BASIC MATHEMATICS FOR RADIO AND ELECTRONICS

F. M. COLEBROOK

B.Sc., D.I.C., A.C.G.I.

Additional chapters by

J. W. HEAD

M.A. (Cantab)

Foreword by

Professor G. W. O. HOWE

D.Sc., LL.D., M.I.E.E.

LONDON: ILIFFE & SONS LTD

NEW YORK: PHILOSOPHICAL LIBRARY

First published	1946
Second edition	1949
Third edition	1957

Published for "Wireless World" by Iliffe & Sons, Ltd., Dorset House, Stamford Street, London, S.E.1

Published in the U.S.A. by Philosophical Library, Inc., 15 East 40th Street, New York 16, N.Y.

Made and printed in England at The Chapel River Press, Andover, Hants

BKS 2841

CONTENTS

FOREWORD

IT is a pleasure to me to contribute a foreword to this book written by an old student of mine. It is well entitled "basic mathematics" for one of the outstanding features of the book is its fundamental character. The laying of sure foundations is very essential in mathematics. As the author says in the introduction, a great many people have been compelled by force of circumstances to take up the study of radio with a very scanty mathematical background, and to these the book should make a very special appeal. In no branch of engineering is a thorough knowledge of basic mathematics more essential than in radio. By its very nature radio involves the study of rapid variations of current and voltage of various degrees of complexity, and a student cannot hope to get a clear understanding of even the simplest problems unless he has a thorough grasp of the fundamentals of differential and integral calculus and of vector algebra. A valuable feature of the book is the gradual development of the subject step by step, and the pains taken at every step to endow the symbols with definite meaning. It is all too easy for a student to attain some facility in the manipulation of symbols without having any real understanding of the physical realities involved. It is important for the student to clothe his symbols with physical reality and to appreciate the implications of each step of the mathematical development, and this the author has done his utmost to encourage.

Although written primarily for radio engineers it would be a great mistake to assume that it is not suited to students of other subjects. The first six of the seven chapters into which the book is divided are quite general, and it is only in the final chapter that the mathematical methods developed in the earlier chapters are applied to radio problems and even these are largely of a general character that will interest students of any branch of physics or engineering.

G. W. O. HOWE

Glasgow University

INTRODUCTION

A VERY long while ago—longer than I care to remember—I was invited to write a series of articles for the *Wireless Engineer*, or rather for *Experimental Wireless* as it was called in those days, on " Mathematics for Wireless Amateurs ". I gladly accepted the invitation because I had already discovered that the best way of learning a subject is to write a book about it. So it proved in this case, and I have ever since been grateful that I was thus constrained to seek out and describe in simple language the true inwardness of the mathematical methods that I had for years been using in my daily work. So also it seems were some readers of the series, judging by kind comments and letters received then and since. I take this opportunity of thanking them, for it is largely due to this encouragement that I have been given the opportunity of turning the series into a book.

The original articles were written for those whose understanding of " wireless ", as we called it then, was limited by a lack of knowledge of mathematics. Hence the "radio" in the title, and those sections which deal particularly with the analysis of alternating-current circuits ; but the really basic ideas of mathematics are common to all its applications, and I therefore hope that the work may be of some use to an even larger class of students—and perhaps even to some teachers. (If this sounds presumptuous, I can say in excuse that there were teachers among those who commended the original series.)

May I anticipate certain criticisms by stating briefly the considerations that have guided both the selection of the material and the manner of its treatment.

Since it was intended to include in a single work all the main parts of elementary mathematics, each of which has usually a textbook all to itself, the choice lay between, on the one hand, a fairly thorough discussion of the basic ideas, with little room left for the detail of their development and application, and, on the other, a condensed statement of rules and formulæ for use in a more or less rule-of-thumb way; but this last is surely a second-class way, suitable only for second-class minds. Even if space had been no object I would still have chosen to emphasise the basic ideas rather than their detailed development and application. In mathematics, as in many other activities, it is the first step which counts. If the basic ideas are well and truly grasped, their application is a happy hunting ground for the adventurous mind, which will indeed learn more by going astray on its own than by being carefully guided along well-worn tracks. (The important thing is, not to avoid mistakes, but never to make the same mistake twice.) What I have done therefore is to select in each branch of the subject those elementary but fundamental ideas which I have found, in many years of practical experience in radio work, to be definitely necessary or specially useful. Further, I have tried at each stage to link up these basic ideas with the real world of sensory experience from which they were, of course, originally derived, however abstract they may appear to be. In this connection I am glad to acknowledge my great indebtedness to Professor G. W. O. Howe, whose early teaching inculcated in me this realistic and critical habit of mind.

Since the original series was written, Mr. Lancelot Hogben has shown how completely an apparently academic, not to say esoteric, subject can be vitalised and humanised by a natural and evolutionary method of description. I mention his work, not to challenge comparison, which would be foolish, but rather to claim a modest kinship, at least in respect of attitude and intention. I commend it to future writers of textbooks, in the spirit of the man who hung up in his chicken-run an ostrich egg marked with the words " Keep your eye on this and do your best ".

Teddington,
Middlesex

F. M. Colebrook

INTRODUCTION

A VERY long while ago—longer than I care to remember—I was invited to write a series of articles for the *Wireless Engineer*, or rather for *Experimental Wireless* as it was called in those days, on " Mathematics for Wireless Amateurs ". I gladly accepted the invitation because I had already discovered that the best way of learning a subject is to write a book about it. So it proved in this case, and I have ever since been grateful that I was thus constrained to seek out and describe in simple language the true inwardness of the mathematical methods that I had for years been using in my daily work. So also it seems were some readers of the series, judging by kind comments and letters received then and since. I take this opportunity of thanking them, for it is largely due to this encouragement that I have been given the opportunity of turning the series into a book.

The original articles were written for those whose understanding of " wireless ", as we called it then, was limited by a lack of knowledge of mathematics. Hence the "radio" in the title, and those sections which deal particularly with the analysis of alternating-current circuits ; but the really basic ideas of mathematics are common to all its applications, and I therefore hope that the work may be of some use to an even larger class of students—and perhaps even to some teachers. (If this sounds presumptuous, I can say in excuse that there were teachers among those who commended the original series.)

May I anticipate certain criticisms by stating briefly the considerations that have guided both the selection of the material and the manner of its treatment.

Since it was intended to include in a single work all the main parts of elementary mathematics, each of which has usually a textbook all to itself, the choice lay between, on the one hand, a fairly thorough discussion of the basic ideas, with little room left for the detail of their development and application, and, on the other, a condensed statement of rules and formulæ for use in a more or less rule-of-thumb way; but this last is surely a second-class way, suitable only for second-class minds. Even if space had been no object I would still have chosen to emphasise the basic ideas rather than their detailed development and application. In mathematics, as in many other activities, it is the first step which counts. If the basic ideas are well and truly grasped, their application is a happy hunting ground for the adventurous mind, which will indeed learn more by going astray on its own than by being carefully guided along well-worn tracks. (The important thing is, not to avoid mistakes, but never to make the same mistake twice.) What I have done therefore is to select in each branch of the subject those elementary but fundamental ideas which I have found, in many years of practical experience in radio work, to be definitely necessary or specially useful. Further, I have tried at each stage to link up these basic ideas with the real world of sensory experience from which they were, of course, originally derived, however abstract they may appear to be. In this connection I am glad to acknowledge my great indebtedness to Professor G. W. O. Howe, whose early teaching inculcated in me this realistic and critical habit of mind.

Since the original series was written, Mr. Lancelot Hogben has shown how completely an apparently academic, not to say esoteric, subject can be vitalised and humanised by a natural and evolutionary method of description. I mention his work, not to challenge comparison, which would be foolish, but rather to claim a modest kinship, at least in respect of attitude and intention. I commend it to future writers of textbooks, in the spirit of the man who hung up in his chicken-run an ostrich egg marked with the words "Keep your eye on this and do your best".

Teddington,
Middlesex

F. M. Colebrook

REVISER'S PREFACE

WHEN I was asked by the publishers to undertake the revision of this book by the late Mr. Colebrook, under whom I once worked at the National Physical Laboratory, Teddington, I very soon became aware that any alterations (except the correction of a few trivial misprints) would merely spoil the original. The first seven chapters have therefore been preserved in their entirety; Section 95 has been somewhat expanded.

Chapters 8 and 9 contain new matter not in the original text for which I alone am responsible. The elements of operational calculus and matrices (with not more than two rows and columns) have been discussed, together with a number of miscellaneous topics, such as numerical computation and normal distributions. The object has been to make the reader's first encounter with these subjects sufficiently encouraging to enable him to face without fear textbooks which deal more adequately with operational calculus, matrices etc. As far as possible, I have tried to make the symbols and notation used consistent with that of the first seven chapters.

Mathematics when unknown and unfamiliar is rather a frightening and unwelcome subject to most engineers. The chief aim of both the original and the revised parts of this book is to remove mathematical inhibitions, and to reveal the power of mathematics to clarify and simplify practical work.

Teddington,
Middlesex

J. W. HEAD

Chapter 1

ELEMENTARY ALGEBRA: THE FUNDAMENTAL IDEAS

1. Symbols

ONE of the most characteristic things about algebra is its appearance, and to a beginner it is certainly not prepossessing. In place of concrete and understandable numbers, to which we have learned to attach definite meanings, we are confronted with such things as

$$a + b = c.$$

The immediate and natural reaction is like that of the little girl who knew her tables up to twelve times and was asked what was three times thirteen—"Don't be silly. It doesn't exist". The writer's earliest recollection of algebra was precisely like that, which shows that the matter was not clearly explained to him, or at least not clearly enough. The whole point is that the letters of ordinary algebra are not really being used as letters at all. They are just symbols which stand for numbers, and it so happens that the letters of the alphabet are very convenient symbols to use because they have an agreed shape and known names. In addition, certain other symbols are used which are either a short way of making statements (for instance, the symbol "=" is only a short way of writing "is the same as" or "is equal to") or else instructions to do certain things with the numbers represented by the letters.

2. Algebra As A Generalisation of Arithmetic

Bearing in mind the real character of the letter symbols used in algebra, we can understand this statement taken from Chrystal's textbook. "Ordinary algebra is simply the general theory of those operations with quantity of which the operations of ordinary arithmetic are a particular case. The fundamental laws of this algebra are therefore to be sought for in ordinary arithmetic." This, then, let us proceed to do.

3. Addition

What do we really mean by the addition of two numbers in ordinary arithmetic? Briefly, a number is a group of ones that we know by name. Adding two numbers means finding the name of the group which contains as many ones as the two groups put together. Thus we know (by memory now, but originally by trial with fingers or beans) that the group two combined with the group three has the same number of ones as the group that we have agreed to call five.

The above example is an ideal case, concerned with pure numbers. In practice, however, we shall not be concerned with pure numbers but with numbers of things—volts, amperes, pounds, shillings and pence, cabbages or kings—and here we come to one of the most important rules in the whole of mathematics.

Things can only be added together in the arithmetical sense if they are things of the same kind.

For instance, three apples can be added to two apples, and the resulting group can be called five apples. But can three apples be added to two oranges? Yes, in the sense that they can all be put into the same dish, but the number five cannot be attached to the group—unless indeed you call them five fruits, but then you are obeying the general rule, for you have obliterated the distinction between the two kinds of things, and have really added three fruits to two fruits; in other words, the things have been regarded as of the same kind. But obviously this can only be done if the distinction between the two kinds is unimportant for the purpose in mind, and in radio problems this never happens.

The above is the essence of what is known by the rather impressive title of "The Theory of Dimensions". It will be considered more fully later on, but for the present it will be enough to realise that if the working out of a given problem in wireless leads to the conclusion

$$L+R=10 \text{ ohms},$$

where L means some number of microhenries and R means some number of ohms, then the result is wrong without any further consideration, because it adds together two numbers of things of different kinds and calls the result a single number of one of the kinds.

In the ordinary arithmetic series of numbers that we know by simple names, one, two, three, four, five, and so on, the successive members get larger and larger in an orderly, uniform, and simple way. For this reason, numbers are a very convenient scale or "yard-stick" for describing size or quantity. This may appear to be an extension of the simple idea of pure number, but in fact it is a direct application of it, the only difference being that each kind of quantity has its own kind of "one" or "unit". Thus a current of 5 amperes means a current having 5 units or "ones" of current, the unit for this kind of quantity being called the ampere. Moreover, just as in pure numbers we have a simple way of describing large numbers in terms of certain special groups of ones—tens, hundreds, thousands and so on, so also in some kinds of quantity we have a series of special names for larger and larger groups of the unit of that quantity—inches, feet, yards, for example; and it is just too bad that our "scales of notation" in quantities of various kinds are not the "decimal" scale that we use for numbers. Perhaps the less said about this the better.

Now we can go on to the finding out of the generalised rules of arithmetical addition, and the easiest way will be to fix on some one particular kind of thing that we can make a picture of either mentally or actually. Then, on the understanding that all our numbers, or symbols standing for numbers, mean numbers of this particular thing, we shall be obeying the fundamental rule about addition, and the conclusions arrived at will apply generally to any form of arithmetical addition. A convenient thing will be a travel or a journey of, say, one inch in some definite direction, this direction being from left to right

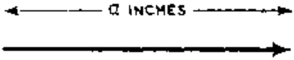


Fig. 1—Representing the number *a* by a journey of *a* inches in a horizontal direction to the *right*

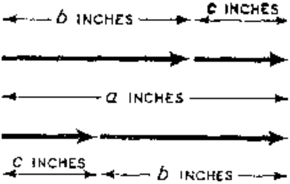


Fig. 2—Showing that a journey *a* can be made up of journey *b* and then journey *c*, or journey *c* first and then journey *b*

parallel to the bottom edge of the page. This may seem a curious thing to choose, but it will be found later that it is very suitable for finding out the rules about subtraction also, which are not so easy to understand as those of addition.

Any number, say three, of these things will mean three journeys of one inch added together, and since the adding of two journeys means starting the second from the finishing point of the first, this will be the same as a journey of three inches in the given direction. In general any number a will mean a journey of a inches in the given direction (Fig. 1). Suppose now that the journey a is carried out in two stages, first a journey b and then a journey c, as shown in the upper part of Fig. 2. Then a can be described as the result of adding the journey c to the journey b, that is,

$$a = b + c.$$

Further, it is clear that it does not matter which of the two journeys b or c is made first (lower part of Fig. 2), so that

$$a = b + c = c + b.$$

In the same way it will be just as easy to show that

$$\begin{aligned} b + c + d &= b + d + c \\ &= c + b + d \\ &= d + c + b, \text{ etc., etc.,} \end{aligned}$$

d being another journey of d inches added to the other two. The same idea could be extended to any number of journeys without alteration and we thus arrive at the second general rule about addition.

A succession of additions will lead to the same result whatever the order in which they are carried out.

4. The Use of Brackets

Returning to the statement

$$a = b + c,$$

this expresses the idea that the two journeys b and c can be considered as a single journey. Similarly, any number of journeys b, c, d, e, etc. can, if desired, be associated together and considered to be a single journey $b + c + d + e +$ etc. Or, again, two of these journeys can be considered as a single journey, if that is convenient for some particular purpose, leaving the others as separate journeys. When it is desired to consider any particular group of journeys (or numbers) as a single thing, that group can be enclosed between two brackets thus $(c+d)$, and that means that we will take this as a whole, without regard to the fact that it is actually made up of two parts. Thus the compound group of numbers a, b, c, d can be written in the form

$$a + b + c + d$$

or in the form

$$a + b + (c + d),$$

the c and d numbers being, so to speak, wrapped up into a single brown paper parcel. This is called the Law of Association for addition, though it is certainly very difficult to see why it should be called a law at all, any more than the wrapping up of things in a brown paper parcel should be called the Law of Brown Paper. However, the idea of associating certain sets of numbers together by means of brackets is a useful one in practice.*

5. The Addition of Double Groups

Before leaving this subject of addition it may be as well to return for a moment to our three apples and two oranges,

* The associative law for addition is usually expressed: $a + (b + c) = (a + b) + c$.

because they can be used to illustrate a very important extension of the idea of addition, one which will prove useful in connection with alternating current circuits.

It has been shown that the group three apples plus two oranges cannot be expressed in any simpler form, since the two parts of the group cannot be combined in the sense of arithmetical addition. Two such double groups can be so combined, however. For instance, three apples plus two oranges combined with four apples plus six oranges can obviously be expressed as one double group, seven apples plus eight oranges, that is, the two sets of apples can be added and the two sets of oranges can be added. To generalise this idea, suppose the letters a, b, c, etc., to represent numbers of apples, and the Greek letters α, β, γ, etc., to represent numbers of oranges, and suppose that the working out of a problem concerned with these double groups of apples and oranges leads to the statement

$$a + \alpha + b + c + \beta + \gamma = d + \delta.$$

Then the number d on the right-hand side must be the result of adding together all the apples on the left-hand side, and similarly the number δ must represent all the oranges on the left-hand side. The statement is therefore equivalent to two separate statements

$$a + b + c = d$$
$$\alpha + \beta + \gamma = \delta.$$

If this simple idea is thoroughly assimilated, the reader will find that he has got a firm footing in the "complex number" or "operator" method of working out alternating current circuits.

6. Subtraction

The idea of subtraction in the ordinary arithmetical sense is one with which life makes us almost painfully familiar. Returning to the ideal case, we know (by memory) that if two ones are subtracted or removed from a group of five ones, then what remains is the group that we have agreed to call three, and if the ones are, say, pounds, then our understanding of the process is intensified in some cases by its emotional associations. For a child, actually carrying out the process with fingers or beans, the matter ends quite definitely when all the five ones have been subtracted, and if he is asked to take away any more after that he will say, quite rightly, " It can't be done ". Later on, however, he will be forced to acquire the idea of a negative number when he finds that seven pounds have been subtracted from his five pounds, leaving him with a debt of two pounds. Mathematically speaking he would now be said to possess minus two (−2) pounds, and assuming him to be an honest man, he will realise that if later on he earns two more pounds, these must be set against his debt of two pounds, leaving him the possessor of exactly nothing. Expressing this mathematically,

$$-2 + 2 = 0$$

and since this is true whatever the magnitude of the debt and of the equal amount that has to be earned to set against it,

$$-a + a = 0$$

where a means any number.

This statement is the most general way of saying what is meant by the negative number $-a$. It is that number which, when added to or combined with the positive number a makes the total result nothing. The actual or practical meaning of the negative number $-a$ will therefore depend on, will in fact be, in the sense indicated above, the reverse of the meaning that we attach to the positive number a.

The form of expression should be noticed carefully. The word " subtraction " does not come into it. We are actually going to deal with subtraction, but it will be found that when this operation is given the wider sense that it has in algebra it will be much clearer and more convenient to think of it as the combination or addition of positive and negative numbers.

What is the meaning of the negative number $-a$ in terms of the things or units that were used in finding out the rules about addition? The positive number a was

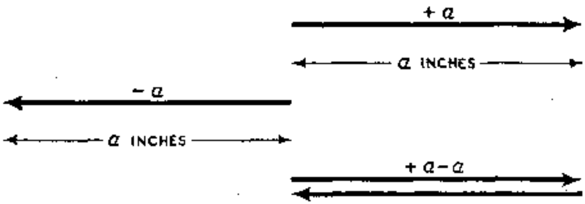


Fig. 3—Showing how the number $-a$ is represented by a journey of a inches in a horizontal direction to the *left*

shown to represent a journey of a inches to the right, and the negative number $-a$ is such that

$$-a + a = 0.$$

In other words, adding the number a to the negative number $-a$ cancels the effect of the negative number, that is, brings us back to the starting point. This becomes quite intelligible if the negative number $-a$ means a journey of a inches to the left. In fact this is the only way to make it intelligibly consistent with the given meaning of $+a$ (see Fig. 3). One of the most important rules of the negative sign now becomes clear. Applying the negative sign to the number a reverses the direction of the journey of a inches which this number represents. Applying the negative sign to the number $-a$ will therefore reverse it again, which brings it back to its original direction, that is,

$$- - a = + a.$$

Putting this shortly in words, *minus minus* is *plus*, and in signs

$$- - = +.$$

Here we have a familiar acquaintance in academic dress, the familiar acquaintance being the phrase, "Two negatives make an affirmative".

7. Some Special Symbols

Before going on to consider the combination of positive and negative numbers in general, it will be convenient to learn a little more shorthand (for that is what the use of mathematical signs really amounts to).

(i) $$a > b$$

means that the number a is greater than the number b. Notice that the thin edge of the wedge points to b, which is, so to speak, the thinner of the two numbers.

(ii) $$a < b$$

means that the number a is smaller than the number b. The same mnemonic* will serve.

There are some variations of these signs which are useful for numbers which can vary in size under given conditions.

(iii) $$a \ngtr b$$

means that a is *not greater than* b, that is, a can be any number equal to or less than b. Notice that the "not" is suggested by the crossing out of the "greater than" sign.

(iv) Similarly $$a \nless b$$

means that a is *not less than* b, that is, can be any number equal to or greater than b.

(v) In the same way $$a \neq b$$

means that a is not equal to b.

(vi) Finally, it is necessary to define a new idea, or a new word, which comes to the same thing. *The difference between two numbers is that number which must be added to the smaller of the two to make it equal to the larger.* Thus the difference between two and five is three, three being the number which must be added to the smaller, two, to make it equal to the larger, five. We are not here concerned with sign at all, nor with the combination of positive and negative numbers. Difference expresses a simple arithmetical relation between the sizes of numbers, regardless of sign. In symbols the difference between a and b is written

$$a \sim b$$

*Mnemonic means "assistance to memory".

and since the difference between a and b, as defined, is the same number as the difference between b and a,

$$a \sim b = b \sim a.$$

8. Combinations of Positive and Negative Numbers

The general case of the combination of positive and negative numbers can now be considered, and with the help of the ideas already acquired it should not prove difficult.

The combination of the positive number a and the negative number $-b$ can be written

$$+a - b,$$

or more shortly still,

$$a - b.$$

In terms of our journey units this means a journey of a inches to the right followed by a journey of b inches to the left. This is illustrated in Fig. 4 (a) for the case $a > b$. The length of the resulting journey will obviously be the difference between a and b, that is $(a \sim b)$, and its direction, that is, sign, will depend on whether we go more to the right than to the left or *vice versa*. In other words, the sign of the resulting journey will be the sign of the greater of the two numbers, a and b, just as in a human partnership the stronger of the two partners will get things his own way. In the case illustrated in Fig. 4 (a), a wins, that is, the sign of the resultant journey is positive. In symbols,

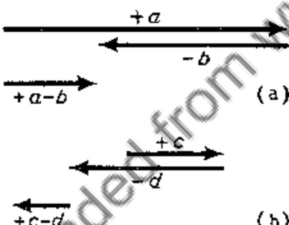


Fig. 4

$$a - b = +(a \sim b).$$

On the other hand, for two other numbers c and d, such that $c < d$ (Fig. 4 (b))

$$c - d = -(c \sim d),$$

or the resultant journey is to the left or negative.

In words, *the combination of a positive and a negative number* *is another number the size of which is the difference of the two numbers, and the sign of which is the sign of the greater of the two numbers.*

A further thing to notice is that it does not matter which of the two journeys is made first. The result will be the same. If this is not immediately obvious it can be made a fact of experience by actually drawing out and measuring several such cases, as in Fig. 5. This conclusion can be stated

$$+a - b = -b + a.$$

Further, in the case of several journeys such as

$$+a - b - c + d - e + f, \text{ etc.,}$$

any arrangement of these can be got by changing round two at a time, and since changing round two at a time does not affect the result, any arrangement of these journeys will have the same result. Finally, as these journey units satisfy the fundamental rule about addition (being things all of the same kind), the above statement may be completely generalised as follows :—

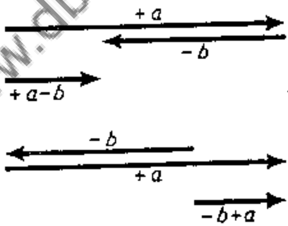


Fig. 5

The result of adding any number of positive and negative numbers is not affected by the order in which the various positive and negative numbers are added.

This is known as the Law of Commutation (*cf.* Section 3).

How is the rule of signs to be applied to a rather complicated group of journeys (or numbers) like the one just considered ? The easiest way is to take all the positive journeys together and all the negative journeys together, just as in making up accounts one adds up all the credits together and then all the debits together.

First, therefore, the journey will be arranged in the form

$$+a+d+f-b-c-e.$$

Now all the positive journeys, or journeys to the right, can be combined together and considered as the single journey to the right,

$$+(a+d+f).$$

Similarly all the negative journeys, or journeys to the left, can be combined together. The result will be a single negative journey the length of which is equal to that of all the separate journeys put together, that is, the sum of the separate negative journeys. This single negative journey is therefore

$$-(b+c+e).$$

The total journey can now be expressed as the combination of these groups, that is,

$$\begin{aligned} +a-b-c+d-e+f \\ = +a+d+f-b-c-e \\ = (a+d+f)-(b+c+e), \end{aligned}$$

and the rule of sign can be applied to the two separate groups. Thus if $(a+d+f)$ is greater than $(b+c+e)$ the sign of the total combination will be positive, and *vice versa*.

9. Brackets and the Negative Sign

It was shown in the preceding section that the three negative journeys $-b$, $-c$, and $-e$ could be combined together into the single negative journey represented by $-(b+c+e)$, and since $(b+c+e)$ is the same as $(+b+c+e)$, then this grouping process is represented by

$$-b-c-e=-(+b+c+e)$$

and since there is no reason why this process should not be reversed

$$-(+b+c+e)=-b-c-e.$$

This shows that if there is a negative sign in front of a bracketed expression, and the brackets are taken away, then all the positive signs inside the brackets must be changed to negative signs.

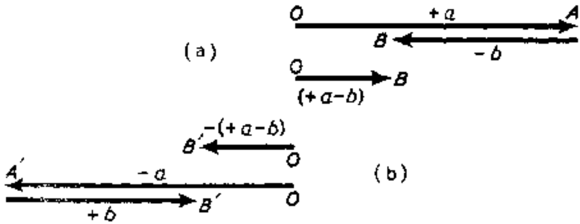


Fig. 6

Consider now the combination of journeys $+ a - b$ shown in Fig. 6 (a). This means the journey O to A and then A to B. Considered as a single journey it is the journey $(+ a - b)$, that is, the journey O to B. Applying the negative sign to this journey reverses its direction and makes it the journey O to B' shown in Fig. 6 (b), that is,

$$OB = (+ a - b)$$
$$OB' = - (+ a - b).$$

If now it is wished to dissociate the journey OB' into two parts, one of which shall be of length a, and the other of length b, there is only one way to do it. The journey a, the longer of the two, must be made to the left, and the journey b to the right. That is,

$$OB' = - a + b.$$

This shows that

$$- (+ a - b) = - a + b.$$

The same argument can be applied if a is less than b.

If the bracket contains a number of positive and negative numbers, these can be grouped together in the manner shown in the preceding section. For instance,

$$a - b - c + d + e - f = a + d + e - b - c - f$$
$$= (a + d + e) - (b + c + f) \text{ by Section 3.}$$

Now representing the group $(a + d + e)$ by the single number or journey p, and the group $(b + c + f)$ by the single number or journey q, then

$$a - b - c + d + e - f = p - q$$

and

$$-(a - b - c + d + e - f) = -(p - q)$$
$$= -p + q \quad \text{as already shown.}$$
$$= -(a + d + e) + (b + c + f)$$
$$= -a - d - e + b + c + f$$
$$= -a + b + c - d - e + f.$$

The general rule should now be clear. *If a negative sign is in front of a bracketed group of numbers, and the brackets are taken away, then all the positive signs inside the brackets must be changed into negative signs, and all the negative signs must be changed into positive signs.*

It should already be obvious that if there is a plus sign in front of the brackets, then the brackets can be taken away without altering the signs inside the brackets at all. This only expresses the idea that a combination of positive and negative numbers can be considered as a single number, for example,

$$a - b - c + d = (a - b - c + d)$$
$$= +(a - b - c + d)$$

or, reversing the process,

$$+(a - b - c + d) = a - b - c + d.$$

10. Brackets Within Brackets

Just as a number of parcels can be wrapped up into a larger parcel, and a number of these larger parcels can be wrapped up together into a still larger parcel, so a number of combined numbers can be grouped together into a combination of combined numbers, and these again can be further grouped together indefinitely. For distinctness, each degree of combining or wrapping up is shown by a different sort of bracket. To take an example—combining the numbers

$$(a - b)$$

and $$-(c + d)$$

gives $$[(a - b) - (c + d)].$$

Combining this with a similarly combined negative number

$$-[(e + f) + (g - h)]$$

gives

$$\{[(a - b) - (c + d)] - [(e + f) + (g - h)]\},$$

and combining this with another simple number k gives

$$k + \{[(a - b) - (c + d)] - [(e + f) + (g - h)]\}.$$

In carrying out the reverse process, one would, in dealing with a parcel of parcels, remove one wrapper at a time, starting with the outside one. Dealing with the above combination of combinations in the same way, and remembering that every sub-combination is to be regarded as a separate single thing, the process becomes

$$\begin{aligned} & k + \{[(a - b) - (c + d)] - [(e + f) + (g - h)]\} \\ &= k + [(a - b) - (c + d)] - [(e + f) + (g - h)] \\ &= k + (a - b) - (c + d) - (e + f) - (g - h) \\ &= k + a - b - c - d - e - f - g + h, \end{aligned}$$

which reveals the fact that k, a, and h are the only positive numbers in the whole group, a point which was not at all obvious in the combined form.

Actually the parcel analogy is not quite complete, because in a case like the one above there is nothing to prevent one from unwrapping some of the inside parcels first without disturbing the outer wrapping. Thus

$$\begin{aligned} & k + \{[(a - b) - (c + d)] - [(e + f) + (g - h)]\} \\ &= k + \{[a - b - c - d] - [e + f + g - h]\}. \end{aligned}$$

One can even rearrange the separate numbers into different inside parcels again, thus

$$k + \{[(a - d) - (c + b)] - [(e - h) + (g + f)]\}.$$

Another inside rearrangement is

$$k + \{[a - (b + c + d)] - [(e + f + g) - h]\},$$

or

$$k + \{[a - (b + c + d)] + [h - (e + f + g)]\},$$

which exhibits the number in a more symmetrical form.

The following is left as a simple exercise for the reader. Show that

$$(a + b + c) - \{(a - b) - [(c - b) + (a - c)]\}$$
$$- [(a + b - c) + (a + b + c) + (c - a - b)] = 0.$$

11. Generalisation of the Letter Symbols

Up to this point it has been assumed that the letter symbols used in algebra represent positive whole numbers, and negative numbers have been represented by prefixing the negative sign to the letter symbol. This was done in order that the laws about association and commutation and the rules relating to the negative sign might be explained as clearly as possible. But now that these things have been explained, there is no reason why the meaning of a letter symbol should be restricted in this way,* and in practice it would be rather inconvenient. In later applications it will be found that the number represented by a given symbol may vary over very wide ranges in accordance with the conditions of the problem, and may be negative for one set of conditions and positive for another. It will be found on examination that the rules developed for the combining of the numbers a, b, c, d, etc., in accordance with the signs written in front of them, will be equally valid whether the actual numbers represented by the letters are taken as positive or negative. For example, on the understanding that the letters mean positive numbers, it was shown that

$$a - (b - c) = a - b + c$$

and this can be confirmed by putting

$$a = 1$$
$$b = 2$$
$$c = 3.$$

Then $(b - c) = 2 - 3 = -1$,

and $a - (b - c) = 1 - -1 = 1 + 1 = 2.$

Also $a - b + c = 1 - 2 + 3 = 2$,

which confirms the statement.

* According to Chrystal this important fact was not generally realised until about 1640.

Now suppose instead that

$$a = 1$$
$$b = 2$$
$$c = -3.$$

Then $(b - c) = 2 - -3 = 2 + 3 = 5,$

and $a - (b - c) = 1 - 5 = -4.$

Also $a - b + c = 1 - 2 + -3 = 1 - 2 - 3 = -4,$

and again the statement is confirmed. As an example of a case in which a single symbol can assume a wide range of values, positive or negative, take

$$a = b - c,$$

where b can be any number from 0 to 100 and c can be the same or any other number from 0 to 100. Here a can be any number from -100 to $+100$.

On the basis of this wider interpretation of the letter symbols, it will be necessary to reconsider the meaning of the "greater than" and "less than" signs (Section 6). When a and b are positive numbers, say $a = 1{,}000$ and $b = 1$, then the statement $a > b$ is quite unambiguous. But suppose $a = -1{,}000$ and $b = 1$. What is meant by "greater than" in a case like this? It is, of course, only a matter of definition and agreement, and it is generally agreed that the "greater than" sign shall mean "more positive than" so that now $b > a$ for the given values. Remember that the wealth of a man who has one pound is greater than that of a man who has a debt of a thousand pounds. That is the sense in which the "greater than" sign is used algebraically. The other signs are similarly interpreted, and need not be considered in detail. If in any case it is desired to indicate the relation between the actual numerical magnitudes of the two numbers without regard to sign, then this must be specifically stated, for example, "$a > b$ numerically". Some writers use for the same purpose the notation

$$|a| > |b|,$$

that is, the symbol is enclosed between strokes to indicate that its numerical magnitude only is being considered.

For the present, therefore, the letter symbols can be taken to mean positive or negative whole numbers. Further extensions of their possible meaning will come in their own time.

12. An Electrical Example

The general meaning of algebraic addition will perhaps be made clearer by a physical example. A man starts from a point 100 feet above sea level and goes for a long walk up hill and down dale, eventually coming back to where he started from. Dividing up his walk into parts where he is either keeping on the same level, going up, or going down, let c_1, c_2, c_3, c_4, etc., completely represent the separate changes of level that he encounters. Then since his total change of level is nothing from start to finish,

$$c_1 + c_2 + c_3 + c_4 + c_5 + \text{etc.} = 0.$$

It is clear that some of these numbers must be positive and some negative, and this is what is implied by the words "completely represent". If an ascent is represented by a positive number, then the algebraic statement is sound if a descent is represented by a negative number. If a_1, a_2, a_3, etc., and d_1, d_2, d_3, etc., be *positive* numbers which represent the actual amounts of the various ascents and descents, then the statement can be put in the form

$$a_1 + a_2 + a_3 + \text{etc.} - d_1 - d_2 - d_3 - \text{etc.} = 0.$$

It is just a matter of definition and agreement which form is used. If the first, which is in a sense more general, then it must be remembered that when actual numbers are put in place of the letters, the numbers must have their proper sign attached to them, according to whether they stand for a rise of so many feet or a fall of so many feet.

The discovery of a scientific law is often, perhaps always, the recognition of an analogy, or an essential similarity between two sets of ideas. If, in the above example, the idea of potential difference is substituted for "change of level", and the journey round and back to the starting point is thought of as going round a closed circuit of electric conductors, then students of electricity will recognise in the example a statement of Kirchhoff's Law about the sum of the voltages in a closed circuit, an ascent representing a forward electro-motive force, and a descent representing a fall of voltage.

13. Understanding or Rule of Thumb

Before going on to the last of the fundamental rules of arithmetic, that is, those concerned with multiplication and division, it may be as well to face, and try to answer, a question that may have arisen in the minds of some readers. "If I learn all the rules, such as the one about 'minus minus equals plus' and so on, and am very careful to go by them always, is it *really* necessary for me to understand them all as completely as you seem to want me to?"

The reader need not wait until he completely understands all the rules before making use of them. But there are several reasons why he should try to understand them as fully as possible.

In the first place, understanding deepens the original impression, and so gives the memory a firmer foothold. Again, understanding gives confidence, real confidence and mastery, and not merely the absence of fear that follows on familiarity. Finally, using rules without understanding them is not treating the mind with proper respect. It is asking it to be satisfied with the position and capabilities of a tram driver, without aspiring to those of the engineer who knows how to lay down fresh tracks and design better machines to run on them.

There is one other thing to be said about this question of understanding the rules, and this introduces what has always seemed to the writer to be one of the most fascinating aspects of mathematics. Borrowing a phrase from the racing stables, the Rules of Mathematics are by Intuition out of Experience. In other words, they result from the mating of certain inborn habits of mind or "modes of perception" with sensory impressions of the outside world. Now many cases will arise where adherence to these rules will take the mind into a region where understanding breaks down, in the sense that no kind of sensory impressions can be found to correspond to the processes of thought. Nevertheless, it will be found that the rules remain valid in passing through this region of darkness, for the processes can be carried still further until they re-emerge into daylight, so to speak, and sensory impressions can again be found to correspond to them. To take an example, which the more advanced readers at least will be able to follow, it is possible, by blindly following the rules, to calculate where the tangents to a circle from a point inside the circle will make contact with the circle. Such tangents cannot be drawn in the ordinary sense of the word, and the points of contact will be " imaginary ". However, continuing the process, the line that passes through these points can also be calculated, still by blindly following the rules, and it will be found that this is a real line, possessing certain useful properties with respect to the point inside the circle. It is difficult to realise the full implications of this process. It suggests that by means of its intuitions the mind can transcend the limitations of sense impressions. It is only in this sense that the mind should be allowed to be satisfied in using rules without understanding them ; and even here the apparently unintelligible parts of the process should not be accepted as an unsolvable mystery but rather as a challenge to further thought. Later on we shall come to a classical example in which a very simple and logical interpretation was found for a set of ideas previously regarded as purely symbolic and untranslatable into any kind of sensory impressions.

14. Multiplication

There are probably many readers who will say or think, " I know all about simple multiplication. Let's get on to the calculus ". The implicit assumption that one needs to know all about multiplication before going on to the calculus is right enough, but such readers are invited to reconsider the first part of the statement. Anyone who knows all about multiplication should be able to give clear answers to the following questions and to many like them :—

1. Since one can multiply volts by amperes and call the product "watts", can one multiply oranges by apples? If not, why not?

2. Since one can multiply inches by inches and call the product "square inches", can one multiply apples by apples and call the product "square apples"? If not, why not?

3. Can one multiply a number by a negative number? (The writer would be prepared to offer long odds against this one being answered correctly.)

The following discussion of the subject is recommended to the attention of all those who feel at all uncertain about the answers to the above questions.

Multiplication in algebra is a generalisation of the corresponding process in arithmetic. Let us therefore go back for a while to the multiplication tables: $3 \times 5 = 15$. Three multiplied by five is fifteen. We learned it by heart a long while ago and have probably never given it another thought. Giving it another thought, we realise that it means

$$3 \times 5 = 3 + 3 + 3 + 3 + 3 = 15,$$

that is, five groups of three ones combined together make up the group that we have agreed to call fifteen. In other words, multiplication *is* addition, a rather special case of addition in which all the numbers added together are the same as one another. We owe a great deal to the ingenious person (probably a Greek or an Arab) who realised that this special case of addition could be written down in a very short way.

Expressing the process in algebra terms, that is, by means of letters which for the present will be taken to mean any *positive* whole numbers,

$$a \times b = a + a + a + a \quad . \quad . \quad .$$

written down b times.

Notice that although a and b are both numbers in the case of pure arithmetical multiplication, there is a difference between them when the process refers to groups of things, which things must of course be things all of the same kind, whether volts, amperes, cabbages or kings, "a" is the number of things in each group, and "b" is the number of groups, so that $a \times b$ means "b" groups of "a" things. On the other hand $b \times a$ means "a" groups of "b" things. We can see that the total number of things is the same in the two cases. Starting with "b" groups of "a" things, take one out of each group. This will give one group of "b" things, and since this process can be repeated until all the "a" things in each group are used up, that is, "a" times, the final result of the rearrangement will be "a" groups of "b" things. In symbols this is

$$(a \times b) = (b \times a).$$

It is one of the most important properties of the process of multiplication, and is called the Law of Commutation. In $(a \times b)$, "a" is called the multiplicand and "b" the multiplier. $(a \times b)$ is called the product of "a" and "b". In practice the explicit multiplication sign or St. Andrew's Cross is often omitted, or replaced by a dot, thus

$$a \times b = a.b = ab.$$

The product $(a \times b)$ can be considered as a single group and as such can be multiplied again, for example, $(a \times b) \times c$. Furthur, it can be shown that

$$(a \times b) \times c = a \times (b \times c).$$

The proof is not immediately obvious, but it will not be given in full as it would take up rather a lot of space. The first step is

$$(a \times b) \times c = (b \times a) \times c.$$

By writing out the right-hand side fully, and by regrouping the symbols, it can be shown that

$$\begin{aligned}(b \times a) \times c &= (b \times c) \times a \\ &= a \times (b \times c),\end{aligned}$$

so that

$$(a \times b) \times c = a \times (b \times c),$$

or

$$(ab)c = a(bc).$$

This is known as the Associative Law for multiplication. By combining this with the Commutative Law it can be shown that

$$abc = bca = cab = cba = \text{etc.},$$

and since the whole argument can then be repeated with $(a \times b \times c) \times d$, and so on indefinitely, we may say at once that *the result of multiplying together any number of positive whole numbers is independent of the order in which the operations are performed.*

Returning to $a \times b$, suppose a is the sum of two numbers c and d, that is,

$$a = c + d$$

Then

$$a \times b = (c + d) \times b$$
$$= (c + d) + (c + d) + (c + d)$$

written down b times.

Therefore by the Associative Law for addition

$$(c + d) \times b = (c + c + c + c + c + \ . \ . \ . \ \text{written down } b \text{ times})$$
$$+ (d + d + d + d + d + \ . \ . \ . \ \text{written down } b \text{ times})$$
$$= (c \times b) + (d \times b),$$

and since

$$(c + d) \times b = b \times (c + d),$$

we have

$$b \times (c + d) = (b \times c) + (b \times d),$$

or

$$b(c + d) = bc + bd.$$

This is known as the Distributive Law, for a fairly obvious reason. The process can clearly be continued. Thus if, in the above,

$$b = e + f$$
$$b(c + d) = (e + f)(c + d)$$
$$= (e + f)c + (e + f)d, \text{ as already shown,}$$
$$= ec + fc + ed + fd.$$

The general character of the process will not be expressed in words. It would take too long. The algebraic symbols tell the whole truth much more concisely.

15. Multiplication and Negative Numbers

In terms of the original definition and of the interpretation which has already been found for $-a$ (see Section 6) the meaning of $-a \times b$, where a and b are positive whole numbers, does not present any difficulty, for

$$(-a) \times b = (-a) + (-a) + (-a) + (-a \ . \ . \ . \ \text{written down } b \text{ times})$$
$$= -(a + a + a + a \ . \ . \ . \text{ etc., written down } b \text{ times}).$$
$$= -(a \times b).$$

But what is $a \times (-b)$? According to the original definition it would be

$$a \times (-b) = a + a + a + a, \text{ etc., written down } -b \text{ times.}$$

But this does not mean anything. It is literally nonsense. Nor can we get over the difficulty by writing

$$a \times (-b) = (-b) \times a = -(b \times a),$$

for this law was only proved for positive numbers. Actually there is no way out of the difficulty, or rather, there is no difficulty. There is simply the plain statement that one cannot multiply a number by a negative number* (see question 3, Section 14). But then, who wants to?

It is true that later on there will often be occasion to pretend to multiply by a negative number. For instance, a number "a" can be multiplied by "b" giving $(a \times b)$. Its sign can then be reversed, giving $-(a \times b)$. To save time this can be described as multiplying by $-b$, but in fact it consists of two quite separate operations.

Actually the writer has never come across any case in which the application of the laws of multiplication to positive or negative numbers indifferently has led to a

* This statement is calculated to raise clouds of dust from the pages of agitated and outraged textbooks. It is possibly a position that could be blown to pieces by the "big guns" of pure mathematics, but this work is being written with a strong practical bias, and the writer hopes to be able to avoid the introduction of incomprehensible ideas into the fundamental definitions. The standard textbooks usually take a perfectly plausible but not very useful way out of this apparent difficulty. Having shown that $a \times b = b \times a$ for positive numbers, they say it *shall* be so for negative numbers also. In other words the symbol $\times$ is so defined. But since we can then no longer understand the process we are not much better off.

false conclusion, but that is almost certainly due to the fact that in every case the apparent process of multiplication by a negative number admits of some alternative explanation of the kind illustrated above. In all that follows, therefore, it will be considered possible to multiply by a negative number on this understanding.

16. The Multiplication of Number Groups. Zero in Multiplication

Consider the product $a \times (b - c)$. In view of the discussion in Section 15, it will be considered permissible, irrespective of the sign of $(b - c)$, that is, whether $b > c$ or $b < c$, to put

$$a \times (b - c) = (b - c) \times a,$$

in which form the right-hand side can be written out in full exactly as in Section 14. In this way it can be shown that

$$(b - c) \times a = (b \times a) - (c \times a),$$

that is,

$$a(b - c) = (b - c)a = ba - ca = ab - ac.$$

If $b > c$, there is another way of arriving at the same result which will give a further conclusion. The group $a \times (b - 1)$ will obviously contain one a less than that represented by the group $a \times b$, so that

$$\{a \times (b - 1)\} + a = (a \times b).$$

Similarly

$$\{a \times (b - 2)\} + (a \times 2) = (a \times b),$$

and so on up to

$$\{a \times (b - c)\} + (a \times c) = (a \times b).$$

Now adding to each equal number the negative number $-(a \times c)$,

$$\{a \times (b - c)\} + (a \times c) - (a \times c) = (a \times b) - (a \times c),$$

and since

$$+(a \times c) - (a \times c) = 0,$$
$$a \times (b - c) = (a \times b) - (a \times c).$$

Now continue the process further, by increasing c until it is equal to b. Then

$$a \times (b - b) = (a \times b) - (a \times b),$$

that is,

$$a \times 0 = 0.$$

Also from the original definition,

$$0 \times a = 0 + 0 + 0 + 0 \text{ etc., . . . written down } a \text{ times}$$
$$= 0.$$

This shows that the symbol 0 obeys the Commutative Law, and that the product of any number with zero is zero.

Returning to the multiplication of number groups, in the product

$$(b - c) \times a = (b \times a) - (c \times a),$$

suppose that a is itself the sum of d and $-e$, that is,

$$a = (d - e).$$

Then

$$\begin{aligned}(b - c) \times (d - e) &= \{b \times (d - e)\} - \{c \times (d - e)\} \\ &= \{(d - e) \times b\} - \{(d - e) \times c\} \\ &= \{(d \times b) - (e \times b)\} - \{(d \times c) - (e \times c)\} \\ &= (d \times b) - (e \times b) - (d \times c) + (e \times c),\end{aligned}$$

that is,

$$(b - c)(d - e) = db - cd - be + ec,$$

in which it is seen that *each number in the first group is multiplied by every number in the second group, signs being combined according to the rules already established* in the section dealing with the combination of positive and negative numbers.

As in the case of addition, the meaning of the letter symbols can now quite legitimately be extended so as to include both positive and negative numbers.

An interesting special case should be noted at this point.

$$\begin{aligned}(x - a)(x - b) &= xx - bx - ax + ab \\ &= xx - (a + b)x + ab,\end{aligned}$$

and if $b = a$,

$$\begin{aligned}(x - a)(x - a) &= xx - (a + a)x + aa \\ &= xx - 2ax + aa.\end{aligned}$$

In this last expression, we have for the first time algebraic symbols and ordinary numbers in association. There is obviously no need whatever for any given expression to be made up either of letters or explicit numbers exclusively. Nothing more need be said about this for it introduces no new ideas, though as a matter of fact it does introduce a new word. A number associated with a letter symbol in the way that -2 is associated with x in $xx - 2x + 4$ is often called the "coefficient" of the given letter. This name is also applied more generally to one member of a product which remains constant while the other is allowed to vary in magnitude under given conditions. Thus if x is allowed to vary, the $-2a$ in $xx - 2ax + aa$ could be called the coefficient of x. It is not an entirely fortunate choice of word, for "efficient" is always used adjectively in common speech, whereas it appears substantively in "co-efficient", the literal meaning of which is "co-worker".

The form of the two products just considered should be very carefully observed, for they play a large part in the practical applications of algebra. It will be seen later that there is a shorter way of writing them, which does not of course affect their general character.

17. Resolution into Factors

The reverse (not the *inverse*, which is something quite different) of the process described in the last paragraph is at least as important as the direct process. It is called "factorisation" or "resolution into factors", and consists of the expression of a more or less complex group of numbers in the form of a product of two or more terms, called "factors". Thus the factors of $xx - (a + b)\, x + ab$ are $(x - a)$ and $(x - b)$. Take for instance $xx - 7x + 12$. This can be written $xx - (4 + 3)\, x + (4 \times 3)$, and, comparing this with the general form, it is clear that the factors of this expression are $(x - 4)$ and $(x - 3)$, that is,

$$xx - 7x + 12 = (x - 4)\,(x - 3).$$

At present the only method available for factorising such expressions is that illustrated in this example, namely, "inspired guesswork", or what the disrespectful or possibly envious poet described as "cunning low, meet for things problematical ". Later on some more certain, though, in consequence, rather less exciting methods will be discovered.

18. The Physical Application of Multiplication

The main interest of those for whom these articles are being written will be the physical application of the processes of mathematics. This is what distinguishes practical or applied mathematics from pure mathematics considered as a science in itself.

Two questions involving the physical meaning of the process of multiplication have already been raised in Section 14, and these questions we are now in a position to answer.

The physical aspect of addition presented no great difficulty. It was clear that things could only be added together in the arithmetical sense if they were things of the same kind. What is the corresponding rule for multiplication? The fact that one cannot multiply three oranges by four apples and equally cannot multiply three apples by three apples suggests that one can multiply neither things of the same kind nor things of different kinds. And yet we talk of multiplying volts by amperes and inches by inches. Does this mean that in some cases one can and in others cannot multiply together things of the same or things of different kinds? That if the result means anything one can do either of these things, and that if it does not one cannot? Some will perhaps be willing to accept this as the true explanation. Actually it is not an explanation at all. It is a mere statement, and not even a true statement.

The pure mathematician would probably dismiss the question as meaningless, saying that multiplication is an operation with pure numbers and that is all there is about it; which is quite true as far as it goes, but not very helpful.

A more useful answer has already been indicated in Section 14. Multiplication is the addition of a number of groups, and, physically, a group means a group of things which must all be of the same kind. Thus one of the factors of a product must be interpretable as a number of groups, and the other as a group of similar things. But how is this to be related to the apparent process of multiplying volts by amperes and calling the product "watts"?

To trace the connection we will take an instance which is sufficiently novel and ridiculous to be free from any preconceptions. Returning to our three apples, let it be supposed that one of these, following a well-known precedent, falls on the head of a philosopher and makes him think. It can even be supposed that the harder it hits him the more it will make him think, so that if it falls from a height of one foot it produces one theory, if from two feet, two theories, and so on. Suppose now that each of these three apples falls on the head of the philosopher from a height of five feet. Each apple now produces five theories. Each apple, so to speak, represents a group of five theories, and to find the total number of theories produced by this profitable though perhaps painful episode, we combine together as many groups of five theories as there are apples; that is, three groups of five theories. Actually, however, life is too short to permit our being as explicit as this on every occasion, and we would quickly form the habit of saying that we multiply the height by the apples and call the product theories, just as we have formed the habit of saying that we multiply the volts by the amperes and call the product watts—a very good habit too, provided we do not make the mistake of thinking that we really mean what we say.

In the electrical example it is a matter of definition and of experience that each ampere in "falling through" a potential difference of five volts would give rise to a group of five watts, and three amperes would therefore give three groups of five watts.

Thus the physical meaning of the process of multiplication is derivable from the ideal or purely mathematical meaning, and the application of multiplication in physics is seen to be of the same essential character as ordinary arithmetical or algebraic multiplication.

Examples 1

1. Distribute (that is, " multiply out ")

   $(a - b)(b - c)(c - a)$.

2. Find the magnitude of the number

   $(xxx - 6xx + 11x - 6)$, when $x = 1, 2, 3$, and 4.

3. For what values of x will $(x - a)(x - b)(x - c)$ be zero ?

4. Find the factors of

   (*a*) $xx - 7x + 10$,
   (*b*) $xx - 3x - 10$,
   (*c*) $xx + 3x - 10$,
   (*d*) $aa - 100ab + 99bb$,
   (*e*) $aa - bb$.

19. Division

Following the same method as for the other fundamental operations, we will approach the general or algebraic idea of division through the familiar ground of arithmetic. What do we really mean when we say that 15 divided by 5 is 3? Actually, of course, it is the mere repetition of a formula by memory, but the real basis of the formula is this: If the group called fifteen is separated out into five equal groups, each of these will be the group that we call three. In other words, the process of dividing 15 by 5 consists of finding that group, five of which combined together will make up the group called 15. Thus there are really three steps in the statement, as the reader will be able to prove to himself if he will think about it carefully enough. The steps are

$$(15 \div 5) \times 5 = 15,$$

because that is what division means, followed by

$$3 \times 5 = 15,$$

as an act of memory, followed by the deduction

$$(15 \div 5) = 3.$$

The important step for our present purpose is the first, because it shows the reader that the formal algebraic definition of division

$$(a \div b) \times b = a$$

is identical in form with that which he already accepts, though perhaps unconsciously, as the definition of division in ordinary arithmetic.

The first thing to notice about the operation of division so defined is that $(a \div b)$ may or may not be a number in the original sense of that word, that is, a characteristic property of a group of things that can be counted. Thus $(15 \div 5)$ is the number 3, but $(4 \div 5)$ is not a number, because there is no simple group 5 of which combined together will make up the group 4.

However, as practical mathematicians, we shall be more interested in quantity than in pure numbers, and in terms of quantitv $(4 \div 5)$ is an idea that presents no difficulty whatever. We can take any quantity, such as a journey of four units (for example, inches) in a given direction and divide it into five equal journeys without turning a hair, let alone splitting one. Then if the number 4 represents the original journey, each of these smaller journeys (call it "a") is such that

$$a + a + a + a + a = 4,$$

that is, $$a \times 5 = 4,$$

so that $$a = (4 \div 5)$$

in terms of our original definition.

It will be convenient to introduce a few more special terms at this point. If $(a \div b) = c$, where c is a number, then a, b, and c are called respectively dividend, divisor, and quotient. The sign $\div$ is not a very convenient one

to use in practice, and $(a \div b)$ is usually written $\frac{a}{b}$ or, to get it on one line, a/b. A group a/b which cannot be expressed as a number, and in which a and b contain no common factor greater than 1 is called a fraction, b being the denominator and a the numerator. In practice these names are often applied to any expression of the form a/b, a useful extension of the meaning of these words—indeed a necessary one, since until a and b are given specified numerical values one cannot say whether a/b is a number or a fraction.

It is important to realise that as far as the interpretation in terms of quantity is concerned there is no essential difference between a whole number and a fraction. In fact the same quantity can be represented by either. Thus a given weight may be represented by the number 7 or the fraction 7/16 according as to whether it is measured in ounces or pounds. Here the representation in the first place by a number and in the second by a fraction is seen to mean nothing more than a difference in the unit of measurement.

One very important conclusion follows at once from this. The Laws of Commutation and Association and the rules relating to sign in the addition and subtraction of numbers were stated and demonstrated in terms of magnitudes—lines or journeys of various lengths represented by symbols which in their turn represented numbers. It would make no difference to the reasoning employed if the units and lengths had been such as to require the use of the fractional form for the numerical description of the various magnitudes. We can therefore proceed at once to the conclusion that the Laws of Commutation and Association and the rules of sign apply without modification to the addition and subtraction of fractions, and the symbols used in the relevant preceding sections can be given this additional freedom of interpretation. They may represent positive or negative whole numbers or fractions.

Giving to the word fraction the extended meaning which has been indicated above, the determining of the rules of algebraic division really amounts to finding out how such quantities enter into the operations of addition, subtraction, and multiplication.

First, however, it will be well to consider how the negative sign will enter into the operation of division as defined above, for it will be a convenience if the letter symbols used in the remainder of this section can be given their unrestricted meaning as positive or negative numbers.

20. The Negative Sign in Division

The statement

$$(-a \div b) \times b = -a$$

presents no difficulty. By interpreting it in terms, for instance, of the journey units used in connection with simple addition and subtraction, it can easily be shown that

$$(-a \div b) = -(a \div b)$$

that is,

$$(-a/b) = -(a/b)$$

This is division *of* a negative number. But is division *by* a negative number a comprehensible operation? If so, what does it mean? In any case, its meaning must be consistent with our definition of division, so we can start with

$$(a \div -b) \times -b = a$$

which brings us back to multiplication by a negative number. Now an explanation has already been given (Section 15) for the apparently incomprehensible operation of multiplication by a negative number, and it has been shown that no mistake will arise from using the formal rule

$$(a \times -b) = -(a \times b).$$

Applying this rule in the above case we have

$$a = (a \div -b) \times -b = -\{(a \div -b) \times b\}.$$

Therefore

$$(a \div -b) \times b = -a.$$

Comparing this with the previous result, that is

$$(-a \div b) \times b = -a,$$

we see that the apparent process of division by a negative number will be consistent with our interpretation of multiplication by a negative number if $(a \div -b)$ is taken to mean the same thing as $-(a \div b)$, so that

$$(a \div -b) = (-a \div b) = -(a \div b).$$

Applying these formal rules to the case $(-a) \div (-b)$ will obviously give

$$(-a) \div (-b) = (a \div b).$$

Thus we have the three general rules for the combination of signs in division :—

Minus divided by Plus is Minus.
Plus divided by Minus is Minus.
Minus divided by Minus is Plus.

These, however, will not impose any additional burden on the memory, for they are the same as those for multiplication.

Now that an intelligible interpretation has been found for division of or by a negative number, together with the rules that apply, the letter symbols may be taken to represent any positive or negative numbers as practical conditions may require. On this basis we can proceed to deduce formulæ appropriate to various operations with the fractional form.

21. Various Operations with Fractions

In the following paragraphs the name "fraction" is taken to mean any combination of the form $(a \div b)$, where a and b are positive or negative numbers. The reader is asked to obliterate from his mind as completely as may be any preconceptions about fractions based on imperfectly or even misunderstood rules remembered from school days, and to think instead of the operation of division as defined completely and exclusively by the statement

$$(a \div b) \times b = a.$$

(a) *Reduction of a fraction to its lowest terms.*

Suppose
$$a \times p = c$$
$$b \times p = d.$$

Then
$$(a \times p) \div (b \times p) = (c \div d).$$

and by the definition of division,

$$(a \times p) = \{(a \times p) \div (b \times p)\} \times (b \times p),$$

that is, $(a \times p) = (c \div d) \times (b \times p)$.

Therefore $(a \times p) = \{(c \div d) \times b\} \times p$,

and $\quad a = (c \div d) \times b.$

But $\quad a = (a \div b) \times b.$

Therefore $(a \div b) = (c \div d)$,

that is, $(a \times p) \div (b \times p) = (a \div b)$,

or
$$ap/bp = a/b.$$

This shows that if the numerator and denominator of a fraction contain a common factor, this factor can be removed from each without altering the magnitude of the fraction. This process is called the reduction of a fraction to its lowest terms.

(b) *Distribution of denominator.*

By the Law of Distribution in multiplication,

$$\{(a \div c) + (b \div c)\} \times c = \{(a \div c) \times c\} + \{(b \div c) \times c\}$$
$$= a + b.$$

Therefore, by definition,

$$(a \div c) + (b \div c) = (a + b) \div c,$$

or $$\frac{a}{c} + \frac{b}{c} = \frac{a+b}{c}.$$

It will be a useful exercise for the reader to prove for himself the statement

$$c \div (a + b) = (c \div a) + (c \div b).$$

If he succeeds the truth is not yet in him, for the proposition is false. The attempt to prove it will assist in impressing its untruth.

(*c*) *The addition of fractions.*

The results stated in (*a*) and (*b*) taken together show how any two fractions can be added together and expressed as a single fraction, for

$$\frac{a}{b} + \frac{c}{d} = \frac{ad}{bd} + \frac{cb}{db} = \frac{ad}{bd} + \frac{bc}{bd} = \frac{ad+bc}{bd}.$$

The extension of this process to give

$$\frac{a}{b} + \frac{c}{d} + \frac{e}{f} = \frac{adf + cbf + ebd}{bdf}$$

is obvious, but the reader is advised to check it for himself. A little further thought will show that these formulæ are illustrations of the fact that only groups of like things can be combined arithmetically.

In virtue of the Law of Association the addition of any number of fractions can be represented as a single fraction by repeating the above process as often as necessary. It should be noted that the numerator of the combined fraction is obtained by combining according to sign the products of each numerator with all the other denominators, the sign being considered to be attached to the numerator in each case. Thus

$$\frac{a}{b} - \frac{c}{d} + \frac{e}{f} = \frac{adf - cbf + ebd}{bdf}.$$

(*d*) *Multiplication of a fraction.*

Since a fraction, interpreted quantitatively, is not essentially different from a whole number, the multiplication of a fraction by a whole number does not involve any new ideas at all. It is worth while to notice, however, that the result of any such multiplication can be represented as a single fraction.

$(a \div b) \times b = a$ by definition.

Therefore $(a \div b) \times b \times c = a \times c$

$(a \div b) \times c \times b = a \times c$ (Law of Commutation),

that is, $\{(a \div b) \times c\} \times b = (a \times c)$.

Therefore $(a \div b) \times c = (a \times c) \div b$

by the definition of division.

In fraction form

$$(a/b) \times c = (ac/b).$$

(*e*) *Multiplication by a fraction.*

Remembering that the fundamental definition of the operation of multiplication is that contained in the statement

$$a \times b = a + a + a + a + a + \text{etc.},\ b \text{ terms in all,}$$

it appears that multiplication *by* a fraction is not a comprehensible operation, and in point of fact it certainly is not. There is nevertheless an operation that can conveniently be described as multiplication by a fraction provided its real character is clearly understood. The fraction $(a \div b)$ can be multiplied by c giving $(a \div b) \times c$. This number or fraction can then be divided by d, giving $\{(a \div b) \times c\} \div d$. Now this process of multiplication by c followed by division by d can be expressed for the sake of abbreviation as multiplication by $(c \div d)$. In fractional form the operations can then be written

$$\frac{a}{b} \times \frac{c}{d}.$$

It will be assumed that whenever the apparent process of multiplication by a fraction arises in practice it will be legitimate to re-interpret the operation in the above manner, that is,

$$\frac{a}{b} \times \frac{c}{d} = \{(a \div b) \times c\} \div d.$$

In fact this assumption is necessary, for only on these lines can a comprehensible meaning be attached to the process. On this understanding we can proceed to find the single fraction that shall have the same magnitude as the "product" of two fractions. Suppose $(e \div f)$ to be this fraction. Then

$$\{(a \div b) \times c\} \div d = (e \div f),$$

and by the definition of division

$$(a \div b) \times c = (e \div f) \times d.$$

Now it has been shown above in para. (*d*) that

$$(a \div b) \times c = (a \times c) \div b,$$

therefore $(a \times c) \div b = (e \div f) \times d,$

and, by definition, $(a \times c) = (e \div f) \times d \times b$

$= (e \div f) \times (b \times d),$

therefore $(a \times c) \div (b \times d) = (e \div f),$

that is,

$$\frac{a}{b} \times \frac{c}{d} = \{(a \div b) \times c\} \div d = (a \times c) \div (b \times d) = \frac{ac}{bd}.$$

The extension of the above reasoning to give

$$\frac{a}{b} \times \frac{c}{d} \times \frac{e}{f} \times \frac{g}{h} \times \text{etc.} = \frac{aceg \text{ etc.}}{bdfh \text{ etc.}}$$

is a straightforward application of the Law of Association.

(*f*) *Division of a fraction by a fraction.*

Strictly speaking, division by a fraction is no more intelligible than multiplication by a fraction. It can easily be shown, however, that its meaning can be derived immediately from that of multiplication by a fraction, if it be assumed, as it obviously must be, that the apparent process of division by a fraction is consistent with the formal definition of division.

Putting $(a \div b) \div (c \div d) = (e \div f),$

then by the definition of division,

$$(a \div b) = (e \div f) \times (c \div d)$$
$$= \{(e \div f) \times c\} \div d \text{ as above.}$$

Therefore $(a \div b) \times d = (e \div f) \times c,$

and, by definition, $\{(a \div b) \times d\} \div c = (e \div f).$

Therefore $(a \div b) \div (c \div d) = \{(a \div b) \times d\} \div c,$

and on the understanding indicated in para. (*e*) above, the right-hand side can be written in the form $(a/b) \times (d/c)$,

giving $$\frac{a}{b} \div \frac{c}{d} = \frac{a}{b} \times \frac{d}{c} = \frac{ad}{bc}.$$

This shows that dividing by a fraction is the same as multiplying by the same fraction inverted. The result of inverting a fraction in this way is called the "reciprocal" of the fraction.

Since $c = c \div 1$, we have as a special case

$$(a/b) \div c = (a/b) \times (1/c) = a/bc.$$

Thus division by a number can be represented as multiplication by a fraction having 1 as numerator and the number as denominator.

22. The Division of Zero by a Number

There is no difficulty about this operation. The fraction $(0 \div a)$, if it exists, must have a meaning consistent with the definition

$$(0 \div a) \times a = 0$$

for any positive or negative value of a. Now, as shown in Section 16,

$$0 \times a = 0.$$

Therefore the statement that

$$(0 \div a) = 0$$

fits in quite satisfactorily with the rest of our rules and definitions.

23. The Division of a Number by Zero, and the Collapse of Mathematics

What is the meaning of $(a \div 0)$? The meaning, if any, must be consistent with the definition

$$(a \div 0) \times 0 = a.$$

But for all positive and negative values of b

$$b \times 0 = 0$$

and the reader can easily prove for himself that this is equally true for any positive or negative fraction. Therefore $(a \div 0)$ must be something essentially different from a whole number or a fraction, and the rules of algebra must *not* be applied to the group $(a \div 0)$. To prove this statement we will proceed to wreck the whole structure of mathematics by assuming that they can be applied to it.

$$\frac{a}{0} + \frac{b}{c} = \frac{(a \times c) + (b \times 0)}{c \times 0} \text{ by Section 21}$$

$$= \frac{ac}{0}.$$

Now let d be any other number different from b. Then

$$\frac{a}{0} + \frac{d}{c} = \frac{a \times c + d \times 0}{c \times 0}$$

$$= \frac{ac}{0}.$$

Therefore

$$\frac{a}{0} + \frac{d}{c} = \frac{a}{0} + \frac{b}{c},$$

$$\frac{d}{c} = \frac{b}{c},$$

and finally

$$b = d.$$

Therefore any number is equal to any other number, and the whole of mathematics is nonsense.

In fact, the group $(a \div 0)$ is a germ of insanity that is liable to infect the most sedate of propositions and set it babbling gibberish. If at any time a set of equations goes thus suddenly mad and announces that the moon is made of green cheese the baneful effect of division by zero can be suspected. The germ may have crept in disguised as something quite harmless and respectable, for instance, division by $(a - b)$ at one point, followed at a later stage by the condition "let $a = b$".

24. The Fundamental Rules of Algebra—General Conclusions

Intelligible interpretations have now been found for positive and negative fractions, and for the addition, subtraction, multiplication and division of positive and negative fractions. We can now proceed to the full generalisation of the letter symbols. In any expression in which letter symbols are associated with all or any of the operations of addition, subtraction, multiplication, or division, the letter symbols may be taken to represent any positive or negative whole numbers (integers) or positive or negative fractions as the conditions of the problem may require.

The writer has deliberately refrained from presenting any tabular summary of the rules and formulæ so far developed. Any reader who is new to the subject is strongly advised to do this for himself, as this will help him in the understanding and memorising of the more important conclusions. He is also recommended to familiarise himself with the technique involved by working out far more examples than can be presented in the limited space available. Plenty can be found in any elementary textbook of the subject. An even better, and certainly more interesting way, is to make up a number of examples for some imaginary pupil.

25. Mathematics As A Thought-Saving Device

Before going on to the interesting elaborations and developments of the fundamental rules of algebra it will be well to say a final word with regard to the character of mathematics in general and the right way of applying it.

The writer has been at some pains to invest the mystic symbols of mathematics with a precise and real, one might almost say homely, significance. This is strictly in accordance with the spirit of modern mathematicians, who are becoming increasingly distrustful of any purely formal symbolism of doubtful interpretation, and who seek to base their science on a few simple ideas about number or magnitude. Clearness in the initial ideas is essential, and this clearness has to be paid for in hard thought; but it is not for a moment suggested that all the ideas involved are to be turned over in the mind every time the fundamental formulæ are applied. In fact, nothing could be farther from the object of mathematics, which is, though it may sound paradoxical, the elimination of thought. This is very clearly stated by Professor A. N. Whitehead, a mathematician who combined the intellect of a scientist with the imagination of a poet and whose book, *An Introduction to Mathematics* (Home University Library), can be very warmly recommended to all who are interested in this subject. The object of the symbolism of mathematics is to enable us to " make transitions in reasoning almost mechanically by the eye, which otherwise would call into play the higher faculties of the brain. It is a profoundly erroneous truism . . . that we should cultivate the habit of thinking of what we are doing. The precise opposite is the case. Civilisation advances by extending the number of important operations that we can perform without thinking about them ". This may sound like a glorification of " rule of thumb ", but it is " rule of thumb " with a difference. To apply a formula without thinking about it is a gesture not unworthy of a mathematician. To apply it without understanding it, or without ever having understood it, is at best a *pis aller*, and at worst an unintelligent faith in Mumbo-Jumbo.

Examples II

1. Show that

$$\frac{3}{10}+\frac{7}{100}+\frac{9}{1,000}+\frac{6}{10,000}+\frac{8}{100,000}$$

$$=\frac{37,968}{100,000}.$$

2. If
$$\frac{1}{b}-\frac{1}{a}=\frac{c}{ab},$$
prove that $ac+bc=aa-bb$.

3. Show that

$$\left\{\left(\frac{1}{a}+\frac{1}{b}+\frac{1}{c}\right)-\left(\frac{a}{bc}+\frac{b}{ac}+\frac{c}{ab}\right)\right\}$$

$$\div\left(1-\frac{b}{a}\right)\left(1-\frac{c}{b}\right)\left(1-\frac{a}{c}\right)=\frac{1}{a-b}+\frac{1}{b-c}+\frac{1}{c-a}.$$

4. Prove that if

$$R_0R_1R_2+R_0R_2R_3+R_0R_3R_1-R_1R_2R_3=0,$$

$$\frac{1}{R_0}=\frac{1}{R_1}+\frac{1}{R_2}+\frac{1}{R_3}.$$

Note.—R_0, R_1, R_2, etc., are simple symbols, like a, b, c, etc. The subscripts are just a convenient way of distinguishing them. The usefulness of such subscripts will appear later.

