

~~OFFICIAL~~



AP3456
The Central Flying School (CFS)
Manual of Flying

Volume 13 – Maths & Physics

CHAPTER 1 - FRACTIONS AND DECIMALS

Fractions

1. **Description.** Vulgar fractions are numerical quantities which are not whole numbers, expressed in terms of a numerator divided by a denominator. There are two types: proper fractions which are less than 1 and improper fractions which are greater than 1. Mixed numbers include whole numbers and vulgar fractions.

2. **Comparing Two or More Fractions.** To compare two or more fractions they must first be given the same denominator.

Example:

Arrange $\frac{5}{7}$, $\frac{11}{14}$, and $\frac{3}{4}$ in order of size.

The lowest common multiple of 7, 14 and 4 is 28.

$$\frac{5}{7} \times \frac{4}{4} = \frac{20}{28} \qquad \frac{11}{14} \times \frac{2}{2} = \frac{22}{28} \qquad \frac{3}{4} \times \frac{7}{7} = \frac{21}{28}$$

In order of size $\frac{5}{7} < \frac{3}{4} < \frac{11}{14}$

3. **Reducing a Fraction to the Lowest Terms.** Reducing a fraction to the lowest terms, or simplest form, means finding the equivalent fraction with the smallest possible numerator and denominator.

Examples:

$$\frac{12}{18} = \frac{2}{3} \text{ after dividing numerator and denominator by 6}$$

$$\frac{20}{35} = \frac{4}{7} \text{ after dividing numerator and denominator by 5.}$$

4. **Addition and Subtraction of Fractions.** If the denominators of the fractions are the same the numerators may simply be added or subtracted.

Example:

$$\frac{3}{5} + \frac{7}{5} + \frac{1}{5} = \frac{3+7+1}{5} = \frac{11}{5} = 2\frac{1}{5}$$

If the denominators are different it is necessary to find the lowest common multiple so that the fractions may be rewritten with the same denominator.

Example:

$$\frac{2}{3} + \frac{4}{5} + \frac{4}{6} \quad \text{The lowest common multiple is 30.}$$

$$\text{The fractions may be expressed as } \frac{20}{30} + \frac{24}{30} + \frac{20}{30} = \frac{64}{30} = 2\frac{2}{15}$$

5. **Multiplication of Fractions.** Fractions may be multiplied by first multiplying the numerators together and then multiplying the denominators.

Example:

$$\frac{10}{15} \times \frac{9}{7} = \frac{10 \times 9}{15 \times 7} = \frac{90}{105} = \frac{6}{7}$$

6. **Division of Fractions.** To divide one fraction by another, the divisor should be inverted and the fractions then multiplied.

Example:

$$\frac{3}{4} \div \frac{5}{7} = \frac{3}{4} \times \frac{7}{5} = \frac{21}{20} = 1 \frac{1}{20}$$

Decimals

7. **Description.** Decimals are fractions in which the denominators are powers of 10. Decimals are written using a decimal point, instead of in the fraction form.

8. **Changing Fractions to Decimals.** A fraction may be converted to a decimal by dividing the numerator by the denominator.

Example:

$$\frac{7}{8} = 7 \div 8 = 0.875$$

It is also possible to convert a fraction to a decimal by expressing the denominator as a power of 10.

Example:

By multiplying numerator and denominator by 4

$$\frac{13}{25} = \frac{52}{100}$$

and so

$$\frac{52}{100} = 0.52$$

9. **Changing Decimals to Fractions in their Lowest Terms.** To change a decimal to a fraction, the decimal should be written as a numerator with a denominator of a suitable power of 10.

Example:

Express 0.68 as a fraction.

$$0.68 = \frac{68}{100} = \frac{17}{25}$$

10. **Addition and Subtraction of Decimals.** When adding or subtracting decimals it is essential to ensure that the decimal points are in line.

Example:

$$212.2 + 14.9 + 6.3 + 0.36$$

$$\begin{array}{r} 212.2 \\ +14.9 \\ + 6.3 \\ + 0.36 \\ \hline 233.76 \end{array}$$

11. **Multiplying Decimals By Powers of 10.** When multiplying by powers of 10, the decimal point is moved one place to the right for each power of 10.

Example:

$$\begin{array}{l} 0.7 \times 10 = 7.0 \\ 0.7 \times 100 = 70.0 \\ 0.7 \times 10,000 = 7000.0 \end{array}$$

12. General Multiplication of Decimals. To multiply decimals the numbers should be multiplied together and then the number of decimal places should be counted and the point set accordingly.

Examples:

$$0.9 \times 0.007 = 0.0063$$

(decimal places $1 + 3 = 4$)

$$2.652 \times 0.04 = 0.10608$$

(decimal places $3 + 2 = 5$)

$$410 \times 0.12 = 49.20$$

(decimal places $0 + 2 = 2$)

13. Division of Decimals by Powers of 10. When dividing decimals by powers of 10 the point should be moved one place to the left for each multiple of 10.

Example:

$$0.7 \div 10 = 0.07$$

$$0.7 \div 100 = 0.007$$

$$0.7 \div 10,000 = 0.00007$$

14. General Division of Decimals. To divide decimals, both numbers should be multiplied by whatever power of 10 is required to convert the denominator into a whole number, then division may be carried out in the normal fashion.

Examples:

$4.55 \div 0.5$, may be written as

$$\frac{4.55}{0.5} \times \frac{10}{10} = \frac{45.5}{5} = 9.1$$

$42.6 \div 0.03$, may be written as

$$\frac{42.6}{0.03} \times \frac{100}{100} = \frac{4260}{3} = 1420$$

$0.0272 \div 0.4$, may be written as

$$\frac{0.0272}{0.4} \times \frac{10}{10} = \frac{0.272}{4} = 0.068$$

15. Significant Figures. If a number is given as an approximation it may be rounded to a multiple of 10. Thus, 76,282 may be given as 76,000 which is accurate to the nearest thousand or to 2 significant figures, the 7 and the 6. To 3 significant figures it would be 76,300 because 76,282 is nearer to 76,300 than to 76,200. The general rule is to consider the next digit to the right of the one to which 'significant figure' accuracy is required. If it is greater than 5, then the previous figure should be increased by one, and the appropriate number of noughts appended. If it is less than 5, then the previous figure should stand, again with the appropriate number of noughts added.

16. Decimal Places. Numbers are often rounded off or given correct to a certain number of decimal places, depending on the degree of accuracy required. A calculator may give pi as 3.141592654 which, for most purposes, will be given to 3 decimal places and written as 3.142.

17. For some fractions, the division never ends, but numbers (or a series of numbers) are repeated:

$$\frac{1}{3} = 0.3333\dots$$

$$\frac{4}{7} = 0.571428571428\dots\text{etc.}$$

Such decimals are called recurring decimals. The repeating pattern can be shown by placing a dot over the first and last digits in the recurring group:

$$\frac{1}{3} = 0.\dot{3}$$

$$\frac{4}{7} = 0.\dot{5}7142\dot{8}$$

CHAPTER 2 - PERCENTAGES AND PROPORTIONS

Definition of Percentage

1. Percent means 'per hundred'. A percentage is a fraction with a denominator of 100. Thus 13 percent means 13 divided by 100 or $\frac{13}{100}$, and is written as 13% or 13pc.

Percentages as Fractions or Decimals

2. To convert a percentage into a fraction or a decimal, it should be divided by 100.

Examples:

$$42\% \text{ expressed as a fraction} = \frac{42}{100} = \frac{21}{50} \text{ or as a decimal } 42\% = 0.42$$

$$26\frac{1}{3}\% \text{ expressed as a fraction} = \frac{26\frac{1}{3}}{100} = \frac{79}{300} \text{ or as a decimal} = 0.263$$

$$9.8\% \text{ expressed as a fraction} = \frac{9.8}{100} = \frac{98}{1000} = \frac{49}{500} \text{ or as a decimal } 0.098.$$

Fractions or Decimals as Percentages

3. To convert a fraction or decimal to a percentage it should be multiplied by 100.

Examples:

$$\frac{3}{5} \text{ as a percentage} = \frac{3}{5} \times 100\% = 60\%$$

$$\frac{7}{8} \text{ as a percentage} = \frac{7}{8} \times 100\% = 87\frac{1}{2}\%$$

$$0.62 \text{ as a percentage} = 0.62 \times 100\% = 62\%$$

Finding a Percentage

4. To find a percentage of a given quantity the quantity should first be multiplied by the required percentage and then divided by 100.

Example:

Find 36% of 180

$$180 \times \frac{36}{100} = 64.8$$

Expressing One Quantity as a Percentage of Another

5. To express one quantity as a percentage of another, first express one as a fraction of the other and then multiply by 100.

Example:

Express 49 miles as a percentage of 392 miles.

$$\frac{49}{392} \times 100 = 12.5\%$$

Percentage Increase or Decrease

6. To increase or decrease an amount by a given percentage the amount should be multiplied by the new percentage.

Examples:

Increase 650 by 6%

$$650 \times \frac{106}{100} = 689 \text{ (or } 650 \times 1.06 = 689)$$

Decrease 650 by 6%

$$650 \times \frac{94}{100} = 611 \text{ (or } 650 \times 0.94 = 611)$$

7. To find an original quantity, given a quantity which has been increased or decreased by a percentage, it is necessary to first divide the quantity by the new percentage, and then multiply by 100.

Example:

After an increase of 8% a quantity is 178, what was the original quantity?

The increased quantity is 108% of the original.

108% of original quantity is 178

$$1\% \text{ of original is } \frac{178}{108}$$

$$\text{So } 100\% \text{ of original is } \frac{178}{108} \times 100 = 164.81$$

Ratios

8. A ratio enables the comparison of two or more quantities of the same kind and is calculated by dividing one quantity by the other.

Example:

$$\text{The ratio of 375 to 500} = \frac{375}{500} = \frac{3}{4} \text{ and is written as } 3:4.$$

$$\text{To find the ratio of 5 km to 700 m: Ratio} = \frac{5000}{700} = 50 : 7$$

9. Division in a Given Ratio. To divide a quantity according to a ratio 3:4:5, the quantity is first divided by 3+4+5, then 3 parts, 4 parts and 5 parts are allocated.

Example:

Divide 2400 in the ratio 3:4:5

$$\frac{2400}{3+4+5} = \frac{2400}{12} = 200$$

Sums are 600, 800 and 1000

10. Increasing and Decreasing in a Given Ratio.

Example:

If fuel consumption of 60 kg per minute is increased in a ratio of 5:4 what is the new consumption?

$$\text{Consumption} = \frac{5}{4} \times 60 = 75 \text{ kg per min.}$$

Scales

11. If a map has a scale of 1:50,000 it means that 1 cm on the map represents 50,000 cm on the ground. In the same way 1 km on the ground is represented by

$$\frac{100,000}{50,000} \text{ cm or } 2 \text{ cm.}$$

The scale 1:50,000 could also be given as '2 cm to 1 km'.

Proportion

12. If two quantities are in direct proportion then an increase in one quantity causes a predictable increase in the other. An inversely proportional relationship means that an increase in one quantity causes a predictable decrease in the other.

Examples:

a. **Direct Proportion.**

If 400 cards cost £28 find the cost of 650 cards.

$$\text{Cost of 650 cards} = \frac{650}{400} \times 28 = £45.50$$

b. **Inverse Proportion.**

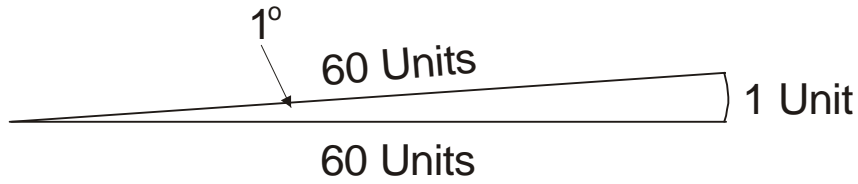
If it takes 6 men 12 days to paint a hangar, how long will it take 9 men?

$$\text{Time for 9 men} = \frac{6}{9} \times 12 = 8 \text{ days}$$

The 1 in 60 Rule

13. The 1 in 60 rule is used as a method of assessing track error and closing angle, and has long been favoured as a mental deduced reckoning (DR) navigation technique because of its flexibility, ease of use and relative accuracy (up to about 40°). The 1 in 60 rule postulates that an arc of one unit at a radius of 60 units subtends an angle of one degree (see Fig 1).

13-2 Fig 1 The 1 in 60 Rule



14. In practical use, this 1 in 60 rule may be applied equally well to a right-angled triangle. It may be accepted that, in a right-angled triangle, if the length of the hypotenuse is 60 units, the number of units of length of the small side opposite the small angle will be approximately the same as the number of degrees in the small angle (see Fig 2).

13-2 Fig 2 Application to a Right-angled Triangle



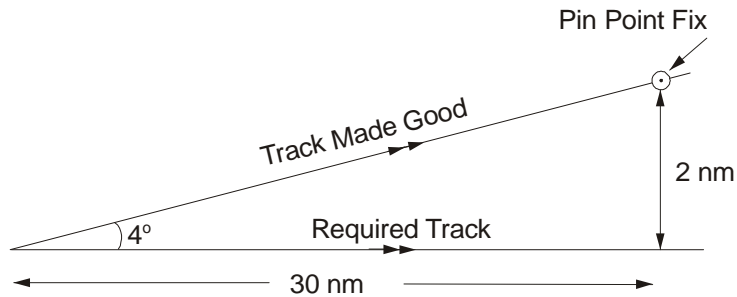
This approximation can be compared with the exact computation below:

Short Side	Sine of Angle	Angle
1 unit	1/60 = .0167	0° 57'
10 units	10/60 = .1667	9° 36'
20 units	20/60 = .3333	19° 28'
30 units	30/60 = .5000	30°
35 units	35/60 = .5833	35° 41'
40 units	40/60 = .6667	41° 49'

15. Furthermore, since the navigator is likely to have distances on the required track marked on his map, the approximation is just as good if the distance gone is measured along the required track (see Fig 3). In either case, the distance gone is compared with the distance off track and the ratio of one to the other is reduced to an angle.

$$\text{Track error (degrees)} = \frac{\text{Distance off Track} \times 60}{\text{Distance along Track}}$$

13-2 Fig 3 Calculation of Track Error



Thus, an aircraft passing over a feature 2 miles port of the required track, after flying 30 miles has a track error of:

$$\frac{2}{30} \times 60 = 4^{\circ}$$

CHAPTER 3 - AVERAGES

Introduction

1. Averages are discussed in some detail in Volume 13, Chapter 16, where it can be seen that an 'average' might mean any one of three quite different values. Of these the most useful and most commonly used is more accurately described as the arithmetic mean.

Arithmetic Mean

2. The arithmetic mean of a set of values is defined as:

$$\frac{\text{The sum of all the values}}{\text{The number of values}}$$

Examples:

a. A rugby scrum has players of weights 92 kg, 89 kg, 86 kg, 94 kg, 97.5 kg, 97 kg, 96 kg, and 95.5 kg. The arithmetic mean (or average) weight of the players may be calculated as follows:

$$\begin{aligned} \text{Average weight} &= \frac{92+89+86+94+97.5+97+96+95.5}{8} \\ &= 93.375 \text{ kg} \end{aligned}$$

b. The times taken to travel to work from Monday to Friday are 1 hr 12 min, 1 hr 18 min, 1 hr 14 min, 1 hr 21 min, and 1 hr 22 min. The average time taken in travelling to work can be calculated as follows:

$$\begin{aligned} \text{Average time} &= \frac{72+78+74+81+82}{5} \text{ mins} \\ &= 1 \text{ hr } 17.4 \text{ mins} \end{aligned}$$

3. The arithmetic mean is useful for presenting large amounts of data in a simplified form, and is most accurate when used in calculations involving data which do not include extreme values. This form of average may also yield data which are capable of further statistical analysis or mathematical treatment. It uses every value in a distribution, and is the most readily understood and commonly accepted representation of the term 'average'.

Limitations in the Use of Arithmetic Mean

4. The arithmetic mean may produce distortions because of extreme values in a distribution.

Example:

The values of stamps to be auctioned are estimated at £15, £17, £23, £24, £20, and £500. The average (arithmetic mean) of their values is given by:

$$\frac{15+17+23+24+20+500}{6} = \text{£}99.83$$

However, it would clearly be misleading to describe the stamps as being of average value of approximately £100. A more accurate and fair description would be that with one exception the average value of the stamps is approximately £20.

5. The arithmetic mean can also produce impossible quantities where data is necessarily in discrete values, (e.g. 1.825 children in an average family).

Weighted Averages

6. When calculating an average from more than one set of data, the figures cannot be combined without giving due regard to the relative sizes of the samples.

Example:

A class of 40 students score an average of 60 marks and a class of 20 students score an average 68 marks. The average might be calculated as:

$$\frac{60+68}{2} = 64 \text{ but this is clearly incorrect.}$$

The marks should be weighted according to the number of students in each group thus:

$$\frac{60(40)+68(20)}{40+20} = \frac{2400+1360}{60} = 62.66$$

This is termed the weighted average, and it gives a more accurate measure in this type of situation. Weighted averages may also be used when it is desired to give certain quantities greater importance than others within a distribution.

CHAPTER 4 - BASIC VECTOR PROCESSES

Introduction

1. Many physical quantities like mass, volume, density, temperature, work and heat, are completely specified by their magnitudes. Such quantities are known as scalar quantities or scalars. Other physical quantities possess directional properties as well as magnitudes, so that each magnitude must be associated with a definite direction in space before the physical quantity can be completely described. It is found that some, though not all, of these directed quantities possess a further common property in that they obey the same triangle (or parallelogram) law of addition. Directed quantities which obey the triangular law of addition are known as vectors.

2. **Definition of a Vector.** Any quantity which possesses both magnitude and direction, and which obeys the triangle law of addition is a vector.

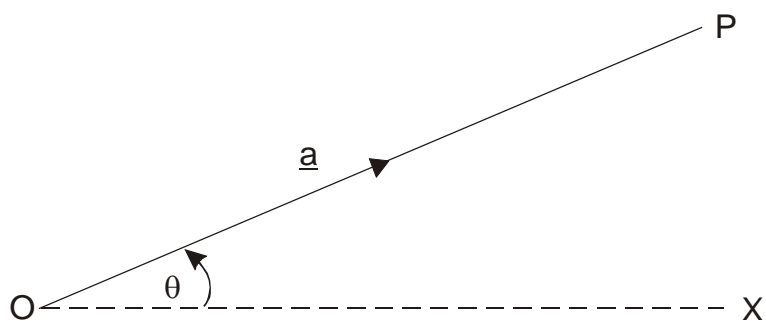
Graphical Representation of Vectors

3. A Vector quantity may be represented by a line having:

- a. Direction.
- b. Magnitude.
- c. Sense.

Fig 1 shows a vector having a direction defined by the angle θ , a magnitude equal to the length OP and a sense indicated by the arrow. The vector OP may be represented by a symbol which may be either in bold type or underlined, eg OP may be represented by \underline{a} . When a vector is shown graphically, a scale should be given.

13-4 Fig 1 Graphical Representation of a Vector



The Resultant Vector

4. The resultant of a system of vectors is that single vector which would have the same effect as the system of vectors.

The Resolution of Vectors into Components

5. It may be convenient to resolve a vector into two components acting at right angles to each other. In Fig 2, vector OP is at an angle of 30° to Ox . The vector may be divided into two components OA and OB at right angles to each other. Resolved graphically, the component OA may be measured as

6.9 units and OB as 4 units. To calculate the magnitudes of OA and OB mathematically, use is made of the trigonometrical ratio of OA to OP, thus:

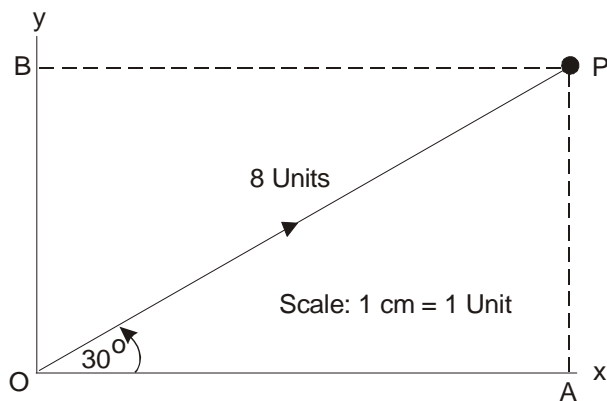
$$\frac{OA}{OP} = \cos 30^\circ$$

$$\therefore OA = OP \cos 30^\circ = 6.93 \text{ units}$$

$$\text{and } \frac{OB}{OP} = \frac{AP}{OP} = \sin 30^\circ$$

$$\therefore OB = OP \sin 30^\circ = 4 \text{ units}$$

13-4 Fig 2 Resolution of a Vector into Components

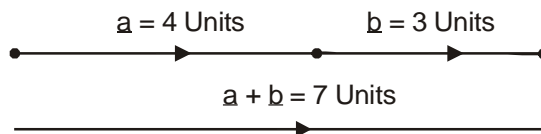


Addition of Vectors

6. **Co-linear Vectors.** The simplest case of the addition of vectors occurs when the vectors are parallel. There are two cases to consider:

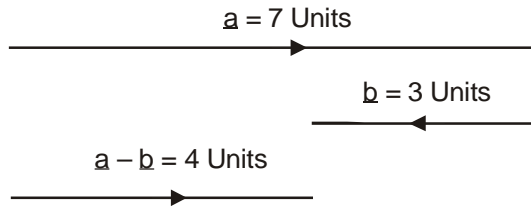
a. **Parallel Vectors Acting in the Same Direction.** Consider two forces acting on a body, one of magnitude 4 units and the other of magnitude 3 units. The two forces act in the same direction. Fig 3 shows the two vectors, \underline{a} of four units and \underline{b} of three units. The sum of the vectors is $\underline{a} + \underline{b} = 4 + 3 = 7$ units.

13-4 Fig 3 Addition of Parallel Vectors Acting in the Same Direction



b. **Parallel Vectors Acting in Opposite Directions.** When two forces acting on a body are parallel and in opposite directions the vector representation is as Fig 4. The forces are of magnitude seven units and three units. The sum of the vectors is $\underline{a} - \underline{b} = 7 - 3 = 4$ units.

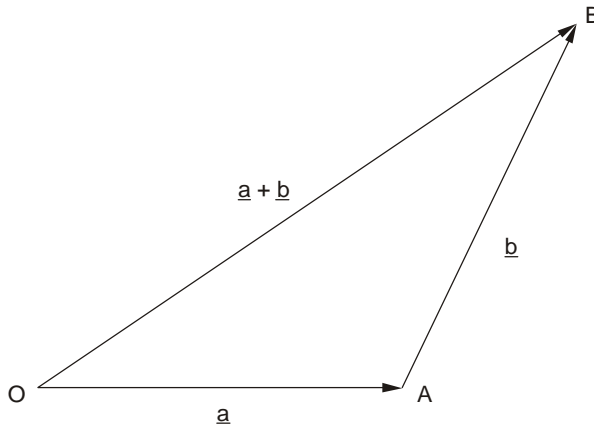
13-4 Fig 4 Addition of Parallel Vectors Acting in Opposite Directions



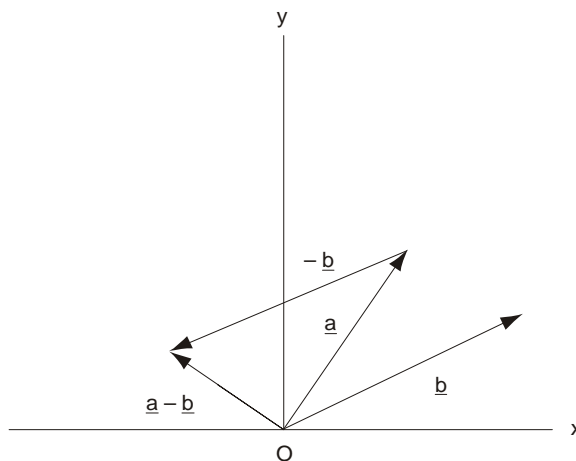
7. **Non-Co-linear Vectors.** Any two vectors may be added together. Fig 5 shows a triangle of vectors. A displacement from O to A is represented by vector \underline{a} , and a further displacement from A to B is represented by vector \underline{b} . The sum of the displacements is equivalent to a displacement from O to B, or $\underline{a} + \underline{b}$.

8. **Vector Difference.** The difference of two vectors may be represented as $\underline{a} + (-\underline{b})$, the vector $-\underline{b}$ being \underline{b} rotated through 180° as shown in Fig 6.

13-4 Fig 5 Triangle of Vectors



13-4 Fig 6 Vector Difference

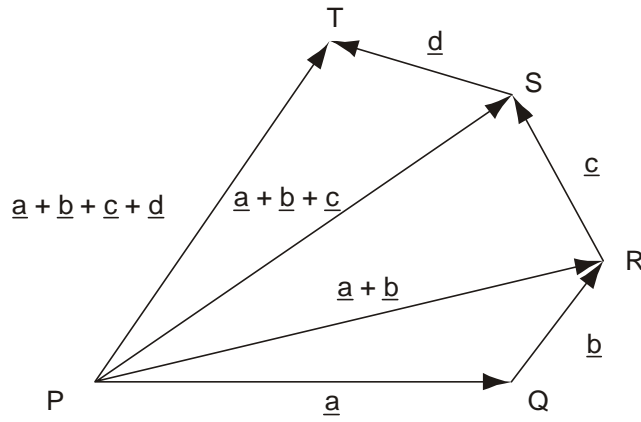


The Polygon of Vectors

9. When more than two vectors are to be resolved they can be added or subtracted two at a time as shown in Fig 7. To resolve $\underline{a} + \underline{b} + \underline{c} + \underline{d}$, PQ is drawn to represent \underline{a} , then from the terminal point of PQ, QR is drawn to represent \underline{b} , and so on until all the vectors are represented. $PR = \underline{a} + \underline{b}$ and, in the

triangle PRS, $\underline{PS} = \underline{PR} + \underline{RS} = \underline{a} + \underline{b} + \underline{c}$. In the triangle PST, $\underline{PT} = \underline{PS} + \underline{ST} = \underline{a} + \underline{b} + \underline{c} + \underline{d}$. Any number of vectors can be summed in this way.

13-4 Fig 7 The Polygon of Vectors



CHAPTER 5 - INDICES AND LOGARITHMS

INDICES

Introduction

1. When a number is successively multiplied by itself, it is said to be raised to a power. Thus:

4×4 is four raised to the power of 2 (or 4 squared)

$5 \times 5 \times 5$ is five raised to the power of 3 (or 5 cubed)

$6 \times 6 \times 6 \times 6 \times 6$ is six raised to the power of 5

This is usually written as the number that is to be multiplied, known as the base, together with the number of times it is to be multiplied as a superscript, known as the index. Thus:

$$4 \times 4 = 4^2 \quad \leftarrow \text{Index}$$

$$\quad \quad \quad \uparrow$$

$$\quad \quad \quad \text{Base}$$

and $6 \times 6 \times 6 \times 6 \times 6 = 6^5$.

The notation is not confined to actual numbers; algebraic symbols and expressions may be similarly expressed. Thus:

$$m \times m \times m \times m = m^4$$

$$\text{and, } (a + 2) \times (a + 2) \times (a + 2) = (a + 2)^3.$$

Multiplication and Division Rules

2. **Multiplication.** Suppose it is necessary to multiply 2^5 by 2^3 . Now, $2^5 = 2 \times 2 \times 2 \times 2 \times 2$ and $2^3 = 2 \times 2 \times 2$.

$$\therefore 2^5 \times 2^3 = (2 \times 2 \times 2 \times 2 \times 2) \times (2 \times 2 \times 2) = 2^8$$

i.e. the result is obtained by adding the indices, e.g. $2^{16} \times 2^8 = 2^{24}$

3. **Division.** If it is necessary to divide, say, 2^5 by 2^3 then this may be written as:

$$\frac{2 \times 2 \times 2 \times 2 \times 2}{2 \times 2 \times 2}$$

Cancelling the terms yields the result 2^2 i.e. the result is obtained by subtracting the indices, e.g. $m^{10} \div m^6 = m^4$. Consider $m^4 \div m^4$. Clearly, any number or expression divided by itself = 1. By the subtraction rule $m^4 \div m^4 = m^0$. Therefore, $m^0 = 1$. Indeed, by the same reasoning, any number or expression raised to the power zero = 1.

Negative Indices

4. Consider $2^5 \div 2^6$. This is equivalent to:

$$\frac{2 \times 2 \times 2 \times 2 \times 2}{2 \times 2 \times 2 \times 2 \times 2 \times 2} = \frac{1}{2}$$

By the division rule $2^5 \div 2^6 = 2^{-1}$. So $2^{-1} = \frac{1}{2}$, i.e. the negative index indicates a reciprocal. Similarly,

for example, $m^{-3} = \frac{1}{m^3}$

Fractional Indices

5. Consider the problem:

“What number, when multiplied by itself = 2?” Expressing this in index form, and using the multiplication rule:

$$2^a \times 2^a = 2^1$$

$$\therefore a + a = 1, \text{ i.e. } 2a = 1$$

$$\therefore a = \frac{1}{2}$$

Thus, the fractional power $\frac{1}{2}$ has the meaning of square root. Similarly, $\frac{1}{3}$ = cube root, $\frac{1}{4}$ = fourth root and so on.

Power of a Power

6. Consider the expression $(2^2)^3$. This is equivalent to:

$$(2 \times 2) \times (2 \times 2) \times (2 \times 2) = 2^6$$

The result is obtained by multiplying the indices. In general terms: $(a^m)^n = a^{mn}$

Standard and Engineering Forms

7. In science and engineering a very wide range of numerical values are frequently encountered. For example, the velocity of light is approximately 300,000,000 metres per second whilst the wavelengths of light are in the approximate range of 0.000000008 metres to 0.000000004 metres. Such very large and small numbers are clearly cumbersome in use and often difficult to comprehend quickly. In order to overcome this difficulty, it is common practice to express numbers in a standard form or in an 'engineering' form, making use of index notation.

8. **The Standard Form.** The standard form of a number consists of only one digit in front of the decimal point which is then multiplied by the appropriate power of 10, i.e. in the form:

$$A \times 10^n$$

where A is between 1.0000 and 9.9999, and the index, n, is the required power of 10. Thus, for example:

$$67.9 \text{ in standard form} = 6.79 \times 10^1$$

$$679 \text{ in standard form} = 6.79 \times 10^2$$

$$300,000,000 \text{ in standard form} = 3 \times 10^8$$

$$0.00679 \text{ in standard form} = 6.79 \times 10^{-3}$$

$$0.000000008 \text{ in standard form} = 8.0 \times 10^{-10}$$

9. **Engineering Form.** Engineering notation is also commonly available on calculators. It differs from the standard form in that the power of 10 is always a multiple of 3. For example:

$$300,000,000 \text{ in engineering form} = 300 \times 10^6$$

$$0.679 \text{ in engineering form} = 679 \times 10^{-3}$$

Summary

10. In summary the rules for the handling of numbers or algebraic expressions in index form are as follows:

$$a^0 = 1$$

$$a^m \times a^n = a^{m+n}$$

$$a^m \div a^n = a^{m-n}$$

$$(a^m)^n = a^{mn}$$

$$a^{-m} = \frac{1}{a^m}$$

$$a^{\frac{1}{m}} = \sqrt[m]{a}$$

LOGARITHMS

Introduction

11. The concept of logarithms is closely associated with the notion of indices. If a positive number, y , is expressed in index form with a base a , i.e.

$$y = a^x$$

then the index, x , is known as the logarithm of y to the base a . Thus:

$$\text{If } y = a^x, \text{ then } x = \log_a y$$

For example:

$$y = 32 = 2^5, \therefore \log_2 32 = 5$$

$$\text{If, } \log_{10} y = 3, \text{ then } y = 10^3$$

Common Logarithms

12. The most commonly used form of logarithms is to the base 10. The abbreviation 'log' is used and unless a base is explicitly stated or otherwise implied then 10 may be assumed. Using index notation any positive integer, N , may be written as:

$$N = 10^x, \text{ then } \log_{10} N = x$$

Values of the common log of any number may be found either from tables or from an electronic calculator. Prior to the widespread use of electronic calculators, logs were used as an aid to calculation. As logs are no more than indices, they obey the same rules as indices. Thus if it is necessary to multiply two numbers this can be achieved by finding the logs of the numbers, adding these and then finding the number corresponding to this log. Similarly, division may be accomplished by subtracting logs, the power of a number can be found by multiplying its log by the power, and the root of a number by dividing its log by the root index.

Naperian, Natural or Hyperbolic Logarithms

13. In many natural processes, the rate of growth or decay of a substance is proportional to the amount of substance present at a given time. It has been found that this relationship can be expressed in terms of a universal constant known as the exponential constant, e . The number, e , is irrational and is the sum of the infinite series:

$$1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \frac{1}{5!} + \frac{1}{6!} + \dots$$

where the symbol ! means factorial, i.e. that number multiplied by all of the positive integers less than itself, e.g. $6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1$.

14. By taking an appropriate number of terms, e can be calculated to any desired level of accuracy. Note that as the terms have factorials of ever-increasing numbers as their denominators then each successive term becomes smaller and the reduction in significance is rapid. As a comparison, the third term is 0.5, the seventh term is 0.00139, and the tenth term is 0.00000276. A value to 4 significant figures can be calculated from the first seven terms as 2.718.

15. Logarithms with e as the base are known as natural or Naperian (occasionally hyperbolic) logarithms. They are frequently encountered in scientific texts and are the only logarithms used in calculus. The abbreviation \ln is generally used. Whereas natural logarithms follow the same rules as common logarithms and can be used for the same purposes, they are rather more difficult to extract from tables. In any case, the use of logarithms to carry out arithmetic has been superseded by the electronic calculator.

Decibels

16. An application of logarithms is encountered in the field of amplification or gain, which is often expressed in units of bels or more normally decibels. If $P(I)$ is the input power into an amplifier and $P(O)$ is the output power, the gain is given by:

$$G = \log \frac{P(O)}{P(I)} \text{ bels} = 10 \log \frac{P(O)}{P(I)} \text{ decibels}$$

The ratio of the powers $\frac{P(O)}{P(I)}$ can be expressed in terms of output and input voltages as:

$$\frac{P(O)}{P(I)} = \frac{V(O)^2}{V(I)^2} = \left(\frac{V(O)}{V(I)} \right)^2$$

$$\therefore G = 10 \log \left(\frac{V(O)}{V(I)} \right)^2$$

$$= 20 \log \frac{V(O)}{V(I)} \text{ decibels}$$

Similarly, in terms of output and input currents:

$$G = 20 \log \frac{I(O)}{I(I)} \text{ decibels}$$

17. Example. If an amplifier has a gain of 30 decibels, calculate the input voltage required to produce an output of 50 volts.

$$\text{Using } G = 20 \log \frac{V(O)}{V(I)}$$

$$30 = 20 \log \frac{50}{V(I)}$$

$$1.5 = \log \frac{50}{V(I)}$$

Taking antilogs:

$$31.62 = \frac{50}{V(I)}$$

$$V(I) = \frac{50}{31.62}$$

$$= 1.581 \text{ V}$$

CHAPTER 6 - GRAPHS

Introduction

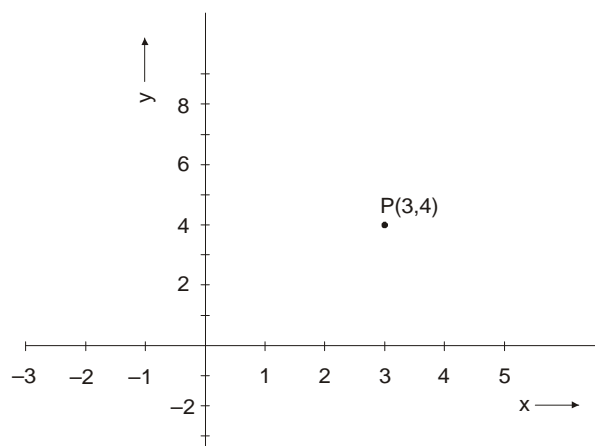
1. The term 'graph' is usually applied to a pictorial representation of how one variable changes in response to changes in another. This chapter will deal with the form of simple graphs, together with the extraction of data from them, and from the 'families of graphs' and carpet graphs that are frequently encountered in aeronautical publications. Although not strictly a 'graph', the Nomogram will also be covered. The pictorial representation of data in such forms as histograms, frequency polygons, and frequency curves will be treated in Volume 13, Chapter 16.

Coordinate Systems

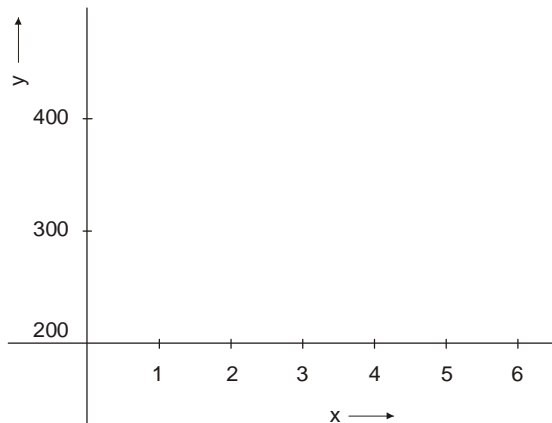
2. Graphs are often constructed from a table of, say, experimental data which gives the value of one variable, x , and the experimentally found value of the corresponding variable, y . In order to construct a graph from this data it is necessary to establish a framework or coordinate system on which to plot the information. Two such coordinate systems are commonly used: Cartesian coordinates and Polar coordinates. Both systems will be described below, but the remainder of this chapter will be concerned only with the Cartesian system.

3. **Cartesian Coordinates.** Cartesian coordinates are the most frequently used system. Two axes are constructed at right angles, their intersection being known as the origin. Conventionally the horizontal ' x ' axis represents the independent variable; the vertical ' y ' axis represents the dependent variable, i.e. the value that is determined for a given value of x . Any point on the diagram can now be represented uniquely by a pair of coordinate values written as (x,y) provided that the axes are suitably scaled. It is not necessary for the axes to have the same scale. Thus, in Fig 1, the point P has the coordinates $(3,4)$, i.e. it is located by moving 3 units along the x axis and then vertically by 4 ' y ' units. It is sometimes inconvenient to show the origin $(0,0)$ on the diagram when the values of either x or y cover a range which does not include 0. Fig 2 shows such an arrangement where the x -axis is scaled from 0 but the corresponding values of y do not include 0. The intersection of the axes is the point $(0,200)$. It should be noted from Fig 1 that negative values of x or y can be shown to the left and below the origin respectively.

13-6 Fig 1 Cartesian Coordinates

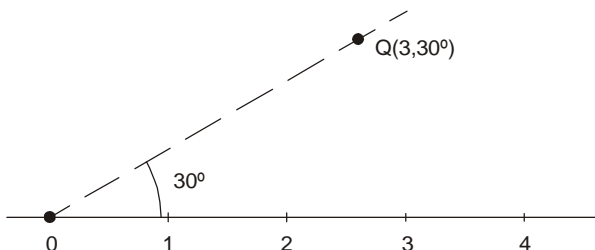


13-6 Fig 2 Cartesian Coordinates - Displaced Origin



4. **Polar Coordinates.** Polar coordinates specify a point as a distance and direction from an origin. Polar coordinates are commonly encountered in aircraft position reporting where the position is given as a range and bearing from a ground beacon; they are also used in certain areas of mathematics and physics. As with Cartesian systems it is necessary to define an origin, but only one axis or reference line is required. Any point is then uniquely described by its distance from the origin and by the angle that the line joining the origin to the point makes with the reference line. The coordinates are written in the form (r, θ) , with θ in either degree or radian measure. Conventionally, angles are measured anti-clockwise from the reference line as positive and clockwise as negative. Fig 3 illustrates the system. Point Q has the coordinates $(3, 30^\circ)$ or $(3, -330^\circ)$ in degree measure; $(3, \frac{\pi}{6})$ or $(3, \frac{-11\pi}{6})$ in radian measure.

13-6 Fig 3 Polar Coordinates



The Straight Line Graph

5. Table 1 shows a series of values of x and the corresponding values of y. Fig 4 shows these points plotted on a graph.

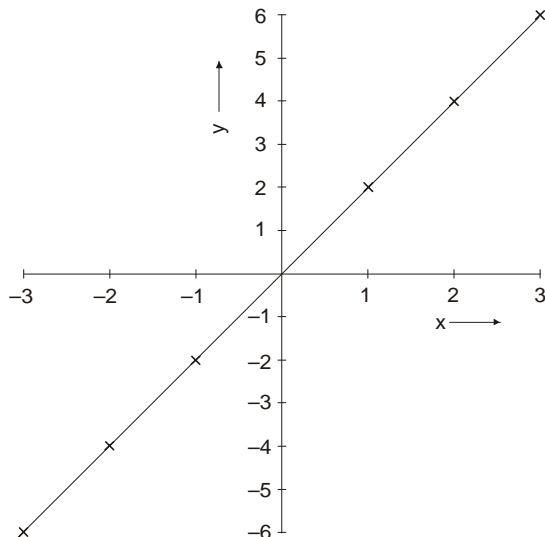
Table 1 Values of x and y

x	-3	-2	-1	0	1	2	3
y	-6	-4	-2	0	2	4	6

It will be seen that all the points lie on a straight line which passes through the origin. It is clear from the table of values that if the value of x is, say, doubled then the corresponding value of y is also doubled. Such a relationship is known as direct proportion and the graphical representation of direct proportion is always a straight line passing through the origin. In general the value of y corresponding to a value of x may be derived by multiplying x by some constant factor, m, ie: $y = mx$. In the example, m has the value 2, i.e. $y = 2x$. Because such a relationship produces a straight-line graph, it is known

as a linear relationship and $y = mx$ is known as a linear equation. Such relationships are not uncommon. For example the relationship between distance travelled, (d), speed, (s), and time, (t) is given by $d = st$. This would be a straight-line graph with d plotted on the y-axis and t on the x-axis.

13-6 Fig 4 Graph of $y = 2x$



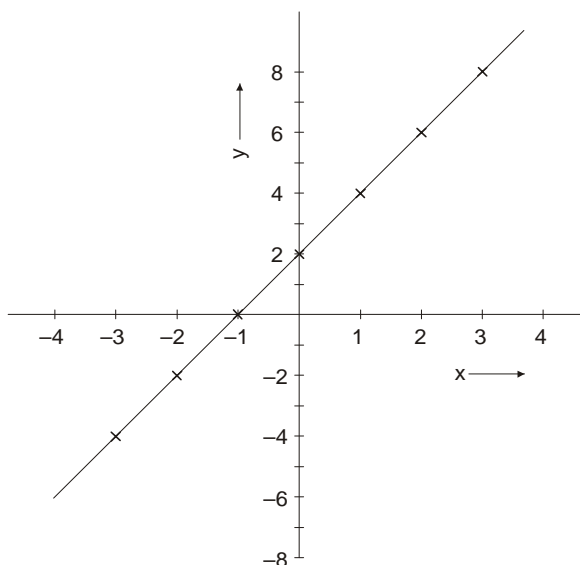
6. It is of course possible for a straight line through the origin to slope down to the right rather than up to the right as in the previous example. In this case positive values of y are generated by negative values of x and the equation becomes: $y = -mx$

7. Consider now the values of x and y in Table 2, and the associated graph, Fig 5.

Table 2 Values of x and y

x	-3	-2	-1	0	1	2	3
y	-4	-2	0	2	4	6	8

13-6 Fig 5 Graph of $y = 2x + 2$



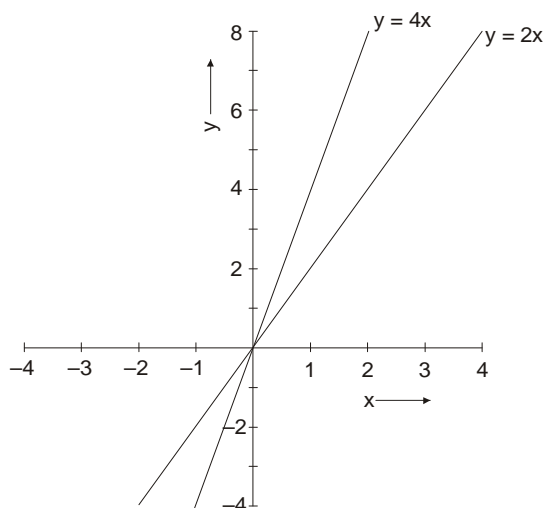
Clearly the graph is closely related to the previous example of $y = 2x$. In essence the line has been raised up the y-axis parallel to the $y = 2x$ line. Investigation of the table of values will reveal that the relationship between x and y is governed by the equation: $y = 2x + 2$, and in general, a graph of this type has the equation: $y = mx + c$, where c is a constant. It will be apparent that the equation $y = mx$ is identical to the equation $y = mx + c$ if a value of 0 is attributed to the constant c . Thus, $y = mx + c$ is the general equation for a straight line, m and c being constants which can be positive, negative or zero. A zero value of m generates a line parallel to the x-axis. The value of c is given by the point at which the line crosses the y-axis and is known as the intercept.

8. **Gradient.** Consider Fig 6 which shows two straight-line graphs: $y = 2x$ and $y = 4x$. Both lines pass through the origin and the essential difference between them is their relative steepness. The line $y = 4x$ shows y changing faster for any given change in x than is the case for $y = 2x$. The line $y = 4x$ is said to have a steeper gradient than the line $y = 2x$. The gradient is defined as the change in y divided by the corresponding change in x , ie $\frac{y}{x}$. Rearranging the general equation for a straight line ($y = mx$), to make m the subject gives $m = \frac{y}{x}$, i.e. the constant m is the gradient of the straight line. As the line $y = mx + c$ has been shown to be parallel to $y = mx$, this clearly has the same gradient, given by the value of m . In the equation:

$$\text{distance} = \text{speed} \times \text{time}$$

'speed' is equivalent to 'm' in the general equation, and it is apparent that the gradient, speed, represents a rate of change - in this case the rate of change of distance with time. This concept of the gradient representing a rate of change will become important when dealing with calculus in Volume 13, Chapter 13.

13-6 Fig 6 Graphs of $y = 2x$ and $y = 4x$



Non-Linear Graphs

9. Not all relationships result in straight-line graphs, indeed, they are a minority. A body falling to earth under the influence of gravity alone falls a distance y feet in time t seconds governed by the equation:

$$y = 16t^2$$

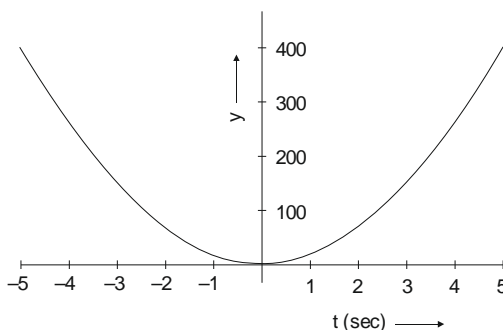
Table 3 shows a range of values of t with the corresponding values of y , and Fig 7 the associated graph.

Table 3 Values of t and y

t	0	1	2	3	4	5
y	0	16	64	144	256	400

Although not relevant in this example, notice that negative values of t produce identical positive values of y to their positive counterparts. The graph is therefore symmetrical about the y-axis and the shape is known as a parabola. The constant in front of the t^2 term determines the steepness of the graph.

13-6 Fig 7 Graph of $y = 16t^2$

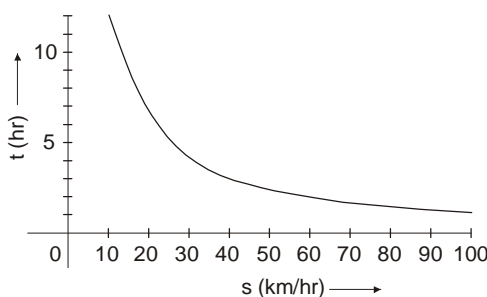


10. Consider now the problem "How long will it take to travel 120 km at various speeds?" This can be expressed as the equation:

$$t = \frac{120}{s}$$

where t = time in hours and s = speed in km/hr. This is an example of inverse proportion, ie an increase in s results in a proportional decrease in t. If values are calculated for s and t, and a graph is plotted, it will have the form illustrated in Fig 8 known as a hyperbola.

13-6 Fig 8 Graph of $t = \frac{120}{s}$ - Inverse Proportion



11. Graphs of $y = \sin x$ and $y = \cos x$ will be encountered frequently. The shapes of the graphs are shown below (Fig 9).

13-6 Fig 9 Graphs of sin and cos

Fig 9a $y = \sin x$

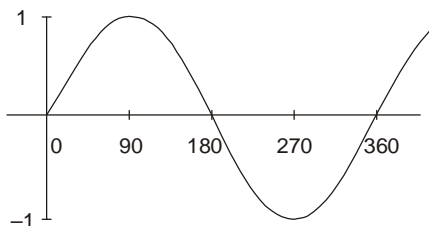
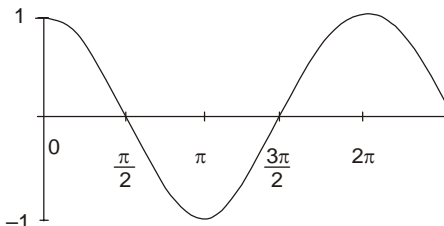


Fig 9b $y = \cos x$



The sine graph is shown with the x-axis scaled in degrees while the cosine graph has the x-axis scaled in radians. Either is correct; the radian form is frequently encountered in scientific texts. Sketches of these graphs are useful when trying to determine the value and sign of trigonometric functions of angles outside of the normal 0° to 90° range. Notice that both graphs repeat themselves after 360° (2π radians).

12. Finally, it is worth considering the graph that describes the relationship:

$$y = e^{ax}$$

where a is a positive or negative constant. This form of equation is very common in science and mathematics and variants of it can be found in the description of radioactive decay, in compound interest problems, and in the behaviour of capacitors. The irrational number 'e' equates to 2.718 to 4 significant figures. The graph of $y = e^x$ is shown in Fig 10a and that of $y = e^{-x}$ in Fig 10b. The significant point about these graphs, which are known as exponential graphs, is that the rate of increase (or decrease) of y increases (or decreases) depending upon the value of y. A large value of y exhibits a high rate of change. It is also worth noting that there can be found a fixed interval of x over which the value of y doubles (or halves) its original value no matter what initial value of y is chosen. This is the basis of the concept of radioactive decay half-life. The interval is equivalent to $\frac{0.693}{a}$ where a is the constant in the equation $y = e^{ax}$.

13-6 Fig 10 Exponential Graphs

Fig 10a $y = e^x$

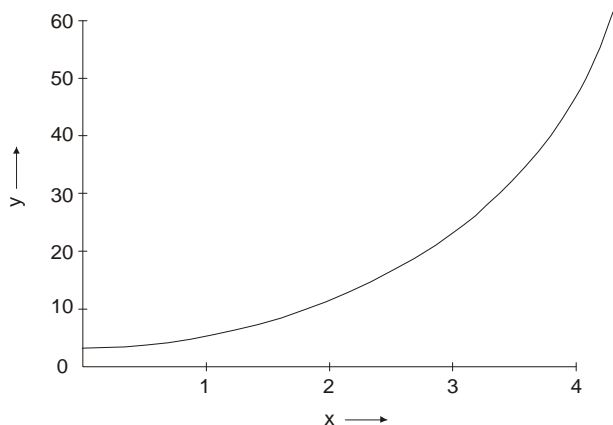
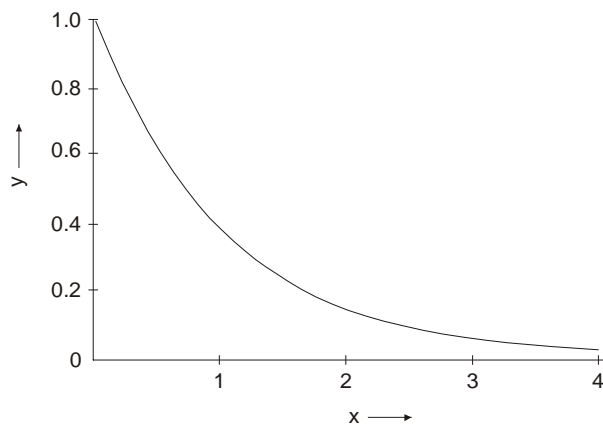


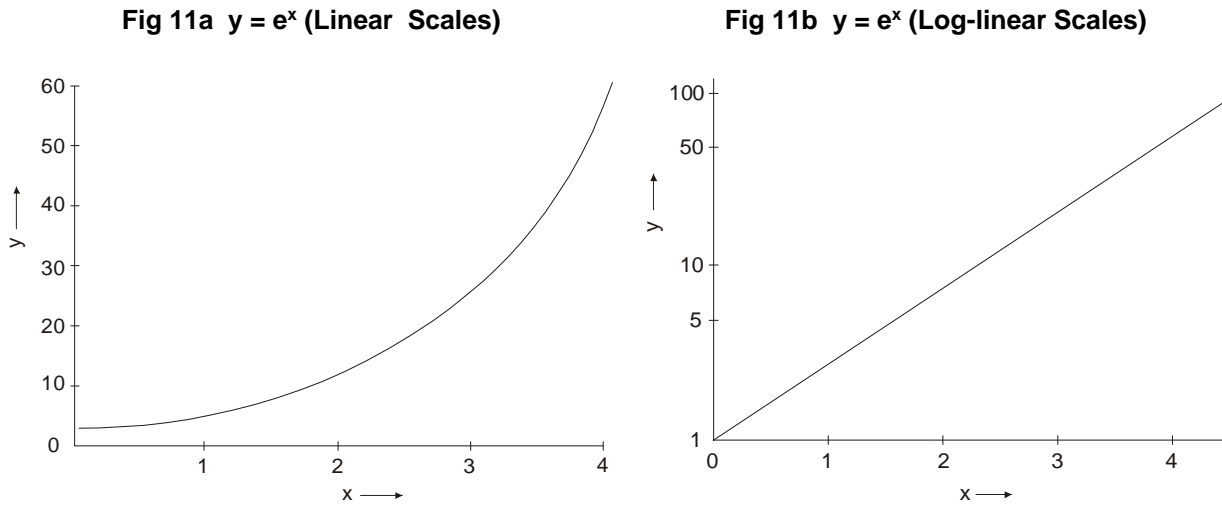
Fig 10b $y = e^{-x}$



13. **Logarithmic Scales.** Clearly plotting and interpreting from exponential graphs can be difficult. The problem can be eased by plotting on a graph where the x-axis is scaled linearly while the y-axis

has a logarithmic scale. This log-linear graph paper reduces the exponential curve to a straight line. A comparison between the linear and log-linear plots of $y = e^x$ is shown in Fig 11.

13-6 Fig 11 Comparison Between Linear and Log-linear Plots



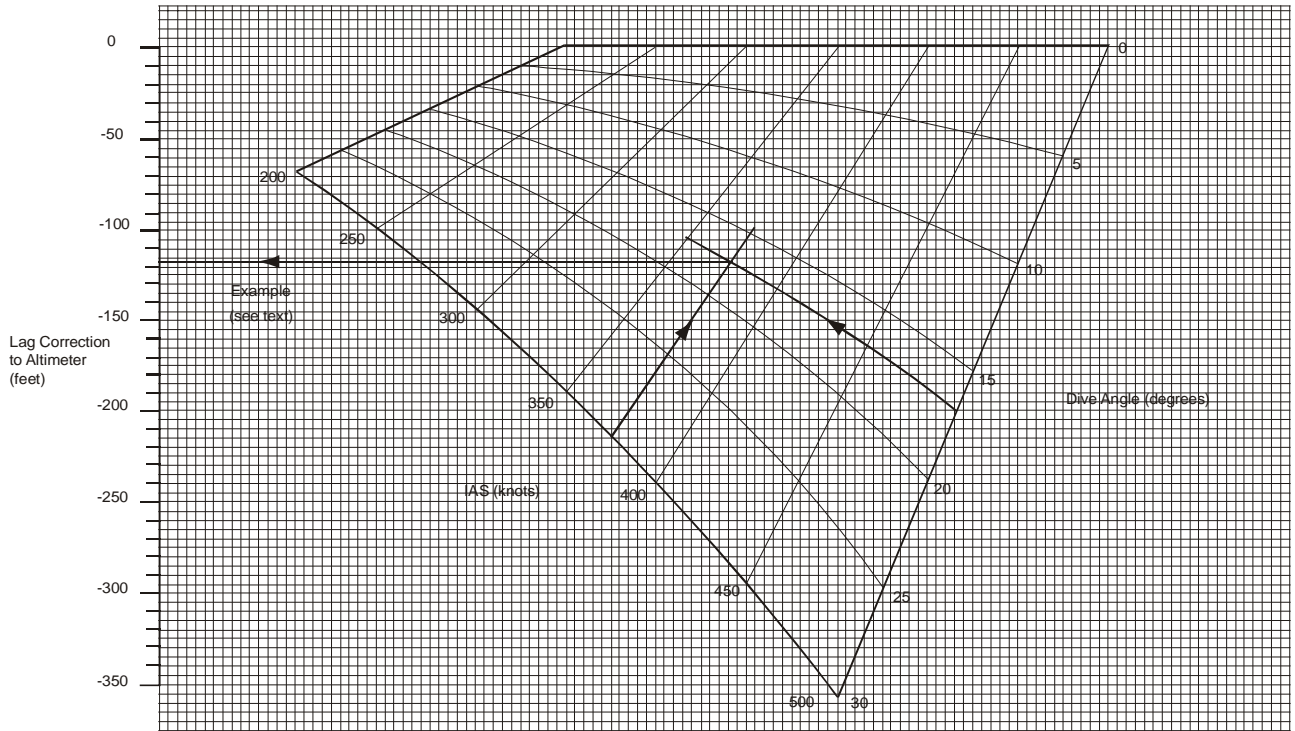
The Presentation and Extraction of Data

14. So far this chapter has been concerned with the mathematical background to simple graphs. More commonly graphs will be encountered and used as a source of data, especially in the field of flight planning and aircraft performance. Whilst occasionally these graphs will be either the simple forms already described or variations on these forms, more often rather complex graphs are used as being the only practical way of displaying complex relations. Two such types of complex graph will be described here in order to establish the method of data extraction. Finally, the nomogram will be discussed.

15. **Carpet Graphs.** An example of a carpet graph is shown in Fig 12.

13-6 Fig 12 Carpet Graph

Altimeter Lag Correction
(for altitudes up to 5000 feet)

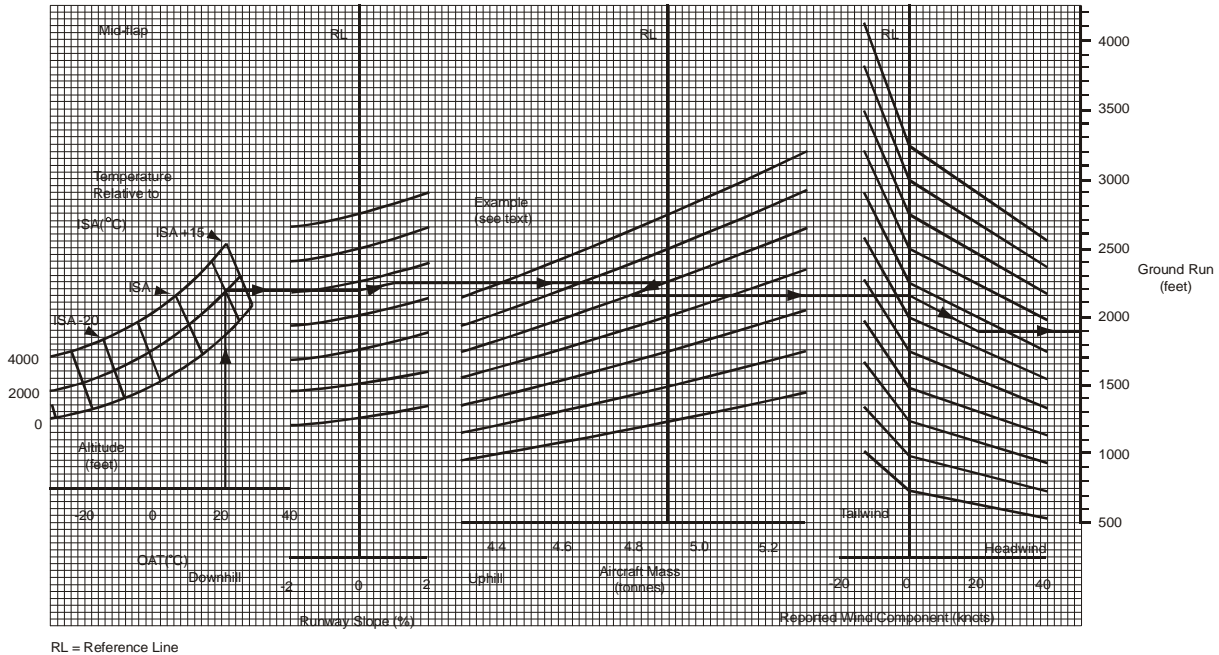


The aim of the graph is to indicate the lag in the altimeter experienced in a dive. Unlike the graphs already discussed where one input, 'x', produced one output, 'y', the carpet graph has two inputs for one output. The output is on the conventional 'y' axis but there is no conventional 'x' axis, rather there are two input axes. On the right hand edge of the 'carpet' diagram are figures for dive angle whilst on the bottom edge are figures for indicated air speed. To use the graph it is necessary to enter with one parameter, say dive angle, and follow the relevant dive angle line into the diagram until it intersects the appropriate IAS line. Intermediate dive angle and IAS values need to be interpolated, thus in the example values of 17° and 375 kt have been entered. From the point of intersection a horizontal line is constructed which will give the required lag correction figure where it intersects the 'y' axis, -118 feet in the example.

16. Families of Graphs. It is often necessary to consider a number of independent factors before coming to an end result. In this situation a family of graphs is frequently used to present the required information. Fig 13 shows such a family designed for the calculation of the aircraft's take-off ground run. Apart from the aircraft configuration which is indicated in the graph title, there are five input parameters. There will very often be a series of related graphs with variations in the title, for example in this case there will be another family of graphs for an aircraft with wing stores. It is clearly important that the correct set is selected. The method of using the graph will be described with reference to the example.

13-6 Fig 13 Family of Graphs

Take-off Ground Run. Clean or Gunpod Only



RL = Reference Line

17. At the left end is a small carpet graph. Starting with the value of outside air temperature (21°) proceed vertically to intersect the altitude line (2,000 feet). Alternatively enter the 'carpet' at the intersection of the altitude and temperature relative to ISA. From this intersection proceed horizontally into the next graph to intersect the vertical reference line, marked RL. From this point parallel the curves until reaching the point representing the value of runway slope as indicated on the bottom scale (1% uphill). From here construct a horizontal line to the next graph reference line. Repeat the procedure of paralleling the curves for aircraft mass (4.8 tonnes) and then proceed horizontally into the last graph for head/tail wind (20 kt head) which is used in the same manner. Finally the horizontal line is produced to the right hand scale where the figure for ground run can be read (1,900 feet).

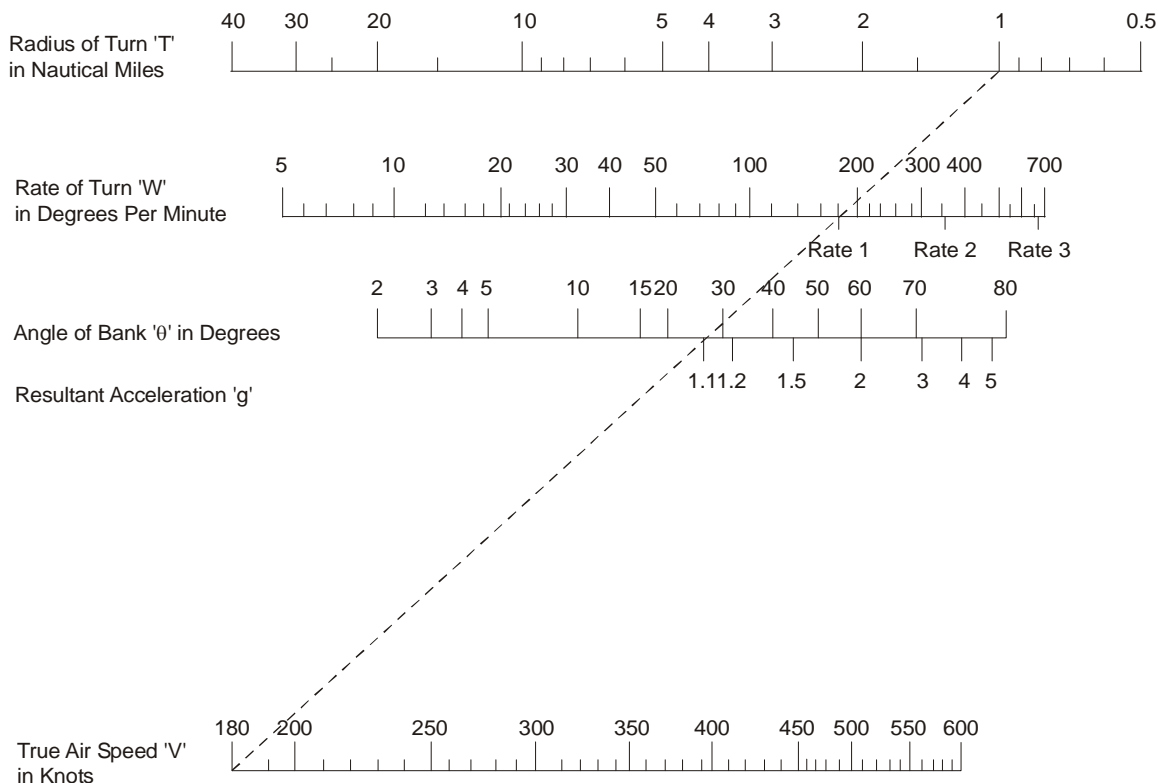
18. **The Nomogram.** The nomogram is not strictly a graph but a diagrammatic way of solving rather complex equations. There are usually two input parameters for which one or two resultant outputs may be derived. Fig 14 shows a nomogram for the determination of aircraft turning performance. The equations involved are:

$$\text{Rate of turn TAS} = \frac{\text{TAS}}{\text{Radius of Turn}}$$

$$= \text{Acceleration} \times \tan \text{Bank Angle}$$

This nomogram consists of four parallel scaled lines. Two known values are joined by a straight line, and the intersection of this line with the other scales gives the unknown values. In the example illustrated, an input TAS of 180 kts with a Rate 1 turn gives a resultant of 1.1 g, and a radius of turn of 1nm.

13-6 Fig 14 A Nomogram



The variables are related by the equation $V_w = \frac{V}{T} = g \tan \theta$

To use the Nomogram join two known values by a straight line and the intersection of this line on its projection with the other scales give the unknown values

Example : TAS (V) = 180 Kt
 Rate of Turn (W) = 1
 Angle of Bank (θ) = 25°
 Using the Nomogram (see dotted line)
 Resultant Acceleration (g) = 1.1
 Radius of Turn (T) = 1 nm

CHAPTER 7 - UNIT CONVERSIONS

SI Units

1. The Systeme International (SI) d'Unites is a metric system based upon seven fundamental units which are:

Length	- metre (m)
Mass	- kilogram (kg)
Time	- second (s)
Electric current	- ampere (A)
Luminous intensity	- candela (cd)
Temperature	- kelvin (K)
Amount of substance	- mole (mol)

Other SI units in common usage are:

Frequency	- hertz (Hz)
Energy	- joule (J)
Force	- newton (N)
Power	- watt (W)
Electric charge	- coulomb (C)
Potential difference	- volt (V)
Capacitance	- farad (F)
Inductance	- henry (H)
Magnetic field	- tesla (T)

Magnitudes

2. Prefixes used with SI Units to indicate magnitudes:

Factor	Name of Prefix	Symbol
10^{-18}	atto	a
10^{-15}	femto	f
10^{-12}	pico	p
10^{-9}	nano	n
10^{-6}	micro	μ
10^{-3}	milli	m
10^{-2}	centi	c
10^{-1}	deci	d
10	deca	da
10^2	hecto	h
10^3	kilo	k
10^6	mega	M
10^9	giga	G
10^{12}	tera	T
10^{15}	peta	P
10^{18}	exa	E

Conversion Factors

3. The following conversion factors have been selected as those most likely to be of general use.

		To Convert →	To →	Multiply By
Length				
4.	0.0394	inches (in)	millimetres (mm)	25.40
	3.2808	feet (ft)	metres (m)	0.3048
	1.0936	yards (yd)	metres (m)	0.9144
	5.399×10^{-4}	nautical miles (nm)	metres (m)	1852.0
	0.6214	miles	kilometres (km)	1.6093

Area				
5.	0.1550	square inches (in ²)	square centimetres (cm ²)	6.4516
	10.7639	square feet (ft ²)	square metres (m ²)	0.0929
	1.1960	square yards (yd ²)	square metres (m ²)	0.8361

Volume				
6.	0.2200	gallons (UK)	litres (l)	4.5460
	0.2643	gallons (US)	litres (l)	3.785
	0.0353	cubic feet (ft ³)	litres (l)	28.3161
	35.3147	cubic feet (ft ³)	cubic metres (m ³)	0.0283
	1.3080	cubic yards (yd ³)	cubic metres (m ³)	0.7646
	0.0610	cubic inches (in ³)	cubic centimetres (cm ³)	16.3871
	1×10^{-3}	cubic metres (m ³)	litres (l)	1000.0

	Multiply By ←	To ←	To Convert	To →	Multiply By
		To Convert	To		

Mass				
7.	0.0353	ounces (oz)	grams (g)	28.3495
	2.2046	pounds (lb)	kilograms (kg)	0.4536
	0.0685	slugs	kilograms (kg)	14.5939

Velocity				
8.	3.2808	feet/second (ft/s)	metres/second (m/s)	0.3048
	1.9685	feet/minute (ft/min)	centimetres/second (cm/s)	0.5080
	0.6214	miles/hour (mph)	kilometres/hour (km/h)	1.6093
	2.2369	miles/hour (mph)	metres/second (m/s)	0.4470
	0.5400	knots (kt)	kilometres/hour(km/h)	1.8520
	0.5921	knots (kt)	feet/second (fps)	1.6889
	1.9426	knots (kt)	metres/second (m/s)	0.5148

Acceleration				
9.	3.2808	feet/second ² (ft/s ²)	metres/second ² (m/s ²)	0.3048
	0.1020	gravitational acceleration (g)	metres/second ² (m/s ²)	9.8067

Force

10.	0.2248	pounds-force (lbf)	newtons (N)	4.4482
	2.2046	pounds-force (lbf)	kilograms-force (kgf)	0.4536
	7.2330	poundals (pdl)	newtons (N)	0.1383
	0.1020	kilograms-force (kgf)	newtons (N)	9.8067
	32.174	poundals (pdl)	pounds-force (lbf)	0.0311

Torque

11.	0.7376	pounds-force feet (lbf ft)	newton metres (Nm)	1.3558
	8.8507	pounds inches (lb in)	newton metres (Nm)	0.1130
	0.1020	kilograms-force metres (kgf m)	newton metres (Nm)	9.8067

Pressure

12.	9.869×10^3	atmospheres (atm)	kilopascals (kPa)	101.30
	0.0680	atmospheres (atm)	pounds-force/inch ² (psi)	14.6960
	0.1450	pounds-force/inch ² (psi)	kilopascals (kPa)	6.8948
	0.0100	bars	kilopascals (kPa)	100.0
	10.00	millibars (mbar)	kilopascals (kPa)	0.1000
	33.86	millibars (mbar)	inches mercury (in Hg)	0.0295
	1.000	newtons/metre ² (N/m ²)	pascals (Pa)	1.000
	25.4	mm mercury (mm Hg)	inches mercury (in Hg)	0.0394
	7.493	mm mercury (mm Hg)	kilopascals (kPa)	0.1334

Multiply By ← To ← To Convert

		To Convert →	To →	Multiply By
13.	0.0624	pounds/foot ³ (lb/ft ³)	kilograms/metre ³ (kg/m ³)	16.0185
	10^{-3}	grams/centimetre ³ (g/cm ³)	kilograms/metre ³ (kg/m ³)	1000.0
	0.0100	pounds/gallon	kilograms/metre ³ (kg/m ³)	99.776
	10.0221	pounds/gallon	kilograms/litre (kg/l)	0.0998

Power

14.	1.3410	horsepower (hp)	kilowatts (kW)	0.7457
	1.8182	horsepower (hp)	foot pounds-force/second (ft lbf/s)	550.0
	0.7376	foot pounds-force/second (ft lbf/s)	watts (W)	1.3558
	0.7376×10^3	foot pounds-force/second (ft lbf/s)	kilowatts (kW)	1.3558×10^{-3}

Energy, Work, Heat

15.	0.7376	foot pounds-force (ft lbf)	joules (J)	1.3558
	0.2388	calories (cal)	joules (J)	4.1868
	9.478×10^{-4}	British thermal units (Btu)	joules (J)	1055.1
	3412.1	British thermal units (Btu)	kilowatt hours (kWh)	2.931×10^{-4}

0.3725	horsepower hours (hph)	megajoules (MJ)	2.6845
1.3410	horsepower hours (hph)	kilowatt hours (kWh)	0.7457
9.478×10^{-3}	therms	megajoules (MJ)	105.51
Multiply By ←	To ←	To Convert	

CHAPTER 8 - PRINCIPLES AND RULES

Introduction and Notation

1. Algebra is that branch of mathematics dealing with the properties of, and relations between, quantities expressed in terms of symbols rather than numbers. The use of symbols allows general mathematical statements to be written down rather than just specific ones. For example, the relationship between °C and °F can be expressed as:

$$F = \frac{9}{5}C + 32$$

$$\text{or } C = \frac{5}{9}(F - 32)$$

Thus, given a value of temperature in either scale, the corresponding value in the other scale can be calculated. This is a considerably more concise method of relating the two scales than having a table showing the equivalent values, which in practice would have to be limited to a specified range of temperatures and with a specified level of precision. The algebraic relationship is in general more accurate than any representation by graph or nomogram.

2. Normally, when an algebraic expression is written down the conventional multiplication sign is omitted, both for brevity and to avoid confusion with the often-used symbol, x . Sometimes a full stop is used instead. The division sign is usually replaced by the solidus ($/$), or by separating the expression to be divided and the divisor by a horizontal line, thus for example:

$$(3x - 6) \div (7x + 3) \text{ may be written as } (3x - 6)/(7x + 3) \text{ or, more commonly, as } \frac{(3x - 6)}{(7x + 3)}.$$

The Laws of Algebra

3. There are several laws of algebra which govern how algebraic expressions may be manipulated:
- a. **Commutative Law.** This law states that additions and subtractions within an expression may be performed in any order. So may divisions and multiplications,
e.g. $x + y = y + x$; $x + y - z = x - z + y$; $xy = yx$; $xyz = zxy = yzx$
 - b. **Associative Law.** This law states that terms in an algebraic expression may be grouped in any order,
e.g. $x + y + z = (x + y) + z = x + (y + z)$; $xyz = x(yz) = (xy)z$
 - c. **Distributive Law.** This law states that the product of a compound expression and a single term is the algebraic sum of the products of the single term with all the terms in the expression,
e.g. $x(y + z) = xy + xz$; $4x(2y - 4z) = 8xy - 16xz$
 - d. **Laws of Precedence.** These laws dictate the order in which algebraic operations should be effected.

First, deal with terms in brackets; then work out multiplications and divisions; finally, work out additions and subtractions.

Operations within brackets are dealt with using the same precedence.

