CONCEPTS OF INFINITY

Inaugural Lecture of the
Professor of Pure Mathematics
delivered at the College
on 20 March 1962

by

PROFESSOR J. D. WESTON
B.Sc. (Eng.), Ph.D., D.Sc.

GWEDDW CREFFT HEB EI DAWN

UNIVERSITY COLLEGE OF SWANSEA

UNIVERSITY COLLEGE OF SWANSEA

CONCEPTS OF INFINITY

Inaugural Lecture of the
Professor of Pure Mathematics
delivered at the College
on 20 March 1962

by

PROFESSOR J. D. WESTON
B.Sc. (Eng.), Ph.D., D.Sc.

PRINTED IN GREAT BRITAIN

CONCEPTS OF INFINITY

IT is well known that mathematical studies have prospered in the University College of Swansea since its foundation; but I believe this is the first inaugural lecture to be given here by a professor of pure mathematics. I should like to be able, on this occasion, to explain what pure mathematics is about, for that is certainly a matter on which there are widespread misunderstandings. But such an objective would require many lectures, and a variety of lecturers. The task I have set myself is therefore a very limited one: I propose to discuss just a few of the ideas that make up the foundations of the subject. I shall try to do this in a way that will illustrate the development of mathematical thought in comparatively recent times.

Although the history of mathematics goes back to classical antiquity and beyond, pure mathematics can hardly be said to have existed before the middle of the nineteenth century. It has grown prodigiously since then, and the knowledge already accumulated greatly exceeds the capacity of any individual mind. To be effective nowadays, a mathematician has to specialize fairly narrowly. My own special interests belong to the branch of pure mathematics known as 'analysis'. This can be loosely described as the theory of infinite processes, and its origins can be traced to the invention of the calculus in the seventeenth century.

In essence, the calculus was a method of representing some very subtle concepts by an algebraic formalism that was rather easy to manipulate. The power of this method was demonstrated by the spectacular achievements of Newton,[1] which marked the beginning of what is now called applied mathematics. But the ease of manipulation

[1] 1642–1727.

was deceptive: few of the early practitioners seem to have understood the conceptual basis of the calculus at all clearly, and some of them were tempted to enlarge the formal structure without paying proper attention to the foundations. This sort of thing happened on a large scale in the eighteenth century, which was an age of decadence for mathematics. Under the dominating influence of Euler,[1] the formalities of the calculus, and related formalities such as those of infinite series, were elaborated with much ingenuity but little regard for logic. In 1734, the philosopher Berkeley[2] published a tract called *The Analyst* in which the unsoundness of much contemporary mathematical thought was clearly exposed; but although this checked some of the wilder excesses, it provoked no significant development of analysis. The time, apparently, was not ripe.

Great advances were made, however, during the nineteenth century, mainly in France and Germany. Among the pioneers were Gauss[3] and Cauchy[4] in the early part of the century, Riemann[5] and Weierstrass[6] in the middle, Dedekind[7] and Cantor[8] towards the end. Of these, Cantor probably had the most original mind, and it may be significant that he suffered a good deal from mental illness. By the end of the century, the scope and power of the calculus had immensely increased, and its foundations were secure. Moreover, through the work of Cantor, the modern theory of infinite sets had begun to evolve. This theory has profoundly influenced the development of analysis in the present century, and I shall try to describe some of its fundamental ideas.

Consider first the sequence of natural numbers—the numbers we use for counting:

$$1, 2, 3, 4, \ldots .$$

[1] 1707–83. [2] 1685–1753. [3] 1777–1855. [4] 1789–1857.
[5] 1826–66. [6] 1815–97. [7] 1831–1916. [8] 1845–1918.

