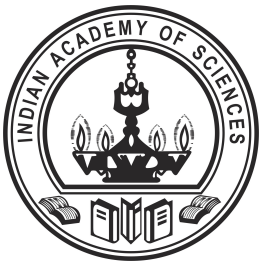
A Miscellany of Mathematical Physics

V. Balakrishnan



Reproduced from
Resonance – journal of science education

Reformatted by :
Sriranga Digital Software Technologies Private Limited,
Srirangapatna.

Printed at:
Tholasi Prints India Pvt Ltd

Published by
Indian Academy of Sciences

Foreword

The Masterclass series of eBooks brings together pedagogical articles on single broad topics taken from *Resonance*, the Journal of Science Education, that has been published monthly by the Indian Academy of Sciences since January 1996. Primarily directed at students and teachers at the undergraduate level, the journal has brought out a wide spectrum of articles in a range of scientific disciplines. Articles in the journal are written in a style that makes them accessible to readers from diverse backgrounds, and in addition, they provide a useful source of instruction that is not always available in textbooks.

The fourth book in the series, 'A Miscellany of Mathematical Physics', is by Prof. V. Balakrishnan. A distinguished theoretical physicist, Prof. Balakrishnan worked at TIFR (Mumbai) and RRC (Kalpakkam) before settling down at IIT Madras, from where he retired as an Emeritus Professor in 2013, after a stint lasting 33 years. Prof. Balakrishnan is widely renowned, in India and abroad, as a stimulating and inspiring teacher at all levels, undergraduate to doctoral. He has also contributed pedagogical articles regularly to *Resonance*, and a selection of these articles, substantially reworked in many cases, comprise the present book.

Prof. Balakrishnan is well known for his research contributions to the areas of stochastic processes, quantum dynamics, non-linear dynamics and chaos. The book, which will be available in digital format, and will be housed as always on the Academy website, will be valuable to both students and experts as a useful handbook on diverse topics in theoretical physics, ranging from the rotations of vectors and matrices to the many avatars of the Dirac delta function.

Amitabh Joshi
Editor of Publications
Indian Academy of Sciences
February 2018

About the Author

Prof V. Balakrishnan – universally abbreviated to Bala or Balki – needs no introduction to professional physicists of his generation – and indeed many later generations, especially those who are theoretically inclined. He also needs no introduction to almost any student who has passed through the Indian Institute of Technology, Madras in the last thirty years, being something of a legend there. Indeed, many would-be engineers were attracted to physics by his courses. Thanks to NPTEL and the internet, his lectures have now reached an even larger constituency, namely engineering and physics students all over the country and ouside. He has thus been an inspiring figure in the research and teaching community for decades.

The target audience of this Masterclass, however, includes young readers who will be encountering him for the first time. The bare facts of his career and achievements have already been mentioned in the preface, so a more personal note may be appropriate here.

I come from roughly one academic generation – five years – after Bala. Long before I heard him in person, research students of the 1969 batch at the Tata institute of Fundamental Research in Mumbai wrote to me in faraway Bangalore of how Bala had made a course on special functions come alive in the complex plane. I later read his research papers in *Pramana*, the journal of physics of the Indian Academy of Sciences. Unlike most such papers, these were written as if they were meant to be read aloud, combining originality with context, motivation, and lucid exposition. When I eventually met him face to face, our discussions initially centred on statistical physics, but as I soon realized, there was very little in any branch of theoretical physics that he had not thought about very deeply.

Bala has a first-hand perspective: the mathematics and physics in any discussion with him is served fresh. This is the hallmark of the best teachers, and of people who have thought and worked long and hard and reconstructed their subjects for themselves. These perspectives go well beyond physics: Bala's sharp insights into events around him, both academic and non-academic, have always been a delight to listeners – as much for what is said as for how it is expressed. All this might leave the reader with a picture of an armchair critical academic but it could not be more wrong. I was fortunate to see another side of Bala at close quarters, when serving under his chairmanship on the board of a charitable trust. His handling of difficult, people, situations and decisions was balanced, humane and principled and I am sure this would have characterized his functioning in numerous other roles as well.

This collection of articles written originally for *Resonance* at various points during the past twenty or so years (but now considerably revised and enhanced) has been woven into chapters. It takes a deceptively relaxed and informal approach to a formidable subject, and serves as a gateway to deeper study. Bala has written and lectured extensively on mathematical methods used by physicists in various forums, and for a beginning student there is no better place to start than this Masterclass.

Rajaram Nityananda
Azim Premji University, Bengaluru

Preface

This book is based on eight articles that were written for *Resonance*. Needless to say, the material has been rearranged, revised, and considerably expanded in some directions, so as to have a degree of cogency. The articles themselves grew out of the notion that the mathematical tools and techniques required by the students of physical sciences can, and should, be introduced to them in a more 'user-friendly' style than is generally the case. The initial introduction should be heuristic, with adequate motivation. The development of the subject matter should help the student not only to learn the techniques, but also to gain insight and the ability to recognise interconnections. Attention should be paid to the natural unfolding of the subject matter; one thing should lead to another. While correctness can never be sacrificed, formal rigour and exactitude need not be at the forefront.

In this book, we take a walking tour — or a casual stroll — through a variety of topics, such as vectors, reciprocal vectors, spherical and hyperspherical coordinates, Cartesian tensors, the true meaning of a vector, linear vector spaces, ket and bra vectors, infinitesimal rotation generators, the finite rotation formula, rotation matrices, the orthogonal group $SO(3)$ and its parameter space, Pauli matrices, the unitary group $SU(2)$ and its parameter space, the Dirac delta function, Green functions, the wave equation, its Fourier transform, and its fundamental solution in spaces of different dimensions. One thing leads to another!

V. Balakrishnan

Contents

What *can* the answer be?*

V. Balakrishnan

Scientific problems are very often first solved by a combination of analogy, educated guesswork and elimination – in short, 'insight'. The refinements that come later do not make this earlier process less important. Rather, they generally serve to highlight its value.

There is no graded set of lessons by which one progressively gains insight. However, a profitable line of approach is to ask, at each stage, what the answer to a problem *could possibly be*, subject to the conditions involved. Techniques such as dimensional analysis, scaling arguments and order-of-magnitude estimates, as well as checks based on limiting values or limiting cases are part of the armoury in this mode of attack. Elementary vector analysis offers a convenient platform to illustrate this approach.

An example from algebra

To set the stage, let us begin with an example in elementary algebra. Consider the determinant

$$\Delta(x_1, x_2, x_3) \quad = \quad \begin{vmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \\ x_1^2 & x_2^2 & x_3^2 \end{vmatrix}. \tag{1}$$

It is straightforward, of course, to find Δ explicitly by expanding the determinant. But the point I wish to make here is that Δ can be evaluated almost *by inspection*, if we note the following facts:

(i) Multiplying each of x_1, x_2 and x_3 by some number λ multiplies the value of Δ by λ^3. Thus Δ is a *homogeneous* function of degree 3.

(ii) Δ vanishes if any two of the x's are equal. Therefore, regarded as a function of x_1, Δ is quadratic with factors $(x_1 - x_2)$ and $(x_1 - x_3)$; and similarly for x_2 and x_3.

(iii) Δ changes sign if any two of the x's are interchanged.

Combining these points, we conclude that Δ *must* be given by

$$\Delta(x_1, x_2, x_3) = C(x_1 - x_2)(x_2 - x_3)(x_3 - x_1) \tag{2}$$

where C is some numerical constant.

*Based on *Resonance*, Vol. **1**, No.8, pp.8-15, 1996.

- In order to find the constant C, we have only to look at a simple special case, e.g., $x_1 = 0,\ x_2 = 1$. It is now trivial to simplify the determinant in eq. (1) directly. The result is

$$\Delta(0, 1, x_3) = x_3(x_3 - 1). \tag{3}$$

  On the other hand, eq. (2) gives

$$\Delta(0, 1, x_3) = C x_3(x_3 - 1). \tag{4}$$

It follows at once that C must be equal to 1. (Alternatively, match the term $+ x_2 x_3^2$ obtained by multiplying together all the diagonal elements of the determinant with the corresponding term $+ C\, x_2 x_3^2$ on the right in eq. (2).) Finally, therefore, we have

$$\Delta(x_1, x_2, x_3) = (x_2 - x_1)(x_3 - x_1)(x_3 - x_2). \tag{5}$$

The factors on the right-hand side of eq. (5) have been written in such a way that *selecting the first term in each bracket yields the product of the diagonal elements of the determinant with the correct sign*.

What is important is that our chain of reasoning permits us to *generalize* this result to the case of the $(n \times n)$ determinant (called the *Vandermonde determinant*)

$$\Delta(x_1, \ldots, x_n) = \begin{vmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{n-1} & x_2^{n-1} & \cdots & x_n^{n-1} \end{vmatrix}, \tag{6}$$

where n is any integer ≥ 2. We can now see that Δ must simply be a product of the $n(n-1)/2$ distinct factors $(x_j - x_k)$ that can be formed from the variables $x_1, \ldots, x_n$. The diagonal element $+ x_2\, x_3^2\, x_4^3 \cdots x_n^{n-1}$ indicates that the sign of each term in Δ is taken care of if we always maintain $j > k$ in each factor of $(x_j - x_k)$. We may therefore deduce the general result

$$\Delta(x_1, \ldots, x_n) = \prod_{1 \leq k \leq j \leq n} (x_j - x_k) \tag{7}$$

without going through a tedious calculation. This is the spirit in which we shall approach the problems that follow.

Some vector identities

Let us now go on to vector analysis. As the first example, we consider the derivation of the well-known identity for the triple cross-product, namely,

$$\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = (\mathbf{a} \cdot \mathbf{c})\mathbf{b} - (\mathbf{a} \cdot \mathbf{b})\mathbf{c} \tag{8}$$

where $\mathbf{a}, \mathbf{b}$ and $\mathbf{c}$ are three arbitrary vectors (in the familiar three-dimensional space). We would like to avoid the 'brute force' method of writing out components, etc., in some particular coordinate system. We therefore proceed as follows. Let $\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = \mathbf{d}$.

What can the answer be?

- If $\mathbf{a}, \mathbf{b}$ and $\mathbf{c}$ are three general non-planar vectors in three-dimensional space, any arbitrary vector can be uniquely written as a linear combination of these three vectors. (They serve to define a set of 'oblique' axes). But $\mathbf{d}$ cannot have any component along $\mathbf{a}$, because it is a *cross* product of *va* with another vector. Therefore, in general, $\mathbf{d}$ must necessarily be expressible as

$$\mathbf{d} = \beta\mathbf{b} + \gamma\mathbf{c} \tag{9}$$

  where β and γ are *scalars*. Note that this argument is valid even in the case of oblique axes, i.e., $\mathbf{b}$ and $\mathbf{c}$ are *not* required to be perpendicular to $\mathbf{a}$.

- $\mathbf{d}$ is of *first order* in each of the vectors $\mathbf{a}, \mathbf{b}$ and $\mathbf{c}$: that is, multiplying any one of them by a constant multiplies the answer by the same constant; further, $\mathbf{d}$ vanishes if any of these three vectors is zero. Therefore β *must* be proportional to $(\mathbf{a} \cdot \mathbf{c})$ and γ to $(\mathbf{a} \cdot \mathbf{b})$, respectively, as these are the only first-order scalars that can be formed from $(\mathbf{a}, \mathbf{c})$ and $(\mathbf{a}, \mathbf{b})$ respectively. Hence

$$\mathbf{d} = \lambda(\mathbf{a} \cdot \mathbf{c})\mathbf{b} + \mu(\mathbf{a} \cdot \mathbf{b})\mathbf{c} \tag{10}$$

  where λ and μ are absolute constants – dimensionless pure numbers – that are *independent* of $\mathbf{a}, \mathbf{b}$ and $\mathbf{c}$.

- But $\mathbf{d}$ changes sign if $\mathbf{b}$ and $\mathbf{c}$ are interchanged, because $\mathbf{c} \times \mathbf{b} = -\mathbf{b} \times \mathbf{c}$. Therefore

$$-\mathbf{d} = \lambda(\mathbf{a} \cdot \mathbf{b})\mathbf{c} + \mu(\mathbf{a} \cdot \mathbf{c})\mathbf{b}. \tag{11}$$

  Comparison with eq. (10) gives $\mu = -\lambda$, so that

$$\mathbf{d} = \lambda[(\mathbf{a} \cdot \mathbf{b})\mathbf{c} - (\mathbf{a} \cdot \mathbf{c})\mathbf{b}]. \tag{12}$$

- Having nearly solved the problem, we may *now* determine λ by looking at an appropriate simple special case because eq. (12) is valid for *all* $\mathbf{a}, \mathbf{b}, \mathbf{c}$. Thus, setting $\mathbf{a} = \mathbf{i}$, $\mathbf{b} = \mathbf{i}$, $\mathbf{c} = \mathbf{j}$ (for instance), we get $\mathbf{d} = -\mathbf{j}$ by direct evaluation of $\mathbf{a} \times (\mathbf{b} \times \mathbf{c})$, while the right-hand side of eq. (12) gives $\mathbf{d} = -\lambda\,\mathbf{j}$. Hence $\lambda = 1$. We thus obtain the general formula quoted in eq. (8).

The arguments used above can be repeated to tackle numerous other cases. Let us consider, for instance, the scalar product $(\mathbf{a} \times \mathbf{b}) \cdot (\mathbf{c} \times \mathbf{d})$, where $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}$ are four arbitrary vectors. We again make the following observations:

- The expression is of first order in each of the four vectors.

- The presence of the cross products $(\mathbf{a} \times \mathbf{b})$ and $(\mathbf{c} \times \mathbf{d})$ implies that there can be no terms proportional to $(\mathbf{a} \cdot \mathbf{b})$ and $(\mathbf{c} \cdot \mathbf{d})$ in the result. Hence the answer *must* be of the form

$$(\mathbf{a} \times \mathbf{b}) \cdot (\mathbf{c} \times \mathbf{d}) = \lambda(\mathbf{a} \cdot \mathbf{c})(\mathbf{b} \cdot \mathbf{d}) + \mu(\mathbf{a} \cdot \mathbf{d})(\mathbf{b} \cdot \mathbf{c}) \tag{13}$$

  where λ and μ are pure numbers.

- As before, since the answer changes sign if **a** and **b** are interchanged, we get $\lambda = -\mu$.
- Finally, the overall constant factor is fixed by looking at a special case, e.g., $\mathbf{a} = \mathbf{c} = \mathbf{i}$, $\mathbf{b} = \mathbf{d} = \mathbf{j}$. This gives $\lambda = 1$. We thus obtain the familiar formula

$$(\mathbf{a} \times \mathbf{b}) \cdot (\mathbf{c} \times \mathbf{d}) = (\mathbf{a} \cdot \mathbf{c})(\mathbf{b} \cdot \mathbf{d}) - (\mathbf{a} \cdot \mathbf{d})(\mathbf{b} \cdot \mathbf{c}). \tag{14}$$

The formulas (8) and (14) are, of course, well known, and several different proofs of their validity can be given. My aim has been to bring out the fact that *general considerations of linearity, symmetry (or antisymmetry), dimensionality, homogeneity, etc. practically determine the final answer in such problems*. This is brought home even more convincingly by the example that follows.

Evaluation of an integral

We will first evaluate the surface integral

$$I_4(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}) = \int d\Omega \, (\mathbf{e}_r \cdot \mathbf{a}) \, (\mathbf{e}_r \cdot \mathbf{b}) \, (\mathbf{e}_r \cdot \mathbf{c}) \, (\mathbf{e}_r \cdot \mathbf{d}), \tag{15}$$

where the unit vector $\mathbf{e}_r$ varies over the surface of a sphere of unit radius centred at the origin. Here $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}$ are four arbitrary constant vectors, which is why I have used the notation I_4. (Such integrals occur in several contexts in physical calculations – for example, in the theory of collisions of particles.) A brute force approach to the evaluation of $I_4(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d})$ is a formidable task, but there is a very 'physical' way of tackling the problem. We may try to simplify the task by choosing spherical polar coordinates with the polar axis along one of the given vectors, say **a**. But this does not help much, because there are *three* other vectors pointing in arbitrary directions. Instead, we note that $I_4(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d})$ (i) is a scalar, (ii) is of first order in *each* of the four vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}$ and vanishes if any one of them is zero, and (iii) is *totally symmetric* under the interchange of any of these vectors among themselves. Therefore $I_4(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d})$ must necessarily be of the form

$$I_4(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}) = \lambda \, [(\mathbf{a} \cdot \mathbf{b})(\mathbf{c} \cdot \mathbf{d}) + (\mathbf{a} \cdot \mathbf{c})(\mathbf{b} \cdot \mathbf{d}) + (\mathbf{a} \cdot \mathbf{d})(\mathbf{b} \cdot \mathbf{c})] \, , \tag{16}$$

where λ is a pure number. The plus signs and the common overall constant λ follow from (iii) above. [I have also used the fact that $(\mathbf{a} \cdot \mathbf{b})(\mathbf{c} \cdot \mathbf{d}) = (\mathbf{c} \cdot \mathbf{d})(\mathbf{a} \cdot \mathbf{b})$, as well as the property $(\mathbf{a} \cdot \mathbf{b}) = (\mathbf{b} \cdot \mathbf{a})$.] Likewise, combinations such as $(\mathbf{a} \times \mathbf{b}) \cdot (\mathbf{c} \times \mathbf{d})$ are not allowed by this symmetry. (iv) The constant λ is now determined by going over to the special case $\mathbf{a} = \mathbf{b} = \mathbf{c} = \mathbf{d} = \mathbf{k}$ (the unit vector along the polar or z-axis). In that case, since $\mathbf{e}_r \cdot \mathbf{k} = \cos\theta$, the integral reduces by direct evaluation to

$$I_4(\mathbf{k}, \mathbf{k}, \mathbf{k}, \mathbf{k}) = \int_{-1}^{1} d(\cos\theta) \int_0^{2\pi} d\varphi \, \cos^4\theta \; = \; \frac{4\pi}{5} \, , \tag{17}$$

on the one hand; on the other hand, eq. (16) gives $I_4(\mathbf{k},\mathbf{k},\mathbf{k},\mathbf{k}) = 3\lambda$. Hence $\lambda = 4\pi/15$, yielding the final answer

$$I_4(\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d}) = (4\pi/15)\ [(\mathbf{a}\cdot\mathbf{b})(\mathbf{c}\cdot\mathbf{d}) + (\mathbf{a}\cdot\mathbf{c})(\mathbf{b}\cdot\mathbf{d}) + (\mathbf{a}\cdot\mathbf{d})(\mathbf{b}\cdot\mathbf{c})]\,. \tag{18}$$

A generalization

A generalization of the result just derived is tempting and possible! We see at once that all the odd numbered integrals $I_1, I_3, I_5 \ldots$ must *vanish identically*, because there is no way that we can form a *scalar* from an *odd* number of vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}, \ldots$ that satisfies both (ii) and (iii) listed above. What about the corresponding general integral of *even* order,

$$I_{2n}(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{2n}) = \int d\Omega \prod_{j=1}^{2n} (\mathbf{e}_r \cdot \mathbf{a}_j), \tag{19}$$

involving the $2n$ arbitrary vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{2n}$? The arguments given earlier now lead us to conclude that the value of the integral must be given by

$$I_{2n}(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{2n}) = \lambda \sum_{\text{perm}}{}' (\mathbf{a}_{j_1} \cdot \mathbf{a}_{j_2})\,(\mathbf{a}_{j_3} \cdot \mathbf{a}_{j_4}) \cdots (\mathbf{a}_{j_{2n-1}} \cdot \mathbf{a}_{j_{2n}}), \tag{20}$$

where λ is a constant, yet to be determined, and $\{j_1, j_2, \ldots, j_{2n}\}$ is a *permutation* of $\{1, 2, \ldots, 2n\}$. (The prime on the summation sign will be explained shortly.) Now, we know that there are $(2n)!$ permutations of the set $\{1, 2, \ldots, 2n\}$. However, the number of terms in the summation in eq. (20) is smaller than $(2n)!$, because of the following restrictions:

(i) $(\mathbf{a}_j \cdot \mathbf{a}_k)$ and $(\mathbf{a}_k \cdot \mathbf{a}_j)$ are not to be counted as two distinct quantities.

(ii) The order of the n individual factors in a quantity of the form $(\mathbf{a}_{j_1}\cdot\mathbf{a}_{j_2}) \cdots (\mathbf{a}_{j_{2n-1}}\cdot\mathbf{a}_{j_{2n}})$ does not matter. Each such product must appear only once, although there are $n!$ permutations of the factors in each case.

These restrictions bring down the number of terms in the summation in eq. (20) from $(2n)!$ to $(2n)!/(2^n n!)$. The prime on the summation sign in eq. (20) is meant to indicate this restricted summation.

Next, we determine the constant λ by directly evaluating I_{2n} from its defining integral (19) in a simple special case. The most convenient choice is obviously $\mathbf{a}_1 = \mathbf{a}_2 \cdots = \mathbf{a}_{2n} = \mathbf{k}$. The integral may then be done by choosing the polar axis to lie along $\mathbf{k}$, so that $\mathbf{e}_r \cdot \mathbf{k} = \cos\theta$. Therefore

$$I_{2n}(\mathbf{k},\mathbf{k},\ldots,\mathbf{k}) = 2\pi \int_{-1}^{1} d(\cos\theta)\,(\cos\,\theta)^{2n} \;=\; \frac{4\pi}{2n+1}\,. \tag{21}$$

Equation (20), on the other hand, gives

$$I_{2n}(\mathbf{k},\mathbf{k},\ldots,\mathbf{k}) = \lambda \sum_{\text{perm}}{}' (\mathbf{k}\cdot\mathbf{k}) = \frac{(2n)!\,\lambda}{2^n\, n!}. \tag{22}$$

Comparing the two expressions, we get

$$\lambda = \frac{2^{n+2}\pi n!}{(2n+1)!}\,. \tag{23}$$

Thus, we arrive at the result

$$I_{2n}(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{2n}) = \frac{2^{n+2}\pi n!}{(2n+1)!} \sum_{\text{perm}}{}' (\mathbf{a}_{j_1} \cdot \mathbf{a}_{j_2})\,(\mathbf{a}_{j_3} \cdot \mathbf{a}_{j_4}) \cdots (\mathbf{a}_{j_{2n-1}} \cdot \mathbf{a}_{j_{2n}}). \tag{24}$$

You may check that the result derived earlier, eq. (18), is recovered correctly on setting $n = 2$.

A further generalization; hyperspherical coordinates

A *further* generalization of the result just derived that suggests itself (and which may indeed occur in actual calculations) is the following. What is the value of the integral

$$I_{n,d}(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n) = \int d\Omega\ (\mathbf{e}_r \cdot \mathbf{a}_1) \ldots (\mathbf{e}_r \cdot \mathbf{a}_n) \tag{25}$$

where $\mathbf{e}_r$ varies over the surface of a unit hypersphere in Euclidean space of $d > 3$ dimensions? The evaluation of this integral gives us an opportunity to introduce and use spherical polar coordinates in dimensions $d > 3$. These coordinates are called *hyperspherical coordinates*.

It is obvious that, regardless of the value of d, the integral continues to be (i) a scalar, (ii) of first order in each of the vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$, and (iii) a totally symmetric function of these n vectors. A moment's thought then shows that $I_{n,d}$ must vanish identically if n is an odd number, for the same reason as before. We are then left with the task of evaluating

$$I_{2n,d}(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{2n}) = \int d\Omega \prod_{j=1}^{2n} (\mathbf{e}_r \cdot \mathbf{a}_j). \tag{26}$$

As in the case $d = 3$, the answer must necessarily be of the form given by eq. (20), namely,

$$I_{2n,d}(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{2n}) = \lambda \sum_{\text{perm}}{}' (\mathbf{a}_{j_1} \cdot \mathbf{a}_{j_2})\,(\mathbf{a}_{j_3} \cdot \mathbf{a}_{j_4}) \cdots (\mathbf{a}_{j_{2n-1}} \cdot \mathbf{a}_{j_{2n}}), \tag{27}$$

where the summation over the permutations of $\{1, 2, \ldots, 2n\}$ is subject to the same conditions as before. Once again, it is convenient to determine the value of λ by (i) setting each a_j equal to the same unit vector, denoted by $\mathbf{k}$. The result is precisely eq. (22), as before. The argument used to compute the number of terms in $\sum'_{\text{perm}}$ is not affected by the dimensionality of the space concerned.

We must now evaluate the integral

$$I_{2n,d}(\mathbf{k}, \mathbf{k}, \ldots, \mathbf{k}) = \int d\Omega\,(\mathbf{e}_r \cdot \mathbf{k})^{2n} \tag{28}$$

over the surface of a hypersphere of unit radius in d-dimensions. For this purpose we must specify hyperspherical coordinates in d dimensions, generalizing the familiar spherical polar coordinates of 3-dimensional space. Let $x_1, x_2, \ldots, x_d$ be the Cartesian coordinates of any point with position vector $\mathbf{r}$ in the space. The hyperspherical coordinates of $\mathbf{r}$ are given by

$$\mathbf{r} = (r, \theta_1, \theta_2, \ldots \theta_{d-2}, \varphi), \tag{29}$$

where r is the radial distance of the point from the origin of coordinates. The ranges of the hyperspherical coordinates are

$$0 \le r < \infty, \;\; 0 \le \theta_k \le \pi \; (1 \le k \le d-2), \;\; 0 \le \varphi < 2\pi. \tag{30}$$

That is, there are $d-2$ 'polar' angles θ_k and a single 'azimuthal' angle φ. These coordinates are related to the Cartesian coordinates of $\mathbf{r}$ according to

$$\left.\begin{array}{lcl} x_1 & = & r\cos\theta_1 \\ x_2 & = & r\sin\theta_1\cos\theta_2 \\ x_3 & = & r\sin\theta_1\sin\theta_2\cos\theta_3 \\ \cdots & = & \cdots\cdots\cdots\cdots\cdots\cdots \\ x_{d-1} & = & r\sin\theta_1\sin\theta_2\ldots\sin\theta_{d-2}\cos\varphi \\ x_d & = & r\sin\theta_1\sin\theta_2\ldots\sin\theta_{d-2}\sin\varphi. \end{array}\right\} \tag{31}$$

The solid angle element $d\Omega$ can be shown to be

$$d\Omega = (\sin\theta_1)^{d-2}\,(\sin\theta_2)^{d-3}\,\cdots\,(\sin\theta_{d-2})\,d\theta_1\,d\theta_2\cdots d\theta_{d-2}\,d\varphi. \tag{32}$$

The next step is to choose $\mathbf{k}$ in the most convenient manner. Clearly, it is simplest to choose it to be the unit vector along the x_1-direction. (This is the analogue of choosing the polar or z-axis along $\mathbf{k}$ in the 3-dimensional case.) Then $(\mathbf{e}_r \cdot \mathbf{k})$ is just $\cos\theta_1$. Equation (28) gives

$$\begin{aligned} I_{2n,d}(\mathbf{k}, \mathbf{k}, \ldots, \mathbf{k}) &= 2\pi \int_0^\pi d\theta_1\,(\sin\theta_1)^{d-2}(\cos\theta_1)^{2n} \\ &\times \prod_{k=2}^{d-2} \int_0^\pi d\theta_k\,(\sin\theta_k)^{d-k-1}. \end{aligned} \tag{33}$$

The definite integrals involved are standard integrals whose values are tabulated. (I shall not digress into the derivation of those results here.) Using these known values and simplifying the expressions obtained in eq. (22) enables us to determine λ. Inserting the result in eq. (27), we arrive at the answer for λ. It is most compactly expressed in terms of the so-called *gamma function*, and is given by

$$\lambda = \frac{\pi^{d/2}}{2^{n-1}\Gamma\left(n + \frac{1}{2}d\right)}. \tag{34}$$

In simpler terms, this formula reduces to the following: for even values of d, we have

$$\lambda = \frac{\pi^{d/2}}{2^{n-1}\left(n + \frac{1}{2}d - 1\right)!}\,. \tag{35}$$

For odd values of d,

$$\lambda = \frac{\pi^{(d-1)/2} 2^{n+d}\left(n + \frac{1}{2}(d-1)\right)!}{(2n + d - 1)!}\,. \tag{36}$$

Insertion of these expressions in eq. (27) completes the solution.

Oblique axes, reciprocal basis, kets and bras*

V. Balakrishnan

In this chapter, we shall see how concepts such as reciprocal basis vectors, dual spaces and co-vectors can be motivated from simple considerations starting from well-known identities in elementary vector analysis.