5. Show that
$$\frac{R_1R_2}{R_1+R_2},$$
where R_1 and R_2 are positive numbers or fractions, is less than the smaller of R_1 and R_2.

6. Simplify $\dfrac{xx-yy}{x-y}$ and $\dfrac{xx-yy}{x+y}$.

7. If $3x + 4 = 2x + 19$, what is the value of x?

8. If $xy = 144$, find the values of $x + y$ when $x = 2, 4, 6, 8, 9, 12, 16, 18, 24, 36, 72$.

INDICES AND LOGARITHMS

WE now come to a very important set of ideas derived directly from the ideas of multiplication and division.

26. Definition of An Index

Consider the product $a \times b \times c \times d \ldots n$ factors in all, where n is of necessity a positive integer, and the other letters have the full significance of algebraic letter symbols. Now suppose that the factors are all equal to each other. Then the product takes the form

$$a \times a \times a \times a \ldots n \text{ factors.}$$

Now just as the sum of n a's was abbreviated into $a \times n$, so the product of n a's similarly lends itself to abbreviation, and the accepted form of abbreviation is

$a \times a \times a \times a \ldots n$ factors $=$ (that is, is written) a^n.

The double symbol on the right is spoken of as "a to the nth" or "the nth power of a". It follows from this definition that n must be a positive integer.

27. The Index Law

From the definition of an index,

$$\begin{aligned} a^m \times a^n &= (a \times a \times a \times a \ldots m \text{ factors}) \\ &\quad \times (a \times a \times a \times a \ldots n \text{ factors}) \\ &= (a \times a \times a \times a \times a \times a \ldots m + n \text{ factors}) \\ &= a^{(m+n)} \end{aligned}$$

that is, $$a^m \times a^n = a^{(m+n)}.$$

The general result symbolised in this formula will be referred to hereafter as the Index Law.

28. The Negative "Index"

The word "index" is put in inverted commas to emphasise the fact that this is only a manner of speaking. An index is a positive integer, but it is very convenient to allow a negative number to masquerade as an index for the following reason.

Consider the fraction a^m/a^n. If $m > n$, a^m can be written $a^m = a^{m-n+n} = a^{m-n} \times a^n$ by the Index Law, so that

$$a^m/a^n = (a^{m-n} \times a^n)/a^n = a^{m-n}.$$

On the other hand, if $m < n$ it can obviously be shown in exactly the same manner that

$$a^m/a^n = 1/a^{n-m}.$$

Now this double result, depending on whether $m > n$ or $m < n$ involves a deal of tiresome conversation and would be a great nuisance in practice. The nuisance can be avoided in this way. Let us write $1/a^n$ in the form a^{-n}. The symbol a^{-n} is in itself quite meaningless, so we are quite at liberty to attach the above meaning to it if we so desire, and if any useful purpose will be served by doing so. Notice carefully, however, that the $-n$ in a^{-n} is not strictly speaking an index, and *a^{-n} is nothing more than a way of writing $1/a^n$.* On this understanding,

$$a^m/a^n = a^m \times a^{-n}$$

Now if $m > n$, $\quad a^m/a^n = a^{m-n}$ as shown above,

that is,

$$a^m \times a^{-n} = a^{m-n}$$
$$= a^{m+(-n)},$$

which shows that the $-n$ can be treated as if it were a genuine index, obeying the Index Law. Moreover, both cases are now included in the one formula, for if $m < n$, then, as shown above,

$$a^m/a^n = 1/a^{n-m},$$

and $1/a^{n-m}$, in virtue of the meaning that we have attached to a negative number written as an index, can be put in the form

$$1/a^{n-m} = 1/a^{-(m-n)}$$
$$= a^{m-n}.$$

Thus the meaning that we have arbitrarily assigned to a negative number written as an index has the advantage that such negative numbers can be treated as if they were genuine indices obeying the Index Law, and an index can if necessary be allowed to assume a negative value without causing us to rewrite the resulting expressions in some other form.

29. THE ZERO "INDEX"

Again, there is strictly speaking no such thing as the zero index. However, consider a^m/a^n when n becomes equal to m. In accordance with the ideas explained in the preceding paragraph

$$a^m/a^m = a^{(m-m)} = a^0$$

and since a^m/a^m is unity, it follows that a^0 is unity for all finite values of a. Thus a^0, an idea which is incomprehensible in itself, is really nothing more than a convenient way of writing a^m/a^m,* and its value is unity for all finite values of a except $a = 0$.

30. REPEATED PRODUCTS OF POWERS

Since $a^m \times a^n = a^{m+n}$ for positive or negative values of m or n, it follows from the Law of Association that

$$a^m \times a^n \times a^p \times a^q \ldots r \text{ factors}$$
$$= a^{m+n+p+q+ \ldots r \text{ terms}},$$

and if

$$m = n = p = \text{etc.},$$

this becomes

$$a^m \times a^m \times a^m \times a^m \ldots r \text{ factors}$$
$$= a^{m+m+m+ \ldots r \text{ terms}},$$

* Some more advanced students will possibly object that a^0 must be something more than the convenient mathematical fiction that it is here stated to be, since it is a form which frequently occurs in practical physics. For instance, take some quantity which changes with time according to the formula $x = x_0a^{bt}$ where x_0, a, and b are constant numbers. When $t = 0$ the value of x is x_0a^{b0}, that is, x_0. It must be remembered, however, that t really represents an *interval* of time and should ideally be written in the form $t_2 - t_1$. The condition $t = 0$ is thus really the condition $t_2 = t_1$, which is in agreement with the above discussion of a^0.

that is, $$(a^m)^r = a^{m \times r} = a^{mr}.$$

As an exercise, the reader is recommended to show from first principles that this is true for negative values of m and r, for example, $-p$ and $-q$, where p and q are positive. This amounts to showing that

$$\frac{1}{\left(\frac{1}{a^p}\right)^q} = a^{pq}.$$

31. Powers of Products and Quotients

By writing out in full, it will be found quite easy to show that for positive or negative values of n

$$(a \times b)^n = a^n \times b^n$$

and

$$(a/b)^n = a^n/b^n.$$

As a particular case of the latter, when $a = 1$

$$(1/b)^n = 1^n/b^n = 1/b^n.$$

By writing $1/b$ as b^{-1} this is seen to be a special case of the result proved in Section 30. Notice further that

$$(1/b)^{-n} = b^n.$$

32. Roots

Consider the statement

$$a \times a \times a \times a \quad . \quad . \quad . \quad n \text{ factors} = a^n = b.$$

Here the number b is described as being made up of n factors each equal to a. Similarly a could be described in terms of b and n as that number which, multiplied together n times, would give the number b. There is, in fact, a recognised way of writing this, that is,

$$a = \sqrt[n]{b}.$$

a is described as the nth root of b. Thus since

$$2 \times 2 \times 2 \times 2 = 2^4 = 16, \quad 2 = \sqrt[4]{16},$$

and 2 is described as the fourth root of 16. The definition of the nth root of b is clearly

$$(\sqrt[n]{b})^n = b.$$

There is no special magic about this way of writing the

nth root of b, and if we wished to emphasise the analogy with the index form it could equally well be written

$$b^{n'},$$

the dash serving to distinguish it from b^n. Using this way of writing, the definition of $a^{n'}$ would be

$$(a^{n'})^n = a.$$

Comparison of this definition with the result obtained above, that is

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

suggests an even better way of writing the nth root of a.

Suppose we write it as $a^{\frac{1}{n}}$ or $a^{1/n}$. Then the definition of $a^{\frac{1}{n}}$ is

$$\left(a^{\frac{1}{n}}\right)^n = a$$

which is in agreement with the formula $(a^m)^n = a^{mn}$, for it can be put

$$\left(a^{\frac{1}{n}}\right)^n = a^{\frac{1}{n} \times n} = a^1 = a.$$

Notice carefully that $a^{\frac{1}{n}}$ is only *a convenient way of writing* $\sqrt[n]{a}$. The $1/n$ is not really an index and the application of the above formula in the manner shown does not follow from the Index Law but from the definition of $a^{\frac{1}{n}}$.

It will be convenient to carry this notation one stage further, for, as the definition of the qth root of a^p we have

$$(\sqrt[q]{a^p})^q = a^p$$

and if $(\sqrt[q]{a^p})$ is written in the form $a^{p/q}$ this becomes

$$(a^{p/q})^q = a^p,$$

which again is in agreement with the formula

$$(a^n)^m = a^{nm}.$$

This suggests that $a^{p/q}$ will prove to be a convenient way of writing the qth root of a^p, so that we have as the definition of $a^{p/q}$ the statement

$$(a^{p/q})^q = a^p.$$

33. Product of Roots

We must now see to what extent the p/q in $a^{p/q}$ can be treated as if it really were an index.

Since $$(a^{p/q})^q = a^p,$$

$$\{(a^{p/q})^q\}^s = a^{ps},$$

that is, $$(a^{p/q})^{qs} = a^{ps}, \qquad \text{by Section 30.}$$

Similarly, $$(a^{r/s})^{qs} = a^{rq}.$$

Therefore $$(a^{p/q})^{qs} \times (a^{r/s})^{qs} = a^{ps} \times a^{rq},$$

that is, $$(a^{p/q} \times a^{r/s})^{qs} = a^{(ps+rq)}.$$

This shows that $(a^{p/q} \times a^{r/s})$ is the qsth root of $a^{(ps+rq)}$ which we have agreed to write in the form $a^{\frac{ps+rq}{qs}}$, so that

$$\left(a^{\frac{p}{q}} \times a^{\frac{r}{s}}\right) = a^{\frac{ps+rq}{qs}} = a^{\left(\frac{p}{q}+\frac{r}{s}\right)}.$$

The suggested form of notation therefore has the great advantage that the formula

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

can be applied when m and n are fractions. A simple extension of the above proof will show that it will still apply even when m and n are negative fractions or when either is negative.

34. Full Generalisation of the Index Formulæ

Assuming only that $a^{p/q}$ is a convenient way of writing the qth root of a^p, that is, on the basis of the definition

$$(a^{p/q})^q = a^p,$$

it will be quite easy to show that the formulæ given in the preceding sections can all be applied when the indices are fractional in form. For instance, to prove that

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

where $m = p/q$ and $n = r/s$, p, q, r, and s being integers, we have, by definition,

$$\left\{\left(a^{\frac{p}{q}}\right)^{\frac{r}{s}}\right\}^s = \left(a^{\frac{p}{q}}\right)^r,$$

and

$$\left\{\left(a^{\frac{p}{q}}\right)^{\frac{r}{s}}\right\}^{qs} = \left\{\left(a^{\frac{p}{q}}\right)^r\right\}^q$$

$$= \left\{\left(a^{\frac{p}{q}}\right)^q\right\}^r$$

$$= (a^p)^r$$

$$= a^{pr}$$

so that by definition

$$\left\{\left(a^{\frac{p}{q}}\right)^{\frac{r}{s}}\right\} = a^{\frac{pr}{qs}}.$$

The proofs for the remaining formulæ will be omitted to save space. They will follow exactly similar lines to that given as an example. To resume, it may now be stated that for positive or negative integral or fractional values of m and n

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

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

$$(ab)^n = a^n b^n$$

$$(a/b)^n = a^n/b^n.$$

The reader should notice very carefully the logical sequence of the above demonstration of the generalisation of the index formulæ. This demonstration is put forward with all due deference as an alternative to the usual text-book treatment, in order to show that the full generalisation required can be obtained without bringing in any incomprehensible ideas or any purely formal symbolism.*

* It is seriously stated in some textbooks that the generalisation of the index formulæ depends on the Principle of the Permanence of Equivalent Forms, that is, "A law of algebra which admits of proof subject to certain limitations is true generally provided the removal of the limitations is not *incompatible* with the truth of the law". This principle appears to the writer to be modelled on the tactics of that sort of cuttle-fish which, when in difficulties, squirts out a cloud of sepia and escapes in the confusion.

35. An Example of the Index Notation

The products considered in Section 16 can now be written

$$(x-a)(x-b) = x^2 - (a+b)x + ab$$

and

$$(x-a)(x-a) = (x-a)^2 = x^2 - 2ax + a^2.$$

Further,

$$(x-a)(x+a) = x^2 - a^2.$$

These three formulæ, which are three special cases of the multiplication of number groups, will prove to be of great use in practical mathematics.

36. Is $a^{\frac{1}{n}}$ always a Number?

The definition of $a^{\frac{1}{n}}$ is

$$\left(a^{\frac{1}{n}}\right)^n = a.$$

We have so far taken it for granted that if a and n are numbers, there is some other number, written $a^{\frac{1}{n}}$, which fulfils the above definition; and so there is in most cases, but not in every case. Suppose a is 64 and n is 3. Then, by definition,

$$(64^{\frac{1}{3}})^3 = 64,$$

and since $4^3 = 64$,

we may say that $64^{\frac{1}{3}} = 4.$

Again, since $5^3 = 125$, $125^{\frac{1}{3}}$ is the number 5. Moreover, since 100 lies between 64 and 125, $100^{\frac{1}{3}}$ will presumably be some number greater than 4 and less than 5, what Barrie would call "four and a bittock", that is, four and a fraction. Actually there is no number between four and five the cube of which is *exactly* equal to 100, and $100^{\frac{1}{3}}$ is for this reason called an "irrational" quantity. (This, by the way, does not mean "unreasonable". It simply means a number which cannot be exactly expressed as the ratio of two whole numbers.) However, by methods to be described later, a number can be found which satisfies the condition to a very high degree of accuracy.

Thus
$$(4.6)^3 = 97.34$$
$$(4.64)^3 = 99.9$$
$$(4.642)^3 = 100.03$$

and so on. Since in real life there is always a limit to the accuracy with which measurements of quantity can be carried out, the distinction between an irrational quantity and a rational quantity is, as far as the physicist or experimenter is concerned, purely academic. Thus for any work of an accuracy of a tenth of 1 per cent, $100^{\frac{1}{3}}$ *is* 4.64.

Consider now another case. What is $25^{\frac{1}{2}}$? By definition,

$$(25^{\frac{1}{2}})^2 = 25.$$

Now $(+5)^2 = 25$,

and also $(-5)^2 = 25$ (see Section 16),

so that $25^{\frac{1}{2}} = +5$

and $25^{\frac{1}{2}} = -5$.

In other words, $25^{\frac{1}{2}}$ not only exists—it leads a double life, a sort of Dr. Jekyll and Mr. Hyde. This duality can be expressed

$$25^{\frac{1}{2}} = \pm 5 \text{ (plus or minus 5)}.$$

Notice further that any even root can be expressed as a square root, for

$$a^{\frac{1}{2p}} = \left(a^{\frac{1}{p}}\right)^{\frac{1}{2}},$$

so that the $2p$th root of a is the square root of the pth root of a. It is clear, therefore, that any even root, if it exists at all, will have at least two real values, differing only in sign. In practice this implicit (that is, hidden) ambiguity would be an inconvenience, since mathematics is, or should be, plain dealing *par excellence*, so it is generally agreed that $a^{\frac{1}{2}}$ shall mean the real positive number which satisfies the definition. The ambiguity can now be made explicit.

For instance, if $x^2 = b$,

then $x = \pm b^{\frac{1}{2}}$ or $\pm\sqrt{b}$.

Notice that this ambiguity of sign will not occur in the case of odd roots.

Thus $-3 \times -3 \times -3 = -27$,

and	$3 \times 3 \times 3 = \quad 27,$
so that	$(27)^{\frac{1}{3}} = \quad 3,$
and	$(-27)^{\frac{1}{3}} = -3.$
Similarly	$32^{\frac{1}{5}} = \quad 2,$
and	$(-32)^{\frac{1}{5}} = -2.$

In words, the odd root of a negative number is a negative number or fraction, and the odd root of a positive number is a positive number or fraction.

One other important case remains. What is the square root of a negative number, $(-9)^{\frac{1}{2}}$ for instance? First let us simplify the matter a little. Since

$$-9 = -1 \times 9,$$

$$(-9)^{\frac{1}{2}} = (-1)^{\frac{1}{2}} \times (9)^{\frac{1}{2}} \text{ (see Section 31)}$$

$$= (-1)^{\frac{1}{2}} \times 3 \text{ or } \sqrt{-1} \times 3.$$

In a similar manner the square root of any negative number or fraction can be reduced to the form $\sqrt{-1} \times a$, and the feature common to all such cases is $\sqrt{-1}$. Also any even root of a negative number will depend on this same thing, for the $2p$th root of -1 is the pth root of the square root of -1. Now there is no number which, multiplied by itself, will give -1, so the symbol $\sqrt{-1}$ or $(-1)^{\frac{1}{2}}$ is, as far as pure number is concerned, meaningless. Later on an interpretation will be found for it in relation to quite a different set of ideas, and the symbol will prove to be of great service in connection with alternating current theory. For the present, however, it will be sufficient to say that the square root of a negative number is non-existent as a number, or, as the mathematicians have not very happily designated it, is an imaginary quantity.

Examples III

1. Show that $$(x^{\frac{1}{2}} + a^{\frac{1}{2}})^2 = x + 2a^{\frac{1}{2}}x^{\frac{1}{2}} + a,$$
   and $$(x^{\frac{1}{2}} - a^{\frac{1}{2}})(x^{\frac{1}{2}} + a^{\frac{1}{2}}) = (x - a).$$

2. Simplify $$\frac{a^{\frac{5}{2}}\, b^{1\frac{1}{8}}\, c^{\frac{3}{2}}}{b^{\frac{7}{8}}\, c^{\frac{3}{8}}} - \frac{a^{\frac{3}{4}}\, b^{\frac{3}{2}}\, c}{a^{\frac{1}{4}}\, b^{\frac{1}{2}}}.$$
   Show that it is equal to
   $$abc(a^{\frac{3}{2}}\, b^{-1\frac{1}{8}}\, c^{\frac{1}{8}} - a^{-\frac{1}{2}}).$$

3. If
   $$a^x = \alpha,$$
   $$a^y = \beta,$$
   $$a^z = \gamma,$$
   and $$\alpha\beta = \gamma,$$
   what is the relation between x, y, and z? Further, if $\gamma = \alpha/\beta$ what is the relation between x, y, and z?

4. Show that
   $$(a + b)^3 = a^3 + 3a^2b + 3ab^2 + b^3,$$
   $$(a - b)^3 = a^3 - 3a^2b + 3ab^2 - b^3,$$
   $$(a^3 - b^3) = (a - b)(a^2 + ab + b^2),$$
   $$(a^3 + b^3) = (a + b)(a^2 - ab + b^2).$$

5. What are the factors of $(a^4 - b^4)$ and of $(a^6 - b^6)$ and $(a^6 + b^6)$?

6. Show that
   $$\left(\frac{x^a}{x^b}\right)^{(a+b)} \times \left(\frac{x^b}{x^c}\right)^{(b+c)} \times \left(\frac{x^c}{x^a}\right)^{(c+a)} = 1.$$

37. Functions: Graphical Representation

Any algebraic letter symbol can be considered to stand for any positive or negative number or fraction. Starting with some symbol, x for instance, other numbers can be built up, the magnitude of which will depend in some specified way on the magnitude of x. The number $3x+4$ is a simple instance of such a built-up number, and, being a number, it can be represented for convenience by some other single symbol, such as y. This number y is then defined by the statement (or equation)

$$y = 3x + 4.$$

In such a connection the number x is termed an "independent variable", the idea being that it is at liberty to wander at its own sweet will over the whole range of magnitude.

The number y on the other hand has no will of its own and has to go where x tells it. This is expressed rather grandiloquently by the phrase "y is a function of x". In mathematical shorthand this is written

$$y = f(x).$$

Notice that the letter symbol f does not in this instance represent a number, but expresses a functional relationship between y and whatever symbol is written inside the brackets. Different functional relationships can of course be represented by other letters, for example, $F(x)$ or $\phi(x)$. Once a particular functional relationship has been specified the same letter should be used throughout any piece of work for that function.

x	y
− 2	− 2
− 1	1
0	4
1	7
2	10

The dependence of y upon x can be emphasised by assigning various values to x and finding the corresponding value of y. Such values can be tabulated as shown on the left.

An extensive set of tables could be drawn up in this way, but such tables would not reveal the distinctive character of the dependence of y on x. This, however, can be made quite clear by means of a method of graphical representation invented by the philosopher Descartes while he was lying in bed one morning (which just shows the unwisdom of too early rising).

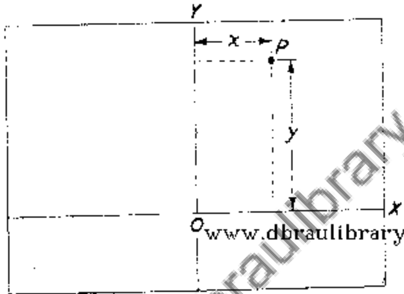


Fig. 7

Draw two lines OX and OY at right angles (Fig. 7). Any related pair of numbers x and y can now be represented by a point such as P, which is situated x units of length perpendicularly to the right (if x is positive) or left (if x is negative) of OY, and y units of length perpendicularly up (if y is positive) or down (if y is negative) from OX. (There is no need for the units of length to be the same in the two directions.) The numbers x and y are called the "co-ordinates" of the point P, which can be referred to as the point (x, y).

In Fig. 8 are shown all the points represented by the pairs of numbers listed in the above table. A new aspect of the matter is at once revealed. All the points are seen to lie on a straight line. Moreover, it will be found that any other related numbers belonging to this set will give rise to points which also lie on this straight line. The straight line can therefore be regarded as a complete representation of the function

$$y = 3x + 4$$

since all the points of the function will be found somewhere on this line. Conversely, any point on the line satisfies the functional relationship, and the value of y for any given value of x could be read off the diagram or a suitable

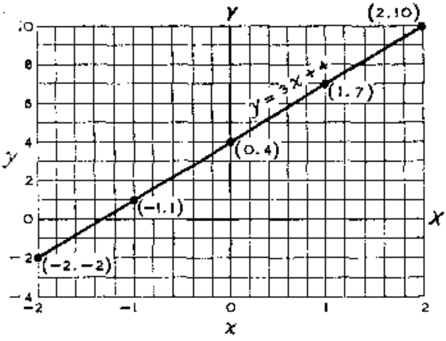


Fig. 8

extension of it. For $x = 1.7$, for instance, the corresponding value of y is seen to be 9.1.

The idea of functional form can now be described. There is no special magic about the numbers 3 and 4 in the above function, so it is inherently probable, and is true in fact, that any other pair of numbers a and b would give rise to a similar picture for the function, that is, a straight line. The function

$$y = ax + b$$

is for this reason called a straight-line function of x.

Obviously much more complicated numbers than this could be built up out of x and other constant numbers. For instance

$$y = 3x^2 + 4x + 5,$$

or, more generally,

$$y = ax^2 + bx + c.$$

The points of any such function for given values of a and b could be similarly plotted on a diagram as in the above simpler case. It would be found that the points do not lie on a straight line for this rather more elaborate function. Nevertheless, by plotting a sufficiently large number of points close together, a dotted line will be produced, through which a smooth curve can be drawn, and it will be found that any other points of the function within the range of the diagram will lie on or very close to this curve, and the greater the number of points calculated and plotted, the more nearly will any other points be found to lie on the curve.

38. The Exponential Function

The above is no more than a very brief introduction to the idea of functions and their graphical representation. From this point there will be considered a particular function, namely

$$y = a^x$$

where a is some constant number.

To fix ideas, let a have some simple value, such as 2. That is

$$y = 2^x.$$

The first thing to notice is that x can have any positive or negative fractional or integral value, for the significance of such a general index has been determined. For instance, if $x = .5$

$$y = 2^{.5} = 2^{\frac{1}{2}} = \sqrt{2} = 1.414,$$

and if $x = -.5$,

$$y = 1/x^{.5} = 1/1.414 = .707.$$

Proceeding in this way, it will be possible to draw up a table of related values of x and y covering any desired range. For instance,

x	y
− 3.0	.125
− 2.5	.177
− 2.0	.250
− 1.5	.355
− 1.0	.500
0	1.00

x	y
1.0	2.00
1.5	2.83
2.0	4.00
2.5	5.66
3.0	8.00

These and similar values can be plotted on a diagram and a curve drawn. The curve is shown in Fig. 9.

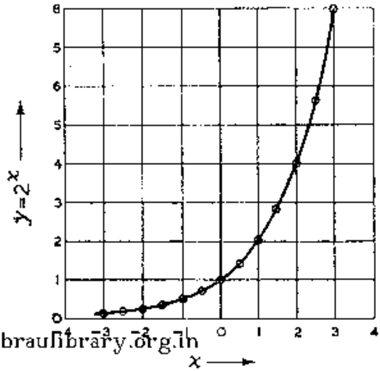


Fig. 9

From such a curve any other corresponding pair of numbers within the given range can be determined fairly accurately (to about 1 per cent or so). For instance, for the point on the curve for which $y = 4.8$, x is seen to be 2.26, so that

$$4.8 = 2^{2.26}$$

(This illustrates a very useful application of the graphical representation of functions. Given that

$$4.8 = 2^x,$$

the determination of x by any method of direct calculation would be a troublesome business, and would puzzle most people.)

Theoretically the above curve could be extended indefinitely beyond the limits shown in either direction, and it would be found to be a smooth continuous curve. This means that for any finite value of y it will be possible to find a value for x such that

$$y = 2^x.$$

There is no special magic about the number 2 which has been used for a in these calculations, and it may therefore be said that in general, for any finite positive value of a, and any finite value of y, it will be possible to find a value for x such that

$$y = a^x.$$

(Notice that a must be positive. The reader is recommended to try to tabulate a set of related values for y and x similar to the above for $y = (-2)^x$.)

39. Logarithms

When x, y, and a are so related that

$$y = a^x,$$

x is called the logarithm of y to the base a, and is written

$$x = log_a y.$$

The two statements

$$y = a^x$$

and

$$x = log_a y$$

therefore mean the same thing. The number corresponding to a given logarithm is called the antilogarithm of the given logarithm. Thus in the above y is called the antilogarithm of x. (These, by the way, are the full ceremonial titles. Log and antilog are what one actually says.)

Having now shown that the logarithm of any number to the base a exists, and having indicated the possibility of determining it by simple arithmetic and drawing, it will be assumed that there is available a set of tables or curves recording the logarithm of all numbers to the given base a, the intervals between whole numbers being subdivided to any desired degree of fineness. For reasons to be given later, 10 is the base chosen in practice, and tables of logarithms and antilogarithms to the base 10 are easily obtainable. What, now, is the use of such tables?

Suppose it is required to find the product of two numbers y_1 and y_2. The logarithms of these numbers can be found in the tables. Call them x_1 and x_2. Then

$$y_1 = a^{x_1},$$

and

$$y_2 = a^{x_2},$$

so that

$$y_1 \times y_2 = a^{x_1} \times a^{x_2} = a^{(x_1+x_2)}.$$

This means that $(x_1 + x_2)$ is the logarithm of $(y_1 \times y_2)$, or that $(y_1 \times y_2)$ is the antilogarithm of $(x_1 + x_2)$. To find the product $(y_1 \times y_2)$, therefore, it is only necessary to find the logarithms of y_1 and y_2, add these together and find the antilogarithm of this total. For instance

$$\begin{aligned} log_{10}\, 3.412 &= \ .53300 \\ log_{10}\, 796 &= 2.90091 \\ \text{Sum} &= 3.43391 \\ antilog_{10}\, 3.43391 &= 2715.8. \end{aligned}$$

Therefore

$$3.412 \times 796 = 2715.8.$$

Thus multiplication is reduced to a simple operation of addition. In a similar manner, division can be simplified to subtraction, for, by an obvious modification of the above proof it can be shown that

$$log_a(y_1/y_2) = log_a y_1 - log_a y_2.$$

Thus to divide 796 by 3.412,

$$\begin{aligned} log_{10}\, 796 &= 2.90091 \\ log_{10}\, 3.412 &= \ .53300 \\ \text{Difference} &= 2.36791 \\ antilog_{10}\, 2.36791 &= 233.29. \end{aligned}$$

Therefore

$$796/3.412 = 233.29.$$

Again, since

$$log_a(y_1 \times y_2) = log_a y_1 + log_a y_2.$$

it follows that

$$log_a(y_1 \times y_2 \times y_3 \ldots\ldots y_n) = log_a y_1 + log_a y_2 + log_a y_3 \ldots\ldots log_a y_n,$$

and if

$$y_1 = y_2 = y_3 = y_4 \text{ etc.} = y,$$

this becomes

$$\begin{aligned} log_a(y \times y \times y \ldots\ldots n \text{ factors}) &= log_a(y^n) \\ &= log_a y + log_a y + log_a y \ldots\ldots n \text{ terms} \\ &= n \times log_a y. \end{aligned}$$

Thus a number can be raised to any integral power by simple multiplication. For instance

$$\begin{aligned} log_{10}\, 3.412 &= .53300 \\ log_{10}\, 3.412^5 &= 5 \times .53300 = 2.66500 \\ antilog_{10}\, 2.66500 &= 462.38. \end{aligned}$$

Therefore

$$3.412^5 = 462.38.$$

Furthermore, the general formula

$$\log_a y^n = n \times \log_a y.$$

proved above for an integral index, is true also for a fractional index. Let $n = p/q$, where p and q are integers. Then, by the definition of a fractional index,

$$(y^{p/q})^q = y^p.$$

Therefore

$$\log_a (y^{p/q})^q = \log_a y^p = p \times \log_a y.$$

But

$$\log_a (y^{p/q})^q = q \times \log_a (y^{p/q}).$$

Therefore

$$q \times \log_a (y^{p/q}) = p \times \log_a y,$$

or

$$\log_a (y^{p/q}) = (p/q) \times \log_a y.$$

In particular the nth root of any number can be calculated by logarithms, for

$$\log_a \sqrt[n]{y} = \log_a y^{1/n}$$
$$= (1/n) \times \log_a y, \text{ or } (\log_a y) \div n.$$

Thus, to find the fifth root of 796,

$$\log_{10} 796 = 2.90091$$
$$\log_{10} \sqrt[5]{796} = 2.90091 \div 5$$
$$= .58018$$
$$\text{antilog}_{10}\, .58018 = 3.8037.$$

Therefore

$$\sqrt[5]{796} = 3.8037.$$

Here an arithmetical process which is too complicated to be admitted into ordinary textbooks is reduced to a simple matter of division.

The full advantage of the logarithmic method of calculation is seen in the determination of a more or less complicated number such as

$$y = \frac{2.43^3 \times 191 \times \sqrt[2]{347}}{\sqrt[4]{1.519^3}}.$$

The determination of this by direct calculation would be a weariness to the flesh, for which weariness the flesh would probably retaliate by slipping into error. Logarithmically, however, the calculation is simplified to

$$log_a y = (3 \times log_a 2.43) + log_a 191 + \tfrac{1}{2} log_a 347 - \tfrac{3}{4} log_a 1.519.$$

The whole of the foregoing propositions with regard to the application of logarithms are independent of the particular base to which the logarithms are referred. The choice of the base to be used in practice is merely a matter of convenience. The base 10 is actually used for the following reason. Any whole number or decimal can be expressed as some number between 1 and 10 multiplied by some power of 10. For instance

$$.143 = 1.43/10 = 1.43 \times 10^{-1},$$
$$3179.8 = 3.1798 \times 1000 = 3.1798 \times 10^3,$$

and so on. Thus any number can be expressed in the form

$$y \times 10^n,$$

where n is a positive or negative integer, and y is some number or fraction between 1 and 10. Now

$$log_{10} y \times 10^n = log_{10} y + log_{10} 10^n$$
$$= log_{10} y + n.$$

Thus all that is needed in a table of logarithms to the base 10 are the logarithms of all numbers between 1 and 10, subdivided decimally to any desired degree (five figures are generally quite sufficient for experimental work). Such logarithms will all lie between 0 and 1, for example,

$$log_{10} 2 = .30103.$$

The whole number to be placed in front of the decimal point will be n, determined as shown above. Thus $log_{10} 200$ would be 2.30103, and $log_{10} .002$, that is, $log_{10} 2 \times 10^{-3}$, would be $\bar{3}.30103$. In practice the negative sign is put as a bar over the whole number, as in this example, to show that it refers only to the whole number and not to the decimal part which follows. It is more convenient to keep the decimal part positive. Thus $\bar{2}.30103$ is the same as $(-2+.30103)$, that is (-1.69897), but the first form is always used in computation.

Logarithms calculated to the base 10 are called "common logarithms", and the base is not expressed. Thus if no other base is specified, it is always assumed to be 10.

There is another set of logarithms which is in occasional use. These are calculated to the base 2.71828, for which apparently arbitrary number the symbol e or the Greek ϵ is always used. There is method in this apparent madness. So much so, in fact, that such logarithms are called "natural logarithms", though at first sight nothing could appear less natural. The natural logarithm of y is written $log_e y$. Actually the use of this other system of logarithms will not necessarily involve a new set of tables, for a logarithm to any one base can be readily converted to the logarithm to some other base in this way. Suppose

$$log_a y = m$$

$$log_b y = n,$$

that is, $$y = a^m = b^n.$$

Suppose further that the logarithm of b to the base a is k,

that is, $$b = a^k.$$

Then

$$y = b^n = (a^k)^n = a^{kn},$$

and since also

$$y = a^m,$$

$$m = kn,$$

so that

$$log_a y = log_a b \times log_b y.$$

In particular,

$$log_e y = log_e 10 \times log_{10} y,$$

and since $log_e 10$ is a constant number (2.3026) the conversion reduces to

$$log_e y = 2.3026 \times log_{10} y,$$

that is, Natural Log. = Common Log. × 2.3026.

If the foregoing description of logarithms and their application is thoroughly understood, the reader should have no difficulty in using this method of computation. It does not pretend to be a complete set of working instructions for the manipulation of log tables, but if the theory is really appreciated, the reader should be able to follow the detailed instructions which are generally included with any such tables (to guard against any misconception on this point, it should be made clear that the tabulated logarithms found in the usual published tables have not actually been determined by the method described in this article, which would not be nearly accurate enough).

Finally, one practical point. The logarithmic method can only be applied to any expression which consists exclusively of products, quotients, and powers. Therefore, before setting out on any series of calculations it is well to arrange them as completely as may be in a form suitable for logarithmic computation. An instance which may frequently occur in connection with alternating current problems is the difference of two squares, that is, $a^2 - b^2$, where a and b have certain specified numerical values. The form $a^2 - b^2$ is not suitable for the use of logarithms, but, as already shown (see Examples I),

$$a^2 - b^2 = (a - b)(a + b),$$

and the form on the right-hand side, consisting of the product of two factors, is better adapted for calculation. For example,

$$\begin{aligned} 874^2 - 27.8^2 &= (874 - 27.8)(874 + 27.8) \\ &= 846.2 \times 901.8. \end{aligned}$$

No general rules can be laid down, but the exercise of a little ingenuity in this matter will often save a great deal of time and labour.

Examples IV

1. Given that $\log 2 = .30103$, show that $\log 5$ is .69897.
2. $Log\ 3 = .4771$. Show that

$$\log 1/3 = \bar{1}.5229,$$
$$\log 81 = 1.9084,$$
$$\log 3\tfrac{1}{3} = .5228,$$
$$\log 3^{30} = 14.313.$$

3. What are the whole numbers in the logs of 21, .021, 99918, .00073 ?
4. Show that 2^{100} is a whole number of 31 figures.
5. Show that $\log (\log 10^{10x}) = 1 + \log x$.
6. Show that $\log_a 1 = 0$ for all finite values of a.
7. Is $\log (a + b) = \log a + \log b$?
8. What is $\log 0$?
9. What is $\log (-1)^n$? Is it equal to $n \log (-1)$? If not, why not?
10. If $y = ka^{bx}$, where a, b, and k are constant numbers, show that $\log y$ is a linear function of x.

EQUATIONS: COMPLEX NUMBERS

40. The Solution of Equations

THE solution of equations is a sort of mathematical detective work. One is given a sufficiency of clues and instructed to find the body, as it were. More prosaically, equations are statements embodying information about certain unknown numbers (usually represented by the symbols x, y, or z) and various other known numbers (a, b, c, etc.), and from this information it is required to find what numbers are represented by the x, y, or z symbols. This is more generally expressed as finding the "values" of the unknown numbers. The use of the word "value" in this connection is sanctioned by general practice, but one has only to speak of x being "more valuable" than y to realise that this special application of the term "value" does violence to the ordinary meaning of the word, which is rather a pity. It is, however, a convenient word to use for the full algebraic significance of a symbol, that is, magnitude and sign.

The simplest equations are those in which only one unknown number, x, has to be determined, and these will first be considered.

41. Equations for One Unknown Number

This subject is best approached as one might approach some new and unexplored city. Having first seen it as a whole from overlooking high ground, one enters and gets to know its main thoroughfares and as many of its byways as one has occasion to use.

The meaning of the word "function" and of the notation

$$y = f(x)$$

was explained in the preceding section. It was shown that $f(x)$ is a number, the magnitude and sign of which depends in some specified manner on the value assigned to x. For instance the form of the function might be specified as

$$y = f(x) = 3x^2 + 5x + 7.$$

The value of y for some particular value of x, 8 for instance, is written $f(8)$, so that for this value of x

$$y = f(8) = (3 \times 8^2) + (5 \times 8) + 7 = 239.$$

In general, if $f(x)$ is specified as

$$y = f(x) = ax^2 + bx + c,$$

the value of y corresponding to some particular numerical value of x, α for instance, would be

$$y = f(\alpha) = a\alpha^2 + b\alpha + c.$$

It was further shown that in general the relation between y and x, where

$$y = f(x),$$

could be exhibited graphically in the form of a picture drawn in accordance with certain agreed rules. The relation is shown in the form of a line which may be a straight line (see Fig. 8), but which will in general be a curved line which may assume an infinite variety of shapes according to the type of the function (see for instance Fig. 9).

In most cases the curve representing the function will cut the axis of x (that is, the line OX in Fig. 8). If it is a straight line it will cut it in one point only, but if it is a curved line it may cut it in several places or perhaps not at all. Thus for some assumed function $y = F(x)$, the curve of which is as shown in Fig. 10, the axis of x is crossed at three points, the x co-ordinates of these points being, say, α, β, and γ. For all points on the x axis, the y co-ordinate is zero, that is, $y = 0$. For all points at which the curve $y = f(x)$ cuts the x axis, the value of x is such that

$$y = f(x) = 0.$$

Thus for the function $y = F(x)$ shown in Fig. 10,

$$F(\alpha) = F(\beta) = F(\gamma) = 0.$$

Any value of x for which

$$f(x) = 0$$

is called a "root" of the equation $f(x) = 0$, so that α, β, and γ are the roots of $F(x) = 0$. (One would like to know why this curious name is used. The name "solution" can also be used, and is preferable in some respects.)

Generally speaking, for any specified function

$$y = f(x),$$

the value of y corresponding to any desired value of x can be determined by simple arithmetic. We are concerned now with the reverse process, that is, given some particular value of y, that is, zero, to find the corresponding value or values of x. This may be, and usually is, a rather more difficult matter. One perfectly general method is obvious from the above discussion, and was illustrated in the preceding section for the case of the function

$$y = a^x.$$

The method is to draw the curve of the function and find the points at which it cuts the x axis. This, however, is likely to be laborious. Moreover, the accuracy of the solution will depend on one's skill in drawing, the sharpness of one's pencil, and various other non-mathematical factors. Nevertheless, it is in many cases the only available general method and one that is quite often used. Further consideration will be given to it later on, but for the present we shall be concerned with simpler and more

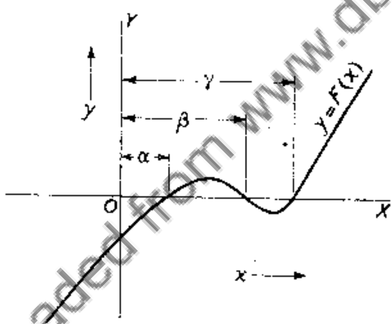


Fig. 10

accurate methods which can be applied to certain common types of equation.

42. The Linear Equation for One Unknown Number

The simplest kind of function is that of which the picture is a straight line (see Fig. 8). All such functions can, by appropriate manipulation, be reduced to the form

$$y = ax + b.$$

Corresponding to such functions is the linear equation

$$ax + b = 0,$$

the characteristic of which is that it contains only known numbers and the first power of the unknown number. As already explained, the solution of this equation is the x co-ordinate of the point in which the line

$$y = ax + b$$

cuts the x axis. Except in the case in which the line is parallel to the x axis (a case of no practical importance) it will always cut the x axis at one point and one point only. Thus there is always one solution to the equation. Once the equation has been reduced to the standard form its solution is simple. Since the numbers $ax + b$ and 0 are equal, the addition of the number $-b$ to each side will not disturb the equality, that is,

$$ax + b - b = 0 - b,$$

or

$$ax = -b.$$

Notice that in effect the b is taken over to the other side of the equation, its sign being changed on the way. In practice one speaks of taking the number or symbol over to the other side. Actually the process consists of the addition of the corresponding number or symbol with the sign reversed to each side of the equation. Since the numbers ax and $-b$ are equal, the division of each by a will leave the resulting numbers or fractions equal, that is,

$$ax/a = -b/a,$$

or

$$x = -b/a.$$

Thus $x = -b/a$ is the solution of the equation $ax + b = 0$, and the whole art of solving a simple linear equation consists of reducing it to the standard form, after which its solution can be written down at once. For instance, given

$$4(3x - 8) = 7 + 4x,$$

then $$12x - 32 = 7 + 4x,$$

and taking the 7 and $4x$ over to the left

$$12x - 4x - 32 - 7 = 0,$$
$$8x - 39 = 0,$$

and comparing this with

$$ax + b = 0,$$

the solution is seen to be

$$x = 39/8 = 4\tfrac{7}{8}.$$

Taking a rather more difficult example,

$$\frac{2}{5x + 7} = \frac{3}{8x + 2},$$

multiply each equal fraction by the number $(5x + 7)(8x + 2)$. Then

$$\frac{2(5x + 7)(8x + 2)}{(5x + 7)} = \frac{3(5x + 7)(8x + 2)}{(8x + 2)},$$

and cancelling out the common factors in the numerator and denominator of each fraction (see Section 21),

$$2(8x + 2) = 3(5x + 7),$$

from which point the solution proceeds exactly as in the first example.

Notice that, in general, if

$$\frac{a}{b} = \frac{c}{d},$$
$$ad = bc.$$

This is known as "cross multiplication", but it really consists of multiplying each equal fraction by the number bd.

This is all that need be said about the simple linear equation. No perfectly general rules can be laid down for the reduction to the form

$$ax + b = 0,$$

but a little common sense and ingenuity are all that is required.

43. The Quadratic Equation

Next in order of complexity to the linear equation comes the quadratic equation, in which the unknown quantity appears in the second power as well as the first. The general type of this equation is

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

corresponding to the function

$$y = ax^2 + bx + c.$$

The reader is recommended to plot out roughly some such function, for example,

$$y = 2x^2 - 5x + 2.$$

It will be found that in every case the curve obtained will resemble one or other of the two curves shown in Fig. 11. Such curves are called parabolic and are of great importance in geometry and in applied science. The equation

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

is sometimes called parabolic also, but the other name is more generally used.

One general characteristic of this sort of equation appears at once from the picture of the typical parabolic function. The curve of such a function will either cut the x axis in two points or not at all. This means that a quadratic equation will either have two real roots or no real roots at all.

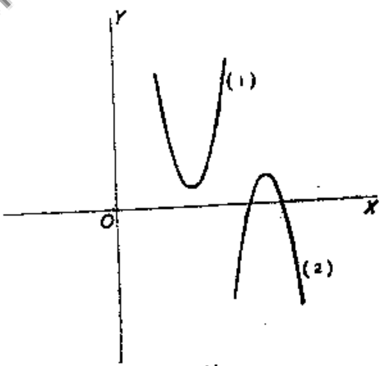


Fig. 11

As in the general

case, one method of solving such an equation would be to plot the curve of the function

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

and find where it cut the x axis, but fortunately there are in this case simpler and more accurate methods available.

The easiest method is one for which the way was prepared in Sections 16 and 17. The multiplication of any two numbers $(x - \alpha)$ and $(x - \beta)$ gives

$$(x - \alpha)(x - \beta) = x^2 - (\alpha + \beta)x + \alpha\beta.$$

Now suppose the numbers α and β are such that

$$\alpha + \beta = - b/a$$

and $$\alpha\beta = c/a.$$

Then $$(x - \alpha)(x - \beta) = x^2 + \frac{b}{a}x + \frac{c}{a},$$

and multiplying each of these equal numbers by a gives

$$a(x - \alpha)(x - \beta) = ax^2 + \frac{ab}{a}x + \frac{ac}{a}$$

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

If therefore we can find two numbers α and β such that

$$\alpha + \beta = - b/a$$

and $$\alpha\beta = c/a,$$

then

$$a(x - \alpha)(x - \beta) = ax^2 + bx + c = 0.$$

Now there are two values of x for which $a\ (x - \alpha)(x - \beta) = 0$. The first is α, for, putting $x = \alpha$,

$$a(x - \alpha)(x - \beta) = a(\alpha - \alpha)(\alpha - \beta)$$
$$= a \times 0 \times (\alpha - \beta)$$
$$= 0,$$

and in a similar manner putting $x = \beta$ will also make the number zero. Therefore α and β are the solutions or roots of

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

For example, consider

$$3x^2 + 54x + 96 = 0.$$

Here $$- b/a = - 18$$

and $$c/a = 32.$$

Also $$(-2) + (-16) = - 18$$

and $(-2) \times (-16) = 32,$
so that $\alpha = -2$
and $\beta = -16$
will satisfy the conditions

$$\alpha + \beta = -b/a = -18$$
$$\alpha\beta = c/a = 32.$$

Therefore the roots of this equation are -2 and -16, and

$$\begin{aligned} 3x^2 + 54x + 96 &= 3(x^2 + 18x + 32) \\ &= 3\{x - (-2)\}\{x - (-16)\} \\ &= 3(x + 2)(x + 16). \end{aligned}$$

This, however, is inspired guesswork, and inspiration is notoriously erratic and capricious. For everyday purposes we need something more certain, a method which, though neither as speedy nor as exhilarating as the flying leap, can always be relied upon to get there. However the equation is solved, whether by flying or walking, the solution will be the same, so that the roots of the equation are α and β, where

$$\alpha + \beta = -b/a$$
and
$$\alpha\beta = c/a.$$

If α and β cannot be guessed from these clues, is there any other way? Yes, there is another way, depending on a trick which is well worth learning for other purposes also. Since

$$\alpha + \beta = -b/a$$

the squares of these equal numbers will also be equal, that is,

$$(\alpha + \beta)^2 = \alpha^2 + 2\alpha\beta + \beta^2 = b^2/a^2.$$

Also
$$4\alpha\beta = 4c/a.$$

Since these pairs of numbers are equal, the differences between the pairs will be equal, that is,

$$\alpha^2 + 2\alpha\beta + \beta^2 - 4\alpha\beta = b^2/a^2 - 4c/a,$$

that is,
$$\begin{aligned} \alpha^2 - 2\alpha\beta + \beta^2 &= b^2/a^2 - 4ac/a^2 \\ &= (b^2 - 4ac)/a^2. \end{aligned}$$

But
$$\alpha^2 - 2\alpha\beta + \beta^2 = (\alpha - \beta)^2.$$

Therefore $$(\alpha - \beta)^2 = (b^2 - 4ac)/a^2,$$

and since these numbers are equal their square roots will also be equal,

that is, $$\alpha - \beta = \sqrt{(b^2 - 4ac)/a^2}$$
$$= \sqrt{(b^2 - 4ac)}/a.$$

Now we know both the sum and difference of α and β, and from these both can be determined, for if

$$\alpha + \beta = -b/a$$

and $$\alpha - \beta = \sqrt{(b^2 - 4ac)}/a,$$

the addition of these pairs of equal numbers will give

$$(\alpha + \beta) + (\alpha - \beta) = 2\alpha = -b/a + \sqrt{(b^2 - 4ac)}/a$$
$$= \frac{-b + \sqrt{b^2 - 4ac}}{a}.$$

Therefore

$$\alpha = \frac{-b + \sqrt{b^2 - 4ac}}{2a}.$$

The subtraction of the pairs gives

$$(\alpha + \beta) - (\alpha - \beta) =$$
$$(\alpha + \beta) - \alpha + \beta = 2\beta = \frac{-b - \sqrt{b^2 - 4ac}}{a}.$$

Therefore

$$\beta = \frac{-b - \sqrt{b^2 - 4ac}}{2a}.$$

Thus the roots of $ax^2 + bx + c = 0$ are α and β, where these numbers have the values given above in terms of the known numbers, a, b, and c.

Writing it out in full

$$ax^2 + bx + c = a\left(x - \frac{-b + \sqrt{b^2 - 4ac}}{2a}\right)$$
$$\left(x - \frac{-b - \sqrt{b^2 - 4ac}}{2a}\right) = 0.$$

For brevity, the two roots are usually combined into one expression by writing

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

which means that

$$x = (-b + \sqrt{b^2 - 4ac})/2a$$

and
$$x = (-b - \sqrt{b^2 - 4ac})/2a$$

are the solutions of the equation. The reader is recommended to confirm the fact that the sum of these roots is $-b/a$ and their product c/a.

As an example, suppose

$$6x^2 + 5.4x - 19.32 = 0,$$

that is,
$$a = 6,$$
$$b = 5.4,$$
$$c = -19.32,$$
$$b^2 = 29.16,$$
$$4ac = -463.68,$$
$$b^2 - 4ac = 492.82,$$
$$\sqrt{b^2 - 4ac} = 22.2.$$

Therefore the roots are

$$x = (-5.4 + 22.2)/12 = 1.4$$

and
$$x = (-5.4 - 22.2)/12 = -2.3,$$

and it can be confirmed by actual multiplication that

$$6x^2 + 5.4x - 19.32 = 6(x - 1.4)(x + 2.3).$$

It was stated above that in certain cases the equation would have no real roots at all. If the general solution given above is really general, this fact should be implicit in it; and so it is, for the roots contain the term $\sqrt{b^2 - 4ac}$, and this will not always exist as a real number. Suppose a and c are of the same sign so that $-4ac$ is a negative number. Then, if $4ac$ is greater than b numerically, $b^2 - 4ac$ will be negative, and $\sqrt{b^2 - 4ac}$ will be "imaginary" (see Section 36). Under this condition there will be no real roots to the equation. For instance

$$29x^2 - 4x + 1 = 0$$

has no real solutions, for

$$b^2 - 4ac = 16 - 116 = -100,$$

and
$$\sqrt{b^2 - 4ac} = \sqrt{-100} = 10\sqrt{-1}.$$

Subject to finding some interpretation for $\sqrt{-1}$ the roots can be left in the form $(4 \pm 10\sqrt{-1})/58$, that is, $(2/29) + (5\sqrt{-1}/29)$ and $(2/29) - (5\sqrt{-1}/29)$. It should be noted and confirmed by actual multiplication and addition that though these are "complex numbers", that is, numbers consisting of two parts, one of which is imaginary, their product and sum are real, being respectively 1/29 and 4/29.

For the solution of a quadratic equation, therefore, it is only necessary to reduce it to the form

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

and the roots can then be found either by guesswork or by means of the formula: The roots of the equation are the numbers $(-b \pm \sqrt{b^2 - 4ac})/2a$. (Whether these numbers are real or complex or imaginary, they are still called the roots of the equation.)

As in the case of the linear equation, however, some little ingenuity may be required for the reduction of any given equation to the standard form, but this is only a matter of practice and experience. To take an example, suppose

$$3(x + 10) = 1/(1 + x).$$

Multiplying each equal number by $(1 + x)$ gives

$$3(x + 10)(1 + x) = (1 + x)/(1 + x),$$

that is, $$3(x^2 + 11x + 10) = 1,$$

or $$3x^2 + 33x + 29 = 0.$$

Again,

$$7\left(x + \frac{1}{x}\right) + 3\left(1 + \frac{1}{x}\right) = -16.$$

Multiplying each equal number by x and rearranging the terms will give

$$7x^2 + 19x + 10 = 0,$$

and a practised eye would see the roots of this to be $-5/7$ and -2.

44. Equations Reducible to Quadratics

Before leaving the subject of quadratic equations it will be well to point out that an equation that looks forbidding and unapproachable at first sight may prove on closer inspection to be quadratic, not in x, but in some simple function of x. An equation

$$ay^2 + by + c = 0$$

can be solved for y, whatever the nature of y may be. If the roots are α and β then

$$y = \alpha, \text{ or } y - \alpha = 0,$$

and $$y = \beta, \text{ or } y - \beta = 0$$

will satisfy the original equation.

If, now, y is some function of x, say $f(x)$, then these become

$$f(x) - \alpha = 0$$

and

$$f(x) - \beta = 0,$$

two further equations in x, which may or may not be solvable, but which will in any case be more manageable than the original equation in x. Take a very simple example:

$$ax^4 + bx^2 + c = 0.$$

This can be written

$$a(x^2)^2 + b(x^2) + c = 0,$$

or, writing y for the function x^2,

$$ay^2 + by + c = 0,$$

If the roots of this equation are α and β, then the original equation is satisfied by

$$y - \alpha = 0$$

and $$y - \beta = 0,$$

that is, $$x^2 - \alpha = 0$$

and $$x^2 - \beta = 0.$$

The factors of the first are

$$(x - \sqrt{\alpha})(x + \sqrt{\alpha}) = 0,$$

and of the second

$$(x - \sqrt{\beta})(x + \sqrt{\beta}) = 0,$$

so that the four roots $x = \pm\sqrt{\alpha}$ and $\pm\sqrt{\beta}$ will satisfy the original equation. Notice that the linear equation containing the first power of x was found to have one root. The quadratic, containing the second power of x, was found to have two roots. The above equation, containing the fourth power, has four roots. This suggests, but does not of course prove, that an equation containing the nth power of x will have n roots. This is so in fact and will be proved later.

The recognition of the quadratic form will not be always so simple as in the above example. Take, for instance,

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

This can be written

$$\left(x^2 + 2 + \frac{1}{x^2}\right) + \left(x + \frac{1}{x}\right) - 6 = 0.$$

Now

$$\left(x^2 + 2 + \frac{1}{x^2}\right) = \left(x + \frac{1}{x}\right)^2,$$

so that the equation is really

$$\left(x + \frac{1}{x}\right)^2 + \left(x + \frac{1}{x}\right) - 6 = 0,$$

and, writing y for $(x + 1/x)$, this becomes

$$y^2 + y - 6 = 0,$$

of which the solutions can be seen to be

$$y + 3 = 0$$

and

$$y - 2 = 0,$$

that is,

$$\left(x + \frac{1}{x}\right) + 3 = 0$$

and

$$\left(x + \frac{1}{x}\right) - 2 = 0.$$

The multiplication of each side of each equation by x will convert these into two simple quadratics in x which can then be solved for x in the ordinary manner. Here again it is the insight born of practice and experience that sees the essential simplicity at the heart of the apparent complexity.

Examples V

1. Solve the equations :—

(*a*) $3(x + 27) = 2(x + 4) + 7(x - 9)$,

(*b*) $\dfrac{2}{x + 3} = \dfrac{5}{7(x - 1) + 2(x + 3) + 1}$,

(*c*) $ax + b = cx + d$,

(*d*) $\dfrac{m}{ax + b} = \dfrac{n}{cx + d}$.

2. Find where the line

$$y = 3x + 4$$

cuts the axis of x.

Find the co-ordinates of the points on this line for which $y = -5$ and $y = 4$.

3. Solve by inspection the equations :—

(*a*) $x^2 - 5x + 6 = 0$,

(*b*) $x^2 + 5x + 6 = 0$,

(*c*) $x^2 - x - 6 = 0$,

(*d*) $x^2 + x - 6 = 0$.

4. Solve the equations :—

(*a*) $\dfrac{1}{x - 1} = x\left(1 + \dfrac{1}{x}\right)$,

(*b*) $(x - 5) + \dfrac{1}{(x - 2)} = \dfrac{1}{(2 - x)}$,

(*c*) $x^2 - 2ax + a^2 - b^2 = 0$,

(*d*) $x^2 - 2bx + b^2 - a^2 = 0$.

5. Solve the equations :—

(*a*) $(x^2 - 4)(x^2 - 5x + 4) = 0$,

(*b*) $(x - a)(x - b)(x - c)(x - d)(x - e) = 0$,

(*c*) $(ax^2 + bx + c)(dx^2 + ex + f) = 0$.

45. The General Equation of the nth Degree

Having disposed of the linear equation and the quadratic equation, the next in order is the cubic, which contains up to the third power of x or, another way of saying the same thing, is of the third degree in x. The typical equation is

$$ax^3 + bx^2 + cx + d = 0.$$

Then comes the quartic, that is,

$$ax^4 + bx^3 + cx^2 + dx + e = 0,$$

and so on indefinitely.

Unfortunately, however, there are no general solutions for these equations. This is specially unfortunate from the point of view of electrical theory, for the analysis of the behaviour of two coupled circuits involves a quartic, or, as it is sometimes called, a bi-quadratic equation, and for this reason practically all textbooks fight shy of it, confining the full discussion to the simpler case of resistance-less circuits, when the bi-quadratic becomes a quadratic in x^2.

Although there are, as stated, no perfectly general solutions for equations of higher degree than the second, such cases are by no means hopeless. Certain forms of higher degree equation can be solved (two bi-quadratics were solved in the preceding paragraph). Also there are certain general considerations relating to equations of any degree which will sometimes point the way to a solution, and which are intrinsically interesting and useful apart from this application. We will therefore discuss briefly the general equation of the nth degree, typified by

$$ax^n + bx^{n-1} + cx^{n-2} + dx^{n-3} + ex^{n-4}, \text{ etc., etc., } + k = 0,$$

where the coefficients a, b, c, etc., can have any sign and any numerical value. First let us see how such a number as that on the left-hand side of this equation can be built up. By multiplying together n simple numbers of the form $x - \alpha$, $x - \beta$, $x - \gamma$, etc., it can be shown that

$$(x - \alpha)(x - \beta)(x - \gamma), \text{ etc. . . . } n \text{ factors}$$

$$= x^n - (\alpha + \beta + \gamma + \ . \ . \ .) x^{n-1}$$
$$+ (\alpha\beta + \beta\gamma + \gamma\delta + \ . \ . \ .) x^{n-2}$$
$$- (\alpha\beta\gamma + \beta\gamma\delta + \ . \ . \ .) x^{n-3}, \text{ etc. } . \ . \ .$$
$$(-1)^n (\alpha\beta\gamma\delta \ . \ . \ .)$$

the coefficient of x^{n-1} consisting of minus the sum of all the α, β numbers, that of x^{n-2} being plus the sum of all the products two at a time of the α, β numbers, and so on. The coefficient of x^0 (that is, the constant term not containing x) will be the sum of all the products n at a time, that is, the product of all the α, β numbers, and its sign will be plus if n is even and minus if n is odd, which fact is conveniently expressed by the factor $(-1)^n$. The reader can easily confirm this general formula by multiplying two, then three, then four, simple factors together and arranging the product in each case in the above manner. For shortness, put

$$(\alpha + \beta + \gamma + \delta + \ . \ . \ .) = -b/a,$$
$$(\alpha\beta + \beta\gamma + \gamma\delta + \ . \ . \ .) = c/a,$$
$$\text{etc., etc.,}$$
$$(\alpha\beta\gamma\delta \ . \ . \ .) = (-1)^n k/a.$$

Then

$$(x - \alpha)(x - \beta)(x - \gamma) \text{ etc. } . \ . \ . \ n \text{ factors}$$
$$= x^n + (b/a) x^{n-1} + (c/a) x^{n-2}, \text{ etc. } . \ . \ . \ k/a,$$

and multiplying each of these equal numbers by a will give

$$a (x - \alpha) (x - \beta) (x - \gamma) \ . \ . \ . \ n \text{ factors}$$
$$= ax^n + bx^{n-1} + cx^{n-2} + dx^{n-3}, \text{ etc. } . \ . \ . \ k.$$

The above complicated number of the nth degree in x can therefore be expressed as the product of n simple numbers of the form $x - \alpha$, and the general equation can be written

$$ax^n + bx^{n-1} + cx^{n-2} + dx^{n-3}, \text{ etc. } . \ . \ . \ k$$
$$= a (x - \alpha)(x - \beta)(x - \gamma) \ . \ . \ . \ n \text{ factors} = 0.$$

As already explained, the equation will be satisfied when any of the individual factors are zero, that is, when $x = \alpha, \beta, \gamma$, etc. The numbers α, β, γ, etc., are therefore the roots of the equation. Thus to solve the general equation of the nth degree, it is only necessary to break it up into its n

simple factors ; but this is easier said than done. In fact, generally speaking, it cannot be done. The only general method of solution is the graphical one already described, and even that is not much use from a theoretical point of view. Suppose x, for instance, represents the resonant frequencies of some coupled circuits, the coefficients of the equation being the electrical constants of the circuits. What one usually wants to know in such a case is the way the roots depend on the coefficients, that is, one wants x explicitly in terms of a, b, c, etc., and the graphical method will only give the solutions for particular numerical cases in a form in which the individual coefficients have quite lost their identity.

However, certain general conclusions of some practical value can be drawn from the above discussion. The most important of these is that the general equation of the nth degree has n roots. Actually the above discussion only shows that an equation with n roots will be of the nth degree, but the converse is also true, though a rigid proof is rather beyond the scope of the present work. The roots will not necessarily be all different. For instance, the roots of

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

are 2, 2, since $x^2 - 4x + 4 = (x - 2)(x - 2)$. This shows that the curve $y = x^2 - 4x + 4$ cuts the x axis in two coincident points. If the left-hand curve in Fig. 11 is moved up in the positive direction of the y axis the points in which it cuts the x axis will obviously come closer and closer together and will eventually coincide. In such a case the curve is said to touch the line tangentially.

Again, the roots will not necessarily be all real. If the upward movement of the curve referred to above is continued beyond the position in which it just touches the x axis, the curve will not cut the x axis at all. The corresponding quadratic equation can still be solved, however, but it will be found that $b^2 - 4ac$ is now negative, so that $\sqrt{b^2 - 4ac}$ is imaginary, and the roots will be of the form $p + q\sqrt{-1}$, $p - q\sqrt{-1}$, as already explained. The curve represented by the general function may have several such bends, and any of these which do not come down below the x axis will give rise to similar pairs of complex roots of the form $p \pm q\sqrt{-1}$. It can be proved that any complex roots of the general equation will necessarily occur in such pairs, so that if $p + q\sqrt{-1}$ is known to be a root, then $p - q\sqrt{-1}$ must also be a root.

The relationships between the roots and the coefficients, that is, the sum of the roots equals $-b/a$, etc., etc., may sometimes be of practical value even though the roots cannot be determined, and should therefore be noted.

The expression of the general function as the product of the root factors shows that if one root of an equation can be found in some way, for example, by "guesstimation", or by plotting the function over a likely range of values, then the degree of the equation can be lowered by one by dividing by the appropriate factor (the method of division is illustrated on page 96). For instance, take

$$x^3 + 6x^2 - 3x - 4 = 0.$$

Here the sum of the coefficients is zero, which means that $x = 1$ satisfies the equation (for the result of putting $x = 1$ is obviously the sum of the coefficients). This means that $(x - 1)$ is a factor of the number on the left-hand side, and by division ✱ it will be found that

$$x^3 + 6x^2 - 3x - 4 = (x - 1)(x^2 + 7x + 4) = 0,$$

so that the roots of the equation are 1, together with the roots of the quadratic

$$x^2 + 7x + 4 = 0,$$

which can be found as described above.

An interesting special case of the general equation is

$$x^n - 1 = 0,$$

or

$$x^n = 1,$$

that is,

$$x = \sqrt[n]{1}.$$

Since this equation must have n roots, we must conclude that there are n nth roots of unity. And so there are.

✱ See special note on "Division in Algebra" at end of this section, on page 96.

For instance, consider

$$x^4 - 1 = 0.$$

Remembering the result already established for the factors of $a^2 - b^2$, that is,

$$a^2 - b^2 = (a - b)(a + b),$$

the equation can be broken up into factors thus :—

$$\begin{aligned} x^4 - 1 &= (x^2 - 1)(x^2 + 1) \\ &= (x - 1)(x + 1)(x - \sqrt{-1})(x + \sqrt{-1}) = 0, \end{aligned}$$

so that the four fourth roots of unity are ± 1 and $\pm \sqrt{-1}$. The significance of this will be made clear when we come to the interpretation of complex numbers.

A great deal more might be said about the general equation of the nth degree, but it has already taken as much space as can be allowed for it. The important things to remember are that it has n roots, that any complex roots will occur in pairs differing only in the sign of the imaginary part, and that each root gives a factor of the equation.

* [*Division in Algebra.*—The division of one more or less complex group of symbols by another in algebra is carried out in very much the same way as the ordinary process of long division in arithmetic. It can be illustrated by the case which arises in the text.

$$\begin{array}{r|l} x - 1) & x^3 + 6x^2 - 3x - 4\ (x^2 + 7x + 4 \\ & \underline{x^3 - x^2} \\ & \quad 7x^2 - 3x \\ & \quad \underline{7x^2 - 7x} \\ & \qquad 4x - 4 \\ & \qquad \underline{\underline{4x - 4}} \end{array}$$

If it is remembered that any ordinary number of several digits is really the sum of various multiples of powers of ten, for example,

$$4352 = (4 \times 10^3) + (3 \times 10^2) + (5 \times 10) + 2,$$

which can be compared with

$$4x^3 + 3x^2 + 5x + 2,$$

the identity of the above algebraic division with arithmetical division will be apparent.]

46. Simultaneous Equations for Two Unknown Numbers

It has been shown that, starting with some number represented by the letter x, some other number can be built up out of x and various other constant numbers, represented by a, b, c, etc., the more or less complicated number which results being called a function of x. Similarly one could start with two numbers represented by the letters x and y and build up some other number of more or less complicated structure, the magnitude of which would depend on the magnitudes and signs assigned to x and y. Such a number could then be described as a function of x and y. Using a similar notation, any such number could be written $F(x, y)$. In a given case, for instance, the function might be defined as

$$F(x, y) = ax^2 + bxy + cy^2 + dx + ey + f,$$

which is the most general form for a function of the second degree in x and y (second degree because no term contains more than two x's or y's multiplied together). In the above case, x and y can be regarded as independent variables, which, between them, fix the value of the function. Suppose, however, that some particular value is assigned to the function, zero, for instance, so that

$$ax^2 + bxy + cy^2 + dx + ey + f = 0,$$

then x and y are no longer independent variables, for if any given value is assigned to either, the other must have such a value or values that the sum of all the terms is zero. Obviously, in the above case, if some given value is assigned to x, k for instance, we have

$$ak^2 + bky + cy^2 + dk + ey + f = 0,$$

which is a quadratic equation for y, so that for any given value of x there will in general be two values of y which will satisfy the original definition. Thus, whatever the form of the function $F(x, y)$, the equation

$$F(x, y) = 0$$

can be regarded as defining a relationship between x and y, just as the equation $y = f(x)$ defines more explicitly a relationship between x and y. By means of the graphical method already described the relationship between the numbers x and y defined by $F(x, y) = 0$ could be represented by means of a line of more or less complicated shape. For instance, the general expression of the second degree written out above, if plotted for any particular set of values for the coefficients, will always give a curve belonging to one of the three kinds of curve that can be obtained by intersecting a cone and a plane. Such curves are called conic sections. The parabola already referred to is one kind of conic section.

Two such functions of any character,

$$F(x, y) = 0,$$
$$G(x, y) = 0,$$

plotted on the same diagram would give two lines or curves which would in general intersect one another at various points, as shown, for instance, on Fig. 12. The numbers x and y defined by any such point of intersection, since they satisfy both the functional relationships, are described as solutions of the simultaneous equations,

$$F(x, y) = 0,$$
$$G(x, y) = 0.$$

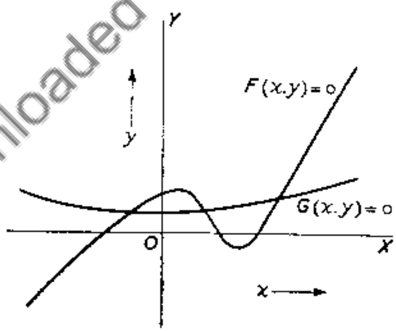


Fig. 12

More generally, any corresponding values of x and y, whether real or complex or imaginary, which satisfy both the functional relationships, are called solutions of the simultaneous equations. Thus a general method for determining the real solutions of such simultaneous equations would be the plotting of both the curves to find the point or points of intersection, though any such graphical method would be very laborious in most cases, and would have the disadvantages of lack of generality and restricted accuracy. However, the above discussion of the graphical point of view brings out one very important point. It was found that for one unknown quantity one equation is sufficient. Now it appears that if two unknown quantities have to be determined two equations are required. This suggests —though, of course, it does not actually prove—that *for the determination of any given number of unknown quantities an equal number of independent equations are required.* This is so in fact, though a perfectly general proof is beyond the scope of this work.

Out of the infinite variety of possible forms of simultaneous equations for two unknown numbers, a few of the most generally useful types will now be considered.

47. Simultaneous Equations of the First Degree for Two Unknowns

This is the simplest case, both equations being rectilinear. As it introduces ideas of wide general utility it will be considered in the general form, though literal (that is, letter) expressions tend to spread themselves out rather and look more fearsome than they are in fact. The general form will be

$$ax + by = c,$$
$$dx + ey = f.$$

These lines, being co-planar (that is, in one plane), will necessarily intersect at one point only. Even if they are parallel? Yes, for parallel lines intersect at infinity, which is only another way of saying that there is no sudden and catastrophic difference between lines which are very nearly parallel and lines which are quite parallel.

The important thing to realise about simultaneous equations is that although in general the x's and y's of the two equations are quite different and have nothing to do with each other, the x's and y's corresponding to a simultaneous solution, that is, to a point of intersection, are of necessity the same for both equations, and can therefore be combined together without making any distinction between them.

The first thing to do is to make one of the unknowns, y for instance, appear in the same way in each equation. If the first equation is multiplied by the coefficient of y in the second, and *vice versa*, this result is obtained right away, for we have

$$eax + eby = ec,$$
$$bdx + bey = bf.$$

Since the x's and y's in each equation are the same as far as the point of intersection is concerned, the equations can be subtracted from each other, giving

$$eax + eby - bdx - bey = ec - bf.$$

Remember that this derived equation is only true for x, y numbers which satisfy *both* the original equations.