4. **Addition and Subtraction.** Within an algebraic expression, like terms, e.g. all x terms, all y terms, all xy terms, all z^2 terms etc, may be collected together and combined into a single term; unlike terms cannot be so combined,
e.g. $3x^2 + 6x - 4y + 2y + 5xy - x^2 + 9xy = 2x^2 + 6x - 2y + 14xy$

whereas, $3x^2 + 6x - 4y + 5xy$ cannot be simplified any further by addition or subtraction of terms.

5. **Multiplication and Division.** If two expressions which are to be multiplied together (or one divided by the other) have the same sign, the result is positive; while if their signs are different, the result is negative. The rules of indices (Volume 13, Chapter 5) similarly apply to algebraic expressions. Thus, for example: $4xy^2 \times 12x^{-3}y^4 = 48x^{-2}y^6$; $25a^4b^6 \div 5a^2b = 5a^2b^5$

6. When multiplying an expression within brackets then all of the terms within the bracket must be multiplied,

e.g. $5(3x^2 - 4y + 5) = 15x^2 - 20y + 25$

7. When two bracketed expressions are to be multiplied together then all of the terms within one set of brackets must be multiplied by all the terms within the other set,

e.g. $(3x^2 + 6)(2x - 4) = 6x^3 - 12x^2 + 12x - 24$

Factorization

8. A factor is a term by which an expression may be divided without leaving a remainder; a common factor is a term which is common to all of the terms of the expression. Thus for example in the expression $bx + by$, b is a common factor and the expression may be rewritten as $b(x + y)$. Similarly, in the expression: $24a^3 + 6a^2 - 12a$

$6a$ is common to all the terms and thus it may be rewritten as:

$$6a(4a^2 + a - 2)$$

9. Often an expression can be arranged into groups of terms where each group has its own factor, e.g. $ax + bx + ay + by$ can be regarded as two pairs of terms, thus:

$$(ax + bx) + (ay + by)$$

then each pair has its own common factor, x and y respectively, and so can be rewritten as:

$$x(a + b) + y(a + b)$$

$(a + b)$ now appears as a common factor and so the expression can be further factorized as:

$$(a + b)(x + y)$$

CHAPTER 9 - EQUATIONS

Introduction

1. An equation is a mathematical statement expressing an equality, i.e. it equates one algebraic expression with another. Equations may range in complexity from simple linear equations which contain only one unknown quantity and whose graphical representation is a straight line, to complex equations containing elements of calculus. This chapter will deal with simple linear, simultaneous and quadratic equations.

Transposition

2. Perhaps the most common use of an equation is in the determination of the value of one parameter given the values of other terms. Thus, for example, given the equation for temperature conversion:

$$F = \frac{9}{5}C + 32$$

if values for °C are given then °F can be determined by substituting the value in the equation, eg for 20 °C:

$$\begin{aligned} F &= \frac{9 \times 20}{5} + 32 \\ &= 36 + 32 \\ &= 68^{\circ}\text{F} \end{aligned}$$

3. However, suppose that it is necessary to find the Celsius equivalent of a Fahrenheit temperature. Clearly the equation needs to be rearranged so that C becomes the subject. In order to achieve this it is important to remember that operations may be carried out on the equation provided that the same process is applied to both sides of the '=' sign. The only forbidden operation is division by zero; multiplication by zero is not forbidden but of course gives the trivial result $0 = 0$.

4. In the example, the first step is to subtract 32 from both sides of the equation:

$$F - 32 = \frac{9}{5}C + 32 - 32$$

Next, both sides of the equation can be multiplied by 5/9, remembering that all terms must be multiplied:

$$\frac{5}{9}(F - 32) = \frac{5}{9} \times \frac{9}{5}C$$

$$\text{Thus } C = \frac{5}{9}(F - 32)$$

5. Frequently, equations will be encountered which contain powers and/or roots. These can be dealt with in an analogous fashion provided again that the same operations are carried out on both sides of the equation.

6. As an example, the periodic time of a simple pendulum, of length L, is given by the equation:

$$T = 2\pi \sqrt{\frac{L}{g}} \text{ seconds}$$

Suppose it is required to make 'L' the subject of the equation. The first operation is to square both sides to remove the square root:

$$T^2 = (2\pi)^2 \frac{L}{g}$$

$$= 4\pi^2 \frac{L}{g}$$

Then divide both sides by $4\pi^2$:

$$\frac{T^2}{4\pi^2} = \frac{L}{g}$$

Finally, multiply both sides by g (and conventionally the subject term is taken to the left side).

$$\text{Thus } L = \frac{gT^2}{4\pi^2}$$

7. Sometimes the way to proceed is not immediately obvious; the relationship between the distance of an object from a lens, (u), the distance of its image, (v), and the focal length of the lens, (f), is given by:

$$\frac{1}{u} + \frac{1}{v} = \frac{1}{f}$$

Suppose it is necessary to make f the subject of this equation. One technique is to initially multiply by the product of the denominators of the left-hand side, ie by uv .

$$\frac{uv}{u} + \frac{uv}{v} = \frac{uv}{f}$$

$$\text{ie } v + u = \frac{uv}{f}$$

Next, taking the 'f' term to the left and inverting both sides.

$$\frac{f}{uv} = \frac{1}{v + u}$$

Finally, multiply again by uv :

$$f = \frac{uv}{v + u}$$

8. However, there are often alternative methods. For example, taking the optical equation again:

$$\frac{1}{u} + \frac{1}{v} = \frac{1}{f}$$

The left-hand side can be combined into one term by using a common denominator, uv , thus:

$$\frac{v + u}{uv} = \frac{1}{f}$$

then inverting both sides

$$f = \frac{uv}{v + u}$$

[Click here to open a short quiz to test your understanding of Transposition.](#)

Simple Linear Equations

9. A simple linear equation is one which has only one unknown quantity and in which the unknown quantity has no power other than 1, e.g. $4x + 6 = 30$. The solution, ie finding the value of x which satisfies the conditions of the equation, is accomplished using the transposition techniques discussed above. Thus, in the example:

$$4x + 6 = 30$$

Subtract 6 from both sides

$$4x = 30 - 6 = 24$$

Divide both sides by 4

$$x = \frac{24}{4} = 6$$

Notice that addition and subtraction is equivalent to transferring the term to the other side of the equation accompanied by a change of sign, eg

Taking $4x + 9 = 3x - 6$

The '3x' term can be transferred to the left-hand side if its sign is changed to '-', and similarly the '9' may be transferred to the right-hand with a sign change, thus:

$$\begin{aligned} 4x - 3x &= -6 - 9 \\ \text{ie } x &= -15 \end{aligned}$$

[Click here to open a short quiz to test your understanding of Linear Equations.](#)

Linear Simultaneous Equations

10. Linear simultaneous equations are independent equations, with no powers other than 1, relating to more than one unknown. All of the equations must be true at the same time. For example:

$$\begin{aligned} x + 3y &= 20 & (1) \\ 9x - y &= 12 & (2) \end{aligned}$$

In general, to find values of all the unknowns which satisfy the equations then it is necessary to have as many independent equations as there are unknowns, for example if there are 5 unknowns then 5 independent equations would be required. For a pair of simultaneous equations with two unknowns there are two methods of solution; elimination and substitution.

11. **Solution by Elimination.** In this method one or both of the equations are manipulated so that the coefficient of one of the unknowns is identical in each equation. One equation is then subtracted from the other to eliminate one unknown resulting in a simple equation in the other unknown which can be solved readily. This value is then substituted back into one of the original equations to generate another readily soluble simple equation. Taking the examples from para 10:

Multiply equation (1) by 9:

$$9x + 27y = 180$$

Subtract equation (2)

$$\begin{array}{r} 9x + 27y = 180 \\ \underline{9x - y = 12} \\ 28y = 168 \end{array}$$

$$28y = 168$$

Divide by 28

$$y = 6$$

Substitute this value for y in equation (2)

$$\begin{aligned}9x - 6 &= 12 \\9x &= 18 \\x &= 2\end{aligned}$$

12. Solution by Substitution. In this method one equation is rearranged to express one unknown in terms of the other. This 'value' is then substituted into the other equation, which reduces to a simple equation in one unknown. After solution of this simple equation the value is substituted into either equation to find the other unknown value. Taking the same example equations:

$$\begin{aligned}x + 3y &= 20 & (1) \\9x - y &= 12 & (2)\end{aligned}$$

Rearrange equation (1) to make 'x' the subject

$$x = 20 - 3y$$

Substitute this into equation (2) and solve for y:

$$\begin{aligned}9(20 - 3y) - y &= 12 \\180 - 27y - y &= 12 \\28y &= 168 \\y &= 6\end{aligned}$$

Substitute this value into equation (1):

$$\begin{aligned}x + 18 &= 20 \\x &= 2\end{aligned}$$

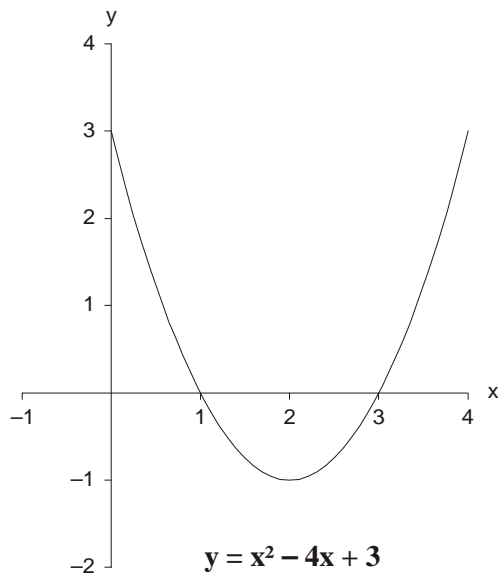
Quadratic Equations

13. A quadratic equation contains the square of the unknown quantity but no higher power. The simplest type has the form $x^2 = n$ where n is a positive number. The solution is a simple matter of finding the square root of the positive number ie $x = \sqrt{n}$, remembering that the result can have a negative or positive value. More commonly, a quadratic equation has the form:

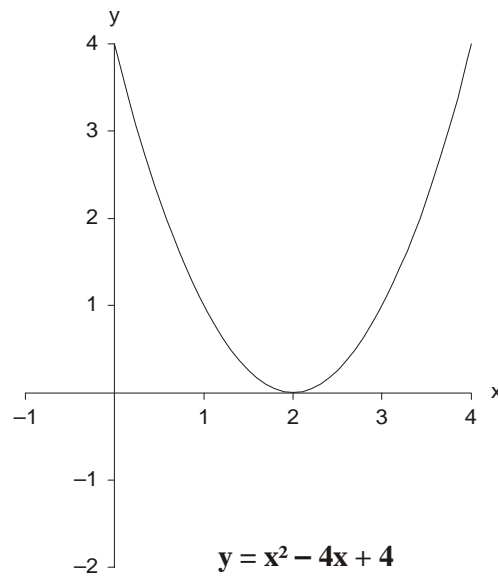
$$ax^2 + bx + c = 0, \text{ where } a, b \text{ and } c \text{ are numbers.}$$

14. It is instructive to examine the graphs representative of quadratic functions; their shape is parabolic. The solutions, or roots, of a quadratic equation are where $y = 0$ on the graph, i.e. where the graph crosses the x-axis. Four examples are illustrated in Fig 1.

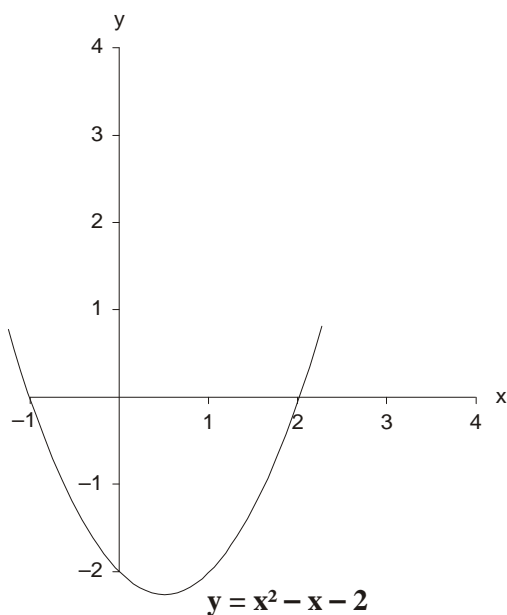
13-9 Fig 1 Graphs of Quadratic Equations



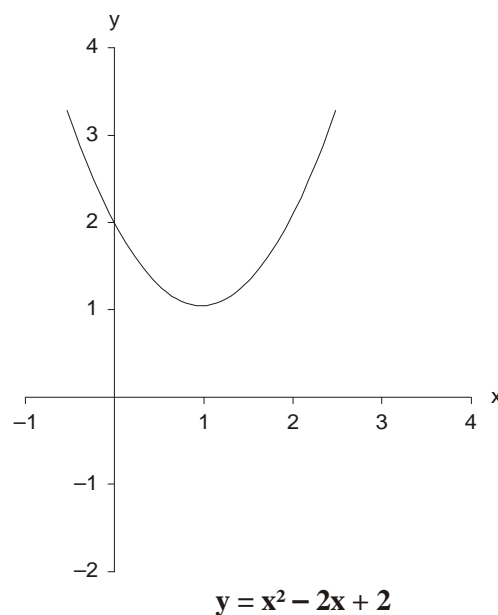
a



c



b



d

15. Fig 1a shows the graph of $y = x^2 - 4x + 3$. It will be seen that there are two positive roots; where $x = 1$ and where $x = 3$, ie where the graph crosses the x-axis. If the function had been $x^2 + 4x + 3$ then the graph would have crossed the x-axis to the left of the origin, ie giving two negative roots, -1 and -3 .

16. Fig 1b shows the graph of $y = x^2 - x - 2$. Here the graph crosses the axis at two points, one to the left and one to the right of the origin, thus there is one positive root and one negative root.

17. Fig 1c shows the graph of $y = x^2 - 4x + 4$. Here the graph does not cross the x-axis, rather the x-axis is a tangent to the curve at the value $x = 2$. Here the two roots are said to be coincident or equal.

18. Fig 1d shows the graph of $y = x^2 - 2x + 2$. In this case the graph does not cross the x-axis at all therefore there are no real roots.

Solving Quadratic Equations

19. Clearly, a quadratic equation can be solved as illustrated above by drawing the graph of the function and establishing where the curve crosses the x-axis. However, this method is tedious and often inaccurate. There are two other methods of solution in common use: factorization and formula.

20. **Factorization.** Some quadratic equations, but not all, are readily solved by the factorization method. The equation must first be arranged so that all the terms are on the left-hand side with just a zero to the right of the '=' sign. The problem then is to find factors of the expression on the left-hand side, remembering that it will not always be possible. Consider the equation:

$$x^2 - 4x + 3 = 0$$

The left-hand side can be factorized as:

$$(x - 1)(x - 3) = 0$$

As the left-hand side is now comprised of the product of two factors equal to zero, then one of the factors at least must equal zero, ie either:

$$x - 1 = 0, \therefore x = 1$$

$$\text{r } x - 3 = 0, \therefore x = 3$$

21. **Formula.** The formula method can be used to solve all quadratic equations that have a solution and, unless the factors are readily apparent, is normally the preferred method of solution. The equation must first be arranged into the form:

$$ax^2 + bx + c = 0$$

The formula to be used is then:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

As an example, take the equation:

$$2x^2 + 3x - 2 = 0$$

In terms of the standard form, $a = 2$, $b = 3$, $c = -2$. Thus, putting these values into the formula:

$$x = \frac{-3 \pm \sqrt{9 - (-16)}}{4}$$

$$x = \frac{-3 + 5}{4} \quad \therefore x = \frac{1}{2}$$

$$\text{or } x = \frac{-3 - 5}{4} \quad \therefore x = -2$$

22. The part of the formula ' $b^2 - 4ac$ ' is known as the discriminant and gives information about the roots of the equation. There are three possible cases:

a. $b^2 > 4ac$. This generates a positive term and so will have two real square roots. Thus, there will be two real roots to the equation. This is the situation shown by Figs 1a and 1b.

b. $b^2 = 4ac$. This is the case illustrated by Fig 1c. $b^2 - 4ac = 0$ and the roots are coincident and equal to $-b/2a$.

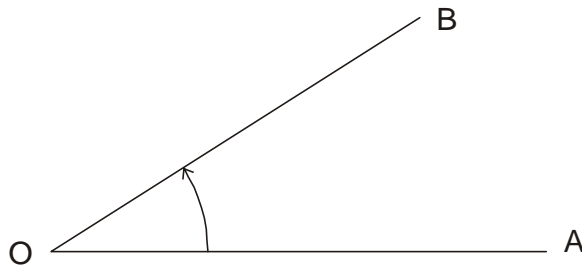
c. $b^2 < 4ac$. This makes $b^2 - 4ac$ negative. There are no real square roots to a negative number and therefore the equation has no real roots. This is the case illustrated in Fig 1d where the graph does not cross the x-axis. Although there are no real solutions in this case, this form of equation has many applications in, for example, control systems and aerodynamics.

CHAPTER 10 - PLANE TRIGONOMETRY

Definitions and Axioms

1. **Angles.** An angle is formed by the intersection of two lines. In Fig 1 AOB is an angle which is formed by a line which starts from the position OA and rotates about O in an anti-clockwise direction to the position OB. In this case O is the 'vertex' of the angle while OB and OA are the 'arms' of the angle. With anti-clockwise rotation, the angle is regarded as positive; if clockwise, the angle is regarded as negative. Unless otherwise stated, it is assumed that rotation will be anti-clockwise.

13-10 Fig 1 An Angle

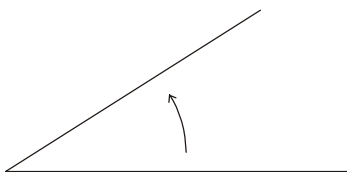


2. **Measurement of Angles.** One complete revolution is divided into 360 degrees ($^{\circ}$). The degree is sub-divided into 60 minutes ($'$), and the minute is sub-divided into 60 seconds ($''$). Ninety degrees constitutes one right-angle.

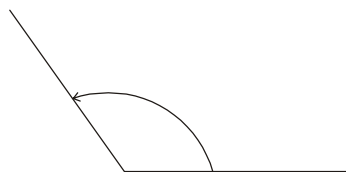
3. **Acute, Obtuse and Reflex Angles.** An acute angle is one which is less than 90° , an obtuse angle is greater than 90° but less than 180° and a reflex angle is greater than 180° (see Fig 2).

13-10 Fig 2 Acute, Obtuse, and Reflex Angles

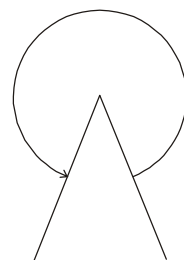
a Acute



b Obtuse



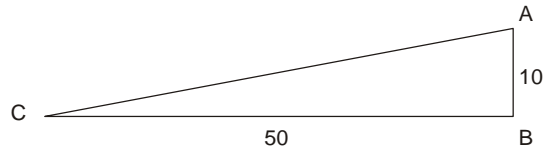
c Reflex



4. **Complementary and Supplementary Angles.** If two angles together make up 90° they are said to be complementary angles and each is the complement of the other. If two angles together make up 180° they are said to be supplementary and each is the supplement of the other.

5. **Slope and Gradient.** The slope of the line CA in Fig 3 is the angle ACB. The gradient of the line CA is the ratio $AB/CB = 1/5$.

13-10 Fig 3 Slope and Gradient



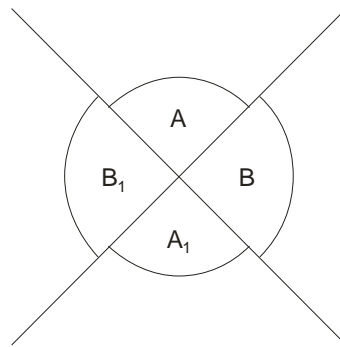
6. **Angles Formed by Two Intersecting Straight Lines.** When two straight lines intersect, as in Fig 4, the sum of the adjacent angles is 180° , and the vertically opposite angles are equal.

13-10 Fig 4 Two Intersecting Straight Lines

$$A + B = B + A_1 = A_1 + B_1 = B_1 + A = 180^\circ$$

$$A = A_1$$

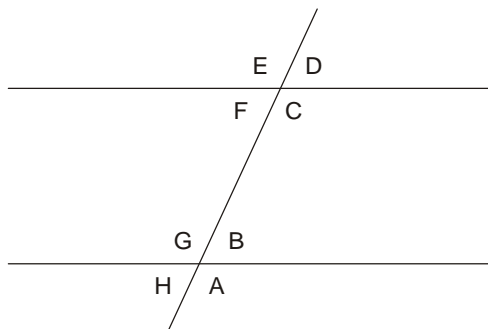
$$B = B_1$$



7. **Parallel Lines Cut by a Transversal.** When a transversal intersects two parallel straight lines, as in Fig 5:

- a. The corresponding angles are equal.
($\angle B = \angle D$, $\angle F = \angle H$, $\angle A = \angle C$, $\angle G = \angle E$)
- b. The alternate angles are equal.
($\angle F = \angle B$, $\angle C = \angle G$)
- c. The sum of the interior angles on the same side of the transversal is equal to 180° .
($\angle C + \angle B = 180^\circ$, $\angle F + \angle G = 180^\circ$)

13-10 Fig 5 Parallel Straight Lines and Transversal

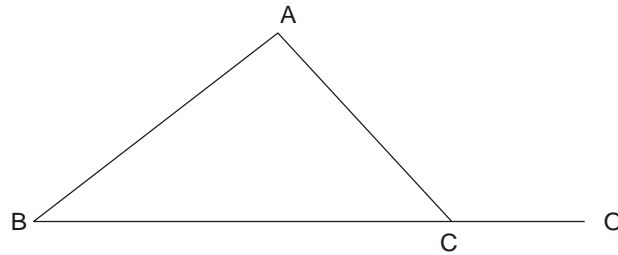


8. **Angle Properties of a Triangle.** The sum of the angles of a triangle is 180° . When one side of a triangle is produced, as in Fig 6, the exterior angle thus formed is equal to the sum of the two interior opposite angles.

13-10 Fig 6 Angle Properties of a Triangle

$$\angle A + \angle B + \angle C = 180^\circ$$

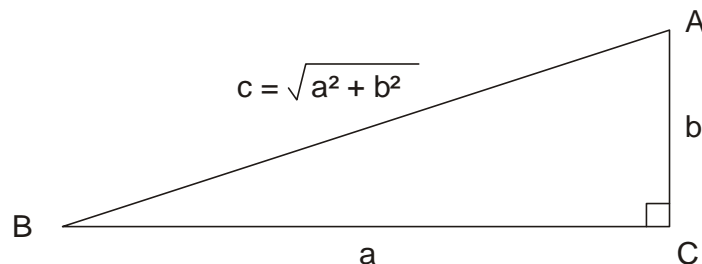
$$\angle ABC + \angle BAC = \angle ACO$$

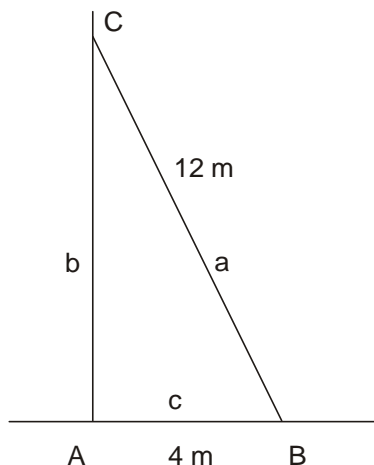


9. **Congruency of Triangles.** Two triangles are congruent if one can be superimposed on the other, so that they exactly coincide with regard to their vertices and their sides. Their areas must consequently be equal. Thus the three sides of one triangle must have the same lengths as the three sides of the other and the angles opposite to the equal sides must be equal. Triangles can be proved congruent when:

- a. The three sides of one are equal to the corresponding sides of the other.
- b. They have two sides and the included angle of one, equal to the corresponding sides and included angle of the other.
- c. They have two angles and a corresponding side equal.

10. **The Theorem of Pythagoras.** The conventional notation used for the solution of triangles is to denote the three angles by the capital letters A, B, C and the sides opposite these angles by the small letters a, b, c. Pythagoras' theorem states that, in any right-angled triangle, the square on the hypotenuse is equal to the sum of the squares on the other two sides, ie in Fig 7, where the angle at C = 90° , $c^2 = b^2 + a^2$. This theorem is of considerable importance and can be used to find one side of a right-angled triangle when the other two are known. For example, if a 12-metre ladder rests against a house so that its foot is 4 metres from the wall, it is possible to calculate how far up the side of the house the ladder will reach (see Fig 8).

13-10 Fig 7 Pythagoras' Theorem

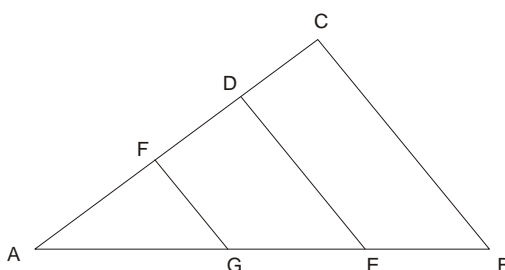
13-10 Fig 8 A Practical Application of Pythagoras' Theorem

From Pythagoras, it is known that

$$\begin{aligned}
 a^2 &= b^2 + c^2 \\
 \therefore b^2 &= a^2 - c^2 \\
 \text{ie } b &= \sqrt{144 - 16} \\
 &= \sqrt{128} \\
 &= 11.3
 \end{aligned}$$

thus, the ladder reaches 11.3 metres up the wall.

11. **Similar Triangles.** If, in two triangles, the three angles of one are equal to the three angles of the other, it does not necessarily follow that they are congruent. Consider Fig 9 in which angle A is common to the three triangles AFG, ADE and ACB.

13-10 Fig 9 Similar Triangles

$$\begin{aligned}
 \angle AFG &= \angle ADE = \angle ACB \text{ (corresponding angles)} \\
 \angle AGF &= \angle AED = \angle ABC \text{ (corresponding angles)}
 \end{aligned}$$

Such triangles are said to be similar. When triangles are equiangular, the ratios of corresponding sides are also equal.

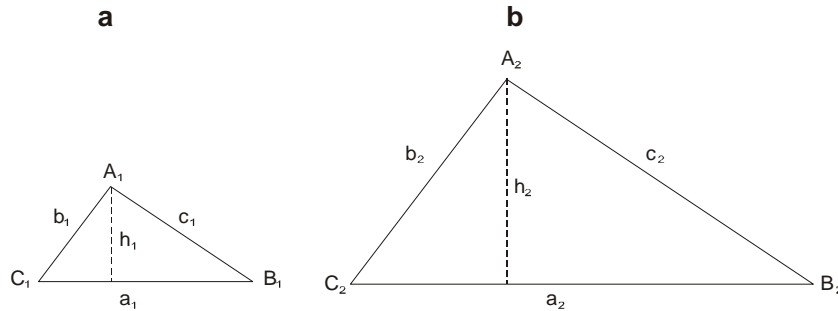
$$\text{Thus, } \frac{AF}{AG} = \frac{AD}{AE} = \frac{AC}{AB}$$

$$\text{and it follows that } \frac{AF}{AD} = \frac{AG}{AE} = \frac{FG}{DE}$$

Note: A similar relation holds good for two polygons which are equiangular.

12. **The Relationship Between Sides and Areas of Similar Triangles.** Triangles $A_1B_1C_1$ and $A_2B_2C_2$ are similar triangles with heights h_1 and h_2 respectively (see Fig 10).

13-10 Fig 10 Areas of Similar Triangles



$$\text{Then } \frac{h_1}{a_1} = \frac{h_2}{a_2} \quad \therefore h_1 = \frac{a_1 h_2}{a_2}$$

$$\text{Also, } \frac{\text{Area of } A_1B_1C_1}{\text{Area of } A_2B_2C_2} = \frac{\frac{1}{2}a_1h_1}{\frac{1}{2}a_2h_2}$$

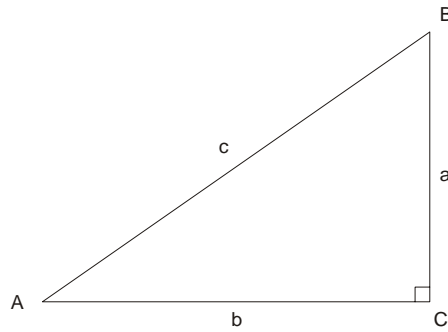
Substituting for h_1

$$\frac{\text{Area of } A_1B_1C_1}{\text{Area of } A_2B_2C_2} = \frac{\frac{1}{2}a_1 \times \frac{a_1 h_2}{a_2}}{\frac{1}{2}a_2 h_2} = \frac{a_1^2}{a_2^2}$$

Similarly the areas can be proved proportional to the squares of b_1 , b_2 and c_1 , c_2 . Hence the areas of similar triangles are proportional to the squares of the corresponding sides.

Trigonometrical Ratios

13. In any right-angled triangle, the side opposite to the right-angle is called the hypotenuse, and the other two sides are called the opposite and adjacent according to their position relative to the angle under consideration. Fig 11 shows a right-angled triangle ABC in which, relative to angle A, the side BC is opposite and the side AC is adjacent. The reverse is true relative to angle B. There are six trigonometrical ratios:

13-10 Fig 11 Right-angled Triangle

$$\text{The sine of an angle (sin)} = \frac{\text{opposite}}{\text{hypotenuse}}$$

$$\text{The cosine of an angle (cos)} = \frac{\text{adjacent}}{\text{hypotenuse}}$$

$$\text{The tangent of an angle (tan)} = \frac{\text{opposite}}{\text{adjacent}}$$

$$\text{The cosecant of an angle (cosec)} = \frac{1}{\sin A}$$

$$\text{The secant of an angle (sec)} = \frac{1}{\cos A}$$

$$\text{The cotangent of an angle (cot)} = \frac{1}{\tan A}$$

14. The trigonometric ratios, in terms of the triangle in Fig 11 are:

$$\sin A = \frac{a}{c} \qquad \sin B = \frac{b}{c}$$

$$\cos A = \frac{b}{c} \qquad \cos B = \frac{a}{c}$$

$$\tan A = \frac{a}{b} \qquad \tan B = \frac{b}{a}$$

$$\text{cosec } A = \frac{c}{a} \qquad \text{cosec } B = \frac{c}{b}$$

$$\sec A = \frac{c}{b} \qquad \sec B = \frac{c}{a}$$

$$\cot A = \frac{b}{a} \qquad \cot B = \frac{a}{b}$$

and it may be deduced that:

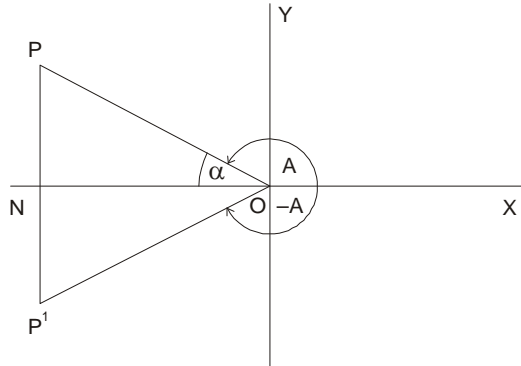
$$\tan A = \frac{\sin A}{\cos A} \qquad \cot A = \frac{\cos A}{\sin A}$$

$$\sin A = \cos (90 - A); \quad \cos A = \sin (90 - A); \quad \tan A = \cot (90 - A); \quad \cot A = \tan (90 - A)$$

15. The Trigonometric Ratios for Angles of any Magnitude. So far, only acute angles have been considered but it is also necessary to be able to find the trigonometrical ratios of obtuse, reflex and sometimes negative angles. Consider a set of rectangular axes OX, OY (see Fig 12). To determine any trigonometrical ratio of any angle, the angle is set up on this system of axes as follows. A radius vector, OP, initially along OX, is considered to turn about O in a counter-clockwise sense through the required angle, A. For a negative angle it turns in the clockwise sense. From P, drop a perpendicular,

PN, on to the x-axis. Any trigonometrical ratio of A is then referred to the right-angled triangle OPN and the acute angle α which OP makes with the x-axis. OP is always taken to be +ve, but ON and PN take the signs which would be attached to them when regarded as the coordinates of the point P.

13-10 Fig 12 The Four Quadrants



Thus, for the angles A and -A in the figure:

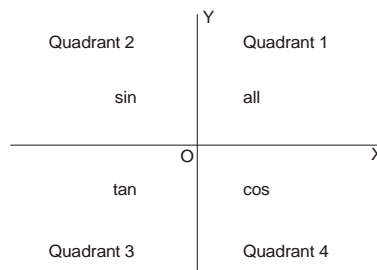
$$\sin A = \frac{PN}{OP}, \text{ and is +ve; } \quad \sin (-A) = \frac{-P'N}{OP'}, \text{ and is -ve}$$

$$\cos A = \frac{-ON}{OP}, \text{ and is -ve; } \quad \cos (-A) = \frac{-ON}{OP'}, \text{ and is -ve}$$

$$\tan A = \frac{PN}{-ON}, \text{ and is -ve; } \quad \tan (-A) = \frac{-P'N}{-ON}, \text{ and is +ve}$$

The reciprocal ratios, cosec, sec and cot, have the same sign respectively as sin, cos and tan. These more general definitions of the trigonometrical ratios, which apply to all angles of any magnitude and sign, are consistent with the former definitions which applied only to acute angles, since, for an acute angle, the radius vector, OP would lie in the first quadrant and ON and OP would then both be +ve. It has been shown that, for angles in the second quadrant, sin is +ve while cos and tan are -ve. Similarly it can be shown that, for angles in the third quadrant, tan is +ve while sin and cos are -ve, and for angles in the fourth quadrant, cos is +ve while sin and tan are -ve. Hence the 'all, sin, tan, cos' rule for determining the sign of a trigonometrical ratio. Fig 13 indicates which functions are positive in each of the quadrants.

13-10 Fig 13 Signs of Trig Functions by Quadrant



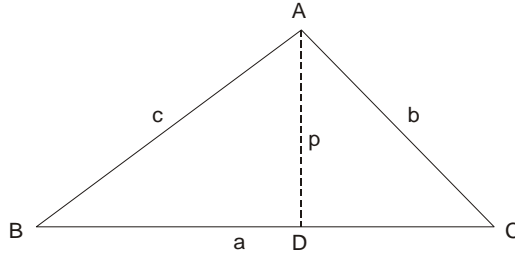
The Sine Rule

16. The three sides and three angles of a triangle are sometimes called its 'elements'. When given a sufficient number of these elements it is possible to find the remainder. Thus for example if two sides

and one angle or one side and two angles are known, and in each case an angle and the side opposite to it are included, the Sine Rule can be used to evaluate the unknown sides and angles.

17. In the triangle ABC at Fig 14, draw AD perpendicular to BC and let AD = p.

13-10 Fig 14 The Sine Rule



$$\frac{p}{c} = \sin B \quad \therefore p = c \sin B; \quad \frac{p}{b} = \sin C \quad \therefore p = b \sin C$$

$$\therefore c \sin B = b \sin C \text{ or } \frac{c}{\sin C} = \frac{b}{\sin B}$$

In a similar way it can be proved that:

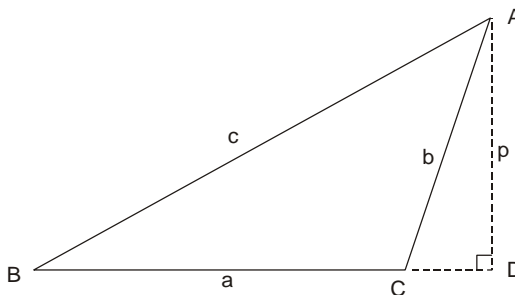
$$\frac{c}{\sin C} = \frac{a}{\sin A}$$

and the Sine Formula is:

$$\frac{a}{\sin A} = \frac{b}{\sin B} = \frac{c}{\sin C}$$

18. In the triangle ABC at Fig 15, where angle C is obtuse, draw AD perpendicular to BC produced and let AD = p.

13-10 Fig 15 Sine Rule - Obtuse Angled Triangle



$$\frac{p}{c} = \sin B \quad \therefore p = c \sin B; \text{ and } \frac{p}{b} = \sin ACD \quad \therefore p = b \sin ACD$$

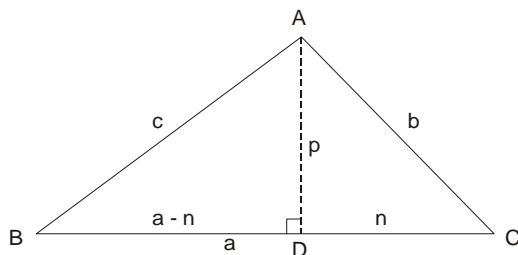
but, $\sin ACD = \sin (180 - ACD) = \sin C$

$$\therefore p = c \sin B = b \sin C, \text{ as before, or } \frac{b}{\sin B} = \frac{c}{\sin C}$$

The Cosine Rule

19. When the Sine Rule is not applicable, as, for instance, when two sides and an included angle are given, the Cosine Rule may be used. Consider the triangle ABC in Fig 16:

13-10 Fig 16 The Cosine Rule

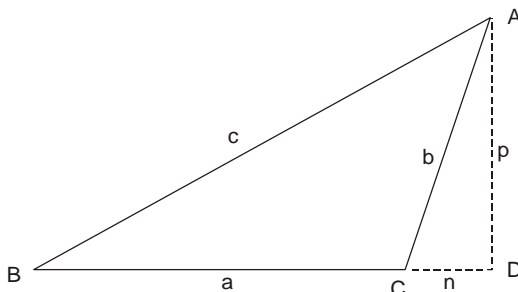


From A draw AD (= p) perpendicular to BC. Let DC = n and BD = a-n. Then, by the principle of Pythagoras,

$$\begin{aligned}
 & p^2 &= b^2 - n^2 \\
 \text{and} & p^2 &= c^2 - (a - n)^2 \\
 \therefore & b^2 - n^2 &= c^2 - (a^2 - 2an + n^2) \\
 \therefore & b^2 - n^2 &= c^2 - a^2 + 2an - n^2 \\
 \therefore & b^2 &= c^2 - a^2 + 2an \\
 \text{but } n & &= b \cos C \\
 \therefore & b^2 &= c^2 - a^2 + 2ab \cos C \\
 \text{and } c^2 & &= a^2 + b^2 - 2ab \cos C
 \end{aligned}$$

20. In the case of triangle ABC, where C is an obtuse angle, as in Fig 17:

13-10 Fig 17 Cosine Rule - Obtuse Angled Triangle



$$\begin{aligned}
 & p^2 &= b^2 - n^2 \\
 \text{and} & p^2 &= c^2 - (a+n)^2 \\
 \therefore & b^2 - n^2 &= c^2 - a^2 - 2an - n^2 \\
 \therefore & b^2 &= c^2 - a^2 - 2an \\
 \text{where } n & &= b \cos ACD \\
 & &= -b \cos ACB \\
 & &= -b \cos C \\
 \therefore & b^2 &= c^2 - a^2 - 2a(-b \cos C) \\
 \therefore & b^2 &= c^2 - a^2 + 2ab \cos C \\
 \therefore & c^2 &= a^2 + b^2 - 2ab \cos C
 \end{aligned}$$

which is identical to the previous formula.

21. By the same method, it can be shown that

$$b^2 = a^2 + c^2 - 2ac \cos B; \quad \text{and } a^2 = b^2 + c^2 - 2bc \cos A$$

Thus, given any two sides and their included angle, the third side can be found. It may then be more convenient to apply the Sine Rule to find any other unknown elements.

22. When three sides a , b , and c are given, the cosine formula may be re-arranged as follows:

$$c^2 = a^2 + b^2 - 2ab \cos C$$

$$2ab \cos C = a^2 + b^2 - c^2$$

$$\cos C = \frac{a^2 + b^2 - c^2}{2ab}$$

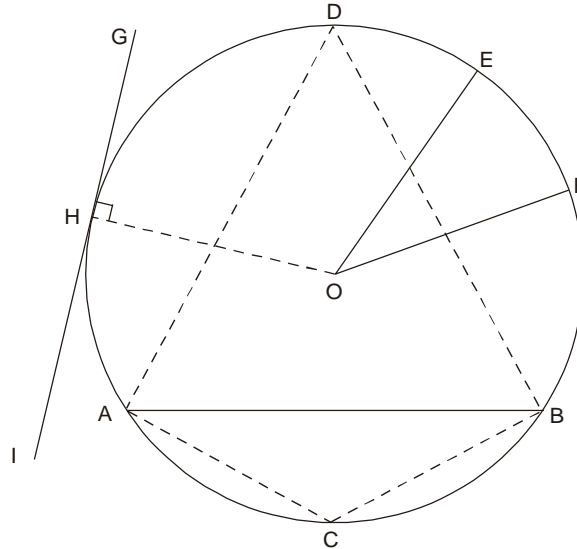
Similarly $\cos A$ and $\cos B$ may be found.

CHAPTER 11 - THE CIRCLE

Definitions

- Some important definitions relating to the circle are explained with reference to Fig 1.

13-11 Fig 1 The Properties of a Circle

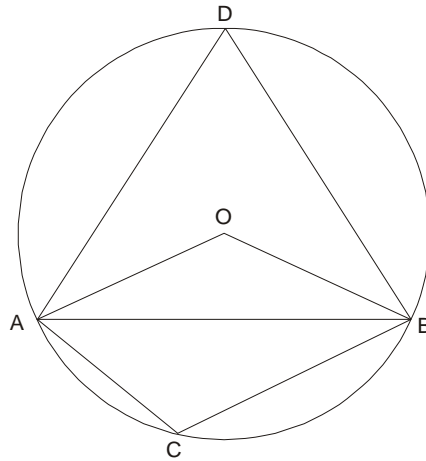


- The Chord of a circle is any straight line which divides the circle into two parts, and is terminated at each end by the circumference. AB in Fig 1 is a chord.
- A Segment of a circle is a figure bounded by a chord and the arc which it cuts off. In Fig 1 the chord AB divides the circle into two segments:
 - The minor segment is ABC.
 - The major segment is ADB.
- An Arc of a circle is a portion of the circumference. EF in Fig 1 is an arc.
- A Sector of a circle is a figure which is bounded by two radii and the arc between them. OEF in Fig 1 is a sector.
- A Tangent to a circle is a straight line which meets the circle in one point called the point of contact, and does not cut the circle when produced. A tangent is at right angles to the radius drawn from the point of contact. GHI in Fig 1 is a tangent.
- The Angle in a Segment is the angle subtended at a point on the arc of a segment by the chord of the segment. In Fig 1 the angle in the major segment is $\angle ADB$ and the angle in the minor segment is $\angle ACB$.
- The Angle at the Centre is the angle subtended at the centre of a circle by a chord or by an arc. $\angle EOF$ in Fig 1 is the angle at the centre subtended by EF.

Theorems

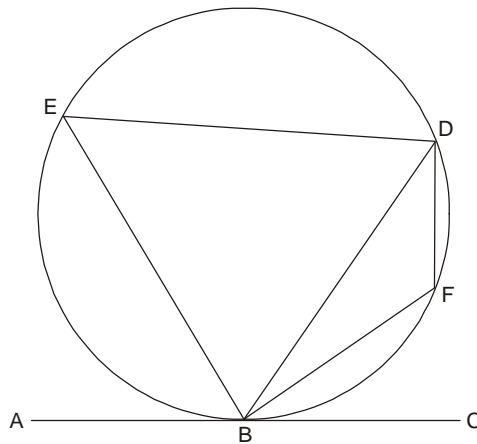
- The angle at the centre of a circle subtended by an arc is double the angle at the circumference subtended by the same arc. In Fig 2 $\angle AOB = 2\angle ADB$ and the reflex $\angle AOB = 2\angle ACB$. Some important results follow from this theorem:

13-11 Fig 2 The Angles in a Circle



- a. All angles in the same segment of a circle are equal.
 - b. The opposite angles of a quadrilateral inscribed in a circle are together equal to 180° ; that is, they are supplementary.
 - c. The angle in a semi-circle is a right angle.
 - d. In equal circles, arcs which subtend equal angles either at the centres or at the circumference are equal.
 - e. In equal circles, chords which cut off equal arcs are equal.
3. The angle between a tangent and a chord drawn through the point of contact is equal to the angle in the alternate segment. In Fig 3 $\angle DBC = \angle DEB$ and $\angle ABD = \angle DFB$.

13-11 Fig 3 The Angle between Tangent and Chord



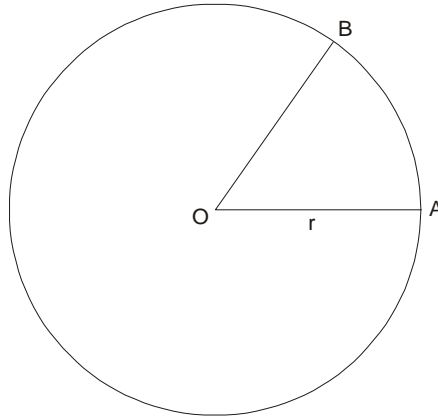
4. The ratio of the circumference of a circle to its diameter is denoted by π , so that $C/d = \pi$ or $C = \pi d$. Since diameter is equal to two times the radius, $C = 2\pi r$.

Circular Measure

5. The magnitude of an angle is commonly expressed in degrees which are obtained by the division of a right angle into 90 parts. There is another method which is of great practical importance and in which the unit employed is an absolute one. Consider Fig 4. Suppose the line OA in Fig 4 rotated

about the point O to the position OB, so that the length of the arc AB is equal to the radius of the circle. The angle AOB subtended by the arc AB is called a radian. The radian is the unit of measurement in circular measure. Hence a radian may be defined as the angle subtended at the centre of a circle by an arc equal in length to the radius.

13-11 Fig 4 The Radian



6. The length of an arc, when the angle is given in radians, can be calculated as follows:

Length of an arc for 1 radian = r

Length of arc for σ radians = $r\sigma$

Arc = $r\sigma$

7. **The Relationship Between Radians and Degrees.** Since an arc of r units in length subtends an angle of one radian, the number of radians subtended by the circumference of a circle is given by the number of times the radius is contained in the circumference, ie $C = 2\pi r$. The number of radians for one revolution = $\frac{2\pi r}{r} = 2\pi$ radians. From this, π radians = 180° and 1 radian = $180/\pi = 57.3^\circ$.

Radians may be converted to degrees by multiplying by π and dividing by 180.

Examples:

Convert to radians 45° , 30° taking π as 3.1416

$$\text{a. } \frac{45 \times 3.1416}{180} = \frac{3.1416}{4} = 0.7854 \text{ rad}$$

$$\text{b. } \frac{30 \times 3.1416}{180} = \frac{3.1416}{6} = 0.5236 \text{ rad}$$

Conversions are easily carried out if an electronic calculator is available which will enter an accurate value for π at the touch of a button.

Angular Rotation

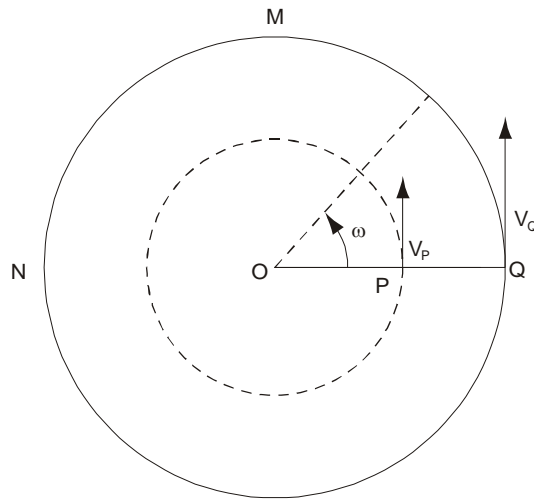
8. A straight strip of tape stuck on the face of a gear wheel from the axis to perimeter, will enable the observation that, in one revolution of a gear wheel, an individual tooth is rotated through 360° . Since $360^\circ = 2\pi$ radians, then one revolution is also 2π radians. In circular measure, therefore, all even

multiples of π correspond to complete revolutions. For example, 4π radians will be 2 revolutions, 6π radians will be 3 revolutions, etc.

9. If a shaft or pulley is rotating at 3 revolutions per second, then the angular rotation must be $3 \times 2\pi$ radians per second. In general terms, a rotation of n revs per second will give an angular velocity of $2\pi n$ radians per second.

10. **The Relationship Between Angular and Linear Velocity.** Let QMN of Fig 5 represent a flywheel which has an angular velocity of ω radians per sec. This means that any radius OQ rotates through an angle of ω radians in 1 sec. Any point P on OQ will also have the same angular velocity. Since arc = $r\theta$, the arc traced out by Q in 1 sec = $\omega \times OQ$, and the arc traced out by P in 1 sec = $\omega \times OP$. In general, if the point is at a distance r from the centre of rotation, the linear velocity of that point will be ωr . If V is the linear velocity of a point, then $V = \omega r$. Although all points on the flywheel have the same angular velocity, the linear velocity of any point will depend on its distance from the centre of rotation.

13-11 Fig 5 Angular and Linear Velocity



CHAPTER 12 - SPHERICAL TRIGONOMETRY

Introduction

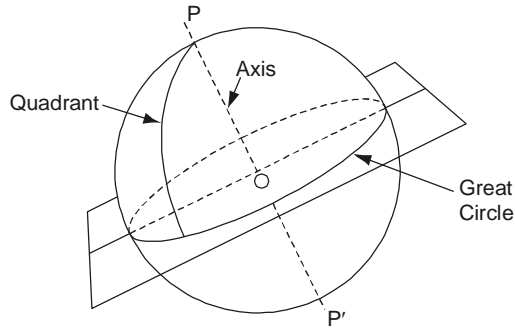
1. Many of the problems encountered in navigation require the solution of triangles on the Earth's surface. In all but very small triangles, the effect of curvature must be taken into account.

GEOMETRY OF THE SPHERE

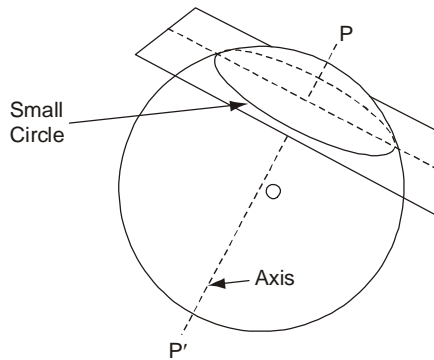
2. The following properties and definitions apply to the geometrical sphere and, although the shape of the Earth is not truly spherical, it is considered to be so for most practical navigational purposes.

- a. The section of the surface of a sphere made by any intersecting plane is a circle.
- b. The axis of a circle on a sphere is the diameter of the sphere at right angles to the plane of the circle (PP' in Figs 1 and 2).
- c. The two points at which the axis of the circle intersects the surface of the sphere are called the poles of the circle (P and P' in Figs 1 and 2).
- d. If the plane passes through the centre of the sphere the section is called a great circle. All other sections are called small circles.
- e. A quadrant is the great circle arc drawn from any point on a great circle to its pole (Fig 1).
- f. Only one great circle passes through two points which are not at opposite ends of a diameter of a sphere. The shorter arc of this great circle is the shortest distance between these two points, measured over the surface of the sphere.

13-12 Fig 1 Great Circle



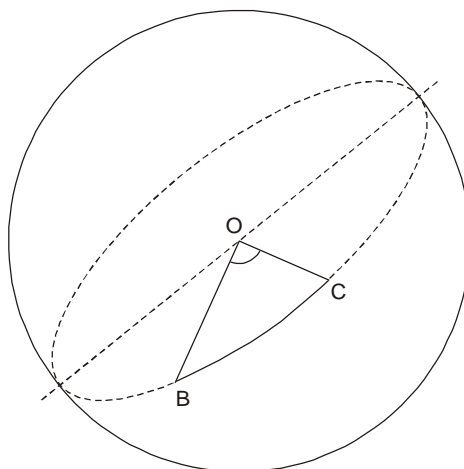
13-12 Fig 2 Small Circle



Spherical Distance

3. The spherical distance between two points on the surface of a sphere is the length of the shorter great circle arc joining them. It is measured by the angle which that arc subtends at the centre of the sphere, expressed in degrees, minutes and seconds, or in radians. In Fig 3, the spherical distance BC is measured as \widehat{BOC} , e.g. $BC = 42^\circ 27'$ means that the arc BC subtends an angle of $42^\circ 27'$ at the centre of the sphere.

13-12 Fig 3 Spherical Distance



- 4. Angular measurement of spherical distance is convenient for the following reasons:
 - a. The actual length of the arc is readily obtained, given the radius of the sphere, since the length of the arc = radius of the sphere \times the angle subtended at the centre (in radians). In the

case of the Earth, by definition, an arc of length 1 nautical mile subtends an angle of 1'. Thus, if B and C are two points on the surface of the Earth, the length of BC is given directly by converting \widehat{BOC} to minutes, e.g.:

$$BC = 42^\circ 27' = 2,547 \text{ nm}$$

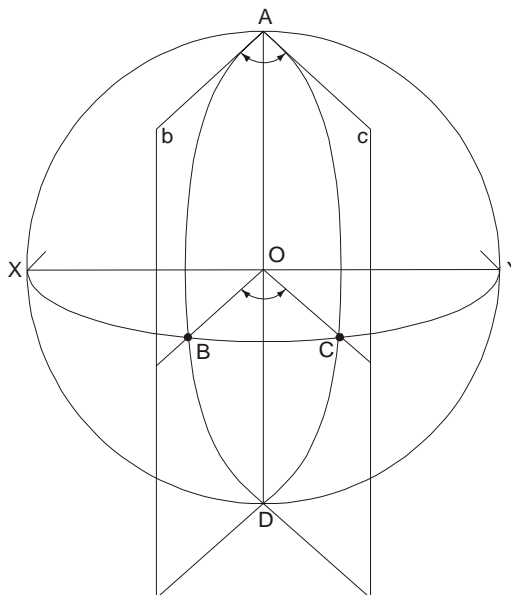
- b. Angular measurement of spherical distance can be made without reference to the size of the sphere; this is of value when dealing with an abstract quantity such as the celestial sphere.
- c. The use of angular measurement of the sides simplifies the solution of the various formulae used in spherical trigonometry.

Spherical Angle

5. A spherical angle is formed at a point where two great circles intersect and is measured by the angle between the great circles at that point. This is the equivalent of measuring the angle between the planes of the two great circles in a plane mutually perpendicular to them both.

6. In Fig 4, \widehat{BAC} is the spherical angle formed by the great circles AB and AC at the point A and is measured by $\widehat{bAc} = \widehat{BOC} = \text{arc BC}$. BOC is a plane perpendicular to the planes of the great circles ABD and ACD, and is contained in the plane of the great circle XBCY. OA is perpendicular to the plane of XBCY and A is a pole of that great circle. By definition, \widehat{BOC} is a measure of the spherical angle at A, from which it follows that the length of arc BC is also a measure of A. Thus, the spherical angle formed at a point may be measured by the arc intercepted between those great circles along the great circle to which that point is a pole.

13-12 Fig 4 Spherical Angle

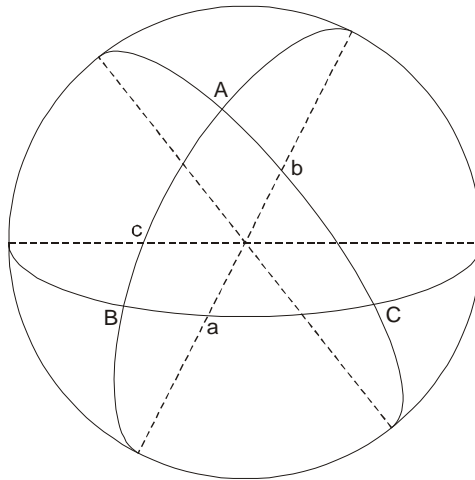


The Spherical Triangle

7. A spherical triangle is formed by the intersection of three great circles on the surface of a sphere. Fig 5 shows such a triangle, ABC. The arcs BC, CA and AB form the sides of the triangle and are denoted by a, b and c respectively. The angles \widehat{BAC} , \widehat{ABC} , and \widehat{BCA} form the angles of the triangle

and are denoted by A , B and C . Note that by this convention, a is the side opposite angle A , b is opposite angle B and c is opposite angle C .

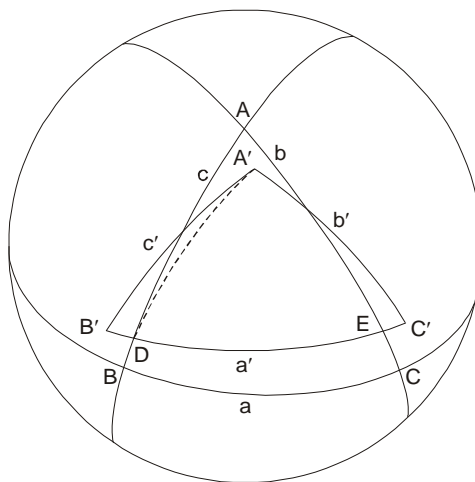
13-12 Fig 5 Spherical Triangle



The Polar Triangle

8. Consider the spherical triangle ABC in Fig 6. Point A' is a pole of the great circle of which side a is a part and is that particular pole which lies on the same side of the great circle as angle A . Similarly, B' and C' are the poles appropriate to sides b and c respectively.

13-12 Fig 6 Polar Triangle



9. The 3 points so obtained are joined by arcs of great circles $B'A'$, $A'C'$ and $C'B'$ giving a second spherical triangle $A'B'C'$. The original triangle is called the primitive and the second, the polar triangle. It should be noted that in many cases the shape of the polar triangle might bear little resemblance to that of its primitive.

10. From para 2c, if A' is the pole of arc a then the arc BA' is a quadrant. But point A' lies on the arc b' , therefore B is also a pole of arc b' . Similarly, A and C are poles of arcs a' and c' respectively; thus triangle ABC is the polar triangle of $A'B'C'$. So, if one triangle should be a polar triangle of another, the latter will be the polar triangle of the former.

Relationship between the Primitive and Polar Triangles

11. In Fig 6, let D and E be the points where the arc a' is intercepted by the arcs c and b respectively. Then, from para 6, since A is a pole of a' , the spherical angle A is measured by the arc DE . But $B'E$ and $C'D$ are both quadrants:

$$\therefore B'E + C'D = 180^\circ$$

$$\text{also } B'E + C'D = B'C' + DE$$

$$\begin{aligned} \therefore DE &= 180^\circ - B'C' \\ &= 180^\circ - a' \end{aligned}$$

or, angle A is the supplement of the angle subtended by arc a' . Similarly, it can be proved that:

$$A' = 180^\circ - a, B' = 180^\circ - b, C' = 180^\circ - c$$

$$\text{and } a' = 180^\circ - A, b' = 180^\circ - B, c' = 180^\circ - C$$

Properties of Spherical Triangles

12. The following statements are rules of spherical trigonometry and are stated without the proof which may be found in any basic spherical trigonometry primer. For a spherical triangle:

- a. By convention, each side is less than 180° .
- b. From a above it follows that each angle must be less than 180° .
- c. Any two sides are together greater than the third side.
- d. If two sides are equal, the angles opposite those sides are equal; conversely, if two angles are equal, then the sides opposite those angles are equal.
- e. The greater side is opposite the greater angle; conversely, the greater angle is opposite the greater side.
- f. The sum of all the sides is less than 360° .
- g. The sum of all the angles lies between 180° and 540° .

Determination of Spherical Triangles

13. The spherical triangle has 6 parts, i.e. 3 sides and 3 angles. In general, if any 3 parts are known the triangle is fixed. The following combination of 3 parts all determine unique triangles:

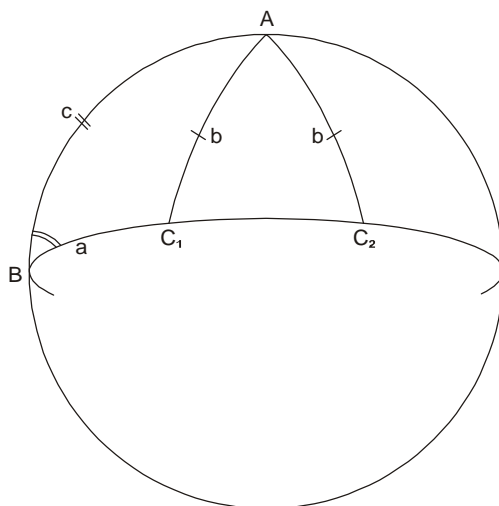
- a. Two sides and the included angle.
- b. Two angles and the included side.
- c. Three sides.
- d. Three angles.

In the following cases, there may be 1 or 2 solutions:

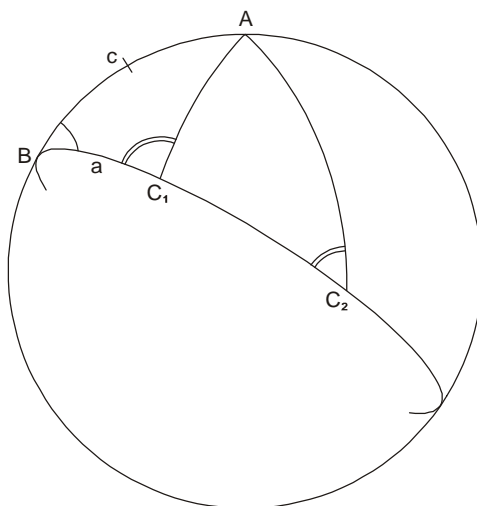
- e. Two sides and a non-included angle.
- f. Two angles and a non-included side.

For example, in Fig 7, given c , b and B , there are 2 possible triangles, ABC_1 and ABC_2 . In Fig 8, given B , C and c , there are again 2 possible triangles, ABC_1 and ABC_2 .

13-12 Fig 7 Two Possible Triangles given Two Sides and a Non-included Angle



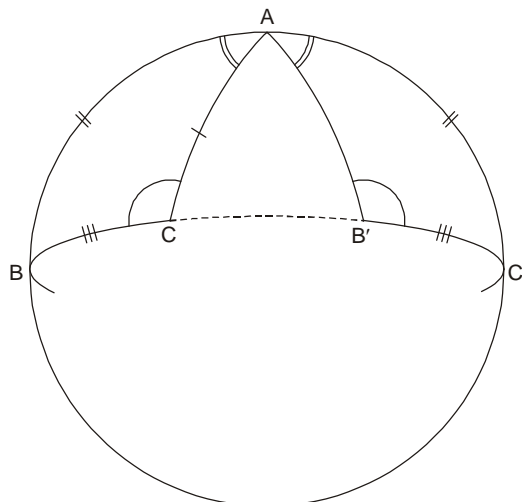
13-12 Fig 8 Two Possible Triangles given Two Angles and a Non-included Side



Symmetrical Equality

14. The 2 triangles in Fig 9 obey all the normal rules of congruency, i.e. the sides and angles of one are equal to the corresponding sides and angles of the other. However, triangle ABC cannot be superimposed on $AB'C'$ since its curvature is in the opposite sense. Hence, the 2 triangles cannot be truly congruent since they fail in this respect. They are, therefore, said to be symmetrically equal.

13-12 Fig 9 Symmetrically Equal Triangles



15. **Points to Note.** The following points of difference between plane and spherical triangles should be noted:

- a. Given 2 angles of a spherical triangle, the third angle is still undetermined since, from para 12g, the sum of the three angles may be anywhere between 180° and 540° . This is in contrast to the plane triangle in which the third angle may be obtained by subtracting the sum of the two known angles from 180° .
- b. Since 3 angles determine a unique spherical triangle (para 13d), it follows that similar triangles do not occur on the same sphere.
- c. In plane trigonometry, the angles of an equilateral triangle are all 60° . In an equilateral spherical triangle, however, whilst the 3 angles are all equal, their value is not restricted to 60° .

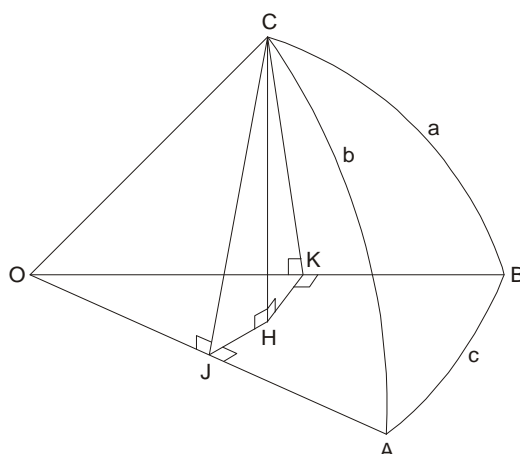
FORMULAE FOR THE SOLUTION OF SPHERICAL TRIANGLES

16. Spherical triangles may be solved by various formulae as listed below.

The Sine Formula

17. From C (Fig 10) drop a perpendicular to the plane OAB, meeting the plane in H. Draw a perpendicular from C to OA, meeting that line in J. Then JH will be perpendicular to OA. Draw a perpendicular from C to OB, meeting that line in K. Then KH will be perpendicular to OB.

13-12 Fig 10 Proof of the Sine Formula



18. From Fig 10:

$$\begin{aligned} CJ &= OC \sin \hat{C}OJ \\ OJ &= OC \cos \hat{C}OJ \\ JH &= CJ \cos \hat{C}JH \\ &= OC \sin \hat{C}OJ \cos \hat{C}JH \\ CH &= OC \sin \hat{C}OJ \sin \hat{C}JH \end{aligned}$$

Let $OC = 1$ unit

$\hat{C}OJ = b$ (angle subtended at the centre)

$\hat{C}JH = A$ (CJ is in the plane of arc b , JH is in the plane of arc c , $\hat{C}JH$ is the angle between the planes)

$\therefore OJ = \cos b$

and $CJ = \sin b$

$JH = \sin b \cos A$

$CH = \sin b \sin A$(1)

Similarly,

$OK = \cos a$

and $CK = \sin a$

$KH = \sin a \cos B$

$CH = \sin a \sin B$(2)

Equating (1) and (2):

$\sin b \sin A = \sin a \sin B$(3)

By repeating this construction in the plane OCB it may be shown that:

$\sin b \sin C = \sin c \sin B$(4)

From (3) and (4) it follows that:

$$\frac{\sin a}{\sin A} = \frac{\sin b}{\sin B} = \frac{\sin c}{\sin C} \dots\dots\dots(5)$$

Or, the sines of the angles are proportional to the sines of the sides opposite. This expression is known as the Sine Formula

The Cosine Formula

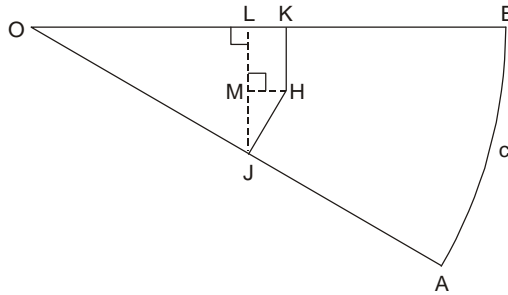
19. Fig 11 shows the plane OAB of Fig 10. By definition $\hat{B}O A = c$ (angle subtended at the centre). Draw JL perpendicular to OB ; draw HM perpendicular to JL .

Then $OK = OJ \cos c + JH \sin \hat{M}JH$.

But $\hat{M}JH = 90^\circ - \hat{M}J O = \hat{B}O A = c$.

$\therefore OK = OJ \cos c + JH \sin c$(6)

13-12 Fig 11 The Plane OAB of Fig 10



Substituting the values for OK, OJ, and JH in (6)

$$\cos a = \cos b \cos c + \sin b \sin c \cos A \dots\dots(7)$$

By repeating the construction in the planes of OCB and OCA, 2 further expressions are obtained, viz:

$$\cos b = \cos a \cos c + \sin a \sin c \cos B \dots\dots(8)$$

$$\cos c = \cos a \cos b + \sin a \sin b \cos C \dots\dots(9)$$

Equations 7, 8 and 9 are of the same form and permit the determination of 1 side knowing the other 2 sides and the included angle.

Examples of the Use of the Sine Formula

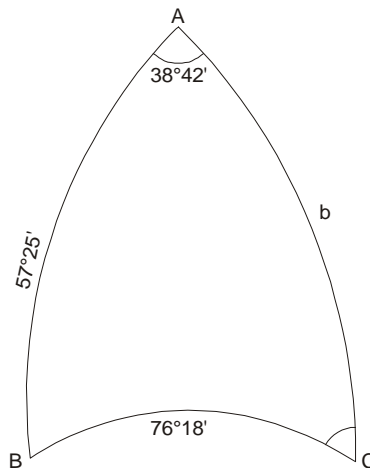
20. The Sine formula can be used in the following cases:

- a. To find a non-included angle, given 2 sides and a non-included angle.
- b. To find a non-included side, given 2 angles and a non-included side.

In accordance with para 13, these cases are ambiguous and may yield 2 possible results, as demonstrated below.

21. **Example 1, Sine Formula.** From Fig 12, find C, given $A = 38^\circ 42'$, $a = 76^\circ 18'$, $c = 57^\circ 25'$.

13-12 Fig 12 Example 1 - Sine Formula



From the Sine formula

$$\frac{\sin C}{\sin c} = \frac{\sin A}{\sin a}$$

$$\sin C = \frac{\sin A \sin c}{\sin a}$$

$$\sin C = \frac{\sin 38^\circ 42' \sin 57^\circ 25'}{\sin 76^\circ 18'}$$

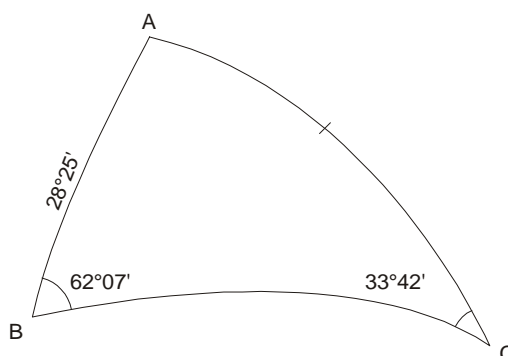
Using calculator or logarithms, $\log \sin C = 1.73422$.

$$\therefore C = 32^\circ 50' \text{ or } 147^\circ 10'$$

In this particular case, one result may be eliminated by applying the rule that the greater side must be opposite the greater angle (para 12e): a is greater than c hence A must be greater than C ; therefore, C cannot have a value $147^\circ 10'$. The requirements are satisfied by 1 triangle only and, therefore, $C = 32^\circ 50'$.

22. **Example 2, Sine Formula.** From Fig 13, find b , given $c = 28^\circ 25'$, $B = 62^\circ 07'$ and $C = 33^\circ 42'$.

13-12 Fig 13 Example 2 - Sine Formula



$$\frac{\sin b}{\sin B} = \frac{\sin c}{\sin C}$$

$$\sin b = \frac{\sin B \sin c}{\sin C}$$

$$= \frac{\sin 62^\circ 07' \sin 28^\circ 25'}{\sin 33^\circ 42'}$$

So, using calculator or logarithms, $\log \sin b = \bar{1}.87973$.

$$\therefore b = 49^\circ 18' \text{ or } 130^\circ 42'$$

In this case, B is greater than C , thus b must be greater than c . Both values of b satisfy this requirement, hence the ambiguity is unresolved, and both results must be accepted.

Examples of the Use of the Cosine Formula

23. The Cosine formula may be used as follows:

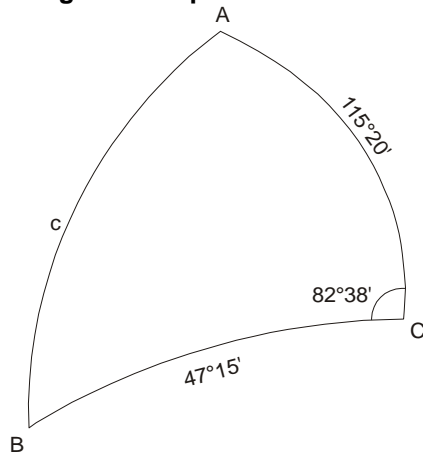
- a. To find the third side, given 2 sides and the included angle.

b. To find the third angle, given 2 angles and the included side (by transposition to the polar triangle).

Both these cases determine unique triangles; hence no ambiguity will arise.

24. **Example 1, Cosine Formula.** From Fig 14, find c , given $a = 47^\circ 15'$, $b = 115^\circ 20'$ and $C = 82^\circ 38'$.

13-12 Fig 14 Example 1 - Cosine Formula



From the Cosine formula:

$$\cos c = \cos a \cos b + \sin a \sin b \cos C$$

$$\cos c = \cos 47^\circ 15' \cos 115^\circ 20' + \sin 47^\circ 15' \sin 115^\circ 20' \cos 82^\circ 38'$$

Now $115^\circ 20'$ is in the second quadrant. It may thus be written that:

$$\cos 115^\circ 20' = -\cos 64^\circ 40'$$

$$\sin 115^\circ 20' = \sin 64^\circ 40'$$

$$\therefore \cos c = -\cos 47^\circ 15' \cos 64^\circ 40' + \sin 47^\circ 15' \sin 64^\circ 40' \cos 82^\circ 38'$$

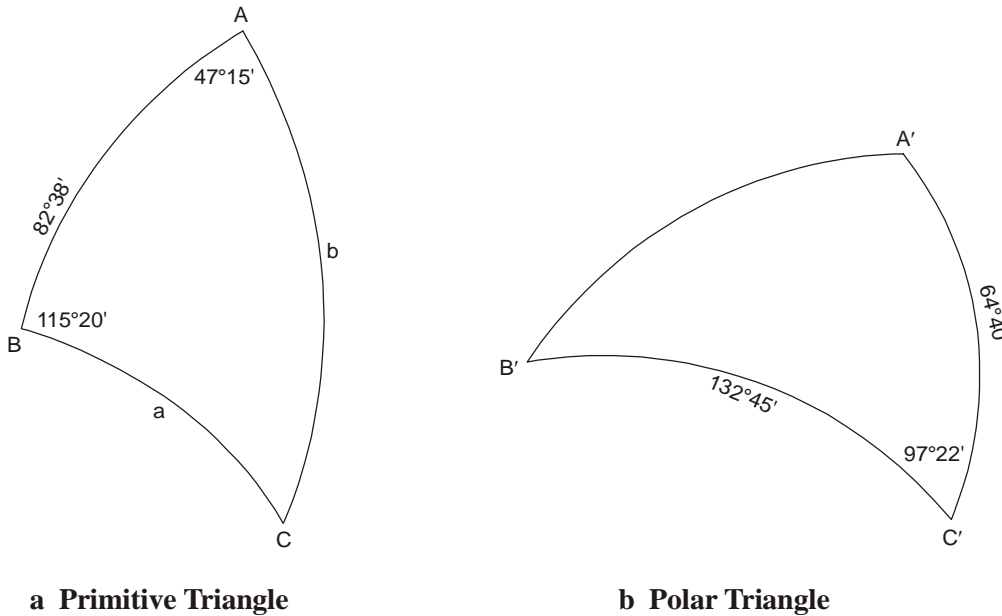
$$= -0.29044 + 0.08510$$

$$\therefore \cos c = -0.20534$$

$$c = 180^\circ - 78^\circ 09' = 101^\circ 51'$$

25. **Example 2, Cosine Formula.** From triangle ABC in Fig 15a, find C, given $A = 47^\circ 15'$, $B = 115^\circ 20'$, $c = 82^\circ 38'$. In this case, two angles and the included side are given and the Cosine formula is not directly applicable. However, the polar triangle may be derived thus; from the rules of para 11:

13-12 Fig 15 Example 2 - Cosine Formula Indirect Application



$$\begin{aligned} a' &= 180^\circ - A = 180^\circ - 47^\circ 15' = 132^\circ 45' \\ b' &= 180^\circ - B = 180^\circ - 115^\circ 20' = 64^\circ 40' \\ C' &= 180^\circ - c = 180^\circ - 82^\circ 38' = 97^\circ 22' \\ c' &= 180^\circ - C'. \end{aligned}$$

The given quantities are now in terms of 2 sides and an included angle and:

$$\begin{aligned} \cos c' &= \cos a' \cos b' + \sin a' \sin b' \cos C' \\ \cos c' &= \cos 132^\circ 45' \cos 64^\circ 40' + \sin 132^\circ 45' \sin 64^\circ 40' \cos 97^\circ 22' \\ \cos c' &= -\cos 47^\circ 15' \cos 64^\circ 40' - \sin 47^\circ 15' \sin 64^\circ 40' \cos 82^\circ 38' \\ &= -0.29044 - 0.08510 \\ &= -0.37554 \\ c' &= 112^\circ 03\frac{1}{2}' \\ \therefore C' &= 180^\circ - 112^\circ 03\frac{1}{2}' = 67^\circ 56\frac{1}{2}' \end{aligned}$$

The Haversine Formula

26. In the preceding examples it has been necessary to consider the sign of the various functions. This is an inconvenience and calculations would be much simplified if only positive values occurred. The Haversine (half-reverse-sine) formula may be used to achieve this object.

