The Lie group approach to solving differential equations

Introduction

Certain ideas recur in many areas of mathematics. One example is groups of symmetries, which appear in the Galois theory of equations and in Lie groups. Lie groups are of great value in physics, where Noether's theorem enables us to derive a conservation law for every case in which a function known as the Lagrangian is invariant under a one-parameter Lie group. The importance of this approach can be seen from the fact that the laws of the conservation of energy, linear momentum and angular momentum are all outcomes of Noether's theorem, though they can of course be derived by simpler methods. The full power of Noether's approach is shown in its applications to quantum field theory, where it can be used to find conserved currents and charges.

Despite their central place in mathematical physics, Lie groups are generally regarded as requiring a long apprenticeship in the theory of differentiable manifolds and topological groups. However, this is not how they arose historically. Sophus Lie was prompted to study these structures in the late nineteenth century. He became convinced that the power of Galois theory in the investigation of the solutions of algebraic equations could be harnessed to the study of differential equations, and could be made to yield equally striking results there. This application has been almost forgotten today, or relegated to the province of specialists, but the purpose of this paper is to argue that an elementary treatment is both possible and enlightening. It is hoped that this treatment will help to motivate students to believe that Lie groups can be not only useful, but natural, objects of study, with an intuitive interpretation in terms of phenomena that can be visualized at a physical level, certainly in the one-parameter case.

This paper is not necessarily meant for consumption neat, as it were, by students. However, it the author's hope that it all or part of it might serve as the basis for classroom exposition by teachers or lecturers at senior sixth form or first year university levels. The content is based on my reading of both classical and modern texts. I have found Cohen [1] to be invaluable. It appears to be close to the spirit of Lie's original papers. A somewhat more concise treatment is available in Ince [2, chapter 4]. The present paper aims to cover roughly the ground of that chapter, but informed by the more modern approach inspired by the notes of an illuminating lecture by Helgason [3]. All of these sources are available online, and can be downloaded free from academic electronic resources.

I have included a discussion section which contains an admittedly conjectural attempt to explain, from a heuristic viewpoint, why Lie's methods work as well as they do.

Motivation

Historically, mathematicians sought in vain to find a general method of calculating explicit solutions for ordinary differential equations (ODEs) of the form $dy/dx = f(x, y)$, or equivalently, $Mdx + Ndy = 0$ (where M, N are functions of x and y). By analogy with the particular devices used to solve algebraic equations up to the fourth degree, mathematicians resorted to looking for particular approaches which would at least suffice to solve specific types. Some varieties, such as linear or homogeneous equations, yielded to fairly simple treatments. These methods fell into one of two broad categories: a change of variables in which an equation would become separable, or the discovery of an "integrating factor" I which would enable $I(Mdx + Ndy)$ to be expressed as a so-called "exact differential", i.e. in the form $d(F(x, y))$, from which the solution could be found immediately as $F(x, y) = \text{const}$.

Lie was inspired by the approach of Evariste Galois, which enabled the solution of an algebraic equation to be understood as the application of properties of the group under which the equation was invariant, in the sense that the solutions of the equation were permuted amongst themselves by elements of the group. Lie discovered that integrating factors for ODEs could be calculated once a group of transformations could be specified under which the ODE was invariant, in the sense that the elements of the group permuted the solution curves of the ODE amongst themselves.

At first sight it may seem that this approach simply replaced one difficult problem, that of finding an integrating factor for an ODE, with another equally difficult one, that of finding a group under which it is invariant. However, it turns out that Lie groups do simplify the problem, provided one approaches it from a different angle.

The key idea is to turn the problem on its head. Although one might begin with an ODE and seek a Lie group which leaves it invariant, it turns out to be more fruitful to begin by seeking Lie groups that act on the real plane, and then finding which ODEs they preserve. Felix Klein famously defined a geometry in terms of the group of transformations which left certain entities invariant. This leads one in a natural way to look at groups of transformations of the real Euclidean plane such as the affine transformations, the rotations about the origin, and the group of dilations (in which the position vectors of points are multiplied by a constant). These are all Lie groups, though in the case of the affine maps one has to take subgroups to obtain one-parameter Lie groups. It turns out that each such group gives rise to not just one, but a whole class of differential equations which are invariant under the group in question. Better still, it is possible to find the general form of these ODEs explicitly from the group itself.

So by starting with Lie groups, one automatically finds the solution to whole categories of ODEs, whereas starting with an ODE one may only succeed in solving that particular equation. The Lie group approach therefore yields great economy of effort. But it does much more than this. It provides an underlying mechanism which unites a number of apparently random and unconnected methods of finding integrating factors and shows *why* these methods work. It can even be used to show that the two broad categories of the separation of variables method and the discovery of an integrating factor, are underlying manifestations of the same principle. This paper proposes to explore at an intuitive level the integrating factor methodology. Identifying in detail how it equates to the separation of variables would take us too far afield, but we will indicate briefly how this arises.

Vector fields

We introduce the notion of a one-parameter Lie group acting on the real plane R^2, and the vector field to which it gives rise, as follows. Imagine that the plane is covered with a thin layer of some fluid, which is flowing across the surface of R^2 like a river running across its flood plain, except that this river is infinitely large so that its flood plain is the whole of R^2. Imagine also that the flow is in a steady state, so that at each fixed point, the velocity is constant over time.

We make no assumptions about the fluid being like water in any physical sense; it is not, for instance, assumed to be incompressible; moreover, we are not interested in the depth, simply the horizontal flow velocity at any point. We do, however require that the fluid moves in a smooth continuous manner. This means that if you drop a cork into the fluid at some point, its subsequent journey is along a curve that is smooth, and that the cork's velocity varies smoothly with time. (The assumptions of smoothness can be made concrete for more sophisticated students by specifying that the curves are C^∞).

At each point of the plane there is a velocity defined, which is constant over time if one focuses on a fixed point of the plane. Since velocity is a vector, each point gives rise to a vector. This association between points and vectors can be regarded as a map defined with domain as the entire plane, which takes a point of the plane onto the vector of the flow velocity at that point. This mapping from the plane to the two dimensional vector space of velocities, is called a **vector field.**

The group of the vector field

The flow clearly defines a series of what in hydrodynamics are called **streamlines**. The streamline through a point P is the path that a small object such as a cork would take if placed at P at some instant, and then allowed to travel with the flow for as long as we like. Intuitively, there is a streamline through each point, and two distinct streamlines cannot intersect: if they did intersect at a point, it would mean that the flow at that point must be in two directions at once, which contracts our assumption that there is a well defined velocity everywhere (streamlines may, however, intersect at points where the velocity is zero. We will usually exclude such points in what follows). Therefore the streamlines are a **congruence of curves**, or a **congruence** for short, defined to be a set of curves such that there is one and only one curve through each point (again, we may have to exclude certain exceptional points).

So far, we have defined streamlines as generated only in one time direction, that of increasing time, but it is easy to see intuitively that they "go backwards" as well. Reversing the flow by taking a flow field the exact negative of the given field would give the other half of each streamline. These definitions are consistent, because if the flow takes a cork from point A to point B in time t, the reverse flow takes it from B to A in time t.