This sequence does not come to an end, for there is no largest number. The totality of natural numbers is thus an example of an *infinite set*. (It is, of course, a set of *ideas*, existing in the mind; there may well be no such thing as an infinite set of physical objects.) In contrast, a *finite* set is a collection of individuals (ideas, or physical objects) which can be counted, in the sense that a different natural number can be assigned to each member of the set, the numbers so assigned being those that are not greater than some particular number. If we have two finite sets, A and B, we can find out whether or not they have the same number of members by counting both sets. But there is another method, which does not involve counting. It consists in pairing members of A with members of B: if this can be done in such a way that each member of A has a unique counterpart in B, and there are no members of B left over, then, and only then, are the sets A and B equally numerous. Thus, while I cannot say how many shoes I possess, for I have not counted them recently, I have no hesitation in saying that the number of left shoes that I have is the same as the number of right shoes. Now this second method of comparison is applicable to infinite sets as well as to finite ones. For example, let A be the totality of natural numbers and let B be the totality of even numbers,

$$2, 4, 6, 8, \ldots .$$

There is an obvious pairing here: the members of B are obtained by doubling those of A. We can say, therefore, that these two infinite sets are equally numerous. Since, in this case, B is a part of A, it appears that a part can be 'equal' to the whole (if we use the word 'equal' to mean 'equally numerous'). This is a characteristic property of infinite sets.

The fact that the members of one set can sometimes be paired with those of another, giving what is technically known as a 'one-to-one correspondence', suggests a method of classifying infinite sets. A set that can be put into one-to-one correspondence with the set of all natural numbers is said to be *countably infinite*; and for convenience the word 'countable' is used to mean 'either finite or countably infinite' (though of course only a finite set can actually be counted). It is quite easy to show that any part of a countable set is countable, and that a set is countable if it can be divided into a countable set of parts each of which is countable. These facts enable us to discern many examples of countably infinite sets. One such example is the set of all the numbers with which elementary arithmetic is concerned—the so-called *rational* numbers, of which I shall have more to say later. But, as Cantor was the first to show, uncountable sets also exist: an example is the set of all sets of natural numbers. The set of all permutations of the *sequence* of natural numbers is also uncountable.

The idea of a sequence, which is one of the primitive notions on which mathematics is founded, involves the recognition of a natural order of precedence among the natural numbers. These numbers constitute, in fact, not merely a set but a set with certain structural properties. By considering some of those properties in isolation, using a typically mathematical process of abstraction, we can arrive at the idea of a 'well-ordered set'. A set A is said to be *well ordered* if a notion of precedence is defined among its members in such a way that the following two conditions are satisfied: first, if x and y are members of A, and x precedes y, then y does not precede x; second, if B is any part of A having more than one member, then there is a member of B which precedes every other member of B. When a set of natural numbers is ordered in the usual way, so that the lesser of two numbers is

considered to 'precede' the greater, we have an example of a well-ordered set. It is this fact that gives validity to the well-known principle of mathematical induction, which is a powerful instrument of deductive reasoning.

A set may have a natural ordering with which it is not well ordered. For example, the set of all positive rational numbers, ordered by magnitude in the usual way, is not well ordered, since it has no smallest member. However, this set, and indeed any countable set, can be well ordered merely by putting it into one-to-one correspondence with a set of natural numbers. In 1904 Zermelo,[1] a German mathematician, published a proof that *every* set can be well ordered. This led to the formulation of an extended principle of mathematical induction. The new principle, known as 'transfinite induction', is indispensable in modern analysis and has important applications in several other branches of mathematics.

For a considerable time, however, the validity of Zermelo's result was a matter of dispute among leading mathematicians. This was because the proof depended in an essential way on the proposition that if an infinite set is divided into infinitely many parts, then there is a set consisting of exactly one member from each of those parts. This proposition, which has become known as 'the axiom of choice', may seem at first to be self-evident; but it was objected to on the grounds that the set whose existence it asserts might require for its specification infinitely many —perhaps uncountably many—individual acts of choice. It was held to be inconceivable that such a procedure could ever be completed. Put in this way, the objection resembles the famous argument put forward by Zeno,[2] in the fifth century B.C., to show that Achilles could never catch up with the tortoise; but it is much harder to refute.