Resolution of a vector along oblique axes

Let us begin with the resolution of an ordinary vector $\mathbf{v}$ in three-dimensional (Euclidean) space, according to

$$\mathbf{v} = \mathbf{i}\, v_x + \mathbf{j}\, v_y + \mathbf{k}\, v_z\,. \tag{1}$$

What are v_x, v_y and v_z in terms of $\mathbf{v}$? Clearly, $v_x = \mathbf{i}\cdot\mathbf{v}$, $v_y = \mathbf{j}\cdot\mathbf{v}$, and $v_z = \mathbf{k}\cdot\mathbf{v}$. Therefore, if we introduce the *projection operator* $P_x = \mathbf{ii}$ (note that there is no dot or cross in between the two vectors !), and 'operate' with it on the arbitrary vector $\mathbf{v}$ by taking the dot product, the result is precisely $\mathbf{ii}\cdot\mathbf{v} = \mathbf{i}(\mathbf{i}\cdot\mathbf{v}) = \mathbf{i}v_x$, the component or part of $\mathbf{v}$ that lies along the unit vector $\mathbf{i}$. Similarly, we have projection operators $P_y = \mathbf{jj}$ and $P_z = \mathbf{kk}$. The *unit operator* (the operator that leaves any vector $\mathbf{v}$ unchanged) is clearly just the sum of *all* the projection operators,** namely,

$$\mathbb{I} = P_x + P_y + P_z = \mathbf{ii} + \mathbf{jj} + \mathbf{kk}. \tag{2}$$

This is called the *resolution* of the identity operator. Thus eq. (1) expresses the fact that

$$\mathbf{v} = \mathbb{I}\,\mathbf{v} = \mathbf{i}(\mathbf{i}\cdot\mathbf{v}) + \mathbf{j}(\mathbf{j}\cdot\mathbf{v}) + \mathbf{k}(\mathbf{k}\cdot\mathbf{v}). \tag{3}$$

We now ask: what is the counterpart of eq. (3) in the case of *oblique* axes defined by three arbitrary, non-coplanar vectors $\mathbf{a}, \mathbf{b}$ and $\mathbf{c}$, instead of the rectangular axes defined by $\mathbf{i}, \mathbf{j}$ and $\mathbf{k}$?

Once again, we can arrive at the answer by figuring out what the answer can possibly be. Writing

$$\mathbf{v} = \alpha\,\mathbf{a} + \beta\,\mathbf{b} + \gamma\,\mathbf{c}, \tag{4}$$

we observe that the coefficient α cannot involve any overlap‡ of $\mathbf{v}$ with either $\mathbf{b}$ or $\mathbf{c}$; β cannot involve any overlap of $\mathbf{v}$ with either $\mathbf{c}$ or $\mathbf{a}$; and γ cannot involve any overlap of $\mathbf{v}$ with either $\mathbf{a}$ or $\mathbf{b}$. This assertion is more or less obvious when the vectors $\mathbf{a}, \mathbf{b}$ and $\mathbf{c}$ are mutually perpendicular. As we are dealing here with oblique axes, however, some elaboration is required. Suppose we

*Based on *Resonance*, Vol. **1**, No.10, pp. 6-13, 1996.

**You are familiar with the dot product $\mathbf{a}\cdot\mathbf{b}$ (also called the scalar product) of two vectors $\mathbf{a}$ and $\mathbf{b}$, as well as the cross product $\mathbf{a}\times\mathbf{b}$ (also called the vector product) of the two vectors. The object $\mathbf{ab}$ is called the *tensor product* of the two vectors. In the present context, it is an operator that acts on a vector to produce another vector.

‡'Overlap' here refers to the dot or scalar product (also called the inner product) of the vectors concerned.

keep α fixed at some value, and let β and γ vary. The varying part of $\mathbf{v}$ then is then restricted to a plane parallel to the plane formed by the vectors $\mathbf{b}$ and $\mathbf{c}$. All vectors in this plane have the same value of α, but their projections on the plane formed by $\mathbf{b}$ and $\mathbf{c}$ vary. Hence α cannot depend on that projection for any given $\mathbf{v}$. It can only depend on the projection of $\mathbf{v}$ onto a vector normal to the plane formed by $\mathbf{b}$ and $\mathbf{c}$, i.e, onto the vector $\mathbf{b}\times\mathbf{c}$. Therefore α *must* be proportional to $[(\mathbf{b}\times\mathbf{c})\cdot\mathbf{v}]$. Similar conclusions hold good for the coefficients β and γ. Hence

$$\mathbf{v} = \lambda\,\mathbf{a}\,[(\mathbf{b}\times\mathbf{c})\cdot\mathbf{v}] + \mu\,\mathbf{b}\,[(\mathbf{c}\times\mathbf{a})\cdot\mathbf{v}] + \nu\,\mathbf{c}\,[(\mathbf{a}\times\mathbf{b})\cdot\mathbf{v}], \tag{5}$$

where the scalar factors λ, μ and ν are yet to be determined. The equivalence of all directions in space ("the isotropy of space") implies that λ, μ and ν must be equal to each other. Setting $\mathbf{v} = \mathbf{a}, \mathbf{b}$ and $\mathbf{c}$ in turn, we find immediately that $\lambda = \mu = \nu = 1/[(\mathbf{a}\times\mathbf{b})\cdot\mathbf{c}]$.[§] Therefore

$$\mathbf{v} = \frac{\mathbf{a}\,[(\mathbf{b}\times\mathbf{c})\cdot\mathbf{v}] + \mathbf{b}\,[(\mathbf{c}\times\mathbf{a})\cdot\mathbf{v}] + \mathbf{c}\,[(\mathbf{a}\times\mathbf{b})\cdot\mathbf{v}]}{[(\mathbf{a}\times\mathbf{b})\cdot\mathbf{c}]}\,. \tag{6}$$

Before discussing the properties of this expansion, let us consider another, equally instructive, way to arrive at it. We begin with the well-known vector identity

$$\mathbf{u}\times(\mathbf{b}\times\mathbf{c}) = \mathbf{b}\,(\mathbf{u}\cdot\mathbf{c}) - \mathbf{c}\,(\mathbf{u}\cdot\mathbf{b}). \tag{7}$$

(Recall that a proof of eq. (7) based on general arguments was given in ch. 1.) Now suppose $\mathbf{u}$ itself is of the form $\mathbf{u} = \mathbf{v}\times\mathbf{a}$. Substitution in eq. (7) gives

$$(\mathbf{v}\times\mathbf{a})\times(\mathbf{b}\times\mathbf{c}) = \mathbf{b}[(\mathbf{v}\times\mathbf{a})\cdot\mathbf{c}] - \mathbf{c}[(\mathbf{v}\times\mathbf{a})\cdot\mathbf{b}]. \tag{8}$$

The vector representing the quadruple cross product on the left-hand side is thus a linear combination of the vectors $\mathbf{b}$ and $\mathbf{c}$. It therefore lies in the plane formed by these two vectors. However, we could as well have written $\mathbf{b}\times\mathbf{c} = \mathbf{d}$, in which case

$$\begin{aligned}(\mathbf{v}\times\mathbf{a})\times(\mathbf{b}\times\mathbf{c}) &= (\mathbf{v}\times\mathbf{a})\times\mathbf{d}\\ &= \mathbf{a}\,(\mathbf{v}\cdot\mathbf{d}) - \mathbf{v}\,(\mathbf{a}\cdot\mathbf{d})\\ &= \mathbf{a}\,[\,\mathbf{v}\cdot(\mathbf{b}\times\mathbf{c})] - \mathbf{v}\,[\mathbf{a}\cdot(\mathbf{b}\times\mathbf{c})].\end{aligned} \tag{9}$$

The *same* vector is therefore a linear combination of the two vectors $\mathbf{a}$ and $\mathbf{v}$, and thus lies in the plane formed by them. As the four vectors $\mathbf{v}, \mathbf{a}, \mathbf{b}$ and $\mathbf{c}$ may be chosen quite arbitrarily, this appears to be paradoxical. However, we must now recall that these are vectors in three-dimensional space, *in which no more than three vectors of a given set of vectors can be linearly independent*, i.e., non-coplanar. In other words, if the vectors $\mathbf{a}, \mathbf{b}$ and $\mathbf{c}$ are a linearly independent set, the fourth vector $\mathbf{v}$ *must* be expressible as a linear combination of these, precisely by equating the expressions found in eqs (8) and (9) and solving for $\mathbf{v}$. The result, after using once again the cyclic symmetry of the scalar triple product and some rearrangement, is precisely eq. (6). This is the counterpart of the resolution in eq. (3) of an arbitrary vector $\mathbf{v}$ in a basis of orthogonal axes.

[§]Here we have used the cyclic symmetry of the scalar triple product, namely, $(\mathbf{a}\times\mathbf{b})\cdot\mathbf{c} = (\mathbf{b}\times\mathbf{c})\cdot\mathbf{a} = (\mathbf{c}\times\mathbf{a})\cdot\mathbf{b}$.

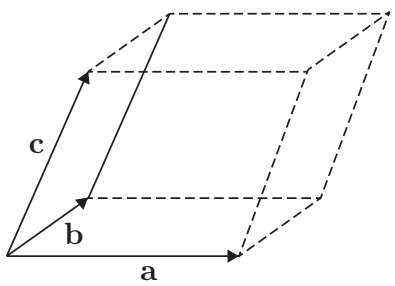


Figure 1: The volume of the parallelepiped formed by the vectors $(\mathbf{a}, \mathbf{b}, \mathbf{c})$ is the magnitude of their scalar triple product.

The reciprocal basis

The answer to the problem of resolving a vector $\mathbf{v}$ in an arbitrary basis $\{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$ is thus

$$\mathbf{v} = \mathbf{a}(\mathbf{A} \cdot \mathbf{v}) + \mathbf{b}(\mathbf{B} \cdot \mathbf{v}) + \mathbf{c}(\mathbf{C} \cdot \mathbf{v}), \tag{10}$$

where

$$\mathbf{A} = \frac{\mathbf{b} \times \mathbf{c}}{V}\,,\ \mathbf{B} = \frac{\mathbf{c} \times \mathbf{a}}{V}\,,\ \mathbf{C} = \frac{\mathbf{a} \times \mathbf{b}}{V}\,, \tag{11}$$

and $V = (\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c}$. The symbol V has been used because $|V|$ is the volume of the parallelepiped formed by the vectors $\mathbf{a}, \mathbf{b}$ and $\mathbf{c}$ (see Figure 1). The set of vectors $\{\mathbf{A}, \mathbf{B}, \mathbf{C}\}$ forms the so-called *reciprocal basis*. The terminology is most familiar in crystallography: if $\mathbf{a}, \mathbf{b}, \mathbf{c}$ are the primitive basis vectors of a lattice, then $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are the basis vectors of the 'reciprocal' lattice. It is immediately verified that

$$\mathbf{A} \cdot \mathbf{a} = \mathbf{B} \cdot \mathbf{b} = \mathbf{C} \cdot \mathbf{c} = 1, \tag{12}$$

which helps explain why the term 'reciprocal basis' is used. Further,

$$\mathbf{A} \cdot \mathbf{b} = \mathbf{A} \cdot \mathbf{c} = \mathbf{B} \cdot \mathbf{a} = \mathbf{B} \cdot \mathbf{c} = \mathbf{C} \cdot \mathbf{a} = \mathbf{C} \cdot \mathbf{b} = 0\,. \tag{13}$$

In fact, the reciprocal basis is *defined* in books on crystallography by eqs (12) and (13): they can be solved for $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$, to obtain precisely the expressions in eq. (11). It is easy to check that the general formula of eq. (10) reduces to eq. (3) in the special case of an orthogonal basis.

In what space does the reciprocal basis $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ 'live'? If the original basis vectors $\mathbf{a}, \mathbf{b}$ and $\mathbf{c}$ have the physical dimensions of length, eqs (11) show immediately that $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$ have the physical dimensions of $(\text{length})^{-1}$. In crystallography and lattice dynamics this fact is used to define a 'reciprocal lattice' in wavenumber space, in which $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ are the primitive lattice vectors. Why does one do this? It is not my intention to go into lattice dynamics or crystal physics here, but two good reasons (among several others) may be cited.

(i) In crystal physics, we have very frequently to deal with periodic functions, i.e., functions that satisfy $f(\mathbf{e}_r) = f(\mathbf{e}_r + \mathbf{R})$ where $\mathbf{R}$ is any lattice vector. That is,

$$\mathbf{R} = m\mathbf{a} + n\mathbf{b} + p\mathbf{c}, \tag{14}$$

where m, n and p take on integer values. Such a function can be expanded in a Fourier series of the form

$$f(\mathbf{e}_r) = \sum_{\mathbf{G}} f_{\mathbf{G}}\, e^{i\,\mathbf{G}\cdot\mathbf{e}_r}. \tag{15}$$

Here, the sum over $\mathbf{G}$ runs over the vectors of the reciprocal lattice, i.e.,

$$\mathbf{G} = h\mathbf{A} + k\mathbf{B} + l\mathbf{C}, \tag{16}$$

where $(h\,, k\,, l)$ are integers.
(ii) The second noteworthy point is that the Bragg condition for diffraction (of X-rays, electrons, neutrons, etc.) from a crystal is expressible in a very simple form in terms of $\mathbf{G}$, namely, $2\mathbf{k}\cdot\mathbf{G} = \mathbf{G}^2$, where $\mathbf{k}$ is the wave vector of the incident beam). Likewise, the Laue conditions for diffraction maxima reduce to just $\mathbf{G}\cdot\mathbf{a} = h$, $\mathbf{G}\cdot\mathbf{b} = k$, $\mathbf{G}\cdot\mathbf{c} = l$. These relations follow directly from eqs (12), (13), and (16).

Ket vectors and bra vectors

We are now at a point where the concepts of ket and bra vectors can be introduced naturally. Going back to eq. (1), we note the following. Any vector $\mathbf{v}$ in three dimensions can be represented in the form of a *column matrix* according to

$$\mathbf{v} = \begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} v_x + \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} v_y + \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} v_z. \tag{17}$$

(Here and in what follows, I freely use the '=' symbol between an abstract quantity and its *representation* in any form.) To save space, let us write $(1\ 0\ 0)^T$ for the (3×1) column matrix with elements $1, 0, 0$. (T stands for 'transpose'.) In this way of representing vectors, therefore,

$$\mathbf{i} = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}^T, \quad \mathbf{j} = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}^T, \quad \mathbf{k} = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix}^T. \tag{18}$$

We could also identify these with unit *ket vectors* denoted by $|e_1\rangle$, $|e_2\rangle$ and $|e_3\rangle$ respectively. Operating on a general vector $\mathbf{v} = (v_x\ v_y\ v_z)^T$, the projection operator $P_x = \mathbf{ii}$ introduced below eq. (1) must yield the component $\mathbf{i}\,v_x = (v_x\ 0\ 0)^T$. This is achieved if we identify P_x with the (3×3) *matrix* $(1\ 0\ 0)^T(1\ 0\ 0)$. In other words, the $\mathbf{i}$ on the *left* in $\mathbf{ii}$ really stands for the column vector $(1\ 0\ 0)^T$ or the *ket* vector $|e_1\rangle$, while the $\mathbf{i}$ on the *right* stands for the *row* vector $(1\ 0\ 0)$ — which it is now natural to identify with the *bra vector* $\langle e_1|$. The operator P_x is therefore

$|e_1\rangle\langle e_1|$. Similarly, $P_y = |e_2\rangle\langle e_2|$ and $P_z = |e_3\rangle\langle e_3|$. The resolution of the identity, eq. (2), now reads

$$|e_1\rangle\langle e_1| + |e_2\rangle\langle e_2| + |e_3\rangle\langle e_3| = \mathbb{I}. \tag{19}$$

The component v_x, which we saw was simply the scalar or dot product $\mathbf{i}\cdot\mathbf{v}$, is now written as the 'inner product' $\langle e_1|e_2\rangle$ where we have used the ket vector $|v\rangle$ to denote the vector $\mathbf{v} = (v_x\ v_y\ v_z)^T$. We can then go on to generalize this idea of ket vectors and their adjoint bra vectors to n-dimensional Euclidean spaces, and then to infinite-dimensional Hilbert spaces. The whole treatment provides an admittedly heuristic, but easily digested, method of introducing the machinery of linear vector spaces (e.g., for quantum mechanics) to students of physics whose background in this regard comprises little more than some familiarity with elementary matrix analysis — the situation most commonly encountered.

Let us now translate our findings for oblique axes to this language of ket and bra vectors. Writing $\mathbf{a}, \mathbf{b}$ and $\mathbf{c}$ as the ket vectors $|a\rangle, |b\rangle$ and $|c\rangle$ respectively, eq. (12) suggests at once that the reciprocal basis vectors $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$ are in fact to be identified with *bra* vectors $\langle A|, \langle B|$ and $\langle C|$, respectively. Equation (12) is the statement that the corresponding inner products are normalized to unity, i.e.,

$$\langle A|a\rangle = \langle B|b\rangle = \langle C|c\rangle = 1. \tag{20}$$

The expansion of an arbitrary vector $\mathbf{v}$ in eq. (10) reads, in this language,

$$|v\rangle = (\langle A|v\rangle)|a\rangle + (\langle B|v\rangle)|b\rangle + (\langle C|v\rangle)|c\rangle. \tag{21}$$

In other words, the resolution of the identity given by eq. (19) for orthogonal coordinates is now replaced by

$$|a\rangle\langle A| + |b\rangle\langle B| + |c\rangle\langle C| = \mathbb{I}. \tag{22}$$

The space spanned by the reciprocal basis vectors $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$ (more generally, by *bra* vectors) may be regarded as a kind of *dual* of the original space spanned by the vectors $\mathbf{a}, \mathbf{b}$ and $\mathbf{c}$.[¶] It turns out that we can prove that the dual space is actually *isomorphic* to the original space, provided the latter is finite-dimensional (in our case, it is three-dimensional). 'Isomorphic to' does not mean 'identical with', of course, but it does mean that the properties of the two space are essentially the same. This isomorphism between a linear vector space and its dual space *may* sometimes be valid even for infinite-dimensional spaces. A common but nontrivial example in physics occurs in elementary quantum mechanics: the position-space wave function of a particle moving in one spatial dimension is a member of the linear vector space of square-integrable functions of a real variable $x \in \mathbb{R}$. Its Fourier transform has a physical interpretation as the corresponding wave function in momentum space. This is also square integrable, and is a member of an *isomorphic* linear vector space of square-integrable functions of a real variable $p \in \mathbb{R}$.

[¶]This statement is a bit loose and glosses over certain technical details, but is quite acceptable at the present level of rigour.

We have seen how 'reciprocal' vectors (in a 'dual' vector space) arise naturally if we work with an oblique set of axes. The distinction between the original space and its dual space exists in any case, but it may be blurred in the case of an orthogonal basis set like $\{\mathbf{i}, \mathbf{j}, \mathbf{k}\}$ in a real vector space because the reciprocal basis appears to coincide with the original basis. When faced with a non-orthogonal basis set, the usual practice in quantum mechanics is to construct an orthogonal basis by, say, the Gram-Schmidt procedure. In crystallography, however, the structure of the lattice may force us to stick to the non-orthogonal basis as the natural and more useful one, supplemented, as we have seen, by the reciprocal basis. It must be remembered that we have been working in three-dimensional Euclidean space for the greater part. What if the number of dimensions we have to deal with is not equal to three ? (For one thing, the 'cross product' of two vector is a vector only in three dimensions !) What if the space itself is curved? Do vectors and reciprocal or bra vectors (living in the dual vector space) have anything to do with the distinction between *contravariant* and *covariant* vectors, (or 'upstairs' and 'downstairs' indices), *tangent* and *cotangent* spaces, and maybe even the Lagrangian and Hamiltonian formalisms in classical mechanics? The answer is "yes", implying that some profound aspects of the physical world are lurking behind the simple geometrical questions we have been discussing. We shall touch upon these matters in the sequel.

Reciprocal basis in two dimensions and other nice things*

V. Balakrishnan

In the preceding chapter, we saw how dual vectors arose very naturally even in elementary vector analysis. At the end of that chapter, I mentioned that dual vectors and the reciprocal basis were very far-reaching concepts. They appear in many different contexts, some of which will be described below. We begin with a situation that might appear to be simpler than the three-dimensional case already worked out – but we are in for a surprise!

Reciprocal basis in two dimensions (2D)

Let us recall briefly the essential result derived in ch. 2: Given any three non-coplanar, i.e., *linearly independent* vectors $\{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$ in the familiar three-dimensional (3D) Euclidean space, the reciprocal basis comprises three vectors $\{\mathbf{A}, \mathbf{B}, \mathbf{C}\}$ such that

$$\mathbf{A} \cdot \mathbf{a} = \mathbf{B} \cdot \mathbf{b} = \mathbf{C} \cdot \mathbf{c} = 1 \tag{1}$$

and, further,

$$\mathbf{A} \cdot \mathbf{b} \;=\; \mathbf{A} \cdot \mathbf{c} \;=\; \mathbf{B} \cdot \mathbf{a} \;=\; \mathbf{B} \cdot \mathbf{c} \;=\; \mathbf{C} \cdot \mathbf{a} \;=\; \mathbf{C} \cdot \mathbf{b} \;=\; 0. \tag{2}$$

The vectors $\{\mathbf{A}, \mathbf{B}, \mathbf{C}\}$ satisfying these conditions are found to be

$$\mathbf{A} = \frac{\mathbf{b} \times \mathbf{c}}{V}\,,\; \mathbf{B} = \frac{\mathbf{c} \times \mathbf{a}}{V}\,,\; \mathbf{C} = \frac{\mathbf{a} \times \mathbf{b}}{V}\,, \tag{3}$$

where $V = (\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c}$. Its magnitude is the volume of the parallelepiped formed by $\mathbf{a}, \mathbf{b}$ and $\mathbf{c}$. The expressions in (3) have a pleasing cyclical symmetry.

We now ask: what about the simpler case of *two* dimensions, i.e., a plane? We now have *two* vectors $\mathbf{a}$ and $\mathbf{b}$ that are not parallel or antiparallel to each other. We want to find two other vectors $\mathbf{A}$ and $\mathbf{B}$ in the same plane such that

$$\mathbf{A} \cdot \mathbf{a} = \mathbf{B} \cdot \mathbf{b} = 1, \quad \text{while} \quad \mathbf{A} \cdot \mathbf{b} = \mathbf{B} \cdot \mathbf{a} = 0. \tag{4}$$

This is easily done if we regard $\mathbf{a}$ and $\mathbf{b}$ as defining the directions of a pair of *oblique* axes in the plane. Then $\mathbf{A}$ and $\mathbf{B}$ must necessarily be linear combinations of the form

$$\mathbf{A} = \alpha_1\, \mathbf{a} + \beta_1\, \mathbf{b}, \quad \mathbf{B} = \alpha_2\, \mathbf{a} + \beta_2\, \mathbf{b}. \tag{5}$$

*Based on *Resonance*, Vol.**2**, No.5, pp.8-14, 1997.

The four constants α_1, β_1, α_2 and β_2 can now be found by taking the dot products of **A** and **B** with **a** and **b** in turn, and imposing the conditions in (4). This involves solving *four* simultaneous equations, which is a bit tedious, although quite straightforward. Is there a simpler, shorter way? One might be tempted to use the following argument:
"Since $\mathbf{A}\cdot\mathbf{b}=0$, **A** must be perpendicular to **b**. Similarly, $\mathbf{B}\cdot\mathbf{a}=0$, so that $\mathbf{B}\perp\mathbf{a}$. Hence **A** cannot have a part proportional to **b**, i.e., $\beta_1=0$. Likewise, $\alpha_2=0$. This leaves only the two constants α_1 and β_2 to be determined."
But this argument is incorrect! It is only valid if **a** and **b** are perpendicular to each other. If they are not, then the condition $\mathbf{A}\cdot\mathbf{b}=0$ does *not* imply that **A** is directed along the other axis, i.e., along **a**. Since **a** itself has a non-zero projection along **b**, the vector **A** cannot be directed exclusively along **a**. It must also have a compensating piece proportional to **b**, so that its *net* perpendicular projection on **b** is zero.

There *is*, however, a way to find **A** and **B** by solving just two equations, rather than four. Any arbitrary vector **v** in the plane can be expanded in the form

$$\mathbf{v}=c_1\,\mathbf{a}+c_2\,\mathbf{b}. \tag{6}$$

Now recall from ch. 2 that the tensor products $\mathbf{a}\,\mathbf{A}$ and $\mathbf{b}\,\mathbf{B}$ also serve as *projection operators* that add up to the unit operator, i.e., $\mathbf{a}\,\mathbf{A}\ +\mathbf{b}\,\mathbf{B}=\mathbf{I}$. Hence

$$\mathbf{v}\equiv\mathbf{I}\cdot\mathbf{v}=\mathbf{a}\,(\mathbf{A}\cdot\mathbf{v})+\mathbf{b}\,(\mathbf{B}\cdot\mathbf{v}). \tag{7}$$

In other words, $c_1=\mathbf{A}\cdot\mathbf{v}$ and $c_2=\mathbf{B}\cdot\mathbf{v}$. Take the dot product of both sides of eq. (6) with **a** and **b** in succession, to get two simultaneous equations for c_1 and c_2. Solve them to obtain c_1 and c_2. You can then simply read off **A** and **B** from the expressions for c_1 and c_2, using the fact that $c_1=\mathbf{A}\cdot\mathbf{v}$ and $c_2=\mathbf{B}\cdot\mathbf{v}$. The final result is

$$\mathbf{A}=\frac{b^2\mathbf{a}-(\mathbf{a}\cdot\mathbf{b})\mathbf{b}}{a^2b^2-(\mathbf{a}\cdot\mathbf{b})^2}\quad\text{and}\quad\mathbf{B}=\frac{a^2\mathbf{b}-(\mathbf{a}\cdot\mathbf{b})\mathbf{a}}{a^2b^2-(\mathbf{a}\cdot\mathbf{b})^2} \tag{8}$$

where $a^2\equiv\mathbf{a}\cdot\mathbf{a}$ and $b^2\equiv\mathbf{b}\cdot\mathbf{b}$, as usual.

Although these expressions are not too complicated, they are not too simple, either. Nor do they have the elegant cyclically symmetrical form of the expressions in the 3D case, eqs (3). This is quite surprising, because we should expect the answer in 2D to be actually simpler than that in 3D. In particular, the denominator $a^2b^2-(\mathbf{a}\cdot\mathbf{b})^2$ in (8) is of *second* order in **a** and **b**, while the denominator in V in (3) is of *first* order in $\mathbf{a},\mathbf{b}$ and $\mathbf{c}$. In 3D, $|V|$ is the volume of the parallelepiped formed by the set $\{\mathbf{a},\mathbf{b},\mathbf{c}\}$. Its analogue in 2D is the *area* $|\mathbf{a}\times\mathbf{b}|$ of the parallelogram formed by the set $\{\mathbf{a},\mathbf{b}\}$. We might therefore expect this area to appear in the denominator in the formulas for **A** and **B**. The problem, however, is that the cross product of two vectors in 2D space, i.e., of two vectors living strictly in a plane, is not a vector! More precisely: if (a_1,a_2) and (b_1,b_2) are the Cartesian components of the 2D vectors **a** and **b**, their 'cross product' $\mathbf{a}\times\mathbf{b}$ has only *one* component, $a_1\,b_2-a_2\,b_1$, instead of the two needed to make a 2D vector. This is the root of the difficulty.

But an interesting observation now comes to our aid. The *square* of $a_1b_2 - a_2b_1$ is just $a^2\,b^2 - (\mathbf{a}\cdot\mathbf{b})^2$, remembering that $a^2 = a_1^2 + a_2^2$ and $b^2 = b_1^2 + b_2^2$. And if $\mathbf{A}$ and $\mathbf{B}$ are written out component-wise, a factor $a_1b_2 - a_2b_1$ cancels out in each case, and we get

$$\begin{aligned} A_1 &= \frac{b_2}{a_1b_2 - a_2b_1}\,, \quad A_2 = \frac{-b_1}{a_1b_2 - a_2b_1}\,, \\ B_1 &= \frac{-a_2}{a_1b_2 - a_2b_1}\,, \quad B_2 = \frac{a_1}{a_1b_2 - a_2b_1}\,. \end{aligned} \tag{9}$$

These expressions do show (at last!) a sort of cyclic symmetry. Let us compare them with what happens in 3D, eqs (3). In that case we have

$$A_1 = \frac{(b_2c_3 - b_3c_2)}{V}\,, \quad A_2 = \frac{(b_3c_1 - b_1c_3)}{V}\,, \quad A_3 = \frac{(b_1c_2 - b_2c_1)}{V}\,, \tag{10}$$

where

$$\begin{aligned} V &= \mathbf{a}\cdot(\mathbf{b}\times\mathbf{c}) \\ &= a_1(b_2c_3 - b_3c_2) + a_2(b_3c_1 - b_1c_3) + a_3(b_1c_2 - b_2c_1), \end{aligned} \tag{11}$$

and similar expressions for the components of $\mathbf{B}$ and $\mathbf{C}$. What is the common feature of the denominators in eqs (9) and (10)-(11)? A moment's thought shows that, in each case, what we have is just the *determinant* formed by writing out the basis vectors in component form, one after the other, i.e.,

$$\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix} \text{ in 2D, and } \begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix} \text{ in 3D.} \tag{12}$$

This is the vital clue – the hidden pattern is now sufficiently revealed for us to guess the complete answer in an *arbitrary* number of dimensions! To do that, it is necessary to introduce, first, an important symbol and a convenient bit of notation.

The Levi-Civita symbol and Einstein's summation convention

We shall use the subscripts i, j, k, ... to denote the various components of a vector. Thus, a_i stands for the i^{th} component of the vector $\mathbf{a}$. Here the subscript or *index* i can take on the values $1, 2, \ldots, d$ in d dimensions.