It is intuitively clear that for any given time t, there is a map defined from R^2 to itself defined by the flow of the fluid over time t. The image of a point P in R^2 can be envisaged as the location of a cork placed in the flow at P at a certain time and carried along in the flow for a time t. This is a "nice" map of R^2 onto itself (technically, a diffeomorphism of R^2). Call this map arising from a flow for time t, g(t). When t is zero, g(0) is clearly the identity map, and for each t, g(-t) is the inverse of g(t). The maps g(t) as t varies form a group under composition, the **one-parameter group** associated with the flow, and this group behaves under composition like the additive group R, because it is clear intuitively that $g(s + t) = g(s) \circ g(t)$. We refer to this group as G.

In the traditional treatments, the vector field is referred to as the "infinitesimal transformation" defined by the group. The group itself is referred to as the one-parameter group "generated" by the infinitesimal transformation. The motivation for this definition is that in a certain sense the vector field generates the group, in the same way that a single element can generate a finite cyclic group. This is because the effect of g(t) is to displace a point sequentially along an infinite number of small paths, each one given by the flow velocity at that point.

In practice, a one-parameter Lie group is usually given in terms of the vector field, that is, the infinitesimal transformations that define it, rather than explicitly as a set of functions g(t) for real t. Certain vector fields are, as we have mentioned, natural and obvious ones to adopt in the plane. The next obvious step is to integrate these infinitesimal transformations and derive the functions g(t). This is generally not difficult. The task is made easier by the fact that the functions g(t) are uniquely defined by their infinitesimal transformations, as is intuitively clear if one thinks of them as fluid motions over a given time interval. However, we will see that in order to use these groups to find integrating factors for ODEs, even this explicit construction of the group is not essential; only the vector field is required. We will, however, carry out the process of constructing the group in the simple instances of Lie groups examined below.

The following examples may help to make the concepts clearer.

Example 1. If the flow is one of constant velocity u parallel to the x-axis, so that the velocity at any point is given by $\mathbf{v}(x, y) = (u, 0)$, the group it generates is very simple: g(t) takes the point (x, y) to the point (x + ut, y), so the group is that of translations parallel to the x-axis. The streamlines are the horizontal lines.

Example 2. Likewise, a flow with constant speed v parallel to the y-axis gives rise to the group of transformations h(t) where h(t) takes (x, y) to (x, y + vt). The streamlines are the vertical lines.

Example 3. If the flow is a "rigid" rotation around the origin with unit angular velocity, the actual velocity at the point (x, y) is (-y, x). The group is most easily described using polar coordinates. The element g(t) takes the point (r, θ) to the point $(r, \theta + t)$. The streamlines are the concentric circles centred on the origin.

Example 4. If the flow is directed radially outwards from the origin with velocity equal to the radius vector, then $\mathbf{v} = (x, y)$. In terms of radial coordinates, $dr/dt = r$, $d\theta/dt = 0$, so $r(t) = r(0)e^t$ and g(t) takes (x, y) to (xe^t, ye^t), the streamlines being the set of straight lines through the origin. Since e^t attains every positive real value, this consists of the group of "dilations" with centre at the origin.

A reader who wishes to follow the argument in the work of, for example, [2], will find a different notation. There, the equivalent of our vector field **v** defined on the plane with components (u, v) at each point, is presented as a set of infinitesimal transformations $u\,\partial/\partial x + v\,\partial/\partial y$. This refers to the fact that $u\,\partial f/\partial x + v\,\partial f/\partial y$ represents the rate of change of an arbitrary scalar function of position, f, when the point at which it is evaluated is carried along by the flow generated by the vector field. Our notation is simpler, and mathematically equivalent.

Effect of the one-parameter group elements on subsets of R^2

Just as each element of the group maps a point of R^2 to another point, so also it maps subsets of R^2 to other subsets. In particular, it maps any curve in R^2 to another curve. Given a congruence of curves C, having just one curve through each point, g(t) will always map a member c of C into another curve in R^2, but in general c will not lie in C. If it is the case for every t that g(t) *does* map a curve of C into another curve of C, we will say that C **admits** the group G. Given an arbitrary curve c in R^2 which intersects each streamline in a single point, there is an obvious way to create a congruence C which admits G. To be explicit, let C be the set of images under g(t) of c, as t varies. C will admit G because if k is a curve in the congruence equal to the image of c under g(t), then for another element g(s) of the group, g(s) will take k into $g(s)(g(t)(\mathsf{c})) = g(s + t)(\mathsf{c})$. We need to insert one caveat. C will be a congruence provided that the flow does not have zero velocity at any point. If, as in examples 3 and 4 above, this condition is violated, we must agree to exclude these points from the analysis.

The congruence of streamlines of G admits G, but trivially, each curve being mapped to itself by each element of G. Other examples are more revealing. In example 1, G acts non-trivially on the set of vertical straight lines, and in example 2, on the set of lines parallel to the x-axis. There are other congruences admitting G, which can all be found by taking a suitable curve and taking all its images under G. For instance with example 2, the parabola $y = x^2$ generates the congruence consisting of the set of curves $y = x^2 + c$. In example 3, the set of lines through the origin admits G, and in example 4, the set of circles concentric with the origin admits the group, though again, other congruences of curves also admit G.

Defining a canonical coordinate on R^2

We can use the idea of congruences admitting G to change the (x, y)-coordinates defining R^2 into a pair of what are called "canonical coordinates" for G. As mentioned earlier, there are two alternative routes to solving simple differential equations, those of an integrating factor and the separation of variables. If we were following the second option, it would be possible to use the canonical coordinates to achieve such a separation. The first step in this procedure would be to obtain an equation for the streamlines in the form s(x, y) = c as c varies, and then using c as one of the coordinates.

Here, however, we will focus only on the second coordinate, t, defined in the following three paragraphs. It turns out that this will provide the key to finding the solutions of certain ODEs, namely those whose solution curves admit G. The curves t = constant will in fact provide precisely the solutions to such ODEs. However, some of the ideas can be developed in the more general context of arbitrary congruences admitting G, and it may be simpler to understand what is going on if the principles are first outlined using this approach, narrowing the application to ODEs only in the later stages.

Take an arbitrary curve c in R^2 which is a section of the streamlines, that is, c intersects each streamline in a single point without cutting any streamline twice. Consider the set C whose members are the images of c under the elements of G. As outlined above, C will necessarily, by its very construction, admit G. There is a natural definition of the points at which t = 0: these are just the points lying on the original curve c, which represents the "starting line" for defining t. One can imagine setting the clock running at t = 0, and then observing the way the curve c is carried forward with time under the motion of the fluid. For each value of t, we will see a new curve, that is a new member of C, the image of c under g(t).

As t varies, the images of c will sweep out an area of the plane. By taking also the images under negative values of t, we can see that the whole plane will, in general, be accounted for in this way (though we may have to exclude some exceptional points or regions). For a given value of t, we can observe the set of points (x, y) lying on the appropriate image curve, g(t)(c). These are all given the value t for their canonical coordinate. Conversely, for a point (x, y), we can find the value of t such that (x, y) lies on g(t)(c). The map from (x, y) to this value of t gives a function t(x, y) on the plane.

It is intuitively clear that this map t(x, y) is in general a well-defined function from R^2 into R. In some cases the function is not well defined, as in example 3. In this case we could take c to be the radius extending from the origin along the positive half of the x-axis, but in that case it is not possible to define t at the origin, and t takes multiple values elsewhere, since under the action of G, the images of the radius will cross any given location at multiple time points. However, in this and similar cases we can define t uniquely if certain points are removed, leaving an open region of R^2.

In example 1, take c to be the y-axis, x = 0. Then the lines of constant t are the lines parallel to this, and for a point (x, y) the t-coordinate is x/u.

