no vector in the set can be written as a linear combination of the others. In  $\mathbb{R}^n$ , if you have more than  $n$  vectors, at least one of those vectors can always be written as a linear combination of the others, meaning they will be linearly dependent. This is because there are only  $n$  independent directions in  $\mathbb{R}^n$

In  $\mathbb{R}^2$ , the dimension of the space is 2, meaning that any valid set of linearly independent vectors can have at most two vectors. This is because two vectors are sufficient to fully describe the space—they form a basis. Any other vector in  $\mathbb{R}^2$  can be expressed as a linear combination of these two vectors.

Once you have two linearly independent vectors, adding any third vector will result in linear dependence, because that third vector will lie in the span of the first two vectors.

- **Discern valid basis:** To give a valid basis for a vector space, we must list a collection of vectors that satisfy
  1. Basis should span the whole space
  2. No redundant basis vectors
- **Abstract vector space example:** Lets define an abstract vector space

$$M_{m \times n}.$$

That houses all  $m \times n$  matrices

To deduce whether this is a valid vector space, we check

1. **Zero vector:** The zero vector (matrix in this case) is definitely in the space

$$\begin{bmatrix} 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \dots & 0 \end{bmatrix}.$$

2. **Closed under multiplication by a scalar:**

$$\begin{aligned} s & \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ \vdots & \vdots & \ddots & \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \\ & = \begin{bmatrix} sa_{11} & sa_{12} & \dots & sa_{1n} \\ \vdots & \vdots & \ddots & \\ sa_{m1} & sa_{m2} & \dots & sa_{mn} \end{bmatrix}. \end{aligned}$$

Is definitely a member of the space

3. **Closed under addition:**

$$\begin{aligned} & \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \vdots & \ddots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} + \begin{bmatrix} b_{11} & \dots & b_{1n} \\ \vdots & \vdots & \ddots \\ b_{m1} & \dots & b_{mn} \end{bmatrix} \\ & = \begin{bmatrix} a_{11} + b_{11} & \dots & a_{1n} + b_{1n} \\ \vdots & \vdots & \ddots \\ a_{m1} + b_{m1} & \dots & a_{mn} + b_{mn} \end{bmatrix}. \end{aligned}$$

Is also definitely a member of the space

Now we must find the basis. A valid assumption would probably be

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

Let's check

$$\begin{aligned} & \begin{bmatrix} a & b \\ c & d \end{bmatrix} \\ & = a \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + b \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \\ & \quad + c \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} + d \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \\ & = \begin{bmatrix} a & b \\ c & d \end{bmatrix}. \end{aligned}$$

So this basis spans the whole span, but what about redundancies? Suppose  $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$  is redundant, then

$$\begin{aligned} \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} & = x \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + y \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \\ & \quad + z \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} x & 0 \\ y & z \end{bmatrix}. \end{aligned}$$

Thus, not redundant, we found the basis.

Since we have four basis matrices, this is a four dimensional space.

- **The trace of a matrix:** The trace of a square matrix is the sum of its diagonal elements. For an  $n \times n$  matrix  $A$ , the trace is defined as:

$$\text{Tr}(A) = \sum_{i=1}^n A_{ii}$$

where  $A_{ii}$  are the diagonal entries of  $A$ . The trace is only defined for square matrices and has several useful properties, such as being invariant under a change of basis.

- **Cyclic property of the trace:** The trace is invariant under cyclic permutations. Observe

$$\text{Tr}(AB) = \text{Tr}(BA).$$

Proof:

$$\begin{aligned} \text{Tr}(AB) &= \sum_{i=1}^n (ab)_{ii} \\ &= \sum_{i=1}^n \sum_{k=1}^n a_{ik} b_{kj}. \end{aligned}$$

**Remark.** For finite sums, the order of summation can be swapped because summation is commutative and associative when dealing with real or complex numbers. This means that the sum of a collection of terms does not depend on the order in which the terms are added. So, rearranging the summation over  $i$  and  $k$  doesn't change the value of the overall sum. Thus

$$\begin{aligned} \text{Tr}(AB) &= \sum_{i=1}^n \sum_{k=1}^n a_{ik} b_{kj} \\ &= \sum_{k=1}^n \sum_{i=1}^n a_{ik} b_{kj} \\ &= \sum_{k=1}^n \sum_{i=1}^n b_{kj} a_{ik} \\ &= \sum_{k=1}^n (ba)_{kk} \\ &= \text{Tr}(BA). \end{aligned}$$

- **Abstract vector space: Trace:** Suppose we define a space

$$U = \{u \in M_{2 \times 2} : \text{tr}(u) = 0\}.$$

It's a useful exercise to show why this is a valid space, but not shown here.

But what is the basis of this space? Let's keep the two valid basis matrices from  $M_{2 \times 2}$ , we then have

$$\begin{aligned} \begin{bmatrix} a & b \\ c & -a \end{bmatrix} &= s_1 \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} + s_2 \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 0 & s_1 \\ s_2 & 0 \end{bmatrix}. \end{aligned}$$

Which does not span the whole space, thus we need to add at least one more. Lets choose  $\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$ . Then

$$\begin{aligned} \begin{bmatrix} a & b \\ c & -a \end{bmatrix} &= s_1 \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} s_2 \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} + s_3 \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \\ &= \begin{bmatrix} s_1 & s_2 \\ s_3 & -s_1 \end{bmatrix}. \end{aligned}$$

Let  $s_1 = a$ ,  $s_2 = b$ , and  $s_3 = c$ , then we see we have the required span, and the dimension of the space is 3.

- **Abstract vector space: Determinant:** Suppose we want to define a space

$$V = \{v \in M_{2 \times 2} : \det(v) = 0\}.$$

Then we check

1. **Zero matrix:** Since the zero matrix from the ambient space has determinant zero, it is also in the subspace. Thus, we have a zero vector
2. **Scalar multiplication:**

$$\begin{aligned} sA &= s \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} sa & sb \\ sc & sd \end{bmatrix} \\ \det(sA) &= sasd - sbsc = s^2 \underbrace{(ad - bc)}_0 = 0. \end{aligned}$$

3. **Addition:** For  $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ , and  $B = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix}$

$$\begin{aligned} A + B &= \begin{bmatrix} a & b \\ c & d \end{bmatrix} + \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} \\ &= \begin{bmatrix} a + \alpha & b + \beta \\ c + \gamma & d + \delta \end{bmatrix} \\ \det(A + B) &= (a + \alpha)(d + \delta) - (b + \beta)(c + \gamma) \\ &= ad + a\delta + \alpha d + \alpha \delta - bc - b\gamma - \beta c - \beta \gamma \\ &= \underbrace{ad - bc}_0 + \underbrace{\alpha \delta - \beta \gamma}_0 + a\delta + \alpha d - b\gamma - \beta c \\ &= a\delta + \alpha d - b\gamma - \beta c. \end{aligned}$$

Although we clearly have terms left, this is not sufficient to assert that determinant is non-zero, lets look at an example where the product of two matrices with zero determinant is non-zero. If  $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ , and  $B = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$ . Then

$$\begin{aligned} A + B &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\ \det(A + B) &= 1. \end{aligned}$$

In this case, we have  $b = 0$ ,  $c = 0$ ,  $d = 0$ ,  $\beta = 0$ ,  $\gamma = 0$ , but  $\alpha = 1$ , and  $\delta = 1$ . So the terms we have left are

$$\begin{aligned} a\delta + 0 - 0 - 0 \\ = a\delta. \end{aligned}$$

Thus, this subspace is **not** closed under addition and is therefore not a valid vector space.

- **Side note:** Let's take a look at the trace equation and the determinant equation for two by two matrices

$$\begin{aligned}\text{tr}(A) &= a + d \quad (\text{linear}) \\ \det(A) &= ad - bc \quad (\text{non-linear}).\end{aligned}$$

This gives us a clue about whether a subspace from the ambient space forms a valid vector space. It turns out that the distinction between linear and non-linear properties plays a critical role in determining whether a set forms a valid subspace of a vector space.

A linear condition is one where the defining property of the vectors (like trace) is preserved under addition and scalar multiplication. This is why a subspace defined by "trace zero" is valid:

- **Trace of the sum:** If the trace of two matrices  $A$  and  $B$  is zero ( $\text{tr}(A) = 0$  and  $\text{tr}(B) = 0$ ), then

$$\text{tr}(A + B) = \text{tr}(A) + \text{tr}(B) = 0.$$

So, the subspace is closed under addition.

- **Trace under scalar multiplication:** If  $\text{tr}(A) = 0$ , then

$$\text{tr}(cA) = c \cdot \text{tr}(A) = 0,$$

so it is closed under scalar multiplication.

Since the trace is a linear function, these properties guarantee that the set of all matrices with trace zero is a subspace of  $M_{2 \times 2}$ .

A non-linear condition is one where the defining property (like determinant) is not preserved under addition or scalar multiplication in a linear way

- **Determinant of the sum:** Even if the determinant of two matrices  $A$  and  $B$  is zero ( $\det(A) = 0$  and  $\det(B) = 0$ ), it is generally not true that

$$\det(A + B) = 0.$$

In fact,  $\det(A + B)$  could be non-zero.

- **Determinant under scalar multiplication:** If  $\det(A) = 0$ , it's true that

$$\det(cA) = c^n \cdot \det(A) = 0$$

(for an  $n \times n$  matrix), but the failure under addition is enough to invalidate the subspace.

Since the determinant is a non-linear function, the set of all matrices with determinant zero is not closed under addition and thus does not form a subspace.

A linear condition typically leads to a valid subspace because it guarantees that the defining property of the subspace will hold under both vector addition and scalar multiplication, which are the core operations in a vector space.

Linear functions (like trace) satisfy:

$$\begin{aligned} f(\mathbf{u} + \mathbf{v}) &= f(\mathbf{u}) + f(\mathbf{v}) \\ f(s\mathbf{u}) &= sf(\mathbf{u}). \end{aligned}$$

These properties naturally preserve the structure of a subspace.

Non-linear functions (like determinant) do not satisfy these properties, and thus sets defined by non-linear conditions fail to be closed under the required operations for subspaces.

Thus, if a condition is linear, then the set of elements (such as matrices or functions) that satisfy this condition will form a valid subspace of a vector space, and thus a vector space in its own right. This is because a linear condition ensures that the set will automatically satisfy the necessary properties of a subspace.

If a condition is non-linear (like the condition that the determinant of a matrix is zero), it is not guaranteed to form a vector space. However, some non-linear conditions might still result in a valid vector space, depending on the specific structure of the set and how the condition interacts with vector addition and scalar multiplication.

- **Abstract vector space example: Polynomials:** Suppose we define a space (quadratic polynomials)

$$\{a + bx + cx^2 : a, b, c \in \mathbb{R}, x \in \mathbb{R}\}.$$

1. Let the zero function be the zero vector ( $f(x) = 0$ )
2.  $s(a + bx + cx^2) = sa + sbx + scx^2$  is in the space
3.  $(a + bx + cx^2) + (\hat{a} + \hat{b}x + \hat{c}x^2) = (a + \hat{a}) + (b + \hat{b})x + (c + \hat{c})x^2$  is in the space

What is the basis? a valid assumption would be  $b_1 = 1, b_2 = x, b_3 = x^2$ . Then,

$$a + bx + cx^2 = s_1(1) + s_2x + s_3x^2.$$

Let  $s_1 = a, s_2 = b, s_3 = c$ , we see that this forms a valid basis, and the dimension of the space is three.

- **Choosing a member of the quadratic polynomial vector space to best approximate  $x^3$ :** Now that we have defined the vector space that houses all quadratic polynomials  $p(x) = a + bx + cx^2$ , we may wonder how might we choose a member of this set to best approximate  $x^3$ . Choosing a member of this vector space essentially let's us choose any  $a, b, c$ , the pair that best minimizes the error.