From this equation $(ea - bd)\,x = ec - bf$,

so that $$x = \frac{ec - bf}{ea - bd}.$$

Now the point having this x co-ordinate lies on both lines. The corresponding y co-ordinate can therefore be found by putting this value for x in either of the original equations. Since, from the first,

$$y = -\frac{a}{b}x + \frac{c}{b},$$

then for the particular value of x above,

$$y = -\frac{a}{b}\left(\frac{ec - bf}{ea - bd}\right) + \frac{c}{b}.$$

This expression for y can be simplified as already shown in connection with the manipulation of fractions, and will give the result

$$y = \frac{af - cd}{ae - bd}.$$

Alternatively, of course, the solution for y could be found by the same process as that used for the determination of x, and in some cases this might be preferable.

To take a simple numerical example of the general method,

$$2x + 3y = 4,$$
$$7x + 8y = 9,$$

therefore

$$16x + 24y = 32,$$
$$21x + 24y = 27,$$

and subtracting the second from the first

$$-5x = 5,$$

and

$$x = -1.$$

Putting this value for x in the first equation gives

$$-2 + 3y = 4,$$

that is,

$$3y = 6,$$

therefore

$$y = 2.$$

It will be found that these values for x and y will satisfy the second equation, which checks the accuracy of the solution.

48. The Special Case of Parallel Lines: The Meaning of "Infinity"

Suppose coefficients of the general equations are such that

$$a/d = b/e,$$

that is,

$$ae - bd = 0.$$

Then the solution for x becomes

$$x = \frac{ce - bf}{ae - bd} = \frac{ce - bf}{0}.$$

Now we have already encountered the group $a/0$, and decided that it was something to which the ordinary laws of mathematics could not be applied without disastrous results. However, some rather more definite ideas on the subject must be introduced at this point, or the above equations will be left without a reasonable solution. Instead of jumping right over the precipice let us walk slowly up to the edge and look over, a very sensible proceeding which is frequently adopted in mathematics when there seems to be trouble ahead. Consider the group a/h, where h is a number which can be made smaller and smaller indefinitely. If

$$h = 1/1{,}000, \quad a/h = 1{,}000a,$$

if $$h = 1/10{,}000, \quad a/h = 10{,}000a,$$

if $$h = 1/100{,}000, \quad a/h = 100{,}000a.$$

As h is made numerically smaller and smaller, a/h becomes numerically larger and larger. By making h sufficiently small, a/h can be made larger than any number we can name, however large it may be. This is expressed mathematically by saying that the limit of a/h, when h tends to zero, is infinity—not a very consistent form of expression perhaps, because infinity means without limit, greater than any limited number. In symbols the idea is written

$$\underset{h \to 0}{lt.} \; a/h = \infty.$$

The important thing to remember is that ∞ is *not* a number, and almost anything is rather more than likely to happen if it is treated as a number.

In the above case therefore, the solution for x is infinity. For instance, if

$$3x + 4y = 5,$$

and $$21x + 28y = 4,$$

the solution for x is

$$x = (140 - 16)/(84 - 84) = \infty,$$

and for y, $$y = (105 - 12)/(84 - 84) = \infty.$$

If these lines are actually plotted on a cartesian diagram, it will be found that they are parallel. This is what is meant by saying that parallel lines meet and intersect at infinity.

49. An Electrical Example

Now, in case the more experimentally minded reader is beginning to lose interest in x's and y's and the like, it will be well to apply the above to the solution of a simple D.C. circuit problem on the basis of Kirchhoff's Laws. For the circuit shown in Fig. 13 the summation of the potential differences will give

$$(i_1 - i_2)\, R_1 = e,$$

and $$(i_2 - i_1)\, R_1 + R_2\, i_2 = 0,$$

that is,

$$i_1 - i_2 = e/R_1,$$

and $$R_1\, i_1 - (R_1 + R_2)\, i_2 = 0.$$

The solution for i_1 is

$$i_1 = \frac{e\,(R_1 + R_2)/R_1}{(R_1 + R_2) - R_1},$$

which can be rearranged to give

$$i_1 = \left(\frac{1}{R_1} + \frac{1}{R_2}\right)e,$$

Fig. 13

that is, $$i_1 = \frac{e}{R},$$

where $$\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2}.$$

The solution of the simultaneous equations given above is therefore one way, though not the shortest, of arriving at the well known expression for the resistance of two resistors in parallel.

50. Simultaneous Equations for Two Unknowns. One Equation Linear

This comes next in order of complexity to the case of two linear equations. The method of solution in this case is fairly obvious and will be best described by means of a practical example. Suppose

$$4x + 2y = 5$$

and

$$3x^2 - 4xy - 4y^2 = 0$$

are given as the two equations. The first is a straight line and the second a conic. In general a straight line will cut a conic in two points, but of course these points may be imaginary and in some cases either or both may be at infinity. That, however, is by the way, and only indicates that there will be two solutions, that is, two pairs of values of x and of y, for the above equations. From the first equation we can find the value of y in terms of x, that is,

$$y = (5 - 4x)/2,$$

and since at any points of intersection the x's and y's will be the same in the two equations, this value of y can be substituted in the second equation, that is,

$$3x^2 - \frac{4x(5 - 4x)}{2} - \frac{4(5 - 4x)^2}{4} = 0,$$

which can be simplified to

$$-5(x^2 - 6x + 5) = 0,$$

that is, $$-5(x - 5)(x - 1) = 0,$$

whence $$x = 1 \text{ or } 5.$$

From the linear equation, if

$$x = 1, \quad y = [5 - (4 \times 1)]/2 = \tfrac{1}{2},$$

and if

$$x = 5, \quad y = [5 - (4 \times 5)]/2 = -7\tfrac{1}{2}.$$

The solutions are therefore

$$\left.\begin{array}{l} x = 1 \\ y = \tfrac{1}{2} \end{array}\right\} \quad \left.\begin{array}{l} x = 5 \\ y = -7\tfrac{1}{2} \end{array}\right\}$$

In general terms, the value of y in terms of x, as determined from the linear equation, is substituted in the other equation, which then becomes an equation in one unknown only, and if this latter equation can be solved, the simultaneous solutions are easily obtained as shown.

51. Simultaneous Equations for Two Unknowns. Both of Second Degree

In this case each equation will in general represent a conic. Two conics will intersect in four points, so that four solutions will have to be obtained, and in the perfectly general case the determination of the solutions will require the solving of a fourth-power equation, which cannot always be done. There are, however, two special cases in which the solution presents no difficulty, and these will now be described very briefly. If one of the equations is homogeneous of the second degree, that is, contains only terms of the second degree (x^2, y^2, or xy), for example—

$$y^2 - 11xy + 24x^2 = 0,$$

then dividing this equation by x^2 will give

$$(y/x)^2 - 11\,(y/x) + 24 = 0,$$

which reduces it to a quadratic in y/x. This can be solved either by inspection or by the formula. In the given case we have by inspection

$$\{(y/x) - 3\}\{(y/x) - 8\} = 0,$$

so that $\quad y = 3x$, or $y = 8x$.

These values for y substituted in the other equation will reduce it to a simple quadratic in x which can be solved in the ordinary way.

Again, if terms of the first degree are absent from *both* equations, as in

$$2x^2 + 3xy + 4y^2 = 2,$$

and

$$5x^2 + 2xy + 2y^2 = 7,$$

then out of these two equations a homogeneous equation can be constructed. Multiplying the first by the constant term of the second, and the second by the constant term of the first, and then subtracting,

$$14x^2 + 21xy + 28y^2 = 14,$$

and

$$10x^2 + 4xy + 4y^2 = 14,$$

so that,

$$4x^2 + 17xy + 24y^2 = 0,$$

and from this point the solution proceeds exactly as in the previous case.

The subject of simultaneous equations for two unknowns is obviously capable of almost indefinite expansion, but most of the cases which permit of simple solution belong to one or other of the types considered above. A little practice and experience will soon enable one to judge whether any given equations are solvable or not. As an instance, an unsolvable pair is included in the examples given below.

Equations having more than two unknown numbers are very frequently met with in electrical theory, such equations being obtained from the application of Kirchhoff's Laws to more or less complicated current networks. Some brief account of the way of dealing with such equations will therefore be given in the next section.

Examples VI

1. Solve $x^3 - 4x^2 - 39x - 54 = 0$, given that -2 is one solution.

2. Solve $x^4 - 36x^3 + 193x^2 - 410x + 600 = 0$, given that $1 - 2\sqrt{-1}$ is one solution.

3. If a, b, and c are the roots of $x^3 + qx^2 + r = 0$, find the value of $(1/a) + (1/b) + (1/c)$.

4. Solve the equations

(*a*)
$$10x + y/3 = 37,$$
$$2x - y = -79,$$

(*b*)
$$\frac{3}{3x - 2y} = \frac{5}{2x + 7y} = 2.$$

5. Solve

(*a*)
$$x + y = 5,$$
$$x^2 + y^2 = 13,$$

(*b*)
$$3x^2 + 2xy - y^2 = 0,$$
$$x^2 + y^2 + 2x = 12,$$

(*c*)
$$x^2 + 3y^2 = 43,$$
$$x^2 + xy = 28.$$

6. Solve

$$x^2 + 3x + y = 5,$$
$$2y^2 + x - 4 = 0.$$

52. Simultaneous Equations for Several Unknown Numbers

It has already been indicated that if several unknown numbers have to be determined an equal number of separate and independent equations will be required. A perfectly general discussion of m equations of the nth degree would take the writer out of his depths and bore the reader intolerably, besides wasting his time. The solution of several first degree equations for an equal number of unknowns is, however, quite another matter, for it is comparatively simple, and of great practical importance, in applied electricity. As already mentioned, such equations arise from the application of Kirchhoff's Laws to direct or alternating current networks, and are therefore of particular interest to radio students.

To save time, space, and trouble, the representative example of three equations for three unknowns will be considered. It will be found that the method is the same as that which has been applied to the solution of two equations for two unknowns. The method is, in fact, capable of indefinite extension. Take as the equations

$$x + 2y + 3z = 4, \quad \ldots \quad \ldots \quad (1)$$
$$2x + 3y + 4z = 5, \quad \ldots \quad \ldots \quad (2)$$
$$3x + 6y + 7z = 6. \quad \ldots \quad \ldots \quad (3)$$

Multiply (1) by 4 and (2) by 3 in order to make the coefficients of z the same in each, so that

$$4x + 8y + 12z = 16 \quad \ldots \quad \ldots \quad (4)$$
$$6x + 9y + 12z = 15 \quad \ldots \quad \ldots \quad (5)$$

and by subtracting (4) from (5),

$$2x + y = -1. \quad \ldots \quad \ldots \quad (6)$$

Another equation in x and y can be obtained from equations (2) and (3) in a similar manner, or from (1) and (3) if this is preferred for any reason. Multiplying (2) by 7 and (3) by 4,

$$14x + 21y + 28z = 35 \quad \ldots \quad \ldots \quad (7)$$
$$12x + 24y + 28z = 24 \quad \ldots \quad \ldots \quad (8)$$

and subtracting (8) from (7),

$$2x - 3y = 11. \quad \ldots \quad \ldots \quad \ldots \quad (9)$$

Also, from (6),

$$2x + y = -1. \quad \ldots \quad \ldots \quad \ldots \quad (6)$$

In this case it happens that the coefficients of x are the same in these two equations. If they were not, the same operation would induce them to be so. As it is, we can simply subtract (6) from (9), giving

$$-4y = 12, \quad \ldots \quad \ldots \quad \ldots \quad (10)$$

or

$$y = -3. \quad \ldots \quad \ldots \quad \ldots \quad (11)$$

Substituting this value for y in (6),

$$2x - 3 = -1, \quad \ldots \quad \ldots \quad \ldots \quad (12)$$

or

$$x = 1. \quad \ldots \quad \ldots \quad \ldots \quad (13)$$

Now, putting $x = 1$ and $y = -3$ in (1),

$$1 - 6 + 3z = 4. \quad \ldots \quad \ldots \quad \ldots \quad (14)$$

Therefore

$$z = 3, \quad \ldots \quad \ldots \quad \ldots \quad (15)$$

and, rounding them up, we have as the simultaneous solutions of the three equations

$$x = 1,$$
$$y = -3,$$
$$z = 3,$$

It may be pointed out as a matter of interest that the cartesian diagram already considered in its simplest form with the two x and y axes at right angles to each other can be completed by the addition of a third or z axis at right angles to both, like the adjacent edges of a cube. Three co-ordinates, x, y, and z will then define a point in space, just as two co-ordinates define a point in a plane in the simpler plane diagram. Further, just as

$$ax + by = c$$

defines a line in the x, y plane, so

$$ax + by + cz = d$$

defines a plane in space. The solution obtained above for the three plane equations gives the co-ordinates of a point common to the three planes, that is, the point of intersection of the three planes. If we could visualise a fourth dimension, the first degree equation in four unknowns could be similarly interpreted, but unfortunately we cannot, so there we have to leave the geometrical aspect of the matter.

That, however, would not prevent us from solving four such equations for four unknowns. From any two, one of the unknowns can be eliminated. From three different pairs, the same unknown can be eliminated three times, giving three equations for three unknowns, which can be solved as above. And so on for any given number of unknowns.

To show how such equations can arise from the application of Kirchhoff's Laws to a network, let us write down the equations for the system shown in Fig. 14, which the reader will recognise as a Wheatstone bridge. The equations are

$$e - Bi_1 - R(i_1 - i_2) - S(i_1 - i_3) = 0,$$
$$R(i_2 - i_1) + Pi_2 + G(i_2 - i_3) = 0,$$
$$G(i_3 - i_2) + Qi_3 + S(i_3 - i_1) = 0,$$

which can be re-arranged rather more tidily as

$$(B + R + S)i_1 - Ri_2 - Si_3 = e,$$
$$-Ri_1 + (R + P + G)i_2 - Gi_3 = 0,$$
$$-Si_1 - Gi_2 + (G + Q + S)i_3 = 0.$$

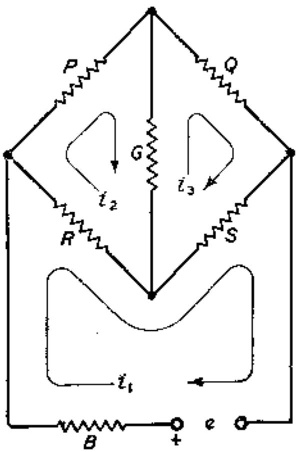


Fig. 14

The reader is strongly advised to solve these equations for the three currents i_1, i_2, i_3, but is warned to secure largish sheets of paper for the purpose, owing to the unrestrained prolixity of literal expressions. In the present instance, particular interest attaches to the current through the branch of resistance G, which represents the resistance of a galvanometer or similar measuring instrument. The reader should be able to show that this current, that is, $i_2 - i_3$, is given by

$$i_2 - i_3 = (RQ - SP)\, e/K$$

where K is the simple but voluminous expression

$$K = BGP + BGQ + BGR + BGS + PQR + PQS + PRS + QRS + GPR + GQS + GPS + + GQR + BPQ + BRS + BSP + BRQ.$$

This shows that the current through G will be zero when

$$RQ - SP = 0,$$

that is, when $$\frac{P}{Q} = \frac{R}{S},$$

which is the well known balance condition for the simple resistance bridge.

Later on it will be found that even the most complicated networks carrying alternating currents can be similarly analysed, though for this purpose several additional ideas will be required.

53. Complex Numbers: The Symbol "i"

The reader has already a bowing acquaintance with $\sqrt{-1}$, and has possibly decided that this somewhat perplexing symbol, being devoid of that comfortable and concrete reality that attaches to real numbers, is not likely to count for much in practical politics. Associated with this symbol, however, are certain ideas which will eventually prove to be of very great value in connection with alternating current problems, so time will not be ill spent in cultivating a somewhat closer acquaintance. It is, moreover, an intrinsically interesting symbol, and, under the pet name of "*i*" enjoyed what Prof. Whitehead described as a *succès de scandale* when it first made its appearance in mathematics.

At present all we know about it is the definition

$$\sqrt{-1} \times \sqrt{-1} = i \times i = i^2 = -1,$$

and it is worth noting that this definition assumes the possibility of associating the symbol i with the idea of multiplication, so that $ai \times bi$ for instance, can be taken to mean

$$a \times i \times b \times i = a \times b \times i \times i = abi^2 = -ab.$$

This is an instance of that possibility envisaged in Section

13, where ideas, in themselves incomprehensible, can be employed without violating one's intellectual self-respect, it being understood that any such incomprehensible operations stand not on their own merits, but on the validity of the conclusions they lead to. Moreover, this is in fact the classical instance referred to in this same paragraph, for it will appear later on that a clear and simple interpretation can be found for these apparently incomprehensible operations.

54. Addition of Complex Numbers

A group, such as $a + bi$ (or $a + ib$, it does not matter which way it is written), is called a complex number. Such numbers have been met with in the solution of quadratic equations in the preceding section. Such complex numbers can be compared with the "double groups" referred to in Section 5, and, in the addition of such complex numbers, the same ideas will apply. That is to say, in $a + ib + c + id$, the numbers a and c can be added arithmetically, and the bi's can be added to the di's, giving $(b + d)$ i's, so that

$$a + ib + c + id = (a + c) + i(b + d),$$

for instance

$$3 + i4 + 7 + i8 = 10 + i12,$$

and, of course, similarly for subtraction. In particular,

$$(a + ib) + (a - ib) = 2a,$$

and

$$(a + ib) - (a - ib) = 2ib.$$

The first of these statements involves the idea that $i(b - b)$, that is, $i \times 0$ is 0, which is seen to be quite reasonable if $i \times 0$ is regarded as $\sqrt{-0}$, for -0 is by definition the same as $+0$.

55. The Complex Zero

Let

$$a + ib = 0,$$

where a and b are real numbers. Subtracting ib from each side gives

$$a = -ib.$$

But this is a contradiction in terms, for a is a number and

ib is not. To make this clearer, squaring each side would give

$$a^2 = (-ib)^2 = (ib)^2 = i^2 \times b^2 = -1 \times b^2 = -b^2,$$

which is impossible, since both a^2 and b^2 are necessarily positive numbers, whatever the signs of a and b may be. However, the statement $a = -ib$ is not completely impossible, because it is true if both a and b are zero, *but only in this case*. The statement

$$a + ib = 0$$

therefore implies $a = 0$ and $b = 0$.

Further, if we are given that

$$x + iy = a + ib,$$

subtracting $a + ib$ from each complex number will give

$$(x - a) + i(y - b) = 0,$$

so that $(x - a) = 0$ and $(y - b) = 0$,

or $x = a$ and $y = b$.

The original equation is therefore really equivalent to two separate equations. This process is referred to as equating the real and the imaginary parts.

56. The Multiplication of Complex Numbers

On the above understanding (Section 53) with respect to the association of i with the idea of multiplication, the multiplication of complex numbers follows the ordinary rules, that is,

$$\begin{aligned}(a + ib)(c + id) &= ac + ibc + aid + ibid \\ &= ac + ibc + iad + iibd \\ &= (ac - bd) + i(bc + ad).\end{aligned}$$

In particular,

$$(a + ib)(a - ib) = a^2 + b^2.$$

Two complex numbers such as these, differing only in the sign of the imaginary part, are said to be mutually conjugate. Notice that the sum and the product of conjugates are wholly real. Notice further that in general the sums or products of complex numbers are other complex numbers, so that if $f(x)$ be any built-up number composed of various integral powers of x associated with constant coefficients (*cf.* the general equation of the *n*th degree), then if x be given a complex value $(a + ib)$, $f(a + ib)$ will be a complex number, that is,

$$f(a + ib) = P + iQ,$$

just as $f(x)$ is a real number if x is given any real value.

57. The Modulus of Complex Numbers

The modulus of $(a + ib)$ is the positive value of the square root of $(a^2 + b^2)$, that is,

$$mod.\ (a + ib) = \sqrt{a^2 + b^2}.$$

For instance, the modulus of $(12 + i5)$ is $\sqrt{12^2 + 5^2}$ that is, 13. Notice that the modulus of a complex number is the same as that of its conjugate.

A very useful property of moduli is that the modulus of the product of two complex numbers is the same as the product of their moduli. This is easily proved, for

$$\begin{aligned} mod.\ (a + ib)(c + id) &= mod.\ [(ac - bd) + i(bc + ad)] \\ &= \sqrt{(ac - bd)^2 + (bc + ad)^2} \\ &= \sqrt{a^2c^2 + b^2d^2 + b^2c^2 + a^2d^2} \\ &= \sqrt{(a^2 + b^2)\ (c^2 + d^2)} \\ &= \sqrt{(a^2 + b^2)}\ \sqrt{(c^2 + d^2)} \\ &= mod.\ (a + ib) \times mod.\ (c + id). \end{aligned}$$

This can obviously be extended to the product of any number of complex numbers, and thence to integral powers of complex numbers, so that

$$mod.\ (a + ib)^n = \{mod.\ (a + ib)\}^n = \{\sqrt{a^2 + b^2}\}^n = (a^2 + b^2)^{n/2}.$$

Further, this can be shown to be true for negative or fractional values of n,

that is, let $\qquad (a + ib)^{1/m} = P + iQ.$

Then, from the definition of the fractional index,

$$(a + ib) = (P + iQ)^m,$$

$$mod.\ (a + ib) = mod.\ (P + iQ)^m = \{mod.\ (P + iQ)\}^m = \{\sqrt{P^2 + Q^2}\}^m.$$

Therefore $\sqrt{a^2 + b^2} = \{\sqrt{P^2 + Q^2}\}^m,$

whence $\{\sqrt{a^2 + b^2}\}^{1/m} = \sqrt{P^2 + Q^2},$

or $\{mod.\ (a + ib)\}^{1/m} = mod.\ (a + ib)^{1/m},$

and similarly for the general fractional index and for negative indices. The modulus of a more or less complicated complex number can thus be written down at sight. For instance,

$$mod.\ \frac{(a + ib)^m\ (c - id)^{p/q}}{(e + if)^r} = \frac{\sqrt{\{a^2 + b^2\}^m}\ \{\sqrt{c^2 + d^2}\}^{p/q}}{\{\sqrt{e^2 + f^2}\}^r}$$

Is the modulus of the sum of two complex numbers the same as the sum of their moduli? It isn't, but the reader is advised to prove this for himself.

58. Application to the General Equation of the *n*th Degree

Complex numbers were first encountered in connection with the solution of quadratic equations, and it was indicated that they would also occur in the solution of the general equation of the *n*th degree. It was further stated that complex roots would always occur in pairs (of conjugates). This can now be proved quite simply. First it is required to show that if

$$(a + ib)^n = P + iQ$$

then $(a - ib)^n = P - iQ.$

If $(a + ib)^n = P + iQ,$

$$\frac{1}{(a + ib)^n} = \frac{1}{P + iQ},$$

therefore $\dfrac{(a - ib)^n}{(a + ib)^n\ (a - ib)^n} = \dfrac{(P - iQ)}{(P + iQ)\ (P - iQ)},$

that is, $\dfrac{(a - ib)^n}{\{(a + ib)\ (a - ib)\}^n} = \dfrac{P - iQ}{P^2 + Q^2}$

or $\dfrac{(a - ib)^n}{(a^2 + b^2)^n} = \dfrac{P - iQ}{P^2 + Q^2}.$

But since

$$(a + ib)^n = P + iQ,\ (\sqrt{a^2 + b^2})^n = \sqrt{P^2 + Q^2},$$

as shown on page 114, so that

$$(a^2 + b^2)^n = P^2 + Q^2.$$

Therefore $$(a - ib)^n = P - iQ.$$

Consider now the general equation of the nth degree, that is,

$$f(x) = ax^n + bx^{n-1} + cx^{n-2} + dx^{n-3} + \text{etc.} \ldots \; k = 0.$$

For any complex value $(\alpha + i\beta)$ of x, each separate power of x will give rise to a complex number of the form $P + iQ$, and the sum of all these will be some complex number, say, $M + iN$. It follows from the above that for the conjugate value $(\alpha - i\beta)$ for x, the sum of all the separate terms will be $M - iN$, that is, if

$$f(\alpha + i\beta) = M + iN,$$
$$f(\alpha - i\beta) = M - iN.$$

Now suppose that $(\alpha + i\beta)$ is known to be a root of the equation, so that

$$f(\alpha + i\beta) = M + iN = 0.$$

Then, from Section 55,

$$M = 0 \text{ and } N = 0,$$

so that $$f(\alpha - i\beta) = M - iN = 0.$$

Therefore $(\alpha - i\beta)$ must also be a root of the equation.

59. The Square Root of a Complex Number

If $(x + iy)$ be the square root of $(a + ib)$, then, by definition,

$$(x + iy)^2 = a + ib.$$

For many purposes it will be necessary to know x and y in terms of a and b, that is, to find the real and imaginary parts of $(a + ib)^{\frac{1}{2}}$. It can be done in this way. Multiply $(x + iy)$ by itself as shown in Section 56.

$$(x + iy)^2 = (x + iy)(x + iy) =$$
$$x^2 - y^2 + 2ixy = a + ib.$$

Therefore, as shown in Section 55,

$$x^2 - y^2 = a$$

and $$2xy = b.$$

These two equations can be solved for x and y by methods already described. In this particular case, however, the work can be shortened by making use of the fact that the square of the modulus of $(x + iy)$ will be equal to the modulus of $(a + ib)$, so that

$$x^2 + y^2 = +\sqrt{a^2 + b^2}.$$

For shortness, and also to avoid typographical complication later on, we shall write r for $+\sqrt{a^2 + b^2}$, so that we have the two equations

$$x^2 + y^2 = r,$$
$$x^2 - y^2 = a,$$

whence, by addition and subtraction,

$$x^2 = \tfrac{1}{2}(r + a),$$
$$y^2 = \tfrac{1}{2}(r - a),$$

that is,

$$x = \pm\sqrt{\tfrac{1}{2}(r + a)},$$
$$y = \pm\sqrt{\tfrac{1}{2}(r - a)}.$$

Thus x and y are determined. In the matter of signs we seem to have an *embarras de richesse*, but the apparent superfluity can be disposed of in this way: Since $2xy = b$, x and y must be of the same sign if b is positive and *vice versa*. Therefore if b is positive

$$\sqrt{a + ib} = \pm\{\sqrt{\tfrac{1}{2}(r + a)} + i\sqrt{\tfrac{1}{2}(r - a)}\},$$

and if b is negative

$$\sqrt{a + ib} = \pm\{\sqrt{\tfrac{1}{2}(r + a)} - i\sqrt{\tfrac{1}{2}(r - a)}\}.$$

To take a simple example, consider $\sqrt{3 + i4}$. Here a is 3 and b is 4, so that r is $+\sqrt{9 + 16} = 5$, $(r - a)$ is 2, and $(r + a)$ is 8.

Therefore

$$\sqrt{3 + i4} = \pm\{\sqrt{\tfrac{1}{2} \times 8} + i\sqrt{\tfrac{1}{2} \times 2}\}$$
$$= \pm(2 + i),$$

and if the reader has any doubts about it he can square $\pm(2 + i)$ to make sure.

There is a peculiar fascination about the subject of complex numbers, but this, unfortunately, is all the space that can be allowed for it at present. Once more the reader is cautioned against dismissing the subject as academic on the grounds that an imaginary quantity cannot have any practical significance. Though academic in appearance, these same ideas can, as it were, doff hood and gown and set about a job of real work.

Examples VII

1. Solve the equations :—

(*a*)
$$x + y + z = 12,$$
$$x + 2y + 3z = 26,$$
$$5x + 2y + z = 28.$$

(*b*)
$$x + y = 2z,$$
$$5x + 4y + 3z = 12,$$
$$21x - 20y - 2z = -1.$$

(*c*)
$$43x + 19y + 10z = 100,$$
$$100x + y - 3z + 30 = 0,$$
$$x - y + z = 10.$$

2. Solve :—

$$w + x + y = 6,$$
$$2x + 3y + 4z = 29,$$
$$4y + 2z - 3w = 17,$$
$$8z - 5w + 7x = 41.$$

3. Solve the equations :—

$$mx + ny + z = (1 + m + n)a + (3 + m + 2n)b,$$
$$2x + 3y - 5z = -7b,$$
$$5x + 8y - 7z = +6a.$$

4. Show that :—

$$\frac{a + ib}{c + id} = \frac{ac + bd}{c^2 + d^2} + i\frac{bc - ad}{c^2 + d^2}.$$

5. Find the moduli of :—

(*a*) $$\frac{(3 + i)(7 - i)}{(2 - i)},$$

(*b*) $$\frac{(3 - 4i)}{(3 + 4i)} + \frac{(12 - 5i)}{(12 + 5i)}.$$

6. Show that :—

$$\sqrt{i} = \pm (1 + i)/\sqrt{2},$$
$$\sqrt{-i} = \pm (1 - i)/\sqrt{2},$$

and hence show that the eight eighth roots of unity are :— $\pm 1, \pm i, (1 \pm i)/\sqrt{2}, \pm (1 - i)\sqrt{2}.$

Chapter 4

CONTINUITY: LIMITS: SERIES

60. The Continuity of Functions

THE idea conveyed by the group of symbols $y = f(x)$ has already been explained, but it has been considered hitherto from what may be described as the static point of view, that is, we have thought of y as a number the value of which depends in some specified manner on the value assigned to the independent variable x. ("Independent variable", by the way, is rather a mouthful. From now on we shall use the older name "argument" instead. It is less explicit, but its meaning should be quite clear at this stage.) Another and rather more important aspect of the matter is suggested by the phrase "the *behaviour* of a function". It is, as it were, the dynamic aspect, and is concerned not so much with individual values of the function but rather with the succession of values corresponding to a continuous variation of the argument. Putting it in graphical terms, we are going to consider the shape and other characteristics of the line which represents the variation of y with x. The ideas we shall meet in doing so are among the most important in the whole of mathematics, and must be taken seriously by anyone who wants to cultivate a mathematical habit of mind as distinct from a specious fluency in the tricks of the trade.

To fix ideas we shall specify the function in the form

$$y = 10^{1/(x-1)} + 1.$$

The graph, or picture, of this function is given in Fig. 15 for a range of values of the argument from $x = -6$ to $x = +8$ (any such range of values is called "an interval" of values of the argument). The most noticeable feature of the curve is that it appears to break up into two parts, the left-hand part terminating abruptly at the point for which $x = 1$, while the right-hand part seems to come flying sheer down out of the blue, after the manner of a dive bomber making an attack, "flattening out" towards the line for which $y = 2$. It is clear that something very drastic happens to y when x is given the value 1, and a more detailed examination of this region will be made later. Everywhere else the curve is a smooth, unbroken, continuous line without any sharp angles or sudden changes in direction, showing that y changes gradually with x without any sudden jumps from a small to a large value or *vice versa* for a small change in x. Variation of this character is described as "continuous", and the function is said to be "continuous" through any such interval. On the other hand, a point at which there is an abrupt transition in value, such as occurs when x is given the value 1, is described as a point of "discontinuity", and the function is said to be "discontinuous" at any such point. It is obvious that in carrying out operations with a function one must look for such points, for certain operations which may be quite legitimate and safe in a region of continuity might lead to disastrous conclusions if the function has a point of discontinuity in the interval concerned. To quote an example from Professor Whitehead, a man who walked over the edge of Shakespeare Cliff on the assumption that the height of the ground above sea level was a continuous function of his distance from Dover, would be dead before he had time to rearrange his ideas on the subject.

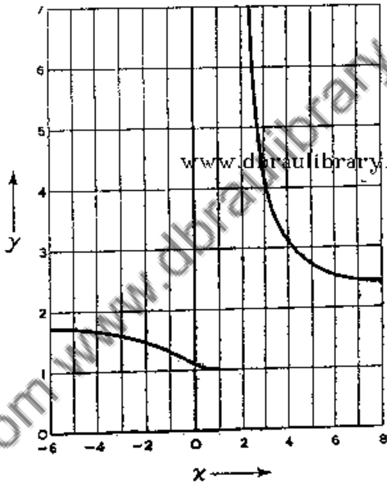


Fig. 15

The continuity or otherwise of a function is therefore a matter of practical importance. That being so, we must try to find some exact description of what we mean by continuity, some test which can be applied over any region in which there is cause to suspect irregularity of conduct on the part of the function; for although the above preliminary account conveys the general idea, those readers who have acquired the fastidiousness in thinking which is essential for mathematics will certainly not be satisfied by so vague and woolly a description. "Gradually", for instance—what does "gradually" mean in terms of the fundamental ideas of mathematics? The practical man might be tempted to reply that the variation of y is gradual over any region in which a small change in x produces a correspondingly small change in y. But what does "small" mean? This is probably where the practical man begins to get annoyed; but we cannot help it. Common sense is *not* enough. It is essential to realise that there is no abstract quality of absolute largeness or smallness in numbers, or indeed in anything else, for all magnitude is relative. Even in ordinary conversation the word "large" is so vague as to be meaningless apart from its context, stated or implicit. When the Englishman says, "That is a large apple", he means that it is large compared with, that is, larger than, the average apple of his experience, and when the American replies, "Large! Call that a large apple?" he is presumably thinking of the pumpkin-sized affairs which he has no doubt could be found growing on his uncle's farm in California, if he had an uncle in California. Here is one and the same thing called large by one person and small by another, and such examples could be multiplied indefinitely to show that "large" and "small" are no more than current and convenient abbreviations for "larger than" and "smaller than". This is equally true of mathematics, which is, after all, no more than an idealisation of experience. Of two numbers, one can be smaller than, equal to, or larger than the other, and those are the only fundamental ideas about number which can be admitted in the description of a mathematical conception. Our description of continuity must, therefore, dispense with any absolute "large" or "small" and employ only those fundamental ideas about number on which the whole science of mathematics is based.

First we need some word which will serve to distinguish a point of a function from a restricted interval of values containing the point, that is, extending on either side of it. The word used is "neighbourhood". Notice particularly that a statement cannot be made about the neighbourhood of a point unless there is a finite interval of values, containing the point, of which the statement is true. Notice further that a statement made about the neighbourhood of a point may or may not be true of the point itself. Nothing is said about that when the neighbourhood is mentioned. It is in fact a very useful feature of the word "neighbourhood" that it distinguishes between the point and some restricted interval of values containing the point. Now take the function we are considering, that is,

$$y = 10^{1/(x-1)} + 1.$$

For the value 3/2 for x it is very easy to show that the value of y is 101; but we cannot say that the function has the value 101 in the neighbourhood of $x = 3/2$, for there is no finite *interval* of values of x containing 3/2 of which this is true. But we can say that in the neighbourhood of $x = 3/2$ y differs from 101 by less than .1, or, another way of saying the same thing, approximates to 101 within a standard of .1, because a finite interval of values of x can be found for which this is true. By a simple calculation which need not be detailed, it can be shown that y approximates to 101 within a standard of .1 for all values of x between 1.4999 and 1.5001. In the present instance we could be more exact still, and say that y approximates to 101 within a standard of .00001 in the neighbourhood of $x = 3/2$, because again a finite interval of values containing 3/2 can be found for which this is true, though of course it will be a very much smaller interval than before. Actually, however small the standard of approximation be taken, it can be shown that y approximates to 101 within that standard in the neighbourhood of $x = 3/2$, and we can say at once that y approximates to 101 in the neighbourhood of $x = 3/2$ within *every* standard. This is exactly what is meant by continuity. Expressed more formally the statement becomes: A function $f(x)$ is continuous for a value a of the argument when in the neighbourhood of the point for which $x = a$ its value approximates to $f(a)$ (that is, its value at a) within every standard. The full beauty of this definition will not perhaps be realised all at once, but it will repay thinking about, for it is a fine example of the precision of mathematical thought. A lawyer experienced in the difficulty of clothing ideas in words would recognise it with delight as a perfect fit.

The function we have been considering will pass this test at every point except that for which $x = 1$, and is therefore said to be everywhere continuous except at $x = 1$. It is not, of course, suggested that every function one encounters must be scrutinised all over with this sort of microscope. One soon becomes able to tell by inspection where critical variation is likely to occur. Such points will, generally, but not invariably, be associated with values of the argument for which zeros or infinities appear in some part of the functional expression. This does not make the matter academic, for though it is true that infinities do not occur in real life they frequently occur in the functions that we use as a convenient approximate representation of some particular slice of real life that we may be contending with. For instance, neglecting the resistance in a high frequency circuit calculation may be both legitimate and convenient in general, but, by freeing the function of its ballast it may introduce the possibility of extravagant acrobatics for certain critical values of the frequency or of the circuit constants, and it is necessary to be prepared for such happenings.

61. LIMITS

We will now consider the behaviour of the above function when x is given the value 1, and in order to see more clearly what is happening at this point we shall examine the region with a magnifying glass. In other words, we shall tabulate values of x and y through a restricted interval of values containing 1 :—

x	y
·9 ·99 ·999	$10^{-10} + 1$ $10^{-100} + 1$ $10^{-1000} + 1$
1·0	?
1·001 1·01 1·1	$10^{1000} + 1$ $10^{100} + 1$ $10^{10} + 1$

This shows that as the value 1 for x is approached from the less-than-1 side, y approximates more and more closely to 1. On the other hand, if the value 1 for x is approached from the greater-than-1 side, y increases continually, and the closer x comes to 1 the greater the value of y. The first set of values would lead one to suppose that y becomes 1 when x is 1, but the second set suggests that y becomes greater than any finite number when x is 1; but the function is single valued everywhere else, that is, for any given value of x there is only one value of y. Why then should it assume a sort of dual personality at this point? The answer is that it does not. At this point, on the contrary, it has no defined value at all, for it becomes

$$y = 10^{1/(1-1)} + 1 = 10^{1/0} + 1$$

and 1/0 is not a number at all. We saw in Section 23 that the whole structure of mathematics would collapse if 1/0 were treated as if it were a number subject to the ordinary laws of arithmetic. This, then, is the first thing to notice about this point—the function has no defined value at all when x is 1. But from the tabulated values it is clear that y has a definite value when $x = 1 - h$, however small h may be *provided it is not actually zero*. Moreover, it is clear that y can be made to approximate to 1 within any desired standard, however small, by assigning a sufficiently small value to h. Under these conditions the function is said to have a finite limit when x tends to 1, although it has no defined value at all when x *is* 1. The idea can be expressed in various ways in symbols; for instance,

$$\underset{h \to 0}{lt.} \; f(1 - h) = 1,$$

which is quite explicit, or again

$$\underset{x \to 1 - 0}{lt.} \; y = 1,$$

or

$$\underset{x \to 1 - 0}{lt.} \; f(x) = 1,$$

where $x \to 1 - 0$ is taken to mean that the value 1 for x is approached from the less-than-1 side. Notice particularly, that in any case such as this, where the function has no defined value for a given value of the argument but nevertheless approaches a finite limit as the argument approaches this value, this finite limit is never actually reached, although the function can be made to approximate to it within any desired standard. This idea of a limit is of the utmost importance in mathematics, and has been the subject of much criticism and research, particularly in recent years. Unless the reader understands it thoroughly he can never hope for anything better than a dangerous rule-of-thumb knowledge of the calculus, so he is advised to go over this part again and again if necessary, until the understanding of it is assured. The formal statement of the idea is usually expressed in some such way as this: A function $f(x)$ is said to have a limit L for a value a of its argument if for every quantity k another quantity h can be found such that when x differs from a by less than h, $f(x)$ differs from L by less than k. This has always seemed to the writer a case where the definition was much harder to understand than the idea it defined. The reader need not stick over this definition, but must make sure of appreciating the idea as illustrated in the above example.

In the example quoted, the limiting value is never actually reached, since the function has no defined value at the point; but the actual definition of a limit does not exclude the possibility that the limit of a function may be the same as its value at the point concerned. In fact, if the reader has really understood the definition of continuity, he will see that it requires that the function shall have a finite limit for the given value of the argument, that limit being the same as the value of the function at the point. In general, however, the rather difficult conception of a limit is substituted for the simpler idea of the value only where it is really necessary, that is where the value does not exist.

So much for the behaviour of the above function when x tends to $(1 - 0)$. The behaviour on the other side of this point is quite different, though a somewhat similar idea is involved. The tabulated values show that y increases very rapidly as x approaches 1 from the greater-than-1 side, and the reader should have no difficulty in seeing that y can be made to exceed any finite number by bringing x sufficiently near to 1. This is conveniently, though perhaps not very happily, expressed by saying that the limit of y under these conditions is infinity, that is,

$$\underset{x \to 1+0}{lt.} \quad y = \infty,$$

which really means that y has no finite limit at all, but continues to increase without limit as x approaches 1.

Yet another kind of limit, the antithesis of infinity, is illustrated by this function (it was chosen, of course, because of its versatility in this direction). What happens when x becomes very large compared with 1? The index of the ten is $1/(x - 1)$, and it is clear that as x becomes larger and larger this index will become smaller and smaller, By making x sufficiently large compared with 1, $1/(x - 1)$ can be made smaller than any fraction of unity, however small. *It can never be made zero by increasing* x, but it can be made to differ from zero by as little as we please by making x large enough compared with 1. Here, then, the conception of limit comes in again, and we say that the limit of $1/(x - 1)$ is zero when x tends to infinity, that is,

$$\underset{x \to \infty}{lt.} \quad 1/(x - 1) = 0.$$

It is sometimes stated in textbooks that there are two kinds of zeros in mathematics, the absolute algebraic zero arrived at as the difference of two equal numbers, that is,

$$a - a = 0,$$

and another kind arrived at, or rather never quite arrived at, as the limit of $1/a$ when a is increased indefinitely, that is,

$$\underset{a \to \infty}{lt.}\ 1/a = 0.$$

The distinction may be important in some connections, though it is rather reminiscent of the waitress who, when asked for coffee without cream, said " I'm sorry, sir, we have no cream ; but you can have it without milk ".

Since
$$\underset{x \to \infty}{lt.}\ 1/(x - 1) = 0.$$

it follows that

$$\underset{x \to \infty}{lt.}\ 10^{1/(x-1)} = 10^0 = 1,$$

and that

$$\underset{x \to \infty}{lt.}\ 10^{1/(x-1)} + 1 = 2,$$

that is,

$$\underset{x \to \infty}{lt.}\ f(x) = 2.$$

Under these conditions the line representing the function is said to approach the line $y = 2$ asymptotically, that is, it gets closer and closer to it as x is increased, but never actually reaches it. It is tangential to it, but the point of contact is at infinity.

In an exactly similar manner it can be shown that

$$\underset{x \to -\infty}{lt.}\ f(x) = 2,$$

so that we have for this function the four limiting conditions :—

$$\underset{x \to -\infty}{lt.}\ f(x) = 2,$$
$$\underset{x \to 1-0}{lt.}\ f(x) = 1,$$
$$\underset{x \to 1+0}{lt.}\ f(x) = \infty,$$
$$\underset{x \to \infty}{lt.}\ f(x) = 2.$$

Now we come to a case which occurs very frequently in practical analysis, and which might cause considerable perplexity to a person who had not assimilated the above ideas on limits. Take the function

$$y = F(x) = \frac{x^2 + 2x - 3}{x^2 + 3x - 4}.$$

In general the value of this function for any value of the argument can easily be calculated by ordinary arithmetic, but when $x = 1$ we have

$$y = F(1) = \frac{1 + 2 - 3}{1 + 3 - 4} = \frac{0}{0}.$$

Now 0/0 is a group to which no meaning can be attached in terms of the fundamental conceptions of arithmetic. What then are we to do about this? In the first place, since the two quadratic expressions vanish when x is 1, it follows that each is divisible exactly by $(x - 1)$ (see Section 45). With this clue it is easy to express the function in the form

$$F(x) = \frac{(x - 1)\,(x + 3)}{(x - 1)\,(x + 4)},$$

and now obviously we can divide the top and bottom of this fraction by $(x - 1)$, so that

$$F(x) = \frac{(x + 3)}{(x + 4)},$$

and therefore when $x = 1$,

$$F(x) = \tfrac{4}{5}.$$

All very plausible, isn't it—and quite wrong. It just shows how careful one has got to be. The top and bottom of the fraction can only be divided by $(x - 1)$ on the condition that $(x - 1)$ is *not* zero, that is, on the condition that x is *not* 1. Otherwise we are dividing top and bottom by zero, which, as we have seen, is definitely not legitimate under any circumstances whatever, so that just precisely the case in which it is essential to divide through by $(x - 1)$ is the one in which it cannot be done. However, let us stop short just on the edge of the precipice, instead of falling over it, that is, put $x = 1 + h$ instead of 1, h being a small quantity compared with 1. Then

$$F(x) = F(1 + h) = \frac{h(4 + h)}{h(5 + h)},$$

and since h is not zero,

$$F(x) = F(1 + h) = \frac{(4 + h)}{(5 + h)},$$

and this is true, however small h may be as long as it is not zero. Now by making h small enough, the fraction can be made to differ from $4/5$ by as little as we please. In other words, the limit of the fraction when x tends to 1 is $4/5$. Therefore, although $F(x)$ has no defined value when x *is* 1, it has a definite limit when x *tends to* 1, that limit being $4/5$, that is,

$$\underset{x \to 1}{lt.} F(x) = 4/5.$$

Moreover, since the whole of the above reasoning can be repeated when $x = (1 - h)$ with the same result, the limit is the same for either direction of approach, that is,

$$\underset{x \to 1 \pm 0}{lt.} F(x) = 4/5.$$

What are we to say about the continuity of the function through this critical point? It is a difficult question to answer, and, as a matter of fact, the writer cannot give a definite answer himself and has not been able to get any authoritative general statement on the point. The difficulty is that the function certainly does not satisfy the continuity test at the point, since it has no defined value; nevertheless, it will be found that the function can be plotted as a perfectly smooth and apparently continuous line through this point, and, moreover, it satisfies the continuity definition if the *limit* when x tends to 1 be substituted in the definition for the *value* when $x = 1$. Actually it is very unlikely that any error will arise in practice from assuming that this function, or any of the very large number of similar functions that are involved in practical mathematics, is continuous through this undefined and indeterminate point, but in the absence of any certainty in the matter any operations which involve the assumption of continuity will have to be carried through with some degree of mental reserve.

So much for " continuity ", " value ", and " limit ",—or, at least, so much for an elementary introduction to these difficult but intriguing ideas. It may have seemed wordy and excessively fine drawn, but it is necessary all the same, for, as Prof. Whitehead has pointed out, " large parts of mathematics as enunciated in the old happy-go-lucky manner were simply wrong ". It is even probable, or at least possible, that the refinements of modern mathematics may prove insufficient in some directions. In any case, an excess of precision, if it is a fault at all, is a fault in the right direction, and is worth pursuing not only for its own sake but for the mental training it involves. There is no room for slipshod thinking in mathematics.

Examples VIII

1. Show that the function

$$y = \frac{a + bx}{c + dx}$$

is discontinuous when $x = -c/d$. Find the limits of y when $x \to -(c/d) \pm 0$ and when $x \to \pm \infty$.

2. Find the points of discontinuity of

$$y = \frac{x^2 - 3x + 2}{x^2 - 7x + 12},$$

and find the limits of y when $x \to 3 \pm 0, 4 \pm 0, \pm \infty$

3. Find the limits when $x \to a \pm 0$ of

$$y = \frac{x^2 - (a + b)x + ab}{x^2 - (a + c)x + ac}.$$

4. Find the limits of

$$y = \frac{x^3 - 6x^2 + 11x - 6}{x^3 - 7x^2 + 14x - 8}$$

when $x \to 1 \pm 0, 2 \pm 0, 4 \pm 0.$

5. Find the limit when $x \to \infty$ of

$$y = \frac{ax^2 + bx + c}{mx^3 + nx^2 + px + q}.$$

6. Find the limit when $x \to a$ of

$$\frac{\sqrt{x} - \sqrt{a} + \sqrt{x - a}}{\sqrt{x^2 - a^2}}.$$

62. Series

The word "series" has essentially the same meaning in mathematics as in ordinary speech. Any ordered succession of numbers

$$a_1, a_2, a_3, a_4, a_5, \ldots . \ a_n$$

is called a series, but a random succession, such as

$$18, 1, 5\tfrac{1}{2}, 7\tfrac{3}{8}, 5003, \text{etc., etc.,}$$

is not of any practical interest. It is form without significance, and these numbers might have been arranged by a trained monkey or a gust of wind. But in the matter of series, as in everything else, any sign of law or design gives significance to the form, and the mind becomes interested. For instance,

$$0/1, 2/3, 4/5, 6/7, \text{etc., etc.,}$$

shows evidence of thought, and nothing less than human could have arranged this series. It is designed according to a definite plan and can be related to the simplest series of all, that of the cardinal numbers, by means of a general formula which expresses this law and summarises the design. The law can be put in the form

$$a_n = \frac{2n - 2}{2n - 1},$$

a_n being the nth term. The first term is obtained by putting $n = 1$, that is, $0/1$, the second by putting $n = 2$, that is, $2/3$, and so on. An ordered series of this kind really grows out of the idea of a function, for the general or nth term is a function of n, that is,

$$a_n = f(n)$$

where the argument n takes in succession the values 1, 2, 3, etc. Moreover, the function need not contain n and numbers only. It may contain one or more letter symbols as well. For instance,

$$1, 2x, 3x^2, 4x^3, \ldots . \ nx^{n-1}$$

is a series the terms of which are functions of x and of n. Thus an infinite variety of form is included in the general idea of a series, and the series form of representation plays a very large part in theoretical and applied mathematics.

63. The Sum of a Series

Next to the form or law of a series, the most important thing to know about it is its sum, that is, the sum of its terms. Think of a number, for instance, say, 21493. This is the sum of a series, a series of powers of ten, for it is only a short way of writing

$$(2\times 10^4) + (1\times 10^3) + (4\times 10^2) + (9\times 10^1) + (3\times 10^0).$$

Again, 3.1415 is a short way of writing

$$(3\times 10^0) + (1\times 10^{-1}) + (4\times 10^{-2}) + (1\times 10^{-3}) + (5\times 10^{-4}),$$

a series which symbolises the fundamental process of approximation. This illustrates the basic character of the summation of a series which we must now consider.

First we shall take the simple case in which there is a limited, that is finite, number of terms. The sum of such a series is of course just what the phrase implies, that is the number which results from the adding together of all the terms of the series, but in this connection the word sum is usually given a rather special sense, and finding the sum of the series means finding some general formula which provides a shorter way of arriving at the result than by the detailed addition of the separate terms. Take for instance a simple "arithmetical progression" as it is called, in which each term differs from the preceding one by a constant number, for example, 3, 7, 11, 15, etc., or, in general terms, a, $a + d$, $a + 2d$, $a + 3d$, ... $a + (n - 1)d$. The sum of n terms is

$$S_n = a + (a + d) + (a + 2d) + (a + 3d) \quad . \; . \; . \; . \quad \{a + (n - 2)d\} + \{a + (n - 1)d\}.$$

In this case a short formula can be found for the sum by means of a trick. Turning round the right-hand side,

$$S_n = \{a + (n - 1)d\} + \{a + (n - 2)d\} + \; . \; . \; . \; . \; (a + d) + a,$$

and adding these two equations term by term gives

$$2S_n = \{2a + (n - 1)d\} + \{2a + (n - 1)d\} + \{2a + (n - 1)d\}, \text{ etc.}, \quad \ldots\ldots \text{ n terms in all.}$$

$$= n\{2a + (n - 1)d\},$$

so that

$$S_n = \frac{n}{2}\ \{2a + (n - 1)d\} = n\left(\frac{a + l}{2}\right)$$

where $$l = a + (n - 1)d,$$

that is, l is the last term.

Notice that $(a + l)/2$ is thus shown to be, as indeed is obvious, the average value of all the terms of the series.

Again, for another simple form, the so-called geometric series in which each term bears a constant ratio to the preceding term, for example, $\frac{1}{4}$, $\frac{1}{2}$, 1, 2, etc., or, in general terms, a, ar, ar^2, ar^3, etc., ... ar^{n-1}, the sum of n terms is

$$S_n = a + ar + ar^2 + ar^3 + ar^4, \text{ etc.}, \quad \ldots\ ar^{n-1}.$$

Therefore

$$rS_n = ar + ar^2 + ar^3 + ar^4, \text{ etc.}, \quad \ldots\ ar^{n-1} + ar^n,$$

and by subtraction,

$$S_n(r - 1) = ar^n - a = a(r^n - 1).$$

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

For instance, the sum of 32 terms of the series $\frac{1}{4}$, $\frac{1}{2}$, 1, 2, etc., would be

$$S_{32} = \frac{\frac{1}{4}(2^{32} - 1)}{(2 - 1)}$$

$$= 2^{30} - \tfrac{1}{4}.$$

This commemorates the tragic fate of the desperate fugitive who offered a blacksmith a farthing for the first, a half-penny for the second, a penny for the third, and so on, for the 32 nails of the four shoes of his horse. The bill came to just under four and a half million pounds.

In addition to the above two simple standard series, the arithmetic* and the geometric, there are various other types which can be summed by means of general formulæ, but these are not of much practical importance, so we shall proceed at once to the much more general propositions relating to the sum of an infinite number of terms of a series.

64. The Sum to Infinity. Convergency and Divergency

First, what is meant by an infinite series? Given a general term

$$a_n = f(n),$$

say,

$$a_n = \frac{1}{2^{n-1}}$$

for instance, then if there is no upper limit specified for n, the number of terms can exceed any finite number, however large, that is, can be infinite. What, then, is the sum to infinity? It cannot mean the result of adding the terms together, for that is an unending process. It is conceivable, however, that S_n may have a finite limit when n is increased indefinitely, that is, it may be possible to find a quantity S such that, by making n sufficiently large, S_n can be made to approximate to S within any standard, however small. The quantity S is then called the sum to infinity of the series. Actually it is the limit of S_n when n tends to infinity, that is,

$$S = \underset{n \to \infty}{lt.} S_n.$$

This limit S will never be reached by the sum of any finite number of terms, but it can be approached to any desired standard of accuracy by taking a sufficiently large number of terms. Take, for example, the series quoted above for which

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

This is a geometric series of which the first term, a, is 1, and the common ratio r, $\frac{1}{2}$. From the formula given

* Pronounced with the accent on "met".

above,

$$S_n = \frac{1 \times (1 - \frac{1}{2}^n)}{(1 - \frac{1}{2})} = 2(1 - \tfrac{1}{2}^n).$$

Now by making n sufficiently large, $(1 - \frac{1}{2}^n)$ can be made to differ from 1 by as little as we please, that is, it can be made to approximate to 1 within any standard, so that

$$S = \underset{n \to \infty}{lt.} S_n = \underset{n \to \infty}{lt.} 2(1 - \tfrac{1}{2}^n) = 2.$$

In general, for the geometric series

$$S_n = \frac{a(1 - r^n)}{1 - r},$$

and provided r is less than 1 numerically, the limit of $1 - r^n$ when r tends to infinity is 1, and

$$S = \underset{n \to \infty}{lt.} S_n = \underset{n \to \infty}{lt.} \frac{a(1 - r^n)}{1 - r}$$

$$= \frac{a}{1 - r}.$$

It should already be clear that not all series will have a finite sum to infinity. In the above case, for instance, if r is equal to or greater than 1,

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

will increase without limit as n tends to infinity. In fact, series can be divided into two classes—those which have and those which have not a finite limit for the sum to infinity. Series of the first kind are called convergent, and play a very large part in applied mathematics. The others are called divergent, and are not of much use to anyone. (Note : A series will be called divergent unless S_n tends to a finite limit as n tends to infinity, in accordance with the definition of limit ; but it does not follow that S_n will increase without limit for all divergent series. For instance, the series $a, -a, a, -a$, etc., which is a geometric series having a as the first term and -1 as the common ratio, gives $S_n = 0$ or $-a$ according as n is even or odd, however large n may be. This kind of divergent series is called " oscillating ".)

It is clear that the important thing to know about an infinite series is whether or not it is convergent. If it is, then although it may not be possible to find any simple expression for the limit of the sum to infinity, as close an approximation to it as may be desired can be found by actually calculating and adding together a sufficiently large number of terms, whereas if the series is divergent, this would be a vain pursuit of something which does not exist. Where a series is a function of n and of some other independent variable x, for instance

$$1, x, x^2, x^3, x^4, \text{etc., etc.} \quad . \quad . \quad . \quad x^{n-1},$$

the sum to infinity may be finite for a certain interval of values of x and infinite or oscillating for others. In fact, we have already shown that this series is convergent for all values of x between plus and minus 1 and divergent for all other values of x. In any case of this kind it will be necessary to know for what interval or intervals of values of the argument the series is convergent.

For these reasons a great deal of research has been directed to the discovering of tests for convergence, tests which can be applied to a series as a preliminary operation, to find out whether and under what conditions a sum to infinity actually exists. The research has so far failed to establish any single test of universal application which separates the sheep from the goats; but a considerable number of tests of very useful even though limited applicability have been developed. A few of the most useful of these will now be described, but for a full account of this rather difficult and voluminous subject some standard textbook, such as *Chrystal's Algebra*, should be consulted.

65. Tests for Convergency

The series will be represented by

$$a_1, a_2, a_3, a_4, \text{etc., etc.,} \quad . \quad . \quad . \quad . \quad a_n,$$

a_n being the nth term. Two perfectly general points should be noted first. The sum of any finite number of terms is finite. Therefore if a series can pass a test for convergence for all terms after a certain point, say the rth term, then it is convergent, even though the terms up to this point do not satisfy the convergency condition. Again, if a given series of positive terms is convergent, any series differing from it only in the sign of some of the terms will also necessarily be convergent.

A first minimum test of convergence is that

$$\underset{n \to \infty}{lt.} a_n = 0,$$

for a series is not convergent unless S_n, S_{n+1}, S_{n+2}, etc., converge to a finite limit S, so that S_n, S_{n+1}, S_{n+2}, etc., differ from S, and therefore from each other, by a quantity which can be diminished without limit by making n sufficiently large. Therefore

$$\underset{n \to \infty}{lt.} (S_{n+1} - S_n) = \underset{n \to \infty}{lt.} a_{n+1} = 0$$

(Note that $\underset{n \to \infty}{lt.} a_{n+1}$ is the same as $\underset{n \to \infty}{lt.} a_n$.) A series of which this is not true cannot be convergent. Unfortunately, very unfortunately in fact, it does not follow that a series is convergent if $\underset{n \to \infty}{lt.} a_n = 0$.

For instance, the series

$$\frac{1}{1}, \frac{1}{2}, \frac{1}{3}, \frac{1}{4} \quad . \quad . \quad . \quad . \quad \frac{1}{n}$$

fulfils the condition, but it can be shown that this series is divergent. The condition is therefore a necessary but not a sufficient one. It is nevertheless a useful point to remember, and will sometimes save further investigation.

A series is convergent if the ratio of each term to the preceding one is numerically less than unity; for

$$a_1 + a_2 + a_3 + a_4 + a_5 \quad . \quad . \quad . \quad + a_n \quad . \quad . \quad .$$

etc., etc., *ad inf.*, can be written in the form

$$a_1 \left\{ 1 + \frac{a_2}{a_1} + \frac{a_3 a_2}{a_2 a_1} + \frac{a_4 a_3 a_2}{a_3 a_2 a_1} + \frac{a_5 a_4 a_3 a_2}{a_4 a_3 a_2 a_1} \right.$$
$$\left. . \quad . \quad . \quad . \quad \text{etc., etc., } ad\ inf. \right\}.$$

Now by hypothesis

$$\frac{a_2}{a_1} < r$$

$$\frac{a_3}{a_2} < r$$

etc., etc., where $r < 1$. Therefore the sum of the series is less than

$$a_1 (1 + r + r^2 + r^3 + r^4 \ . \ . \ . \ \text{etc., etc., } ad\ inf.).$$

But

$$1 + r + r^2 + r^3 + r^4 \ . \ . \ . \ \text{etc., etc., } ad\ inf.$$
$$= 1/(1 - r),$$

since r is less than 1. The sum of the series is therefore less than $a_1/(1 - r)$ and is therefore convergent.

Notice that this proof will break down unless

$$\frac{a_{n+1}}{a_n} < 1 \text{ numerically}$$

however large n may be. The series will therefore not be convergent unless

$$\underset{n \to \infty}{lt.} \frac{a_{n+1}}{a_n} < 1 \text{ numerically.}$$

Conversely, if the above condition is fulfilled, the series is convergent; for let

$$\underset{n \to \infty}{lt.} \frac{a_{n+1}}{a_n} < r, \text{ where } r < 1.$$

Then, by definition, if m is made large enough,

$$\frac{a_{m+1}}{a_m} \sim r < \epsilon,$$

however small ϵ may be, and if ϵ is made small enough,

$$r + \epsilon < 1,$$

since

$$r < 1.$$

If, therefore,

$$\underset{n \to \infty}{lt.} \frac{a_{m+1}}{a_n} < r,$$

where $r < 1$, we can find some fixed term, say the mth, from and after which a_{m+1}/a_m is less than 1. The series will therefore be convergent as already shown.

In addition to the above, there are two comparison theorems which are of use. Given two series

$$a_1, a_2, a_3, a_4, a_5 \;\ldots\; a_n \;\ldots\; \text{etc., etc., } ad\ inf.$$

and

$$b_1, b_2, b_3, b_4, b_5, \;\ldots\; b_n \;\ldots\; \text{etc., etc., } ad\ inf.$$

of positive terms, then if the a series is convergent and each term of the b series is less than the corresponding term of the a series, the b series will also be convergent. This is sufficiently obvious without any formal proof.

Another comparison, not so obvious, is this. If the ratio of corresponding terms is always finite, the series will be either both convergent or both divergent. Suppose the a series to be convergent, its sum to infinity being S_a, and let k be the largest value of the ratio b_n/a_n of any two corresponding terms. Then the sum of the b series is less than kS_a, and the b series is therefore convergent. If, on the other hand, the a series is divergent, its sum to infinity being infinite, and r is the smallest value of the ratio of two corresponding terms b_n/a_n, the sum of the b series is greater than r times the sum of the a series, and the b series is therefore divergent.

A useful test series to which the above theorems can be applied is

$$\frac{1}{1}, \frac{1}{2^k}, \frac{1}{3^k}, \frac{1}{4^k}, \frac{1}{5^k} \;\ldots\ldots\; \text{etc., etc., } ad\ inf.,$$

which can be shown to be convergent if k is numerically greater than 1, and divergent if k is equal to or numerically less than 1.

All the above tests of convergence have been stated in relation to series of positive terms, but in view of the general statement made above about series of which the terms vary in sign, there should be no difficulty in applying them to such series. It should be pointed out, however, there are series which are divergent if all the terms are positive but convergent if they are alternatively positive and negative. In fact it can easily be shown that a series having alternately positive and negative terms is convergent if each term is less than the preceding, and if the terms decrease without limit in absolute magnitude.

66. Some Important Series

One of the most useful series in the whole of mathematics is

$$1,\ mx,\ \frac{m(m-1)}{1.2}x^2,\ \frac{m(m-1)(m-2)}{1.2.3}x^3, \text{ etc., etc.}$$

This, for a reason which will appear later, is called the Binomial Series. The general or nth term is

$$a_n = \frac{m(m-1)(m-2)(m-3)\ \ldots\ (m-n+2)}{1.2.3.4\ \ldots\ (n-1)}x^{n-1},$$

and the $(n+1)$th term is therefore

$$a_{n+1} = \frac{m(m-1)(m-2)(m-3)\ \ldots\ (m-n+1)}{1.2.3.4\ \ldots\ n}x^n.$$

This, by the way, introduces a new notation. The denominator in a_{n+1} is the product of all the whole numbers from 1 to n. This is a group which very frequently occurs in mathematics, especially in series. It is written for shortness $\underline{|n}$, and is called "factorial n". There is an alternative and more recent notation in which it is written $n!$ This has the advantage that it is all on one line. It is still called "factorial n" in this form, though one of the brighter sort of mathematicians always refers to it as "n By Jove!" The name has not been adopted in teaching circles because it is liable to disturb a class.

Returning to our series, the first thing to notice is that if m is any positive whole number the series terminates at the $(m+1)$th term, because the $(m+2)$th term contains the factor

$$m-(m+2)+2=0,$$

and all subsequent terms will contain this factor also.

If m is negative or fractional, however, the series is infinite, that is, does not terminate. Under what conditions will it be convergent? In the absence of any obvious answer the first thing to investigate is the ratio of a_{n+1} to a_n. This is

$$\frac{(m - n + 1)}{n} x = -x\left(1 - \frac{m + 1}{n}\right).$$

This ratio can be made to approximate to $-x$ within any desired standard by sufficiently increasing n, however large m may be. Thus, if x is less than 1 numerically there will in every case be a term in the series from and after which the ratio of a_{n+1} to a_n is less than 1. The series is therefore convergent provided x is less than 1 numerically.

Another important series is

$$1, \frac{x}{1!}, \frac{x^2}{2!}, \frac{x^3}{3!}, \frac{x^4}{4!}, \frac{x^5}{5!} \cdot \cdot \cdot \frac{x^{n-1}}{(n-1)!} \cdot \cdot \cdot \text{ etc., etc., } ad\ inf.$$

This is called the Exponential Series. In this case

$$\frac{a_{n+1}}{a_n} = \frac{x^n}{n!} \times \frac{(n-1)!}{x^{n-1}} = \frac{x}{n}.$$

Now, however large x may be, there will always be some term in the series from and after which x/n is less than 1 numerically. This series is therefore convergent for *all* values of x.

The series

$$x, -\frac{x^2}{2}, \frac{x^3}{3}, -\frac{x^4}{4}, \frac{x^5}{5} \cdot \cdot \cdot (-1)^{n+1}\frac{x^n}{n} \cdot \cdot \cdot$$

etc., etc., *ad inf.*

is called the Logarithmic Series. Here

$$\frac{a_{n+1}}{a_n} = -\frac{nx}{n+1} = -x\left(\frac{n}{n+1}\right),$$

and this will always be less than 1 numerically provided x is less than 1 numerically. This series is therefore convergent for all values of x between plus and minus 1.

There is, of course, a host of other series which play a large part in applied mathematics, but these will be considered individually as they occur.

Examples IX

1. Show that the series of which the nth term is $an + b$ is arithmetic. What is its first term, and the sum of the second fifty terms ?

2. The first term of an arithmetic series is $n^2 - n + 1$, and the common difference is 2. Show that the sum of n terms is n^3, and thence show that

$$1^3 = 1$$
$$2^3 = 3 + 5$$
$$3^3 = 7 + 9 + 11$$
$$4^3 = 13 + 15 + 17 + 19$$
etc., etc.

3. Prove that if the sum of n terms of a series is $a(r^n - 1)$, the series is geometric. Find the first term and the common ratio.

4. Prove that in an infinite geometric series (common ratio less than 1) the ratio of any term to the sum of all that follow it is constant.

5. Show that the series of which the general term is $1/(n^2 - x)$ is convergent except when x is the square of an integer.

6. Discuss the convergency of the series having as nth terms :—
   (a) $1/(x + n - 1)$
   (b) $(1 + n)/(1 + n^2)$
   (c) $\dfrac{3.5.7 \quad \ldots\ldots \quad (2n + 1)}{4.7.10 \quad \ldots\ldots \quad (3n + 1)} x^n$

67. THE BINOMIAL THEOREM

The Binomial Series has already been introduced. It is

$$1,\ mx,\ \frac{m(m-1)x^2}{2!},\ \frac{m(m-1)(m-2)x^3}{3!}\ \text{etc., etc.,}$$

the general or *n*th term being

$$\frac{m(m-1)(m-2)(m-3)\ .\ .\ .\ (m-n+2)x^{n-1}}{(n-1)!}.$$

The further discussion of this series will lead us to one of the most famous theorems in the whole of mathematics and certainly one of the most useful—the Binomial Theorem. (It is, incidentally, one of the oldest theorems, for, according to Hogben, its foundations were laid by Omar Khayyám.)

The name sounds rather impressive, and the series itself looks very complicated and mathematical; but, after all, it is only a number, or rather a set of numbers. To make sure that our feet are still on solid ground, let us materialise this airy spirit and give it a substantial form by putting $m = 5$ and $x = 2$. The numbers then become

$$1,\ 5 \times 2,\ \frac{5 \times 4 \times 4}{2},\ \frac{5 \times 4 \times 3 \times 8}{3 \times 2}$$

$$\frac{5 \times 4 \times 3 \times 2 \times 16}{4 \times 3 \times 2},\ \frac{5 \times 4 \times 3 \times 2 \times 1 \times 32}{5 \times 4 \times 3 \times 2}$$

that is, $\quad 1,\ 10,\ 40,\ 80,\ 80,\ 32.$

There are only six terms to the series in this case, for the seventh and all subsequent terms contain the factor 0 in the numerator. The sum of the six terms is 243.

So much by way of reassurance, in case it was necessary. Now we can return to the symbols and try to find some simple formula for the sum of the series when *m* is a positive integer. We have already seen that in all such cases the series is finite and terminates at the $(m + 1)$th term. The sum of the series is clearly a function of *x* and of *m*, so we can write

$$f(x,m) = 1 + mx + \frac{m(m-1)x^2}{2!} + \frac{m(m-1)(m-2)x^3}{3!} \text{ etc., etc.}$$

Now multiply each side by $(1 + x)$ and arrange the right-hand side in ascending powers of x, as in the original series. The multiplication is quite a simple and straightforward process and the reader will have no difficulty in showing that

$$\{f(x,m)\}(1 + x) = 1 + (m+1)x + \frac{(m+1)mx^2}{2!} + \frac{(m+1)\,m\,(m-1)}{3!}x^3, \text{ etc., etc.}$$

Now the right-hand side is the original series but with $(m + 1)$ written everywhere instead of m, that is, it is $f(x, 1 + m)$,* so that

$$\{f(x, m)\}(1 + x) = f(x, 1 + m).$$

It follows from this that

$$f(x,m) = (1 + x)^m,$$

but it is rather a long jump so we shall come to it in smaller steps. Since m is any positive integer, put

$$m + 1 = r.$$

Then $$m = r - 1,$$

and we have $$(1 + x)f(x,r-1) = f(x,r)$$

or $$\frac{f(x,r)}{(1+x)} = f(x, r-1).$$

Since this is a perfectly general formula, we may say that

$$\frac{f(x,r)}{(1+x)^2} = \frac{f(x,r-1)}{(1+x)} = f(x,r-1-1) = f(x,r-2).$$

The process can be continued, giving

* To make this entirely complete and convincing it should be proved for the general term in each case. This can be done but is a rather lengthy business. The reader should be able to do it for himself.

$$\frac{f(x,r)}{(1+x)^3} = f(x, r-3),$$

and so on, up to r times, which will lead to

$$\frac{f(x,r)}{(1+x)^r} = f(x,r-r) = f(x,0).$$

But

$$f(x,0) = 1.$$

Therefore

$$\frac{f(x,r)}{(1+x)^r} = 1,$$

so that

$$(1+x)^r = f(x,r) = 1 + rx + \frac{r(r-1)x^2}{2!} + \frac{r(r-1)(r-2)x^3}{3!} \text{ etc., etc.,}$$

which is what we set out to prove. The symbol m has got changed on the way, but that doesn't matter.

