INTRODUCTION

The physicist and the engineer feel at a loss when a process of mathematical analysis carries them out of contact with physical reality, for they like each symbol to be physically identifiable and each step to be guided by physical intuition. Mathematicians with a geometrical turn of mind feel the same revulsion against pure analysis, and there is a famous example of this in the deliberate revolt of Poinsot, in his study of the motion of rigid bodies, against the uncompromising attitude of Lagrange, who excluded all diagrams from his *Mécanique analytique*.

But there are complex situations which baffle the intuition, as in the theory of elasticity when we try to keep track of six components of stress and three components of displacement, and in such cases there seems to be nothing for it but to throw oneself into the mathematical formulae and hope for the best without intuitive guidance. It was to escape this fate that Professor W. Prager and I evolved what we called the method of the hypercircle in function-space (Prager and Synge (1)†).

In the hypercircle method as applied to elasticity we substitute for the direct intuition of stress and displacement the intuition of Euclidean geometry, extended from three dimensions to an infinity of dimensions, an extension much less troublesome than one would at first suppose. To each state of stress of the body (six functions of three coordinates) there corresponds a single point of function-space, and this space is endowed with a metric in a way which seems very natural physically, the square of the distance of a point from the origin of function-space (the state of zero stress) being twice the strain energy of the state corresponding to that point. This representation of states of stress by points of function-space would be merely a rather trivial game were it not for the fact that the geometrical picture fits together in a remarkable way, the minimum principles which hold in elastic equilibrium taking on a simple geometrical interpretation analogous to the fact that the perpendicular dropped from a point on a plane is the shortest distance to that plane. The method is called the method of the hypercircle because a certain geometrical figure (called a hypercircle by analogy with the circle of ordinary geometry) appears in the theory as the locus of possible positions of the point corresponding to the unknown solution of a problem of elastic equilibrium.

† See Bibliography at end for all references.

It was the complexity of the theory of elasticity which forced us to invent this geometrical approach, but Professor Prager rightly suggested at the time that the method of the hypercircle had a much wider range of applicability. It can, indeed, be used for many boundary value problems (Synge (1)), including the comparatively simple ones associated with Laplace's equation in a plane (torsion, electrostatic capacity, etc.), the essential condition being that the analytic problem can be presented as the geometrical problem of finding the point of intersection of two orthogonal linear subspaces in a suitably chosen function-space (the analogues of two perpendicular straight lines in ordinary space). This is linked with the possibility of deriving the partial differential equations of the problems considered from variational principles (McConnell (1)).

A wide class of boundary value problems of mathematical physics may thus be given a geometrical form, a wedding of analysis to geometry which is pleasant to contemplate. It is by no means a sterile marriage, for though this geometrical picture of two intersecting subspaces with the goal at the point of intersection may not tell us how to attain that goal, it does suggest that we should advance towards it in both of the subspaces instead of in one only, a great improvement on standard procedures because the use of both subspaces enables us to estimate our error accurately at any stage of the advance.

Ordinarily one works in one linear subspace. For example, to solve the torsion problem for a square section one works in the linear subspace of harmonic functions, setting up a series of such functions with coefficients chosen to satisfy the boundary conditions. For a square this is an excellent plan, because we get an exact solution in this way. But our success is due to the simplicity of the square, and we cannot get the solution in this simple way for a more general section, say a square with one corner knocked off.

The hypercircle method is applied in detail to the torsion problem in Chapter 4, and here it is enough to say that, when we advance towards the solution through both the linear subspaces, we get a controlled approximation. Knowing that exact formal solutions can be obtained only in a few special cases, we are prepared from the first to break off with an approximation, but when we do break off we know just where we stand in the sense that we know how far we are away from the solution in terms of the metric of function-space, or equivalently in terms of an integral of the square of the error.

The method of the hypercircle makes contact on one side with the theory of Hilbert spaces and on the other with equations in finite differences and the relaxation method of Sir Richard Southwell.