To get a quantity for the error  $E(p)$ , we can square the difference  $x^3 - p(x)$ , and then integrate over the unit interval  $[0, 1]$ . By integrating the squared difference, we are finding the average. Recall the mean value theorem

$$\frac{1}{b-a} \int_a^b f(x) dx.$$

By integrating over the unit interval, the term outside the integral collapses to one, which makes things easy. Thus, we have

$$\int_0^1 (x^3 - (a + bx + cx^2))^2 dx.$$

If  $a, b, c = 0$ , then

$$\text{ave}(E(p)) = \sqrt{\int_0^1 (x^{3-0})^2 dx} = \sqrt{\frac{1}{7}x^7 \Big|_0^1} = \frac{1}{\sqrt{7}} \approx 0.37796.$$

But this is clearly a bad approximation. We must find an  $a, b, c$  such that the average error  $\text{ave}(E(p))$  is as close to zero as possible (minimized).

To find  $a, b, c$ , we define the error function  $E(a, b, c)$ . Then, we find

$$\begin{aligned} E(a, b, c) &= \int_0^1 (x^3 - a + bx + cx^2)^2 dx \\ &= -\frac{1}{2}a - \frac{2}{5}b - \frac{1}{3}c + \frac{1}{7} + \frac{2}{3}ac + \frac{1}{2}bc + \frac{1}{5}c^2 + ab + \frac{1}{3}b^2 + a^2. \end{aligned}$$

To minimize, first find

$$\begin{pmatrix} \frac{\delta E}{\delta a} \\ \frac{\delta E}{\delta b} \\ \frac{\delta E}{\delta c} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

We have

$$\begin{aligned} \frac{\delta E}{\delta a} &= -\frac{1}{2} + 2a + b + \frac{2}{3}c \\ \frac{\delta E}{\delta b} &= -\frac{2}{5} + a + \frac{2}{3}b + \frac{1}{2}c \\ \frac{\delta E}{\delta c} &= -\frac{1}{3} + \frac{2}{3}a + \frac{1}{2}b + \frac{2}{5}c. \end{aligned}$$

Thus,

$$\begin{pmatrix} 2 + 1 + \frac{2}{3} \\ 1 + \frac{2}{3} + \frac{1}{2} \\ \frac{2}{3} + \frac{1}{2} + \frac{2}{5} \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Solving this system yields

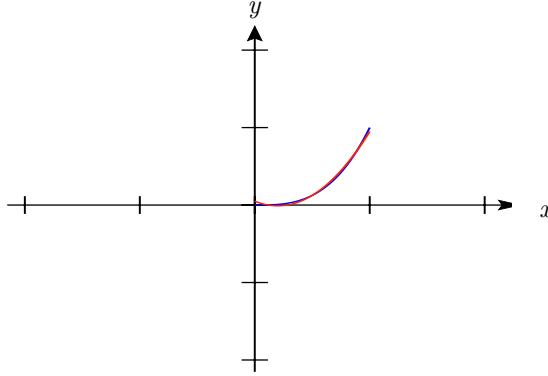
$$\begin{aligned} a &= \frac{1}{20} \\ b &= -\frac{3}{5} \\ c &= \frac{3}{2}. \end{aligned}$$

Getting the Hessian matrix and substituting int the critical point yields

The principal minors are

$$\begin{aligned} H_1 &= 2 > 0 \\ H_2 &= \frac{1}{3} > 0 \\ H_3 &= \frac{1}{270} > 0. \end{aligned}$$

Thus, the point  $\frac{1}{20}, -\frac{3}{5}, \frac{3}{2}$  is a local min and the function  $p(x) = \frac{1}{20} - \frac{3}{5}x + \frac{3}{2}x^2$  best approximates  $x^3$ .



- **Isomorphic vector spaces:** Two vector spaces are isomorphic if there is a one-to-one correspondence (a bijection) between them that preserves the structure of vector addition and scalar multiplication. This means that if vector spaces  $V$  and  $W$  are isomorphic, there exists a map (called an isomorphism)  $\phi : V \rightarrow W$  such that:

1.  $\phi$  is bijective: Every element in  $W$  has a unique preimage in  $V$ , and every element in  $V$  is mapped to a unique element in  $W$ .

$$\forall w \in W, \exists v \in V \text{ such that } \phi(v) = w$$

2.  $\phi$  preserves addition: For any two vectors  $u, v \in V$ ,

$$\phi(u + v) = \phi(u) + \phi(v)$$

3.  $\phi$  preserves scalar multiplication: For any scalar  $c \in \mathbb{F}$  and any vector  $v \in V$ ,

$$\phi(cv) = c\phi(v)$$

**Example:**  $M_{2 \times 2}$  is isomorphic to  $\mathbb{R}^4$

A matrix in  $m_{2 \times 2}$  can be written as

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

Where  $a, b, c, d \in \mathbb{R}$ . This matrix can be uniquely represented as a 4-dimensional vector:

$$(a, b, c, d) \in \mathbb{R}^4.$$

The correspondence between the matrix and the 4-dimensional vector is a linear bijection that preserves both vector addition and scalar multiplication. Thus, there is a one-to-one correspondence between  $M_{2 \times 2}$  and  $\mathbb{R}^4$ , and the two spaces are isomorphic.

Moreover, The isomorphism between  $M_{2 \times 2}$  (the space of  $2 \times 2$  matrices) and  $\mathbb{R}^4$  (the 4-dimensional real vector space) can be described by a linear map that transforms a matrix into a 4-dimensional vector by simply mapping the matrix entries to the components of the vector.

Let's define the linear map  $\phi : M_{2 \times 2} \rightarrow \mathbb{R}^4$ .

For any matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in M_{2 \times 2},$$

the corresponding vector in  $\mathbb{R}^4$  under the map  $\phi$  would be:

$$\phi(A) = (a, b, c, d) \in \mathbb{R}^4.$$

This map simply "flattens" the matrix into a 4-tuple of real numbers, with the components arranged in a consistent order (for example, row by row or column by column). In this case, we are mapping the entries of the matrix row by row.

Conversely, given any vector  $(a, b, c, d) \in \mathbb{R}^4$ , the corresponding matrix in  $M_{2 \times 2}$  under the inverse map  $\phi^{-1}$  would be:

$$\phi^{-1}(a, b, c, d) = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

### Properties of the Map:

- **Bijectivity:** Every matrix corresponds to a unique vector, and every vector corresponds to a unique matrix.

- **Additivity:** If  $A_1 = \begin{pmatrix} a_1 & b_1 \\ c_1 & d_1 \end{pmatrix}$  and  $A_2 = \begin{pmatrix} a_2 & b_2 \\ c_2 & d_2 \end{pmatrix}$ , then

$$\phi(A_1 + A_2) = \phi\left(\begin{pmatrix} a_1 + a_2 & b_1 + b_2 \\ c_1 + c_2 & d_1 + d_2 \end{pmatrix}\right) = (a_1 + a_2, b_1 + b_2, c_1 + c_2, d_1 + d_2),$$

which is the same as  $\phi(A_1) + \phi(A_2)$ .

- **Scalar Multiplication:** For any scalar  $\lambda \in \mathbb{R}$ ,

$$\phi(\lambda A) = \phi\left(\begin{pmatrix} \lambda a & \lambda b \\ \lambda c & \lambda d \end{pmatrix}\right) = (\lambda a, \lambda b, \lambda c, \lambda d),$$

which is the same as  $\lambda\phi(A)$ .

Thus,  $\phi$  is a linear isomorphism between  $M_{2 \times 2}$  and  $\mathbb{R}^4$ .

**Note:** We only have isomorphism if the dimensions are the same.

- **Dot product in function spaces:** The dot product in a function space is defined as the product of a pair of two basis vectors, integrated over  $[0, 1]$ .
- **Invariant subspace:** An invariant subspace is a subspace of a vector space that remains unchanged when a linear operator is applied to it. More formally, if  $V$  is a vector space and  $T : V \rightarrow V$  is a linear operator, a subspace  $W \subseteq V$  is called an invariant subspace under  $T$  if, for every vector  $w \in W$ , the image  $T(w) \in W$ . In other words, applying  $T$  to any vector in  $W$  results in another vector that is still within  $W$ .

Formally,  $W$  is an invariant subspace under  $T$  if

$$T(W) \subseteq W.$$

In other words,  $\forall w \in W, T(w) \in W$

- **Eigenspaces:** The eigenspace corresponding to an eigenvalue  $\lambda$  of a linear operator  $T$  is the set of all eigenvectors associated with  $\lambda$ , along with the zero vector:

$$E_\lambda = \{v \in V \mid T(v) = \lambda v\}.$$

Each eigenvalue  $\lambda$  has its own eigenspace.

**Claim:** An Eigenspace is a vector space

**Proof.**

1. (**zero vector**): If  $v = 0$ , then  $T(0) = \lambda 0$ , since a linear map always maps the zero vector to the zero vector.  $0 = \lambda 0 \implies 0 = 0$ . Thus,  $0 \in E_\lambda$
2. (**Scalar closure**):  $se = s\lambda v = \lambda(sv)$ , thus  $se \in E_\lambda$
3. (**Addition closure**):  $e + \hat{e} = \lambda e + \lambda \hat{e} = \lambda(e + \hat{e})$ , thus  $e + \hat{e} \in E_\lambda$

- **More on function spaces:** Suppose we define a vector space  $P_4$ , which contains all polynomials of degree four or less. It has a basis

$$1, x, x^2, x^3, x^4.$$

Thus, is dim 5. The verification of this being a legitimate vector space is not shown. Suppose we define two subspaces of  $P_4$ ,  $E$  and  $O$ . Where  $E$  is the vector subspace of  $P_4$  that contains all even functions, and  $O$  is the vector subspace of  $P_4$  that contains all the odd functions. Recall that a function is even iff  $f(-x) = f(x)$ , and a function is odd iff  $f(-x) = -f(x)$ . Both subspaces can be shown that they are indeed vector subspaces. In both subspaces, we use the zero function from  $P_4$ , because the zero function is defined here to be both even and odd.

$E$  has basis

$$1, x^2, x^4.$$

Which is dim 3. Note that all constant functions are even, because they satisfy the property of even functions  $f(-x) = f(x)$ .

$O$  has basis

$$x, x^3.$$

Which is dim 2. Let's define a mapping

$$\begin{aligned} L : P_4 &\rightarrow P_4 \\ L(p(x)) &= p'(x). \end{aligned}$$

Which can easily be shown to be linear. Note that this map is not surjective or injective, and thus not bijective. The codomain can not be filled, because no functions of degree four can be reached by differentiating polynomials of degree four or less. Not injective because all constant functions yield the same derivative (0).

Now, we apply the differentiation operator  $L(p(x)) = p'(x)$  to each basis element:

- $L(1) = 0$  because the derivative of a constant is 0.
- $L(x) = 1$ .
- $L(x^2) = 2x$ .
- $L(x^3) = 3x^2$ .
- $L(x^4) = 4x^3$ .

Next, express each of these derivatives as a linear combination of the basis elements  $\{1, x, x^2, x^3, x^4\}$ :

- $L(1) = 0 = 0 \cdot 1 + 0 \cdot x + 0 \cdot x^2 + 0 \cdot x^3 + 0 \cdot x^4$ .
- $L(x) = 1 = 1 \cdot 1 + 0 \cdot x + 0 \cdot x^2 + 0 \cdot x^3 + 0 \cdot x^4$ .
- $L(x^2) = 2x = 0 \cdot 1 + 2 \cdot x + 0 \cdot x^2 + 0 \cdot x^3 + 0 \cdot x^4$ .
- $L(x^3) = 3x^2 = 0 \cdot 1 + 0 \cdot x + 3 \cdot x^2 + 0 \cdot x^3 + 0 \cdot x^4$ .
- $L(x^4) = 4x^3 = 0 \cdot 1 + 0 \cdot x + 0 \cdot x^2 + 4 \cdot x^3 + 0 \cdot x^4$ .

We now construct the matrix of the linear map  $L$  with respect to the basis  $\mathcal{B}$ . Each column of the matrix corresponds to the image of one of the basis elements, written as a linear combination of the basis elements

$$[L]_{\mathcal{B}} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 4 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

- The first column corresponds to  $L(1) = 0$ .
- The second column corresponds to  $L(x) = 1$ .
- The third column corresponds to  $L(x^2) = 2x$ .
- The fourth column corresponds to  $L(x^3) = 3x^2$ .
- The fifth column corresponds to  $L(x^4) = 4x^3$ .