In example 2, a similar result can be found, with the x and y-axes interchanged.

In example 3, if c is the line (in polar coordinates) $\theta = 0$, the value of t at (r, θ) is just θ; in this case we must exclude the origin and confine the range of values of θ to the semi-closed interval $[0, 2\pi)$.

In example 4, if c is the circle of unit radius centred at the origin, the value of t at (r, θ) is $\ln(r)$.

Finding grad(t) determined by a congruence that admits G

Given that students may not have encountered the idea of a gradient before, it may be worth introducing the topic in heuristic terms. Imagine a relief map which is intended to give the height h(x, y) above sea level at points (x, y) within the area of the map. In practice, this is done by drawing contour lines on the map indicating the locus of points (x, y) where $h(x, y) = c$, the values of c varying in a regular stepwise manner, with a constant difference between neighbouring values. In this context, grad(h) at each point is defined to be the vector which points in the direction of the line of steepest ascent at that point, and whose magnitude is the gradient, in the usual sense, of a path taken in this direction of steepest ascent. One can estimate the direction of grad(h) at any point by taking the direction at right angles to the contour lines in the neighbourhood of that point, and its magnitude is then inversely proportional to the distance between the contour lines there. Anyone familiar with maps will know that "crowded" contour lines mean tough ascents (or descents), whereas widely spaced lines indicate relatively flat terrain. It is not too hard to show that the rate of ascent along a path in a certain direction is equal to the projection of the gradient vector in that direction, and from this it follows that the x and y components of grad(h) are $\partial h/\partial x$ and $\partial h/\partial y$.

This and the following sections are devoted to showing how a function t, related to the group parameter, can be defined for the plane when we are given a congruence C which admits G. In fact, we will not need to know the whole family C of curves explicitly; we need only have the direction of the normal to the curve of the congruence which passes through any given point. From this, we may find an expression for grad(t). We then show how this can be solved for t as a function of x and y using integration: "solution by quadratures" in the classical terminology. In the case where C is the set of solution curves of an ODE compatible with the group, this will yield the solutions of the ODE, but for the moment we do not confine ourselves to this case.

If a curve is used to define the function t as described in the previous section, then the function will clearly depend on the choice of the curve c which is taken as the locus of points for which $t = 0$. However, there is one very useful property which always holds whatever curve is chosen. Clearly, if we have an expression for c, we will expect to be able to find an expression for the curves of the congruence that it generates: they are just the images of c under the actions of G, and from this, t can be calculated. But it turns out that remarkably, we do not need to have an explicit form for the curve c to calculate t. We do not in fact require the "global" detail embodied in the form of an equation for the whole curve. It is sufficient to have some "local" information, in terms of an expression for the *normal* to the curve at the relevant points of the plane.

Looking ahead to the application to ODEs, this fact is important for in this case, the congruence of interest is that of the solution curves to the ODE in question. Given the differential equation, it is a simple matter to find the direction of the normal to the solution curve through each point; this local information is available to us from the equation in the form $Mdx + Ndy = 0$: the vector (M, N) is normal to the vector (dx, dy) and therefore normal to the solution curve through (x, y). (This will not of course in general be the *unit* normal). It is finding the global description, the curve itself, that is the problem.

And yet calculating t(x, y) gives precisely a set of solution curves for the ODE, in the form t(x, y) = const, so that the Lie group provides the mechanism to lift us from the local to the global level. This can be done even if the Lie group is defined in terms of its infinitesimal transformations, that is, its vector field. “Solving” the group by finding the form of its finite (non-infinitesimal) maps g(t) from the vector field is not required. To anticipate, finding the solution curves for the ODE still requires an operation which lifts us from the local to the global level, but this can be done by carrying out two integrations, which is conceptually much simpler than attempting to find the form of the integral curves direct.

We have claimed to be able to find the value of t at each point given the normal to the congruence curve through that point. As an intermediate step, we will find the value of grad(t), the gradient function of t, at each point. Given the gradient, it is relatively straightforward to find t, using a method which is given in the following section, involving the two quadratures mentioned above.

The first step is to observe that since by their definition, the curves of the congruence C are the curves of constant t, representing the “contour lines” of constant t, grad(t) at a given point must be normal to the curve passing through that point. Therefore, if we are given an expression for the direction of the normal to the curve, we already have the *direction* of grad(t). However, this is of little use without also knowing its *magnitude*. For this, we need one additional equation involving grad(t). Fortunately, such an equation can be found. In what follows, we give two derivations of it.

Let **v** be the velocity vector at any point defined by the group action. It is known that for a particle moving with a velocity **v**, the rate at which the value of a function φ(x, y) is changing for the particle as it moves across different values of x and y, is **v**.grad(φ). The rate at which t is changing for a particle of the fluid moving along a streamline is therefore **v**.grad(t). But this must equal one, because the way the function t is defined, the rate at which it is changing for a particle moving with the fluid is the rate at which time is changing per unit of time, which of course is unity, expressing the undoubted truth that time passes at the rate of one second per second. Therefore **v**.grad(t) = 1, and this is the second equation required to determine grad(t).

However, since this derivation is somewhat abstract, we give a geometrical picture which may make the situation clearer.

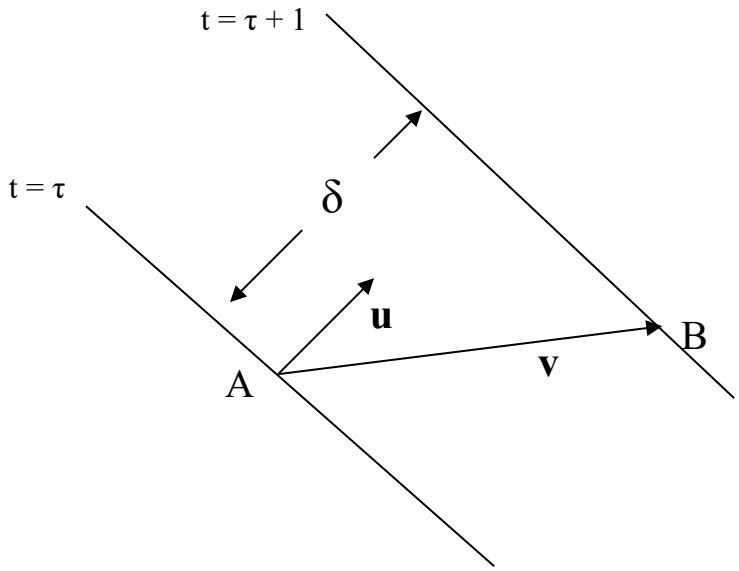

Figure 1: heuristic proof that grad(t) . **v** = 1

Suppose we are given two congruence curves for values of t separated by one unit, represented by the lines in figure 1. For simplicity, assume that v is constant and that the curves are straight lines. This assumption will be dropped below as we go to the limit where the time difference tends to zero, but the calculation is simplified and the essential points made clearer if the time interval is unity. If point A is carried under the flow to point B in unit time, then the distance vector AB is clearly equal to the velocity vector **v**.

grad(t) is proportional to the unit normal to the curve at A, represented by the vector **u**. If the two curves are separated by a perpendicular distance δ, then by the definition of the grad function, grad(t) has magnitude 1/δ. Thus, grad(t) = **u**/δ, or **u** = δ grad(t).

But δ is equal to the component of the vector AB in the direction of **u**, so that δ = **u** . AB, the scalar product of the two vectors, and since AB = **v**, in fact δ = **u** . **v**.