27. The expression $\frac{1 - \cos A}{2}$ is known as the haversine of an angle A, written hav A and has special properties. Thus:

When	$A = 0^\circ$	$\cos A = 1$	and	$\text{hav } A = 0$
	$A = 90^\circ$	$\cos A = 0$	and	$\text{hav } A = \frac{1}{2}$
	$A = 180^\circ$	$\cos A = -1$	and	$\text{hav } A = 1$
	$A = 270^\circ$	$\cos A = 0$	and	$\text{hav } A = \frac{1}{2}$
	$A = -90^\circ$	$\cos A = 0$	and	$\text{hav } A = \frac{1}{2}$
	$A = -180^\circ$	$\cos A = -1$	and	$\text{hav } A = 1$
	$A = -270^\circ$	$\cos A = 0$	and	$\text{hav } A = \frac{1}{2}$

$$A = -360^\circ \quad \cos A = 1 \quad \text{and} \quad \text{hav } A = 0$$

The value of the haversine never exceeds 1 and is always positive irrespective of whether the angle is positive or negative and since:

$$\text{hav } a = \frac{1 - \cos a}{2} \quad \left(\text{and } \text{hav } A = \frac{1 - \cos A}{2} \right)$$

$$\cos a = 1 - 2 \text{hav } a \quad (\text{and } \cos A = 1 - 2 \text{hav } A)$$

Substituting for $\cos a$ and $\cos A$ in the Cosine formula (para 19):

$$\cos a = \cos b \cos c + \sin b \sin c \cos A$$

$$1 - 2 \text{hav } a = \cos b \cos c + \sin b \sin c (1 - 2 \text{hav } A)$$

$$\text{or } 1 - 2 \text{hav } a = \cos b \cos c + \sin b \sin c - 2 \sin b \sin c \text{hav } A$$

$$\text{Now: } \cos b \cos c + \sin b \sin c = \cos (b - c)$$

$$\text{and } \cos (b - c) = 1 - 2 \text{hav } (b - c)$$

$$\therefore 1 - 2 \text{hav } a = 1 - 2 \text{hav } (b - c) - 2 \sin b \sin c \text{hav } A$$

$$\text{From which: } \text{hav } a = \text{hav } (b - c) + \sin b \sin c \text{hav } A$$

28. Since the haversine is always positive, the value of $\text{hav } (b - c)$ is positive no matter what values are assigned to b and c and the equation may be simplified by writing $(b \sim c)$, meaning the difference between b and c . So:

$$\text{hav } a = \text{hav } (b \sim c) + \sin b \sin c \text{hav } A \dots\dots\dots(10)$$

Similarly, it may be shown that:

$$\text{hav } b = \text{hav } (a \sim c) + \sin a \sin c \text{hav } B$$

$$\text{and } \text{hav } c = \text{hav } (a \sim b) + \sin a \sin b \text{hav } C$$

By convention, only angles and sides up to 180° are considered. Terms $\sin a$, $\sin b$ and $\sin c$ will, therefore, always be positive; hence every term in the Haversine formula is positive.

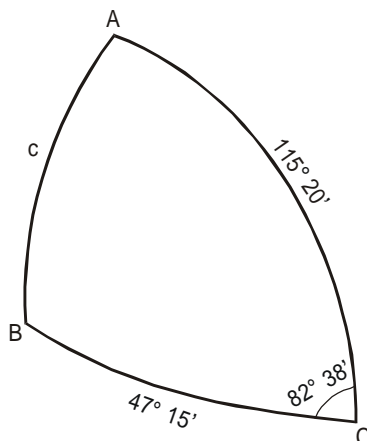
Example of the Use of the Haversine Formula

29. The Haversine formula is used to:

- a. Find the third side, given 2 sides and the included angle.
- b. Find the third angle, given 2 angles and the included side (by transposition to the polar triangle).

From triangle ABC in Fig 16, find c given $a = 47^\circ 15'$, $b = 115^\circ 20'$ and $C = 82^\circ 38'$.

13-12 Fig 16 Application of the Haversine Formula



$$\text{hav } c = \text{hav } (47^\circ 15' \sim 115^\circ 20') + \sin 47^\circ 15' \sin 115^\circ 20' \text{hav } 82^\circ 38'$$

$$\text{hav } c = \text{hav } 68^\circ 05' + \sin 47^\circ 15' \sin 64^\circ 40' \text{hav } 82^\circ 38'$$

$$= 0.31337 + 0.28931$$

$$\therefore c = 101^\circ 51'$$

The Cosecant Formula

30. Given the 3 sides of a spherical triangle, the 3 angles may be determined by substitution in the Haversine formula. This is transposed for convenience to give the Cosecant formula:

$$\text{hav } a = \text{hav } (b \sim c) + \sin b \sin c \text{hav } A$$

$$\text{hav } A = \frac{\text{hav } a - \text{hav } (b \sim c)}{\sin b \sin c}$$

$$\therefore \text{hav } A = \text{hav } a - \text{hav } (b \sim c) \text{ cosec } b \text{ cosec } c \dots (11)$$

The other 2 forms are:

$$\text{hav } B = \text{hav } b - \text{hav } (a \sim c) \text{ cosec } a \text{ cosec } c, \text{ and}$$

$$\text{hav } C = \text{hav } c - \text{hav } (a \sim b) \text{ cosec } a \text{ cosec } b$$

From these equations, all 3 angles may be determined. Alternatively, given 3 angles, the 3 sides can be obtained by transposition to the polar triangle.

Example of the Use of the Cosecant Formula

31. From the triangle ABC in Fig 17a, find side a given $A = 82^\circ 30'$, $B = 60^\circ 52'$ and $C = 45^\circ 02'$. The corresponding sides of the polar triangle (Fig 17b) are:

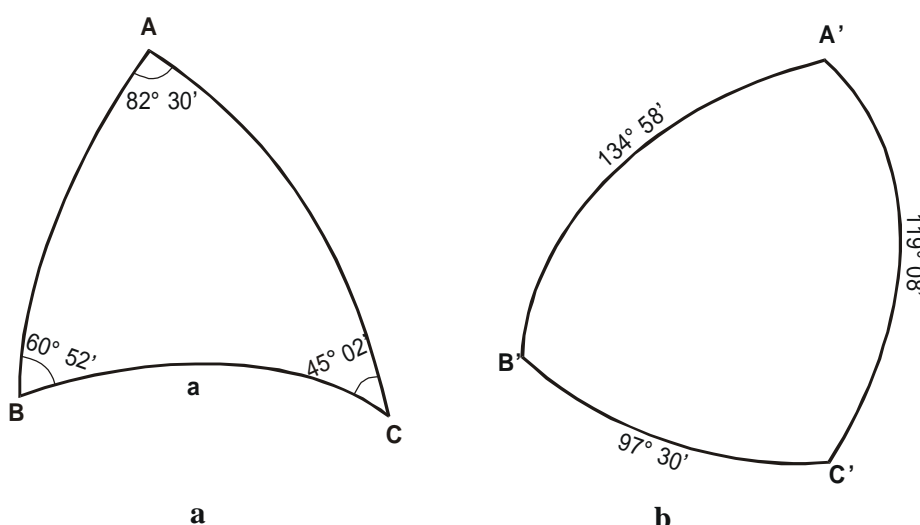
$$a' = 180^\circ - 82^\circ 30' = 97^\circ 30'$$

$$b' = 180^\circ - 60^\circ 52' = 119^\circ 08'$$

$$c' = 180^\circ - 45^\circ 02' = 134^\circ 58'$$

$$A' = 180^\circ - a$$

13-12 Fig 17 Application of the Cosecant Formula



Then:

$$\begin{aligned} \text{hav } A' &= [\text{hav } a' - \text{hav } (b' \sim c')] \text{ cosec } b' \text{ cosec } c' \\ &= [\text{hav } 97^\circ 30' - \text{hav } (119^\circ 08' \sim 134^\circ 58')] \times (\text{cosec } 119^\circ 08' \text{ cosec } 134^\circ 58') \\ &= [\text{hav } 97^\circ 30' - \text{hav } 15^\circ 50'] \times (\text{cosec } 60^\circ 52' \text{ cosec } 45^\circ 02') \\ &\quad (\because 60^\circ 52' \text{ is } 180^\circ - 119^\circ 08' \text{ and } 45^\circ 02' \text{ is } 180^\circ - 134^\circ 58') \\ &= [0.56526 - 0.01897] \times (\text{cosec } 60^\circ 52' \text{ cosec } 45^\circ 02') \end{aligned}$$

Using calculator or logarithms, $\log \text{hav } A' = \bar{1}.94642$

$$\therefore A' = 140^\circ 10'$$

and $a = 180^\circ - 140^\circ 10' = 39^\circ 50'$

Sides b and c can be found in a similar manner.

The Half-log Haversine Formula

32. By rewriting equation (11), substituting $\frac{1 - \cos a}{2}$ for hav a and $\frac{1 - \cos(b \sim c)}{2}$ for hav (b ~ c), the following expression is obtained:

$$\begin{aligned} \text{hav } A &= \frac{1}{2}[1 - \cos a - (1 - \cos(b \sim c))](\text{cosec } b \text{ cosec } c) \\ &= \frac{1}{2} \cos[(b \sim c) - \cos a](\text{cosec } b \text{ cosec } c) \end{aligned}$$

$$\begin{aligned} \text{Now: } \cos(b \sim c) - \cos a &= -2 \sin \left[\frac{(b \sim c) + a}{2} \right] \sin \left[\frac{(b \sim c) - a}{2} \right] \\ &= 2 \sin \left[\frac{a + (b \sim c)}{2} \right] \sin \left[\frac{a - (b \sim c)}{2} \right] \end{aligned}$$

$$\text{But: } \text{hav } \theta = \frac{1 - \cos \theta}{2} = \frac{2 \sin^2 \frac{\theta}{2}}{2} = \sin^2 \frac{\theta}{2}$$

$$\text{So: } \sin \frac{\theta}{2} = \sqrt{\text{hav } \theta}$$

Therefore:

$$\sin \left[\frac{a + (b \sim c)}{2} \right] = \sqrt{\text{hav } [a + (b \sim c)]}$$

$$\text{and } \sin \left[\frac{a - (b \sim c)}{2} \right] = \sqrt{\text{hav } [a - (b \sim c)]}$$

Therefore:

$$\text{hav } A = \sqrt{\text{hav } [a + (b \sim c)]} \times \sqrt{\text{hav } [a - (b \sim c)]} \times (\text{cosec } b \text{ cosec } c)$$

Similarly:

$$\text{hav } B = \sqrt{\text{hav } [b + (a \sim c)]} \times \sqrt{\text{hav } [b - (a \sim c)]} \times (\text{cosec } a \text{ cosec } c)$$

$$\text{hav } C = \sqrt{\text{hav } [c + (a \sim b)]} \times \sqrt{\text{hav } [c - (a \sim b)]} \times (\text{cosec } a \text{ cosec } b)$$

This equation is easier to manipulate than the Cosecant formula since a straight multiplication is the only operation required. A calculator or logarithms will produce the relevant results. It should be noted that:

$$\log \sqrt{\text{hav } [a + (b \sim c)]} = \frac{1}{2} \log \text{hav } [a + (b \sim c)]$$

An alternative to this formula is called the **All Natural Haversine** formula where:

$$\text{hav } A = \frac{\text{hav } a - \text{hav } (b \sim c)}{\text{hav } (b + c) - \text{hav } (b \sim c)}$$

Example of the Use of the Half-log Haversine Formula

33. Using triangle ABC in Fig 17 again, find side a given A = 82° 30', B = 60° 52' and C = 45° 02'. The corresponding sides of the polar triangle are:

$$a' = 180^\circ - 82^\circ 30' = 97^\circ 30'$$

$$b' = 180^\circ - 60^\circ 52' = 119^\circ 08'$$

$$c' = 180^\circ - 45^\circ 02' = 134^\circ 58'$$

$$b' \sim c' = 15^\circ 50'$$

$$\text{hav } A' = \sqrt{\text{hav}(97^\circ 30' + 15^\circ 50')} \times \sqrt{\text{hav}(97^\circ 30' - 15^\circ 50')} \times (\text{cosec } 119^\circ 08' \text{ cosec } 134^\circ 58')$$

$$= \sqrt{\text{hav}(113^\circ 20')} \times \sqrt{\text{hav}(81^\circ 40')} \times (\text{cosec } 60^\circ 52' \text{ cosec } 45^\circ 02')$$

$$(\because 60^\circ 52' \text{ is } 180^\circ - 119^\circ 08' \text{ and } 45^\circ 02' \text{ is } 180^\circ - 134^\circ 58')$$

Using calculator or logarithms, $\log \text{hav } A' = \bar{1}.94642$

$$\therefore A' = 140^\circ 10'$$

and $a = 180^\circ - 140^\circ 10'$

$$= 39^\circ 50'$$

sides b and c can be found in similar manner.

The Four Parts Formula

34. A case that arises frequently in practice is that in which, given 3 consecutive parts of a spherical triangle (say 2 sides and an included angle), it is required to know a further angle. This could be solved by using the Haversine and Cosecant formulae in succession, but such a method is laborious.

35. Consider any spherical triangle ABC in which a, c and B are known.

Then: **cos a = cos b cos c + sin b sin c cos A.....(12)**

$$\cos b = \cos a \cos c + \sin a \sin c \cos B$$

$$\sin b = \frac{\sin a}{\sin A} \sin B$$

Substituting for sin b and cos b in equation (12):

$$\begin{aligned} \cos a &= \cos c (\cos a \cos c + \sin a \sin c \cos B) + \sin c \cos A \frac{\sin a}{\sin A} \sin B \\ &= \cos a \cos^2 c + \cos c \sin a \sin c \cos B + \sin a \sin c \sin B \cot A \end{aligned}$$

$$\cos a (1 - \cos^2 c) = \sin a \sin c (\cos c \cos B + \sin B \cot A)$$

Now: $1 - \cos^2 c = \sin^2 c$

So: $\cos a \sin^2 c = \sin a \sin c (\cos c \cos B + \sin B \cot A)$

$$\therefore \cot a \sin c = \cos c \cos B + \sin B \cot \dots \dots \dots (13)$$

Likewise: $\cot a \sin b = \cos b \cos C + \sin C \cot A$
 $\cot b \sin c = \cos c \cos A + \sin A \cot B$
 $\cot c \sin a = \cos a \cos B + \sin B \cot C$
 $\cot b \sin a = \cos a \cos C + \sin C \cot B$
 $\cot c \sin b = \cos b \cos A + \sin A \cot C$

36. The following rules, using Fig 18 as an example, may assist in memorizing these formulae:

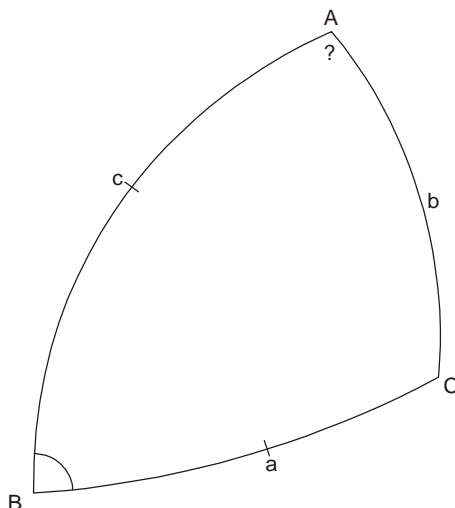
- a. The four parts follow consecutively around the spherical triangle, eg A, c, B, a in the example.
- b. c and B are known as inner parts.
- c. A (to be found) and a are known as outer parts.
- d. The equations always follow the form cot sin cos, cos sin cot.
- e. The sides always appear in the first 3 terms, the angles always appear in the last 3 terms.
- f. Each inner part appears twice in the equation.
- g. Each outer part appears only once.
- h. The outer parts appear at each end of the equation.

Applying these rules to Fig 18:

The sequence is: $\cot \sin = \cos \cos + \sin \cot$

B and c must both appear twice, a appears only once and at the left-hand end since it is a side and must be in the first 3 terms; Therefore, A must be at the right-hand end. Thus, $\cot a \sin c = \cos c \cos B + \sin B \cot A$ which matches equation (13) and is correct.

13-12 Fig 18 Illustration of Four Parts Formula



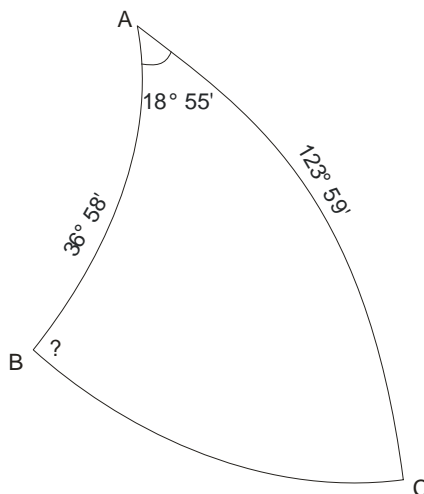
Example of the Use of the Four Parts Formula

37. The Four Parts formula can be used as follows:

- a. To find an angle, given 2 sides and the included angle.
- b. To find a side, given 2 angles and the included side.

38. In the spherical triangle ABC at Fig 19, find B given that $A = 18^\circ 55'$, $b = 123^\circ 59'$, $c = 36^\circ 58'$.

13-12 Fig 19 The Four Parts Formula – Example



B and b are the outer parts.

$$\therefore \cot b \sin c = \cos c \cos A + \sin A \cot B$$

$$\cot B = \frac{\cot b \sin c - \cos c \cos A}{\sin A} = \frac{\cot 56^\circ 01' \sin 36^\circ 58' - \cos 36^\circ 58' \cos 18^\circ 55'}{\sin 18^\circ 55'}$$

Using calculator or logarithms

$$\cot B = \frac{-0.40536 - 0.75584}{\sin 18^\circ 55'} = \frac{-1.16120}{\sin 18^\circ 55'} = -0.55411$$

$$\therefore B = -15^\circ 36'$$

$$\therefore B = 180^\circ - 15^\circ 36' = 164^\circ 24'$$

Tangent Formula or Napier’s Analogies

39. Although the lengthy working is omitted, from the Sine and Cosine formulae it can be proven that for any spherical triangle ABC:

$$\tan \frac{1}{2}(A - B) = \cot \frac{1}{2}C \frac{\sin \frac{1}{2}(a - b)}{\sin \frac{1}{2}(a + b)} \dots\dots\dots(14)$$

Also

$$\tan \frac{1}{2}(A + B) = \cot \frac{1}{2}C \frac{\cos \frac{1}{2}(a - b)}{\cos \frac{1}{2}(a + b)} \dots\dots\dots(15)$$

40. From any polar triangle A'B'C':

$$C = (180^\circ - c')$$

$$\therefore \cot \frac{1}{2}C = \cot (90^\circ - \frac{1}{2}c') = \tan \frac{1}{2}c'$$

$$(A - B) = (180^\circ - a') - (180^\circ - b') = -(a' - b')$$

$$\tan [\frac{1}{2}(A - B)] = \tan [-\frac{1}{2}(a' - b')] = -\tan \frac{1}{2}(a' - b')$$

Similarly, $(a - b) = -(A' - B')$

$$\sin [\frac{1}{2}(a - b)] = \sin [-\frac{1}{2}(A' - B')] = -\sin \frac{1}{2}(A' - B')$$

$$\frac{1}{2}(a + b) = \frac{1}{2}(180^\circ - A' + 180^\circ - B') = 180^\circ - \frac{1}{2}(A' + B')$$

$$\sin [\frac{1}{2}(a + b)] = \sin [180^\circ - \frac{1}{2}(A' + B')] = \sin \frac{1}{2}(A' + B')$$

Substituting these values in equation (14):

$$-\tan \frac{1}{2}(a' - b') = \tan \frac{1}{2}c' - \frac{\sin \frac{1}{2}(A' - B')}{\sin \frac{1}{2}(A' + B')}$$

$$\therefore \tan \frac{1}{2}(a' - b') = \tan \frac{1}{2}c' \frac{\sin \frac{1}{2}(A' - B')}{\sin \frac{1}{2}(A' + B')}$$

Thus, in any spherical triangle ABC:

$$\tan \frac{1}{2}(a - b) = \tan \frac{1}{2}c \frac{\sin \frac{1}{2}(A - B)}{\sin \frac{1}{2}(A + B)} \dots\dots\dots(16)$$

Similarly, from equation (15):

$$\tan \frac{1}{2}(a + b) = \tan \frac{1}{2}c \frac{\cos \frac{1}{2}(A - B)}{\cos \frac{1}{2}(A + B)} \dots\dots\dots(17)$$

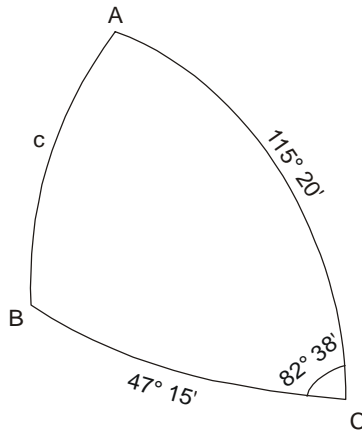
Examples of the Use of the Tangent Formulae

41. The tangent formulae are used:

- a. Given 2 sides and the included angle, to find the other 2 angles.
- b. Given 2 angles and the included side, to find the other 2 sides.
- c. Given 2 sides and the angles opposite, to find the other unknowns.

42. **Example 1, Tangent Formula.** In Fig 20 (a repeat of Fig 16), find A and B, given a = 47° 15', b = 115° 20' and C = 82° 38'.

13-12 Fig 20 Example 1 of the Tangent Formulae



$$\tan \frac{1}{2}(A - B) = \cot \frac{82^\circ 38'}{2} \frac{\sin \frac{1}{2}(47^\circ 15' - 115^\circ 20')}{\sin \frac{1}{2}(47^\circ 15' + 115^\circ 20')}$$

A useful feature of the tangent formulae emerges here. Since b > a, sin 1/2(a - b) is negative. However, with b > a, it follows that B > A and tan (A - B) is also negative. The negative sign appears on both sides of the equation and may, thus, be disregarded. It is therefore permissible to write:

$$\tan \frac{1}{2}(B - A) = \cot \frac{1}{2}C \frac{\sin \frac{1}{2}(b - a)}{\sin \frac{1}{2}(b + a)}$$

That is to say, in the application of the tangent formulae to any example the order may be changed so that the smaller quantity is subtracted from the larger and negative angles do not occur. This operation must be performed throughout all terms in the equation. Then:

$$\tan \frac{1}{2}(B - A) = \cot 41^{\circ} 19' \frac{\sin 34^{\circ} 02 \frac{1}{2}'}{\sin 81^{\circ} 17 \frac{1}{2}'}$$

$$\tan \frac{1}{2}(B + A) = \cot 41^{\circ} 19' \frac{\cos 34^{\circ} 02 \frac{1}{2}'}{\cos 81^{\circ} 17 \frac{1}{2}'}$$

Using a calculator or logarithms:

$$\frac{1}{2}(B - A) = 32^{\circ} 47 \frac{1}{2}'$$

$$\frac{1}{2}(B + A) = 80^{\circ} 52 \frac{1}{2}'$$

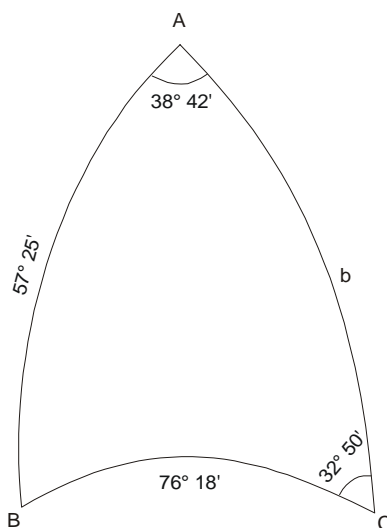
$$B = \frac{1}{2}(B - A) + \frac{1}{2}(B + A)$$

$$A = \frac{1}{2}(B + A) - \frac{1}{2}(B - A)$$

$$\therefore A = 113^{\circ} 40' \text{ and } B = 48^{\circ} 05'$$

43. **Example 2, Tangent Formula.** In Fig 21, find B, given $A = 38^{\circ} 42'$, $a = 76^{\circ} 18'$, $C = 32^{\circ} 50'$ and $c = 57^{\circ} 25'$.

13-12 Fig 21 Example 2 of the Tangent Formulae



$$\tan \frac{1}{2}(A - C) = \cot B \frac{\sin \frac{1}{2}(a - c)}{\sin \frac{1}{2}(a + c)}$$

$$\cot \frac{1}{2}B = \frac{\tan \frac{1}{2}(A - C) \sin \frac{1}{2}(a + c)}{\sin \frac{1}{2}(a - c)} = \frac{\tan \frac{1}{2}(38^{\circ} 42' - 32^{\circ} 50') \sin \frac{1}{2}(76^{\circ} 18' + 57^{\circ} 25')}{\sin \frac{1}{2}(76^{\circ} 18' - 57^{\circ} 25')}$$

$$= \frac{\tan 2^{\circ} 56' \sin 66^{\circ} 51 \frac{1}{2}'}{\sin \frac{1}{2}(76^{\circ} 18' - 57^{\circ} 25')}$$

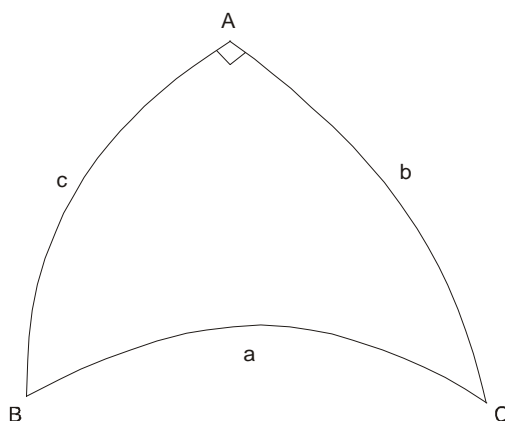
From which, using a calculator or logarithms:

$$B = 147^{\circ} 57'$$

Right-angled Spherical Triangles

44. Consider a triangle ABC in Fig 22 in which the spherical angle A = 90°.

13-12 Fig 22 Right-angled Spherical Triangle



Then, by the Four Parts formula:

$$\cot a \sin c = \cos c \cos B + \sin B \cot A$$

$$\text{but: } \cot A = \cot 90^{\circ} = 0$$

$$\therefore \cot a \sin c = \cos c \cos B$$

$$\text{and: } \cos B = \cot a \tan c$$

$$\text{Now: } \cos B = \sin (90^{\circ} - B)$$

$$\text{and: } \cot a = \tan (90^{\circ} - a)$$

$$\therefore \sin (90^{\circ} - B) = \tan (90^{\circ} - a) \tan c \dots \dots \dots (18)$$

From the Cosine formula:

$$\cos a = \cos b \cos c + \sin b \sin c \cos A$$

$$\text{but: } \cos A = 0$$

$$\text{so: } \cos a = \cos b \cos c$$

$$\therefore \sin (90^{\circ} - a) = \cos b \cos c \dots \dots \dots (19)$$

By taking each form of the Cosine and Four Parts formulae in turn a series of expressions can be obtained as follows:

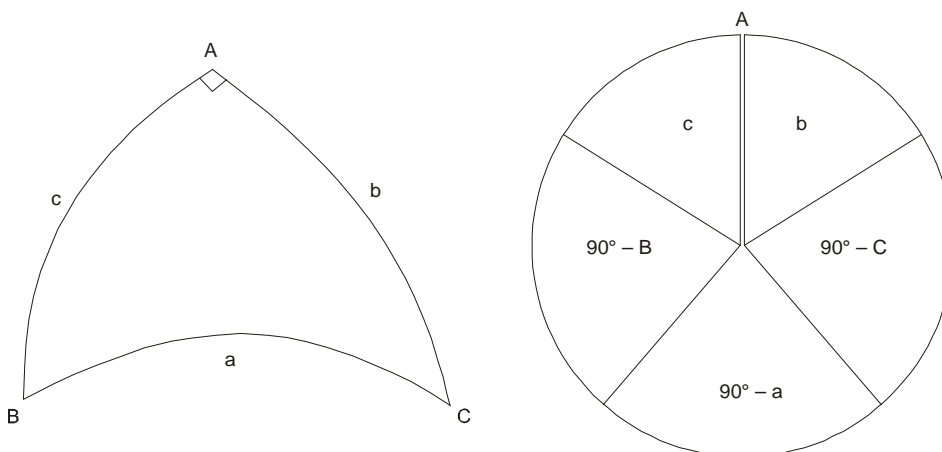
$$\begin{aligned} \sin(90^\circ - C) &= \tan(90^\circ - a) \tan b \\ \sin(90^\circ - C) &= \cos(90^\circ - B) \cos c \\ \sin(90^\circ - B) &= \cos(90^\circ - C) \cos b \\ \sin(90^\circ - A) &= \tan(90^\circ - B) \tan(90^\circ - C) \\ \sin c &= \tan(90^\circ - B) \tan b \\ \sin c &= \cos(90^\circ - a) \cos(90^\circ - C) \\ \sin b &= \tan(90^\circ - C) \tan c \\ \sin b &= \cos(90^\circ - B) \cos(90^\circ - a) \end{aligned}$$

Napier’s Rules of Circular Parts

46. The foregoing rules are difficult to memorize and are conveniently summarized in Napier’s Rules of Circular Parts. Fig 23 shows a right-angled spherical triangle with the appropriate circular parts written alongside. Note that:

- a. The parts are written down in the order in which they appear in the triangle.
- b. The right angle is not counted as a circular part and is represented in the diagram by the double line.
- c. The circular parts corresponding to the other 2 angles are the complements of those angles.
- d. The circular part corresponding to the side opposite the right angle is the complement of that side.

13-12 Fig 23 Diagram for Napier’s Rules of Circular Parts for a Right-angled Spherical Triangle



47. Provided that the circular parts are written down in accordance with the above principles, any one of the formulae in para 45 may be derived on sight from the following 2 rules:

- a. The sine of the middle part is equal to the product of the tangents of the adjacent parts.
- b. The sine of the middle part is equal to the product of the cosines of the opposite parts.

For example, select any part as the middle part. Let this be c in Fig 23.

Then: b and $(90^\circ - B)$ are the adjacent parts

$$\therefore \sin c = \tan b \tan (90^\circ - B)$$

$(90^\circ - a)$ and $(90^\circ - C)$ are the opposite parts

$$\therefore \sin c = \cos (90^\circ - a) \cos (90^\circ - C)$$

Examples of the Use of Napier’s Rules.

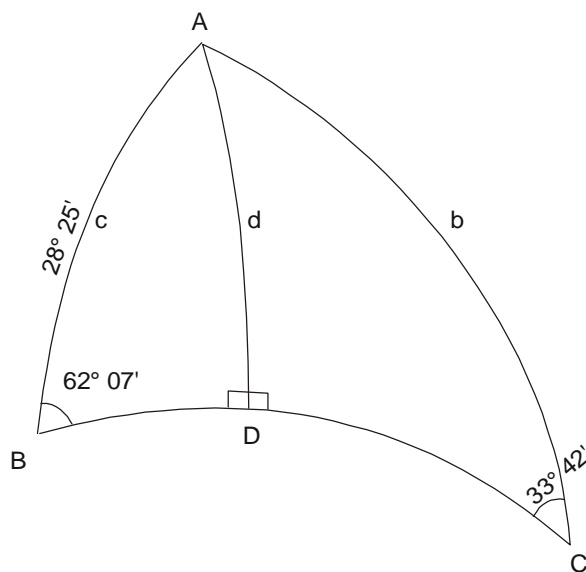
48 Napier’s Rules are especially useful when it is required to solve the spherical triangle for any other part, given:

- a. Two sides and a non-included angle.
- b. Two angles and a non-included side.

This is done by dividing the triangle into 2 right-angled triangles and applying the preceding rules. In para 22, an attempt was made to solve such a case by use of the Sine formula and it was apparent that ambiguity could arise. This, unfortunately, is also possible using Napier’s Rules.

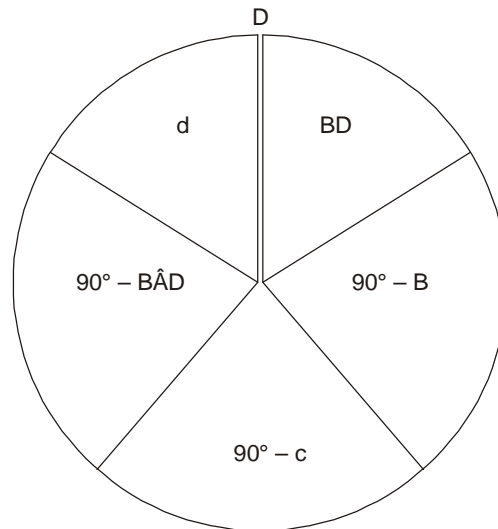
49. **Example Using Napier’s Rules.** In the spherical triangle ABC in Fig 24, find side b , given that $B = 62^\circ 07'$, $C = 33^\circ 42'$ and $c = 28^\circ 25'$. To simplify the calculation, first construct AD, a perpendicular drawn from A to arc BC. Let this be designated d . In order to obtain b , side d is required.

13-12 Fig 24 Example of Napier’s Rules



From Napier’s Rules, using the circular parts diagram for triangle ABD at Fig 25:

13-12 Fig 25 Circular Parts for Triangle ABD



$$\begin{aligned}\sin d &= \cos (90^\circ - B) \cos (90^\circ - c) \\ &= \cos 27^\circ 53' \cos 61^\circ 35'\end{aligned}$$

Using a calculator or logarithms

$$d = 24^\circ 52\frac{1}{2}' \text{ or } 155^\circ 07\frac{1}{2}'$$

But c is opposite the right-angle, so d cannot possibly be larger than c .

$$\text{Hence: } d = 24^\circ 52\frac{1}{2}'$$

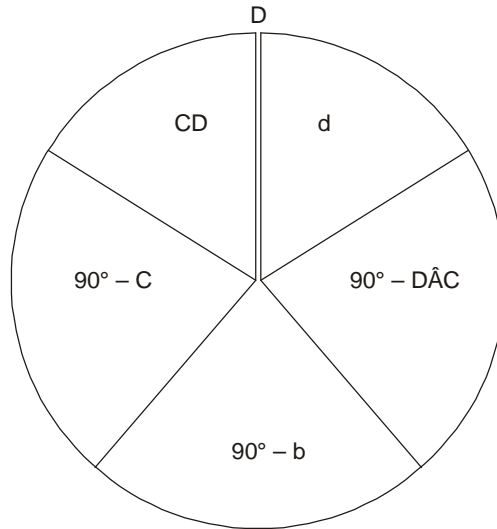
Continuing, from Napier's Rules using the circular parts diagram for triangle ADC at Fig 26:

$$\begin{aligned}\sin d &= \cos (90^\circ - C) \cos (90^\circ - b) \\ &= \sin C \sin b \\ \sin b &= \sin d \operatorname{cosec} C \\ &= \sin 24^\circ 52\frac{1}{2}' \operatorname{cosec} 33^\circ 42'\end{aligned}$$

Using a calculator or logarithms:

$$b = 49^\circ 18' \text{ or } 130^\circ 42', \text{ both of which are valid.}$$

13-12 Fig 26 Circular Parts for Triangle ADC



Right-sided Spherical Triangles

50. A spherical triangle (Fig 27) in which 1 side has a value of 90° (sometimes called a 'quadrantal triangle') may be solved by Napier's Rules because, if a triangle is right-sided, it follows that its polar triangle is right-angled. To save the labour of conversion to the polar form in such cases, the following rules for the circular parts of right-sided triangles are stated without proof:

- a. The parts are written down in the order in which they appear in the spherical triangle.
- b. The right side is not counted as a circular part.
- c. The circular parts corresponding to the other 2 sides are the complements of those sides.
- d. The circular part corresponding to the angle opposite the right side is the complement of that angle.

13-12 Fig 27 Right-sided Spherical Triangle

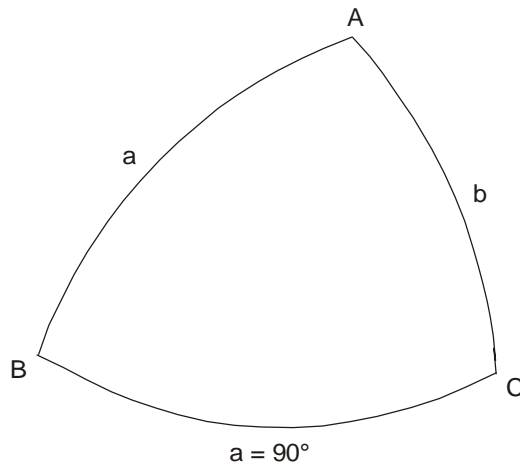
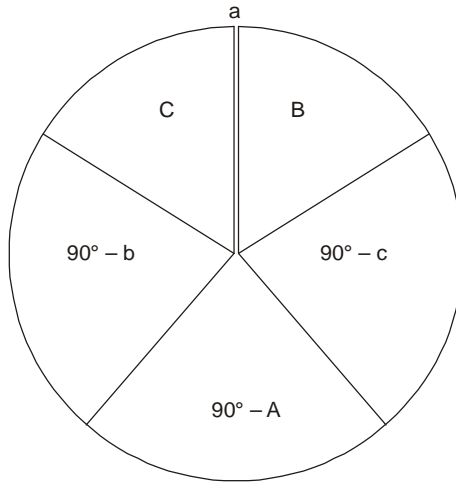


Fig 28 shows the circular parts diagram in this case.

13-12 Fig 28 Circular Parts Diagram for Right-sided Triangle



51. Napier's Rules for right-sided spherical triangles are the same as those given in para 47, viz:

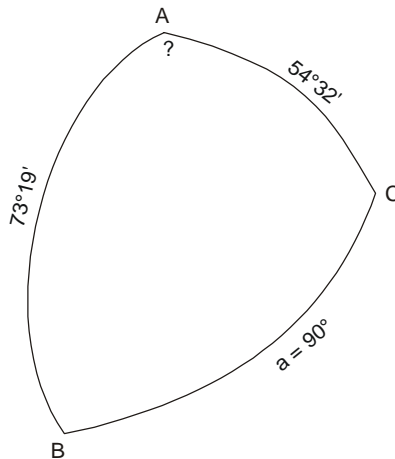
- a. The sine of the middle part is equal to the product of the tangents of the adjacent parts.
- b. The sine of the middle part is equal to the product of the cosines of the opposite parts.

The exception is that when the adjacent or opposite parts are both sides or both angles, a negative sign is added to the equation.

e.g. $\sin(90^\circ - A) = -\tan(90^\circ - b) \tan(90^\circ - c)$, but $\sin(90^\circ - b) = +\cos(90^\circ - c) \cos B$

52. **Example of a Right-sided Spherical Triangle.** In Fig 29, find A, given $a = 90^\circ$, $c = 73^\circ 19'$ and $b = 54^\circ 32'$.

13-12 Fig 29 Right-sided Triangle Example



By Napier's Rules:

$$\sin(90^\circ - A) = -\tan(90^\circ - b) \tan(90^\circ - c)$$

$$\therefore \cos A = -\tan 35^\circ 28' \tan 16^\circ 41'$$

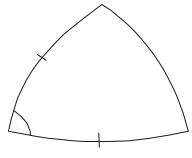
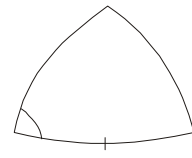
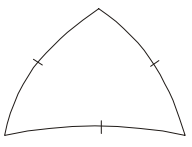
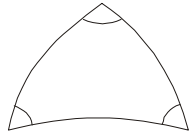
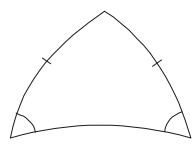
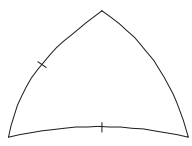
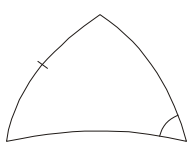
Using a calculator or logarithms:

$$A = 102^\circ 19.7'$$

Summary of Formulae

53. Table 1 summarizes the use of the various formulae covered in this chapter.

Table 1 Summary of Formulae

Case	Formulae to Use	Giving
	1 Haversine Four Parts Tangent	Third side Either angle Both angles
	2 Four Parts Tangent Convert to polar form and use Haversine	Either side Both sides Third angle
	3 Half-log Haversine Cosecant All Natural Haversine	Any angle Any angle Any angle
	4 Convert to polar form and use Half-log Haversine All Natural Haversine Cosecant	Any side Any side Any side
	5 Tangent Divide into right-angle triangles and apply Napier's Rules	{ Third angle Third side Third angle Third side
	6 Sine Divide into right-angle triangles and apply Napier's Rules	Opposite angle { Either angle Third side
	7 Sine Divide into right-angle triangles and apply Napier's Rules	Opposite angle { Either side Third angle

CHAPTER 13 - FUNCTIONS AND LIMITS

Functions

1. In Volume 13, Chapter 6, it was shown that the relationship between two variables, x and y say, can be expressed in an equation such as $y = mx + c$. The principle is not confined to the linear relationship but may also be extended to such equations as:

$$y = \sin x, y = e^x, \text{ etc.}$$

Since values are attributed to x it is known as the independent variable; the corresponding values of y may then be determined, and y is therefore known as the dependent variable.

2. The dependence of y upon x is expressed mathematically in the phrase 'y is a function of x' and is usually written as $y = f(x)$, in which $f(x)$ is a shorthand way of indicating some expression in terms of x . Thus, in the expression $y = x^2 - 4x + 3$, $f(x)$ is $x^2 - 4x + 3$; similarly, in $y = \sin 2x$, $f(x)$ is $\sin 2x$; and in $y = e^{2x}$, $f(x)$ is e^{2x} . In each case by plotting the graphs of these functions a smooth curve is obtained whose shape depends upon the nature of $f(x)$.

3. In each of the above examples an explicit statement has been made, i.e. y is equal to some function of x . Such functions are known as explicit functions.

4. It is however possible to write a function, such as $9x + 6xy + 4y^2 = 1$, in which, although there is no direct statement of y in terms of x , it is evident nevertheless that corresponding values of y could be determined by giving values to x . Such a function is known as an implicit function.

Gradients

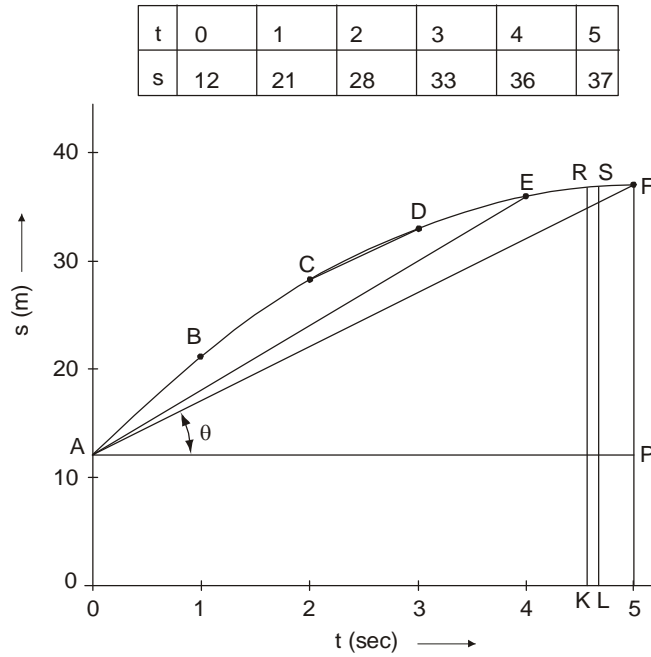
5. Suppose that an object is moving in a straight line in such a way that its distance, s metres, from a fixed point on the line at any time, t seconds after it started moving, is governed by the equation:

$$s = 12 + 10t - t^2, \text{ i.e. } s = f(t)$$

By giving a series of values to t and calculating the corresponding values of s then a graph of the function can be plotted showing how s changes as t changes. Such a graph is shown in Fig 1.

6. Information about the speed at which the object is moving can be obtained from this graph by constructing chords. For example, over the period of 5 seconds, the increase in s is indicated by $PF = 25$ metres and the object's average speed over the period is therefore $25/5$ m/sec = 5 m/sec. Letting $\angle FAP$ be called θ , then $\tan \theta = PF/AP = 25/5$. So, the average speed during the 5 secs is given by the slope, or gradient, of the chord AF . Similarly the object's average speed over the first 4 secs is given by the gradient of the chord $AE = 24/4 = 6$ m/sec. Thus, it can be inferred that the average speed over any selected period of time will be given by the gradient of the chord spanning that part of the curve. For example, the average speed of the object during the third second of its movement is given by the gradient of the chord CD , i.e. 5 m/sec.

13-13 Fig 1 Distance/Time Graph $s = 12 + 10t - t^2$



7. Whereas, for reasonably long intervals of time, it is possible to measure the gradient of the chord directly from the graph, if it becomes necessary to determine the gradient over a short period, such as KL, then this method will be difficult and inaccurate. However, it is possible to obtain the desired result by using the actual function:

$$s = 12 + 10t - t^2$$

As an example, suppose that it is required to find the average speed of the object over the period of time from $t = 3$ secs to $t = 3.1$ secs.

$$\begin{aligned} \text{After 3.1 secs, } s &= 12 + 10(3.1) - (3.1)^2 \text{ m} \\ &= (43 - 9.61) \text{ m} = 33.39 \text{ m} \end{aligned}$$

$$\text{After 3 secs, } s = 12 + 30 - 9 \text{ m} = 33 \text{ m}$$

Therefore, in 0.1 secs the object covered 0.39 m at an average speed of 3.9 m/sec.

8. By shortening the interval of time to 0.01 secs, i.e. from 3 to 3.01 secs, and substituting these figures in the function, the average speed becomes 3.99 m/sec. Taking an even shorter interval from 3 secs to 3.001 secs yields an average speed of 3.999 m/sec. If the same exercise is repeated for time intervals just prior to 3 secs, the following results are obtained:

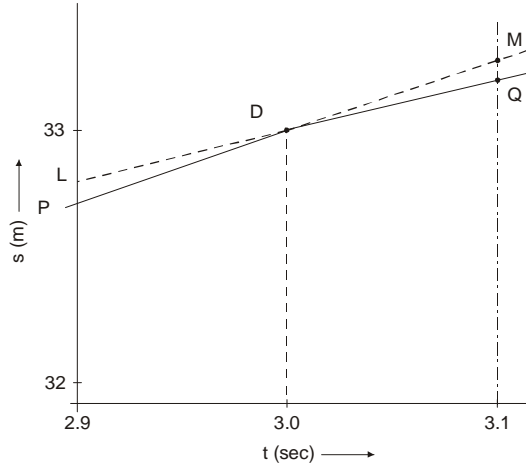
- a. From 2.9 to 3 secs: 4.1 m/sec
- b. From 2.99 to 3 secs: 4.01 m/sec
- c. From 2.999 to 3 secs: 4.001 m/sec

From the figures, it can be inferred that at the precise time of 3 secs the actual or instantaneous speed was 4 m/sec.

9. Fig 2 shows a magnified section of the graph with just two of the chords drawn. The gradient of the chord PD represents the average speed between 2.9 and 3.0 secs; the gradient of DQ represents the average speed between 3.0 and 3.1 secs. The chord PD has been extended to M and the chord QD projected back to L. Between 2.9 and 3.1 secs, the chord PM rotates about an axis through D until

it is aligned with LQ. At some instant during this rotation, the chord will take up the position of the tangent to the curve at D. It can be inferred that this will occur at the time $t = 3$ secs; thus the gradient of the tangent at a point on a distance/time graph measures the actual speed at that instant, i.e. the rate of change of s compared with the rate of change of t at that instant.

13-13 Fig 2 Two Chords on Magnified Section of $s = 12 + 10t - t^2$



Infinitesimals and Limits

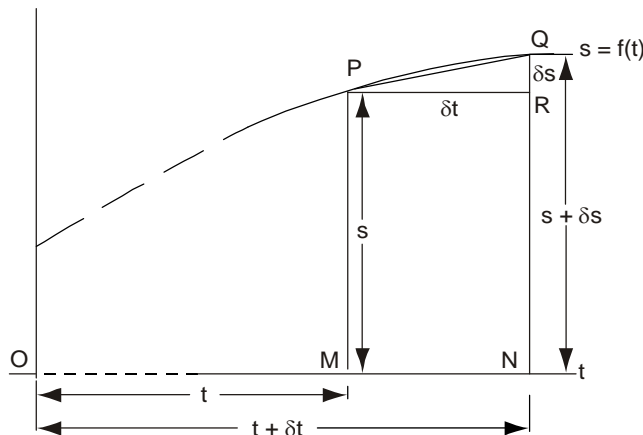
10. A shorter method of arriving at this conclusion without using specified intervals was devised by Newton. He suggested that a small increase in any quantity like s might be indicated by a special symbol δs (delta s) which has no specified size but represents a minutely small change in s . A similar small change in t would be denoted by δt , and in x by δx , etc.

11. Fig 3 shows a section of the curve

$$s = 12 + 10t - t^2$$

PM represents the distance covered, s , at time OM (t). QN represents the distance ($s + \delta s$) covered in time ON ($t + \delta t$). In both cases, δs and δt are very small. Thus, the graph shown is a greatly magnified portion of a very small arc of the curve. The gradient of the chord PQ represents the average speed between time t and $(t + \delta t)$ and can be measured as $QR/PR = \delta s/\delta t$.

13-13 Fig 3 Section of the Curve $s = 12 + 10t - t^2$



12. Since Q is on the curve:

$$s + \delta s = 12 + 10(t + \delta t) - (t + \delta t)^2$$

$$= 12 + 10t + 10\delta t - t^2 - 2t\delta t - (\delta t)^2 \dots (1)$$

and for P

$$s = 12 + 10t - t^2 \dots (2)$$

Subtracting (2) from (1)

$$\delta s = 10\delta t - 2t\delta t - (\delta t)^2$$

Dividing by δt

$$\delta s / \delta t = 10 - 2t - \delta t$$

Thus, a formula has been derived for calculating the average speed over any period of time however small.

For example, between $t = 3$ and $t = 3.0001$ secs, ie $\delta t = 0.0001$ secs:

$$\delta s / \delta t = 10 - 6 - 0.0001 = 3.9999 \text{ m/sec}$$

If a value of 0.000001 secs had been used in the formula, then $\delta s / \delta t$ would have been 3.999999.

13. Thus, it will be seen that in the expression:

$$\delta s / \delta t = 10 - 2t - \delta t$$

if the value of δt is allowed to grow smaller and smaller, i.e. approaches zero, then $\delta s / \delta t$ approaches the value $10 - 2t$. This is written as:

$$\lim_{\delta t \rightarrow 0} \frac{\delta s}{\delta t} = 10 - 2t$$

This is read as 'The limit of delta s by delta t, as delta t tends to zero, equals $10 - 2t$ '.

14. If the value of 3 secs is now substituted into this expression, then $10 - 2t = 4$, which is the value for the gradient that was deduced earlier, i.e. the actual speed at the instant of $t = 3$ secs. To indicate that this is the actual gradient at an instant then:

$$\lim_{\delta t \rightarrow 0} \frac{\delta s}{\delta t} \text{ is replaced by } \frac{ds}{dt}$$

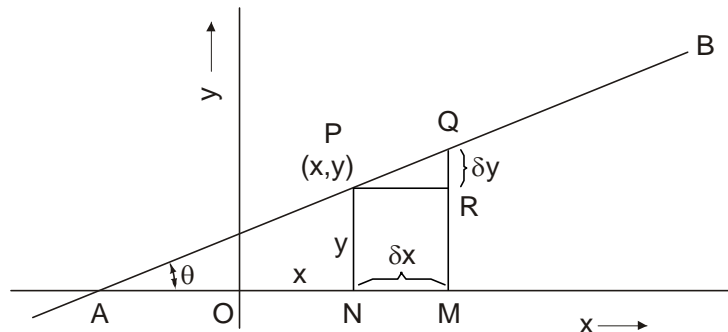
Thus, in summary, if an object is moving so that the distance s metres covered in time t seconds is a function of the time, (i.e. $s = f(t)$ and $f(t) = 12 + 10t - t^2$), then its speed at any time, t , is equal to the gradient of the tangent to the distance/time graph at time t and is defined by ds/dt which may be calculated from the expression $ds/dt = 10 - 2t$. The notation ds/dt is, therefore, a measure of the rate at which s is changing compared with the rate at which t is changing at an instant of time ' t '.

CHAPTER 14 - DIFFERENTIATION

Gradients

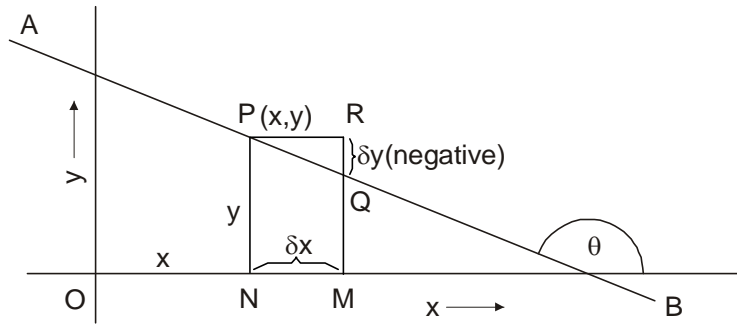
1. In Volume 13, Chapter 13 it was shown that given the relationship between distance gone and time it was possible to find the rate of change of distance with time, i.e. speed, either over a specified interval or at a particular instant, by determining the gradient of the appropriate chord or tangent. The technique is not restricted to the distance/time problem but has a general applicability whenever one parameter is changing in response to changes in another.
2. For example, when we say that a train passes us at 60 mph, we do not mean that it has travelled 60 miles in the last hour nor that it will travel 60 miles in the next hour. We mean that it will travel about 1 mile in the next minute or, better still, about half a mile in the next 30 seconds or, with still greater probability, about 88 feet in the next second. To find the speed of the train at the instant at which it passes us, we must measure the distance it goes in as small an interval of time as possible and then work out the average speed for this short interval. The shorter the interval, the closer will our answer be to the train's actual speed at that instant. In practice, there will always be an error in the measurement of instantaneous speed but we can calculate instantaneous speed with complete accuracy by means of differentiation provided we know enough about the motion.
3. In moving along the straight line AB (Fig 1), starting at P(x,y), an increase NM (= PR) in x produces an increase RQ in y. The ratio of the increase in y to the increase in x (ie RQ/PR) is called the gradient of the slope of the line AB. Clearly, the gradient is equal to $\tan \theta$.

13-14 Fig 1 Gradient of a Straight Line



4. A small interval in the x-axis, like NM, is usually denoted by δx , pronounced delta-x and must be thought of as a single symbol, the x never being separated from the variable. If y is given in terms of, or as a function of, x, i.e. $y = f(x)$, then any change in the value of x produces a change in the value of y. The symbol δy is used to denote the increment in y caused by the increment δx . Notice the difference in the definitions of δx and δy because x is the independent and y the dependent variable. In Fig 1, $\delta x = NM = PR$, $\delta y = RQ$ and the gradient of AB is $\delta y/\delta x$.
5. In Fig 2, $\delta x = NM = PR$ and δy , the corresponding increase in y, is $-RQ$. δy is negative because an increase in x causes a decrease in y. The resulting gradient $\delta y/\delta x$ is, therefore, negative.

13-14 Fig 2 Negative Gradient

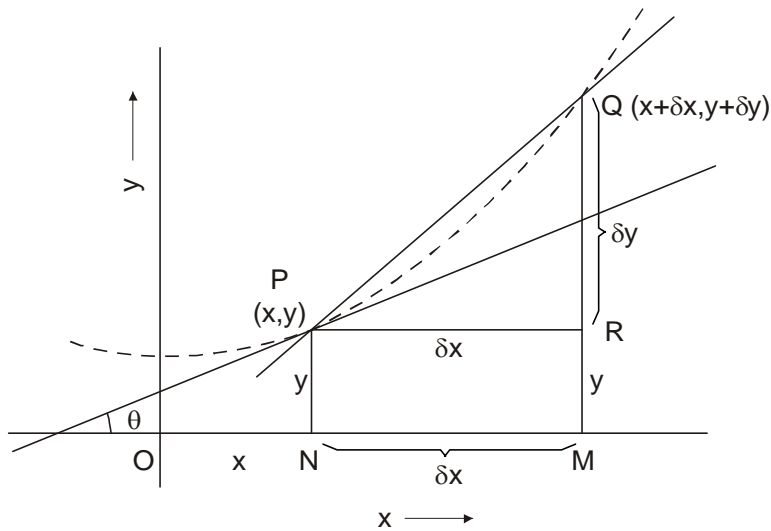


6. **Gradient of a Curve.** The gradient of a curve at any point is defined as the gradient of the tangent to the curve at that point. In Fig 3, let $P(x,y)$ be any point on the curve. Let $NM = \delta x$, then the corresponding increase in y is RQ and so $\delta y = RQ$. Then $\delta y/\delta x =$ the gradient of the chord PQ and represents the average gradient of the curve between the points P and Q . As $\delta x \rightarrow 0$ (meaning δx tends towards 0, i.e. becomes closer to 0), the value of $\delta y/\delta x$ changes and, at the same time, the chord PQ approaches its limiting position, namely the tangent to the curve at P . Hence the gradient of the curve at P is described as:

$$P = \lim_{\delta x \rightarrow 0} \frac{\delta y}{\delta x}$$

and this limit can be denoted as $\frac{dy}{dx}$, pronounced "dee-y by dee-x". The value is, of course, given also by $\tan \theta$, where θ is the angle between the tangent and OX in Fig 3.

13-14 Fig 3 Gradient of a Curve



7. **The Gradient of $y = x^3$.** As an example, Fig 4 shows the graph representing $y = x^3$. Let $P(x,y)$ be any point on the curve. NM represents a small change δx , in x , and RQ represents the consequential change δy , in y . Thus, Q is the point $(x + \delta x, y + \delta y)$. As both P and Q lie on the line then:

$$\text{for } P, \quad y = x^3 \quad (1)$$

$$\begin{aligned} \text{and for } Q, \quad y + \delta y &= (x + \delta x)^3 \\ &= x^3 + 3x^2\delta x + 3x(\delta x)^2 + \delta x^3 \end{aligned} \quad (2)$$

Subtracting (2) – (1)

$$\delta y = 3x^2\delta x + 3x(\delta x)^2 + \delta x^3$$

Dividing by δx

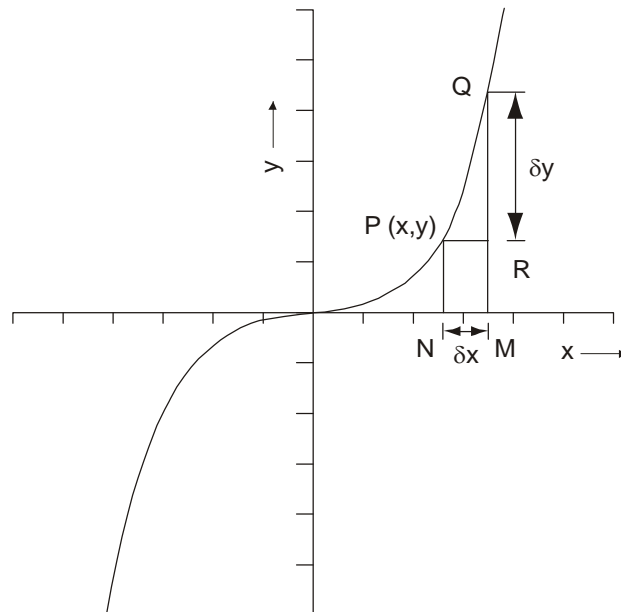
$$\frac{\delta y}{\delta x} = 3x^2 + 3x\delta x + \delta x^2$$

Then by definition (para 6) the gradient of the tangent to the curve:

$$\frac{dy}{dx} = \lim_{\delta x \rightarrow 0} \frac{\delta y}{\delta x}$$

i.e. $\frac{dy}{dx} = 3x^2$ as the $3x\delta x$ and δx^2 terms are eliminated (because δx becomes 0).

13-14 Fig 4 Fig 1 $y = f(x) = x^3$



Clearly, the gradient varies from point to point, as can be seen from the graph. The way in which the gradient varies is given by $\frac{dy}{dx}$ i.e. by the function $3x^2$. Thus the value of the gradient can be determined by substituting the appropriate value of x into the expression $3x^2$.

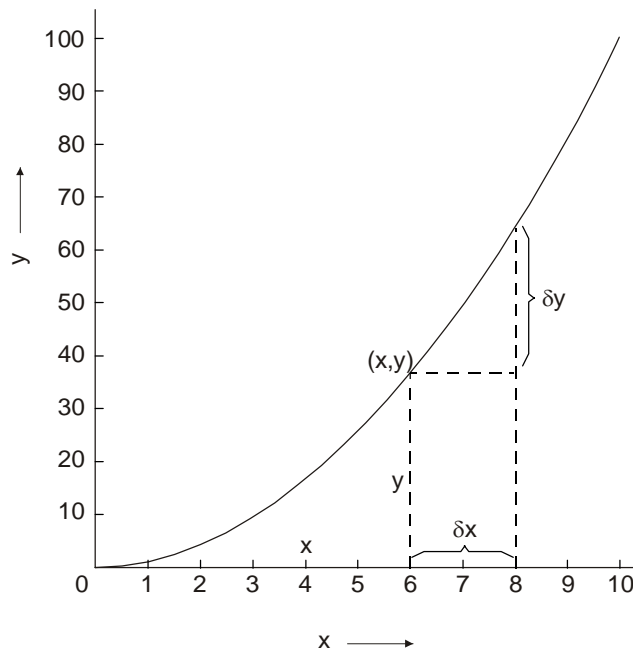
e.g. gradient at $x = 0$, is 0

gradient at $x = 1$, is 3

gradient at $x = 2$, is 12

8. **The Gradient of $y = x^2$.** Fig 5 shows the graph representing $y = x^2$.

13-14 Fig 5 Fig 2 $y = f(x) = x^2$



If (x, y) is any point on the curve then:

$$y = x^2 \quad (3)$$

If x increases by δx so that y increases by δy then

$$\begin{aligned} y + \delta y &= (x + \delta x)^2 \\ &= x^2 + 2x\delta x + (\delta x)^2 \end{aligned} \quad (4)$$

Subtracting (4) – (3) gives

$$\delta y = 2x\delta x + (\delta x)^2$$

and $\frac{\delta y}{\delta x} = 2x + \delta x$

from which the gradient of $y = x^2$ at (x, y) , ie $\frac{dy}{dx}$, is $2x$.

The Differential Coefficient (Derivative)

9. $\frac{dy}{dx}$ is called the differential coefficient of y with respect to x , or the derivative of y with respect to x .

The process of obtaining $\frac{dy}{dx}$ is called differentiating y with respect to x . Sometimes $\frac{dy}{dx}$ is written $\frac{d}{dx}(y)$

in which $\frac{d}{dx}$ is an operator, like the symbol $\sqrt{\quad}$, and means simply "the derivative of". Thus, the expressions:

$$\frac{d}{dx}(x^2 + 5x)$$

$$\text{or } \frac{d(x^2 + 5x)}{dx}$$

$$\text{or } \frac{dy}{dx} \text{ where } y = x^2 + 5x,$$

all mean the same thing, namely:

$$\lim_{\delta x \rightarrow 0} \frac{\delta y}{\delta x} \text{ when } y = x^2 + 5x$$

Differentials

10. Although $\frac{\delta y}{\delta x}$ is a quotient (i.e. it stands for $\delta y \div \delta x$), $\frac{dy}{dx}$ is not. Once the quotient $\frac{\delta y}{\delta x}$ has been

obtained, $\frac{dy}{dx}$ can be found as the limiting value of this quotient. Strictly speaking $\frac{dy}{dx}$ ought to be

regarded as a single symbol like δy or δx . However, although it is important to remember that $\frac{dy}{dx}$ is

obtained as a limit and is not strictly a quotient, it is often convenient to treat it as if it is. For example having found that when $y = x^2$, $\frac{dy}{dx} = 2x$, this result could be written as $dy = 2x dx$ or $d(x^2) = 2x dx$. In

this notation, dy , dx and $d(x^2)$ are best regarded as infinitesimal increments in y , x , and x^2 and are called the differentials of those quantities.

The General Case

11. The arguments in paras 7 and 8 to obtain the gradient or differential coefficients of $y = x^3$ and $y = x^2$, can be generalized for the case of $y = f(x)$. Remembering that if y is given as a function of x , i.e. $y = f(x)$, then any change in the value of x produces a change in the value of y . So, if (x, y) is any point on a curve then:

$$y = f(x) \quad (5)$$

and an increase in x , i.e. δx causes an increase in y , i.e. δy , such that:

$$y + \delta y = f(x + \delta x) \quad (6)$$

Subtracting (6) – (5):

$$\delta y = f(x + \delta x) - f(x)$$

Hence,
$$\frac{\delta y}{\delta x} = \frac{f(x + \delta x) - f(x)}{\delta x}$$

and so
$$\frac{dy}{dx} = \lim_{\delta x \rightarrow 0} \frac{f(x + \delta x) - f(x)}{\delta x}$$

Successive Differentiation

12. When $y = f(x) = x^3$ was differentiated the result was

$$\frac{dy}{dx} = \text{or} \left[\frac{d}{dx} (x^3) \right] = 3x^2$$

which is itself a function of x , sometimes expressed as $f'(x)$. This can itself be differentiated and can be shown to be:

$$\frac{d}{dx} \left(\frac{dy}{dx} \right) = 6x$$

The expression $\frac{d}{dx} \left(\frac{dy}{dx} \right)$ is usually written as $\frac{d^2 y}{dx^2}$ or as $f''(x)$. Similarly the result of further

differentiation of $6x$ would be written as $f'''(x)$ or $\frac{d^3 y}{dx^3} = 6$, and $f''''(x)$ or $\frac{d^4 y}{dx^4} = 0$.

Standard Derivatives

13. Examination of the successive differentiation above, known as differentiation from first principles, reveals a pattern from which a general rule can be derived. In practice, there are a number of rules which allow the derivatives of certain functions to be determined without recourse to formal working. For example, the results in para 12 show that:

$$\frac{d}{dx} ax^n = nax^{n-1}$$

where a and n are constants which may be positive or negative, fractions or integers. This formula may be applied to each level of differentiation to reach the result. For example, using the formula:

Where $y = x^2$, $\frac{dy}{dx} = 2x$ (i.e. $2 \times 1 \times x^1$) and $\frac{d}{dx}(2x) = 2$ (i.e. $1 \times 2 \times x^0$)

14. **Sum of Terms.** Where $f(x)$ is the sum of a number of terms eg:

$$y = ax^3 + bx^2 + cx + d$$

where a , b , c , and d are constants, then $f'(x)$ is the sum of the derivatives of each individual term. Thus, in this case:

$$f'(x) = 3ax^2 + 2bx + c$$

Note that the derivative of a constant = 0.

15. **Product Rule.** It may be that it becomes necessary to differentiate an expression which is the product of two functions of the same variable, e.g.:

$$y = (x + 1)(x^2 - 3)$$

In this case it would be possible to multiply out the expression without much difficulty and then differentiate the sum of the terms as outlined in para 14. However this may not be convenient, especially if there are several functions rather than just two. In this situation the product rule can be used. Let one function be u and the other v , so $y = uv$. Then, by the product rule:

$$\frac{dy}{dx} = \frac{udv}{dx} + \frac{vdu}{dx}$$

i.e. the result is the first function multiplied by the derivative of the second function, plus the second function multiplied by the derivative of the first function. Thus in the example:

$$y = (x + 1)(x^2 - 3)$$

Let, $(x + 1) = u$ and $(x^2 - 3) = v$

$$\text{Then: } \frac{dy}{dx} = \frac{udv}{dx} + \frac{vdu}{dx}$$

$$\text{Thus: } \frac{dy}{dx} = (x + 1) \times 2x + (x^2 - 3) \times 1$$

$$= 2x^2 + 2x + x^2 - 3$$

$$= 3x^2 + 2x - 3$$

This method can be extended to cover more than two factors.

$$\text{Thus, } \frac{d(uvw)}{dx} = uv \frac{dw}{dx} + wu \frac{dv}{dx} + vw \frac{du}{dx}.$$

For example:

$$\frac{d}{dx}(x + 1)(x + 2)(x + 3) = (x + 1)(x + 2) + (x + 1)(x + 3) + (x + 2)(x + 3)$$

$$= x^2 + 3x + 2 + x^2 + 4x + 3 + x^2 + 5x + 6$$

$$= 3x^2 + 12x + 11$$

16. **Function of a Function - The Chain Rule.** Consider an expression such as:

$$y = (3x^2 + 2)^2$$

Here, y is a function of $(3x^2 + 2)$ and $(3x^2 + 2)$ is a function of x . As with the product rule, in some cases the function may be simplified into the sum of several functions which may then be differentiated individually. However this will be tedious if the power is greater than, say, 2. The chain rule can be used to solve this problem as follows:

$$\text{Let, } 3x^2 + 2 = u$$

$$\text{Then, } y = u^2$$

$$\therefore \frac{dy}{du} = 2u$$

$$\text{and as } u = 3x^2 + 2$$

$$\frac{du}{dx} = 6x$$

$$\frac{dy}{du} \times \frac{du}{dx} = \frac{dy}{dx} = 2u \times 6x$$

Then substituting back for u :

$$\begin{aligned} \frac{dy}{dx} &= 2(3x^2 + 2) \times 6x \\ &= (6x^2 + 4) \times 6x \\ &= 36x^3 + 24x \end{aligned}$$

17. **Quotient of 2 Functions.** If $y = u/v$ where u and v are functions in x then:

$$\frac{dy}{dx} = \frac{v \frac{du}{dx} - u \frac{dv}{dx}}{v^2}$$

As an example consider the function:

$$y = \frac{(x^2 + 1)}{(3x + 2)}$$

$$\text{Then, } u = (x^2 + 1) \quad \therefore \frac{du}{dx} = 2x$$

$$\text{And, } v = (3x + 2) \quad \therefore \frac{dv}{dx} = 3$$

$$\begin{aligned} \text{Thus, } \frac{dy}{dx} &= \frac{(3x + 2) \times 2x - (x^2 + 1) \times 3}{(3x + 2)^2} \\ &= \frac{6x^2 + 4x - 3x^2 - 3}{(3x + 2)^2} \\ &= \frac{3x^2 + 4x - 3}{(3x + 2)^2} \end{aligned}$$

Summary

18. A summary of the results derived above together with other standard derivatives is shown in Table 1.

Table 1 Standard Differential Coefficients

Type of Function	Standard Type	Standard Differential Coefficient	Comments
Standard	$y = f(x)$	$\frac{dy}{dx}$	
Algebraic	$y = ax^n$	nax^{n-1}	Reduce the power by 1 and multiply by the original power.
Trigonometric	$y = \sin x$ $y = \cos x$ $y = \tan x$	$\cos x$ $-\sin x$ $\sec^2 x$	
Logarithmic	$y = \log_e x$	$\frac{1}{x}$	
Exponential	$y = e^{kx}$	ke^{kx}	Multiply the original function by the differential coefficient of its index.
Sum of two or more functions	$u + v$	$\frac{du}{dx} + \frac{dv}{dx}$	The differential coefficient of a sum is the sum of the differential coefficients.
Product of two functions	uv	$u \frac{dv}{dx} + v \frac{du}{dx}$	Multiply each function by the differential coefficient of the other and add the results.
Quotient of two functions	$\frac{u}{v}$	$\frac{v \frac{du}{dx} - u \frac{dv}{dx}}{v^2}$	
Function of a function	$F[f(x)]$	$\frac{df}{d[f(x)]} \times \frac{df(x)}{dx}$	Use the chain rule.

CHAPTER 15 - INTEGRATION

PRINCIPLES OF INTEGRATION

Introduction

1. Differentiating means solving the problem:

$$\text{Given } y = f(x), \text{ find } \frac{dy}{dx}.$$

The reverse problem is:

$$\text{Given } \frac{dy}{dx} = f(x), \text{ find } y.$$

This reverse process is called integration.

Indefinite Integrals

2. Consider the following:

$$\frac{dy}{dx} = ax^n \quad (1)$$

From the discussion on differentiation in Volume 13, Chapter 14, it will be apparent that the function whose derivative with respect to x is ax^n is of the form:

$$y = bx^{n+1} \quad (2)$$

If (2) is differentiated with respect to x the result is:

$$\frac{dy}{dx} = (n + 1)bx^n \quad (3)$$

Comparing (1) and (3), they will be the same if $(n + 1)b = a$, ie if $b = \frac{a}{n + 1}$. Substituting this value of b in (2):

$$y = \frac{ax^{n+1}}{n + 1} \quad (4)$$

Although (4) is certainly one solution to the problem, it is not a unique solution. Since the derivative of any constant is zero, the derivative of (4) will be unchanged if any constant, c , is added to the right-hand side of the equation. Therefore, the general solution is:

$$y = \frac{ax^{n+1}}{n + 1} + c \quad (5)$$

Because of the presence of the arbitrary constant, c , (5) is known as the indefinite integral of (1).

3. **Integration Symbol.** It is convenient to have a symbol to denote the indefinite integral of a function, thus (5) may be rewritten as:

$$\int ax^n dx = \frac{ax^{n+1}}{n + 1} + c \quad (6)$$

In this notation the \int and the dx are used as brackets to denote that everything between them is to be integrated with respect to x . The quantity so bracketed is known as the integrand. Thus ax^n is the integrand of $\int ax^n dx$, while the right-hand side of (6) is the integral. Formula (6) holds for all values of n , integral and fractional, positive and negative, with the single exception of $n = -1$. This case will be dealt with later.

4. **The Constant of Integration.** When a function is differentiated, the result represents the gradient of the graph of that function. Consequently as the integration process is the reverse of differentiation, an integral represents a function with the given gradient. Recalling that the equation of a straight line is $y = mx + c$ it will be remembered that the coefficient of x , i.e. m , equates to the gradient of the line. There is, however, an infinite family of parallel lines, all with the same gradient, m , varying in the value of the constant c . Thus the knowledge of the gradient is insufficient to describe uniquely a particular straight line. So when a function is integrated an arbitrary constant must be included to take account of the infinite number of 'parallel' functions. As an example consider the following function:

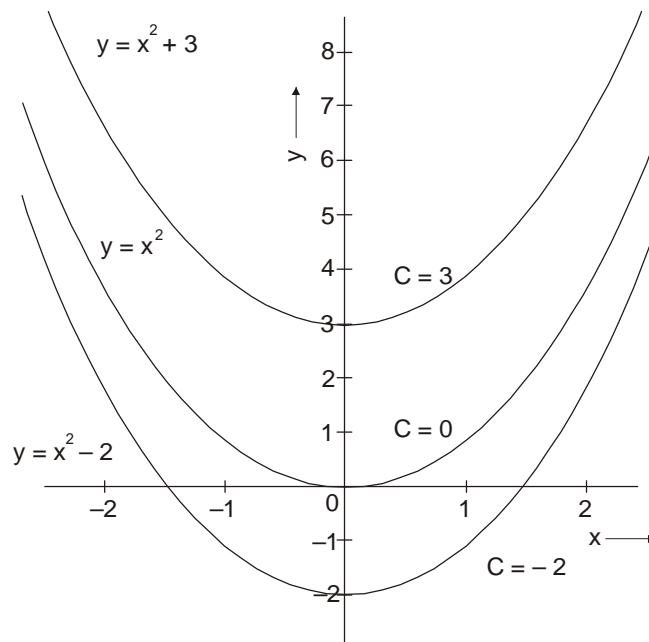
$$y = \int 2x \, dx$$

To perform the integration the power of x has to be increased by 1, and then the integrand has to be divided by the new power. Finally the arbitrary constant must be added. Thus:

$$y = x^2 + c$$

Fig 1 shows the graphs of $y = x^2 + c$ for a variety of values of c .