This matrix represents the differentiation operator on the vector space  $P_4$  in the basis  $\{1, x, x^2, x^3, x^4\}$ .

Any polynomial  $p(x) \in P_4$  can be written as a linear combination of the basis elements  $\{1, x, x^2, x^3, x^4\}$ , so we can express the polynomial as a vector of coefficients. For example, if:

$$p(x) = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4,$$

then the polynomial corresponds to the vector:

$$\mathbf{p} = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix}.$$

To apply the linear map  $L$ , you multiply the matrix representing  $L$  by the vector of coefficients for  $p(x)$ .

If  $[L]_{\mathcal{B}}$  is the matrix of the linear map  $L$ , and  $\mathbf{p}$  is the vector of coefficients for  $p(x)$ , then the result is:

$$L(p(x)) = [L]_{\mathcal{B}} \cdot \mathbf{p}.$$

A linear map from a vector space to itself can be represented as a square matrix whose size corresponds to the dimension of the space. In this case, since  $L$  maps  $P_4$  to itself, and  $P_4$  has dimension 5, the matrix representing  $L$  must be  $5 \times 5$ .

Now we define a few more maps, first  $L : E \rightarrow O$ , which has a matrix

$$\begin{pmatrix} 0 & 2 & 0 \\ 0 & 0 & 4 \end{pmatrix}.$$

Then  $L : P_4 \rightarrow P_3$ , which has matrix

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 4 \end{pmatrix}.$$

Then,  $L : O \rightarrow E$

$$\begin{pmatrix} 1 & 0 \\ 0 & 3 \\ 0 & 0 \end{pmatrix}.$$

Then,  $K : P_3 \rightarrow P_4$ , where  $K(p(x)) = \int_0^x p(u) du$ , defining the integral in this way leads to no constant of integration. Having a constant of integration would lead to a map that is no longer linear, because  $K(0) \neq 0$

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & \frac{1}{4} \end{pmatrix}.$$

Now, we take a look at  $L \circ K : P_3 \rightarrow P_3$

$$\begin{aligned} & \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 4 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & \frac{1}{4} \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = I. \end{aligned}$$

Thus, we say that  $L$  is a left inverse of  $K$

What about  $K \circ L : P_4 \rightarrow P_4$ ?

$$\begin{aligned} & \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & \frac{1}{4} \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 4 \end{pmatrix} \\ &= 0 \neq I. \end{aligned}$$

Thus,  $L$  is not a right inverse of  $K$

- **A quick remark:** We remark that to express a map as a matrix it must satisfy

1. Linear

## 2. Known basis

- **More on the dot product of function spaces:** For two functions  $f$  and  $g$  in a function space, the dot product (also referred to as the inner product) is typically defined as an integral of their product over a specific interval. Specifically, for real-valued functions  $f$  and  $g$  on an interval  $[a, b]$ , the inner product is defined as:

$$\langle p, q \rangle = \int_a^b p(x)q(x) dx.$$

If we define the interval of our function space  $x \in [0, 1]$ , for example  $P_2$ ,  $x \in [0, 1]$ . Then the inner product is

$$\langle p, p \rangle = \int_0^1 p(x)q(x) dx.$$

- **Properties of the inner product:**

- $\langle p, 0 \rangle = \int_a^b 0p(x) dx = 0 \int_a^b p(x)q(x) dx = 0$
- $\langle q, p \rangle = \int_a^b q(x)p(x) dx = \int_a^b p(x)q(x) dx = \langle p, q \rangle$ . Thus, we say the inner product is symmetric
- $\langle p, p \rangle = \int_a^b p^2(x) dx \geq 0$  and only equal to zero if  $p = 0 \forall x \in [a, b]$
- $\langle sp, q \rangle = \langle p, sq \rangle = s \langle p, q \rangle$ . Note that  $s$  cannot depend on  $x$
- $\langle p + q, r \rangle = \langle p, r \rangle + \langle q, r \rangle$

- **Norm of the inner product:** Recall in vector calculus

$$\begin{aligned} \mathbf{u} \cdot \mathbf{u} &= |\mathbf{u}|^2 \\ \implies \sqrt{\mathbf{u} \cdot \mathbf{u}} &= |\mathbf{u}|. \end{aligned}$$

Given item three above,  $\langle p, p \rangle = \int_a^b p^2(x) dx$ . Thus,  $|\langle p, p \rangle| = \sqrt{\int_a^b p^2(x) dx}$

**Example:** Suppose we consider the function space  $P_2$  over  $[0, 1]$

$$|1 + x^2| = \sqrt{\int_0^1 (1 + x^2)^2 dx} = \sqrt{\frac{28}{15}}.$$

- **Angle between function space vectors:** Recall from vector calculus

$$\begin{aligned} v \cdot w &= |v||w|\cos(\theta) \\ \implies \cos(\theta) &= \frac{v \cdot w}{|v||w|}. \end{aligned}$$

Suppose we have  $\langle 1, x \rangle$ , in  $P_2$  over  $[0, 1]$

$$\langle 1, x \rangle = \cos^{-1} \left( \frac{\int_0^1 1 \cdot x dx}{\sqrt{\int_0^1 1^2 dx} \sqrt{\int_0^1 x^2 dx}} \right).$$

- **General Inner product:** If  $V$  is a vector space, then the inner product is an operation

$$\langle v_1, v_2 \rangle : V \times V \rightarrow \mathbb{R}.$$

With norm  $|v| = \sqrt{\langle v, v \rangle}$ . And with properties described above. Note that only the zero vector has norm zero.

- **Cauchy Schwarz inequality:** The Cauchy-Schwarz inequality states that for any two vectors  $u$  and  $v$  in an inner product space, the absolute value of their inner product is less than or equal to the product of their norms:

$$|\langle v, w \rangle| \leq |v||w|.$$

Equality holds if and only if  $u$  and  $w$  are linearly dependent.

A consequence of this is the inequality

$$-1 \leq \frac{\langle \mathbf{v}, \mathbf{w} \rangle}{\|\mathbf{v}\|\|\mathbf{w}\|} \leq 1$$

As we saw above,  $\theta = \cos^{-1} \left( \frac{\langle \mathbf{v}, \mathbf{w} \rangle}{\|\mathbf{v}\|\|\mathbf{w}\|} \right)$

We note that  $\theta$  must satisfy  $0 \leq \theta \leq \pi$ . When  $\theta = 0$ , the two vectors are parallel. When  $\theta = \frac{\pi}{2}$ , the two vectors are orthogonal. When  $\theta = \pi$ , the two vectors are anti-parallel

- **Inner product space:** An inner product space, also called an *inner space*, is a vector space equipped with an additional structure called an inner product. The inner product is a way to define a notion of "angle" and "length" in the vector space, generalizing the dot product in Euclidean space.
- **Recall: Vector projection:** The vector projection of  $\mathbf{v}$  onto  $\mathbf{u}$  has the same initial point as  $\mathbf{u}$  and  $\mathbf{v}$  and the same direction as  $\mathbf{u}$ , and represents the component of  $\mathbf{v}$  that acts in the direction of  $\mathbf{u}$  :

$$\text{proj}_{\vec{u}} \vec{v} = \frac{\vec{v} \cdot \vec{u}}{\|\vec{u}\|^2} \vec{u}.$$

We say "The vector projection of  $\vec{v}$  onto  $\vec{u}$ "

### 6.1.11 Kernel, Image, Orthogonality, and the Rank Nullity theorem

- **Gram-schmidt orthogonalization:** The Gram-Schmidt process is a method used to take a set of linearly independent vectors and convert them into an orthogonal (or orthonormal) set of vectors. It's useful in linear algebra for generating an orthogonal basis from a given set of vectors in an inner product space. Here's how the process works step by step:

Start with the first vector:

$$\mathbf{u}_1 = \mathbf{v}_1$$

This becomes the first orthogonal vector.

For each subsequent vector  $\mathbf{v}_k$ ,

subtract the projection of  $\mathbf{v}_k$  onto the previously found orthogonal vectors:

$$\mathbf{u}_k = \mathbf{v}_k - \sum_{j=1}^{k-1} \text{proj}_{\mathbf{u}_j}(\mathbf{v}_k)$$

where  $\text{proj}_{\mathbf{u}_j}(\mathbf{v}_k)$  is the projection of  $\mathbf{v}_k$  onto  $\mathbf{u}_j$ :

$$\text{proj}_{\mathbf{u}_j}(\mathbf{v}_k) = \frac{\langle \mathbf{v}_k, \mathbf{u}_j \rangle}{\langle \mathbf{u}_j, \mathbf{u}_j \rangle} \mathbf{u}_j$$

If an orthonormal set is desired, normalize each orthogonal vector:

$$\mathbf{e}_k = \frac{\mathbf{u}_k}{\|\mathbf{u}_k\|}$$

**Example 1:** Consider in  $\mathbb{R}^2$ , the vectors  $v_1 = (1, 0)$ , and  $v_2 = (1, 1)$ . We take the first vector and label it  $\tilde{b}_1$ . If we find the vector is of unit length, we then remove the tilde and declare it a basis vector. If it is not unit length, we normalize. Recall to normalize a vector, we divide by its norm. Since the norm of  $\tilde{b}_1$  is one,  $b_1 = 1$

Then, for the second vector,

$$\begin{aligned} \tilde{b}_2 &= (1, 1) - \text{proj}_{b_1}(v_2)b_1 = (1, 1) - \frac{(1, 0) \cdot (1, 1)}{1^2}(1, 0) \\ &= (1, 1) - (1, 0) = (0, 1). \end{aligned}$$

Thus,  $\tilde{b}_2 = (0, 1)$ , we then check its norm.  $|\tilde{b}_2| = \sqrt{1^2} = 1$ , since the norm is one,  $b_2 = (0, 1)$  and we are done.

- **Graham-schmidt process with inner spaces:** Suppose we have the inner space  $P_1$ , where  $x \in [0, 1]$ , a standard choice of basis vectors is  $1, x$ . Then by the process described above,

$$\tilde{b}_1 = 1 \quad |1| = \sqrt{\int_0^1 1^2 dx} = 1.$$

Thus,  $b_1 = 1$ . Note that we are not talking about real numbers here, we are talking about 1 the function. Furthermore,

$$\begin{aligned}\tilde{b}_2 &= x - \langle 1, x \rangle 1 \\ &= x - \sqrt{\int_0^1 x dx} \\ &= x - \frac{1}{2}.\end{aligned}$$

The norm of  $x - \frac{1}{2}$  is

$$\left| x - \frac{1}{2} \right| = \sqrt{\int_0^1 \left( x - \frac{1}{2} \right)^2 dx} = \frac{1}{\sqrt{12}}.$$

Since this is not of unit length, we divide by the norm to get

$$b_2 = \sqrt{12} \left( x - \frac{1}{2} \right).$$

We have constructed an orthonormal basis for the inner product space  $P_1$ , consisting of the functions  $b_1 = 1$  and  $b_2 = \sqrt{12} \left( x - \frac{1}{2} \right)$ . These functions are orthogonal to each other, and each has unit length with respect to the inner product defined by integration over  $[0, 1]$ .