Now consider the set $\{\epsilon_{ij}\}$ of $2^2 = 4$ quantities in 2D, defined as follows: $\epsilon_{12} = +1$, $\epsilon_{21} = -1$, and $\epsilon_{11} = \epsilon_{22} = 0$. Its counterpart in 3D is ϵ_{ijk}, defined as follows :

$$\epsilon_{ijk} = \begin{cases} +1, & \text{if } ijk \text{ is an } \textit{even}\text{permutation of 123} \\ -1, & \text{if } ijk \text{ is an } \textit{odd}\text{permutation of 123} \\ 0, & \text{in all other cases.} \end{cases} \tag{13}$$

Here, a permutation of the natural order 123 is said to be even [respectively, odd] if it is made up of an even [resp., odd] number of interchanges or *transpositions* of two indices at a time. Thus, of the $3^3 = 27$ quantities ϵ_{ijk},

$$\begin{array}{rcrcrcr} \epsilon_{123} & = & \epsilon_{312} & = & \epsilon_{231} & = & +1, \\ \epsilon_{213} & = & \epsilon_{321} & = & \epsilon_{132} & = & -1, \end{array} \tag{14}$$

and the remaining 21 components are equal to 0. It is evident that $\epsilon_{ijk} = 0$ is zero whenever at least two of the indices take on the same value, such as ϵ_{112} or ϵ_{333}. The generalization to d dimensions is immediate! In d dimensions, each of the indices $i, j, k, l, \ldots$ can take on values from 1 to d. Then

$$\epsilon_{ijkl\ldots} = \begin{cases} +1, & \text{if } ijkl\cdots \text{ is an even permutation of } 1234\cdots d \\ -1, & \text{if } ijkl\cdots \text{ is an odd permutation of } 1234\cdots d \\ 0, & \text{whenever any two indices are equal.} \end{cases} \tag{15}$$

$\epsilon_{ijkl\ldots}$ is called the Levi-Civita (or totally antisymmetric) symbol in d dimensions. Its great utility will become clear shortly. But first, a remark:

***Remark*:** In the special case of 3D, and *only* in this case, is the definition of Levi-Civita symbol given in (13) entirely equivalent to the following definition, found in many textbooks:

$$\epsilon_{ijk} = \begin{cases} +1, & \text{if } ijk \text{ are in cyclic order of } 123 \\ -1, & \text{if } ijk \text{ are in anticyclic order of } 123 \\ 0, & \text{in all other cases.} \end{cases} \tag{16}$$

While (16) is correct, it can be misleading, because it cannot be extended as it stands to any other dimension, including 2D ($\epsilon_{12} = +1$, but $\epsilon_{21} = -1$ although 21 is a cyclic permutation of 12). The correct general definition for any $d \geq 2$ is that given in (15).

Tullio Levi-Civita (1873–1941), mathematician

Abraham Pais, in his superb biography of Einstein (*Subtle is the Lord*), from which the quotations here are taken, speaks of "a noble line of descendence" in the works of Gauss, Riemann, Christoffel, Ricci and Levi-Civita, one of whose culmination points was Einstein's General Theory of Relativity (GTR). In 1917, Levi-Civita introduced in a mathematically rigorous manner the concept of *parallel transport*, a fundamental notion in tensor calculus and differential geometry. His correspondence with Einstein early in 1915 helped Einstein in his final formulation of GTR later that year – he was "happy to have finally found a professional who took a keen interest in his work", and in a grateful letter to Levi-Civita, said, "... It is therefore doubly gladdening to get to know better a man like you."

Among other uses, the Levi-Civita symbol helps us write down the volume of the parallelepiped formed by the basis vectors $\mathbf{a}, \mathbf{b}, \ldots$ in any number of dimensions, i.e., the value of the determinant formed by the components of the vectors. We see at once that in 2D, this is

$$(a_1 b_2 - a_2 b_1) = \sum_{i=1}^{2} \sum_{j=1}^{2} \epsilon_{ij}\, a_i\, b_j\,. \tag{17}$$

Similarly, in 3D,

$$V = \mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = \sum_{i=1}^{3} \sum_{j=1}^{3} \sum_{k=1}^{3} \epsilon_{ijk}\, a_i\, b_j\, c_k\,. \tag{18}$$

The notation is simplified considerably if we adopt a *convention*: if an index is *repeated* (i.e., if it appears twice in any expression), *it is automatically summed over all the values it can take*. This summation convention was introduced by Einstein himself in 1916. Besides significantly reducing the 'clutter' in mathematical expressions, it has a great advantage. It gives us a way of making an important consistency check on calculations involving tensors:

(i) Every index symbol that appears *once* on the left-hand side of any equation must do so on the right-hand side as well. (This is called a *free* index.)

(ii) Any index symbol that appears *twice* in an expression is a *dummy* index, and is to be summed over all its possible values.

(iii) *No* index symbol can appear in any expression *more* than twice. If it does so, there's a mistake somewhere! (In particular, if an expression requires the use of two or more dummy indices, you must use distinct symbols for the purpose.)

Taking a cue from the 2D and 3D cases, we have the following result: In d dimensions, the volume of the hyperparallelepiped formed by the basis vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}, \ldots$ is simply

$$V_d = \epsilon_{ijkl\ldots}\, a_i\, b_j\, c_k\, d_l \cdots\,, \tag{19}$$

where each subscript must be summed over from 1 to d. Note that this formula is also applicable in the cases $d = 2$ (i.e., in 2D) and $d = 3$ (i.e., in 3D), as you have already seen in eqs (17) and (18).

Finally, here is another important special property of three dimensions: It is only in 3D that the cross product of two vectors is itself a vector! (I will qualify this remark in Ch. 4, in the interests of technical accuracy.) It is easy to check that the k^{th} component of the vector formed by the cross product of two vectors $\mathbf{a}$ and $\mathbf{b}$ in 3D is just $\epsilon_{ijk}\, a_i\, b_j$. This quantity has precisely one free index (namely, k), as required for a vector. On the other hand, in 2D the cross product $\epsilon_{ij}\, a_i\, b_j$, which has no free index left at all, and is thus a scalar rather than a vector. In $d > 3$, the cross product is $\epsilon_{ijkl\ldots}\, a_i\, b_j$. But this quantity has two or more free indices $(k, l, \ldots)$, and hence denotes a *tensor* of rank 2 or more — to be precise, of rank $d - 2$. Since any two

(non-collinear) vectors define a plane, a geometrical way of saying all this is as follows : in d-dimensional space, we have d independent mutually orthogonal directions and ${}^dC_2 = d(d-1)/2$ independent orthogonal planes. Only in 3D are these two numbers equal to each other! This is one of the main reasons why 3D is so special.

We have now set up all the machinery needed to find the reciprocal basis in an arbitrary number of dimensions. This will be our first task in the next chapter.

The importance of good notation

I end this chapter with a few lines on an apparently trivial, but in reality quite important, matter.

Mathematical *notation* is generally regarded as a rather irrelevent or unimportant. It is often so – and yet, proper notation is so essential for clear understanding! Moreover, there are many striking instances when adopting a good notation has helped vitally in the development of the subject. Newton, when he invented the differential calculus, which he originally called *fluxions*, used $\dot{y}, \ddot{y}, \ldots$ to denote successive derivatives. It is easy to see that this notation rapidly leads to problems with higher-order derivatives, partial derivatives, and so on. In contrast, to quote E.T. Bell in *Men of Mathematics*, "$\cdots$ the more progressive Swiss and French, following the lead of Leibniz, and developing his incomparably better way of merely *writing* the calculus, perfected the subject", and thus made it "$\cdots$ a simple, easily applied implement of research $\cdots$".

Two other instances readily come to mind in which a happy choice of notation even acts as an automatic check against mistakes: Dirac's bra and ket notation for linear vector spaces, which was introduced in ch. 2, and the Einstein summation convention in tensor analysis. (If an index symbol appears twice in an expression, it is to be summed over all its allowed values. If it appears more than twice, there's a mistake somewhere!) Einstein himself appears to have been pleased with his innovation, for he jested to a friend that he had "made a great discovery in mathematics; we have suppressed the summation sign every time that the summation must be made over an index that appears twice $\cdots$".

At an even more fundamental level, we have instances where notation is crucial to the very existence of a subject — consider, for example, the place-value system of writing numbers (including fractions and irrational numbers), introduced in ancient India. Going further on the same trail, consider the idea of *exponents*, that lets us handle numbers of arbitrarily large or small magnitudes in a very compact and efficient manner. Can one even imagine where mathematics would be, if it were not for these examples of ingenious 'good notation'?

Reciprocal basis in *d* dimensions and other ramifications*

V. Balakrishnan

Reciprocal basis in *d* dimensions

In the preceding chapters, we have posed and solved the problem of finding the reciprocal basis corresponding to any given basis set of vectors in 2-dimensional (2D) and 3-dimensional (3D) space. We are ready, now, to tackle the problem of finding a reciprocal basis in Euclidean space of d dimensions (dD). Consider a set of d vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}, \ldots$ that form a *basis* in dD. This means that the set satisfies two *distinct* properties (i.e., neither of them implies the other):

(i) They are *linearly independent*. That is, none of them can be written as a linear combination of the rest of the set.

(ii) They *span* the space. That is, any arbitrary vector in the space can be written as a linear combination of the vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}, \ldots$.

Note that we do not require the vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}, \ldots$ to be mutually perpendicular, nor do we require each of the vectors to be a unit vector.** In other words, we have a general set of oblique axes in the space. The task is to find the corresponding reciprocal basis, i.e., the set of d vectors $\mathbf{A}, \mathbf{B}, \mathbf{C}, \ldots$ such that

$$\mathbf{A} \cdot \mathbf{a} = \mathbf{B} \cdot \mathbf{b} = \mathbf{C} \cdot \mathbf{c} = \cdots = 1, \tag{1}$$

and all other scalar products of a vector from the reciprocal basis and a vector from the original basis vanish. That is,

$$\mathbf{A} \cdot \mathbf{b} = \mathbf{A} \cdot \mathbf{c} = \cdots = 0, \quad \mathbf{B} \cdot \mathbf{a} = \mathbf{B} \cdot \mathbf{c} = \cdots = 0, \text{ etc.} \tag{2}$$

Recall the corresponding solutions in 2D and 3D, respectively. These have been derived in chapters 2 and 3. The Levi-Cita symbol and the summation convention help us express the solutions in very compact form. Repeating them for ready reference, the components of the vectors of the reciprocal basis are given by

$$A_i = \frac{\epsilon_{ij}\, b_j}{\epsilon_{pq}\, a_p\, b_q}, \quad B_j = \frac{\epsilon_{ij}\, a_i}{\epsilon_{pq}\, a_p\, b_q} \quad \text{in 2D}, \tag{3}$$

*Based on *Resonance*, Vol.**2**, No.7, pp.20-26, 1997.

**If those two conditions are also satisfied, we have an *orthonormal* basis, as you know.

and by

$$A_i = \frac{\epsilon_{ijk}\, b_j\, c_k}{\epsilon_{pqr}\, a_p\, b_q\, c_r}\,, \quad B_j = \frac{\epsilon_{ijk}\, a_i\, c_k}{\epsilon_{pqr}\, a_p\, b_q\, c_r}\,, \quad C_k = \frac{\epsilon_{ijk}\, a_i\, b_j}{\epsilon_{pqr}\, a_p\, b_q\, c_r} \quad \text{in 3D.} \tag{4}$$

You can easily verify that these expressions satisfy the requirements in eqs (1) and (2). It is now straightforward to see that that the answer in any dimension d is simply

$$A_i = \frac{\epsilon_{ijkl\ldots}\, b_j\, c_k\, d_l \cdots}{\epsilon_{pqrs\ldots}\, a_p\, b_q\, c_r\, d_s \cdots}\,, \; B_j = \frac{\epsilon_{ijkl\ldots}\, a_i\, c_k\, d_l \cdots}{\epsilon_{pqrs\ldots}\, a_p\, b_q\, c_r\, d_s \cdots}\,, \; C_k = \frac{\epsilon_{ijkl\ldots}\, a_i\, b_j\, d_l \cdots}{\epsilon_{pqrs\ldots}\, a_p\, b_q\, c_r\, d_s \cdots}\,, \; \cdots \tag{5}$$

and so on down the line, for each of the d vectors $\mathbf{A}, \mathbf{B}, \mathbf{C}, \ldots$. Once again, it is easy to check that the conditions (1) and (2) are met by these solutions. Recall, also, that the denominator in each of the expressions above is just V_d, whose magnitude is the volume of the (hyper)parallelepiped formed by the basis vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}, \ldots$.

I have already mentioned in ch. 2 that the notion of a reciprocal basis is a fundamental one in crystallography. You are now ready to begin the study of crystallography in d dimensions!

Crystal classes in d dimensions

Crystallography involves a fascinating branch of mathematics, the *theory of discrete groups*. The total number of *crystallographic space groups* (which is related to the kinds of crystalline symmetry possible) is 17 in 2D and 230 in 3D. The corresponding quantity in $d > 3$ dimensions is of interest in group theory and in certain applications of mathematical physics. It is 4894 in 4D, and increases very rapidly as d increases.

The Gram determinant

We noted in ch. 3 that, in 2D, $(a_1 b_2 - a_2 b_1)^2 = a^2 b^2 - (\mathbf{a}\cdot\mathbf{b})^2$. But this is the same as saying that

$$\begin{vmatrix} a_1 & a_2 \\ b_1 & b_2 \end{vmatrix}^2 = \begin{vmatrix} a_1 & a_2 \\ b_1 & b_2 \end{vmatrix} \begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix} = \begin{vmatrix} \mathbf{a}\cdot\mathbf{a} & \mathbf{a}\cdot\mathbf{b} \\ \mathbf{b}\cdot\mathbf{a} & \mathbf{b}\cdot\mathbf{b} \end{vmatrix}. \tag{6}$$

Note how the square on the left-hand side has been written as the product of a determinant and its transpose. It is immediately clear that a similar relationship is valid in $d \geq 3$ dimensions as well.

$$\begin{vmatrix} a_1 & a_2 & a_3 & \cdots & a_d \\ b_1 & b_2 & b_3 & \cdots & b_d \\ c_1 & c_2 & c_3 & \cdots & c_d \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \cdot & \cdot & \cdot & \cdots & \cdot \end{vmatrix} \begin{vmatrix} a_1 & b_1 & c_1 & \cdots & \cdot \\ a_2 & b_2 & c_2 & \cdots & \cdot \\ a_3 & b_3 & c_3 & \cdots & \cdot \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_d & b_d & c_d & \cdots & \cdot \end{vmatrix} = \begin{vmatrix} \mathbf{a}\cdot\mathbf{a} & \mathbf{a}\cdot\mathbf{b} & \mathbf{a}\cdot\mathbf{c} & \cdots & \cdot \\ \mathbf{b}\cdot\mathbf{a} & \mathbf{b}\cdot\mathbf{b} & \mathbf{b}\cdot\mathbf{c} & \cdots & \cdot \\ \mathbf{c}\cdot\mathbf{a} & \mathbf{c}\cdot\mathbf{b} & \mathbf{c}\cdot\mathbf{c} & \cdots & \cdot \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \cdot & \cdot & \cdot & \cdots & \cdot \end{vmatrix}. \tag{7}$$

The quantity on the left-hand side is of course V_d^2. The determinant on the right-hand side, formed by taking the scalar products of the d vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}, \ldots$ among themselves, is called the *Gram determinant* G_d. Since $G_d = V_d^2$, it follows that $G_d \geq 0$. The equality sign is valid if and only if $V_d = 0$. But that would imply that the vectors $\mathbf{a}, \mathbf{b}, \ldots$ are not linearly independent of each other, which cannot be the case for a set of vectors forming a basis in the space.

More generally, let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r$ be an arbitrary set of r vectors in a space of dimensionality d. Let G_r be their Gram determinant. Then:

- $G_r \geq 0$.
- The vectors of the set are linearly independent *if and only if* $G_r \neq 0$. That is, the non-vanishing of their Gram determinant is a necessary and sufficient condition for linear independence.
- If $r > d$, $G_r = 0$, i.e., the vectors *cannot* be linearly independent.
- If $r = d$ and $G_r \neq 0$ (so that the vectors are linearly independent), the set can form a basis in the space.
- If $r < d$, the set cannot form a basis in the space even if $G_r \neq 0$, because the vectors do not span all of the space.

Let $\mathbf{v}_1$ and $\mathbf{v}_2$ be any two vectors (in a space of dimensionality d), and let θ be the angle between them. Let v_1 and v_2 be the magnitudes of the two vectors. We then have

$$(\mathbf{v}_1 \cdot \mathbf{v}_1)(\mathbf{v}_2 \cdot \mathbf{v}_2) - (\mathbf{v}_1 \cdot \mathbf{v}_2)^2 = v_1^2 v_2^2 (1 - \cos^2 \theta) \geq 0, \tag{8}$$

because $|\cos \theta| \leq 1$ for any angle θ. The second equality sign in (8) applies if and only if $\mathbf{v}_1$ and $\mathbf{v}_2$ are collinear. When extended to any two vectors in what is called a general *linear vector space*, (8) is known as the *Cauchy–Bunyakovsky–Schwarz inequality*. This important inequality is useful, for instance, in establishing the Heisenberg uncertainty relation between two observables in quantum mechanics. The corresponding extension of the inequality $G_r \geq 0$ to a general linear vector space is thus a generalization of the Cauchy–Bunyakovsky–Schwarz inequality to an arbitrary number of vectors. The way we have arrived at it here brings out its geometrical interpretation.

What is a vector?

In all the foregoing, I have used the terms *scalar* and *vector* without formal definition, in the common or usual sense familiar to us from high school: a scalar only has a numerical value, while a vector has both a magnitude and a direction. This is not a very satisfactory definition even in high school! Direction with respect to which set of coordinate axes? If the answer is, 'Some given, fixed set of axes', then why is it that this set is never specified right at the beginning of each book dealing with vectors? And how is it that the same equation,

say $\mathbf{F} = m\mathbf{a}$, makes sense whether it is written down in Mumbai or Mogadishu, although it is unlikely that the orientations of the coordinate frames chosen in the two cases will be the same?

The correct way to define scalars, vectors, tensors, etc. is via their transformation properties under changes of coordinate frames. Once this is done, any equation or relationship involving only those quantities whose transformation properties are specified is guaranteed to remain the same in form (*form-invariant*) for two users who use different coordinate frames. That is why books dealing with vector equations do not bother to specify any special set of coordinate axes right in the beginning! Even this definition is not adequate. What *sort(s)* of coordinate transformations are we talking about?

The usual (high school!) scalars and vectors that we have considered are actually defined with respect to the set of *rotations* of the coordinate axes. The value of a scalar thus defined (e.g., the distance of a point from the origin of coordinates) does not change at all under such a rotation. A vector comprises a set of numbers (called its components) that do change under a rotation of the coordinate axes – but they do so *in precisely the same manner as the coordinates of any point themselves change*. Indeed, this is the very *definition* of a vector of the usual kind. *Tensors* of higher rank $(2, 3, \ldots)$ are defined in an analogous manner; they have (slightly) more involved transformation properties under rotations of the coordinate axes. In technical terms: the scalars and vectors I have used so far (except for the general cases mentioned briefly on occasion) are actually scalars and vectors *under the group of proper rotations in d-dimensional Euclidean space*. Now that we have become quite familiar with scalars and vectors of this kind, the statement just made should be much easier to digest. As my whole aim has been to provide a simple, heuristic approach to some aspects of vector analysis, we have preferred to mention these issues at this stage, rather than to open the discussion with them. I have also glossed over many mathematical technicalities wherever these have not been directly relevant to the point being made. For instance, I have not made a careful distinction between the elements of a linear vector space and those of the dual vector space, i.e., between vectors and co-vectors, or – in a different language – between vectors and one-forms. Nor have I presented a rigorous version of the somewhat loose statement that the cross-product of two vectors is itself a vector in (Euclidean) 3D space. While there has admittedly been a lack of mathematical rigour in this sense, I believe that the treatment given suffices for our present purposes.

Extensions and generalizations

The elementary concepts I have tried to describe in this chapter and the preceding ones have wide-ranging extensions and generalizations, with remarkably diverse applications. I can only mention some of these here.

The *form invariance* of physical laws from one coordinate frame to another is a cornerstone of all modern physics. In a vastly generalized and extended form, it is, in fact, *the prime guiding principle* in all modern physics! (Examples include the general theory of relativity and quantum field theories describing the interactions of elementary particles, but we shall not

go into this here.) This means that the laws of physics *must* be specified in terms of quantities whose transformation properties are prescribed – e.g., scalars, vectors, tensors, etc. – quantities *which carry their own dictionaries*, so to speak, so that different users related to each other by these transformations can simultaneously and unambiguously use the same laws. This is why we write Newton's second law of motion, or Maxwell's equations of electromagnetism, for instance, in terms of vectors.[‡]

We can see immediately, now, that it might be both necessary and possible to have scalars, vectors, ... (or their equivalents) under *other* sets of transformations than just rotations of the coordinate axes, and in spaces that are more complicated than the Euclidean spaces we have used in the foregoing. We believe, for instance, that in regions where space-time is essentially *flat*, i.e., in the absence of very intense gravitational fields that give space-time a significant curvature, the laws of physics are form invariant under *Lorentz transformations* (which include rotations of the spatial coordinate axes), rather than just rotations of the axes. Moreover, the space-time geometry is not strictly Euclidean. We must therefore deal with scalars, vectors and other such objects defined with respect to the set (more specifically, the group) of Lorentz transformations in what is known as a non-Euclidean 4D *space-time manifold*. In the presence of gravitational fields, this manifold itself becomes curved in a specific mathematical sense. The set of transformations under which physical laws are required to remain form invariant is now even more general. The distinction between vectors and their corresponding dual vectors is now non-trivial, and not just a matter of using oblique axes in a flat (Euclidean) space. To help keep this in mind, the indices are written as *superscripts* for vectors and *subscripts* for their dual counterparts. In tensor analysis, the traditional names for these quantities are *contravariant vectors* and *covariant vectors*. In mathematics, they are simply *vectors* and *co-vectors*, which is much better terminology.

In classical dynamics, too, vectors and co-vectors play a crucial role. In the Lagrangian formalism, we describe a system in terms of a set of generalized coordinates and the corresponding generalized velocities. In the Hamiltonian formalism, the latter are replaced by generalized momenta. Roughly speaking, the shift from velocities to momenta corresponds to going from a vector space to its dual vector space. Pursuing this further, we arrive at the modern mathematical description of Hamiltonian dynamics, using differential geometry and topology. But that is another story.

In quantum mechanics, as already mentioned already in ch. 2, we describe a system by a *state vector* or ket vector $|\psi\rangle$ in a particular kind of linear vector space called a *Hilbert space*. (The corresponding dual is the bra vector $\langle\psi|$.) It turns out, however, that the multiplication of $|\psi\rangle$ by any complex number of the form $e^{i\theta}$ (where θ is any real number) does not lead to a new state, so that $|\psi\rangle$ and $e^{i\theta}\ |\psi\rangle$ describe the same state. In terms of the corresponding bra vectors, this means that $\langle\psi|$ and $\langle\psi|\ \mathrm{e}^{-i\theta}$ are equivalent. Consequently, the appropriate vector space in quantum mechanics is a so-called *projective* Hilbert space rather than the original

[‡]And not because each vector equation stands for three equations (one for each component), thus saving space in books!

Hilbert space itself, and the descriptor of a state is the object $|\psi\rangle\langle\psi|$ rather than the ket $|\psi\rangle$ or bra $\langle\psi|$ by itself. (In this combination, the factors $e^{i\theta}$ and $e^{-i\theta}$ cancel each other out.) By now, we can recognise an object like $|\psi\rangle\langle\psi|$. Like the tensor product **aA**, it is an *operator*. Its formal name in quantum mechanics is the *density operator*. The most general descriptor of a quantum mechanical system is its density operator ρ (which does not, in the general case, have the form $|\psi\rangle\langle\psi|$). The evolution of the system with time is governed by the so-called *Liouville equation* for $\partial\rho/\partial t$. This, too, is another story.

These remarks are meant merely to give an idea of the generality of the concepts of the reciprocal basis and dual vectors, and to whet your appetite for more!

How is a vector rotated?*

V. Balakrishnan

Introduction

In the preceding chapters, we have seen how very general arguments such as linearity, homogeneity and symmetry can be used to derive a number of results in elementary vector and tensor analysis. Moreover, generalizations of many of these results to an Euclidean space of an arbitrary number of dimensions could be made quite readily. Some of the special features of three dimensions (3D) were also brought out. Most notably, the cross-product of two vectors is again a vector only in 3D. As I have already emphasized, this property is connected to the fact that the number of independent, mutually orthogonal *planes* through the origin, given by ${}^dC_2 = d(d-1)/2$, is equal to the number of independent *axes*, d, if and only if $d = 3$. In this chapter, we shall see how this fact can be exploited to deduce, in a simple way, an explicit formula for the action of an arbitrary rotation of the coordinate frame upon any given vector. This formula is of fundamental importance, because of the very definition of a vector: it is a quantity whose components transform, under a rotation of the axes, in precisely the same manner as the position coordinates of an arbitrary point transform.

The approach will again be of the 'What can the answer possibly be?' type. The only place where the argument needs to be supported by a more detailed calculation will be pinpointed. Even this is instructive, as it highlights a very basic property of rotations. The entire argument is quite short, but I will describe it somewhat elaborately for the sake of clarity and ease of understanding. The formula to be deduced is called the *finite rotation formula*. In the next chapter, a more formal treatment of rotations will be given, involving the generators of infinitesimal rotations. As a by-product, that approach will enable us to derive the finite rotation formula rigorously.

Numerous physical quantities (velocity, linear momentum, acceleration, force, electric field, angular momentum, ...) are vectors in the elementary sense of the term: like the position vector of a point in space, they have a magnitude and a direction. The numerical values of these, namely, the values of the *components* of a vector, naturally depend on the particular coordinate frame chosen. However, as I have explained in ch. 4, the whole point of writing relationships between physical quantities in the form of a vector (more generally, tensor) equations is as follows: These relations are independent of the particular coordinate frame one may choose. Their form is the same in all frames obtainable from each other by rotations of the set of axes by arbitrary amounts, in arbitrary directions. A general formula that connects a vector in a given frame to its transformed version in a rotated frame is therefore rather useful. Note that the formula is applicable to any vector quantity, regardless of its physical dimensions in terms of M, L and T (mass, length and time).

*Based on *Resonance*, Vol.**4**, No.10, pp.61-68, 1999.

Rotating a coordinate frame in 3D

Suppose S and S' are two coordinate frames related to each other by a rotation about the origin in S. It is obvious that the same point remains the origin in S'. Let $\mathbf{A}$ denote a vector as described in S, and $\mathbf{A}'$ its counterpart in S'. Given $\mathbf{A}$, we want to find $\mathbf{A}'$.

The most common context in which such a problem arises is that of the rotational dynamics of a rigid body. In that case, one has a fixed reference frame called *space-fixed axes* (which we may identify with S), and another reference frame attached to the rigid body, called *body-fixed axes* (which we may identify with S'). The orientation of the rigid body at any instant of time is then specified by the orientation of the body-fixed axes relative to the space-fixed axes. In turn, this is quantified by the rotation needed to put the set of axes in S into coincidence with those in S', i.e., the rotation that 'takes' S to S'.

A standard way of describing or implementing such a rotation proceeds in three successive steps via the use of three *Euler angles* α, β and γ, as follows.

(i) Rotate the frame S about its z-axis by an angle α.

(ii) Rotate the resulting frame about its (new) x-axis by an angle β.

(iii) Rotate the resulting frame about its (new) z-axis by an angle γ.

It can be shown that any orientation of S' with respect to S, no matter what it is, is guaranteed to be reached in this manner by suitable choices of the angles α, β and γ. The sequence of rotations, however, is not unique: there are other sequences of three rotations that will take S to any given S'.** Although the specification of a rotation via a set of Euler angles is very useful in practice, especially in engineering mechanics, it is not very convenient for our present purposes.

There is an alternative way to specify a rotation of the coordinate frame about the origin of coordinates, also discovered by Euler. (It is called *Euler's Theorem* in rigid body dynamics.) This description of a general rotation has a direct physical interpretation. According to Euler's Theorem, any orientation S' can be reached by rotating the frame S through an angle ψ about a direction given by a unit vector $\mathbf{n}$ (see Figure 1). That is, the axis of rotation is $\mathbf{n}$, while the amount of rotation (in the plane normal to $\mathbf{n}$) is ψ. Now, any unit vector $\mathbf{n}$ is itself specified by two angles in S: a polar angle θ and an azimuthal angle ϕ. Therefore the rotation as a whole is again specified by three angles (θ, ϕ, ψ), instead of three Euler angles (α, β, γ). Let us denote the corresponding *rotation operator* that is to act on vectors by $R(\mathbf{n}, \psi)$. In other words, R acts on any vector $\mathbf{A}$ (specified in S) to produce another vector $\mathbf{A}'$, according to the formal equation

$$\mathbf{A}' = R(\mathbf{n}, \psi)\,\mathbf{A}. \tag{1}$$

The task is to derive an explicit formula for $\mathbf{A}'$.