13-15 Fig 1 Graphs of $y = x^2 + c$



For any given value of x all of the curves have the same gradient, i.e. they all satisfy the condition $dy/dx = 2x$. In order to determine which graph is the solution to the particular problem then more information is required. For example, it may be known that $y = 3$ when $x = 0$, hence $3 = 0 + c$, thus $c = 3$. Therefore, the required solution is $y = x^2 + 3$.

5. As a practical example suppose that a body moves with an acceleration of 3ms^{-2} and it is necessary to find an expression for its velocity after t seconds:

$$\frac{dv}{dt} = 3, \text{ i.e. } v = \int 3dt = 3t + c$$

The reason that this is an inadequate description of the velocity is that, although acceleration information was provided, no information was given concerning the initial velocity of the body.

Therefore, no definite value for the velocity at any given time can be deduced. If the initial velocity was, say, 2 ms^{-1} then the velocity at any time, t , becomes $3t + 2 \text{ ms}^{-1}$ (i.e. $c = 2$).

Standard Integrals

6. Just as with differentiation, there are a number of standard integrals which are used. In general an unfamiliar expression must be converted into a standard form, or a variation or a combination of standard forms, before the integration can be accomplished. Similar rules to those used in differentiating apply in integrating; thus the integral of a sum of a set of functions becomes the sum of the integrals of each individual function. If an integrand has a constant then this is taken out before integration is performed, thus:

$$\int 3x^2 \, dx = 3 \int x^2 \, dx$$

Usually, products or quotients must be simplified into simple functions before integration can take place. Thus, for example:

$$\begin{aligned} & \int (x + 2)(x - 3) \, dx \\ &= \int (x^2 - x - 6) \, dx \\ &= \frac{x^3}{3} - \frac{x^2}{2} - 6x + c \end{aligned}$$

and

$$\begin{aligned} & \int \frac{x(x-1)}{x^{\frac{1}{2}}} \, dx \\ &= \int \frac{x^2 - x}{x^{\frac{1}{2}}} \, dx \\ &= \int \left[x^{\frac{3}{2}} - x^{\frac{1}{2}} \right] \, dx \\ &= \frac{2x^{\frac{5}{2}}}{5} - \frac{2x^{\frac{3}{2}}}{3} + c \end{aligned}$$

7. In paragraph 3 it was shown that the integral of a simple function in x , ax^n is given by:

$$\frac{ax^{n+1}}{n+1}$$

However, it was stated that this formula did not apply when $n = -1$. This is because the denominator of the expression would become $-1 + 1 = 0$, and dividing by zero has no real meaning. The paradox can be resolved by recalling that differentiating $\log_e x$ yields $\frac{1}{x}$, therefore the converse means:

$$\int \frac{1}{x} dx = \log_e x$$

8. A list of the more common standard integrals is shown in Table 1.

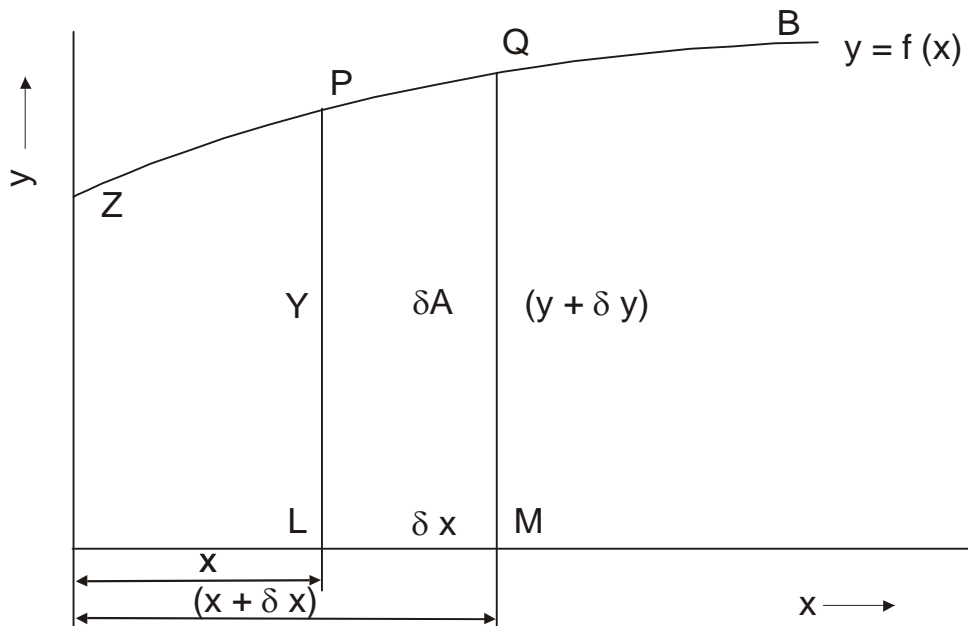
Table 1 Some Standard Integrals

$\frac{dy}{dx}$	$y = \int \frac{dy}{dx} dx$	Comments
x^n	$\frac{x^{n+1}}{n+1}$	Increase the index by 1, and divide by the new index.
$\cos x$	$\sin x$	Inverse of differentiation.
$\sin x$	$-\cos x$	Inverse of differentiation.
\sec^{2x}	$\tan x$	
e^{kx}	$\frac{e^{kx}}{k}$	Write down the function e^{kx} and divide it by the differential coefficient of the index of e .
$\frac{1}{x}$	$\log_e x$	

Definite Integrals

9. Fig 2 shows an arc, ZB, of the curve $y = f(x)$. P is the point (x, y) and Q the point $[(x + \delta x), (y + \delta y)]$, where δx and δy are very small quantities.

13-15 Fig 2 Area under the Curve of $f(x)$



10. The elemental strip LPQM is part of the area (A) between the curve and the axes of x and y. Let the area, LPQM, be denoted as δA . The mean height of the arc PQ lies between y and $y + \delta y$. Suppose it equals $y + K\delta y$, where $K < 1$, then the area $LPQM = \delta A = \delta x(y + K\delta y)$

$$\therefore (\delta A/\delta x) = y + K\delta y$$

In the limit as $\delta x \rightarrow 0$ then $(\delta A/\delta x) \rightarrow (dA/dx)$ and $\delta y \rightarrow 0$

$$\therefore \frac{dA}{dx} = y, \text{ and } \int \frac{dA}{dx} dx = \int y dx$$

i.e. $A = \int y dx = \int f(x) dx$

Suppose, $\int f(x) dx = F(x) + c$, then, $A = F(x) + C$

11. If it is required to find the area between the curve, the x axis, and the ordinates at $x = a$ and $x = b$, i.e. the area DCBA in Fig 3 then:

For the ordinate at $x = b$,

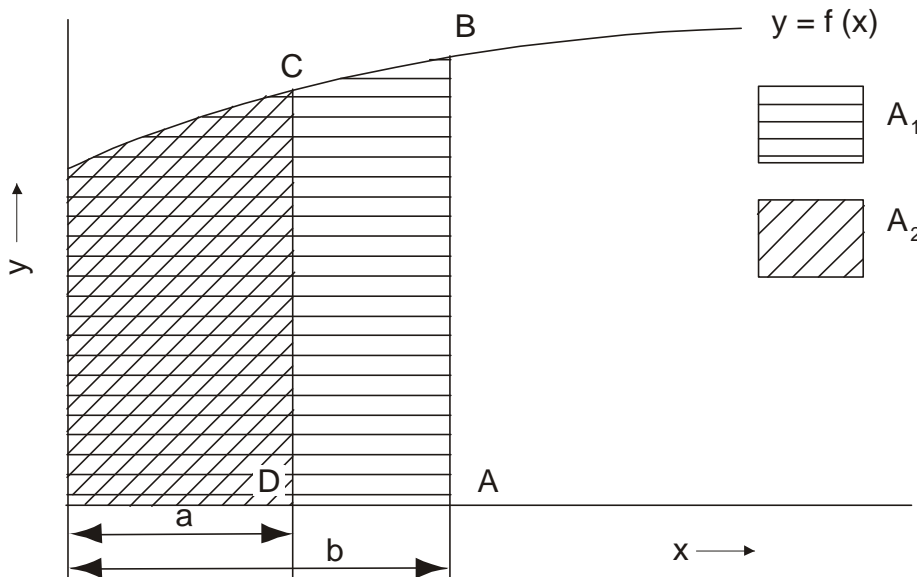
$$A_1 = F(b) + C$$

and at $x = a$,

$$A_2 = F(a) + C$$

Subtracting these, $DCBA = A_1 - A_2 = F(b) - F(a)$

13-15 Fig 3 Area under the Curve - The Definite Integral



This is written as:

$$\int_a^b f(x)dx = F(b) - F(a)$$

or, in words, "the integral $f(x)dx$ between the limits $x = a$ and $x = b$ ". 'a' and 'b' are called, respectively, the lower and upper limits of the value of x. Notice that the constant of integration has disappeared; this is because it would appear in both $F(b)$ and in $F(a)$ and is thus cancelled in the subtraction. Because such integrals are evaluated between defined limits, they are called definite integrals.

12. In summary, the method is as follows:

- a. Integrate the function, omitting the constant of integration.
- b. Substitute the value of the upper limit for x ; repeat for the value of the lower limit. Subtract the results to give $F(b) - F(a)$.

Example:

$$\int_1^2 x^3 dx = \left[\frac{x^4}{4} \right]_1^2 = \left(\frac{2^4}{4} \right) - \left[\frac{1^4}{4} \right] = 3.75$$

APPROXIMATE NUMERICAL INTEGRATION

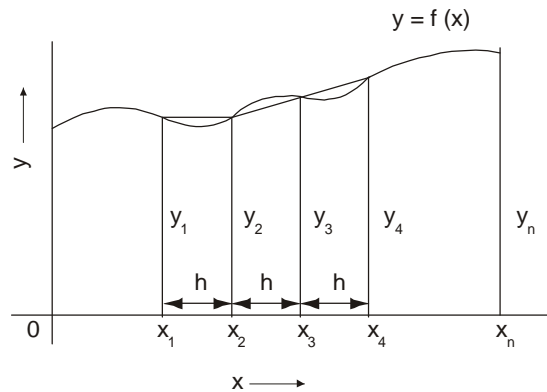
Introduction

13. The integration process is often complex, but, provided that the requirement is for a numerical answer to definite integration, then there are a number of methods available which yield approximate results, most of which are suitable for computer implementation if necessary. Two such methods, the trapezoidal rule and Simpson's rule, will be described.

Trapezoidal Rule

14. In the trapezoidal rule, the x axis is divided into equal intervals, h , and the top of each arc section is approximated by the chord, as in Fig 4. Thus, a series of trapezia are formed whose top coordinates have the values y_1, y_2 , etc.

13-15 Fig 4 The Trapezoidal Rule



15. The area of the first trapezium is: $\frac{1}{2}h(y_1 + y_2)$, and so:

$$\begin{aligned} \int_{x_1}^{x_n} y dx &= \frac{1}{2}h(y_1 + y_2) + \frac{1}{2}h(y_2 + y_3) + \dots + \frac{1}{2}h(y_{n-1} + y_n) \\ &= h(\frac{1}{2}y_1 + y_2 + y_3 + \dots + y_{n-1} + \frac{1}{2}y_n) \end{aligned}$$

In general, a small value of h will give a better solution than a large one, but the best procedure is to repeat the computation with successively smaller values of h until two results agree within the required level of precision.

16. **Example.** Compute $\int_{0.5}^1 x^{\frac{1}{2}} dx$ using the trapezoidal rule.

First compute with $h = 0.1$, say:

$$X_1 = 0.5; \quad \frac{1}{2}y_1 = 0.3535$$

$$X_2 = 0.6; \quad y_2 = 0.7746$$

$$X_3 = 0.7; \quad y_3 = 0.8367$$

$$X_4 = 0.8; \quad y_4 = 0.8944$$

$$X_5 = 0.9; \quad y_5 = 0.9487$$

$$X_6 = 1.0; \quad \frac{1}{2}y_6 = \underline{0.5000}$$

$$\text{Sum} = 4.3079$$

$$\int_{0.5}^1 x^{\frac{1}{2}} dx = 0.1 \times 4.3079 = 0.43079$$

Repeat with $h = 0.05$

$$X_1 = 0.5; \quad \frac{1}{2}y_1 = 0.3535$$

$$X_2 = 0.55; \quad y_2 = 0.7416$$

$$X_3 = 0.6; \quad y_3 = 0.7746$$

$$X_4 = 0.65; \quad y_4 = 0.8062$$

$$X_5 = 0.7; \quad y_5 = 0.8367$$

$$X_6 = 0.75; \quad y_6 = 0.8660$$

$$X_7 = 0.80; \quad y_7 = 0.8944$$

$$X_8 = 0.85; \quad y_8 = 0.9220$$

$$X_9 = 0.9; \quad y_9 = 0.9487$$

$$X_{10} = 0.95; \quad y_{10} = 0.9747$$

$$X_{11} = 1.0; \quad \frac{1}{2}y_{11} = \underline{0.5000}$$

$$\text{Sum} = 8.6184$$

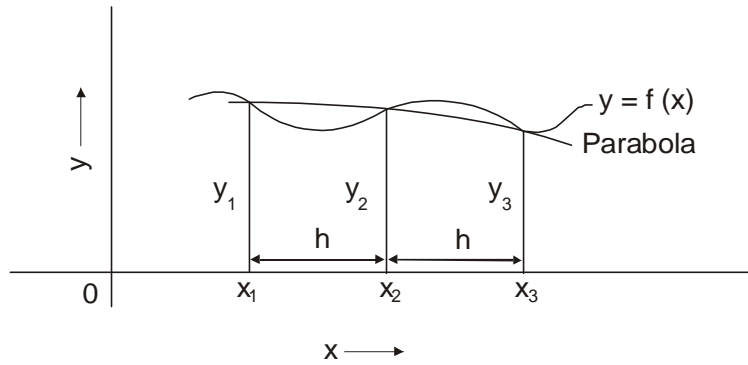
$$\int_{0.5}^1 x^{\frac{1}{2}} dx = 0.05 \times 8.6184 = 0.43092$$

To three decimal places, the result is 0.431 which compares very well with the correct value of 0.43096.

Simpson's Rule

17. In the trapezoidal rule the curve $y = f(x)$ is approximated by a series of straight lines. It can, however, be approximated by any suitable curve and in the case of Simpson's rule, a parabola is used. Rather than joining pairs of points, a parabola is traced through three points on the line as shown in Fig 5.

13-15 Fig 5 Simpson's Rule - Fitting a Parabola through Three Points



18. The result for an integration interval divided into 2 parts with 3 ordinates is:

$$\int_{x_1}^{x_3} y \, dx = \frac{1}{3}h(y_1 + 4y_2 + y_3)$$

19. **Example.** Compute $\int_{0.5}^1 x^{\frac{1}{2}} \, dx$ using Simpson's rule with $h = 0.25$.

$$x_1 = 0.5; \quad y_1 = 0.7071$$

$$x_2 = 0.75; \quad y_2 = 0.8660$$

$$x_3 = 1.00; \quad y_3 = 1.0000$$

and the integral is given by:

$$\frac{1}{3} \times 0.25(0.7071 + 4 \times 0.8660 + 1.0000) = 0.4309$$

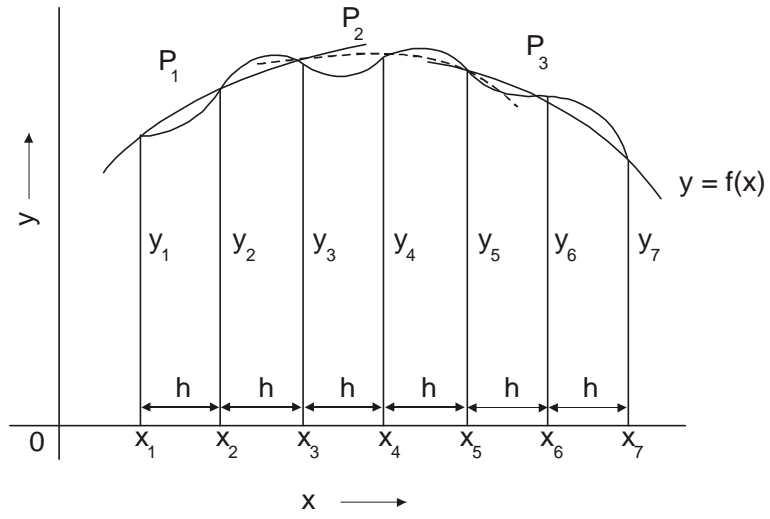
which is a better result than that given by the trapezoidal rule with $h = 0.10$.

20. Simpson's rule will usually give a more accurate result than the trapezoidal rule for the same interval, h , but it is often necessary to sub-divide the curve into more than one set of three ordinates. A different parabola is fitted over each section. For example, in Fig 6, seven ordinates are used, the parabolas are P_1 , P_2 , P_3 and the integral is given by:

$$\frac{1}{3}h(y_1 + 4y_2 + y_3) + \frac{1}{3}h(y_3 + 4y_4 + y_5) + \frac{1}{3}h(y_5 + 4y_6 + y_7) = \frac{1}{3}h(y_1 + 4y_2 + 2y_3 + 4y_4 + 2y_5 + 4y_6 + y_7)$$

The principle is capable of extension to any ODD number of ordinates.

13-15 Fig 6 Simpson's Rule - Seven Ordinates



CHAPTER 16 - THE SCOPE OF STATISTICAL METHOD

Introduction

1. The word statistics is used in two distinct ways. It is used to mean either sets of figures, usually tabulated, in which sense it is short for statistical data, or to mean the methods whereby the significant details may be extracted from such sets of figures. In this sense, it is short for statistical method, and it is with this meaning that this section is concerned.
2. A good definition of the subject is: a body of methods for making wise decisions in the face of uncertainty. Defined in this way the subject may be regarded as an extension of the idea of common sense, which is the name each person gives to his own method of making wise decisions in everyday matters. The fact that the answers provided by common sense on the one hand and by statistical method on the other often seem to be poles apart, is attributable either to the inadequacies of common sense or to an incorrect use of statistical method.
3. As a broad generalization, it may be stated that statistics takes over from common sense where the complexity of the problem warrants it, and where the quantities involved can be expressed in the form of numbers. If these conditions are fulfilled then the use of statistical method will give an economy of effort and a precision which is unobtainable in any other way.
4. There are very few aspects of human activity in which uncertainty does not play a part, so that the potential uses of statistical method are very large in number. In general, practical problems have solutions which are more or less probable, and probability theory forms the basis of statistics. An understanding of at least the elementary ideas of probability is, therefore, a prerequisite for the understanding of statistical method.
5. It is important to notice that the word, "wise" appears in the definition of statistics, and not the word "right". That the latter word is inadmissible follows, of course, from the fact that we are accepting uncertainty as a basic ingredient of the problem. The wise decision that is made will be based on the most probable occurrence, but the most probable occurrence is not bound to occur. Our decisions, therefore, will sometimes turn out to be wrong, no matter how elegant the mathematics used in the solution of the problem, and this fact must be accepted.
6. Quite frequently, decisions will prove to be wrong because they were based on inadequate data, and the point must be made that statistical analysis does not bring anything out of the data that is not already there. Statistics provides an objective way of testing the data and of obtaining answers free from personal prejudice and preconceived notions. The use of statistical method, in other words, makes it possible to interpret correctly the influence of chance in the evidence available.
7. It is clearly important that any experiment or series of trials should be designed to provide relevant evidence in sufficient quantity to do what is required. The only safe way to ensure this is to bring the statistician into the enquiry from the very beginning, so that he will not only analyse the data but in fact will also state what data should be collected. Far too often the statistician is called in too late, so that the investigator is faced with the invidious choice of either giving incomplete answers or of repeating all or part of the investigation in order to obtain adequate data. This procedure is likely to be far more costly in the long run than a properly planned attack on the problem in the first place. The best results are invariably obtained by the statistician and subject specialist working together from the beginning of the enquiry.

The General Method of Statistics

8. Although there are many distinct statistical techniques, they all have certain features in common. The following remarks are of general application.

9. **The Presentation of Data.** The raw data of statistics consists of a more or less haphazard collection of numerical data. An important first step before any statistical analysis can be started is to tabulate the data in some way which is meaningful for the investigation in hand. Suitable tabulation will, in fact, probably suggest the profitable line of attack, a process which may be aided by a graphical presentation of the data.

10. **Population.** Any collection of data which is to be analysed by a statistical process represents a finite selection of values from a practically if not theoretically, infinite population. In this context, the word population is part of statistical jargon, and refers to the totality of values which it is possible to conceive within the restrictions laid down for the data. It is important to note that the population may be one of people, of measurements, of bombs dropped on a target under specified conditions, or of any data whatsoever that may be given numerical values.

11. **Sample.** The concept of population is always to some extent an abstraction, since it is not possible to obtain access to all members of it (if it were, statistical analysis would be unnecessary). The data that is available is a sample taken from the population being studied. A problem will always be posed in terms of populations and not of samples. The solution to the problem, however, must proceed through an analysis of samples. For this to be a valid procedure it is clearly necessary that the sample should be a fair representation of the population that is being studied. It may then be assumed that parameters calculated from the sample are reliable estimates of the corresponding parameters of the population, and that conclusions based on the sample will be valid for the population also.

12. **Sampling.** It will be clear from para 11 that the technique of sampling, whereby the sample to be used in the investigation is to be chosen, plays a vital part in statistical work. Every collection of data is a fair sample from some population, but the important thing is to ensure that the collection selected is a fair sample of a specified population. This is by no means an easy thing to ensure. Broadly, two different sampling techniques are used, the choice in a particular case being dependent upon the extent of one's knowledge of the system studied. If a great deal is known about the population it may be possible to select a sample which conforms to the same pattern as the population; this technique is used extensively in public opinion polls. If it is not possible to do this, or if one is uncertain about the completeness of one's knowledge, then a technique of random selection, in which every member of the population has an equal chance of being selected for the sample, will be used. This has the extreme merit, if done correctly, of removing unsuspected bias, which may arise in any subjective method of sampling. A common method of ensuring randomness in the sample is through the use of special tables of random numbers, which have been thoroughly tested and found free from bias.

13. **Statistical Significance.** It is not unusual in statistical work to find that when a random sample is taken with the aim of providing a hypothesis it turns out that the sample data does not wholly support that hypothesis. The difference could be due to:

- a. The hypothesis being wrong, or
- b. The sample being biased.

Clearly, tests are needed to determine which is the more likely possibility. Tests of significance are very important to statisticians but are outside the scope of the chapter.

14. **Proof and Disproof.** It is perhaps clear already from the foregoing discussion that statistical method never provides a definite proof of any hypothesis, though it may provide very strong evidence indeed in favour of it. Any process which involves extrapolation from sample to population, and usually from past to future time, must involve uncertainty, and no matter how improbable an event there is always the possibility that it will happen. The gibe that "you can prove anything by statistics" shows a complete misunderstanding of the methods of statistical inference. Nothing is "proved" or "disproved" by statistics.

Some Uses of Statistics

15. The point has been made that the opportunities for the application of statistical method are virtually unlimited, so that any list of uses will necessarily be incomplete. The following selection of topics, however, gives a good idea of the great scope of statistical method.

16. **The Measurement of the Inexplicable.** Many problems are concerned with quantities which do not take unique values under the conditions which it is possible to specify. In such cases, residual uncertainties may be very important, and it is necessary to be able to assess their likely magnitude. Probability theory provides a way of doing this. We may distinguish two important types of problem:

- a. Those in which the uncertainties give rise to a random departure from a true value or from a desired result. We are here concerned with the determination of errors.
- b. Other cases, in which there is no true or desired value, but in which random variability is an essential feature of the system. Measurements of biological quantities and of human attainment come into this category.

The emphasis here is on the use of statistics to calculate meaningful parameters which may be used to describe the population. Some distribution of values is obtained, and, in practice, the essential features of this distribution are established by comparison with a similar theoretical distribution.

17. **The Identification of Important Factors.** The problem here is to determine which of several factors that might be expected to affect performance do in fact have a significant effect. In some cases, the important factors may be seen without the aid of statistical analysis. In other cases, because of interactions between factors, the issue may be by no means clear, and statistical techniques then provide a way of separating the effects of individual factors and thus establishing their relative importance.

Some Misuses of Statistics

18. No discussion of the uses of statistics would be complete unless accompanied by a discussion of misuses, for abuse of statistical method occurs all too frequently. It is because of this that statistics are viewed with such suspicion by so many people.

19. In general we may lay down the principle that from a given set of data one set of conclusions relative to a particular issue will be more likely than any other set. But the fact is that other conclusions, incompatible with the first, will frequently be drawn, often because a certain conclusion is desired and sometimes, as in misleading advertising, because no other conclusion is acceptable. It is often difficult or impossible for a person not intimately engaged in an investigation to trace invalid reasoning, and hence the belief arises that a judicious use of statistics will enable conflicting conclusions to be drawn from the same set of data. Again, the possibilities are legion, but the following examples illustrate some of them.

20. **Deceptive Presentation.** Cases of presentation with intent to deceive are common features of everyday life, and very often take the form of the omission of relevant data. Thus, a poster supporting an anti-immunization campaign announced boldly that in a certain period of time 5,000 cases of diphtheria occurred among immunized children. The public was expected to infer that immunization failed, but the poster did not disclose the highly relevant information that in the same period 75,000 cases occurred among non-immunized children, nor that non-immunized children were 6 times as likely to get diphtheria and 30 times as likely to die from it.

21. **Sampling Errors.** The importance of sampling has already been emphasized, and it must be obvious that bad sampling will lead to invalid conclusions. For example, taking the announcement of births from the columns of the "Times" gave a sex ratio of 1,089 males per 1,000 females. However, the Registrar General found during the same period a ratio of 1,050 females per 1,000 males. The reason for the discrepancy is no doubt that the sample from the "Times" was not a fair sample of the whole population, perhaps because parents are more inclined to announce the births of their sons and heirs than of their daughters.

22. **False Correlations.** The technique of correlation is very easily misapplied, and correlation between two variables should not be sought unless there are reasons, stemming from knowledge of the system studied, to expect it. It is easy to think of variables which, though unrelated, will show strong correlation, often because both happen to vary in a certain way with the passage of time. Such correlations are called nonsense correlations.

23. **Statistics versus Experience.** It is the subject expert who will make the decision, making full use both of his own specialist knowledge and of the results of the statistical analysis. If a statistical inference runs counter to what the specialist expects, then he should query it. The inference may be correct, but a little probing will be most worthwhile.

CHAPTER 17 - DESCRIPTIVE STATISTICS

Introduction

1. Statistics is concerned with the mathematical analysis of numerical data. The numerical data is in the form of a set of observations of the variable (or variables) under consideration. A variable (or variate) is a quantity which assumes different measurable values, e.g. height, weight, examination marks, temperature, intelligence, length of life, etc. Any set of observations of the variable is considered, for statistical purpose, as a sample drawn from some infinitely large population. When each and every member of a population has an equal chance of being selected for a sample, the sample is called a random sample. The principle task of statistical analysis is to deduce the properties of the population from those of a random sample. In this chapter, we discuss how samples and populations can be described; in particular, we will look at averages and the bunching of the samples and population about these averages.

AVERAGES

Types of Average

2. **The Arithmetic Mean.** The arithmetic mean is commonly referred to as 'the average'. The mean is the sum of all the values of a variable divided by the number of variables. The algebraic form of expression is:

$$\bar{X} = \frac{X_1 + X_2 + X_N}{N} = \frac{\sum X}{N}$$

where: \bar{X} = the mean

X_1, \dots, X_N = the values of the different variables in a distribution

N = the number of variables

Σ = the symbol instructing the addition of all of the values

Example:

A sample consists of the data 5, 8, 9, 6, 12, 14.

$$\text{The mean is } \bar{X} = \frac{5+8+9+6+12+14}{6} = \frac{54}{6} = 9$$

3. **The Median.** If the sample observations are arranged in order from the smallest to the largest, the median is the middle observation. If there are two middle observations, as in the case of an even number of observations, the median is halfway between them.

Examples:

a. Given sample 1, 14, 9, 6, 12. Arranged in order 1, 6, 9, 12, 14, the median is 9.

b. Given sample 20, 7, 11, 10, 13, 17. Arranged in order 7, 10, 11, 13, 17, 20, the median is 12.

4. **The Mode.** The mode is the observation which occurs most frequently in a distribution. If each observation occurs the same number of times, there is no mode. If two or more observations occur the same number of times, and more frequently than any other observations, then the sample is said to be multi-modal.

Examples:

- a. Given sample 16, 13, 18, 16, 17, 16, the mode is 16.
- b. Given sample 4, 7, 4, 9, 3, 7, then the modes are 4 and 7.
- c. Given sample 3, 7, 12, 11, 16, 20 there is no mode.

The mode is seldom used but has been included for completeness.

MEASURES OF DISPERSION

Introduction

5. Knowledge of the average of a distribution provides no information about whether figures in a distribution are clustered together or well spread out. For example, two groups of students have examination marks of 64%, 66%, 70%, and 80%, for the first group and 37%, 61%, 88% and 94% for the second group. Both groups have a mean mark of 70% but the marks of the second group have a much greater dispersion than those of the first group. Clearly, it would be useful to have a way of measuring dispersion (or variance) and expressing it as a simple figure. The most commonly used measures are:

- a. Range.
- b. Quartile Deviation.
- c. Standard Deviation.
- d. CEP (used in particular applications).

Range

6. Range is the difference between the highest and lowest values. Unfortunately, range is too much influenced by extreme values so that one value differing widely from the remainder in a group could give a distorted picture of the distribution. Range also fails to indicate the clustering of values into particular groups or areas.

Quartile Deviation

7. Quartiles are the values of the items one quarter and three quarters of the way through a distribution. If the top and bottom quarters are cut off, extreme values are discarded and a major disadvantage of range as a measure of dispersion is avoided.

$$\text{Quartile Deviation} = \frac{\text{Third Quartile} - \text{First Quartile}}{2}$$

As with Range, this method fails to indicate clustering.

Standard Deviation

8. Standard deviation is the most important of the measures of dispersion. The standard deviation (σ) is found by adding the square of the deviations of the individual values from the mean of the distribution, dividing the sum by the number of items in the distribution, and then finding the square root of the quotient. (Scientific calculators include a facility for finding σ and other statistical parameters from sets of data.)

$$\sigma = \sqrt{\frac{\sum(x - \bar{x})^2}{N}}$$

The more that values of individual items differ from the mean, the greater will be the square of these differences, giving rise to a large measure of dispersion. The main disadvantage of using standard deviation as a measure of dispersion therefore is that it can give disproportionate weight to extreme values because it squares the deviations eg a value twice as far from the mean as another is weighted by a factor of 4, (2^2). Nevertheless, standard deviation is the best and most useful measure of dispersion within a set of observations.

Circular Error Probable (CEP)

9. A term commonly encountered in weapon effects planning is Circular Error Probable (CEP). The CEP is the radius of the circle, centered on the mean impact point, within which 50% of weapons fall. Strictly, CEP should only be used when the distribution of impacts is known to be circular, but this restriction is often ignored in practice. As a rough guide, the CEP is approximately 1.18 times the standard deviation of the linear errors in weapon impacts, although this convention is only justifiable if the errors in range and deflection are normally distributed (see paras 25-27).

FREQUENCY DISTRIBUTIONS

Introduction

10. It is very difficult to learn anything by examining unordered and unclassified data. Table 1 displays such raw data.

Table 1 Weekly Mileages Recorded by 60 Salesmen

504	592	671	498	601	532
623	548	467	487	399	482
507	477	501	562	555	642
477	522	627	556	622	521
429	491	497	510	603	547
535	517	612	491	432	508
577	444	556	639	444	723
562	685	432	642	562	662
688	492	486	467	474	433
417	512	563	612	375	578

Raw data is simply a list of data as received, in this case from sixty individual salesmen. Little of use can be learned from data presented in this form.

Ungrouped Frequency Distribution

11. The first step in making the raw data more meaningful is to list the figures in order from the lowest mileage to the highest. At the same time, it may be convenient to annotate those figures that occur more than once with the frequency of occurrence. The result is the distribution shown in Table 2. The symbol for frequency is ' f '. The sum of the frequencies ($\sum f$) must equal the total number of items making up the raw data.

Table 2 Ungrouped Frequency Distribution

Mileage	f	Mileage	f	Mileage	f	Mileage	f	Mileage	f
375	1	482	1	508	1	555	1	622	1
399	1	486	1	510	1	556	2	623	1
417	1	487	1	512	1	562	3	627	1
429	1	491	1	517	1	563	1	639	1
432	2	492	1	521	1	577	1	642	2
433	1	497	1	522	1	578	1	662	1
444	2	498	1	532	1	592	1	671	1
467	2	501	1	535	2	601	1	685	1
474	1	504	1	547	1	603	1	688	1
477	2	507	1	548	1	612	2	723	1

Grouped Frequency Distribution

12. Though the ungrouped frequency distribution is an improvement in the presentation, there are still too many figures for the mind to grasp the information effectively. More simplification is necessary in order to compress the data. This can be done by grouping the figures and showing the frequency of the group occurrence. The result is shown in Table 3.

Table 3 Grouped Frequency Distribution

Mileage	f
375 – under 425	3
425 - under 475	9
475 - under 525	18
525 - under 575	12
575 - under 625	9
625 - under 675	6
675 - under 725	3
Total Records	60

13. **Effect of Grouping.** As a result of grouping, it is possible to see from Table 3 that mileages cluster around 475-525. Although grouping highlights the pattern of a distribution it does lead to the loss of information about where in the group the 18 occurrences lie. The increased significance of the table has therefore been paid for but the cost is worthwhile. The loss of information also means that calculations made from grouped frequency distribution cannot be exact.

Class Limits

14. The boundaries of a class are called the class limits. Care must be taken in deciding class limits to ensure that there is no overlapping of classes or gaps between them. For example if the class limits in table 3 had been 375-425 and 425-450 which group would a mileage of 425 have gone into? Likewise, if the class limits had been 375-424 and 425-449 a mileage of $424\frac{1}{2}$ would have no group to fit into.

15. **Discrete and Continuous Data.** In defining class limits, it should be remembered that discrete data increases in jumps. For instance, data relating to the number of children in families will be in whole units because $1\frac{1}{2}$ and $2\frac{1}{4}$ children are not possible. Continuous data, however, may include fractions.

16. **Class Intervals.** Class intervals define the width of a class. If the class intervals are equal, the distribution is said to be an equal class interval distribution.

17. **Unequal Class Intervals.** Some data is such that if equal class intervals were used a very few classes would contain all the occurrences whilst the majority would be empty. An example of this is the distribution of salaries using a class interval of £1000. In this situation, the class intervals should be arranged so that over-full classes are subdivided and near empty ones grouped together.

18. **Choice of Classes.** The construction of a grouped frequency distribution always involves a decision as to what classes to use. The following suggestions should be borne in mind:

- a. Class intervals should be equal wherever possible.
- b. Class intervals of 5, 10, or multiples of 10 are more convenient than, say, 7 or 11.
- c. Classes should be chosen so that occurrences within the classes tend to balance around the mid point.

Construction of a Grouped Frequency Distribution

19. To construct a grouped frequency distribution directly from raw data the following steps should be taken:

- a. Pick out the highest and lowest figures (375 and 723) and on the basis of these decide upon the list and the classes.
- b. Take each figure in the raw data and insert a check mark (1) against the class into which it falls.
- c. Total the check marks to find the frequency of each class (see Table 4).

Table 4 Direct Construction of Grouped Frequency Distribution

Class	Check Marks	f
375 - under 425	111	3
425 - under 475	1111 1111	9
475 - under 525	1111 1111 1111 111	18
525 - under 575	1111 1111 11	12
575 - under 625	1111 1111	9
625 - under 675	1111 1	6
675 - under 725	111	3
	Total Records	60

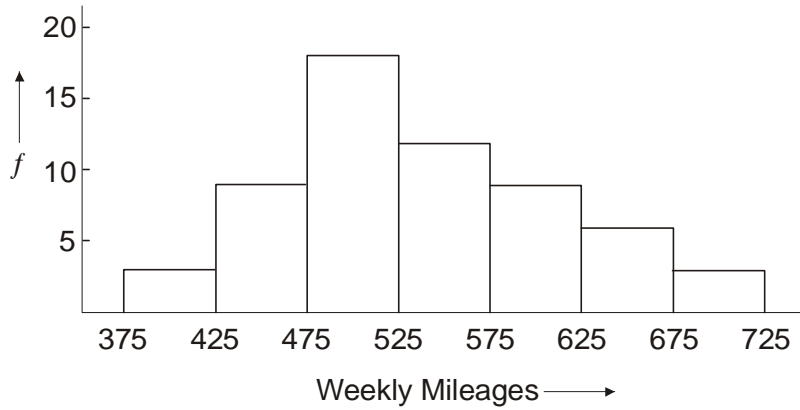
GRAPHS OF OBSERVATIONS

The Histogram

20. A Histogram is a graph of a frequency distribution. It is shown in Fig 1 and is constructed as follows:

- a. The horizontal axis is a continuous scale running from one end of the distribution to the other. The axis should be labelled with the name of the variable and the unit of the measurement.
- b. For each class in the distribution a vertical column is constructed with its base extending from one class limit to the other and its area proportional to the frequency of the class.

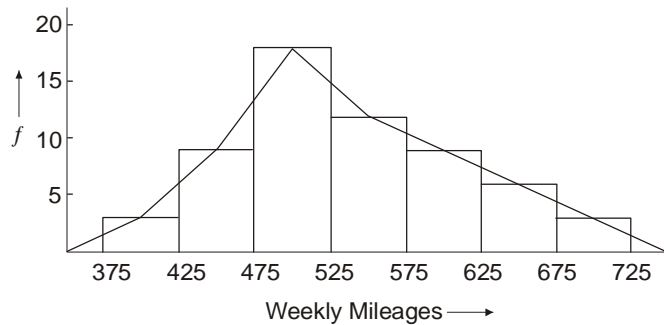
13-17 Fig 1 Histogram of Data from Table 3



The Frequency Polygon

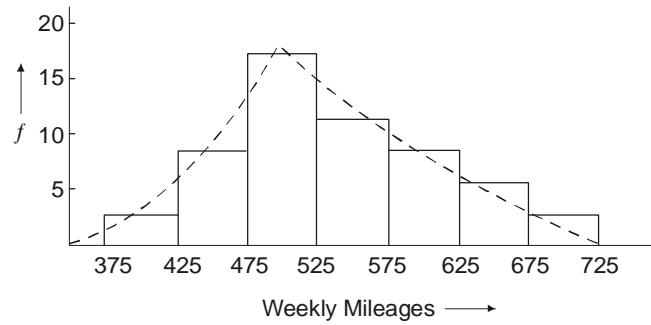
21. If the mid points of the tops of the blocks of a frequency histogram are joined by straight lines, the resulting figure is a frequency polygon as illustrated in Fig 2. The area enclosed by the horizontal axis, the polygon, and any two ordinates represents approximately the number of observations in the corresponding range. The total area enclosed by the polygon represents the total number of observations if the frequency scale is used, and unity if the probability scale is used.

13-17 Fig 2 Frequency Polygon



The Frequency Curve

22. If the number of observations is greatly increased and the size of the class interval is correspondingly reduced, then the frequency histogram and the frequency polygon will tend to a smooth curve as in Fig 3. By increasing the number of observations indefinitely, the whole population instead of just a sample will have been analysed. Thus, the smooth curve obtained by this process represents the distribution of the complete population in the same way as the histogram or frequency polygon represents the distribution of the sample.

13-17 Fig 3 Frequency Curve

23. If the original sample is random, and large enough, then it is unlikely that the smooth curve will be very different from drawing a smooth curve through the mid points of the histogram blocks. Such a smooth curve is known as a frequency curve if the frequency scale is used or a probability curve if the probability scale is used. The area under the curve between any two ordinates represents, as accurately as any estimate can, the number of observations within the corresponding range, while the same area for a probability curve represents the probability that a single observation, taken at random from the complete population, will lie within the corresponding range. The latter idea is more useful since it applies to the whole population and is no longer confined to the sample.

24. Histograms frequently display a pattern in which there is a high column in the centre with decreasing columns spread symmetrically either side. If the class interval is small enough, the frequency curve looks like the cross section of a bell. This pattern occurs frequently in statistical work.

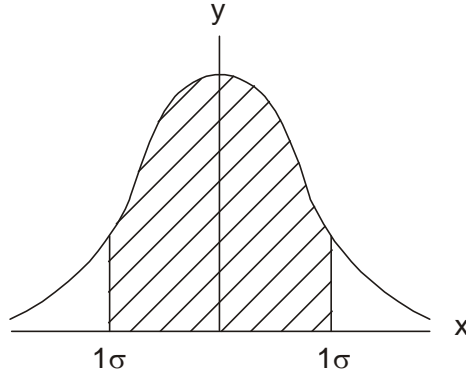
The Normal Curve of Distribution

25. The task of handling statistical data can be simplified if a mathematical curve can be found approximating to that which would be produced by plotting the actual data on a graph. By substituting the mathematical graph for the real one, it is possible to make calculations to reveal facts about the distribution of the raw data which would otherwise have been difficult to determine. The normal curve of distribution satisfies this requirement. It has the following features:

- a. It is symmetrical
- b. It is bell shaped
- c. Its mean lies at the peak of the curve
- d. The two tails continuously approach, but never cross, the horizontal axis (x) (see Fig 4).

The formula for the curve can be ignored because any mathematical data relating to it can be found in mathematical tables.

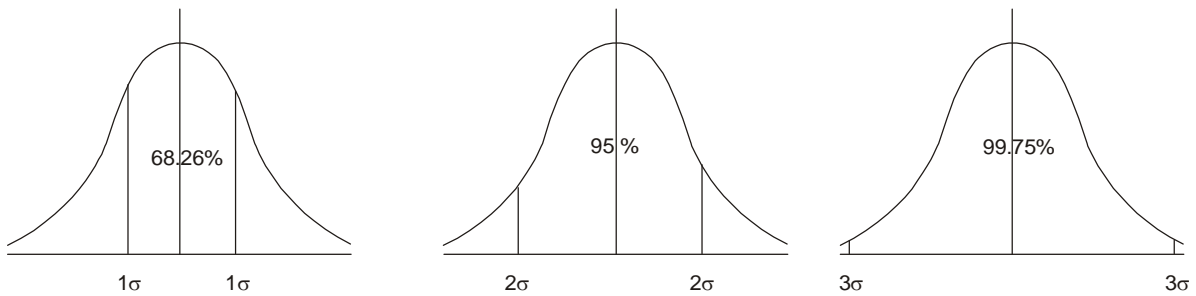
13-17 Fig 4 Normal Curve of Distribution



26. **Areas Below the Normal Curve of Distribution.** The 'y' axis in Fig 4 is at the peak of the curve and passes through the mean value of the distribution. The total area beneath the curve is unity, representing the fact that a random variable is certain to lie between $+\infty$ and $-\infty$. If 1σ lengths are now marked off on the x-axis from the mean value of the curve, the area enclosed by the curve and the 1σ boundaries is 68.26% of the total area. Tables are available to give the areas lying under the curve between any two lines on the x-axis designated in terms of σ .

27. **Areas and Frequencies.** Areas under the normal distribution curve are proportional to frequencies. An area of 95% of the total area is equivalent to a frequency figure which indicates that 95% of all occurrences lie between the two 2σ values. Similarly, 99.75% of all occurrences fall within the 3σ values (see Fig 5). The height of the curve at a particular point has no practical relevance - areas under the curve must always be used to give the frequency of occurrence in a specified range of values.

13-17 Fig 5 Approximate Areas Beneath the Normal Curve



CHAPTER 18 - ELEMENTARY THEORY OF PROBABILITY

Mathematical Definition of Probability

1. Suppose that a certain experiment can have just n possible results, all of them equally likely (e.g. in drawing a card from a pack there are just 52 possible results and all of them are equally likely) and suppose that a certain event can occur as m of these n possible results (e.g. picking an ace can occur as 4 of the 52 possible results). Then the probability of the event occurring in any one experiment is defined as m/n .

2. Thus the probability of drawing an ace is $4/52$ or $1/13$. We may also say that the chances are 1 in 13 or that the odds are 12 to 1 against.

3. If the probability that an event will occur is m/n , then the probability that it will fail to occur is $(n-m)/n = 1 - m/n$. Thus if p is the probability of success and q the probability of failure, we have

$$q = 1 - p \quad \text{or} \quad p = 1 - q \quad \text{or} \quad p + q = 1$$

4. It is certain that an event will either occur or fail to occur. The probability of either a success or a failure is $n/n = 1$. Hence probability = 1 denotes certainty and, similarly, probability = 0 denotes impossibility.

Interdependence of Events

5.

a. **Mutually Exclusive Events.** Two events are mutually exclusive if the occurrence of one prevents the occurrence of the other. If a die is thrown, the occurrence of a 6 prevents the occurrence of a 5.

b. **Independent Events.** Two events are independent if the occurrence or non-occurrence of one has no effect on the probability of the occurrence of the other. If two dice are thrown, the result of one throw has no effect on the result of the other.

c. **Dependent Events.** Two events are dependent when the occurrence or non-occurrence of one event has some effect on the probability of occurrence of the other. If two cards are drawn from a pack, the probability that the second card is an ace is $3/51$ or $4/51$ depending on whether the first card was or was not an ace.

Notice that mutually exclusive events occur as alternative results of the same experiment whereas independent and dependent events occur as the simultaneous or consecutive results of different experiments.

Calculation of Probabilities

6. **Theorem I - Addition of Probabilities.** If two events are mutually exclusive, then the probability of either one or the other event occurring is the sum of the probabilities of the individual events.

Proof. Let the probability of E_1 be m_1/n and let the probability of E_2 be m_2/n . Then out of n equally likely events, m_1 are E_1 and m_2 are E_2 , ie $m_1 + m_2$ events out of n are either E_1 or E_2 . Hence the probability of either E_1 or E_2 occurring is:

$$\frac{m_1 + m_2}{n} = \frac{m_1}{n} + \frac{m_2}{n}$$

7. Theorem II - Multiplication of Probabilities.

a. **Independent Events.** If two events are independent, then the probability of both one and the other happening is the product of the probabilities of the individual events.

b. **Dependent Events.** If two events are dependent, then the probability of both the first event and the second event happening is the product of the probability of the first event and the conditional probability of the second event on the assumption that the first event has happened.

Proof. Let the probability of E_1 be m_1/n_1 and let the probability of E_2 be m_2/n_2 . If E_1 and E_2 are independent, then the probability of E_2 is independent of the success or failure of E_1 , but if they are dependent then the probability m_2/n_2 is to be regarded as the conditional probability of E_2 on the assumption that E_1 has happened. Then the total number of possible results of the two experiments together is $n_1 n_2$. Of these possible results, E_1 and E_2 can occur together in $m_1 m_2$ ways. Hence probability of both E_1 and E_2 occurring is:

$$\frac{m_1 m_2}{n_1 n_2} = \frac{m_1}{n_1} \times \frac{m_2}{n_2}$$

8. Conclusion both Theorem I and Theorem II extend to any number of events.

Example 1. A card is drawn from a pack. What is the probability of it being either an ace, the king of clubs or a red queen?

The probability of drawing an ace is $\frac{4}{52}$

The probability of drawing the king of clubs is $\frac{1}{52}$

The probability of drawing a red queen is $\frac{2}{52}$

Hence the required probability is:

$$\frac{4}{52} + \frac{1}{52} + \frac{2}{52} = \frac{7}{52}$$

Example 2. Two dice are thrown and a card is drawn from a pack. What is the probability that both dice will show sixes and that the card will be the ace of spades?

The probability of a six in one throw is $\frac{1}{6}$ and the probability of drawing the ace of spades is $\frac{1}{52}$

Hence, the required probability is:

$$\frac{1}{6} \times \frac{1}{6} \times \frac{1}{52} = \frac{1}{1872}$$

Example 3. Two dice are thrown. What is the probability of the total throw being 10? The possible successes are (4, 6), (5, 5), (6,4). The result of one die is independent of the result of the other.

$$\text{Probability of (4, 6) is } \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

$$\text{Probability of (5, 5) is } \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

$$\text{Probability of (6, 4) is } \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

Any combination of the pair excludes every other combination. Hence the required probability is:

$$\frac{1}{36} + \frac{1}{36} + \frac{1}{36} = \frac{1}{12}$$

Example 4. Two cards are drawn from a pack. What is the probability that they will both be aces?

The probability that the first card is an ace is $\frac{4}{52}$ and the conditional probability that the second card shall also be an ace is $\frac{3}{51}$. Hence, the required probability is:

$$\frac{1}{13} \times \frac{1}{17} = \frac{1}{221}$$

Example 5. Five balls are drawn from a bag containing 6 white balls and 4 black balls. What is the probability that 3 white balls and 2 black balls are drawn?

Although the rules for combining probabilities are important, it sometimes pays to work from first principles, i.e. direct from the mathematical definition of probability as in this example.

5 balls can be selected from 10 in $\binom{10}{5}$ ways, that is in

$$\frac{10.9.8.7.6}{1.2.3.4.5} = \frac{2.9.7.2}{1} = 2.9.7.2 \text{ ways}$$

This is the total number of possible selections.

Three white balls can be selected from 6 in $\binom{6}{3}$ ways, i.e.

$$\frac{6.5.4}{1.2.3} = 5.4 \text{ ways}$$

2 black balls can be selected from 4 in $\binom{4}{2}$ ways, ie

$$\frac{4.3}{1.2} = 2.3 \text{ ways}$$

Hence, the number of ways in which 3 white balls and 2 black balls can be selected is 5.4.2.3, from which the required probability is:

$$\frac{5.4.2.3}{2.9.7.2} = \frac{10}{21}$$

Limitations of Mathematical Definition

9. It is not always possible to use the mathematical definition of probability because the definition requires that the experiment should have a finite number of possible different results. In practice, there are often an infinite number of possible results (e.g. in dropping a bomb on a target). Moreover, the definition depends on the results being equally likely, which to a certain extent begs the question unless we can satisfy ourselves intuitively that the results are equally likely. Thus, when the number of possible results is infinite or when we cannot be sure that they are all equally likely, we cannot use the mathematical definition, and instead we use the following one.

Statistical Definition of Probability

10. Suppose that a certain experiment is carried out n times and that a certain event occurs as m of the n results. Then the probability of the event occurring as the result of any one experiment is defined as:

$$p = \lim_{n \rightarrow \infty} \frac{m}{n}$$

In practice, we cannot let $n \rightarrow \infty$, but instead we take n to be as large as we conveniently can. The resulting value obtained for the probability is then the best available estimate in cases where the mathematical definition cannot be used.

11. It can be shown that there is little theoretical difference between the two definitions and that therefore, the theorems proved on the basis of the mathematical definition still hold for probabilities obtained statistically.

Example 6. The operational requirement for a guided weapon demands that the weapon should have a reliability of 90%. If the weapon can be broken down into 120 functionally-tested components of equal complexity and reliability, determine the reliability demanded from each component.

The reliability of a weapon or a component is the probability that the weapon or component will be completely serviceable. In this case, the statistical definition of probability clearly applies. Now if R_C is the reliability of a component, then the probability that 120 such components will all be serviceable together may be obtained using the Multiplication Theorem for independent events.

$$0.90 = R_C^{120}, \text{ or } R_C = \sqrt[120]{0.90} = 0.9991$$

Thus, each component must have a reliability of 99.91 %.

Probability of At Least One Success

12. It is important to distinguish between:

- a. The probability of at least one success.
- b. The probability of one success only.

Suppose it is required to get a minimum of one hit on a target. It is not sufficient to calculate the probability of one hit only because this would exclude the possibility of 2, 3, or more hits which must also count as successes. The criterion of success is at least one hit.

Example 7. Three ballistic missiles are launched at a certain target. From the statistical analysis of performance trials of the missile, it is estimated that the probability of hitting the target with a single missile is $\frac{1}{6}$. Calculate the chance of scoring: a. One hit only; b. At least one hit.

For a: The chance of a hit with the first missile but not with the other two is:

$$\frac{1}{6} \times \frac{5}{6} \times \frac{5}{6} = \frac{25}{216}$$

The chance of a hit with the second missile but not with the other two is:

$$\frac{5}{6} \times \frac{1}{6} \times \frac{5}{6} = \frac{25}{216}$$

The chance of a hit with the third missile but not with the other two is:

$$\frac{5}{6} \times \frac{5}{6} \times \frac{1}{6} = \frac{25}{216}$$

Hence the chance of one hit only is:

$$\frac{25}{216} + \frac{25}{216} + \frac{25}{216} = \frac{75}{216}$$

For b: In the same way, the chance of two hits only is:

$$\frac{1}{6} \times \frac{1}{6} \times \frac{5}{6} \times 3 = \frac{15}{216}$$

and the chance of three hits is:

$$\frac{1}{6} \times \frac{1}{6} \times \frac{1}{6} = \frac{1}{216}$$

Hence the chance of at least one hit is:

$$\frac{75}{216} + \frac{15}{216} + \frac{1}{216} = \frac{91}{216}$$

Alternatively, since the chance of a hit with one missile is $\frac{1}{6}$, the chance of missing with one missile is $1 - \frac{1}{6} = \frac{5}{6}$. Thus, the chance of missing with all three missiles is $\left(\frac{5}{6}\right)^3 = \frac{125}{216}$, and the chance of failing to miss with all three missiles, that is the chance of at least one hit, is:

$$1 - \frac{125}{216} = \frac{91}{216}$$

To generalize, let the chance of success in one attempt be p , then the chance of failure in one attempt is $(1 - p)$, ie the chance of failure in n attempts = $(1 - p)^n$, and finally, the chance of at least one success in n attempts is $1 - (1 - p)^n$.

Example 8. If the probability of obtaining a hit with a single missile is assessed as $\frac{1}{20}$, how many missiles must be launched to give a 75% chance of at least one hit?

If n is the number of missiles which must be launched we require that:

$$1 - \left(1 - \frac{1}{20}\right)^n = 0.75$$

$$\text{ie } \left(\frac{19}{20}\right)^n = 0.25$$

$$\text{or } 0.95^n = 0.25$$

Taking logarithms,

$$n \log 0.95 = \log 0.25$$

$$\therefore n = \frac{\log 0.25}{\log 0.95} = 27$$

Example 9. The data given below refers to an interceptor fighter armed with 2 air-to-air guided weapons. Determine the overall effectiveness of the weapon system.

Aircraft serviceability	0.90
Aircraft reliability in flight	0.80
Missile reliability	0.70
Missile lethality	0.50

The probability that the aircraft will be both serviceable and reliable in flight, and therefore able to deliver the weapon is:

$$0.90 \times 0.80 = 0.72$$

The probability that a single weapon will inflict the required damage on the target is:

$$0.70 \times 0.50 = 0.35$$

Thus, the probability that at least one weapon will inflict the required damage is:

$$1 - (1 - 0.35)^2 = 0.58$$

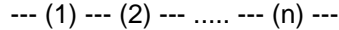
i.e. the overall effectiveness is: $0.72 \times 0.58 = 0.42$ or 42%

Reliability

13. The accurate assessment of the reliability of complex and expensive systems or equipment can be a vitally important factor in planning the purchase or deployment of resources. In order to quantify reliability for analysis, it is necessary to be able to attach a numerical value to it. Reliability is defined as the probability that an item will not fail during a given period of time. The probability of an item not failing is denoted by p . The probability of an item failing is denoted by q . It follows that $p+q = 1$. It is important that when a figure for reliability is quoted the time period to which it relates should also be quoted. It should be noted that reliability is a probability and is therefore expressed as a fraction of 1 or a percentage.

14. **Combination.** The overall reliability (R) of a combination of components may be calculated on the assumption that the quantities p and q are independent of the other components in the system. This being so, probabilities must be combined by the multiplication rule to give the probability that all will occur together.

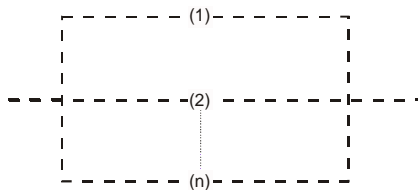
a. **Components in Series.**



The system will survive only if all the components survive.

$$\therefore R = p_1 \times p_2 \times \dots \times p_n$$

b. **Components in Parallel.**



The system will fail only if all the components fail.

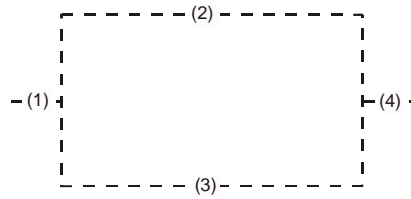
$$\therefore 1 - R = q_1 \times q_2 \times \dots \times q_n$$

$$\text{or } R = 1 - (q_1 \times q_2 \times \dots \times q_n)$$

$$\text{or } R = 1 - (1 - p_1)(1 - p_2) \dots (1 - p_n)$$

A parallel system may be referred to as a redundant system.

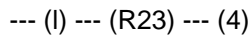
c. **Components may be combined partly in series and partly in parallel.**



These systems may be reduced to a system of units in series by obtaining the overall reliability of each parallel branch. Thus in the above example, if R₂₃ is the overall reliability of the parallel components 2 and 3

$$R_{23} = 1 - q_2 \times q_3$$

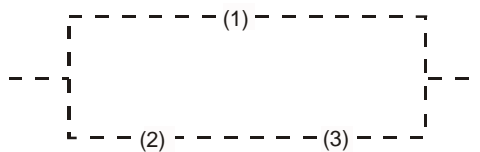
and the system reduces to:



whence, $R = p_1 \times R_{23} \times p_4$

$$= p_1 \times p_4 \times [1 - (1 - p_2)(1 - p_3)]$$

15. Where redundant components are used on a mutually exclusive basis either one is used or another, and in this case the addition rule for combining the probabilities must be used. For example, a navigation system consists of mode 1 (reliability p_1) and mode 2 (p_2) which will only be switched in by a switch, of reliability p_3 , if mode 1 fails. The system will survive only if mode 1 survives or if mode 1 fails and the switch operates and mode 2 survives. In terms of reliabilities, the switch and mode 2 are only called upon to survive if mode 1 fails.



$$\therefore R = p_1 + 1(1 - p_1)(p_3)(p_2)$$

The term $(1 - p_1)(p_3)(p_2)$ is the reliability contribution of the standby equipment. The overall reliability of the system can be worked out by the series/parallel calculation, giving:

$$R = 1 - (1 - p_1)(1 - p_3p_2) = p_1 + p_2p_3 - p_1p_2p_3.$$

CHAPTER 19 - THE NATURE OF HEAT

Temperature and Heat

- Heat is a form of energy possessed by a body by virtue of its molecular agitation. The heat content of an object is not measured simply by its temperature; heat content is also a function of mass. One unit of heat is the kilocalorie (kcal), and may be defined as the quantity of heat required to raise the temperature of 1 kilogram of pure water from 14.5 °C to 15.5 °C at a pressure of 1 atmosphere. The temperature range needs to be specified, as the quantity of heat required to raise the temperature by 1 °C depends slightly on which °C is chosen.
- Because heat is a form of energy, it may equally, and perhaps more properly, be expressed in the SI unit of energy, the joule (J). Indeed, by international agreement the kilocalorie is now defined as 4186.8 joules.

Specific Heat

- Materials other than water require different quantities of heat to change their temperature by 1 °C. All materials may thus be attributed a value, known as the specific heat, which reflects this variation and is defined as the amount of heat required to raise the temperature of 1 kilogram of the substance by 1 °C. This may be expressed in the equation:

$$Q = mct \text{ joules}$$

- where
- Q = quantity of heat
 - m = mass of substance (kg)
 - c = specific heat ($\text{J kg}^{-1} \text{ } ^\circ\text{C}^{-1}$)
 - t = change in temperature ($^\circ\text{C}$)

- The value of specific heat depends upon the external conditions under which the heat is applied. Two variables are normally taken into consideration, leading to two values, one at constant pressure and one at constant volume. In the case of solids and liquids which are generally heated at constant pressure, then only the constant pressure value is normally quoted, and in any case the difference between the two values is negligible for all normal purposes. However, in the case of gases the two specific heats are quite different.

Change of State

- When a material changes state from a solid to a liquid or from a liquid to a gas, or vice versa, then energy must either be added to the substance or be released from the substance. For example, in order to change 1 kg of ice into water, approximately $335 \times 10^3 \text{ J}$ of heat needs to be added. During the change of state the temperature does not rise, i.e. 1 kg of ice at 0 °C changes to 1 kg of water at 0 °C. Conversely, to freeze 1 kg of water then approximately $335 \times 10^3 \text{ J}$ of heat needs to be removed. The heat which is required to change the state of a substance, without any temperature change, is known as latent heat. Where the change is between solid and liquid it is known as the latent heat of fusion; where the change is between liquid and gas it is known as latent heat of vaporization. In both cases, the values quoted refer to 1 kg of the substance at the normal melting and boiling points. (Heat energy which causes a change of temperature without giving rise to a change of state is defined as sensible heat.)

6. **Supercooling.** If a liquid is cooled slowly and is kept motionless, its temperature can be reduced to well below its normal freezing point. This is known as supercooling. A supercooled liquid is in an unstable state and any disturbance will cause some of the liquid to solidify, thereby releasing latent heat. The temperature of the supercooled liquid is raised to its freezing point by the release of this latent heat and the normal process of solidification takes place.

Heat Transfer

7. Heat may be transferred from one place to another by three mechanisms:
- a. **Conduction.** If one end of, say, a metal rod is heated, then the atoms and electrons at that end will acquire higher kinetic and potential energy than those in other parts of the rod. In random collisions, these energetic atoms and electrons will transfer energy to their neighbours which in turn will become more energetic, collide with their neighbours, and transfer some of their energy. Thus in this way the thermal energy which was applied at one end of the rod will be transferred along the rod. This process is known as conduction.
 - b. **Convection.** Convection is the transfer of heat from one place to another occasioned by the movement of the heated substance. Convection only occurs in liquids and gases.
 - c. **Radiation.** In radiation, the heat is transferred in the form of electromagnetic waves. The intervening medium plays no part in the process and indeed radiation can take place in a vacuum. Heat is radiated more efficiently by dull surfaces and dark colours than by polished ones and light colours; thus, the most efficient radiating surface is matt black. Correspondingly, radiated heat energy is most efficiently absorbed by matt black surfaces, and most efficiently reflected by light shiny ones.

CHAPTER 20 - TEMPERATURE AND EXPANSION

The Concept of Temperature

1. Temperature may be defined as the degree of hotness of a body measured according to some fixed scale. The sense of touch readily distinguishes between hot and cold bodies and thus between higher and lower temperatures. Temperature may also be regarded in another way; if two bodies are placed in thermal contact, then heat will flow from the body at higher temperature to that at a lower temperature.
2. The direction in which the heat flows does not depend on the quantity of heat in either body. For example, the water in a tank may be at a lower temperature than a hot soldering iron, but owing to its greater volume it may contain a greater quantity of heat. If the soldering iron is immersed in the water, heat will pass from the iron to the water. Once the temperatures are equal no more heat will be transferred.

Temperature Scales

3. The temperature of an object is measured by a thermometer which makes use of those properties of liquids, gases and other substances, which vary continuously with temperature and are independent of previous treatment. Common methods involve the measurement of the expansion of solids, liquids or gases as they are heated, (liquid-in-glass thermometers and bi-metallic strips), the measurement of gas pressure as a gas is heated under constant volume (constant volume gas thermometer), and the change in colour of emitted light as an object is heated (optical pyrometer).
4. All of these thermometers require to be calibrated according to a defined scale. Historically two fixed temperature points have been defined; zero degrees Celsius (originally centigrade), and 100 degrees Celsius, or their Fahrenheit equivalents of 32 °F and 212 °F.
5. The lower point was defined by the temperature at which pure water and ice exist in thermal equilibrium; the upper point at which pure water and steam exist in thermal equilibrium. Both points were defined at a pressure of 1 atmosphere ($1.01325 \times 10^5 \text{ Nm}^{-2}$).
6. Temperature scales are now defined relative to the Kelvin scale. The temperature at which all molecular agitation due to heat energy ceases is defined as Absolute Zero. This corresponds to a temperature of approximately -273° on the Celsius scale and it is assigned a value of 0 K. The size of the degree Kelvin is identical to that of the degree Celsius, thus the conventional fixed points of 0 °C and 100 °C are 273 K and 373 K respectively. Conventionally, the degrees sign is omitted when referring to Kelvin temperatures.
7. **Temperature Conversion.** Conversion of temperature values between Celsius and Fahrenheit scales may be accomplished using the following two equations:

$$(X \text{ }^\circ\text{C} \times \frac{9}{5}) + 32 = Y \text{ }^\circ\text{F}$$

$$\frac{5}{9} \times (Y \text{ }^\circ\text{F} - 32) = X \text{ }^\circ\text{C}$$

Temperature Measurement

8. Temperature is measured using a thermometer (or a pyrometer for high temperatures), various types of which are described in the following paragraphs.

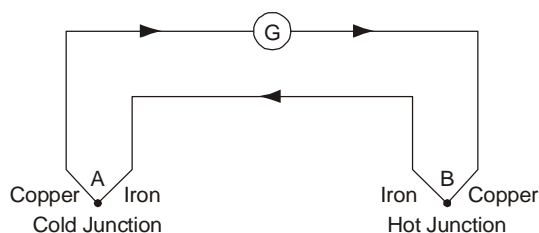
9. **Liquid-in-glass Thermometers.** The liquid-in-glass thermometer is the simplest instrument for the measurement of temperature. It depends on the fact that as the temperature of a liquid changes, so the volume of that liquid expands or contracts. For most purposes, mercury is used as the liquid. However, as mercury solidifies at $-39\text{ }^{\circ}\text{C}$, there is a limit to the lower end of its useful range. Alcohol may be used for lower temperature measurement, but it is limited at the higher temperatures as it has a boiling point of $78\text{ }^{\circ}\text{C}$.

10. **Bi-metallic Strip.** The bi-metallic strip thermometer consists of two strips of dissimilar metals welded together, and usually formed into a helix. One end of the helix is fixed while the other is free to rotate. As the two metals have different coefficients of expansion, they will expand or contract at different rates as temperature changes. This will be manifested in the helix coiling and uncoiling in response to temperature changes. A pointer is attached to the free end of the helix and this moves over a graduated scale. This type of thermometer is frequently used for outside air temperature measurement. A non-coiled version is often used as the sensing element in thermostatic controls in which the bending of the bi-metallic strip makes or breaks an electrical circuit.

11. **Pyrometers.** Conventional thermometers are not suitable for the measurement of very high temperatures, such as those found in jet pipes. The instrument used for high temperature measurement is called a pyrometer and three types are described below:

- a. **Thermocouple.** The principle of operation of the thermocouple is illustrated in Fig 1. A and B are junctions of dissimilar metals, G is a sensitive galvanometer. If the temperature of the two junctions is different, a current will flow from the iron to the copper at the colder junction and from the copper to the iron at the hotter junction. The size of the current is measured by the galvanometer; a higher current indicates a greater temperature difference between the two junctions. The advantage of this system is that the cold junction and the galvanometer can be remote from the hot, sensing, junction. The main disadvantage of the system is that it has to be calibrated, both to relate the current to the temperature difference and to determine the cold junction temperature, so that actual, rather than relative, temperatures can be determined. The two metals used in the junctions can be varied to suit the temperature range to be measured. Thermocouples are typically used in the measurement of jet pipe temperatures (see also Volume 5, Chapter 26).

13-20 Fig 1 Thermocouple



- b. **Optical Radiation Pyrometer.** The radiation pyrometer relies on the principle that materials change colour and brightness as they are heated. A simple type of optical pyrometer is used to measure the temperature of kilns and furnaces. The instrument has an electrically heated filament which is placed between the eye of the observer and the bright interior of the furnace.

The current flowing through the filament is adjusted until the filament brightness matches that of the furnace interior. As with the thermocouple, calibration against known temperature sources enables the measured current to be related to temperature. For automatic use, in a more hostile environment such as an aero-engine, more sophisticated instruments are available. These use photo-voltaic cells and amplifiers to measure the emitted radiation which is then converted to a measured temperature.

c. **Resistance Wire.** The electrical resistance pyrometer relies on the fact that the electrical resistance of materials varies with temperature. The resistance of metals increases with temperature increases while the resistance of non-metals decreases with temperature increases. Over moderate temperature ranges the resistance change is proportional to temperature change.

The Behaviour of Gases

12. It can be shown that, for a given amount of gas (say n moles), the pressure (p), the volume (V), and the temperature (T), of the gas are related by the ideal gas equation:

$$pV = nRT, \text{ where } R \text{ is the universal gas constant.}$$

13. This equation incorporates two laws as follows:

a. **Boyle's Law.** This law asserts that if the temperature of a gas is kept constant then the product of the pressure and the volume remains constant as a given amount of gas is compressed or expanded:

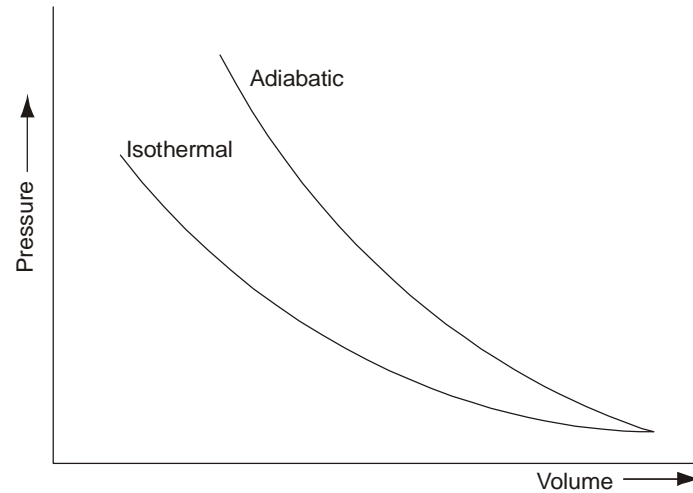
$$\text{i.e. } pV = \text{constant.}$$

b. **Charles' Law.** This law asserts that if the pressure of a gas is kept constant then the ratio of the volume to the temperature remains constant as a given amount of gas is heated or cooled:

$$\text{i.e. } \frac{V}{T} = \text{constant}$$

This behaviour of gases is used as the basis for the two types of gas thermometer: the constant volume and the constant pressure thermometers.

14. **Isothermal and Adiabatic Changes.** Fig 2 illustrates the two ways in which the pressure of a gas changes as the volume is changed. The difference depends upon whether the temperature changes concurrently. If a gas is compressed slowly such that there is time for the energy transferred to it to be dissipated through the container, then the temperature of the gas will remain constant, and the change of state is said to be isothermal. On the other hand if no energy transfer between the gas and the surroundings is permitted, then its temperature will rise and the change in state is termed adiabatic.

13-20 Fig 2 Isothermal and Adiabatic Changes**Expansion of Solids and Liquids**

15. Solids expand as they are heated. The thermal expansion of a solid is usually best described by the increase in linear dimension. The increment is directly proportional to the temperature change and to the original length. Thus:

$$\Delta L = \alpha L \Delta T$$

where L = original length
 ΔL = change of length
 ΔT = change in temperature

The constant, α , is called the coefficient of linear expansion. Its value for any material varies slightly with the initial temperature.

16. The volumetric expansion of a solid can be expressed in an analogous equation:

$$\Delta V = \beta V \Delta T$$

where V = original volume
 ΔV = change in volume
 ΔT = change in temperature

and the constant, β , is the coefficient of volume expansion ($\beta = 3\alpha$).

17. Most liquids expand when they are heated, so that their density reduces with increasing temperature. However, water exhibits a somewhat unusual variation. From 0 °C to 4 °C the volume decreases, non-linearly, as temperature increases. Above 4 °C, the volume increases with temperature. Thus, water has its maximum density at 4 °C.

CHAPTER 21 - THE NATURE OF LIGHT

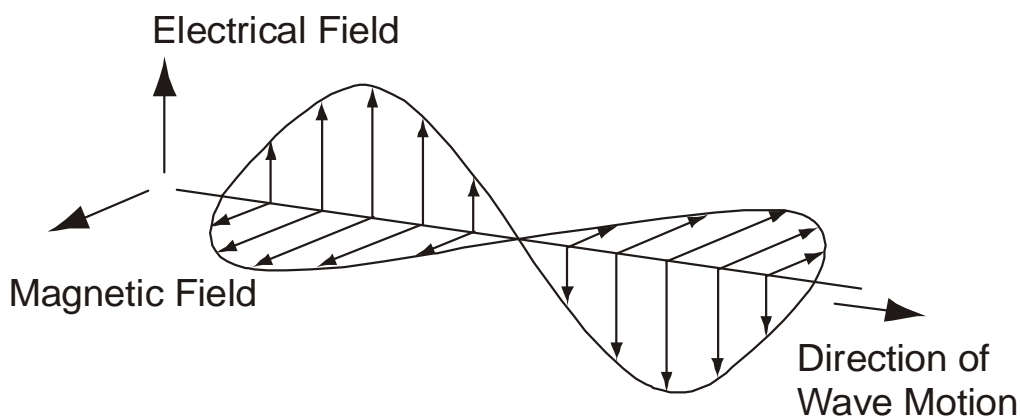
Introduction

1. The true nature of light is a question which has taxed scientists for many generations. Most theories treat light as the transport of energy, either as a wave motion or as a stream of particles. Each of these ideas can be used to explain certain phenomena associated with light but neither can satisfactorily explain them all. For example, the wave theory can explain why crossed light beams do not scatter each other, whereas the particle concept would not allow this to happen. Conversely the wave theory cannot be used to explain the photoelectric effect - only the particle theory is satisfactory. Neither idea really defines what light actually is; rather each is a model which can be used to describe and predict the behaviour of light under a particular set of circumstances. Thus, it is necessary to choose the appropriate model for the task in hand. In general, where the behaviour of light in motion is being studied, the wave model is more useful, while the particle model is to be preferred when studying the interaction of light with other matter, eg absorption and emission.

The Wave Model

2. When a pebble is tossed into a pond it sets into motion the water particles with which it comes into contact. These particles set neighbouring particles in motion and so on until the disturbance reaches the edge of the pond. In fact, any particular particle only oscillates vertically about a mean position. It does not itself move to the edge of the pond; only the disturbance moves through the water. This is a typical characteristic of wave motion. The oscillations are at right angles to the direction of propagation of the wave, and such a wave motion is known as a transverse wave. Light waves are also transverse waves, but rather than representing the motion of particles in a medium, the wave represents the variations in electric and magnetic field strength. Neither of these fields can exist without the other and they are interdependent. The fields are mutually perpendicular, and both are at right angles to the direction of wave motion. This arrangement is illustrated in Fig 1. Unlike water waves, which are constrained to move along the water surface, light waves can propagate in any direction and are not dependent upon the medium; indeed, they can move through a vacuum.