If $m = 5$ and $x = 2$, the sum of the series is therefore $(1+2)^5 = 3^5 = 243$, a result which has already been demonstrated above.

So much for positive integral values of m; but the series is not inherently limited in the values that m may take, so it will be necessary to carry the investigation one stage further and find what the series means when m is negative or fractional.

It has already been shown that in such cases the series is infinite, and that it is convergent when x is less than 1 numerically, which condition will be assumed in all that follows.

Without assuming anything at all about u and v, take the two binomial series

$$f(x,u) = 1 + ux + \frac{u(u-1)x^2}{2!} + \frac{u(u-1)(u-2)x^3}{3!} \text{ etc., etc.,}$$

$$f(x,v) = 1 + vx + \frac{v(v-1)x^2}{2\,!} + \frac{v(v-1)(v-2)x^3}{3\,!} \text{ etc., etc.,}$$

multiply them together, and arrange the product in ascending powers of x. As before, the operation should really be carried out for the general term, but this would take rather too much of our limited space. Taking any finite number of terms, the reader will have no difficulty in showing that the product can be put in the form

$$f(x,u)\,f(x,v)$$
$$= 1 + (u+v)x + \frac{(u+v)(u+v-1)x^2}{2\,!} + \frac{(u+v)(u+v-1)(u+v-2)x^3}{3\,!} \text{ etc., etc.}$$
$$= f(x,u+v).$$

This is the important step, and the full interpretation of the series is implicit in this equation, for we can use it to show that

$$f(x,m) = (1+x)^m$$

even when m is negative or fractional, provided x is less than 1 numerically.

Suppose first that u is a positive integer, and that v is a negative integer equal to u in magnitude (that is, $v = -u$). Then since

$$f(x,u) = (1+x)^u,$$

u being a positive integer, and since it has been shown that

$$f(x,u)\,f(x,v) = f(x,u+v)$$

for any values of u and v, then if $v = -u$,

$$f(x,u)\,f(x,v) = (1+x)^u f(x,-u)$$
$$= f(x,u+v) = f(x,u-u) = f(x,0).$$

But
$$f(x,0) = 1.$$

Therefore $$(1 + x)^u f(x, -u) = 1,$$

or $$f(x, -u) = 1/(1 + x)^u = (1 + x)^{-u},$$

which establishes the result for a negative value of m.

From the general result

$$f(x,u)\, f(x,v) = f(x,u + v)$$

it is easy to show (as on p. 57) that

$$\{f(x,u)\}^q = f(x,uq),$$

where u has any value, and q is a positive integer.

Since u can have any value, let it be a fraction p/q, so that uq is a positive integer p. Then the equation

$$\{f(x,u)\}^q = f(x,uq)$$

becomes

$$\{f(x,p/q)\}^p = f(x,p) = (1 + x)^p.$$

Therefore

$$f(x,p/q) = (1 + x)^{p/q},$$

which proves the result for a fractional index.

To sum up,

$$(1 + x)^m = 1 + mx + \frac{m(m-1)x^2}{2\,!} + \frac{m(m-1)(m-2)x^3}{3\,!} + \text{etc., etc.,}$$

$$+ \frac{m(m-1)(m-2)\ \ldots\ (m-n+2)x^{n-1}}{(n-1)\,!} \ldots$$

for all values of x if m is a positive integer, and for all values of m provided x is less than 1 numerically. This is known as the Binomial Theorem.

It is obvious that there are many useful and important applications for this result. Take for instance the general solution of a quadratic equation:—

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

which can be put in the form

$$x = \frac{-b}{2a}\left\{1 \pm \left(1 - \frac{4ac}{b^2}\right)^{\frac{1}{2}}\right\}.$$

Now if $4ac$ is less than b^2 numerically, so that $b^2/4ac$ is less than 1 numerically, putting $x = -4ac/b^2$ and $m = \frac{1}{2}$ in the Binomial expansion gives

$$\left(1 - \frac{4ac}{b^2}\right)^{\frac{1}{2}}$$

$$= 1 - \frac{2ac}{b^2} - \frac{2a^2c^2}{b^4} - \frac{4a^3c^3}{b^6} - \frac{10a^4c^4}{b^8} - \text{etc., etc., } ad\ inf.,$$

so that the solutions are

$$x = -\frac{b}{2a} \pm \frac{b}{2a} \mp \frac{c}{b} \mp \frac{ac^2}{b^3} \mp \frac{2a^2c^3}{b^5} \mp \frac{5a^3c^4}{b^7} \text{ etc., etc., } ad\ inf.$$

In terms of actual numerical values this series solution may not be any simpler for computation than the original form of solution, but since the successive terms will decrease in magnitude more or less rapidly according to the magnitude of $4ac/b^2$, this form of statement facilitates a process of approximation, and if $4ac/b^2$ is very small compared with 1, so that powers above the second can be neglected, it gives a very closely approximate solution in a very simple form.

Again, the binominal theorem is very useful for the approximate calculation of certain numerical expressions. The nth root of a number, for instance, can be obtained by a general method which can best be explained by a simple illustration. Find to four places of decimals the fifth root of 3131. First find by guessing and trial the nearest whole number—in this case it is 5, for $5^5 = 3125$. Then

$$3131 = 5^5 + 6 = 5^5(1 + 6/3125).$$

Notice as a further simplification that

$$6/3125 = 6/5^5 = 6 \times 2^5/10^5 = 6 \times 32/10^5.$$

Then $$\sqrt[5]{3131} = 5[1 + (6 \times 32 \times 10^{-5})]^{\frac{1}{5}}$$

$$= 5\left\{1 + \frac{6 \times 32 \times 10^{-5}}{5} - \frac{4 \times 6^2 \times 32^2 \times 10^{-10}}{25 \times 2} + \text{etc., etc.}\right\}$$

$$= 5(1 + \cdot 00038)$$

taking the first two terms only (a little consideration will show that the third term will not affect the fourth place of decimals). Finally

$$\sqrt[5]{3131} = 5.0019.$$

Certain power calculations can be very considerably simplified in a similar manner, for example $(3.03)^{10}$. This can be put in the form

$$(3.03)^{10} = 3^{10}(1 + 10^{-2})^{10}$$

$$= 3^{10}[1 + (10 \times 10^{-2}) + (45 \times 10^{-4}) + (120 \times 10^{-6}) + (210 \times 10^{-8}) \text{ etc., etc.}]$$

$$= 3^{10} \times 1.10462$$

$$= 59049 \times 1.10462$$

correct to five figures. The last multiplication will be a rather long business, but the whole calculation will be very much shorter than the direct working out of the original expression.

68. The Exponential Series

Apart from these immediate and practical uses, the Binomial Theorem plays a large part in the development of other important series. This is illustrated in the Exponential Series, of which mention has been made already.

It has been shown that the series of numbers

$$1, x, \frac{x^2}{2!}, \frac{x^3}{3!}, \frac{x^4}{4!}, \frac{x^5}{5!} \quad . \quad . \quad . \quad \frac{x^{n-1}}{(n-1)!}$$

known as the Exponential Series, is convergent for all values of x. The series is really a special case of the Binomial Series, and can be derived in this way.

$$\left(1+\frac{1}{n}\right)^{nx}=1+\frac{nx}{n}+\frac{nx(nx-1)}{2!\,n^2}$$

$$+\frac{nx(nx-1)(nx-2)}{3!\,n^3}+\text{etc., etc.}$$

$$=1+x+\frac{x(x-1/n)}{2!}$$

$$+\frac{x(x-1/n)(x-2/n)}{3!}+\text{etc., etc.}$$

Now by sufficiently increasing n, the series on the right can be made to differ by as little as we please from the series

$$1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+\text{etc., etc., } ad\ inf.$$

In other words, this series is the limit of the original series when n tends to infinity, and we have

$$\underset{n\to\infty}{lt.}\left(1+\frac{1}{n}\right)^{nx}=1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+\frac{x^4}{4!}\ \cdots$$

$$+\frac{x^{n-1}}{(n-1)!}\ \cdots\ \text{etc., etc., } ad\ inf.$$

(This, by the way, though it is given in this form in some textbooks, cannot be regarded as a rigid proof. The limit of the sum of an infinite number of quantities is *not necessarily* the same as the sum of their limits, as the above proof assumes. However, it serves to demonstrate the connection between the Binomial and the Exponential Series. A completely rigid proof would take rather more space than can be allowed to it.)

An important special case of this series is that in which $x = 1$. The series then becomes

$$\underset{n\to\infty}{lt.}\left(1+\frac{1}{n}\right)^{n}=1+1+\frac{1}{2!}+\frac{1}{3!}+\frac{1}{4!}+\ \cdots$$

$$+\frac{1}{(n-1)!}\ \cdots\ \text{etc., etc., } ad\ inf.$$

The number represented by the series on the right-hand side plays a very large part in physics, particularly in electricity. It is called ϵ, and its magnitude to ten places of decimals is 2.7182818285. Readers have already made the acquaintance of this number in the section dealing with logarithms, where it was introduced as the base of the system of Naperian or Natural Logarithms. It was remarked at the time that nothing could seem more arbitrary and unnatural than this awkward looking number, but we see now that there is at least nothing arbitrary about it, and a little further consideration will show that there is very good reason for calling it " natural ". The reason is that it symbolises a process of growth or change which is of very frequent occurrence in natural phenomena.

Consider, for instance, what happens when a condenser of capacity C is given a charge of amount Q_0 and then allowed to discharge through a resistance R as shown in Fig. 16. The charge on the positive plate of the condenser will flow away in the form of a current through the resistance, and the magnitude of this current will depend on the potential of the condenser, that is, on the charge on the condenser. Thus the condenser discharges at a rate which is proportional to the charge, or, in other words, the charge disappears at a rate which is proportional to itself. This means that the rate of discharge will not be constant for any finite interval of time, but decreases continually as the charge leaks away. The determination of the charge left on the condenser after any given interval thus appears to be a very difficult matter. In fact, it cannot be solved directly without the aid of the Differential Calculus. The following method can be made to give the right answer, however, and is a very good example of the part played by ϵ in all such phenomena.

We will assume that the rate of discharge varies not continuously, but by sudden steps. That is, we will assume that the condenser discharges for an interval of time δt at the rate corresponding to the initial conditions, and then for a second interval δt it discharges at the rate corresponding to the

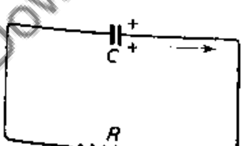


Fig. 16

conditions at the end of the first interval, and so on. The initial potential of the condenser is Q_0/C, and the initial current is therefore this potential divided by the resistance, that is, Q_0/CR. Since a current is the rate of flow of electricity, the quantity of electricity that leaves the condenser at this rate in the time δt is $Q_0\delta t/CR$, and the charge left on the condenser at the end of the first interval (call it Q_1) is $Q_0 - Q_0\delta t/CR$, that is,

$$Q_1 = Q_0 - Q_0\delta t/CR = Q_0(1 - \delta t/CR).$$

Similarly, Q_1, being the charge at the beginning of the second interval, the charge at the end of the second interval will be

$$Q_2 = Q_1(1 - \delta t/CR) = Q_0(1 - \delta t/CR)^2$$

and so on. After n such intervals the charge will be

$$Q_n = Q_0(1 - \delta t/CR)^n.$$

Assuming that we want to determine the charge after an interval t we can consider that this interval is divided up into n smaller intervals δt, that is, $t = n\delta t$, and writing Q_t for this remaining charge

$$Q_t = Q_0 (1 - \delta t/CR)^n = Q_0(1 - t/nCR)^n.$$

But this is admittedly an approximate solution. The rate of discharge does not remain constant during the interval δt, however short that interval may be ; but it is clear that the shorter the interval, the more correct the solution will be. That is, for a given interval t, the larger n becomes, the more nearly correct will be the solution. In fact the approximation can be made as close as we please by sufficiently increasing n, and the exact solution is therefore the limit of the above expression when n tends to infinity, that is,

$$Q_t = Q_0 \underset{n \to \infty}{lt.} (1 - t/nCR)^n.$$

This can also be written

$$Q_t = Q_0 \underset{n \to \infty}{lt.} \{(1 - t/nCR)^{-nCR/t}\}^{-t/CR},$$

and, writing $1/m$ for $- t/nCR$ (notice that m and n will tend to infinity together),

$$Q_t = Q_0 \underset{m \to \infty}{lt.} \{(1 + 1/m)^m\}^{-t/CR}$$
$$= Q_0 \epsilon^{-t/CR}.$$

The reader should put in some actual values in order to get some idea of the scale of the phenomenon. How long will it take the condenser to discharge completely, that is, for what value of t is $Q_0 \epsilon^{-t/CR}$ zero? The answer is that there is no finite value for t which makes Q_t zero. The condenser is *never* discharged. Think of it! All the condensers in the world are still trying to get rid of their last charge and not succeeding. Actually, of course, the charge falls to an immeasurably small quantity in a very short time, fractions of a second in general, and can be made smaller than any given quantity by making t sufficiently large. Mathematically speaking, complete discharge of the condenser is represented by

$$\underset{t \to \infty}{lt.} Q_t = \underset{t \to \infty}{lt.} Q_0 \epsilon^{-t/CR} = 0.$$

Nature abounds in instances similar to the above, where a quantity changes at a rate which is proportional to the magnitude of the quantity. In all such cases ϵ, generally with a negative index, will appear in the mathematical representation of the process. In fact ϵ turns up nearly as often as π, which is saying a good deal.

69. Elementary Algebra: Conclusions

So much by way of an introduction to Algebra (for even at the risk of discouraging the reader it is well to remind him that it is only a bucketful out of the ocean).

One final word of advice will not be out of place before we leave this part of the subject and proceed on the next stage of the journey towards the calculus. Mathematics, like mankind, is a mystic duality of body and soul. Algebra is the soul of arithmetic. Its intimate association with, and, in a certain sense, dependence upon, concrete reality should never be forgotten. A comprehensible numerical or physical interpretation is the ultimate sanction of any algebraical operation, and only within the limits of this sanction can the wonderful labour and thought-saving devices of algebraic symbolism be employed with perfect confidence. Even practised mathematicians are liable to be pulled up short by the sudden materialisation of a grinning absurdity out of a mist of ill-defined symbols, as when, for instance, to quote an example that once came to the writer's notice, a few pages of apparently unimpeachable analysis led to the conclusion that the height of the Heaviside layer could be expressed as a complex number. As the Duchess would have said to Alice, the moral of that is—take care of your grounds, and the sense will take care of itself.

Examples X

1. Expand to the first five terms

   (*a*) $1/(1 + x)$
   (*b*) $1/(a^5 - x^5)^{\frac{1}{5}}$
   (*c*) $(1 - x)^{-3}$.

2. Show that the *n*th terms of $(1 - x)^{-n}$ and $(1 + x)^{2n-2}$ are equal.

3. Find $\sqrt[6]{719}$ to four places of decimals by the Binomial Theorem. (Note $3^6 = 729$.)

4. Find $\sqrt[7]{108}$ to four places of decimals by the Binomial Theorem. (Note $2^7 = 128$.)

5. Show that

$$a^x = 1 + x \log_e a + \frac{x^2(\log_e a)^2}{2\,!} + \frac{x^3(\log_e a)^3}{3\,!} + \text{etc., etc., } ad\ inf.$$

6. Show that

$$\frac{e^x + e^{-x}}{2} = 1 + \frac{x^2}{2\,!} + \frac{x^4}{4\,!} + \frac{x^6}{6!} + \frac{x^8}{8\,!} \text{ etc., etc., } ad\ inf.$$

$$\frac{e^x - e^{-x}}{2} = x + \frac{x^3}{3\,!} + \frac{x^5}{5!} + \frac{x^7}{7\,!} + \frac{x^9}{9!} \text{ etc., etc., } ad\ inf.$$

7. Show that

$$\frac{1}{e} = \frac{2}{3\,!} + \frac{4}{5!} + \frac{6}{7!} + \frac{8}{9!} + \text{etc., etc., } ad\ inf.$$

Chapter 5

GEOMETRY AND TRIGONOMETRY

70. The Straight Line

GEOMETRY is the study of space and spatial relationships. As usually taught in schools (or rather to avoid possible injustice, as it was taught to the writer*) nothing could seem more academic or remote from real life; but in fact no other branch of mathematics plays a larger part in a wider range of human activities. One has only to think of surveying, architecture, navigation, astronomy and so on to realise this. The coils and condensers and aerials of radio are themselves solid configurations in space and their functioning is in many cases determined by their "geometry". Moreover, apart from these direct and obvious embodiments of geometrical elements, there is so close a connection between the ideas of pure algebra and those of pure geometry that some knowledge of the latter is an essential part of the equipment of anyone who wishes to apply these precise and beautiful methods of thinking to practical problems.

For the sake of brevity and compactness, the subject will be developed in terms of the "vector" ideas which have proved of such immense value in their application to the analysis of radio and electrical problems.

Geometry has the whole of space for its domain. Indeed, as an abstract subject it is not even limited to space of three dimensions; but two dimensions will suffice for the scope of this book. We shall consider plane relationships only.

The simplest geometrical element is a point, and the next simplest a straight line. But what is a straight line? Euclid defined it, over two thousand years ago, as the shortest distance between two points. This has always seemed to the writer to be more like a deduction than a definition, and it certainly is not easy to use as the founda-

* The reviser's experience was much happier.

tion of a logical system ; but it is also not easy to find a better one. However, let us think of the practical definitions which are used by many persons every day. There are several of them. When a surveyor measures an angle with a theodolite, he is virtually defining a straight line as the path of a ray of light. A builder assumes that a plumb-line is a straight line. To a draughtsman, a straight line is a line drawn against a straight-edge—and in this combination we can find a good working definition, for the line can be used to test the straightness of the straight-edge and therefore of itself. Firstly the edge can be slid along the line in both directions. If the two can be made to fit for any amount of sliding, the edge is either straight or has uniform curvature. But if it can also be made to fit in the same way when the straight-edge is turned over so that its under surface is now uppermost, then there cannot be any curvature, uniform or otherwise, and the edge and line are both straight.

This appears to be a rather crude definition, and the writer felt very nervous about putting it forward ; but he was reassured by finding that a real mathematician, Henri Poincaré, writes in his "Science and Method" that the shortest path definition " does not satisfy me at all ", and then goes on to give, at least for purposes of education, what is virtually the same as the above definition, adding that it is in effect defining a straight line as an axis of rotation.

This may seem to be making a great and unnecessary fuss about a simple idea that anyone can understand without even thinking about it, but the plain fact is that a straight line is not by any means a simple idea, as " number " is, and it is as well to realise this from the start. This is particularly necessary for those whose desire for understanding may carry them to the region of " relativity ", or even beyond.

71. Angles

A single geometrical straight line, defined as above, but assumed to be unlimited in extent, has no describable characteristics except straightness and extension ; but two such lines which intersect, that is, have one point in common, present much more food for thought. To start with, they define a plane, and in all that follows this plane will be represented by the plane of the paper. Further, each has direction relative to the other.

This relative direction is called the angle between the lines, and is conveniently described by means of letters in either of the ways illustrated in Fig. 17. The angle marked with a star will be called the angle AOB (written $A\hat{O}B$) or θ. Where a single symbol is used (generally a Greek letter) it must be inscribed in the angle as shown.

It is important to realise at the outset that the letter notation of geometry has a dual character. In the first place it is used simply to identify certain elements for reference. In the second, it functions as an algebraic symbol, standing for a number which measures in some specified manner the magnitude of the element. Thus the θ of Fig. 17 serves both to identify the angle AOB and to specify its amount, or magnitude. How this amount is to be measured will be considered later.

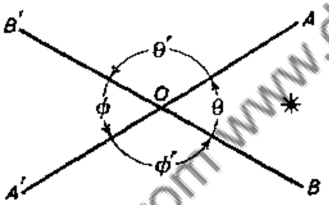


Fig. 17

Apart from any system of measurement, an angle can be thought of as an amount of turning. Thus the line OB has to be turned through an angle θ about O as pivot in order to give it the same direction as OA. The two alternative directions of rotation (anti-clockwise and clockwise) suggest at once a sign distinction. It is almost universally agreed to consider an anti-clockwise rotation as describing a positive angle (as in Fig. 17) and a clockwise rotation as describing a negative angle. This conforms with the algebraic significance of the negative sign, for

$$\theta + (-\theta) = 0, \text{ algebraically and geometrically.}$$

It is obvious that the amount of turning represented by $\theta + \theta'$, that is, the amount of turning that brings OB into line with OB', BB' being a straight line, is a constant for all such pairs of lines. The angles θ and θ' are called "complementary", each being called the complement of the other. It follows that

$$\theta + \theta' = \theta' + \phi,$$

so that

$$\theta = \phi.$$

The angles θ and ϕ, θ' and ϕ' are called opposite angles.

If the line OA is rotated about O in a positive direction, θ will increase and θ' decrease, and since the sum of the two is constant a condition will be reached when

$$\theta = \theta' \quad \text{(Fig. 18).}$$

The line OA is then said to be perpendicular to BB' and the angle θ (or θ') is called a "right angle".

The right angle is the basis of one system of measuring angles. It is divided into 90 equal rotations, each of which is called a degree. A degree is further sub-divided into 60 minutes and a minute into 60 seconds. In this system a complete rotation is four right angles or 360 degrees, and a half rotation 180 degrees. Thus complementary angles are such that

$$\theta + \theta' = 180^\circ,$$

and in practice the definition is extended so as to include the case in which either angle is greater than 180°, the other being correspondingly negative.

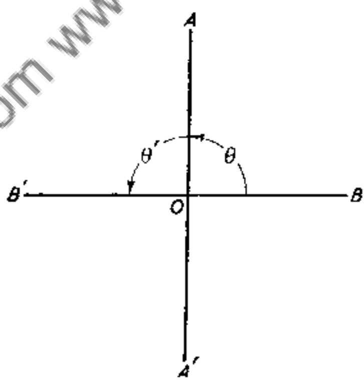


Fig. 18

At this point it may well be asked "Why in the name of all that is arbitrary and inconvenient, pick on 360 as the number of degrees in a complete revolution?" Hogben says that the Babylonians are probably to blame for this. They thought that the solar year was 360 days. Later on another system for the measurement of angles will be described. This alternative system, though numerically not much more convenient than the degree system, has a more logical basis. It cannot be fully explained, however, until we have some understanding of the great principle of geometrical similarity on which it is based.

72. The Triangle

Suppose now that a third line is introduced into our infinite plane (see Fig. 19). Each of the three lines will then acquire the additional characteristic of relative position as well as relative direction. This is not immediately

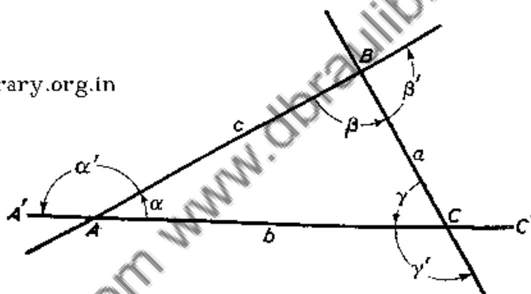


Fig. 19

relevant, but is introduced to emphasise the essentially relative character of position, and the fact that two other lines are required to specify it.

The figure bounded by these three coplanar straight lines is called a plane triangle. The sides can be described by means of the letter pairs *AB*, *BC*, *CA*, or more conveniently by the single letters shown in the figure.

Notice first that if the line *A'C'* is turned about *A* through the angle α in the direction shown, then about *B* through the angle β and then about *C* through the angle γ, it will make a half revolution, the points *A'* and *C'* reversing their position with respect to *B*. Therefore

$$\alpha + \beta + \gamma = 180°.$$

In words, the sum of the internal angles of a triangle is two right angles. Further, since

$$\gamma' + \gamma = 180°,$$
$$\gamma' = \alpha + \beta$$

and similarly for the other pairs of internal angles. The angles α', β' and γ' are called the external angles of the triangle.

73. Parallels

Suppose the line BC (Fig. 19) is rotated in a positive direction about B. The point C will move along $A'C'$, as shown in Fig. 20.

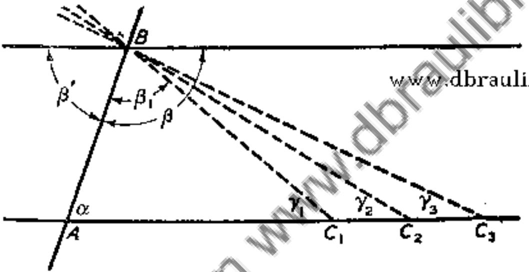


Fig. 20

Then

$$\alpha + \beta_1 + \gamma_1 = 180°$$
$$\alpha + \beta_2 + \gamma_2 = 180°$$
$$\alpha + \beta_3 + \gamma_3 = 180°$$

and so on.

But

$$\gamma_1 = \gamma_2 + C_1\hat{B}C_2.$$

Therefore

$$\gamma_1 > \gamma_2.$$

Similarly

$$\gamma_2 > \gamma_3$$

and so. Thus, as C moves away from A, the corresponding γ gets smaller and smaller and smaller. In fact, by turning the line so that C continues to move away from A,

γ can be made to approximate to 0 *within every standard* (see Section 61). Thus, in the limit

$$\alpha + \beta = 180°,$$

and the lines AC and BC are said to be parallel. This is what is meant by saying that parallel lines meet at infinity.

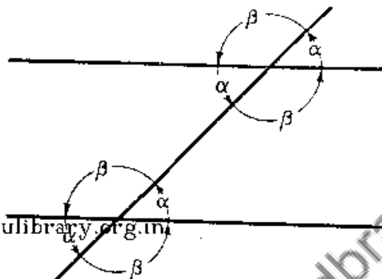


Fig. 21

Further, since

$$\alpha + \beta = 180°$$

and

$$\beta' + \beta = 180°$$

then

$$\beta' = \alpha.$$

These are obviously general relationships, true for any pair of parallel lines and a third cutting through them both. The angular equalities are as indicated in Fig. 21.

74. Congruence of Triangles

By the complete congruence of geometrical figures is meant the congruence in magnitude of all the corresponding elements, sides, angles, etc. A practical criterion of congruence is that the one figure, placed on top of the other, will coincide with it at every point.

The two triangles ABC and DEF of Fig. 22 are equal in every respect if

$$AB = DE, \qquad A\hat{B}C = D\hat{E}F,$$
$$BC = EF, \qquad B\hat{C}A = E\hat{F}D,$$
$$CA = FD, \text{ and } C\hat{A}B = F\hat{D}E,$$

but obviously not all these conditions are necessary. For instance, if two pairs of angles are given equal, the remaining pair must also be equal (see Section 72). What are the minimum conditions that will ensure complete equality? There are three separate combinations of such conditions.

(*a*) Two sides and the included angle, that is,

$$AB = DE,$$
$$AC = EF,$$
$$A\hat{B}C = D\hat{E}F.$$

(*b*) One side and two angles not opposite to this side, that is,

$$AC = DF,$$
$$C\hat{A}B = F\hat{D}E,$$
$$B\hat{C}A = E\hat{F}D.$$

and therefore also $\mathrm{A\hat{B}C = D\hat{E}F}$.

(*c*) Three sides, that is,

$$AB = DE,$$
$$BC = EF,$$
$$CA = FD.$$

The first two are very easily proved by considering the second triangle to be, as it were, lifted and placed down on the first. The proofs are too simple to justify giving space to them. The third cannot be proved in this way, and will be deferred till later.

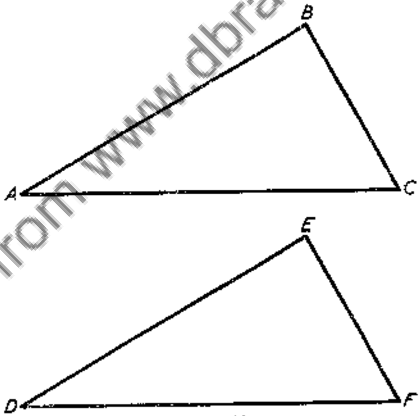


Fig. 22

75. Geometrical Similarity

The above discussion of equality brings us to one of the most practically useful and important ideas in the whole of geometry, the idea of geometrical similarity. It is probably not generally realised that the whole science of trigonometry, with its applications to surveying, navigation, and astronomy, is based on this one principle. The idea is concerned with the relation between equiangular triangles, that is, triangles of which the corresponding angles are equal. Such triangles are not necessarily equal in every respect, as Fig. 23 will show. There is, however, a definite metrical relation between such triangles, the relation being

$$\frac{a}{d} = \frac{b}{e} = \frac{c}{f}.$$

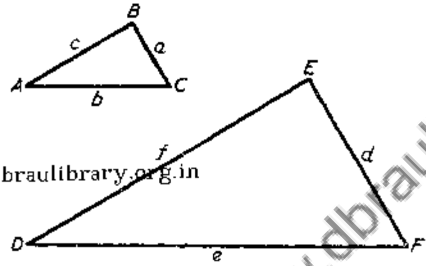


Fig. 23

The proof is not easy, but an outline of it will be given on account of its importance in all that follows.

Consider first the diagram of Fig. 24, in which AB is parallel to CD and AC to BD. It follows from the properties of parallels that the angles marked with the same Greek letter are equal. The triangles ABC and DBC, having the side BC in common and these angles equal, are equal in every respect (see Section 74). Therefore

$$AB = CD \text{ and } AC = BD.$$

This is the first step. Now look at Fig. 25 where B is the middle point of AC, BD parallel to CE, and DF parallel to AC. By applying the result just proved it is easy to show that $AD = DE$. This result can be extended as shown in Fig. 26. If the line AB is divided into n equal parts by the parallels to BC, then AC will also be divided into n equal parts or "segments". Returning now to Fig.

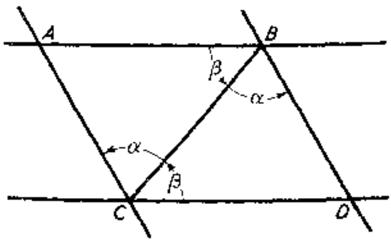


Fig. 24

23, the smaller of the two triangles can be drawn inside the other, as shown in Fig. 27, the line BC being parallel to EF, in virtue of the angular equalities assumed. If the ratio of AC to AF is expressed in its lowest terms as the fraction m/n, AF can be divided into n equal segments by lines parallel to, and including, BC. The line AC will be divided into n equal segments by n of these lines. It follows from the preceding that AB and AE will be similarly divided by these parallels, so that

$$\frac{AB}{AE} = \frac{m}{n} = \frac{AC}{AF}.$$

Similarly, by redrawing the smaller triangle so that B coincides with E, it could be shown that

$$\frac{AB}{AE} = \frac{BC}{EF} = \frac{CA}{FA} = \frac{m}{n},$$

that is,

$$\frac{a}{d} = \frac{b}{e} = \frac{c}{f}.$$

It follows from the properties of fractions already proved that

$$\frac{a}{b} = \frac{d}{e} \quad \text{and} \quad \frac{b}{c} = \frac{e}{f},$$

both of which equalities are included in the form

$$a : b : c = d : e : f.$$

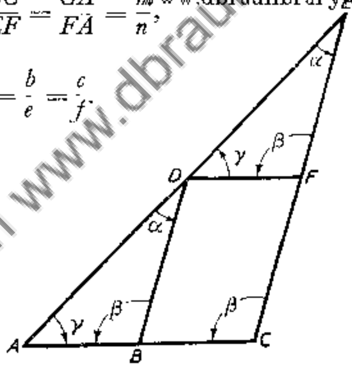


Fig. 25

This means that the *relative* magnitudes of the sides of a triangle depend only on the shape of the triangle, and will be the same for all triangles having two angles of the one equal to the two corresponding angles of the

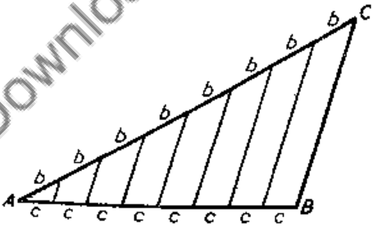


Fig. 26

other. That is the important conclusion to remember. Two triangles related in this way are said to be similar. The trigonometrical ratios which will now be described are no more than an expression of this similarity in the particular case of right-angled triangles.

76. The Trigonometrical Ratios

Consider the right-angled triangle shown in Fig. 28. Any other right-angled triangle containing the angle of θ will be of the same shape as this. Therefore the relative magnitude of the sides, that is, a/b, c/b, and a/c, and the reciprocals of these ratios or numbers, will depend only on θ. They can therefore be considered as functions of θ and tabulated for various values of θ. Special names, tabulated below, are given to these numbers.

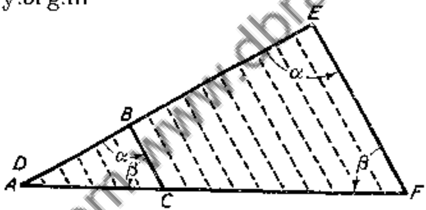


Fig. 27

Ratio	*Name*	*Abbreviations*
height/hypotenuse	sine	sin
base/hypotenuse	cosine	cos
height/base	tangent	tan
hypotenuse/height	cosecant	cosec
hypotenuse/base	secant	sec
base/height	cotangent	cot

Apart from the reciprocal relationship, it is clear that these ratios, being functions of θ, must also be functions of one another, and will therefore possess certain interrelationships. One is immediately obvious from their definition, for

$$\frac{\sin}{\cos} = \frac{\text{height}}{\text{hypotenuse}} \Big/ \frac{\text{base}}{\text{hypotenuse}} = \frac{\text{height}}{\text{base}} = \tan.$$

Another important relationship will be proved later.

In practice the above definitions of the trigonometrical ratios are extended, as shown in Fig. 29, to cover any value of θ up to four right angles. The positive angle θ is the amount of anti-clockwise rotation of AC from AX, and $\sin\theta$ is BC/AC, etc., etc. Similarly for the remaining quadrants.

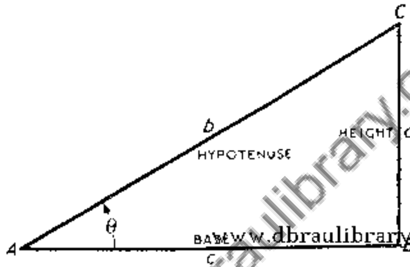


Fig. 28

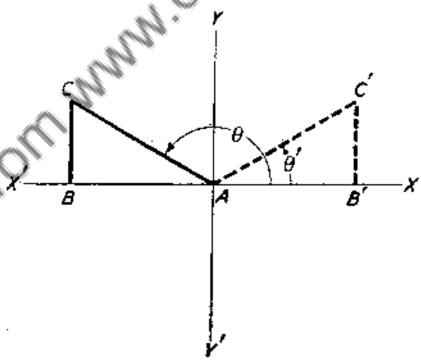


Fig. 29

Without further qualifications, however, an angle such as θ', where $AB = AB'$ and $BC = B'C'$ in magnitude, would have the same ratios as θ. The ambiguity is avoided by a sign convention relating to the constituent lines of the ratios. The sloping line is considered to have no sign at all but all other lines are considered to be measured away from the centre A and are taken as positive in the directions AX and AY and negative in the directions AX' and AY'. Thus (Fig. 30), if the lines have the magnitudes shown by the small letters, BC and ED are interpreted as the number a, BG and EF as the number $-a$, and so on.

This convention results in the following signs for the ratios in the various quadrants :—

sin cos tan	+ − −	+ + +
sin cos tan	− − +	− + −

The table need not be memorised, as it is easier to apply the sign convention direct in any given case.

Remembering that a negative angle is a rotation in a clockwise direction, it is easy to see that

$$\sin\,\theta = -\sin\,(-\theta),$$
$$\cos\,\theta = \cos\,(-\theta),$$

and therefore

$$\tan\,\theta = -\tan\,(-\theta).$$

Further, there will be certain simple relationships between the ratios of angles that differ by positive or negative multiples of a right angle. For instance,

$$\sin\,\theta = \cos\,(90^\circ - \theta),$$
$$\cos\,\theta = \sin\,(90^\circ - \theta),$$
$$\sin\,\theta = -\cos\,(90^\circ + \theta), \text{ etc., etc.}$$

These, again, need not be memorised, as it is much easier to draw the appropriate lines in a quadrant diagram in any given case.

Given that

$$\sin\,\theta = n,$$

θ may be described as the angle having the sine n. This is written, conventionally

$$\theta = \sin^{-1} n.$$

It should be noticed that whereas $\sin\,\theta$ is a single-valued function of θ, that is, given θ there is only one value for its sine, $\sin^{-1} n$ on the other hand is many valued, has in fact an infinity of values, for if θ be the smallest angle having the given sine, $\theta + (r \times 360°)$, r being any integer, will all have the same sine. So will $(180° - \theta) + r \times 360°$. Similarly for the other "inverse" functions.

There is no need to illustrate the sine and cosine of an angle graphically, for these curves will already be familiar to all students of electricity in connection with wave forms of alternating currents. Before leaving this part of the subject, however, some important special cases should be noted.

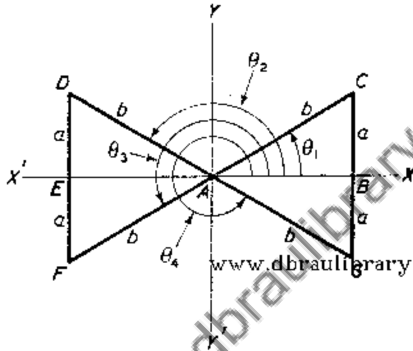


Fig. 30

Referring to Fig. 28,

$$\sin\theta = a/b, \quad \cos\theta = c/b.$$

Now as θ tends to zero, b tends to equality with c, and a tends to zero, so that

$$\sin 0 = 0, \quad \cos 0 = 1, \quad \tan 0 = 0.$$

Similarly

$$\sin 90° = 1, \quad \cos 90° = 0, \quad \tan 90° = \infty,$$
$$\sin 180° = 0, \quad \cos 180° = -1, \quad \tan 180° = 0,$$

etc., etc.

77. THE CIRCULAR MEASURE OF ANGLES

An isosceles triangle is one having two sides equal. It is easy to show from the results of Section 74 that the angles opposite these sides will be equal, and therefore that all such triangles with a given angle included between the equal sides will be similar. n such triangles of equal size and with vertical angle $360°/n$ can obviously be combined into a figure such as that shown in Fig. 31. Such a figure is called a regular n-sided polygon (that illustrated is an octagon). From the preceding discussion of similarity it follows that for any given number of sides n the ratio b/s, and therefore nb/s is constant, that is, independent of the size of the figure, that is,

$$nb/s = k_n,$$

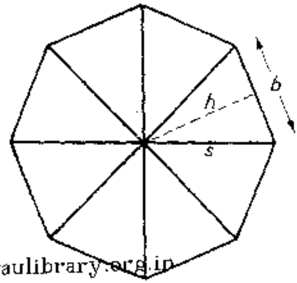


Fig. 31

where k_n depends only on n. Now by sufficiently increasing n the figure can be made to differ by as little as we please from a circle. A circle is in fact the limiting case when n is made infinite, nb then becomes the periphery or circumference of the circle, and s its radius. Thus the ratio

$$\text{circumference/radius} = k_\infty$$

is constant for all circles. The constant is the number 6·2831 . . . . . , usually written 2π. The symbol 2π is used both for shortness and because π is what is called an incommensurable number, that is, it cannot be completely represented by any decimal. It follows that the ratio of any given fraction of the circumference to the radius is also constant for all circles. Thus in Fig. 32,

$$\frac{\text{length of arc } a_1}{r_1} = \frac{\text{length of arc } a_2}{r_2}$$
$$= \text{constant for all circles,}$$

the magnitude of the constant depending only on the angle θ. The ratio arc/rad. is therefore a natural measure of the angle and is in fact called the circular measure of the angle. Unit angle on this basis will be that for which

$$\text{arc/rad} = 1$$

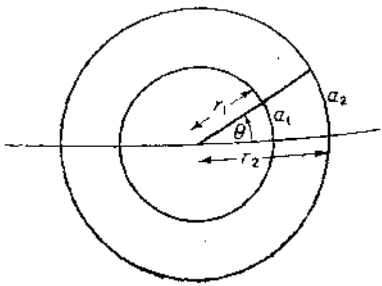


Fig. 32

that is, the angle subtended by an arc formed by bending the radius round the circumference. This angle is called one radian, and all angles in this system are expressed in terms of radians. Thus an angle 1.3, that is, 1.3 radians, is one subtended by an arc 1.3 times the radius in length.

The relation between the two sets of units is obvious, for half the circumference subtends 180°, and the ratio of this arc to the radius is as shown above, π, or 3·14159 . . . . . ., so that

$$180^\circ = \pi \text{ radians},$$

which gives 57° 17′ 44·8″ as 1 radian. This is a cumbersome sort of relationship, but conversion is very rarely called for so it does not really matter. In general terms we have

$$\theta \text{ radians} = \left(\frac{180}{\pi}\,\theta\right)^\circ.$$

A right angle is clearly $\pi/2$ radians, and the angles of a right-angled triangle can therefore be expressed in circular measure as $\pi/2$, θ, and $(\pi/2 - \theta)$. Supplementary angles (Section 71) are defined by

$$\theta + \theta' = \pi,$$

and in practice the definition is extended to the case in which either of these angles is numerically greater than π, the other being correspondingly negative. In the familiar group of symbols $\sin 2\pi ft$ or $\sin \omega t$, where $\omega = 2\pi f$ (f being the frequency in cycles per second), ωt is an angle which increases at the rate ω, or $2\pi f$, radians per second.

78. Area

Area, or amount of surface, is a fundamental conception or thing of its own kind, which cannot be described in terms of anything else, as one soon discovers by trying to do so. Like all the fundamental physical quantities, its magnitude can only be expressed in terms of itself, that is, in terms of its own unit. Thus an area of ten units means an area having ten times as much surface as some area which it has been agreed to call a unit area. Humanity has always been vitally concerned with area and a multiplicity of practical units has arisen in consequence, but they all have this feature in common—they are expressed in terms of the amount of surface of a square having a side of specified length, and for this reason they nearly all bear names, such as square mile, square centimetre, etc., which indicate the length of the side of the square. This choice of the unit shape is quite arbitrary—a circle of specified radius would serve the same purpose, but the accepted shape has the advantage that the area of a rectangle of sides a and b units of length is arrived at by the simplest possible calculation on this system. It is in fact ab units of area, as can easily be demonstrated by a little simple drawing, the unit of area being the square unit of length, whatever that may be. This, however, must not be taken to mean that "area is length multiplied by length" a statement which is completely unintelligible except as a conveniently abbreviated expression of the ideas which have just been described (compare this with the discussion in Section 18 on the physical aspect of multiplication).

(*a*) *Area of a triangle*

Referring to Fig. 33, the area of the triangle ABC is $\frac{1}{2}ah$ where a is the length of the side BC. This can be demonstrated by completing the rectangles $EBDA$, $AFCD$. It can be shown, as in Section 74, that the lines AB, AC divide these rectangles into congruent triangles, whence the above result follows.

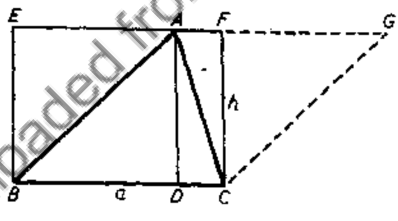


Fig. 33

(*b*) *Area of a parallelogram*

In a similar manner it can be shown that the area of the parallelogram $ABCG$ is ah, for the diagonal AC divides it into two equal triangles.

(c) *Area of a circle*

Referring to Fig. 31, the area of each of the isosceles triangles into which the polygon is divided is $\frac{1}{2}bh$, so that the area of the whole polygon is $\frac{1}{2}nbh$. If the number of sides is increased indefinitely the figure becomes a circle, nb becoming the circumference and h the radius. The area of the circle is therefore half the product of the length of the circumference and that of the radius. As already shown, the circumference is 2π times the radius, so that the area of the circle is $\frac{1}{2}(2\pi r)r$, that is, πr^2.

Examples XI

1. Two angles of a triangle are $17°$ and $83°$. What is the third?

2. An irregular pentagon is a five sided figure with unequal sides. Prove that the sum of the internal angles is $540°$. What would it be for an n-sided figure?

3. The sides a, b and c of the triangle in Fig. 23 are 4, 5 and 9 inches long respectively. The side f of the triangle DEF is 10 inches. How long are e and d?

4. Show that $\operatorname{cosec}\theta \tan\theta = \sec\theta$,
$$\sec\theta \cot\theta = \operatorname{cosec}\theta.$$

5. Given that $\sin 50° = 0{\cdot}766$,
$\cos 50° = 0{\cdot}643$,

find the values of
$\tan 50°$,
$\sec 50°$,
$\operatorname{cosec} 50°$,
$\cot 50°$,
$\sin 130°$,
$\sin 140°$,
$\cos 220°$,
$\tan 320°$,
$\tan -50°$,
$\sin 40°$,
$\sec -40°$.

6. Find the value of $(\sin 50°)^2 + (\cos 50°)^2$. (Note: This is usually written $\sin^2 50° + \cos^2 50°$.)

7. Taking $22/7$ as a sufficiently close approximation for π, find in degrees and minutes to the nearest minute
$$1\tfrac{1}{21} \text{ radians},$$
$$\tfrac{1}{7} \text{ radians}.$$
Find the magnitude in radians of $10°$, $1260°$.

8. Find the area of (i) a regular octagon with sides

10 inches long, (ii) a sector of a circle, radius 5 cm, length of arc 10 cm.

9. If the unit of area were defined as the area of a circle of unit radius, what would be the areas of the figures in question 8?

79. Vectors

The idea "vector" is still hedged about with that vague apprehension and dislike that attaches to the unfamiliar, but one cannot get far in electrical theory nowadays without it. The trouble is that although the essential idea itself is simple enough, its application calls for new habits of thought, a new mental technique. That means hard thinking, and hard thinking is the hardest of hard work. The technique, however, is worth all it costs in that way, and there can be no doubt that the vector notation will play a large and increasing part in the mathematics and mathematical physics of the future. Even to-day it is unusual to find any analysis of alternating current circuit problems which is not expressed in terms of the symbol "j", and that symbol, as the following sections will show, is very intimately connected with the vector idea, though it is admittedly open to more than one interpretation (and a few misinterpretations).

It was pointed out in Section 71 that a single straight line in an infinite plane cannot in any useful sense of the word be said to possess direction, for direction is essentially a relation to some other straight line. We will therefore take as the domain of our present thinking an infinite plane and an infinite line in that plane. By the direction of any other line in the plane will then be meant its direction relative to this given line. For practical purposes, the infinite plane can be taken as that of the paper one is writing on, and the reference line can be any line parallel to the bottom edge of the paper.

On this understanding any line in the plane of the paper will have direction, and any finite segment of the line will have both magnitude and direction. The name "vector" is given to any line regarded in this way as a combination of magnitude and direction. In general, any physical quantity whatever which possesses both magnitude and direction is called a vector quantity. Thus velocity, force, acceleration, etc., are vector quantities, and as such are capable of representation by means of vectors (not necessarily co-planar in any given system). Most of the quantities with which physics, and more particularly electricity is concerned are of this character, whence the fundamental importance of the vector idea and of its technique. As distinct from a vector quantity, any quantity which has magnitude only is called a scalar quantity, or, shortly, a scalar. Density, temperature, energy, etc., are examples. The distinction is well marked in the case of weight and mass. The latter is a scalar, and the former, being the gravitational force associated with the mass, is a vector.

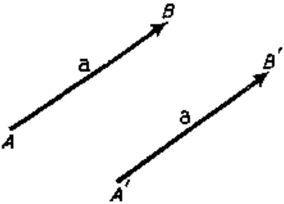


Fig. 34

For the present, however, we shall not be concerned with vector quantities in general, but simply with the co-planar line vectors defined as above. It should be noted that no mention is made of position in the definition. Position plays a secondary part in vector analysis, and where it does enter into any given problem it will arise as a consequence of the other two attributes or will be otherwise specified. In general any two lines such as AB, $A'B'$, in Fig. 34, which are equal in magnitude and direction (the latter being indicated by an arrow head as shown) are vectorially identical. Following a well-established typographical practice, a line of length a will be printed in bold face type ($\mathbf{a}$) if it is being considered as a vector. Alternatively, $\overrightarrow{AB}$ will be taken to mean the line AB considered as a vector. The magnitude of any given vector $\mathbf{a}$ will be indicated either by using the same letter in ordinary type, or by $|\mathbf{a}|$.

80. The Addition of Vectors

A vector can be regarded as a displacement or step of specified amount and direction. The obvious interpretation of the addition of two vectors is the combination of the two displacements as shown in Fig. 35, the *sum*, as distinct from the *process* of addition, being the single displacement which has the same total effect as the two

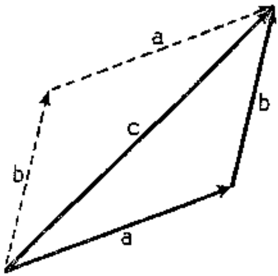


Fig. 35

displacements made in succession. Thus the sum of **a** and **b** is **c**. From the properties of a parallelogram it follows that **b** + **a**, shown by the dotted lines in Fig. 35, is the same as **a** + **b**, so that the process of addition of vectors obeys the commutative law, which brings it into line with the same process in ordinary algebra. The extension of the above, and of the commutative law, to the addition of any number of vectors is obvious.

81. Vector Interpretation of the Negative Sign

As far as possible the symbolism of vector algebra will be made analogous to that of scalar algebra. Since therefore

$$a + (-a) = 0$$

in scalar algebra, let us use the same form vectorially and interpret $(-\mathbf{a})$, at present undefined, to suit this condition. If

$$\mathbf{a} + (-\mathbf{a}) = 0,$$

then $-\mathbf{a}$ is a displacement which cancels the displacement **a**, that is, $-\mathbf{a}$ is a vector of the same magnitude as **a** but opposite to it in direction. The subtraction of vectors then becomes an operation of essentially the same character as addition, for

$$\mathbf{a} - \mathbf{b} = \mathbf{a} + (-\mathbf{b}).$$

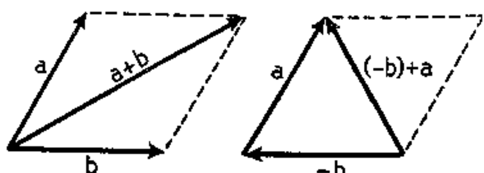


Fig. 36

The process is illustrated in Fig. 36. Notice that $\mathbf{a} + \mathbf{b}$ and $\mathbf{a} - \mathbf{b}$ are the diagonals of a parallelogram having $\mathbf{a}$ and $\mathbf{b}$ as sides.

82. The Multiplication of a Vector by a Number

By analogy with the corresponding scalar operation

$$\mathbf{a} + \mathbf{a} + \mathbf{a} + \mathbf{a} + \text{etc.} \; . \; . \; . \; . \; . \; . \; . \; n \text{ terms}$$

will be written $\mathbf{a} \times n$, or more conveniently $n\mathbf{a}$. Thus $n\mathbf{a}$ is a vector of the same direction as $\mathbf{a}$ but n times as long.

83. The Idea of Operator and Operand

The group $n\mathbf{a}$ can be regarded as symbolising a definite operation on the vector $\mathbf{a}$, which in this relation can be called the "operand". If n is a positive number or fraction, the operation consists of the multiplication of the magnitude of $\mathbf{a}$ by n without changing its direction. If n is a negative number or fraction, say, $-m$ where m is positive, then $n\mathbf{a}$ represents the somewhat more elaborate operation of multiplying the magnitude of $\mathbf{a}$ by m and reversing its direction. Regarded in this way, the symbol n is called an "operator". It is an essential feature of an operator that its effect shall be independent of its operand, that is, the relation of $n\mathbf{a}$ to $\mathbf{a}$ does not depend in any way on $\mathbf{a}$. This idea is more than a mere pedantic elaboration of terminology. Other forms of operator, which play a very large part in alternating current theory, will be introduced in later sections.

84. Unit Vectors

A unit vector is a vector of unit length. It follows from Sections 82 and 83 above that any vector whatever can be expressed in terms of a positive or negative number and some unit vector. Thus the vector $\mathbf{a}$ can be expressed in the form $a\mathbf{a}_1$, where $\mathbf{a}_1$ is the unit vector in the direction of $\mathbf{a}$. Thus $m\mathbf{a}_1$ and $n\mathbf{a}_1$ are vectors of magnitudes m and n and of the same or opposite direction according as m and n are of the same or opposite sign. Notice that

$$m\mathbf{a}_1 + n\mathbf{a}_1 = (m + n)\mathbf{a}_1.$$

In all that follows the symbol ν will be used for the unit vector in the direction of the bottom edge of the page from left to right.

85. The Scalar Product of Vectors

In the arithmetical sense of the term, the multiplication of two vectors is not an intelligible process at all, but there is a quantity which involves two vectors in a manner similar to multiplication, and to this the name scalar product is given. It is defined in this way. The scalar product of two vectors $\mathbf{a}$ and $\mathbf{b}$ is the scalar quantity $ab \cos \theta$, a and b being the magnitudes of the vectors and θ the angle between their positive directions. It is written $\mathbf{a} \cdot \mathbf{b}$, so we have

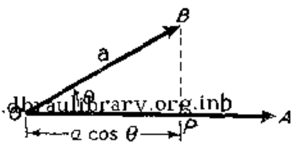


Fig. 37

$$\mathbf{a} \cdot \mathbf{b} = ab \cos \theta.$$

The character of this product will be made clear by an inspection of Fig. 37. If BP is drawn perpendicular to OA, then

$$\frac{OP}{OB} = \cos \theta,$$

that is,
$$OP = OB \cos \theta = a \cos \theta,$$

so that
$$\mathbf{a} \cdot \mathbf{b} = ab \cos \theta = OA \cdot OP.$$

The scalar product is thus the product of the magnitude of one vector and of the "projection" of the other on it.

Notice that a scalar product has sign as well as magnitude. This is because $\cos \theta$ has sign. Thus the scalar product of the vectors $\mathbf{a}$ and $\mathbf{b}$ shown in Fig. 38 will be a negative quantity, for in this case θ is in the second quadrant and its cosine is therefore negative.

Notice further that there is no question of extending this notation to more than two vectors. $\mathbf{a} \cdot \mathbf{b}$ is a scalar quantity, so there is no scalar product of $(\mathbf{a} \cdot \mathbf{b})$ and a third vector $\mathbf{c}$.

This conception of scalar product may at first sight

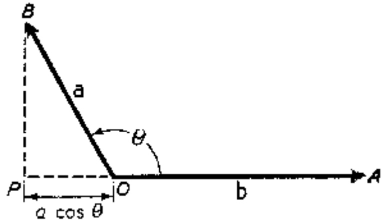


Fig. 38

seem rather arbitrary, but it has as a matter of fact a very definite physical significance. To take one of the many instances of its physical interpretation, if **a** represents the displacement of a body under the action of a force represented by the vector **b**, then **a** · **b** is the work done by the force on the body. Another application of a rather different character is of particular interest to wireless amateurs. Let **e** be a vector of magnitude $\hat{e}$, making with ν an angle θ. Further, suppose that θ is proportional to time, increasing at ω radians per second, its magnitude being ϕ when $t = 0$,

that is, $$\theta = \omega t + \phi.$$

Then $$\mathbf{e} \cdot \nu = \hat{e} \cos (\omega t + \phi).$$

Thus a sine wave of e.m.f. can be represented as the scalar product with a fixed unit vector of another vector of constant length rotating with constant angular velocity (ω).

It will now be shown that scalar multiplication obeys the same formal laws as ordinary multiplication. It follows from the definition that

$$\mathbf{a} \cdot \mathbf{b} = \mathbf{b} \cdot \mathbf{a},$$

so that a scalar product obeys the law of commutation.

Further it is easy to show that scalar multiplication of vectors obeys the distributive law, that is,

$$\mathbf{c} \cdot (\mathbf{a} + \mathbf{b}) = (\mathbf{a} + \mathbf{b}) \cdot \mathbf{c} = \mathbf{a} \cdot \mathbf{c} + \mathbf{b} \cdot \mathbf{c}.$$

The demonstration is illustrated in Fig. 39. We have

$$\begin{aligned}(\mathbf{a} + \mathbf{b}) \cdot \mathbf{c} &= OQ \times OC \\ &= (OP + PQ) \times OC \\ &= OP \times OC + PQ \times OC \\ &= \mathbf{a} \cdot \mathbf{c} + \mathbf{b} \cdot \mathbf{c}.\end{aligned}$$

It follows directly from this that

$$(\mathbf{a}+\mathbf{b})\cdot(\mathbf{c}+\mathbf{d})=\mathbf{a}\cdot\mathbf{c}+\mathbf{a}\cdot\mathbf{d}+\mathbf{b}\cdot\mathbf{c}+\mathbf{b}\cdot\mathbf{d},$$

and similarly for the scalar products of vectors expressed as the sums or differences of any number of component vectors. Operations of scalar multiplication of vectors can therefore be carried out in just the same way as the ordinary multiplication of similar number groups.

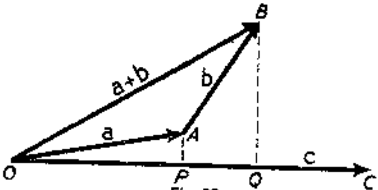


Fig. 39

86. Two Important Special Cases

(*i*)
$$\mathbf{a}\cdot\mathbf{a}=a\times a\times\cos 0=a^2.$$

This can be written

$$\mathbf{a}^2=a^2.$$

Thus the scalar square of a vector is the square of its magnitude.

(*ii*) If **a** and **b** are perpendicular to each other, then

$$\mathbf{a}\cdot\mathbf{b}=a\times b\times\cos\pi/2=ab\times 0=0.$$

The converse of this is also true with a reservation.

If
$$\mathbf{a}\cdot\mathbf{b}=ab\cos\theta=0,$$
then
$$a=0,$$
or
$$b=0,$$
or
$$\cos\theta=0,$$
that is,
$$\mathbf{a}=0,$$
or
$$\mathbf{b}=0,$$
or the vectors are mutually perpendicular.

This will prove to have an important application to alternating current theory in the following form. Suppose that i_1, i_2, i_3, etc., be a number of alternating currents which meet at a branch point of a network of conductors as shown in Fig. 40. Then by Kirchhoff's first law, the sum

$$i_1 + i_2 + i_3 + \text{etc.} = 0$$

at every instant. Now as shown in above, each of these currents can be represented in the form $\mathbf{i}_1 \cdot \nu$, $\mathbf{i}_2 \cdot \nu$, etc., where $\mathbf{i}_1$, $\mathbf{i}_2$ are vectors of constant magnitude rotating with constant angular velocity. Therefore

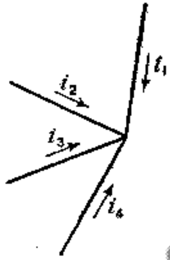


Fig. 40

$$\mathbf{i}_1 \cdot \nu + \mathbf{i}_2 \cdot \nu + \mathbf{i}_3 \cdot \nu + \text{etc.} = 0,$$

that is, $(\mathbf{i}_1 + \mathbf{i}_2 + \mathbf{i}_3 + \text{etc.}) \cdot \nu = 0.$

Therefore the vector $(\mathbf{i}_1 + \mathbf{i}_2 + \mathbf{i}_2 + \text{etc.})$ is zero or else is perpendicular to ν at every instant. The second condition cannot be fulfilled at every instant, for the vectors are assumed to be rotating with constant angular velocity. Therefore

$$(\mathbf{i}_1 + \mathbf{i}_2 + \mathbf{i}_3 + \text{etc.}) = 0,$$

so that Kirchhoff's law applies not only to the instantaneous values of the currents which meet at a branch point, but also to the rotating vectors which, as described more fully later on, are used to represent these currents.

87. An Important Geometrical Proposition

The vectors $\mathbf{c} = \mathbf{a} + \mathbf{b}$, $\mathbf{a}$, and $\mathbf{b}$ form a triangle as shown in Fig. 41. From Section 85,

$$\mathbf{c}^2 = (\mathbf{a} + \mathbf{b})^2 = \mathbf{a}^2 + \mathbf{b}^2 + 2\,\mathbf{a} \cdot \mathbf{b}.$$

Putting in the scalar magnitudes of these scalar products and squares,

$$c^2 = a^2 + b^2 + 2ab\cos\theta$$
$$= a^2 + b^2 - 2ab\cos\gamma,$$

since $\theta + \gamma = \pi$.

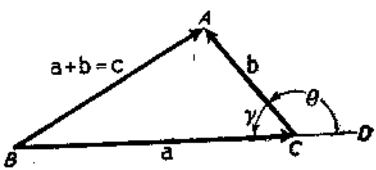


Fig. 41

88. Applications and Deductions

(*a*) Given two sides of a triangle (a and b) and the angle (γ) between them, the magnitude of the third side can be calculated from the formula

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

(*b*) Given the three sides of a triangle, a, b, and c, the angle γ between a and b is given by

$$\cos \gamma = (a^2 + b^2 - c^2)/2ab$$

which, together with the fact that γ must be less than 180°, defines γ completely.

It follows from this that if two triangles are equal as to the lengths of their sides they are equal in every respect (see Section 74).

(*c*) The special case known as Pythagoras Theorem.

If $A\hat{C}B$ is a right angle (Fig. 42) then

$$\cos \gamma = 0,$$

and

$$c^2 = a^2 + b^2.$$

Thus, in a right-angled triangle, the square on the hypotenuse is equal to the sum of the squares on the other two sides.

This leads to the functional relationship between the trigonometrical ratios of an angle to which a reference was made in anticipation in Section 76. Since (Fig. 42)

$$a^2 + b^2 = c^2,$$

$$(a^2/c^2) + (b^2/c^2) = 1$$

or

$$(a/c)^2 + (b/c)^2 = 1,$$

that is,

$$(\cos \beta)^2 + (\sin \beta)^2 = 1$$

or, as it is usually written,

$$\cos^2\beta + \sin^2\beta = 1.$$

This relationship can of course be expressed in a variety of ways. For instance division by $\cos^2\beta$ gives

$$1 + \tan^2\beta = \sec^2\beta$$

and so on.

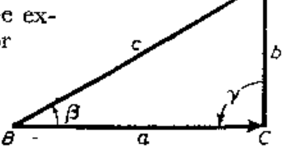


Fig. 42

Examples XII

1. The vectors **a** and **b** are respectively 5 cm and 10 cm in length, and the angle between them is 30°. What are the values of

$$\mathbf{a} \cdot \mathbf{b},$$
$$|\mathbf{a} + \mathbf{b}|,$$
$$|\mathbf{a} - \mathbf{b}|,$$
$$(\mathbf{a} + \mathbf{b}) \cdot (\mathbf{a} - \mathbf{b})\ ?$$

Show by calculation that

$$(\mathbf{a} + \mathbf{b})^2 = a^2 + \sqrt{3}\, ab + b^2,$$
$$(\mathbf{a} - \mathbf{b})^2 = a^2 - \sqrt{3}\, ab + b^2.$$

2. A rhombus is a four-sided figure with all its sides of equal length. Show, by vectors, that its diagonals are mutually perpendicular.

3. Show that $\mathbf{a}^2 = \text{const.} = r$

defines a circle of radius r if one end of **a** is fixed.

4. Show that the area of a parallelogram having the vectors **a** and **b** as sides is

$$\sqrt{(ab)^2 - (\mathbf{a} \cdot \mathbf{b})^2}.$$

89. The Operator "j"

The symbol "j" is very widely employed in alternating current analysis but there is a divergence of opinion as to its essential character and occasionally discussions arise as to the legitimacy of certain applications of it or as to the interpretation of expressions in which it appears.

Of course any system of ideas which, to put it colloquially, "delivers the goods" and which is self-consistent, can be used as the basis of a system of symbolic logic. The operator interpretation here described is not put forward as the inspired word and only true gospel. All that is claimed for it is that it is clear and self-consistent and avoids some of the difficulties of interpretation associated with the imaginary quantity and complex number nomenclature.

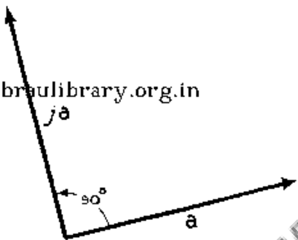


Fig. 43

Two types of operator with a vector operand have already been introduced. The first is simple scalar multiplication represented by the a in $a\mathbf{a}_1$, which increases the length of its operand without altering its direction. The second is the operator -1 which reverses the direction of its operand without changing its magnitude. A third will now be introduced. The operator j rotates its vector operand through $\pi/2$ (90°) in a positive direction in a given plane without changing its magnitude (the plane of the operation will be taken throughout as the plane of the paper). The vectors $\mathbf{a}$ and $j\mathbf{a}$ will therefore be as shown in Fig. 43.

It follows from the definition of j that

$$j(j\mathbf{a}) = -\mathbf{a}.$$

For shortness $j(j\mathbf{a})$ can be written $jj\mathbf{a}$ and still more compactly as $j^2\mathbf{a}$, but the real significance of j^2 should be borne in mind. Then

$$j^2(\ ) = -1\ (\ ).$$

Similarly $$j^3 = -j,$$
and $$j^4 = 1,$$
so that the relationship between "powers of j" and j is the same as that between powers of $\sqrt{-1}$ and $\sqrt{-1}$. Whether or no this establishes any identity between j and $\sqrt{-1}$ is irrelevant to the present purpose. The important point is the effect of powers of j, that is, of successive operations with j.

The operator j can obviously be combined with a scalar number or multiplier, and the association is commutative, for ajv is the same vector as jav, from which it follows that the operator aj is the same in effect as ja.

90. The Operator $(a + jb)$

The obvious interpretation of $(a + jb)v$ is $av + jbv$. This is illustrated in Fig. 44. Notice that

$$(a + jb)v = (jb + a)v$$

by elementary geometry. In the above, a and b can be any positive or negative numbers or fractions.

It follows from para. (c) of Section 88 that (Fig. 44)

$$BA^2 = BC^2 + CA^2,$$

that is, writing r for the magnitude of BA,

$$r^2 = a^2 + b^2,$$

or

$$r = +\sqrt{a^2 + b^2},$$

while θ is defined by

$$\sin\theta = b/r, \quad \cos\theta = a/r, \quad \tan\theta = b/a,$$

which make θ quite definite for given signs and magnitudes of a and b. The effect of the operator $(a + jb)$ is thus seen to be a rotation through θ specified as above, combined with a multiplication by $r = \sqrt{a^2 + b^2}$, and this

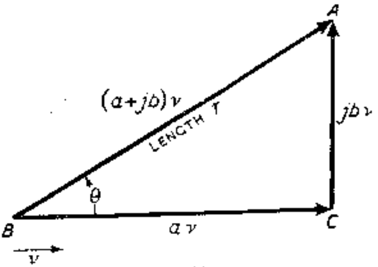


Fig. 44

effect is quite independent of the actual operand, which may be a unit vector as in Fig. 44, or any other vector whatever. Further, by a suitable choice of a and b, r and θ can be made to assume any values whatever, and any vector in the plane of the operation can be represented in the form $(a + jb)\nu$. The operator $(a + jb)$ can therefore be regarded as the most general possible form of coefficient in any given plane. Its importance for students of electricity lies in the fact (to be demonstrated later) that any alternating current impedance can be expressed in this form. The fairly detailed study of this coefficient and its combinations is therefore a practical necessity. The calculus of such coefficients will prove to be identical in form with that of complex numbers, but its full geometrical significance will probably be missed if this is simply taken for granted.

The effect of the operator can be specified exactly in terms of a scalar product. If the vector $\mathbf{u}$ is of length u and direction ψ with respect to ν, then

$$\mathbf{u} \cdot \nu = u \cos \psi,$$

and from the preceding paragraph it follows that

$$(a + jb)\,\mathbf{u} \cdot \nu = ru \cos (\psi + \theta),$$

a result which should be noted carefully because of its later significance in alternating current analysis.

91. Addition of Operators

By applying the laws of commutation and association in the addition of vectors,

$$\begin{aligned}\{(a + jb) + (c + jd)\}\nu &= (a + jb)\nu + (c + jd)\nu \\ &= a\nu + jb\nu + c\nu + jd\nu \\ &= a\nu + c\nu + jb\nu + jd\nu \\ &= (a + c)\nu + (jb + jd)\nu \\ &= \{(a + c) + j(b + d)\}\nu,\end{aligned}$$

that is, the operator $(a + jb) + (c + jd)$ is equivalent to the operator $(a + c) + j(b + d)$.

The reader is advised to interpret all these steps geometrically. It appears that the addition of operators follows the same rules as that of complex numbers. The result can obviously be extended to the addition (or subtraction) of any number of operators.

92. Equality of Operators

The equation

$$(a + jb)\nu = (c + jd)\nu$$

clearly implies the operator equality

$$(a + jb) = (c + jd).$$

If to each of the equal vectors in the first equation we add the vector $(-jb - c)\nu$, we get

$$(a - c)\nu = (d - b)j\nu,$$

that is, a vector equal to another which is perpendicular to it; but this is impossible by the definition of a vector unless each is zero.

Therefore if $\quad (a + jb) = (c + jd),$

then $\quad a - c = 0$ and $d - b = 0,$

that is, $\quad a = c$ and $b = d.$

This process can be compared with the separate equating of the real and imaginary parts of equal complex numbers. In operators the components can be referred to as the a and b parts, or, perhaps better, as the "axial" and "non-axial" parts. Neither of these alternatives is quite satisfactory, but either is better than the customary "real" and "imaginary" which are misleading and confusing.

As a special case of the above, if

$$a + jb = 0,$$

then $\quad a = 0$ and $b = 0.$

93. Multiplication and Division of Operators

By the geometry of similar triangles it is easy to show that

$$\begin{aligned} a\{(c + jd)\nu\} &= a(c\nu + jd\nu) \\ &= ac\nu + ajd\nu \\ &= ac\nu + jad\nu, \end{aligned}$$

and similarly,

$$b\{(c + jd)v\} = bcv + jbdv,$$

and
$$jb\{(c + jd)v\} = jbcv + jjbdv$$
$$= jbcv - bdv,$$

and, by addition,

$$(a + jb)\{(c + jd)v\} = acv + jadv + jbcv - bdv$$
$$= \{(ac - bd) + j(ad + bc)\}v.$$

The intermediate brackets on the left-hand side can be omitted, though they are required for a comprehensible interpretation. Then

$$(a + jb)(c + jd)v = \{(ac - bd) + j(ad + bc)\}v,$$

or, considering the operators only,

$$(a + jb)(c + jd) = (ac - bd) + j(ad + bc),$$

which again is similar to the corresponding operation with complex numbers.