Mathematicians will recognize at once that (when the metric is positive-definite, as it is throughout Part II) the function-space is a Hilbert space, shorn of those refinements which we do not need because we are not concerned with existence theorems; we know or assume that the solution exists, and are interested only in finding it.† The Hilbert space is further simplified by our concentration on real functions; this makes our function-space much easier to think about than the Hilbert space of quantum mechanics.

But Part III introduces a less familiar type of function-space. It is not a Hilbert space because the metric is now indefinite, and its geometry is analogous, not to Euclidean geometry, but to the geometry of Minkowski in the space-time of special relativity. The method of the hypercircle (now the method of the pseudohypercircle) softens and serves less firmly as a guide to approximate solutions, arithmetical bounds being no longer available and minimum principles changing to stationary principles. This type of function-space occurs in the geometrization of problems of forced vibrations, mechanical and electromagnetic. Free vibrations are touched on only lightly and the determination of eigenvalues is not discussed, for although a geometrical approach is powerful here, it is not the geometry appropriate to the method of the book.

The method of the hypercircle may be called a relaxation method because points in the two linear subspaces in which we advance towards the solution represent solutions of problems (sometimes physically artificial) in which some of the conditions of the original problem are relaxed. Equations in finite differences are by no means an essential part of the method of the hypercircle; they come in when we use a certain technique, the method of pyramid functions, to get points on these subspaces of relaxation. This technique enables us to handle bounding curves of any form in plane problems.

As commonly used, equations in finite differences are substituted for the differential equation of the problem, and one proceeds with a general confidence that the solution of the equations in finite differences will not be far off the solution of the differential equation, provided the grid is fine. The method of the hypercircle is much more precise. We never replace a differential equation by equations in finite differences,

† For an introduction to the mathematical theory of Hilbert space, see Stone (1) or Halmos (1).

but use the latter within the framework of the hypercircle method, the solution of the equations in finite differences giving us accurately, not the solution to our original problem, but the centre and radius of the hypercircle on which the point representing that solution lies. The finer the grid, the closer we are to the solution, but with any grid at all we know how far we are away from the solution. Thus the torsional rigidity of a regular hexagon is given on p. 268 with an accuracy of 0·4 %, and we know exactly what we mean by this statement of error, upper and lower bounds being established. We might say that the method of the hypercircle, when combined with pyramid functions, is a refined finite difference method with rigorously controlled error.

It is impossible to write a book which carries every reader along at the right speed; it is bound to be too slow for some and too fast for others. I have thought it wiser to err on the side of slowness, particularly at the beginning, for if the reader fails to get hold of the geometry of function-space with a feeling of security in its use, then the book has failed in its purpose. I have filled in a good deal of detail in the calculations, knowing that one can grasp the full significance of a general argument only by seeing cases worked out in full detail. My own mind works that way, and I assume that it is true of many others. A few exercises are inserted at the end of each section, mostly of a very simple nature; they serve to keep one's feet on the ground.

This book has been written on the assumption that frequent appeal to geometrical intuition, not for proof but for suggestion, will please the reader as much as it pleases the author. Those who prefer to take their analysis neat will find other treatments of the problem of bounding the solutions of boundary value problems elsewhere, without diagrams or geometrical ideas, notably in papers by Diaz (1, 2, 3) and Cooperman (1).

We must reconcile ourselves to the fact that, in mathematics, there is no single universal mode of thought which reveals in a flash the inner meaning of an argument so that it becomes easy and almost self-evident to everyone. What is natural and easy for one man is often artificial and difficult for another. Anyone who believes in the simplicity and uniqueness of mathematical thought would be well advised to read Hadamard's (1) book on *The Psychology of Invention in the Mathematical Field*; incidentally, his remarks (*op. cit.* p. 88) on Hilbert's use of diagrams in setting up the logical principles of geometry are apposite to the question of diagrams of function-space.