We remark that this orthonormal basis spans the same as the original basis,  $1, x$

- **More Graham-schmidt on inner spaces:** Suppose we now use the inner space  $P_2$ , which is the set of all polynomials of degree two or less. With standard basis  $1, x, x^2$ . By the last example, we know that  $b_1$  and  $b_2$  will be 1 and  $\left( x - \frac{1}{2} \right)$  respectively. Since we introduced a third basis vector  $x^2$ , we must compute  $\tilde{b}_3$

$$\tilde{b}_3 = x^2 - \text{proj}_{b_1}(x^2) - \text{proj}_{b_2}(x^2)$$

Where

$$\begin{aligned}\text{proj}_{b_1}(x^2) &= \frac{\langle x^2, 1 \rangle}{|1|^2} 1 = \int_0^1 x^2 dx = \frac{1}{3} \\ \text{proj}_{b_2}(x^2) &= \frac{\langle x^2, \sqrt{12} \left( x - \frac{1}{2} \right) \rangle}{|\sqrt{12} \left( x - \frac{1}{2} \right)|^2} \sqrt{12} \left( x - \frac{1}{2} \right) \\ &= \sqrt{12} \int_0^1 x^2 \left( x - \frac{1}{2} \right) dx \sqrt{12} \left( x - \frac{1}{2} \right) \\ &= \left( x - \frac{1}{2} \right).\end{aligned}$$

Thus

$$\begin{aligned}\tilde{b}_3 &= x^2 - \frac{1}{3} - \left( x - \frac{1}{2} \right) \\ &= x^2 - x + \frac{1}{6}.\end{aligned}$$

With

$$\begin{aligned} |\tilde{b}_3| &= \sqrt{\int_0^1 \left(x^2 - x + \frac{1}{6}\right)^2 dx} = \frac{1}{\sqrt{180}} \\ \therefore b_3 &= \sqrt{180} \left(x^2 - x + \frac{1}{6}\right) \\ &= \sqrt{5}(6x^2 - 6x + 1). \end{aligned}$$

Note that this orthonormal set of basis vectors  $\{1, \sqrt{3}(2x - 1), \sqrt{5}(6x^2 - 6x + 1)\}$  for  $P_2$  is precisely the preferred basis we saw when approximating polynomials with quadratics.

Suppose we wish to best approximate  $x^3$  over  $x \in [0, 1]$  with a quadratic. We simply add together the projections of  $x^3$  on each orthonormal basis in  $P_2$ . Thus,

$$\begin{aligned} f_{\text{approx}} &= \text{proj}_{b_1}(x^3) + \text{proj}_{b_2}(x^3) + \text{proj}_{b_3}(x^3) \\ &= \int_0^1 x^3 dx \cdot 1 + \sqrt{3} \int_0^1 x^3(2x - 1) dx \cdot \sqrt{3}(2x - 1) \\ &\quad + \sqrt{5} \int_0^1 x^3(6x^2 - 6x + 1) dx \cdot \sqrt{5}(6x^2 - 6x + 1) \\ &= \frac{1}{4} + \frac{9}{20}(2x - 1) + \frac{1}{4}(6x^2 - 6x + 1) \\ &= \frac{3}{2}x^2 - \frac{3}{5}x + \frac{1}{20}. \end{aligned}$$

Which is precisely the same function we derived using the standard basis for  $P_2$  and the various calculus techniques.

**Note:** The projection calls for division by the square of the norm of the basis vector, but since the basis vectors are normal, we omit it from the calculation.

- **Linear Operations:** A linear map from a vector space to itself,  $L : V \rightarrow V$  is called a *linear operation*, or just an *operation*
- **Defining linear maps on vector spaces, null and image space:** Suppose we have the function space  $P_4$ , all polynomials of degree four or less, along with  $E \subset P_4$  and  $O \subset P_4$ , the even and odd functions

If we define a map

$$L : V \rightarrow W.$$

Where  $L$  is linear, and  $V, W$  are linear. Let's define a subspace

$$N_L = \{v \in V : L(v) = 0_w\} \subset V.$$

This subspace can be shown to be a vector space. Formally, this vector subspace is called the null space of  $L$ , or  $\ker(L)$  (kernel of  $L$ ).

We can check that it is a valid vector space by checking that it has a zero vector, and it is closed under addition and multiplication by a scalar.

1.  $L(0_v) = 0_w$ , so  $0_v \in N_L$
2.  $L(sv) = sL(v) = s0_w = 0_w \in N_L$
3.  $L(v + \hat{v}) = L(v) + L(\hat{v}) = 0_w + 0_w = 0_w \in N_L$

Let's define a new subspace

$$R_L = \{w \in W : w = L(v)\}.$$

In other words, all vectors  $w \in W$  such that  $w$  is the image of a vector in  $v$ . If  $L$  is onto, then every vector in the codomain  $W$  is the image of a vector in  $v$ , thus  $R_L$  will be all vectors in  $W$ . If  $W$  has dimension  $n$  and  $W$  is onto, then  $\dim(R_L) = \dim(W)$ . If  $L$  is not onto, then  $\dim(R_L) < \dim(W)$

If  $L$  is not onto, there exist some vectors in  $W$  that are not the image of any vector in  $V$ . This means that  $L$  does not "cover" the whole space  $W$ , and the image of  $L$  (i.e., the range  $R_L$ ) is a proper subspace of  $W$ . Since subspaces have lower dimension than the spaces they are part of,  $\dim(R_L) < \dim(W)$ .

We can show that  $R_L$  is a vector space by showing

1. **Zero vector:**  $0_w = L(0_v), \implies 0_w \in R_L$
2. **Closed under addition:**  $w + w' = L(v) + L(v') = L(v + v')$ . Since  $v + v' \in V$  this criteria is satisfied and  $w + w' \in R_L$
3. **Closed under scalar multiplication:**  $sw = sL(v) = L(sv)$ . Since  $V$  is given to be a vector space, it must be that  $sv \in V$ , thus this criteria is satisfied

Thus,  $R_L$  is a vector space.

**Note:**  $R_L$  is called the image (or range) space of  $L$ , denoted  $\text{Im}(L)$ .

If a map is surjective, the the image is the codomain.

- **More the image of a map (range space):** The image of a map  $L : V \rightarrow W$ ,  $\text{Im}(L)$  is often referred to as the *column* space of  $L$ , if  $L$  is represented by a matrix.

For a linear map  $L : V \rightarrow W$ , when  $L$  is represented by a matrix  $A$ , each column of  $A$  shows how  $L$  transforms a specific basis vector of  $V$ . The span of these columns, i.e., the column space, contains all possible outputs (or images) of the linear map.

Thus, the column space of the matrix is exactly the set of all vectors in  $W$  that can be reached by applying  $L$  to vectors in  $V$ . This makes the column space equivalent to the image (or range) of the linear map  $L$ .

- **Image of the transpose (row space):** If a linear map is represented by a matrix, then we saw above the image of the map is the column space

We also see that the image of the transpose of the map is then the row space. The image of  $A^T$ ,  $\text{Im}(A^T)$ , is the span of the columns of  $A^T$ , which corresponds to the row space of  $A$  (since the rows of  $A$  become the columns of  $A^T$ ). Thus

$$\begin{aligned} \text{Col}(A^T) &= \text{Row}(A) \\ \mathcal{C}(A^T) &= \mathcal{R}(A). \end{aligned}$$

- **Defining linear maps on vector spaces, eigenspaces:** For the linear operation  $L : V \rightarrow V$ . Define a subspace The column space of a matrix represents the span of its columns, which correspond to the images of the basis vectors of the domain under the linear map.

$$E_\lambda = \{v \in V : L(v) = \lambda v\} \subset V.$$

Where  $\lambda$  is fixed. Is it a vector space?

1. **Zero vector:**  $L(0) = \lambda 0 = 0 \in E_\lambda$
2. **Closed under addition:**  $L(sv) = sL(v) = s(\lambda v) = \lambda(sv) \in E_\lambda$
3. **Closed under addition:**  $L(v + \hat{v}) = L(v) + L(\hat{v}) = \lambda v + \lambda \hat{v} = \lambda(v + \hat{v}) \in E_\lambda$

Thus,  $E_\lambda$  is a vector space. We call such vector spaces *eigenspaces* if the dimension is greater than zero. If the dimension is zero, then the only vector in the space is the zero vector. An eigenspace must contain at least one nonzero eigenvector. If the only vector in the set is the zero vector, it is not considered an eigenspace for any eigenvalue.

- **Perpendicular inner spaces (Orthogonal complement):** Suppose we have an ambient space  $V$ , then we take a subset of  $V$ , namely  $U \subset V$ ,  $U \neq \emptyset$ . We define

$$U^\perp = \{v \in V : \langle v, u \rangle = 0, \forall u \in U\}.$$

**Example:** Suppose we have the ambient space  $\mathbb{R}^2$ , then we define  $U = \{(1, 0)\}$ .  $U^\perp$  is then

$$U^\perp = \{(x, y) \in \mathbb{R}^2 : \langle (x, y), (1, 0) \rangle = 0\}.$$

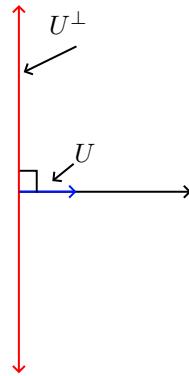
Thus, are vectors  $(x, y) \in \mathbb{R}^2$  must satisfy

$$\begin{aligned} \langle (x, y), (1, 0) \rangle &= 0 \\ \implies 1 \cdot x + 0 \cdot y &= 0 \\ \implies x &= 0. \end{aligned}$$

From this, we can state

$$U^\perp = \{(0, y) : y \in \mathbb{R}\}.$$

Geometrically,  $U^\perp$  is the  $y$ -axis.



But is  $U^\perp$  a vector space?

1. let  $y = 0$ , then  $(0, y) = (0, 0) \in U^\perp$
2.  $s(0, y) = (0, sy)$ . Since  $s \in \mathbb{R}$  and  $y \in \mathbb{R}$ ,  $sy \in \mathbb{R}$ , thus  $(0, sy) \in U^\perp$
3.  $(0, y) + (0, \hat{y}) = (0, y + \hat{y})$ , since  $y + \hat{y} \in \mathbb{R}$ ,  $(0, y + \hat{y}) \in \mathbb{R}$

Thus,  $U^\perp$  is a vector space. ■

- **Claim:  $U^\perp$  is always a vector space:** If  $U$  is a subset of a inner space, then  $U^\perp$  will be a vector space.

**Proof:**

1. The zero vector is orthogonal to all vectors in the inner product space. So, for any  $u \in U$ , we have  $\langle 0, u \rangle = 0$ . Hence,  $0 \in U^\perp$ .
  2. Let  $v \in U^\perp$ , and take any scalar  $s \in \mathbb{F}$  (where  $\mathbb{F}$  is the field over which the vector space is defined). Then  $\langle sv, u \rangle = s\langle v, u \rangle = s(0) = 0$ , so  $sv \in U^\perp$ .
  3. Let  $v, w \in U^\perp$ . Then for any  $u \in U$ , we have  $\langle v+w, u \rangle = \langle v, u \rangle + \langle w, u \rangle = 0+0 = 0$ . Hence,  $v+w \in U^\perp$ .
- **Perpendicular inner space examples:** Suppose in the ambient space  $\mathbb{R}^3$  we define a subset  $U = \{(1, 2, 3)\}$ . Then,

$$U^\perp = \{(x, y, z) \in \mathbb{R}^3 : \langle (x, y, z), (1, 2, 3) \rangle = 0\}.$$

Implies

$$\begin{aligned} (x, y, z) \cdot (1, 2, 3) &= 0 \\ \implies x + 2y + 3z &= 0. \end{aligned}$$

We see that this defines a vector plane in  $\mathbb{R}^3$ , with normal vector  $\vec{n} = (1, 2, 3)$ . Thus, the members of  $U^\perp$  are the members of the ambient space that are perpendicular to the vector  $(1, 2, 3)$ .

Consider a different example. Again, let the ambient space be  $\mathbb{R}^3$ . Then define  $U = \{(1, -1, 1), (-1, 1, -1)\}$ . Then

$$U^\perp = \{(x, y, z) \in \mathbb{R}^3 : \langle (x, y, z), u \rangle = 0 \ \forall u \in U\}.$$

Note that  $U$  is dim 1, we see that

$$\begin{aligned} (1, -1, 1) &= \lambda(-1, 1, -1) \\ \implies 1 &= -\lambda \implies \lambda = -1 \\ -1 &= \lambda \\ 1 &= -\lambda \implies \lambda = -1. \end{aligned}$$

Since the second vector can be expressed as a linear combination of the first, it is linearly dependent, and the space  $U$  is therefore spanned by a single vector, and then has dim 1.