The above two results for addition and multiplication can be extended to division, powers, etc., precisely as for the same operations with real or complex numbers described in Section 53 *et seq.*, and no new rules have to be learnt. The interpretations will be sufficiently obvious in most cases. For instance the equation

$$\mathbf{u} = \frac{1}{a + jb}\mathbf{v}$$

really means that **u** is that vector which, operated on with $(a + jb)$, gives **v**,

that is,
$$(a + jb)\mathbf{u} = \mathbf{v},$$

Operating on each of these equal vectors with $(a - jb)$ gives

$$(a - jb)(a + jb)\mathbf{u} = (a^2 + b^2)\mathbf{u} = (a - jb)\mathbf{v},$$

or
$$\mathbf{u} = \frac{(a - jb)}{(a^2 + b^2)}\mathbf{v},$$

which is the process corresponding to the rationalisation of complex numbers.

The interpretation of roots of operators will require a little more care, and will be considered more fully later on.

Before leaving this part of the subject, it will be well to point out the distinction between $(a + jb)(c + jd)v$, which implies successive operations on v, and $(a + jb)v \cdot (c + jd)v$, which is the scalar product of two vectors. By applying the results of Section 86 it is easy to show that the latter is simply the number $(ac + bd)$. This latter form will be met later in connection with expressions for the power in alternating current circuits.

94. Alternative Form for $(a + jb)$

Referring back to Fig. 44,

$$a = r\cos\theta, \text{ and } b = r\sin\theta,$$

so that

$$(a + jb) = r(\cos\theta + j\sin\theta)$$

and since the effect of $(a + jb)$ is multiplication by r and rotation through θ in a positive direction it is clear that the operator $(\cos\theta + j\sin\theta)$ represents the rotation through θ (Fig. 45).

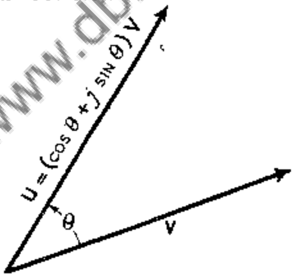


Fig. 45

So far θ has been taken to be inherently positive, but this restriction can be removed, for if

$$\mathbf{u} = (\cos\theta + j\sin\theta)\mathbf{v}$$

be written in the form

$$\mathbf{v} = \frac{1}{\cos\theta + j\sin\theta}\mathbf{u}$$

it is clear that $1/(\cos\theta + j\sin\theta)$ means a rotation of amount θ in a negative direction, that is, a rotation $-\theta$. But by the preceding section,

$$\frac{1}{\cos\theta + j\sin\theta} = \frac{\cos\theta - j\sin\theta}{\cos^2\theta + \sin^2\theta}$$

$$= \cos\theta - j\sin\theta$$

$$= \cos(-\theta) + j\sin(-\theta),$$

so that $\cos(-\theta) + j\sin(-\theta)$ effects a rotation of $-\theta$. Therefore for positive or negative values of θ, $(\cos\theta + j\sin\theta)$ effects a rotation θ.

Successive rotations of θ_1 and θ_2, these being positive or negative, are obviously equivalent to a single rotation $(\theta_1 + \theta_2)$. Therefore

$$(\cos\theta_1 + j\sin\theta_1)(\cos\theta_2 + j\sin\theta_2)$$
$$= \cos(\theta_1 + \theta_2) + j\sin(\theta_1 + \theta_2).$$

This remarkable formula, which, in its operator interpretation is self-evident, embodies what is known in pure mathematics as De Moivre's Theorem. Many results of far reaching importance can be deduced from it. Some of these will now be considered.

95. The Addition Formulæ of Trigonometry

$\cos(\theta_1 + \theta_2) + j\sin(\theta_1 + \theta_2)$

$= (\cos\theta_1 + j\sin\theta_1)(\cos\theta_2 + j\sin\theta_2)$

$= \cos\theta_1\cos\theta_2 - \sin\theta_1\sin\theta_2 + j(\sin\theta_1\cos\theta_2 + \cos\theta_1\sin\theta_2)$

by multiplication, as in Section 93. Equating separately the a and b parts,

$$\cos(\theta_1 + \theta_2) = \cos\theta_1\cos\theta_2 - \sin\theta_1\sin\theta_2,$$
$$\sin(\theta_1 + \theta_2) = \sin\theta_1\cos\theta_2 + \cos\theta_1\sin\theta_2.$$

From these two important formulæ, many others can be derived as special cases. For instance, if $-\theta_2$ be written for θ_2,

$$\cos(\theta_1 - \theta_2) = \cos\theta_1\cos\theta_2 + \sin\theta_1\sin\theta_2,$$
$$\sin(\theta_1 - \theta_2) = \sin\theta_1\cos\theta_2 - \cos\theta_1\sin\theta_2,$$

and if in the first pair we put $\theta_1 = \theta_2 = \theta$,

$$\cos 2\theta = \cos^2\theta - \sin^2\theta$$
$$= 2\cos^2\theta - 1$$
$$= 1 - 2\sin^2\theta$$
$$(\text{since } \cos^2\theta + \sin^2\theta = 1).$$

Also $$\sin 2\theta = 2\sin\theta\cos\theta.$$

Similarly formulæ for the tangent and cotangent of sums and differences, and for the ratios of 3θ, 4θ, etc., can be obtained by inserting appropriate special values in the original formulæ. The product formulæ

$$2 \sin \theta_1 \sin \theta_2 = \cos(\theta_1 - \theta_2) - \cos(\theta_1 + \theta_2), \text{ etc.,}$$

should be noted, as they have many applications in wireless telegraphy and telephony. They are most useful in the form obtained by putting $\theta_1 + \theta_2 = x$, $\theta_1 - \theta_2 = y$ so that $\theta_1 = \frac{x+y}{2}$, $\theta_2 = \frac{x-y}{2}$. We then find

$$\sin x + \sin y = 2 \sin \frac{x+y}{2} \cos \frac{x-y}{2}$$

$$\sin x - \sin y = 2 \cos \frac{x+y}{2} \sin \frac{x-y}{2}$$

$$\cos x + \cos y = 2 \cos \frac{x+y}{2} \cos \frac{x-y}{2}$$

$$-\cos x + \cos y = 2 \sin \frac{x+y}{2} \sin \frac{x-y}{2}$$

Only the formulæ for $\sin(\theta_1 + \theta_2)$ and $\cos(\theta_1 + \theta_2)$ need be remembered by heart, as long as the existence of the formulæ immediately above, and the substitution giving θ_1, θ_2 in terms of x, y and vice versa, are noted. The signs of the x and y formulæ sometimes are found difficult; they are true whatever the values and signs of x and y, but it is easiest to think as if x is greater than y and both are positive and less than $\frac{1}{2}\pi$ (90°). Then the last formula, the most difficult, makes sense because if $\frac{1}{2}\pi > x > y > 0$, $1 > \cos y > \cos x > 0$; the left-hand side $-\cos x + \cos y$ is therefore positive, while the right-hand side is obviously positive.

These "addition formulæ" can also be used to simplify expressions like

$$C = a \cos \theta_1 + b \sin \theta_1$$

for which the appropriate trick is to let

$$r = (a^2 + b^2)^{\frac{1}{2}}$$

and then let θ, as in Fig. 44 (Section 89), be the angle for which

$$\sin \theta = b/r, \cos \theta = a/r, \tan \theta = b/a$$

We then find

$$C = r\left\{\frac{a}{r}\cos\theta_1 + \frac{b}{r}\sin\theta_1\right\}$$
$$= r\{\cos\theta\cos\theta_1 + \sin\theta\sin\theta_1\}$$
$$= r\cos(\theta - \theta_1).$$

so that the maximum value of C is r and it occurs when $\theta = \theta_1$.

It is sometimes also useful to write the expression

$$D = a\sin x + b\sin y$$

in the equivalent form

$$D = (a+b)\sin\frac{x+y}{2}\cos\frac{x-y}{2} + (a-b)\cos\frac{x+y}{2}\sin\frac{x-y}{2}$$

particularly if x and y, or a and b, or both pairs, are nearly equal. The second form of D is easily deduced from the formulæ already given for $\sin x \pm \sin y$, and expressions of the form

$$\alpha\cos x + \beta\cos y$$

can be similarly treated, whatever the signs of x, y, a, b, α and β.

96. The Exponential Form for $(\cos\theta + j\sin\theta)$

Since n equal rotations of θ are equal to a single rotation $n\theta$, n being a positive integer,

$$\cos n\theta + j\sin n\theta = (\cos\theta + j\sin\theta)^n = \cos^n\theta\,(1 + j\tan\theta)^n.$$

Now put

$$n\theta = \phi,$$

then

$$\cos\phi + j\sin\phi = \cos^n\frac{\phi}{n}\left(1 + j\tan\frac{\phi}{n}\right)^n.$$

This formula will remain true however large n may be, so that

$$\cos\phi + j\sin\phi = \lim_{n\to\infty}\cos^n\frac{\phi}{n}\left(1 + j\tan\frac{\phi}{n}\right)^n.$$

Now by sufficiently increasing n, $\tan\frac{\phi}{n}$ can be made to differ from ϕ/n by as little as we please, and $\cos^n\frac{\phi}{n}$ can be made to differ from 1 by as little as we please. This can easily be appreciated by drawing a diagram showing θ, $\tan\theta$, and $\cos\theta$ for a very small angle θ. Therefore, remembering the definition of "limit" (Section 61),

$$\underset{n\to\infty}{lt.}\ \cos^n\frac{\phi}{n}\left(1+j\tan\frac{\phi}{n}\right)^n = \underset{n\to\infty}{lt.}\left(1+j\frac{\phi}{n}\right)^n$$

It has been shown that the multiplying together of operators follows the same rules as the multiplication of real or complex numbers. Therefore the product $\left(1+j\frac{\phi}{n}\right)^n$ can be written down by the Binomial Theorem for a positive integral index (Section 67), and the limit when n tends to infinity can be found in exactly the same way as is shown in full in Section 68. The result is

$$\cos\phi + j\sin\phi =$$

$$1 + j\phi + \frac{(j\phi)^2}{2!} + \frac{(j\phi)^3}{3!} + \text{etc., } ad\ inf.$$

The infinite series on the right will be written $S(j\phi)$ for shortness. If x is any real number,

$$S(x) = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \text{etc., } ad\ inf. = \epsilon^x$$

(Section 68), where ϵ is the number 2·71828 . . . . But this does not mean that $S(j\phi)$ is similarly $\epsilon^{j\phi}$, for how can any number be multiplied by itself $j\phi$ times? Nevertheless $\epsilon^{j\phi}$ will be adopted as a convenient short way of writing $S(j\phi)$, or the operator $(\cos\phi + j\sin\phi)$ which has the same effect, and this practice will not lead to any errors for it has already been shown that

$$(\cos\theta_1 + j\sin\theta_1)(\cos\theta_2 + j\sin\theta_2)$$
$$= \cos(\theta_1+\theta_2) + j\sin(\theta_1+\theta_2),$$

which, in the exponential form, becomes

$$\epsilon^{j\theta_1}\ \epsilon^{j\theta_2} = \epsilon^{j(\theta_1+\theta_2)},$$

which is in formal agreement with the index law. The $j\phi$ in $\epsilon^{j\phi}$ does actually function as an index as far as combinations of such quantities are concerned, and the exponential form of writing is therefore a convenient and legitimate way of expressing this fact. Remember, however, that the index behaviour is derived first and independently, and is not a deduction from the exponential form. On this understanding, any operator $(a + jb)$ can be written in the form

$$(a + jb) = r(\cos\theta + j\sin\theta) = r\epsilon^{j\theta},$$

where r and θ are as specified in Section 90. The convenience of this last form for calculation will be apparent later on. Remembering the effect of the operator $r\epsilon^{j\theta}$, it is easy to see that $r\epsilon^{j\theta}\nu$ is a vector of magnitude r making an angle θ with the direction of ν, so that

$$r\epsilon^{j\theta}\nu\cdot\nu = r\cos\theta.$$

In particular, the vector $\hat{e}\epsilon^{j\omega t}\nu$ is of magnitude $\hat{e}$ and makes with ν an angle which increases at the rate ω radians per second (t representing the time in seconds from some definite zero or starting point), that is, it is a vector of constant magnitude rotating with constant angular velocity, and

$$\hat{e}\epsilon^{j\omega t}\nu\cdot\nu = \hat{e}\cos\omega t.$$

Such an operator can therefore be used to specify an alternating e.m.f. in the vector form $\hat{e}\epsilon^{j\omega t}\nu$. In practice it will not be necessary to write in the unit vector of reference ν, but its implicit existence as the operand of $\hat{e}\epsilon^{j\omega t}$ should be borne in mind, and will generally facilitate the interpretation of vector calculations. The significance of this will appear more fully later on.

97. Series Form for Sine and Cosine

Another very interesting result is derived from De Moivre's Theorem as follows :—

$\cos\theta + j\sin\theta$

$$= 1 + j\theta + \frac{(j\theta)^2}{2!} + \frac{(j\theta)^3}{3!} + \frac{(j\theta)^4}{4!} + \frac{(j\theta)^5}{5!} + \text{etc.,} \quad \textit{ad inf.}$$

$$= 1 + j\theta - \frac{\theta^2}{2!} - \frac{j\theta^3}{3!} + \frac{\theta^4}{4!} + \frac{j\theta^5}{5!} + \text{etc., } \textit{ad inf.}$$

$$= \left(1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \text{etc., } ad\ inf.\right) + j\left(\theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \text{etc., } ad\ inf.\right).$$

Therefore equating separately the a and b parts of these operators

$$\cos\theta = 1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \text{etc., } ad\ inf.$$

$$\sin\theta = \theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \text{etc., } ad\ inf.$$

These series are rapidly convergent for small values of θ and give as convenient approximations for such small values

$$\cos\theta = 1 - \theta^2/2,$$
$$\sin\theta = \theta - \theta^3/6.$$

98. The Generalisation of the Sine and Cosine

The sine and cosine of θ have a simple and intelligible geometrical meaning when θ is any real number of radians, positive, negative, or fractional, and we have seen that they can also be represented as a series of powers of θ; but an important thing to notice about the series is that they have simple and intelligible meanings, even if θ is a vector operator of the form $z = a + jb$, for they then represent the sums of a number of operators, each of which can be reduced to the form $a + jb$. Such series will arise in practice, and it will then be convenient, and will save a lot of writing, to denote them by the abbreviations $\sin z$ and $\cos z$. This, as far as we are concerned, is all that is meant by the so-called generalisation of the circular functions to include "complex" arguments.

99. The Exponential Form of the Sine and Cosine

It follows directly from De Moivre's Theorem that

$$\sin\theta = \frac{\epsilon^{j\theta} - \epsilon^{-j\theta}}{2j},$$

$$\cos\theta = \frac{\epsilon^{j\theta} + \epsilon^{-j\theta}}{2},$$

with similar formulæ for the other circular trigonometrical ratios. This result is not of much practical value, and is mentioned chiefly as an introduction to the next Section.

100. Hyperbolic Functions

Consider now the following exponential expressions, similar to those in the preceding section, that is,

$$\frac{\epsilon^x - \epsilon^{-x}}{2}$$

and

$$\frac{\epsilon^x + \epsilon^{-x}}{2}.$$

If x is any real number, positive, negative, or fractional,

$$\frac{\epsilon^x - \epsilon^{-x}}{2} = x + \frac{x^3}{3!} + \frac{x^5}{5!} + \frac{x^7}{7!} + \ldots\ldots$$

and

$$\frac{\epsilon^x + \epsilon^{-x}}{2} = 1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \frac{x^6}{6!} + \ldots\ldots$$

and, as we know already, the series are convergent.

These functions prove to be quite useful in connection with the theory of cables and transmission lines. They are called respectively the hyperbolic sine and hyperbolic cosine of x (because they have to the curve known as the rectangular hyperbola a relation somewhat similar to that of the sine and cosine to the circle).

The function

$$\frac{\epsilon^x - \epsilon^{-x}}{\epsilon^x + \epsilon^{-x}},$$

which is clearly the ratio of the hyperbolic sine to the hyperbolic cosine, is called the hyperbolic tangent. But these full ceremonial titles are too long for currency and are abbreviated as shown in the following list :—

hyperbolic sine $x = \sinh x$ (pronounced " shine x ")

$$= \frac{\epsilon^x - \epsilon^{-x}}{2},$$

hyperbolic cosine $x = \cosh x$ (pronounced " cosh x ")

$$= \frac{\epsilon^x + \epsilon^{-x}}{2},$$

hyperbolic tangent $x = \tanh x$ (pronounced " than x "—*th* as in *th*ink)

$$= \frac{\epsilon^x - \epsilon^{-x}}{\epsilon^x + \epsilon^{-x}},$$

with similar definitions for the other derived functions sech x (pronounced " sheck x "), cosech x and coth x.

From the above definitions it will be quite easy to prove a number of formulæ analogous to those of trigonometry,

for example, $\cosh^2 x - \sinh^2 x = 1.$

So far, we have been talking pure and simple algebra ; but just as it was found possible to give an intelligible meaning to sin θ for complex or " operator " values of the argument, so also sinh z, where $z = (a + jb)$, is simply *a short way of writing the sum of the series of separate operations represented by* z, $z^3/3\,!$, $z^5/5\,!$, etc.

If this is clearly understood, the reader will have no qualms about such operators as $\tanh (a + jb)$ when he meets them later in cable theory. He should even be able to show that, far from being unintelligible mathematical abstractions, they correspond to definite physical realities—the summation of electric waves that are being continually reflected to and fro along the same path.

The generalised interpretation of the circular and hyperbolic functions, taken in conjunction with the exponential formulæ, should enable the reader both to show, and to understand, a number of connections between the two sets of functions for operator arguments, for example,

$$\sin jb = j \sinh b,$$
$$\cos jb = \cosh b,$$
$$\tan jb = j \tanh b,$$
etc., etc.

101. Roots of Operators

By analogy with the ordinary arithmetical meaning of " root ", the nth root of the operator $r(\cos \theta + j \sin \theta)$

will be taken to mean any operator which, repeated n times, that is, raised to the nth power, has the same effect as $r(\cos\theta + j\sin\theta)$. We shall now see that if n is an integer there will be n different operators which fulfil this condition. First let $r^{1/n}$ be written for that positive number which, raised to the nth power, gives r. For given values of r and n there is only one such number. Now if m is any integer,

$$\left[r^{1/n}\left\{\cos\left(\frac{\theta}{n} + \frac{2m\pi}{n}\right) + j\sin\left(\frac{\theta}{n} + \frac{2m\pi}{n}\right)\right\}\right]^n$$
$$= r\{\cos(\theta + 2m\pi) + j\sin(\theta + 2m\pi)\},$$

as already shown in Section 96. By drawing a simple quadrant diagram it will be easy to see that for all integral values of m

$$\cos(\theta + 2m\pi) = \cos\theta, \text{ and } \sin(\theta + 2m\pi) = \sin\theta,$$

so that

$$\left[r^{1/n}\left\{\cos\left(\frac{\theta}{n} + \frac{2m\pi}{n}\right) + j\sin\left(\frac{\theta}{n} + \frac{2m\pi}{n}\right)\right\}\right]^n$$
$$= r(\cos\theta + j\sin\theta),$$

and all the operators obtained by giving m any integral value in

$$r^{1/n}\left\{\cos\left(\frac{\theta}{n} + \frac{2m\pi}{n}\right) + j\sin\left(\frac{\theta}{n} + \frac{2m\pi}{n}\right)\right\}$$

are nth roots of $r(\cos\theta + j\sin\theta)$. This seems rather a lot, but actually they are not all different. For instance if n is 3, then putting 0, 1, 2, 3 for m gives

$$r^{\frac{1}{3}}\{\cos(\theta/3) + j\sin(\theta/3)\},$$
$$r^{\frac{1}{3}}\left\{\cos\left(\frac{\theta}{3} + \frac{2\pi}{3}\right) + j\sin\left(\frac{\theta}{3} + \frac{2\pi}{3}\right)\right\},$$
$$r^{\frac{1}{3}}\left\{\cos\left(\frac{\theta}{3} + \frac{4\pi}{3}\right) + j\sin\left(\frac{\theta}{3} + \frac{4\pi}{3}\right)\right\},$$
$$r^{\frac{1}{3}}\left\{\cos\left(\frac{\theta}{3} + 2\pi\right) + j\sin\left(\frac{\theta}{3} + 2\pi\right)\right\}.$$

But the last is a repetition of the first. It will be found that putting higher values for m, or giving it any negative integral value, will only give a repetition of one or other of the first three roots. In general, there are only *n different* values of the nth root of any operator. Any reader to whom this is new should draw out the various operators in a simple case, say, for $n = 3$. The geometrical significance of this many-valuedness will then be apparent. The value obtained by putting $m = 0$ is called the "principal value", and is sometimes written $\{r(\cos\theta + j\sin\theta)\}^{1/n}$, the alternative form

$$\sqrt[n]{r(\cos\theta + j\sin\theta)}$$

being used to indicate all or any of the roots. The many-valuedness of the root sometimes enters into calculations, and should always be borne in mind. The following are some important special cases.

(*a*) If $\theta = 0$,

$$\sqrt[n]{r} = r^{1/n}\left(\cos\frac{2m\pi}{n} + j\sin\frac{2m\pi}{n}\right).$$

$$m = 0, 1, 2, \ldots (n-1),$$

the principal value being $r^{1/n}$. If n is 2,

$$\sqrt{r} = r^{\frac{1}{2}}(\cos 0 + j\sin 0)\,;\; r^{\frac{1}{2}}(\cos\pi + j\sin\pi) = = +r^{\frac{1}{2}}\,;\; -r^{\frac{1}{2}}.$$

that is, $\sqrt{r} = \pm r^{\frac{1}{2}}$.

(*b*) Putting $\theta = \pi$,

$$\sqrt[n]{r(\cos\pi + j\sin\pi)} = \sqrt[n]{-r}$$

$$= r^{1/n}\left\{\cos\frac{(2m+1)\pi}{n} + j\sin\frac{(2m+1)\pi}{n}\right\},$$

and the principal value is

$$r^{1/n}\left\{\cos\frac{\pi}{n} + j\sin\frac{\pi}{n}\right\}.$$

For instance, if n is 2,

$$\sqrt{-r} = r^{\frac{1}{2}}\left\{\cos\frac{\pi}{2} + j\sin\frac{\pi}{2}\right\}$$

$$\text{or}\quad r^{\frac{1}{2}}\left\{\cos\frac{3\pi}{2} + j\sin\frac{3\pi}{2}\right\}$$

that is, $\sqrt{-r} = \pm jr^{\frac{1}{2}}$.

102. Generalisation of the Index Formula for Operators

It has already been shown that for any positive or negative integral value of n

$$\{r(\cos\theta + j\sin\theta)\}^n = r^n(\cos n\theta + j\sin n\theta).$$

Also, by the preceding section, the same is true for the principal value when n is $1/m$, m being a positive integer. This leads without difficulty to the complete generalisation that for any positive or negative integral or fractional value of n, the principal value of

$$\{r(\cos\theta + j\sin\theta)\}^n \quad \text{is} \quad r^n(\cos n\theta + j\sin n\theta).$$

Space cannot be spared for a detailed development of this generalisation, which follows exactly the same lines as the generalisation of the interpretation of an index given in Sections 27 *et seq.*

103. Calculations with Operators

We have seen that any operator $(a + jb)$ can be put in the form $r\epsilon^{j\theta}$, where $r^2 = a^2 + b^2$ and θ is one of the angles having b/a as tangent. Referring to the quadrant diagram (Fig. 30) it will be seen that if a and b are positive, θ will be an angle in the first quadrant, that is, less than 90°. If b is positive and a negative θ will lie in the second quadrant, and so on. Writing $|b|$ and $|a|$ for the magnitudes of b and a, then the angle $\alpha = \tan^{-1}|b|/|a|$ is between 0 and 90°, and can be determined from the ordinary table of tangents, and θ will be $+\alpha$, $\pi - \alpha$, $\pi + \alpha$, or $-\alpha$, according to whether it is in the first, second, third or fourth quadrant. An example will perhaps make this clearer. For the operator $3 - j4$, r is 5 and θ is an angle in the fourth quadrant. Now $\tan^{-1} 4/3$ is 53.1°, so that θ is $-53.1°$. For the operator $-3 + j4$, on the other hand, θ will be in the second quadrant, so that it is given by $(180 - 53.1)°$, that is, 126.9°.

By using the form $r\epsilon^{j\theta}$, the single operator which is equal to, that is, has the same effect as, any given combination of operators, can be written down at once. For instance, if

$$a + jb = r\epsilon^{j\theta},$$
$$c + jd = s\epsilon^{j\phi},$$

then

$$(a + jb)^2/(c + jd) = (r\epsilon^{j\theta})^2/s\epsilon^{j\phi} = r^2\epsilon^{j2\theta}/s\epsilon^{j\phi}$$
$$= (r^2/s)\epsilon^{j(2\theta-\phi)},$$

which can then be put in the form $A + jB$ if desired, where

$$A = (r^2/s)\cos(2\theta - \phi),$$
$$B = (r^2/s)\sin(2\theta - \phi).$$

A simple type of calculation which will occur frequently in alternating current calculations is exemplified by the following.

Given that $\mathbf{e} = \hat{e}\epsilon^{j\omega t}\nu,$

and that $\mathbf{i} = \mathbf{e}/(R + jX)$

R and X being numbers, find $\mathbf{i}\cdot\nu$.

Notice that

$$\mathbf{e}\cdot\nu = \hat{e}\epsilon^{j\omega t}\nu\cdot\nu = \hat{e}\cos\omega t$$

is the instantaneous value of an alternating e.m.f. It will be shown later that $\mathbf{i}\cdot\nu$ is similarly the instantaneous value of the "steady state" alternating current produced by this e.m.f. in a circuit of impedance $(R + jX)$. Expressing this impedance in the form

$$R + jX = Z\epsilon^{j\phi},$$

where $Z^2 = R^2 + X^2$, and $\tan\phi = X/R$,

then

$$\mathbf{i} = \frac{\mathbf{e}}{R + jX} = \frac{\hat{e}\epsilon^{j\omega t}}{Z\epsilon^{j\phi}}\nu = \frac{\hat{e}}{Z}\epsilon^{j(\omega t-\phi)}\nu,$$

so that

$$\mathbf{i}\cdot\nu = (\hat{e}/Z)\epsilon^{j(\omega t-\phi)}\nu\cdot\nu = (\hat{e}/Z)\cos(\omega t - \phi).$$

An alternative form for this result may sometimes be more convenient. The reciprocal of $(R + jX)$ can be expressed in the form

$$\frac{1}{(R + jX)} = \frac{R - jX}{R^2 + X^2} = \frac{R}{Z^2} - \frac{jX}{Z^2} = \frac{R}{Z^2} + \frac{X}{Z^2}\epsilon^{-j\pi/2}$$

since $-j = \epsilon^{-j\pi/2}$.

Therefore

$$\begin{aligned}\mathbf{i} \cdot \nu &= (R/Z^2)\, \hat{e}\epsilon^{j\omega t}\, \nu \cdot \nu + (X/Z^2)\, \hat{e}\epsilon^{j\omega t}\epsilon^{-j\pi/2}\, \nu \cdot \nu \\ &= (R/Z^2)\, \hat{e} \cos \omega t + (X/Z^2)\, \hat{e} \cos (\omega t - \pi/2) \\ &= (R/Z^2)\, \hat{e} \cos \omega t + (X/Z^2)\, \hat{e} \sin \omega t.\end{aligned}$$

So much for a brief introductory survey of the main features of vectors and vector operators. The more detailed study of the subject will be confined to those regions where it comes into immediate contact with the field of alternating current phenomena, in the discussion of which it will be found to infuse a beautiful simplicity; but first we must see how it comes to be associated with what would at first sight seem to be a quite different set of ideas. The nature of the connection has already been foreshadowed, but for its complete delineation some little knowledge of the differential and integral calculus will be required, and to this branch of mathematics the next part of this book will be devoted.

Examples XIII

1. Prove the following formulæ :—

$$\cosh^2 x - \sinh^2 x = 1,$$
$$\cosh^2 x + \sinh^2 x = \cosh 2x,$$
$$2 \sinh x \cosh x = \sinh 2x,$$
$$\sinh (x + y) = \sinh x \cosh y + \cosh x \sinh y.$$

2. Prove the following formulæ :—

$$\sin (a + jb) = \sin a \cosh b + j \cos a \sinh b,$$
$$\cos (a + jb) = \cos a \cosh b - j \sin a \sinh b,$$
$$\sin (a + jb) = \frac{\sin 2a + j \sinh 2b}{\cos 2a + \cosh 2b},$$
$$\tanh (a + jb) = \frac{\sinh 2a + j \sin 2b}{\cosh 2a + \cos 2b}.$$

3. Show that $\dfrac{a - jb}{a + jb} = \epsilon^{-2j\theta}$, where $\tan \theta = b/a$.

4. Hence show that if a is constant and b variable, and z is given by

$$z = \frac{1}{2a}\left(1 + \frac{a - jb}{a + jb}\right),$$

the locus of z is a circle of radius $1/2a$.

Finally, show that $z = \dfrac{1}{a + jb}$.

5. Show that $\dfrac{1 + r\epsilon^{-j\theta}}{1 + r\epsilon^{j\theta}} = \epsilon^{-2j\phi}$,

where $$\tan \phi = \frac{r \sin \theta}{1 + r \cos \theta}.$$

6. Confirm the following transformations:

$$\frac{1 - r\epsilon^{j\theta}}{1 + r\epsilon^{j\theta}} = \frac{1 + r^2}{1 - r^2} - \frac{2r}{1 - r^2}\,\epsilon^{j(\theta - 2\phi)},$$

and $\dfrac{1 - r\epsilon^{j\theta}}{1 + r\epsilon^{j\theta}} = j\,\{\cot \theta - \operatorname{cosec} \theta\, \epsilon^{j(\theta - 2\phi)}\}$,

where $\tan \phi = \dfrac{r \sin \theta}{1 + r \cos \theta}$.

Hence show that the locus of the operator $\frac{1 - r\epsilon^{j\theta}}{1 + r\epsilon^{j\theta}}$ is a circle if r is constant and θ varies, and also if θ is constant and r varies.

7. Show that tanh $(a + jb)$ can be put in the form

$$\frac{1 - r\epsilon^{j\theta}}{1 + r\epsilon^{j\theta}}.$$

8. Express

$$z = \frac{(a + jb)(c + jd)}{(e + jf)}$$

in the forms $A + jB$, and $r\epsilon^{j\theta}$.

9. One of the sides of an equilateral triangle is the vector **a**. Express the other sides in terms of **a** and vector operators. Hence show that

$$1 + \epsilon^{j2\pi/3} + \epsilon^{j4\pi/3} = 0,$$

and confirm by calculation.

Chapter 6

THE DIFFERENTIAL AND INTEGRAL CALCULUS

104. The Object and Scope of the Differential Calculus

THE Differential Calculus is concerned with the systematic study of variation, and its field of application is therefore co-extensive with the whole domain of natural phenomena. As Professor Whitehead has expressed it, ". . . the fundamental idea of change, which is at the basis of our whole perception of phenomena, immediately suggests the inquiry as to rate of change. . . . Thus the differential calculus is concerned with the very key of the position from which mathematics can be successfully applied to the course of nature".

105. Rate of Change

The first requirement is to make quite clear what is meant by "rate of change". Rate is the same word as ratio, so that "rate of change" can be paraphrased as "ratio of changes", in which form its meaning is already much clearer. It implies two quantities, one of which changes in consequence of a change in the other. In more technical language, it implies a function and an independent variable. The rate of change of the function is thus the ratio of the change in the function to the change in the variable which produces it.

The most suitable example to take will be that one the contemplation of which first evoked in the brain of Newton the ideas on which the present form of the calculus is based—a body moving in space, or, to make it a little less abstract, a train moving along rails. The distance (measured along the rails from some fixed point) which is travelled by the moving train is, in the full mathematical sense of the word, a function of time. In all such physical problems, " time " means, of course, an interval of time measured from some arbitrary " zero " of time (anyone who served in the war will know only too well what " zero hour " means). To fix ideas in the above instance, we will take as the origin of distance some fixed point on the rails, and as the origin or " zero " of time the actual clock time at which the engine passes this fixed point. The distance from the fixed point will be represented by the symbol s, which stands for a number (of miles, feet, inches, or whatever unit is most convenient) and the time will be represented by the symbol t, also a number (of hours, minutes, seconds, or whatever time unit is most convenient). Then, in general,

$$s = f(t),$$

which is a short way of saying " the distance depends upon the time ". Suppose we are told that in a given case the form of the function is linear (see Section 37), that is,

$$s = a + bt,$$

a and b being constant numbers. In the interval of time between t and $t + \delta t$ (δt being considered as a single symbol meaning " a change of t "), s will increase by an amount which we will represent by δs. The relation between $s + \delta s$ and $t + \delta t$ is that given above, that is,

$$\begin{aligned} s + \delta s &= a + b(t + \delta t) \\ &= a + bt + b\delta t, \end{aligned}$$

and since

$$s = a + bt,$$

$$\delta s = b\delta t.$$

Therefore the ratio of the change of s to the change of t is

$$\delta s/\delta t = b,$$

b representing a certain number of miles per hour or feet per second or whatever the selected units may be. It is in fact the speed or velocity of the train, and since it does not depend either on t or on δt the movement of the train is completely described by the single constant b.

But now suppose that the form of the function is given as

$$s = a + bt + ct^2.$$

Then in precisely the same way as before it will be found that

$$\delta s = b\,\delta t + 2ct\,\delta t + c\,\delta t^2 = (b + 2ct + c\delta t)\delta t,$$

so that the rate of change of s is

$$\delta s/\delta t = b + 2ct + c\delta t.$$

This depends not only on t, the beginning of the interval δt, but also on δt, the length of the interval. If the interval δt could be reduced to zero we could say that at the instant t the train was travelling at a speed $(b + 2ct)$ miles per hour, or whatever the units might be. But that is just what we cannot do, for each side of the equation can only be divided by δt as long as δt is *not* zero (see page 49). Moreover, considering the matter physically, one cannot measure the distance δs travelled in zero time.

This is the difficulty which confronted Newton. He probably solved it for himself so completely that he lost sight of the difficult character of the ideas involved. Be that as it may, he did not resolve the difficulty in language sufficiently to prevent confusion of thought on the part of some of his disciples. The more discerning mathematicians were greatly worried by this difficulty for a long while after Newton. It did not worry the less discerning ones, for they blotted it out under a cloud of bad philosophy. It appeared that the quantity δt had to be both zero and not zero. "Fancy that!" they said, "What a wonderful quantity it must be!" and gave it a wonderful name, calling it an "infinitesimal". But, unfortunately, as Napoleon remarked on one occasion, "You can call a thing what you like, but you cannot prevent it from being what it is". The infinitesimal in its original form was a disappointing child, and died comparatively young.

Actually, of course, there is no need for any mysticism in this matter. The difficulty is completely resolved by means of the conception of "limit" described briefly in Section 61. The reader is strongly advised to read this section again, or, better still, to re-write it for himself in his own words. That is always the best way of learning a new set of ideas.

As long as δt remains finite, however small it may be, then

$$\delta s/\delta t = (b + 2ct) + c\delta t.$$

Now by taking δt sufficiently small, $\delta s/\delta t$ can be made to differ from $(b + 2ct)$ by less than any assigned amount, that is, it can be made to approximate to $(b + 2ct)$ within every standard. This is expressed symbolically

$$\underset{\delta t \to 0}{lt.} (\delta s/\delta t) = b + 2ct.$$

The expression on the left is inconveniently long to write down, and it is commonly abbreviated to ds/dt. Thus, if

$$s = a + bt + ct^2$$
$$ds/dt = b + 2ct.$$

The symbol ds/dt is called " the differential coefficient of s with respect to t ". Notice the " symbol " ds/dt—not the " fraction " ds/dt, because it is *not* a fraction. The parts ds and dt considered separately are quite undefined. The ds is *not* the limit of δs when δs tends to zero, because then ds would be zero. Similarly for dt. The symbol ds/dt is always to be considered as single and undecomposable like any other simple algebraic symbol, x for instance. It is no more than a convenient abbreviation for

$$\underset{\delta t \to 0}{lt.} \delta s/\delta t.$$

There are various other ways of writing the differential coefficient which will perhaps be met with later, but this one is universally accepted, and will be used exclusively for the present.

In the present instance, the number ds/dt is the instantaneous velocity of the train at the instant t. It is called instantaneous because there is no finite interval of time for which it remains constant. At a particular instant t_1 its magnitude would be $(b + 2ct_1)$. The magnitude of ds/dt at the instant t_1 can conveniently be written ds/dt_1, or alternatively $(ds/dt)_{t=t_1}$. The first is preferable for compactness, but the second is more explicit.

The differentiation of s with respect to t leads to the equation

$$ds/dt = b + 2ct.$$

This is known as a differential equation, this name being applied to any equation in which differential coefficients appear. By the solving of a differential equation is meant the reversing of the process that has just been described, that is, deriving from the differential equation the ordinary algebraic equation to which it corresponds. Notice that in the present case for given values of b and c the differential coefficient would be the same whatever the magnitude of a in the original equation. There are therefore an infinite number of solutions of the above differential equation, and additional information (generally referred to as a " boundary condition ") is required to make the solution complete and unique for the given case considered. Suppose for instance we are told that

$$dy/dx = px + q,$$

p and q being known numbers, and that y is known to have the value y_0 when x is zero. By analogy with the equation that has just been considered, the solution of the above differential equation is

$$y = (p/2)x^2 + qx + K,$$

where K is any independent constant number. But we are told that the relation between y and x is such that when x is zero y is y_0. Putting these related values in the above equation

$$y_0 = (p/2)0 + q0 + K = K$$

so that the complete solution for y is

$$y = (p/2)x^2 + xq + y_0,$$

being the only expression for y which satisfies both the differential equation and the given " boundary condition ".

This, however, is by way of a digression. It is put in simply to show what is meant by a differential equation and by the solution of a differential equation, and is included at this place because it follows so naturally from the above introduction to the differential coefficient. It is with the latter that we are more concerned at the moment and the next thing to do will be to generalise this important idea.

106. General Definition of "Differential Coefficient"

Given that y is a function of x, the differential coefficient of y with respect to x at any value of x in the neighbourhood of which the function is finite and continuous is defined by

$$\frac{dy}{dx} \text{ or } \frac{d}{dx} f(x) = \underset{h \to 0}{lt.} \frac{f(x+h) - f(x)}{h}$$

The reader should have no difficulty in seeing that this is only the general statement of the process which was illustrated for a particular function in the preceding section.

A further illustration of its application is given in Section 109.

107. Geometrical Interpretation of the Differential Coefficient

The curve shown in Fig. 46 is assumed to represent the variation of $f(x)$ with x, that is, if $OS = x$, $SP = f(x)$. An increase of x from OS to OT produces in $f(x)$ an increase

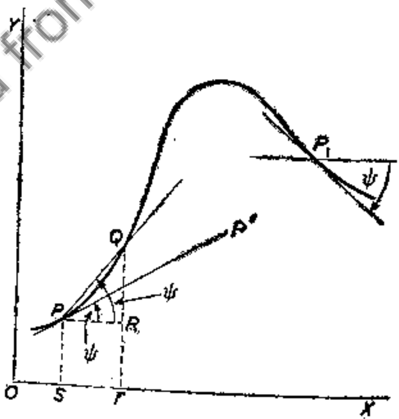


Fig. 46

represented by RQ. Therefore putting h for the magnitude of PR,

$$\frac{f(x + h) - f(x)}{h} = \frac{RQ}{PR} = \tan \psi.$$

As h tends to zero, Q moves down the curve towards P. In the limiting position when Q tends to coincidence with P, the chord PQ becomes what is known as the tangent at P. Therefore if PP' in Fig. 46 represents the tangent at P to the curve which represents the function, then

$$\frac{d f(x)}{dx} = \underset{h \to 0}{lt.} \frac{f(x + h) - f(x)}{h} = \tan \psi.$$

where ψ is the inclination of the tangent at P with respect to the x axis. For this reason the differential coefficient is sometimes referred to as the slope of the function.

108. The Sign of the Differential Coefficient

If at the point x, $f(x)$ is increasing with x (as at the point P in Fig. 46), then

$$f(x + h) - f(x)$$

will be a positive number, and the differential coefficient is therefore positive. Conversely, if the function is decreasing with x, as at the point P_1 in Fig. 46, then the differential coefficient will be negative in sign, corresponding to the fact that the tangent makes a negative angle with the x axis.

109. The Differentiation of Positive Integral Powers of x

The differentiation of $y = x^n$, where n is a positive integer, is a good illustration of the application of the above general definition of the differential coefficient. By the Binomial Theorem,

$$(x + h)^n = x^n + nx^{n-1}h + \frac{n(n-1)}{2} x^{n-2}h^2$$

$$+ \frac{n(n-1)(n-2)}{2.3} x^{n-3}h^3 \text{ etc.},$$

there being $n + 1$ terms in the expansion. Therefore as long as h is finite,

$$\frac{(x + h)^n - x^n}{h} = nx^{n-1} + \frac{n(n-1)}{2} x^{n-2}h$$

$+$ terms containing h^2 and higher powers of h.

By taking h sufficiently small, the right-hand side can be made to differ from nx^{n-1} by less than any finite quantity, however small. Therefore

$$dy/dx = dx^n/dx = \underset{h \to 0}{lt.} \frac{(x + h)^n - x^n}{h} = nx^{n-1}.$$

For instance $dx^2/dx = 2x$; $dx/dx = 1$. Notice that the above determination of dx^n/dx depends on the fact that the limit of the sum of a *finite* number of terms is the sum of the limits of the terms. The reader is cautioned against assuming that the limit of the sum of an *infinite* number of terms is equal to the sum of the limits of the terms. It may or may not be so, and frequently isn't.

If the above brief introduction to the differential calculus is thoroughly understood, the reader will have no difficulty in understanding the subsequent sections of this chapter, which will consist mainly of applications and developments of these few comparatively simple ideas.

110. Rules for Differentiating Combinations of Functions

This section deals with technique rather than principles. As such it is likely to be dullish reading but will repay attention.

(i) *Differentiation of the sum of a number of functions.*

Suppose it is required to differentiate

$$y = 4x^3 + 3x^2 + 2x + 10.$$

The best method is essentially that indicated by Æsop in the fable about the bundle of sticks. The above function can be regarded as a sum of the simpler functions $4x^3$, $3x^2$, etc. Now it is fairly easy to show from the definition of a limit that the limit of the sum of a finite number of terms is equal to the sum of the limits of the separate terms.

From this it follows at once that if u, v, w, etc., are functions of x, and

$$y = u + v + w + \text{etc.},$$

then

$$\frac{dy}{dx} = \frac{du}{dx} + \frac{dv}{dx} + \frac{dw}{dx} + \text{etc.}$$

Applying this to the example given,

$$dy/dx = 12x^2 + 6x + 2.$$

[*Disappearance of the constant term.* Notice that in the above differentiation the 10 disappears. The disappearance of any constant term is inherent in the process of differentiation. This important fact must be borne in mind when the process is reversed. For example, if n is a positive integer, it has been shown that the d.c. of x^n is nx^{n-1}. But so also is the d.c. of $x^n + C$, where C is any constant number whatever. If, therefore, we are told that y is a function of x such that

$$dy/dx = nx^{n-1},$$

then we can only infer that

$$y = x^n + C,$$

where C is some unknown "arbitrary" constant for the determination of which further information is required.]

(ii) *The differentiation of a product of functions*

If $f(x)$ and $g(x)$ are two functions of x, and

$$y = f(x)g(x),$$

then, by definition,

$$\frac{dy}{dx} = \lim_{h \to 0} \frac{f(x+h)g(x+h) - f(x)g(x)}{h}$$

$$= \lim_{h \to 0} \frac{f(x+h)g(x+h) - f(x)g(x+h) + f(x)g(x+h) - f(x)g(x)}{h}$$

$$= g(x)\frac{df(x)}{dx} + f(x)\frac{dg(x)}{dx},$$

or, putting this in a form which is rather easier to remember,

$$\frac{d(uv)}{dx} = u\frac{dv}{dx} + v\frac{du}{dx}.$$

Consider for example the two functions x^n and x^{-n}, n being a positive integer. The product of these functions is 1, and the d.c. of the product is therefore zero, that is,

$$\frac{dx^{-n}x^n}{dx} = x^{-n}\frac{dx^n}{dx} + x^n\frac{dx^{-n}}{dx} = 0,$$

or $$nx^{n-1}x^{-n} + x^n\frac{dx^{-n}}{dx} = 0,$$

therefore $$\frac{dx^{-n}}{dx} = -nx^{n-1}x^{-n}x^{-n}$$

$$= -nx^{-n-1},$$

which shows that the formula for the d.c. of x^n is true for negative indices.

It is easy to extend the above result to products of more than two terms. It will be found that

$$\frac{d(uvw)}{dx} = uv\frac{dw}{dx} + uw\frac{dv}{dx} + vw\frac{du}{dx},$$

and so on.

(iii) *The differentiation of a quotient*

This can be derived from the preceding, just as the idea of a quotient is derived from that of a product.

If $y = u/v$ then $u = vy$.

The right-hand side can be differentiated by the rule for a product, and in this way it is easy to show that

$$\frac{dy}{dx} = \frac{1}{v^2}\left\{v\frac{du}{dx} - u\frac{dv}{dx}\right\}.$$

(iv) *The differentiation of a function of a function*

This sounds tautological, for a function of a function of x is a function of x. However, in a case such as

$$y = 3(x^2 + 2x + 5)^2 + 7(x^2 + 2x + 5) + 8,$$

it will generally be more convenient to treat y as a function of the variable $(x^2 + 2x + 5)$, which variable is itself a function of x. The function $y = \log(\sin x)$ is another instance. The general form is

$$u = f(x),$$

and

$$y = \phi(u) = \phi\{f(x)\}.$$

Suppose x increases to $x + h$, in consequence of which u increases to $u + k$ and y to $y + m$. Then

$$\frac{du}{dx} = \underset{h \to 0}{lt.} \frac{k}{h},$$

and

$$\frac{dy}{du} = \underset{k \to 0}{lt.} \frac{m}{k},$$

Further, since the functions are assumed to be continuous,
k will tend to zero as h tends to zero, so that

$$\frac{dy}{du} = \underset{h \to 0}{lt.} \frac{m}{k}.$$

Therefore

$$\frac{dy}{du}\frac{du}{dx} = \underset{k \to 0}{lt.} \frac{m}{k} \underset{h \to 0}{lt.} \frac{k}{h}.$$

It is easy to show from the definition of limit that the product of the limits of two terms is equal to the limit of the product of the terms, so that

$$\frac{dy}{du}\frac{du}{dx} = \underset{h \to 0}{lt.} \left\{\frac{m}{k}\frac{k}{h}\right\}$$

$$= \underset{h \to 0}{lt.} \frac{m}{h},$$

since the quantities m, k, and h are not zero. But

$$\underset{h \to 0}{lt.} \frac{m}{h} = \frac{dy}{dx},$$

Therefore

$$\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx}.$$

For instance, taking the first example quoted above,

$$y = 3u^2 + 7u + 8,$$

and

$$u = x^2 + 2x + 5,$$
$$dy/du = 6u + 7,$$
$$du/dx = 2x + 2.$$

Therefore

$$dy/dx = \{6(x^2 + 2x + 5) + 7\}(2x + 2)$$
$$= 12x^3 + 36x^2 + 98x + 74.$$

An important special case is that in which y is that function of x which makes

$$y^q = x^p, \text{ that is, } y = x^{p/q}.$$

The differentiation of y^q by the above rule gives

$$q y^{q-1}(dy/dx) = px^{p-1}.$$

By substituting in this the value for y in terms of x and rearranging, it is easy to show that

$$dy/dx = (p/q)x^{(p/q)-1}.$$

We can therefore say that

$$dx^n/dx = nx^{n-1}$$

for all real values of n, positive or negative, integral or fractional.

III. Standard Forms

In order to acquire fluency in the applications of the calculus it is advisable to learn off by heart the differential coefficients of a number of the most common functions, just as one learns off by heart the multiplication tables at an earlier stage of one's mathematical education. With these standard forms and the above rules for dealing with simple combinations of functions the differentiation of any ordinary function is a comparatively simple matter. The more important of these standard forms will now be detailed.

(i) x^n

It has already been shown that

$$dx^n/dx = nx^{n-1}$$

for all values of n.

(ii) ϵ^x

On account of its great importance, this case will be given in full. If $y = \epsilon^x$ then, by definition,

$$\frac{dy}{dx} = \underset{h \to 0}{lt.} \frac{\epsilon^{x+h} - \epsilon^x}{h}$$

$$= \underset{h \to 0}{lt.} \epsilon^x \frac{\epsilon^h - 1}{h}$$

$$= \epsilon^x \underset{h \to 0}{lt.} \frac{\epsilon^h - 1}{h}.$$

Now

$$\frac{\epsilon^h - 1}{h} = 1 + \frac{h}{2}\left\{1 + \frac{h}{3} + \frac{h^2}{3 \cdot 4} + \frac{h^3}{3 \cdot 4 \cdot 5} + \text{etc., } ad\ inf.\right\}.$$

For values of h less than 1 the sum to infinity of the series in the brackets is less than

$$1 + h + h^2 + h^3 + h^4 + \text{etc., } ad\ inf. = 1/(1 - h)$$

(by Section 66). Therefore the series in the brackets can be put equal to $K/(1 - h)$, where K is less than 1 as long as h is less than 1. Therefore

$$\underset{h \to 0}{lt.} \frac{\epsilon^h - 1}{h} = \underset{h \to 0}{lt.} \left\{1 + \frac{Kh}{2(1 - h)}\right\}$$

$$= 1,$$

so that $$dy/dx = d\epsilon^x/dx = \epsilon^x = y$$

(It might appear that this could be proved more simply by differentiating term by term the series for ϵ^x. The series so obtained, however, is not necessarily equal to the differential coefficient of the sum of the original series, for the sum of the limits of an infinite number of terms is not necessarily equal to the limit of the sum. It generally is, but it quite often is not, and the assumption may never be made without question.)

Note that if

$$y = a\epsilon^{mx} = a(\epsilon^x)^m,$$

then by the rule for the differentiation of a function of a function

$$dy/dx = am(\epsilon^x)^{m-1}\epsilon^x = am\epsilon^{mx} = my.$$

Thus the function $a\epsilon^{mx}$ has the remarkable property that its rate of change is proportional to itself. Further, it is the only known function which has this property. In other words the most general solution of the differential equation

$$dy/dx = my$$

is

$$y = a\epsilon^{mx},$$

where a is an arbitrary constant number for the determination of which further information is required. This is the reason why the curious and rather awkward number $\epsilon = 2.71828$ . . . is always turning up in applied mathematics and physics.

The rather more general case

$$y = a^x$$

can be derived at once by writing the constant a in the form ϵ^m, that is, m is $\log_\epsilon a$. Then

$$y = a^x = (\epsilon^m)^x = \epsilon^{mx}$$

and

$$dy/dx = m\epsilon^{mx} = m(\epsilon^m)^x = a^x \log_\epsilon a.$$

(iii) $\text{Log}_\epsilon\, x$

This case can be derived from the preceding, for if

$$y = \log_\epsilon x,$$
$$x = \epsilon^y,$$

and the differentiation of both sides (the right-hand side being a function of a function of x) gives

$$1 = \epsilon^y\,(dy/dx),$$

so that

$$dy/dx = d \log x/dx = 1/\epsilon^y = 1/x.$$

An obvious extension is

$$\frac{d \log f(x)}{dx} = \frac{1}{f(x)}\,\frac{d f(x)}{dx}$$

(iv) Sin x

The reader should have no difficulty in showing from the trigonometrical formulæ derived in Section 95 that

$$\sin A - \sin B = 2 \cos \frac{(A+B)}{2} \sin \frac{(A-B)}{2},$$

so that

$$\sin (x+h) - \sin x = 2 \cos \left(x + \frac{h}{2}\right) \sin h/2.$$

Therefore, if $y = \sin x$,

$$\frac{dy}{dx} = \underset{h \to 0}{lt.} \cos \left(x + \frac{h}{2}\right) \frac{\sin (h/2)}{h/2}$$

$$= \cos x,$$

for, as is easily shown from the series for $\sin \theta$, the limit of $(\sin \theta)/\theta$ when θ tends to zero is 1.

In a precisely similar manner it can be shown that

$$d (\cos x)/dx = - \sin x.$$

Notice that

$$d (\sin mx)/dx = m \cos mx,$$

by the rule for the differentiation of a function of a function.

The other trigonometrical functions are derived from these two, the sine and the cosine, and the differential coefficients can therefore be calculated from the above rules relating to combinations of functions. Space will not permit of their being detailed individually, but they are listed below for reference.

Function.	Differential Coefficient.
$\tan x$	$\sec^2 x$
$\cot x$	$-\operatorname{cosec}^2 x$
$\sec x$	$\sec x \tan x$
$\operatorname{cosec} x$	$-\operatorname{cosec} x \cot x$

So much for what may be called the ABC of the calculus. Not enough, perhaps, some may think. It certainly is rather concentrated, but the essence of the matter is there. Familiarity with the ideas involved can only be had by practice, and then more practice, and then some more. A few examples are given, but many more need to be worked by a beginner. A good plan is to express a function in two ways and differentiate each form. Work can be made more or less self checking in this way. (Examples : $(a + x)^3$ and $a^3 + 3a^2x + 3ax^2 + x^3$; $\tan 2x$ and $(2 \sin x \cos x)/(\cos^2 x - \sin^2 x)$; and so on.)

112. Successive Differentiation

This does not introduce any new ideas, but only some more " shorthand ". If y is a function of x, then in general dy/dx will also be a function of x, and as such can be differentiated with respect to x, giving

$$\frac{d}{dx}\left(\frac{dy}{dx}\right).$$

Since this is rather cumbersome to write, it is abbreviated to

$$d^2y/dx^2,$$

the d's and dx's being, so to speak, multiplied together as if they were numbers (which of course they are not. It is merely a convenience of notation). The process can obviously be extended, giving d^3y/dx^3, d^4y/dx^4, etc., d^ny/dx^n being referred to as the nth differential coefficient of y with respect to x, or sometimes as the nth derivative. As an example, if

$$y = ax^3,$$
$$dy/dx = 3ax^2,$$
$$d^2y/dx^2 = 6ax,$$
$$d^3y/dx^3 = 6a,$$

and
$$d^4y/dx^4 = 0,$$

so that the process terminates. On the other hand a function such as $\sin x$ can be differentiated for ever. In this matter the trigonometric functions have a peculiar property which can be illustrated by

$$\begin{aligned} y &= a \sin mx + b \cos mx \\ dy/dx &= am \cos mx - bm \sin mx \\ d^2y/dx^2 &= -am^2 \sin mx - bm^2 \cos mx \\ &= -m^2y. \end{aligned}$$

No other function can be found which has this property that the second differential coefficient is equal to the function multiplied by a negative number. In other words the most general solution of the differential equation

$$d^2y/dx^2 = -m^2y$$

is
$$y = a \sin mx + b \cos mx,$$
where a and b are constant numbers which can only be determined by additional information. Suppose, for instance, we are given that
$$d^2y/dx^2 = -169\,y \text{ (that is, } -13^2 y),$$
$$y = 10 \text{ when } x = 0,$$
$$dy/dx = 26 \text{ when } x = 0.$$
Then the general solution is
$$y = a \sin 13x + b \cos 13x,$$
so that
$$dy/dx = 13a \cos 13x - 13b \sin 13x.$$
Therefore when x is 0 we have
$$(y)_0 = 10 = b,$$
$$(dy/dx)_0 = 26 = 13a, \text{ that is, } a = 2,$$
giving as the complete particular solution
$$y = 2 \sin 13x + 10 \cos 13x.$$

113. Partial Differentiation

Here again there is no new idea but only additional notation. As already pointed out, two or more independently variable numbers can be combined in a variety of ways to give another number. For instance, if x and y are independent variables,
$$z = ax^2 + bxy + cx^2$$
is a function of the two variables x and y. Such a function could be represented geometrically by taking x and y as rectangular co-ordinates and z as the vertical height above the x, y plane at the point x, y. The equation would then define a surface. Now in general, the rate of change of z, that is, the slope of the surface, will depend on the direction in which it is measured, a fact which is demonstrated afresh every time the driver of a waggon zig-zags up a hill which is too steep for his horses. The term "differential coefficient" is therefore indefinite unless the direction is specified in some way. There are two directions which naturally suggest themselves, and which are generally of more interest than any others—the directions

of the x and y axes. Moving in the direction of the x axis means that y is kept constant (if this is not immediately obvious, the reader should draw the axes. Then it will be). As long as y is kept constant, z is in effect a function of the single variable x and has a differential coefficient with respect to x, that is, a slope in the direction of the x axis. That is what is meant by dz/dx in such a case, only it is written $\partial z/\partial x$ in order to distinguish it from the case in which z is a function of x only, in the full ordinary sense of that phrase. It is called the partial differential coefficient of z with respect to x. Similarly for $\partial z/\partial y$. For instance, in the above case, that is,

$$z = ax^2 + bxy + cy^2,$$

$$\partial z/\partial x = 2ax + by,$$

since cy^2 is by definition a constant as far as this rate of change is concerned. Similarly,

$$\partial z/\partial y = bx + 2cy.$$

Both these partial differential coefficients will, in general, be functions of x and of y, as they are in the above case, and will therefore themselves have further partial differential coefficients, defined in the same way. Thus

$$\frac{\partial}{\partial x}\left(\frac{\partial z}{\partial x}\right),$$

which is written for shortness $\partial^2 z/\partial x^2$, is $2a$, and

$$\frac{\partial}{\partial y}\left(\frac{\partial z}{\partial x}\right),$$

which is abbreviated to $\partial^2 z/\partial y \partial x$, is b. Notice that $\partial^2 z/\partial x \partial y$ and $\partial^2 z/\partial y \partial x$ have different meanings as defined above. It can be shown, however, that if they both exist they will be equal, as they are in the present instance. The proof is rather beyond the scope of this work.

The reader should have no difficulty in showing that if z is a function of a function of x and y, that is,

$$z = f(u)$$

where $$u = \phi(x, y),$$

then $$\frac{\partial z}{\partial x} = \frac{df(u)}{du} \frac{\partial u}{\partial x},$$

and $$\frac{\partial z}{\partial y} = \frac{df(u)}{du} \frac{\partial u}{\partial y}.$$

The proof follows exactly the same lines as for the rule for the differentiation of a function of a function, and is omitted to save space. Consider, for instance, the anode current of a triode valve, which, using the usual symbols, can be expressed in the form

$$i_a = f(v_a + \mu v_g + a),$$

where a is a constant. The quantities v_a and v_g are independent variables, and have probably been varied independently by readers of this book on many occasions. The relation can be put in the form

$$i_a = f(V), \text{ where } V = v_a + \mu v_g + a,$$

which gives i_a as a function of a function of the two variables. The slope of the anode-current—grid-voltage characteristic is

$$\frac{\partial i_a}{\partial v_g} = \frac{df(V)}{dV} \frac{\partial V}{\partial v_g} = \mu \frac{df(V)}{dV},$$

and of the anode-current—anode-voltage characteristic

$$\frac{\partial i_a}{\partial v_a} = \frac{df(V)}{dV} \frac{\partial V}{\partial v_a} = \frac{df(V)}{dV},$$

so that $$\frac{\partial i_a}{\partial v_g} = \mu \frac{\partial i_a}{\partial v_a}.$$

Notice that $df(V)/dV$ is the slope of what is sometimes called the " lumped volts " characteristic.

114. Critical Values

Given a circuit or some other combination of apparatus the performance of which depends on, or, in other words, is a function of, some variable condition of operation, it is generally desirable and sometimes very important to know what condition of operation will give the best performance. Suppose, for instance, that a battery is to supply electrical

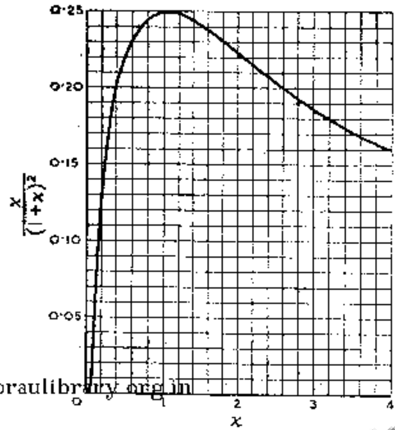


Fig. 47

power to some variable load resistance. What will be the magnitude of the resistance which will absorb the maximum power from a battery of given characteristics? Such problems are of frequent occurrence in applied electricity, and the technique of the differential calculus finds one of its most valuable applications in the solution of such problems.

First, let us examine a little more closely the example quoted above in order to get a clearer idea of the nature of the problem. If the "open circuit" e.m.f. of the battery is e volts, its internal resistance R_0 ohms, and the resistance of the load R ohms, then the current will be,

$$i = e/(R_0 + R)$$

amperes, by Ohm's Law. The power absorbed by the load will be

$$p = i^2R = \frac{R}{(R + R_0)^2}e^2,$$

and is thus a function of R for given constant values of e and R_0. It is therefore also a function of the ratio R/R_0, and for the present purpose is more conveniently expressed in terms of this ratio. Putting x for R/R_0,

$$p = \frac{x}{(1 + x)^2} \quad \frac{e^2}{R_0},$$

and the variation of the power p with R is seen to be essentially the variation with x of the function $x/(1 + x)^2$.

Calling this function y, the variation of y with x is shown in Fig. 47. This is the simplest form of "efficiency curve", and in most practical cases the variation of electrical efficiency will be of this character. It appears from the diagram that y reaches a pronounced maximum value of $\frac{1}{4}$ when x is 1, that is, when $R = R_0$. Therefore in this particular case the maximum power obtainable from the battery is $e^2/4R_0$, and the "optimum" load corresponding to this output is a load equal to the internal resistance of the battery.

Now let us see how this same conclusion could be reached without the trouble of drawing the curve of the function, by applying the technique of the differential calculus.

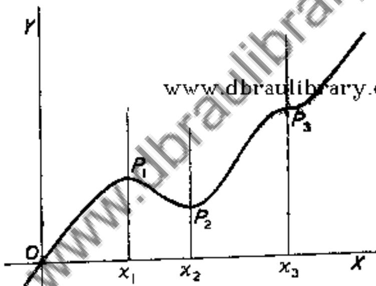


Fig. 48

In Fig. 48 the point P_1 is assumed to correspond to the maximum value reached by y in the range 0 to x_1 of x, y being a continuous function of x in the range illustrated in the diagram. Up to P_1, y increases with x so that dy/dx is positive. From P_1 to P_2, y decreases with x so that dy/dx is negative. The point P_1 therefore separates positive and negative values of dy/dx. It will be assumed that the variation of dy/dx is continuous. Then the point of separation of positive and negative values is the value

$$dy/dx = 0$$

(that is, the tangent at P_1 is parallel to the x axis). Notice that from 0 to x_2, dy/dx decreases continuously. This means that $d(dy/dx)/dx$, or d^2y/dx^2 is negative throughout this range and therefore negative at P_1. Thus for the maximum value at P_1 (corresponding to the value x_1 of x),

$$dy/dx = 0,$$

and $$d^2y/dx^2 < 0 \text{ (that is, is negative),}$$

or, another way of writing it,

$$dy/dx_1 = 0,$$

and $$d^2y/dx_1^2 < 0.$$

In a precisely similar manner, the point P_2 will be a minimum if

$$dy/dx_2 = 0,$$

and $$d^2y/dx_2^2 > 0 \text{ (that is, is positive).}$$

It is important to notice that the condition $dy/dx = 0$ alone will not necessarily determine a maximum or minimum value of y. At P_3, for instance, the tangent is parallel to the axis of x, so that $dy/dx = 0$. However, dy/dx is positive on both sides of P_3. Therefore zero is a minimum value of dy/dx, and d^2y/dx^2 passes through zero at P_3. This is known as a point of inflection.

The example already considered will serve as an illustration of a maximum value.

If $$y = x/(1 + x)^2,$$

$$dy/dx = \{(1 + x)^2 - 2x(1 + x)\}/(1 + x)^4,$$

which reduces to

$$dy/dx = (1 - x)/(1 + x)^3,$$

so that

$$dy/dx = 0 \text{ when } x = 1.$$

Also it is easy to show that

$$d^2y/dx^2 = -2(2 - x)/(1 + x)^4,$$

which is negative when $x = 1$. Therefore y passes through a *maximum* value ($\frac{1}{4}$) when $x = 1$.

The following is a very useful practical point in connection with maximum and minimum values. It frequently happens that the quantity to be investigated can be regarded as a function of a function, for example,

$$y = \phi(u),$$

where

$$u = f(x).$$

The critical condition is then

$$\frac{dy}{dx} = \frac{d\phi(u)}{du} \cdot \frac{du}{dx} = 0,$$

so that either

$$\frac{d\phi(u)}{du} = 0$$

or

$$\frac{du}{dx} = 0,$$

and in the majority of cases it will be the second condition that counts. Take, for instance, the case

$$y = 1/\sqrt{R^2 + (\omega L - 1/\omega C)^2},$$

where R, L, and ω are constant numbers, and C is variable. This is in effect

$$y = 1/\sqrt{u},$$

where

$$u = R^2 + (\omega L - 1/\omega C)^2.$$

There is no critical condition for dy/du, but

$$du/dC = 2(\omega L - 1/\omega C)/\omega C^2,$$

which will be zero when $\omega L = 1/\omega C$. There is no need to write out the whole differentiation in full in this or similar cases.

This is the straightforward or rule of thumb solution. But a person who works by his head and not by rule of thumb could go straight to an even better solution without knowing anything about the calculus. Putting

$$x = \omega L - 1/\omega C,$$

then if C varies, x varies too. We want to know when

$$\frac{1}{\sqrt{R^2 + x^2}}$$

is a maximum if x is allowed to vary. This will be a maximum when $\sqrt{R^2 + x^2}$ is a minimum and $\sqrt{R^2 + x^2}$ will be a minimum when $R^2 + x^2$ is a minimum. But x^2 can never be negative. Therefore its minimum value is 0. Thus the minimum of $\sqrt{R^2 + x^2}$ occurs when $x = 0$, and this gives a maximum value to $1/\sqrt{R^2 + x^2}$. This maximum value will thus occur when C or L or ω are given such values that

$$\omega L - 1/\omega C = 0.$$

Examples XIV

1. Find the first, second, and third derivatives of

   i. $ax^2 + bx + c$.

   ii. $a + b/x + c/x^2$.

   iii. $30 \sin x + 15 \sin 2x + 10 \sin 3x$.

   iv. $\epsilon^{ax} \sin bx$.

   v. a^x.

   vi. $a^x \log (\sin x)$.

2. Solve the equation

$$dQ/dt = -Q/CR$$

where C and R are numbers, given that Q is 10 when t is 0.

3. Solve the equation

$$d^2i/dt^2 = -25\ m^2i,$$

where m is a number, given that i is 0 when t is 0, and that i is 10 when t is $\pi/2m$.

4. Find

$$\partial z/\partial x,\ \partial z/\partial y,\ \partial^2 z/\partial x^2,\ \partial^2 z/\partial y^2,\ \partial^2 z/\partial x \partial y,\ \partial^2 z/\partial y \partial x$$

for

$$z = ax^2 + bxy + cy^2$$

and

$$z = \epsilon^{ax + by} \sin xy.$$

5. Discuss the critical values of

   i. $x/(1 + 2ax + x^2)$.

   ii. $x/(1 + x^2)$.

6. The instantaneous rate of motion of a particle moving in a straight line is given as $50 + 100t$ cm/sec. How far will it travel in an hour from its position when $t = 0$?

7. The distance, in cm from a fixed point, travelled by a particle moving in a straight line is given by $10 + 500t - 5t^2$, where t is in seconds. At what distance from the starting point will it come to rest?

How long will it take to return to its starting point from the point at which it comes to rest? (Note: starting point—not fixed point.)

8. If $xy = c =$ constant, find the critical value of $x + y$. Is it a minimum or a maximum?

9. If
$$\frac{d^2y}{dx^2} = k^2y,$$
show that either
$$y = A\epsilon^{kx} + B\epsilon^{-kx}$$
or
$$y = C \sinh kx + D \cosh kx$$
will satisfy the equation. Express C and D in terms of A and B.

115. Vector Functions and the Differentiation of Vectors

Now we come to a section which links up directly with alternating current phenomena, and thus with wireless telegraphy.

Let **v** be a vector of magnitude v and direction θ relative to the fixed unit vector of reference ν parallel to the bottom edge of the paper, that is,

$$\mathbf{v} \cdot \nu = v \cos \theta.$$

Now either or both of v and θ may depend in some specified manner on some independent variable, t for instance (time), being functions of the independent variable in the ordinary sense of the word. Thus if v is $f(t)$ and $\theta = \phi(t)$,

$$\mathbf{v} \cdot \nu = f(t) \cos \phi(t),$$

and this equation completely defines the vector. Two important special cases are

$$\mathbf{v} \cdot \nu = v_0 \cos \omega t,$$

and

$$\mathbf{v} \cdot \nu = v_0 \epsilon^{-kt} \cos \omega t.$$

In the first case the vector is constant in magnitude and rotates with constant angular velocity (ω radians per second), and in the second case the magnitude of the vector decreases exponentially while it rotates with constant angular velocity ω. These vectors are illustrated in Figs. 49 and 50. The locus of the end of the first vector is a circle and that of the second an equiangular spiral. The reason for the latter name will appear later. It has already been pointed out that a vector of the first type can be used to represent an alternating current or potential

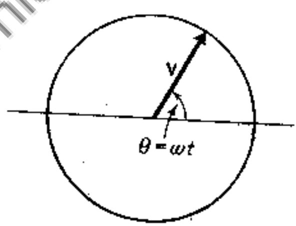


Fig. 49

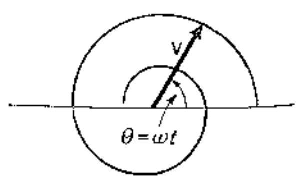


Fig. 50

difference. Similarly a vector of the second type can be used to represent what is known as a " damped oscillation ", the word " damped " being used in the sense of " decreasing "—presumably by derivation from the effect of water on a fire.