Substituting δ.grad(t) for **u**, we have δ = δ grad(t) . **v**, and therefore grad(t) . **v** = 1.

If instead of taking a unit time interval we take a small interval Δt, in the limit the assumptions that **v** is constant and that the curves are linear are valid, and the derivation goes through with factors 1/Δt which cancel, giving the same result. If a more rigorous proof is needed, this can be provided by using a Taylor's expansion of t(x + uΔt, y + vΔt), where u, v are the components of **v**, and dropping terms of higher order in Δt. Here it is necessary to use the fact that the components of grad(t) are $\partial t/\partial x$ and $\partial t/\partial y$.

Now suppose that we have an expression **w** = **w**(x, y) giving vectors which are normal to the set of curves of the congruence at each point (as noted, these arise naturally in the ODE application). We do not require **w** to be a unit vector. Since grad(t) is normal to the curve through a given point, grad(t) = μ.**w**, for some μ (μ will in general depend on x and y). We also know that grad(t) . **v** = 1. So μ**w** . **v** = 1, and μ = 1/(**w** . **v**).

$$\text{Finally, } \operatorname{grad}(t) = \mathbf{w}/(\mathbf{w} \cdot \mathbf{v}) \qquad (1)$$

Since by assumption both **w** and **v** are known, we have an expression for grad(t). (1) will be crucial to finding the integrating factor for an ODE whose solution curves admit G.

Incidentally, grad(t) is unchanged if we take a different curve k from the congruence as the arbitrary starting point, the locus of points for t = 0. This is because the change is equivalent to adding an arbitrary constant to the values of t for the original curve, which does not alter grad(t).

The scalar factor μ introduced above is referred to as an "integrating factor" in the classical texts (see [2, p. 27]).

Finding t from grad(t)

This section gives the final step in the challenge of "solving for t", that is, of finding the value of the function t everywhere, at least up to the addition of an arbitrary constant (and with the exclusion, if necessary, of certain exceptional points or areas in which the solution may not be well defined). In fact the method of calculating t from grad(t) is perfectly general.

Reverting to the cartographic analogy with which the concept of the gradient was introduced, let h be the height above some datum point, and suppose that one is given the values of grad(h) everywhere on some map. One is required to find h, at least up to an additive constant. It is possible, given grad(h), to reconstruct h in various ways. One might, for example, take line elements orthogonal to grad(h) at each point and join them up in curves, which would represent the contour lines, and then ensure that the spacing of the lines was inversely proportional to the magnitude of grad(h). However, this is not very precise mathematically, and a better method is to use a technique one might call "walking the grid".

Start at some datum point, for example the lower left hand corner of the map, which we take as the origin (0, 0). Given that we know grad(h) and therefore $\partial h/\partial x$ and $\partial h/\partial y$ at each point, we know how the ground slopes up or down, as we move due East and due North. $\partial h/\partial x$ is the slope at each point as we go due East, and likewise $\partial h/\partial y$ is the slope going due North. We want to calculate the height at any given point (x, y), within the area of the map (ensuring that both x and y are non-negative).

Starting at (0, 0), go due East until you reach the point (x, 0), integrating the value of $\partial h/\partial x$ along the path with respect to x as you go. Integrating the slope in the direction of travel gives you the total change in height, by the fundamental theorem of calculus, so the definite integral will give you the value of h(x, 0) – h(0, 0).

Now turn and proceed due North along the path (x, λy), for $0 \leq \lambda \leq 1$, until you reach the point (x, y), integrating the value of $\partial h/\partial y$ with respect to y as you go. As before, the integral of the slope will give the change in height, so you have found h(x, y) – h(x, 0).

Adding this to h(x, 0) – h(0, 0) gives h(x, y) – h(0, 0), and this can be calculated for any point (x, y). So h(x, y) is determined uniquely, up so some fixed constant h(0, 0). In the application where we are trying to find t from a known functional expression for grad(t), this will give the value of t throughout the region of interest, up to an arbitrary constant. Taking a particular curve of the congruence and defining it to be the locus t = 0 will determine the value of t for each point uniquely.

Solution curves of first order ODEs with solution curves admitting G

Before applying Lie group theory to solving some first order ODEs, we need to sketch an intuitive theory of the solution of such equations.

The general form of a first order ODE is f(x, y, dy/dx) = 0. This can often be solved for dy/dx, to give an equation of the form dy/dx = F(x, y). This expresses the fact that if a particular solution passes through the point (x, y), then the slope of the tangent to the solution curve at that point is F(x, y).

Intuitively, we can find the whole solution curve that passes through a specific point by drawing a little line element (dx, dy) parallel to the tangent to the curve there, that is, a line element such that dy/dx = F(x, y). Moving along this line element from the point (x, y) one arrives at the "next" infinitesimally close point of the curve, at (x + dx, y + dy). The equation tells us that the tangent to the solution curve at this point should be a pair (δx, δy) such that δx/δy = F(x + dx, y + dy), and moving along this we arrive at a further point of the solution curve, at (x + dx + δx, y + dy + δy). (This is essentially the argument given in [2, p.13]).

Though not rigorous, this argument should make it plausible that at least for "nice" functions F, there is in general just one solution curve that passes through each point in R^2. In such cases, the set of solutions C is therefore a congruence in R^2.

Now for the payoff. We will show that if we can find a group G such that C admits G, then we can find an explicit solution to the ODE. In fact, this will amount to finding what in traditional language is called an "integrating factor" for F.

At first sight, it may seem that this gets us no further forward, because in order to prove that C admits G, we apparently need to know what curves C consists of, and we can only know that if we have already solved the ODE. However, it is a remarkable fact that for a given G and an ODE dy/dx = F(x, y), we can tell whether C admits G without knowing anything about the individual solution curves.

We will show with an example how this may be done. First, however, let us assume that the ODE admits G and show how this enables us to find the solutions (we will use the phrase "the ODE admits G" as shorthand for "the set of solution curves of the ODE admits G").

An ODE of the form Mdx + Ndy = 0 can be written (M, N).(dx, dy) = 0, where the "." refers to the scalar, or inner, product between the two vectors. If (dx, dy) is viewed as a tangent element of the integral curve at the point (x, y), then this equation tells us that the vector (M, N) is normal to the solution curve at that point.

We have assumed that the ODE admits the group G. Therefore, we know that there exists a way of defining t on the plane such that the curves in the congruence of solution curves are precisely the curves t = const. Since the vector (M, N) is normal to the solution curve passing through a given point, we saw above in equation (1) that

grad(t) = **w**/(**w** . **v**), where in this case **w** = (M, N). If the vector **v** is written **v** = (u, v), then

$$\text{grad}(t) = (M, N)/(uM + vN) \qquad (2)$$

Since all the quantities on the right hand side of this equation are known, grad(t) is also known, from which t can be found by integration, and the solutions to the ODE expressed in the form t(x, y) = const.

Classical texts refer to 1/(uM + vN) as the "integrating factor" for the equation. To see why this term is justified, consider the ODE in the form Mdx + Ndy = 0. Multiplying by the integrating factor yields

$$(Mdx + Ndy)/(uM + vN) = 0.$$

The expression on the left of this equation is the scalar product of the expression on the right hand side of (2) with the line-element vector (dx, dy).