We require

$$\begin{aligned} (x, y, z) \cdot (1, -1, 1) &= 0 \\ (x, y, z) \cdot (-1, 1, -1) &= 0 \end{aligned}$$

Which implies the system of linear equations

$$\begin{cases} x - y + z = 0 \\ -x + y - z = 0 \end{cases}.$$

We can see that adding equation one to equation two leads to the discovery that equation two is redundant, and thus

$$x - y + z = 0.$$

Is the only constraint. Furthermore, we have  $\vec{n} = (1, -1, 1)$

Suppose we change the second vector in  $U$  above to  $(-1, 2, -1)$  instead of  $(-1, 1, -1)$ . Then the system becomes

$$\begin{cases} x - y + z = 0 \\ -x + 2y - z = 0 \end{cases}.$$

Using gaussian elimination, we find

$$\begin{cases} x + z = 0 \\ y = 0 \end{cases}.$$

Both equations must hold for a vector  $v \in \mathbb{R}^3$  to be in  $U^\perp$ . If we let  $z = t$ , the n

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -t \\ 0 \\ t \end{pmatrix} = t \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}.$$

Which describes a line in  $\mathbb{R}^3$  that passes through the origin

- **Dimensions of  $U$  and  $U^\perp$ :** Suppose we have the ambient space  $\mathbb{R}^3$ , where  $U = \{(1, 2, 3)\} \subset R^3$ . Then  $U^\perp = \{(x, y, z) \in \mathbb{R}^3 : \langle (x, y, z), (1, 2, 3) \rangle = 0\}$ . Thus,

$$(x, y, z) \cdot (1, 2, 3) = 0 \implies x + 2y + 3z = 0.$$

Which describes a 2 dimensional plane in  $\mathbb{R}^3$

The dimension of  $U$  is one because it is spanned by a single vector. It is the set of all points that lie on this single vector. The dimension of  $U^\perp$  is two because it describes a two dimensional plane. A plane is described by two independent vectors.

In general, we can say that the dimension of the ambient space is the dimension of the subspace (in this case  $U$ ) plus the dimension of the orthogonal complement. From this, we gather that the dimension of the orthogonal complement must be the dimension of the ambient space minus the dimension of the subspace.

The orthogonal complement represents all directions "left over" after accounting for the directions already captured by the subspace  $V$ . If  $V$  is  $k$ -dimensional, then the orthogonal complement  $V^\perp$  must account for the remaining  $n - k$  dimensions in order to span the full  $n$ -dimensional space.

Suppose in the ambient space  $\mathbb{R}^3$ , we define the subspace  $U = \{(1, 0, 2), (1, 0, -2)\}$ . Because these two vectors are independent, the space is spanned by two vectors and the dimension of the subspace is two. From the claim above, we know the orthogonal complement must be dimension one. If  $U^\perp$  is given by

$$\begin{aligned} U^\perp &= \{(x, y, z) \in \mathbb{R}^3 : \langle (x, y, z), u \rangle = 0 \forall u \in U\} \\ &\implies x + 2y = 0 \\ &\quad x - 2y = 0. \end{aligned}$$

The only solution to this system is when  $x = 0, z = 0$ . Which means the vectors in  $U^\perp$  are

$$U^\perp = \{(0, y, 0) : y \in \mathbb{R}\}.$$

Since this set forms a line in  $\mathbb{R}^3$ ,  $U^\perp$  is one-dimensional, which is what we expected. We can see that  $U^\perp$  is one-dimensional by written  $(0, y, 0)$  as  $y(0, 1, 0)$ . Thus, we see that the space is spanned by just a single vector.

- **Orthogonality and Direct Sum:** One important property of orthogonal complements is that the entire space  $\mathbb{R}^n$  can be written as the direct sum of the subspace  $V$  and its orthogonal complement  $V^\perp$ . This means:

$$\mathbb{R}^n = V \oplus V^\perp.$$

In other words, every vector in  $\mathbb{R}^n$  can be uniquely written as the sum of one vector from  $V$  and one vector from  $V^\perp$ .

Since these two subspaces are orthogonal and span the entire space, their dimensions must add up to the dimension of the ambient space  $\mathbb{R}^n$ .

$$\dim(V) + \dim(V^\perp) = n.$$

- **Orthogonal complement of the kernel:** Recall that the kernel of a linear map acting on vector spaces,  $L : V \rightarrow W$  is the set of vectors  $v \in V$  such that  $L(v) = 0$ . Formally, for a linear map  $L : V \rightarrow W$ ,  $\ker(L) = \{v \in V : L(v) = 0\}$ . It represents the set of vectors in the domain space that get annihilated by the map. If a map is injective (one-to-one), then the only member of the kernel is the zero vector. Note that the kernel of a linear map is also well defined for linear operations of the form  $L : V \rightarrow V$ . Ie a mapping of a vector space onto itself.

We can find the orthogonal complement of the null space,

$$N^\perp(L) = \ker^\perp(L) = \{v \in V : \langle v, n \rangle = 0 \ \forall n \in N(L)\}.$$

If the kernel only contains the zero vector (the map is injective), then the orthogonal complement would be the whole space  $V$ , because every vector is perpendicular to the zero vector, even the zero vector itself.

- **Vector plane:** A vector plane (or simply a plane) refers to a two-dimensional flat surface that extends infinitely in all directions within a higher-dimensional space, such as three-dimensional space ( $\mathbb{R}^3$ ). It can be thought of as a set of all possible linear combinations of two linearly independent vectors. Formally, a vector plane can be described as the span of two non-parallel vectors  $\mathbf{v}_1$  and  $\mathbf{v}_2$  in a vector space. If these vectors belong to  $\mathbb{R}^3$ , their span is the set of all vectors of the form:

$$\mathbf{r} = c_1 \mathbf{v}_1 + c_2 \mathbf{v}_2$$

where  $c_1$  and  $c_2$  are scalars (real numbers). This defines a plane that passes through the origin.

We can find the spanning vectors of a plane given by

$$ax + by + cz = 0.$$

By first finding its normal vector  $\vec{n} = (a, b, c)$  which is normal to all points on the plane. we need to find two linearly independent vectors that lie on the plane and are perpendicular to  $\vec{n}$

To find vectors that satisfy the plane equation, we can make some simple choices by assigning values to two of the coordinates and solving for the third.

For the plane  $2x + 3y + z = 0$ , let's find two vectors that lie on this plane.

- **First vector:** Choose  $x = 1$  and  $y = 0$ . Substituting these into the plane equation:

$$2(1) + 3(0) + z = 0 \Rightarrow 2 + z = 0 \Rightarrow z = -2$$

So one vector on the plane is:

$$\mathbf{v}_1 = (1, 0, -2)$$

- **Second vector:** Now choose  $x = 0$  and  $y = 1$ . Substituting these into the plane equation:

$$2(0) + 3(1) + z = 0 \Rightarrow 3 + z = 0 \Rightarrow z = -3$$

So another vector on the plane is:

$$\mathbf{v}_2 = (0, 1, -3)$$

These two vectors,  $\mathbf{v}_1 = (1, 0, -2)$  and  $\mathbf{v}_2 = (0, 1, -3)$ , are linearly independent and lie on the plane, so they span the plane.

- **Affine plane, Affine space:** An affine space is a geometric structure that is similar to a vector space but lacks a fixed origin. In an affine space, points can be connected by vectors, but there is no specific "starting point" or origin like in a vector space.

Affine spaces allow for translations, you can shift the entire plane by adding a vector to all points, and the result will still be the same affine plane.

In  $\mathbb{R}^3$ , an affine plane is essentially a shifted version of a vector plane (which passes through the origin).

A typical affine plane in  $\mathbb{R}^3$  can be described by an equation of the form:

$$ax + by + cz = d.$$

where  $a$ ,  $b$ , and  $c$  are constants that define the orientation of the plane, and  $d$  is a constant that determines its position relative to the origin.

If  $d = 0$ , the plane passes through the origin, and it becomes a vector plane. When  $d \neq 0$ , it is a parallel shift of the vector plane, making it an affine plane.

An affine plane in  $\mathbb{R}^n$  has the same properties as a Euclidean plane in terms of being flat and infinite in two directions.

However, it doesn't have a fixed origin, so you cannot perform operations like vector addition directly between points in an affine plane.

Instead, you can measure distances, angles, and define lines in an affine plane, similar to Euclidean geometry.

Since the affine plane does not pass through the origin, it must be described by a point on the plane and two direction vectors that span the plane. A general form for a plane in  $\mathbb{R}^3$  is:

$$\mathbf{r} = \mathbf{r}_0 + c_1\mathbf{v}_1 + c_2\mathbf{v}_2$$

where  $\mathbf{r}_0$  is a point on the plane, and  $\mathbf{v}_1$  and  $\mathbf{v}_2$  are two direction vectors as before.

Consider the affine plane equation:

$$2x + 3y + z = 6$$

**Find a point on the plane:**

- Set  $x = 0$  and  $y = 0$ :

$$2(0) + 3(0) + z = 6 \Rightarrow z = 6$$

So, one point on the plane is  $\mathbf{r}_0 = (0, 0, 6)$ .

**Find two spanning vectors:**

- To find the first vector, set  $x = 1$  and  $y = 0$ :

$$2(1) + 3(0) + z = 6 \Rightarrow 2 + z = 6 \Rightarrow z = 4$$

So, another point on the plane is  $(1, 0, 4)$ . The vector from  $\mathbf{r}_0$  to this point is:

$$\mathbf{v}_1 = (1 - 0, 0 - 0, 4 - 6) = (1, 0, -2)$$

- To find the second vector, set  $x = 0$  and  $y = 1$ :

$$2(0) + 3(1) + z = 6 \Rightarrow 3 + z = 6 \Rightarrow z = 3$$

So, another point on the plane is  $(0, 1, 3)$ . The vector from  $\mathbf{r}_0$  to this point is:

$$\mathbf{v}_2 = (0 - 0, 1 - 0, 3 - 6) = (0, 1, -3)$$

- **Kernal, Image, and the transpose:** For a linear map  $L : V \rightarrow W$ , where  $V, W$  are vector spaces

$$\begin{aligned} \text{Ker}(L) &\subset V & \text{Im}(L) &\subset W \\ \text{Ker}(L^T) &\subset W & \text{Im}(L^T) &\subset V. \end{aligned}$$

We also state: if  $V, W$  are inner spaces, then they have a orthogonal complement, and

$$\begin{aligned} \text{Ker}^\perp(L) &\subset V & \text{Im}^\perp(L) &\subset W \\ \text{Ker}(L^T)^\perp &\subset W & \text{Im}^\perp(L^T) &\subset V. \end{aligned}$$

- **Transpose in inner spaces:** Suppose we have a linear map  $L : V \rightarrow W$ , where  $V, W$  are inner spaces. We know  $L : V \rightarrow W$ , and  $L^T : W \rightarrow V$ . Over in  $W$ , we have the inner product  $\langle L(v), w \rangle_W$ , and in  $V$  we have  $\langle v, L^T(w) \rangle_V$ . We assert

$$\langle L(v), w \rangle_W = \langle v, L^T(w) \rangle_V.$$

- **Definition of the transpose:** The **adjoint** (or transpose) of a linear map  $L : V \rightarrow W$  between two inner product spaces  $V$  and  $W$ , denoted  $L^T : W \rightarrow V$ , is the unique linear map that satisfies the following condition for all  $v \in V$  and  $w \in W$ :

$$\langle L(v), w \rangle_W = \langle v, L^T(w) \rangle_V$$

In other words, the adjoint  $L^T$  transfers the action of  $L$  across the inner product while preserving the result.

- **Adjoint in linear operations**  $L : V \rightarrow V$
- **Relating the orthogonal complement to the transpose of the image (Row space):** We assert  $\text{Ker}^\perp(L) = \text{Row}(L) = \text{Im}(L^T)$ . If the linear map is injective, then the kernel only has the zero vector. Thus, it seems trivial that all members of the row space are orthogonal to the kernel. When the kernel (or null space) of a matrix is not just the zero vector, the relationship between the row space and the kernel becomes more interesting, but the key idea still holds: the row space is orthogonal to the entire kernel, not just the zero vector.