Leaving these special cases for a moment, consider the perfectly general case illustrated in Fig. 51, where the variation of the vector is such that its end point moves along the dotted line. Let OV and OV' represent the vector at the instants t and $t + \delta t$. If $\delta\mathbf{v}$ be the change in $\mathbf{v}$ in the interval δt, then OV' is the vector $\mathbf{v} + \delta\mathbf{v}$, whence it follows that VV' represents the vector $\delta\mathbf{v}$. Now the differential coefficient of $\mathbf{v}$ with respect to t is defined in exactly the same way as in the corresponding case of a scalar function, that is,

$$d\mathbf{v}/dt = \underset{\delta t \to 0}{lt.} \delta\mathbf{v}/\delta t.$$

The first thing to notice is that the d.c. of a vector is a vector, since $\delta\mathbf{v}$ is a vector. It therefore has both magnitude and direction. Further, VV' is a chord of the locus, and it is easy to see that in the limit when t tends to zero this chord will coincide in direction with the tangent to the locus at the point V. The direction of the vector $d\mathbf{v}/dt$ is thus the direction of the tangent to the locus of $\mathbf{v}$ at the instant t.

For the complete specification of $d\mathbf{v}/dt$, that is, for the determination of the scalar product $(d\mathbf{v}/dt) \cdot v$, it is necessary to know both its magnitude and direction. These will obviously depend on the nature of the time variation of the magnitude and direction of the vector $\mathbf{v}$, and the vector $d\mathbf{v}/dt$ can be expressed very simply in terms of these two separate variations. On OV' mark the point P such that $OP = OV$ in magnitude. Then PV' represents the change in the magnitude of $\mathbf{v}$.

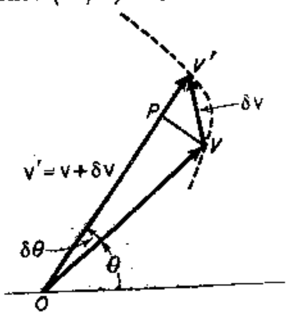


Fig. 51

Further, the angle POV or $\delta\theta$ represents the change in the direction of $\mathbf{v}$. The vector $\delta\mathbf{v}$ or VV' is the sum of the vectors VP and PV'. Let $\mathbf{v}_1$ be the unit vector in the direction of $\mathbf{v}$, that is,

$$\mathbf{v} = v\mathbf{v}_1 \text{ or } \mathbf{v}_1 = \mathbf{v}/v.$$

The magnitude of VP is approximately $v\delta\theta$ and its direction is approximately perpendicular to $\mathbf{v}$. The unit vector perpendicular to $\mathbf{v}$ is $j\mathbf{v}_1$, or $j\mathbf{v}/v$. As a vector, therefore, VP can be written approximately

$$VP = v\delta\theta\, j\mathbf{v}/v.$$

Again the magnitude of the vector PV' is δv, and if $\delta\theta$ is very small its direction will be approximately that of $\mathbf{v}$. As a vector, therefore, it is approximately $\delta v\mathbf{v}/v$. We have thus the approximate equation

$$\delta\mathbf{v} = \delta v(\mathbf{v}/v) + jv\delta\theta(\mathbf{v}/v),$$

and dividing through by δt

$$\frac{\delta\mathbf{v}}{\delta t} = \frac{\delta v}{\delta t}\frac{\mathbf{v}}{v} + jv\frac{\delta\theta}{\delta t}\frac{\mathbf{v}}{v}$$

$$= \left(\frac{1}{v}\frac{\delta v}{\delta t} + j\frac{\delta\theta}{\delta t}\right)\mathbf{v}.$$

So far this is an approximation only. But notice that all the approximations are such that the statements become more and more correct as δt decreases in magnitude, and become exact in the limit when δt tends to zero. All the statements could be made exact with vanishing differences, but this rigid demonstration would take up rather a lot of valuable space. In the limit when δt tends to zero we have

$$\frac{d\mathbf{v}}{dt} = \left(\frac{1}{v}\frac{dv}{dt} + j\frac{d\theta}{dt}\right)\mathbf{v},$$

which determines $d\mathbf{v}/dt$ completely if dv/dt and $d\theta/dt$ are known. In general, $d\mathbf{v}/dt$ is thus expressible in the form $(a + jb)\mathbf{v}$, where a and b are known functions of t.

The expression assumes a very simple form in the two special cases mentioned above. In the first, since the vector is constant in magnitude, dv/dt is zero. Also, since $\theta = \omega t$, $d\theta/dt = \omega$ and is a constant. Therefore for a vector of constant magnitude rotating with constant angular velocity ω,

$$d\mathbf{v}/dt = j\omega\mathbf{v},$$

a vector perpendicular to $\mathbf{v}$ and ω times as large. Geometrically, this expresses the fact that the tangent to a circle is perpendicular to the radius. Notice that

$$\frac{d^2\mathbf{v}}{dt^2} = j\omega\,\frac{d\mathbf{v}}{dt} = (j\omega)^2\mathbf{v} = -\,\omega^2\mathbf{v},$$

and, in general,

$$d^n\mathbf{v}/dt^n = (j\omega)^n\mathbf{v}.$$

For the second case, that of a vector of exponentially decreasing magnitude rotating with constant angular velocity, $d\theta/dt$ is ω as before,

and since $\quad v = v_0\epsilon^{-kt},$

$$dv/dt = -\,kv_0\epsilon^{-kt} = -\,kv,$$

so that $\quad \dfrac{1}{v}\dfrac{dv}{dt} = -\,k.$

Therefore $\quad \dfrac{d\mathbf{v}}{dt} = (-\,k + \omega j)\mathbf{v},$

and, in general, $\quad \dfrac{d^n\mathbf{v}}{dt^n} = (-\,k + \omega j)^n\mathbf{v}.$

(Notice that since $d\mathbf{v}/dt = (-\,k + \omega j)\mathbf{v}$ the tangent to the locus at $\mathbf{v}$ makes with $\mathbf{v}$ a constant angle $\tan^{-1} -\omega/k$. This is the reason for the name " equiangular spiral " given to this locus.)

These two special cases have very important applications to alternating current theory, and later sections will be devoted to the development of these applications. The matter must be left for the present in favour of a brief account of the companion subject of the differential calculus, that is, the integral calculus.

116. The Integral Calculus: Indefinite Integration

It is rather unfortunate that the word " integration " is used in two different senses, but this will not matter very much as long as the two ideas are clearly distinguished right from the start. We shall take the simpler of the two first, generally called for the sake of distinction "indefinite integration", though there is in fact nothing really indefinite about it. Integration in this sense is simply the inverse of differentiation. The integral with respect to x of any given function of x is the most general function of x of which it is the differential coefficient. The integral of $f(x)$ with respect to x is written

$$\int f(x)\,dx,^{*}$$

and its definition is

$$\frac{d}{dx}\left\{\int f(x)\,dx\right\} = f(x).$$

Any function of x which fulfils this definition can be called *an* integral of $f(x)$, but *the* integral will be taken to mean the most general function which fulfils the definition. For instance $x^{n+1}/(n+1)$ is *an* integral of x^n, for

$$\frac{d}{dx}\left(\frac{x^{n+1}}{n+1}\right) = x^n,$$

but the most general function which satisfies the condition is

$$\frac{x^{n+1}}{n+1} + C,$$

where C is any constant number whatever, so that this will be regarded as *the* integral of x^n. In general, any two integrals of the same function can only differ by a constant, in view of the definition given above, and the simplest form, together with an arbitrary constant, will be taken as the integral.

So far so good. It all seems plain sailing. Any table of differential coefficients will immediately furnish an equal number of integrals. For instance, since

* At this point the reader may want to raise an agitation against this apparent violation of the integrity of dy/dx, which he has been told to regard as an inseparable unity. But the "dx" under the integral sign is not, in fact, the dx from dy/dx trying to lead a separate existence. It is merely an agreed shorthand for "with respect to x". Certain textbooks use a notation which implies the separate existence of dy and dx, but this is liable to be misleading, and the writer has carefully avoided it.

$$\frac{d\epsilon^x}{dx} = \epsilon^x, \qquad \int \epsilon^x dx = \epsilon^x + C.$$

Again, $$d \log_\epsilon x/dx = 1/x,$$

whence $$\int \frac{1}{x}\, dx \text{ or } \int \frac{dx}{x} = \log_\epsilon x + C,$$

and so on for all the standard forms of differential coefficient which have already been discussed. Space cannot be spared for the enumeration of them, but the reader is advised to make himself familiar with the more important standard forms.

Outside the comparatively few standard forms, however, the difficulties begin. Differentiation is a comparatively simple matter. There is the fundamental formula to start with, and rules for combinations of functions to simplify its application. For the inverse process, however, there is practically speaking no guide at all and no such rules for dealing with combinations, no rules, that is to say, which will inevitably succeed. What then is there to help? Only inspired guesswork. That, of course, lends a certain fascination to the business, but its practical limitations need hardly be pointed out.

However, there are one or two general propositions which may simplify matters a little, and these we shall briefly pass in review.

117. A Constant Factor

It is easy to show that a constant factor can be placed outside the sign of integration, that is,

$$\int a\, f(x)\, dx = a \int f(x)\, dx,$$

for

$$\frac{d}{dx}\int a\, f(x)\, dx = a\, f(x)$$

by definition, and by the rules of differentiation

$$\frac{d}{dx}\, a \int f(x)\, dx = a\, \frac{d}{dx}\int f(x)\, dx$$
$$= a\, f(x).$$

118. The Integration of the Sum or Difference of Functions

This is comparable with the corresponding proposition in differentiation, and can be immediately derived therefrom. If P and Q are functions of x, then by definition

$$d/dx\int(P \pm Q)\,dx = P \pm Q,$$

also, by the rules of differentiation,

$$\frac{d}{dx}\left\{\int P\,dx \pm \int Q\,dx\right\} = \frac{d}{dx}\int P\,dx \pm \frac{d}{dx}\int Q\,dx$$
$$= P \pm Q,$$

therefore

$$\int(P \pm Q)\,dx = \int P\,dx \pm \int Q\,dx.$$

The proposition can obviously be extended to the sum or difference of any *finite* number of functions. As an example,

$$\frac{1}{x^2 - a^2} = \frac{1}{2a}\left\{\frac{1}{x - a} - \frac{1}{x + a}\right\}$$

by elementary algebra. Therefore

$$\int\frac{dx}{x^2 - a^2} = \frac{1}{2a}\left\{\int\frac{dx}{x - a} - \int\frac{dx}{x + a}\right\}$$
$$= \frac{1}{2a}\{\log_e(x - a) - \log_e(x + a)\} + C$$
$$= \frac{1}{2a}\log_e\frac{x - a}{x + a} + C,$$

C being an "arbitrary constant" of integration. This example also illustrates the application of the bundle-of-sticks idea to integration. Where possible, a complicated function should be separated out into the sum or difference of a number of simpler functions.

119. Changing the Variable

(i) Suppose $f(x)$ be expressible in the form $\phi(u)(du/dx)$, where u is some other function of x.

For example, $\sec^2 x/(a^2 - b^2\tan^2 x)$. Let $u = b \tan x$.

Then $du/dx = b \sec^2 x$, as already shown, so that

$$\frac{\sec^2 x}{a^2 - b^2 \tan x} = \frac{1}{b}\,\frac{1}{a^2 - u^2}\,\frac{du}{dx}.$$

Now it is easy to show that

$$\int \phi(u)\,\frac{du}{dx}\,dx = \int \phi(u)\,du,$$

for the R.H.S., which we shall call F for short, is a function of a function of x, so that

$$\frac{dF}{dx} = \frac{dF}{du}\,\frac{du}{dx}$$

Also, by the definition of integration,

$$dF/du = \phi(u).$$

Therefore

$$dF/dx = \phi(u)(du/dx) = f(x),$$

whence, by definition,

$$F = \int f(x)\,dx.$$

For the above example,

$$\int \frac{\sec^2 x\,dx}{a^2 - b^2 \tan^2 x} = \int \frac{1}{b}\,\frac{1}{a^2 - u^2}\,\frac{du}{dx}\,dx$$

$$= \int \frac{1}{b}\,\frac{1}{a^2 - u^2}\,du$$

$$= \frac{1}{2ab}\log_e \frac{a-u}{a+u} + C \quad \text{(See Sec. 118)}$$

$$= \frac{1}{2ab}\log_e \frac{a - b\tan x}{a + b\tan x} + C.$$

Thus, inspired guesswork is only required to furnish the substitution $u = b \tan x$, and thereafter all is plain sailing. This is characteristic of a large number of processes of integration.

(ii) The substitution of a single letter for a group, as in the above, seems a reasonable method of simplification. In some cases, however, the reverse process can be used with advantage, that is, the substitution of some simple group for the variable x. Thus in $\int f(x)\,dx$, if we put

$x = \phi(u)$, then $f(x)$ becomes a function of the variable u, say $F(u)$, and the integral becomes $\int F(u)\,dx$. Now it can be shown, much as in the previous case, that

$$\int F(u)\,dx = \int F(u)\frac{dx}{du}\,du,$$

and this form will be much simpler than the original if the substitution has been well chosen. For example, in $f(x) = 1/\sqrt{a^2 - x^2}$ let $x = a \sin u$. Then

$$\frac{1}{\sqrt{a^2 - x^2}} = \frac{1}{\sqrt{a^2 - a^2 \sin^2 u}} = \frac{1}{a \cos u}.$$

Also $dx/du = a \cos u$, therefore

$$\int \frac{dx}{\sqrt{a^2 - x^2}} = \int \frac{1}{a \cos u}\, a \cos u\, du = \int du$$

$$= u + C = \sin^{-1} x/a + C.$$

Trigonometrical substitutions of this kind will nearly always afford a simplification in binomial surd functions such as

$$\sqrt{a^2 - x^2},\ \sqrt{a + x}/\sqrt{a - x},$$

and so on.

120. Integration by Parts

Another very useful dodge is derived from the differential formula

$$\frac{d(uv)}{dx} = u\frac{dv}{dx} + v\frac{du}{dx}.$$

It applies to cases in which the function to be integrated can be put in the form

$$I = \int f(x)\,dx = \int P.R\,dx,$$

P and R being functions of x, one of which at least, say R, is easily integrable. Suppose

$$\int R\,dx = Q, \text{ that is, } R = dQ/dx.$$

Then

$$I = \int P\frac{dQ}{dx}\,dx.$$

Now it can be shown that

$$\int P \frac{dQ}{dx} dx = PQ - \int Q \frac{dP}{dx} dx,$$

for, differentiating this equation, and remembering the definition of an integral,

$$P \frac{dQ}{dx} = \frac{dPQ}{dx} - Q \frac{dP}{dx},$$

which is the "differential of a product" formula already quoted above. As an example,

$$\int x \cos x \, dx = \int x(d \sin x/dx) \, dx$$

$$= x \sin x - \int \sin x \, dx$$

$$= x \sin x + \cos x + C.$$

Or, again,

$$\int x\epsilon^x \, dx = \int x(d\epsilon^x/dx) \, dx = x\epsilon^x - \int \epsilon^x dx$$

$$= x\epsilon^x - \epsilon^x + C.$$

So much for a brief outline of the subject of "indefinite" integration. The above formulæ are practically all one has to go on. The rest is inspired guesswork of an intuitive kind, but fortunately the intuitive faculty required increases with practice and experience. A few examples are given at the end of this section, but far more should be worked by any serious student, for integration, like genius, is nine parts perspiration to one of inspiration. Examples are easily made up, and the work can be made self-checking by differentiation of the result.

121. Definite Integration

Now we come to the more practically important of the two ideas associated with the word "integration". What we are concerned with now is the evaluation of expressions such as

$$\underset{\delta x \to 0}{lt.} \{f(a) + f(a + \delta x) + f(a + 2\delta x) + \text{etc.} \; . \; . \; . \; .$$

$$f(b - \delta x) + f(b)\} \, \delta x = \underset{\delta x \to 0}{lt.} \sum_{x=a}^{x=b} f(x)\delta x,$$

that is, the limit when δx tends to zero of the sum of all terms such as $f(x)$ δx when x increases by steps of δx from a lower value a to an upper value b.

But first, readers will probably want to know how such cumbersome-looking expressions come into practical politics at all. Let Fig. 52 represent $f(x)$ plotted against x for the range a to b of x. It will be assumed that there is no minimum or maximum value of $f(x)$ in this range, that is, $f(x)$ either decreases or increases uniformly from a to b. If $f(x)$ is not in fact of this character the range can be divided up into subranges in each of which the limitation applies, and the following discussion can then be applied to each of these separately. Suppose we require to calculate the area included between the ordinates at a and b and the curved line representing $f(x)$.

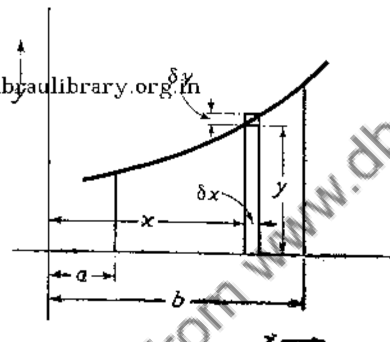


Fig. 52

One method that suggests itself is to divide up the area into n strips each of width δx. The area of any strip, such as that shown in the figure, lies between that of the shorter and that of the taller of the two rectangles, that is, between $y\delta x$ and $(y + \delta y)\delta x$, and the corresponding limits of the total area will be the sums of these expressions for all the strips, that is, the total area will lie between

$$\sum_{x=a}^{x=b} y \, \delta x \text{ and } \sum_{x=a}^{x=b} (y \, \delta x + \delta y \, \delta x),$$

the difference between these limits being

$$\sum_{x=a}^{x=b} \delta y \delta x = \delta x \sum_{x=a}^{x=b} \delta y = \delta x \{f(b) - f(a)\}$$

(since the sum of all the separate increments of y is the difference between the ordinates at a and b, that is, $f(b) - f(a)$).

It is clear that, by making δx sufficiently small, either calculation will give the area required to a high degree of accuracy. Further, since the difference between the two is $\{f(b) - f(a)\}\delta x$, which tends to zero as δx tends to zero, it follows that the area is given *exactly* by

$$\underset{\delta x \to 0}{lt.} \sum_{x=a}^{x=b} y\delta x \quad \text{or} \quad \underset{\delta x \to 0}{lt.} \sum_{x=a}^{x=b} f(x)\delta x.$$

Here, then, is one way in which the expression given at the beginning of this section will arise in practice.

Again, suppose we are told that the velocity of a moving body is known as a certain function of time, say $f(t)$, and we are asked to calculate the distance it will travel in the interval between the instants $t = a$ and $t = b$. There is no question of simply multiplying the time interval $b - a$ by the velocity, because the latter is not constant. As in the above case, however, an approximation could be obtained by dividing the interval into a large number of smaller equal intervals δt, calculating the velocity $f(t)$ at the beginning of each interval and multiplying by δt to get the distance travelled in the short interval. Upper and lower limits for the distance travelled could then be calculated for the whole interval precisely as shown above, and the difference between these could be made as little as desired by sufficiently decreasing δt. The exact result would be, as before,

$$\underset{\delta t \to 0}{lt.} \sum_{t=a}^{t=b} f(t)\ \delta t.$$

There is, therefore, very good reason for trying to find some means of evaluating this limit of a sum, and a combination of the ideas of the differential calculus and of indefinite integration will show how this can be done.

It will be assumed that the function $f(x)$ and its integral are finite and continuous over the range a to b of x. Further, let $F(x)$ be the integral of $f(x)$, that is, $f(x)$ is the differential coefficient of $F(x)$ with respect to x. The range a to b is divided into the n intervals δx, that is, $n\delta x = b - a$. By definition,

$$\underset{\delta x \to 0}{lt.} \frac{F(x + \delta x) - F(x)}{\delta x} = f(x).$$

Therefore for any value of δx greater than zero in magnitude,

$$\frac{F(x + \delta x) - F(x)}{\delta x} = f(x) + h,$$

where h is a quantity which tends to zero when δx tends to zero. This can be written

$$F(x + \delta x) - F(x) = f(x)\delta x + h\delta x.$$

By hypothesis, this is true for all values of x between a and b, whence

$$F(a + \delta x) - F(a) = f(a)\delta x + h_1\delta x,$$

$$F(a + 2\delta x) - F(a + \delta x) = f(a + \delta x)\delta x + h_2\delta x,$$

$$F(a + 3\delta x) - F(a + 2\delta x) = f(a + 2\delta x)\delta x + h_3\delta x,$$

etc. etc. etc.

$$F(b - \delta x) - F(b - 2\delta x) = f(b - 2\delta x)\delta x + h_{n-1}\delta x,$$

$$F(b) - F(b - \delta x) = f(b - \delta x)\delta x + h_n\delta x.$$

By addition,

$$F(b) - F(a) = \sum_{x=a}^{x=b-\delta x} f(x)\delta x + \delta x \sum_{n=1}^{n=n} h_n.$$

Therefore

$$\underset{\delta x \to 0}{lt.} \sum_{x=a}^{x=b-\delta x} f(x)\delta x = F(b) - F(a) - R,$$

where

$$R = \underset{\delta x \to 0}{lt.}\ \delta x \sum_{n=1}^{n=n} h_n.$$

Now the quantities h_n are finite by hypothesis. Let h be the largest value reached by any of them for any value of δx between δx and zero. Then

$$|\delta x \sum_{n=1}^{n=n} h_n| \not> |n\,\delta x\,h|$$

(the vertical strokes mean that magnitude only is being considered, and $\not>$ means " not greater than "). Therefore, since $n\delta x = b - a$,

$$|\underset{\delta x \to 0}{lt.}\ \delta x \sum_{n=1}^{n=n} h_n| \not> |\underset{\delta x \to 0}{lt.}\ (b-a)h| = |(b-a)\underset{\delta x \to 0}{lt.}\ h| = 0$$

since the limit of all the h quantities is zero when δx tends to zero. Therefore, finally, since R is zero,

$$\underset{\delta x \to 0}{lt.} \sum_{x=a}^{x=b-\delta x} f(x)\delta x = \underset{\delta x \to 0}{lt.} \sum_{x=a}^{x=b} f(x)\delta x = F(b) - F(a).$$

The expression on the left is usually written in the more compact form

$$\int_a^b f(x)\,dx,$$

and is called the " definite integral " (or, in practice, just " the integral ") of function x with respect to x from a to b. Thus we have

$$\int_a^b f(x)\,dx = F(b) - F(a),$$

where $F(x)$ is the integral of $f(x)$ with respect to x in the first sense of the word,

that is, $$F(x) = \int f(x)\,dx.$$

In the actual calculation of definite integrals, $F(b) - F(a)$ is written

$$\Big[F(x)\Big]_a^b,$$

so we have

$$\int_a^b f(x)\,dx = \Big[\int f(x)\,dx\Big]_a^b = \Big[F(x)\Big]_a^b = F(b) - F(a).$$

As an example,

$$\int_a^b dx/x = \Big[\log x\Big]_a^b = \log b - \log a = \log (b/a).$$

Notice that there is no need to include the arbitrary constant of integration in $F(x)$, for it would automatically disappear in taking the difference of the limiting values. Notice, further, that

$$\int_a^b f(x)\, dx$$

is not, in general, a function of x, but is a function of the limits a and b. It will only be a function of x if x, or any term depending on x, appears in the limits.

122. The Mean Value of a Function

Another important application of definite integration is the determination of the mean value of a function over a certain range of the variable. In terms of area, the mean value of the function is the height of the rectangle of base $(b - a)$ the area of which is equal to that enclosed by the curve $y = f(x)$ and the ordinates at a and b. In terms of the variable velocity example given above, it would mean the equivalent constant velocity, equivalent in the sense that the moving body would travel the same distance in the same time. In general terms, therefore, the definition of y_m, the mean value of $f(x)$ over the range a to b of x, is

$$(b - a)y_m = \int_a^b f(x)\, dx, \text{ or } y_m = \frac{1}{b - a}\int_a^b f(x)\, dx.$$

Two important special cases are (i) the mean value of an alternating current $i = \hat{\imath} \sin \omega t$ over a period (that is, $2\pi/\omega$), and (ii) the mean value of the square of the same alternating current over the same period (see Examples, page 252).

123. Periodic Functions and Fourier's Theorem

No account of the basic mathematical ideas of importance in radio would be complete without some mention of Fourier's Theorem—one of the most beautiful and useful of all the great mathematical discoveries. Unfortunately, it is, mathematically speaking, rather too difficult for a work as elementary as this, and it cannot, therefore, be dealt with as a mathematical subject, but we have now covered enough of the ideas involved to be able at least to describe the theorem, and show its particular usefulness in electrical theory.

First, it will be well to point out that is is *not* a theorem about periodic functions, for it is often so presented as to give that impression. However, it is chiefly in connection with periodic functions of time that it comes into electrical theory, and it will be presented in that form.

A periodic function of time, $f(t)$, is such that for all values of t

$$f(t + T) = f(t),$$

where T is a constant. If T is the smallest constant for which this is true, it is called the period (or periodic time) of the function. Fig. 53 shows two examples of such a function.

The simple sine wave $f(t) = A \sin (\omega t + \theta)$ is itself a periodic function, with period $2\pi/\omega$, and the essence of Fourier's Theorem is that *any* periodic function, subject to certain stipulations about finite value and continuity, which are, in fact, always complied with by functions which represent any real, natural, periodic phenomena, can be represented as an infinite sum of simple sine functions, that is,

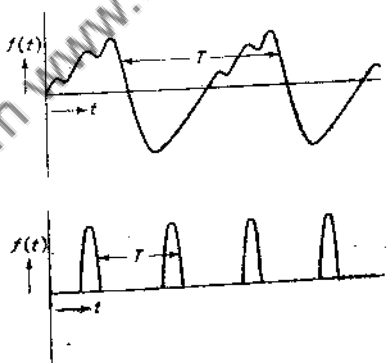


Fig. 53

$$f(t) = A_0 + A_1 \sin (\omega t + \theta_1) + A_2 \sin(2\omega t + \theta_2) + A_3 \sin (3\omega t + \theta_3) + \text{etc., } ad\ inf.,$$

where $\omega = 2\pi/T$, and $A_0 A_1 A_2$, $\theta_1 \theta_2 \theta_3$, etc., are constants.

In such a series, A_0 is called " the constant term ", for a sufficiently obvious reason, $A_1 \sin(\omega t + \theta_1)$ is called "the fundamental", and the next, and succeeding terms are called the 2nd, 3rd, etc., "harmonics".

As already stated, we cannot here demonstrate this expansion, or derive it from first principles, but we can at least show how to determine the constants, assuming the expansion to be valid.

Consider first the average value of the function over a period, that is,

$$\frac{1}{T}\int^{T} f(t)\, dt.$$

If the series truly represents the function, then the average value of the series will be the same as the average value of the function. But the average value of the series is A_0, since the average value of all the other terms is zero. Therefore

$$A_0 = \frac{1}{T}\int_0^T f(t)\, dt.$$

(In most radio applications, but not always, A_0 will be zero, that is, the function will be symmetrical in area about the axis.)

The rest of the process is somewhat analogous to using a stroboscope to analyse a complex rotating system. Multiply the function and the series by $\sin \omega t$, and again equate the average values of the two. Then by $\cos \omega t$, and do the same. Then by $\sin 2\omega t$, and by $\cos 2\omega t$. Better still, do the whole job in one go (or rather, two goes) by multiplying by $\sin n\omega t$ and $\cos n\omega t$. In calculating the average values of the series when so multiplied, it will be necessary to demonstrate the following results :—

$$\frac{1}{T}\int_0^T \sin^2 n\omega t\, dt = \frac{1}{T}\int_0^T \cos^2 n\omega t\, dt = \tfrac{1}{2},$$

and

$$\frac{1}{T}\int_0^T \sin m\omega t \sin n\omega t\, dt = 0,$$

$$\frac{1}{T}\int_0^T \cos m\omega t \cos n\omega t\, dt = 0,$$

$$\frac{1}{T}\int_0^T \sin n\omega t \cos m\omega t \, dt = 0,$$

where m and n are any *unequal* positive whole numbers.

These results are covered by the preceding section, and are included in the examples at the end of this chapter. With this outline for guidance, the reader should have no difficulty in showing that

$$A_n \cos \theta_n = \frac{2}{T}\int_0^T f(t) \sin n\omega t \, dt,$$

and $$A_n \sin \theta_n = \frac{2}{T}\int_0^T f(t) \cos n\omega t \, dt.$$

Finally, assuming for the moment that these definite integrals can be evaluated, and calling them S_n and C_n for short,

from $$A_n \cos \theta_n = S_n,$$

and $$A_n \sin \theta_n = C_n,$$

we get $$A_n^2 = S_n^2 + C_n^2,$$

and $$\tan \theta_n = C_n/S_n,$$

so that A_n and θ_n, $n = 1, 2, 3$ etc., are known.

The integrals can, in fact, be evaluated in various ways. It may be, for example, that $f(t)$ is itself a simple, known function of t, so that the integrals can be calculated. In general, however, the integrals can be evaluated by graphical, or similar, computational methods. Space cannot be spared for a detailed description of any of these, which are, moreover, very fully covered by many existing textbooks (incidentally, in some twenty years of work in radio research the writer has never had occasion to use any of them, and he rather suspects that they are not, in fact, very much used outside class rooms. Where it is necessary to measure the "harmonic content" of complex waves, there are various kinds of electrical and mechanical instruments for doing so—but this is rather beyond our present scope).

The really important point for our present purposes is that any periodic alternating current or voltage that is likely to be met with in practice can, as shown above, be represented as a sum of a series of simple sine waves—theoretically an infinite series, but in many cases a comparatively small number of terms will give a very close approximation. In consequence, circuit calculations with such irregular waves can be carried out almost as simply as for pure sine waves, and the behaviour of circuits and apparatus to such complex kinds of excitation can be understood and mastered in a way that would otherwise be impossible.

Here ends the account of the mathematical foundations of alternating current analysis. The remaining chapter will be devoted entirely to specific applications to actual problems. The fundamental ideas have been presented in a very condensed form, but the necessarily limited space available has precluded a very detailed exposition. This very limitation can, however, be turned to good account by any serious student of the subject, who will find in the development of the detail the best possible means of familiarising himself with the important fundamental ideas.

Examples XV

1. Given that $\mathbf{v} \cdot \nu = v_0 \epsilon^{kt} \cos(\omega t + \psi)$ find $d\mathbf{v}/dt$ and $d^2\mathbf{v}/dt^2$ in terms of $\mathbf{v}$ and a vector operator. Also find $(d\mathbf{v}/dt) \cdot \nu$ and $(d^2\mathbf{v}/dt^2) \cdot \nu$.

2. Find the following integrals :—

   i. $\int \frac{dx}{ax + b}$,

   ii. $\int \frac{(ax + b)\,dx}{cx + d}$,

   iii. $\int \sec x \tan x \, dx$,

   iv. $\int \frac{dx}{a^2 + x^2}$ (put $x = a \tan \theta$),

   v. $\int \frac{\epsilon^x dx}{1 + \epsilon^{2x}}$ (put $\epsilon^x = u$).

3. Integrate by parts :—

   i. $x^2 \log_\epsilon x$,

   ii. $(\log_\epsilon x)^2$,

   iii. $\tan^{-1} x$.

4. Show that :—

   i. $\int_a^b f(x)\,dx = -\int_b^a f(x)\,dx$,

   ii. $\int_a^c f(x)\,dx = \int_a^b f(x)\,dx + \int_b^c f(x)\,dx$.

5. Find the value of

   i. $\int_0^\pi \sin \theta \, d\theta$,

   ii. $\int_0^{2\pi} \sin \theta \, d\theta$,

   iii. $\int_0^{2\pi} \sin^2 \theta \, d\theta$.

Remember that $\sin^2\theta = \frac{1}{2}(1 - \cos 2\theta)$.

iv. $\int_0^{2\pi} \sin m\theta \sin n\theta \, d\theta, \; m \neq n,$

v. $\int_0^{2\pi} \sin m\theta \cos n\theta \, d\theta, \; m \neq n.$

6. Show that the curve $x^2 + y^2 = a^2$ is a circle. Find the area included between the x axis and that part of the curve for which y is positive, and hence show that the area of the whole figure is πa^2.

7. A periodic function $f(t)$ is defined by the following :—

from $t = 0$ to $t = T/2$, $f(t) = A$ (const.),

from $t = T/2$ to $t = T$, $f(t) = -A$.

Find the Fourier series for $f(t)$.

8. Show that if $f(t) = -f\left(t + \frac{T}{2}\right)$, $f(t)$ has no even harmonics.

9. Given that $f_1(t) = f_1(t + T_1)$,

and $f_2(t) = f_2(t + T_2)$,

and $F(t) = f_1(t) \; f_2(t)$,

is $F(t)$ its periodic function ? If so, what is its period ?

Chapter 7

THE APPLICATION OF MATHEMATICAL IDEAS TO RADIO

124. Electric Oscillations and Waves

THE remainder of this book will be devoted to showing how the basic mathematical ideas so far described can be applied to the description and analysis of the electric waves, the apparatus, and the circuits used in the various applications of radio-frequency electricity. For this purpose it will be necessary to assume that the reader has already, or is acquiring elsewhere, a sufficient knowledge of the corresponding basic electrical ideas, for in the limited space available it will only be possible to give these a bare minimum of description—just enough to link them up with the mathematical symbols and methods involved. Also, of course, from so vast a field of application it will only be possible to select a few elementary but fundamental examples, but it is hoped that these will serve to illustrate the way in which the mathematical technique is used over the whole field.

First we will consider some simple types of electric waves, represented in their simplest form by alternating currents in wires.

The simplest kind of alternating current is that which can be represented as a pure sine or cosine function of time, that is,

$$i = \hat{\imath} \cos \phi = \hat{\imath} \cos (\omega t + \theta).$$

The angle $\phi = \omega t + \theta$ (radians is called the "phase" of the oscillation (the angle θ is frequently so called, but in fact it is a particular value of the phase—the value when $t = 0$. It is more suitably called the initial phase).

We know already that the period T and frequency f of the oscillating current are given by

$$\omega = \frac{2\pi}{T} = 2\pi f,$$

but it is important to notice that f can also be expressed as

$$f = \frac{1}{2\pi}\frac{d\phi}{dt},$$

and that, in consequence

$$\phi = \int 2\pi f\, dt = 2\pi ft + \theta = \omega t + \theta.$$

The definition of frequency as $1/2\pi$ times the rate of change of phase is a useful generalisation, because it leads to the idea of "instantaneous frequency" as applied to certain important cases where the frequency is not, strictly speaking, a constant, but is a slowly varying function of time.

Suppose, for example, that the "instantaneous frequency" is given by

$$f = f_0(1 + m \sin pt).$$

Then

$$\frac{1}{2\pi}\frac{d\phi}{dt} = f_0(1 + m \sin pt),$$

and

$$\phi = 2\pi f_0 \int (1 + m \sin pt)\, dt$$

$$= 2\pi f_0\left(t - \frac{m}{p}\cos pt\right) + \theta,$$

so that

$$i = \hat{\imath} \cos\left(\omega_0 t - \frac{m}{p}\,\omega_0 \cos pt + \theta\right),\ \omega_0 = 2\pi f_0,$$

which radio students will recognise as what is called a "frequency-modulated" current.

Returning now to the original single-frequency current, suppose this to be flowing in a pair of lines as shown in Fig. 54. The current has the same phase at all points along the line; but it need not have. For example, θ might be a simple function of x, the distance along the line, say, $-\beta x$, that is, uniformly decreasing. Then

$$i = \hat{\imath} \cos(\omega t - \beta x).$$

The current has the same amplitude $\hat{\imath}$ all along the line, but, because of the uniformly decreasing phase, its instantaneous value will be as shown in Fig. 55. At a

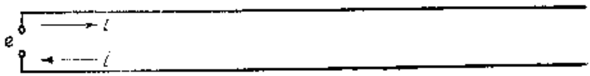


Fig. 54

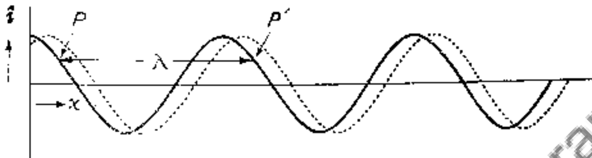


Fig. 55

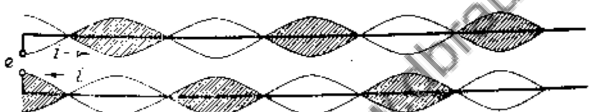


Fig. 56

short time δt later it will be as shown by the dotted line. If one could actually see the current, say, as a luminous haze round the wires, red for positive and blue for negative, the appearance would be somewhat as in Fig. 56, and the whole picture would seem to be streaming away to the right (in fact, of course, there is nothing streaming away to the right, no *thing*, that is. It is a state or condition which " moves ", that is, is acquired by successive parts of the line at successive instants).

What is the velocity of the wave? Is it dx/dt? No, because x is simply the distance along the line and is not a function of t; but we can fix on some point P, which is assumed to move along the line at such a rate that the current at the point P is always the same in magnitude, that is, is always at the same phase. If x_0 be distance of this point at the instant t, then

$$\omega t - \beta x_0 = \text{const} = \phi,$$

that is,
$$\beta x_0 = \omega t - \phi,$$
$$x_0 = \frac{\omega t - \phi}{\beta},$$
$$\frac{dx_0}{dt} = \frac{\omega}{\beta},$$

and this, obviously, is the rate at which the wave travels along the lines. The distance between two successive points such as P and P' in Fig. 55, at which the current has the same phase, is called the wavelength, λ,

that is, if
$$\omega t - \beta x = \phi,$$
$$\omega t - \beta(x + \lambda) = \phi - 2\pi,$$
that is,
$$-\beta\lambda = -2\pi,$$
or
$$\beta = 2\pi/\lambda.$$

The wave can therefore be written
$$i = \hat{\imath} \cos(\omega t - 2\pi x/\lambda),$$
or
$$i = \hat{\imath} \cos 2\pi (ft - x/\lambda).$$

Thus, the expression
$$i_1 = \hat{\imath} \cos(\omega t - \beta x)$$
represents a current-wave travelling in the positive x direction. Similarly,
$$i_2 = \hat{\imath} \cos(\omega t + \beta x)$$
represents a wave travelling in the opposite or negative x direction.

If both currents are made to flow in the line at the same time, the instantaneous value of the current will be the sum of the two (*not* the difference), that is,
$$i = i_1 + i_2 = \hat{\imath}\{\cos(\omega t - \beta x) + \cos(\omega t + \beta x)\}$$
$$= 2\hat{\imath} \cos \omega t \cos \beta x \quad \text{(see page 192)}$$
$$= (2\hat{\imath} \cos \beta x) \cos \omega t.$$

In this case, the phase (ωt) is constant all the way along the line. This time it is the amplitude which varies along the line. At some instant t the instantaneous value of the current would, for example, be as in Fig. 57, and at a short time δt later, as shown by the dotted line. The haze-cloud picture that we imagined for the travelling wave

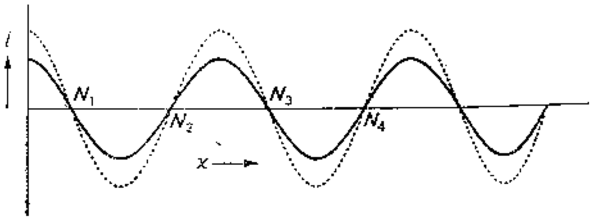


Fig. 57

would, in this case, at a given moment, look as in Fig. 56, but this time there would not appear to be any sideways travel, in either direction. Instead, the apparent movement would be all up and down as the haze-clouds waxed and waned, all disappearing completely into the lines twice every cycle (when $\cos \omega t = 0$ and there is no current anywhere) and re-emerging having changed colour meanwhile, red to blue and blue to red.

In Fig. 57, the points N_1, N_2, N_3, are clearly those for which $\cos \beta x = 0$, that is, $\beta x = \pi/2, 3\pi/2, 5\pi/2$, or $x = \lambda/4$, $3\lambda/4$, $5\lambda/4$, so that the points are half a wavelength apart. They are called nodes.

The two kinds of waves

$$i = \hat{\imath} \cos (\omega t - \beta x)$$

and

$$i = \hat{\imath} \cos \beta x \cos \omega t$$

are called respectively progressive and stationary waves. Both play a large part in radio engineering.

125. The Fundamental Laws of Current Networks

The whole theory of electric current networks, whatever be the nature of the conducting elements of the networks or of the currents flowing in them, is based on two remarkable generalisations known as Kirchhoff's first and second laws—remarkable because of their almost axiomatic simplicity and the wealth of information and deduction derivable from their application.

The first law—the algebraic sum of the currents which meet at any point in a network of conductors is zero.

Notice the word "algebraic", which here, as always, means that sign must be taken into account. By "sign" is meant the sense of the current relative to the point considered. The usual convention in this matter is that a current will be reckoned positive if it is flowing towards the point, and negative if it is flowing away from the point. Thus the application of the law to the element of a network shown in Fig. 58 leads to the equation

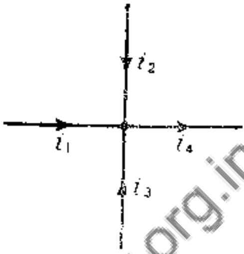


Fig. 58

$$i_1 + i_2 + i_3 - i_4 = 0.$$

"But", says the reader, "how can I tell, in the case of a really complicated network, which way the currents are flowing?" The answer is: "You cannot, but it does not matter, because the combination of correct analysis with known data (such as the nature and disposition of the acting electromotive forces) will automatically confirm or correct the assigned directions". Thus, if in the above example the current i_2 is in fact flowing away from the point, its evaluation will lead to a negative number, -10, for instance, showing that it is a current of magnitude 10 flowing in a direction opposite to that indicated.

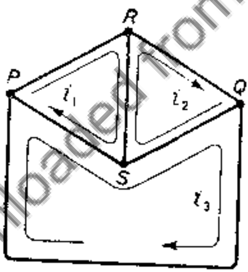


Fig. 59

An alternative and preferable manner of representing the flow of current in a network is that illustrated in Fig. 59, where the actual currents are regarded as due to the superposition of the circuital currents shown. Thus the current in PR is i_1, in RQ i_2, and in RS $(i_1 \sim i_2)$. This form of representation saves the writing of several current equations, for it actually assumes and embodies the first law. The sum of the currents meeting at the point R, for instance, is $i_1 - i_2 - (i_1 - i_2)$, that is, 0.

Physically, the law states that there is no accumulation of electricity at any meeting point of conductors in a network. In actual fact, there will be local accumulations of electricity for an exceedingly short period after a circuit has been closed, just as water released into a system of pipes will first fill up the pipes before settling down to a steady flow, but, in general, this initial period will be negligibly short in duration, and the law applies exactly to the final steady state.

Kirchhoff's second law relates to the sum of the e.m.f.s and the counter-e.m.f.s or potential drops in a closed circuit. Consider, for instance, the passage of a current of magnitude i through a resistance R ohms, illustrated in Fig. 60. By Ohm's Law, a knowledge of which is assumed, the magnitude of the potential drop between the points a and b is iR volts, and the current flows " downhill ", as one would expect it to. That is, a is at a higher potential than b. Thus, if the direction of the e.m.f. be taken as positive, the appropriate sign to attribute to the potential difference iR is negative. Conversely, any voltage in the closed circuit containing R which would tend to maintain the current in the same—that is, positive—direction can reasonably be given a positive sign. Allocating signs in this manner, one can form the algebraic sum of all the e.m.f.s and voltage drops in any closed circuit, and Kirchhoff's second law states that this sum is zero. There should be no difficulty in appreciating the physical significance of the second law, for it means no more than this—that a man who sets out from his home on a roundabout journey up hill and down dale, and then comes home again, must of necessity in the course of his wanderings have gone uphill and downhill to exactly the same extent, since he has finished up at the level from which he started.

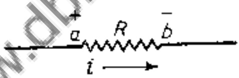


Fig. 60

Applying the law by way of illustration to the circuit shown in Fig. 61, which represents a battery of e.m.f. e volts and internal resistance R_0 ohms supplying current to a resistance R ohms, we have

$$e + e_0 + e_R = 0,$$

that is,

$$e - iR_0 - iR = 0$$

or

$$i = e/(R_0 + R).$$

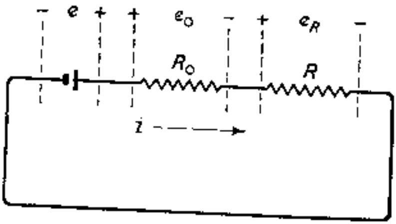


Fig. 61

In the case of a varying current, for example, the high-frequency sine-wave alternating current of radio, other "back e.m.f.s", or opposing voltages, will come into play in addition to those due to the resistances of the conductors involved. It must be assumed that these ideas are already familiar to the reader, but a brief statement of them will be given for the sake of completeness.

126. Inductance

A pure inductance opposes to a varying current i a back e.m.f. e_L proportional to the rate of change of the current, that is, proportional to di/dt. The unit of inductance is so chosen that the back e.m.f. in volts is $-L(di/dt)$, L being the inductance in henries. A negative sign is attributed to it for the same reason as in the case of a pure resistance.

127. Capacitance

A pure capacitance in a circuit carrying a varying current i (see Fig. 62) will oppose to the current a back e.m.f. proportional to the quantity of electricity stored in the condenser. The unit of capacitance is so chosen that this back e.m.f. in volts is q/C, C being the capacitance in farads, and q the quantity of electricity in coulombs stored on the positive electrode of the condenser, that is, leaving the matter of sign for the present

$$|e_c| = q/C.$$

Now, since the rate of change of q (that is, dq/dt) is the rate of flow of electricity along the conductor, that is, i amperes (or i coulombs per sec), we have

$$i = dq/dt.$$

Therefore

$$\frac{d|e_c|}{dt} = \frac{d}{dt}\left(\frac{q}{C}\right) = \frac{1}{C}\frac{dq}{dt} = \frac{i}{C}.$$

Since the direction of the potential difference e_c relative to that of the current is such as to oppose the current, a negative sign is attributed to it in the above equation, giving

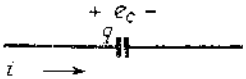


Fig. 62

$$\frac{de_c}{dt} = -\frac{i}{C}.$$

128. Vectorial Representation of Back Voltages

It was shown in Section 85 that a sine-wave alternating current can be represented in the form

$$\mathbf{i} \cdot \nu = \hat{\imath} \cos \omega t,$$

where $\mathbf{i}$ is a vector of constant magnitude $\hat{\imath}$, rotating with constant angular velocity, ω, and where ν is a constant unit vector of reference parallel to the bottom edge of the paper (in the writer's opinion, this form of statement is preferable to the more usual description of the instantaneous value of the current as the " projection " of the rotating vector on a given time axis, since it permits the relationship between the vector and the current to be stated as an exact equation, as above).

(*a*) If such a current flows through a resistance R ohms, the back voltage e_R is given, as already shown, by

$$e_R = -Ri,$$

and since e_R is thus a simple multiple of i it will of necessity be a sine-wave alternating potential difference of the same frequency as i, and can therefore be represented by a rotating vector $\mathbf{e}_R$ in the same manner. Expressing both the current and the back voltage in vector form, we have

$$\mathbf{e}_R \cdot \nu = -R(\mathbf{i} \cdot \nu) = -R\mathbf{i} \cdot \nu.$$

This can be put in the form

$$(\mathbf{e}_R + R\mathbf{i})\cdot\nu = 0.$$

Since this is true at every instant, it follows that the vector $(\mathbf{e}_R + R\mathbf{i})$ is either zero at every instant or else is perpendicular to ν at every instant. The second condition is obviously not fulfilled. Therefore

$$\mathbf{e}_R + R\mathbf{i} = 0, \text{ or } \mathbf{e}_R = -R\mathbf{i}.$$

This shows that the vector representing e_R is R times $\mathbf{i}$ in magnitude and opposite to it in direction, as in Fig. 63.

(*b*) For the back e.m.f. generated in a pure inductance L we have

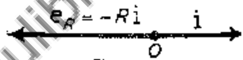


Fig. 63

$$e_L = -L(di/dt),$$

and since the differential coefficient of a sine wave is a sine wave (or a cosine wave, which comes to the same thing) of the same frequency it follows that e_L can also be represented by a rotating vector of the same angular velocity as $\mathbf{i}$. Hence we have the scalar-product equation

$$\mathbf{e}_L\cdot\nu = -L\frac{d(\mathbf{i}\cdot\nu)}{dt}.$$

Now it is easy to show that as ν is a constant vector

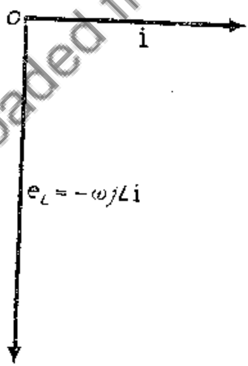


Fig. 64

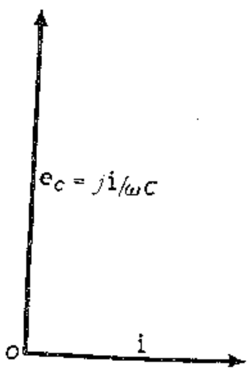


Fig. 65

$$\frac{d(\mathbf{i}\cdot\nu)}{dt} = \frac{d\mathbf{i}}{dt}\cdot\nu$$

Further, it has been shown (see Section 115), that for a vector of this character,

$$d\mathbf{i}/dt = \omega j\mathbf{i}.$$

Therefore

$$\mathbf{e}_L \cdot \nu = -\omega jL\mathbf{i}\cdot\nu,$$

whence, as in case (a), $\mathbf{e}_L = -\omega jL\mathbf{i}$. The relation between the vectors $\mathbf{e}_L$ and $\mathbf{i}$ is therefore as shown in Fig. 64.

(c) With the above two examples, the reader should have no difficulty in showing that the vector representing the back e.m.f. due to a condenser of capacity C is given by

$$\omega j\mathbf{e}_C = -\mathbf{i}C,$$

that is,

$$\mathbf{e}_C = -\mathbf{i}/\omega jC = j\mathbf{i}/\omega C.$$

This is illustrated in Fig. 65.

129. The Characteristics and Behaviour of an Oscillatory Circuit

The circuit shown in Fig. 66 will be recognised at once by all students of radio as probably the most important single circuit in the whole subject. It will now be shown that by means of the mathematical ideas already described, we can find exact answers to such questions as these :—

1. What is the nature and magnitude of the current produced in the circuit by a sine-wave alternating electromotive force ? How does it depend on the frequency and on the magnitudes of the circuit constants, L, C, and R ?

2. What is meant by the " resonance " of such a circuit ?

3. What is the magnitude of the potential difference across the inductance when the circuit is tuned to resonance ? How does this depend on the resistance ?

4. What is the distribution in space and time of the electrical energy in the circuit ?

5. Why is it called an oscillatory circuit?

It was shown in Section 86 that Kirchhoff's law relating to the zero sum of currents meeting at a point in a network is equally true of the rotating vectors which can be used to represent any such set of alternating currents of the same frequency. In precisely the same way it can be shown that Kirchhoff's second law, referring to the sum of the voltages round a closed circuit applies equally to the vectors representing such voltages. Applying this second law to the circuit shown in Fig. 66, gives at once the vector equation

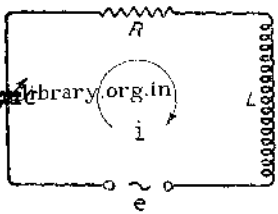


Fig. 66

$$\mathbf{e} + \mathbf{e}_R + \mathbf{e}_L + \mathbf{e}_C = 0.$$

On the instant of closing the circuit certain transient phenomena will occur due to the fact that no system of finite mass (and inductance is the electrical equivalent of mechanical mass) can pass instantaneously from one equilibrium condition to another. This aspect of the matter will be considered more fully later on. When a steady state has been reached (generally in a fraction of a second) the current flowing in the circuit will be of the same character and frequency as the e.m.f. since no other than a current of this character can give rise to back e.m.f.s which will exactly balance the driving e.m.f. at every instant. Therefore, assuming the e.m.f. to be

$$\mathbf{e} \cdot v = \hat{e} \cos \omega t,$$

the current will be of the same frequency $\frac{\omega}{2\pi}$, and can thus be represented by a vector $\mathbf{i}$ of constant magnitude $\hat{\imath}$ and constant angular velocity ω. The expressions already determined for the back e.m.f.s can therefore be substituted at once in the equation, giving

$$\mathbf{e} - R\mathbf{i} - j\omega L\,\mathbf{i} - \frac{\mathbf{i}}{j\omega C} = 0,$$

that is,

$$\left(R + j\omega L + \frac{1}{j\omega C}\right)\mathbf{i} = \mathbf{e},$$

or

$$\left\{R + j\left(\omega L - \frac{1}{\omega C}\right)\right\}\mathbf{i} = \mathbf{e}.$$

It will be convenient to put the single symbol X for the quantity $(\omega L - 1/\omega C)$, that is,

$$(R + jX)\,\mathbf{i} = \mathbf{e},$$

or

$$\mathbf{i} = \frac{\mathbf{e}}{(R + jX)}.$$

This is the complete solution for the "steady-state" alternating current in the circuit, and for most practical purposes it is the best way of representing it. In fact, most modern radio engineers habitually work in terms of such rotating vectors and associated vector operators, without converting back to the scalar instantaneous values at all; but for the present educational purpose it will be well to complete the story in this respect.

In the first place, the "impedance operator" $R + jX$ can (see Section 96) be expressed in the form

$$R + jX = Z\epsilon^{j\phi},$$

where

$$Z^2 = R^2 + X^2,$$

and

$$\phi = \tan^{-1} X/R.$$

Then

$$\mathbf{i} = \frac{\mathbf{e}}{Z\epsilon^{j\phi}}.$$

Now, as already explained, the effect of the operator $1/Z\epsilon^{j\phi}$ is to divide the magnitude of its operand, $\hat{e}$, by Z, and to rotate it through an angle $-\phi$. Therefore, since

$$\mathbf{e}\cdot\nu = \hat{e}\cos\omega t,$$

$$\frac{1}{Z\epsilon^{j\phi}}\ \mathbf{e}\cdot\nu = \frac{\hat{e}}{Z}\cos(\omega t - \phi).$$

Therefore the instantaneous value of i, which is $\mathbf{i}\cdot\nu$, is given by

$$i = \mathbf{i}\cdot\nu = \frac{1}{Z\epsilon^{j\phi}}\ \mathbf{e}\cdot\nu = \frac{\hat{e}}{Z}\cos(\omega t - \phi),$$

where $$Z = \sqrt{R^2 + X^2}$$
$$= \sqrt{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2};$$

and the phase angle ϕ is given by
$$\tan \phi = \frac{\omega L - 1/\omega C}{R}.$$

Here, then, is one form of answer to the first of the questions listed on page 263; but another form may also be useful. Since

$$\frac{1}{R + jX} = \frac{1}{R + jX} \cdot \frac{R - jX}{R - jX},$$
$$= \frac{R - jX}{R^2 + X^2} = \frac{R}{Z^2} - j\frac{X}{Z^2},$$

$$\mathbf{i} = \frac{R}{Z^2}\mathbf{e} - j\frac{X}{Z^2}\mathbf{e},$$

therefore $$i = \mathbf{i} \cdot \nu = \frac{R}{Z^2}(\mathbf{e} \cdot \nu) - \frac{X}{Z^2}(j\mathbf{e} \cdot \nu)$$
$$= \frac{R}{Z^2}\hat{e}\cos \omega t + \frac{X}{Z^2}\hat{e}\sin \omega t.$$

This exhibits the current as the sum of two components of which one is in phase with the e.m.f. and the other is 90° out of phase, that is, in "quadrature" with the e.m.f.

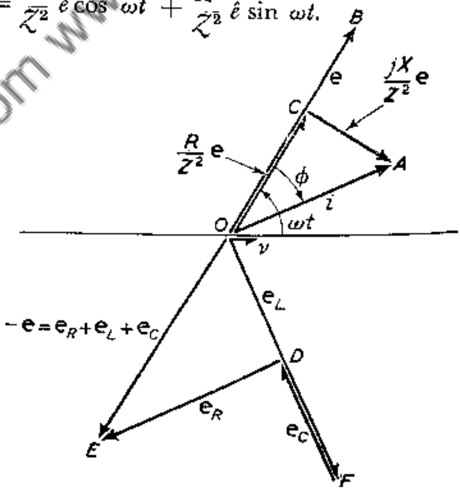


Fig. 67

One advantage of the vector method of analysis is the ease with which the results can be shown in pictorial form. Thus all the vector quantities involved in the above calculation are as shown in Fig. 67.

130. The Impedance Operator

The operator group $(R + jX)$ is known as the "impedance operator" (now frequently abbreviated, in conversation at least, to "impedance") of the circuit, and the angle $\phi = \tan^{-1} X/R$ is called the phase angle of the impedance. R and X are called the resistive and reactive components of the impedance. In general, the relation between the current and e.m.f. vectors for any single-frequency circuit, however complicated, will be an operator of the form $(R + jX)$, since any combination of such operators can ultimately be reduced to a single operator of this kind. In such cases, the R and X terms respectively will often be called the resistive and reactive components, even though they may be composite and complicated in structure, and the resistive component may, for example, contain terms other than pure resistances.

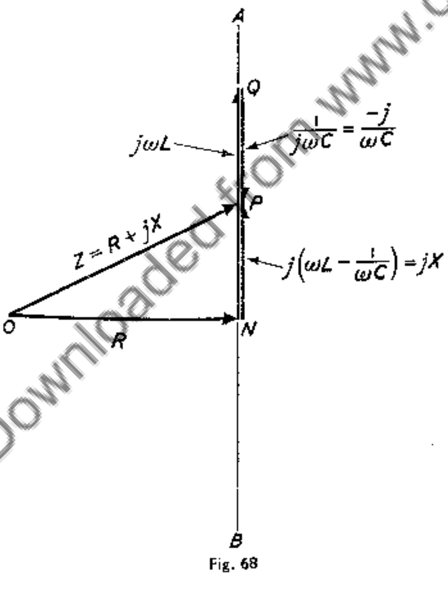


Fig. 68

In this connection, by the way, it should be noted that the circuit in Fig. 66 shows the resistance as a separate circuit element. In most practical cases an oscillatory or tuned circuit of this kind consists simply of a coil and a condenser, and the resistance involved is chiefly that of the coil, and is, as it were, distributed along the length of the wire making up the coil. It may be taken as a fact of experience that, at any given frequency, such an item will behave as if it consisted of a pure inductance in series with a pure resistance (this will, at least, be very nearly true for frequencies not too near the so-called natural frequency of the coil—but this is taking us rather far from our immediate object).

Returning to the impedance operator $R + jX$, where $X = (\omega L - 1/\omega C)$, it is important to notice that this also lends itself to pictorial representation. The accepted convention for this purpose is illustrated in Fig. 68. It is clearly equivalent to drawing the vector $(R + jX)\nu$, where ν is the agreed unit vector of reference, but as long as it is always borne in mind that $R + jX$ is not itself a vector, but a vector operator, there is no need to include the symbol ν in the diagram.

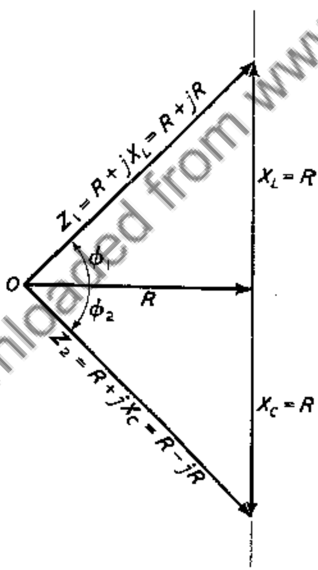


Fig. 69

This diagram enables us to see very clearly what will happen if everything in the system is kept constant except the capacitance of the condenser, which is shown in the circuit diagram as variable. Clearly the component $1/j\omega C$ will vary in length as C is varied, and point P, which is the termination of the operator Z, will move up and down the line AB. The line AB is called the locus of Z for variation of C. Also, when $X = 0$, P coincides with N, and $Z = R$, and thus reaches a minimum value. Under these conditions, i reaches a maximum value and is equal to e/R, so that

$$i = \frac{\hat{e}}{R} \cos \omega t.$$

This is part of the answer to the second question on page 263. Resonance occurs when the condenser is adjusted so that $X = 0$, that is,

$$\omega L = 1/\omega C,$$

or

$$\omega^2 LC = 1,$$

and the current then reaches a maximum value given by $\mathbf{i} = \mathbf{e}/R$, and comes into phase with $\mathbf{e}$.

The brightest readers will already have seen why this is only part of the answer to the second question. They will have noticed that if everything is kept constant except ω, the locus of Z will be the same straight line AB, and that $X = 0$ is also the condition for resonance with respect to variation of frequency (in practice this will be very nearly true, but not exactly, because R will vary also as frequency is varied, though not nearly as rapidly as X. However, this is rather outside our present scope. It is just a caution).

Notice, by the way, the special cases when C is adjusted as shown in Fig. 69 so that $X_L = + R$ and $X_c = - R$. Under these conditions $\phi = \pm 45°$ and, from the property of right-angled triangles, $|Z| = \sqrt{2}\, R$, so that

$$i = \frac{\hat{e}}{R\sqrt{2}} \cos (\omega t \pm \pi/4),$$

which is $1/\sqrt{2}$ times the resonance value. Radio students will recognise this as the basis of the reactance-variation method of measuring radio-frequency resistance, for

$$\omega L - \frac{1}{\omega C_1} = R,$$

and

$$\omega L - \frac{1}{\omega C_2} = - R,$$

which lead to

$$R = \frac{1}{2\omega} \frac{\delta C}{C^2},$$

or

$$\frac{\omega L}{R} = \frac{2C}{\delta C}.$$

where δC is written for $C_1 - C_2$, and C is the resonance value given by $\omega^2 LC = 1$.

Before leaving this resonance part of the story, we shall find the answer to the third question. At resonance

$$\mathbf{i} = \frac{\mathbf{e}}{R},$$

Therefore

$$\frac{\mathbf{e}_L}{\mathbf{e}} = j\frac{\omega L\mathbf{i}}{\mathbf{e}} = j\frac{\omega L}{R},$$

or

$$\left|\frac{\mathbf{e}_L}{\mathbf{e}}\right| = \frac{\omega L}{R}.$$

This quantity expresses the "magnification" of the e.m.f. $\mathbf{e}$ into a large potential difference $\mathbf{e}_L$ by means of the resonance effect, and is therefore an important practical characteristic of the coil. Indeed it is a measure of the "goodness" of the coil for most purposes, and is so used. It is commonly called the Q factor, or, colloquially, "the Q". Since, at resonance, $\omega L = 1/\omega C$,

$$\frac{\omega L}{R} = \frac{1}{\omega CR} = \frac{1}{R}\sqrt{\frac{L}{C}}.$$

These alternative expressions involve C as well, and are better described as the Q of the circuit. Note, by the way, that if we include the resistance in the coil—as in practice we must—the ratio for the coil takes the form

$$\frac{\sqrt{R^2 + \omega^2 L^2}}{R} = \frac{\omega L}{R}\left(1 + \frac{R^2}{\omega^2 L^2}\right)^{\frac{1}{2}}.$$

As shown on page 148, this can be expressed approximately as

$$\frac{\omega L}{R}\left(1 + \frac{1}{2}\frac{R^2}{\omega^2 L^2}\right).$$

In general, $\omega L/R$ will be of the order 100, so that $\frac{1}{2}R^2/\omega^2 L^2$ is of the order 10^{-4}, and can be neglected compared with 1. But do not fall into the trap of assuming that R can always be neglected in this way. We have already seen that, at resonance, R is the most important term in the whole system.

131. Free Oscillation

Instead of assuming a certain applied e.m.f. and then finding the resulting current in the circuit, we could reverse the process, that is, assume a certain current and then find out what e.m.f. is required to balance the consequent back voltages and thus maintain the current.

Suppose, for example, that the current, instead of being a continuous alternation $\hat{i}\cos\omega t$, has a form that has already been briefly mentioned (on page 253), that is,

$$i = \hat{i}\epsilon^{-kt}\cos(\omega t + \phi),$$

which, as there shown, can be represented by a vector of exponentially decreasing magnitude $\hat{i}\epsilon^{-kt}$, rotating with uniform angular velocity ω. By applying the results for the differentiation of such a vector (given in Section 115) to the determination of the potential differences across the various circuit elements, as in Section 128, it will be quite easy to show that

$$\mathbf{e}_R = -R\mathbf{i},$$
$$\mathbf{e}_L = (k - \omega j)L\mathbf{i},$$

and $$\mathbf{e}_C = \frac{\mathbf{i}}{(k - \omega j)C}.$$

Therefore the e.m.f. required to balance these is given by

$$\mathbf{e}_R + \mathbf{e}_L + \mathbf{e}_C + \mathbf{e} = 0$$

or $$\left\{R - (k - \omega j)L - \frac{1}{(k - \omega j)C}\right\}\mathbf{i} = \mathbf{e},$$

and since $$\frac{1}{k - \omega j} = \frac{k + \omega j}{k^2 + \omega^2},$$

$$\left[R - k\left\{L + \frac{1}{(\omega^2 + k^2)C}\right\} + j\left\{\omega L - \frac{\omega}{(\omega^2 + k^2)C}\right\}\right]\mathbf{i} = \mathbf{e}.$$

Now suppose that ω is so chosen that

$$\omega L - \frac{\omega}{(\omega^2 + k^2)C} = 0,$$

which will be true if $\omega = 0$, or if

$$\omega^2 + k^2 = \frac{1}{LC},$$

that is, if $$\omega^2 = \frac{1}{LC} - k^2.$$

Then the equation for the required e.m.f. becomes

$$\left[R - k\left\{ L + \frac{1}{(\omega^2 + k^2)C} \right\} \right] \mathbf{i} = \mathbf{e},$$

or $$(R - 2kL)\,\mathbf{i} = \mathbf{e}.$$

Now suppose in addition that k is given the value

$$k = 2L/R.$$

Then the equation becomes

$$(0)\,\mathbf{i} = \mathbf{e},$$

or $$\mathbf{e} = 0.$$

What does this mean physically ? It means that if by any means we can start up in the circuit a current

$$i = \hat{\imath}\, \epsilon^{-kt} \cos(\omega t + \phi),$$

where $$\omega = \sqrt{\frac{1}{LC} - k^2} = \sqrt{\frac{1}{LC} - \frac{R^2}{4L^2}}$$

and $$k = \frac{R}{2L},$$

then *no applied e.m.f. is required to maintain it.* It will go on quite happily for ever without any external assistance ; but it also means more than this, for suppose we can start *any* current flowing in the circuit at the instant $t = 0$, say by suddenly charging the condenser with an external battery which is then removed. For physical reasons, the current cannot vanish instantaneously. It must continue to flow for a finite time, however short; but during this time, however short, it must satisfy Kirchhoff's second law, and it can do so if it takes the form we have assumed. (We have not, by the way, shown that this is the *only* form which will satisfy all the requirements. It is so in fact, but the proof is rather beyond the scope of an elementary discussion.)

This kind of current is known as a " damped oscillation " because although, formally speaking, it goes on for ever,

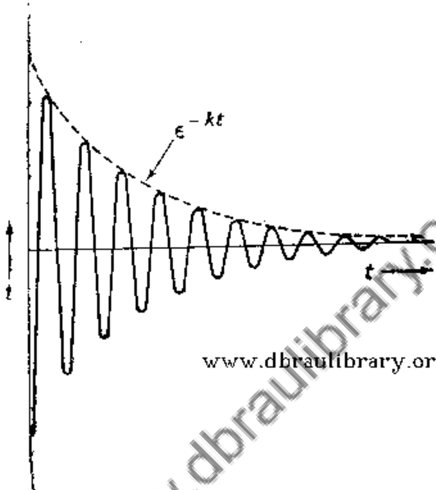


Fig. 70

it is in fact very quickly "damped down" to a very small amplitude, because of the term ϵ^{-kt}, as shown in Fig. 70 (there is, in any case, a physical lower limit due to the fact that electricity is not a continuous and infinitely sub-divisible fluid, but has an atomic structure). Such an oscillation is also called a "free oscillation" for a sufficiently obvious reason. Generally speaking, all natural free oscillations are damped oscillations.

Here, then, we have the answer to question No. 5. The circuit is called "oscillatory" because, when electrically disturbed, it tends to oscillate at its own natural free frequency.

Notice, by the way, that if a pure sine-wave e.m.f. is suddenly applied to the circuit at the instant $t = 0$, a free oscillation may, and generally will, be set up in addition to the continuous or "forced" oscillation already determined for this case. Since the sum of the back e.m.f.s is zero for such a free oscillation, and is **e** for the forced oscillation, the combined sum is **e** and thus satisfies Kirchhoff's second law. However, this "transient condition", as it is called, will very quickly give way to the sustained or "quasi-stationary" state owing to the damping down of the free oscillation.

Why, physically, is the free oscillation damped down? Because the current, flowing in the resistance, generates, or is converted into, heat. It cannot go on doing this at any finite rate for ever. If it could, there would be no fuel problem. The process can only continue until the energy given to the circuit to start it oscillating has all been dissipated as heat. This leads us to what may be called the "dynamics", or energy relationships of the oscillating circuit, that is, to question No. 4.

132. The Dynamics of Electrical Oscillation

The reader is assumed to understand that if a current of instantaneous value i amperes flows in a circuit element across which the consequent instantaneous value of the potential difference is v volts, then the instantaneous rate at which electrical energy is being absorbed by the element or converted into some other form of energy (for example, heat or motion) is given in watts by the product iv. Similarly, if this current is maintained by an electromotive force e, the latter is supplying energy at the instantaneous rate ie watts.

Applying this to the oscillatory circuit excited by the continuous e.m.f. $\hat{e} \cos \omega t$, we have, as shown on page 266,

$$i = \frac{R}{Z^2}\ \hat{e} \cos \omega t + \frac{\omega L \hat{e}}{Z^2} \sin\ \omega t - \frac{(1/\omega C)}{Z^2}\ \hat{e} \sin\ \omega t.$$

Therefore

$$ie = \frac{R}{Z^2}\hat{e}^2 \cos^2 \omega t + \frac{\omega L \hat{e}^2}{Z^2} \sin \omega t \cos \omega t - \frac{(1/\omega C)}{Z^2}\ \hat{e}^2 \sin \omega t \cos \omega t,$$

and since

$$\hat{i}^2 = \frac{\hat{e}^2}{Z^2},$$

$$\cos^2 \omega t = \frac{1}{2} + \frac{1}{2} \cos 2\omega t,\ \sin\ \omega t \cos\ \omega t = \frac{1}{2} \sin 2\omega t,$$

this can be written

$$ie = \frac{\hat{i}^2 R}{2} + \frac{\hat{i}^2 R}{2} \cos 2\omega t + \omega L\ \frac{\hat{i}^2}{2} \sin 2\omega t - \frac{1}{\omega C}\ \frac{\hat{i}^2}{2} \sin 2\omega t.$$

This shows how the instantaneous power is distributed among the various elements of the circuit, and the equation tells an interesting story.