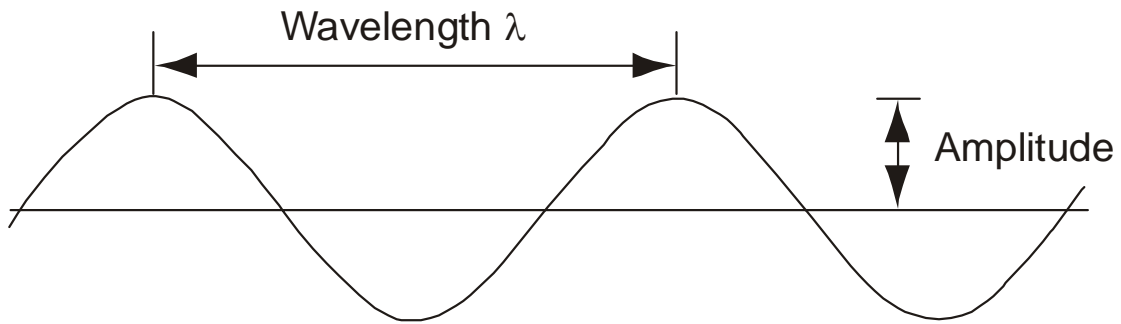
13-21 Fig 1 Light Wave



3. The waves generated by tossing a pebble into a pond are small compared with, for example, sea waves. Also, whereas the distance between the crests of the pond waves may be only a few centimetres, with sea waves the separation may be several metres. Similar variations occur in light waves and these parameters are summarized in Fig. 2, where it will be seen that light waves have a sinusoidal form. The distance between the subsequent crests of a wave (or between any other

corresponding points) is known as the wavelength and is usually given the symbol λ . The vertical size of the wave is measured from the mean level and is known as the amplitude. The time taken for corresponding points on a wave to pass a fixed point is known as the period, and the number of corresponding points passing in unit time is the frequency (f). The wave travels at a speed (C), which depends upon the medium through which it is passing. In free space the speed of light is approximately 3×10^8 metres per second (m/s or ms^{-1}) or 186,000 miles per second. From this, it will be seen that speed, frequency, and wavelength are related by the equation: $C = f\lambda$.

13-21 Fig 2 Wave Parameters



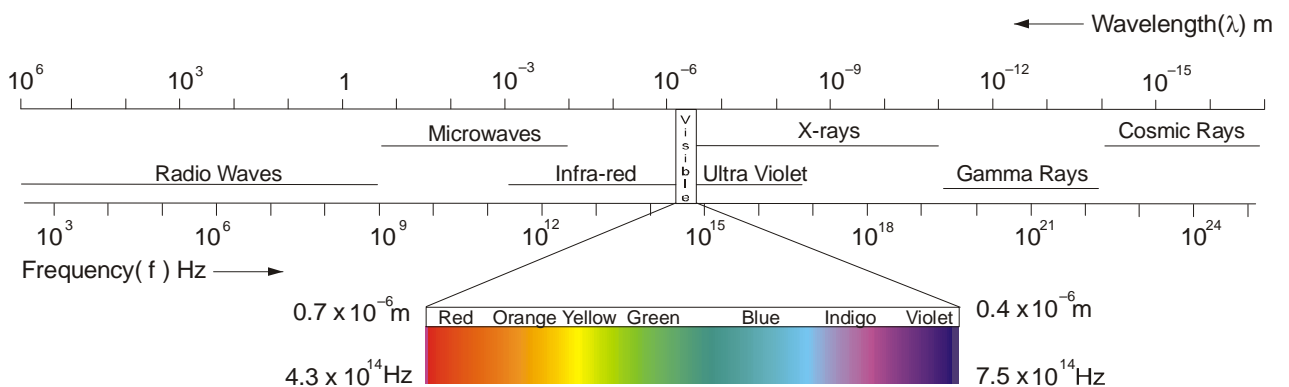
Polarization

4. In general, the electric and magnetic fields of a light wave are free to vibrate in any of the infinite planes at right angles to the direction of propagation. Any ordinary light source consists of the superposition of a number of plane waves each with a random plane of vibration. Such a light beam is said to be unpolarized. If, however, the electric field is constrained to lie in a particular plane then the light is said to be polarized. Polarization can be achieved by passing the light through a suitable filter.

The Electromagnetic Spectrum

5. Light waves can be shown to have essentially the same characteristics as many other types of radiation such as radio waves, microwaves and X-rays. Indeed, all of these are examples of electromagnetic radiation, differing primarily in their frequencies. The electromagnetic spectrum describes a large range of such waves, illustrated in Fig 3, in which visible light occupies only a very small band. White light consists of a graded spectrum containing the colours Red, Orange, Yellow, Green, Blue, Indigo and Violet.

13-21 Fig 3 The Electromagnetic Spectrum

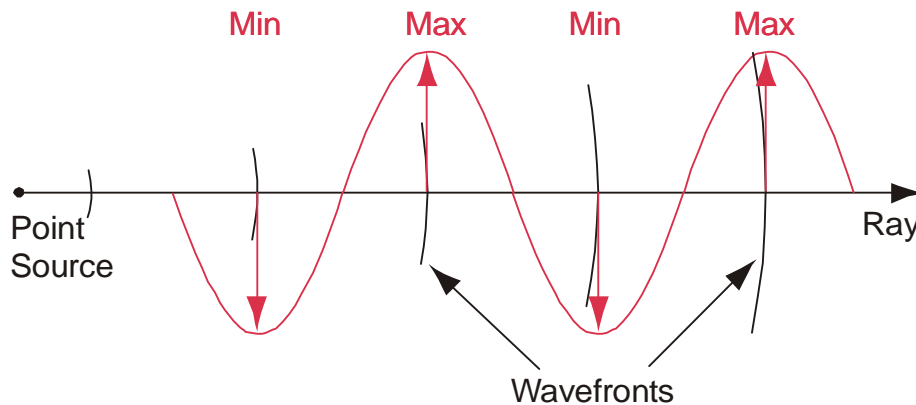


Wave-fronts and Rays

6. From a point source, light is propagated in all directions. Travelling out from the source, wave maxima will be encountered at regular intervals, corresponding to the wavelength. Similarly, at different positions, but still with the same spacing, minima will occur and the same will apply for any intermediate value of amplitude. Lines can be drawn joining points of equal amplitude, analogous to contour lines, and these are known as wave-fronts. As the radiation occurs in three dimensions these wave-fronts are in fact spherical surfaces, but it is often satisfactory to treat the radiation as if it were planar in which case the wave-fronts reduce to circles. As a further practical simplification, if the wave-front is at a large distance from the source, and if only a small sector is investigated, then the wave-front may be approximated by a straight line.

7. The direction of propagation of light is at 90° to the wave-fronts and a line representing this direction is known as a ray. Rays are often used in diagrams where only the direction of propagation is of concern. Wave-fronts and rays are illustrated in Fig 4.

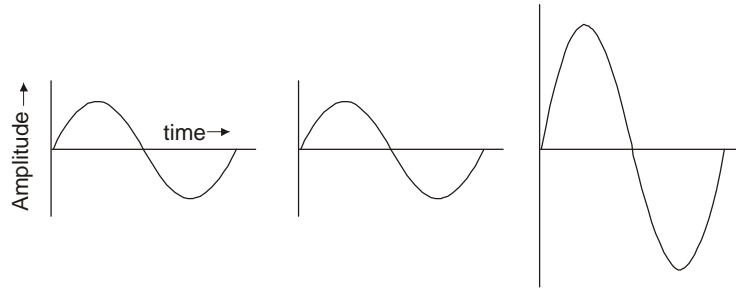
13-21 Fig 4 Wave-fronts and Rays



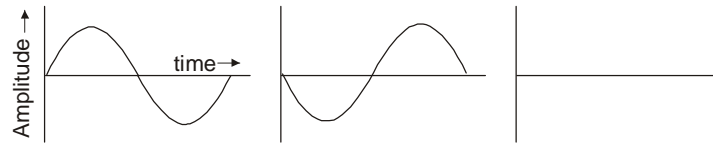
Superposition and Interference

8. When two or more sinusoidal waves act at the same place then the result can be found by adding algebraically the individual amplitudes. Four examples of this principle of superposition are shown in Figs 5, 6, 7 and 8. In Fig 5, the two waves have the same amplitude and frequency and are in phase, ie the maxima and minima of each wave occur at the same time and place. In this case the waves reinforce each other, thus resulting in a doubling of the amplitude. This is known as constructive interference. In Fig 6, the same two waves are completely out of phase. At every point the amplitudes have the same magnitude but the opposite sign, and so cancel each other - a situation known as destructive interference. In Fig 7, three waves having the same frequency but different phases add to produce another sinusoid with the same frequency. Fig 8 shows the general case where two waves have different amplitudes and frequencies.

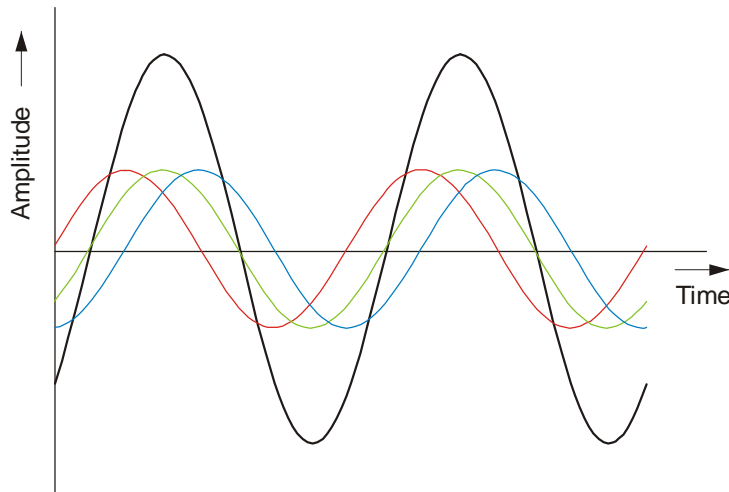
13-21 Fig 5 Two Sinusoidal Waves, Same Amplitude, Frequency and Phase



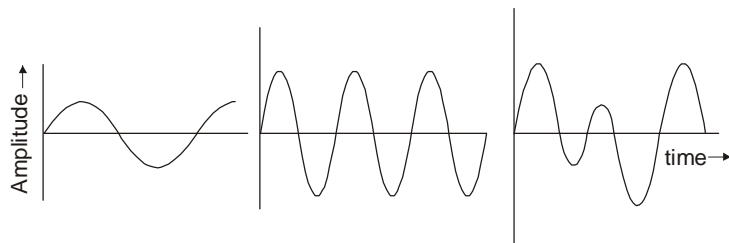
13-21 Fig 6 Two Sinusoidal Waves, Same Amplitude and Frequency, 180° Out of Phase



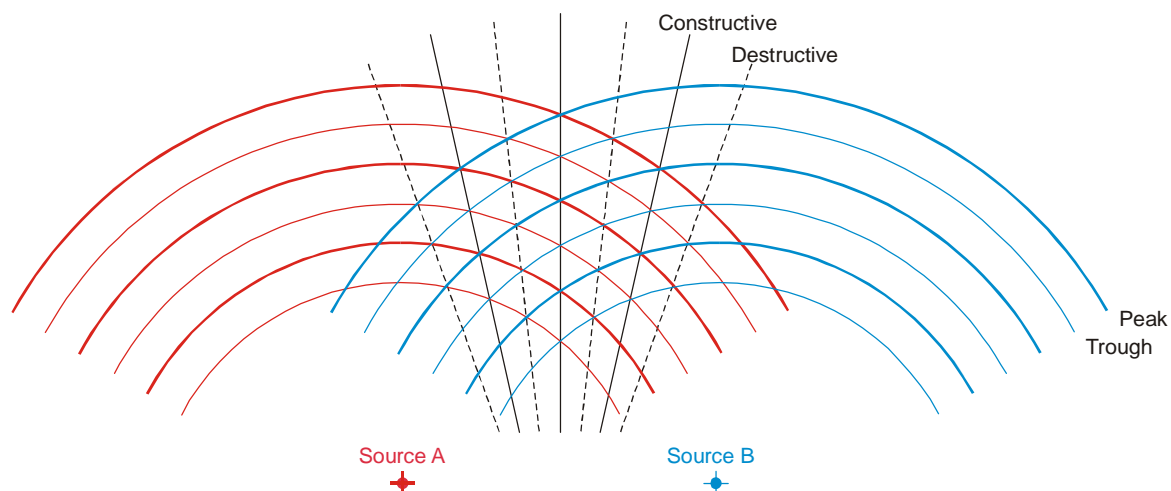
13-21 Fig 7 Three Sinusoidal Waves, Same Amplitude, and Frequency, Different Phases



13-21 Fig 8 Two Sinusoidal Waves, Different Amplitude and Frequency



9. If two close sources radiate light of the same frequency and if the two radiations are in phase then, as shown in Fig 9, at some points there will be constructive interference whilst at others there will be destructive interference. The result is a radially symmetrical interference pattern.

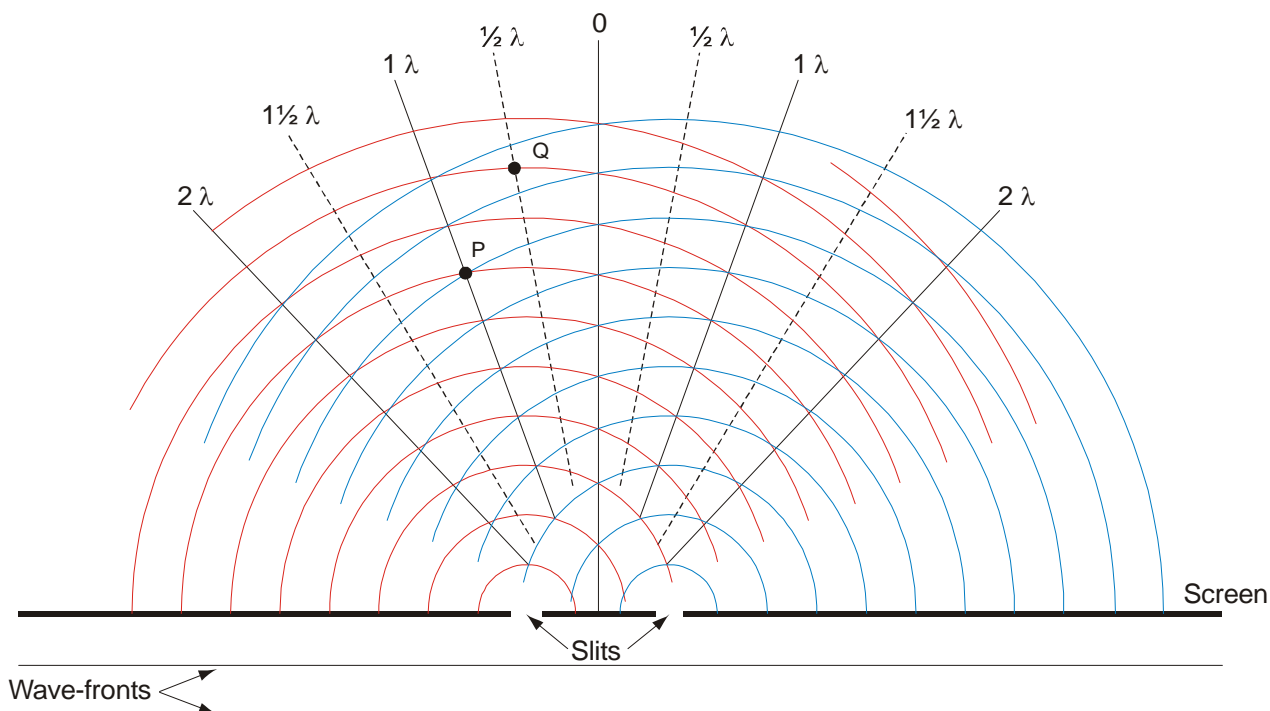
13-21 Fig 9 Constructive and Destructive Interference**Diffraction**

10. It is a characteristic feature of waves that they will deflect around the edges of obstacles placed in their path and spread into the shadow zone behind the obstacle. This effect can often be seen where for example water waves pass through a narrow harbour entrance or impinge upon a breakwater. There is no complete 'shadow' behind the wall, rather the waves appear to bend around the obstacle. This phenomenon is known as diffraction. In general, the amount of diffraction is related to the wavelength and the size of the gap through which the waves must pass. The deviations of the wave are quite large when the size of the obstacle or gap is of the same order as the wavelength. Light, behaving as a wave form, will also experience diffraction but as the wavelength of light is very small the effect is only pronounced when the obstacle or gap is very small.

11. The effect of diffraction is closely associated with the ideas of constructive and destructive interference developed in paragraphs 7 to 9. An insight into the effect can be gained by studying the passage of light through a pair of closely spaced narrow slits.

12. Fig 10 shows plane waves arriving at a screen in which there are two narrow slits which are perpendicular to the page. The two slits act as sources of light and therefore two series of circular wave-fronts can be constructed, one set from each slit. In the diagram the wave-fronts represent the crests of the waves. As the slits were illuminated by the same incident light the light leaving each slit has the same wavelength, amplitude and phase. The principle of superposition can be used to predict the effects which will be observed. At point P crests of waves from each slit arrive simultaneously and therefore constructive interference takes place and a reinforcement of amplitude will be seen. The same argument can be applied to any point where crests intersect. Such points have been joined by bold lines in the diagram and strong waves would be expected to be seen radiating along these lines. If the distance between point P and each slit is measured it will be seen to be different by one wavelength (1λ). The same is true for any point along the bold line through P. Other lines of constructive interference will each demonstrate a different value for the difference in distance and these values have been indicated in the diagram. In each case, the value is an integral multiple of a wavelength ($n\lambda$).

13-21 Fig 10 Diffraction by Two Slits

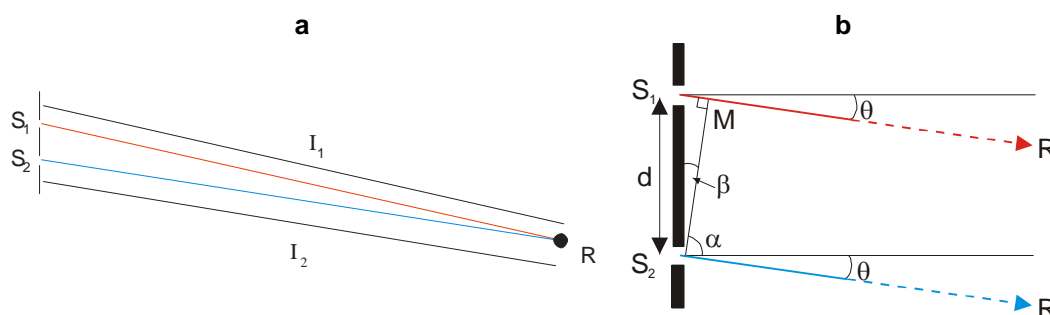


13. Point Q corresponds to the point of intersection of a crest from one slit and a trough from the other and therefore destructive interference will occur. The dashed lines represent the lines along which destructive interference occurs. In this case, the difference in distance from any point on a dashed line to each slit is an integral multiple of half a wavelength ($n\lambda/2$).

14. Points like P and Q represent the extremes of complete constructive interference and complete destructive interference. In between these points will be points where the interference is partially constructive, ie the resultant amplitude is greater than zero but less than twice the amplitude of each wave.

15. Consider now Fig 11a, in which point R is a large distance from an opaque screen in which there are two slits S_1 and S_2 . There is a small angle between the lines joining the slits to point R, however, if the point is sufficiently far away then this angle becomes so small that it may be safely ignored, and the lines can be considered to be parallel. This situation is reflected in Fig 11b which is a magnification of the area close to the screen in which the lines joining the slits to R are drawn parallel. The lines are inclined at an angle θ to lines normal to the screen.

13-21 Fig 11 Determination of Angles for Constructive and Destructive Interference



16. A line S_2M is drawn which is perpendicular to the lines from each slit to R. Point M and the slit S_2 are the same distance from R and the difference between distances I_1 and I_2 is equal to the distance between slit S_1 and M.

17. It has been shown in paragraphs 11 and 12 that the factor determining whether constructive or destructive interference occurs at R is the difference between the distances l_1 and l_2 . Thus, it is now necessary to relate the distance S_1M to the angle θ .

	angle θ + angle $\alpha = 90^\circ$
and	angle β + angle $\alpha = 90^\circ$
Thus,	angle $\beta = \text{angle } \theta$
Now,	$\sin \beta = S_1M/d$
Thus,	$\sin \theta = S_1M/d$
So,	$S_1M = d \sin \theta$

For constructive interference to occur the difference between the distances l_1 and l_2 must be equal to an integral number of wavelengths so:

$$l_1 - l_2 = S_1M = d \sin \theta = n\lambda$$

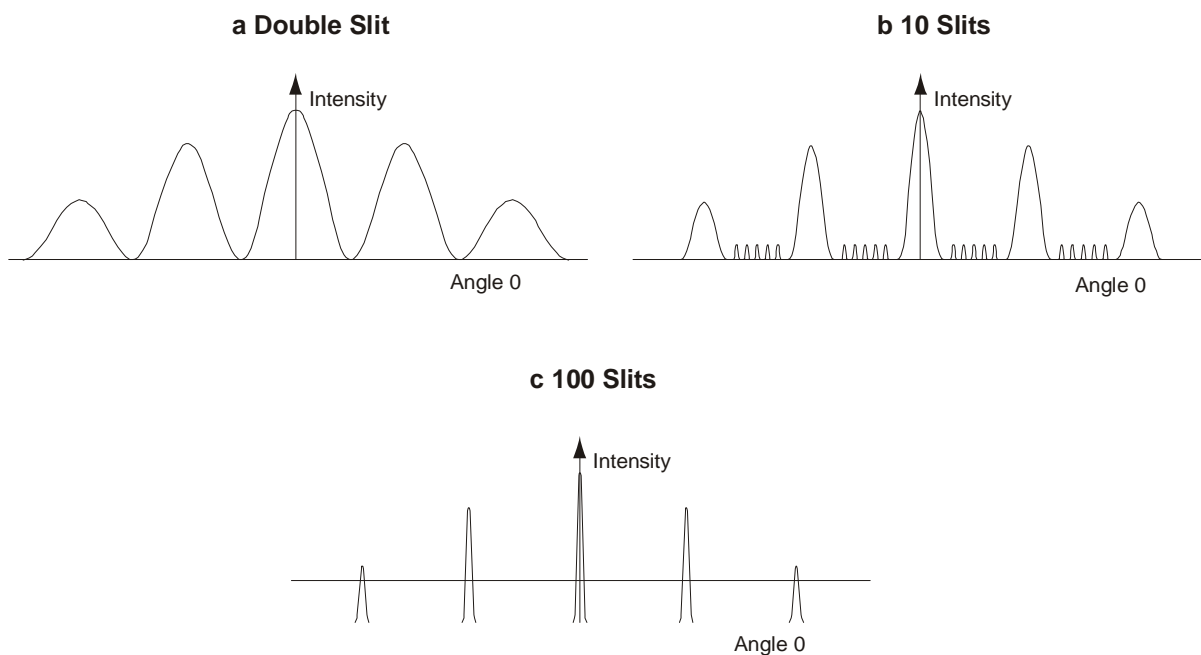
or, $\sin \theta = n\lambda/d$, where n is any integer.

This equation determines the angles at which constructive interference occurs in terms of wavelength and slit spacing. It can be shown that the equivalent equation for destructive interference is:

$$\sin \theta = (n + \frac{1}{2})\lambda/d.$$

18. As the number of equally spaced slits in the screen (and hence interference sources) is increased, the destructive interference regions widen, as shown in Fig 12. A screen with many equally spaced slits is known as a diffraction grating.

13-21 Fig 12 Diffraction Patterns



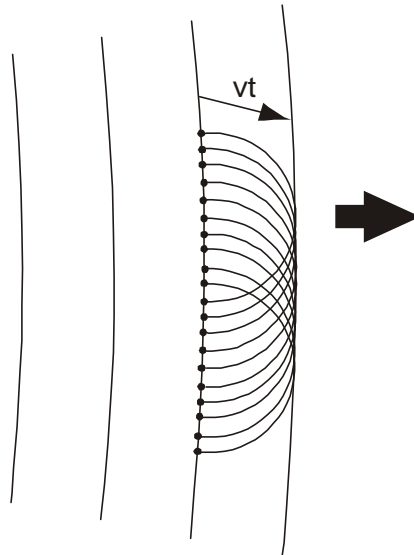
Huygens' Principle

19. A knowledge of the manner in which a wave-front propagates is necessary in order to explain the phenomena of reflection and refraction. In 1690 the Dutch physicist Huygens proposed the following method:

To find the change of position of a wave-front in a small interval, t , draw many small spheres of radius $[\text{wave speed}] \times t$ with centres on the old wave-front. The new wave-front is the surface of tangency to those spheres.

It should be remembered that this is only a model which enables predictions to be made, it is not meant to be a description of reality. The small spheres employed in this construction are known as wavelets. Clearly most diagrams are constrained to showing phenomena in two dimensions only in which case the spherical wavelets reduce to circles. Fig 13 shows how Huygens' principle is used to predict the new position of a planar wave-front after a short interval.

13-21 Fig 13 Huygens' Principle of Wave-front Propagation



Reflection

20. The case of a plane wave incident on a plane mirror is shown in Fig 14. Fig 14a shows the wave-fronts approaching a reflecting surface. One edge of the leading wave-front is just touching the surface at point P. The situation a short time later is shown in Fig 14b where some Huygens' wavelets have been constructed (the portion of the wavelets below the surface have been omitted as irrelevant). The new wave-front touches the surface at point P'. To the right of P' the new wave-front is parallel to the old wave-front as it has not yet been reflected. In order to find the position of the new wave-front to the left of P', a straight line is drawn starting from P', and tangential to the wavelet centred on P (Huygens' Principle). This straight line represents the part of the wave that has been reflected. The two right-angled triangles are congruent as they have a common side, PP', and the short sides, PQ and P'Q' are equal (being radii of the wavelets). Thus, the angles θ and θ' are equal. In studying reflection, it is usually more convenient to deal with rays rather than with wave-fronts, since they more readily show the direction of propagation.

13-21 Fig 14 Reflection Using Huygens' Principle

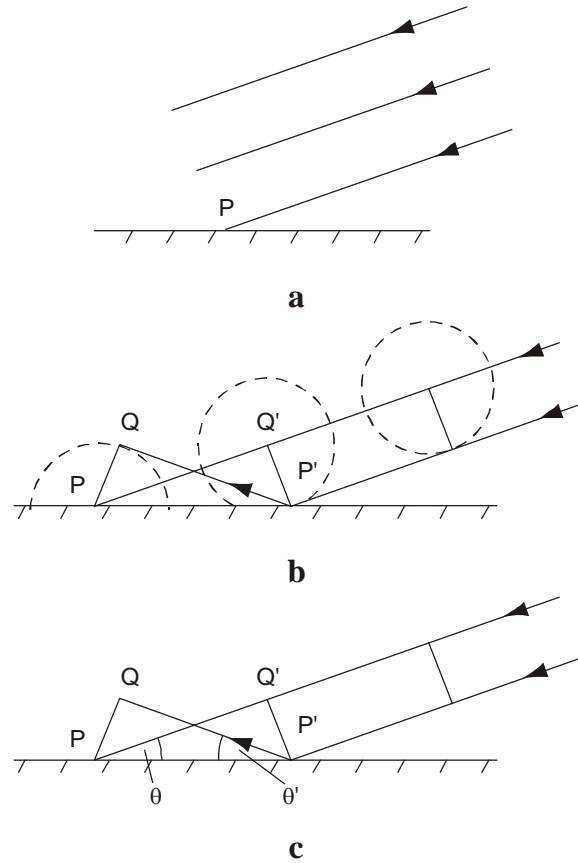
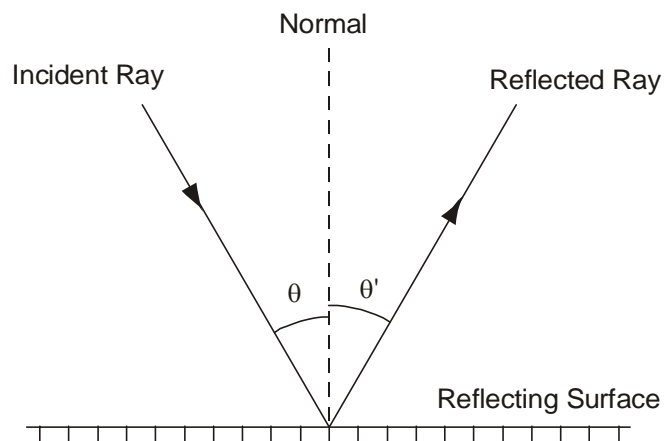


Fig 15 shows the reflection of the rays corresponding to the wave-fronts of Fig 14c.

13-21 Fig 15 Reflection of a Ray



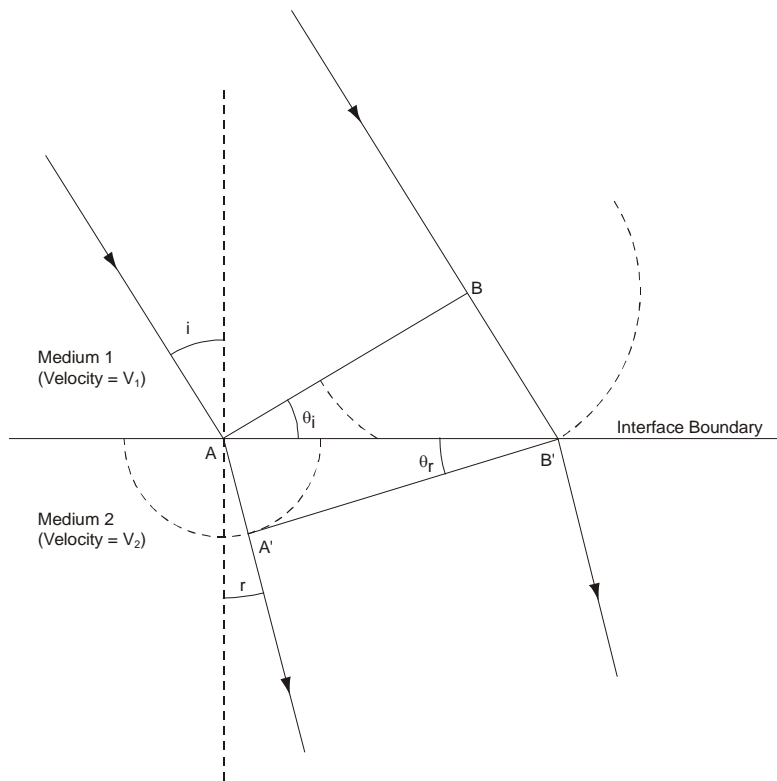
In summary there are two Laws of Reflection as follows:

- a. The angle of incidence equals the angle of reflection.
- b. The incident and reflected rays and the normal to the reflecting surface lie in the same plane.

Refraction

21. Huygens' Principle can also be used to predict the behaviour of light when it is transmitted at a boundary between two mediums rather than being reflected. Consider Fig 16. AB represents a wave-front arriving at such an interface at an angle θ_i , the point A arriving before point B. To find the position of the wave-front after a short time interval, t , Huygens' wavelets are constructed emanating from points A and B. From point B, which can be assumed to be in air, the light travels at velocity v_1 , and a wavelet can be drawn representing the time taken for the light to travel from B to B'. In the same time interval, the light from A is travelling in the transparent medium (glass say) at a slower velocity (v_2). Thus, the same time interval will correspond to a smaller wavelet. The new wave-front is now drawn from point B' to be tangential to the wavelet centred on A. Thus, the light wave has been deviated as it changes from one medium to another in which the velocity of light is different. This phenomenon is known as refraction.

13-21 Fig 16 Refraction Using Huygens' Principle



22. From Fig 16:

$$\begin{aligned} & \text{and} \quad BB' = v_1 t \\ & \text{and} \quad AA' = v_2 t \\ \text{Thus,} \quad & \frac{BB'}{AB'} = \frac{v_1}{v_2} \dots\dots\dots(1) \end{aligned}$$

$$\begin{aligned} \sin \theta_i &= \frac{BB'}{AB'} \\ \text{and} \quad \sin \theta_r &= \frac{AA'}{AB'} \end{aligned}$$

$$\text{Hence,} \quad \frac{\sin \theta_i}{\sin \theta_r} = \frac{BB'}{AA'}$$

Noting that $\theta_i = i$ and $\theta_r = r$ and comparing with equation (1):

$$\frac{\sin i}{\sin r} = \frac{v_1}{v_2} \dots\dots\dots(2)$$

23. It is normal to express the velocities of light in the two media as fractions of c , the velocity of light in a vacuum. Hence, $v_1 = c/n_1$, and, $v_2 = c/n_2$. The numbers n_1 and n_2 are known as the refractive indices of medium 1 and medium 2 respectively, and equation (2) may be rewritten as:

$$\frac{\sin i}{\sin r} = \frac{n_2}{n_1} \dots\dots\dots(3)$$

If the incident wave is travelling in a vacuum, (or, for nearly all practical purposes, in air), then $n_1 = 1$ and equation (3) reduces to:

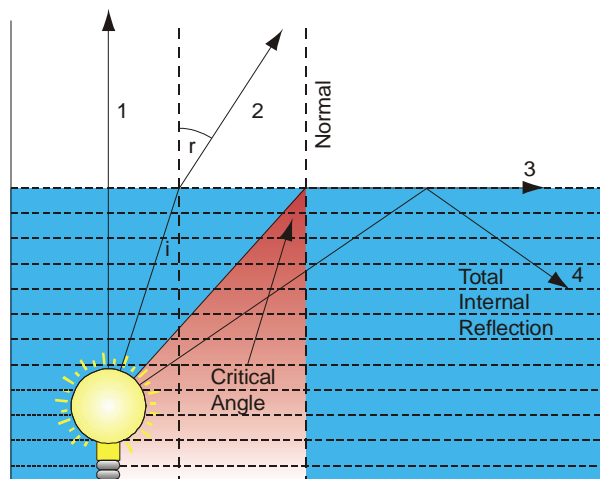
$$\frac{\sin i}{\sin r} = n$$

where n is the refractive index of the second medium. This relationship between the angle of incidence, angle of refraction and the refractive indices of the media is known as Snell's Law of refraction. When a wave passes from a medium of high velocity to one of lower velocity then it is refracted towards the normal and conversely when passing from a 'slower' medium to a 'faster' medium it is refracted away from the normal. As with reflection the two rays and the normal to the interface all lie in the same plane. It is also evident that there is a corresponding change in wavelength.

Total Internal Reflection

24. Consider rays of light travelling from a slower velocity medium, such as water to a faster velocity medium (say air) as illustrated in Fig 17. Ray 1 emitted normal to the surface of the water continues into the air normal to the surface whereas ray 2, at an angle of incidence i , is refracted through the angle r .

13-21 Fig 17 Internal Reflection



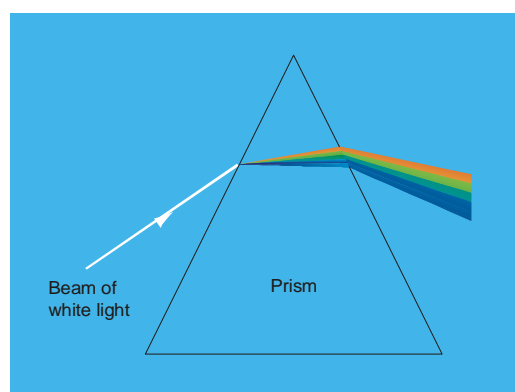
As i is increased, r eventually becomes 90° (ray 3). The value of the angle of incidence obtained in this case is called the critical angle. The change from refraction to reflection is not sudden, for some reflection always takes place at the surface of separation when i is less than critical. However, when i is greater than the critical angle all of the incident light is reflected at the boundary. This behaviour is known as total internal reflection. The phenomenon has a number of practical applications, perhaps the most important of which has been the development of the optical fibre in which the light is transmitted along the fibre experiencing total internal reflection when it impinges upon the fibre walls. Since a refracted ray cannot be deviated more than 90° from the normal to the interface, when $r = 90^\circ$, $\sin i$ has its maximum possible value and, as $\sin r = 1$:

$$\sin i = \frac{v_1}{v_2}$$

Dispersion

25. Unlike reflection, refraction is frequency dependent. This is because the velocity of a wave in a medium changes as the frequency of the wave changes. Thus the 'bending power' of a given material is dependent upon the frequency; an effect known as dispersion. Thus, for example if a ray of light containing a mix of frequencies is refracted by a medium then each of the component frequencies, or colours, will emerge at a different angle. Traditionally a prism has been used to demonstrate this effect and to analyse the component frequencies of a light source. Fig 18 illustrates the arrangement. The incident light ray experiences refraction at each glass/air interface with those components at the red end of the spectrum experiencing less refraction than those at the violet end.

13-21 Fig 18 Dispersion of Light by a Prism

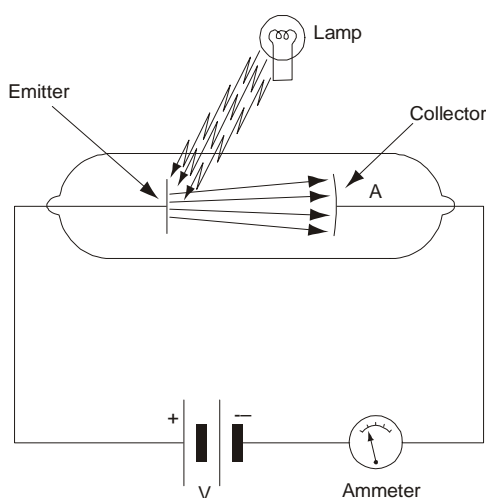


The Particle Model - The Photoelectric Effect

26. Although the wave model has provided a good description of many of the ways in which light behaves, it fails to explain the photoelectric effect.

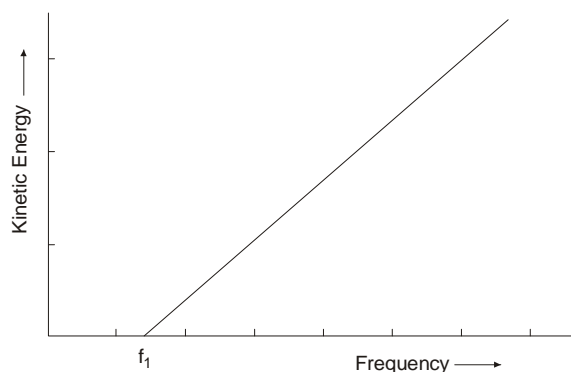
27. The photoelectric effect involves the conversion of light into electricity and is used in solar cells and photographic light meters for example. A simple device to illustrate the effect is shown in Fig 19. Light from a lamp illuminates a metal electrode enclosed in an evacuated tube. Electrons are ejected from this electrode, travel to the collecting electrode (A) and then flow around the circuit in which an ammeter can measure the current. The kinetic energy of the ejected electrons can be determined by applying a potential difference between the emitting and collecting electrodes using an adjustable source. With the polarity as shown the collector exerts a repulsive force on the electrons. A potential can therefore be applied which will just stop the flow of electrons from the emitter to the collector. This potential is known as the stopping voltage.

13-21 Fig 19 Apparatus for the Investigation of the Photoelectric Effect



28. It can be shown experimentally that the kinetic energy of the electrons increases linearly with the frequency of the incident light as shown in Fig 20. Below a certain frequency, the light is incapable of ejecting electrons; this frequency, f_1 , in Fig 20 is called the threshold frequency.

13-21 Fig 20 Variation of Kinetic Energy with Frequency



29. The energy of waves depends upon intensity and not frequency and, therefore, the wave model would predict that the kinetic energy of the ejected electrons would increase with increasing intensity. Thus, the wave model is at variance with the experimental result.

30. The particle model assumes that monochromatic light of frequency f comprises identical particles each carrying energy hf where h is a constant known as Planck's constant. The particles of light (known as photons) collide with electrons in the metal and the energy hf carried by a photon is transferred to an electron. When the frequency of the light is below the threshold frequency, the energy carried by the photon is insufficient to free even the weakest bound electrons from the metal and all of the photon's energy is converted into heat. Once the frequency exceeds the threshold frequency then the energy is sufficient to free electrons from the metal and also to give the electrons some kinetic energy. The higher the frequency, the higher the photon energy, and thus the higher will be the ejected electron's kinetic energy.

The Measurement of Light

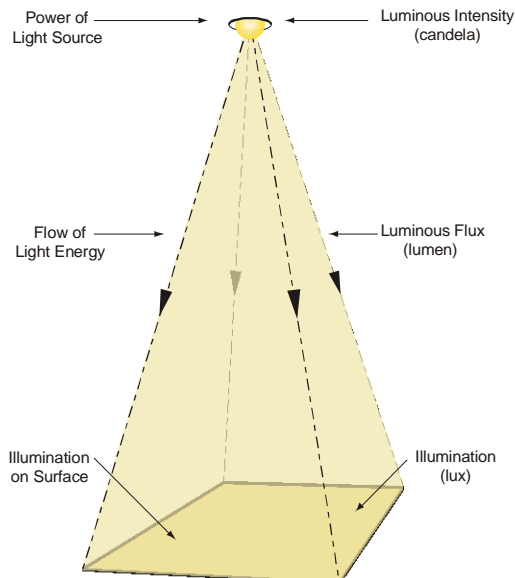
31. Visible light is part of the electromagnetic spectrum, just like radar and X-rays (see Fig 3). As with all electromagnetic radiation, it can be measured in terms of both frequency (4.3×10^{14} Hz to 7.5×10^{14} Hz) and wavelength (0.4×10^{-6} m to 0.7×10^{-6} m). However, valid though these measurements are, they do not convey information about how light is perceived by the human eye. The concepts of power, brightness, and illumination need to be considered.

32. It is beyond the scope of AP 3456 to discuss photometrics (the study and measurement of visible light) in great detail. The definitions given in the following paragraphs represent a simplified approach to light which should suffice for aircrew purposes.

33. The 'power' of traditional electric filament light bulbs was usually stated in watts, the most common types being 60 watts and 100 watts. However, the watt rating only refers to how much power the bulbs consume, not how much light they give out. Over 80% of the energy consumed is used to heat the filament to make it glow and emit light. Modern compact fluorescent (CF) bulbs generate light by 'gas discharge' and are much more energy efficient. For the same light output, CF bulbs consume about a quarter of the energy of filament bulbs. Typical light output for a conventional 100-watt bulb is about 1750 lumens. This same output can be generated by a CF bulb rated at 27 watts.

34. The 'lumen' is the SI unit of light flow or 'luminous flux' and is the most common measurement of light output. The relationship between lumens, lux and candela is shown in Fig 21. The candela is the SI unit of luminous intensity and, in very simple terms, is the amount of light generated by a 'standard' candle. A typical 100-watt incandescent filament light bulb has a luminous intensity of about 120 candelas. Light level, illumination or illuminance is measured in lux (or millilux, where 1 lux = 1000 millilux). In simple terms, 1 lux is the light level obtained when a candle is held one meter from a subject in a darkened room. If the candle is held one foot away from the subject, the light level obtained is obviously higher (this is the old imperial measure of illuminance, known as one 'foot-candle', which is approximately equal to 10 lux). To light a surface of one square meter evenly at 1 lux requires 1 lumen of total light, i.e. 1 lux = 1 lumen/m².

13-21 Fig 21 Illustration of Common Lighting Measurements



35. Outside on a clear summer day, in the UK, the light level is about 10,000 lux. Outdoor light levels for different conditions are shown in Table 1.

Table 1 Typical Outdoor Light Levels

Condition	Illumination (lux)
Full daylight	10,000
Overcast day	1,000
Very dark day	100
Twilight	10
Full Moon	0.1
Starlight	0.001 (1 millilux)
Overcast night	0.0001

36. In buildings, light levels are considerably reduced, and artificial lighting may be required depending on the tasks being undertaken. Table 2 shows recommended light levels for various indoor situations.

Table 2 Recommended Indoor Light Levels

Location/Activity	Illumination (lux)
Warehouses, Homes, Theatres	150
Normal Office Work, Library, Showrooms, Laboratories	500
Supermarkets, Mechanical Workshops	750
Operating Theatres, Normal Drawing Work	1,000
Detailed Drawing Work, Very Detailed Mechanical Work	1500 - 2000

CHAPTER 22 - MIRRORS AND LENSES

Introduction

1. This chapter will deal with the formation of images by mirrors and lenses. Images will be described as upright or inverted, magnified or diminished, real or virtual, and sometimes reversed. Whereas most of these terms are straightforward, the terms 'real' and 'virtual' need some explanation.

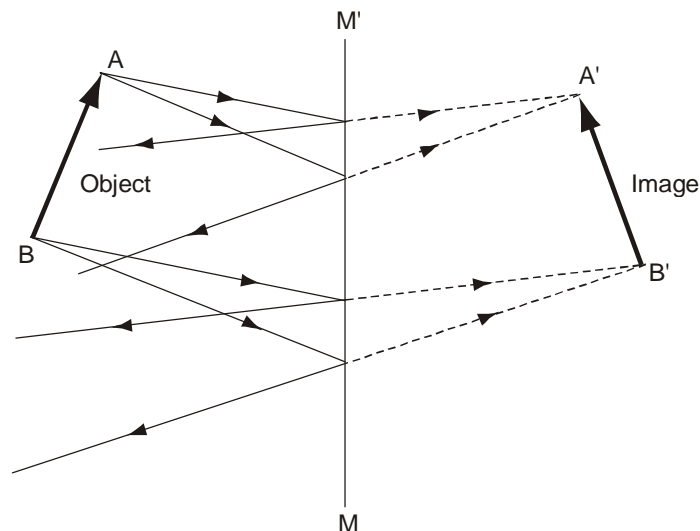
2. If rays of light coming from a point (the object) are caused to converge to a second point, the second point is called the image and is a real image. If, however, rays of light coming from a point are made to appear to diverge from a second point, the second point is a virtual image. It will become apparent that an image formed by a mirror will be real if object and image are on the same side of the mirror, whereas with a lens the image is real if it occurs on the opposite side of the lens to the object. Further differences are that a real image can be projected on to a screen whereas a virtual image cannot, and real images are inverted whilst virtual images are upright.

MIRRORS

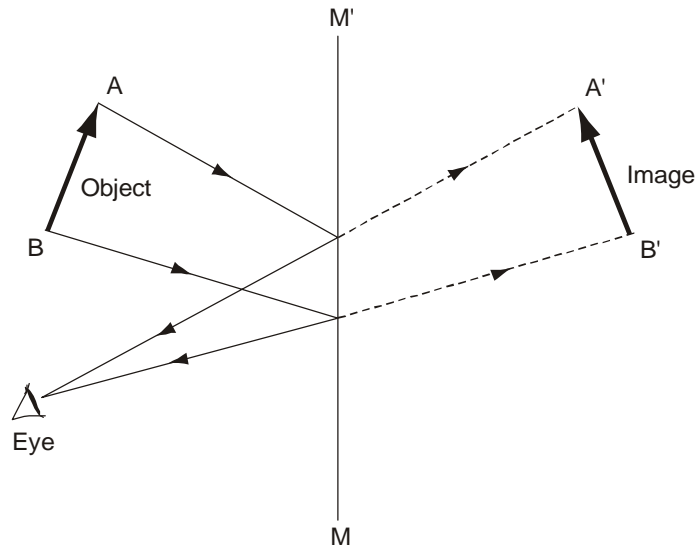
Plane Mirrors

3. Fig 1 shows a straight-line object, AB, being reflected in a plane mirror MM'. The image of AB can be determined by considering its end points. Two incident rays are drawn from each point to the mirror and produced according to the laws of reflection. Both pairs of reflected rays are then extended behind the mirror. Thus, the reflected pair of rays coming from A converge at A' and the pair from B converge at B'. If the points A' and B' are joined by a straight line then the complete image of AB is produced.

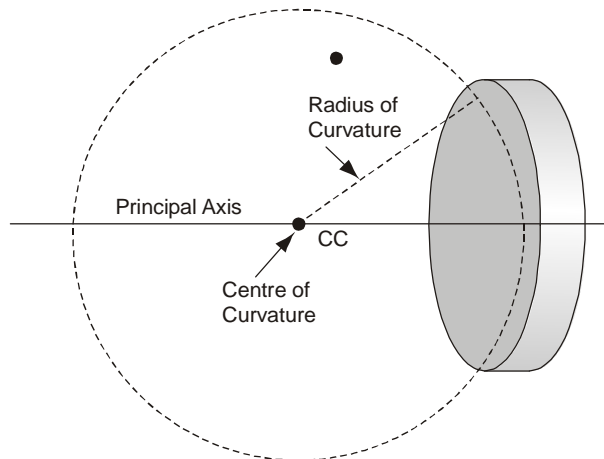
13-22 Fig 1 Reflection of an Object in a Plane Mirror



4. Fig 2 shows the paths of the rays of light required for an eye to see the image in a mirror. Measurement will show that the image is as far behind the mirror as the object is in front and that the size of the image is the same as that of the object. The image is reversed and virtual. Everyday experience in the use of plane mirrors will confirm that the image is upright.

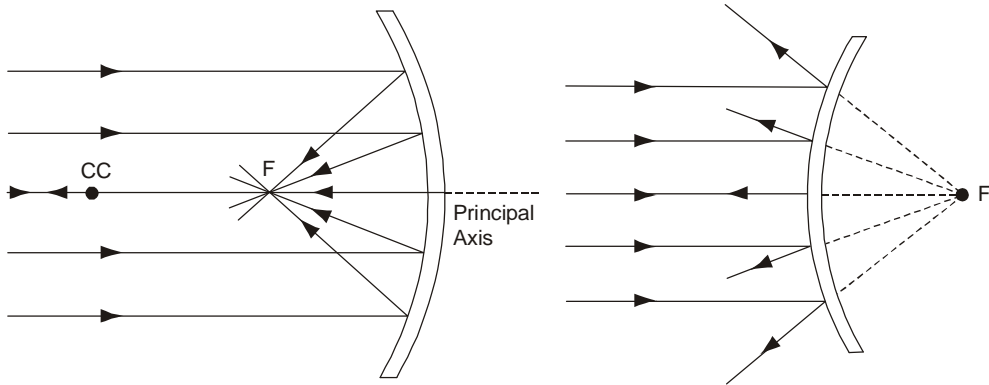
13-22 Fig 2 Viewing an Image in a Plane Mirror**Curved Mirrors**

5. The commonest types of curved mirrors are those consisting of a portion of the surface of a sphere; they may be either concave or convex. The centre of the sphere, of which the curved mirror is a part, is called the centre of curvature, and its radius is called the radius of curvature. A line drawn from the centre of curvature to the centre of the mirror surface is called the principal axis. These terms are illustrated in Fig 3.

13-22 Fig 3 Features of a Spherical Mirror

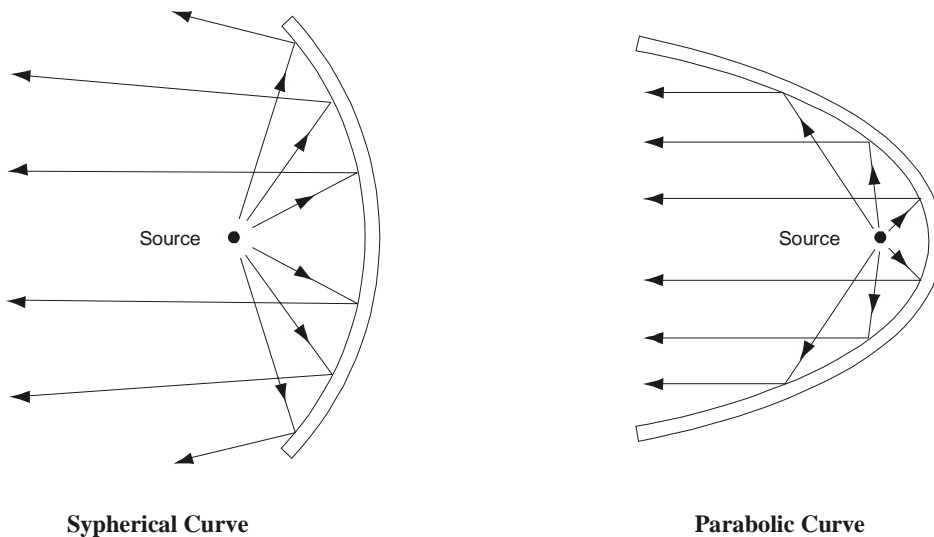
6. Consider Fig 4. If the rays of light falling on a curved mirror are parallel to the principal axis, they are reflected from a concave mirror so that they converge at one point, and from a convex mirror such that they appear to diverge from one point. This point is called the principal focus (F). The distance from the principal focus to the centre of the mirror is called the focal length and is approximately equal to half the radius of curvature.

13-22 Fig 4 Spherical Mirror - Reflection of Rays Parallel to the Principal Axis



7. If a source of light is at or near the principal focus of a spherical concave mirror, the light rays striking the mirror near its centre are reflected parallel to the principal axis. The rays striking the edge of the mirror are reflected so that they diverge. Parallel rays could be produced from all points on the mirror by altering the shape to a parabola, as shown in Fig 5.

13-22 Fig 5 Comparison of Spherical and Parabolic Curved Mirrors



Images in Curved Mirrors

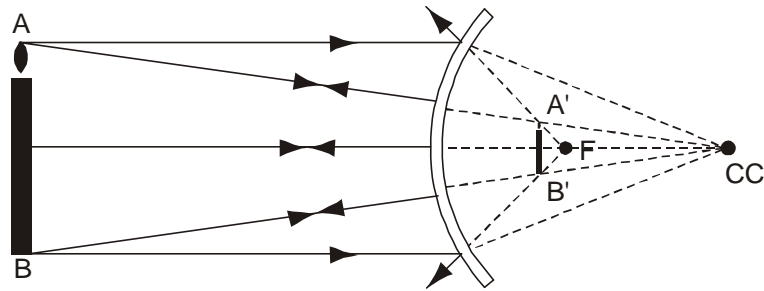
8. The position and size of an image produced by a curved mirror can be determined by the technique of ray tracing. The behaviour of certain rays from an object can be easily determined and drawn as follows:

- a. Rays parallel to the principal axis will be reflected through, or appear to diverge from, the principal focus.
- b. Rays passing through the centre of curvature will be reflected along the same line.

9. Fig 6 illustrates the principle applied to an object reflected in a convex mirror. F is the principal focus and C is the centre of curvature. Rays from points A and B on the object, parallel to the principal axis, are reflected as if they diverged from F. Rays from A and B, which would pass through C if extended behind the

mirror, are reflected back along the same path. The image is virtual, upright and smaller than the object. These characteristics are true regardless of the object's distance from the mirror.

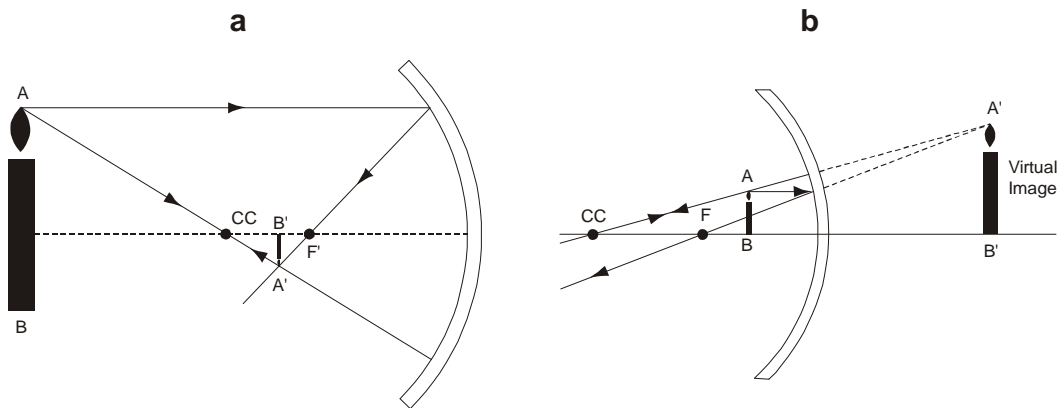
13-22 Fig 6 Formation of Image in a Convex Mirror



10. Images in concave mirrors are real unless the object is placed between the principal focus and the mirror. The image increases in size as the object is brought from infinity towards the mirror, attaining the same size as the object when the object reaches the centre of curvature and being magnified when the object is inside the centre of curvature.

11. Fig 7 shows how ray tracing can be used to determine the position and size of an image produced by a concave mirror. In Fig 7a the object is outside the centre of curvature; in Fig 7b the object is inside the principal focus.

13-22 Fig 7 Formation of Images in Concave Mirrors



12. The position of the image produced by a spherical curved mirror can be determined from the equation:

$$\frac{1}{v} + \frac{1}{u} = \frac{1}{f}$$

where u = object distance, v = image distance and f = focal length.

13. In order for the equation to differentiate between real and virtual images a sign convention is necessary. The 'real is positive' convention is normally used, and the following rules apply:

- a. A concave mirror has a positive focal length.
- b. A convex mirror has a negative focal length.

- c. Real objects are assigned a positive u value.
- d. Real images have positive v values.
- e. Virtual images have negative v values.

The formula will automatically generate the correct sign for any derived distance.

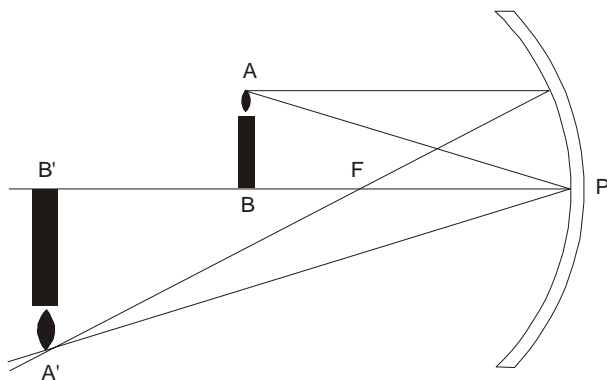
14. **Magnification.** The linear magnification due to a mirror is the ratio of the height of the image to the height of the object. In Fig 8, where $A'B'$ is the image of AB and the triangles ABP and $A'B'P$ are similar:

$$m = \frac{A'B'}{AB} = \frac{PB'}{PB} = \frac{v}{u}$$

Thus when the image is further from the mirror than the object is, the magnification will be greater than one, and vice versa. When a real object produces a real image the magnification is positive whilst if a virtual image is produced the magnification is negative. By rearranging the formula in para 12, it can be shown that magnification can be expressed in terms of v and f , or u and f as follows:

$$m = \frac{v-f}{f}, \text{ or } m = \frac{f}{u-f}$$

13-22 Fig 8 Magnification in a Concave Mirror

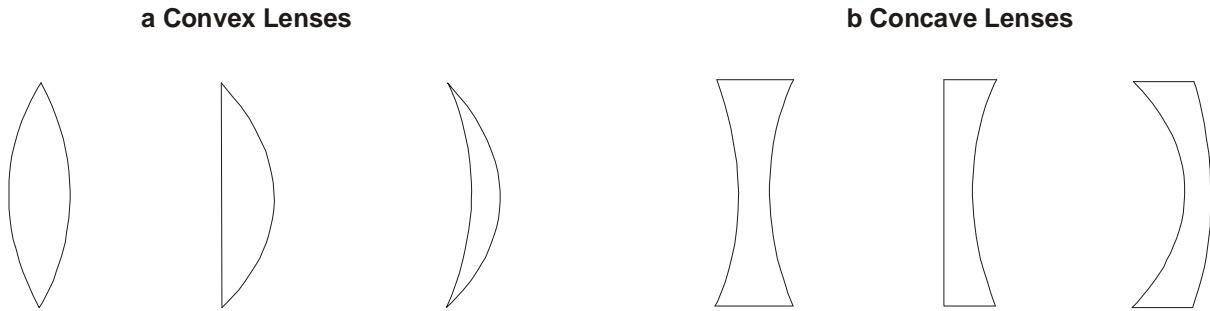


LENSES

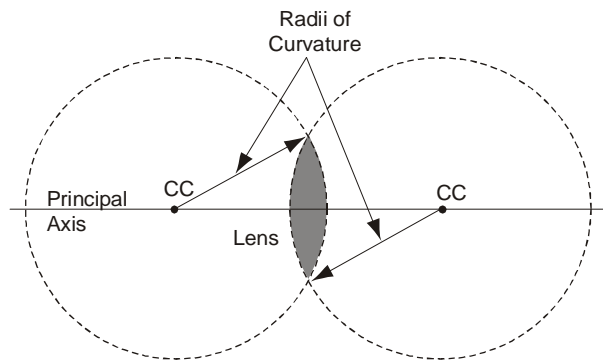
Description

15. A lens is a portion of a transparent medium bounded by two curved surfaces. Most lenses are made of glass or plastic, and their surfaces are portions of spheres or cylinders. Only spherical lenses (of which there are two basic types), will be described here:

- a. **Convex Lenses.** These are thicker at the centre than at the edges and are known as converging lenses (Fig 9a).
- b. **Concave Lenses.** These are thinner at the centre than at the edges and are known as diverging lenses (Fig 9b).

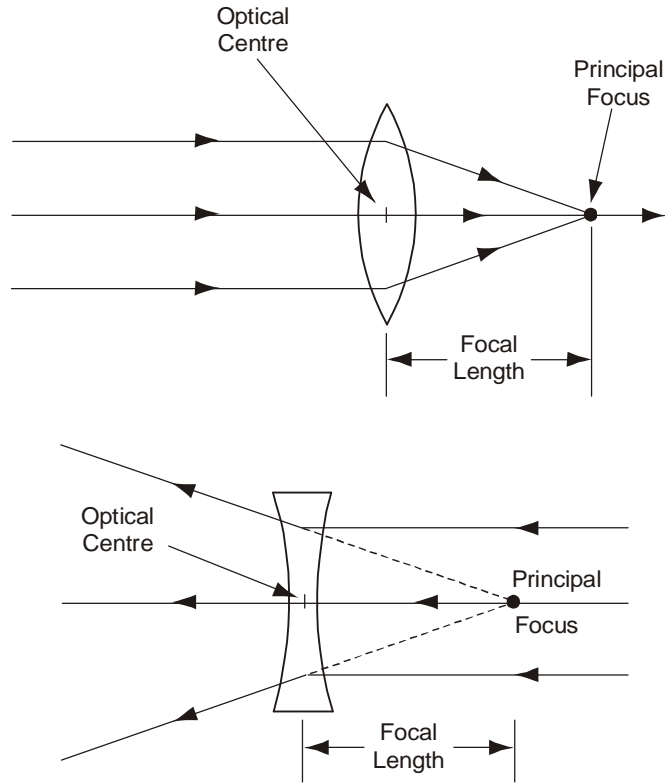
13-22 Fig 9 Convex and Concave Lenses

16. Lenses have two surfaces, each of which may be considered to be part of a spherical surface, and therefore have a centre of curvature. A straight line joining the two centres of curvature is called the principal axis and is perpendicular to the surfaces where it passes through them as, shown in Fig 10.

13-22 Fig 10 Principal Axis of Convex Lens

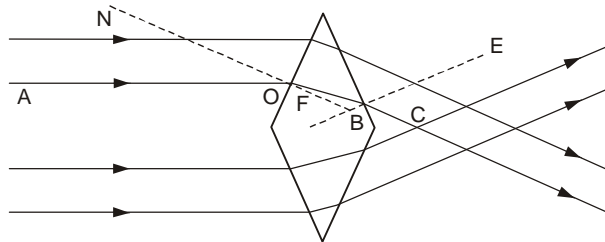
17. The principal focus is the point on the principal axis to which all rays which are close to and parallel to the axis converge, or from which they appear to diverge, after refraction. The optical centre is a fixed point for any particular lens and coincides with the geometric centre of a symmetrical lens (see para 20). The distance from the principal focus to the optical centre of a lens is called the focal length. These terms are illustrated in Fig 11.

13-22 Fig 11 Lens Terminology

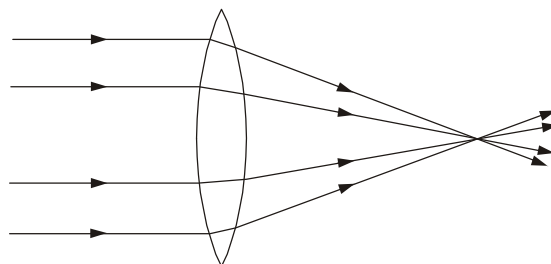


18. Refraction by Convex Lenses. A convex lens is approximately the same as two prisms placed base to base. Fig 12 shows parallel rays of light falling on a pair of prisms. At O the ray AO is refracted towards the normal NF. As it leaves the prism at B it is refracted away from the normal BE along the line BC. The feature to be noted is that light is bent towards the base or thicker part of the prism. Similarly, when light rays parallel to the principal axis fall on a convex lens they are refracted towards the thick part of the lens as shown in Fig 13.

13-22 Fig 12 Refraction by Two Prisms

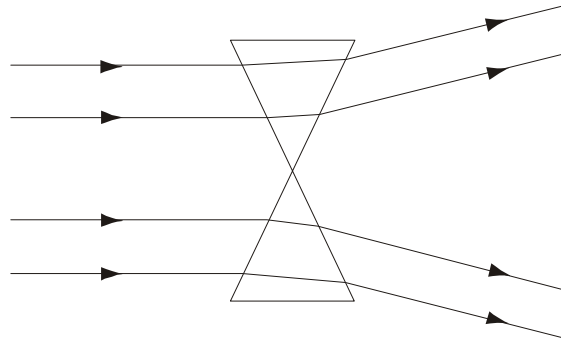


13-22 Fig 13 Refraction by Convex Lens

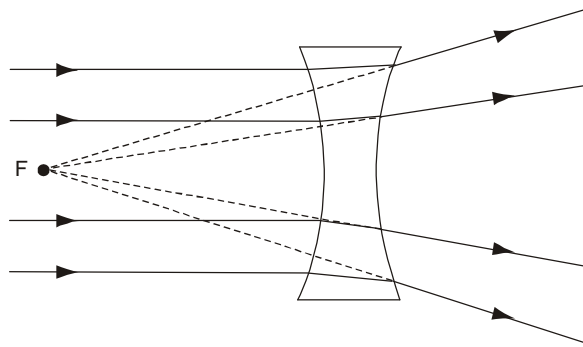


19. **Refraction by Concave Lenses.** A concave lens is approximately the same as the prism arrangement shown in Fig 14. The parallel rays of light are refracted towards the base of each prism and therefore diverge. Similarly, the lens in Fig 15 causes light rays parallel to the principal axis to diverge, apparently from a point F which is called the virtual focus.

13-22 Fig 14 Refraction by Two Prisms



13-22 Fig 15 Refraction by a Concave Lens

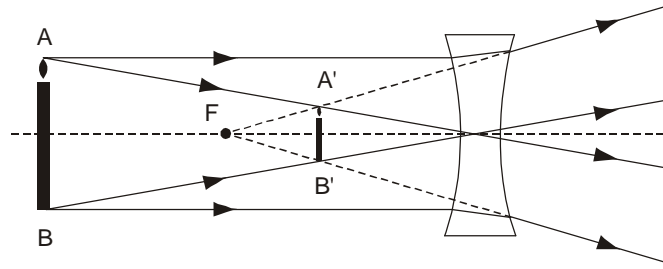


Images Formed by Lenses

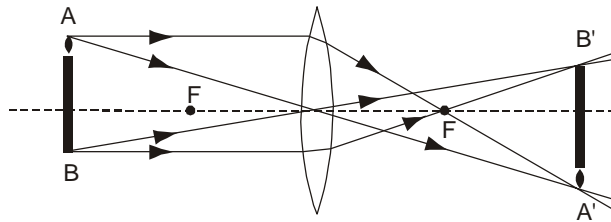
20. The ray tracing technique can equally be applied to the formation of images by lenses. The rays of use are similar to those employed in the case of mirrors. A ray parallel to the principal axis will emerge to pass through the principal focus in the case of a converging lens or will appear to have passed through the principal focus in the case of a diverging lens. As light paths are reversible, a ray which passes through the principal focus before entering a convex lens, or which would have passed through the principal focus had it not been intercepted by a concave lens, will emerge parallel to the principal axis. Finally a ray coincident with the principal axis will be undeviated; in practice provided that the thickness and diameter of the lens are small compared with its focal length, and provided that the location of the object or image point is not too far from the axis, then this non deviation rule can be generalized to include all rays passing through the optical centre.

21. Fig 16 shows the formation of an image by a concave lens. The image is always virtual, upright and smaller than the object. Fig 17 illustrates the formation of a real image by a convex lens and Fig 18 shows the formation of a virtual image by a convex lens.

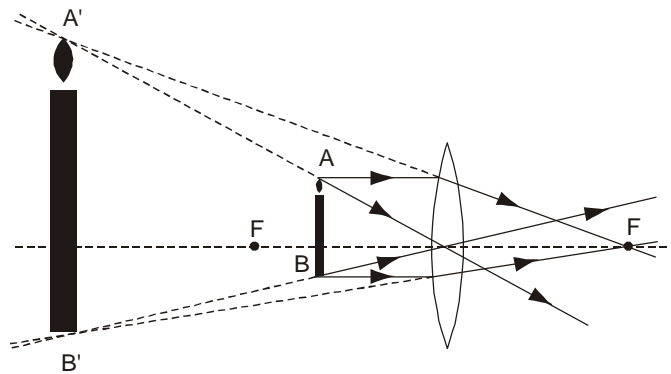
13-22 Fig 16 Formation of Virtual Image of Concave Lens



13-22 Fig 17 Formation of Real Image by Convex Lens



13-22 Fig 18 Formation of Virtual Image by Convex Lens



22. The formula: $\frac{1}{v} + \frac{1}{u} = \frac{1}{f}$ is equally applicable to lenses as it is to mirrors, using the same sign convention; convex lenses having positive focal lengths, concave lenses having negative focal lengths. The magnification formula is also the same as for mirrors.

Lens Power

23. A thick lens with sharply curved surfaces bends light rays more than a thin flat lens does; it has a shorter focal length. The ability of a lens to refract light rays is a measure of its power. The power is measured in dioptres (symbol D) and if the focal length (f) is measured in metres then:

$$D = \frac{1}{f}$$

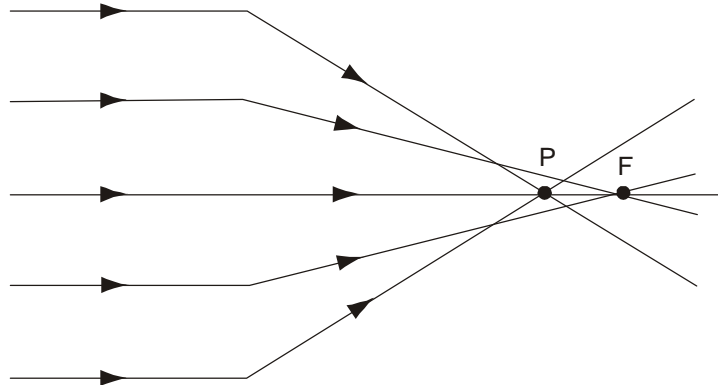
The power of a convex lens is positive and that of a concave lens is negative.

24. If two lenses are placed in contact then the resultant power can be obtained by summing the powers of the individual components lenses.

Lens Defects

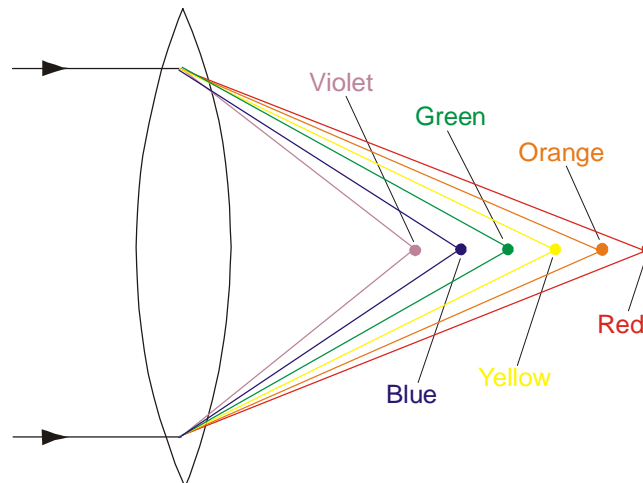
25. **Spherical Aberration.** Particularly if a lens has a wide aperture, rays parallel to the principal axis and passing through the periphery of the lens converge to a point which is nearer to the lens than the point to which a narrow central beam of parallel rays converge. At P in Fig 19, the central parts of the object are blurred whilst the peripheral portions are distinct. At F the situation will be the reverse. The distance PF is called the longitudinal spherical aberration. Spherical aberration is more pronounced in thick lenses with short focal lengths than in thin lenses with long focal lengths.

13-22 Fig 19 Spherical Aberration



26. **Chromatic Aberration.** Since the refractive index of a prism or lens is greater for violet light than for red light, the lens may be considered as having a different focal length for each colour as shown in Fig 20.

13-22 Fig 20 Chromatic Aberration



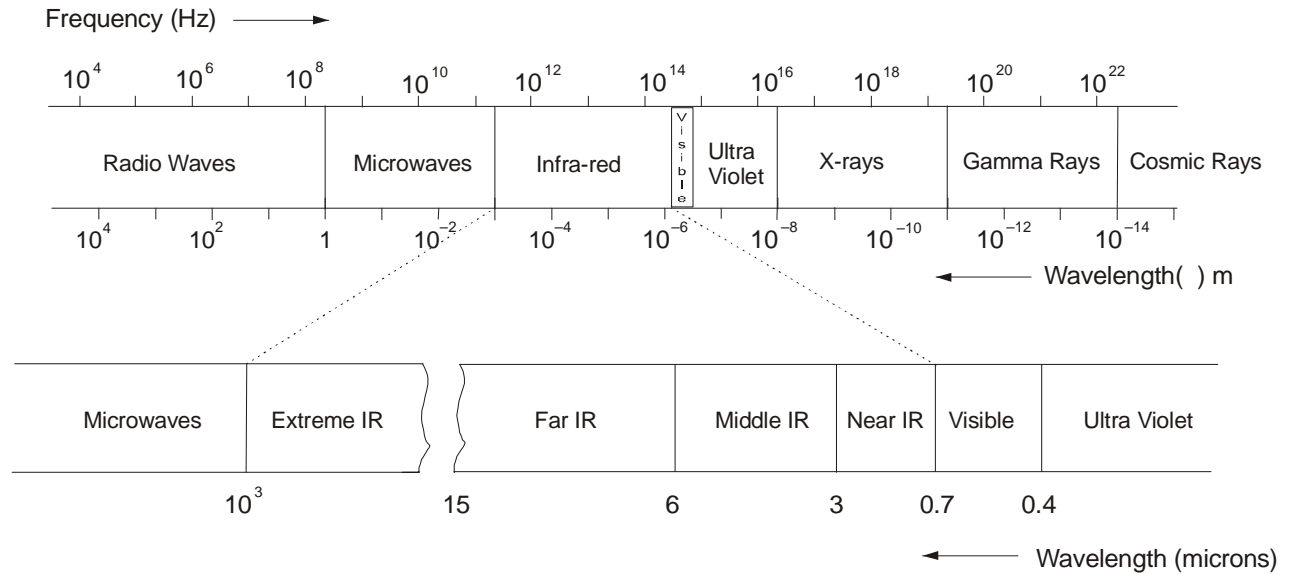
27. **Correcting Aberrations.** Spherical aberration can be prevented by placing an adjustable diaphragm in front of the lens thus eliminating peripheral light rays. Alternatively a compound lens can be used to correct for both spherical and chromatic aberration.