When we think about a matrix  $A$  acting on a vector  $\mathbf{x}$ , we are performing matrix multiplication. This operation is done by taking the dot product of the vector  $\mathbf{x}$  with each row of the matrix  $A$ . Each row of the matrix is treated as a vector in this context. Now, if  $\mathbf{x}$  is in the kernel of  $A$ , this means that multiplying  $A$  by  $\mathbf{x}$  results in the zero vector, i.e.,  $A\mathbf{x} = \mathbf{0}$ . This tells us that for each row  $\mathbf{r}_i$  of the matrix  $A$ , the dot product between  $\mathbf{x}$  and  $\mathbf{r}_i$  must be zero. In other words, for every row  $\mathbf{r}_i$ ,

$$\langle \mathbf{x}, \mathbf{r}_i \rangle = 0$$

This equation must hold for all the rows of the matrix. Therefore, if every vector in the kernel has a dot product of zero with every row of the matrix, it implies that the row vectors are orthogonal to the vectors in the kernel. This is what we mean when we say that the row space (which is the space spanned by the rows) is orthogonal to the kernel.

If a vector is both in the row space and the kernel of a matrix, it must be the zero vector.

By definition, the row space and the kernel (null space) of a matrix are orthogonal subspaces. This means that every vector in the row space is orthogonal to every vector in the kernel. In other words, the dot product between any vector from the row space and any vector from the kernel is zero.

Since any non-zero vector  $v$  cannot be orthogonal to itself, the only vector that can be orthogonal to itself is the zero vector,  $v = 0$

- **Proving  $\text{ker}^\perp(L) = \text{Im}(L^T)$ :**

**Claim:**  $\text{ker}^\perp(L) = \text{Im}(L^T)$

Let  $L$  be a linear map action on two inner spaces,  $L : V \rightarrow W$ . Then,

$$L : V \rightarrow W \quad L^T : W \rightarrow V.$$

Also

$$\begin{aligned}\text{Ker}(L) &\subset V, \text{ Im}(L) \subset W \\ \text{Ker}(L^T) &\subset W, \text{ Im}(L^T) \subset V \\ \text{Ker}^\perp(L) &\subset V, \text{ Im}^\perp(L) \subset W \\ \text{Ker}^\perp(L^T) &\subset W, \text{ Im}^\perp(L^T) \subset V \\ &\vdots\end{aligned}$$

And, we saw above, by definition of the adjoint

$$\langle L(v), w \rangle_W = \langle v, L^T(w) \rangle_V.$$

We start by showing that  $\text{Im}(L^T) \subset \text{ker}^\perp(L)$ . If we can show this, and  $\text{ker}^\perp(L) \subset \text{Im}(L^T)$ . Then the two sets must be equal. To show that  $\text{Im}(L^T) \subset \text{ker}^\perp(L)$ , we choose an element  $v \in \text{Im}(L^T)$ , and show that it must be in  $\text{ker}^\perp(L)$ . We know that if  $v \in \text{Im}(L^T)$ , then  $v = L^T(w)$ . By definition of the orthogonal complement

$$\text{ker}^\perp(L) = \{v \in V : \langle v, u \rangle = 0 \quad \forall u \in \text{ker}(L)\}.$$

Thus, for  $v$  to be in the orthogonal complement of the kernel of  $L$ , it must be that  $\langle v, u \rangle = 0$  for all vectors in the kernel.

$$\begin{aligned}\langle v, u \rangle &= 0 \\ \implies \langle L^T(w), u \rangle &= 0 \\ \implies \langle w, L(u) \rangle &= 0 \\ \implies \langle w, 0 \rangle &= 0 \\ \implies 0 &= 0 \\ \therefore v &\in \text{ker}^\perp(L).\end{aligned}$$

**Lemma 1 (Orthogonal Complement of a Subset Relationship).** If  $A \subset B$ , then  $B^\perp \subset A^\perp$ . This "reversal" of subset direction is due to how orthogonality works; elements in  $B^\perp$  are orthogonal to everything in  $B$ , so they must also be orthogonal to everything in  $A$ , making them elements of  $A^\perp$ .

**Lemma 2 (Double Orthogonal Complement Property).** In a finite-dimensional space, the double orthogonal complement of a set is the closure of the original set itself, i.e.,  $(\text{Ker}(L)^\perp)^\perp = \text{Ker}(L)$  and similarly for other subspaces. This property lets you "recover" the original space by taking the orthogonal complement twice, provided the space is closed (which holds in finite dimensions).

Since the first claim holds true, we must now show that  $\text{ker}^\perp(L)\text{Im}(L^T)$ . However, we can instead show  $\text{ker}^{\perp\perp}(L) \subset \text{Im}^\perp(L^T)$ . By taking the orthogonal complement of both sides. But, we must reverse the direction of the subset. Thus, we have

$$\text{Im}^\perp(L^T) \subset \text{ker}(L).$$

Let  $v$  be a member of  $\text{Im}(L^T)^\perp$ . Then,  $\langle v, u \rangle = 0 \quad \forall u \in \text{Im}(L^T)$ . Let  $w \in W$  be arbitrary. We know that if  $u \in \text{Im}(L^T)$ , then  $u = L^T(w)$ , thus

$$\begin{aligned}\langle v, u \rangle &= 0 \\ \implies \langle v, L^T(w) \rangle &= 0 \\ \implies \langle L(v), w \rangle &= 0.\end{aligned}$$

Let  $w = L(v)$ , then

$$\begin{aligned}\langle L(v), L(v) \rangle &= 0 \\ \implies |L(v)|^2 &= 0.\end{aligned}$$

Thus,  $L(v)$  must be the zero vector,  $|L(v)|^2 = 0 \implies 0 = 0$ , therefore  $v \in \ker(L)$ .

- **Corollary from the proof above:** We have shown that

$$\ker(L)^\perp = \text{Im}(L^T).$$

From this, we gather

$$\begin{aligned}\ker(L^T)^\perp &= \text{Im}(L^{TT}) \\ \ker(L^T)^{\perp\perp} &= \text{Im}(L)^\perp.\end{aligned}$$

This corollary states that any vector perpendicular to a vector in the image of  $L$ , must be in the kernel of the transpose of  $L$ . The vectors that get annihilated by the transpose are orthogonal to the range space.

### 6.1.12 More on dimensions and the rank nullity theorem

- **Dimension of the null space and the range space:** We know

$$\begin{aligned}\ker(L) &= \{v \in V : L(v) = 0\} \subset V \\ \text{Im}(L) &= \{w \in W : w = L(v)\} \subset W.\end{aligned}$$

But what are the dimensions of these spaces? We assert

$$\begin{aligned}0 \leq \dim(\ker)(L) &\leq \dim(V) \\ 0 \leq \dim(\text{Im})(L) &\leq \dim(W).\end{aligned}$$

- **Rank nullity theorem:** The **Rank-Nullity Theorem** states that for a linear map  $L : V \rightarrow W$  between two vector spaces  $V$  and  $W$ , the dimension of the domain  $V$  is the sum of the rank and the nullity of  $L$ . Mathematically, it is expressed as:

$$\dim(V) = \dim(\ker(L)) + \dim(\text{Im}(L))$$

Where:

- $\dim(V)$  is the dimension of the domain vector space  $V$ ,
- $\dim(\ker(L))$  is the **nullity** of  $L$ , i.e., the dimension of the **null space** (the set of vectors in  $V$  that map to the zero vector in  $W$ ),
- $\dim(\text{im}(L))$  is the **rank** of  $L$ , i.e., the dimension of the **image** (the set of all vectors in  $W$  that are the image of some vector in  $V$ ).

In simpler terms, the dimension of the vector space  $V$  is the sum of the number of independent vectors that are mapped to zero and the number of independent vectors that are mapped to non-zero vectors.

Thus, the dimensions of the kernel and image of a linear map must sum to the dimension of the domain.

**Example:** Suppose we have the map  $L : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ . Then

$$\begin{aligned}\dim(\ker)(L) &\in \{0, 1, 2, 3\} \\ \dim(\text{Im})(L) &\in \{0, 1, 2\}.\end{aligned}$$

But, since the two dimensions must sum to 3, the dimension of the kernel cannot be one. From this we gather that the map cannot be injective, if the kernel has dimension greater than zero, there are multiple vectors that get sent to zero by the map, and thus the map is not one-to-one.

Suppose  $\dim(\ker)(L) = 1$ , then  $\dim(\text{Im})(L) = 2$ . Since the dimension of the range space equals the dimension of the codomain, the map is surjective. Since the dimension of the kernel is still greater than zero, the map is not injective.

Suppose  $\dim(\ker)(L) = 2$ , then  $\dim(\text{Im})(L) = 1$ . Thus, not injective or surjective. This is also the case for  $\dim(\ker)(L) = 3$ ,  $\dim(\text{Im})(L) = 0$ . In this case, everything in the domain is sent to zero. This map is known as the *zero map*, and its image is known as the *trivial image*. Also, the kernel is the entire vector space  $V$ . This map is *fully compressive*, collapsing the entire structure of  $V$  into just the zero vector in  $W$ .

- **Rank nullity theorem with orthogonality:** Recall we proved the theorem

$$\begin{aligned}\ker^\perp(L) &= \text{Im}(L^T) \\ \ker(L^T) &= \text{Im}^\perp(L).\end{aligned}$$

Suppose  $L : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ ,  $L^T : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ . Then,

$$\dim(\ker)(L^T) + \dim(\text{Im})(L^T) = 2.$$

- **Rank nullity theorem example with matrix:** Suppose we have a linear map  $L : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ , define  $A = \begin{pmatrix} 0 & 1 & 2 \\ -1 & 0 & 1 \end{pmatrix}$ . To find the dimension of the kernel.

$$\begin{pmatrix} 0 & 1 & 2 \\ -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\implies \begin{cases} y + 2z = 0 \\ -x + z = 0 \end{cases}.$$

Using gaussian elimination, we find

$$\begin{aligned} x - z &= 0 \implies x = z \\ y + 2z &= 0 \implies y = -2z. \end{aligned}$$

Let  $z = \alpha$ , then

$$\begin{aligned} x &= \alpha \\ y &= -2\alpha. \end{aligned}$$

And

$$\vec{r} = \begin{pmatrix} \alpha \\ -2\alpha \\ \alpha \end{pmatrix} = \alpha \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}.$$

Thus, the dimension of the kernel is one as it is described by a single vector. Since the kernel has dimension one, the image must have dimension two.

Since the kernel is spanned by the vector  $\begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}$ , this is a decent choice of basis, we could make it a unit vector by dividing by its norm. Thus,

$$b_1 = \frac{1}{\sqrt{6}} \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}.$$

Would be a basis for the kernel. Furthermore, since the image has dimension two, the image spans the entire codomain and the map is therefore onto. The standard basis for  $\mathbb{R}^2$  is  $(1, 0), (0, 1)$ .

- **Rank nullity theorem with matrix:** Suppose we have the linear map  $L : \mathbb{R}^5 \rightarrow \mathbb{R}^3$ , described by the matrix

$$A = \begin{pmatrix} -1 & 0 & 1 & 2 & 3 \\ 2 & -1 & 0 & 1 & -1 \\ 1 & -1 & 1 & 3 & 2 \end{pmatrix}.$$

Then by the guassian elimination process, we find the kernel described by the equations

$$\begin{aligned} x_1 &= x_3 + 2x_4 + 3x_5 \\ x_2 &= 2x_3 + 5x_4 + 5x_5. \end{aligned}$$

Since in this case we have three free variables,  $x_3, x_4, x_5$ , the dimension of the kernel is three, and the dimension of the image is therefore two.