Notice first that for all the circuit elements the energy terms are proportional to the square of the current amplitude. Again, the energy changes in the coil and condenser are in antiphase. While the one is gaining energy the other is losing it. Moreover, when $\omega L = 1/\omega C$, the coil and condenser form a balanced system, in that the one gains energy and the other loses it at exactly the same rates. This seems to be the real physical significance of resonance, for under these conditions the e.m.f. does not have to supply any of this oscillating or "reactive" energy. Notice again that even in the general case, there is no net supply of energy to either coil or condenser (remember that this refers to a coil assumed to have no resistance—the resistance is considered as a separate element), for the average value of $\sin 2\omega t$ is zero. This means that the energy supplied to either during any half period of $\sin 2\omega t$ is recovered during the next half period, which means that energy is alternately stored and released. Thus the total energy acquired by the inductance in the interval $t = 0$ to $t = T/4$ is

$$W = \int_0^{T/4} \frac{\omega L \hat{i}^2}{2} \sin 2\omega t \, dt$$

$$= \frac{\omega L \hat{i}^2}{2} \left[-\frac{1}{2\omega} \cos 2\omega t \right]_0^{T/4}$$

$$= \frac{1}{2} L \hat{i}^2.$$

This is a perfectly general result. A coil of inductance L carrying a current i has a stored (magnetic) energy of amount $\frac{1}{2} Li^2$.

Similarly, for the condenser, it can be shown in the same way that, for the same period of time,

$$W = -\frac{1}{\omega^2 C} \frac{\hat{i}^2}{2}.$$

This, however, is better expressed in terms of $\hat{v}_C$, the amplitude of the potential difference across the condenser, which is given by

$$\hat{v}_C = -\frac{\hat{i}}{\omega C}.$$

so that

$$W = -\frac{1}{2} C\hat{v}^2.$$

This is also a general result, that is, a condenser of capacitance C charged to a potential v has a stored (electric) energy $\frac{1}{2}Cv^2$.

Now consider the resistive part of the circuit. Take the term $\frac{1}{2}\hat{i}^2R \cos 2\omega t$. This has positive and negative values. Does this mean that there are times when the resistance is returning energy to the circuit? This would indeed be something new and startling in electrical theory! No, the point is that in this case the separation into two terms obviously does not correspond to any physical reality, since the resistance is itself a single element. The power supply to the resistance is really the single term $\hat{i}^2R \cos^2 \omega t$, which is always positive. The separation into two terms is simply a convenient way of showing that the average value is $\hat{i}^2R/2$, since the average value of the other term is zero.

Thus, briefly, the dynamics of the circuit can be described as an oscillation of energy between the inductance and the condenser, combined with a pulsating dissipation of energy in the resistance. It presents a very complete analogy with mechanical oscillation, in which there is an alternation of energy between the kinetic (magnetic) and potential (electric) forms, with a pulsating dissipation of energy by friction (resistance).

It has already been pointed out that the "goodness" of a tuned circuit for most radio applications can conveniently be measured by the quantity $Q = \omega L/R$. We can now see that this is in effect the ratio of the oscillating energy to the dissipated energy, that is,

$$Q = \frac{\omega L}{\mathrm{R}} = \omega \; \frac{\frac{1}{2}L\hat{i}^2}{\frac{1}{2}R\hat{i}^2}.$$

The $\omega L/R$ definition only applies to certain simple special cases, but the definition in terms of energy is quite general for all kinds of electrical oscillators, and has recently been adopted as an agreed general definition of Q.

133. The Measurement of Alternating Current, Voltage and Power

From the preceding section, it is clear that a direct current i flowing in a resistance R will be dynamically equivalent to an alternating current of amplitude $\hat{\imath}$ in the same resistance, provided

$$i^2R = \frac{\hat{\imath}^2R}{2},$$

that is, provided $$i = \frac{\hat{\imath}}{\sqrt{2}},$$

The quantity $\hat{\imath}/\sqrt{2}$ is called the "effective value" or the "root-mean-square value" (r.m.s. value) of the alternating current $i = \hat{\imath}\cos(\omega t + \phi)$, and is usually written I. It is, of course, so called because

$$I = \frac{\hat{\imath}}{\sqrt{2}} = \left\{\frac{1}{T}\int_0^T i^2 dt\right\}^{\frac{1}{2}}.$$

The reader should have no difficulty in showing that if i is an alternating current of complex waveform represented by

$$i = \hat{\imath}_1\cos(\omega t + \phi_1) + \hat{\imath}_2\cos(2\omega t + \phi_2) + \hat{\imath}_3\cos(3\omega t + \phi_3) + \text{etc., etc.,}$$

$$I^2 = \frac{1}{T}\int_0^T i^2 dt = \frac{\hat{\imath}_1{}^2 + \hat{\imath}_2{}^2 + \hat{\imath}_3{}^2 + \text{etc., etc.}}{2}.$$

Alternating current is therefore measured on instruments in which the deflecting force is proportional to the square of the current, and in which the moving system has so much inertia that it cannot follow the instantaneous (double-frequency) fluctuations of i^2 and therefore registers its average value. Alternatively, and more usually at radio frequencies, the measurement will depend on measuring i^2R, for example, the thermal e.m.f. in a thermojunction, or the change with temperature of the length or the resistance of a wire.

The effective or root-mean-square value of an alternating voltage is similarly defined.

Finally, the average value of the electrical power supplied by an e.m.f. $e = \hat{e} \cos \omega t$, which maintains a current $i = \hat{\imath} \cos (\omega t - \phi)$, is

$$W = \frac{1}{T}\int^{T}\{\hat{\imath} \cos (\omega t - \phi)\, \hat{e} \cos \omega t\}\, dt$$

$$= \frac{\hat{\imath}\,\hat{e}}{2} \cos \phi = IE \cos \phi.$$

In this expression, $\cos \phi$ is called the "power-factor".

134. Combinations of Impedances

There is another very important general proposition to demonstrate in connection with the simple tuned circuit, but it will first be convenient to note some general properties of series and parallel combinations of impedances.

For three impedances connected as shown in Fig. 71, that is, in series, the application of Kirchhoff's first law, and other propositions already established, will lead to

$$\mathbf{i} = \mathbf{e}/z,$$

where

$$z = z_1 + z_2 + z_3,$$

that is, the effective total impedance is the sum (operator sum, that is) of the separate impedances. Thus, expressing z as $R + jX$,

$$R = R_1 + R_2 + R_3,$$

and

$$X = X_1 + X_2 + X_3.$$

Note also that, since $\mathbf{v}_1 = \mathbf{i}z_1$, $\mathbf{v}_2 = \mathbf{i}z_2$, etc.,

$$\frac{\mathbf{v}_1}{\mathbf{e}} = \frac{z_1}{z},\ \frac{\mathbf{v}_2}{\mathbf{e}} = \frac{z_2}{z}, \text{ etc.,}$$

which is often useful in circuit analysis.

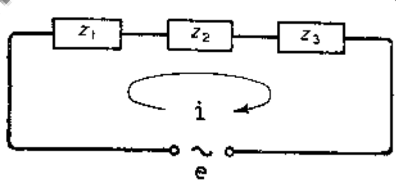


Fig. 71

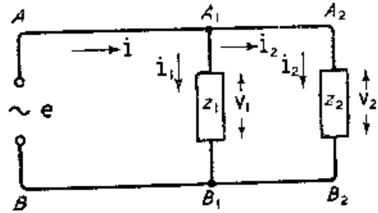


Fig. 72

Now consider two impedances z_1 and z_2 connected in parallel as shown in Fig. 72.

From Kirchhoff's second law, applied to the closed circuits $A\,A_1\,B_1\,B$ and $A\,A_2\,B_2\,B$,

$$\mathbf{e} + \mathbf{v}_1 = \mathbf{e} + \mathbf{v}_2 = 0,$$

that is, $$-\mathbf{v}_1 = -\mathbf{v}_2 = \mathbf{e},$$

or $$\mathbf{i}_1 z_1 = \mathbf{i}_2 z_2 = \mathbf{e}.$$

Therefore $$\mathbf{i}_1 = \frac{\mathbf{e}}{z_1} \text{ and } \mathbf{i}_2 = \frac{\mathbf{e}}{z_2}.$$

Now, applying Kirchhoff's first law to the currents at the point A_1,

$$\mathbf{i} = \mathbf{i}_1 + \mathbf{i}_2,$$

that is, $$\mathbf{i} = \frac{\mathbf{e}}{z_1} + \frac{\mathbf{e}}{z_2} = \mathbf{e}\left(\frac{1}{z_1} + \frac{1}{z_2}\right).$$

Therefore, putting

$$\mathbf{i} = \frac{\mathbf{e}}{z},$$

$$\frac{1}{z} = \frac{1}{z_1} + \frac{1}{z_2}.$$

The generalisation to any number of impedances is sufficiently obvious, and need not be detailed.

For the case of two general impedances in parallel there is a very simple geometrical construction which is useful in visualising the combination. It is illustrated in Fig. 73. The construction is obvious except that the triangle

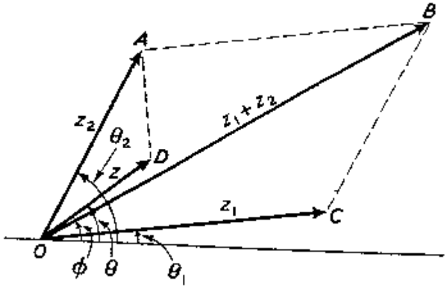


Fig. 73

OAD is drawn similar to OBC.

$$\frac{z_2}{z} = \frac{OA}{OD}\,\epsilon^{j(\theta_2-\theta)},$$

$$\frac{z_1+z_2}{z_1} = \frac{OB}{OC}\,\epsilon^{j(\phi-\theta_1)}.$$

But, in virtue of the similarity of the triangles,

$$\frac{OA}{OD} = \frac{OB}{OC}, \text{ and } \theta_2 - \theta = A\hat{O}D = B\hat{O}C = \phi - \theta_1.$$

Therefore $$\frac{z_2}{z} = \frac{z_1+z_2}{z_1}$$

or $$\frac{1}{z} = \frac{z_1+z_2}{z_1 z_2} = \frac{1}{z_1} + \frac{1}{z_2}.$$

If z_1 and z_2 are mutually perpendicular, the construction is a special case of the above and takes the simple form shown in Fig. 74.

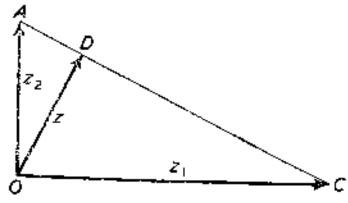


Fig. 74

135. The Rejector Circuit

The oscillatory circuit may be excited by an e.m.f. applied across the circuit as shown in Fig. 75. The branch impedances are now in parallel, and, from the preceding section,

$$\mathbf{i}_0 = \left(j\omega C + \frac{1}{z}\right)\mathbf{e},$$

where $\quad z = R + j\omega L,$

or $\quad \mathbf{i}_0 = \left\{\frac{R}{Z^2} + j\left(\omega C - \frac{\omega L}{Z^2}\right)\right\}\mathbf{e}.$

where $\quad Z = R^2 + \omega^2 L^2$

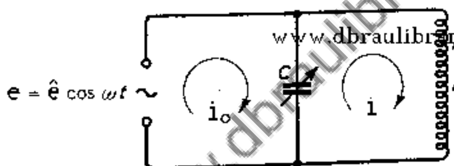


Fig. 75

Here again there is clearly a resonant or critical condition, at the frequency corresponding to

$$\omega C - \frac{\omega L}{Z^2} = 0,$$

or

$$\omega^2 LC = \frac{1}{\left(1 + \frac{R^2}{\omega^2 L^2}\right)}$$

(this is not exactly the same as the series resonant frequency, for which $\omega^2 LC = 1$, but since R^2 is usually very small indeed compared with $\omega^2 L^2$, the difference is generally negligible).

It is easy to see that this resonance corresponds to a *minimum* value of the current $\mathbf{i}_0$. In fact, at resonance

$$\mathbf{i}_0 = \frac{R}{Z^2}\mathbf{e} = \frac{\mathbf{e}}{Z^2/R},$$

and the circuit behaves to the applied e.m.f. as if it were a pure resistance $R_e = Z^2/R$, which will in general be a large quantity—anything from ten thousand ohms to a megohm. Also, paradoxical as it may seem, the smaller the series resistance of the circuit the larger this effective external-circuit resistance. In this connection the circuit is called a "rejector circuit", and the resistance R_e is called the "rejector circuit resistance", or, more generally though less justifiably, the "dynamic resistance". Note that since, at resonance,

$$\omega C = \frac{\omega L}{Z^2} \quad \text{or } Z^2 = \frac{L}{C},$$

$$\frac{Z^2}{R} = \frac{L}{CR}.$$

And, if R^2 is small compared with $\omega^2 L^2$,

$$\frac{Z^2}{R} \approx \frac{\omega^2 L^2}{R}.$$

This transformation of a low real resistance into a high effective resistance must, of course, be associated with a corresponding transformation of current satisfying the energy equality

$$\frac{i_0^2}{2} R_e = \frac{i^2}{2} R,$$

that is,

$$\frac{R_e}{R} = \frac{i^2}{i_0^2}.$$

This is easily confirmed by applying Kirchhoff's second law to the *L-C-R* circuit, that is,

$$\frac{\mathbf{i} - \mathbf{i}_0}{j\omega C} + z\,\mathbf{i} = 0,$$

whence

$$\frac{\mathbf{i}}{\mathbf{i}_0} = \frac{1}{1 + j\omega C z},$$

and since, at resonance,

$$\omega C = \omega L/Z^2,$$

$$\frac{\mathbf{i}}{\mathbf{i}_0} = \frac{1}{1 + j\omega L/(R - j\omega L)} = \frac{R - j\omega L}{R},$$

so that, $$\frac{i^2}{i_0^2} = \frac{Z^2}{R^2} = \frac{Z^2/R}{R} = \frac{R_e}{R}.$$

This is very nearly equivalent to

$$\frac{i}{i_0} = \frac{\omega L}{R},$$

which shows that the current in the tuned circuit is very much larger than that in the external circuit.

This "transformation ratio" as between the tuned circuit and external currents, with the possibility of controlling this by means of adjustment of the circuit constants, is a matter of considerable practical importance. We shall see in the next section how it enables optimum efficiency conditions to be established by "matching" a tuned circuit to a source of e.m.f. which, as in all practical cases, is associated with an effective internal impedance having a resistive component.

136. The "Matching" of an Oscillatory Circuit to a Source of E.M.F.

In most of the practical applications of the simple oscillatory circuit, the arrangement will be essentially as shown in Fig. 76, that is, the circuit is connected to the source of e.m.f. through an impedance. Moreover, by one of various possible means, the reactive part of the impedance can be controlled. An important example is the connection of such a circuit to an aerial by means of a circuit which includes a variable condenser.

In most cases it will be desired to realise the maximum value of the resonant voltage v for a given e.m.f.

The analysis will be given very briefly, since the essential steps have already been exemplified, but more space will be given to the physical interpretation, which is nearly always the most useful and interesting part of the story.

Putting $R_0 + jX_0 = z_0$, and z_e for the parallel circuit impedance,

$$\frac{\mathbf{v}}{\mathbf{e}} = \frac{z_e}{z_0 + z_e} = \frac{1}{z_0} \cdot \frac{1}{\frac{1}{z_0} + \frac{1}{z_e}}$$

$$= \frac{1}{z_0} \cdot \frac{1}{\dfrac{R_0}{Z_0^2} + \dfrac{R}{Z^2} + j\left(\omega C - \dfrac{\omega L}{Z^2} - \dfrac{X_0}{Z_0^2}\right)},$$

where $$Z^2 = R^2 + \omega^2 L^2$$
and $$Z_0^2 = R_0^2 + X_0^2.$$

Resonance of **v**/**e** with respect to C will occur when

$$\omega C - \frac{\omega L}{z^2} - \frac{X_0}{z_0^2} = 0,$$

and $$\left(\frac{\mathbf{v}}{\mathbf{e}}\right)_{\text{res.}} = \frac{1}{z_0} \cdot \frac{1}{\dfrac{R_0}{Z_0^2} + \dfrac{R}{Z^2}}.$$

Therefore $$\left(\frac{v}{e}\right)_{\text{res.}} = \frac{1}{Z_0} \cdot \frac{1}{\dfrac{R_0}{Z_0^2} + \dfrac{R}{Z^2}} = \frac{1}{\dfrac{R_0}{Z_0} + \dfrac{Z_0}{R_e}},$$

where $$R_e = \frac{Z^2}{R},$$

and is the "rejector circuit resistance" of the preceding section.

The problem now is to find how this resonant voltage ratio depends on the adjustable element X_0, and whether there is any optimum adjustment. However, with the ratio expressed in the above form, the problem is a very simple one, for if X_0 is varied, Z_0 will also vary, and we can therefore consider the "optimisation" with respect to Z_0. Now the denominator consists of the sum of two terms, R_0/Z_0 and Z_0/R_e, the product of which is constant with respect to Z_0. Therefore (see Example 8, page 229) their sum will be a minimum when they are equal, that is, the optimum condition is defined by

$$\frac{R_0}{Z_0} = \frac{Z_0}{R_e}$$

or $$Z_0^2 = R_0^2 + X_0 = R_e R_0$$

or $$X_0^2 = R_0(R_e - R_0),$$

which shows that there is no optimum value for X_0 unless $R_e > R_0$.

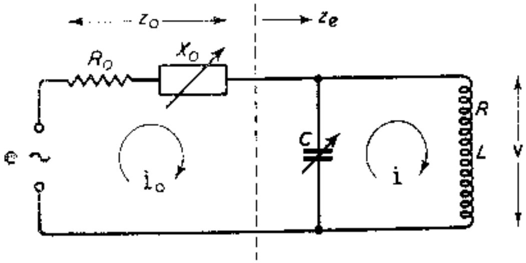


Fig. 76

The corresponding maximum value of the resonant voltage ratio is given by

$$\left(\frac{\hat{v}}{\hat{e}}\right)_{res.\ max.} = \frac{1}{2}\,\frac{Z_0}{R_0} = \frac{1}{2}\sqrt{\frac{R_e}{R_0}} \approx \frac{1}{2}\,\frac{\omega L}{R}\sqrt{\frac{R}{R}}\,.$$

This process of adjustment is called "matching" the tuned circuit to the source, and similar matching processes play a very large part in radio. The reason for the name and the physical meaning of the process will become clear if we find the value of z_e under the above tuned and matched condition. We have

$$\frac{1}{z_e} = \frac{1}{z} + j\omega C = \frac{R}{Z^2} - j\,\frac{\omega L}{Z^2} + j\omega C,$$

and since for the tuned value of C,

$$j\omega C = \frac{j\omega L}{Z^2} = j\,\frac{X_0}{Z_0^{\,2}},$$

$$\frac{1}{z_e} = \frac{R}{Z^2} + j\,\frac{X_0}{Z_0^{\,2}}.$$

But under the matched condition,

$$\frac{R}{Z^2} = \frac{R_0}{Z_0^{\,2}},$$

therefore $\quad \dfrac{1}{z_e} = \dfrac{R_0 + jX_0}{Z_0^{\,2}} = \dfrac{1}{R_0 - jX_0},$

therefore $\quad z_e = R_0 - jX_0.$

Thus, the matched condition is effectively as shown in Fig. 77. This, as far as the e.m.f. is concerned, is a series

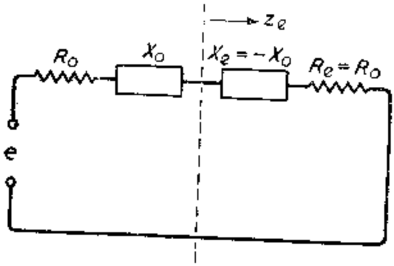


Fig. 77

tuned circuit with zero total reactance and two equal resistances R_0. That is to say, the transformation process described in the preceding section enables the effective resistance of the oscillatory circuit to be matched to, that is, adjusted to equality with, the internal resistance R_0 of the source, while at the same time the reactance of the source is tuned out by an equal and opposite reactance provided by the oscillatory circuit.

The most important point to notice about the matched system is that the total electrical power being dissipated in the system is equally divided between the source and the oscillatory circuit, which is the useful load. It will be found that this "dynamic balance" is a perfectly general characteristic of all such matching processes, however they are effected, and is their real physical significance.

Chapter 8

HEAVISIDE'S TECHNIQUE FOR SOLVING MORE DIFFICULT ELECTRICAL PROBLEMS

137. The Object of Heaviside's Technique

In Chapter 7 we have considered mainly what happens when the driving voltage is a sine-wave alternating current of frequency $\omega/(2\pi)$, and we have assumed that a steady state has been reached, except for a brief investigation of "free oscillations", which once started can be maintained without any external driving e.m.f. We have regarded the currents and voltages involved as rotating vectors, and all the vectors have rotated at the same rate.

The object of the present chapter is to explain the technique introduced by Heaviside for use in more difficult cases. This technique was used by Heaviside in a heuristic or exploratory manner, and full mathematical justification was not available for twenty years or so. This need not deter us from following in the footsteps of Heaviside. If we can start with an act of faith in Heaviside and his assumptions (listed below), we can calculate the current I flowing in a circuit in which we are interested.

It is comparatively easy to prove by other methods, should we wish to do so, that I is the correct current; the difficulty, initially, is to have any idea of the sort of current to expect. At this stage we can think of Heaviside's technique as an extraordinarily powerful means of enlightened guesswork as to the behaviour of electrical circuits.

Heaviside was a very extraordinary and unusual man, but it is important to note that he was a practical engineer, not a professional mathematician. Indeed, for many years he was despised and rejected by contemporary Cambridge mathematicians. Heaviside's calculus came into being because he wanted to solve practical problems, although an advanced modern textbook on the subject may look like the work of a pure mathematician.

138. Heaviside's Assumptions

In order to understand Heaviside's assumptions, which are in effect a generalisation of the idea of "impedance" and of Ohm's Law, we shall have to be willing to make frequent excursions in both directions between the familiar world of time and space and a new world, where the most conspicuous symbol is the "Heaviside operator", p.

Our assumptions are concerned with the structure of the "p-world", and the means of communication in both directions between it and the time-world. The complete solution of a problem usually first requires a journey into the p-world, then algebraic manipulation within that world, and finally a translation of the results of this manipulation into terms of the time-world; but we shall see that in some cases, such as when we are considering stability, all the information we really require is available before the return to the world of time begins.

Our first assumption is that to any function of time (t) *supposed to be zero for negative values of t and to start suddenly when $t = 0$*, there corresponds a unique function of p, and this correspondence works both ways. An elementary list of pairs of corresponding functions is tabulated below:—

Time $f(t)$	a	$t^n/n!$	ϵ^{at}	$\epsilon^{at}-1$	$\sin at$
p-world $f[p]$	a	$1/p^n$	$p/(p-a)$	$a/(p-a)$	$pa/(p^2+a^2)$

$\cos at$	$\epsilon^{-at}\cos bt$	$\epsilon^{-at}\sin bt$
$p^2/(p^2+a^2)$	$p(p+a)/\{(p+a)^2+b^2\}$	$pb/\{(p+a)^2+b^2\}$

$t^{n-1}\epsilon^{-at}/(n-1)!$
$p/(p+a)^n$

Thus if a d.c. voltage V is suddenly applied to a circuit by the closing of a switch when $t = 0$, this is represented in the p-world also by V (see the first entry of the table; a is assumed to be a real constant throughout). If as a result of the p-world manipulations to be explained later, we find that the current is represented in the p-world by say $V/(p-a)$, then the current that is actually flowing is $(V/a)(\epsilon^{at}-1)$, from the fourth entry in the table; this current starts suddenly when $t = 0$.

This assumption establishes two-way communication between the two worlds (and the list given can of course be vastly extended, though it will be sufficient for understanding the fundamental ideas). The remaining assumptions hold *within the p-world* and it is because of their simplicity and generality within that world that Heaviside's technique works and leads to useful results.

Our second assumption, then, is that within the p-world, Ohm's Law applies universally, that is to say, the *back* voltage E_t produced when a current I_t flows in a network equivalent to a single impedance Z_p is given by

$$E_p = -Z_p \times I_p$$

where E_p is the p-function corresponding in our list to E_t, and I_p is the p-function corresponding in our list to I_t. Thus if we know I_t, we find I_p from our list, multiply it by Z_p which is a property of the network (it is in fact a generalisation of the "impedance operator" we have already met in Section 130) and then look for the product in the bottom row of our list to obtain E_t from the corresponding entry in the top row. If E_p is not in the list, algebraic manipulation will be required to break E_p up into a number of terms each of which is in the list.

It remains to determine Z_p. It will be sufficient if we determine the impedances of simple circuit elements, because our third assumption is that impedances in series or parallel can be combined in exactly the same way as in Section 134, that is to say, if a circuit consists of two elements having impedances Z_1 and Z_2 in series, these may be replaced by a single element having impedance $Z_1 + Z_2$, and if the same two elements are in parallel, they can be replaced by a single element of impedance Z where

$$(1/Z) = (1/Z_1) + (1/Z_2).$$

We have already noticed (in Section 126) that the *back* e.m.f. e_L due to a pure inductance is $-L(di/dt)$ when a varying current i flows through it. We now assume that the proper way to express this in the language of the p-world is to say that the impedance of the inductance is pL ohms when the inductance is L henrys. There is thus a close connection between p and d/dt.

The corresponding assumption for a resistance R ohms is simply that its impedance is R ohms, while for a capacitance of C farads the corresponding assumption is that the impedance is $1/(pC)$ ohms. This assumption implies that the operator $1/p$ corresponds to integration in the time world in much the same way as p itself corresponds to differentiation. But it is not worth while to pursue this correspondence at this stage. p and d/dt are in different worlds and they should remain in their own worlds; no attempt to communicate between the two worlds should be made except by means of the list of corresponding items already given, and the general rules for extending that list which follow. These general rules complete our equipment for investigating the effect of input voltages such as unit-step voltages which are not pure sine-wave alternating voltages, and enable us to consider the initial "transient" conditions as well as the ultimate "steady-state" conditions. The general rules are

(*i*) Within the p-world, p may be manipulated algebraically.

(*ii*) Both Kirchhoff's laws apply within the p-world as well as within the time-world.

(*iii*) If V_t corresponds in the list to V_p, then

(*a*) $\dfrac{dV_t}{dt}$ corresponds to $pV_p - p[V_t]_{t=0}$

$\dfrac{d^2V_t}{dt^2}$ corresponds to $p^2V_p - p^2[V_t]_{t=0} - p\left[\dfrac{dV_t}{dt}\right]_{t=0}$

$\dfrac{d^3V_t}{dt^3}$ corresponds to $p^3V_p - p^3[V_t]_{t=0} - p^2\left[\dfrac{dV_t}{dt}\right]_{t=0}$

$- p\left[\dfrac{d^2V_t}{dt^2}\right]_{t=0}$

and so on; the general result can be expressed

$$\frac{d^n V_t}{dt^n} \text{ corresponds to } p^n V_p - \sum_{s=0}^{n-1} p^{\,n-s} \left[\frac{d^s V_t}{dt^s}\right]_{t=0}$$

(b) $$\underset{t\to 0}{lt.}\, V_t = \underset{p\to\infty}{lt.}\, V_p$$

(c) $$\underset{t\to\infty}{lt.}\, V_t = \underset{p\to 0}{lt.}\, V_p$$

(d) $$\epsilon^{-\alpha t} V_t \text{ corresponds to } \frac{p}{p+\alpha}\, V_{p+\alpha}$$

The result (*iii*) (*a*) will be found particularly useful in connection with the solution of linear differential equations (Section 147). For a "dead" circuit, the result can be simplified to: $\frac{d^n V_t}{dt^n}$ corresponds to $p^n V_p$.

(*iii*) (*b*) and (*c*) often give us sufficient information about V_t and its derivatives if we know V_p, so that V_t need not be explicitly determined. In (*iii*) (*d*) the symbol $V_{p+\alpha}$ means that p is replaced by $p + \alpha$ in V_p.

It is perhaps worth noting in passing that if in our list $f(t)$ corresponds to f_p, then*

$$f_p = p \int_0^\infty e^{-pt} f(t)\,dt$$

provided that certain mild restrictions are satisfied by $f(t)$ and f_p. We shall not in fact use this general relation,

* For the benefit of the reader who is familiar with Laplace transforms, we note that these are closely allied to Heaviside's calculus. The Laplace transform ("one-sided") $L[f(t)]$ of a function $f(t)$ of t is defined by

$$L[f(t)] = \int_0^\infty e^{-st} f(t)\,dt$$

and $L[f(t)]$ is thus a function of s. If we multiply $L[f(t)]$ by s, and then replace s by p, we obtain f_p. It is highly desirable to keep different letters s and p for the two kinds of expression. In this book we have arbitrarily chosen to use only the Heaviside calculus in order to avoid confusing the reader. We could have chosen to use only the Laplace system; it is largely though not entirely a matter of personal preference. Impedances and transfer functions are the same in the two systems except insofar as p has to be replaced by s or vice versa.

as we shall only be concerned with straightforward cases.

We are now in a position to apply the technique to specific examples; in one case the input voltage is not a pure sine-wave alternating voltage, so that we could not have used the methods of Chapter 7, while in the other, we shall see how Heaviside's technique confirms what we already knew about steady-state conditions, and clearly disentangles steady-state and transient terms.

139. Application to the Series RC and Oscillatory Circuits

For our first example, consider the circuit of Fig. 66 with the inductance short-circuited; let e_p be the p-world equivalent of the source voltage, and i_p the p-world equivalent of the current. Then $-Ri_p$ is the p-world equivalent of the *back* voltage in the resistance and $\frac{1}{pC}$ is the impedance of the capacitor; the *back* voltage in the capacitance has the p-world equivalent $-i_p/(pC)$. Hence from Ohm's or Kirchhoff's laws we find

$$e_p - i_p\left(R + \frac{1}{pC}\right) = 0$$

so that

$$i_p = \frac{e_p}{R + (1/pC)} = \frac{pC\,e_p}{1 + pCR}$$

If e_t is a voltage e suddenly applied at time $t = 0$ and thereafter maintained constant, $e_p = e$ from the list of Section 138, and hence

$$i_p = \frac{e}{R}\,\frac{p}{p + \alpha}$$

where $\alpha = 1/(CR)$. It follows from the third entry in the list, with $-\alpha$ for a, that

$$i_t = \frac{e}{R}\,\epsilon^{-\alpha t}$$

In some cases we can overcome gaps in the list if we can expand i_p in a series of descending powers of p. Thus if we write

$$\frac{p}{p+\alpha} = \frac{1}{1+\frac{\alpha}{p}} = 1 - \frac{\alpha}{p} + \frac{\alpha^2}{p^2} - \dots$$

we obtain, by repeatedly using the second entry in the list,

$$i_t = \frac{e}{R}\left[1 - \alpha t + \frac{\alpha^2 t^2}{2!} - \dots\right]$$

which agrees with our previous result. At low frequencies we are usually interested in having α so small [CR so large] that the response is practically perfect, that is to say, the time terms in i_t are negligible.

Expansions of i_p in descending powers of p can often be obtained in more complicated circuits where compensation is required. By means of such expansions we may be able to determine values of the elements under our control so that the coefficient of t in i_t is zero, and perhaps the coefficients of higher powers of t also.

Our second example is to consider the effect on the complete series oscillatory circuit of Fig. 66 of suddenly applying the voltage e at time $t = 0$, and subsequently maintaining this voltage constant. Since the circuit consists of three elements in series, it is equivalent to a single impedance Z_p which is the sum of the impedances of the individual elements. Hence

$$Z_p = pL + R + (1/pC)$$

The p-world counterpart of the input voltage is e, and we therefore have, from the generalisation of Ohm's Law, that if the p-world counterpart of the current is I_p, the *back* voltage due to this current has the p-world counterpart

$$E_p = -Z_p \times I_p$$

and by Kirchhoff's second law

$$E_p + e = 0$$

It follows that

$$I_p = e/Z_p = \frac{pe}{Lp^2 + Rp + (1/C)}$$

We therefore have complete information about I_t provided that we can in some way reduce the expression for I_p to the sum of one or more terms in our list which have known counterparts in the time world. If, as is customary, we write

$$\omega_0^2 = 1/(LC); \ \alpha = R/(2L)$$

we have

$$I_p = \frac{e}{L} \cdot \frac{p}{p^2 + 2\alpha p + \omega_0^2} = \frac{e}{L} \cdot \frac{p}{(p+\alpha)^2 + (\omega_0^2 - \alpha^2)}$$

There are now two cases to consider, according as $\omega_0 > \alpha$ $[R^2 < 4L/C]$ or $\omega_0 < \alpha$ $[R^2 > 4L/C]$. In the former case, if we put ω for $(\omega_0^2 - \alpha^2)^{\frac{1}{2}}$, we have, from the list of Section 138

$$I_p = \frac{e}{L\omega} \cdot \frac{p\omega}{(p+\alpha)^2 + \omega^2} \text{ whence } I_t = \frac{e}{L\omega} \epsilon^{-\alpha t} \sin \omega t \text{ and}$$

in fact I_t is very similar to the current found in Section 131 which required no voltage to maintain it. The significance of this will be more easily understood in the light of the remaining examples.

In this case $(\omega_0 > \alpha)$ the denominator of I_p is a quadratic in p which has no real linear factors; in the contrary case, however, we must write k for $(\alpha^2 - \omega_0^2)^{\frac{1}{2}}$, so that $k < \alpha$, and I_p becomes

$$I_p = \frac{ep}{L} \cdot \frac{1}{(p + \alpha + k)(p + \alpha - k)}$$

We cannot use the list of Section 138 to deal with this expression as it stands, but we can break up

$$1/(p + \alpha + k)\ (p + \alpha - k)$$

into " partial fractions ", thus

$$\frac{1}{(p+\alpha+k)(p+\alpha-k)} = \frac{1}{(p+\alpha+k)\{-\alpha-k+\alpha-k\}}$$

$$+\frac{1}{\{-\alpha+k+\alpha+k\}(p+\alpha-k)} = \frac{1}{2k}\left[\frac{1}{p+\alpha-k} - \frac{1}{p+\alpha+k}\right]$$

and thus I_p reduces to the difference between two listed terms

$$I_p = \frac{e}{2kL}\left[\frac{p}{p+\alpha-k} - \frac{p}{p+\alpha+k}\right]$$

so that

$$I_t = \frac{e}{2kL}\left[\epsilon^{-(\alpha-k)t} - \epsilon^{-(\alpha+k)t}\right].$$

The process of breaking up the expression

$$1/\{(p+\alpha+k)(p+\alpha-k)\}$$

into "partial fractions" so that I_p becomes a series of terms in our list is of great importance. We are at liberty to keep the factor p in I_p apart from the expression to be broken up if we wish, since the breaking-up process is purely algebraic; alternatively, we would have obtained the same answer if we had broken up $p/\{(p+\alpha+k)(p+\alpha-k)\}$ into partial fractions. We have already come across partial fractions in a different connection in Section 118, where we wished to integrate the expression $1/(x^2-a^2)$ and similarly broke that up into "partial fractions".

The rule for determining the partial fractions is quite simple, and is indicated in the derivation above. If we require the "partial fraction" with a particular denominator $(p+\xi)$, we put $-\xi$ for p in all the rest of the expression to be split up, numerator and denominator, *except* the factor $(p+\xi)$ in the denominator. If the factors of the expression to be put in partial fractions are not all real as here, some adaptation of this technique is required, as will be indicated in the following examples, but it is always possible to obtain the correct result by proceeding as if the denominator of the expression to be put in partial fractions had factors which involve only real numbers. If subsequently complex numbers have to be substituted for some of them, this may complicate the algebra, but does not affect the general nature of the result, or the validity of the calculations for the factors which involve only real numbers.

We have so far omitted the border-line case of so-called "critical damping" when $R^2 = 4L/C$, $\omega = k = 0$, and $\alpha = R/(2L)$. In this case

$$I_p = \frac{e}{L} \cdot \frac{p}{(p+\alpha)^2}$$

and this is covered by the last entry in our list, so that

$$I_t = \frac{e}{L} \cdot t\epsilon^{-\alpha t}$$

This case, where some of the factors of the expression for which "partial fractions" are required are not distinct, is more difficult, and in general is outside our scope; this example merely indicates a possible way round the difficulty. Replacing equal factors by distinct factors having a small difference will very often indicate adequately what happens. Indeed, in the case considered, the three expressions derived for I_t differ very little for small and moderate t if k is small or if ω is small.

Now let us consider the case when the input voltage, instead of being e starting at $t = 0$, is $e \sin \Omega t$ starting at $t = 0$. The object of this example is to show how Heaviside's calculus confirms what we have already learnt elsewhere, and also gives us further information. We shall consider first the case in which $R^2 > 4L/C$ so that the α, k notation applies. We now have that the counterpart in the p-world of the input voltage is $ep\Omega/(p^2+\Omega^2)$ instead of e, while the impedance is unchanged. It follows that

$$I_p = \frac{e}{L} \cdot p \cdot \frac{p\,\Omega}{(p^2+\Omega^2)(p+\alpha+k)(p+\alpha-k)}$$

It is advisable, when possible, to keep a factor p outside the expression to be put in partial fractions as we have done, since so many of the listed expressions contain p as a factor. Now corresponding to each of the *real* factors $(p+\alpha+k)$ and $(p+\alpha-k)$ of the denominator of I_p, we can expect a partial fraction of the form $A/(p+\alpha+k)$ and $B/(p+\alpha-k)$ where A and B are obtained by putting $-\alpha-k$ and $-\alpha+k$ respectively for p in all of the expression

$$X(p) = \frac{p\Omega}{(p^2+\Omega^2)(p+\alpha+k)(p+\alpha-k)}$$

except the factor which would make the expression infinite. With the factor $(p^2+\Omega^2)$ in the denominator, the procedure is slightly different, since the expression $p^2+\Omega^2$ has no real linear factors in p. We can always assume that any factor of $X(p)$ which has degree greater than 1 gives rise to a partial fraction having that denominator and a numerator of degree one less, so that in this case the partial fraction to be expected can be taken as

$$\frac{Cp+D}{p^2+\Omega^2}$$

We thus expect to be able to find A, B, C, D so that

$$X(p) = \frac{A}{p+\alpha+k} + \frac{B}{p+\alpha-k} + \frac{Cp+D}{p^2+\Omega^2}$$

and we have already discussed finding A, B.

To find C, D, multiply through by $p^2+\Omega^2$ and then put $p = j\Omega$. We have

$$(p^2+\Omega^2)\,X(p) = \frac{p\Omega}{(p+\alpha+k)(p+\alpha-k)}$$

or $$\frac{p\Omega}{(p+\alpha+k)(p+\alpha-k)} =$$

$$\left(\frac{A}{p+\alpha+k} + \frac{B}{p+\alpha-k}\right)(p^2+\Omega^2) + Cp + D$$

so that, since $p = j\Omega$ implies $p^2+\Omega^2 = 0$, we have

$$\frac{j\Omega^2}{(j\Omega+\alpha+k)(j\Omega+\alpha-k)} = Cj\Omega + D$$

and $C\Omega$ is the imaginary part and D the real part of the complex number on the left-hand side. It follows that

$$I_t = \frac{e\Omega}{L}\left\{Ae^{(\alpha+k)t} + Be^{-(\alpha-k)t} + C\cos\Omega\, t + \frac{D}{\Omega}\sin\Omega\, t\right\}$$

where A, B, C, D have known values. Now it turns out, as it should, that C and D have just the values required to give I_t the "steady-state" terms already encountered in Section 129; I_t has, however, two additional terms which decay exponentially with time. These are the "transient" terms.

If now $R^2 < 4L/C$, it turns out that the only difference is that the "transient" terms become of the form $Ke^{-\alpha t}\cos(\omega t + \phi)$ already encountered in Section 131, and we now notice that in all the examples we have studied, transient terms have been one or more terms of the form $Ae^{\alpha t}$ where α is a real or complex quantity which, when substituted for p in Z_p, makes Z_p zero. If α has a positive real part, the corresponding transient term will grow with time and the system will be unstable. If Z_p is only quadratic in p, the two α's are easy to determine; we have, however, seen in Section 45 that it is difficult to solve an equation of degree higher than 2. Fortunately, we can determine whether any α's that make Z_p zero have positive real parts without finding all these α's. This is discussed in the next section.

140. Conditions for Stability

Let us now suppose that Z_p is quintic in p, that is to say

$$Z_p = a_5p^5 + a_4p^4 + a_3p^3 + a_2p^2 + a_1p + a_0$$

This is merely to fix our ideas; we shall later see how to adjust them when Z_p is of lower or higher degree. First, let us separate the terms of even degree from those of odd degree, so that we write

$$Z = p[a_5p^4 + a_3p^2 + a_1] + [a_4p^4 + a_2p^2 + a_0]$$

Now solve the quadratic equation

$$a_4p^4 + a_2p^2 + a_0 = 0$$

and let us call the roots $p^2 = -\alpha_1$, $p^2 = -\alpha_2$ $(\alpha_1 < \alpha_2)$;

α_1 and α_2 are thus determined only by the terms of even degree in Z_p. Correspondingly, solve the quadratic equation

$$a_5p^4 + a_3p^2 + a_1 = 0$$

and let us call the roots of this second equation $p^2 = -\beta_1$ and $p^2 = -\beta_2$ ($\beta_1 < \beta_2$). Then β_1 and β_2 are determined only by the terms of odd degree in Z_p. There will be stability for the system associated with impedance Z_p only if (*i*) all the a coefficients have the same sign [which we shall take to be positive]

(*ii*) α_1, α_2, β_1 and β_2 are real and positive

(*iii*) $\alpha_1 < \beta_1 < \alpha_2 < \beta_2$

The proof of this statement is somewhat outside our present scope, but it is a very powerful result which is easy to use because although Z_p has been assumed quintic, we have only to solve two quadratic equations to determine stability, and stability is out of the question if either of those quadratics has complex roots in p^2. In the above form, the result is easily applied when Z_p has a higher degree. If, for example, Z_p was of degree 8, the terms of even degree, treated as above, would form a quartic in p^2, so there would be four roots α_1, α_2, α_3 and α_4 (in ascending order) where in the above discussion we had two, and the terms of odd degree, treated as above, would form a cubic in p^2, so there would be three roots β_1, β_2 and β_3 (in ascending order) where in the above discussion we had only two. Conditions (*i*) and (*ii*) would be unaltered except that *all* the α's and β's must now be real, while condition (*iii*) becomes

$$\alpha_1 < \beta_1 < \alpha_2 < \beta_2 < \alpha_3 < \beta_3 < \alpha_4$$

Unfortunately it is not possible to express the condition for stability in terms of the coefficients a when the degree of Z_p is 5 or more, but if Z_p is of degree less than 5 the general result already given is equivalent to:

Z_p quadratic ($a_3 = a_4 = a_5 = 0$): All coefficients of the same sign is sufficient.

Z_p cubic ($a_4 = a_5 = 0$): $a_1a_2 > a_0a_3$.

Z_p quartic ($a_5 = 0$): $a_1a_2a_3 > a_4a_1^2 + a_0a_3^2$

141. Conditions for Adequate Damping

In the case of amplifiers (especially feedback amplifiers) we may be uncertain about stability; in the case of a passive network we know in advance that there must be stability. There may, however, be inadequate damping. If we can express our requirements in the form that all transient terms must die away at least as rapidly as ϵ^{-at} (a real and positive) then we can apply the test of the last section to see whether the damping is adequate without having to find all the values of p which make Z_p zero. For if the damping is to be at least as rapid as ϵ^{-at}, it is not sufficient that a real or complex quantity α which makes Z_p zero has a negative real part; it must have a real part less than $-a$, and therefore $(\alpha + a)$ must have a negative real part. In Z_p we therefore replace p by $(q - a)$, and work out the resulting q-equation, expanding powers of $(q - a)$ where necessary by means of the binomial theorem (Section 67). Thus if $Z_p = 1 + 3p + 4p^2 + 2p^3$ the condition of stability (for a cubic) is easily met since $a_1a_2 = 12$, $a_0a_3 = 2$; but if we required damping at least as rapid as $\epsilon^{-\frac{1}{2}t}$, we would replace p by $(q - \frac{1}{2})$ and evaluate

$$Z_q = 1 + 3(q - \tfrac{1}{2}) + 4(q - \tfrac{1}{2})^2 + 2(q - \tfrac{1}{2})^3$$
$$= \tfrac{1}{4} + \tfrac{1}{2}q + q^2 + 2q^3$$

and for this expression, $a_1a_2 = \frac{1}{2} = a_0a_3$, so that it is on the borderline; actually

$$Z_p = (1 + p)(1 + 2p + 2p^2)$$

so that Z_p is zero when $p = -1$ or $p = -\frac{1}{2} \pm \frac{1}{2}i$; the complex roots correspond to a transient term of the form $A\epsilon^{-\frac{1}{2}t}\cos(\frac{1}{2}t + \phi)$, and are therefore damped at the minimum rate.

142. Further Examples

Consider the circuit of Fig. 78. First of all, we must find the relation between V_p and i_p in the p-world in the general case. To do this, we observe that the right-hand inductance L and the terminating resistance R are in series, and therefore equivalent to a single impedance $pL + R = Z$. This impedance Z is in parallel with the

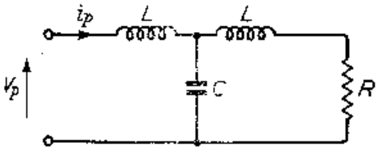


Fig. 78

capacitance C, so that the two together are equivalent to a single impedance z where

$$\frac{1}{z} = pC + \frac{1}{pL + R}$$

Finally, the left-hand inductance L is in series with z, so that the two together are equivalent to a single impedance ζ where

$$\zeta = pL + z = pL + \frac{pL + R}{p^2LC + pCR + 1}$$

and now the application of Ohm's Law to the circuit replaced by this single impedance ζ gives

$$V_p - \zeta i_p = 0$$

which reduces to

$$i_p = \frac{V_p}{\zeta} = \frac{V_p(p^2LC + pCR + 1)}{p^3L^2C + p^2LCR + 2pL + R}$$

We shall consider the case where a step-voltage V is suddenly applied at time $t = 0$, so that $V_p = V$. We now have to express i_p in partial fraction form, and in order to do this we first require to know the real factor $(p + \alpha)$ of the denominator. In a numerical case we could probably find this with sufficient accuracy by drawing a graph of the denominator as a function of p for negative values of p. (In Section 148 below, when dealing with numerical computation, we discuss how the accuracy of an approximation to the root of an equation can be improved where necessary.) In general, we know that there must be at least one real positive value of α. The remaining factor of the denominator will then be

$$X(p) = L^2Cp^2 + (LCR - \alpha L^2C)p + (R/\alpha)$$

since this expression, when multiplied by $(p + \alpha)$ gives the correct p^3, p^2 and numerical terms for the denominator of i_p, and will necessarily give the correct p-term as well if α has been found with sufficient accuracy. There will be a term in i_p of the form $A/(p + \alpha)$ where A is now found by putting $-\alpha$ for p in the numerator and in the factor $X(p)$ of the denominator. Now consider the expression

$$i_p - \frac{A}{p+\alpha} = \frac{V(p^2LC + pCR + 1) - AX(p)}{(p+\alpha)\,X(p)}$$

When we substitute the value for A found already, namely

$$\frac{V(\alpha^2LC - \alpha CR + 1)}{\alpha^2L^2C - (LCR - \alpha L^2C)\alpha + (R/\alpha)}$$

it will be found that a factor $(p + \alpha)$ cancels in the numerator and denominator, leaving us an expression of the form

$$\frac{Rp + Q}{X(p)}$$

If $X(p)$ has real linear factors in p, so that $X(p) = L^2C(p + \beta)(p + \gamma)$, this last expression can be further reduced to

$$\frac{Q - R\beta}{(\gamma - \beta)(p+\beta)} + \frac{Q - R\gamma}{(\beta - \gamma)(p+\gamma)}$$

or alternatively, since the denominator is now

$$L^2C(p + \alpha)(p + \beta)(p + \gamma)$$

we can say immediately

$$i_p = \frac{A}{p+\alpha} + \frac{V(\beta^2LC - \beta CR + 1)}{(\alpha - \beta)(p+\beta)(\gamma - \beta)L^2C}$$

$$+ \frac{V(\gamma^2LC - \gamma CR + 1)}{(\alpha - \gamma)(\beta - \gamma)(p+\gamma)L^2C}$$

It is well worth verifying that in a particular case the same result is obtained by both methods. In the case when $X(p)$ has real linear factors, if we abbreviate i_p to

$$i_p = \frac{A}{p+\alpha} + \frac{Y}{p+\beta} + \frac{Z}{p+\gamma}$$

it follows from the list in Section 137 that

$$i_t = \frac{A}{\alpha}(1 - e^{-\alpha t}) + \frac{Y}{\beta}(1 - e^{-\beta t}) + \frac{Z}{\gamma}(1 - e^{-\gamma t})$$

and if β and γ happen to be complex (which means that Y and Z are also complex) this result is still formally correct, but it is easier to combine the last two terms of i_p in the form obtained earlier

$$\frac{Rp+Q}{X(p)} = \frac{Rp+Q}{L^2C\{(p+\xi)^2+\omega^2\}}$$

We now choose λ, μ, ν so that this shall equal

$$\frac{\lambda p(p+\xi)+\mu p\omega+\nu\{(p+\xi)^2+\omega^2\}}{L^2C\{(p+\xi)^2+\omega^2\}}$$

so that the last two terms of i_t become

$$\frac{1}{L^2C}\{\lambda e^{-\xi t}\cos\omega t+\mu e^{-\xi t}\sin\omega t+\nu\}$$

We find, from the coefficients of p^2, p and 1 in the numerator,

$$\lambda+\nu=0;\ \ \lambda\xi+\mu\omega+2\nu\xi=R;\ \ \nu(\xi^2+\omega^2)=Q$$

so that
$$\nu=-\lambda=Q/(\xi^2+\omega^2)$$

$$\mu=\left\{R-\frac{Q\xi}{\xi^2+\omega^2}\right\}/\omega$$

A further example involving a circuit not initially "dead" is discussed in Section 147 below; this example illustrates the application of Heaviside's technique to linear differential equations with constant coefficients which need not necessarily be of electrical origin.

The following examples show the power of the method described in Section 140 to determine conditions of stability. First, consider a system equivalent to a single impedance

$$Z_p = p^5 + 5p^4 + 3p^3 + 20p^2 + 2p + 18{\cdot}75$$

Here we have, in the notation of Section 140, $\alpha_1 = 1{\cdot}5$, $\alpha_2 = 2{\cdot}5$, $\beta_1 = 1$, $\beta_2 = 2$. Hence the condition $\alpha_1 < \beta_1 < \alpha_2 < \beta_2$ is not met and the system is unstable (or, the equation $Z_p = 0$ has a root in p with positive real part). If now, on the other hand,

$$Z_p = 5p^5 + p^4 + 20p^3 + 3p^2 + 18{\cdot}75p + 2$$

we have $\alpha_1 = 1$, $\alpha_2 = 2$, $\beta_1 = 1{\cdot}5$, $\beta_2 = 2{\cdot}5$ so that $\alpha_1 < \beta_1 < \alpha_2 < \beta_2$ and the system is stable.

Now consider the system for which

$$Z_p = 5p^5 + p^4 + 20p^3 + 3p^2 + 18{\cdot}75p + \lambda$$

(This type of Z_p can occur with feedback amplifiers.) What restrictions are placed on λ by stability requirements?

Here α_1 and α_2 depend on λ, but $\beta_1 = 1{\cdot}5$ and $\beta_2 = 2{\cdot}5$ are fixed. α_1 or α_2 is equal to β_1 if $2{\cdot}25 - 3 \times 1{\cdot}5 + \lambda = 0$, or $\lambda = 2{\cdot}25$; α_1 or α_2 is equal to β_2 if $6{\cdot}25 - 3 \times 2{\cdot}5 + \lambda = 0$, or $\lambda = 1{\cdot}25$. Hence for stability, λ must lie between $1{\cdot}25$ and $2{\cdot}25$. If $\lambda = 2{\cdot}25$, $\alpha_1 = \alpha_2 = \beta_1 = 1{\cdot}5$; for $\lambda > 2{\cdot}25$, α_1 and α_2 are complex; if $\lambda = 1{\cdot}25$, $\alpha_1 = 0{\cdot}5$ and $\alpha_2 = 2{\cdot}5$. For smaller λ, instability arises because α_2 is greater than β_2.

Chapter 9

MISCELLANEOUS TECHNIQUES

143. General Objective of Chapter

We have now assembled sufficient mathematical equipment to be able to solve simple radio problems and understand the kind of solution that is to be expected in many other cases. There are, however, a number of subjects, such as matrices, which, although not inevitably compulsory for a radio engineer, nevertheless greatly assist his work if he can grasp the elements of them.

The aim of this chapter is to discuss in outline a few such topics, so that they need not be any longer a hindrance when the radio engineer suddenly finds himself confronted by them. Textbooks can be found easily which deal in much greater detail with all these subjects, and others for which there is no space here; it is, however, the first step in a new direction which is often critical from the point of view of confidence and morale. We have, therefore, tried to collect a few useful signposts for subjects bordering on radio.

144. Matrices: The Rules of Matrix Algebra

The first of these subjects is matrices. In its most general form, this is undoubtedly a subject best left to pure mathematicians, but the matrices relevant to radio have seldom more than two rows and columns, and those discussed here never have. This enormously simplifies the whole subject, so that an engineer can learn the necessary technique for radio purposes in a few hours.

Consider then a four-terminal network as in Fig. 79 enclosed within a "black box" B. We are not allowed to ask questions as to what the box contains, only about what it does; this is known to us merely through observation of the input voltage V_1, the output voltage V_2, the input current I_1 and the output current I_2. We shall

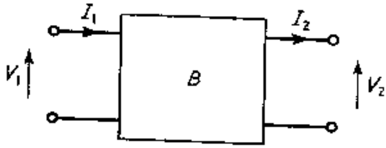


Fig. 79

consistently use the sign convention of Fig. 79, so that I_1 is positive if current is flowing *into* the network in the direction of the arrow, but I_2 is positive if current is flowing *out of* the network in the direction of the arrow. V_1 and V_2 are both positive if the upper terminal is at a higher potential than the corresponding lower terminal. If the network is reversed, V_1 and V_2 are exchanged, but I_2 becomes $-I_1$ and I_1 becomes $-I_2$. Assuming that the "black box" B is a *linear* network, the principle of superposition applies, and there will be two linear relations between the variables V_1, I_1, V_2 and I_2. We can write these relations down by expressing any two of the variables in terms of the other two; thus, if we express V_1 and I_1 in terms of V_2 and I_2, we have

$$V_1 = a_{11}V_2 + a_{12}I_2$$
$$I_1 = a_{21}V_2 + a_{22}I_2$$

where a_{11}, a_{12}, a_{21} and a_{22} are coefficients depending on the nature of the network, but not on V_1, I_1, V_2 and I_2.

We must regard the two equations as holding entirely within the p-world, and V_1, I_1, V_2 and I_2 as the p-world counterparts of the input and output current and voltage; the a's may thus be functions of p. Under steady-state conditions at a fixed frequency $\omega/(2\pi)$, p may be replaced by $j\omega$, so that the a's will depend upon the frequency. The coefficient a_{12} may be regarded as an impedance, while the coefficient a_{21} may correspondingly be regarded as an admittance, or the reciprocal of an impedance. a_{11} and a_{22} are in effect pure numbers, in spite of the fact that they may depend upon p.

Now in matrix language, the two relations expressing V_1, I_1 in terms of V_2, I_2 are replaced by a single relation which is written $\begin{pmatrix} V_1 \\ I_1 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} V_2 \\ I_2 \end{pmatrix}$

Initially we regard this new way of writing the relation as a mere shorthand. It is only a useful shorthand because it is possible to build up an algebra of these new symbols

$$\begin{pmatrix} V_1 \\ I_1 \end{pmatrix}, \quad \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \text{ etc.}$$

by means of which they can be combined. In Section 138, when we first encountered the p-world, we likewise constructed an algebra of impedances by initial arbitrary assumptions that Ohm's Law held universally in the p-world, that a single resistance, inductance and capacitance each had a known impedance, and that there were known rules for the combination of impedances in series or in parallel.

We are thus becoming gradually more accustomed to the fact that progress in the understanding of radio and electrical problems can often by made by initially taking upon trust certain definitions and rules, and obeying them implicitly. Our faith in the usefulness of these rules and arbitrary definitions is strengthened as we find that they first give us results whose correctness is easily seen by other methods, and finally increase our understanding of situations very difficult to disentangle by other methods.

The above preliminary remarks should be sufficient preparation for the following arbitrary rules of matrix algebra:

(*a*) Expressions like

$$\begin{pmatrix} V_1 \\ I_1 \end{pmatrix}, \quad \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \text{ and } (P\ Q)$$

are called respectively column, square and row matrices. Row matrices will not in practice concern us here; the definition and one or two fundamental properties are merely included for the sake of completeness.

(*b*) The quantities like V_1, I_1, a_{11}, etc., in the above matrices are called elements; their position in the matrix as well as their magnitude is significant. Two matrices are equal if and only if they are both column matrices or both square matrices, and elements in a corresponding position are equal. The significance of the "equation" we have already written, namely

$$\begin{pmatrix} V_1 \\ I_1 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} V_2 \\ I_2 \end{pmatrix}$$

is not apparent until we can define a procedure for multiplying two matrices so that the right-hand side can be replaced by a single matrix. This multiplication procedure will be given shortly.

(*c*) Two matrices may be added if they are of the same kind (i.e. both column or both square) and the sum is the matrix obtained by adding elements in corresponding positions in the two matrices.

(*d*) A matrix can easily be multiplied by a number; it is merely necessary to multiply every element of the matrix by that number. But the multiplication of two matrices follows a more difficult rule. The important point is that not all pairs of matrices can be multiplied, and that if a matrix A can be multiplied by B, it does not necessarily happen that B can be multiplied by A, or that if this can happen, the two products are equal. The rule for multiplying two square matrices is

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \times \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{pmatrix}$$

whereas

$$\begin{pmatrix} b_{11}b_{12} \\ b_{21}b_{22} \end{pmatrix} \times \begin{pmatrix} a_{11}a_{12} \\ a_{21}a_{22} \end{pmatrix} = \begin{pmatrix} a_{11}b_{11} + a_{21}b_{12} & a_{12}b_{11} + a_{22}b_{12} \\ a_{11}b_{21} + a_{21}b_{22} & a_{12}b_{21} + a_{22}b_{22} \end{pmatrix}$$

A square matrix can be multiplied by a column matrix if the square matrix comes first; the result is a column matrix analogous to the above when $b_{12} = b_{22} = 0$, namely

$$\begin{pmatrix} a_{11}a_{12} \\ a_{21}a_{22} \end{pmatrix} \times \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} a_{11}c_1 + a_{12}c_2 \\ a_{21}c_1 + a_{22}c_2 \end{pmatrix}$$

and similarly a row matrix can be multiplied by a square matrix if the row matrix comes first. A row matrix can be multiplied by a column matrix only if the row matrix comes first, and the result is simply a number

$$(r_1 r_2) \times \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = r_1 c_1 + r_2 c_2.$$

(*e*) It is sometimes necessary to raise a matrix to a power, and the usual procedure for doing this, applicable to matrices having any number of rows and the same number of columns, is outside our present scope. However, the following ingenious special method, applicable to matrices having two rows and columns only, is due to W. Proctor Wilson, and has recently been published.* The key point in this method is that the matrix product

$$\begin{pmatrix} \dfrac{\cosh(mu+v)}{\cosh v}, & \mu\dfrac{\sinh mu}{\cosh v} \\ \dfrac{\sinh mu}{\mu \cosh v}, & \dfrac{\cosh(mu-v)}{\cosh v} \end{pmatrix} \begin{pmatrix} \dfrac{\cosh(nu+v)}{\cosh v}, & \mu\dfrac{\sinh nu}{\cosh v} \\ \dfrac{\sinh nu}{\mu \cosh v}, & \dfrac{\cosh(nu-v)}{\cosh v} \end{pmatrix}$$

reduces to the specially simple form

$$\begin{pmatrix} \dfrac{\cosh([m+n]u+v)}{\cosh v}, & \mu\dfrac{\sinh(m+n)u}{\cosh v} \\ \dfrac{\sinh(m+n)u}{\mu \cosh v}, & \dfrac{\cosh([m+n]u-v)}{\cosh v} \end{pmatrix}$$

and therefore, by first putting $m = n = 1$, then $m = 2$, $n = 1$, then $m = 3$, $n = 1$, then $m = 4$, $n = 1$ and so on, we can deduce

$$\begin{pmatrix} \dfrac{\cosh(u+v)}{\cosh v} & \mu\dfrac{\sinh u}{\cosh v} \\ \dfrac{\sinh u}{\mu \cosh v} & \dfrac{\cosh(u-v)}{\cosh v} \end{pmatrix}^m =$$

$$\begin{pmatrix} \dfrac{\cosh(mu+v)}{\cosh v}, & \mu\dfrac{\sinh mu}{\cosh v} \\ \dfrac{\sinh mu}{\mu \cosh v}, & \dfrac{\cosh(mu-v)}{\cosh v} \end{pmatrix}$$

We shall not prove this result here. It depends upon the

* "Some matrix theorems," *Electronic & Radio Engineer*, Vol. 34, No. 6, 1957.

repeated use of "sums and products" formulæ for hyperbolic sines and cosines, namely

$$\sinh x + \sinh y = 2 \sinh \frac{x+y}{2} \cosh \frac{x-y}{2}$$

$$\sinh x - \sinh y = 2 \cosh \frac{x+y}{2} \sinh \frac{x-y}{2}$$

$$\cosh x + \cosh y = 2 \cosh \frac{x+y}{2} \cosh \frac{x-y}{2}$$

$$\cosh x - \cosh y = 2 \sinh \frac{x+y}{2} \sinh \frac{x-y}{2}$$

similar to the corresponding trigonometrical formulæ. These formulæ can be deduced directly from the definitions of hyperbolic sines and cosines in Section 100. The important point is that we can find parameters derivable from the elements of a 2×2 matrix so that the mth power of that matrix can be written down explicitly. The usual method of raising a matrix to a power involves explicitly the difficult concept of latent roots, which we have been at pains to avoid here.

However neat the above result might be, it would be useless unless we could determine the parameters u, v, μ for a given matrix

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

First, let Δ be the quantity $a_{11}a_{22} - a_{21}a_{12}$ which is called the "determinant" of the matrix A. Take out $\Delta^{\frac{1}{2}}$ as a numerical factor from the matrix, so that

$$A = \Delta^{\frac{1}{2}} \begin{pmatrix} \frac{a_{11}}{\Delta^{\frac{1}{2}}} & \frac{a_{12}}{\Delta^{\frac{1}{2}}} \\ \frac{a_{21}}{\Delta^{\frac{1}{2}}} & \frac{a_{22}}{\Delta^{\frac{1}{2}}} \end{pmatrix}$$

The reason for doing this is that the determinant of

$$M = \begin{pmatrix} \dfrac{\cosh(u+v)}{\cosh v} & \mu \dfrac{\sinh u}{\cosh v} \\ \dfrac{1}{\mu} \dfrac{\sinh u}{\cosh v} & \dfrac{\cosh(u-v)}{\cosh v} \end{pmatrix}$$

happens to be equal to 1 (this can be proved by means of the formulæ involving hyperbolic sines and cosines already mentioned). We now equate corresponding elements of M and $A/\Delta^{\frac{1}{2}}$. This apparently gives us four equations for three unknowns u, v, μ, but any three equations derived from these four will be sufficient. By further use of the formulæ involving hyperbolic sines and cosines already mentioned, we can simplify two of these equations to

$$\cosh u = \frac{a_{11} + a_{22}}{2\Delta^{\frac{1}{2}}}; \quad \mu^2 = \frac{a_{12}}{a_{21}}$$

and knowing u and μ, v is then easily deduced from

$$a_{12}/\Delta^{\frac{1}{2}} = \mu \sinh u/\cosh v$$

and we thus finally obtain the result

$$A^n = \Delta^{n/2} \begin{pmatrix} \dfrac{\cosh(nu+v)}{\cosh v} & \mu \dfrac{\sinh nu}{\cosh v} \\ \dfrac{\sinh nu}{\mu \cosh v} & \dfrac{\cosh(nu-v)}{\cosh v} \end{pmatrix}$$

where u, v, μ have the values just mentioned. This last formula holds even for non-integral values of n: this is relevant in connection with the theory of cables and transmission lines.

(*f*) A matrix $[A]$ is multiplied by a number k if all its elements are multiplied by k (as already mentioned under (*d*) above). The matrix thus obtained is the same as if the rule for matrix multiplication given in (*d*) above is applied to the matrix multiplications

$$[A] \times [K] \text{ or } [K] \times [A]$$

where

$$[K] = \begin{pmatrix} k & 0 \\ 0 & k \end{pmatrix}$$

and this is one of the exceptional cases where the order of a matrix multiplication does not matter. We have already introduced the "determinant" of a matrix in (*e*) above in connection with matrix powers. If two matrices are multiplied, the determinant of the product matrix is the product of the determinants of the separate matrices and therefore in particular if $[A]$ is multiplied by k, the determinant of $[A]$ is multiplied by k^2.

(*g*) The matrices

$$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \text{ and } \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

are called respectively the null matrix and the unit matrix; the latter is usually denoted by I. They correspond in matrix algebra to the numbers 0 and 1 in ordinary algebra.

The above is a collection of the rules of matrix algebra which we shall require, but before we can apply them, we need to know how matrices can be combined when the networks with which they are associated are connected in various ways. For it was the possibility of regarding a complicated network as equivalent to a single impedance derived by the laws of series and parallel combination given in Section 138 that enabled us to use the generalisation of Ohm's Law so effectively. Similarly, with matrices, it is the possibility of reducing the complicated relations within a four-terminal network as in Fig. 79 to a single matrix relation between input and output current and voltage that makes matrices such a powerful tool for radio and electrical investigation. The laws of combination are given in the next section.

145. Matrices : The Laws of Combination

We have already noted in connection with Fig. 79 the possibility of expressing the relations between input and output current and voltage in the form

$$V_1 = a_{11}V_2 + a_{12}I_2$$
$$I_1 = a_{21}V_2 + a_{22}I_2$$

for which the matrix "shorthand" is

$$\begin{pmatrix} V_1 \\ I_1 \end{pmatrix} = \begin{pmatrix} a_{11}a_{12} \\ a_{21}a_{22} \end{pmatrix} \begin{pmatrix} V_2 \\ I_2 \end{pmatrix}$$

The matrix

$$A = \begin{pmatrix} a_{11}a_{12} \\ a_{21}a_{22} \end{pmatrix}$$

is usually known as the "*A*-matrix" of the network. If the network is passive and obeys the law of reciprocity,

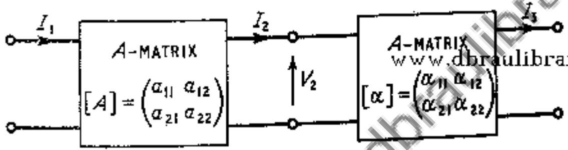


Fig. 80

the determinant of the *A*-matrix is always unity. Now suppose that we have two networks connected in cascade, as in Fig. 80. Then it can be shown by substitution that

$$\begin{pmatrix} V_1 \\ I_1 \end{pmatrix} = \begin{pmatrix} a_{11}a_{12} \\ a_{21}a_{22} \end{pmatrix} \begin{pmatrix} V_2 \\ I_2 \end{pmatrix} \text{ and } \begin{pmatrix} V_2 \\ I_2 \end{pmatrix} = \begin{pmatrix} \alpha_{11}\alpha_{12} \\ \alpha_{21}\alpha_{22} \end{pmatrix} \begin{pmatrix} V_3 \\ I_3 \end{pmatrix}$$

implies

$$\begin{pmatrix} V_1 \\ I_1 \end{pmatrix} = [A\alpha] \begin{pmatrix} V_3 \\ I_3 \end{pmatrix}$$

where $[A\alpha]$ is the matrix product $[A] \times [\alpha]$ obtained by the product rule (*d*) of Section 144.

Next, suppose that we have the two networks connected in "parallel-parallel", as in Fig. 81.

Then first we need to rearrange our network equations connecting input and output current and voltage for a single network as in Fig. 79, so that instead of expressing V_1 and I_1 in terms of V_2 and I_2, we express I_1 and I_2 in terms of V_1 and V_2.