**There are, in fact, 12 different Euler angle conventions possible.

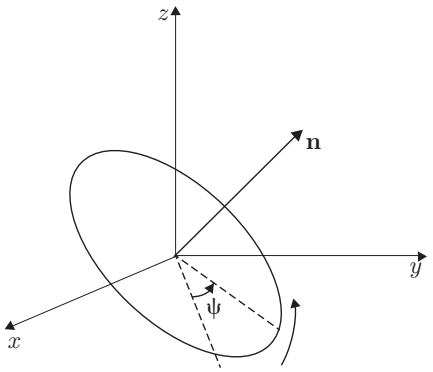


Figure 1: A rotation of the coordinate frame about the origin is specified by the axis of rotation **n** and the angle of rotation ψ.

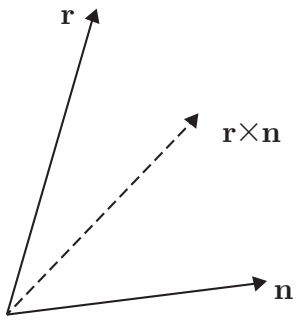


Figure 2: The position vector $\mathbf{r}'$ of a point in the rotated frame must be a linear combination of its original position vector **r**, the axis of rotation **n**, and the vector $\mathbf{r} \times \mathbf{n}$.

What can A′ possibly be?

It is obvious that $\mathbf{A}'$ will be a function of ψ, **n** and of course **A** itself. The only *vectors* on which $\mathbf{A}'$ can depend are **A** and **n**. We may therefore expect $\mathbf{A}'$ to have components along each of these two vectors. Moreover, **A** and **n** together determine a certain plane, and the vector $(\mathbf{A}\times\mathbf{n})$ is normal to this plane. Therefore $\mathbf{A}'$ might have a component in this direction as well. In other words, **A**, **n** and $(\mathbf{A}\times\mathbf{n})$ form a triad in general, i.e., a set of oblique axes, along which $\mathbf{A}'$ can be resolved. (Figure 2 depicts this triad in the case when $\mathbf{A} = \mathbf{r}$, the position vector of an arbitrary point.) Thus $\mathbf{A}'$ *must* be a linear combination of the three vectors **A**, **n** and $(\mathbf{A} \times \mathbf{n})$. Moreover, each term in the expansion must be linear in **A** (equivalently, in physical terms: each term must have the same physical dimensions as **A**). This implies that the part along **n** must actually be

proportional to $(\mathbf{n}\cdot\mathbf{A})\mathbf{n}$, which is the portion of $\mathbf{A}$ along $\mathbf{n}$. Therefore $\mathbf{A}'$ *must* be of the form

$$\mathbf{A}' = f\mathbf{A} + g\,(\mathbf{n}\cdot\mathbf{A})\,\mathbf{n} + h\,(\mathbf{A}\times\mathbf{n}), \tag{2}$$

where f, g and h are *scalar* quantities, yet to be determined.

Now, the effect on a vector $\mathbf{A}$ of rotating the coordinate *frame* S about the axis $\mathbf{n}$ through an angle ψ is the same as that of leaving the frame fixed, but rotating the *vector* about the same axis through an angle $-\psi$, as follows.The tips of the vectors $\mathbf{A}$ and $\mathbf{A}'$ lie on the rim of a cone whose axis is along $\mathbf{n}$. The magnitudes of $\mathbf{A}$ and $\mathbf{A}'$ are equal, i.e., $|\mathbf{A}| = A = |\mathbf{A}'|$. The angle between $\mathbf{n}$ and $\mathbf{A}$ is equal to the angle between $\mathbf{n}$ and $\mathbf{A}'$. (This is the half-angle of the cone.) Hence $\mathbf{n}\cdot\mathbf{A} = \mathbf{n}\cdot\mathbf{A}'$. The angles $\measuredangle(\mathbf{n},\mathbf{A})$ and ψ are the only two angles in the problem. The scalars f, g and h can only depend on these angles. But a rotation is a *linear* transformation when it acts on any vector $\mathbf{A}$, as implied by eq. (1). This means that f, g and h cannot themselves have any sort of dependence on the particular vector $\mathbf{A}$ on which the rotation acts. It follows that f, g and h can only depend on ψ. Equation (2) therefore reduces to

$$\mathbf{A}' = f(\psi)\,\mathbf{A} + g(\psi)\,(\mathbf{n}\cdot\mathbf{A})\,\mathbf{n} + h(\psi)\,(\mathbf{A}\times\mathbf{n}). \tag{3}$$

We can go a little farther. It is clear, intuitively, that a rotation is a *continuous* transformation: that is, a rotation about the axis $\mathbf{n}$ through an angle ψ can be implemented by a succession of infinitesimal rotations about the dame axis. f, g and h must therefore be continuous, differentiable functions of ψ. Moreover, a rotation about any axis by a multiple of 2π clearly brings S back its original orientation. $\mathbf{A}'$ must coincide with $\mathbf{A}$ in this case. Hence $f(\psi), g(\psi)$ and $h(\psi)$ must be *periodic* functions of ψ with a period of 2π, i.e.,

$$f(\psi + 2\pi) = f(\psi), \quad g(\psi + 2\pi) = g(\psi), \quad h(\psi + 2\pi) = h(\psi). \tag{4}$$

We conclude that f, g and h are continuous, differentiable functions of $\cos\psi$ and $\sin\psi$.

Determination of $f(\psi), g(\psi)$ and $h(\psi)$

We note first that $\mathbf{A}$ will remain unaltered if the axis of rotation is along $\mathbf{A}$ itself, whatever be the value of ψ. (Any vector along the axis of rotation is obviously left unchanged by the rotation.) In this situation $(\mathbf{n}\cdot\mathbf{A})\,\mathbf{n} = \mathbf{A}$, and $\mathbf{n}\times\mathbf{A} = 0$. Therefore we must have

$$\mathbf{A} = [f(\psi) + g(\psi)]\,\mathbf{A} \tag{5}$$

for every vector $\mathbf{A}$, i.e., $f(\psi) + g(\psi) = 1$ in *general*. Equation (3) reduces to

$$\mathbf{A}' = f(\psi)\,\mathbf{A} + [1 - f(\psi)]\,(\mathbf{n}\cdot\mathbf{A})\,\mathbf{n} + h(\psi)\,(\mathbf{A}\times\mathbf{n}). \tag{6}$$

But the vector $\mathbf{A}$ must remain unchanged if $\psi = 0$ (no rotation occurs at all), whatever be $\mathbf{n}$. Hence

$$\mathbf{A} = f(0)\,\mathbf{A} + [1 - f(0)]\,(\mathbf{n}\cdot\mathbf{A})\,\mathbf{n} + h(0)\,(\mathbf{A}\times\mathbf{n}). \tag{7}$$

Once again, this must hold good for any arbitrary direction **n** and any vector **A**, which is possible only if

$$f(0) = 1 \quad \text{and} \quad h(0) = 0. \tag{8}$$

Now square each side of eq. (6), and recall that $|\mathbf{A}| = |\mathbf{A}'| = A$. Using the identities

$$\begin{aligned} \mathbf{A} \cdot (\mathbf{A} \times \mathbf{n}) &= \mathbf{n} \cdot (\mathbf{A} \times \mathbf{n}) = 0, \\ (\mathbf{A} \times \mathbf{n}) \cdot (\mathbf{A} \times \mathbf{n}) &= A^2 - (\mathbf{A} \cdot \mathbf{n})^2, \end{aligned} \tag{9}$$

we get

$$A^2 = \left\{ f^2(\psi) + h^2(\psi) \right\} A^2 + \left[1 - \left\{ f^2(\psi) + h^2(\psi) \right\} \right] (\mathbf{n} \cdot \mathbf{A})^2. \tag{10}$$

This relation can be valid for *all* **n** and **A** if and only if

$$f^2(\psi) + h^2(\psi) = 1. \tag{11}$$

Equations (4), (8) and (11) suggest strongly[‡] that

$$f(\psi) = \cos \psi \quad \text{and} \quad h(\psi) = \pm \sin \psi. \tag{12}$$

To decide whether $h(\psi) = +\sin \psi$ or $-\sin \psi$, it is convenient to look at the special case when **n** is perpendicular to **A** and, moreover, the angle of rotation is an *infinitesimal* one, $\delta\psi$. Then $\cos(\delta\psi) \approx 1$ and $\sin(\delta\psi) \approx \delta\psi$, to first order in the infinitesimal $\delta\psi$. From Figure 2, we see that

$$\mathbf{A}' - \mathbf{A} = -(\mathbf{n} \times \mathbf{A})\, \delta\psi \approx (\mathbf{A} \times \mathbf{n}) \sin(\delta\psi), \tag{13}$$

so that

$$\mathbf{A}' = \mathbf{A} + (\mathbf{A} \times \mathbf{n}) \sin(\delta\psi). \tag{14}$$

From this we deduce that $h(\psi) = \sin\psi$ in general, because a finite rotation ψ may be regarded as being built up of a succession of incremental angles $\delta\psi$.

With the functions $f(\psi) = \cos\psi$ and $h(\psi) = \sin\psi$ at hand, we are ready to write down the general formula for the transformation of an arbitrary vector **A** under rotation of the coordinate frame about an axis **n**, through an angle ψ:

$$\mathbf{A}' = R(\mathbf{n}, \psi)\, \mathbf{A} = (\cos\psi)\, \mathbf{A} + (1 - \cos\psi)\, (\mathbf{n} \cdot \mathbf{A})\, \mathbf{n} + (\sin\psi)\, (\mathbf{A} \times \mathbf{n}). \tag{15}$$

This is the finite rotation formula[§] for a vector.

At this stage, it is necessary to point out the weak link in this admittedly non-rigorous derivation of the formula. I have stated that eqs (4), (8) and (11) suggest strongly the solutions $\cos\psi$ and $\sin\psi$ for the functions $f(\psi)$ and $h(\psi)$, respectively. Obviously, these are not unique solutions. For instance, the functions

$$f(\psi) = \cos(\sin\psi), \quad h(\psi) = \sin(\sin\psi) \tag{16}$$

[‡]The phrase 'suggest strongly' requires some elaboration. This follows shortly.

[§]Also called Rodrigues' rotation formula.

or the functions

$$f(\psi) = \cos\left[\frac{1}{3}(\sin\psi + \sin 2\psi)\right],\ h(\psi) = \sin\left[\frac{1}{3}(\sin\psi + \sin 2\psi)\right] \tag{17}$$

also satisfy the same conditions, as do an infinite number of other possibilities. The functions $\cos\psi$ and $\sin\psi$ are merely the simplest of these possibilities. They also happen to be the correct solutions, as we shall show by a proper derivation of the finite rotation formula in ch. 6. The rigorous derivation of the results $f(\psi) = \cos\psi$, $h(\psi) = \sin\psi$ depends on the only property of rotations that we have not fully exploited so far — namely, that a rotation through a finite angle ψ can be achieved by a succession of n rotations through infinitesimal angles $\delta\psi$ such that, in the limit $n \to \infty$ and $\delta\psi \to 0$, the product $n\,\delta\psi \to \psi$. (In technical language, 'a proper rotation is continuously connected to the identity transformation'.) *All* other properties of rotations in 3D have been used in the simple step-by-step derivation just described: first and foremost, the existence of an axis of rotation $\mathbf{n}$ corresponding to any arbitrary rotation; next, the linear and homogeneous nature of the transformation (every term on the right in eq. (2) is linear in $\mathbf{A}$); its distance-preserving nature ($|\mathbf{A}| = |\mathbf{A}'|$); and finally, the fact that a rotation by a multiple of 2π is equivalent to no rotation at all as far as scalars and vectors are concerned.

Remarks on rotations in dimensions $d \neq 3$

The formula in eq. (15) gives an explicit representation of the rotation operator $R(\mathbf{n}, \psi)$ in 3D in the form in which it acts on an arbitrary vector. Symbolically,

$$R(\mathbf{n}, \psi) = (\cos\psi)\,\mathbb{I} + (1 - \cos\psi)\,\mathbf{n}\,\mathbf{n}\cdot - (\sin\psi)\,\mathbf{n}\times \tag{18}$$

Any vector on which R acts is to be inserted to the right of the operator on the right-hand side. The symbol $\mathbb{I}$ stands for the unit operator. Any vector on which it acts is left unchanged. $\mathbf{n}\,\mathbf{n}\cdot$ is a projection operator of the kind introduced in ch. 2. When it acts on a vector $\mathbf{A}$ from the left, it yields $\mathbf{n}\,(\mathbf{n}\cdot\mathbf{A})$.[¶] An explicit representation such as this for the action of *finite* (as opposed to infinitesimal) transformations is obviously very useful, but is not as readily available for rotations in dimensions $d \neq 3$.

Consider, first, a rotation about the origin of coordinates in 2D, i.e., in a plane. The rotation takes place about a *point*. There is no *axis* about which the rotation of the coordinate frame occurs. The effect of a rotation of the (Cartesian) axes by an angle ψ is easily deduced. Using elementary trigonometry, we find that the position vector $\mathbf{r} = (x, y)$ of an arbitrary point is transformed to $\mathbf{r}' = (x', y')$, where

$$x' = x\cos\psi + y\sin\psi,\quad y' = -x\sin\psi + y\cos\psi. \tag{19}$$

I have already emphasised that a vector is *defined* as a quantity that transforms (under a rotation of the coordinate frame) in precisely the same way as the position coordinates transform.

[¶] Objects like $\mathbb{I}$ and $\mathbf{n}\,\mathbf{n}$ are sometimes referred to as *dyads*, especially in the older literature on vector analysis.

Therefore, given *any* vector $\mathbf{A}$, eqs (19) specify the components (A'_x, A'_y) of the transformed vector $\mathbf{A}'$ in terms of the components (A_x, A_y) of the original vector $\mathbf{A}$. However, there is no way of writing $\mathbf{A}'$ in *vector* form in a manner analogous to eq. (15). This is because in 2D there is no vector like the unit vector $\mathbf{n}$ in 3D characterizing the rotation. We can, however, invoke the antisymmetric Levi-Civita symbol ϵ_{ij} defined in ch. 3, and express the *components* of $\mathbf{A}'$ in the form

$$A'_i = (\cos\,\psi)\, A_i + (\sin\,\psi)\, \epsilon_{ij}\, A_j, \tag{20}$$

where the indices i and j run over the values 1 and 2, and a summation over the repeated index j is implied. The analogous formula in 3D is, from eq. (15),

$$A'_i = (\cos\,\psi) A_i + (1 - \cos\psi)\, n_i\, n_j\, A_j + (\sin\,\psi)\, \epsilon_{ijk}\, A_j\, n_k. \tag{21}$$

One might be tempted to jump to the following conclusion: In 2D, a rotation leaves a point (dimension = 0) unchanged; in 3D, it leaves a line or axis (dimension = 1) unaffected. By extrapolation, one might imagine that in d-dimensional space, an arbitrary rotation would leave a subspace of dimensions $(d-2)$ unchanged, thus affecting only some plane in the space. But this is incorrect, as the following argument shows.

Any rotation in d dimensions can be expressed as the result of *successive* 'planar' rotations, i.e., linear, homogeneous, distance-preserving transformations, each of which acts on just two of the coordinates, leaving the others unchanged. In d dimensions, the coordinate axes taken pairwise define $d(d-1)/2$ different orthogonal planes. A rotation in d dimensions is specified by $d(d-1)/2$ angles (analogous to the three Euler angles in 3D). However, a general rotation in $d \geq 4$ dimensions *cannot* be reduced to a rotation in some single 'tilted' plane in space, leaving a $(d-2)$-dimensional subspace unaltered. It is easy to see why: If the latter were possible, such a rotation would be specifiable by just d parameters: namely, $(d-1)$ parameters to specify the unit vector normal to the plane concerned, together with 1 angle to specify the angle through which the plane is rotated. But when $d > 3$, this is *less* than the actual number of parameters $[d(d-1)/2]$ it takes to specify a general rotation in the space. Therefore a general rotation is not reducible to a rotation in some plane in the space. As a consequence, formulas as simple as those of eqs (20) and (21) are thus not available for these transformations in spaces of dimensions greater than 3. There are, however, many interesting properties of rotations in $d > 3$ that have been established by mathematicians. For instance, every rotation in $d = 4$ can be implemented in at least one way by rotations in 2 different mutually orthogonal planes in the space.

Another interesting fact is worth pointing out. We may ask whether a general rotation in $d \geq 3$ leaves at least a *direction* (a 1-dimensional subspace) unchanged: in other words, can we associate an *axis* with every rotation? The answer reveals a deep and profound difference between even and odd dimensional spaces. It is 'no' if d is even, and 'yes' if d is odd. For example, let x_i $(i = 1, 2, 3, 4)$ be the cartesian coordinates in 4D. Consider a rotation in the (x_1, x_2) plane by an angle α_1, followed by a rotation in the (x_3, x_4) plane by an angle α_2. We

are then guaranteed that the net rotation leaves no nonzero vector unchanged, i.e., no axis of rotation exists for this particular rotation.

There is an easy way to see how this general result comes about. A ('proper') rotation R in d dimensions can be implemented (represented) by a $(d \times d)$ real orthogonal matrix with determinant equal to $+1$. All d eigenvalues of such a matrix must lie on the unit circle in the complex plane, in complex conjugate pairs. The product of all the eigenvalues must be equal to $+1$, the determinant of the matrix. (In the example given above, the eigenvalues of R are $e^{\pm i\alpha_1}$ and $e^{\pm i\alpha_2}$.) If $d = 2k+1$, there are in general k such pairs $e^{\pm i\alpha_1}, \ldots, e^{\pm i\alpha_k}$, and a final eigenvalue $+1$. Hence there must be at least one non-zero vector $\mathbf{n}$ in the space, such that $R\,\mathbf{n} = 1\,\mathbf{n}$. That is, there is an *axis* of rotation for every rotation in any odd-dimensional space. A moment's thought shows that this axis is *unique* when $d = 3$, i.e., exactly one eigenvalue is equal to $+1$ for every rotation matrix in this case.

In the next chapter, we will return to rotations in 3-dimensional space, and derive an expression for the matrix representing an arbitrary rotation. As a by-product, this will yield a rigorous derivation of the finite rotation formula.

Rotation matrices and the rotation group

V. Balakrishnan

Recall that a rotation of the coordinate frame about the origin in 3-dimensional space can be specified by an axis of rotation, **n**, and an angle of rotation, ψ, as shown in Figure 1, ch. 5.) The rotation, denoted by $R(\mathbf{n}, \psi)$, acts on the position vector **r** of any arbitrary point to yield the position vector $\mathbf{r}' = R(\mathbf{n}, \psi)$ in the rotated coordinate frame. As emphasised in preceding chapters, any arbitrary vector **A** also transforms, *by definition*, in precisely the same way: that is, $\mathbf{A}' = R(\mathbf{n}, \psi)\mathbf{A}$. In ch. 5, I have described an heuristic way of arriving at an expression for $\mathbf{A}'$ (the finite rotation formula). In this chapter, we shall derive this formula rigorously. On the way, we will discuss the generators of rotations, the algebra satisfied by these generators, and other related matters.

Generators of infinitesimal rotations

Let us write the position vector **r** of a point in the form of a (3×1) column matrix with elements x_1, x_2, x_3 (the Cartesian coordinates of the point concerned). $R(\mathbf{n}, \psi)$ is then represented by a (3×3) matrix with elements $R_{jk}(\mathbf{n}, \psi)$, where the indices j and k run over the values 1, 2 and 3. This is called a rotation matrix. Our focus will be on the set of (3×3) rotation matrices, parametrized by **n** and ψ. This set of matrices forms the so-called standard representation of the 3-dimensional rotation group. As we shall see, the matrices will turn out to be orthogonal (i.e., $RR^{\mathrm{T}} = I_3$ where T stands for the transpose and I_3 is the (3×3) identity matrix) and unimodular (i.e., the determinant of R is equal to $+1$). Such matrices comprise the *special orthogonal group* in 3 dimensions, $SO(3)$. More will be said about this group in the sequel.

Let $(\mathbf{i}, \mathbf{j}, \mathbf{k})$ denote the unit vectors along the three Cartesian axes, as usual. We begin by writing down the rotation matrix corresponding to a special case: a rotation about **k** (i.e., a rotation in the (x_1, x_2) plane) by an angle ψ. This leaves the coordinates of points on the x_3 axis unchanged. As you know from elementary coordinate geometry, we have in this case

$$R(\mathbf{k}, \psi) = \begin{pmatrix} \cos\psi & \sin\psi & 0 \\ -\sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{1}$$

By cyclic permutation, the rotation matrices corresponding to rotations about **i** and **j** are easily written down. They are

$$R(\mathbf{i}, \psi) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\psi & \sin\psi \\ 0 & -\sin\psi & \cos\psi \end{pmatrix} \tag{2}$$

and

$$R(\mathbf{j},\psi) = \begin{pmatrix} \cos\psi & 0 & -\sin\psi \\ 0 & 1 & 0 \\ \sin\psi & 0 & \cos\psi \end{pmatrix}. \tag{3}$$

You can easily check that $RR^{\mathrm{T}} = I_3$ and $\det R = 1$ in each of these cases.

The next step is to show that each of these rotation matrices is the *exponential* of another matrix, being generated as the limit of a succession of infinitesimal rotations about the same axis. This is easy to do because we already have the closed-form expressions for finite rotations in eqs (1)–(3). Take, for instance, $R(\mathbf{k},\psi)$. This rotation can be achieved by n successive rotations through an infinitesimal angle $\delta\psi$ about the x_3-axis, such that $n\,\delta\psi = \psi$. $R(\mathbf{k},\delta\psi)$ is easy to write down, because $\cos\delta\psi \simeq 1$ and $\sin\delta\psi \simeq \delta\psi$ correct to first order in $\delta\psi$. Taking out the (3×3) unit matrix, which corresponds to the identity transformation (i.e., no rotation at all), we get, to first order in $\delta\psi$,

$$R(\mathbf{k},\delta\psi) = I_3 + i\,J_3\,(\delta\psi), \tag{4}$$

where

$$J_3 = \begin{pmatrix} 0 & -i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{5}$$

Note that the elements of the matrix J_3 are pure numbers that are *independent* of the angle of rotation. A factor i has also been separated out in the definition of J_3 in order to ensure that J_3 is a *hermitian* matrix, i.e., $J_3 = J_3^{\dagger}$ (its complex conjugate transpose).

The matrix corresponding to rotation about the x_3-axis by a finite angle ψ is then given by the n-fold product of $R(\mathbf{k},\delta\psi)$ with itself, i.e.,

$$R(\mathbf{k},\psi) = [R(\mathbf{k},\delta\psi)]^n = \left[I_3 + i\,J_3\,(\delta\psi)\right]^n. \tag{6}$$

Now set $\delta\psi = \psi/n$ and let $n\to\infty$. Then

$$R(\mathbf{k},\psi) = \lim_{n\to\infty}\left(I_3 + \frac{i\,J_3\,\psi}{n}\right)^n = e^{i\,J_3\,\psi}. \tag{7}$$

Note that the exponential of any square matrix M of finite order has exactly the same power series expansion as the exponential of any complex number: that is,

$$e^M = \sum_{n=0}^{\infty} M^n/n! \tag{8}$$

Repeating the foregoing procedure for $R(\mathbf{i},\psi)$ and $R(\mathbf{j},\psi)$, we get

$$R(\mathbf{i},\psi) = e^{i\,J_1\,\psi} \quad \text{where} \quad J_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -i \\ 0 & i & 0 \end{pmatrix} \tag{9}$$

and

$$R(\mathbf{j},\psi) = e^{i J_2 \psi} \quad \text{where} \quad J_2 = \begin{pmatrix} 0 & 0 & i \\ 0 & 0 & 0 \\ -i & 0 & 0 \end{pmatrix}. \tag{10}$$

The hermitian matrices J_1, J_2 and J_3 are called the *generators* of infinitesimal rotations about the x_1, x_2 and x_3 axes, respectively. Exponentiating them as in eqs (9), (10) and (7) yields the corresponding matrices for rotations by a finite angle ψ. I leave it as an exercise for the reader to start with (9), (10) and (7), use the expansion in eq. (8) for the exponential of a matrix, and recover eqs (2), (3), and (1).

The most important property of these generators is as follows. Recall that the *commutator* of two square matrices A and B of the same order is defined as

$$[A, B] \equiv AB - BA. \tag{11}$$

It is easily checked (do so!) that the generators satisfy the *commutation relations*

$$[J_1, J_2] = i\, J_3, \quad [J_2, J_3] = i\, J_1, \quad [J_3, J_1] = i\, J_2. \tag{12}$$

Using ϵ_{klm}, the Levi-Civita symbol in 3 dimensions (defined in ch. 3), these relations can be written more compactly as

$$[J_k\,,\, J_l] = i\epsilon_{klm}\, J_m, \tag{13}$$

where the indices run over the values 1, 2 and 3, and a summation over the repeated index m is implied. Equation (13) is called the *angular momentum algebra*. The reason for this name lies in quantum mechanics, where hermitian matrices (more generally, hermitian operators) are associated with physical observables. The generators of infinitesimal rotations are the matrices associated with the components of the angular momentum.

The general rotation matrix: formal expression

We have seen that the matrices corresponding to rotations about the three Cartesian axes can be written as exponentials of the corresponding infinitesimal generators. What about a *general* rotation by an angle ψ, about an axis $\mathbf{n}$ pointing in an arbitrary direction in space?

A problem arises now. Rotations through finite angles about two different axes do not commute with each other! The final orientation of a coordinate frame after two successive rotations about different axes of rotation depends on the order in which the rotations are performed. This is easy to visualize physically.* In mathematical terms, the corresponding rotation matrices do not commute with each other. For example,

$$R(\mathbf{i},\alpha)R(\mathbf{j},\beta) \neq R(\mathbf{j},\beta)R(\mathbf{i},\alpha), \tag{14}$$

*For instance, rotate a cube by $\frac{1}{2}\pi$ about two mutually perpendicular axes in succession. Now carry out the rotations in the opposite order, and observe that the final orientation of the cube is different in the two cases.

precisely because $J_1 J_2 \neq J_2 J_1$. On the other hand, *infinitesimal* rotations about different axes *do* commute with each other, correct to first order in the infinitesimal angles. If (n_1, n_2, n_3) are the Cartesian components of the unit vector $\mathbf{n}$, then

$$\begin{aligned} R(\mathbf{n}, \delta\psi) &\simeq (I_3 + iJ_1 n_1\, \delta\psi)(I_3 + iJ_2 n_2\, \delta\psi)(I_3 + iJ_3 n_3\, \delta\psi) \\ &\simeq [I_3 + i(J_1 n_1 + J_2 n_2 + J_3 n_3)\, \delta\psi] \\ &\equiv I_3 + i\,(\mathbf{J}\cdot\mathbf{n})\delta\psi. \end{aligned} \tag{15}$$

This statement requires elaboration. First, the order of the brackets in the right-hand side of the first equation in (15) does not matter, *correct to first order* in $\delta\psi$. Second, the generators (J_1, J_2, J_3) have been treated like the components of a vector $\mathbf{J}$ in the final equation in (15). *This is more than mere notation*, as you will see later on, in ch. 7. With this result at hand, it follows that

$$R(\mathbf{n}, \psi) = \lim_{n\to\infty} \left(I_3 + \frac{i\,(\mathbf{J}\cdot\mathbf{n})\psi}{n} \right)^n = e^{i\,(\mathbf{J}\cdot\mathbf{n})\,\psi}. \tag{16}$$

Thus, the rotation matrix for any arbitrary rotation is also expressible as the exponential of a linear combination of the three generators.