The controversy over the axiom of choice arose at a

[1] 1871–1953. [2] 495–435.

time when mathematicians were very sensitive about the need for care in defining sets. Several disconcerting paradoxes had been noticed, and these suggested that there might be serious flaws in the foundations of Cantor's theory. Some mathematicians, indeed, were disposed to reject the theory altogether. The trouble was that what appeared to be a proper definition of a set sometimes turned out not to be. For example, there are natural numbers that can be expressed in fewer than fifteen words of the English language, and we can consider (or so it would seem) the set that consists of all such numbers: this is surely a finite set, since the English language does not have an infinity of words (we could fix the number of words available by deciding on a particular dictionary). It should therefore make sense to speak of 'the smallest natural number that cannot be expressed in fewer than fifteen English words'; but I have used only fourteen words to express this number, so we have a contradiction. The only conclusion we can draw is that, contrary to appearance, our 'set' has not been properly defined. It is not surprising that in the presence of such anomalies the axiom of choice was viewed with suspicion: how could one be sure that its unrestricted use might not lead to a nasty contradiction?

This was one of the questions that gave rise to the modern science of mathematical logic; but it was a long time before a satisfactory answer was given. As an interim measure, mathematical proofs were classified as 'constructive' or 'non-constructive', those that depended on the axiom of choice being regarded as non-constructive. Many mathematicians thought it prudent to give constructive proofs whenever they could manage to do so, and some went so far as to regard non-constructive proofs as entirely inadmissible. However, bold exploitation of the axiom of choice also took place, notably in Poland,

where the new method of transfinite induction was used to great effect in the development of analysis.

Eventually, in 1940, the celebrated logician Kurt Gödel produced what most analysts have interpreted as a vindication of the axiom of choice. He showed that if, by legitimate reasoning, one could deduce a contradiction in the theory of sets with the help of the axiom of choice, then one could also deduce such a contradiction without invoking this axiom; in other words, the axiom of choice is logically consistent with those propositions in the theory of sets that can be established without its aid. That it is also independent of those propositions, in the sense that it cannot be deduced from them, was proved in 1956 by a young American logician, Elliott Mendelson. Thus, while constructive proofs may still be preferable to non-constructive ones when they are available, it must be expected that some mathematical propositions will be provable only by non-constructive methods. Some of the most important propositions in analysis seem to be of this kind.

The ultimate ingredients of analysis are the natural numbers, but it is not so much the numbers themselves as certain relations between them that are of interest. Relations that can be expressed in terms of addition and multiplication are particularly important, and the concept of number can be usefully generalized by means of such relations. For example, the idea of a *ratio* is expressible in terms of multiplication, for when we say that p bears the same ratio to q as r does to s we mean that the product of p and s is the same as the product of q and r. Ratios can be added, multiplied, and compared as to magnitude, according to definitions that naturally suggest themselves; the ratios then form an arithmetical system which can be regarded as an enlargement of the system of natural numbers (since each natural number can be identified with the ratio it bears to the number 1). In the enlarged system the operation of division can be carried out without restriction: a ratio of ratios is a ratio. By considering differences of ratios one can define an even larger system, that of the rational numbers. The positive rational numbers can be identified with the ratios, and with each of these we associate the 'opposite' negative number, the system being completed by the special number zero. Except that we cannot divide by zero, all four of the fundamental operations of arithmetic have free play within the system of rational numbers, subject to the algebraic laws that are implicit in the definitions of these operations.

The Greeks, with their love of proportion, were deeply interested in the arithmetical properties of ratios. From the time of Pythagoras,[1] they knew that the number 2, among other natural numbers, cannot be expressed as the product of two equal ratios. However, in Greek geometry it was implicitly assumed that length and area could be measured by numbers belonging to some arithmetical system that included the ratios, and that the operations of addition and multiplication in this system corresponded to certain geometrical constructions. From these rather sweeping assumptions it was easy to deduce that if the length of one side of a square were represented by the number 1, then the length of each diagonal would have to be represented by a number whose product with itself is 2. (I am alluding, of course, to the famous theorem about 'the square on the hypotenuse'.) It was therefore necessary to postulate the existence of 'numbers' which, though not ratios, were yet comparable with ratios and subject to the same kind of arithmetic. This was done explicitly by Eudoxus[2] (a pupil of Plato[3]), who proposed a system of postulates which seems to have

[1] 569–500. [2] 408–355. [3] 429–348.

been the foundation of subsequent Greek thought on this subject. It has often been said that the Greeks 'discovered' irrational numbers, but this is scarcely correct: they discovered a need for such numbers, and they met the need by formulating a hypothesis. Their hypothesis remained unverified—and apparently unquestioned—for some 2,000 years.