Let  $x_3 = \alpha, x_4 = \beta, x_5 = \gamma$ , then

$$\begin{aligned} x_1 &= \alpha + 2\beta + 3\gamma \\ x_2 &= 2\alpha + 5\beta + 5\gamma. \end{aligned}$$

Thus, a vector in the kernel can be expressed as

$$\vec{r} = \alpha \begin{pmatrix} 1 \\ 2 \\ 1 \\ 0 \\ 0 \end{pmatrix} + \beta \begin{pmatrix} 2 \\ 5 \\ 0 \\ 1 \\ 0 \end{pmatrix} + \gamma \begin{pmatrix} 3 \\ 5 \\ 0 \\ 0 \\ 1 \end{pmatrix}.$$

Since a vector in the kernel is described by three independent vectors, the surface is spanned by three vectors, and the dimension is therefore three, which describes a 3-dimensional surface embedded in  $\mathbb{R}^5$ .

- **The image:** We can find a description of the image by row reducing the matrix associated with a linear map. The number of leading zeros in the rows can give us not only the dimension of the image, but also a basis by using the columns from the original matrix where the leading ones appear in the row reduced matrix. Consider an example.

Suppose we have the matrix

$$A = \begin{pmatrix} 1 & 2 & 0 & 1 & 0 \\ 1 & 2 & 1 & 2 & 1 \\ 2 & 4 & 1 & 3 & 1 \\ 3 & 6 & 2 & 5 & 1 \end{pmatrix}.$$

Which represents a linear map  $L : \mathbb{R}^5 \rightarrow \mathbb{R}^4$  with respect to the standard basis. Row reducing this matrix yields

$$\begin{pmatrix} 1 & 2 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

As we can see, we have three leading ones, in columns 1,3, and 5. Thus, the dimension of the image is three. Furthermore, we can use these columns from the original matrix as a basis for the image. Thus,

$$\begin{pmatrix} 1 \\ 1 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

Is a basis for the image. The other columns are linearly dependent on these three vectors. Furthermore, by the rank-nullity theorem

$$\begin{aligned} \dim(\ker) &= \dim(V) - \dim(Im) \\ \implies \dim(\ker) &= 5 - 3 = 2. \end{aligned}$$

We can find a description for the kernel by taking a further look at the row reduced matrix above, we have

$$\begin{aligned}x_1 &= -2x_2 - x_4 \\x_3 &= -x_4 \\x_5 &= 0.\end{aligned}$$

We see that we have two free variables  $x_2$  and  $x_4$ . Thus, we see from this that the dimension is two, which verifies what we gathered from the rank-nullity theorem. We have

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} -2x_2 - x_4 \\ x_2 \\ -x_4 \\ x_4 \\ 0 \end{pmatrix}.$$

To form a basis, we need to find two vectors from the information above. First, let

$x_2 = 1, x_4 = 0$ . Which gives the vector  $v_1 = \begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$ , let  $x_2 = 0$ , and  $x_4 = 1$ , which

gives  $v_2 = \begin{pmatrix} -1 \\ 0 \\ -1 \\ 1 \\ 0 \end{pmatrix}$ . These two vectors form a basis for the kernel. Thus, a vector in the kernel is expressed as

$$n = x_2 \begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} + x_4 \begin{pmatrix} -1 \\ 0 \\ -1 \\ 1 \\ 0 \end{pmatrix}.$$

To find a description of the orthogonal complement, we can instead find a description of the row space  $\text{Im}(L^T)$ , since we know  $\ker(L)^\perp = \text{Im}(L^T)$ . Since the dimension of the kernel is two, the dimension of the orthogonal complement of the kernel is three.

- **Rank theorem:** Let  $A$  be a matrix that represents a linear map  $L$ , the rank of the matrix is given by

$$\text{Rank}(A) = \dim(\text{Im})(L).$$

Furthermore

$$\text{Rank}(A) = \text{Rank}(A^T).$$

**6.1.13 Linear algebra with complex numbers**

- **Intro:** Matrices generally stay the same when we introduce complex numbers. Consider the matrix  $\begin{pmatrix} 1 & i \\ 0 & 1 \end{pmatrix}$

$$\begin{pmatrix} 1 & i \\ 0 & 1 \end{pmatrix}^T = \begin{pmatrix} 1 & 0 \\ i & 1 \end{pmatrix}$$
$$\begin{pmatrix} 1 & i \\ 0 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & -i \\ 0 & 1 \end{pmatrix}.$$

### 6.1.14 Proofs

- **Linear maps  $L : \mathbb{R} \rightarrow \mathbb{R}^2$ :** We know for a map to be linear,

$$L : \vec{v} \rightarrow \vec{v}.$$

Which is  $L : \mathbb{R} \rightarrow \mathbb{R}^2$

Then has the properties

$$\begin{aligned} L(a\vec{v}) &= aL(\vec{v}) \\ L(\vec{v} + \vec{w}) &= L(\vec{v}) + L(\vec{w}). \end{aligned}$$

These properties also hold when two different vector spaces are involved, ex:  $L : \vec{v} \rightarrow \vec{w}$

Suppose we have

$$L_1(x) = (a_1x, a_2x).$$

Let's show that  $L_1(ax) = aL_1(x)$

$$\begin{aligned} L_1(ax) &= (a_1(ax), a_2(ax)) \\ &= ((a_1a)x, (a_2a)x) \\ &= (a(a_1x), a(a_2x)) \\ &= a(a_1x, a_2x) \\ &= aL_1(x). \end{aligned}$$

Next, let's show  $L_1(x + x') = L_1(x) + L_1(x')$

$$\begin{aligned} L_1(x + x') &= (a_1(x + x'), a_2(x + x')) \\ &= (a_1x + a_1x', a_2x + a_2x') \\ &= (a_1x, a_2x) + (a_1x' + a_2x') \quad (\text{Vector addition}) \\ &= L_1(x) + L_1(x'). \end{aligned}$$

- **Linear maps  $L : \mathbb{R}^2 \rightarrow \mathbb{R}$ :**

For the first property,  $L(ax) = aL(x)$ , we must show that  $L(a(x_1, x_2)) = aL(x_1, x_2)$

$$\begin{aligned} L(x_1, x_2) &= a_1x_1 + a_2x_2 \\ \implies L(a(x_1, x_2)) &= L(ax_1, ax_2) = a_1(ax_1) + a_2(ax_2) \\ &= a(a_1x_1) + a(a_2x_2) \\ &= a(a_1x_1 + a_2x_2) \\ &= aL(x_1, x_2). \end{aligned}$$

For the second property,  $L(x_1 + x_2) = L(x_1) + L(x_2)$ , we show that  $L((x_1, x_2) + (x'_1, x'_2)) = L(x_1) + L(x_2)$ . If we define two vectors  $\vec{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$  and  $\vec{x}' = \begin{bmatrix} x'_1 \\ x'_2 \end{bmatrix}$ , then we have  $L(\vec{x} + \vec{x}')$ .

Using vector addition we then have  $L(x_1 + x'_1, x_2 + x'_2)$ , then

$$\begin{aligned} L(\vec{x} + \vec{x}') &= L(x_1 + x'_1, x_2 + x'_2) \\ &= a_1(x_1 + x'_1) + a_2(x_2 + x'_2) \\ &= a_1x_1 + a_1x'_1 + a_2x_2 + a_2x'_2 \\ &= a_1x_1 + a_2x_2 + a_1x'_1 + a_2x'_2 \\ &= L(x_1, x_2) + L(x'_1, x'_2) = L(\vec{x}) + L(\vec{x}'). \end{aligned}$$

- **Linear maps  $L : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ :**

First, we show  $L : \mathbb{R}^2 \rightarrow \mathbb{R}^2 = L(x_1, x_2) = (ax_1 + bx_2, cx_1 + dx_2)$  has the property  $L(kx) = kL(x)$ . We have

$$\begin{aligned} L(k(x_1, x_2)) &= L(kx_1, kx_2) \\ &= (a(kx_1) + b(kx_2), c(kx_1) + d(kx_2)) \\ &= (k(ax_1) + k(cx_2), k(cx_1) + k(dx_2)) \\ &= (k(ax_1 + bx_2), k(cx_1 + dx_2)) \\ &= k(ax_1 + bx_2, cx_1 + dx_2) \\ &= kL(x_1, x_2). \end{aligned}$$

Next, we show  $L(x + x') = L(x) + L(x')$ . Thus, we show  $L((x_1, x_2) + (x'_1, x'_2)) = L(x_1, x_2) + L(x'_1, x'_2)$ . If  $L(x_1, x_2) = (ax_1 + bx_2, cx_1 + dx_2)$ , then  $L((x_1, x_2) + (x'_1, x'_2)) = L((x_1 + x'_1, x_2 + x'_2)$ , which is then

$$\begin{aligned} L((x_1 + x'_1, x_2 + x'_2)) &= (a(x_1 + x'_1) + b(x_2 + x'_2), c(x_1 + x'_1) + d(x_2 + x'_2)) \\ &= (ax_1 + ax'_1 + bx_2 + bx'_2, cx_1 + cx'_1 + dx_2 + dx'_2) \\ &= (ax_1, cx_1) + (ax'_1, cx'_1) + (bx_2, dx_2) + (bx'_2, dx'_2) \\ &= (ax_1, cx_1) + (bx_2, dx_2) + (ax'_1, cx'_1) + (bx'_2, dx'_2) \\ &= (ax_1 + bx_2, cx_1 + dx_2) + (ax'_1 + bx'_2, cx'_1 + dx'_2) \\ &= L(x_1, x_2) + L(x'_1, x'_2). \end{aligned}$$

- **Find the inverse matrix:** We want a matrix  $\begin{bmatrix} w & x \\ y & z \end{bmatrix}$  such that

$$\begin{bmatrix} w & x \\ y & z \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = I.$$

Thus,

$$\begin{bmatrix} aw + by & ax + bz \\ cw + dy & cx + dz \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Which implies

$$\begin{aligned} aw + by &= 1 \\ cw + dy &= 0 \\ ax + bz &= 0 \\ cx + dz &= 1. \end{aligned}$$

Solve the second equation for  $w$  and plug into first

$$\begin{aligned} w &= -\frac{dy}{c} \\ \implies a\left(-\frac{dy}{c}\right) + by &= 1 \\ \implies -\frac{ady}{c} + \frac{bcy}{c} &= 1 \\ \implies \frac{y(-ad + bc)}{c} &= 1 \\ \implies y &= \frac{c}{-ad + bc} \\ \implies y &= -\frac{c}{ad - bc}. \end{aligned}$$

Plug expression for  $y$  into  $w = -\frac{dy}{c}$

$$\begin{aligned} w &= -\frac{d \left( -\frac{c}{ad-bc} \right)}{c} \\ &= \frac{d}{ad-bc}. \end{aligned}$$

Solve the third equation for  $x$  and plug into fourth

$$\begin{aligned} x &= -\frac{bz}{a} \\ \implies c \left( -\frac{bz}{a} \right) + dz &= 1 \\ \implies z &= \frac{a}{ad-bc}. \end{aligned}$$

Plug expression for  $z$  into  $x = -\frac{bz}{a}$

$$\begin{aligned} x &= -\frac{b \left( \frac{a}{ad-bc} \right)}{a} \\ &= \frac{-b}{ad-bc}. \end{aligned}$$

Thus,

$$\begin{aligned} \begin{bmatrix} w & x \\ y & z \end{bmatrix} &= \begin{bmatrix} \frac{d}{ad-bc} & \frac{-b}{ad-bc} \\ \frac{-c}{ad-bc} & \frac{a}{ad-bc} \end{bmatrix} \\ &= \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}. \end{aligned}$$

- Let  $A$  be an  $n \times n$  real matrix. It could happen that  $A^T = A^{-1}$ .