CHAPTER 23 - INFRA-RED RADIATION

Characteristics of Infra-red Radiation

1. Infra-red (IR) radiation is electro-magnetic radiation and occupies that part of the electro-magnetic spectrum between visible light and microwaves. The IR part of the spectrum is sub-divided into Near IR, Middle IR, Far IR and Extreme IR. The position and division of the IR band, together with the appropriate wavelengths and frequencies, is shown in Fig 1.

13-23 Fig 1 Infra-red in the Electromagnetic Spectrum



Note: One micron (μ) = 10^{-6} metres and is now known as one micro-metre in the SI system.

2. All bodies with a temperature greater than absolute zero (0 K, -273°C) emit IR radiation and it may be propagated both in a vacuum and in a physical medium. As a part of the electro-magnetic spectrum it shares many of the attributes of, for example, light and radio waves; thus it can be reflected, refracted, diffracted and polarized, and it can be transmitted through many materials which are opaque to visible light.

Absorption and Emission

3. **Black Body.** The radiation incident upon a body can be absorbed, reflected or transmitted by that body. If a body absorbs all of the incident radiation then it is termed a 'black body'. A black body is also an ideal emitter in that the radiation from a black body is greater than that from any other similar body at the same temperature.

4. **Emissivity (ϵ).** In IR, the black body is used as a standard and its absorbing and emitting efficiency is said to be unity; i.e. $\epsilon = 1$. Objects which are less efficient radiators, ($\epsilon < 1$), are termed 'grey bodies'. Emissivity is a function of the type of material and its surface finish, and it can vary with wavelength and temperature. When ϵ varies with wavelength the body is termed a selective radiator. The ϵ for metals is low, typically 0.1, and increases with increasing temperature; the ϵ for non-metals is high, typically 0.9, and decreases with increasing temperature.

Spectral Emittance

5. **Planck's Law.** A black body whose temperature is above absolute zero emits IR radiation over a range of wavelengths with different amounts of energy radiated at each wavelength. A description of this energy distribution is provided by the spectral emittance, W_λ , which is the power emitted by unit area of the radiating surface, per unit interval of wavelength. Max Planck determined that the distribution of energy is governed by the equation:

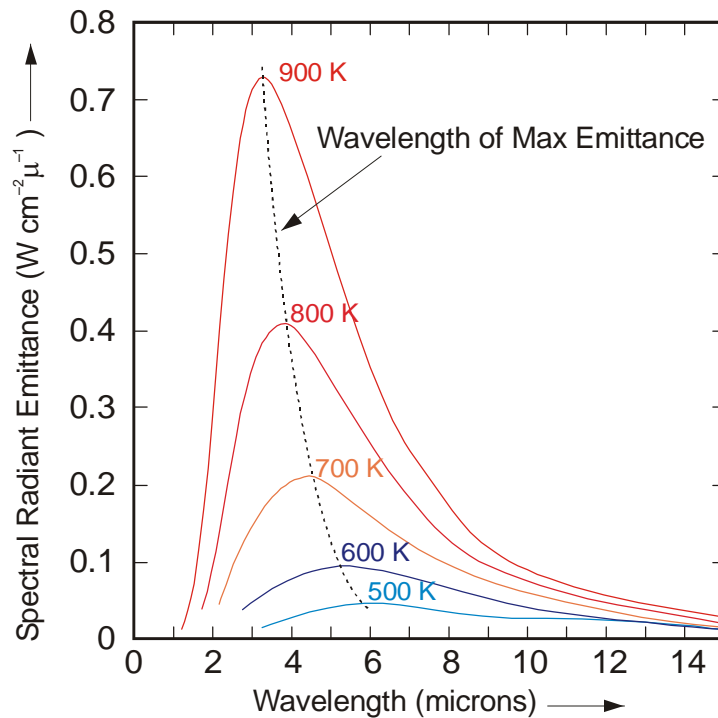
$$W_\lambda = \frac{2\pi c^2 h}{\lambda^5} \left(e^{\frac{hc}{kT\lambda}} - 1 \right)^{-1}$$

where

- λ = Wavelength
- h = Planck's constant
- T = Absolute temperature
- c = Velocity of light
- k = Boltzmann's constant

6. **Temperature/Emittance Relationship.** This rather complex relationship is best shown graphically, as in Fig 2 in which the spectral emittance is plotted against wavelength for a variety of temperatures. It will be seen that the total emittance, which is given by the area under the curve, increases rapidly with increasing temperature and that the wavelength of maximum emittance shifts towards the shorter wavelengths as the temperature is increased.

13-23 Fig 2 Distribution of IR Energy with Temperature



7. **Stefan-Boltzmann Law.** The total emittance of a black body is obtained by integrating the Max Planck equation which gives the result:

$$W = \sigma T^4$$

where W = Total emittance
 σ = Stefan Boltzmann constant
 T = Absolute temperature

For a grey body, the total radiant emittance is modified by the emissivity, thus:

$$W = \epsilon \sigma T^4$$

8. **Wien's Displacement Law.** The wavelength corresponding to the peak of radiation is governed by Wien's displacement law which states that the wavelength of peak radiation (λ_m), multiplied by the absolute temperature is a constant. Thus:

$$\lambda_m T = 2900 \mu\text{K}$$

By substituting $\lambda_m = 2900/T$ into Planck's expression it is found that:

$$W\lambda_m = 1.3 \times 10^{-15} T^5 \text{ expressed in Watts cm}^{-2}\mu^{-1}$$

ie the maximum spectral radiant emittance depends upon the fifth power of the temperature.

Geometric Spreading

9. The laws so far discussed relate to the radiation intensity at the surface of the radiating object. In general, radiation is detected at some distance from the object and the radiation intensity decreases with distance from the source as it spreads into an ever-increasing volume of space. Two types of source are of interest; the point source and the plane extended source.

10. **Point Source.** A point source radiates uniformly into a spherical volume. In this case the intensity of radiation varies as the inverse square of the distance between source and detector.

11. **Plane Extended Source.** When the radiating surface is a plane of finite dimensions radiating uniformly from all parts of the surface then the radiant intensity received by a detector varies with the angle between the line of sight and the normal to the surface. For a source of area A the total radiant emittance is WA . The radiant emittance received at a distance d and at an angle θ from the normal is given by:

$$\frac{WA}{2\pi d^2} \cos\theta$$

IR Sources

12. It is convenient to classify IR sources by the part they play in IR systems; ie as targets, as background, or as controlled sources. A target is an object which is to be detected, located or identified by means of IR techniques, while a background is any distribution or pattern of radiation, external to the observing equipment, which is capable of interfering with the desired observations. Clearly what might be considered a target in one situation could be regarded as background in another. As an example terrain features would be regarded as targets in a reconnaissance application but would be background in a low-level air intercept situation. Controlled sources are those which supply the power required for active IR systems (e.g. communications), or provide the standard for calibrating IR devices.

Targets

13. **Aircraft Target.** A supersonic aircraft generates three principle sources of detectable and usable IR energy. The typical jet pipe temperature of 773 K produces a peak of radiation, (from Wien's law), at 3.75μ . The exhaust plume produces two peaks generated by the gas constituents; one at 2.5 to 3.2μ due to carbon dioxide, the other at 4.2 to 4.5μ due to water vapour. The third source is due to leading edge kinetic heating giving a typical temperature of 338 K with a corresponding radiation peak at about 7μ .

14. **Reconnaissance.** Terrestrial IR reconnaissance and imaging relies on the IR radiation from the Earth which has a typical temperature of 300 K. The peak of radiation corresponding to this temperature is about 10μ and so systems must be designed to work at this wavelength.

Background Sources

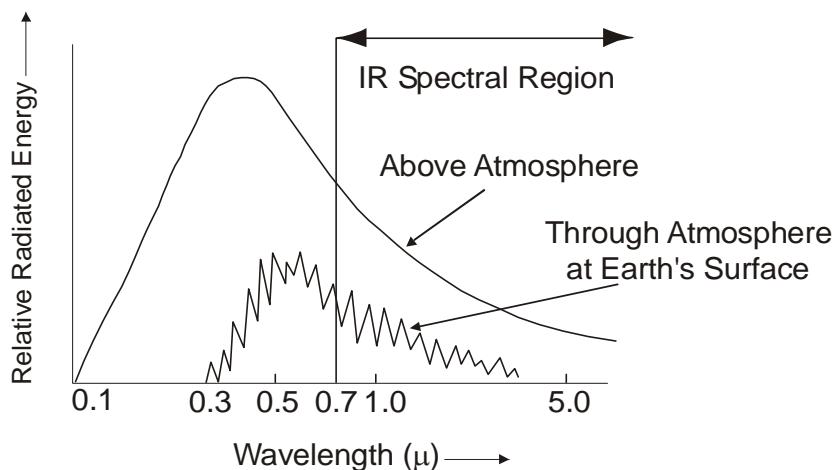
15. Regardless of the nature of the target source, a certain amount of background or interfering radiation will be present, appearing in the detection system as noise. The natural sources which produce this background radiation may be broadly classified as terrestrial or atmospheric and celestial.

16. **Terrestrial Sources.** Whenever an IR system is looking below the horizon it encounters the terrestrial background radiation. As all terrestrial constituents are above absolute zero they will radiate in the infra-red, and in addition IR radiation from the sun will be reflected. Green vegetation is a particularly strong reflector which accounts for its bright image in IR photographs or imaging systems. Conversely, water, which is a good reflector in the visible part of the spectrum, is a good absorber of IR, and therefore appears dark in IR images.

17. **Atmospheric and Celestial Sources.** Whenever an IR device looks above the horizon the sky provides the background radiation. The radiation characteristics of celestial sources depend on the source temperature together with modifications by the atmosphere.

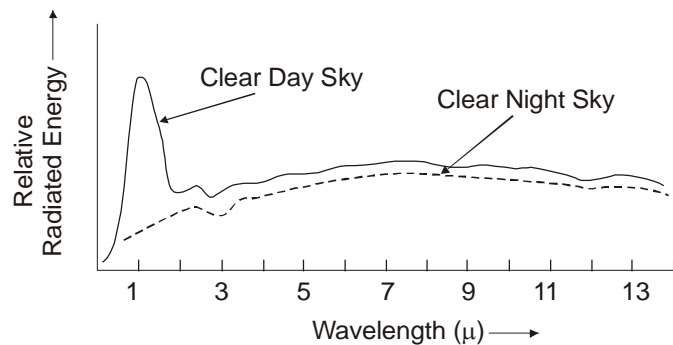
- a. **The Sun.** The sun approximates to a black body radiator at a temperature of 6,000 K and thus has a peak of radiation at 0.5μ , which corresponds to yellow-green light. The distribution of energy is shown in Fig 3 from which it will be seen that half of the radiant power occurs in the infra-red. The Earth's atmosphere changes the spectrum by absorption, scattering and some re-radiation such that although the distribution curve has essentially the same shape, the intensity is decreased and the shorter, ultraviolet, wavelengths are filtered out. The proportion of IR energy remains the same or perhaps may be slightly higher. Sunlight reflected from clouds, terrain and sea shows a similar energy distribution.
- b.

13-23 Fig 3 Spectral Distribution of Solar Radiation



- b. **The Moon.** The bulk of the energy received from the moon is re-radiated solar radiation, modified by reflection from the lunar surface, slight absorption by any lunar atmosphere and by the Earth's atmosphere. The moon is also a natural radiating source with a lunar daytime surface temperature up to 373 K and lunar night time temperature of about 120 K. The near sub-surface temperature remains constant at 230 K, corresponding to peak radiation at 12.6μ .
- c. **Sky.** Fig 4 shows a comparison of the spectral distribution due to a clear day and a clear night sky. At night, the short wavelength background radiation caused by the scattering of sunlight by air molecules, dust and other particles, disappears. At night there is a tendency for the Earth's surface and the atmosphere to blend with a loss of horizon since both are at the same temperature and have similar emissivities.
- d.

13-23 Fig 4 Spectral Energy Distribution of Background Radiation from the Sky



- d. **Clouds.** Clouds produce considerable variation in sky background, both by day and by night, with the greatest effect occurring at wavelengths shorter than 3μ due to solar radiation reflected from cloud surfaces. At wavelengths longer than 3μ , the background radiation intensity caused by clouds is higher than that of the clear sky. Low bright clouds produce a larger increase in background radiation intensity at this wavelength than do darker or higher clouds. As the cloud formation changes the sky background changes and the IR observer is presented with a varying background both in time and space. The most serious cloud effect on IR detection systems is that of the bright cloud edge. A small local area of IR radiation is produced which may be comparable in area to that of the target, and also brighter. Early IR homing missiles showed a greater affinity for cumulus cloud types than the target aircraft. Discrimination from this background effect requires the use of spectral and spatial filtering.

IR Transmission in the Atmosphere

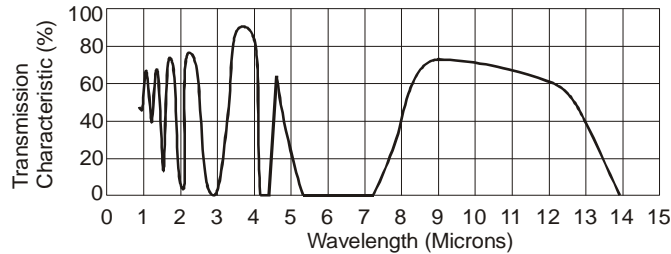
18. **Atmospheric Absorption.** The periodic motions of the electrons in the atoms of a substance, vibrating and rotating at certain frequencies, give rise to the radiation of electro-magnetic waves at the same frequencies. However, the constituents of the Earth's atmosphere also contain electrons which have certain natural frequencies. When these natural frequencies are matched by those of the radiation which strikes them, resonance absorption occurs and the energy is re-radiated in all directions. The effect of this phenomenon is to attenuate certain IR frequencies. Water vapour and carbon dioxide are the principle attenuators of IR radiation in the atmosphere. Figs 5a, 5b and 5c show the transmission characteristics of the atmosphere at sea-level, at 30,000 ft and at 40,000 ft.

19. **Scattering.** The amount of scattering depends upon particle size and particles in the atmosphere are rarely bigger than 0.5μ , and thus they have little effect on wavelengths of 3μ or greater. However, once moisture condenses on to the particles to form fog or clouds, the droplet size can range between 0.5 and 80μ , with the peak of the size distribution between 5 and 15μ . Thus fog and cloud particles are comparable in size to IR wavelengths and transmittance becomes poor. Raindrops are considerably

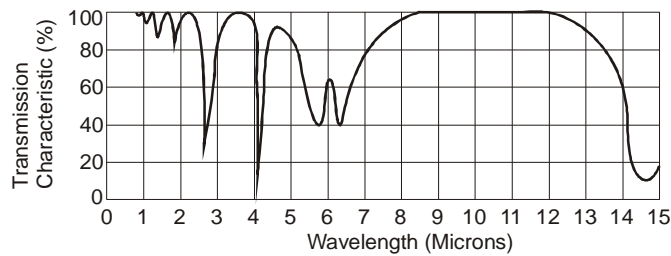
larger than IR wavelengths and consequently scattering is not so pronounced. Rain, however, tends to even out the temperature difference between a target and its surroundings.

13-23 Fig 5 Atmospheric Transmittance vs. Altitude

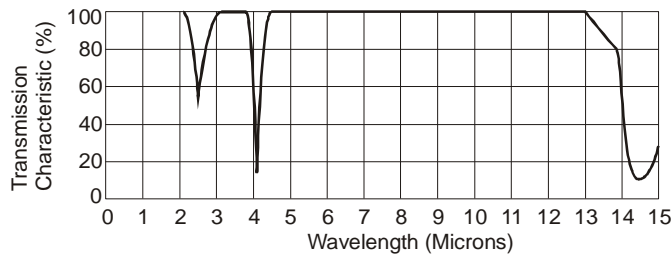
a Transmittance at Sea Level



b Transmittance at 30,000 ft



c Transmittance at 40,000 ft



20. **Scintillation.** Where a beam of IR passes through regions of temperature variation it is refracted from its original direction. Since such regions of air are unstable, the deviation of the beam is a random, time varying quantity. The effect is most pronounced when the line of sight passes close to the earth and gives rise to unwanted modulations of the signal, and incorrect direction information for distant targets.

CHAPTER 24 - LASERS

Introduction

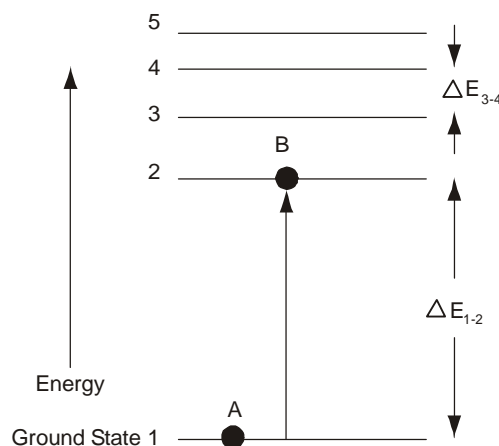
1. The laser is a device that emits an extremely intense beam of energy in the form of electromagnetic radiation in the near ultra-violet, visible, or infra-red part of the electromagnetic spectrum. The word LASER is an acronym derived from the definition of its function, Light Amplification by Stimulated Emission of Radiation. The word has been so integrated into the English language that it is no longer written in capital letters, as are most acronyms. Indeed, in technical circles, its use has spawned a verb, to lase, which describes the action of using a laser. Unlike the radiation from other sources, laser light is monochromatic (single wavelength), coherent (all waves in phase), and highly collimated (near parallel beam). Since the first laser was constructed in 1960 in California, the development has been rapid and uses have been found in a wide variety of civil and military spheres, including surgery, communications, holography and target marking and range-finding. In order to understand the principle of operation of a laser it is first necessary to appreciate some aspects of atomic structure and energy levels.

Atomic Energy Levels

2. The atom consists of a central nucleus containing positively charged protons and neutral neutrons. Surrounding the nucleus are negatively charged electrons. The number of protons and electrons are equal thus resulting in a net zero charge on the atom as a whole. The electrons have a certain energy level due to the sum of their kinetic energy and electrostatic potential energy. However electrons within an atom are constrained to exist in one of a series of discrete energy levels. In normal circumstances the electrons will adopt the minimum energy levels permitted and the atom is then said to be in its ground state. In order for electrons to enter a higher energy level, energy in one form or another has to be supplied. If such a transition to a higher energy level occurs the atom is said to be in an excited state.

3. Conventionally, the energy states of an atom can be shown on an energy level diagram as shown in Fig 1. The horizontal lines represent the permitted energy levels, increasing upwards, separated by varying energy differences, ΔE . The horizontal extent of the lines has no significance. The base line is the ground state i.e. the lowest energy level in which atoms will normally be found (A in Fig 1). By supplying energy, it may be possible to excite an atom (B in Fig 1) into a higher energy level. This process is known as absorption.

13-24 Fig 1 Atomic Energy Levels



Emissions

4. **Spontaneous Emission.** An atom in an excited state is unstable and will have a tendency to revert to the ground state. In doing so, it will emit the excess energy as a single quantum of energy known as a photon, a process known as spontaneous emission. If a large population of atoms are excited into higher states, as for example in a fluorescent lighting tube, then they will occupy a wide band of energy levels. On undergoing spontaneous emissions, some will revert to the ground state directly whilst others will drop via intermediate levels. In either case photons will be emitted with a wide range of energy levels corresponding to the various energy level differences. The frequency of the emitted energy is determined by the Planck-Einstein equation:

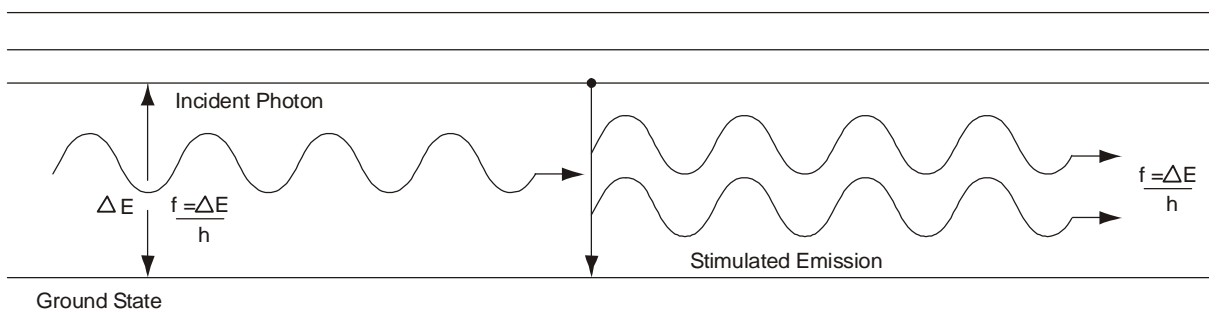
$$E = hf$$

where E is the photon energy, f is the frequency and h is Planck's constant.

Thus in a fluorescent tube, as there are a wide variety of energy level transitions, there will be a wide variety of frequencies in the emitted light giving the impression of white light. It should be noted that what transitions occur and when they occur is a random process. Equally the direction in which the emitted photon is radiated is also random. Thus the radiation generated by spontaneous emission is isotropic (i.e. radiating in all directions), non-coherent and covers a wide frequency band.

5. **Stimulated Emission.** As early as 1917 Einstein predicted on theoretical grounds that the downward transition of an atom could be stimulated to occur by an incident photon of exactly the same energy as the difference between the energy levels. It is this type of emission that is exploited in lasers. This process is shown in Fig 2. It should be noted that the incident photon is not absorbed and so for each incident photon, two photons are emitted, each of which can stimulate further emissions providing that there are atoms in the higher energy level. Furthermore, these emitted photons have the same energy, and therefore frequency, the same phase and are emitted in the same direction as the incident photons. These are, of course, the characteristics of laser radiation.

13-24 Fig 2 Stimulated Emission

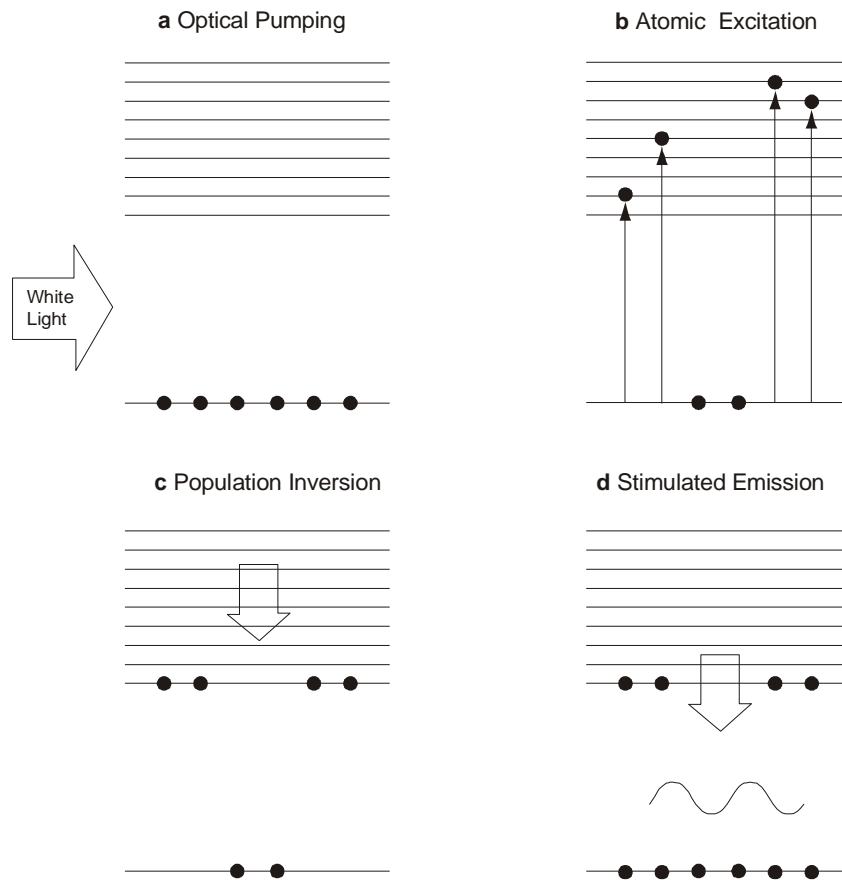


6. **Population Inversion.** In the normal course of events, however, most atoms are in the ground state and so incident photons are more likely to excite a ground state atom than to induce stimulated emission. It is therefore necessary to ensure that there are more atoms in the appropriate higher energy level than in the ground state, a situation known as a population inversion. The process by which this is achieved will be described with reference to the ruby laser which was the first lasing medium to be used.

7. **Optical Pumping.** Fig 3a illustrates the normal configuration with respect to the chromium atoms within a ruby crystal. The diagram shows a number of atoms in the ground state and a number of as yet unoccupied higher energy levels. It should be noted that the energy levels in Fig 3 refer to the energy of the atom as a whole and not to the energy levels of the constituent electrons. At the start of the process the ruby is subjected to a burst of intense white light generated by a system similar to a

photographic electronic flash gun. As the white light comprises a wide range of frequencies then a whole range of energies will be imparted to the ground state atoms. Some of these atoms will therefore be excited to a range of higher energy levels (Fig 3b); a process known as optical pumping.

13-24 Fig 3 The Stages of Operation in a Ruby Laser



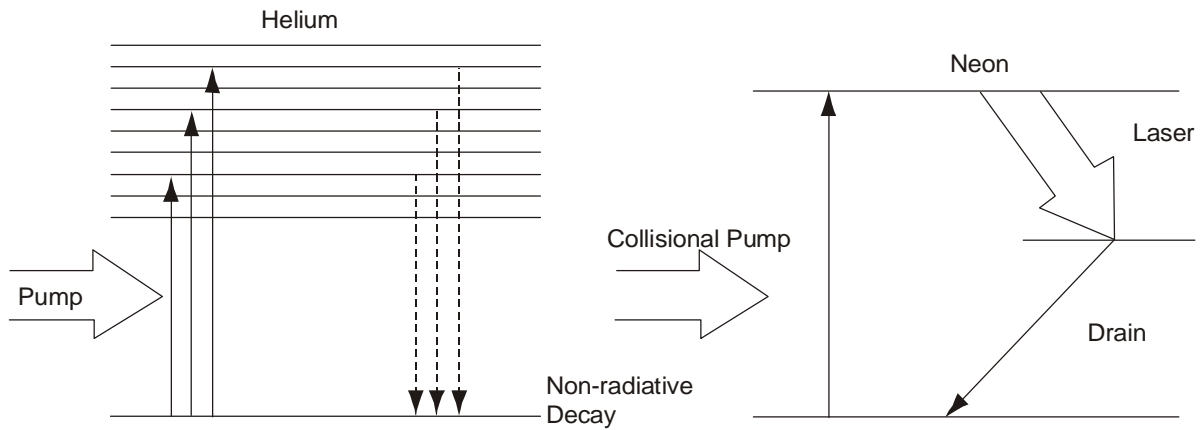
8. **The Metastable State.** From these higher energy levels spontaneous emissions will occur but whereas some will be due to transitions to the ground state, in the case of chromium the majority will decay to an intermediate level known as a metastable state as shown in Fig 3c from which atoms may emit photons at random. Nevertheless, this state, apart from being a preferential level, has the additional feature that atoms tend to remain there for a longer time (by a factor of some 1000s) than they do in any other level other than the ground state. In this way a population inversion is achieved i.e. there are more atoms in the metastable state than in the ground state.

9. **Lasing Action.** Inevitably at some time an atom in the metastable state will make a spontaneous transition to the ground state with the emission of a photon. This photon can now do one of two things; it can either excite a ground state atom into a higher level or it can stimulate an excited atom in the metastable state to make a transition to the ground state. Since a population inversion has been achieved, then on balance it is more likely to stimulate emission than to be absorbed by a ground state atom (Fig 3d). Thus lasing will be initiated. At the end of this process all of the atoms will be back in the ground state ready for further optical pumping to start the cycle again.

10. **Other Techniques.** Optical pumping is not the only means of achieving a population inversion. The helium-neon laser, for example, uses a different method. The medium in this case is a mixture of helium and neon gases of which the neon is responsible for lasing. Energy is input to the helium by means of an electrical discharge and the energized helium atoms transmit their excess energy not by radiation but in collisions with neon atoms. The neon atoms are excited to a high energy level such

that there is a population inversion between this level and an intermediate level rather than with respect to the ground state. Stimulated emission therefore occurs between these two higher levels. This process is illustrated in Fig 4. Atoms in the bottom lasing level eventually decay spontaneously back to the ground state.

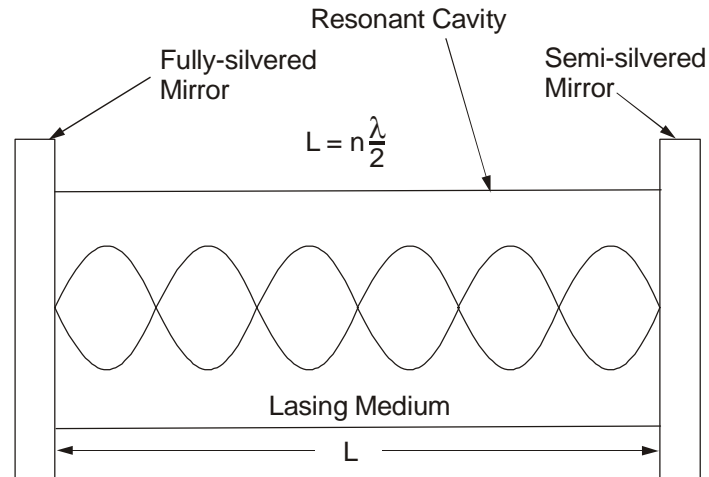
13-24 Fig 4 The Processes in a Helium-Neon Laser



11. The system so far described produces monochromatic and coherent radiation, however it is not very intense and is not emitted as a beam. This is because the stimulating photons are incident upon the atoms from random directions and so the emitted photons follow, likewise, random directions. In addition there will of course be a proportion of random spontaneous emissions. It is therefore necessary to ensure that as many photons as possible are travelling in the required direction. This is achieved by having the lasing medium within an optical resonant cavity.

The Laser

12. The features of the working laser are shown in Fig 5. The optical resonant cavity is achieved by placing mirrors at each end of the lasing medium. These mirrors are separated by an integral number of $\frac{1}{2}$ -wavelengths of the laser radiation and are accurately aligned perpendicular to the laser axis. One of the mirrors is semi-transparent. Photons travelling normal to the mirrors will be reflected backwards and forwards through the cavity and in the process will stimulate further emissions which will radiate in the same direction. The $n \times \frac{1}{2}$ -wavelength nature of the mirror separation ensures that the radiation stays in phase. Off axis radiation will soon be lost to the system through the side walls allowing the axial radiation to increase rapidly in relation to the non-axial radiation. The semi-silvered mirror allows the highly directional beam to leave the cavity.

13-24 Fig 5 Laser Schematic

13. **Q-switching.** A typical ruby laser as described will have a nominal output power of several kW and a pulse length in the order of a millisecond. For many applications it would be beneficial to increase the power by reducing the pulse length. The technique used to achieve this is known as Q-switching. Between the lasing medium and the fully silvered mirror is a glass cell containing a green dye. Although the lasing action starts once the pumping commences, the green dye absorbs the red laser light preventing the build up of energy in the resonant cavity. In doing so the molecules in the dye are raised to an excited state. The concentration of the dye is arranged so that the dye molecules are all excited coincidentally with the maximum number of atoms of chromium being in the metastable level. At this point the dye becomes transparent to the laser wavelength and there is then a very rapid build up of lasing action. The pulse of laser radiation is delivered in about 10 nanoseconds, before the dye molecules return to the ground state and shut off the laser. The output power can be increased to the order of hundreds of mW by this technique.

CHAPTER 25 - THE NATURE OF SOUND

Introduction

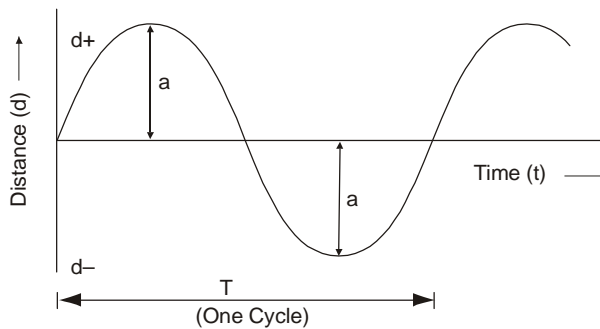
1. **A Definition of Sound.** Sound is the name given to the sensation perceived by the human ear. Every sound is produced by the vibration of the object from which it originates; thus, any explanation of the nature of sound must include a discussion of vibratory motion. The audible sound spectrum is generally acknowledged to cover the frequency range from 15 to 20,000 hertz.

2. **Simple Harmonic Motion.** The simplest form of vibratory motion can be represented by the oscillation of a pendulum bob swinging through a small angle. If the displacement of the bob from the central position is plotted on a graph against time, the variation of the displacement with time gives rise to a sine curve as shown in Fig 1. This motion is called simple harmonic motion, and could equally describe the vibration of a tuning fork. The displacement of the bob can be described by the equation:

$$d = a \sin \frac{2\pi t}{T} \dots\dots\dots(1)$$

where a and T are constants as shown in Fig 1.

13-25 Fig 1 Simple Harmonic Motion



3. **Frequency, Amplitude and Period.** In Fig 1, the cycle represents one complete swing of the pendulum. The frequency of the oscillation is defined as the number of cycles per second (1 cycle per second = 1 hertz (Hz)). During one cycle, the bob twice attains maximum deflection from the central position. The maximum distance displaced is called the amplitude, and in Fig 1 this is shown by the constant a. The period of a vibration is the time it takes to complete one cycle. The frequency can be expressed as:

$$f = \frac{1}{T} \dots\dots\dots(2)$$

where f is the frequency and T is the period in seconds.

4. **Fourier’s Theorem.** The vibration of a tuning fork is the nearest audible equivalent to the oscillation of a pendulum. It can be shown that any vibratory motion which repeats itself regularly can be represented as the resultant or combination of simple harmonic frequencies of suitably chosen amplitudes. These frequencies must also be integral multiples of the frequency with which the motion repeats itself (Fourier’s Theorem). Hence, equation (1) is the basis for describing the vibrations of all sounding objects.

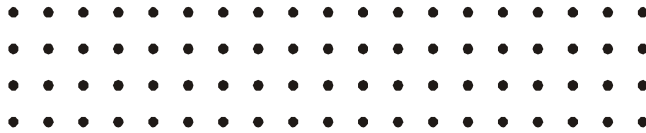
5. **Medium of Transmission.** If a sound source is placed in an airtight chamber which is slowly evacuated of air, the sound will gradually die away as the vacuum increases. Eventually the sound will

cease, although it can be seen that the source is still vibrating. Such an experiment can be used to demonstrate that a material medium such as air, water, wood, glass, or metal is required for the sound to be transmitted.

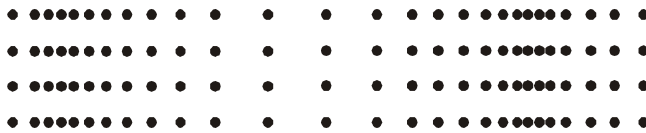
The Propagation of Sound

6. **Transverse Waves and Longitudinal Waves.** The wave motion observed when the surface of a pond is disturbed is called transverse wave motion, because the particles of water oscillate at right angles to the direction of propagation of the waves. Sound waves, however, are propagated as longitudinal waves. In this kind of wave motion the particles oscillate, each about a fixed point, in the direction of propagation of the waves. In Fig 2, the undisturbed particles of a medium are represented by equally spaced dots. A similar set of particles is shown in Fig 3 being disturbed by the passage of a sound wave through the medium. Each particle is displaced to the right and left of its undisturbed position as the wave passes through the medium. If the displacement of a single particle is plotted on the vertical axis of a time graph the familiar sine wave form of simple harmonic motion is produced as shown in Fig 4.

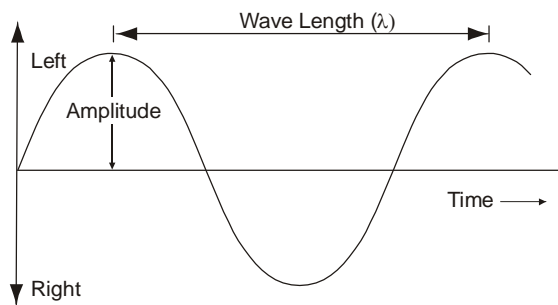
13-25 Fig 2 Undisturbed Particles in a Medium



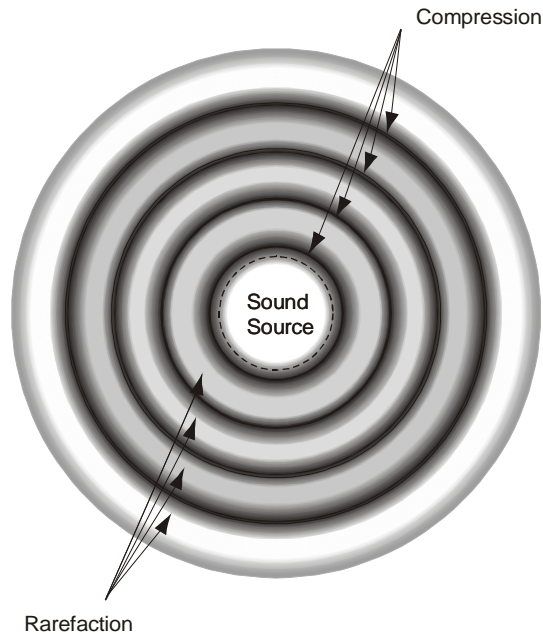
13-25 Fig 3 Passage of a Sound Wave Through a Medium



13-25 Fig 4 Longitudinal Wave Plotted in Graphical Form



7. **Pressure Variations.** When the particles of a medium are displaced by the passage of a sound wave, there is a consequent local variation in pressure. It is these small changes in pressure which actuate the human ear and mechanical devices such as microphones. Fig 5 shows the pressure variations that accompany the passage of a sound wave; they consist of alternate compressions and rarefactions.

13-25 Fig 5 Pressure Variation in a Medium

The Properties of Sound Waves

8. **Reflection.** Like light waves, sound waves are reflected from a plane surface such that the angle of reflection is equal to the angle of incidence. It can also be shown that sound waves come to a focus when they are incident on a concave reflector.

9. **Reverberation.** If sound is generated within a large enclosed space it can be heard directly from the source and indirectly from reflected and diffused (multiple reflection) sound waves. The indirect sounds are called reverberations and continue for a finite time after the sound source has been silenced.

10. **Refraction.** Sound waves travel faster in warm air than in cold air and are therefore refracted when there is a temperature gradient. Refraction also occurs in water and other media because of changes in the velocity of sound.

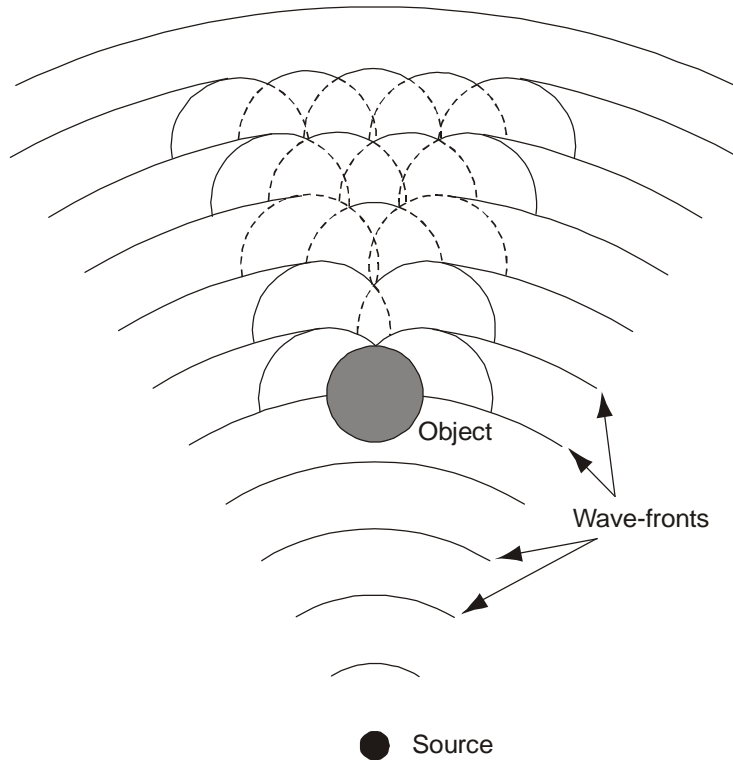
11. **Interference.** If two sound sources are of the same frequency and intensity, and are initially in phase (i.e. coherent) they will interfere with each other and will cancel or reinforce according to the path difference. If the path difference is an odd number of half wavelengths, cancellation occurs and no sound is heard. If it is an even number of half wavelengths, the sounds will reinforce and a louder sound is heard.

12. **The Wave-front.** If a single pulse of noise, such as an explosion, occurs in a medium, the paths followed by the sound waves can be traced by placing a number of recording microphones in the vicinity and noting the time taken for the sound to reach each microphone. If the microphones are located at specified ranges from the source, the points in space reached simultaneously by the sound can be plotted. These points are considered to lie on a surface called the wave-front, and, in a homogeneous medium, the direction of propagation is perpendicular to this surface. It can also be observed that the sound is propagated outwards at a constant velocity.

13. **Diffraction.** It is a common experience that it is possible to hear sound even when the source is behind an obstruction. This 'bending' of sound waves (or indeed any other type of wave) around such

an obstacle is known as diffraction. Although the mathematical treatment is rather complex, a satisfactory explanation of the phenomenon can be made using Huygens' principle. Huygens' principle states that all points on a wave front can be considered as point sources from which secondary wavelets are generated. After a time interval, a new position of the wave-front will be established as the surface of tangency to these secondary wavelets. The way that this principle accounts for the 'bending' of sound around an obstacle is shown in Fig 6.

13-25 Fig 6 Huygens' Principle – 'Bending' Sound around an Obstacle



The Velocity of Sound

14. Since the pressure variations produced by a sound wave are so rapid that no transfer of heat can occur, the process is considered to be adiabatic. The velocity of sound (*c*) in a gas is found to be given by:

$$c = \sqrt{\frac{\gamma p}{\rho}} \dots\dots\dots(3)$$

where γ is the ratio of the specific heat at constant temperature to that at constant volume, p is the pressure, and ρ is the density.

15. Since $\frac{p}{\rho} = RT$ in an ideal gas, where R is the specific gas constant and T is the temperature in K, equation (3) can be rewritten as:

$$c = \sqrt{\gamma RT} \dots\dots\dots(4)$$

Therefore in a given gas, since γ and R are constants, $c \propto R\sqrt{T}$ within the range in which the gas obeys the ideal gas equation. A working expression for the speed of sound in air at a temperature of t °C is given by the equation:

$$c_t = (330 + 0.61t)ms^{-1}$$

In equation (3) both γ and $\frac{p}{\rho}$ are constant for a given gas at a specific temperature, and from this it can be deduced that the velocity of sound in air is independent of pressure.

16. The velocity of sound in water is covered in Volume 13, Chapter 28.

The Intensity of Sound

17. The intensity of sound at any place is defined as the rate of flow of energy across unit area perpendicular to the direction of propagation. If a sound source is emitting J joules of energy per second uniformly in all directions it can be calculated that the energy passing through unit area is proportional to the inverse square of the distance from the source. The intensity of sound is further attenuated by the absorption of energy by the medium through which it is propagated.

The Doppler Effect

18. The Doppler effect occurs when there is a relative velocity between the sound source and the observer.

13-25 Fig 7 The Doppler Effect

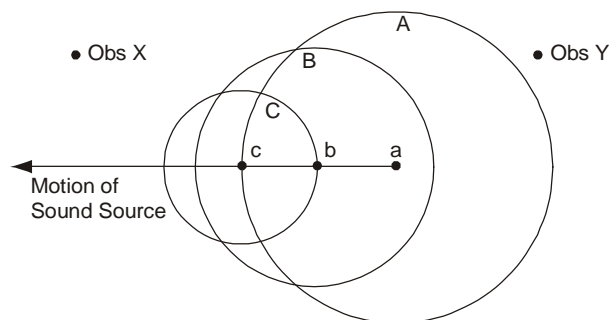


Fig 7 shows how the change of wavelength and hence frequency occurs when a source of sound is moving either towards or away from the observer. The circles represented by A, B, C etc correspond to successive wave-fronts generated at a, b, c etc by the moving source. It is clear that to the observer at X, passage of the wave-fronts will be more frequent (ie the observer will hear a higher pitched sound than was generated), while to the observer at Y, passage of the wave fronts will be less frequent (i.e. the observer will hear a lower pitched sound).

19. The velocity of sound in air of uniform temperature is constant, irrespective of any movement of the source or the observer. However, any movement of the source will alter the wavelength of a sound in air and hence change the pitch of the sound heard by the observer. If the component of velocity of the source towards the observer is V_s , then the frequency, f' , of the note heard by the observer is given by:

$$f' = f_0 \cdot \frac{c}{c - V_s}$$

where f_0 = frequency of the note if heard from a stationary source, and c = velocity of sound in air. Any movement of the observer alters the velocity of the sound relative to the observer, and this also results in a change of pitch. If V_0 is the component of velocity of the observer towards the source, then the frequency, f' , of the note heard by the observer is given by:

$$f' = f_0 \cdot \frac{c + V_0}{c}$$

If both the source and the observer are moving then the frequency, f'' , of the note heard by the observer is given by:

$$f'' = f_0 \cdot \frac{c + V_0}{c - V_s}$$

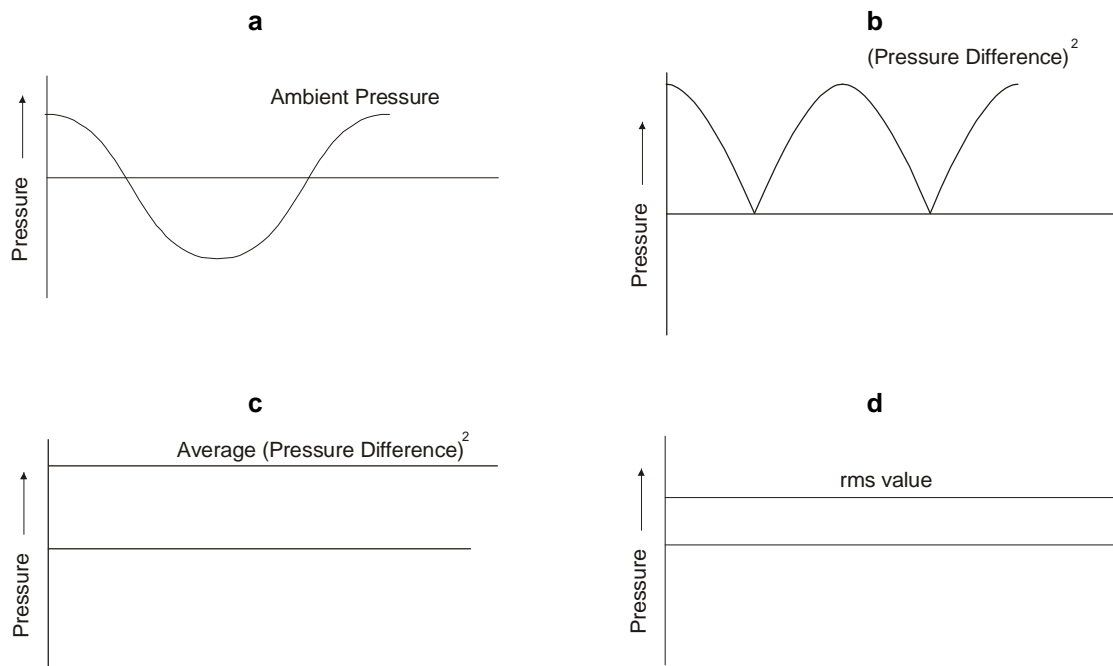
CHAPTER 26 - ACOUSTIC MEASUREMENT

Introduction

1. Sound consists of small variations in pressure above and below the ambient pressure with respect to time. It may be depicted diagrammatically as a sine curve reflecting the gradual increase in pressure to a maximum and its subsequent decrease to a minimum. The maximum excursion of pressure difference from the ambient level is called the amplitude, but for practical purposes some kind of average value over time is required that gives equal weighting to both the rarefaction and compression phases. This 'average' value is known as the root-mean-square (RMS) value, and its derivation is described below.

2. **Root-mean-square Value.** Fig 1a shows a graph of the sinusoidal sound pattern with amplitude on the y-axis. The first stage, (Fig 1b), is to square all the values of amplitude which has the effect of making the negative values positive. The squared values are then averaged, (Fig 1c), to produce a level value. Finally, the square root of the average is taken (Fig 1d). As with all other pressures, the units are newtons per square metre (Nm^{-2}) or Pascals (Pa).

13-26 Fig 1 Root-mean-square Derivation Sound Intensity



Sound Intensity

3. Sound intensity is a measure of the power transmitted per unit area, the area being at right angles to the direction in which the sound is propagating. For unimpeded sound, away from the source, the intensity is proportional to the square of the pressure i.e.:

$$I = kp^2$$

where k is a constant determined by the medium through which the sound is travelling (e.g. for air at atmospheric pressure and $20\text{ }^\circ\text{C}$, $k = 1/410$). The term 'sound intensity' is used in some reference books, but it is not used as a measure in underwater acoustics.

Sound Pressure Level

4. In underwater acoustics, measurements are always in terms of sound pressure level (SPL). Sound pressure levels cover a very wide range of values, for example the range from the threshold of hearing to the onset of pain is from 2×10^{-5} Pa to 20 Pa. Using a linear scale over this range would be cumbersome, and so a logarithmic (decibel) scale is used relative to a datum pressure level. The datum pressure that is chosen for sound in air is 2×10^{-5} Pa, which is equivalent to the lowest sound pressure at 1,000 Hz, detectable on average by people with normal hearing. In underwater measurement, the datum pressure level is the micropascal, (1×10^{-6} Pa).

5. As the energy contained in a wave is proportional to the amplitude squared, and if p is the sound pressure to be measured, then if the transmission was to take place in water:

$$\begin{aligned} \text{SPL} &= 10 \log \frac{p^2}{(1 \times 10^{-6})^2} \text{ dB} \\ &= 20 \log \frac{p}{1 \times 10^{-6}} \text{ dB} \end{aligned}$$

6. Note that a significant change in sound pressure level is represented by only a small change in decibel notation. If the pressure from a noise source is doubled, this equates to $10 \log 2$ in dB, ie 3 dB. Thus, if the noise level from, say, a jet aircraft with four engines running was 140 dB, the same aircraft running on two engines would generate a level of 137 dB. A one or two dB difference in the measurement of acoustic pressure may therefore be significant in target detection.

CHAPTER 27 - OCEANOGRAPHY

Introduction

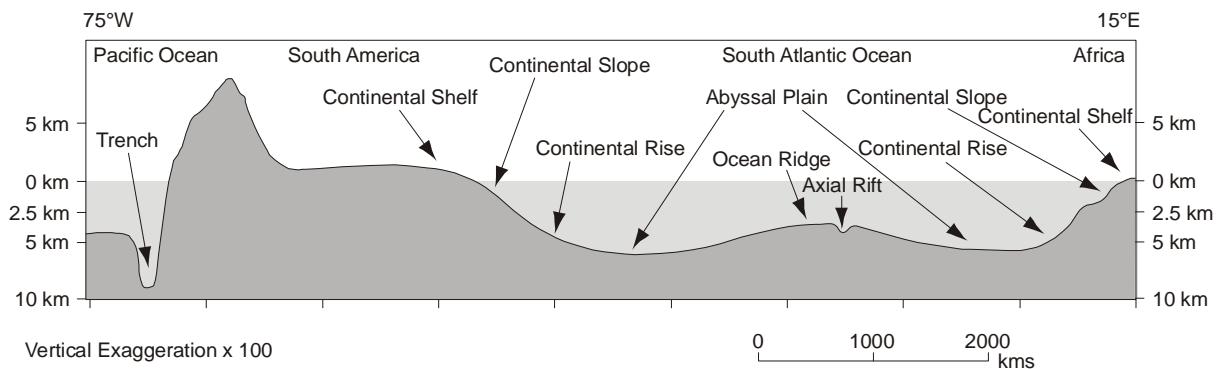
1. Maritime Patrol Aircraft are tasked with the location of submarines using acoustic sensor equipment. The aim of this chapter is to outline those aspects of the oceanic environment which have a bearing on the transmission of sound. Details of how acoustic systems are affected by this environment are covered in Volume 13, Chapter 28.

GEOLOGY AND STRUCTURE

General

2. The oceans are not simply depressions in the Earth's surface filled with water, rather they have a fairly complex structure. Fig 1 shows a cross section of the Earth's crust between Africa and South America and just into the eastern Pacific Ocean.

13-27 Fig 1 Cross-section of the Earth's Crust between Africa and South America



3. In the centre of the Atlantic Ocean is what amounts to a mountain chain - the Mid-Atlantic Ridge. This is part of a ridge system which extends from North of Iceland southwards into the South Atlantic, eastwards into the Indian Ocean, then northwards again in the East Pacific before disappearing under the North American continent. Its total length is over 50,000 km and its width between 1,000 and 4,000 km. The ridge generally rises to about 2 km, occasionally 5 km, above the flanking plain, and the crest is normally about 2½ km below sea level. The Mid-Atlantic Ridge has an axial rift and it is here that new ocean crust is created. This crust formation leads to the Atlantic widening at a rate of 2 to 4 cm/year.

4. Adjacent to, and either side of, the ridge lie the abyssal plains which are in general relatively flat areas. This flatness is primarily due to a thick layer of sediment overlaying the rough topography which has been generated at the ridge. In places this flatness is punctuated by abyssal hills and seamounts. The majority of these are ocean floor volcanoes and most do not rise above sea level; where they do, they form oceanic islands. The Pacific Ocean is particularly well endowed with seamounts some of which are very large. Mauna Loa for example is about 100 km across at its base and rises to about 9 km above the plain, ie about as high as Mt. Everest. Many of these seamounts have flat tops and are known as guyots. It is believed that this flat-topped effect was caused by wave erosion at a time when the volcano was at or above sea level; with the passage of time, they have subsided below the surface.

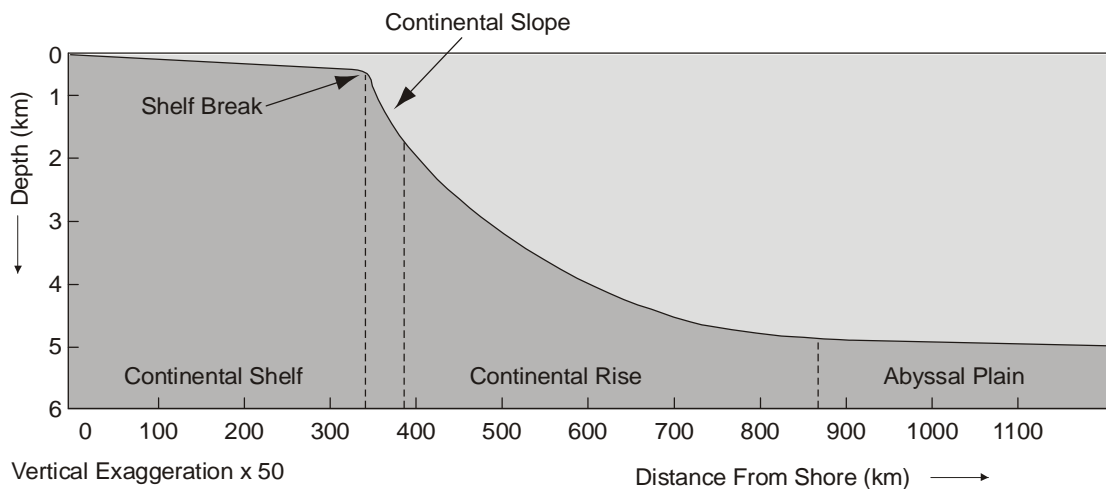
5. The margins of the Atlantic and Pacific Oceans are rather different. The Pacific margin is characterized by seismic and volcanic activity whereas the Atlantic margin is aseismic. This is because the Pacific margin coincides with the edge of a crustal plate, whereas in the case of the

Atlantic the ocean and the adjacent continents are on the same plate. This difference gives rise to characteristically different ocean floor topography.

Atlantic Margin

6. Fig 2 shows the typical shape of an Atlantic style ocean margin. The continental shelf is geologically part of the continent rather than the ocean; the underlying rock type is not oceanic. The shelf is normally coated with sediments derived from the adjacent land. The width of the shelf can extend to as much as 1,500 km although for the most part, it is less than this, typically a few hundred kilometres. The surface is generally flat with an average gradient of only 0.1° . At the shelf break the water depth varies between 20 and 500 m with an average of around 130 m.

13-27 Fig 2 Typical Topography of an Atlantic Style Ocean Margin



7. **Continental Slope.** The continental slope marks a sharp increase in gradient to about 4° on average. The width of the slope is between 20 and 100 km and the base of the slope lies at a depth of between 1.5 and 3.5 km. Frequently the slope is cut by submarine canyons along which sediment is transported to the deep ocean. These canyons, which are somewhat similar to V-shaped river valleys on land, often start on the continental shelf and commonly coincide with the mouths of major rivers. The end of the continental slope marks the end of the continental crust and the beginning of the oceanic crust.

8. **Continental Rise.** The continental rise has a much gentler gradient than the slope, about 1° , and is built up of the sediments which have flowed down the slope. Typically, the continental rise extends for about 500 km and reaches a depth of some 4 km before merging into the abyssal plain.

Pacific Margin

9. The most significant difference between an Atlantic and a Pacific type margin is the presence of a very deep trench in the latter at the outer edge of the continental slope. In general, the shelf is narrower in the Pacific, about 50 km. The shelf break tends to be more abrupt and the slope often has a steeper gradient, up to about 10° in places. The continental rise is missing and is replaced by a trench which marks the site where the oceanic crust is being subducted beneath the continental crust. It is this subduction which gives rise to seismic and volcanic activity. The trenches can reach depths of over 11 km, and a depth of 8 km would not be untypical. In places the trench may be partly filled with land derived sediments.

SEAWATER

Physical Properties

10. **Salinity.** Salinity is a measure of the dissolved solids in seawater. It is usually expressed in values of parts per thousand by weight (i.e. gm kg^{-1}) and the symbol o/oo is used. Surface salinity represents a balance between an increase due to evaporation and freezing on the one hand, and a decrease due to precipitation, ice melting, and river influx on the other hand. Despite these effects, salinity values do not vary greatly; between 30 and 40 o/oo at the extremes. Minimum values occur in coastal regions with large river discharges and the maximum values occur in the Red Sea and Persian Gulf. In the open ocean surface salinity values are even more conservative, varying between about 33 and 37 o/oo. The maximum values tend to occur around the tropics, where the effect of evaporation is at its highest. In low and middle latitudes salinity decreases with depth in the first 600 to 1000 m, a zone known as a halocline, below which it becomes virtually constant at 34.5 to 35 o/oo.

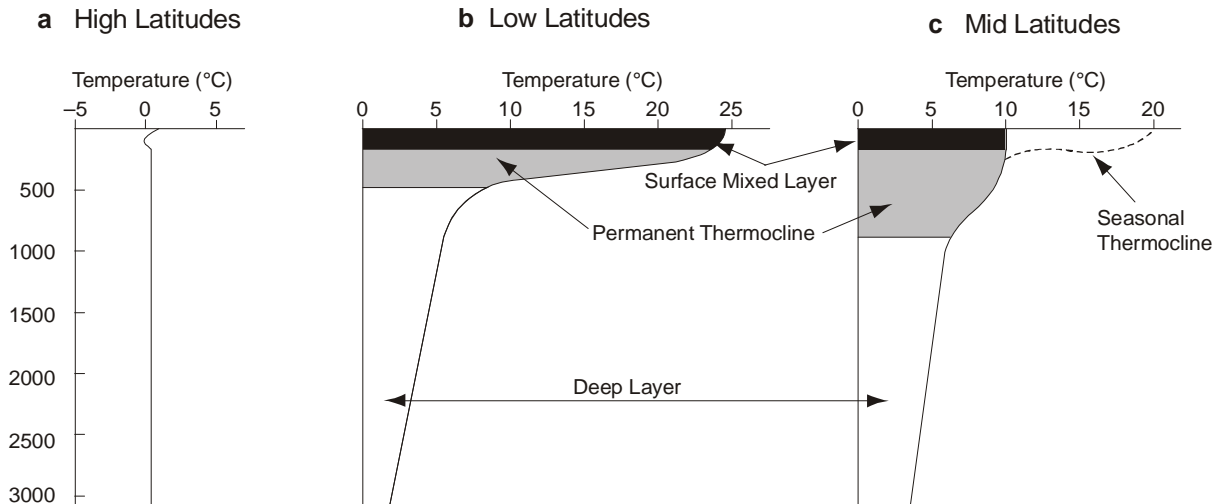
11. **Pressure.** Assuming constant density and a constant value for g (the acceleration due to gravity), then pressure varies virtually linearly with depth. For 99% of the oceans, density remains within $\pm 2\%$ of its mean value, and variations in g are very much smaller than this so the linear relationship is a reasonable model. The change is approximately 10^5 Nm^{-2} (1 bar or 1 atmosphere) per 10 m (33 ft).

12. **Temperature.** The only agent for heating the oceans is incoming solar energy and the majority of this is absorbed within metres of the ocean surface. All of the infra-red radiation is absorbed within one metre and only about 2% of the incident energy reaches 100 m. A small amount of heat is transmitted to depth by conduction but mixing caused by wind and waves is the main mechanism for the transfer of heat. This turbulence generates a mixed surface-layer which may be up to 200 m thick depending upon surface conditions, and therefore upon the season.

13. **Permanent Thermocline.** Below this mixed layer, and down to about 1,000 m, the temperature falls rapidly. This region is known as the permanent thermocline. Underneath this, seasonal variations are virtually non-existent and there is a much shallower temperature gradient with temperature falling to between 0.5 °C and 1.5 °C. This 3-layer model of the ocean temperature structure is illustrated in Fig 3 showing the variations with latitude.

14. **Season and Latitude Effects.** In high latitudes (above 60° N) the permanent thermocline is missing. In mid-latitudes the 'classic' profile may be amended by a seasonal thermocline when surface waters are heated in spring and summer (Fig 3).

13-27 Fig 3 Depth/Temperature Profiles for Different Latitudes

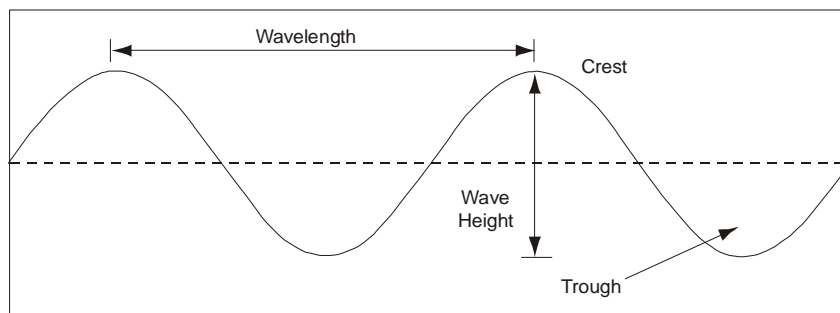


15. **Diurnal Effects.** Diurnal variations in temperature are insignificantly small, usually less than 0.3 °C in the open oceans although perhaps up to 3 °C in shallow coastal waters.

Waves

16. **Formation.** Waves are another manifestation of the transfer of wind energy to the surface water. The precise mechanism of this transfer is complex and, as yet, poorly understood. In the area of formation, waves that are generated will consist of a superimposition of waveforms of varying frequency and amplitude depending upon the wind speed and the fetch (the length of sea area affected). Higher wind speeds lead to higher waves, but eventually a state of equilibrium will be reached with the excess energy being dissipated in, for example, white capping. The influence of waves is only felt close to the surface and waves generated by a severe storm would be virtually unnoticeable below about 100 m. The characteristics of waves are illustrated in Fig 4.

13-27 Fig 4 Characteristics of an Ocean Wave



Wave height - the linear distance between crest and trough
 Wave length - the linear horizontal distance between corresponding points on consecutive waves
 Wave period - the length of time for one complete cycle to pass a fixed point

17. **Swell.** Moving away from the area of generation the short period waves are dissipated, and only the long period waves remain. These are known as swell and can travel large distances from the generating area, typically thousands of kilometres.

18. **Sea Bed Effects.** When waves enter shallow water, they are affected by the seabed once the depth is less than ½ the wavelength. The front of the wave becomes steeper and eventually breaks.

19. **Beaufort Scale.** The relationship between wind speed and sea state is expressed in the Beaufort Scale which may be used for estimating wind speed on land and at sea. It should be noted that, at sea, the scale is valid only for waves generated locally, and providing adequate time has elapsed and

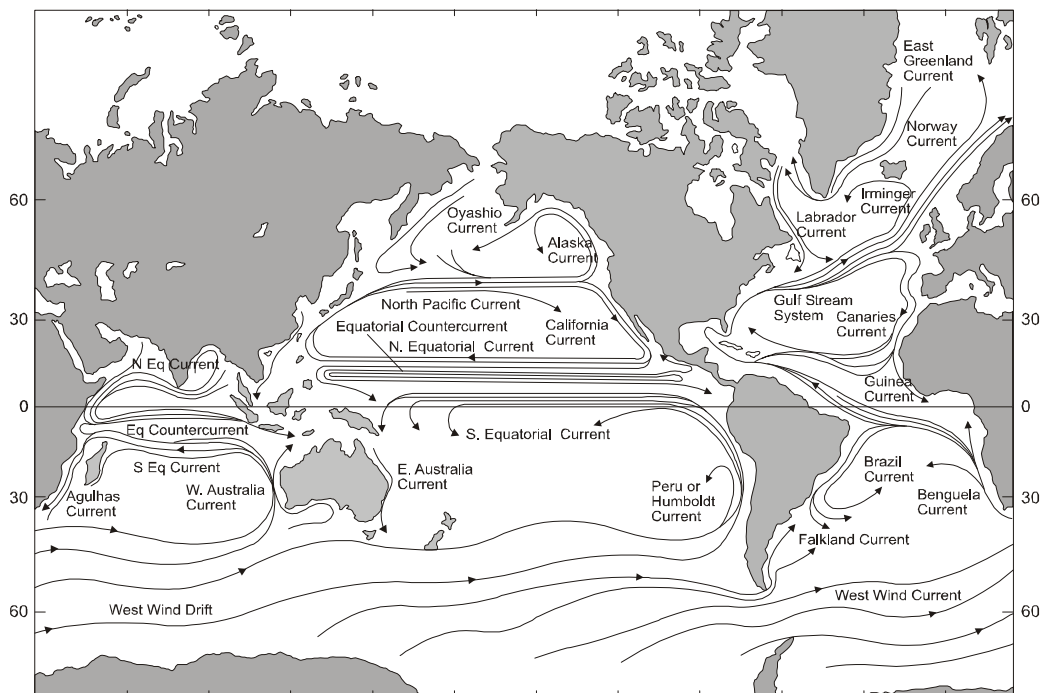
there has been an adequate fetch for a fully developed sea to become established. The Beaufort Scale is explained in Table 1 at the end of the Chapter.

Currents

20. **Deep Currents.** As has been seen, sea water is not homogenous - its temperature may vary considerably, and there may be minor variations in salinity. In the Antarctic and around Greenland large masses of water are generated which are cold due to the interaction with the atmosphere, saline due to the freezing-out of a proportion of the fresh water, and therefore somewhat denser than average. This cold, dense water sinks and moves towards and across the Equator at depth. Clearly there will be boundaries between this water and the other water masses that it encounters with different temperature and salinity characteristics. Similar boundaries exist between other water masses with different properties. Examples are to be found where water from the Baltic enters the North Sea and where Mediterranean water enters the Atlantic Ocean. These boundaries are known as fronts and are analogous to meteorological fronts where different air masses meet.

21. **Surface Currents.** Sea currents are generated by winds when some of the energy of air movement is transferred by friction to the ocean surface. This leads to generalized zonal currents similar to the idealized coriolis-induced global patterns of surface winds illustrated in meteorological texts, and although this simple model is considerably modified by the presence of land masses, the influence of these winds is clearly discernable in Fig 5. As at depth there will be fronts between water masses of differing characteristics, reflecting their different origins and histories. Whereas some of these fronts will be more or less permanent, others will be of a temporary nature primarily due to seasonal effects.

13-27 Fig 5 Generalized Surface Currents



The Biological Environment

22. **Life Forms.** Ninety-eight percent of marine species belong to the benthic, or seabed-living system. But, it is the remaining 2% of species, the pelagic or free-swimming system, which is of concern here. The life forms can be divided into two main groups: the plankton and the nekton (there is a third small group, the pleuston, which includes the jellyfish, which live at the sea surface, but they are of no relevance here).

23. **The Plankton.** The plankton are the group of organisms, both plant and animal, whose powers of locomotion are insufficient to prevent them being transported by currents. They range in size from less than 2 μm up to just over 2 cms. As a large proportion of the plankton (the phytoplankton) is responsible for photosynthesis they must exist in that part of the ocean to which light can penetrate, typically the top 100 m or so. Virtually all plankton exhibit a diurnal variation of depth, rising near the surface at night and descending again just before sunrise.

24. **The Nekton.** The nekton are those creatures which can swim with sufficient power to be more or less independent of water movement. This group includes fish, cephalopods, and whales in the open ocean. Often, they exhibit the same diurnal variation in depth as the plankton on which they feed.

25. **The Deep Scattering Layer.** The diurnal depth variation of most plankton and nekton has already been described and this event occurs within the top 100 m or so of the ocean. However, it is worth noting that there are some mainly carnivorous plankton and nekton which migrate between much greater depths. Typically, they will spend the day at between 400 and 1000 m and the night at about 200 to 300 m. The significance of this group of organisms is that they can affect echo sounders and SONAR equipment. It is thought that the reflection of sound waves is caused primarily by small fish possessing air bladders and by prawn-like creatures. This phenomenon is known as the deep scattering layer and can be observed in all oceans.

Table 1 Sea State and Beaufort Wind Scale

Sea State	Beaufort No	Name	Wind Speed		Characteristics	Wave Height (m)	Wind Gusts
			knots	ms ⁻¹			
0	0	Calm	1	0.0-0.2	Sea like mirror. Complete calm. Smoke rises vertically.	0	-
	1	Light Air	1-3	0.3-1.5	Ripples with appearance of scales, no foam crests. Wind direction shown by smoke drifts but not by wind vanes.	0.1-0.2	-
1	2	Light Breeze	4-6	1.6-3.3	Small wavelets, crests have glassy appearance but do not break. Wind felt in face, leaves rustle, ordinary vanes moved by wind.	0.3-0.5	-
2	3	Gentle Breeze	7-10	3.4-5.4	Large wavelets, crests begin to form, scattered white horses. Leaves and small twigs in constant motion, wind extends light flag.	0.6-1.0	-
3	4	Moderate Breeze	11-16	5.5-7.9	Small waves becoming longer, fairly frequent white horses. Wind raises dust and loose paper; small branches are moved.	1.5	-
4	5	Fresh Breeze	17-21	8.0-10.7	At sea, moderate waves assume longer form, many white horses and chance of spray, crested wavelets on inland lakes. Small trees in leaf begin to sway.	2.0	-
5	6	Strong Breeze	22-27	10.8-13.8	Large waves, extensive white foam crests, spray probable. Large branches in motion, umbrellas used with difficulty, whistling in telegraph wires.	3.5	-
6	7	Near Gale	28-33	13.9-17.1	Sea heaps up, white foam begins to be blown in streaks. All trees in motion, inconvenience felt when walking against wind.	5.0	-
7	8	Gale	34-40	17.2-20.7	Moderately high waves of greater length, crests break into spindrift, foam well blown in streaks. Twigs break off, wind impedes progress.	7.5	43-51
8	9	Strong or Severe Gale	41-47	20.8-24.4	High waves, dense streaks of foam, spray may affect visibility, sea begins to roll. Slight structural damage occurs (chimney pots and slates removed).	9.5	52-60
9	10	Storm	48-55	24.5-28.4	Very high waves, overhanging crests, sea surface white, heavy rolling and visibility reduced. Trees uprooted, considerable structural damage occurs.	12.0	61-68

	11	Violent Storm	56-64	28.5-32.7	Exceptionally high waves, small and medium sized ships lost to view in troughs. Widespread wind damage on land.	15.0	69-77
	12	Hurricane	64+	32.7+	Air filled with foam and spray, sea completely white with driving spray. Visibility greatly reduced. Structural damage to disaster level.	15+	78+

CHAPTER 28 - SOUND IN THE SEA

Introduction

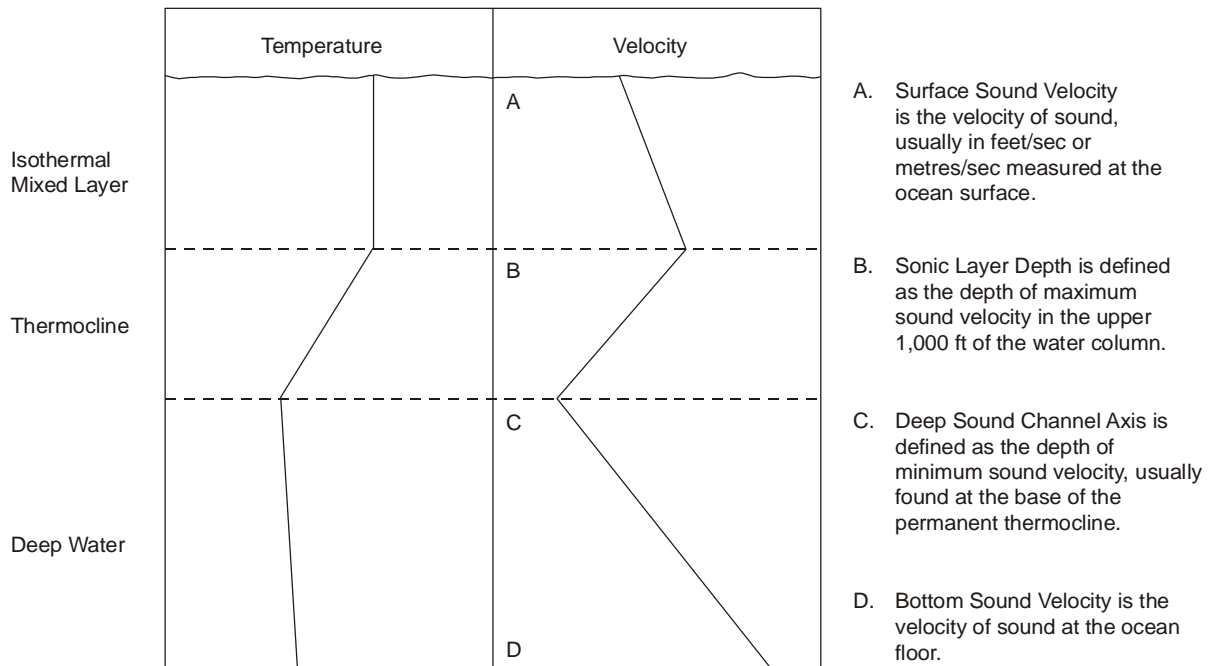
1. Due to the rapid attenuation of other energy forms in the sea, e.g. light, the use of sound is one of the few methods available for underwater detection. A study of the behaviour of sound in the sea is therefore important for operators of maritime patrol aircraft.
2. Unlike light, which is a form of electromagnetic energy, sound involves the vibration of the medium through which it is transmitted. So, whereas light can pass through a vacuum, sound cannot, generally travelling best through solids and liquids and somewhat less well through gases.
3. The wavelengths of sound which are relevant in the ocean range from about 50 metres to 1 millimetre. Assuming a velocity of sound in sea water of $1,500 \text{ ms}^{-1}$, this corresponds to frequencies between 30 Hz and 1.5 MHz. Variations in the velocity of sound in sea water and their effect on propagation are in fact very important and will be examined in some detail.