Greek arithmetic was concerned only with what we now call 'positive' numbers. Ideas of negative numbers, and of zero, came much later: they seem to have filtered into European thought, from Hindu and Islamic sources, during the Middle Ages, when algebra was beginning to take shape. The early algebraists, however, were interested in the manipulation of symbols rather than the clarification of concepts. They contributed to the development of mathematics mainly by inventing notational devices which served to mechanize some of the processes of arithmetical reasoning. The system of numbers in which this reasoning was supposed to operate was an extension of the Greek system, and no less hypothetical.

The rise of analysis in the nineteenth century sharpened the need for an arithmetical system larger than that of the rational numbers. The required system had to conform to a rather exacting specification. It had to resemble the system of rational numbers in providing full scope for the fundamental operations of arithmetic, and in being so ordered that every number other than zero would be either positive (that is, greater than zero) or negative (less than zero); and this ordering had to be compatible with the arithmetical structure, in the sense that sums and products of positive numbers should always be positive. But beyond these requirements a crucially important condition had to be satisfied; namely, that if one were to divide the system into two parts, A and B, in such a way as to make every member of A less than every member of B, then either A would have a greatest member or B would have a least member. That a system having all these properties could actually exist was not known for certain until 1872, when Dedekind showed that certain infinite sets of rational numbers could be used to construct exactly what was needed. Dedekind's system is known as 'the continuum', and its individual members are called 'real numbers'. An alternative method of constructing the continuum was given a few years later by Cantor, who made a very detailed study of its fundamental properties.

One of the more obvious properties of the continuum has to do with the idea of 'bounds' for sets of numbers. A number b is said to be an *upper bound* for a set S of numbers if there is no member of S greater than b. Now, in the continuum, any set that has upper bounds must have a *least* upper bound, and this fact has many important consequences. For example, consider the set that consists of every real number whose product with itself is less than 2: this set has a least upper bound, and it is easy to show that the product of this bound with itself is exactly 2. Thus the continuum includes some at least of the irrational numbers needed in Greek geometry. In fact it includes all such numbers.

Suppose now that we have in mind some countable set C (perhaps a set of natural numbers) and a prescription that assigns what I shall call a 'value' to each member of C, each value being a real number which is either positive or zero. Then we can assign a *total* value to each finite part of C by adding together the values attached to the individual members of that part. (Different members may have the same value, and it is to be understood that such a value would be repeated an appropriate number of times in the summation—as in the adage '2 and 2 make 4'.) The totals thus obtained, by considering all the finite

parts of C, form a set of real numbers which may or may not have upper bounds. If C happens to be a finite set, such upper bounds certainly exist, and the least of them is the total value of the whole set C. However, if C is an infinite set its total value cannot be defined arithmetically, since addition is essentially a finite process: even a high-speed computer can never perform an infinity of operations. Nevertheless, if the set of totals for the finite parts of C has upper bounds, then, in the continuum, it has a least upper bound; and it is natural to *define* the total value of C to be this least upper bound. By so doing we are able to transcend arithmetic. We have here one of the simplest examples of an 'infinite process'; it is a process of summation involving a countable infinity of terms, none of which is negative.