The general rotation matrix: explicit form

The next task is to examine whether this expression can be simplified to yield an *explicit* (3×3) rotation matrix whose elements are functions of $\mathbf{n}$ and ψ. It is not immediately clear that this can be done, because

$$e^{i\,(J_1 n_1 + J_1 n_1 + J_1 n_1)\,\psi} \neq e^{i\,J_1 n_1\,\psi}\, e^{i\,J_2 n_2\,\psi}\, e^{i\,J_3 n_3\,\psi}. \tag{17}$$

Now, there is no closed-form expression for the exponential of a general (3×3) matrix, unlike the case of a general (2×2) matrix. The matrix $M \equiv i(\mathbf{J}\cdot\mathbf{n})$, however, is a rather special kind of matrix. It has a form that enables the matrix $R(\mathbf{n}, \psi) = e^{M\psi}$ to be computed fairly easily. Using the expressions in eqs (9), (10) and (5) for the generators, we find

$$M = i\,J_k n_k = \begin{pmatrix} 0 & n_3 & -n_2 \\ -n_3 & 0 & n_1 \\ n_2 & -n_1 & 0 \end{pmatrix}. \tag{18}$$

The general element of M is given by

$$M_{ij} = \epsilon_{ijk}\, n_k\,. \tag{19}$$

A bit of 'index manipulation' now follows! The general element of the *square* of M is given by

$$\left(M^2\right)_{ij} = M_{ik}\, M_{kj} = \epsilon_{ikl}\, n_l\, \epsilon_{kjm}\, n_m$$

$$\begin{aligned} &= -\epsilon_{kil}\,\epsilon_{kjm}\,n_l\,n_m = (\delta_{im}\,\delta_{jl} - \delta_{ij}\,\delta_{lm})\,n_l\,n_m \\ &= n_i\,n_j - \delta_{ij}. \end{aligned} \tag{20}$$

I have used here the well-known identity that relates the product of two Levi-Civita symbols with one common ('contracted') index to a combination of Kronecker deltas. With eqs (19) and (20) at hand, we get

$$\left(M^3\right)_{ij} = (n_i\,n_k - \delta_{ik})\,\epsilon_{kjl}\,n_l = -\epsilon_{ijl}\,n_l. \tag{21}$$

Note that the term $\epsilon_{kjl}\,n_i\,n_k\,n_l$ vanishes identically, because ϵ_{kjl} is antisymmetric in the indices k and l, while $n_i\,n_k\,n_l$ is symmetric in these two indices.**

Equation (21) implies that $M^3 = -M$. Therefore $M^4 = -M^2$, $M^5 = M$, and so on. It follows at once that $e^{M\psi}$ can be written as a linear combination of the matrices I_3, M and M^2. We get

$$\begin{aligned} e^{M\psi} &= I_3 + M\left(\psi - \frac{\psi^3}{3!} + \cdots\right) + M^2\left(\frac{\psi^2}{2!} - \frac{\psi^4}{4!} + \cdots\right) \\ &= I_3 + M\,\sin\,\psi + M^2\,(1 - \cos\,\psi). \end{aligned} \tag{22}$$

This is the formula sought for the general rotation matrix $R(\mathbf{n}, \psi) = e^{M\psi}$. Putting in the expressions in eqs (19) and (20) for the matrix elements of M and M^2, the final result is the following. The matrix element of a general rotation matrix $R(\mathbf{n}, \psi)$ is given by

$$R_{ij}(\mathbf{n}, \psi) = \delta_{ij}\,\cos\,\psi + n_i\,n_j\,(1 - \cos\,\psi) + \epsilon_{ijk}\,n_k\,\sin\,\psi. \tag{23}$$

The simplicity and elegance of this formula are striking! It is convenient, on occasion, to express the direction cosines n_1, n_2 and n_3 in terms of the spherical polar angles θ and φ of the unit vector $\mathbf{n}$, according to

$$n_1 = \sin\,\theta\,\cos\,\varphi, \quad n_2 = \sin\,\theta\,\sin\,\varphi, \quad n_3 = \cos\,\theta. \tag{24}$$

The finite rotation formula for a vector follows directly from eq. (23). If $\mathbf{A}$ is a vector with components $A_i\,(i = 1, 2, 3)$ in the original coordinate frame, and $\mathbf{A}'$ is the same vector in the rotated frame, the components of the transformed vector are given in terms of the original components by

$$\begin{aligned} A_i' &= R_{ij}(\mathbf{n}, \psi)\,A_j \\ &= A_i + (1 - \cos\,\psi)(A_j\,n_j)\,n_i + (\sin\,\psi)\,\epsilon_{ijk}\,A_j\,n_k. \end{aligned} \tag{25}$$

Recognising that $A_j\,n_j = \mathbf{A}\cdot\mathbf{n}$ and that $\epsilon_{ijk}\,A_j\,n_k$ is just the i^{th} component of the vector $\mathbf{A}\times\mathbf{n}$, we have recovered the finite rotation formula written down in ch. 5: namely,

$$\mathbf{A}' = \mathbf{A} + (1 - \cos\,\psi)\,(\mathbf{A}\cdot\mathbf{n})\,\mathbf{n} + (\sin\,\psi)\,(\mathbf{A}\times\mathbf{n}). \tag{26}$$

**The antisymmetric factor changes sign under the interchange of the dummy or contracted indices k and l, while the symmetric part does not do so. Hence the term is the negative of itself, implying that it must be identically equal to zero.

Properties of the rotation matrix

We are ready, now, to deduce numerous properties of rotation matrices, and hence those of rotation transformations themselves.

(i) R is an *orthogonal* matrix. This can be established in many ways. Here is a simple one. Since $\epsilon_{ijk} = -\epsilon_{jik}$, M is an antisymmetric matrix, i.e., $M^{\mathrm{T}} = -M$. Therefore

$$R^{\mathrm{T}} = \left(e^{M\psi}\right)^{\mathrm{T}} = e^{M^{\mathrm{T}}\psi} = e^{-M\psi} = R^{-1}. \tag{27}$$

Thus, $R^{\mathrm{T}}R = I_3$, so that $(R^{\mathrm{T}}R)_{jk} = \delta_{jk}$. The orthogonality of $R(\mathbf{n}, \psi)$ ensures that distance of any point from the origin is preserved under a rotation of the coordinate axes: we have

$$\begin{aligned} x_i' \, x_i' &= R_{ij} \, x_j \, R_{ik} \, x_k = R^{\mathrm{T}}_{ji} \, R_{ik} \, x_j \, x_k \\ &= (R^{\mathrm{T}}R)_{jk} \, x_j \, x_k = \delta_{jk} \, x_j \, x_k = x_j \, x_j. \end{aligned} \tag{28}$$

(ii) R is a *unimodular* matrix. Since $R^{\mathrm{T}}R = I_3$, we have $(\det R)^2 = 1$. Hence $\det R = \pm 1$. We have been concerned with *proper rotations* throughout, i.e., rotations that are continuously connected to the identity transformation. In other words, the final orientation of the coordinate frame can be attained by continuously rotating the original frame from its starting orientation. For such rotations, $\det R = +1$.[‡] The physical significance of the property $|\det R| = 1$ is that volume elements are unchanged under the transformation. The physical significance of the property $\det R = +1$ is that the *handedness* of the coordinate axes is preserved by the transformation. That is, a right-handed (respectively, left-handed) triad of coordinate axes remains right-handed (respectively, left-handed) after the rotation.

(iii) Every $R(\mathbf{n}, \psi)$ has an *inverse* transformation. As you have seen, $R^{\mathrm{T}}(\mathbf{n}, \psi) = R^{-1}(\mathbf{n}, \psi)$. It is evident, both physically and analytically, that $R^{-1}(\mathbf{n}, \psi) = R(\mathbf{n}, -\psi)$. (It is obvious that a rotation can be 'undone' by rotating the frame in the opposite sense through the same angle, about the same axis.)

(iv) A unique *axis of rotation* can be associated with every rotation. In fact, we started with this premise (it is part of Euler's rotation theorem). In mathematical terms, this means that every rotation matrix $R(\mathbf{n}, \psi)$ has an *eigenvalue* equal to $+1$, the corresponding eigenvector being $\mathbf{n}$. This is easily checked: from eq. (23), it follows easily that $R_{ij} \, n_j = n_j$.

[‡]When $\det R = -1$, we have an *improper* rotation. Such a transformation involves an odd number of reflections over and above a possible proper rotation.

(v) The *eigenvalues* of a rotation matrix are easily determined. All you have to do is to find the eigenvalues of M, and use the fact that, if λ is an eigenvalue of M, then $e^{\lambda\psi}$ is an eigenvalue of R. One of the eigenvalues of M is 0, corresponding to the eigenvalue $+1$ of R (as we have already found). The other two eigenvalues of M are ± 1, corresponding to the eigenvalues $e^{\pm i\psi}$ of R. (Check it out!) This result is in accord with a general result: all the eigenvalues of a real orthogonal matrix (more generally, a *unitary* matrix) must lie on the unit circle in the complex plane.

The parameter space of the rotation group

We come now to an important aspect of rotation matrices in 3 dimensions, i.e., the set of unimodular, real, orthogonal, (3×3) matrices. These matrices comprise a *group*, because they satisfy the properties required of the elements of a group:

(i) Each matrix in the set has an inverse.

(ii) There is an identity element in the set, the unit matrix I_3, corresponding to no rotation at all.

(iii) The product of two matrices in the set is again a matrix of the same kind: for, if R_1 and R_2 are unimodular orthogonal matrices, then

$$\det(R_1 R_2) = (\det R_1)(\det R_2) = 1. \tag{29}$$

Moreover,

$$(R_1 R_2)^{\mathrm{T}} = R_2^{\mathrm{T}} R_1^{\mathrm{T}} = R_2^{-1} R_1^{-1} = (R_1 R_2)^{-1}. \tag{30}$$

(iv) Matrix multiplication is, of course, associative:

$$R_1(R_2 R_3) = (R_1 R_2) R_3. \tag{31}$$

The set of matrices therefore constitutes a group, the special orthogonal group $SO(3)$.

The elements of the group are *parametrized* (that is, its elements are labelled) by a unit vector $\mathbf{n}$ and an angle ψ. Now, a unit vector has two independent parameters, since its components must satisfy the constraint $n_1^2 + n_2^2 + n_3^2 = 1$. Alternatively, we may specify the spherical polar coordinates of the (tip of) the unit vector $\mathbf{n}$, namely, $(1, \theta, \varphi)$, involving once again two independent parameters, θ and φ. The angle of rotation, ψ, is another independent parameter. Each $R(\mathbf{n}, \psi)$ is therefore specified by three independent parameters, making $SO(3)$ a 3-parameter group. The parameters are continuous variables, making it a *continuous group*. Finally, the group elements are *analytic* functions of the parameters, owing to the exponential form $R(\mathbf{n}, \psi) = e^{i J_k n_k \psi}$. $SO(3)$ is therefore a 3-parameter *Lie group*, generated by the three generators J_1, J_2, J_3 that satisfy the *Lie algebra* given by eq. (13). The natural question that arises now is: what sort of space is the 3-dimensional space spanned by these parameters?

All the group parameters of $SO(3)$ are bounded, being direction cosines and angles: for instance, each $|n_i| \leq 1$, and $0 \leq \psi < 2\pi$. The parameter space is therefore a *bounded* space. Imagine a solid sphere (in mathematical terminology, a *ball*), which is a 3-dimensional object. Let the centre of the ball represent the identity element of $SO(3)$. Any other point of the ball is specified by a direction and a distance from the origin. The direction can be specified by the unit vector **n**, while the value of the angle ψ can specify the distance from the origin. As the range of ψ is given by $0 \leq \psi < 2\pi$, you might expect that the parameter space of $SO(3)$ is the interior of a solid ball of radius 2π. Each point in this ball represents an element of $SO(3)$, i.e., a rotation in 3-dimensional space.

Surprisingly, this argument is incorrect! Refer to Figure 1 of ch. 5 once more. Observe the following:

(i) One half of the plane normal to the axis **n** is covered by rotations about **n** as ψ ranges from 0 to π.

(ii) The other half of this plane is covered by rotations about the *oppositely* directed axis $-\mathbf{n}$ as ψ ranges over the same interval, 0 to π.

(iii) Finally, substituting the value $\psi = \pi$ in the general formula (23), we find

$$R_{ij}(\mathbf{n}, \pi) = \delta_{ij} + 2n_i n_j = R_{ij}(-\mathbf{n}, \pi). \tag{32}$$

(iv) In other words, it suffices to let ψ run from 0 to π. Moreover, it is a *fundamental property of* 3*-dimensional space* that a rotation of the coordinate frame about any axis **n** through an angle π leads to the *same* final orientation of the frame as a rotation about the oppositely-directed axis through an angle π.

This seemingly trivial fact[§] has profound consequences. To start with, it has a drastic effect on the nature of the parameter space of $SO(3)$. This space is now a solid sphere (or ball) of radius π, rather than 2π. Further, every point on its surface is *identified* with, or mathematically the same as, its antipodal point! It is as if an invisible wire connected and 'short-circuited' every pair of antipodal points on the surface of the ball. The mathematical name for this space is the *real projective space* RP^3.

Figure 1 is a partial depiction of the space. It does not (cannot!) show that each point on the surface of the ball is identical to its antipodal point. It is not possible to make an actual model of such a space in the 3-dimensional space in which we live. That is, RP^3 cannot be *embedded* in 3-dimensional Euclidean space.[¶] In the next chapter, I will consider another way of representing rotations in 3-dimensional space, and show how it is related it to $SO(3)$ and its parameter space. I will also indicate the deep physical consequences of these mathematical

[§]Check it out using, say, a cube or a brick-shaped object.

[¶]You will need a Euclidean space of at least 5 dimensions to do so. Further elaboration of this point is well beyond the scope of the present discussion.

considerations, in order to highlight how important this topic is for our understanding of the physical universe.

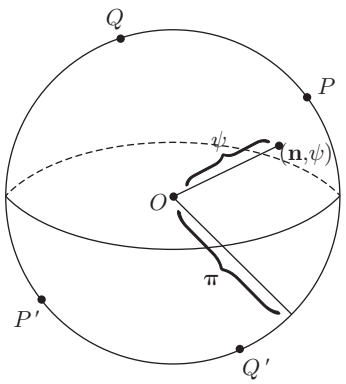


Figure 1: The ball of radius π representing the parameter space of $SO(3)$. The antipodal points P and P' on the surface of the ball represent a single point in the parameter space. Likewise, the antipodal points Q and Q' represent a single point in the space.

More fun with the rotation group

V. Balakrishnan

Introduction

In the preceding chapter, you have seen that the group of rotations of the coordinate frame in 3 dimensions can be represented by unimodular, orthogonal, (3×3) matrices, the Lie group $SO(3)$. At a deeper level, these matrices actually comprise the defining *representation* of the *abstract group* of rotations, which is isomorphic to, and conveniently denoted by, $SO(3)$. The abstract group of rotational transformations in 3 dimensions has many (in fact, an infinite number of) representations. Unimodular orthogonal (3×3) matrices comprise just one (albeit a very 'prominent' one) of these representations. In this chapter, I will introduce and discuss another, equally important, way of representing the rotation group. This representation involves another group of matrices: unimodular, *unitary* (2×2) matrices, comprising the group $SU(2)$. This group is intimately related to the notion of spin $\frac{1}{2}$ in quantum mechanics. (This is, as you know, the spin quantum number of electrons, protons, neutrons and quarks.) The relationship between $SU(2)$ and $SO(3)$, and that between their respective parameter spaces, is of fundamental importance in physics. It underlies integer and half-odd-integer spin quantum numbers, the existence of bosons and fermions, and so on. It is not my intention to consider these matters in detail here, or to go into group representations *per se*. What follows is a discussion of the essential aspects in simple terms.

Pauli matrices

When we write a general (2×2) matrix as

$$M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \tag{1}$$

what, exactly, do we mean? (2×2) matrices constitute a linear vector space. Regarded as an element of this space, eq. (1) stands for the expansion

$$M = a\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + b\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} + c\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} + d\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}. \tag{2}$$

The four matrices on the right-hand side constitute the so-called *natural basis* in the linear vector space (2×2) matrices. There is, however, another basis that is, in many ways, more convenient than the natural basis. This alternative basis consists of the (2×2) unit matrix I_2 and the three *Pauli matrices*, defined as

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},\ \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix},\ \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{3}$$

A general (2×2) matrix like M can then be written as a linear combination

$$M = \alpha_0 I_2 + \alpha_1 \sigma_1 + \alpha_2 \sigma_2 + \alpha_3 \sigma_3 \,. \tag{4}$$

The simplest way to establish this assertion is to compare (4) with (1). This yields a set of four equations expressing a, b, c, d as linear combinations of $\alpha_0, \alpha_1, \alpha_2, \alpha_3$. These are trivially solved to obtain unique solutions for the latter as linear combinations of the former.

The Pauli matrices σ_k $(k = 1, 2, 3)$ have some remarkable properties:

(i) σ_k is hermitian ($\sigma_k = \sigma_k^\dagger$).

(ii) $\mathrm{Tr}\, \sigma_k = 0$ (σ_k is traceless), $\det \sigma_k = -1$.

(ii) $\sigma_k^2 = I_2$, the (2×2) unit matrix. Hence $\sigma_k^{-1} = \sigma_k$.

(iii) σ_k has eigenvalues 1 and -1, with normalized eigenvectors $\binom{1}{0}$ and $\binom{0}{1}$.

(iv) $\sigma_k \sigma_l = i \epsilon_{klm} \sigma_l$.

(v) It follows from (v) that the anticommutator

$$[\sigma_k \,,\, \sigma_l]_+ = \sigma_k \sigma_l + \sigma_l \sigma_k = 0 \quad (k \neq l). \tag{5}$$

(vi) It also follows that the commutator

$$[\sigma_k \,,\, \sigma_l] = \sigma_k \sigma_l - \sigma_l \sigma_k = 2\, i\, \epsilon_{klm}\, \sigma_m. \tag{6}$$

Therefore, setting $J_k = \frac{1}{2} \sigma_k$, we have the *commutation relations*

$$[J_k \,,\, J_l] = i\, \epsilon_{klm}\, J_m. \tag{7}$$

But these commutation relations are exactly the relations satisfied by the infinitesimal generators of the rotation group in 3 three dimensions, as you have seen in ch. 6. (This is the reason I have used the same symbol, J_k, in both cases.) Previously, those generators were represented by three hermitian (3×3) matrices. Now, we find that the same 'angular momentum algebra' is satisfied by the three Pauli matrices multiplied by $\frac{1}{2}$. This suggests strongly that the matrices $\frac{1}{2}\sigma_k$ $(k = 1, 2, 3)$ also generate a representation of the rotation group $SO(3)$, this time in terms of (2×2) matrices. We will see, shortly, how this comes about.

Representing rotations by (2 × 2) matrices

It is helpful to recapitulate, in brief, what was established in ch. 6. Let the position vector of a point be written as a column vector whose elements are the Cartesian components of $\mathbf{r}$. Consider a proper rotation of the coordinate axes in 3-dimensional space, about an axis along

the unit vector $\mathbf{n}$, and through an angle ψ. Then, the effect of the rotation is to transform $\mathbf{r}$ to $\mathbf{r}' = R(\mathbf{n}, \psi)\,\mathbf{r}$, where $R(\mathbf{n}, \psi)$ is a (3×3) matrix. This matrix has the form $R(\mathbf{n}, \psi) = e^{i\,(\mathbf{J}\cdot\mathbf{n})\,\psi}$, where $\mathbf{J} = (J_1, J_2, J_3)$ is the set of generators of infinitesimal rotations about the three Cartesian axes. They are certain (3×3) hermitian matrices that satisfy the angular momentum algebra $[J_k\,,\,J_l] = i\,\epsilon_{klm}\,J_m$. Exponentiating them produces the rotation matrix $R(\mathbf{n}, \psi)$. Owing to the special form of the matrix $i(\mathbf{J}\cdot\mathbf{n}) = iJ_k\,n_k$, the exponential $e^{i\,(\mathbf{J}\cdot\mathbf{n})\,\psi}$ can be computed in closed form to yield $R(\mathbf{n}, \psi)$ explicitly as a function of the rotation parameters $\mathbf{n}$ and ψ. The rotation $R(\mathbf{n}, \psi)$ turns out to be a unimodular, orthogonal matrix. Its action on $\mathbf{r}$ gives the finite rotation formula for $\mathbf{r}'$. The same transformation rule under a rotation of the coordinate frame is valid for any vector, by definition. The set of rotation matrices $\{R(\mathbf{n}, \psi)\}$ constitutes the special orthogonal group $SO(3)$.

We have now found that the generators of infinitesimal rotations can also be represented by the matrices $\frac{1}{2}\sigma_k$. In ch. 6, we saw how a finite rotation could be built up by a sequence of infinitesimal rotations, leading to the exponentiation of the generators. This argument is independent of the actual representation of the generators. In analogy with the notation $\mathbf{J} = (J_1, J_2, J_3)$, let us define

$$\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \sigma_3). \tag{8}$$

The (2×2) matrix

$$U(\mathbf{n}, \psi) = e^{i\,(\boldsymbol{\sigma}\cdot\mathbf{n})\psi/2} \tag{9}$$

must then represent a rotation about $\mathbf{n}$ through the angle ψ. The symbol U has been used in order to distinguish this matrix from the (3×3) matrix $R(\mathbf{n}, \psi)$, and also to remind us of the kind of matrix it is, as will be seen shortly. The exponential on the right-hand side if eq. (9) can be computed even more easily than the way we simplified exponential $e^{i(\mathbf{J}\cdot\mathbf{n})\psi}$ in ch. 6. Note, first, that

$$(\boldsymbol{\sigma}\cdot\mathbf{n})^2 = \sigma_1^2 n_1^2 + \sigma_2^2 n_2^2 + \sigma_3^2 n_3^2 = I_2, \tag{10}$$

because of the properties (iii) and (vi) of the Pauli matrices together with the fact that $\mathbf{n}$ is a unit matrix. Therefore

$$(\boldsymbol{\sigma}\cdot\mathbf{n})^{2n} = I_2, \quad \text{while} \quad (\boldsymbol{\sigma}\cdot\mathbf{n})^{2n+1} = (\boldsymbol{\sigma}\cdot\mathbf{n}). \tag{11}$$

It follows that

$$\begin{aligned} U(\mathbf{n}, \psi) &= I_2 \left\{ 1 - \frac{(\frac{1}{2}\psi)^2}{2!} + \frac{(\frac{1}{2}\psi)^4}{4!} - \cdots \right\} + i\,(\boldsymbol{\sigma}\cdot\mathbf{n}) \left\{ \frac{(\frac{1}{2}\psi)}{1!} - \frac{(\frac{1}{2}\psi)^3}{3!} + \cdots \right\} \\ &= I_2\,\cos\left(\tfrac{1}{2}\psi\right) + i\,(\boldsymbol{\sigma}\cdot\mathbf{n})\,\sin\left(\tfrac{1}{2}\psi\right). \end{aligned} \tag{12}$$

This is *not* an orthogonal matrix! But the Pauli matrices are hermitian. Therefore

$$U^\dagger(\mathbf{n}, \psi) = I_2\,\cos\left(\frac{1}{2}\psi\right) - i\,(\boldsymbol{\sigma}\cdot\mathbf{n})\,\sin\left(\frac{1}{2}\psi\right)$$

$$= \quad e^{-i\,(\mathbf{n}\cdot\boldsymbol{\sigma})\,\psi/2} = U^{-1}(\mathbf{n},\psi). \tag{13}$$

Thus $UU^\dagger = U^\dagger U = I_2$. That is, $U(\mathbf{n},\psi)$ is a *unitary* matrix, rather than an orthogonal one. Writing out the expression in eq. (12) explicitly as a matrix, we get

$$U(\mathbf{n},\psi) = \begin{pmatrix} \cos\,\frac{1}{2}\psi + in_3\,\sin\,\frac{1}{2}\psi & (n_2 + in_1)\,\sin\,\frac{1}{2}\psi \\ (-n_2 + in_1)\,\sin\,\frac{1}{2}\psi & \cos\,\frac{1}{2}\psi - in_3\,\sin\,\frac{1}{2}\psi \end{pmatrix}. \tag{14}$$

Remembering that $n_1^2 + n_1^2 + n_1^2 = 1$, it is easy to check that

$$\det\, U(\mathbf{n},\psi) = 1. \tag{15}$$

Thus, $U(\mathbf{n},\psi)$ is also an unimodular matrix, just like $R(\mathbf{n},\psi)$. It is sometimes useful to write $U(\mathbf{n},\psi)$ in terms of the spherical polar coordinates $(1,\theta,\varphi)$ of the unit vector $\mathbf{n}$. Equation (14) becomes

$$U(\mathbf{n},\psi) = \begin{pmatrix} \cos\,\frac{1}{2}\psi + i\cos\,\theta\,\sin\,\frac{1}{2}\psi & i\,e^{-i\varphi}\sin\,\theta\,\sin\,\frac{1}{2}\psi \\ i\,e^{i\varphi}\sin\,\theta\,\sin\,\frac{1}{2}\psi & \cos\,\frac{1}{2}\psi - i\cos\,\theta\,\sin\,\frac{1}{2}\psi \end{pmatrix}. \tag{16}$$

We have now found a possible representation of rotations in 3-dimensional space by a (2×2) unimodular, unitary matrices. But there remain several basic questions. How can a (2×2) matrix like U act on 3-component vectors? How general is the foregoing representation of rotations in 3-dimensional space? Is (14) the most general kind of (2×2) unimodular, unitary matrix? Does every matrix of this kind represent a rotation in 3-dimensional space? Conversely, is every rotation represented by a unique matrix of this kind? Let us now take up these questions one by one.

Action of $U(\mathbf{n},\psi)$ on a 3-vector

We must find the rule according to which $U(\mathbf{n},\psi)$ acts on the coordinates (x_1,x_2,x_3) of any point $\mathbf{r}$ in 3-dimensional space. This rule must correctly reproduce the finite rotation formula for $\mathbf{r}' = R(\mathbf{n},\psi)\,\mathbf{r}$ derived in ch. 6, namely,

$$\mathbf{r}' = (\cos\,\psi)\,\mathbf{r} + (1-\cos\,\psi)\,(\mathbf{r}\cdot\mathbf{n})\,\mathbf{n} + (\sin\,\psi)\,\mathbf{r}\times\mathbf{n}. \tag{17}$$

Only then can we assert that $U(\mathbf{n},\psi)$ does indeed represent a rotation in 3-dimensional space.

Here is how this is done. Instead of representing the Cartesian coordinates of any point $\mathbf{r}$ as a column vector, consider the (2×2) matrix

$$\mathbf{r}\cdot\boldsymbol{\sigma} = x_1\,\sigma_1 + x_2\,\sigma_2 + x_3\,\sigma_3 = \begin{pmatrix} x_3 & x_1 - ix_2 \\ x_1 + ix_2 & -x_3 \end{pmatrix}. \tag{18}$$

The action of $U(\mathbf{n},\psi)$ is then as follows. If $\mathbf{r}\mapsto\mathbf{r}'$ under the rotation, the transformation rule for $\mathbf{r}\cdot\boldsymbol{\sigma}$ is

$$\mathbf{r}\cdot\boldsymbol{\sigma} \mapsto \mathbf{r}'\cdot\boldsymbol{\sigma} = U(\mathbf{n},\psi)\,(\mathbf{r}\cdot\boldsymbol{\sigma})U^\dagger(\mathbf{n},\psi). \tag{19}$$

Verifying that this transformation law leads to the finite rotation formula requires a bit of algebra. I will outline the steps, leaving it to the reader to fill in the rest. Use the last equation in (12) for U and the first equation in (13) for $U^\dagger$ in the right-hand side of eq. (19), and simplify the resulting expression. Doing this component by component is tedious. It is helpful to use, instead, a couple of identities involving the Pauli matrices. (Again, I leave it to the reader to derive these relations from the properties of Pauli matrices listed earlier.) Let **a** and **b** denote ordinary 3-dimensional vectors. Then

$$(\mathbf{a}\cdot\boldsymbol{\sigma})\,(\mathbf{b}\cdot\boldsymbol{\sigma}) \;=\; (\mathbf{a}\cdot\mathbf{b})\,I_2 + i\,(\mathbf{a}\times)\cdot\boldsymbol{\sigma} \tag{20}$$

$$[(\mathbf{a}\cdot\boldsymbol{\sigma})\,,(\cdot\boldsymbol{\sigma})] \;=\; 2i\,(\mathbf{a}\times)\cdot\boldsymbol{\sigma} \tag{21}$$

After simplification, we arrive at the result

$$\begin{aligned}\mathbf{r}'\cdot\boldsymbol{\sigma} \;=\;& (\cos\psi)\,(\mathbf{r}\cdot\boldsymbol{\sigma}) + (1-\cos\psi)\,(\mathbf{r}\cdot\mathbf{n})\,(\mathbf{n}\cdot\boldsymbol{\sigma}) \\ &+(\sin\psi)\,(\mathbf{r}\times\mathbf{n})\cdot\boldsymbol{\sigma}\end{aligned} \tag{22}$$

This is precisely the relation obtained if we take the dot product of both sides of the finite rotation formula (17) with $\boldsymbol{\sigma}$. Equation (22) also holds good, of course, if **r** is replaced by any vector **A**. We conclude that any rotation of the coordinate frame about the origin in 3-dimensional space can be represented by a matrix like $U(\mathbf{n},\psi)$.

$U(\mathbf{n},\psi)$ is the most general matrix of its kind

Recall that $R(\mathbf{n},\psi)$ is a unimodular, orthogonal matrix with real elements. It involves 3 parameters, the unit vector **n** and the angle ψ. This is the number of parameters required to specify any rotation of the coordinate frame in 3 dimensions.

The (2×2) matrix $U(\mathbf{n},\psi)$ also involves the same three parameters. It is unimodular, but not orthogonal. Its elements are complex numbers, and it is a unitary matrix. The question is whether it is the most general matrix with those properties.