The Velocity of Sound in the Sea

4. In Volume 13, Chapter 25 a simple equation relating the velocity of sound in air to specific heat, pressure and density was given. Unfortunately, this simple relationship does not apply in liquids and although a mathematical treatment can be derived, it is rather complex. However, it is possible to deal empirically with the variation of sound velocity with salinity, pressure and temperature. An increase in any one factor will increase the sound velocity if all other factors remain constant. In line with current practice, changes are shown in units of feet, pounds per square inch (psi) and $^{\circ}\text{F}$ rather than in SI units. The velocity of sound in sea water varies between $4,700 \text{ fts}^{-1}$ and $5,100 \text{ fts}^{-1}$.
5. **Salinity.** A 1 o/oo increase in salinity increases the sound velocity by 4.27 fts^{-1} . Unfortunately, there is no method of readily determining salinity from a maritime patrol aircraft but assuming a constant value of 35 o/oo is normally sufficient. However, caution is needed in those areas where salinity values are liable to be markedly different from the average, e.g. in coastal and ice melt areas.
6. **Pressure.** An increase in pressure of 44 psi, corresponding to a depth change of about 100 ft, will increase the sound velocity by about 1.7 to 1.8 fts^{-1} . Fortunately, pressure increases with depth in a linear fashion and so velocity changes are easily predicted with fair accuracy.
7. **Temperature.** Of the 3 factors being considered, changes in temperature have the greatest effect on sound velocity. A $0.5 \text{ }^{\circ}\text{C}$ ($\approx 1 \text{ }^{\circ}\text{F}$) increase in water temperature will cause an increase in sound velocity of between 4 and 8 fts^{-1} .
8. **Sound Velocity Profile.** Fig 1 shows how sound velocity varies with depth in a typical 'Three-layer Ocean'. This variation with depth is known as a sound velocity profile (SVP). The following points can be deduced:
 - a. There is an increasing velocity down through the surface mixed layer due to the effect of increasing pressure (Temperature constant).
 - b. There is decreasing velocity with depth in the thermocline due to decreasing temperature having a greater effect than increasing pressure.

- c. There is increasing velocity down through the deep water due to increasing pressure (temperature constant).

13-28 Fig 1 Sound Velocity Profile for a Typical 'Three-layer Ocean'



9. **Temperature Gradient and Sound Velocity.** Temperature/depth profiles are readily obtained by MPA crews and the following relationship between temperature profile and sound velocity can be expected:

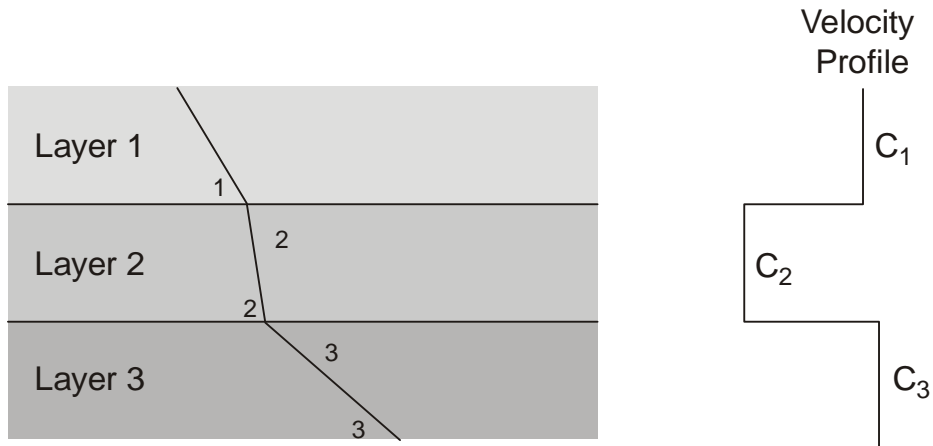
- If temperature decreases with depth at a rate of 0.1 to 0.2 °C/100 ft (0.2 to 0.4 °F/100 ft) its effect will be offset almost exactly by the effect of the increasing pressure with depth. This will result in an ISOVELOCITY condition (i.e. no change in velocity with depth).
- If temperature decreases with depth at a rate of less than 0.1 to 0.2 °C/100 ft (0.2 to 0.4 °F/100 ft) then sound velocity will increase with depth and a POSITIVE velocity gradient will exist.
- If temperature decreases with depth at a rate greater than 0.1 to 0.2 °C/100 ft (0.2 to 0.4 °F/100 ft) then sound velocity will decrease with depth and a NEGATIVE velocity gradient will exist.
- If temperature remains constant with depth, this is termed ISOTHERMAL and there will be velocity gradient of +2 (i.e. an increase in velocity of 2 fts⁻¹ per 100 ft increase in depth).

Refraction

10. When sound leaves a source, it can be considered to move along paths known as rays - analogous to light rays. As with light (see Volume 13, Chapter 21, Para 21), if the medium of transmission is homogenous, then these rays will be straight lines but if the ray moves into an area where the velocity of sound is different then the ray will be bent, a phenomenon known as refraction.

11. The effect of refraction on a ray passing through layers of different velocity is illustrated in Fig 2 where the total ray path will be seen to be made up of a series of straight-line segments.

13-28 Fig 2 Refraction in a Layered Medium



12. A useful mnemonic is the 'HALT' rule which states that rays will be refracted away from depths of relatively high velocity and towards depths of relatively low velocity:

High Away, Low Towards

Thus, in Fig 2 the ray moving from layer 1 to layer 2 (relatively low velocity) will be refracted towards layer 2 i.e. $\theta_2 > \theta_1$. On reaching layer 3 (relatively high velocity) the ray will be refracted away from layer 3 i.e. $\theta_3 < \theta_2$.

13. Applying this rule to the typical SVP in Fig 1, it will be seen that sound rays in the mixed surface layer will be refracted upwards away from the Sonic Layer Depth and sound rays in the thermocline will be refracted downwards until they pass the depth of the Deep Sound Channel Axis when they will be refracted upwards again. The direction of propagation of a sound ray will therefore be determined by the ambient SVP.

14. **Snell's Law.** The degree to which a ray is bent when moving from an area of one velocity to another is determined by Snell's Law which states that:

In a medium comprising discrete layers each of different sound velocity, the angles $\theta_1, \theta_2, \theta_3$ etc of a ray incident on and leaving a boundary between layers are related to the sound velocities, C_1, C_2, C_3 etc of these respective layers such that:

$$\frac{\cos \theta_1}{C_1} = \frac{\cos \theta_2}{C_2} = \frac{\cos \theta_3}{C_3} = \text{a constant, for any given ray.}$$

15. **Limiting Ray.** When sound rays are refracted within a sea layer, one ray will eventually just graze or be tangential to the boundary with an adjacent layer, or to the sea surface or ocean bottom. This ray, which will continue to be refracted within the layer, is known as the limiting ray. Rays which approach the layer boundary at a more perpendicular angle, will pass through it (or be reflected or absorbed).

Shadow Zones

16. The HALT rule predicts that rays will be refracted away from depths of sound velocity maxima. As a result, there often exists at these depths a region into which very little acoustic energy can penetrate; such

regions are termed 'Shadow Zones'. The existence of such a zone, and its relative position, depends upon the SVP. Minor amounts of energy will enter the shadow zone due to diffractive and scattering effects. The nature of shadow zones in a variety of SVPs is shown in Fig 3 a-d.

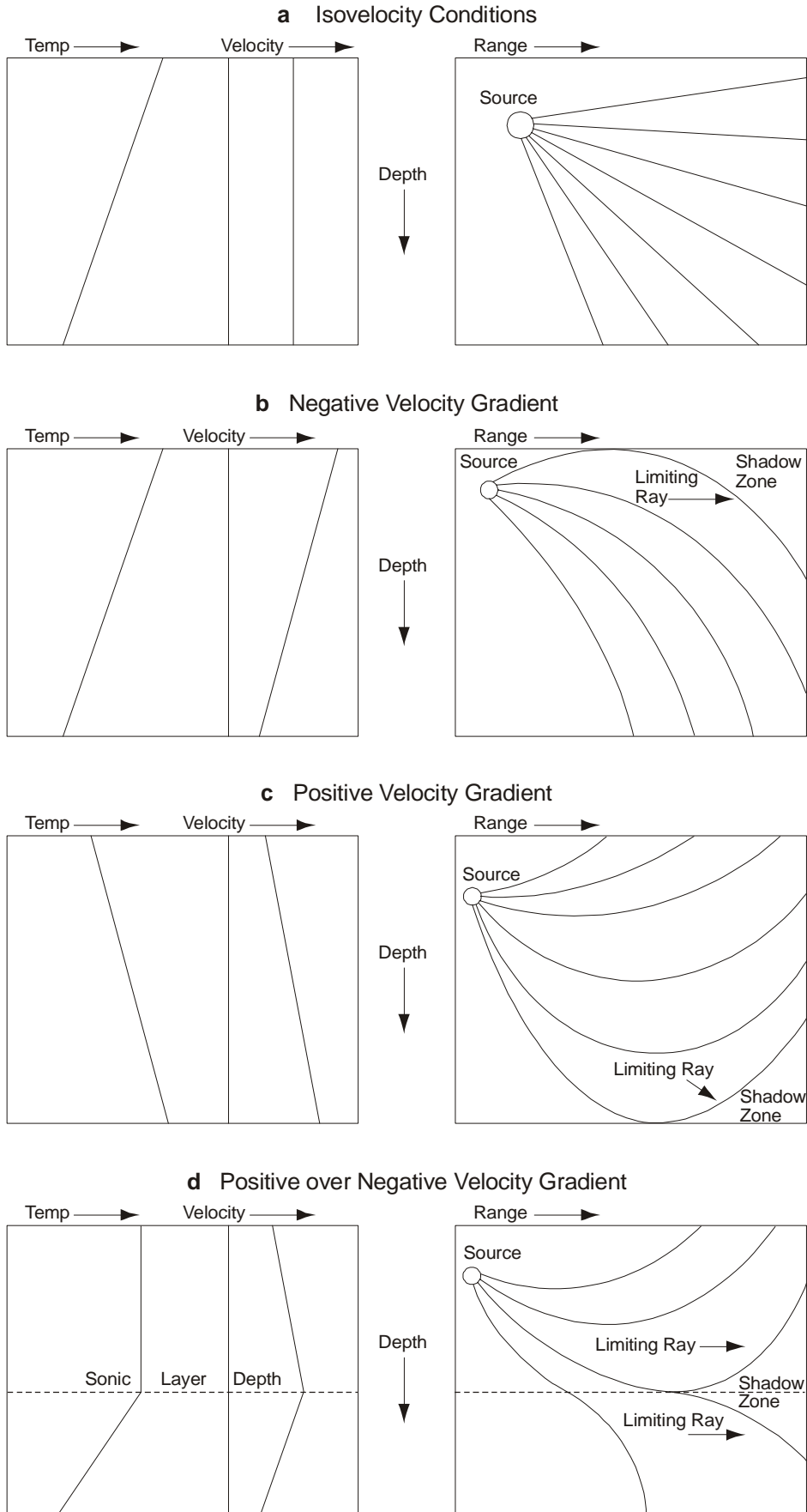
- a. **Isovelocity Conditions (Fig 3a).** In isovelocity conditions, the increase in sound velocity due to increasing pressure with depth is offset by the effect of decreasing temperature with depth. Since there is no variation in velocity there is no refraction and sound rays travel in straight lines from the source.
- b. **Negative Velocity Gradient (Fig 3b).** In this situation, the temperature decreases at a rate sufficient to overcome the pressure effect and so velocity decreases with depth. Sound rays are refracted downwards away from the higher velocity area. A shadow zone exists above and beyond the limiting ray.
- c. **Positive Velocity Gradient (Fig 3c).** In the situation where temperature is constant or increasing with depth, or at least does not fall at a rate to override the pressure effect, then a positive velocity gradient will be established. Sound rays will be refracted upwards away from the higher velocity area and a shadow zone exists below and beyond the limiting ray.
- d. **Positive over Negative Velocity Gradient (Fig 3d).** The combination of a positive velocity gradient above a negative velocity gradient produces a split beam effect at the depth of the velocity gradient change. The limiting ray and all rays above it are refracted upwards away from the higher velocity area. All rays below the limiting ray are similarly refracted downwards. A shadow zone will exist beyond the limiting ray(s).

Sound Transmission Modes (Paths)

17. Clearly sound energy can travel directly underwater from the source to the detector, affected only by 'bending' due to refraction. In general, however, direct rays achieve only relatively short ranges. The range is normally determined by the limiting ray and will depend primarily on the Sonic Layer Depth, the SVP below the surface mixed layer, and the source/receiver depth combination:

- a. **Effect of Sonic Layer Depth.** The deeper the Sonic Layer Depth, the greater will be the Direct Path range. This is because the limiting ray has a greater vertical distance to travel before being refracted upwards and those rays at a greater angle than the critical angle ray will also have a greater vertical distance to travel before passing through the layer.
- b. **Effect of SVP Below the Surface Mixed Layer.** The greater the negative velocity gradient below the surface mixed layer, the more pronounced will be the refraction and so the more reduced will be the direct path range.
- c. **Effect of Receiver/Source Depth Combination.** If the receiver and source are in different layers then this will curtail the direct path range to an extent, particularly if the receiver is shallower than the source. The greatest direct path range is normally achieved when both source and receiver are in an isothermal layer. In general, the direct path range depends upon the velocity gradient.

13-28 Fig 3 Velocity Gradients

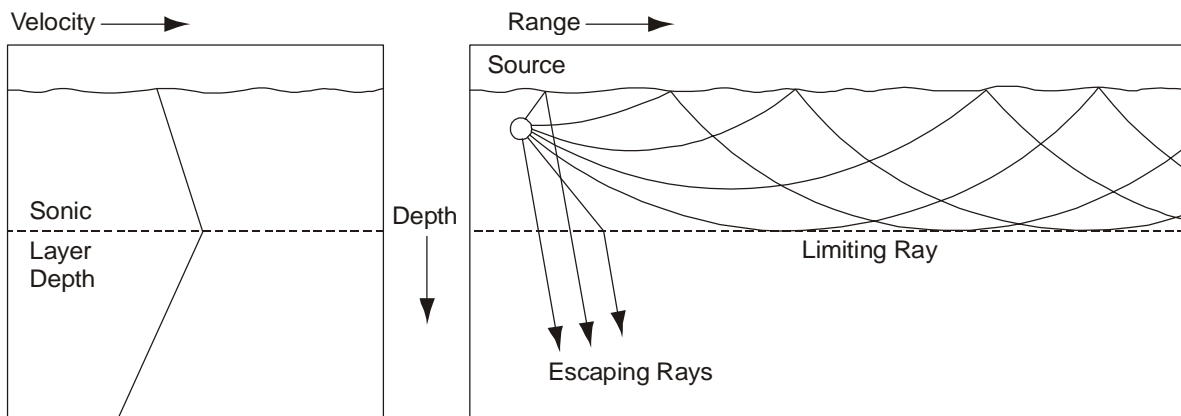


18. **Phenomena.** The variations in sound velocity with depth in the ocean lead to a number of effects which allow much greater sound detection ranges to be achieved than would be possible if the direct path only was available. The infinite variety of sound velocity profiles implies an infinite variety of transmission modes. However, it is possible to group them all into four basic phenomena; other modes tend to be just variations on these four. It should be remembered though, that unless transmitter and receiver are very close together, then the received sound has very probably been transmitted by a combination of more than one mode.

19. **Surface Duct (Surface Sound Channel).** This transmission mode, sometimes known as 'The Mixed Layer Sound Channel', concerns only acoustic energy contained within the surface layer. The prime requisite for the existence of a surface duct is a layer with a positive velocity gradient. Sound rays in the layer are refracted up towards the surface (away from the higher velocity), reflected off the sea surface downwards and are then refracted upwards again and so on. This trapping of the sound energy within the duct can result in extended ranges. The quality of sound transmission via this mode is dependent upon the Sonic Layer depth, the sea state, and the source depth. Whether this mode can be used depends primarily upon whether the sensor can be placed in the layer and also upon the signal frequency. Surface Duct propagation is illustrated in Fig 4, and its usefulness may be affected by the following factors:

- a. **Sonic Layer Depth.** An effective Surface Duct cannot exist unless the Sonic Layer depth is greater than 50 feet. The deeper the Sonic Layer depth the greater the range in the duct as the sound rays have further to travel before being reflected off the surface. Each time a sound ray is reflected off the surface some energy is lost by scattering so for any given range the reflection losses are less for a deep layer.
- b. **Sea State.** Wind, waves and swell all increase absorption, reflection and scattering losses as well as increasing the ambient noise by inducing air bubbles into the water. These factors will all reduce the range achieved and the effectiveness of the duct.
- c. **Velocity Gradient in the Surface Layer.** In order to have an effective surface duct the sound velocity gradient within the surface layer must be positive. The more positive the gradient the less sound energy that can escape and the better the detection ranges. Isovelocity conditions represent the limit of a positive gradient and result in straight-line propagation, the limiting type of Surface Duct.

13-28 Fig 4 Surface Duct



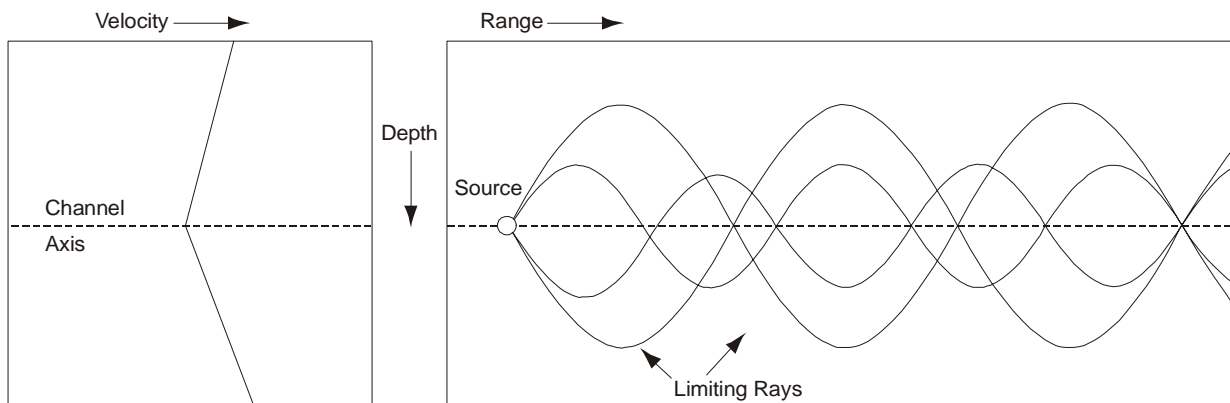
20. **Sound Channels.** The pre-requisite for a Sound Channel is a region of decreasing sound velocity overlaying a region of increasing sound velocity. Applying the HALT rule will show that sound rays will tend to be refracted towards the depth of minimum sound velocity. This depth is known as the Sound Channel Axis. A Sound Channel is illustrated in Fig 5. The following two types of sound channel are recognized:

a. **Deep Sound Channel.** This is also known as the SOFAR Channel (Sound Fixing and Ranging), and is the most common type of Sound Channel. It results from the negative velocity gradient of the thermocline overlaying the positive velocity gradient established by the pressure effect in deep water. The axis of the Deep Sound Channel is found at the depth of minimum sound velocity, usually at the base of the permanent thermocline. The SOFAR axis depth varies geographically. In the Atlantic Ocean, it varies between 1,300 ft in high latitudes (above 60° N) and 6,600 ft in low latitudes (about 35° N). The depth also decreases uniformly with longitude away from the Greenwich Meridian in the North Atlantic.

b. **Shallow Sound Channel.** This is sometimes known as a Depressed Sound Channel or Sub-Surface Duct. This type of sound channel occurs in the upper layers of water above the permanent thermocline but is somewhat transient in nature. In order to be effective, the axis depth of a Shallow Sound Channel must be between 150 and 600 feet below the surface. The channel must be at least 50 feet thick and the limiting rays must be at an angle of at least 2° above and below the channel axis. Shallow Sound Channels exist in the Mediterranean and the North-East Atlantic at certain times of the year and are present in the lower latitudes of the Atlantic all year. In the North Atlantic, they are only established during the summer months, with an axis depth of about 450 feet.

The usability of both types of sound channel is determined by the proximity of the receiver to the axis depth and by the signal frequency.

13-28 Fig 5 Sound Channel



21. **Convergence Zone.** Sound rays leaving a source near the surface which penetrate below the surface layer, will be refracted downwards until they pass the depth of the deep sound channel axis whence they begin to be refracted upwards towards the surface. Provided that the water depth is great enough a sufficient number of rays will eventually reach the near surface depths as a ring of focused energy around the source known as the annulus. This is illustrated in Fig 6. The likelihood of Convergence Zone propagation can be determined from a sound velocity profile using the following parameters which are shown graphically in Fig 7:

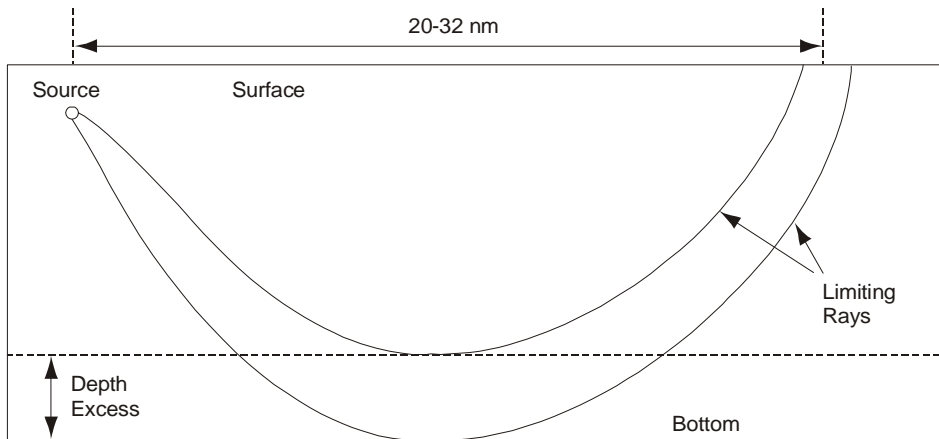
a. **Velocity at Source (Vs).** This is the sound velocity at the source depth.

b. **Velocity at Bottom (V_b)**. This is the sound velocity at the ocean bottom depth and is used to determine whether or not sound rays will be refracted back upwards towards the surface. Before any rays will be refracted upwards, the velocity at the bottom (V_b) must exceed the sound velocity at the source depth (V_s).

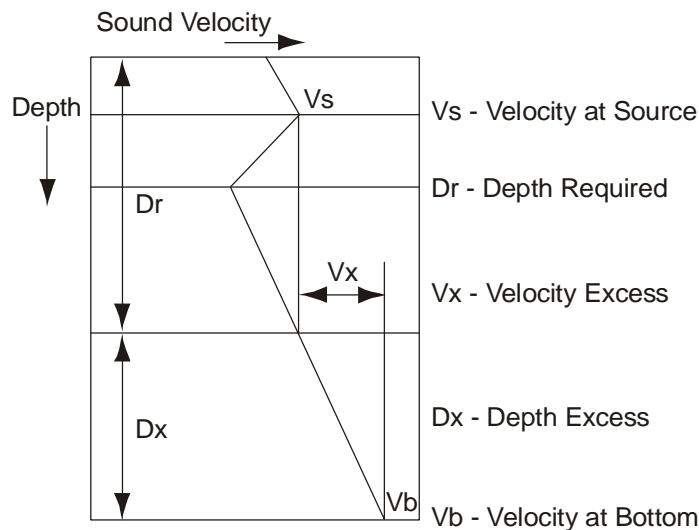
c. **Velocity Excess (V_x)**. If V_b is greater than V_s , then the difference is known as the 'Velocity Excess'. For a reliable Convergence Zone, this velocity excess must be approximately 33 fts^{-1} .

Convergence Zone range in the UK MPA's normal operating areas is between 20 and 32 nautical miles. It is unusual for Convergence Zone propagation to take place in warm or moderately warm water in depths of less than 1,200 fathoms. However, Convergence Zone can exist in water depths of less than 300 fathoms in certain conditions. Convergence zone width is directly proportional to the depth excess and can be roughly estimated at 10% of the range interval. It is possible to have a number of Convergence Zones around a single source. The first zone will be the strongest in terms of intensity. The second and third zones occur, respectively, at twice and three times the range of the first, providing that the source is strong enough.

13-28 Fig 6 Convergence Zone Propagation



13-28 Fig 7 Convergence Zone Velocity Profile



22. **Bottom Bounce Mode**. Sound rays which leave a source at angles greater than that of the Limiting Ray will eventually strike either the ocean bottom or the sea surface. The 'Bottom Bounce Mode' refers to those rays which are reflected back and forth between these two boundaries. Whereas

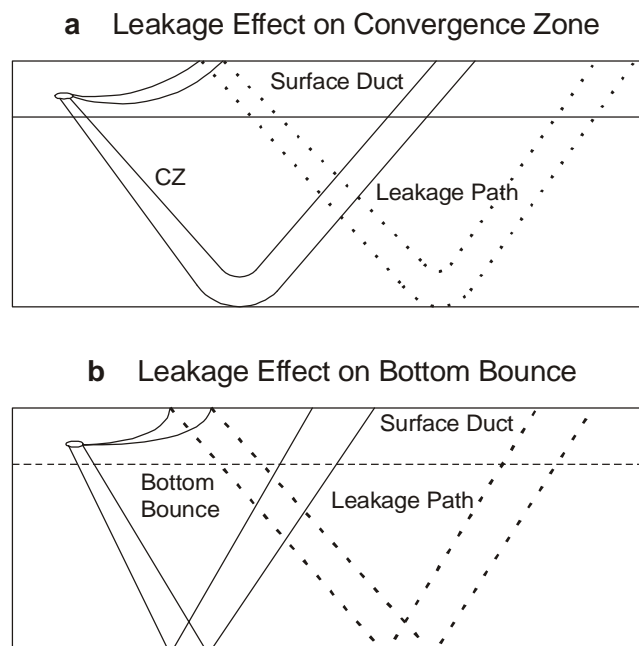
some energy will be reflected off the boundary surface (either ocean bottom or sea surface), a proportion will be lost to sensors due to scattering and absorption. The quality of this transmission mode is primarily dependent on factors such as boundary surface roughness, bottom type, water depth, and signal frequency. Bottom Bounce Mode is particularly important in shallow water or where the Sonic Layer depth is less than 50 feet. In fact, sound propagation for frequencies between 50 Hz and 1,500 Hz in shallow water is dominated by the 'Bottom Bounce' transmission mode.

Leakage Paths

23. In general sound channels or ducts are low loss propagation paths. At low frequencies, however, it is found that losses increase suddenly and the duct 'breaks down'. The frequency at which this effect occurs is known as the cut-off frequency. In simple terms the wavelength becomes too big to fit into the duct. The leakage path is in the direction of propagation and the main cause of the leakage is diffraction together with some scattering losses. In Volume 13, Chapter 25, when diffraction was being considered, it was shown that according to Huygens' principle secondary wavelets are produced from each point on a wave-front. Although the vast majority of the sound energy is propagated in the forward direction, these wavelets have a component of their energy to the side and this energy can 'leak' into shadow zones or out of a sound channel. Leakage paths tend to complicate the relatively simple propagation paths that have been outlined. Two examples are:

- a. **Leakage Effect on Convergence Zone.** Fig 8a illustrates the effect of leakage on convergence zone propagation. Whereas some energy will enter the convergence zone path immediately, some may remain trapped in the surface duct for some range before leaking into a different convergence zone path. This anomalous propagation can lead to confusion in recognizing the convergence zone and the width of the annulus.
- b. **Leakage Effect on Bottom Bounce.** Fig 8b illustrates the effect of leakage on bottom bounce propagation. Again, some energy is transmitted directly to the bottom while some remains trapped in the surface duct for a while before entering a bottom bounce path.

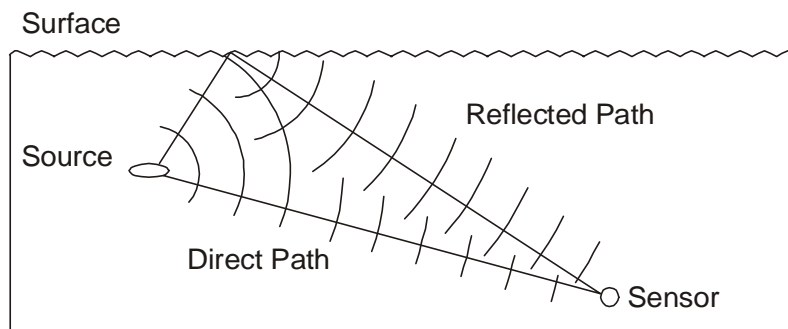
13-28 Fig 8 Leakage Effects



Coherency Effect (Lloyd's Mirror)

24. Clearly, it is possible for sound from one source to arrive at a receiver having followed a variety of paths. In general, any two sound paths will be of different lengths and so it is highly likely that the phase of the sound arriving by one path will be different to that of the sound arriving by another. If the received sound waves are in phase, then there will be an enhancement of the signal. If, however, the phases are different then there will be a reduction in the received signal with complete cancellation if the waves are 180° out of phase. Commonly this effect is a result of interference between a direct wave and a surface reflected wave or, in shallow water, between a direct wave and a bottom-reflected wave. A similar effect can occur if some sound is reflected by an object in the water and therefore arrives out of phase with the direct wave (secondary source polarisation). Lloyd's mirror effect is illustrated in Fig 9.

13-28 Fig 9 Lloyd's Mirror Effect



Diurnal (Afternoon) Effect

25. Solar heating near the surface water can cause short-term variations in the vertical temperature structure which can, in turn, bring about changes in sound propagation paths. These diurnal variations can lead to shortening of detection ranges. As the surface water is heated during the day, a negative temperature profile builds up causing refraction of the sound energy and the establishment of new shadow zones. As the surface water cools again, the temperature gradient and its associated effects disappear.

Noise

26. **Ambient Noise.** Ambient noise is that unwanted sound which is inherent in the sea independent of either the target or the aircraft. It can, of course, mask the noise generated by a target which the sensors are attempting to detect. Ambient noise can be predicted by graphical or computer methods (both of which depend upon a mix of theoretical and empirical data), or it can be measured using an ambient noise meter. There are three main sources of ambient noise in the areas where the UK MPA normally operate:

- a. **Shipping Traffic.** Shipping tends to generate the main component of ambient noise in deep-water areas, especially at lower frequencies (below 500 Hz). It is most intense in the vicinity of major shipping lanes and results from the sound of machinery, electrical systems, hydraulics, hydrodynamic flow, and propeller cavitation.
- b. **Sea State.** This noise source is the result of sea surface agitation causing the entrapment and subsequent escape of air bubbles. This noise is dominant at higher frequencies (above 500 Hz). Sea state noise is increased by strong surface winds especially in shallow water.

c. **Biological.** Some crustaceans, fish and marine mammals contribute to ambient noise by producing sounds associated with their normal life activities. Biological sounds are present over a large frequency spectrum and are extremely difficult to predict with any accuracy due to such factors as changing biological cycles and movement of the source organisms; there are often seasonal and diurnal cycles. In general, biological noise is greatest in shallow water, particularly in tropical and sub-tropical seas.

27. **Self-noise.** Self-noise is a noise generated by a maritime patrol aircraft's own mechanical, avionic, and electrical system.

Sound Transmission Losses

28. Not all of the sound generated by a source reaches a sensor. There are two general ways in which acoustic energy may be dissipated during propagation: spreading and attenuation.

29. **Spreading Losses.** Spreading is a geometrical phenomenon whereby a fixed amount of energy is distributed over an ever-increasing area as it moves away from the source. There are two types of spreading loss which may be encountered.

a. **Spherical Spreading.** If the sound velocity in the ocean is constant in both the horizontal and vertical directions, then the sound energy will radiate equally in all directions from the source. At any given distance, r , from the source, the power of the source, W , will be spread over the surface of a sphere of radius r . This surface area is $4\pi r^2$. The sound intensity, I , which is the sound power per unit area, will be given by:

$$I = \frac{W}{4\pi r^2} \text{ Wm}^{-2}$$

This is an inverse square law, i.e. intensity is inversely proportional to the square of the range. Notice that it is frequency independent. Working in Sound Pressure Levels (SPLs) in dBs, then the spreading loss (spherical) = $20 \log r$. If the range is doubled from r to $2r$ then:

$$\begin{aligned} \text{Change in SPL} &= 20 \log r - 20 \log 2r \\ &= 20 \log [r \div 2r] \\ &= 20 \log 0.5 = 20 \times -0.3 \\ &= -6 \text{ dB} \end{aligned}$$

i.e. a doubling of distance results in a 6dB loss.

b. **Cylindrical Spreading.** Cylindrical spreading occurs when the acoustic energy is trapped between two parallel boundaries such as the upper and lower bounds of a sound channel, surface duct or a convergence zone. The energy is constrained to spread in only two dimensions in a cylindrical manner. In this case, at any given distance from the source, r , the power of the source, W , will be spread over the surface of a cylinder of radius r . The area in this case will be $2\pi r$ and the variation of sound intensity with range is given by:

$$I = \frac{W}{2\pi r} \text{ Wm}^{-2}$$

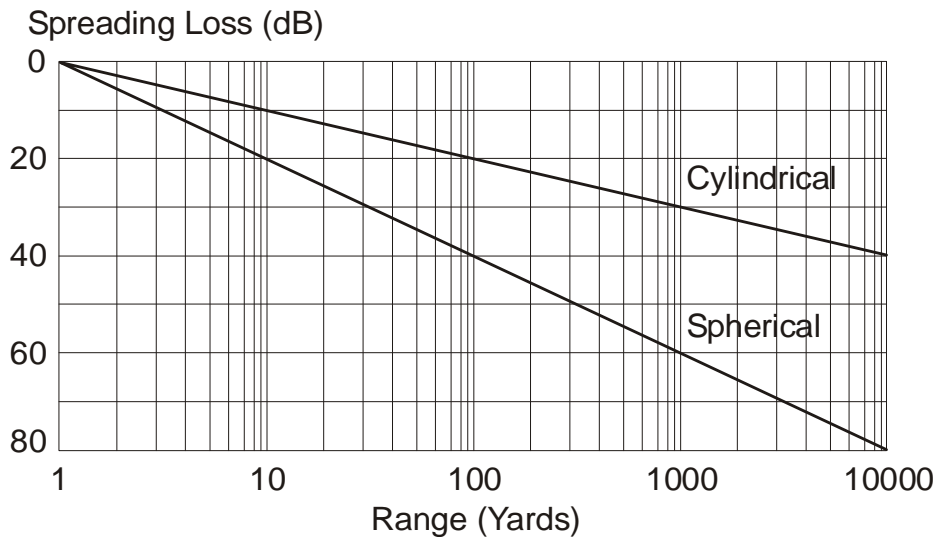
The spreading loss (cylindrical) in dBs = $10 \log r$, so looking again at the loss involved in a doubling of range:

$$\begin{aligned} \text{Change in SPL} &= 10 \log r - 10 \log 2r \\ &= 10 \log [r \div 2r] \\ &= 10 \log 0.5 = 10 \times -0.3 \\ &= -3 \text{ dB} \end{aligned}$$

i.e. doubling of distance results in a 3dB loss.

30. **Spreading Loss Comparisons.** Fig 10 shows a comparison between the different types of spreading loss. Since cylindrical loss is less severe than spherical loss, ducted modes of sound propagation often yield the highest probabilities of signal detection. However, most of the time the received signal has suffered some combination of the spreading losses. For example, since direct path makes up the initial portion of ALL transmission modes, they all exhibit some degree of spherical spreading.

13-28 Fig 10 Spreading Loss Graph



31. **Attenuation Losses.** Attenuation is the term applied to the linear decrease in acoustic energy per unit area of a wave-front as the distance from the source increases. Attenuation loss includes the effects of absorption, scattering and diffractive leakage.

a. **Absorption Loss.** Absorption loss involves the conversion of acoustic energy into heat by molecular action. Absorption loss is directly proportional to the range from the source. Increases in salinity and decreases in temperature increase absorption losses. Unlike spreading losses, absorption depends upon the frequency, varying approximately as the square of the frequency.

b. **Scattering Loss.** Scattering is the random reflection of acoustic energy from the ocean surface, the ocean bottom, or from suspended particles (volume scattering). Factors influencing the degree of each type of scattering are:

(1) **Surface Scattering.** The severity of surface scattering loss is dependent upon wave height, signal frequency, and angle of incidence of the sound energy.

(2) **Bottom Scattering.** The severity of bottom scattering is dependent upon the bottom roughness, sediment particle size, signal frequency, and angle of incidence of the sound energy.

(3) **Volume Scattering.** The severity of the scattering loss due to volume scattering is the most difficult to predict. It is dependent upon the ratio of the particle size responsible for the reflection to the signal wavelength. It is also dependent on the type of particle (i.e. solid or fluid). The most important contributor to volume scattering is biological in nature. Apart from the effect of the Deep Scattering Layer, volume scattering strength tends to decrease with depth.

The prediction of actual losses due to scattering is problematical, but there are some general observations that can be made: higher scattering losses are associated with higher frequencies, rougher scattering surfaces and larger scattering particles. Fluids such as bubbles are generally more effective volume scatterers than solid particles.

CHAPTER 29 - STATICS

Introduction

1. Statics is the study of forces in equilibrium. A particle is in equilibrium when all the forces acting upon it are balanced; it may be stationary, or it may be in a state of unchanging motion. The term force and some associated terms are discussed below.

Forces, Moments and Couples

2. **Force.** A force is that quantity which when acting on a body which is free to move, produces an acceleration in the motion of that body. For example, if a stationary football is kicked, it moves, ie it experiences an acceleration; the kick is the applied force. If a moving football is kicked, it changes speed, or direction, or both, so again it experiences an acceleration. However, just because a body is not accelerating does not mean that there are no forces acting upon it. Although that may be the case, it may equally be the situation that there are several forces acting such that their resultant is zero. For example, an aircraft which is flying at constant altitude, speed, and direction has many forces acting upon it, principally lift, thrust, weight, and drag.

3. **Weight.** Any body on or near the Earth is subject to the attraction of Earth's gravity, and if a body is released in this situation it will accelerate towards the centre of the Earth. The value of the acceleration, which is usually given the symbol g , is approximately 9.8 ms^{-2} but varies slightly both geographically and with altitude. Since gravity imposes an acceleration on a body, it must be a force, and it is this force which is known as weight.

4. **Vector Representation of a Force.** Force is a vector quantity, ie it has both magnitude and direction. A vector may be represented by a straight line drawn to scale and marked with an arrow. The direction of the line represents the direction of the vector and its length represents the magnitude. Vectors may be added together by means of the triangle, parallelogram, and polygon laws to give a single resultant and a single vector may be resolved into two or more components. A more complete treatment of vectors is to be found in Volume 13, Chapter 4.

5. **Moment or Torque.** The moment of a force about a point, or torque, is the tendency of the force to turn the body to which it is applied about that point. The magnitude of the moment or torque is the product of the force and the perpendicular distance from the point to the line of the force. The moment of a force about an axis is the product of the force and the length of the perpendicular common to the line of the force and the axis.

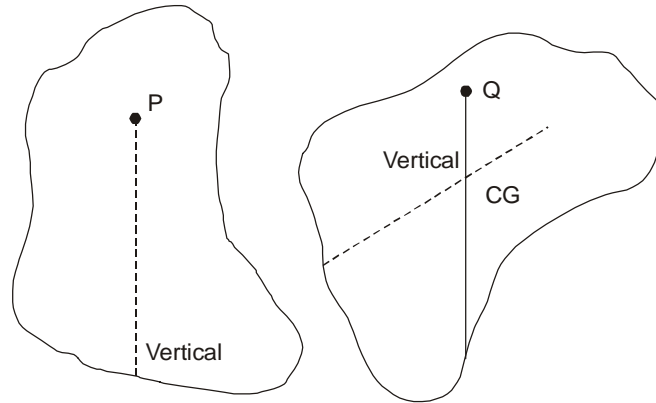
6. **Couple.** A system of two equal and parallel forces, acting in opposite directions but not in the same line, is known as a couple. The moment of a couple about any point in the plane of the forces is constant and equal to the product of one of the forces and the perpendicular distance between their lines of action.

Centre of Gravity

7. In any rigid extended body there is a unique point at which the total gravitational force, the weight, appears to act. This point is known as the centre of gravity.

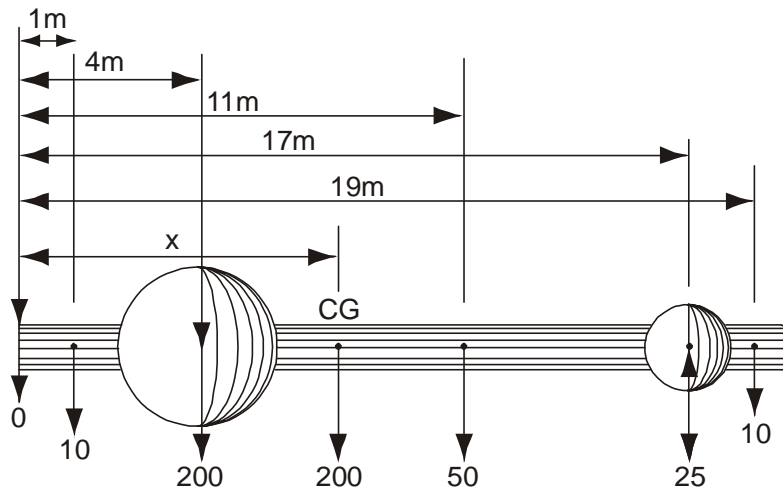
8. The position of the centre of gravity of a flat body can be determined by suspending it at any point, P and marking the vertical, then suspending it at a second point, Q, and again marking the vertical. The centre of gravity is at the intersection of the two lines (Fig 1).

13-29 Fig 1 Determining the Centre of Gravity of a Flat Plate



9. The position of the centre of gravity of a composite body, which can be treated as a number of symmetrical parts, can be found as shown in the following example. A body consists of two uniform spheres mounted on a uniform cylindrical bar as shown in Fig 2.

13-29 Fig 2 Determining the Centre of Gravity of a Symmetrical Body



Taking 0 as the reference point, the parts can be treated as three cylinders with centres of gravity distances 1, 11, and 19, metres respectively from the datum, and weights 10, 50, and 10 units; and two spheres of weights 200 and 25 units, with centres of gravity 4 and 17 metres respectively from the datum. Let the centre of gravity, through which the total weight (295 units) of the body acts, be x metres from 0. Taking moments about 0:

$$(10 \times 1) + (200 \times 4) + (50 \times 11) + (25 \times 17) + (10 \times 19) = 295x$$

$$\therefore 295x = 10 + 800 + 550 + 425 + 190 = 1975$$

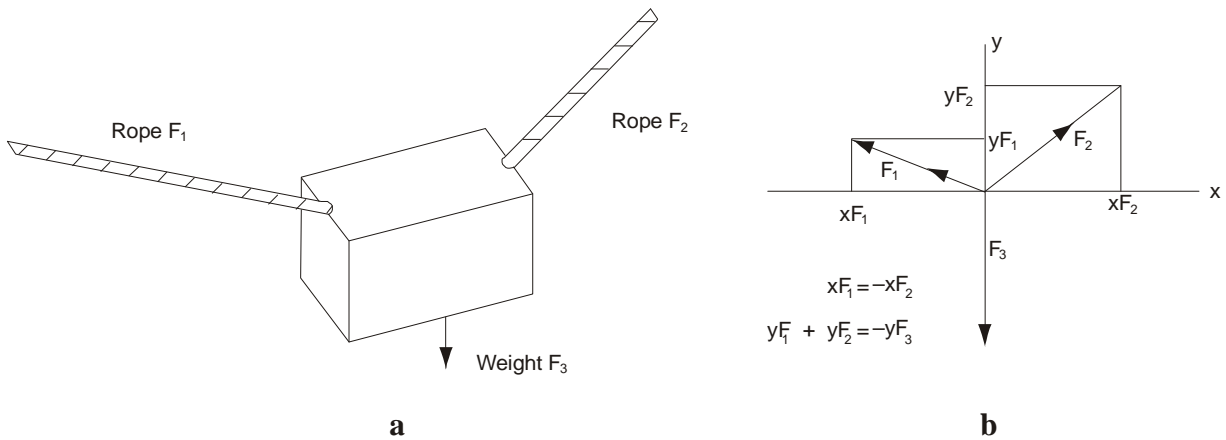
$$\therefore x = 6.69 \text{ metres.}$$

Equilibrium

10. Two types of equilibrium condition may be recognized, translational equilibrium and rotational equilibrium.

11. **Translational Equilibrium.** An object is said to be in translational equilibrium if it has constant velocity (including velocity equal to zero). In order to achieve this condition, the vector sum of all of the forces acting upon the object must be zero. It is usually convenient to consider the components of the forces in three orthogonal directions (x, y, and z axes). In this case the algebraic sum of the x, y and z components must each equal zero. As an example, consider Fig 3a which shows a uniform rectangular body being supported by two ropes. Three forces are acting on the body; the tension in the ropes, F_1 , and F_2 , and the weight, F_3 . All three forces may be regarded as coplanar and as acting through the centre of gravity and this simplified situation is shown in the vector diagram, Fig 3b. For translational equilibrium the x component of F_1 must be equal and opposite to the x component of F_2 ; F_3 has no x component. This equality can be shown by constructing verticals from the ends of the vectors to the x-axis. Also, the sum of the y components of F_1 and F_2 must be equal and opposite to F_3 . The y components may be determined by constructing horizontals from the ends of the vectors to the y-axis.

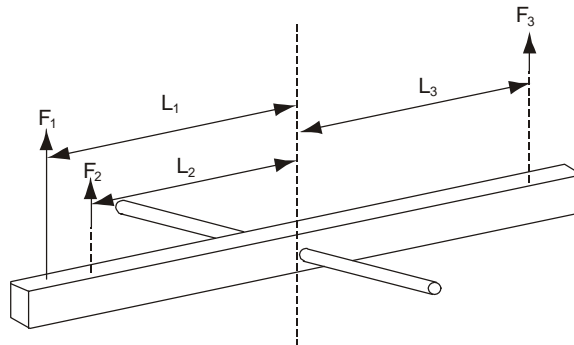
13-29 Fig 3 Translational Equilibrium



12. **Rotational Equilibrium.** An object is in rotational equilibrium when it rotates about an axis of constant direction at a constant angular speed, including zero angular speed. The condition of rotational equilibrium is achieved if the vector sum of all the torques about the axis is zero, i.e. the sum of the clockwise moments has the same magnitude as the sum of the anticlockwise moments.

13. As an example, consider the uniform bar shown in Fig 4, which is free to rotate about the axle. Three forces, F_1 , F_2 , and F_3 , are shown acting on the bar at distances L_1 , L_2 , and L_3 respectively from the axle. For rotational equilibrium, the relationship: $F_1L_1 + F_2L_2 - F_3L_3 = 0$ must be satisfied.

13-29 Fig 4 Rotational Equilibrium



Stability

14. Three conditions of stability with respect to equilibrium may be recognized as follows:

- a. **Stable Equilibrium.** An object is in stable equilibrium if any small displacement caused by external forces or torques tends to be self-correcting. An example would be a marble in the bottom of a round-bottomed cup. If an external force disturbs its equilibrium it will, after a few oscillations, settle back to its original position.
- b. **Unstable Equilibrium.** An object is in unstable equilibrium if any small displacement tends to be escalating. A marble balanced on the end of a finger would be an example. Any small displacement would cause the marble to move to a totally different position.
- c. **Neutral Equilibrium.** Neutral equilibrium is an intermediate state between stable and unstable equilibrium. Consider a marble at rest on a flat horizontal surface. A small displacement will cause the marble to move but the resulting position is unchanged from an equilibrium point of view from the original position. It will have no tendency either to be displaced further, or to return to its original position.

Friction

15. When two solid surfaces which are in contact move, or tend to move, relative to each other, a force acts in the plane of contact of the surfaces in a direction opposing the motion. This force is known as friction and is a result of interactions between the molecules of the two surfaces.

16. **Dynamic Friction.** If an object is sliding over a surface at a constant velocity, then the friction is an example of dynamic friction and there will be some conversion of the kinetic energy of the object into heat. The force of dynamic friction, f_d , depends upon the material of the two surfaces, on their smoothness and on the component of the force, F , that presses the two surfaces together. The frictional force is practically independent of the relative velocity of the two surfaces. It has been found that:

$$f_d = u_d F$$

where u_d is a constant for a given pair of surfaces and is known as the coefficient of dynamic friction. The value of u_d varies from about 0.06 for a smooth steel surface sliding on ice, to 0.7 for rubber sliding over dry concrete.

17. **Static Friction.** If an object is placed on a plane whose angle of inclination can be increased, it is found that the object remains stationary until a certain angle of inclination is reached whereupon the object starts to move. The force preventing the object from moving is known as static friction. As the plane's inclination is increased then the value of the component of the weight parallel to the plane increases. As the friction force is equal and opposite to this force (otherwise the object would move), then it too must increase as the inclination is increased until it reaches a critical angle. As with dynamic friction the maximum force of static friction, $f_{s \text{ max}}$ depends upon the materials, the smoothness of the surfaces in contact and on the component of the force, F , pressing the two surfaces together. For any pair of surfaces, it is found that:

$$f_{s \text{ max}} = u_s F$$

where u_s is constant for any two materials of specified smoothness and is called the coefficient of static friction. It varies from about 0.1 for steel on ice to about 1 for rubber on dry concrete. The coefficient of static friction is always higher than that of dynamic friction, which is why objects tend to start moving with a jerk.

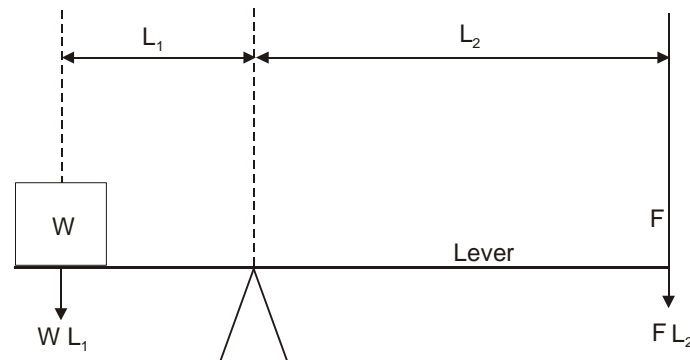
18. **Applicability.** It should be noted that the relationships described above are derived from observations rather than from any theoretical understanding of the mechanisms causing friction. Thus, the equations do not have universal applicability; deviations occur, for example, at extreme speeds, when the surfaces in contact are very small and when the force pressing the surfaces together is very large.

Machines

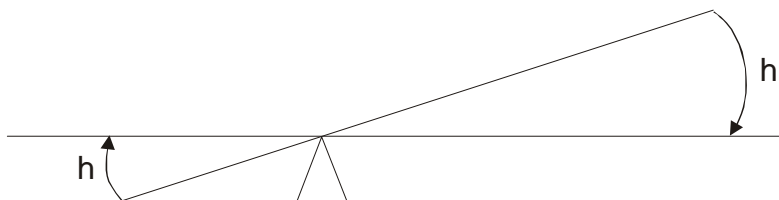
19. A machine is any device which enables energy to be used in a convenient way to perform work. Typical examples of simple machines are levers, winches, inclined planes, pulleys and screws. Machines do not save work; in general, they allow a smaller force to be applied in order to achieve a result, but the smaller force must be applied over a greater distance. As examples of machines, the lever and a pulley system will be reviewed.

20. **The Lever.** A lever can be described as a rigid beam supported at a point or fulcrum that is fixed, and about which the beam can turn. The arrangement is shown in Fig 5 where the purpose of the lever is to lift a load, of weight W . A force F is applied at the opposite end such that the lever is maintained in a horizontal position. In this situation, the system is in rotational equilibrium and so the moments about the pivot must be equal and opposite. The anticlockwise moment is WL_1 whilst the clockwise moment is FL_2 . Therefore, if L_2 is greater than L_1 the force required to balance the load is less than the load. However, as shown in Fig 6, in order to raise the load over a distance, h , the smaller force must be applied over a greater distance, h_1 .

13-29 Fig 5 Lever in Rotational Equilibrium



13-29 Fig 6 Lever - Smaller Force over Greater Distance



21. **A Two-pulley System.** A two-pulley system is shown in Fig 7 in which a force, F , is applied through a distance, a , in order to raise a weight, W , through a distance, b . From the principle of the conservation of energy (see Volume 13, Chapter 31, Para 16), the potential energy gained by the load will equal the work done by the effort, ignoring incidental energy losses (eg friction). Therefore:

$$Wb = Fa, \text{ or } \frac{W}{F} = \frac{a}{b}$$

As with the lever, a load can be lifted with a smaller force, but that force must be applied over a greater distance.

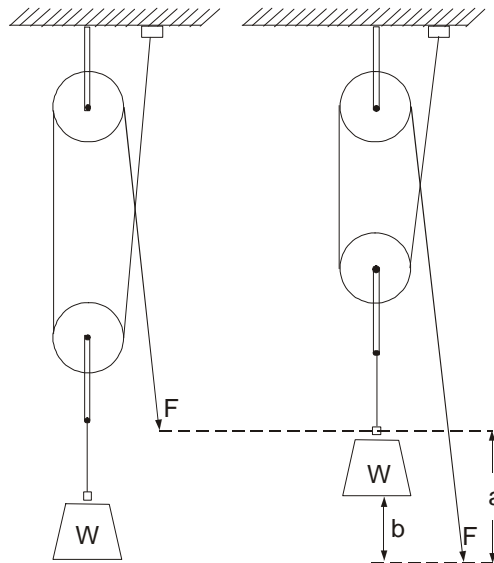
22. **Mechanical Advantage.** The ratio $\frac{W}{F}$ is known as the mechanical advantage. For the pulley system shown, if $a = 2b$, the mechanical advantage is 2.

23. **Efficiency.** Efficiency is related to the work done in moving an object. Work is described fully in Volume 13, Chapter 31, Para 10 and is the product of a force and the distance moved in the direction of that force. In the example:

$$\text{Efficiency} = \frac{\text{Work done on load}}{\text{Work done by effort}} = \frac{Wb}{Fa}$$

Efficiency is usually expressed as a percentage. Since friction has to be overcome, the efficiency of a machine is always less than 100%.

13-29 Fig 7 A Two-pulley System



CHAPTER 30 - KINEMATICS

Introduction

1. Kinematics is the study of motion without reference to the forces involved. In this chapter, linear and angular motion will be examined.

LINEAR MOTION

Speed and Velocity

2. Speed is the ratio of the distance covered by a moving body, in a straight line or in a continuous curve, to the time taken. The velocity of a body is defined as its rate of change of position with respect to time, the direction of motion being specified. If the body is travelling in a straight line, it is in linear motion, and if it covers equal distances in equal successive time intervals it is in uniform linear motion.

3. For uniform velocity, where s is the distance covered in time t , the velocity v is given by:

$$v = \frac{s}{t}$$

4. In the more general case, the instantaneous velocity v_i is given by:

$$v_i = \frac{ds}{dt}$$

Acceleration

5. The acceleration of a body is its rate of change of velocity with respect to time. Any change of either speed or direction of motion involves an acceleration; a retardation is merely a negative acceleration.

6. When the velocity of a body changes by equal amounts in equal intervals of time it is said to have a uniform acceleration, measured by the change in velocity in unit time.

7. If the initial velocity u of a body in linear motion changes uniformly in time t to velocity v , its acceleration a is given by:

$$a = \frac{(v - u)}{t}$$

Vector Representation

8. Velocity and acceleration are vector quantities and the laws of vector addition may be applied. Thus, the resultant velocity of a body having two separate velocities (e.g. an aircraft flying in wind) may be found by the parallelogram law. In addition, a single velocity may be resolved into two or more components.

Relationship between Distance, Velocity, Acceleration, and Time

9. Some useful formulae can be derived relating distance, velocity, acceleration, and time. When dealing only with motion in a straight line, the directional aspects of velocity and acceleration can be ignored, apart from the use of positive and negative signs to indicate forward and backward motion.

10. Consider a body with initial velocity, u , which in time t attains, under uniform acceleration a , a final velocity v . Suppose that the distance covered during this time is s . It has been stated that:

$$a = \frac{(v - u)}{t}$$

$\therefore at = v - u$, or

$$v = u + at \dots\dots\dots(1)$$

The mean or average velocity of the body is $(u + v)/2$. Therefore, the distance covered in time t will be given by:

$$s = \frac{(u + v)}{2} \cdot t \dots\dots\dots(2)$$

Substituting for v from equation (1),

$$s = \frac{1}{2} (u + u + at) \cdot t, \text{ or}$$

$$s = ut + \frac{1}{2} at^2 \dots\dots\dots(3)$$

From equation (1), $t = (v - u)/a$. Substituting for t in equation (2),

$$s = \frac{1}{2} \frac{(u + v)(v - u)}{a}, \text{ or}$$

$$v^2 = u^2 + 2as \dots\dots\dots(4)$$

11. Note that if the distance travelled in time t is denoted by s , velocity can be obtained by differentiation:

$$v = \frac{ds}{dt}$$

Similarly, the acceleration a is given by:

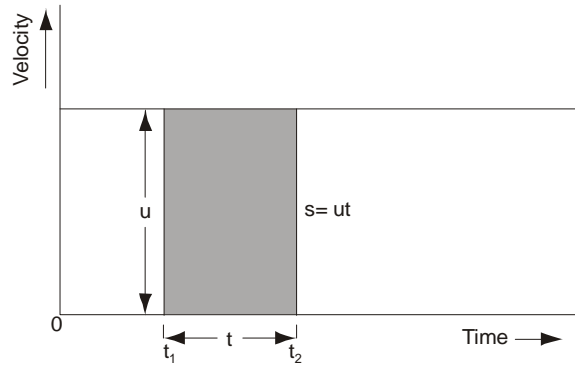
$$a = \frac{dv}{dt} = \frac{d^2s}{dt^2}$$

Conversely, given an expression for acceleration, integration will give an expression for velocity, and further integration an expression for displacement.

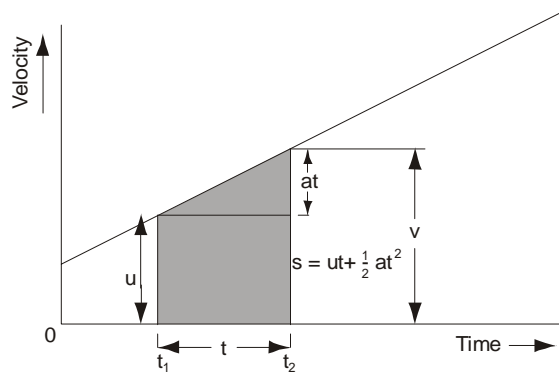
12. *Velocity-Time Graphs.* The linear motion of a body can be illustrated by means of velocity-time graphs, examples of which are shown in Fig 1. In each case the distance travelled by the body between times t_1 and t_2 is represented by the area under the corresponding part of the graph (shaded in Fig 1). The instantaneous acceleration at any time is given by the slope of the curve at that point.

13-30 Fig 1 Velocity - Time Graph

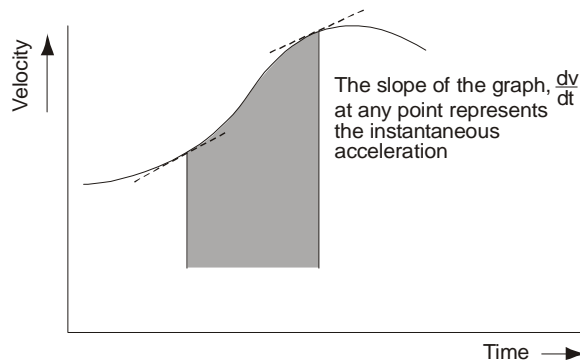
a Uniform Velocity



b Uniform Acceleration



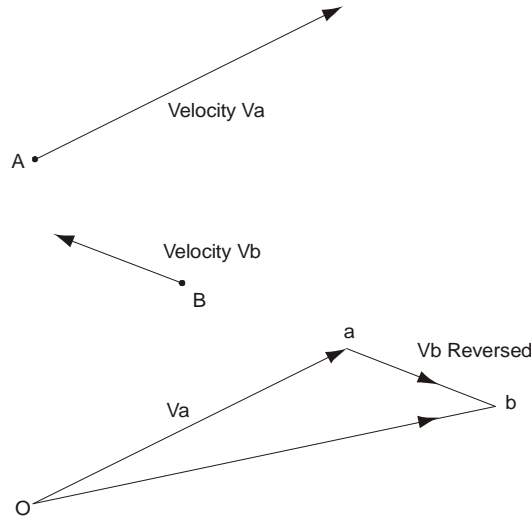
c Changing Acceleration



Relative Velocity

13. It is sometimes necessary to determine the velocity with which one moving body appears to be moving with respect to another. This is known as the relative velocity. For linear motion this type of problem may be solved graphically, as in Fig 2, by drawing from an origin a vector representing the velocity of body A, and from the end of this vector drawing a vector to represent the velocity of B reversed. In Fig 2 the third side of the triangle, ob, represents the velocity of A relative to B.

13-30 Fig 2 Relative Velocity



ROTARY MOTION

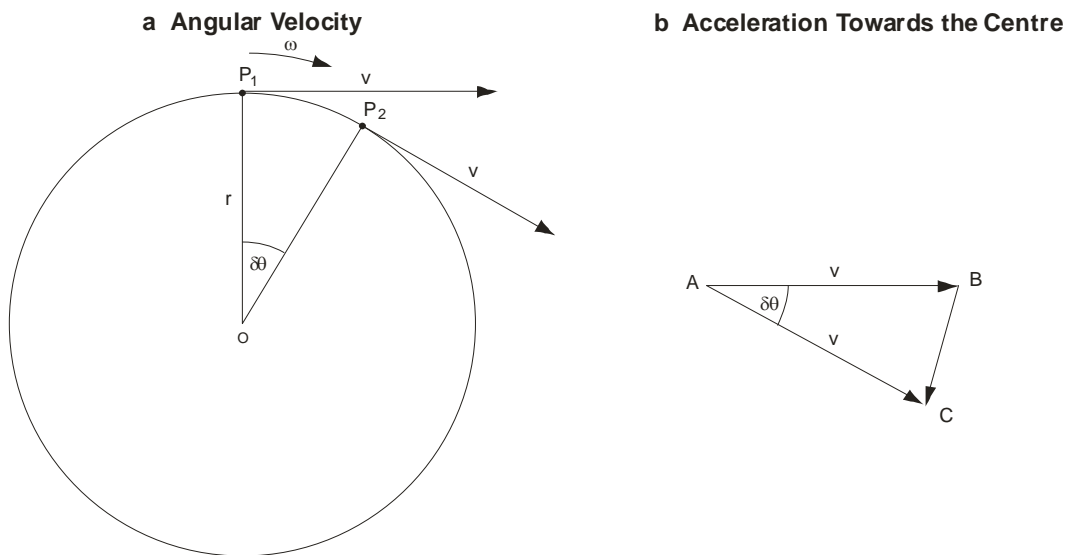
Angular Velocity

14. So far, only motion in a straight line has been examined. Consider now a point P which moves in a circle of radius r at constant speed v. Its angular velocity ω is given by $d\theta/dt$ where θ is the angular displacement in radians (Fig 3a). The speed v is given by:

$$v = r \cdot \frac{d\theta}{dt} = r\omega$$

15. Although the velocity of the point is constant in magnitude, it is constantly changing in direction, and therefore, by definition, P is subject to an acceleration. Consider the point P at a certain instant (position P₁ in Fig 3a), and also after a small interval of time δt (position P₂). It is clear that the vectors representing motion at these two instants are of the same length, but are in different directions. They are represented by AB and AC in Fig 3b, BC representing the change in velocity in the time δt . The acceleration is given by $BC/\delta t$.

13-30 Fig 3 The Acceleration Experienced in Uniform Circular Motion



16. Since δt is small, the angular displacement, $\delta\theta$, is also small, and BC may be considered to be almost perpendicular to both AB and AC, that is, directed towards the centre of the circle.

$$BC = v \cdot \delta\theta$$

$$\text{but, } \delta\theta = \frac{P_1P_2}{r}$$

If the time interval δt is very small, the straight line P_1P_2 is very nearly equal to the distance P_1P_2 along the arc, which is $v \cdot \delta t$.

$$\therefore \delta\theta = \frac{v \cdot \delta t}{r} \text{ radians}$$

$$\therefore BC = \frac{v^2 \delta t}{r}$$

$$\text{Acceleration} = \frac{BC}{\delta t}$$

$$= \frac{v^2}{r} \text{ or } v\omega, \text{ or } \omega^2 r$$

all towards the centre.

17. To summarize, a body moving at constant speed v in a circle of radius r has a constant acceleration of v^2/r , directed towards the centre of the circle.

18. The following formulae for circular motion, similar to those for linear motion, may be derived:

For uniform angular velocity, $\theta = \omega t$.

For uniform angular acceleration,

$$\omega_2 = \omega_1 + \alpha t \quad (\text{cf, } v = u + at)$$

$$\theta = \frac{(\omega_1 + \omega_2)t}{2}$$

$$\text{and } \theta = \omega_1 t + \frac{1}{2} \alpha t^2 \quad (\text{cf, } s = ut + \frac{1}{2} at^2)$$

$$\text{and } \omega^2 = \omega_1^2 + 2\alpha\omega \quad (\text{cf, } v^2 = u^2 + 2as)$$

where

ω_1 = initial velocity in radians per sec

ω_2 = velocity in radians per sec after t sec

α = angular acceleration in radians per sec²

θ = angle through which turned, in radians

Relationship between Angular and Linear Velocity

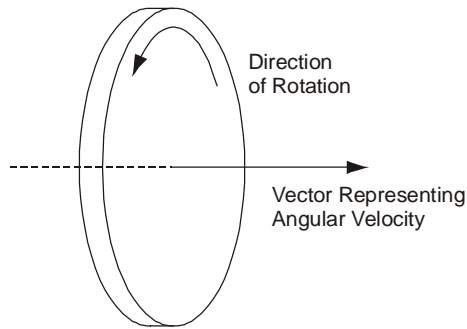
19. Consider again the point P, rotating with uniform angular velocity ω and radius r (Fig 3a). The length of the arc P_1P_2 is the radius multiplied by the angle in radians, i.e. arc $P_1P_2 = r\theta$. But, for uniform angular velocity, $\theta = \omega t$, hence arc $P_1P_2 = r\omega t$ and the linear velocity of P is $r\omega t/t$, or ωr . Similarly, the linear acceleration of P is equal to the angular acceleration times the radius; its direction is radially towards the centre of rotation. Summarizing, if v = linear velocity, ω = angular velocity, a = linear acceleration, and α = angular acceleration, then:

$$v = \omega r, \text{ or } \omega = \frac{v}{r} \text{ and } a = \alpha r, \text{ or } \alpha = \frac{a}{r}$$

Vector Representation of Angular Quantities

20. Angular velocity and acceleration are vector quantities. By convention, they are represented by vectors; the vector length represents the magnitude of the quantity and its direction is perpendicular to the plane of rotation, i.e. parallel to the axis of rotation. The direction of the arrow is such that, on looking in its direction, the rotation is clockwise (Fig 4). This convention is known as the right-handed screw law. Such vectors can be combined and resolved according to the normal principles of vectors.

13-30 Fig 4 Vector Representation of Angular Quantities



CHAPTER 31 - DYNAMICS

Introduction

1. Dynamics is the study of motion related to force. In Volume 13, Chapter 30, it was shown how a body moves; in this chapter it will be shown why the body moves in that way.

QUANTITIES

Mass

2. When an object is at rest, a force is necessary to make it move; similarly, a body in motion needs a force to be applied in order to change its motion. This reluctance to any change in motion is called inertia and the property that gives rise to inertia is mass. It can be shown that inertia is directly proportional to mass.

3. The mass of a body may be defined as the quantity of matter in the body. The unit of mass is the kilogram (kg) and the standard kilogram is a cylinder of platinum-iridium alloy kept at the Bureau International des Poids et Mésures.

Density

4. The density of a substance is the mass per unit volume of that substance. The unit of density is the kilogram per cubic metre (kg m^{-3}). Relative density (or specific gravity) is the ratio of the density of a substance at a stated temperature to the density of water at 4°C .

Momentum

5. The momentum of a body is defined as the product of its mass and its velocity. As the definition includes the velocity term, momentum is a vector quantity and a change in either speed or direction constitutes a change in momentum. The unit of momentum is the kilogram metre per second (kg ms^{-1}).

Force

6. **Newton's laws of motion:**

- a. **First Law.** A body remains in a state of rest or uniform motion (i.e. no acceleration) in a straight line unless acted upon by an external force.
- b. **Second Law.** The rate of change of momentum is proportional to the applied force, and the change of momentum takes place in the direction of the applied force.
- c. **Third Law.** Every action is opposed by an equal and opposite reaction.

7. By selecting a unit of force as that force which gives unit acceleration to unit mass the second law may be written as:

$$F = ma$$

The unit of force is the newton (N) which is the force necessary to induce an acceleration of 1ms^{-2} in a mass of 1 kilogram.

Conservation of Momentum

8. Consider two bodies, A and B, travelling in the same direction, which collide, the duration of the collision being the short time t . Throughout the collision, each will experience a force equal and opposite to that experienced by the other (Newton's third law). The impulse of the force is Ft , and is the same for each body. Thus, the change of momentum will be the same for each body. If at the time of collision body A was overtaking body B, it is apparent that the effect of the impact will be to decrease the momentum of A and increase that of B, and the total momentum of the system of two bodies will be unchanged.

9. The Law of Conservation of Momentum. The effects of the interaction of parts of a closed system are summarized in the Law of Conservation of Momentum, which states that the total momentum in any given direction before impact is equal to the total momentum in that direction after impact.

Work

10. A force is said to do work when its point of application moves, and the amount of work done is the product of the force and the distance moved in the direction of the force.

11. The unit of work is the joule (J) which is the work done when a force of 1 newton moves 1 metre in the direction of the force.

12. The gravitational force acting on a body is the product of its mass (m) and the acceleration of gravity (g), i.e. mg . Therefore, the work done against gravity in raising the body through a vertical distance, h , is mgh .

Energy

13. The energy possessed by a body is its capacity to do work; the unit of energy is the joule. A body may possess this capacity by virtue of:

- a. Its position, when the energy is called potential energy (PE).
- b. Its motion, when the energy is called kinetic energy (KE).

14. Consider a mass, m , projected vertically upwards from the ground with initial velocity, u . It is acted upon (downwards) by gravity, and will attain a height, h , which can be determined by substitution in the formula $v^2 = u^2 + 2as$, thus:

$$0 = u^2 + 2(-g)h$$

the velocity being zero at the highest point

$$\therefore h = u^2/2g$$

The work done reaching a height h is mgh . Substituting for h ,

$$\text{work done} = mgu^2/2g = \frac{1}{2}mu^2$$

$$= \text{work done in coming to rest from velocity } u$$

$$= \text{KE.}$$

15. At the highest point, where the velocity is zero, the kinetic energy is zero. If, however, the mass is allowed to fall, it will, at the moment of striking the ground, have regained its original velocity, and thus

its original kinetic energy. The energy which it possessed at the highest point is potential energy and equals its original kinetic energy.

16. Consider also a body acted upon by a constant force F .

$$\text{Then } F = ma = m \left(\frac{v^2 - u^2}{2s} \right)$$

where u , v , and s have the conventional significance.

$$\text{Thus: } Fs = \frac{1}{2}m(v^2 - u^2)$$

The left-hand side is the work done by the force, while the right-hand side is the resulting change in kinetic energy. The general expression for the kinetic energy of a body moving with velocity v is:

$$\text{KE} = \frac{1}{2}mv^2$$

This illustrates the principle of the conservation of energy - energy can be neither created nor destroyed, though it may be converted from one form to another. It should be noted that kinetic and potential energies are examples of one form of energy, namely mechanical energy; a change may involve other forms of energy such as chemical, heat, light, sound, magnetic or electrical energy.

Power

17. Power is the rate of doing work and has the unit, watt (W), equal to 1 joule per second.

CIRCULAR MOTION

Centripetal Force

18. It was shown in Volume 13, Chapter 30 that a body travelling with uniform speed in a circle has an acceleration of v^2/r towards the centre of the circle. The force producing this acceleration is termed centripetal force, and for a body of mass m the centripetal force is mv^2/r towards the centre of the circle. It should be noted that in the case of a body travelling in a circular path at the end of a string, while the mass is experiencing centripetal force towards the hand, there is an equal and opposite reaction on the hand holding the string, known as centrifugal force. Centrifugal force exists only as an equal and opposite reaction to centripetal force.

Moment of Inertia

19. Consider now a rigid body, such as a flywheel, free to rotate about an axis through its centre, at angular velocity ω . When the wheel is rotating all the particles of the wheel have the same angular velocity, but their linear velocities will depend on their individual distances from the axis, those on the rim moving much faster than those near the axis.

20. A particle of mass m , distance r from the centre, and with linear velocity v , has kinetic energy $\frac{1}{2}mv^2$ or $\frac{1}{2}mr^2\omega^2$. The total kinetic energy of all the particles is $\frac{1}{2}\omega^2 \Sigma mr^2$.

21. The sum (expressed as Σ) of the products mr^2 for all the particles in a rigid body rotating about a given axis is called the Moment of Inertia (I).

$$I = \Sigma mr^2$$

22. Moment of inertia can be considered as the rotational equivalent of mass and, just as in linear motion, if a force is applied to a body, $\text{force} = \text{mass} \times \text{acceleration}$, so in circular motion, if a torque is applied to a wheel,

$$\text{torque} = \text{moment of inertia} \times \text{angular acceleration.}$$

23. **Radius of Gyration.** The radius of gyration is a useful concept whereby the mass $M (= \Sigma m)$ of the wheel is considered to be concentrated in a ring of radius k from the axis, such that $\Sigma mr^2 = Mk^2$. 'k' is then known as the radius of gyration, and $I = Mk^2$.

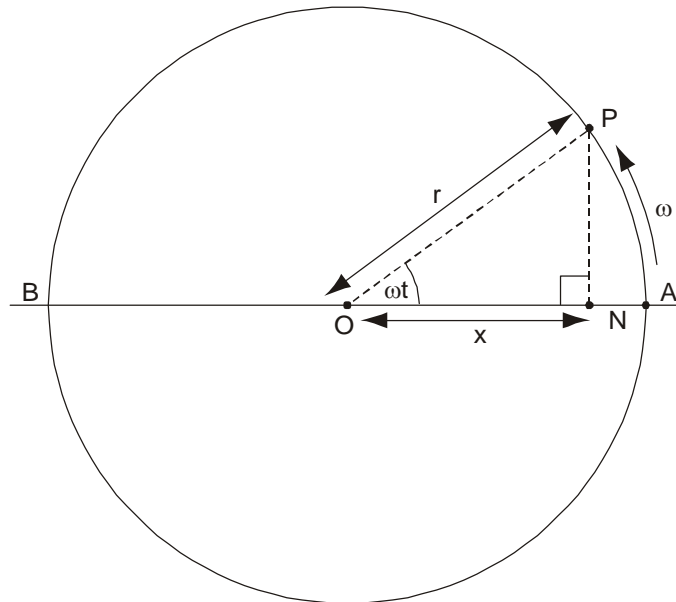
SIMPLE HARMONIC MOTION

General

24. A type of motion which is frequently encountered is simple harmonic motion. This is defined as the motion of a body which moves in a straight line so that its acceleration is directly proportional to the distance from a fixed point in the line, and always directed towards that point. A mass suspended from a spiral spring which is given a small vertical displacement from its equilibrium position and then released will oscillate with simple harmonic motion.