To consider a specific example of this kind of summation, suppose that C is the set of all natural numbers, and let the value 1 be assigned to the number 1, $\frac{1}{2}$ to the number 2, $\frac{1}{4}$ to the number 3, and so on, each value being obtained from the preceding one by halving. It is then a fairly simple matter to show that the total value is 2. In this case each individual term is rational, and the total is rational. But there are cases in which the terms are rational and the total is irrational. For example, take C to be the set of natural numbers as before, let the value 1 be assigned to the number 1, and let each subsequent term be obtained by dividing its predecessor by a number that is 1 at the first step and is then increased by 1 for each subsequent step: thus the first term is 1, the second term is also 1, the third is $\frac{1}{2}$, the fourth is $\frac{1}{6}$, the fifth is $\frac{1}{24}$, and so on. It is then easy to show that the total is an irrational number between 2 and 3. In this case, however, and in others like it, we can find rational numbers as close to the total as we please by adding finitely many of the terms, appropriately chosen. (Any irrational number can be approximated arbitrarily closely by rational numbers, and this fact is very important in some of the applications of analysis—particularly in applications that involve the use of digital computers, for such machines cannot handle numbers that are not rational.)

The process of summation that I have been describing is subject to an awkward limitation. In order that the total value of the set C should exist in the continuum, it is necessary that the set of totals for the finite parts of C should have upper bounds in the continuum. That this condition is not always satisfied can be seen by considering the case in which C is an infinite set and the value given to each of its members is 1. The limitation can be overcome by augmenting the continuum in a very simple way. We take any mathematical entity which is not a number (for example, it might be the set of all natural numbers) and call it 'infinity'. We then decree that infinity shall be deemed to be 'greater than' any real number. The augmented continuum consists of the real numbers together with infinity, and in this system *every* set has at least one upper bound. A set of real numbers which has no upper bound in the continuum evidently has infinity as its least upper bound in the augmented continuum. The restriction on our infinite process of summation is thus removed. In the case where every term is 1, and in many other cases, the total is infinity.

This device for providing the whole continuum with an upper bound is rather more than the mere linguistic trick that it might seem to be. The entity that we call infinity is not a number, but it can be treated in several respects as though it were; in fact it can be allowed to take part in a limited kind of arithmetic. For instance, it is natural to define the result of 'adding' infinity to any real number, or to itself, to be infinity. Moreover, our infinite process of summation can be generalized by

allowing one or more of the terms to be infinity, the total also being infinity in such a case, by definition. This generalized process obeys laws similar to those of ordinary addition: for instance, if the countable set C is divided into parts in any way, and the total values of the separate parts are then summed, the resulting 'grand total' is always the same.

The usefulness of these notions can be illustrated by considering the important concept of *length*. For this purpose it is helpful to visualize the continuum as a line running from left to right: the real numbers are then thought of as *points* on the line, the phrase 'greater than' being interpreted as 'to the right of'. A set of real numbers is called an *interval* if there are no gaps in it: that is to say, if there is no number which does not belong to the set yet lies between two numbers that do belong. An interval may, for example, consist of all the numbers that lie between two given numbers, together, perhaps, with one or both of the given numbers: in such a case, by subtracting the lesser of the given numbers from the greater, we obtain a positive real number which is called the 'length' of the interval. Another possibility is that an interval may consist of a single point, in which case we define its length to be zero. The length of any interval that belongs to neither of these two categories is defined to be infinity; for example, the whole continuum is an interval whose length is infinity, and so is the set of all positive real numbers. Having thus assigned a length to every interval, we can inquire into the possibility of extending the definition of length so that it applies to a larger class of sets.

A set that is made up of finitely many non-overlapping intervals can be considered to have as its length the sum of the lengths of the component intervals. It is of course necessary to prove that this definition is not ambiguous, by showing that different ways of dissecting the set into intervals will always give the same total length: it is not difficult to do this. Let us agree to call a set 'simple' if it is made up in this way. A little investigation shows that the class of all simple sets has three properties which I shall now describe. First, if we have a set belonging to the class and we add a given real number to each of its points (a procedure which can be visualized as a shifting of the set bodily to the right or to the left), then the set so obtained belongs to the class and has the same length as the original set. Secondly, if a set belonging to the class is partly included in another such set, then both the included part and the excluded part are sets that belong to the class. Thirdly, if a set can be dissected into two parts that belong to the class and have no point in common, then the whole set belongs to the class and its length is the sum of the lengths of the separate parts. Any larger class of sets to which the concept of length might be usefully extended ought to have at least these three properties.