To answer this question, you must start with a general (2×2) matrix with complex elements, say

$$W = \begin{pmatrix}\alpha & \beta\\ \gamma & \delta\end{pmatrix}. \tag{23}$$

Since each of the elements is a complex number in general, W has 8 independent real parameters. The hermitian conjugate of W is

$$W^\dagger = \begin{pmatrix}\alpha^* & \gamma^*\\ \beta^* & \delta^*\end{pmatrix}. \tag{24}$$

Imposing the unitarity condition $WW^\dagger = I_2$ yields the four relations

$$\left.\begin{aligned} |\alpha|^2 + |\beta|^2 &= 1, \\ |\gamma|^2 + |\delta|^2 &= 1, \\ \alpha\,\gamma^* + \beta\,\delta^* &= 0, \\ \alpha^*\,\gamma + \beta^*\,\delta &= 0. \end{aligned}\right\} \tag{25}$$

Imposing the condition of unimodularity gives one more relation,

$$\det\, W = \alpha\delta - \beta\gamma = 1. \tag{26}$$

Hence there are 5 relations between the 8 parameters in W, leaving 3 independent parameters. This is precisely the number of parameters possessed by an element of the rotation group in 3 dimensions. You can use eqs (25) and (26) to eliminate γ and δ in favour of α and β. The result is the most general form of a (2×2) unimodular, unitary matrix. I will now call it U instead of W. It is given by

$$U = \begin{pmatrix} \alpha & \beta \\ -\beta^* & \alpha^* \end{pmatrix}, \quad \text{where } |\alpha|^2 + |\beta|^2 = 1. \tag{27}$$

Writing it out even more explicitly in terms of the real and imaginary parts of $\alpha = \alpha_1 + i\alpha_2$ and $\beta = \beta_1 + i\beta_2$, we have, finally,

$$U = \begin{pmatrix} \alpha_1 + i\alpha_2 & \beta_1 + i\beta_2 \\ -\beta_1 + i\beta_2 & \alpha_1 + i\alpha_2 \end{pmatrix}, \tag{28}$$

where

$$\alpha_1^2 + \alpha_2^2 + \beta_1^2 + \beta_2^2 = 1. \tag{29}$$

But the general rotation matrix $U(\mathbf{n}, \psi)$ in Eq, (14) is *precisely* of this form! We conclude that every (2×2) unimodular, unitary matrix represents a rotation of the coordinate frame in 3-dimensional space. This is the converse of the conclusion stated at the end of the preceding section.

The three rotation generators themselves form a vector

I have used the boldface symbol $\mathbf{J}$ for the triad (J_1, J_2, J_3) of (3×3) matrices generating the rotation matrix $R(\mathbf{n}, \psi)$, as if J_k itself was a component of a vector. Likewise, I have used $\boldsymbol{\sigma}$ for the triad of Pauli matrices $(\sigma_1, \sigma_2, \sigma_3)$ which generates (when multiplied by $\frac{1}{2}$) the rotation matrix $U(\mathbf{n}, \psi)$. I stated in ch. 6 that this was more than mere convenient notation. It is now time to explain why the notation is justified.

But first, a word regarding representations (of the elements of a group, or of the generators of a continuous group). We have seen that the group of rotations in 3-dimensions can be generated by the matrices (J_1, J_2, J_3) defined in ch. 6, as well as by the matrices $(\frac{1}{2}\sigma_1, \frac{1}{2}\sigma_2, \frac{1}{2}\sigma_3)$

introduced in this chapter. The common feature is that both sets of matrices satisfy the same *algebra*, the angular momentum algebra. In general, the generators of a group of transformations, and hence its elements, can have an infinite number of different representations. Which representation one uses depends on the nature of the objects on which the transformations act. The representation can even be infinite-dimensional: this is required, for example, when we consider transformations of the state vector of a quantum mechanical system in an infinite-dimensional Hilbert space.

Returning to the case of interest to us here, the rotation group $SO(3)$, let us (for simplicity of notation) denote its generators by (J_1, J_2, J_3) and elements by $U(\mathbf{n}, \psi)$, *regardless of any particular representation*. Consider, for instance, a quantum mechanical system. A rotation of the coordinate axes about the direction $\mathbf{n}$ through an angle ψ then *induces* a *unitary* transformation $U(\mathbf{n}, \psi) = e^{iJ_k n_k \psi}$ that acts on the state vector of the system. In general, J_k (and hence U) are *operators* that act on the vectors in the Hilbert space of the system. The dimensionality of the representation used for J_k (and hence U) depends on the dimensionality of the Hilbert space of the system.

We need to show that $\mathbf{J}$ is a vector under rotations of the coordinate axes, *independent* of any particular representation for J_1, J_2 and J_3. This means that the only input we can use is the *algebra* satisfied by the three operators, namely, $[J_k, J_l] = i\,\epsilon_{klm}\,J_m$. Moreover, $\mathbf{J}$ is not an 'ordinary' vector. It is an *operator-valued* vector, i.e., each J_k is an operator. The transformation law applicable to it is not the direct counterpart of $\mathbf{r}' = R(\mathbf{n}, \psi)\,\mathbf{r}$, namely, $\mathbf{J}' = R(\mathbf{n}, \psi)\,\mathbf{J}$. The transformation rule (under a rotation of the spatial coordinate frame) for any operator W is

$$\begin{aligned} W \mapsto W' &= U^\dagger(\mathbf{n}, \psi)\, W\, U(\mathbf{n}, \psi) \\ &= e^{-i\,(\mathbf{J}\cdot\mathbf{n})\,\psi}\, W\, e^{i\,(\mathbf{J}\cdot\mathbf{n})\,\psi}. \end{aligned} \tag{30}$$

Applying this rule to the operator J_k, we have

$$\begin{aligned} J_k' &= e^{-i\,(\mathbf{J}\cdot\mathbf{n})\,\psi}\, J_k\, e^{i\,(\mathbf{J}\cdot\mathbf{n})\,\psi} \\ &= e^{-i\,J_l n_l \psi}\, J_k\, e^{i\,J_m n_m \psi}. \end{aligned} \tag{31}$$

We must now simplify the expression on the right-hand side of (31), paying attention to the fact that J_k and J_l do not commute with each other when $k \neq l$. At first sight, this appears to be a formidable task. But an important operator identity called *Hadamard's Lemma* comes to our aid. Let X and Y be two operators (acting on the same space, and with the same domain and range), and let $[X, Y] = C_1$. Define the nested commutators

$$C_{r+1} = [X, C_r], \quad r = 1, 2, \ldots . \tag{32}$$

Hadamard's Lemma then states that, if λ is any constant,

$$e^{\lambda X}\, Y\, e^{-\lambda X} = Y + \sum_{r=1}^{\infty} \frac{C_r\, \lambda^r}{r!}\,. \tag{33}$$

The proof of this lemma is not difficult, but I will not give it here. Apply the lemma to the final expression in (31), with the identifications

$$\lambda = -i\psi,\ X = J_l n_l = J_m n_m,\ Y = J_k. \tag{34}$$

Now use the generator algebra $[J_k\,,\ J_l] = i\,\epsilon_{klm}\,J_m$ repeatedly, to show that

$$C_1 = C_3 = \cdots = i\,\epsilon_{klm}\,J_m, \tag{35}$$

and

$$C_2 = C_4 = \cdots = J_k - n_k\,n_l\,J_l. \tag{36}$$

You can now sum the infinite series in Hadamard's lemma easily. The result is

$$J_k' = (\cos\,\psi)\,J_k + (1 - \cos\,\psi)\,(J_l\,n_l)\,n_k + (\sin\,\psi)\,\epsilon_{klm}\,J_l\,n_m. \tag{37}$$

In vector form, this is

$$\mathbf{J}' = (\cos\,\psi)\,\mathbf{J} + (1 - \cos\,\psi)\,(\mathbf{J}\cdot\mathbf{n}) + (\sin\,\psi)\,\mathbf{J}\times\mathbf{n}. \tag{38}$$

But this is identical in form to the finite rotation formula for the coordinate $\mathbf{r}$ of a point in three-dimensional space, eq. (17). This proves the assertion that the generators of infinitesimal rotations in 3-dimensional space themselves constitute (equivalently, transform like) a vector. In short, denoting (J_1, J_2, J_3) by $\mathbf{J}$ is more than mere notation.

The parameter spaces of $SU(2)$ and $SO(3)$

We have seen that the set of (3×3) unimodular, orthogonal matrices with real elements constitutes a 3-parameter Lie group, the special orthogonal group $SO(3)$. Similarly, the set of (2×2) unimodular, unitary matrices with complex elements constitutes a 3-parameter group, the special unitary group $SU(2)$. Since the dependence on the parameters is analytic in nature, $SU(2)$ is also a Lie group. Its general element is a matrix of the form given by eqs (28) and (29). Alternatively, in the context of the (2×2) matrix representation of the group of rotations in 3-dimensional space, the general element of $SU(2)$ is a matrix of the form $U(\mathbf{n}, \psi)$ as given by eq. (14).

Consider, now, the parameter space of $SU(2)$. Equation (29) is one constraint among the four parameters $\alpha_1, \alpha_2, \beta_1$ and β_2. As you know, the equation $x_1^2 + x_2^2 = 1$ specifies the circumference of a unit circle in a 2-dimensional plane, while $x_1^2 + x_1^2 + x_1^2 = 1$ is the equation to the surface of a unit sphere in 3-dimensional space. By analogy, eq. (29) is the 'surface' of a ball of unit radius in 4-dimensional Euclidean space. It is a 3-dimensional space called the 3-sphere, denoted by S^3. In mathematics, S^d denotes the d-dimensional sphere or d-sphere. It is the d-dimensional 'surface' of a ball of unit radius in $(d + 1)$-dimensional Euclidean space.*

*Thus, the circumference of a unit circle in a plane is the 1-sphere S^1. The surface of an ordinary ball of unit radius in our familiar 3-dimensional space is the 2-sphere S^2. (The unit radius condition is not necessary as far as topology is concerned.) The embedding of these spaces in a higher-dimensional Euclidean space helps us visualise these spaces, but it is not necessary. The spaces can be studied in their own right.

The space S^d (for any $d \geq 1$) is *connected*: that is, you can start at any point in the space and trace a continuous path to any other point in the space without leaving the space. The space S^d (for any $d \geq 2$) is *simply connected*. That is, any closed path or loop lying entirely in the space can be shrunk continuously to a point without leaving the space.** This property of simple connectedness makes the space topologically simple in a certain sense.

Let us turn to the real orthogonal group $SO(3)$. We have seen in ch. 6 that the parameter space of $SO(3)$ is the real projective space RP^3. This is a ball of radius π with antipodal points on its surface identified: that is, every pair of antipodal points on the surface of the ball is actually a single point in the space. (Recall Figure 1, ch. 6.) This fact complicates the topology of the parameter space. It can be shown that the RP^3 is connected, but not simply connected. It is *doubly connected*. The proof will not be given here, but a heuristic way of seeing this is as follows. It is clear that one can trace a loop that lies entirely inside the ball, and that every such loop can be shrunk continuously to a point. On the other hand, consider the following path (refer again to Figure 1, ch. 6). Start at the origin O of the ball and move out along a continuous line to the point P on its surface. Re-enter the ball from the antipode P' of the point P (remember that P and P' are supposed to be a single point of the parameter space). Move inward along a continuous line from P' to O. The result looks like a directed line from P' to its diametrically opposite point P, passing through the centre O of the ball. But this is, in fact, a *closed* path in the parameter space! If you attempt to move P into the ball, the loop breaks. If you attempt to move it on the surface towards P' in an attempt to 'close' the path and shrink it, P' moves away, always remaining antipodal to the position of P. The only way to shrink the closed path $O \to P \to P' \to O$ is to trace out another such loop, say $O \to Q \to Q' \to O$, and then performing suitable deformations of the pair of closed paths. Further details would take us too far afield.

How are the groups $SO(3)$ and $SU(2)$ related?

Finally, we turn our attention to the relationship between the special orthogonal group $SO(3)$ and the special unitary group $SU(2)$. Each of these matrix groups provides a representation of the group of rotation transformations in 3-dimensional space.

Let us go back to eq. (19), which tells us that the effect of a rotation of the coordinate frame on any vector $\mathbf{A}$ in 3 dimensional space is to transform it to another vector $\mathbf{A}'$, such that

$$\mathbf{A}' \cdot \boldsymbol{\sigma} = U(\mathbf{n}, \psi)\, (\mathbf{A} \cdot \boldsymbol{\sigma})\, U^{\dagger}(\mathbf{n}, \psi). \tag{39}$$

Here $U(\mathbf{n}, \psi)$ is an element of $SU(2)$, i.e., it is a (2×2) unimodular, unitary matrix. But the form of eq. (39) shows that we have some freedom in choosing the transformation matrix $U(\mathbf{n}, \psi)$. Suppose we replace the matrix $U(\mathbf{n}, \psi)$ with the matrix $e^{i\gamma}\, U(\mathbf{n}, \psi)$, where γ is a real constant. Then $U^{\dagger}(\mathbf{n}, \psi)$ gets replaced by $e^{-i\gamma}\, U^{\dagger}(\mathbf{n}, \psi)$. It is obvious that $(\mathbf{A}' \cdot \boldsymbol{\sigma})$, and hence $\mathbf{A}'$, remains unaffected.

**Interestingly enough, S^1 is not simply connected. In fact, its connectivity is infinite, in a certain specific sense. I will not digress into this aspect here.

But we have found that any (2×2) matrix that represents a rotation in 3 dimensions must be unitary. Since U is already unimodular, we must have

$$\det \left[e^{i\gamma} U(\mathbf{n}, \psi) \right] = e^{2i\gamma} \det U = e^{2i\gamma} = 1. \tag{40}$$

Therefore $e^{i\gamma} = \pm 1$. The implication is that there are *two* different matrices, $U(\mathbf{n}, \psi)$ and $-U(\mathbf{n}, \psi)$, both of them belonging to $SU(2)$, corresponding to every rotation matrix $R(\mathbf{n}, \psi)$ belonging to $SO(3)$. Several important consequences follow, which I will merely state. What follows should also help you to become familiar with some mathematical terminology, as it is introduced here in the setting of the specific example of the rotation group.

(i) For every set of parameters $\{\mathbf{n}, \psi\}$, the pair of $SU(2)$ elements $\{U(\mathbf{n}, \psi), -U(\mathbf{n}, \psi)\}$ is mapped to the $SO(3)$ element $R(\mathbf{n}, \psi)$. The pair of $SU(2)$ elements $\{I_2, -I_2\}$ is mapped to I_3, the identity element of $SO(3)$.

(ii) There is a 2-to-1 mapping or *homomorphism* of the group $SU(2)$ to the group $SO(3)$.

(iii) The set whose general element is the *pair* $\{U(\mathbf{n}, \psi),\ -U(\mathbf{n}, \psi)\}$ constitutes a group called the *quotient group* $SU(2)/\mathbb{Z}_2$.[‡]

(iv) There is a 1-to-1 mapping or *isomorphism* of this quotient group to $SO(3)$, which is written as

$$SU(2)/\mathbb{Z}_2 \sim SO(3). \tag{41}$$

In the local neighbourhood of their respective identity elements, the groups $SU(2)$ and $SO(3)$ 'look alike': their generators satisfy the same angular momentum algebra. In technical terms, they have the same Lie algebra of generators. But globally, the two groups are quite different from each other. The topologies of their parameters spaces are different. $SU(2)$, with its simply-connected parameter space, is the so-called *universal covering group* of $SO(3)$.

The mathematical properties that I have touched upon here have a bearing on several fundamental physical phenomena. Among these are the quantization of angular momentum, integer and half-odd-integer spin quantum numbers, and the existence of bosons and fermions that satisfy (owing to the spin-statistics theorem of quantum field theory) Bose–Einstein and Fermi–Dirac statistics, respectively. Is it not remarkable that all this, and more, originates in the study of rotations of the coordinate axes in 3-dimensional space?

[‡]This is read as "$SU(2)$ modulo $\mathbb{Z}_2$".

The many *avatars* of the Dirac delta function*

V. Balakrishnan

A filter for continuous functions

Suppose $f(x)$ is a function that is defined, say, for all values of the real variable x, and that it is finite everywhere. Can we construct some sort of filter or 'selector' that, when operating on this function, *singles out* the value of the function at any prescribed point x_0? In other words, how can we filter out just the quantity $f(x_0)$?

A hint is provided by the discrete analogue of this question. Suppose we have a sequence $(a_1, a_2, \dots) = \{a_j | j = 1, 2, \dots\}$. How do we select a particular member a_i from the sequence? By summing over all members of (i.e., scanning!) the sequence with a selector called the Kronecker delta, denoted by δ_{ij} and defined as

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j. \end{cases} \tag{1}$$

It follows immediately that

$$\sum_{j=1} \delta_{ij}\, a_j = a_i\,. \tag{2}$$

Further, we have the 'normalization' $\sum_j \delta_{ij} = 1$ for each value of i, and also the symmetry property $\delta_{ij} = \delta_{ji}$.

Reverting to the continuous case, we must replace the summation over j by an integration over x. The role of the specified index i is played by the specified point x_0. The analogue of the Kronecker delta is written like a function, retaining the same symbol δ for it. (Presumably, this was Dirac's reason for choosing this notation for the delta function.) So we seek a 'function' $\delta(x - x_0)$ such that

$$\int_{-\infty}^{\infty} dx\, \delta(x - x_0)\, f(x) = f(x_0)\,. \tag{3}$$

Exactly as in the discrete case of the Kronecker delta, we impose the normalization and symmetry properties

$$\int_{-\infty}^{\infty} dx\, \delta(x - x_0) = 1 \quad \text{and} \quad \delta(x - x_0) = \delta(x_0 - x)\,. \tag{4}$$

The requirements in eqs (3) and (4) may be taken to define the Dirac delta function. The form of eq. (3) suggests that $\delta(x - x_0)$ is more like the kernel of an integral *operator* than a conventional function. We will return to this aspect subsequently.

*Based on *Resonance*, Vol.**8**, No.8, pp.48-58, 2004.

What can $\delta(x - x_0)$ possibly look like? A naive way of answering this question is as follows. Take a rectangular window of width 2ε and height $1/(2\varepsilon)$, so that the area of the window is unity. Place it with its bottom edge on the x axis and slide it along this axis, as shown in Figure 1.

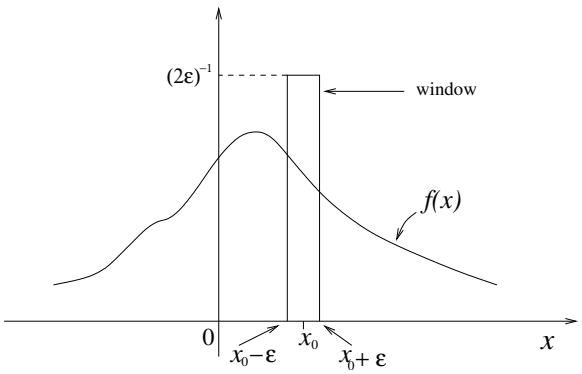


Figure 1 The rectangular window tends to $\delta(x - x_0)$ as $\varepsilon \to 0$.

When the window is centred at the chosen point x_0, the integral of $f(x)$ multiplied by this window function is simply $(1/2\varepsilon)\int_{x_0-\varepsilon}^{x_0+\varepsilon} dx\, f(x)$. This does not quite select $f(x_0)$ alone, of course. But it will do so if we take the limit $\varepsilon \to 0$. In this limit, the width of the window becomes vanishingly small. Simultaneously, its height becomes arbitrarily large, so as to capture all of the ordinate in the graph of $f(x)$, no matter how large the value of $f(x_0)$ is. A possible explicit form for the Dirac delta function $\delta(x - x_0)$ is therefore given by

$$\delta(x - x_0) = \begin{cases} \lim\limits_{\varepsilon\to 0} 1/(2\varepsilon)\,, & \text{for } x_0 - \varepsilon < x < x_0 + \varepsilon \\ 0\,, & \text{for all other } x\,. \end{cases} \tag{5}$$

This cannot be a stand-alone definition. If it is taken literally, then, formally, $\delta(x - x_0)$ must be zero for all $x \neq x_0$, while it must be infinite for $x = x_0$. An explicit form such as eq. (5) for the delta function must be interpreted in the light of eq. (3). The delta function is always to be understood as something that makes sense when it occurs in an integral like eq. (3), i.e., when it acts on ordinary functions like $f(x)$ and an integration is carried out. It is immediately clear that the so-called Dirac delta function cannot be a function in the conventional sense. In particular, $\delta(x - x_0)$ must be *singular* (formally infinite) at $x = x_0$, that is, at the point where its argument is zero.

Mathematically, an explicit form for the Dirac delta function is properly given in terms of a *sequence* or family of conventional functions, rather than the window representation in eq. (5). It can then be arranged that, in a suitable limit, the sequence approaches a quantity that has all the properties desired of the delta function. An infinite number of such sequences may be constructed. For instance, take a family of functions $\phi_\varepsilon(x - x_0)$ parametrized by a positive constant ε, and with the following properties: each member of the family

(i) has a peak at x_0,

(ii) is symmetric about the point x_0, and

(iii) has a definite integral from $-\infty$ to ∞ whose value is unity.

Matters are arranged such that, as the parameter ε is made smaller and smaller, the height of the peak in $\phi_\varepsilon(x)$ increases while its width simultaneously decreases, always keeping the total area under the curve equal to unity. Then $\lim_{\varepsilon\to 0} \phi_\varepsilon(x - x_0)$ represents the delta function $\delta(x - x_0)$. Let us now write down the simplest choices for such sequences. For ease of writing, let us set $x_0 = 0$. One of the simplest possibilities is the family of *Lorentzians*, given by

$$\phi_\varepsilon(x) = \frac{\varepsilon}{\pi(x^2 + \varepsilon^2)} \,. \tag{6}$$

Then $\lim_{\varepsilon\to 0} \phi_\varepsilon(x)$ is a representation of the Dirac delta function $\delta(x)$, with the properties specified in eqs (3) and (4). Some other popular choices for $\phi_\varepsilon(x)$ are the following:

$$\frac{e^{-|x|/\varepsilon}}{2\varepsilon} \,, \quad \frac{e^{-x^2/4\varepsilon}}{2\sqrt{\pi\varepsilon}} \,, \quad \frac{\mathrm{sech}^2(x/\varepsilon)}{2\varepsilon} \,, \quad \frac{\sin(x/\varepsilon)}{\pi x} \,, \cdots \tag{7}$$

It is instructive to sketch these functions schematically, and to check out what happens as smaller and smaller values of ε are chosen. As an amusing exercise, think up your own sequence of functions that leads to the delta function as a limiting case.

What is the point of all this? Before going on to answer this question, it is helpful to re-write the last of the functions in (7) as follows. Putting $\varepsilon = 1/L$, we get

$$\begin{aligned} \delta(x) &= \lim_{L\to\infty} \frac{\sin(L\,x)}{\pi\,x} = \frac{1}{2\pi} \lim_{L\to\infty} \int_{-L}^{L} dk\, e^{ikx} \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} dk\, e^{ikx}. \end{aligned} \tag{8}$$

This turns out to be perhaps the most useful way of representing the delta function. Since $|e^{ikx}| = 1$, it is obvious that the last integral in eq. (8) is not absolutely convergent. Nor is the integral well-defined in the ordinary sense, because $\sin kx$ and $\cos kx$ do not have definite limits as $k \to \pm\infty$. These are just further reminders of the fact that the delta function is not a conventional function, as I have already emphasized. If you are familiar with Fourier transforms, you will recognize that the last equation above seems to suggest that the Fourier transform of the Dirac delta function is just unity. This is indeed so. It suggests, too, that one way of *defining* singular functions like the delta function might be via their Fourier transforms: for example, we could define $\delta(x)$ as the inverse Fourier transform of a constant – in this case, just unity. The Fourier *avatar* of the Dirac delta function (and its higher-dimensional generalizations) is one of its most familiar manifestations.

A brief history of the δ-function

The Dirac delta function has quite a fascinating history. The book by Lützen, cited in the Bibliography, is an excellent source of information. The delta function seems to have made its first appearance in the early part of the 19th century, in the works of Poisson (1815), Fourier (1822), and Cauchy (1823, 1827). Poisson and Cauchy essentially used arguments that implied that the Lorentzian representation of the delta function, namely, eq. (6), had the selector property stated in eq. (3). Fourier, in his fundamental work *Théorie Analytique de la Chaleur*, showed (in connection with Fourier series expansions of periodic functions) that the series

$$\frac{1}{2\pi} + \frac{1}{\pi}\sum_{n=1}^{\infty} \cos\, n(x - x_0)$$

had precisely this sort of selector property, i.e., it was a representation of $\delta(x - x_0)$ in the fundamental interval $(x - x_0) \in [-\pi, \pi]$. His arguments essentially amount to the last of the representations in (7) for $\delta(x)$. These early works did not aim at mathematical rigour in the current sense of the term. Subsequently, Kirchhoff (1882, 1891) and Heaviside (1893, 1899) gave the first mathematical definitions (again non-rigorous, by modern standards) of the delta function. Kirchhoff was concerned with the fundamental solution of the three-dimensional wave equation, while Heaviside introduced the function in his *Operational Calculus*. He pointed out that $\delta(x)$ could be regarded as the derivative of the Heaviside or unit step function $\theta(x)$, defined as

$$\theta(x) = \begin{cases} 1, & \text{for } \; x > 0 \\ 0, & \text{for } \; x < 0. \end{cases} \tag{9}$$

After Heaviside, the delta function was freely used – in particular, in connection with Laplace transforms, especially by electrical engineers (e.g., Van der Pol, 1928). Dirac (1926, 1930) introduced it in his classic and fundamental work on quantum mechanics, essentially as the continuous analogue of the Kronecker delta.** He also wrote down a list of its important properties – much the same list that standard textbooks now carry. Over and above eqs (3) and (4), the delta function has the properties

$$\left.\begin{array}{rclcrcl} x\,\delta(x) & = & 0, & \quad \delta'(-x) & = & -\delta'(x), \\ x\,\delta'(x) & = & -\delta(x), & \quad \delta(ax) & = & (1/|a|)\,\delta(x) \end{array}\right\} \tag{10}$$

where a is any real number, and so on. Again, these equations are to be understood as valid when multiplied by suitable smooth functions and integrated over x. Dirac also listed the useful but not immediately obvious property

P. A. M. Dirac, *The physical interpretation of the* [sic] *quantum mechanics*, *Proc. Roy. Soc.* A **113, 621-641 (1926). So heady was the progress in that incredible period marking the birth of the subject that the definite article preceding 'quantum mechanics' was presumably no longer needed by 1930, as the title of the book P.A.M. Dirac, *The Principles of Quantum Mechanics*, Oxford University Press, 1930 (original edition) shows!

$$\delta\left(x^2 - x_0^2\right) = \frac{\delta(x + x_0) + \delta(x - x_0)}{2|x_0|} \tag{11}$$

where x_0 is any real number.

The use of the delta function became more and more common after the appearance of Dirac's work. Other singular functions also made their appearance, as early versions of quantum field theory began to take shape in the works of physicists such as Jordan, Pauli and Heisenberg. Around the same time, mathematicians began attempts to define such singular quantities in a rigorous manner. The delta function and other such singular objects were recognized to be what are called *generalized functions* or *distributions*, rather than functions in the conventional sense. The first rigorous theory was given by Bochner in 1932. Soon afterwards, Sobolev (1935) gave the rigorous definition of distributions as *functionals*, and the way had been paved for a definitive mathematical theory. This was achieved by Schwartz (1945-50), and comprehensively treated in his *Théorie des Distributions*, Vol. 1 (1950) and Vol. 2 (1951). For lack of space, we will not go further into these aspects, other than to repeat that we now have a completely rigorous mathematical theory of distributions.‡

Why does the δ-function appear in physical problems?

We can now turn to the question of why the delta function appears so naturally in physical problems. Consider, for example, the basic problem of electrostatics: given a static charge density $\rho(\mathbf{r})$ in free space, what is the corresponding electrostatic potential $\phi(\mathbf{r})$ at any arbitrary point $\mathbf{r} = (x, y, z)$? From Maxwell's equations, we know that ϕ satisfies Poisson's equation, namely,

$$\nabla^2\phi(\mathbf{r}) = -\rho(\mathbf{r})/\epsilon_0\,, \tag{12}$$

where ϵ_0 is the permittivity of the vacuum. What does one do in the case of a *point* charge q located at some point $\mathbf{r}_0 = (x_0, y_0, z_0)$? A point charge is an idealization in which a *finite* amount of charge q is supposed to be packed into *zero* volume. The charge density must therefore be infinite at the point $\mathbf{r}_0$, and zero elsewhere. The delta function comes to our aid. We may write, in this case,

$$\rho(\mathbf{r}) = q\,\delta(x - x_0)\,\delta(y - y_0)\,\delta(z - z_0) \equiv q\,\delta^{(3)}(\mathbf{r} - \mathbf{r}_0), \tag{13}$$

where the *three-dimensional* delta function $\delta^{(3)}$ is short-hand for the product of the three delta functions in the equation above. It is easy to verify that this expression for $\rho(\mathbf{r})$ has all the properties required of a point charge at the point $\mathbf{r}_0$. This illustrates how (and why) the delta function frequently appears on the right-hand side of fundamental equations in mathematical

‡A very accessible source of information is M. J. Lighthill, *Introduction to Fourier Analysis and Generalized Functions* (see the Bibliography). Originally published in 1958, this little classic has been reprinted several times. The book's dedication is as succinct as its text, and says: "To Paul Dirac, who saw that it must be true, Laurent Schwartz, who proved it, and George Temple, who showed how simple it could be made".

physics. It turns out that it also appears as the singular part of fundamental *solutions* to basic equations such as the wave equation.

It is worth noting that representations of higher-dimensional delta functions like $\delta^{(3)}$ are easily written down. For instance, the three-dimensional counterpart of eq. (8) above is just

$$\begin{aligned} \delta^{(3)}(\mathbf{r}) &= \frac{1}{(2\pi)^3}\int_{-\infty}^{\infty} dk_1 \int_{-\infty}^{\infty} dk_2 \int_{-\infty}^{\infty} dk_3\, e^{i\,(k_1 x + k_2 y + k_3 z)} \\ &\equiv \frac{1}{(2\pi)^3}\int d^3\mathbf{k}\, e^{i\,\mathbf{k}\cdot\mathbf{r}}. \end{aligned} \tag{14}$$

The notation used in the final equation is self-explanatory.