25. Simple harmonic motion can be illustrated by considering the projection of a point moving at uniform speed in a circular path on to a diameter of that circle (Fig 1).

13-31 Fig 1 Simple Harmonic Motion



Let the point P start at position A, and move with constant angular velocity ω , in a circle of radius r . Let PN be the perpendicular from the point to the diameter AOB. Then while the point P moves round the circle from A through B back to A, N moves through O to B and back to A. The time for one complete cycle of oscillation is $2\pi/\omega$ and this is called the periodic time or simply the period of the oscillation. The frequency of oscillation is the number of complete cycles in unit time, and frequency = $1/\text{periodic time} = \omega/2\pi$. The distance ON at any time, t , is given by $x = r \cos \omega t$. The instantaneous velocity of N is given by:

$$\frac{dx}{dt} = -r\omega \sin \omega t \quad \text{and the acceleration by:}$$

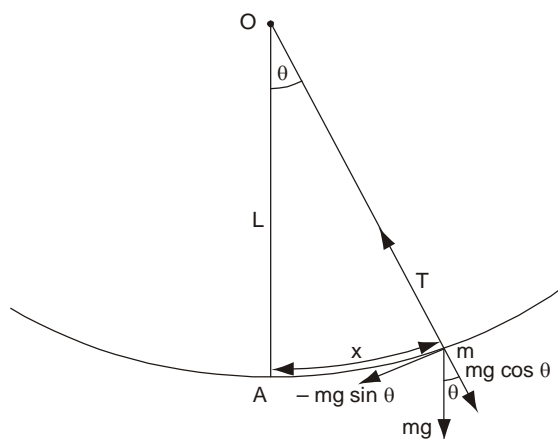
$$\frac{d^2x}{dt^2} = -r\omega^2 \cos \omega t = -\omega^2 x$$

Since the angular velocity ω is constant, the acceleration is proportional to the distance ON, thus the movement of N accords with the definition of simple harmonic motion. Note that the expression is negative, since the acceleration is measured in the opposite direction to that of the displacement ON (i.e. towards O).

The Simple Pendulum

26. The simple pendulum provides an example of simple harmonic motion, provided that its amplitude of oscillation is small. Consider a mass, m , suspended from a light inextensible cord of length L , displaced through a small angle θ (in radians) from the vertical (Fig 2).

13-31 Fig 2 The Simple Pendulum



27. The forces acting on the mass are its weight, mg , acting vertically downwards, and the tension, T , in the cord. The weight can be resolved into two components, $mg \cos \theta$ equal and opposite to T , and the restoring force, $-mg \sin \theta$ acting in the opposite direction to the direction of displacement of the mass from its central position. Using the equation:

$$\text{force} = \text{mass} \times \text{acceleration},$$

$$-mg \sin \theta = ma$$

$$\therefore a = -g \sin \theta, \text{ or, for small angles, } a = -g\theta \text{ (approximately)}$$

$$\text{but, } \theta = \frac{x}{L}$$

$$\therefore a = -\frac{gx}{L}$$

28. As g and L are constant for a particular location and a particular pendulum, the acceleration is proportional to the displacement x , and acts in the direction of the force towards the mid point A , thus (for small displacements) satisfying the conditions for simple harmonic motion.

29. By comparing the results of paras 25 and 27 it can be seen that the period of oscillation of a simple pendulum, for small displacements, is given by:

$$\text{period} = 2\pi \sqrt{\frac{L}{g}} \text{ sec},$$

and the frequency by:

$$\text{frequency} = \frac{1}{2\pi} \sqrt{\frac{g}{L}} \text{ hertz.}$$

It will be noted that, for small angles of swing, the period of oscillation of a simple pendulum is independent of the mass, and of the amplitude of swing.

The Compound Pendulum

30. For a compound, or rigid, pendulum, the period is given by:

$$\text{period} = 2\pi \sqrt{\frac{I}{mgh}}$$

where I is the moment of inertia about the axis of rotation, and h is the distance from the centre of gravity to the pivot.

A Mass Supported by a Spring

31. The comparable equation for the period of oscillation of a mass m kilograms suspended from a spring of stiffness e kilograms per metre is:

$$\text{period} = 2\pi\sqrt{\frac{m}{e}} \text{ sec.}$$

CHAPTER 32 - HYDRAULICS

Introduction

1. Hydraulic systems provide a means of transmitting a force by the use of fluids. They are concerned with the generation, modulation and control of pressure and flow of the fluid to provide a convenient means of transmitting power for the operation of a wide range of aircraft services. A typical aircraft hydraulic system will be used for operating flying controls, flaps, retractable undercarriages and wheelbrakes. Hydraulic systems can transmit high forces with rapid, accurate response to control demands.

Definition of Terms

2. The following definitions need to be understood:

a. **Pressure.** Pressure is the force per unit area exerted by a fluid on the surface of a container. Pressure is measured in bars, pascals (Pa), or newtons per square metre (N/m²).

$$1 \text{ bar} = 100,000 \text{ Pa} = 100,000 \text{ N/m}^2$$

b. **Force.** The force exerted on a particular surface by a pressure is calculated from the formula:

$$\text{Force} = \text{Pressure} \times \text{Surface Area.}$$

c. **Fluid.** A fluid is a liquid or gas which changes its shape to conform to the vessel that contains it.

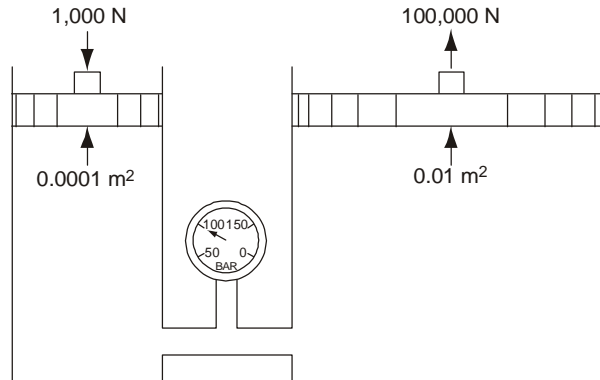
d. **Hydraulic Fluid.** Hydraulic fluid is an incompressible oil. In aircraft systems, low flammability oils are used, the boiling and freezing points of which fall outside operating parameters.

Transmission of Force and Motion by Fluids

3. When a force is applied to one end of a column of confined fluid a pressure is generated which is transmitted through the column equally in every direction. Fig 1 illustrates a simple hydraulic system. Two cylinders are connected by a tube. The cylinders contain pistons of surface areas 0.01 m² and 0.0001 m² giving a piston area ratio of 100:1. The system is filled with hydraulic fluid and fitted with a pressure gauge. A force of 1,000 newtons (N) is applied to the small piston. The force will produce a pressure (P) in the fluid so:

$$P = \frac{\text{Force (F)}}{\text{Piston Area (A)}} \quad \text{i.e.} \quad P = \frac{1000}{0.0001} = 10,000,000 \text{ Pa or } 10 \text{ MPa, which is } 100 \text{ bar.}$$

The system pressure will appear on the gauge and will be felt on every surface within the system. Thus at the large piston a force (F) will be exerted where $F = P \times A = 10 \times 10^6 \times 0.01 = 100,000 \text{ N}$. The force applied at the small piston is therefore increased on delivery by the large piston by a factor of 100, i.e. in the same ratio as that of piston area. This is sometimes referred to as force multiplication.

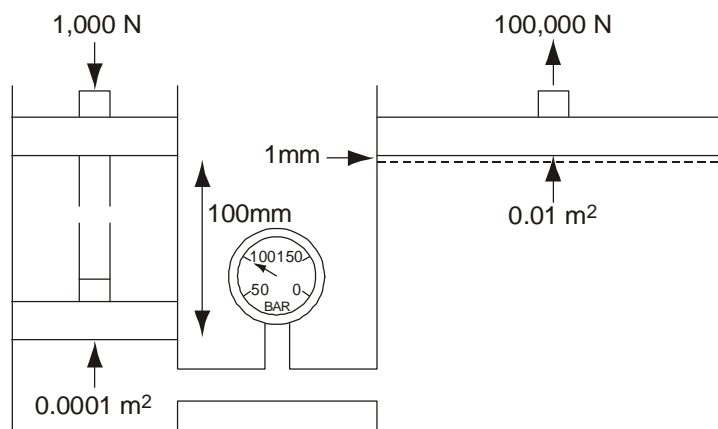
13-32 Fig 1 A Simple Hydraulic System

4. If the small piston is now moved down in its cylinder through a distance of 100 mm, as illustrated in Fig 2, the large piston will move upwards through a distance inversely proportional to the piston area ratio, ie through 1 mm. The work done by the small force is transmitted hydraulically and equals the work expended in moving the greater force through a smaller distance,

$$\text{i.e.} \quad \text{force 1} \times \text{distance 1} = \text{force 2} \times \text{distance 2}$$

$$\text{or} \quad 1000 \text{ N} \times 100 \text{ mm} = 100,000 \text{ N} \times 1 \text{ mm}$$

This is the principle of the hydraulic lever and is the operating principle of any hydraulic system.

13-32 Fig 2 Relative Movement in a Simple Hydraulic System

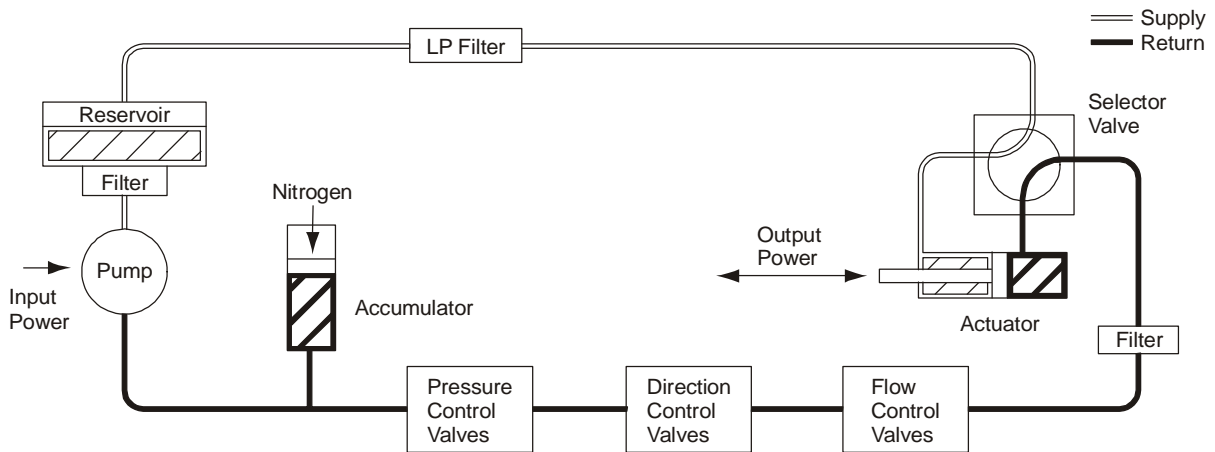
Components of a Typical Hydraulic System

5. In a simple hydraulic system, force applied by hand or foot to a small 'master' piston is transmitted to a larger 'slave' piston to operate the required service. In more complex, high performance systems, the master piston is replaced by hydraulic pumps and the slave by actuators driving each of the powered services. A typical system illustrated in Fig 3 will operate at between 200 and 300 bar, and it will include the following components:

- a. **Hydraulic Pump.** The pump generates hydraulic pressure and delivers it to the pressure lines in the system. It will usually be either engine driven or electrically powered.
- b. **Valves.** Non-return valves control the direction of fluid flow, pressure relief valves the level of power produced, and selector valves the amount of fluid flow to related actuators.

- c. **Actuators.** Actuators convert the hydraulic power into usable mechanical power at the point required.
- d. **Hydraulic Fluid.** The fluid provides the means of energy transmission as well as lubrication, and cooling of the system.
- e. **Connectors.** The connectors link the various system components. They are usually rigid pipes, but flexible hoses are also used.
- f. **Reservoir.** Fluid is stored in a system reservoir in sufficient quantity and quality to satisfy system requirements. The fluid becomes heated by operation of the system, and the reservoir performs the secondary functions of cooling the fluid and of allowing any air absorbed in the fluid to separate out.
- g. **Filters.** Hydraulic system components are readily damaged by solid particles carried in the fluid, and several stages of filtration are included in a system to prevent debris passing from one component to the next. The filters perform as useful tell-tales of fluid contamination. In addition, samples are taken periodically from the fluid and analysed to detect trends in acid and other trace element levels.
- h. **Accumulator.** An accumulator is a cylinder containing a floating piston. On one side of the piston is nitrogen at system pressure, and on the other hydraulic fluid from the pressure line. When the hydraulic pressure is increased, the nitrogen is compressed. The compressed nitrogen then acts as a spring and can damp out system pressure ripples. It also acts as a reserve of fluid and an emergency power source.

13-32 Fig 3 Typical Practical Hydraulic System



CHAPTER 33 - INTRODUCTION TO GYROSCOPES

Introduction

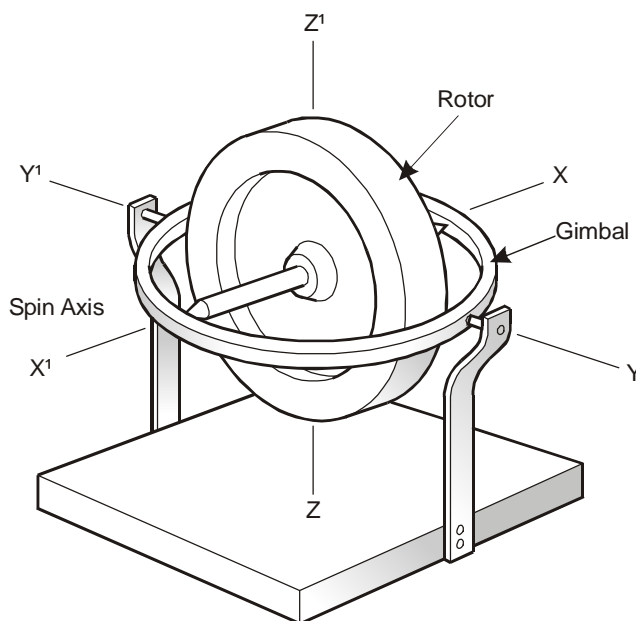
1. Modern technology has brought about many changes to the gyroscope. The conventional spinning gyroscope is still in current use for flight instruments in smaller and simpler aircraft. More sophisticated aircraft however, make use of devices which are termed 'gyros', but this is because of the tasks they perform rather than their manner of operation. Gyroscopes can therefore be categorised as:

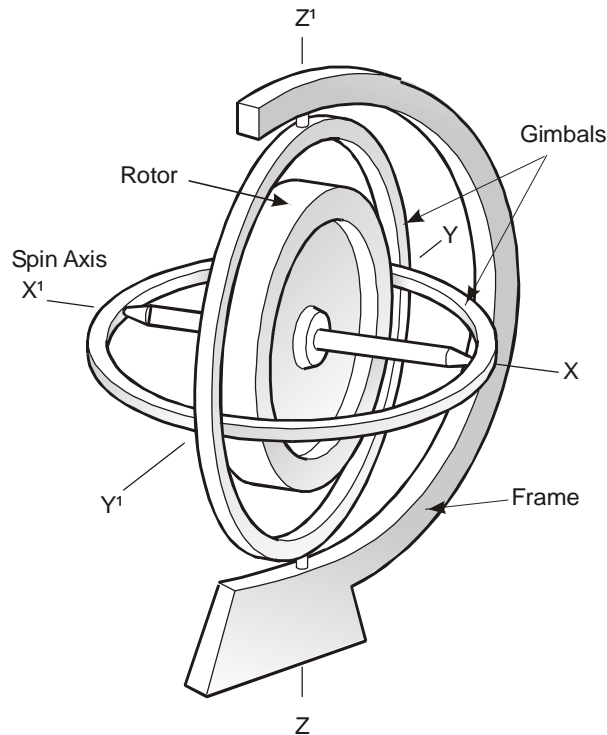
- a. Spinning Gyroscopes.
- b. Optical Gyroscopes.
- c. Vibrating Gyroscopes.

This chapter will concentrate for the most part on the spinning gyroscope.

2. A conventional gyroscope consists of a symmetrical rotor spinning rapidly about its axis and free to rotate about one or more perpendicular axis. Freedom of movement about one axis is usually achieved by mounting the rotor in a gimbal, as in Fig 1 where the gyro is free to rotate about the YY^1 axis. Complete freedom can be approached by using two gimbals, as illustrated in Fig 2.

13-33 Fig 1 One-degree-of-freedom Gyroscope



13-33 Fig 2 Two-degrees-of-freedom Gyroscope

3. The physical laws which govern the behaviour of a conventional gyroscope are identical to those which account for the behaviour of the Earth itself. The two principal properties of a gyro are rigidity in inertial space and precession. These properties, which are explained later, are exploited in some heading reference and inertial navigation systems (INS) and other aircraft instruments which are described in Volume 5, Chapter 10.

Definition of Terms

4. The following fundamental mechanical definitions provide the basis of the laws of gyro dynamics:

a. **Momentum.** Momentum is the product of mass and velocity (mv).

b. **Angular Velocity.** Angular velocity (ω) is the tangential velocity (v) at the periphery of a circle, divided by the radius of the circle (r), so $\omega = \frac{v}{r}$. Angular velocity is normally measured in radians per second.

c. **Moment of Inertia.** Since a rotating rigid body consists of mass in motion, it possesses kinetic energy. This kinetic energy can be expressed in terms of the body's angular velocity and a quantity called 'Moment of Inertia'. Imagine the body as being made up of an infinite number of particles, with masses m_1, m_2 , etc, at distances r_1, r_2 , etc from the axis of rotation. In general, the mass of a typical particle is m_x and its distance from the axis of rotation is r_x . Since the particles do not necessarily lie in the same plane, r_x is specified as the perpendicular distance from the particle to the axis. The total kinetic energy of the body is the sum of the kinetic energy of all its particles:

$$K = \frac{1}{2} m_1 r_1^2 \omega^2 + \frac{1}{2} m_2 r_2^2 \omega^2 + \dots$$

$$= \sum_x \frac{1}{2} m_x r_x^2 \omega^2$$

Taking the common factor $\frac{1}{2}\omega^2$ out of the expression gives:

$$K = \frac{1}{2} \omega^2 (m_1 r_1^2 + m_2 r_2^2 + \dots)$$

$$= \frac{1}{2} \omega^2 (\sum_x m_x r_x^2)$$

The quantity in parenthesis, obtained by multiplying the mass of each particle by the square of the distance from the axis of rotation and adding these products, is called the Moment of Inertia of the body, denoted by I:

$$I = m_1 r_1^2 + m_2 r_2^2 + \dots = \sum_x m_x r_x^2$$

In terms of the moment of inertia (I), the rotational kinetic energy (K) of a rigid body is

$$K = \frac{1}{2} I \omega^2$$

d. **Angular Momentum.** Angular Momentum (L) is defined as the product of Moment of Inertia and Angular Velocity, ie $L = I\omega$.

e. **Gyro Axes.** In gyro dynamics it is convenient to refer to the axis about which the torque is applied as the input axis and that axis about which the precession takes place as the output axis. The third axis, the spin axis, is self-evident. The XX^1 , YY^1 and ZZ^1 axes shown in the diagrams are not intended to represent the x, y and z axes of an aircraft in manoeuvre. However, if the XX^1 (rotational) axis of the gyro is aligned with the direction of flight, the effects of flight manoeuvre on the gyro may be readily demonstrated in similar fashion to the instrument descriptions in Volume 5, Chapter 20.

Classification of Gyroscopes

5. Conventional gyroscopes are classified in Table 1 in terms of the quantity they measure, namely:
 - a. **Rate Gyroscopes.** Rate gyroscopes measure the rate of angular displacement of a vehicle.
 - b. **Rate-integrating Gyroscopes.** Rate-integrating gyroscopes measure the integral of an input with respect to time.
 - c. **Displacement Gyroscopes.** Displacement gyroscopes measure the angular displacement from a known datum.

Table 1 Classification of Gyros

Type of Gyro	Uses in Guidance and Control	Gyro Characteristics
Rate Gyroscope	Aircraft Instruments	Modified single-degree-of-freedom gyro.
Rate-integrating Gyroscope	Older IN Systems	Modified single-degree-of-freedom gyro. Can also be a two-degree-of-freedom gyro.
Displacement Gyroscope	Heading Reference Older IN Systems Aircraft Instruments	Two degrees of freedom. Defines direction with respect to space, thus it is also called a space gyro, or free gyro.

6. It should be realized, however, that the above classification is one of a number of ways in which gyroscopes can be classified. Referring to Table 1, it will be seen that a displacement gyroscope could be classified as a two-degrees-of-freedom gyro or a space gyro. Note also that the classification of Table 1 does not consider the spin axis of a gyroscope as a degree of freedom. In this chapter, a degree of freedom is defined as the ability to measure rotation about a chosen axis.

LAWS OF GYRODYNAMICS

Rigidity in Space

7. If the rotor of a perfect displacement gyroscope is spinning at constant angular velocity, and therefore constant angular momentum, no matter how the frame is turned, no torque is transmitted to the spin axis. The law of conservation of angular momentum states that the angular momentum of a body is unchanged unless a torque is applied to that body. It follows from this that the angular momentum of the rotor must remain constant in magnitude and direction. This is simply another way of saying that the spin axis continues to point in the same direction in inertial space. This property of a gyro is defined in the First Law of Gyrodynamics.

The First Law of Gyrodynamics

8. The first law of gyrodynamics states that:

"If a rotating body is so mounted as to be completely free to move about any axis through the centre of mass, then its spin axis remains fixed in inertial space however much the frame may be displaced."

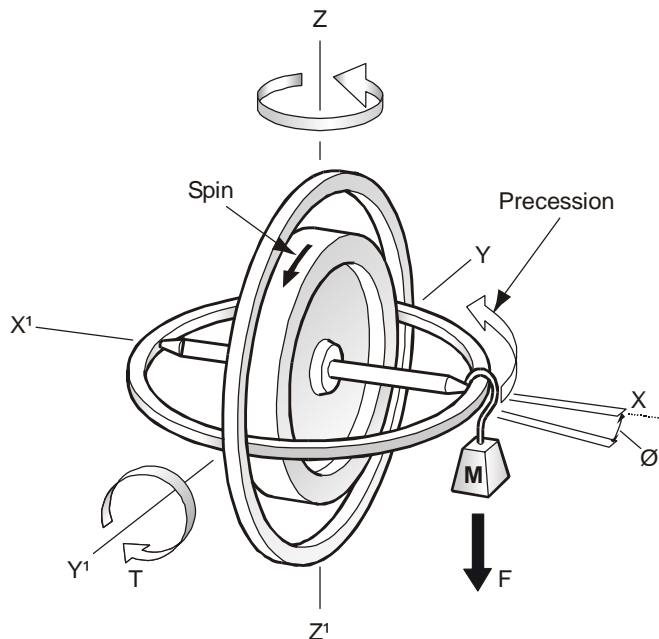
9. A space gyroscope loses its property of rigidity in space if the spin axis is subjected to random torques, some causes of which will be examined later.

Precession

10. Consider the free gyroscope in Fig 3, spinning with constant angular momentum about the XX^1 axis. If a small mass M is placed on the inner gimbal ring, it exerts a downward force F so producing a torque T about the YY^1 axis. By the laws of rotating bodies, this torque should produce an angular acceleration about the YY^1 axis, but this is not the case:

- a. Initially, the gyro spin axis will tilt through a small angle (θ in Fig 3), after which no further movement takes place about the YY^1 axis. The angle θ is proportional to T and is a measure of the work done. Its value is almost negligible and will not be discussed further.
- b. The spin axis then commences to turn at a constant angular velocity about the axis perpendicular to both XX^1 and YY^1 , ie the ZZ^1 axis. This motion about the ZZ^1 axis is known as precession and is the subject of the Second Law of Gyrodynamics.

13-33 Fig 3 Precession



The Second Law of Gyrodynamics

11. The second law of gyrodynamics states that:

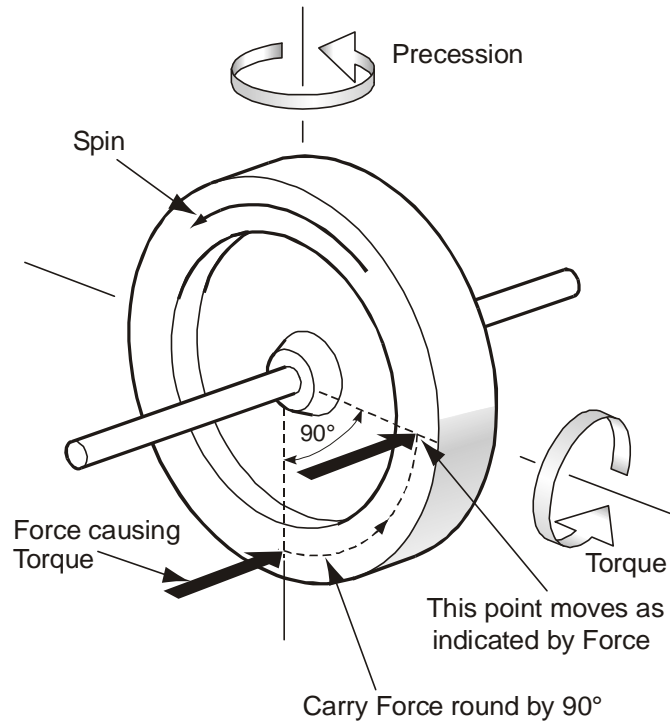
"If a constant torque (T) is applied about an axis perpendicular to the spin axis of an unconstrained, symmetrical spinning body, then the spin axis will precess steadily about an axis mutually perpendicular to the spin axis and the torque axis. The angular velocity of precession (Ω) is given by $\Omega = \frac{T}{I\omega}$."

12. Precession ceases as soon as the torque is withdrawn, but if the torque application is continued, precession will continue until the direction of spin is the same as the direction of the applied torque. If, however, the direction of the torque applied about the inner gimbal axis moves as the rotor precesses, the direction of spin will never coincide with the direction of the applied torque.

Direction of Precession

13. Fig 4 shows a simple rule of thumb to determine the direction of precession:

- a. Consider the torque as being due to a force acting at right angles to the plane of spin at a point on the rotor rim.
- b. Carry this force around the rim through 90° in the direction of rotor spin.
- c. The torque will apparently act through this point and the rotor will precess in the direction shown.

13-33 Fig 4 Determining Precession

CONSERVATION OF ANGULAR MOMENTUM

Explanation

14. In linear motion, if the mass is constant, changes in momentum caused by external forces will be indicated by changes in velocity. Similarly, in rotary motion, if the moment of inertia is constant, then the action of an external torque will be to change the angular velocity in speed or direction and, in this way, change the angular momentum. If, however, internal forces (as distinct from external torques) act to change the moment of inertia of a rotating system, then the angular momentum is unaffected. Angular momentum is the product of the moment of inertia and angular velocity, and if one is decreased so the other must increase to conserve angular momentum. This is the Principle of Conservation of Angular Momentum.

15. Consider the ice-skater starting her pirouette with arms extended. If she now retracts her arms she will be transferring mass closer to the axis of the pirouette, so reducing the radius of gyration. If the angular momentum is to be maintained then, because of the reduction of moment of inertia, the rate of her pirouette must increase, therefore:

- a. If the radius of gyration of a rotating body is increased, a force is considered to act in opposition to the rotation caused by the torque, decreasing the angular velocity.
- b. If the radius of gyration is decreased, a force is considered to act assisting the original rotation caused by the torque, so increasing the angular velocity.

Cause of Precession

16. Consider the gyroscope rotor in Fig 5a spinning about the XX^1 axis and free to move about the YY^1 and ZZ^1 axes. Let the quadrants (1, 2, 3 and 4) represent the position of the rotor in spin at one instant during the application of an external force to the spin axis, producing a torque about the YY^1 axis. This torque is tending to produce a rotation about the YY^1 axis while at the same instant the rotor

spin is causing particles in quadrants 1 and 3 to recede from the YY^1 axis, increasing their moment of inertia about this axis, and particles in quadrants 2 and 4 to approach the YY^1 axis decreasing their moment of inertia about this axis. Particles in quadrants 1, 2, 3 and 4 tend to conserve angular momentum about YY^1 , therefore:

- a. Particles in quadrants 1 and 3 exert forces opposing their movement about YY^1 .
- b. Particles in quadrants 2 and 4 exert forces assisting their movement about YY^1 .

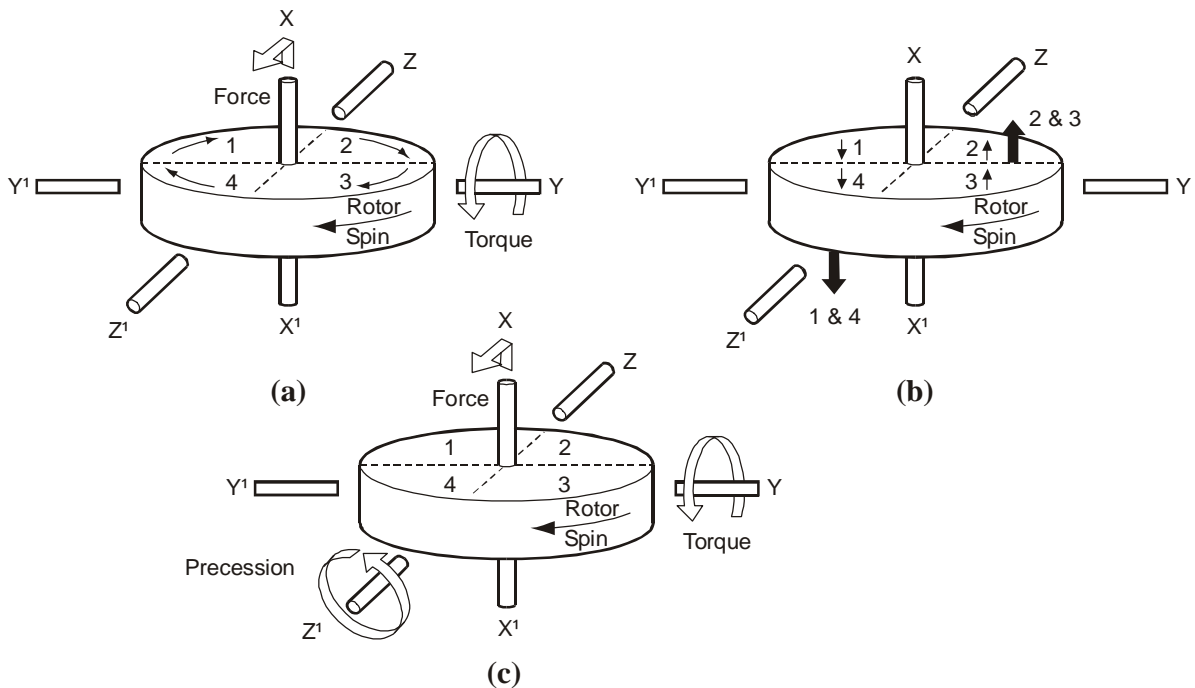
17. Hence, 1 and 4 exert forces on the rotor downwards, whilst 2 and 3 exert forces upwards. These forces can be seen to form a couple about ZZ^1 , (Fig 5b), causing the rotor to precess in the direction shown in Fig 5c.

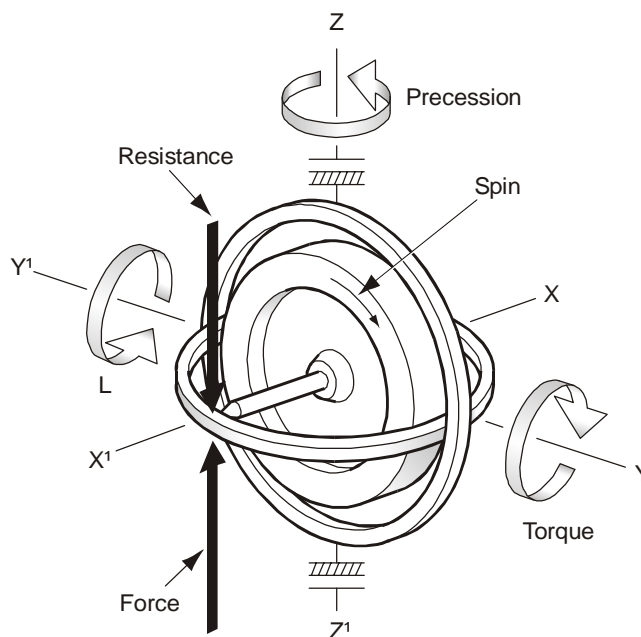
Gyroscopic Resistance

18. In demonstrating precession, it was stated that, after a small deflection about the torque axis, movement about this axis ceased, despite the continued application of the external torque. This state of equilibrium means that the sum of all torques acting about this axis is zero. There must, therefore, be a resultant torque L , acting about this axis which is equal and opposite to the external torque, as shown in Fig 6. This resistance is known as Gyroscopic Resistance and is created by internal couples in a precessing gyroscope.

19. Consider now the gyroscope in Fig 5c spinning about an axis XX^1 and precessing about the ZZ^1 axis under the influence of a torque T , about the YY^1 axis. The rotor quadrants represent an instant during the precession and spin. Using the argument of para 16, the particles in quadrants 1 and 3 are approaching the ZZ^1 axis and exerting forces acting in the direction of precession, while in quadrants 2 and 4 the particles are receding from the ZZ^1 axis and exerting forces in opposition to the precession. The resultant couple is therefore acting about the YY^1 axis in opposition to the external torque. This couple is the Gyroscopic Resistance. It has a value equal to the external torque thus preventing movement about the YY^1 axis.

13-33 Fig 5 Instant of Spin and Precession



13-33 Fig 6 Gyroscopic Resistance

20. Gyroscopic Resistance is always accompanied by precession, and it is of interest to note that, if precession is prevented, gyroscopic torque cannot form, and it is as easy to move the spin axis when it is spinning as when it is at rest. This can be demonstrated by applying a torque to the inner gimbal of a gyroscope with one degree of freedom. With the ZZ¹ axis locked, the slightest touch on the inner gimbal will set the gimbal ring (and the rotor) moving.

Secondary Precession

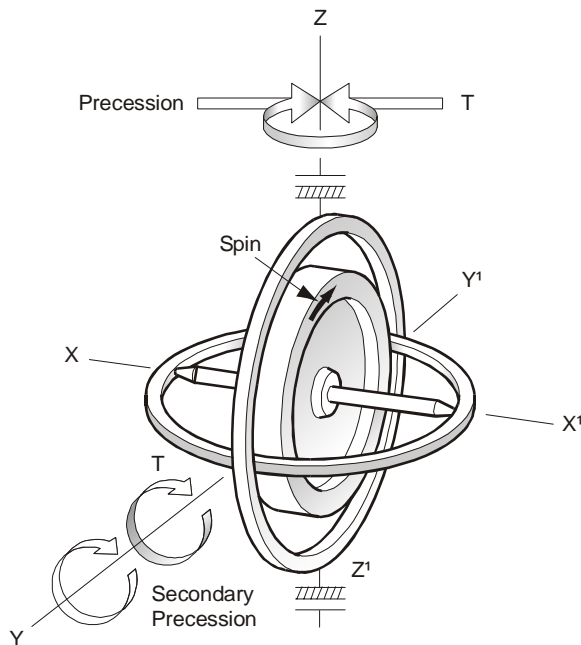
21. If a sudden torque is applied about one of the degrees of freedom of a perfect displacement gyroscope the following phenomena should be observed:

- a. Nodding, or nutation occurs. Here it is sufficient to note that nutation occurs only for a limited period of time and eventually will cease completely. Additionally, nutation can only occur with a two-degree-of-freedom gyro and, to a large extent, it can be damped out by gyro manufacturers.
- b. A deflection takes place about the torque axis, (dip), which remains constant provided that the gyro is perfect, and the applied torque is also constant.
- c. The gyro precesses, or rotates, about the ZZ¹ axis.

22. If, however, an attempt is made to demonstrate this behaviour, it will be seen that the angle of dip will increase with time, apparently contradicting sub-para 21b.

23. To explain this discrepancy, consider Fig 7. If the gyro is precessing about the ZZ¹ axis, some resistance to this precession must take place due to the friction of the outer gimbal bearings. If this torque T is resolved using the rule of thumb given in para 13, it will be seen that the torque T causes the spin axis to dip through a larger angle. This precession is known as secondary precession.

13-33 Fig 7 Precession Opposed by Secondary Precession



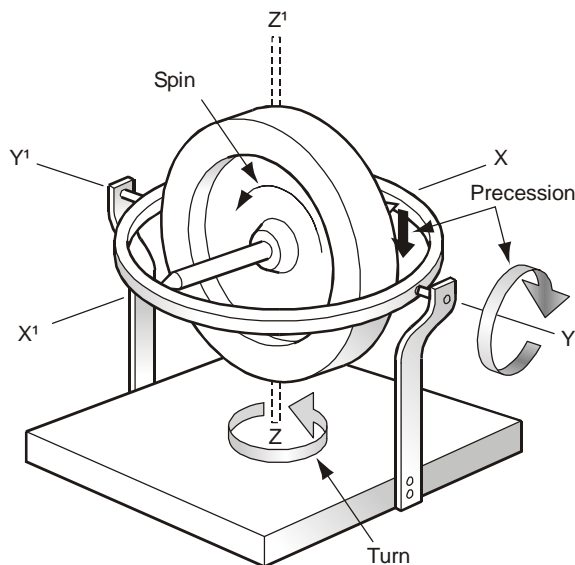
24. Secondary precession can only take place when the gyro is already precessing, thus its name. Note also that secondary precession acts in the same direction as the originally applied torque.

THE RATE GYROSCOPE

Principle of Operation

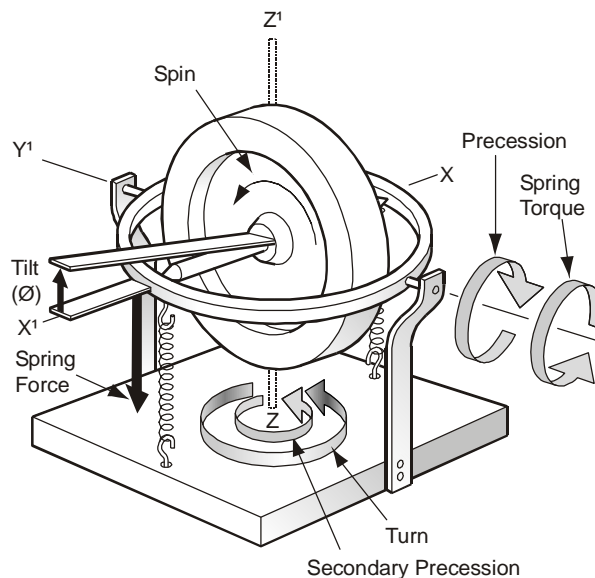
25. Fig 8 shows a gyroscope with freedom about one axis YY^1 . If the frame of the gyro is turned about an axis ZZ^1 at right angles to both YY^1 and XX^1 , then the spin axis will precess about the YY^1 axis. The precession will continue until the direction of rotor spin is coincident with the direction of the turning about ZZ^1 .

13-33 Fig 8 Gyro with One degree of Freedom – Precession



26. Suppose the freedom of this gyroscope about the gimbal axis is restrained by the springs connecting the gimbal ring to the frame as in Fig 9. If the gyroscope is now turned about the ZZ^1 axis, precession about the YY^1 axis is immediately opposed by a torque applied by the springs. It has been shown that any torque opposing precession produces a secondary precession in the same direction as the original torque (see para 24). If the turning of the frame is continued at a steady rate, the precession angle about the YY^1 axis will persist, distending one spring and compressing the other, thereby increasing the spring torque. Eventually, the spring torque will reach a value where it is producing secondary precession about ZZ^1 equal to, and in the same direction as, the original turning. When this state is reached, the gyroscope will be precessing at the same rate as it is being turned and no further torque will be applied by the turning. Any change in the rate of turning about the ZZ^1 axis will require a different spring torque to produce equilibrium, thus the deflection of the spin axis (\emptyset in Fig 9) is a measure of the rate of turning. Such an arrangement is known as a Rate Gyroscope, and its function is to measure a rate of turn, as in the Rate of Turn Indicator.

13-33 Fig 9 Rate Gyroscope



27. The relationship between the deflection angle and rate of turn is derived as follows:

Spring Torque is proportional to \emptyset or

Spring Torque = $K\emptyset$ (where K is a constant)

At equilibrium:

Rate of Secondary Precession = Rate of Turn

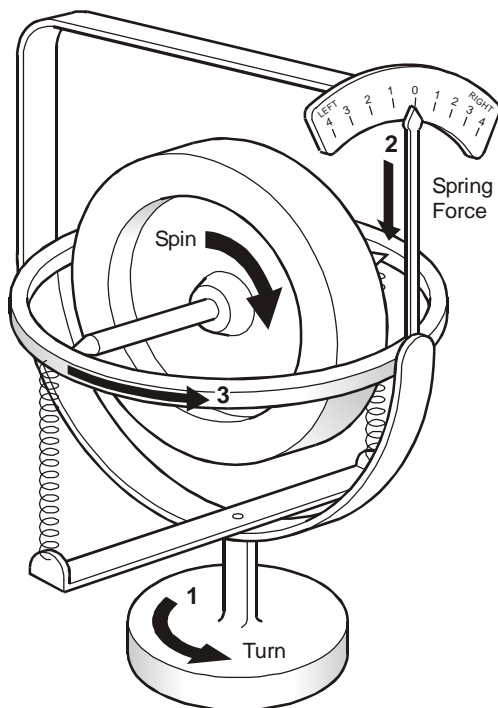
ie $\frac{K\emptyset}{I\omega} = \text{Rate of Turn}$

$\therefore \emptyset$ is proportional to Rate of Turn $\times I\omega$

($I\omega$ is the angular momentum of the rotor and is therefore constant).

The angle of deflection can be measured by an arrangement shown at Fig 10 and the scale calibrated accordingly.

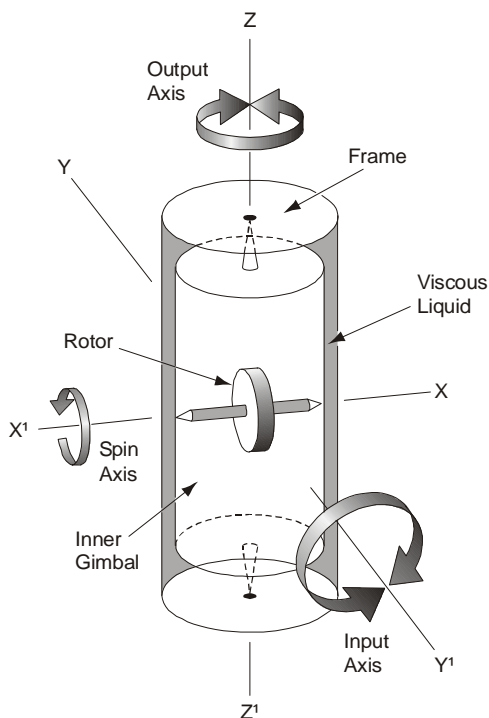
13-33 Fig 10 Rate of Turn Indicator



THE RATE-INTEGRATING GYROSCOPE

Principle of Operation

28. A rate-integrating gyroscope is a single degree of freedom gyro using viscous restraint to damp the precessional rotation about the output axis. The rate-integrating gyro is similar to the rate gyro except that the restraining springs are omitted and the only factor opposing gimbal rotation about the output axis is the viscosity of the fluid. Its main function is to detect turning about the input axis (YY^1 in Fig 11), by precessing about its output axis (ZZ^1 in Fig 11).

13-33 Fig 11 Simple Rate-integrating Gyroscope

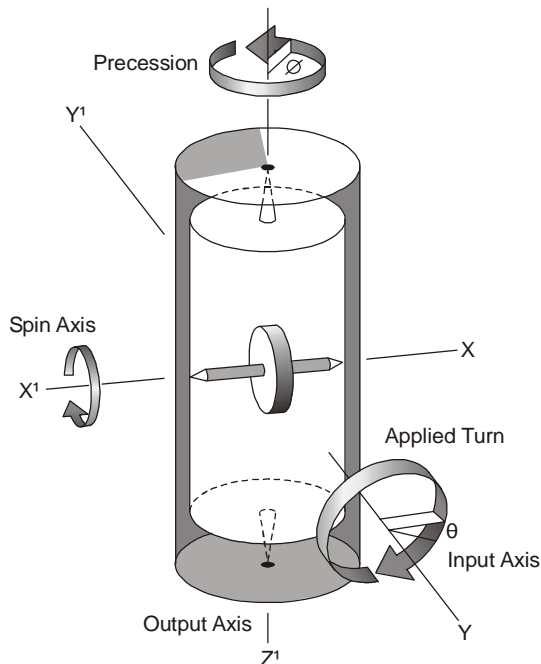
29. The rate-integrating gyro was designed for use on inertial navigation stable platforms, where the requirement was for immediate and accurate detection of movement about three mutually perpendicular axes. Three rate-integrating gyros are used, each performing its functions about one of the required axes. These functions could be carried out by displacement gyros, but the rate-integrating gyro has certain advantages over the displacement type. These are:

- a. A small input rate causes a large gimbal deflection (gimbal gain).
- b. The gyro does not suffer from nutation.

30. Fig 11 shows a simple rate-integrating gyro. It is basically a can within which another can (the inner gimbal) is pivoted about its vertical (ZZ') axis. The outer can (frame) is filled with a viscous fluid which supports the weight of the inner gimbal so reducing bearing torques. The rotor is supported with its spin (XX') axis across the inner gimbal. In a conventional non-floated gyro, ball bearings support the entire gimbal weight and define the output axis. In the floated rate-integrating gyro the entire weight of the rotor and inner gimbal assembly is supported by the viscous liquid, thereby minimizing frictional forces at the output (ZZ') axis pivot points. The gimbal output must, however, be defined and this is done by means of a pivot and jewel arrangement. By utilizing this system for gimbal axis alignment, with fluid to provide support, the bearing friction is reduced to a very low figure.

31. The gyroscope action may now be considered. If the whole gyro in Fig 12 is turned at a steady rate about the input axis (YY'), a torque is applied to the spin axis causing precession about the output axis (ZZ'). The gimbal initially accelerates (precesses) to a turning rate such that the viscous restraint equals the applied torque. The gimbal then rotates at a steady rate about ZZ' , proportional to the applied torque. The gyro output (an angle or voltage) is the summation of the amount of input turn derived from the rate and duration of turn and is therefore the integral of the rate input. (Note that the rate gyro discussed in paras 25 to 27 puts out a rate of turn only). The movement about the output axis may be made equal to, less than, or greater than movements about the input axis by varying the viscosity of the damping fluid. By design, the ratio between the output angle (ϕ) and the input angle (θ) can be arranged to be of the order of 10 to 1. This increase in sensitivity is called gimbal gain.

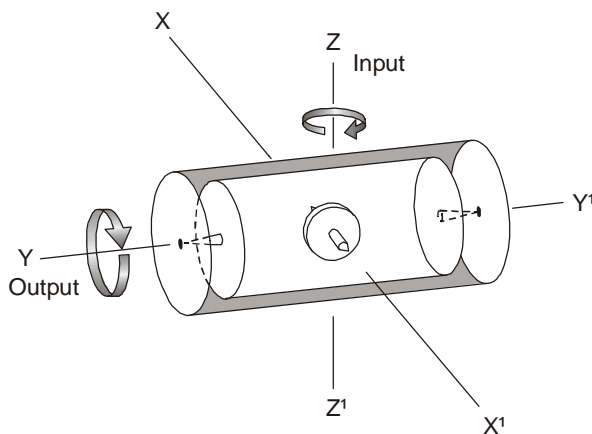
13-33 Fig 12 Function of Rate-integrating Gyroscope



32. A gyro mounted so that it senses rotations about a horizontal input axis is known as a levelling gyro. Two levelling gyros are required to define a level plane. Most inertial platforms using conventional gyros align the input axis of their levelling gyros with True North and East.

33. Motion around the third axis, the vertical axis, is measured by an azimuth gyro, ie one in which the input axis is aligned with the vertical, as in Fig 13.

13-33 Fig 13 Rate-integrating Azimuth Gyroscope



THE DISPLACEMENT GYROSCOPE

Definition

34. A displacement gyro is a two-degree-of-freedom gyro. It can be modified for a particular task, but it always provides a fixed artificial datum about which angular displacement is measured.

Wander

35 Wander is defined as any movement of the spin axis away from the reference frame in which it is set.

36. **Causes of Wander.** Movement away from the required datum can be caused in two ways:

a. Imperfections in the gyro can cause the spin axis to move physically. These imperfections include such things as friction and unbalance. This type of wander is referred to as real wander since the spin axis is actually moving. Real wander is minimized by better engineering techniques.

b. A gyro defines direction with respect to inertial space, whilst the navigator requires Earth directions. In order to use a gyro to determine directions on Earth, it must be corrected for apparent wander due to the fact that the Earth rotates or that the gyro may be moving from one point on Earth to another (transport wander).

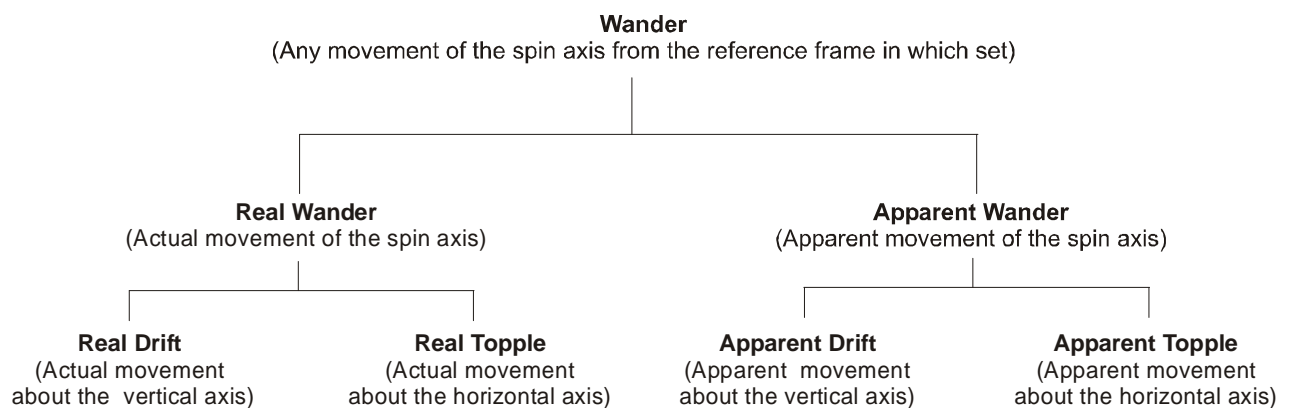
37. **Drift and Topple.** It is more convenient to study wander by resolving it into two components:

a. **Drift.** Drift is defined as any movement of the spin axis in the horizontal plane around the vertical axis.

b. **Topple.** Topple is defined as any movement of the spin axis in the vertical plane around a horizontal axis.

38. **Summary.** Table 2 summarizes the types of wander. From para 36 it should be apparent that the main concern when using a gyro must be to understand the effects of Earth rotation and transport wander on a gyro.

Table 2 Types of Wander



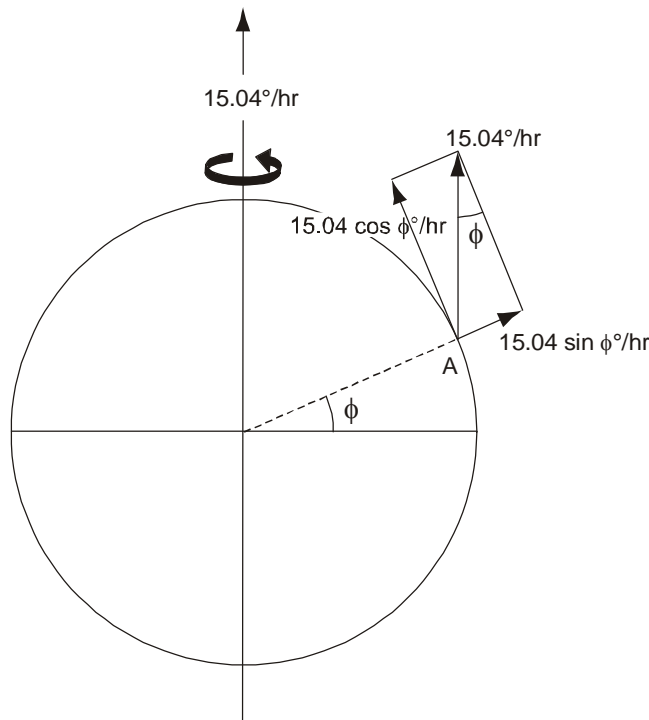
Earth Rotation

39. In order to explain the effects of Earth rotation on a gyro it is easier to consider a single-degree-of-freedom gyro, since it has only one input and one output axis. The following explanation is based on a knowledge of rotational vector notation.

40. Consider a gyro positioned at a point A in Fig 14. It would be affected by Earth rotation according to how its input axis was aligned, namely:

- a. If its input axis was aligned with the Earth's spin axis, it would detect Earth rate (Ω_e) of 15.04 °/hr.
- b. **Azimuth Gyro.** If its input axis was aligned with the local vertical it would detect $15.04 \times \sin \phi$ °/hr, where ϕ = latitude. Note that, by definition, this is drift.
- c. **North Sensitive Levelling Gyro.** If its input axis were aligned with local North, it would detect $15.04 \times \cos \phi$ °/hr. Note that, by definition, this is topple.
- d. **East Sensitive Levelling Gyro.** Finally, if the input axis were aligned with local East, that is, at right angles to the Earth rotation vector, it would not detect any component of Earth rotation.

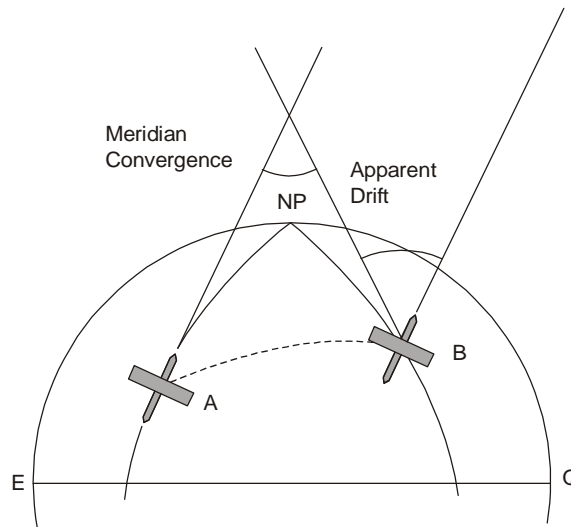
13-33 Fig 14 Components of Earth Rate



Transport Wander

41. If an azimuth gyro spin axis is aligned with local North (i.e. the true meridian) at A in Fig 15 and the gyro is then transported to B, convergence of the meridians will make it appear that the gyro spin axis has drifted. This apparent drift is in addition to that caused by Earth rotation. The gyro has not in fact drifted; it is the direction of the True North which has changed. However, if the gyro is transported North-South, there is no change in the local meridian and therefore, no apparent drift. Similarly, as all meridians are parallel at the Equator, an East-West movement there produces no apparent drift. Transport rate drift thus depends on the convergence of the meridians and the rate of crossing them; i.e. the East-West component of ground speed (U). The amount of convergence between two meridians (C) is $ch \text{ long} \times \sin \text{ lat}$. Any given value of U thus produces an increase in apparent gyro drift as latitude increases.

13-33 Fig 15 Apparent Drift



The amount of drift due to transport rate may be found as follows:

$$C \text{ (}^\circ/\text{hr)} = [\text{ch long/hr}] \times \sin \phi.$$

$$\text{Now, ch long/hr} = \frac{\text{ch Eastings (nm/hr)}}{60} \times \sec \phi$$

and, since $1^\circ = 60 \text{ nm}$ and $\text{ch Eastings (nm/hr)} = U$

$$C = \frac{U}{60} \times \sec \phi \times \sin \phi \text{ (}^\circ/\text{hr)}$$

$$\text{but, } \sec \phi \times \sin \phi = \frac{1}{\cos \phi} \times \sin \phi$$

$$= \frac{\sin \phi}{\cos \phi} = \tan \phi$$

$$\therefore C = \frac{U}{60} \times \tan \phi \text{ (}^\circ/\text{hr)}$$

This can be converted to radians/hour by multiplying by $\frac{\pi}{180}$

$$\therefore C = \frac{U}{60} \times \tan \phi \times \frac{\pi}{180} = U \times \tan \phi \times \frac{\pi}{60 \times 180}$$

Now an arc of length 60 nm on the Earth's surface subtends an angle of 1° ($\pi/180^\circ$) at the centre of the Earth

$$\therefore R \times \frac{\pi}{180} = 60 \text{ where } R = \text{Earth's radius}$$

$$\text{or, } \frac{1}{R} = \frac{\pi}{60 \times 180}$$

Substituting into the above equation for Meridian Convergence (radians/hour)

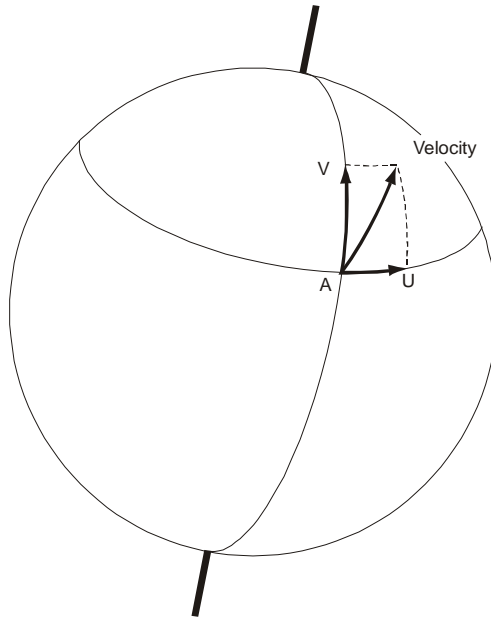
$$C = U \times \tan \phi \times \frac{1}{R}$$

$$\text{or, } C = \frac{U}{R} \times \tan \phi \text{ (radians/hour)}$$

42. Consider now two levelling gyros, whose input axes are North and East respectively, and whose output axes are vertical.

- a. The East component of aircraft velocity in Fig 16 will be sensed by the North gyro as a torque of $\frac{U}{R}$ about its input axis. If the gyro is not corrected for this transport wander, it is said, by definition, to topple.
- b. Similarly, due to the effect of aircraft velocity North, the East gyro will topple at the rate of $\frac{V}{R}$.

13-33 Fig 16 Transport Wander



Apparent Wander Table

43. All of the equations derived in the study of Earth rate and transport wander rate are summarized in Table 3. The units for Earth rate can be degrees or radians, whilst for transport wander they are radians.

44. **Correction Signs.** The correction signs of Table 3 apply only to the drift equations, and they should be applied to the gyro readings to obtain true directions. These correction signs will be reversed for the Southern Hemisphere.

Table 3 Components of Drift and Topple – Earth Rate and Transport Wander Rate

	Input Axis Alignment			Correction Sign
	Local North	Local East	Local Vertical	
Earth Rate degrees (or radians) per hour	$\Omega_e \cos \phi$	Nil	$\Omega_e \sin \phi$	+
Transport Wander radians per hour	$\frac{U}{R}$	$\frac{-V}{R}$	$\frac{U}{R} \tan \phi$	+E -W
	Topple		Drift	

Ω_e = Angular Velocity of the Earth

R = Earth's Radius

ϕ = Latitude

U = East/West component of groundspeed

V = North/South component of groundspeed

Practical Corrections for Topples and Drift

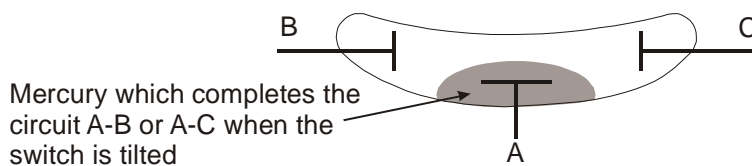
45. If all the corrections of Table 3 were applied to three gyros with their input axes aligned to true North, true East and the local vertical, true directions would be defined continuously, and in effect the gyros would have been corrected for all apparent wander. However, these corrections make no allowance for the real wander of a gyro and consequently an error growth proportional to the magnitude of the real drift and topple will exist. As a rough rule of thumb, an inertial platform employing gyros with real drift rates in the order of 0.01°/hr will have a system error growth of 1 to 2 nm/hr CEP.

46. Flight instruments, on the other hand, employ cheaper, lower quality gyros whose drift rates may be in the order of 0.1°/hr. If these real drift rates were not compensated for, system inaccuracies would be unacceptably large. For this reason, some flight instruments make use of the local gravity vector to define the level plane, thus compensating for both real and apparent drifts.

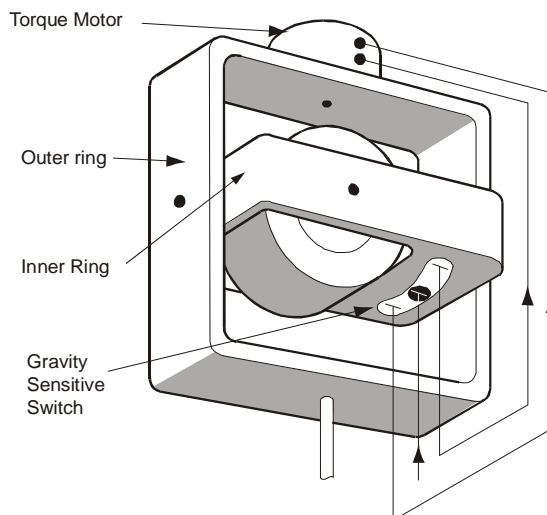
47. Specifically, gyro wander may be corrected in the following ways:

- a. **Topple.** Topple is normally corrected for in gyros by the use of either gravity switches (see Figs 17 and 18), or by case levelling devices (see Fig 19). These devices sense movement away from the vertical and send appropriate signals to a torque motor until the vertical is re-established. The levelling accuracy of these methods is approximately 1°.
- b. **Drift.** Drift corrections can be achieved by:
 - (1) Calculating corrections using Table 3 and applying them to the gyro reading.
 - (2) Applying a fixed torque to the gyro so that it precesses at a rate equal to the Earth rate for a selected latitude. Although this method is relatively simple, it has the disadvantage that the compensation produced will only be correct at the selected latitude.
 - (3) Applying variable torques, using the same approach as in (2) above, but being able to vary the torque according to the latitude. These azimuth drift corrections make no allowance for real drift, which can only be limited by coupling the azimuth gyro to a flux valve.

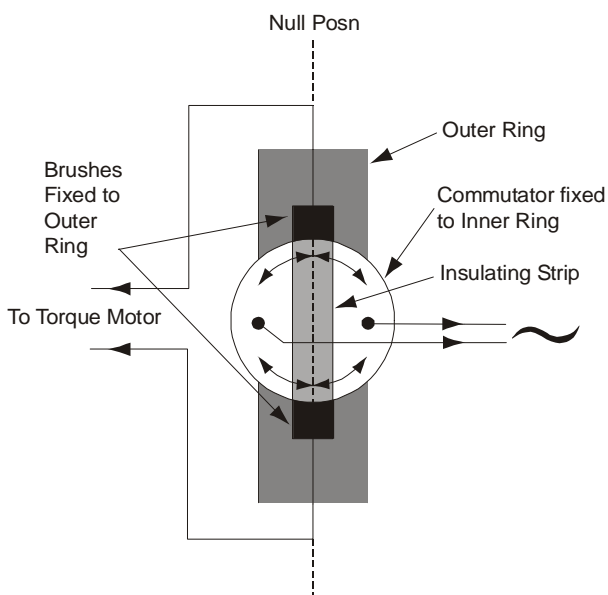
13-33 Fig 17 Gravity Sensitive Switch



13-33 Fig 18 Gravity Levelling



13-33 Fig 19 Synchro Case Levelling Device



48. To complete this study of the displacement gyro, it remains to mention a limitation and an error peculiar to this type of gyro, namely gimbal lock and gimbal error.

Gimbal Lock

49. Gimbal lock occurs when the gimbal orientation is such that the spin axis becomes coincident with an axis of freedom. Effectively the gyro has lost one of its degrees of freedom, and any attempted movement about the lost axis will result in real wander. This is often referred to as toppling, although drift is also present.

Gimbal Error

50. When a 2-degree-of-freedom gyroscope with a horizontal spin axis is both banked and rolled, the outer gimbal must rotate to maintain orientation of the rotor axis, thereby inducing a heading error at the outer gimbal pick-off. The incidence of this error depends upon the angle of bank and the angular difference between the spin axis and the longitudinal axis and, as in most systems, the spin axis direction is arbitrary relative to North, the error is not easily predicted. Although the error disappears

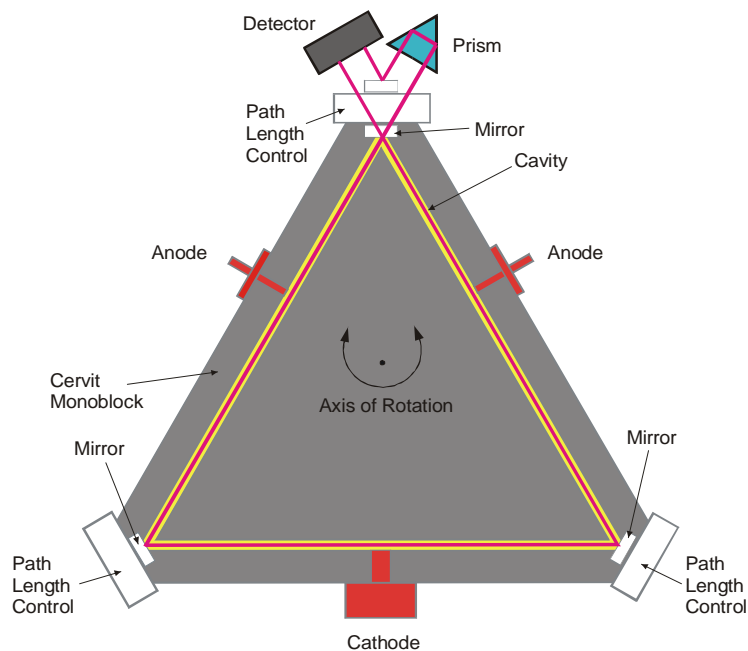
when the aircraft is levelled, it will have accumulated in any GPI equipment, producing a small error in computed position.

OTHER TYPES OF GYRO

The Ring Laser Gyro

51. **Introduction.** The ring laser gyro is one of the modern alternatives to conventional gyros for a number of applications including aircraft inertial navigation systems and attitude/heading reference systems. The ring laser gyro has no moving parts and is not a gyroscope in the normal meaning of the word. The ring laser gyro is, however, a very accurate device for measuring rotation and became the system of choice for use with strapdown inertial navigation systems. A schematic diagram of a ring laser gyro is at Fig 20.

13-33 Fig 20 Schematic Diagram of a Ring Laser Gyro



52. **Principle of Operation.** In ring laser gyros, the rotating mass of the conventional gyro is replaced by two contra-rotating beams of light. The main body of the gyro consists of a single piece (or 'monoblock') of a vitreous ceramic of low temperature coefficient (typically 'Cervit' or 'Zerodour'). A gas-tight cavity is accurately machined into the monoblock and this cavity is then filled with an inert gas, typically a mixture of Helium and Neon. A DC electrical discharge ionizes the gas and causes the lasing action. Two beams of light are produced, flowing in opposite directions in the cavity. Mirrors are used to reflect the beams around the enclosed area, producing a 'laser-in-a-ring' configuration. The frequency of oscillation of each beam corresponds to the cavity resonance condition. This condition requires that the optical path length of the cavity be an integral number of wavelengths. The frequency of each beam is therefore dependant on the optical path length.

53. **Effect of Movement.** At rest, the optical path length for each beam is identical; therefore, the frequencies of the two laser beams are the same. However, when the sensor is rotated about the axis perpendicular to the lasing plane, one beam travels an increased path length, whilst the other travels a reduced path length. The two resonant frequencies change to adjust to the longer or shorter optical path, and the frequency difference is directly proportional to the rotation rate. This phenomenon is

known as the Sagnac effect. The frequency difference is measured by the beaming of an output signal for each wave on to photo detectors spaced one quarter of a wavelength apart, causing an optical effect known as an interference fringe. The fringe pattern moves at a rate that is directly proportional to the frequency difference between the two beams. It is converted to a digital output, where the output pulse rate is proportional to the input turn rate, and the cumulative pulse count is proportional to the angular change. This effect can be quantified using simplified maths, where it can be shown that the frequency difference Δf of the two waves is:

$$\Delta f = \frac{4A\Omega}{\lambda L}$$

Where

- A is the area enclosed by the path.
- λ is the oscillating wavelength.
- L is the length of the closed path.
- Ω is the rate of rotation.

54. **Gyro Control.** The reason why the two beams have to occupy the same physical cavity is the sensitivity of laser light to cavity length. If they did not, a temperature induced difference in path length could result in a large frequency mismatch between the two beams. The path length control mechanism is used to alter the intensity of the laser and thus control expansion due to excess heat. To help avoid perceived differences in path length due to flow of the Helium/Neon gas mix, two anodes are used to balance any flow caused by ionization.

55. **Error Sources.** Ring laser gyros are subject to a number of errors, the most notable of which are:

a. **Null Shift.** Null shift arises due to a difference in path length as perceived by the two opposite beams, thus producing an output when no rotation exists. The major causes of perceived path length difference are:

(1) Differential movement of the gas in the cavity.

(2) Small changes in the refractive index of the monoblock material as the direction of travel of the laser light changes.

b. **Lock-in.** Lock-in occurs when the input rotation rate of the gyro is reduced below a critical value causing the frequency difference between the clockwise and anti-clockwise beams to drop to zero. One of the main causes of this phenomenon is backscatter of light at the mirrors. Some of the clockwise beam is reflected backwards, thereby contaminating the anti-clockwise beam with the clockwise frequency. Similarly, backscattering of the anti-clockwise beam contaminates the clockwise beam. With low input rotational rates, the two beams soon reach a common frequency which renders detection of rotation impossible. Several methods are used to ensure that lock-in is minimized. One method is to physically dither the gyro by inputting a known rotation rate in one direction, immediately followed by a rotation rate in the opposite direction. As the dither rate is known, it can be removed at the output stage. The dither ensures that the two frequencies are kept far enough apart to avoid lock-in.

56. **Advantages of the Ring Laser Gyro.** The main advantages of the ring laser gyro are:

a. Its performance is unaffected by high 'g'.

b. It has no moving parts and therefore has high reliability and low maintenance requirements.

- c. It has a rapid turn-on time.

57. **Disadvantages of the Ring Laser Gyro.** The technical problems associated with ring laser gyros can all be overcome. However, solution of these problems inevitably increases costs which are already very high due to the complex, 'clean-room' manufacturing facilities needed to provide:

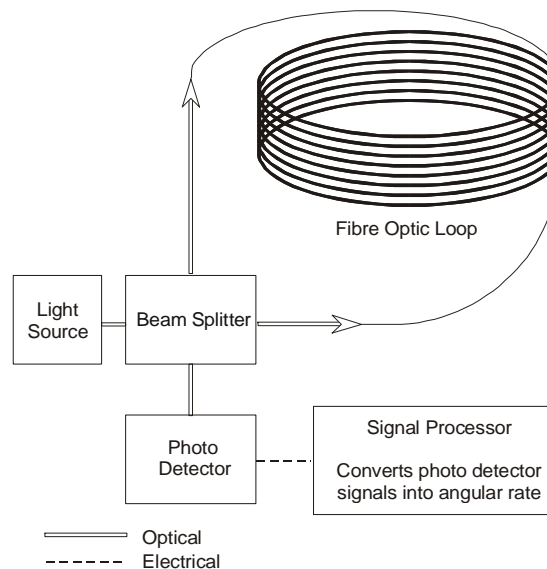
- a. Precision machining and polishing.
- b. High quality mirrors.
- c. Very good optical seals.
- d. A carefully balanced mix of Helium and Neon, free of contaminants.

58. **Summary.** While the ring laser gyro represents a major advance over the traditional spinning gyro, it is only one of a number of possible alternatives. The search for new gyroscopic devices continues, driven by considerations of both cost and accuracy.

Fibre Optic Gyros

59. As previously outlined, the major disadvantage of ring laser gyros is their high cost due to the precise engineering facilities required to manufacture them. The fibre optic gyro (see Fig 21), first tested in 1975, works on the same principal as the ring laser gyro (the Sagnac effect) but no longer relies on a complex and costly block and mirror system since it uses a coil of fibre optic cable.

13-33 Fig 21 Fibre Optic Gyro

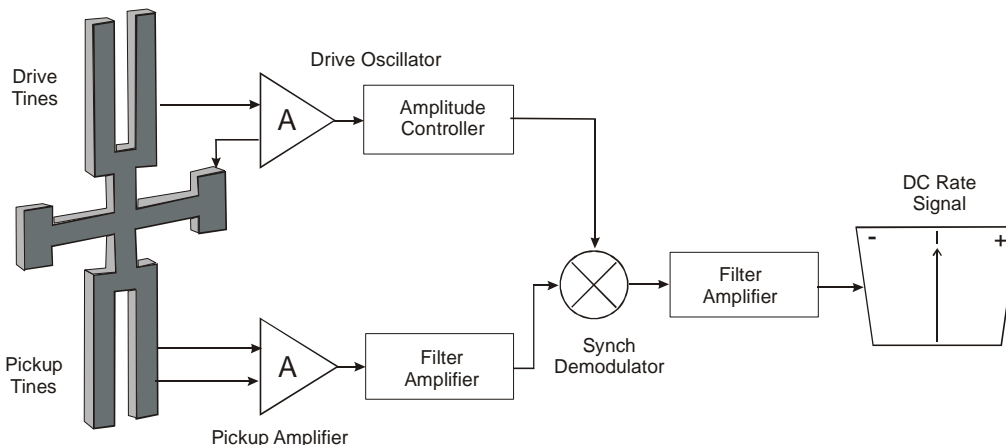


Vibrating Gyros

60. Vibrating gyros work by exploiting the Coriolis effect and, while not yet as accurate as optical gyros, are much smaller and cheaper to produce.

61. The 'GyroChip' (Fig 22) uses a vibrating quartz tuning fork as a Coriolis sensor, coupled to a similar fork as a pickup to produce the rate output signal. The piezoelectric drive tines are driven by an oscillator to vibrate at a precise amplitude, causing the tines to move toward and away from one another at a high frequency. This vibration causes the drive fork to become sensitive to angular rate about an axis parallel to its tines, defining the true input axis of the sensor.

13-33 Fig 22 The 'GyroChip' Vibrating Gyro



62. Vibration of the drive tines causes them to act like the arms of a spinning ice skater, where moving them in causes the skater's spin rate to increase and moving them out causes a decrease in rate. An applied rotation rate causes a sine wave of torque to be produced, resulting from 'Coriolis Acceleration', in turn causing the tines of the Pickup Fork to move up and down (not toward and away from one another) out of the plane of the fork assembly.

63. The pickup tines thus respond to the oscillating torque by moving in and out of plane, causing electrical output signals to be produced by the Pickup Amplifier. These signals are amplified and converted into a DC signal proportional to rate by use of a synchronous switch (demodulator) which responds only to the desired rate signals. The DC output signal of the 'GyroChip' is directly proportional to input rate, reversing sign as the input rate reverses, since the oscillating torque produced by Coriolis reverses phase when the input rate reverses.