PART I

NO METRIC

CHAPTER 1

GEOMETRY OF FUNCTION-SPACE WITHOUT A METRIC

1·1. INTRODUCTORY IDEAS

Representation of numbers

Logic rules mathematics, but we are human beings and the ways in which we see or understand things in mathematics are not always logical ways. The fascination of mathematics lies in the interplay of intuition and logic—the discipline of intuition by logic; neither intuition nor logic alone suffices.

Logically, the concept of a real number is independent of the idea of a point on a line. But mathematicians habitually think of real numbers as points on a line; in complicated situations this representation is essential, for we are human beings with limited

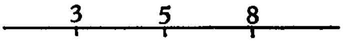

Fig. 1·11. Good representation.

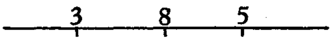

Fig. 1·12. Bad representation.

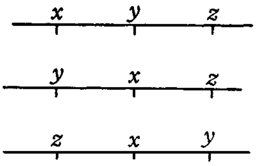

Fig. 1·13. Representations of three unknown real numbers.

facilities of thought (otherwise all mathematics would be obvious to us immediately). The extension of the representation into the complex plane is equally essential. These are things we cannot do without, although we are at all times ready to admit the ultimate authority of logic.

Even in the representation of real and complex numbers there are snares for the unwary. Suppose we are asked to think of the numbers 3, 5, 8. Immediately we think of a representation as in Fig. 1·11. No one would be so foolish as to make a representation as in Fig. 1·12; the order is wrong. But suppose we are asked to think of three *unknown* real numbers, x, y, z. Three different representations suggest themselves (Fig. 1·13); there are others also, in some of which two or all of the three points coincide. Which are good representations and which are bad? If we were given more information (e.g. $x < y < z$) then we might be able to decide, but if no such information is available, all the representations are bad, because each of them commits us further than we ought to be committed at this stage of ignorance.

In mathematics we are constantly being asked to find an unknown number or a set of unknown numbers. To think of the number or numbers, we crave a representation. Any representation we make is dangerous—it may indicate something that is actually false. It is like forming a detailed mental image of John Smith when he appears on page one of a novel, only to find that this mental image is entirely wrong when we read the description of him on page two.

In disgust at the unreliability of representations based on insufficient data, some people try to get on without them. When they have to deal with three numbers x, y, z, they do not attempt to represent them, preferring to carry out the formal manipulation of symbols according to established rules. But those who follow this cautious policy forego a powerful aid to thought, and the wisest course seems to be one of compromise, in which we use representations in a fluid and tentative way, preferring representations which do not say too much.

To sum up: if we are asked to solve a problem involving unknown real numbers, we have two options:

(i) Carry out formal manipulations according to established rules, and make no representation at all.

(ii) Make tentative representations as guides for thought, modifying them as further information becomes available. The standard representation of real numbers is by points on a line, and that of complex numbers by points in a plane (Argand diagram).

Representation of functions

In the problems with which we shall be concerned, we sometimes seek unknown numbers, but usually it is unknown *functions* that we seek. What is a function, and how are we to represent it?

In the eighteenth century a function of x meant a formula involving the letter x (like $1+x^2$ or $\sin x$). Now we say that $f(x)$ is a function of x if there exists a rule by which a value of $f(x)$ corresponds to each value of x in an assigned range.

As for the representations of functions, three are familiar:

(i) a formula,
(ii) a graph,
(iii) a tabulation of values.

Let us now suppose that we have before us a problem in which we are required to find an unknown function $f(x)$ which satisfies a given differential equation for some range of values of x, with sufficient data concerning the end-values of the function and its derivatives to make the solution unique. This is the situation we shall face again and again in this book, complicated by the presence of several unknowns in some cases, by the presence of several independent variables instead of the single x, and by the change from an ordinary differential equation to one or more partial differential equations.

Suppose that the problem is an easy one. By familiar manipulations (not bothering about a representation) we solve the equation and get a formula for the function $f(x)$. From it we can prepare a graph and a tabulation of values. These representations are quite satisfactory, revealing the true properties of the solution.