If we take the scalar product of (dx, dy) with the left side of (2), and use the fact that grad(t) = (∂t/∂x, ∂t/∂y), we obtain the expression ∂t/∂x.dx + ∂t/∂y.dy, which is just the derivative dt.

$$\text{So } (Mdx + Ndy)/(uM + vN) = dt \qquad (3)$$

This means that the expression (Mdx + Ndy)/(uM + vN) is in classical language an "exact differential", and the line elements (dx, dy) comprising the solution curve lie on the curve given by dt = 0, i.e. t = const. The two approaches are equivalent, as they must be, and the expression (Mdx + Ndy)/(uM + vN), which is the scalar product of two vectors, conceals the fact that it is the first of these two vectors – the gradient – that is really of interest to us. The usual method of integration of an exact differential involves precisely the same procedure as we have sketched above for the calculation of grad(t).

The use of our notation appears preferable because the expression (3), involving the additional terms dx and dy, is apt to cause confusion, especially when we have been using them as the components of the line element of a solution curve. When the exact differential is integrated to find the solution curves t = const, dx and dy are taken to be perfectly general terms. It seems better to isolate the gradient function as a vector and then integrate it separately. We refer to the classical expressions simply to make the connection between the two treatments, to make it easier to understand the older treatments (see [1, 2]), and to emphasise that our approach is essentially identical.

Demonstration of the method

General methods exist for determining whether a particular ODE admits a given group as we will show below, but meanwhile it may be helpful to show how this may be possible in particular cases from elementary considerations, and how the method will then enable us to apply the integration procedure to find the family of solution curves.

The group of dilations

Consider the case of the group of dilations in example 4. We will show that any ODE of the form dy/dx = f(y/x), which is the definition of the so-called homogeneous equations, is invariant under G.

Let us take at random a line element AB in the plane corresponding to the ODE, at the point A (see Figure 2). For the purposes of this illustration, I will assume A is located in the top right-

hand quadrant, and that the slope of the line element is negative, simply to make for a clearer diagram.

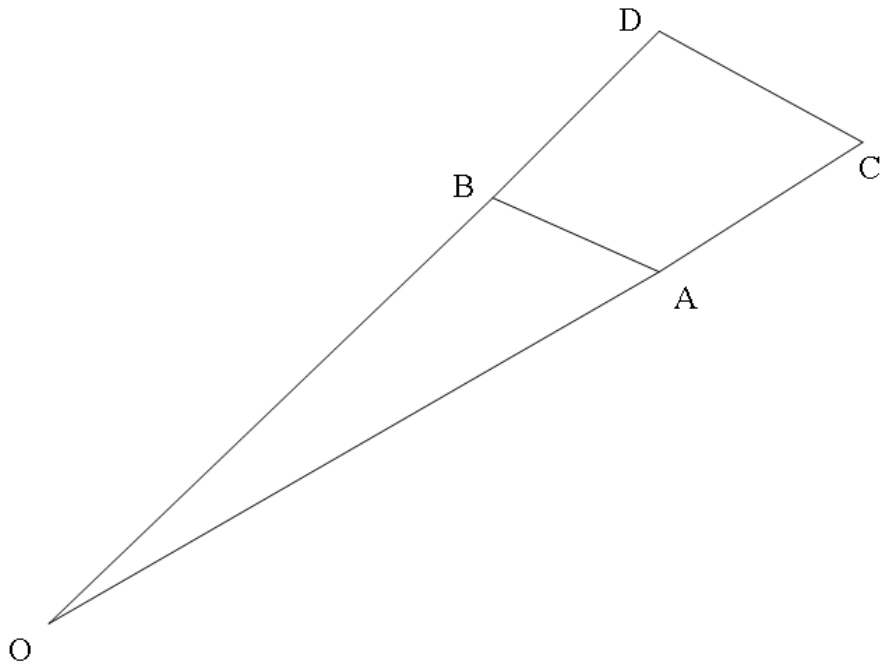

Figure 2: finding equations invariant under the group of dilations

Imagine the angle AOB to be small, so that the side AB approximates to a small line element. Consider the image of triangle OAB under the action of an element g of the group of dilations. O is fixed by g, and the lengths of sides OA, OB are multiplied by the same factor under g. Therefore the image of OAB under g will be a triangle OCD which is similar to OAB, and so CD, the image of line element AB, will be parallel to AB.

Suppose now that the line element AB is part of the solution curve through A for the ODE considered earlier. Then the slope of AB is f(y/x), where A is the point (x, y). Let C be the point (x', y'). Then y/x = y'/x', since O, A, C are collinear, and the slope of CD is equal to the slope of AB as shown above. Therefore

$$f(y/x) = f(y'/x'),\ dy/dx = dy'/dx',$$

and if dy/dx = f(y/x) then dy'/dx' = f(y'/x').
Therefore the line-element AB is carried by g into another line element CD of the solution curve, and the ODE admits G.

Converting the ODE to the form Mdx + Ndy = 0, it takes the form

$$-f(y/x).dx + dy = 0.$$

Recalling that in Example 4 the vector field of G is given by **v** = (x, y) at the point (x, y), we find that the integrating factor given by the formula arising from the group action is the inverse

of the scalar product $(-f(y/x), 1)\cdot(x, y) = (y - xf(y/x))$, so that the formula (2) for grad(t) becomes

$$\text{grad}(t) = (-f(y/x).(y - xf(y/x))^{-1}, (y - xf(y/x))^{-1})$$

and this gives the solutions in form t(x, y) = const, where

$$\partial t/\partial x = -f(y/x).(y - xf(y/x))^{-1} \quad \text{and}$$

$$\partial t/\partial y = (y - xf(y/x))^{-1}$$

The solution can be found by two integrations.

Helgason [3] cites this example and asks the reader to find the solution for the specific case $f(z) = z^2 + 2z$. I will show how this can be found as an illustration of the method. (Note that [3] uses the notation U(x, y) in place of our t(x, y). [2] uses the notation $\Omega(x, y)$ to represent this concept, and uses U to represent the infinitesimal transformation of the Lie group).

For $f(z) = z^2 + 2z$, the equations for t become

$$\partial t/\partial x = (2x + y)/(x^2 + xy) \quad \text{and}$$

$$\partial t/\partial y = -x/(xy + y^2)$$

We will apply the method of finding t at a general point suggested above. We employ the terminology (X, Y) for the point at which we wish to calculate t. The obtrusive use of capitals is adopted here to avoid confusion with the variables used in the integrations used to find the form of t. The first step is to integrate the first of these two expressions, that for $\partial t/\partial x$, with respect to x, to find the value of the definite integral between the points (0, 0) and (X, 0), and then integrate the second term, for $\partial t/\partial y$, between the points (X, 0) and (X, Y).

In this case, the procedure gives problems if the lower limit of the integral involves values of x or y = 0, giving expressions containing ln(0). But this is not a problem for us since the value of t is only determined up to an additive constant, so we can take the initial point to be another value, say (1, 1), for which the integral is well behaved.

The indefinite integral of the expression for $\partial t/\partial x$ for which y = 1 is $\ln(x^2 + x)$.

So its integral between (1, 1) and (1, X) is

$$\ln(X^2 + X) - \ln(2)$$

To find t(X, Y) we have now to integrate $\partial t/\partial y$ between (X, 1) and (X, Y) and add it to the expression just obtained.

The indefinite integral is $\ln(X + y) - \ln(y)$, as can be easily checked, giving for the definite integral of $\partial t/\partial y$ the expression

$$\ln\{(X + Y)/(XY + Y)\}$$

Adding these two expressions yields $t = \ln\{(X^2 + XY)/Y\} - \ln(2)$,

And the general form obtained by setting this equal to an arbitrary constant and simplifying is

$$X^2 + XY = cY$$

Whence $Y = X^2/(c - X)$, the result cited in [3].