We find

$$I_1 = y_{11}V_1 + y_{12}V_2$$
$$I_2 = y_{21}V_1 + y_{22}V_2$$

or, in matrix notation,

$$\begin{pmatrix} I_1 \\ I_2 \end{pmatrix} = [Y] \begin{pmatrix} V_1 \\ V_2 \end{pmatrix}$$

where

$$y_{11} = a_{22}/a_{12} \qquad y_{12} = a_{21} - \frac{a_{22}a_{11}}{a_{12}}$$

$$y_{21} = \frac{1}{a_{12}} \qquad y_{22} = -\frac{a_{11}}{a_{12}}$$

and

$$[Y] = \begin{pmatrix} y_{11} y_{12} \\ y_{21} y_{22} \end{pmatrix}$$

Notice that if the network is passive and obeys the law of reciprocity, so that the determinant of the A-matrix is unity, then $y_{12} = -y_{21}$. The matrix $[Y]$ is called the admittance or Y-matrix of the network of Fig. 79. If in Fig. 81, the upper network has admittance matrix Y and the lower one has admittance matrix

$$[\eta] = \begin{pmatrix} \eta_{11} & \eta_{12} \\ \eta_{21} & \eta_{22} \end{pmatrix}$$

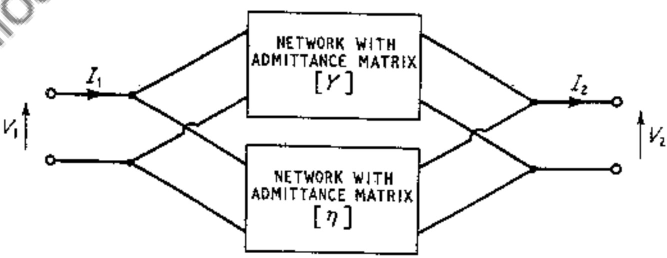


Fig. 81

then for the combined network connected in "parallel-parallel" as in Fig. 81 the admittance matrix is obtained by adding the admittance matrices of the separate networks, that is to say

$$\begin{pmatrix} V_1 \\ I_1 \end{pmatrix} = \begin{pmatrix} y_{11} + \eta_{11} & y_{12} + \eta_{12} \\ y_{21} + \eta_{21} & y_{22} + \eta_{22} \end{pmatrix} \begin{pmatrix} V_2 \\ I_2 \end{pmatrix}$$

The other important combination of networks is "series-series" as in Fig. 82. In this case our original equations

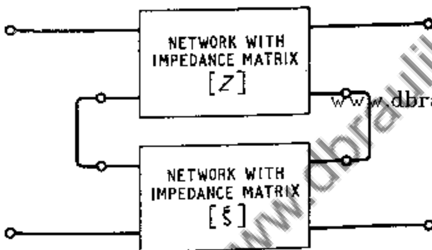


Fig. 82

for the single network of Fig. 79 need rearranging so that V_1 and V_2 are expressed in terms of I_1 and I_2; the result can be written

$$\begin{pmatrix} V_1 \\ V_2 \end{pmatrix} = [Z] \begin{pmatrix} I_1 \\ I_2 \end{pmatrix}$$

and the matrix this time is called the Z or impedance matrix. For networks connected as in Fig. 82, it is the Z-matrices which should be added to give V_1, V_2 in terms of I_1, I_2. The derivation of the Z-matrix from the A-matrix is similar to the derivation of the Y-matrix from the A-matrix, and is therefore not given in detail; it is more convenient to summarise the results of such changes from one kind of matrix to another for a single network by means of the Table overleaf. $|A|$ means the determinant of the A-matrix, $|Y|$ that of the Y-matrix, and $|Z|$ that of the Z-matrix.

	A	Y	Z
A	$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$	$\frac{1}{a_{12}}\begin{pmatrix} a_{22} & -\lvert A\rvert \\ 1 & -a_{11} \end{pmatrix}$	$\frac{1}{a_{21}}\begin{pmatrix} a_{11} & -\lvert A\rvert \\ 1 & -a_{22} \end{pmatrix}$
Y	$\frac{1}{y_{21}}\begin{pmatrix} -y_{22} & 1 \\ -\lvert Y\rvert & y_{11} \end{pmatrix}$	$\begin{pmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{pmatrix}$	$\frac{1}{\lvert Y\rvert}\begin{pmatrix} y_{22} & -y_{12} \\ -y_{21} & y_{11} \end{pmatrix}$
Z	$\frac{1}{z_{21}}\begin{pmatrix} z_{11} & -\lvert Z\rvert \\ 1 & -z_{22} \end{pmatrix}$	$\frac{1}{\lvert Z\rvert}\begin{pmatrix} z_{22} & -z_{12} \\ -z_{21} & z_{11} \end{pmatrix}$	$\begin{pmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \end{pmatrix}$

Thus if we know that the A-matrix of a network is

$$[A] = \begin{pmatrix} a_{11}a_{12} \\ a_{21}a_{22} \end{pmatrix}$$

the Y-matrix is found in terms of the a's by means of the elements in the A-row and the Y-column of the table, and so on.

We are now in a position to build up the matrix of a complicated four-terminal network by breaking it up into a number of component parts connected in series-series, parallel-parallel or in cascade, and this will be done for a few simple but important networks in the next section. The Table just given enables us to find all the matrices, A, Y and Z, of any network for which we know one of those matrices, so that only one matrix will be found for each of the networks.

146. Matrices of Elementary Networks

Consider first the network of Fig. 83, in which there is a single series element having impedance Z and no shunt element. Then it is clear from Ohm's and Kirchhoff's Laws that

$$I_1 = I_2$$
$$V_1 - V_2 = I_2 Z$$

so that the A-matrix is

$$\begin{pmatrix} V_1 \\ I_1 \end{pmatrix} = \begin{pmatrix} 1 & Z \\ 0 & 1 \end{pmatrix}\begin{pmatrix} V_2 \\ I_2 \end{pmatrix}$$

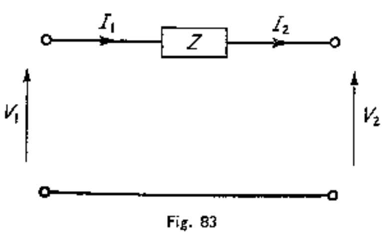


Fig. 83

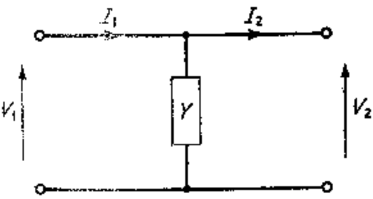


Fig. 84

Similarly, for the single shunt element of Fig. 84, having admittance Y (impedance $1/Y$), the A-matrix is found from

$$\begin{pmatrix} V_1 \\ I_1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ Y & 1 \end{pmatrix} \begin{pmatrix} V_2 \\ I_2 \end{pmatrix}$$

More complicated networks can be built up by placing the networks of Figs. 83 and 84 in cascade in a suitable order; we shall obtain the A-matrix for a general T-network and a general π-network only, and briefly consider ladder networks.

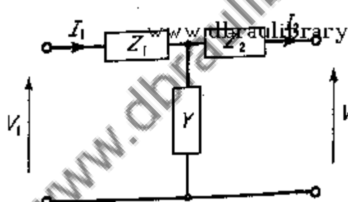


Fig. 85

Thus the T-network of Fig. 85 can be regarded as three networks in cascade, namely (*i*) a network like Fig. 83 with Z_1 instead of Z, (*ii*) a network like Fig. 84 (it is convenient to use admittances for all shunt elements and impedances for all series elements, and this will be done automatically henceforward), (*iii*) a network like Fig. 83 with Z_2 instead of Z.

Now $$\begin{pmatrix} 1 & Z_1 \\ 0 & 1 \end{pmatrix} \times \begin{pmatrix} 1 & 0 \\ Y & 1 \end{pmatrix} = \begin{pmatrix} 1 + YZ_1 & Z_1 \\ Y & 1 \end{pmatrix}$$

and

$$\begin{pmatrix} 1+YZ_1 & Z_1 \\ Y & 1 \end{pmatrix} \begin{pmatrix} 1 & Z_2 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1+YZ_1 & Z_1+Z_2+YZ_1Z_2 \\ Y & 1+YZ_2 \end{pmatrix}$$

so that, for the T-network of Fig. 85

$$\begin{pmatrix} V_1 \\ I_1 \end{pmatrix} = \begin{pmatrix} 1+YZ_1 & Z_1+Z_2+YZ_1Z_2 \\ Y & 1+YZ_2 \end{pmatrix} \begin{pmatrix} V_2 \\ I_2 \end{pmatrix}$$

Correspondingly, the π-network of Fig. 86 is also equivalent to three networks in cascade, namely (*i*) a network like Fig. 84 with the admittance Y replaced by Y_1, (*ii*) a network like Fig. 83, (*iii*) a network like Fig. 84 with Y_2 instead of Y. We thus find, by matrix multiplication as before, that for the π-network of Fig. 86,

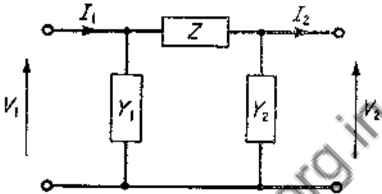


Fig. 86

$$\begin{pmatrix} V_1 \\ I_1 \end{pmatrix} = \begin{pmatrix} 1 + Y_1 Z & Z \\ Y_1 + Y_2 + ZY_1Y_2 & 1 + Y_2 Z \end{pmatrix} \begin{pmatrix} V_2 \\ I_2 \end{pmatrix}$$

Finally, consider a ladder network like that of Fig. 87, having n sections. The A-matrix for one section (Fig. 88) is already known, being the product of the A-matrix for Fig. 83 followed by the A-matrix for Fig. 84, namely

$$[A] = \begin{pmatrix} 1 + YZ & Z \\ Y & 1 \end{pmatrix}$$

Hence the A-matrix for the whole ladder is obtained directly by Wilson's method [Section 144, (*e*)]. Its determinant is obviously unity as it should be, so we put

$$\mu = (Z/Y)^{\frac{1}{2}}, \ \cosh u = 1 + \tfrac{1}{2}YZ, \ \cosh v = (1 + \tfrac{1}{4}YZ)^{\frac{1}{2}}$$

and find that v happens to be $\frac{1}{2}u$. [This can be deduced again from the "sums and products" formulæ for hyperbolic sines and cosines given in Section 144 (*e*).]

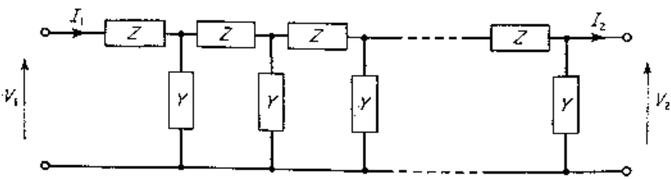


Fig. 87

The A-matrix for the whole ladder network is thus

$$[A]^n = \begin{pmatrix} \dfrac{\cosh(n+\frac{1}{2})u}{\cosh \frac{1}{2}u}, & \mu \dfrac{\sinh nu}{\cosh \frac{1}{2}u} \\ \dfrac{\sinh nu}{\mu \cosh \frac{1}{2}u} & \dfrac{\cosh(n-\frac{1}{2})u}{\cosh \frac{1}{2}u} \end{pmatrix}$$

Having thus obtained the A-matrix for the whole of the network of Fig. 87, any termination at either end can be regarded as a simple network in cascade with the network of Fig. 87, and we can therefore easily deduce the A-matrix for the whole system including terminations.

The above examples must be regarded as illustrating rather than fully exploiting the power of matrix algebra to simplify the theory of four-terminal networks. The A-matrix associated with a given four-terminal network gives us immediately almost all the essential information about the network. In the A-matrix for Fig. 84, the element a_{12} was the series impedance Z and the element a_{21} was the shunt admittance Y. Now in general there are many circuits which can be regarded as "equivalent" to a given four-terminal network, but the A-matrices of all these equivalent circuits will all have the same elements. The element a_{12} may, by analogy with Fig. 84, be regarded as the "series-ness" of the network, and the element a_{21} may be regarded as the "shunt-ness". If the network is passive, there is only one other essential element, since the fourth element of the A-matrix is determined by the other three. If a four-terminal network is symmetrical, as in Fig. 85 when $Z_1 = Z_2$, or Fig. 86 when $Y_1 = Y_2$, we find that $a_{11} = a_{22}$. An unsymmetrical network can be regarded as a symmetrical network preceded or followed in cascade by an ideal transformer of ratio $(a_{22}/a_{11})^{\frac{1}{2}}$; this

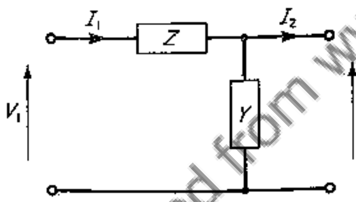


Fig. 88

quantity can therefore be regarded as the "transformer-ness" of the network. Thus the four elements of the A-matrix of a network give us uniquely three essential items of information, namely its "series-ness", "shunt-ness" and "transformer-ness". Any of the other matrices would serve equally well, since the Table of Section 145 enables us to convert readily from one matrix to another.

147. Linear Differential Equations with Constant Coefficients

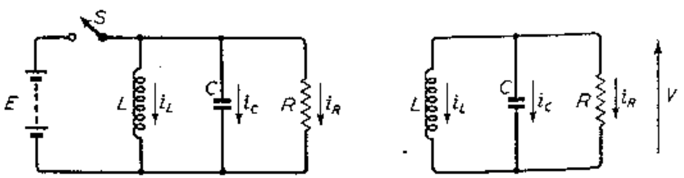
This subject is closely related to the "p-world" considered in Section 138, and to illustrate and explain this relation, we first shall consider in detail an electrical example.

Consider the circuit of Fig. 89 (*a*), which is a highly idealised form of the circuit commonly used in the line timebase of a television receiver. A practical circuit has many added complications which make it less amenable to calculation, but it functions basically in the manner of the idealised circuit.

Initially the circuit is "dead", but on closing the switch S, the battery E is connected, and currents i_L, i_c and i_r flow respectively in L, C and R. Since we are ignoring all resistance in series with L, C, and E (which in practice would inevitably be present in some degree) we have, by Ohm's Law

$$E - pL.i_{Lp} = 0$$

for the inductance, if i_{Lp} is the p-world equivalent of the



(a) (b)

Fig. 89

current i_L. Consequently,

$$i_{Lp} = E/pL$$

and so

$$i_L = Et/L$$

Now for the resistance, clearly $i_R = E/R$, while for the capacitance, writing i_{cp} for the p-world equivalent of i_c, we have

$$i_{cp} - pCE = 0$$

but here we are unable to determine i_c directly by means of the list in Section 138. If i_{cp} had been CE, i_c would have been a step-function like i_R. Now we have already seen that multiplication by p is closely associated with time differentiation, and no change occurs in a step-function except the very sudden one when $t = 0$. We can therefore expect that i_c will be zero at all times except $t = 0$, but that it will behave violently at $t = 0$; theoretically, it becomes infinite at $t = 0$, but in practice various circuit resistances which we here ignore would make i_c finite and last for a finite time instead of for zero time.

After the switch has been closed, C thus becomes suddenly charged to the voltage E, R draws a steady current E/R and L takes a current tE/L at time t. This current, rising linearly with time, forms a part of the scan stroke. When it has reached a particular value i_0, the switch S is opened, and the circuit becomes that of Fig. 89 (b), in which we regard zero time as the instant of *opening* the switch; initially, therefore, in Fig. 89 (b), i_L has the value i_0, and the voltage across C has the value $v_0 = E$.

Let the voltage across the circuit at any later time t be V. Then initially V is positive (if E is positive) and this is indicated in Fig. 89 (b); the arrow indicates that the upper horizontal line has the higher potential, assuming V to be positive. Now if we apply Ohm's Law and the fundamental rules of Sections 126 and 127 with the sign convention of Fig. 89 (b), we find

$$V = Ri_r = L\frac{di_L}{dt}$$

and

$$\frac{dV}{dt} = \frac{i_c}{C}$$

Also, by Kirchhoff's Law

$$i_L + i_c + i_r = 0$$

Hence
$$i_L + C\frac{dV}{dt} + \frac{V}{R} = 0$$

or
$$i_L + CL\,\frac{d^2 i_L}{dt^2} + \frac{L}{R}\frac{di_L}{dt} = 0$$

In a case like this it is safer to stick to the fundamentals of Sections 126 and 127 and not venture into the p-world until we have expressed the circuit relations in the form of a differential equation with known initial conditions; here these conditions are $i_L = i_0$ when $t = 0$, and $V = v_0$ or

$$\left[\frac{di_L}{dt}\right]_{t=0} = \frac{v_0}{L}$$

Now let i_{Lp} be the p-world counterpart of i_L. Then the counterparts of

$$\frac{L}{R}\frac{di_L}{dt} \text{ and } CL\,\frac{d^2 i_L}{dt^2}$$

are respectively (see Section 138)

$$\frac{L}{R}\left\{pi_{Lp} - p[i_L]_{t=0}\right\} = \frac{L}{R}\left\{pi_{Lp} - pi_0\right\}$$

and
$$LC\left\{p^2 i_{Lp} - p^2[i_L]_{t=0} - p\left[\frac{di_L}{dt}\right]_{t=0}\right\} =$$

$$LC\left\{p^2 i_{Lp} - p^2 i_0 - \frac{pv_0}{L}\right\}$$

so that the p-world counterpart of the differential equation for i_L is

$$i_{Lp}\left\{LCp^2 + \frac{L}{R}p + 1\right\} = p^2 LCi_0 + pCv_0 + \frac{pL}{R}i_0$$

so that
$$i_{Lp} = \frac{p^2 i_0 + \dfrac{p}{CR}i_0 + \dfrac{pv_0}{L}}{p^2 + \dfrac{p}{CR} + \dfrac{1}{LC}}$$

Now let

$$\omega_0^2 = 1/(LC) \quad \alpha = 1/(2CR) \quad \omega = (\omega_0^2 - \alpha^2)^{\frac{1}{2}}$$

since in practice ω_0 is usually greater than α. Then we can write

$$i_{Lp} = \frac{i_0 p(p+\alpha) + \frac{1}{\omega}\left(\alpha i_0 + \frac{v_0}{L}\right)p\omega}{(p+\alpha)^2 + \omega^2}$$

so that, from the list of Section 138

$$i_L = i_0 \epsilon^{-\alpha t}\cos \omega t + \frac{1}{\omega}\left(\alpha i_0 + \frac{v_0}{L}\right)\epsilon^{-\alpha t}\sin \omega t.$$

In a practical case we might have $L = 40$ mH, $R = 200$ kΩ, $C = 50$ pF, $i_0 = 0{\cdot}3$ A, $v_0 = 300$ V.

Hence $\quad \omega_0 = 1/\sqrt{(LC)} = 1/\{40 \times 10^{-3} \times 50 \times 10^{-12}\}^{\frac{1}{2}}$

$$= 7{\cdot}0711 \times 10^5$$

$$\alpha = 1/(2CR) = 1/\{10^{-10} \times 2 \times 10^5\} = 5 \times 10^4$$

$$\omega = \sqrt{(\omega_0^2 - \alpha^2)} = \sqrt{(50 \times 10^{10} - 25 \times 10^8)} =$$

$$\sqrt{(49{\cdot}75 \times 10^{10})} = 7{\cdot}0534 \times 10^5$$

$$\alpha/\omega = 0{\cdot}070887$$

$$v_0/\omega L = 0{\cdot}010633$$

so $i_L = 0{\cdot}3\,\epsilon^{-\alpha t}\cos \omega t +$

$$(0{\cdot}3 \times 0{\cdot}070887 + 0{\cdot}010633)\,\epsilon^{-\alpha t}\sin \omega t$$

$$= 0{\cdot}3\,\epsilon^{-\alpha t}\cos \omega t + 0{\cdot}031899\,\epsilon^{-\alpha t}\sin \omega t$$

The voltage across the circuit is

$$v = L\frac{di_L}{dt}$$

Now let ϕ be the angle whose tangent is ω/α and, therefore, whose cosine is α/ω_0 and sine is ω/ω_0 (85° 57′ in our numerical case). Then

$$v = -L\left\{ i_0\alpha\epsilon^{-\alpha t}\cos\omega t + i_0\omega\epsilon^{-\alpha t}\sin\omega t + \frac{\alpha}{\omega}\left(\alpha i_0 + \frac{v_0}{L}\right)\epsilon^{-\alpha t}\sin\omega t - \left(\alpha i_0 + \frac{v_0}{L}\right)\epsilon^{-\alpha t}\cos\omega t \right\}$$

$$= -L\omega_0\left\{ i_0\epsilon^{-\alpha t}\cos(\omega t - \phi) + \frac{1}{\omega}\left(\alpha i_0 + \frac{v_0}{L}\right)\epsilon^{-\alpha t}\sin(\omega t - \phi) \right\}$$

and if $t = 0$, this reduces to v_0 as it should when we substitute for $\cos\phi$ and $\sin\phi$. We also notice that we could obtain di_L/dt from i_L by multiplying by $-\omega_0$ and replacing ωt by $\omega t - \phi$, but leaving αt alone. This suggests, as we can verify by plodding through the differentiation, that

$$\frac{dv}{dt} = L\omega_0{}^2\left\{ i_0\epsilon^{-\alpha t}\cos(\omega t - 2\phi) + \frac{1}{\omega}\alpha i_0 + \frac{v_0}{L}\epsilon^{-\alpha t}\sin(\omega t - 2\phi) \right\}$$

so that $\frac{dv}{dt} = 0$, and v has an extreme value when

$$\tan(\omega t - 2\phi) = -\frac{i_0}{\frac{\alpha i_0}{\omega} + \frac{v_0}{\omega L}}$$

If the term $v_0/\omega L$ could be neglected, this would give

$$\omega t - 2\phi = -\phi$$

but in general $\omega t - 2\phi$ is a negative angle between 0 and $-\phi$ and much nearer $-\phi$, so ωt is somewhat greater than ϕ, and in our numerical example will not be very different from $\frac{\pi}{2}$ when v has its maximum numerical value. The actual value of ωt is 87° 58′, or 1·53531 radians [and ωt must be expressed in radians if we want to find t or $\epsilon^{-\alpha t}$, though we may use degrees for $\cos(\omega t - \phi)$ and $\sin(\omega t - \phi)$].

It follows that

$$v_{max} = -0{\cdot}04 \times 7{\cdot}0711 \times 10^5 \times \epsilon^{-\frac{\alpha}{\omega}1.53531}$$
$$\times \{0{\cdot}3 \cos 2^\circ 1' + 0{\cdot}031899 \sin 2^\circ 1'\}$$
$$= -7591 \text{ V}$$

Thus the peak voltage has the opposite polarity to that of the initial voltage and is very much greater; it occurs approximately a quarter of a cycle after the start of the oscillation, when the energy initially stored in L has been transferred, apart from some loss in R, into C.

If left to itself the circuit will oscillate with i_L and v varying according to the equations we have obtained until all the energy has been dissipated in R. In practice, the circuit is allowed to oscillate freely for about half a cycle only, until v has once again its initial value. The switch S of Fig. 89 (*a*) is then closed once more. To find the time when the switch must be closed, we have to solve the equation

$$v_0 = -L\omega_0 \left\{ i_0 \epsilon^{-\alpha t} \cos(\omega t - \phi) + \frac{1}{\omega}\left(\alpha i_0 + \frac{v_0}{L}\right) \epsilon^{-\alpha t} \sin(\omega t - \phi)\right\}$$

in our numerical case. This appears to be insuperably difficult, but we do know that the relevant ωt cannot differ greatly from π since otherwise the right-hand side of this equation will be numerically much greater than the left-hand side. We can obtain an adequate approximation by expanding $\cos(\omega t - \phi)$ and $\sin(\omega t - \phi)$, replacing $\cos \omega t$ by -1, $\sin \omega t$ by $(-\omega t + \pi)$ which we shall call $-x$, and $\epsilon^{-\alpha t}$ by $\epsilon^{-\frac{\alpha\pi}{\omega}}\left(1 - \frac{\alpha x}{\omega}\right)$ and neglecting x^2, and the equation will become linear in x; alternatively, we can calculate some values of the right-hand side of the equation for ωt in the neighbourhood of π, and plot a curve. Having determined x and hence t as $(\pi + x)/\omega$, or having determined t graphically, we can substitute in the equation already obtained for i_L. i_L is negative at this time, that is to say, it is in the direction opposite to the arrow in Fig. 89.

When the switch S is closed, the voltage across C is equal to that of the battery for we have chosen the time of closing the switch to make this so. The current therefore flows mainly into the battery, but some is taken by R.

The current then decays linearly to zero to form the first part of the scan, and when it reaches zero we are back again at our starting point. There are just two differences at the start of this second cycle of events. The switch S is already closed and C is already charged to the battery voltage. As a result there is no initial rush of current into C.

We started by assuming an inert circuit and first closed the switch somewhere towards the middle of the scan state. Afterwards we just open the switch once per cycle to initiate the flyback and close it again after the flyback to start the scan.

In practice, the switch is replaced by a pair of valves, usually a pentode and a diode, and they are both cut-off during flyback. During scan their internal resistances modify matters appreciably. The analysis given to illustrate the application of mathematics to a practical problem is much nearer the truth during flyback than during the scan period.

Our second example, typical of non-electrical problems, is to solve the equation

$$\frac{d^2x}{dt^2} + 3\frac{dx}{dt} + 2x = \sin 4t$$

given that $x = 0$ and $dx/dt = 1$ when $t = 0$. First, suppose that $\bar{x}$ is the p-world counterpart of x according to the list in Section 138. Then the rule for obtaining the p-world counterpart of d^rx/dt^r is also given in Section 138, so we have, since $x(0) = 0$ and $\left(\frac{dx}{dt}\right) = 1$

$\frac{dx}{dt}$ has the counterpart $p\bar{x} - p.0$

$\frac{d^2x}{dt^2}$ " " " $p^2\bar{x} - p^2.0 - p.1$

The p-world counterpart of $\sin 4t$ is obtained direct from the list, and is $4p/(p^2 + 16)$ so the p-world counterpart of the whole equation is

$$(p^2\bar{x} - p) + 3p\bar{x} + 2\bar{x} = 4p/(p^2 + 16)$$

or

$$\bar{x} = \frac{4p}{(p^2 + 3p + 2)(p^2 + 16)} + \frac{p}{p^2 + 3p + 2}$$

We now require to split up $\bar{x}$ into the sum of a number of terms which are counterparts of known time-functions, in the same way as we did in connection with the series oscillatory circuit in Section 139. As in that case, it is convenient to keep a factor p outside the expressions to be split up into partial fractions. We find

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

and

$$\frac{1}{(p+1)(p+2)(p^2+16)} = \frac{1}{(p+1)(-1+2)(1+16)} + \frac{1}{(-2+1)(p+2)(4+16)} + \frac{Ap+B}{p^2+16}$$

Multiply through this last equation by $(p^2 + 16)$ to get A and B: we have

$$\frac{1}{(p+1)(p+2)} = \frac{p^2+16}{(p+1).17} - \frac{p^2+16}{(p+2).20} + Ap + B$$

Now put $p = 4j$ so that $p^2 + 16 = 0$, and we have

$$\frac{1}{(4j+1)(4j+2)} = 4jA + B$$

This procedure is analogous to that used in Section 139 when we applied a voltage $e \sin \Omega t$ to the circuit of Fig. 66, and required the quantities there called C and D. Since the left-hand side can be handled as in Section 58, we hav

$$\frac{1}{(4j+1)(4j+2)} = \frac{(1-4j)(2-4j)}{(1+4j)(1-4j)(2+4j)(2-4j)}$$

$$= \frac{2-12j-16}{(1^2+4^2)(2^2+4^2)} = \frac{-14-12j}{340}$$

so the real part, B, is $-7/170$ and from the imaginary part, $4jA$, we find $A = -3/340$. We have now succeeded in splitting the expression we had for $\bar{x}$, the p-world counterpart of x, as required; we obtain

$$\bar{x} = 4p\left[\frac{1}{17(p+1)} - \frac{1}{20(p+2)} - \frac{3p}{340(p^2+16)} - \frac{7}{170(p^2+16)}\right]$$

$$+ p\left[\frac{1}{p+1} - \frac{1}{p+2}\right]$$

and, collecting the terms in $p/(p+1)$ and $p/(p+2)$, we find,

$$\bar{x} = \frac{21p}{17(p+1)} - \frac{6p}{5(p+2)} - \frac{3p^2}{85(p^2+16)} - \frac{7}{170}\cdot\frac{4p}{p^2+16}$$

so that, from our list in Section 127,

$$x = \frac{21}{17}e^{-t} - \frac{6}{5}e^{-2t} - \frac{3}{85}\cos 4t - \frac{7}{170}\sin 4t$$

and we can easily verify that this expression really does satisfy the original differential equation and gives the correct initial values of x and dx/dt.

For the benefit of those who may have come across the more usual way of solving linear differential equations by means of "particular integral" and "complementary function", and who may expect the solution of a linear differential equation of the second order to contain two arbitrary constants, we must point out that the solution given above has automatically determined the values of the arbitrary constants required to satisfy the initial conditions. This is just one example of the power of Heaviside's method, and explains why we have attached so much importance and devoted so much space to it. At first it may seem formidable in its unfamiliarity, but it repays abundantly the initial effort required to learn it. The reviser of this book regrets that he did not come across Heaviside's technique earlier, and hopes to save some at least of his readers from a similar misfortune.

148. Numerical Computation

There are several textbooks on this subject available for computers, that is to say, people whose main occupation consists in making numerical calculations and handling various types of mathematical processes and formulæ. It is not the object of this section to compete with any such textbook, but rather to lighten the labour of an engineer who occasionally and incidentally may have to make a reasonably straightforward computation, but who may waste much time in the process if left to fend entirely for himself. In many organisations nowadays, particularly in large ones, desk calculating machines and even electronic computers are available, and people trained primarily as mathematicians are finding that their services are now much more appreciated than, say, twenty years ago by organisations whose primary object is engineering.

If a mathematician or computer spends much of his time in dealing with numerical computation, he cannot help getting to know many useful and labour-saving tricks which would only be learnt painfully by an engineer who might have to use some of them for only one particular calculation, and would gladly forget them a week later. Hence the first suggestion to be put forward on this subject is that mathematicians and computers should be consulted whenever possible.

Even the arrangement of a calculation, and the order in which the operations are performed, may vitally affect the time taken, and in some cases the difference in this with and without a desk machine is much greater than in others. For a calculation which consists entirely of products, powers and quotients, where the final answer is only required to two or three significant figures, a slide-rule is quite adequate; excessive errors are likely to accumulate in calculations where there is addition and subtraction, particularly where differences between nearly equal quantities are involved, even if the slide-rule is not inadequate because four or five significant figures are required.

For a large-scale computation, the possibility of using an electronic computer must not be overlooked; there are a number of processes, such as the solution of n linear simultaneous equations in n unknowns when n is large (say 10 or more) where the "programme," needed for an electronic computer is perfectly well known, and where the machine time required is astonishingly small, and the cost reasonable even at prices of the order of £30 per hour of machine time. An electronic computer may be the best and cheapest way to obtain an answer in any calculation which can be made iterative or repetitive, even if from the point of view of a human calculator, the process might seem to be very roundabout and tedious.

Another possibility, which should not be overlooked where the same type of calculation has to be repeated over and over again but the accuracy of the final answer need only be within a few per cent, is the design of a nomogram, which performs the calculation geometrically in the general case. Once the nomogram is there, an individual example of the calculation is worked out instantly by means of a ruler. Thus if we construct a nomogram for the formula

$$V = \pi DN/12$$

where V ft/min is the cutting speed of a lathe spindle, D in. the diameter of the part to be turned and N rev/min the speed of rotation, when we require the value of N given say $V = 100$ ft/min and $D = 1{\cdot}6$ in., we need only join the point marked $V = 100$ ft/min on the V-scale of the nomogram to the point marked $D = 1{\cdot}6$ on the D-scale of the nomogram. The line joining these points will meet the N-scale of the nomogram in a point which determines the value of N.

The design of nomograms of this kind is much easier than one might expect. There is no further space available to discuss the matter here, but the nomogram for $V = \pi DN/12$ and a full discussion of the construction of nomograms is given elsewhere.*

Once we have realised that there is a number of calculations which ought mainly to be handled by or under the guidance of mathematicians and expert computers, or by means of nomograms, and that engineers should consult mathematicians whenever possible, and not overlook the possibility of using an electronic computer, there remain many calculations which the engineer can tackle for himself with a little encouragement and help, but for which he is likely to take several times as long as he need take. Suppose for example we wish to calculate for various values of ω

$$\tau(\omega) = \frac{1}{1 + \sqrt{2}\omega + \omega^2} + \frac{1}{1 - \sqrt{2}\omega + \omega^2}$$

then we first notice that $\tau(\omega)$ is unchanged if $-\omega$ is put for ω, in other words, $\tau(\omega)$ is an "even" function of ω, and if we happened to know a series (cf. Chapter 4) for $\tau(\omega)$ in powers of ω, only even powers of ω would be present; for such a function $d\tau/d\omega$ is necessarily zero when $\omega = 0$ [provided that, as in this case, the function is continuous and has a derivative when $\omega = 0$]. But also

$$\begin{aligned}(1 + \sqrt{2}\omega + \omega^2)(1 - \sqrt{2}\omega + \omega^2) \\ = (1 + \omega^2)^2 - (\sqrt{2}\omega)^2 \\ = 1 + 2\omega^2 + \omega^4 - 2\omega^2 = 1 + \omega^4\end{aligned}$$

so that $\tau(\omega)$ can be rewritten

$$\tau(\omega) = \frac{(1 - \sqrt{2}\omega + \omega^2) + (1 + \sqrt{2}\omega + \omega^2)}{1 + \omega^4} = 2\,\frac{1 + \omega^2}{1 + \omega^4}$$

Thus, in this case, a little preliminary manipulation of the function has greatly simplified it. This will not always

* *Abacs or Nomograms*, by A. Giet. English edition published by Iliffe & Sons, London, 1956. This book is written by an engineer for engineers, unlike most books on the subject, which take a more mathematical viewpoint.

happen; what matters is that we should be aware that it can happen, and that we should not take for granted that the form in which we first derive an expression is the best for computation of it. Time spent in investigating the effect of simple manipulations like this is seldom wasted. As we now have $\tau(\omega)$ explicitly in terms of ω^2, it is worth considering values of ω in any particular region of interest which make ω^2 rather than ω a simple number, e.g. $\omega^2 = 0{\cdot}2, 0{\cdot}4, 0{\cdot}6, 0{\cdot}8$ and 1 if we are particularly interested in the region $0{\cdot}4 < \omega < 1$.

In the case of a calculation like

$$y = x^2 + x + 3 + 2\sqrt{(x^2 - x + 7)}$$

it pays to have numbered columns for the various quantities involved, and to work vertically, not horizontally, as long as we want, say, a general idea of the values of y for integral values of x between 1 and 10 (or any other set of equally-spaced values); if later we require isolated additional points, the advantages of working vertically will be less marked. Thus our columns in this case might be

1	**2**	**3**	**4**	**5**	**6**
x	x^2	x^2+x+3	x^2-x+7	$2\sqrt{(x^2-x+7)}$	y
		[**1**+**2**+3]	[**2**−**1**+7]	[2√**4**]	[**3**+**5**]

Knowing x, we write down x^2 (Column **2**) and while this column is being written, the object of the calculation and every subsequent step of it can be completely forgotten; we merely concentrate on correctly writing down the squares of integers, and if we have made a mistake because the telephone went, we shall notice this immediately, because this particular process of squaring occurs so very frequently.

Next, we have to concentrate, to the exclusion of all else, on the addition of the first two columns in each row, and adding 3 to the result, writing the answer down in column **3**. Similarly, for column **4**, we concentrate on subtracting column **1** from column **2** and adding the number 7. For column **5** we concentrate on taking the square root and doubling; the entries in column **4** will all be two-figure numbers and therefore tabular entries in four- or five-figure square-root tables. It may be easier to have an intermediate column **4a** in which entries in column **4**, being whole numbers, are multiplied by 4; column **5** will then be the square root of column **4a** (which will be a three-figure number and therefore still a tabular entry). This suggestion may seem trivial and unnecessary; it is made because of the reviser's bitter experience that the more one can avoid thought during the actual carrying out of a computation, the more likely one is to avoid making stupid mistakes. Finally, the entries in columns **3** and **5** of each row are added to give the required value of y.

Next, suppose that we require to three significant figures the values of

$$z = (100 + x)^{\frac{1}{2}} - (100 - x)^{\frac{1}{2}}$$

for x between 0 and 10. If x is 5, five-figure tables give $105^{\frac{1}{2}} = 10{\cdot}247$ and $95^{\frac{1}{2}} = 9{\cdot}7468$, and the difference, 0·500, is reliable only to the third place of decimals, that is, to the third significant figure in spite of the fact that we used five-figure tables. If we had used a slide-rule, we would only have been sure of the first significant figure of z.

Since z involves powers, one possibility is to use the binomial theorem (Section 67) but we have to be a little careful since we are dealing with a non-integral power. We cannot apply the binomial theorem direct to $(100+x)^{\frac{1}{2}}$, but must first write it in the form $100^{\frac{1}{2}}(1 + 0{\cdot}01x)^{\frac{1}{2}}$ when, as here, x is between $+100$ and -100. [If x had been outside these limits we should have had to write

$$(100 + x)^{\frac{1}{2}} = x^{\frac{1}{2}}\left(1 + \frac{100}{x}\right)^{\frac{1}{2}}$$

and the series obtained would have been in descending, not ascending, powers of x.] We now find

$$(100+x)^{\frac{1}{2}} = 10\left[1 + \tfrac{1}{2}(0{\cdot}01x) + \frac{\frac{1}{2}\cdot(-\frac{1}{2})}{2!}(0{\cdot}01x)^2 + \ \ldots\right]$$

and similarly

$$(100-x)^{\frac{1}{2}}=10\left[1-\tfrac{1}{2}(0{\cdot}01x)+\frac{\frac{1}{2}\cdot(-\frac{1}{2})}{2!}(0{\cdot}01x)^2-\ \ldots\ \right]$$

Hence, subtracting,

$$z=20\left[\tfrac{1}{2}\ (0{\cdot}01x)\ +\ \frac{\frac{1}{2}\cdot(-\frac{1}{2})\cdot(-\frac{3}{2})}{3!}(0{\cdot}01x)^3+\ldots\ \ldots\right]$$

since the even powers of x cancel. For the range of x considered, the x^3 term is relatively unimportant and the next term in z, which involves x^5, is completely negligible. If x has the extreme value 10, z is $20[0{\cdot}05 + 0{\cdot}0000625] = 1{\cdot}00125$ and therefore the first term gives z correct to 4 significant figures over the whole range of interest. z thus varies linearly with x to a high degree of accuracy. We shall see later that most expressions involving x can be regarded as varying linearly with x for a sufficiently short range of x.

In the case of the particular expression z we have just been discussing, there is another useful trick which enables us to obtain z quickly and accurately for any value of x. We have

$$\{(100+x)^{\frac{1}{2}}+(100-x)^{\frac{1}{2}}\}\,z$$

$$=\{(100+x)^{\frac{1}{2}}+(100-x)^{\frac{1}{2}}\}\,\{(100+x)^{\frac{1}{2}}-(100-x)^{\frac{1}{2}}\}$$

$$=(100+x)-(100-x)=2x$$

so that

$$z=\frac{2x}{(100+x)^{\frac{1}{2}}+(100-x)^{\frac{1}{2}}}$$

In this form of z, we can safely use a slide-rule for determining $(100+x)^{\frac{1}{2}}$ and $(100-x)^{\frac{1}{2}}$ if we require z to three significant figures, as we are now concerned with the sum of these two square roots instead of the difference.

In general, it is easier to handle an awkward number like $\sqrt{3}$ if it is in the numerator than if it is in the denominator, thus $(1/\sqrt{3})$ is usually better written as $\sqrt{3}/3$ and

($1/(a+b\sqrt{c})$ is usually better written as $(a-b\sqrt{c})/\{a^2-b^2c\}$. Sometimes it is highly desirable to obtain accurately the value of a square root. It is worth mentioning in passing that if a is an approximation to $\sqrt{N}$, then

$$\tfrac{1}{2}\left(a+\frac{N}{a}\right)$$

is a very much better one, correct to roughly twice as many significant figures.

Finally, suppose that we have to find a positive value of k satisfying the equation

$$k\left(1-\frac{1\cdot 58}{0\cdot 175+0\cdot 58k^2}\right)=0\cdot 625$$

At first we are tempted to give up the unequal struggle and say that it is impossible to do this because clearly the above equation, when multiplied through by $(0\cdot 175 + 0\cdot 58k^2)$, will be cubic in k. But a closer inspection shows that as k increases, the second factor of the left-hand side continually increases, since the subtracted expression $1\cdot 58/(0\cdot 175 + 0\cdot 58k^2)$ continually decreases, while the factor k outside continually increases. The left-hand side, however, remains negative until $(0\cdot 175 + 0\cdot 58k^2)$ reaches the value $1\cdot 58$, which happens when $k^2 = (1\cdot 58 - 0\cdot 175)/0\cdot 58$ or $k = 1\cdot 56$. For greater values of k, the left-hand side is positive and increases steadily with k. It becomes nearly equal to k when k is large. As it is a continuous function of k, it can therefore pass through the value $0\cdot 625$ only once. We must now evaluate the left-hand side for some trial values of k until we have one value of k for which the left-hand side is slightly above $0\cdot 625$ and one value of k for which the left-hand side is slightly below $0\cdot 625$. We find $k = 2$ makes the left-hand side $0\cdot 733$ while $k = 1\cdot 9$ makes it $0\cdot 577$. As the left-hand side is varying rapidly with respect to k, and is not near a maximum or minimum value, we assume that it is varying linearly with k, and that a good approximation to the correct value of k will be

$$k = 1\cdot9 + \frac{0\cdot625 - 0\cdot577}{0\cdot733 - 0\cdot577} \times (2 - 1\cdot9)$$

$$= 1\cdot929$$

Substituting 1·929 for k on the left-hand side, it reduces to 0·623. In this case we can therefore take 1·929 as a sufficiently accurate value of k; if we required a more accurate value still, this would be

$$1\cdot929 + \frac{0\cdot625 - 0\cdot623}{0\cdot623 - 0\cdot577} \times (1\cdot929 - 1\cdot9) = 1\cdot930$$

Thus we have found k by means of the fact that most functions can be regarded as varying linearly over short ranges. This fact is essentially a consequence of what is known as Taylor's Theorem, that if $f(x)$ is continuous and possesses derivatives up to and including the nth in the neighbourhood of $x = a$, then

$$f(x) = f(a) + \frac{(x-a)}{1!}f'(a) + \frac{(x-a)^2}{2!}f''(a) + \quad \cdots$$

$$+ \frac{(x-a)^{n-1}}{(n-1)!}f^{(n-1)}(a) + \frac{(x-a)^n}{n!}f^n\ (a + \theta x - \theta a)$$

where $f'(a)$ is the first derivative, $f''(a)$ the second and $f^{(r)}(a)$ the rth derivative of $f(x)$ when $x = a$. θ is an unknown fraction between 0 and 1, but in practical use of the theorem one arranges that $(x - a)^n/n!$ shall be sufficiently small for this last term to be negligible. If then the term $f''(a)(x - a)^2/2!$ is negligible, and later terms still more so, $f(x)$ does vary linearly; it is thus merely necessary to make $(x - a)$ sufficiently small. From this theorem follows also Newton's method of approximating to the root of any equation (which need not be algebraic); if the equation is $f(x) = 0$ and a is an approximation to a root of that equation, a being real or complex, then

$$A = a - \frac{f(a)}{f'(a)}$$

is in general a better approximation. This result enables us to solve many equations for which an approximate solution is known, for example, because we have some idea of the solution to be expected from experimental results or the previous solution of a similar equation.

If $f'(a) = 0$ and $f''(a)$ is not zero, so that $f(x)$ has either a maximum or a minimum at $x = a$, Taylor's Theorem shows that in the immediate neighbourhood of $x = a$, the difference $f(x) - f(a)$ is proportional to $(x - a)^2$. This is thus the exceptional case when we cannot regard $f(x)$ as varying linearly over a sufficiently short range.

149. "Least Squares". Random Variations. Standard Deviation

It often happens in experimental work that we have reason to believe that there is a linear relation between two variables x and y,* and that we seek to determine this relation explicitly. If we plot the corresponding values of x and y obtained from the experiment, it is unlikely that the n points will lie exactly upon a straight line, but we shall fairly easily be able to determine by eye a straight line l which lies evenly among the points (Fig. 90). If the dependent variable y is plotted vertically, we could determine the distance d_r of the rth point above or below l (measured in y-units) and evaluate $d_1{}^2 + d_2{}^2 + \ldots + d_n{}^2$, a quantity which we shall call S. Now it is possible to calculate m and c so that if the d's are measured with respect to the straight line l' whose equation is

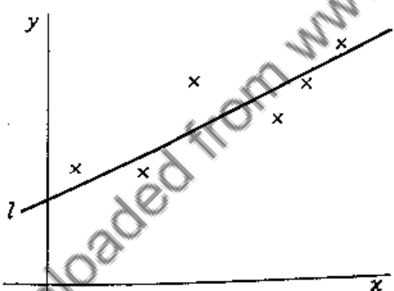


Fig. 90

$$y = mx + c$$

* x need not be the original variable X measured; but may be some function of it, such as X^2 or log X and similarly for y.

then S is less for l' than for any other line, including l. It is this property which gives the line l' its name of "least-squares line". The equations for m and c are

$$\sum_{r=1}^{n} y_r = m \sum_{r=1}^{n} x_r + nc$$

$$\sum_{r=1}^{n} x_r y_r = m \sum_{r=1}^{n} x_r^2 + c \sum_{r=1}^{n} x_r$$

where the symbol $\sum_{r=1}^{n}$ means that the variable which

follows is evaluated for $r = 1, 2, \ldots n$ and the values obtained are added. These equations determine m and c uniquely, so there is one definite line l', to which a line l obtained by eye is likely to be a good approximation. The value of $\{S/(n-1)\}^{\frac{1}{2}}$ for the line l' is a rough measure of the "goodness of fit" of l'.

Now the n points will not be at equal vertical distances from the line l', and in fact d_1, d_2, etc., will vary in a random manner. There will probably be many different causes of this variation, no one of which contributes very much to it. We can therefore expect the values of the d's to be "normally distributed" about zero, and we shall consider briefly the properties of such distribution.

In a normal distribution, there are theoretically an infinite number of examples of the variable in question (in this case d) and they may be concentrated near the mean value or widely scattered, but very few examples have extreme values in either direction. Suppose that we consider a set of N values of a random variable x which is normally distributed, and that the mean value of x is x_0. Then the number of members of the set for which theoretically we should expect x to lie between k and $(k + h)$ is

$$\frac{Nh}{\sigma}\left[\frac{1}{\sqrt{2\pi}} \; \epsilon^{-(k-x_0)^2/2\sigma^2}\right]$$

provided that N is sufficiently large, and h is sufficiently small in relation to σ. The quantity within brackets is tabulated by various authors as a function of $(k - x_0)/\sigma$. The total number m of members of the set for which we should expect x to lie between k and k^1 $(k^1 > k)$ is

$$m = \frac{N}{\sqrt{2\pi}} \int_{(k-x_0)/\sigma}^{(k^1-x_0)/\sigma} \epsilon^{-x^2/2}\, dx$$

and the integral is also tabulated, so that we can evaluate m immediately from the tables given k, k^1, x_0, σ and N. σ is a parameter called the "standard deviation" of the normal distribution, and measures the scatter of the distribution. Only 2·275% of the members of the distribution have values of x exceeding $x_0 + 2\sigma$, 0·135% have values of x exceeding $x_0 + 3\sigma$ and about 0·003% have values of x exceeding $x_0 + 4\sigma$; the distribution is symmetrical about the mean value x_0.

Now we have to consider the reverse problem; given a set of N values of a randomly-varying quantity $x_1, x_2 \ldots x_N$, which we expect to be part of an infinite normally distributed set, what are the most likely values of x_0 and σ when N is reasonably large (which in practice usually means at least 5)? If N is small, say between 5 and 10, it will be easiest to evaluate x_0 and σ directly from the formulæ

$$x_0 = \frac{1}{N} \sum_{r=1}^{N} x_r \qquad \sigma^2 = \frac{1}{N-1} \sum_{r=1}^{N} (x_r - x_0)^2$$

and it is worth noting that σ^2 can be written in the form

$$\sigma^2 = \frac{1}{N-1} \sum_{r=1}^{N} \Big[(x_r - \xi) + (\xi - x_0) \Big]^2$$

$$= \frac{1}{N-1} \left\{ \sum_{r=1}^{N} (x_r - \xi)^2 + 2(\xi - x_0)[Nx_0 - N\xi] + N(\xi - x_0)^2 \right\}$$

$$= \frac{1}{N-1}\left[\sum_{r=1}^{N} (x_r - \xi)^2 - N(x_0 - \xi)^2\right]$$

where we can give ξ any value we please. The nearest round number to the mean x_0 is usually the most convenient value for ξ, as then σ^2 involves only the direct squaring of reasonably small numbers, with perhaps only three significant figures. If N is larger than say 10, the direct calculation of x_0 and σ as above becomes very laborious. It can be simplified if the values of x are "grouped" as indicated immediately below; some loss of accuracy may be risked by doing this, but it is usually not serious for practical purposes. Suppose we have the following 67 values of x in ascending order:

4·7, 6, 6·5, 7·5, 8·1, 8·9, 9·2, 9·5, 9·7, 10·1, 10·2, 10·3, 10·3, 10·6, 10·8, 10·9, 11·1, 11·2, 11·5, 11·7, 11·9, 12·1, 12·2, 12·3, 12·4, 12·6, 12·7, 12·9, 13·1, 13·1, 13·2, 13·4, 13·5, 13·6, 13·7, 13·8, 13·8, 13·9, 14·2, 14·3, 14·5, 14·5, 14·6, 14·7, 14·7, 14·9, 15, 15·2, 15·3, 15·6, 15·8, 15·9, 16·1, 16·3, 16·6, 16·7, 17, 17·2, 17·5, 17·9, 18·1, 18·5, 19, 19·6, 20·3, 21, 22·5.

Then we simplify the calculation by treating all values of x between 4 and 6 as if they had an "effective" value 5, all values of x between 6 and 8 as if they had an "effective" value 7, all values of x between 8 and 10 as if they had an "effective" value 9, and so on. The only exception is that when a "border line" value like 6 occurs (as it does in the list) it is regarded as half in the class between 4 and 6 regarded as 5, and half in the class between 6 and 8 regarded as 7. The number of values between 4 and 6 is thus $1\frac{1}{2}$, that between 6 and 8 is $2\frac{1}{2}$, that between 8 and 10 is 5, and so on, as listed in the column headed F (frequency of occurrence). At this stage we would have a table of groups and the number in each group as below:

x	4–6	6–8	8–10	10–12	12–14	14–16	16–18	18–20	20–22	22–24
F	$1\frac{1}{2}$	$2\frac{1}{2}$	5	12	17	14	8	4	2	1
Effective x	5	7	9	11	13	15	17	19	21	23

↑

We now say that the mean is roughly in the neighbourhood of the group indicated by the arrow, which is called the "working mean".

It is not important if our guess is inaccurate at this stage. We now count how many places each group is to the left or right of the arrow, and call this number of places *d*, positive for places to the right of the arrow and negative for places to the left of it. We now turn the whole arrangement vertically, and to the values of "effective *x*", *F* and *d* already discussed, we add columns of values of *Fd* and Fd^2 as below:

Effective x	*F*	*d*	*Fd*	Fd^2
5	$1\frac{1}{2}$	−4	−6	24
7	$2\frac{1}{2}$	−3	$-7\frac{1}{2}$	$22\frac{1}{2}$
9	5	−2	−10	20
11	12	−1	−12	12
			−total $-35\frac{1}{2}$	*Working Mean*
→13	17	0		
15	14	1	14	14
17	8	2	16	32
19	4	3	12	36
21	2	4	8	32
23	1	5	5	25
			+total 55	
Total	67		*Net total* $+19\frac{1}{2}$	$217\frac{1}{2}$

The *Fd* and Fd^2 columns explain themselves; it saves effort to add up the (*Fd*) entries with negative sign and those with positive sign separately, but Fd^2 is necessarily positive, so they can all be added. We thus find that the mean is $19{\cdot}5/67 = 0{\cdot}291$ rows below the working mean. The 17 members of the working-mean group are all treated as if *x* was 13 for them, and the "group interval" is

2 units of x, so the actual mean is

$$13 + 2 \times 0{\cdot}291 = 13{\cdot}582$$

To determine σ, we have that the sum of the squares Fd^2 would have been $217\frac{1}{2}$ "square group intervals" if the actual mean had been 13 (the mean value of the working-mean group), but it needs a correction—the subtraction of $(19\frac{1}{2})^2/67 = 5{\cdot}7$—because this is not the case. The sum of the values of Fd^2 has thus the corrected value 211·8. Dividing this by $N-1$, that is 66, gives $\sigma^2=3{\cdot}21$, $\sigma=1{\cdot}79$ group intervals, or $1{\cdot}79\times2=3{\cdot}58$ units of x. We have thus obtained σ by means of a reasonably simple calculation. There are various corrections which should strictly be applied to σ, but for engineering purposes we usually only require the value of σ roughly, and it will therefore be sufficient to mention only the most important of these corrections, which is that if σ turns out to be appreciably less than 4 group intervals, as here, the grouping is too coarse; if the group interval had been halved (so that values of x between 4 and 5 were taken as all equal to 4·5, values between 5 and 6 as all equal to 5·5, and so on) it would have been better.

One other statistical quantity should be briefly mentioned: the correlation coefficient. If we have two normally-distributed variables x and y, and a large number N of pairs of corresponding values, we may wish to know whether there is any connection between the variations of x and y. The "correlation coefficient" between x and y is defined as

$$\rho = \frac{\sum_{r=1}^{N} (x-x_0)(y-y_0)}{\left\{\sum_{r=1}^{N} (x-x_0)^2\right\}^{\frac{1}{2}} \left\{\sum_{r=1}^{N} (y-y_0)^2\right\}^{\frac{1}{2}}}$$

where x_0, y_0 are the mean values of x, y respectively. If ρ is zero [or does not differ significantly from zero] the variations of x and y can be regarded as independent; at the other extreme, if y and x are linearly related, ρ will be ±1. ρ cannot exceed 1. If we have a number of pairs of corresponding values of x and y, say at least ten, the best available estimate of ρ is obtained by substituting in the above formula for ρ as if the N values belonged to normal distributions for x and y. A value of ρ has to be high to be statistically significant if there are only a few observations. The actual calculation can be simplified by grouping in much the same way as the calculation of standard deviation already given.

The above is only a bare outline of the statistical quantities of the greatest practical importance, their meaning, and how to estimate them. Statisticians have various tests by which they determine whether a statistical quantity is significant, and such tests are outside our present scope. For practical engineering purposes, the standard deviation σ can be regarded as a rough and ready measure of the scatter of a randomly varying quantity, and a correlation coefficient should be regarded as likely to have little practical significance unless it is of the order of 0·8 or more.

150. Conclusion. What You Can and Cannot Do with Mathematics

Professor Hogben says that "mathematics is the language of size" and that as such it is "an essential part of the equipment of an intelligent citizen". His choice of the word "citizen" instead of "man" is deliberate, in accordance with his new and distinctive contention—that mathematics is a necessary part of political education. It is so because it provides for all men, of all kinds and nationalities, a precise, impersonal, and objective way of describing the facts of the material world. It is, so far, the only effective international language.

But is it not also something more than a logical and efficient size-language? Is it not also a tool, a technique, a means of exploration and discovery? Yes, and no. Here it is necessary to finish on a note of caution, as a safeguard against possible disillusionment. It is true that a mathematical description can be, and indeed should always be, so formulated as to clarify and illuminate the physical system described, and may be used to reveal previously unsuspected relationships with other systems and ideas—to show, for example, that the falling of an apple and the swinging of the planets in their orbits, both depend on the same law, or that light is an electromagnetic wave. The contemplation of such tremendous achievements might well lead to the idea that mathematics is a key to all the mysteries of the universe. But that would be an illusion. Unfortunately it seems to be a popular illusion.

When a conjuror takes a rabbit out of a hat, we know that at some time and in some way, even though we did not see it done, the rabbit must have been put into the hat, and so it is with mathematics. Nothing comes out of a mathematical analysis that was not, implicitly or explicitly, put into it. It is, as it were, a passive network, with no internal source of energy, and the output from it cannot exceed the input, dynamically speaking, however much it may otherwise have been transformed.

Mathematics, by itself, is not enough. The magic lies in the combination of mathematics and experimental science. It happens that radio offers what is probably one of the best of all the examples of this, the transformation of Faraday's experimental laws into the theory of electromagnetic radiation by the mathematical genius of Clerk Maxwell. No other name could more fittingly end a book on mathematics and radio.

ANSWERS TO EXAMPLES

I

1. $ab^2 - a^2b + bc^2 - b^2c + a^2c - ac^2$.
2. 0, 0, 0, 6.
3. a, b, c.
4. $(x - 5)(x - 2)$; $(x - 5)(x + 2)$; $(x + 5)(x - 2)$; $(a - 99b)(a - b)$; $(a - b)(a + b)$.

II

2. Multiply each side by $ab(a + b)$.
3. Multiply numerator and denominator of left-hand side by abc, and multiply each fraction on right-hand side by $\frac{(a-b)(b-c)(c-a)}{(a-b)(b-c)(c-a)}$.
4. Divide both sides by $R_0R_1R_2R_3$.
5. $\frac{R_1 + R_2}{R_1R_2} = \frac{1}{R_1} + \frac{1}{R_2}$, and is therefore greater than either of these. Therefore $\frac{R_1R_2}{R_1 + R_2}$ is less than R_1 or R_2.
6. $x + y$; $x - y$.
7. 15.
8. 74, 40, 30, 26, 25, 24, 25, 26, 30, 40, 74. Notice that $x + y$ is least when $x = y$.

III

2. $a^{5/2}\, b^{5/16}\, c^{9/8} - a^{1/2}\, b\, c$.

3. $x + y = z$; $x - y = z$.

5. $(a - b)\ (a + b)\ (a^4 - a^2b^2 + b^4)$, because $a^6 - b^6 = (a^2)^3 - (b^2)^3$.

6. $\left(\dfrac{x^a}{x^b}\right)^{a+b} = x^{(a^2-b^2)}$ etc., etc.

IV

1. log 5 = log 10 − log 2.

2. log 1/3 = log 1 − log 3 ; log 81 = 4 log 3 ; log 3 1/3 = log 10 − log 3 ; log 3^{30} = 30 log 3.

3. 1, $\bar{2}$, 4, $\bar{4}$.

4. 100 log 2 = 30.103 = 10^{30} × some number between 1 and 10.

5. log $10^{10x} = 10x$; log $10x = 1 + \log x$.

6. $a^\circ = 1$.

7. No.

8. Minus infinity.

9. There is no negative real value of $y = a^x$ and log $(-1)^n$ is meaningless unless n is an even integer.

10. Log $y = \log k + (b \log a)\ x$.

V

1. (a) $x = 22\frac{2}{3}$; (b) $x = 1\frac{2}{13}$; (c) $x = (d - b)/(a - c)$; (d) $x = (nb - md)/(mc - na)$.

2. − 4/3, − 3, 0.

3. (a) 2 and 3 ; (b) − 2 and − 3 ; (c) − 2 and 3 ; (d) 2 and − 3.

4. (a) $\pm\sqrt{2}$; (b) 3 and 4 ; (c) $(a - b)$ and $(a + b)$; (d) $(b - a)$ and $(b + a)$.

5. (a) ±2, 1, 4 ; (b) a, b, c, d, e.

(c) $\dfrac{-b \pm \sqrt{b^2 - 4ac}}{2a}$ and $\dfrac{-e \pm \sqrt{e^2 - 4df}}{2d}$.

VI

1. $-2, -3, 9$.
2. $(1 - 2\sqrt{-1})$, $(1 + 2\sqrt{-1})$, 4, 30. (Since $(1 - 2\sqrt{-1})$ is a root $(1 + 2\sqrt{-1})$ must also be. The corresponding factor is $(x^2 - 2x + 5)$. The other factor is $120x^2 - 240x + 600$.)
3. 0. (See Section 45.)
4. (*a*) $x = 1, y = 81$.
   (*b*) $x = 0.62, y = 0.18$.
5. (*a*) $x = 3, y = 2$ or $x = 2, y = 3$. (Square 1st and subtract 2nd. This gives $2xy = 12$.)
   (*b*) $(2, -2)$; $(-3, 3)$; $(1, 3)$; $(-6/5, -18/5)$. Factors of 1st eqn. are $(3x - y)$ and $(x + y)$.
   (*c*) $(4, 3)$; $(-4, -3)$; $(7\sqrt{3}/2, -5\sqrt{3}/6)$; $(-7\sqrt{3}/2, 5\sqrt{3}/6)$.
6. This is the unsolvable pair referred to in Section 51.

VII

1. (*a*) $x = 3, y = 4, z = 5$; (*b*) $x = y = z = 1$; (*c*) $x = y = 0, z = 10$.
2. $w = 1, x = 2, y = 3, z = 4$.
3. $x = a + b, y = a + 2b, z = a + 3b$.
5. (*a*) 10 ; (*b*) 112/65.
6. $\pm 1, \pm i$, and $\sqrt{\pm i}$ are clearly possible values of the eighth root of unity, and since they give eight different values they must be all the possible values.

VIII

1. When $x = -c/d, y = (ad - bc)/0$.
2. ∞, ∞ ; $b/d, b/d$.
3. $(a - b)/(a - c)$.
4. $y = \dfrac{(x-1)}{(x-1)}\dfrac{(x-2)}{(x-2)}\dfrac{(x-3)}{(x-4)}$. Limits are 2/3, 1/2, ∞.

5. 0.

6. Break up into

$$\frac{\sqrt{x}-\sqrt{a}}{\sqrt{x^2-a^2}}+\frac{\sqrt{x-a}}{\sqrt{x^2-a^2}}.$$ Put $x = a \pm h$, $h \to 0$.

Limit $= 0 + 1/\sqrt{2a}$.

IX

1. $a_{n+1} - a_n = a =$ const. ; $a + b$; $3775a + 50b$.

3. nth term $= S_{n+1} - S_n$; common ratio is r.

4. Ratio is $(1 - r)/r$.

5. Convergent by comparison with series $1/n^2$. If x is r^2 the rth term is infinite.

6. (a) Divergent for all values of x (cf. $1/n$ series).

 (b) Divergent (cf. $1/n$ series).

 (c) Convergent if $|x| < 3/2$. (Note ratio of $(n+1)$th term to nth.)

X

1. (a) $1 - x + x^2 - x^3 + x^4$.

 (b) $a^{-1} + \frac{1}{5}x^5 a^{-6} + \frac{3}{25}x^{10}a^{-11} + \frac{11}{125}x^{15}a^{-16} + \frac{44}{625}x^{20}a^{-21}$.

 (c) $1 + 3x + 6x^2 + 10x^3 + 15x^4$.

2. Because $n(n+1)(n+2) \; . \; . \; . \; . \; (2n-2) = (2n-2)(2n-3) \; . \; . \; . \; . \; n$.

3. 2.9931.

4. 1.9520.

5. Put $a = \epsilon^m$, that is, $m = \log_\epsilon a$.

7. Expand ϵ^{-1} and group terms in pairs.

XI

1. 80°.
2. Divide figure into five triangles by lines to the corners from any internal point. Take sum of angles of these triangles. Hence sum of internal angles of pentagon = 5 × 2 rt. angles − 4 rt. angles.
3. 4 4/9, 5 5/9 (inches).
5. Tan 50° = 1.191 ; sec 50° = 1.555 ; cosec 50° = 1.305; cot 50° = 0.840 ; sin 140° = 0.643 ; cos 220° = − 0.765 ; tan 320° = − 0.840 ; tan − 50° = − 1.191 ; sin 40° = 0.643 ; sec − 40° = sec 40° = 1.305.

6. One.
7. 60° ; 8°11′ ; 0.17 rad. ; 0.22 rad.
8. 88.8 units of area ; 25 units of area.
9. Unit increased in ratio $\pi : 1$ therefore number of units decreased in same ratio, that is, 28.26, 7.95.

XII

1. 43.3 cm² ; 14.55 cm ; 6.197 cm ; 75 cm².
2. If the sides of the rhombus are the vectors **a** and **b**, the diagonals are (**a** + **b**) and (**a** − **b**). But $(\mathbf{a}+\mathbf{b})(\mathbf{a}-\mathbf{b}) = \mathbf{a}^2 - \mathbf{b}^2 = 0$ since $|\mathbf{a}| = |\mathbf{b}|$.
3. The vector is of const. length r, therefore if one end fixed, other lies on circle.
4. $\sqrt{(ab)^2 - (\mathbf{a}\cdot\mathbf{b})^2} = ab\sqrt{1-\cos^2\theta} = ab \sin\theta$; θ is angle between a and b.

XIII

1. Prove by converting to exponential forms.
2. Prove the tan and tanh forms by "rationalising" the denominator and using the results $\cos^2 a + \sin^2 a = 1$; $\cos^2 a - \sin^2 a = \cos 2a$, etc.
3. $a \pm jb = r\epsilon^{\pm j\theta}$ where $r^2 = a^2 + b^2$ and $\tan\theta = b/a$.

4. $z = \frac{1}{2a} + \frac{1}{2a}\epsilon^{-2j\theta}$

$1 + \frac{a - jb}{a + jb} = \frac{2a}{a + jb}.$

5. Express $\epsilon^{j\theta}$ as $\cos\theta + j\sin\theta$; similarly for $\epsilon^{-j\theta}$.

7. $\tanh(a + jb) = \frac{\epsilon^{a+jb} - \epsilon^{-a-jb}}{\epsilon^{a+jb} + \epsilon^{-a-jb}} = \frac{1 - \epsilon^{-2a}\,\epsilon^{-2jb}}{1 + \epsilon^{-2a}\,\epsilon^{-2jb}}$

which is of the given form, with $r = \epsilon^{-2a}$ and $\theta = -2b$.

8. $A = \frac{e(ac - bd) + f(bc + ad)}{e^2 + f^2}$

$B = \frac{e(bc + ad) - f(ac - bd)}{e^2 + f^2}$

$r^2 = \frac{(a^2 + b^2)(c^2 + d^2)}{e^2 + f^2}$

$\theta = \tan^{-1} b/a + \tan^{-1} d/c - \tan^{-1} f/e.$

9. $\epsilon^{j2\pi/3}\,\mathbf{a}$ and $\epsilon^{j4\pi/3}\,\mathbf{a}$. The vector sum of the three sides is zero.

XIV

1. i. $2ax + b$, $2a$, 0.

ii. $-(b/x^2) - 2(2c/x^3)$; $(2b/x^3) + (6c/x^4)$; $(-6b/x^4) - (24c/x^5)$.

iii. $30\cos x + 30\cos 2x + 30\cos 3x.$
$-30\sin x - 60\sin 2x - 90\sin 3x.$
$-30\cos x - 120\cos 2x - 270\cos 3x.$

iv. $\epsilon^{ax}(a\sin bx + b\cos bx)$;
$\epsilon^{ax}\{(a^2 - b^2)\sin bx + 2ab\cos bx\}$;
$\epsilon^{ax}\{(a^3 - 3ab^2)\sin bx + (3a^2b - b^3)\cos bx\}.$

v. $a^x\log_\epsilon a$; $a^x(\log_\epsilon a)^2$; $a^x(\log_\epsilon a)^3$.

vi. $a^x\{\log_\epsilon a \cdot \log_\epsilon(\sin x) + \cot x\}$;

$a^x\left[\log_\epsilon a\left\{2\cot x + \log_\epsilon a\log_\epsilon(\sin x)\right\} - \operatorname{cosec}^2 x\right]$;

$a^x\left[\log_\epsilon a\left\{3\log_\epsilon a\cot x + \log_\epsilon^2 a\log_\epsilon(\sin x) - 3\operatorname{cosec}^2 x\right\} + \cot x\operatorname{cosec}^2 x\right].$

2. $Q = 10\epsilon^{-t/CR}$.

3. $i = 10 \sin 5mt$.

4. i. $2ax + by$; $bx + 2cy$; $2a$; $2c$; b ; b.

ii. $\epsilon^{(ax+by)}(a \sin xy + y \cos xy)$;
$\epsilon^{(ax+by)}(b \sin xy + x \cos xy)$;
$\epsilon^{(ax+by)}\{(a^2 - y^1) \sin xy + 2ay \cos xy\}$;
$\epsilon^{(ax+by)}\{(b^2 - x^2) \sin xy + 2\, bx \cos xy\}$;
$\epsilon^{(ax+by)}\{(ab - xy) \sin xy + (1 + ax + by) \cos xy\}$.
The same.

5. Max. when $x = 1$ and min. when $x = -1$; the same.

6. 6482 km.

7. Putting s for the distance, $ds/dt = 500 - 10t = 0$ when $t = 50$ secs.
From $t = 0$ to $t = 50$, $s = 0{\cdot}125$ km ; 100 secs.

8. $x + y$ has the minimum value $2\sqrt{c}$ when $x = y$.

9. $A - B = C$; $A + B = D$.

XV

1. $(k + \omega j)\mathbf{v}$; $\{(k^2 - \omega^2) + 2\omega jk\}\mathbf{v}$;
$\sqrt{k^2 + \omega^2}\, v_0\epsilon^{kt} \cos(\omega t + \psi + \tan^{-1} \omega/k)$;
$(k^2 + \omega^2)\, v_0\epsilon^{kt} \cos(\omega t + \psi + 2 \tan^{-1} \omega/k)$.

2. i. $\dfrac{\log_\epsilon (ax+b)}{a} +$ const.

ii. Break up into $\dfrac{a}{c}\left\{1 - \dfrac{d}{cx + d}\right\} + \dfrac{b}{cx + d}$.

Integral is $\dfrac{ax}{c} + \dfrac{bc - ad}{c^2} \log_\epsilon (cx + d) +$ const.

iii. Sec x + const.

iv. $-\dfrac{1}{a} \cot^{-1}\left(\dfrac{x}{a}\right) +$ const.

v. $\text{Tan}^{-1}\, \epsilon^x +$ const.

3. i. $\log_\epsilon x(x^2dx)$; $(x^3/9)(3 \log_\epsilon x - 1) +$ const.

ii. $(\log_e x)^2\,(dx)$; $x(\log_e x)^2 - 2\int \log_e x\,dx$.

Then $\log_e x(dx)$; $x \log x - \int x \frac{1}{x}\,dx$.

Final result $x(\log_e x)^2 - 2x(\log_e x) + 2x + \text{const.}$

iii. $\tan^{-1}x(dx)$; $x \tan^{-1}x - \frac{1}{2}\log_e(1 + x^2) + \text{const.}$

4. Put $\int f(x)\,dx = F(x)$.

5. (i) 2. (ii) 0. (iii) π ; put $\sin^2\theta = \frac{1}{2}(1 - \cos 2\theta)$.

(iv) 0 ; put $\sin m\theta \sin n\theta = \frac{1}{2}\{\cos(m - n)\,\theta - \cos(m + n)\theta\}$.

(v) 0.

7. $f(t) = \frac{4A}{\pi}\left\{\sin \omega t + \frac{1}{3}\sin 3\omega t + \frac{1}{5}\sin 5\omega t + \ldots\right\}$

Note : $\frac{2}{T}\int_0^T f(t) \sin n\omega t\,dt = \frac{2}{T}\int_0^{T/2} A \sin n\omega t\,dt$

$- \frac{2}{T}\int_{T/2}^{T} A \sin n\omega t\,dt.$

8. The even harmonics repeat *without* change of sign every half period.

9. Yes. If T is the period, $f_1(t + T)f_2(t + T) = f_1(t)f_2(t)$ is satisfied if $T = mT_1 = nT_2$, and since T is the least value which satisfies these equalities it is the least common multiple of T_1 and T_2.

INDEX

INDEX

In this index the first figure of each item is the page number on which the required section begins and the second figure [in square brackets] is the actual section number.