But suppose that the problem is not an easy one, and that we despair of finding a formula for the solution $f(x)$. Nor can we make a graph or tabulation of values. Nevertheless, we hope to find out *some* facts about $f(x)$. But while we are doing this, how are we to *think* about this unknown function? To think about anything, we must have a representation for it.

We try first to represent $f(x)$ by a formula. What formula? Since we do not know the function, we can merely write $f(x)$ and perform formal manipulations with this symbol. This is like manipulating an unknown number as the letter x, and for some purposes this is the simplest and best thing to do.

What about a graph? What sort of graph should we draw, the function $f(x)$ being unknown? Any graph we draw will have properties—positive slope here, negative slope there, and so on. These properties may be the properties of $f(x)$, but it is probable that they will not. In fact, a graph commits us too much, and a tabulation of values of an unknown function would be absurd, saying far too much about a function of which we are so ignorant.

Thus we have on the one hand a symbol $f(x)$ which says too little and on the other hand a graph or tabulation which says too much. Is there a middle way? For our purposes a function-space representation provides the middle way, and to it we shall now proceed.

The idea of function-space

Let $f(x)$ and $g(x)$ be two functions of x for the range $x_1 \leqslant x \leqslant x_2$. Suppose we know that these functions exist (perhaps as solutions of differential equations), but suppose we do not know what the functions are, i.e. we have no formulae, no graphs, no tabulations of values. We are dealing, in fact, with two *unknown* functions.

We take a sheet of paper and mark two points at random, labelling one point $f(x)$ and the other $g(x)$. This is merely a scheme for mental concentration; when we think of either function, we associate it in our minds with the corresponding point.

Now we bring in a third function, the *zero function* which vanishes for all values of x in the range. For it we mark a third point O on the paper. We have now a *representation* of three functions by three points on a plane (Fig. 1·14).

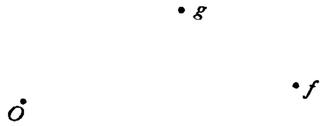

Fig. 1·14. Representation of functions by points.

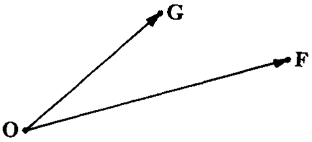

Fig. 1·15. Representation of functions by vectors.

Next we join the zero point (O) to the other points and put arrows on the joins as shown in Fig. 1·15. Regarding these directed joins as vectors, we change the notation, using heavy capital letters, with **O** for the zero vector. We have then the following correspondence between vectors and functions:

$$\mathbf{O} \leftrightarrow 0, \quad \mathbf{F} \leftrightarrow f(x), \quad \mathbf{G} \leftrightarrow g(x).$$

We shall employ the symbol $\leftrightarrow$ to indicate correspondences of this type.

We have used very little of our paper, and the rest of it is available for the representation of other functions. How should we proceed? Remember that we are not now in the domain of logical deductions —we are playing with the problem of representation. In this spirit we consider the vector $\mathbf{F}+\mathbf{G}$ (obtained by the usual parallelogram law) and ask: To what function shall we make $\mathbf{F}+\mathbf{G}$ correspond?

An obvious suggestion is that we should let it correspond to the function $f(x)+g(x)$, and further that we should let the vector $a\mathbf{F}$ (where a is any number, or scalar) correspond to the function $af(x)$. Thus we commit ourselves to the correspondence

$$a\mathbf{F}+b\mathbf{G} \leftrightarrow af(x)+bg(x), \tag{1·101}$$

a and b being any two real numbers, positive or negative.

As a and b take all real values, the extremity of the vector $a\mathbf{F}+b\mathbf{G}$ covers the whole plane, as indicated in Fig. 1·16. This means that we have, in the plane, representations of all functions of the form $af(x)+bg(x)$, where $f(x)$ and $g(x)$ are two definite (if unknown) functions and a and b arbitrary constants, with a one-to-one correspondence between vectors and functions.