It can be shown without too much difficulty by examination of Figure 2 that any ODE that admits the group of dilations must be of the form $dy/dx = f(y/x)$ for some function f, so confining ourselves to equations of this form exploits the full power of the method of Lie groups in this case.

The group of rotations

It is simpler to use the notation in [2] and refer to the line-element dy/dx henceforth as p. Now consider the group of example 3, that is, the group G of all rotations about the origin. What can we say about the general form of an ODE which admits this group?

One way to approach this problem is to observe that any non-identity element of G maps OA onto a different ray OB. It can never map a point of the ray OA to a different point on OA. Therefore, if we are trying to construct an ODE on which G acts invariantly, we are free to specify the value of p along a ray. For the sake of clarity, suppose we choose the positive x-axis as our ray OA. Define p along OA to be a smooth function f(x) of x.

It is simpler to work from now on in polar coordinates. We have specified p along OA as a function f(r) of the distance r. The line element $p = f(r)$ at $(r, 0)$ has a direction, namely the angle $\arctan(p) = \arctan\{f(r)\}$. Consider some different ray OB. There is a unique element g of G which rotates OA into OB, through an angle θ say. Evidently g takes the point $(r, 0)$ at a distance r from the origin on OA, to the point (r, θ) on OB. What does it do to the line element at $(r, 0)$? Clearly it rotates it by the same angle, θ.

Therefore the angle of the line-element at (r, θ) must be $\arctan(p) = \arctan\{f(r)\} + \theta$.
But in Cartesian coordinates, $\theta = \arctan(y/x)$

$$\text{So } \arctan(p) = \arctan\{f(r)\} + \arctan(y/x)$$

Taking tangents of both sides and using the usual expansion for $\tan(A + B)$, we get

$$p = (f(r) + y/x)/(1 - f(r)y/x)$$

Substituting for r in terms of x and y, and writing F for the function given by taking the square root followed by f, and simplifying,

$$p = (xF(x^2 + y^2) + y)/(x - yF(x^2 + y^2)) \qquad (4)$$

Rearranging and solving for the expression $F(x^2 + y^2)$, this can be written equivalently as

$$F(x^2 + y^2) = (xp - y)/(x + yp)$$

which is the form for the general equation invariant under the rotation group (as derived in [2, p. 111]).

In order to solve such equations, the previous method can be used to find an expression for grad(t), or in terms of the presentation in [2], to find an integrating factor.

Our expression (2) gives

$$\text{grad}(t) = (M, N)/(uM + vN).$$

In this case, for the rotation group $(u, v) = (-y, x)$. When the equation (4) is recast in the form $Mdx + Ndy = 0$, we find that $M = -(xF(x^2 + y^2) + y)$, $N = (x - yF(x^2 + y^2))$, and $uM + vN$ simplifies to $(x^2 + y^2)$, so the integrating factor is $(x^2 + y^2)^{-1}$ (as in [2, p.111]).

The solution curves in the form t = const. could now be found in general terms by integrating with respect to x and y as outlined above. However, in this case it is more illuminating to solve for t from grad(t) by using polar coordinates.

We will write $F(x^2 + y^2)$ in the form f(r).

$\partial t/\partial r$ is the component of grad(t) in the direction of the unit radius vector
$\mathbf{r}/r = (x/r, y/r)$, i.e., it is the scalar product grad(t).(x/r, y/r).

Using (2), this in turn is the scalar product $((-xf(r) - y)/r^2,\ (x - yf(r))/r^2).(x/r, y/r)$,

which simplifies to $-f(r)/r$.

$1/r.(\partial t/\partial\theta)$ is the component of grad(t) along the orthogonal unit vector in the direction of increasing θ, and since this unit vector is $(-y/r, x/r)$, we have

$$1/r.(\partial t/\partial\theta) = ((-xf(r) - y)/r^2,\ (x - yf(r))/r^2).(-y/r, x/r).$$

The terms in f(r) cancel and we find that $\partial t/\partial\theta = 1$.
It turns out that we have separated the variables in polar coordinates (a more advanced treatment would show that r and θ are canonical coordinates for the rotation group and that this outcome is a predictable consequence of that fact).

Solving for t is now comparatively simple: in fact, $t = \theta - \int f(r)/r.dr$, and the solution curves are

$$\theta = \int f(r)/r.dr + \text{const.}$$

As an illustration, suppose f(r) is identically equal to a constant α. We find the solution curves to be

$$r = ce^{\theta/\alpha}.$$

This is a family of equiangular spirals, which can also be seen from first principles, since the condition on f(r) means that the line-elements along the x-axis are all at a constant angle α to the horizontal, and therefore the line elements along a radius vector are therefore all at that

same angle to the radius vector. It is well known that this is sufficient to specify the family of equiangular spirals.

The extended group G'

We have already shown the effectiveness of the method, and how in particular cases it may be possible to derive the general ODE admitting G from first principles, given G. We will now sketch the method whereby, given G, one may find the form of any ODE admitting it, and therefore soluble by quadratures. We supply no more than the barest outline of how this is carried out in practice: for demonstrations of its use, we refer the reader to [1] and [2].

We may visualize the effect of g(t) on R^2 by imagining that a little cork is placed at a given point P, and noting the point P' that it has reached after a time t. This point can be described via the formula P' = g(t)(P). Similarly, the effect of g(t) on a curve c can be visualized as the final location of a whole series of corks placed along that curve, when the corks flow along the streamlines for time t. Under this flow process, the tangent to c at P will be transformed into a tangent to the image of c at P'. It will not of course in general be a tangent in the literal sense of the word, because it will no longer be a straight line, but it will at least be a curve that is tangent to the image of c at P'. In fact, if two curves c and d are tangent to one another at P, it is not hard to see, at least heuristically, that their images must be tangent to one another at P'.

This means that the elements of G not only act on the points of the plane, but they also act on **line-elements** of the plane. A rigorous approach would require us to define a line-element as an equivalence class of curves tangent to a give curve at P; the action of elements of G on the set of such line elements would be well defined since G preserves tangency. The images of the curves under elements of G would transform line-elements at P into line elements at P'. But at our present basic level, similar to [2, p. 13], we merely define a line-element at a point as being a “direction” at a point which may be written in the form of an infinitesimal vector (dx, dy). If this is tangent to a curve c at (x, y), then the image of the line element under g(t) is an infinitesimal vector tangent to the image of c at P'.

We will not require to know the magnitude of a line element (indeed, the magnitude of an infinitesimal vector makes little sense) but simply its direction. In fact, the classical approach (see [1, 2]) refers to the line elements as dy/dx, in which only the ratio and not the magnitudes of the elements are relevant. This is really a misuse of mathematical language, because the line elements are not referring to the derivative of a function y dependent on x, but merely to a direction defined at points of the plane. In order to remove the unwanted association between line elements and the operation of differentiation, the classical treatments adopt the convention of referring to line elements using the variable p.

It is not hard to see that as the group G acts on the elements of the plane, it also acts on the line-elements. The group G, conceived as acting on the line-elements in addition to the points of the plane, is referred to as the extended group, and written G'. A very approximate idea of the action of G' can be obtained by imagining a line-element to be a tiny straw placed in the flow, which is not only carried from one position to another but also perhaps rotated by the action of the flow, a consequence of the flow vector being slightly different at the two ends of the straw.