Why did Dirac need the delta function?

The delta function appeared in Dirac's work on quantum mechanics in an *avatar* somewhat different from the ones mentioned in the foregoing.

Recall the notions of a linear vector space, basis vectors, etc., introduced in chapters 2–4. Ordinary three-dimensional (Euclidean) space is a linear vector space (LVS). Any vector in it can be expanded uniquely as a linear combination of the three unit vectors **i**, **j** and **k**. This is because these three vectors are linearly independent of each other, and they span the space: i.e., they form a basis in the LVS. Moreover, $\mathbf{i}\cdot\mathbf{i} = 1$, $\mathbf{i}\cdot\mathbf{j} = 0$, etc. In other words, this is an *orthonormal* basis. Using the superior notation ($\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$) for (**i**, **j**, **k**), all these relations can be compressed into the relation $\mathbf{e}_i\cdot\mathbf{e}_j = \delta_{ij}$ $(i, j = 1, 2, 3)$ in terms of the Kronecker delta. This can be generalized to any d-dimensional LVS: an orthonormal basis $\{\mathbf{e}_i\}$ satisfying

$$\mathbf{e}_i\cdot\mathbf{e}_j = \delta_{ij} \quad \text{(orthonormality)} \tag{15}$$

where $i, j = 1, 2, \ldots, d$ can always be found in the LVS.

What happens if the dimensionality $d \to \infty$? Some subtleties arise. But the preceding discussion goes through, provided care is taken to ensure that certain desirable properties survive – e.g., the vectors in the LVS must have finite magnitudes, and the *triangle inequality* must be satisfied by the magnitudes (or *norms*) of any two vectors and their sum or resultant vector. *Function spaces* provide simple examples of such infinite-dimensional LVS's – for instance, the space of all square-integrable functions of x in some interval $[a, b]$. Naturally, the basis is then an infinite set of suitable functions. A common example is the set of Legendre polynomials $\{P_\ell(x)\}$ where $x \in [-1, 1]$ and $\ell = 0, 1, \ldots$ *ad inf.* The notion of the dot product or inner product of two vectors must also be generalized appropriately. I shall not go into this detail here.

But a new possibility arises when the dimensionality of an LVS is infinite. The vector space may have a basis that is *uncountably* infinite, i.e., a so-called *continuous* basis. Instead of a set $\{\mathbf{e}_i\}$ where i is a discrete index, we have a set $\{\mathbf{e}(\xi)\}$ where ξ is a continuous variable, taking values in some range. It is convenient to use a more abstract notation here, namely, Dirac's bra

and ket notation, because the vectors of the LVS may be quite different in nature than ordinary vectors in two- or three-dimensional Euclidean space. For instance, they may be functions, or matrices of finite or infinite order. (Recall that the bra and ket notation has already been introduced briefly in ch. 2.) Thus, the vector $\mathbf{e}(\xi)$ is simply written as $|\,\xi\,\rangle$. The corresponding dual vector, or bra vector, is $\langle\,\xi\,|$. Passing over some technical details, the orthonormality condition for a continuous basis formally reads

$$\langle\,\xi\,|\,\xi'\,\rangle = \delta(\xi - \xi') \quad \text{(orthonormality)}. \tag{16}$$

Thus, the Dirac delta function replaces the Kronecker delta.

This is the context in which Dirac required the delta function.[§] In quantum mechanics, a system is described by its so-called state vector. This is an element of a certain LVS called the Hilbert space of the system. The classical dynamical observables of the system are replaced by operators that act on the elements of its Hilbert space. It turns out to be convenient to choose the eigenstates of (a subset of) these physical operators as the possible basis sets in the Hilbert space. Moreover, the system gets into these eigenstates when measurements of the corresponding physical observables are made. Certain fundamental operators such as the position operator or linear momentum operator of a particle moving in a given force field turn out to have continuous sets of eigenvalues. Their eigenstates then constitute continuous basis sets, with the orthonormality condition as in eq. (16). An example is provided by the case of a particle moving in one dimension under the influence of a constant force. Its energy E can then be shown to have a continuous set of possible values. The state $|\,E\,\rangle$ of the particle corresponding to a definite value E of its total energy is a member of a continuous basis set of states, satisfying the orthonormality condition $\langle\,E,|\,E'\,\rangle = \delta(E - E')$. Likewise, the position basis and momentum basis for a quantum mechanical particle moving in three-dimensional space satisfy the respective orthonormality relations

$$\langle\,\mathbf{r}\,|\,\mathbf{r}'\,\rangle = \delta^{(3)}(\mathbf{r} - \mathbf{r}') \;\text{ and }\; \langle\,\mathbf{p}\,|\,\mathbf{p}'\,\rangle = \delta^{(3)}(\mathbf{p} - \mathbf{p}'). \tag{17}$$

The delta function and the unit operator

The orthonormality relation in a continuous basis, eq. (16), leads naturally to another important aspect of the delta function: its relation with the unit operator in function space.

Consider a linear vector space (LVS) with a continuous basis $\{|\,\xi\,\rangle\}$ labelled by a real variable $\xi \in \mathbb{R}$. In addition to the orthonormality condition (16), the basis vectors satisfy another relation. Recalling the notion of a *projection operator* introduced in ch. 2, the object $|\,\xi\,\rangle\langle\,\xi\,|$ is a projection operator corresponding to the label ξ. The sum of all the projection operators

[§]Dirac's formulation of quantum mechanics is arguably an intellectual achievement on par with Newton's laws of motion and Einstein's theory of general relativity. The biography of Dirac by G. Farmelo, cited in the Bibliography, makes fascinating reading.

in the basis must be equal to the unit operator. Since ξ is a continuous variable, the sum is replaced by an integral. Hence

$$\int_{-\infty}^{\infty} d\xi \, |\, \xi \,\rangle\langle\, \xi \,| = \mathbb{I} \quad \text{(completeness)}. \tag{18}$$

This property is called *completeness* (of the basis). It is a consequence of the fact that the basis vectors *span* the LVS.

The notion of a continuous basis $\{|\, \xi \,\rangle\}$ leads to an interpretation of what a function like $f(\xi)$ stands for. Let $|\, f \,\rangle$ be any vector in the LVS. Then, operating on $|\, f \,\rangle$, the two sides of eq. (18) give

$$|\, f \,\rangle = \mathbb{I}\,|\, f \,\rangle = \int_{-\infty}^{\infty} d\xi \, |\, \xi \,\rangle\langle\, \xi \,|\, f \,\rangle = \int_{-\infty}^{\infty} d\xi \, f(\xi)\,|\, \xi \,\rangle, \tag{19}$$

where

$$f(\xi) \equiv \langle\, \xi \,|\, f \,\rangle. \tag{20}$$

In other words, the function $f(\xi)$ is just a convenient notation for the coefficient in the expansion of a vector $|\, f \,\rangle$ in the continuous basis labelled by ξ!

The orthonormality relation (18) also leads to yet another interpretation of the Dirac delta function. Consider the general 'matrix element' of the unit operator by sandwiching it between the bra $\langle\, \xi_1 \,|$ and the ket $|\, \xi_2 \,\rangle$. The result is

$$\langle\, \xi_1 \,|\, \mathbb{I} \,|\, \xi_2 \,\rangle = \langle\, \xi_1 \,|\, \xi_2 \,\rangle = \delta\,(\xi_1 - \xi_2). \tag{21}$$

That is, the Dirac delta function is simply the representative (or matrix element) of the unit operator in function space – precisely as the Kronecker delta δ_{ij} is the $(ij)^{\text{th}}$ matrix element of the unit matrix in the discrete case!

Finally, the completeness relation shows us what the inner product of two vectors $|\, f \,\rangle$ and $|\, g \,\rangle$ belonging to the LVS looks like. Sandwich the operators in eq. (18) between $\langle\, f \,|$ and $|\, g \,\rangle$. Then, since $f(\xi) = \langle\, \xi \,|\, f \,\rangle$ and $g(\xi) = \langle\, \xi \,|\, g \,\rangle$, we have

$$\langle\, f \,|\, g \,\rangle = \int_{-\infty}^{\infty} d\xi \, \langle\, f \,|\, \xi \,\rangle\langle\, \xi \,|\, g \,\rangle = \int_{-\infty}^{\infty} d\xi \, f^*(\xi)\, g(\xi). \tag{22}$$

In the final equation, we have used the property $\langle\, f \,|\, \xi \,\rangle = \langle\, \xi \,|\, f \,\rangle^*$ of the inner product in a complex LVS. (The asterisk denotes the complex conjugate.)

The foregoing brief account should convince you of the power and usefulness of both the Dirac delta function and the Dirac bra-and-ket notation.

The delta function and Green functions

We have seen that the delta function is essentially the representation of the unit operator in a continuous basis (or a function space). This fact has a direct bearing on the *Green function*

method for the solution of linear differential (and partial differential) equations. This is a good place to discuss this aspect in brief. What follows is a 'bare essentials' sketch, without any of the necessary conditions, qualifying clauses, etc. It is meant merely to convey the general idea. For simplicity of notation, I shall revert to the case of a single independent variable, say x, ranging over some interval $[a, b]$. Extension to several independent variables (e.g., $\mathbf{r}$, or $\mathbf{r}$ and t) is straightforward.

Consider a linear inhomogeneous differential equation of the form

$$D_x\, f(x) = g(x), \tag{23}$$

where $g(x)$ is a given function and D_x is some differential operator in x, i.e., it is some function of $x, d/dx, d^2/dx^2, \ldots$. It is required to find the unknown function $f(x)$. Clearly, a formal solution is $f(x) = D_x^{-1}\, g(x)$, where D_x^{-1} denotes the *inverse* of the operator D_x (provided the inverse exists, etc.). But this is not the most general solution of eq. (23). The *homogeneous* equation $D_x\, h(x) = 0$ may have one or more linearly independent solutions $h_i(x)$, where $i = 1, 2, \ldots$. These can be added to the solution of the inhomogeneous equation in any linear combination. The general solution of eq. (23) is therefore

$$f(x) = D_x^{-1}\, g(x) + \sum_i c_i\, h_i(x), \tag{24}$$

where the coefficients c_i are constants. It is trivially verified that eq. (23) is recovered on applying the operator D_x to both sides of eq. (24). You will recognise the first term on the right-hand side of (24) as the *particular integral*, and the second term as the *complementary function*. The constants c_i are determined by imposing the boundary or initial conditions, as the case may be.

In the present context, it is the PI that is of interest to us. Let us therefore focus on this part of the solution. What can it be, in general? You might guess that the inverse of a differential operator is an *integral* operator of some sort: namely, that $D_x^{-1}\, g(x)$ is of the form

$$D_x^{-1}\, g(x) = \int_a^b dx'\, G(x, x')\, g(x'), \tag{25}$$

where the function $G(x, x')$ is yet to be found. Applying the operator D_x to both sides of eq. (25), we get

$$g(x) = \int_a^b dx'\, D_x\, G(x, x')\, g(x'). \tag{26}$$

This equation must hold good for any $g(x)$. This is only possible if $G(x, x')$ itself satisfies the differntial equation

$$D_x\, G(x, x') = \delta(x - x'). \tag{27}$$

$G(x, x')$ is called the Green function of the operator D_x.

But we know that $\delta(x - x')$ is just the matrix element $\langle\, x \,|\, \mathbb{I} \,|\, x \,\rangle$ of the unit operator. Equation (27) is therefore a relation between corresponding matrix elements of operators. Hence, it represents a relation between the operators themselves. What is this relation? The differential operator D_x is the representation, in the continuous basis labelled by x, of an abstract operator D that acts on vectors in the LVS. This means that D_x and D are related as follows. If $|\, f \,\rangle$ is *any* element of the LVS, then

$$\langle\, x \,|\, \mathsf{D} \,|\, f \,\rangle = D_x \langle\, x \,|\, f \,\rangle \equiv D_x\, f(x). \tag{28}$$

Similarly, there exists an operator G that acts on vectors in the LVS, such that

$$\langle\, x \,|\, \mathsf{G} \,|\, x' \,\rangle = G(x, x'). \tag{29}$$

It follows at once that eq. (27) expresses the fact that

$$\langle\, x \,|\, \mathsf{DG} \,|\, x \,\rangle = \langle\, x \,|\, \mathbb{I} \,|\, x \,\rangle. \tag{30}$$

Since this is true for all x and x' in $[a, b]$, we must have

$$\mathsf{D}\,\mathsf{G} = \mathbb{I}, \quad \text{or} \quad \mathsf{G} = \mathsf{D}^{-1}. \tag{31}$$

This is the precise sense in which the Green function of a differential operator is just its inverse.

In Chs. 9 and 10, we will derive the fundamental Green function of the wave operator in $d + 1$ dimensions (d spatial dimensions and 1 time dimension), and discuss its properties for different values of d.

The Poisson-Dirac connection

I have mentioned earlier that Poisson was responsible for what was perhaps the first recognizable use of the Dirac delta function. Poisson and Dirac seem to be linked in more ways than one. The most profound of these links is this: Dirac showed that *Poisson brackets* in classical dynamics become the *commutators* of the corresponding operators in quantum mechanics, divided by the constant factor $i\hbar$. It is therefore appropriate to end this chapter with another fascinating link between the names of Dirac and Poisson. There is a very useful and remarkable result in Fourier analysis called the *Poisson summation formula*. In its simplest form, this says that if $\widetilde{f}(k)$ is the Fourier transform of $f(x)$, then

$$\sum_{n=-\infty}^{\infty} f(n) = \sum_{n=-\infty}^{\infty} \widetilde{f}(2\pi n)\,. \tag{32}$$

A very elegant and simple way of deriving this formula makes use of the so-called *Dirac comb*: an array of Dirac delta functions located at the integers. It can be shown that

$$\sum_{n=-\infty}^{\infty} \delta(x - n) = \sum_{n=-\infty}^{\infty} e^{2\pi n i x}\,, \tag{33}$$

i.e., the Dirac comb is identically equal to a sum of exponentials! The latter can be reduced to the expression $1 + 2\sum_1^\infty \cos(2\pi n x)$. The cosines in the sum interfere destructively with each other, leaving behind just the sharp δ-function spikes at integer values of x. With the help of eq. (33), the Poisson summation formula is established quite easily. We have thus come full circle, moving from Poisson to Dirac and returning to Poisson with the help of Dirac!

Signal Propagation: 1. Solution of the Wave Equation*

V. Balakrishnan

The best of all possible worlds

In Voltaire's classic satire *Candide*, the preceptor Pangloss is an unquestioning optimist who keeps insisting that we live in 'the best of all possible worlds', in spite of the most harrowing adversities he and his companions face. Pangloss' unbridled optimism is foolish, if not dangerously stupid, in the light of the events that take place – so much so, that 'panglossian' has come to describe a hopelessly idealistic view held in spite of direct evidence to the contrary.

At a more decidable level, however, we may ask whether the physical universe in which we live is, at least in some limited sense, 'the best of all possible worlds'. But this sense must be specified more precisely, because it is not my intention to discuss any version of the so-called Anthropic Principle here. It is therefore necessary make the question sharper and our objective much more modest. Is there an attribute of the physical universe that is optimal in some fundamental respect?

In this chapter and the next, we shall see that the three-dimensional nature of space is such an attribute. It will be shown that it is impossible to send sharp signals in one- and two-dimensional spaces, in contrast to three-dimensional space. This result, sometimes referred to as the strict form of *Huygens' Principle*, means that the very possibility of communication, and hence the transmission of information from one location to another, is sensitively dependent on the dimensionality of space. In this sense, we are indeed fortunate to live in a space of three dimensions.**

Propagation of a sharply pulsed signal

For the sake of definiteness, I choose a specific mathematical model of signal propagation. Even though it is a 'bare minimum' sort of model, it is general enough to establish the primary result. And this result remains unaffected in essence by the addition of various details and modifications. Substantiating the last statement would lead us into lengthy digressions, and so we shall not attempt to do so here.

The signal we would like to transmit is a sharp pulse that has a definite beginning and a definite end in both space and time: in technical terms, it must be *localized* in space and time. This is essential for us to make out unambiguously that it emanated at *this* place at *this* time, and reached *that* place at *that* time. Let us denote the signal (or rather, some observable property

*Based on *Resonance*, Vol.**9**, No.6, pp.30-38, 2004.

**Although some may feel that mobile phones have made this a mixed blessing, at best!

of the signal, such as its amplitude) by $u(\mathbf{r}, t)$. The propagation of the signal in space-time is governed by the wave equation

$$\left(\frac{1}{c^2}\frac{\partial^2}{\partial t^2} - \nabla^2\right) u(\mathbf{r}, t) = \rho(\mathbf{r}, t) \tag{1}$$

where $\rho(\mathbf{r}, t)$ is a specified function of space and time that represents the *source* of the signal, and c is the speed of the signal. The differential operator on the LHS is called the *wave operator* or the *d'Alembertian*. It is the analogue, in space-time, of the Laplacian operator in space. In writing down eq. (1), I have in mind an electromagnetic signal propagating in free space or vacuum. To keep matters as simple as possible, I assume that the signal is described by a single scalar function u rather than the electric and magnetic field vectors that an actual electromagnetic signal would comprise. It is also assumed that the space in which the signal propagates is ordinary Euclidean space of infinite extent. This helps us avoid complications arising from boundary conditions. As the objective is to analyse and compare signal propagation in spaces of different dimensions, the symbol d will be used for the number of spatial dimensions: d may be 1, 2 or 3. Subsequently, we shall also take a look at what happens for $d > 3$.

In order to focus on the essentials, we consider a sharply pulsed point source of unit strength that is switched on at some point $\mathbf{r}_0$ at the instant of time t_0. We want to find the resulting signal $u(\mathbf{r}, t)$ at an arbitrary point $\mathbf{r}$ at an arbitrary instant of time t. Here $\mathbf{r}_0$ and $\mathbf{r}$ stand for d-dimensional position vectors. (When $d = 1$, $\mathbf{r}$ reduces to a single coordinate, x.) $u(\mathbf{r}, t)$ is then a solution of the equation

$$\left(\frac{1}{c^2}\frac{\partial^2}{\partial t^2} - \nabla^2\right) u(\mathbf{r}, t) = \delta^{(d)}(\mathbf{r} - \mathbf{r}_0)\, \delta(t - t_0)\,. \tag{2}$$

Here δ denotes the Dirac delta function, and $\delta^{(d)}$ its generalization to d dimensions, as defined in ch. 8. The natural boundary condition on $u(\mathbf{r}, t)$ is simply $u(\mathbf{r}, t) \to 0$ as $r \to \infty$, where r stands for $|\mathbf{r}|$, as usual. Moreover, as there is no disturbance anywhere before the source is switched on, we have $u(\mathbf{r}, t) = 0$ and $\partial u(\mathbf{r}, t)/\partial t = 0$ for all $t < t_0$, at all points. This requirement is called *causality*, which means that the effect cannot *precede* its cause. It implies that the solution $u(\mathbf{r}, t)$, which is naturally also dependent on t_0 and $\mathbf{r}_0$, must necessarily have a particular form. This is given by

$$u(\mathbf{r}, t) = \theta(t - t_0)\, K(\mathbf{r}, t\,;\, \mathbf{r}_0, t_0)\,, \tag{3}$$

where $\theta(t - t_0)$ denotes the unit step function (= 1 and 0, respectively, for $t > t_0$ and $t < t_0$.) Owing to the presence of the step function, it is clear that the quantity $K(\mathbf{r}, t\,;\, \mathbf{r}_0, t_0)$ is in fact left unspecified for $t < t_0$. But then this is irrelevant for the physical, causal, solution we seek here. In fact, causality imposes an even stronger constraint, as we shall see. Since the disturbance propagates with a finite speed c, we expect that the signal cannot reach a point $\mathbf{r}$ before the instant $t_0 + |\mathbf{r} - \mathbf{r}_0|/c$, because $|\mathbf{r} - \mathbf{r}_0|/c$ is the time it takes to propagate from $\mathbf{r}_0$ to $\mathbf{r}$. This property, too, will emerge automatically in the solution.

Jean le Rond d'Alembert (1717–1783)

The French mathematician, natural philosopher and encyclopaedist Jean le Rond d'Alembert was among the first to understand the significance of, and study in some detail, several important differential equations of mathematical physics. Among other results, he showed that the general solution of the one-dimensional wave equation

$$\frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = 0$$

is of the form

$$u(x,t) = f_1(x+ct) + f_2(x-ct).$$

This corresponds to the superposition of two different waveforms or pulses moving, respectively, to the left and right with speed c. It is the forerunner of the *method of characteristics* for a class of partial differential equations, the so-called hyperbolic partial differential equations. D'Alembert's name is associated with many other discoveries as well, such as d'Alembert's Principle in Mechanics, d'Alembert's scheme in games of chance such as roulette, and d'Alembert's paradox: he showed that, in the streamlined, irrotational flow of a non-viscous fluid past a solid obstacle, the net drag force on the solid *vanishes*, contrary to what one would guess off-hand.
D'Alembert (along with the philosopher Denis Diderot) spent a good deal of time and effort on a massive project, the great French Encyclopaedia. (Speaking of Diderot, there is no evidence for the veracity of the hilarious popular story about the great mathematician Euler confounding Diderot with his spoof of a 'mathematical proof' for the existence of God!) D'Alembert seems to have been a 'straight shooter'; according to W. W. Rouse Ball, "d'Alembert's style is brilliant but not polished, and faithfully reflects his character, which was bold, honest and frank. $\cdots$ with his dislike of sycophants and bores it is not surprising that during his life he had more enemies than friends."

From our discussion in ch. 8, you will recognise readily that eq. (2) is precisely the equation satisfied by the *Green function* corresponding to the wave operator. The specific solution $u(\mathbf{r}, t)$ in which we are interested here is the *retarded* or *causal Green function* corresponding to this operator, together with the natural boundary condition already stated. The function K is called the *causal propagator*.

We now have a well-defined mathematical problem. Although its solution is a standard exercise, it does involve a rather surprising number of subtleties – especially in the incorporation of the conditions that enable us to arrive at a physically acceptable solution. These finer (but important) points are so often slurred over or misrepresented in otherwise respectable texts, that it seems to be worthwhile to spell them out with some care, at the risk of appearing to dwell on technicalities.

Fourier transform

To proceed, we could substitute in eq. (2) the form given in eq. (3) for $u(\mathbf{r}, t)$, and work with the resulting partial differential equation for K. But it is just as convenient to work directly with eq. (2). We note that the variables $t - t_0$ and $\mathbf{r} - \mathbf{r}_0$ merely represent shifts of t and $\mathbf{r}$ by constant amounts, for any given t_0 and $\mathbf{r}_0$. This suggests immediately that we change variables from $\mathbf{r}$ and t to $\mathbf{R} \equiv \mathbf{r} - \mathbf{r}_0$ and $\tau \equiv t - t_0$, respectively. Equation (2) becomes

$$\left(\frac{1}{c^2}\frac{\partial^2}{\partial\tau^2} - \nabla_{\mathbf{R}}^2\right) u = \delta^{(d)}(\mathbf{R})\,\delta(\tau). \tag{4}$$

I have used the obvious notation $\nabla_{\mathbf{R}}^2$ for the Laplacian operator with respect to $\mathbf{R}$. The causality condition is $u = 0$ and $\partial u/\partial\tau = 0$ for all $\tau < 0^-$. The boundary condition is $u \to 0$ for $R \to \infty$, where $R \equiv |\mathbf{R}|$. It is evident that, under these circumstances, $u(\mathbf{r}, t)$ is in fact a function of $\mathbf{R}$ and τ. In anticipation of this, we have retained the symbol u for the unknown function in eq. (4). Note that in a region of *finite* extent, in the presence of boundary conditions at finite values of r, the dependence of u on $\mathbf{r}$ and $\mathbf{r}_0$ *cannot* be reduced in general to a dependence on the difference $\mathbf{r} - \mathbf{r}_0 = \mathbf{R}$ alone.

Now, the Fourier transform of $u(\mathbf{R}, \tau)$ with respect to both $\mathbf{R}$ and τ is defined as

$$\widetilde{u}(\mathbf{k}, \omega) = \int d^d\mathbf{R} \int_{-\infty}^{\infty} d\tau\, e^{-i(\mathbf{k}\cdot\mathbf{R} - \omega\tau)}\; u(\mathbf{R}, \tau). \tag{5}$$

Here. $d^d\mathbf{R}$ is the volume element in the space of the variable $\mathbf{R}$. Since $\mathbf{R}$ is just a shift of the vector $\mathbf{r}$ by a constant vector $\mathbf{r}_0$, it is clear that $d^d\mathbf{R} = d^d\mathbf{r}$, the volume element in the d-dimensional physical space $\mathbb{R}^d$. The integration is over all of this space. The inverse Fourier transform that yields $u(\mathbf{R}, \tau)$ in terms of $\widetilde{u}(\mathbf{k}, \omega)$ is given by

$$u(\mathbf{R}, \tau) = \int \frac{d^d\mathbf{k}}{(2\pi)^d} \int_{-\infty}^{\infty} \frac{d\omega}{2\pi}\, e^{i(\mathbf{k}\cdot\mathbf{R} - \omega\tau)}\; \widetilde{u}(\mathbf{k}, \omega). \tag{6}$$

Here $d^d\mathbf{k}$ denotes the volume element $dk_1\, dk_2 \ldots dk_d$ in the space $\mathbb{R}^d$ of the d-dimensional vector $\mathbf{k}$. Again, the integration is over all of this space. Using the Fourier representation of the δ-function (i.e., the fact that the Fourier transform of the delta function is just unity), we have

$$\delta^{(d)}(\mathbf{R})\,\delta(\tau) = \int \frac{d^d\mathbf{k}}{(2\pi)^d} \int_{-\infty}^{\infty} \frac{d\omega}{2\pi}\, e^{i(\mathbf{k}\cdot\mathbf{R} - \omega\tau)}. \tag{7}$$

Now substitute eqs (6) and (7) in eq. (4). But

$$\nabla_{\mathbf{R}}^2\, e^{i\mathbf{k}\cdot\mathbf{R}} = -k^2\, e^{i\mathbf{k}\cdot\mathbf{R}} \quad \text{and} \quad \frac{\partial e^{-i\omega\tau}}{\partial\tau^2} = -\omega^2 e^{-i\omega\tau}, \tag{8}$$

where k stands for $|\mathbf{k}|$ as usual. We then get

$$\int \frac{d^d\mathbf{k}}{(2\pi)^d} \int_{-\infty}^{\infty} \frac{d\omega}{2\pi}\, e^{i(\mathbf{k}\cdot\mathbf{R}-\omega\tau)} \{(\omega^2 - c^2k^2)\,\widetilde{u}(\mathbf{k},\omega) + c^2\} = 0, \tag{9}$$

The set of functions $e^{i(\mathbf{k}\cdot\mathbf{R}-\omega\tau)}$, where each Cartesian component of $\mathbf{k}$ and τ can take on all real values, forms a complete orthonormal basis[‡] in the space of integrable functions of $\mathbf{R}$ and τ. The left-hand side in eq. (9) can therefore only vanish if the coefficient of $e^{i(\mathbf{k}\cdot\mathbf{R}-\omega\tau)}$ for *each* value of $\mathbf{k}$ and ω, i.e., the expression in curly brackets, itself vanishes. We thus arrive at the solution for $\widetilde{u}(\mathbf{k},\omega)$, namely,

$$\widetilde{u}(\mathbf{k},\omega) = -\frac{c^2}{\omega^2 - c^2k^2} \tag{10}$$

for all $\mathbf{k}$ and ω. The idea behind the introduction of the Fourier transform should now be obvious now: it converts the partial differential equation for $u(\mathbf{R},\tau)$ into a trivially-solved algebraic equation for $\widetilde{u}(\mathbf{k},\omega)$.

Putting this expression for $\widetilde{u}(\mathbf{k},\omega)$ back in Eq. (6), we have the *formal* solution

$$u(\mathbf{R},\tau) = -c^2 \int \frac{d^d\mathbf{k}}{(2\pi)^d}\, e^{i\mathbf{k}\cdot\mathbf{R}} \int_{-\infty}^{\infty} \frac{d\omega}{2\pi}\, \frac{e^{-i\omega\tau}}{(\omega^2 - c^2k^2)}\,. \tag{11}$$

But this does not make sense as it stands, because the integral over ω diverges owing to the vanishing of the denominator of the integrand at $\omega = -c|k|$ and again at $\omega = c|k|$. The dilemma is resolved by invoking the physical requirement of causality, as we shall see now.

The causal solution

The trick is to carry out the integration over ω in eq. (11) by adding a well-chosen zero to the integral, so as to convert it to an integral over a closed contour in the complex ω-plane. The latter is then evaluated by applying *Cauchy's Residue Theorem*.

Let Ω be a large positive constant. Consider a closed contour comprising a straight line from $-\Omega$ to $+\Omega$ along the real axis in the ω-plane, and a semicircle of radius Ω that takes us back from $+\Omega$ to $-\Omega$ in either the upper or lower half-plane. The limit $\Omega \to \infty$ is to be taken after the contour integral is evaluated. If the contribution from the semicircle vanishes in the limit $\Omega \to \infty$, the original line integral from $-\infty$ to $+\infty$ over ω is guaranteed to be precisely equal to the integral over the closed contour.