The third property is somewhat stronger than it looks: we can easily deduce, by mathematical induction, that if a set can be dissected into *finitely many* non-overlapping parts each of which belongs to a class having this property, then the whole set belongs to the class and its length is the sum of the lengths of the separate parts. For this reason the property is described as 'finite additivity'. For certain purposes of analysis, however, it is very desirable to have an even stronger property, known as 'countable additivity': this is defined in the same way as finite additivity except that we replace the phrase 'finitely many' by 'countably many', using our infinite process of summation to interpret the phrase 'the sum of the lengths' in the case of a set which is divided into a countable infinity of parts.

It is clear that if the concept of length can be extended so as to meet the requirement of countable additivity, then every countable set of real numbers must have zero length, since each individual point has zero length. It follows that the requirement could not be met if the continuum were countable, since the length of the continuum is infinity. But in fact the continuum is not countable, and the requirement of countable additivity *can* be met. The first definition of length that was satisfactory in this respect was given in 1898 by the French mathematician Emile Borel[1] (who later took a prominent part in the debate on the axiom of choice). The class of sets to which Borel extended the concept of length is in fact the smallest class that contains the intervals and satisfies the other requirements. Sets that belong to this class are known as 'Borel sets'. Every countable set of real numbers is, of course, a Borel set whose length is zero; in particular, every set of rational numbers has zero length. Thus, although every interval of positive length contains infinitely many rational numbers, such an interval owes its length entirely to the irrational numbers that it contains.

In 1902 a further extension of the concept of length was made by Lebesgue,[2] a French mathematician whose work forms one of the corner-stones of modern analysis. The sets that Lebesgue considered are called 'measurable' sets. The class of all measurable sets is larger than that of the Borel sets, but in a sense it is not much larger, for every measurable set can be dissected into two parts of which one is a Borel set and the other is a measurable set of zero length. This small difference turned out to be of great importance for the solution of certain problems that had arisen in the development of the calculus. These problems concerned the process known as 'integration'. The modern theory of this process, created during the

[1] 1871–1956. [2] 1875–1941.

present century, is one of the most sublime achievements of the human intellect.

I shall not attempt now to explain precisely what the measurable sets are, or why they are so important; but I have said enough to indicate that the measurable sets of zero length have a special significance. Much of this significance comes from the fact that if a set is measurable and has zero length then every part of it is a measurable set of zero length; whereas certain Borel sets of zero length have parts which are not Borel sets. This deficiency in the system of Borel sets was revealed by Lebesgue in 1905. In the same year the Italian analyst Vitali[1] gave the first proof of the existence in the continuum of sets that are not measurable. His proof shows, in fact, that it is fundamentally impossible to extend the concept of length from the class of all simple sets to the class of all sets in the continuum in such a way as to satisfy the requirement of countable additivity. This is obviously an important result, if only because it prevents those who know about it from wasting their time in trying to devise a definition of length that would apply to every set in the continuum. It is interesting to note, therefore, that Vitali's proof depended on the axiom of choice, and that no-one has succeeded in showing that non-measurable sets exist without making some appeal to this axiom.

In conclusion, I should like to mention some facts that make a strong contrast with what I have just said. If we are willing to accept finite additivity instead of countable additivity (a severe restriction), then it *is* possible to extend the concept of length from the class of all measurable sets to the class of all sets in the continuum. A method of proving this was described in 1923 by Banach,[2] one of the leaders of the Polish school of analysis that was beginning to flourish strongly at about that time. Banach's

[1] 1875–1932. [2] 1892–1941.

method, however, involved the use of transfinite induction, and no constructive proof of the result has yet been found. (I should add that this particular result is by no means the most important of Banach's contributions to analysis; but it does answer an interesting philosophical question, and the method has other applications.)

I have come to the end of this lecture, and I know that I have not done much to illuminate the nature of pure mathematics. But one thing at least must I think be clear: the mere existence of infinite sets of numbers is a guarantee that the subject will never have been fully explored. The scope for mathematical research is, and will always be, infinite.

PRINTED IN GREAT BRITAIN
AT THE UNIVERSITY PRESS, OXFORD
BY VIVIAN RIDLER
PRINTER TO THE UNIVERSITY