The classical treatment proceeds by deriving an equation for the action of group elements on the line-lements p. This may be found by considering the action of an “infinitesimal

transformation" acting over a small time period under which the flow takes the point (x, y) to (x + uδt, y + vδt).

A line-element can be visualized as the point pair (x, y) and (x + dx, y + dy). The action of G' is defined to be the image of this point pair under the action of G on the two points of which it is composed.

The infinitesimal flow just described takes the neighbouring point (x + dx, y + dy) to the point (x + dx + u'δt, y + dy + v'δt), where the velocities u', v' are evaluated not at (x, y) but at (x + dx, y + dy). If p can be conceived as the point pair comprising (x, y) and (x + dx, y + dy), the image p' of p is therefore the point pair comprising (x + uδt, y + vδt) and (x + dx + u'δt, y + dy + v'δt). The value of p' is then found by taking the difference between the y-coordinates of these points and dividing by the difference betwen their x-coordinates.

This is a tedious but essentially elementary exercise in evaluating first order expansions, discarding all terms of higher order. It uses the identity $u' = u + \partial u/\partial x.dx + \partial u/\partial y.dy$, and a similar expression for v'. We find

$$p' - p = \delta t\{\partial v/\partial x + p(\partial v/\partial y - \partial u/\partial x) - p^2\partial u/\partial y\}, \text{ whence}$$

$$\delta p/\delta t = \partial v/\partial x + p(\partial v/\partial y - \partial u/\partial x) - p^2\partial u/\partial y = w, \text{ say.}$$

Looking at the group G in terms of its action on points in the plane, if the velocity vector at some point is (u, v) then the rate of change of the position vector (x, y) can be written as $\delta x/\delta t = u$, $\delta y/\delta t = v$. In the same way, the rate of change of p is $\delta p/\delta t = w$, and we may regard the triple (u, v, w) as a vector field acting on the points of the extended (x, y, p)-space in the same way as the original vector field acts on the points of the real plane. Integrating the flow along the streamlines in this three dimensional space then gives the action of the one-parameter extended group G'.

Suppose now we have an ODE in the form p = f(x, y). G' acts on both sides of this equation. The question of whether the ODE admits G can now be found by evaluating the action of an infinitesimal tranformation in G' on x, y and p, giving new elements x', y' and p', and checking whether or not it is the case that p' = f(x', y') or not. If this equation holds then the ODE admits G, otherwise not.

Finding the general form of the ODE that admits G

We supplement this sketch by the merest suggestion of how the final step can be taken forward. Consider the three dimensional (x, y, p)-space where p is the third, vertical coordinate. In this space, a first order ODE of the form p = f(x, y) is in general represented by a surface, namely the graph of the function f. A given point of this space can be seen in two ways. It is both a point in (x, y, p)-space and also a point of the plane coupled with a line-element, corresponding to the given value of p. If these two aspects are identified, then since any element g' of the extended group G' acts on line-elements, it also acts on points of (x, y, p)-space.

Just as G gives rise to streamlines in the plane, which are the orbits of points under the group G, the group G' generates similar streamlines in (x, y, p)-space. The projections of these onto the (x, y)-plane are just the original streamlines generated by G. Just as in the case of the plane, no two G'-streamlines cross, and in fact the streamlines fill the space in a non-overlapping manner. They form, in other words, a partition of the space. One might imagine

this as being like an infinitely large optical fibre cable running throughout the space, with each orbit under G' comprising one fibre.

The payoff for this geometrical approach is that one can now describe precisely what it means for an ODE to be invariant under G. Take a specific ODE, $p = f(x, y)$, with its associated surface in (x, y, p)-space. Take a point (x_0, y_0, p_0) on this surface. The action of a group element g' on this point is represented by shifting the point (x_0, y_0, p_0) a certain distance along the streamline for G' that passes through this point.

Suppose the shifted point has coordiates (x_1, y_1, p_1). This point lies on the surface if and only if $p_1 = f(x_1, y_1)$, that is, if and only if the group element g' takes the original line-element of the ODE represented by (x_0, y_0, p_0) onto another line-element for that ODE. Varying the element g', we can see that the elements of G' take a point on the surface to other points on the surface, if and only if the G'-streamline passing through the original point lies wholly within the surface.

This argument carries through whatever the original point (x_0, y_0, p_0), and it shows that the ODE is invariant under the group G if and only if the ODE surface consists of nothing but G'-streamlines: it is the union of a set of streamlines. Seen another way, for an invariant ODE, a streamline either lies wholly within the surface of the ODE, or does not intersect it at all.

Using the optical fibre analogy, an ODE invariant under the group can be created by taking a subset of the fibres given by G' in the (x, y, p)-space and taking the union of all the points on the fibres, provided this union gives a surface of points (x, y, p) corresponding to a one-valued function $p = f(x, y)$, and provided this surface satisfies the usual smoothness requirements for an ODE. A general surface in (x, y, p)-space will cut across the fibres: this cut surface will be a two-dimensional array of the cut ends of the optical fibres. By contrast an invariant ODE will give a surface which does not cut any fibres. The appearance of the fibres on such a surface is not an array of "cut ends", but a series of fibres lying wholly within the surface.

We are now reduced to finding ways to generate suitable sets of streamlines to generate our surface. To do this rigorously would require some strenuous mathematics, but given that we are attempting to provide a heuristic rather than a perfectionistic proof, the following outline may suffice. Suppose then that we have an invariant ODE and its corresponding surface in (x, y, p)-space. It consists of the union of a set of streamlines. What can we say about this set? Avoiding technicalities involving manifolds we can describe the surface heuristically as a two-dimensional "thing", and the individual streamlines as one-dimensional "things". How do we choose a set of one-dimensional things that add up to a two-dimensional thing? The same way that we would choose a set of zero-dimensional things (points) to add up to a one-dimensional thing (a curve or path): by taking a one-parameter set of them. That is, if the streamlines generated by G' that fill (x, y, p)-space are regarded as "points", we want to somehow generate a "curve" out of these points.

Using the fibre analogy for these streamlines, imagine cutting across the fibres to examine them. We have a two-dimensional array of "cut ends". Now draw a one-dimensional smooth curve across the cut ends, giving a one-parameter set of fibres, namely those whose cut ends lie on the curve. If we now take this set of fibres for our surface, we can guarantee at least that it satisfies the conditions for G' to leave it invariant.

The problem now reduces to this: how do we generate all the possible one-parameter sets of streamlines of G'? The answer was hinted at in the cut fibre bundle analogy. The whole set of streamlines is a two-dimensional “thing”. If we consider the surface obtained by cutting across the fibres, and identify a “cut end” with the fibre of which it is a part, then the streamlines can be concretely represented as this two-dimensional surface. If we can actually find two parameters that represent these two dimensions, say α and β, then each streamline is represented by a pair of numbers (α, β). Suppose we now draw a curve on this (α, β)-space of streamlines. This will give a one-parameter subset. The points lying on this subset of streamlines in (x, y, p)-space will now be a surface, and provided it satisfies certain smoothness conditions it will correspond to an invariant ODE.

The question now resolves into: how do we describe a general curve on (α, β)-space? The answer given in [2] is that we take a smooth function H and look at the streamlines corresponding to the graph of H, i.e. those for which $\beta = H(\alpha)$. This is certainly a one-parameter set of streamlines. Admittedly this procedure may not give all the one-parameter subsets but it will sweep up a useful chunk of them. The final step is to find out how to parametrize the streamlines.