Now, for $\tau < 0$, this semicircle *must* lie in the *upper* half-plane in ω (the region in which $\mathrm{Im}\,\omega > 0$), because it is only in this region that the factor $\exp(-i\omega\tau)$ in the integrand vanishes exponentially as $\Omega \to \infty$. The addition of the semicircle to the contour would then simply add a vanishing contribution to the original line integral that we want to evaluate. Therefore, if no singularities of the integrand lie on the real axis or in the upper half-plane in ω, the contour

[‡]Recall the notion of a basis set and related aspects introduced in earlier chapters.

integral is guaranteed to vanish identically for $\tau < 0$. But this is precisely what causality requires: namely, that $u(\mathbf{r}, t)$ remain equal to zero for all $t < t_0$, that is, for all $\tau < 0$.

On the other hand, for $\tau > 0$, i. e., for $t > t_0$, we do expect to have a signal that does not vanish identically. But now the semicircle closing the contour *must* lie in the *lower* half-plane (i.e., the region in which $\mathrm{Im}\,\omega < 0$), because it is only then that the contribution from the semicircle to the contour integral vanishes as $\Omega \to \infty$. Therefore, if all the singularities of the integrand are in the lower half-plane, all our requirements are satisfied.

This is ensured by displacing each of the poles of the integrand at $\omega = -ck$ and $\omega = +ck$ by an infinitesimal *negative* imaginary quantity $-i\epsilon$ where $\epsilon > 0$, and then passing to the limit $\epsilon \to 0$ after the integral is evaluated. In general, of course, each of the two poles of the integrand can be displaced so as to lie in the upper or lower half-plane. This leads to four possible ways of making the divergent integral finite. It is easy to see that any two of these are linearly independent solutions, the other two being linear combinations of the former pair. The particular way of displacing the poles (called an '$i\epsilon$-prescription') that I have used above is tailored to ensure that the correct causal solution is picked up from among the set of possible solutions.

- In general, such $i\epsilon$-prescriptions are a way of incorporating boundary conditions (here, initial conditions) into the solutions of differential equations.

The causal solution to eq. (2) or (4) is therefore given by the modified version of eq. (11) that reads

$$u(\mathbf{R}, \tau) = -c^2 \lim_{\epsilon \downarrow 0} \int \frac{d^d\mathbf{k}}{(2\pi)^d}\, e^{i\mathbf{k}\cdot\mathbf{R}} \times \int_{-\infty}^{\infty} \frac{d\omega}{2\pi}\, \frac{e^{-i\omega\tau}}{(\omega + ck + i\epsilon)(\omega - ck + i\epsilon)}\,. \tag{12}$$

But, as discussed above,

$$\begin{aligned} \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} \frac{e^{-i\omega\tau}}{(\omega + ck + i\epsilon)(\omega - ck + i\epsilon)} \quad &= \lim_{\Omega\to\infty} \int_{-\Omega}^{\Omega} \frac{d\omega}{2\pi}\, \frac{e^{-i\omega\tau}}{(\omega + ck + i\epsilon)(\omega - ck + i\epsilon)} \\ &= \lim_{\Omega\to\infty} \int_{C_\pm} \frac{d\omega}{2\pi}\, \frac{e^{-i\omega\tau}}{(\omega + ck + i\epsilon)(\omega - ck + i\epsilon)}\,, \end{aligned} \tag{13}$$

where $C_\pm$ denotes the closed contours shown in Figure 1. As explained above, we must use C_+ for $\tau < 0$ and C_- for $\tau > 0$. However, C_+ does not enclose any singularity of the integrand. Hence the corresponding integral vanishes, precisely as we want it to. In the case of C_-, the integral is $-(2\pi i)$ times the sum of the residues of the integrand at the two poles enclosed: the extra minus sign arises because C_- is traversed in the clockwise sense. We thus obtain, after simplification,

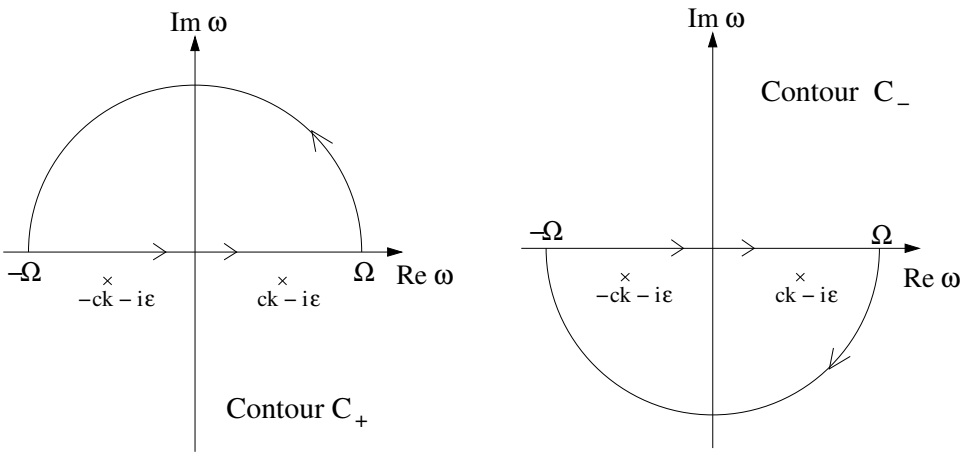


Figure 1 The closed contours C_+ and C_- for evaluating the integral in eq. (13).

$$u(\mathbf{R},\tau) = c\,\theta(\tau)\int \frac{d^d\mathbf{k}}{(2\pi)^d}\,\frac{\sin\,c\tau k}{k}\,e^{i\mathbf{k}\cdot\mathbf{R}}\,. \tag{14}$$

Note how the factor $\theta(\tau)$ required by causality has emerged automatically in the solution for $u(\mathbf{R},\tau)$.

For completeness, I mention in passing that solutions corresponding to the other possible $i\epsilon$-prescriptions mentioned earlier do play a role in physics – for instance, in the so-called *Feynman propagator* in quantum field theory.

Returning to the causal solution of interest to us, it remains to compute the integral in eq. (14). It becomes necessary, now, to distinguish between the solutions obtained for different values of d. The solution will therefore be written henceforth as $u^{(d)}(\mathbf{R},\tau)$ instead of $u(\mathbf{R},\tau)$. In the next chapter, we shall deduce and analyse the explicit form of $u^{(d)}(\mathbf{R},\tau)$ for individual values of d, to bring out the special features of the solution in each case.

Signal Propagation: 2. Effect of dimensionality*

V. Balakrishnan

In the preceding chapter, we have shown that the fundamental, causal solution to the wave equation

$$\left(\frac{1}{c^2}\frac{\partial^2}{\partial t^2} - \nabla^2\right) u^{(d)}(\mathbf{r}, t) = \delta^{(d)}(\mathbf{r} - \mathbf{r}_0)\,\delta(t - t_0) \tag{1}$$

that vanishes as $r \to \infty$ is given by

$$u^{(d)}(\mathbf{R}, \tau) = c\,\theta(\tau) \int \frac{d^d\mathbf{k}}{(2\pi)^d}\,\frac{\sin\, c\tau k}{k}\, e^{i\mathbf{k}\cdot\mathbf{R}}\,, \tag{2}$$

where d is the number of spatial dimensions, $\mathbf{R} \equiv \mathbf{r} - \mathbf{r}_0$ and $\tau \equiv t - t_0$. This solution is, as we know, the causal Green function of the wave operator in infinite d-dimensional space. We now simplify and analyse the solution for different values of d. It follows from eq. (1) that the physical dimensions of $u^{(d)}$ are $L^{2-d}\,T^{-1}$. This serves as a useful check on the solutions for different values of d.

The case $d = 1$

The case of a single spatial dimension is somewhat distinct from the others, and simpler, too.

Recall that the symbol k in the factor $\sin(c\tau k)/k$ in eq. (2) stands for $|\mathbf{k}|$. In the case $d = 1$, therefore, we should remember to write $|k|$ instead of just k in this factor. But $\sin(c\tau|k|)/|k| = \sin(c\tau k)/k$. Further, $\mathbf{k}\cdot\mathbf{R}$ is just kX in this case, where $X = x - x_0$. Therefore

$$u^{(1)}(X, \tau) = \int_{-\infty}^{\infty} \frac{dk}{2\pi}\,\frac{\sin\, c\tau k}{k}\, e^{ikX}. \tag{3}$$

It is clear from this expression that $u^{(1)}(-X, \tau) = u^{(1)}(X, \tau)$, i. e., that $u^{(1)}(X, \tau)$ is in fact a function of $|X|$. (This fact will become explicit in the solution to be derived, eq. (8) below.) Using $e^{ikX} = \cos kX + i \sin kX$, we see that the contribution from the $\sin kX$ term vanishes because the corresponding integrand is an odd function of k. Thus

$$\begin{aligned} u^{(1)}(X, \tau) &= c\,\theta(\tau) \int_0^{\infty} \frac{dk}{\pi k}\, \sin(c\tau k)\, \cos(kX) \\ &= c\,\theta(\tau) \int_0^{\infty} \frac{dk}{2\pi k}\{\sin(c\tau + X)k + \sin(c\tau - X)k\}. \end{aligned} \tag{4}$$

*Based on *Resonance*, Vol.**9**, No.7, pp.8-17, 2004.

We now need the standard integral

$$\int_0^\infty dk \frac{\sin\, bk}{k} = \frac{\pi}{2}\varepsilon(b) \tag{5}$$

where b is any real number, and $\varepsilon(b)$ is the (discontinuous) signum function

$$\varepsilon(b) = \frac{b}{|b|} = \begin{cases} +1 & \text{for} \quad b > 0 \\ -1 & \text{for} \quad b < 0. \end{cases} \tag{6}$$

Equation (4) then yields

$$u^{(1)}(X,\tau) = \frac{c}{4}\,\theta(\tau)\,\{\varepsilon(c\tau + X) + \varepsilon(c\tau - X)\}. \tag{7}$$

Simplifying the expression in the curly brackets, we find that it vanishes for $X < -c\tau$ and for $X > c\tau$, and is equal to 2 when $-c\tau < X < c\tau$. Therefore

$$u^{(1)}(X,\tau) = \frac{c}{2}\,\theta(\tau)\,\theta(c\tau - |X|)\,. \tag{8}$$

(u^1 has the correct physical dimensions, LT^{-1}.) The step function $\theta(c\tau - |X|)$ ensures that the signal does not reach any point x until time $t_0 + |x - x_0|/c$, as required by causality. The presence of this step function makes the other step function, $\theta(\tau)$, redundant from a physical point of view. However, it is present in the formal mathematical solution for the quantity $u^{(1)}(X,\tau)$.

There is another noteworthy aspect of the solution. Although an observer at an arbitrary point x starts receiving the signal at time $t_0 + |x - x_0|/c$, he does not receive a *pulsed* signal, even though the sender sent out such a signal. In fact, the signal received *persists* thereafter for all time, without diminishing in strength! This last feature is peculiar to $d = 1$. Let us now see what happens in higher dimensions.

The case $d = 2$

It is helpful to take note, first, of an important feature of $u^{(d)}(\mathbf{R},\tau)$ when $d \geq 2$. The expression in eq. (2) is a *scalar*: by this we mean that it is unchanged under rotations of the spatial coordinate axes about the origin. This remains true for all integer values of $d \geq 2$. This assertion may seem to be more-or-less obvious, because $\mathbf{k}\cdot\mathbf{R}$ is, after all, a scalar product of two d-dimensional vectors. But it must be proved properly. I leave the proof to the reader, after pointing out that two factors play a role in such a proof. (i) The region of integration in eq. (2) is *all* of $\mathbf{k}$-space, and this is invariant under rotations of the coordinate axes in that space. (ii) The volume element $d^{(d)}\mathbf{k}$ also remains unchanged under rotations of the axes.

As a result of this rotational invariance, $u^{(d)}(\mathbf{R},\tau)$ is actually a function of $R\,(\equiv |\mathbf{R}|)$ and τ. The consequence of this is that we can choose the orientation of the axes in $\mathbf{k}$-space according to our convenience, without affecting the result.

Turning now to the $d = 2$ case, it is evidently most convenient to work in plane polar coordinates (k, φ), choosing the k_1-axis along the vector $\mathbf{R}$. Then, setting $d = 2$ in eq. (2),

$$u^{(2)}(R,\tau) = c\,\theta(\tau)\int_0^\infty \frac{k\,dk}{(2\pi)^2}\,\frac{\sin\,c\tau k}{k}\int_0^{2\pi} d\varphi\, e^{ikR\,\cos\,\varphi}. \tag{9}$$

The definite integral over φ is a known function. It is 2π times $J_0(kR)$, the Bessel function of the first kind and order 0. Hence

$$u^{(2)}(R,\tau) = c\,\theta(\tau)\int_0^\infty \frac{dk}{2\pi}\,\sin\,(c\tau k)\,J_0(kR). \tag{10}$$

The final integral over k is again a known integral. It is equal to $(c^2\tau^2 - R^2)^{-1/2}$ provided $c^2\tau^2 > R^2$, and zero otherwise. Since we are concerned here with the physical region in which both τ and R are non-negative, the solution is given by

$$u^{(2)}(R,\tau) = \frac{c\;\theta(\tau)}{2\pi}\,\frac{\theta(c\tau - R)}{\sqrt{c^2\tau^2 - R^2}}\,. \tag{11}$$

($u^{(2)}$ has the physicsl dimensions T^{-1}, as required.) The signal thus reaches any point $\mathbf{r}$ only at time $t_0 + |\mathbf{r} - \mathbf{r}_0|/c$, in accordance with causality and the finite velocity of propagation of the disturbance. But once again, the signal is no longer a sharply pulsed one: it persists for all $t > t_0 + |\mathbf{r} - \mathbf{r}_0|/c$, with a strength that slowly decays like $1/t$ at very long times ($\tau \gg R/c$).

The case $d = 3$

In stark contrast to the cases $d = 1$ and $d = 2$, something entirely different happens in 3-dimensional space. Setting $d = 3$ in eq. (2), we have

$$u^{(3)}(R,\tau) = c\,\theta(\tau)\int \frac{d^3\mathbf{k}}{(2\pi)^3}\,\frac{\sin\,c\tau k}{k}\,e^{i\mathbf{k}\cdot\mathbf{R}}\,. \tag{12}$$

As you would expect by now, the integral is evaluated by exploiting rotational invariance. It is clear that we must use spherical polar coordinates (k, θ, φ) in $\mathbf{k}$-space, and, moreover, choose the polar axis along the vector $\mathbf{R}$. This immediately enables us to carry out the integration over the azimuthal angle φ, obtaining a factor 2π. It is instructive to write out all the subsequent steps in this instance, because they (or their variants) appear in more than one context in physical applications.

$$\begin{aligned} u^{(3)}(R,\tau) &= \frac{c\,\theta(\tau)}{(2\pi)^2}\int_0^\infty dk\,k^2\,\frac{\sin\,c\tau k}{k}\int_{-1}^{1} d(\cos\,\theta)\,e^{ikR\,\cos\,\theta} \\ &= \frac{2c\,\theta(\tau)}{(2\pi)^2 R}\int_0^\infty dk\,\sin\,(c\tau k)\,\sin\,(kR) \end{aligned}$$

$$= \frac{c\,\theta(\tau)}{(2\pi)^2 R}\int_0^\infty dk\ \{\cos(c\tau - R)k - \cos(c\tau + R)k\}$$

$$= \frac{c\,\theta(\tau)}{2(2\pi)^2 R}\int_{-\infty}^\infty dk\ \{\cos(c\tau - R)k - \cos(c\tau + R)k\}$$

$$= \frac{c\,\theta(\tau)}{2(2\pi)^2 R}\,\mathrm{Re}\int_{-\infty}^\infty dk\ \left\{e^{i(c\tau - R)k} - e^{i(c\tau + R)k}\right\}$$

$$= \frac{c\,\theta(\tau)}{4\pi R}\,\mathrm{Re}\ \{\delta(c\tau - R) - \delta(c\tau + R)\}\,. \tag{13}$$

But the delta functions are real quantities. Moreover, we are concerned with the region in which both τ and R are non-negative. The term $\delta(c\tau + R)$ therefore drops out, and the solution reduces to

$$u^{(3)}(R,\tau) = \frac{c\,\theta(\tau)\,\delta(c\tau - R)}{4\pi R} = \frac{\theta(\tau)\,\delta(\tau - R/c)}{4\pi R}. \tag{14}$$

Thus, almost miraculously, the signal remains a delta function pulse that reaches (and passes) an observer at any point $\mathbf{r}$ at precisely the instant $t_0 + |\mathbf{r} - \mathbf{r}_0|/c$. There is no after-effect that lingers on, unlike the situation in $d = 1$ and $d = 2$.

The solution (14) has the physical dimensions $L^{-1}T^{-1}$, as required. The amplitude of the pulse drops with distance like $1/R$, exactly the way the Coulomb potential drops off. In fact, this is yet another unique feature of the solution in $d = 3$. Formally, if the limit $c \to \infty$ is taken in eq. (1), the wave operator reduces to the negative of the Laplacian operator. We might therefore expect the solution for $u(\mathbf{r}, t)$ to reduce to the corresponding Green function for $-\nabla^2$. In three dimensions, this is precisely $1/(4\pi R)$. This fact is very familiar to you from electrostatics. The potential $\phi(\mathbf{r})$ due to a point charge q located at $\mathbf{r}_0$ satisfies the corresponding Poisson equation

$$-\nabla^2\phi(\mathbf{r}) = \rho(\mathbf{r})/\epsilon_0 = (q/\epsilon_0)\,\delta^{(3)}(\mathbf{r} - \mathbf{r}_0). \tag{15}$$

With the boundary condition $\phi \to 0$ as $r \to \infty$, the solution to this equation is just Coulomb's Law, namely, $\phi(\mathbf{r}) = -q/(4\pi\epsilon_0 R)$ where $R = |\mathbf{r} - \mathbf{r}_0|$. This reduction of the fundamental solution of the inhomogeneous wave equation to that of Poisson's equation in the limit $c \to \infty$ does *not* occur in $d = 1$ or $d = 2$.

Dimensions $d > 3$

Now that we have uncovered a very important feature of three-dimensional space that is absent in 1 and 2-dimensional spaces, it is natural to ask if this feature is unique to $d = 3$. Surprisingly, it is not: the propagation of sharp signals is possible in all *odd*-dimensional spaces with $d \geq 3$, while it fails for all *even* values of d. In other words, the signal received at any point $\mathbf{r}$ lingers on for all $t > t_0 + |\mathbf{r} - \mathbf{r}_0|/c$ in $d = 2, 4, \ldots$, while it is sharply pulsed, arriving and passing on at time $t_0 + |\mathbf{r} - \mathbf{r}_0|/c$ with no after-effect, in $d = 3, 5, \ldots$. There is, however, one feature that

is absolutely *unique* to $d = 3$: this is the only case in which the original δ-function pulse is transmitted without any *distortion*, namely, as a δ-function pulse.

One way to establish these results is to start with eq. (2), and to use hyperspherical coordinates in d dimensions.** Then

$$\mathbf{k} = (k, \theta_1, \theta_2, \ldots, \theta_{d-2}, \varphi), \tag{16}$$

where

$$0 \le k < \infty,\ 0 \le \theta_i \le \pi,\ 0 \le \varphi < 2\pi. \tag{17}$$

Once again, we may choose the k_1 axis to lie along the vector $\mathbf{R}$, which permits us to carry out the integrations over $\theta_2, \ldots, \theta_{d-2}$ and φ. The result is

$$u^{(d)}(\mathbf{R}, \tau) = (\text{const.})\, \theta(\tau) \int_0^\infty dk\, k^{d-2}\, \sin(c\tau k) \times \int_0^\pi d\theta_1\, (\sin\, \theta_1)^{d-2}\, e^{ikR\, \cos\, \theta_1}, \tag{18}$$

where the constant‡ depends on d. But this is clearly a laborious method of finding $u^{(d)}(\mathbf{R}, \tau)$, especially as the integrations over θ_1 and k have yet to be carried out.

There is a more elegant and powerful way to solve the problem. This is based on the *relativistic invariance* of the wave operator and the solution sought. A detailed account of this would take us too far afield. I shall therefore restrict myself to a short description of this approach, to enable you to get some feel for the underlying mechanism responsible for the basic difference between the cases of even and odd d. The discussion will not be fully rigorous, as I shall not elaborate on certain technical details that warrant a more careful examination.

The wave operator $(1/c^2)\,\partial^2/\partial t^2 - \nabla^2$ can be verified to be invariant (i.e., unchanged in form) under Lorentz transformations in $(d + 1)$-dimensional space-time.§ As a consequence of this invariance, the specific solution we seek can also be shown to be Lorentz-invariant. In the present context, this means that we can always evaluate the integrals involved in eq. (2) by first transforming to an inertial frame in which the four-vector $(c\tau, \mathbf{R})$ has only a time-like component, i.e., it is of the form $(c\tau', \mathbf{0})$, where $c^2\tau^2 - R^2 = c^2\tau'^2$. This can only be done for a so-called *time-like* four vector, i.e., one for which $c^2\tau^2 - R^2 > 0$. It cannot be done for a light-like four-vector (i.e., when $c^2\tau^2 - R^2 = 0$) or a space-like four-vector (i.e., when $c^2\tau^2 - R^2 < 0$). (This is the technical point I gloss over, with the remark that our conclusions will not be affected by it.) After the integrals required are evaluated, we can transform back to the original frame by replacing $c\tau'$ with $(c^2\tau^2 - R^2)^{1/2}$. It must also be mentioned that $\tau > 0$ implies $\tau' > 0$, because the sign of the time component of a four-vector remains unchanged under the set of Lorentz transformations with which we are concerned.¶ Denoting the corresponding signal by $u^{(d)}(\tau')$, eq. (2) simplifies to on carrying out all the angular integrations in d-dimensional space. The

**Recall that hyperspherical coordinates have been introduced and discussed in ch. 1.

‡You can calculate this constant using the expression for the solid angle in d dimensions (see ch. 1). The answer is $2\pi^{(d-1)/2}/\Gamma(\frac{1}{2}(d-1))$, where Γ denotes the gamma function.

§That is, d space dimensions and 1 time dimension.

¶Once again, this is only true for a time-like or light-like four-vector, but not a space-like one.

constant on the right-hand side in the last equation depends on d. This representation shows us, in very clear fashion, how the cases of odd and even d differ from each other. When d is odd, the integrand is an even function of k, and hence the integral can be converted to one that runs from $-\infty$ to ∞. The result can then be shown to be essentially a derivative of a certain order of the delta function $\delta(c^2\tau'^2)$, i.e., a sharply em localized, pulsed signal. (The order of the derivative increases with increasing d.) On the other hand, when d is even, this cannot be done, and the integral leads to an *extended* function of $c^2\tau'^2$. This dissection lays bare the precise mathematical distinction that lies at the root of the physical differences in signal propagation in odd and even dimensional spaces, respectively. In fact, the formal solution for $u^{(d)}(\tau')$ can be shown to be essentially the derivative of order $(d-3)/2$ of $\delta(c^2\tau'^2)$ in *all* cases. When d is even, this is a so-called *fractional derivative*, which is a non-local object — in physical terms, an extended function.

$$\begin{aligned} u^{(d)}(\tau') &= c\,\theta(\tau') \int \frac{d^d\mathbf{k}}{(2\pi)^d}\,\frac{\sin\, c\tau' k}{k} \\ &= (\text{const.})\,\theta(\tau') \int_0^\infty dk\, k^{d-2}\,\sin\,(c\tau' k), \end{aligned} \tag{19}$$

The form of the result in eq. (19) suggests even more. Since the second derivative of the sine function is again a sine function (apart from a minus sign), it follows that the solution in $(d+2)$ spatial dimensions can be obtained from that in D space dimensions by a simple trick. We find

$$u^{(d+2)}(\tau') = -\frac{1}{2\pi c^2 d}\,\frac{\partial^2\, u^{(d)}(\tau')}{\partial \tau'^2}\,. \tag{20}$$

This shows how the solutions in $d = 5, 7, \ldots$ can be generated from that in $d = 3$, while those in $d = 4, 6, \ldots$ can be generated from that in $d = 2$. The detailed working out of these solutions is left to the interested reader.

A final remark, before we pass on to more general considerations. How widely applicable are the conclusions at which we have arrived? Basically, there are two important additional aspects of wave or signal propagation that can be adjusted so as to modify the basic result. The first is *dispersion*. Sinusoidal waves of different wavelengths will, in general, propagate with different speeds in a medium. The precise manner in which the frequency and wavelength of waves in a medium are related to each other is called a dispersion relation. Such relations can be quite complicated. The second aspect is *nonlinearity*. The simple wave equation we have used, eq. (1), is *linear* in $u^{(d)}$. On the other hand, physical situations often call for nonlinear equations. The interplay between dispersion and nonlinearity can be extremely intricate and interesting, and a vast variety of new phenomena can arise as a result. Among these are the so-called *solitary waves* and propagating *solitons*, which represent very robust pulsed disturbances.

General remarks on dimensionality

I conclude this discussion with a few remarks of a very general nature about the dimensionality of space.

We have seen how 3-dimensional space has some special properties that are not shared by a space of any other dimensionality. Are there other such properties? Answers to this question can be given at many levels. Here is a general observation: One- and two-dimensional spaces are, in some sense, too 'simple' for anything too complicated to be possible; on the other hand, four- and higher-dimensional spaces are again too 'roomy' for anything very complicated to occur. In even more general terms, this roominess permits the undoing of complications like knots, for instance. From the point of view of topology, 3 and 4 dimensions permit complexities that are not shared by spaces of other dimensions. 'Low-dimensional topology' is a branch of mathematics in its own right! The famous Poincaré Conjecture regarding compact 3-dimensional manifolds, made in 1904, stood unproved for about a hundred years till it was settled, although its higher dimensional counterparts were established many decades earlier.[‖]

Finally, it must be recognized that our sensory organs and the information-processing hardware and software in our brains are designed so specifically for $(3+1)$-dimensional space-time, that we literally take this dimensionality to be a fundamental and self-evident fact of nature. In actuality, however, there are very deep unanswered questions about the nature of space and time. These questions are connected to questions about quantum mechanics, general relativity and the origin of the universe. We do not know for sure whether, at the very smallest time scales and length scales, the number of space dimensions is three or more; whether the number of dimensions is itself a dynamic, 'emergent' property of the universe; whether space-time coordinates must be supplemented with certain other kinds of variables to specify a point in the 'true' arena in which phenomena occur; and even whether space-time is ultimately continuous or discrete (granular). One thing does appear to be fairly certain, though:

- It is very probable that, sooner or later, our long-standing ideas and preconceptions about the nature of space and time will have to be revised significantly at the most fundamental level.

[‖]For a semi-popular account of the Poincaré Conjecture and its resolution, see the book by D. O'Shea cited in the Bibliography.

Suggested Reading

[1] G. B. Arfken, H. J. Weber and F. E. Harris, *Mathematical Methods for Physicists*, 7th Edition, Academic Press, 2012.

[2] V. Balakrishnan, *Mathematical Physics*, Ane Books, 2018.

[3] G. Barton, *Elements of Green's Functions and Propagation – Potentials, Diffusion and Waves*, Clarendon Press, 1989.

[4] E. T. Bell, *Men of Mathematics*, Simon & Schuster, 1965.

[5] M. L. Boas, *Mathematical Methods in the Physical Sciences*, 3rd Edition, Wiley, 2006.

[6] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vols. 1 & 2, Wiley-VCH, 1989.

[7] P. Dennery and A. Krzywicki, *Mathematics for Physicists*, Dover, 1996.

[8] P. A. M. Dirac, *The Principles of Quantum Mechanics*, 4th Edition, Clarendon Press, 1982. (An inexpensive Indian edition is currently available.)

[9] D. G. Duffy, *Green's Functions with Applications*, 2nd Edition, Chapman & Hall/CRC, 2015.

[10] G. Farmelo, *The Strangest Man: The Hidden Life of Paul Dirac, Mystic of the Atom*, Basic Books, 2011.

[11] P. R. Garabedian, *Partial Differential Equations*, 2nd Edition, Chelsea Publishing, 1998.

[12] H. Goldstein, J. L. Safko and C. P. Poole, *Classical Mechanics*, 3rd Edition, Pearson, 2001.

[13] M. D. Greenberg, *Applications of Green's Functions in Science and Engineering*, Dover, 2015.

[14] P. R. Halmos, *Finite-Dimensional Vector Spaces*, Springer, 1993.

[15] M. J. Lighthill, *Introduction to Fourier Analysis and Generalized Functions*, Cambridge University Press, 1960.

[16] J. Lützen, *The Prehistory of the Theory of Distributions*, Springer-Verlag, 1982.

[17] D. O'Shea, *The Poincaré Conjecture: In Search of the Shape of the Universe*, Walker Books, 2007.

[18] D. H. Sattinger and O. L. Weaver, *Lie Groups and Algebras with Applications to Physics, Geometry, and Mechanics*, Springer, 1993.

[19] G. B. Whitham, *Linear and Nonlinear Waves*, Wiley, 1999.