**Claim:** If  $A^T = A^{-1}$ , then  $\det(A) = 1$ .

**Proof:** Let's first consider what we know about the inverse of a matrix,

$$\begin{cases} A^{-1}A = I \\ AA^{-1} = I \end{cases}.$$

Furthermore, we know some facts about the determinant, namely

$$\det(AB) = \det(A)\det(B).$$

Combining these facts, it must be true that

$$\det(A)\det(A^{-1}) = \det(AA^{-1}) = \det(I) = 1.$$

But  $A^{-1} = A^T$ , thus

$$\det(A)\det(A^T) = 1.$$

Since  $\det(A) = \det(A^T)$ , we have

$$\begin{aligned} \det(A)\det(A) &= 1 \\ \implies (\det(A))^2 &= 1 \\ \therefore \det(A) &= \pm 1. \end{aligned}$$

Consider the family of  $2 \times 2$  matrices with determinant one, the rotation matrix

$$\begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}$$

$$\begin{aligned} \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}^{-1} &= \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix} \\ \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}^{-1} T &= \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix}^{-1}. \end{aligned}$$

We have established the proposition... If  $A^T = A^{-1}$  then  $\det(A) = \pm 1$ . But what about the converse, if  $\det(A) = \pm 1$ , does  $A^{-1} = A^T$ . We suspect this may not be true, so we try to prove by counter example. Consider the matrix  $\begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix}$

$$\begin{aligned} \begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix}^{-1} &= \begin{pmatrix} 1 & -x \\ 0 & 1 \end{pmatrix} \\ \begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix}^T &= \begin{pmatrix} 1 & 0 \\ x & 1 \end{pmatrix}. \end{aligned}$$

Thus, the statement is false when  $x \neq 0$ , and the converse must not be true.

- $V$  is a vector space. Let  $B = \{v_1, \dots, v_n\} \subset V$  be a collection of vectors that spans  $V$ . Assume there is no collection of vectors with fewer than  $n$  members that also spans  $V$ . Show that this assumption is true.

To prove this, we need to show that the collection of vectors  $B = \{v_1, \dots, v_n\}$  is linearly independent. If  $B$  spans  $V$  and has the minimal number of vectors required to span  $V$ , then  $B$  must be a basis of  $V$ , meaning it is both linearly independent and spans  $V$ .

Suppose  $a_1v_1 + \dots + a_nv_n = 0_V$ , which should only happen if  $a_1, \dots, a_n = 0$ . Assume, for contradiction, that one of the coefficients is nonzero; suppose  $a_1 \neq 0$ . Then we can divide by  $a_1$  to rewrite this as

$$\begin{aligned} v_1 + \frac{a_2}{a_1}v_2 + \dots + \frac{a_n}{a_1}v_n &= 0 \\ \implies v_1 &= -\frac{a_2}{a_1}v_2 - \dots - \frac{a_n}{a_1}v_n. \end{aligned}$$

This implies that  $v_1$  is a linear combination of the other vectors in  $B$ , which contradicts our assumption that  $B$  is the minimal spanning set. Therefore, all coefficients  $a_1, \dots, a_n$  must be zero, proving that  $B$  is linearly independent.

## 6.2 By The Book

### 6.3 More on calculus with linear algebra

- **Hessian matrix:** The Hessian matrix is a square matrix of second-order partial derivatives of a scalar-valued function. It is used in multivariable calculus to study the local curvature of a function and plays a critical role in optimization problems for determining the nature of critical points (local minima, local maxima, or saddle points). For a function  $f(x_1, x_2, \dots, x_n)$  of  $n$  variables, the Hessian matrix  $H(f)$  is the  $n \times n$  matrix of second-order partial derivatives, given by:

$$H(f) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}$$

The Hessian matrix provides information about the curvature of the function in all directions. Each element  $H_{ij}$  in the matrix represents the second-order partial derivative of  $f$  with respect to variables  $x_i$  and  $x_j$ .

- **Local min and max of a three variable function  $f(x, y, z)$ , second derivative test:** We use the hessian matrix. First we find the critical points

$$\nabla f = \left( \frac{\delta f}{\delta x_1}, \frac{\delta f}{\delta x_2}, \dots, \frac{\delta f}{\delta x_n} \right) = \mathbf{0}.$$

Then, we form the Hessian matrix

$$H(f) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}.$$

Substitute the critical points into the Hessian matrix to get a matrix of constants

- If the Hessian is positive definite (all eigenvalues are positive), the critical point is a local minimum.
- If the Hessian is negative definite (all eigenvalues are negative), the critical point is a local maximum.
- If the Hessian is indefinite (some eigenvalues are positive and others are negative), the critical point is a saddle point.
- If any eigenvalue is zero, the test is inconclusive.

Alternatively, you can check definiteness by evaluating the leading principal minors (determinants of the upper-left submatrices):

- If all principal minors are positive, the Hessian is positive definite (local minimum).
- If the signs alternate, the Hessian is negative definite (local maximum).
- If the signs are mixed, the critical point is a saddle point.

**6.4 Just Formulas and Theorems**

# Combinatorics

## 7.1 Introduction

- **What is combinatorics?**: Combinatorics is a collection of techniques and a language for the study of finite or countably infinite discrete structures. Given a set of elements and possibly some structure on that set, typical questions are
  - Does a specific arrangement of the elements exists?
  - How many such arrangements are there?
  - What properties do these arrangements have?
  - Which one of the arrangements is maximal, minimal, or optimal according to some criterion?
- **Counting the number of subsets for a set**: Let  $[n] = \{1, 2, \dots, n\}$ , and let  $f(n)$  be the number of subsets of  $[n]$ . Then  $f(n) = 2^n$ . For any particular subset of  $[n]$ , each element is either in that subset or not. Thus, to construct a subset, we have to make one of two choices for each element of  $[n]$ . Furthermore, these choices are independent of each other. Hence, the total number of choices, and consequently the total number of subsets is

$$\underbrace{2 \times 2 \times \dots \times 2}_n = 2^n.$$

- **Number of subsets without consecutive integers**: For a sequence  $[n] = \{1, \dots, n\}$  we can count the number of subsets given by  $f(n)$ , that do not contain consecutive integers with the recurrence relation

$$f(n) = f(n-1) + f(n-2).$$

We consider two cases

1.  $n$  is not included in the subsets
2.  $n$  is included in the subsets. In this case, we build the subsets considering the subsequence  $[n-2] = \{1, \dots, n-2\}$ . Note that if we include  $n$ , we must exclude  $n-1$ , because  $n-1$  and  $n$  are consecutive, this will become clear in the upcoming example.

Consider the sequence  $[n] = \{1, 2, 3, 4\}$ . By the relation above,

$$f(4) = f(3) + f(2).$$

Before we are able to compute this, we must define our base cases.

$$f(n) = \begin{cases} 3 & \text{if } n = 2 \\ 2 & \text{if } n = 1 \end{cases}.$$

If  $n = 2$ , we have  $\{1, 2\}$ , and the allowed subsets are  $\emptyset, \{1\}, \{2\}$ . If we have  $n = 1$ , the subsets are  $\{\emptyset, \{1\}\}$ . Thus

$$\begin{aligned} f(4) &= f(3) + f(2) = f(2) + f(1) + f(2) \\ &= 3 + 2 + 3 = 8. \end{aligned}$$

Let's explicitly break up the given sequence so we can see what's going on. In the first case,  $n$  is excluded, thus the sequence becomes  $\{1, 2, 3\}$ . If  $n$  is included, the sequence becomes  $\{1, 2\}$ , where we build the subsets of  $\{1, 2\}$ , and then add 4 to each one. Thus,

$$\begin{aligned}\{1, 2, 3\} + \{1, 2\} &= \{1, 2, 3\} + \emptyset + \{1\} + \{2\} \\ &= \{1, 2, 3\} + \{4\} + \{1, 4\} + \{2, 4\}.\end{aligned}$$

Since the sequence  $\{1, 2, 3\}$  is not a base case, we must split this one up as well, we have

$$\begin{aligned}\{1, 2, 3\} &= \{1, 2\} + \{1\} + \{4\} + \{1, 4\} + \{2, 4\} \\ &= \emptyset + \{1\} + \{2\} + \emptyset + \{1\} + \{4\} + \{1, 4\} + \{2, 4\} \\ &= \emptyset + \{1\} + \{2\} + \{3\} + \{1, 3\} + \{4\} + \{1, 4\} + \{2, 4\}.\end{aligned}$$

Thus, we conclude all "good" subsets of  $[n]$  either have  $n$  or don't have  $n$ . The ones that don't have  $n$  are exactly the "good" subsets of  $[n - 1]$ . The "good" subsets of  $[n]$  that include  $n$  are exactly the "good" subsets of  $[n - 2]$  together with  $n$ . Thus  $f(n) = f(n - 1) + f(n - 2)$  ■

## 7.2 Induction and recurrence relations

- **Principal of Mathematical Induction:** Given an infinite sequence of propositions

$$P_1, P_2, P_3, \dots, P_n, \dots.$$

In order to prove that all of them are true, it is enough to show two things

1. **The base case:**  $P_1$  is true
2. **The inductive step:** For all positive integers  $k$ , if  $P_k$  is true, then so is  $P_{k+1}$

**Example:** Show that

$$1 + 2 + 3 + \dots + n = \frac{n(n+1)}{2}.$$

**Base case:**

$$1 = \frac{1(1+1)}{2} = \frac{2}{2} = 1.$$

**Inductive step:**  $P_k$  is given by

$$1 + 2 + 3 + \dots + k = \frac{k(k+1)}{2}.$$

$P_{k+1}$  is given by

$$1 + 2 + 3 + \dots + k + k + 1 = \frac{k+1(k+2)}{2}.$$

If  $1 + 2 + 3 + \dots + k = \frac{k(k+1)}{2}$ , then

$$\begin{aligned} 1 + 2 + 3 + \dots + k + k + 1 &= \frac{k+1(k+2)}{2} \\ \frac{k(k+1)}{2} + k + 1 &= \frac{k+1(k+2)}{2} \\ \frac{k(k+1) + 2k + 2}{2} &= \frac{k^2 + 3k + 2}{2} \\ \frac{k^2 + 3k + 2}{2} &= \frac{k^2 + 3k + 2}{2}. \end{aligned}$$

Thus, we have showed that  $P_k \implies P_{k+1}$ .  $\blacksquare$ .

**Note:** Our aim is not to directly prove  $P_{k+1}$ , but to prove that  $P_k$  implies  $P_{k+1}$ . In the inductive step we assume  $P_k$  to be true, then show under this assumption,  $P_{k+1}$  is also true.

- **Understanding gauss's formula for the sum of the first  $n$  natural numbers:** Suppose we want to find the sum  $1 + 2 + 3 + \dots + (n-1) + n$ . We could have discovered the formula that we proved above by first writing the sum twice

$$\begin{aligned} 1 + 2 + 3 + \dots + (n-1) + n \\ n + (n-1) + (n-2) + \dots + 2 + 1. \end{aligned}$$

The sum of the two numbers in each column is  $n+1$ , and there are  $n$  columns, so the total sum is  $n(n+1)$ , it then follows that the actual sum is  $\frac{1}{2}n(n+1)$

- **Triangular numbers:** The sequence of integers

$$\begin{array}{ll}
 1 & 3 = 1 + 2 \\
 6 = 1 + 2 + 3 & \\
 10 = 1 + 2 + 3 + 4 & \\
 15 = 1 + 2 + 3 + 4 + 5 & \\
 \dots &
 \end{array}$$

Are called *triangular numbers*. If you were to make a triangle of dots out of the sum, where the highest number is the base, the second highest is the layer ontop of the base, etc, you would form a triangle.

- **Strong induction:** Given an infinite sequence of propositions

$$P_1, P_2, P_3, \dots, P_n.$$

In order to demonstrate that all of them are true, it is enough to know two things.

1. **The base case:**  $P_1$  is true
2. **The inductive step:** For all integers  $k \geq 1$ , if  $P_1, P_2, P_3, \dots, P_k$  are true, then so is  $P_{k+1}$

- **Pingala-fibonacci numbers:** Define a sequence of positive integers as follows:  $F_0 = 0, F_1 = 1$ , and for  $n = 2, 3, \dots$  we have

$$F_n = F_{n-2} + F_{n-1}.$$

This sequence is also known as *the fibonacci sequence*.

- **Lucas numbers:** Change the initial values on the fibonacci sequence. Let  $L_0 = 2, L_1 = 1$ , and  $L_n = L_{n-2} + L_{n-1}$ . Then, we get the *Lucas numbers*

$$2, 1, 3, 4, 7, 11, 18, 29, 47, \dots$$