The group G' generates a vector field in (x, y, p)-space, which was written above as having components (u, v, w) where u, v are the velocity components given by G and w is given by $\partial v/\partial x + p(\partial v/\partial y - \partial u/\partial x) - p^2\partial u/\partial y$. Therefore the motion of a point of (x, y, p)-space can be described as moving in the direction given by this vector. We could say that if this motion is given by the infinitesimal translation (dx, dy, dp), then (dx, dy, dp) is a vector proportional to (u, v, w), or in the notion used in [2],

$$dx/u = dy/v = dp/w \qquad (5)$$

Note: if any of u, v or w is zero, this involves an illicit division. However, we can regard this terminology as simply a shorthand for the statement regarding the proportionality of vectors, given in the previous sentence.

The classical method (see [2, pp. 104-105]) proceeds as follows.

Taking the first pair of equations in (5), solve the equation $dx/u = dy/v$, i.e. $dy/dx = v/u$, and obtain a solution in the form $U = \alpha$. U will be a function of x and y, but not p.

Now obtain a distinct solution to the equations (5): in practice this will be done by solving the pair $dx/u = dp/w$, or the pair $dy/v = dp/w$. The solution has the form $V = \beta$, where V will now involve p explicitly.

Ince [2] gives little motivation for what follows, but essentially the argument appears to be that for each value of α and β, the corresponding equations $U = \alpha$ and $V = \beta$ represent surfaces in (x, y, p)-space. Since the equations (5) represent the streamlines under G' in this space, these surfaces consist of unions of streamlines. The intersection of two such surfaces gives a single streamline, and the parameters α, β specify this streamline. They can therefore be regarded as its coordinates, and as the parameters vary, they give rise to the whole set of streamlines. We are looking for "one-dimensional" sets of streamlines, which correspond to ODEs invariant under G, as remarked above. Such one-dimensional sets arise when the two parameters α and β do not vary independently, in other words, when they are functionally dependent on one another, say $\beta = H(\alpha)$ for some function H.

By virtue of the equations $U = \alpha$ and $V = \beta$, this translates into the equation

$$V = H(U) \tag{6}$$

or $V - H(U) = 0$, which is the form derived in [2, p. 105] (there the notation uses the letters u, v rather than U, V, but in this paper the lower case letters have been pre-empted above by their use in describing the velocity field). The function H is arbitrary, and this gives rise to the family of ODEs admitting G.

The method can best be demonstrated by applying it to the earlier examples.

Example 1.

In this case, it is evident that G' acts as the identity on line-elements p, so the streamlines in the (x, y, p)-plane are horizontal lines parallel to the x-axis. (u, v, w) = (u, 0, 0) so the equations (5) have the form $dx/u = dy/0 = dp/0$.

Taking the first two equations we find y = const., and taking the first and third gives p = const. The condition (6) now gives $p = H(y)$, which describes line-elements that depend only on y, and are therefore invariant under translation horizontally.

Example 2.

This, by a similar derivation, gives equations of the general form $p = H(x)$.

Example 3.

For the rotation group, we have (u, v) = (-y, x) and so $w = 1 + p^2$.

Therefore (5) yields

$$dx/-y = dy/x = dp/(1 + p^2).$$

From this, we can derive a solution in the form

$$(xp - y)/(x + yp) = F(x^2 + y^2)$$

(see [2, pp. 110-111] for details). We have already derived this form above from first principles in a previous section.

Example 4.

In this case (u, v, w) = (x, y, 0) and the integration is simpler. The first pair of equations in (††) gives y/x = const, and the final one, dp = 0, gives p = const, so we obtain the result as $p = F(y/x)$ (see [2, pp. 108-109]).

Discussion

Why does the approach work? It might at first appear that when attempting to solve a difficult problem, it can only make the problem harder if one imposes additional constraints on the possible solutions. Paradoxically, it can, on the contrary, sometimes make it easier. In the case of an ODE, the difficulty is that there are *too many* functional forms for the solution curves. All these functions may give rise to the same set of solution curves, but in order to solve the ODE, we must come up with some particular one of these functions. For example, the equation $dy/dx = 1$ is satisfied by functions of the obvious form $y - x = c$, but it is also satisfied by $\ln(y - x) = c$, $e^{(y-x)} = c$ and many others. We might liken the problem to that faced by a predator seeking its prey among a shoal of fish, a flock of birds or a herd of animals. There are too many functional forms for the solution in the one case, and too many targets in the other (which is precisely why prey animals adopt this flocking behaviour, to baffle predators). To be successful, a predator must somehow identify a particular individual for pursuit. The Lie group performs the same task: it singles out a particular function which gives the solution curves to an ODE which admits the group.

Bertrand Russell defined the analytic method as assuming the answer to a problem to be known, and then acting on the consequences of this assumption. Applied to an ODE, this means assuming that a solution of the ODE is known in the form $f(x,y) = \text{const.}$, and then finding properties that f must satisfy. If sufficiently many such properties are specified, then f might be unique, and in that case there is a good chance that it can be calculated. However, if there are too few constraints as to the form of f, this may be difficult. In this case, making the problem apparently harder by specificing that f must satisfy additional constraints, may limit the possible functions to lie within a small set, whose nature can be explicitly described. If there are enough constraints such that the form of f is specified uniquely, then one can often deduce sufficient information about this unique form, to enable it to be evaluated. All that is needed then is to verify that one or all of these explicit functions satisfies the conditions of the original problem.

But this involves narrowing down the choice of functions so that the conditions specify a function uniquely. The conditions may be too onerous, and such a function may not exist. But if the limitation is just right, the narrowing down process may give rise to a function which is both uniquely specified, and discoverable. This is the case with Lie's approach. Here, specifying that a solution is invariant under a certain Lie group implies, as we have seen, that it must satisfy sufficiently many constraints to be unique up to the addition of a constant. All that is then needed is to show that the function does indeed satisfy the original equation. For an arbitrary Lie group this will not in general be the case but if the group leaves the equation invariant, then an invariant solution must indeed satisfy the equation. And this is the heart of the method.

To attain a deeper understanding of the effect of the Lie condition requires more mathematical sophistication than can be expected in the student audience for this approach, but it might be worth indicating for readers of this paper who would appreciate the context. Very briefly, given a one-parameter Lie group G acting on R^2, one can look on the plane as having an "R-action" defined upon it. Namely, given a point (x, y), the element τ of R acts to take it to the point $(g(\tau)x, g(\tau)y)$. There is also a natural R-action on R itself, which takes r to $r + \tau$. An element of R acts on points of R^2 by "translating" them along streamlines, and acts on points of R by translation in the normal sense of the word.

The Lie constraint can now be stated as follows: limit the acceptable solutions f(x, y) = const. of the ODE to those in which f, considered as a map from R^2 into R, is a *homomorphism* for this R-action.

This is just a restatement of the condition that:

$$\text{If } f(x, y) = c, \text{ then } f(g(\tau)x, g(\tau)y) = c + \tau.$$

It is not hard to see that this is equivalent to the definition of f given above as t, which can be restated, taking c = 0, in the form

$$f(g(t)x, g(t)y) = t(x, y).$$

References

1. A. Cohen, *An introduction to the Lie theory of one parameter groups, with applications to the solution of differential equations*, D. C. Heath and Co. (1911).
2. E. L. Ince, *Ordinary differential equations*, Longmans, Green and Co. (1927).
3. S. Helgason, Sophus Lie's approach to differential equations. IAP lecture (2006).