LEGIBILITY NOTICE

A major purpose of the Technical Information Center is to provide the broadest dissemination possible of information contained in DOE's Research and Development Reports to business, industry, the academic community, and federal, state and local governments.

Although a small portion of this report is not reproducible, it is being made available to expedite the availability of information on the research discussed herein.

ORNL/TM-10655
Dist. Category UC-20

ORNL/TM--10655

DE88 006047

Fusion Energy Division

NONLINEAR DIFFERENTIAL EQUATIONS

L. Dresner

Date published: January 1988

Prepared by the
OAK RIDGE NATIONAL LABORATORY
Oak Ridge, Tennessee 37831
operated by
MARTIN MARIETTA ENERGY SYSTEMS, INC.
for the
U.S. DEPARTMENT OF ENERGY
under contract DE-AC05-84OR21400

CONTENTS

ABSTRACT

This report is the text of a graduate course on nonlinear differential equations given by the author at the University of Wisconsin–Madison during the summer of 1987. The topics covered are

- direction fields of first-order differential equations,
- the Lie (group) theory of ordinary differential equations,
- similarity solutions of second-order partial differential equations,
- maximum principles and differential inequalities,
- monotone operators and iteration,
- complementary variational principles, and
- stability of numerical methods.

The report should be of interest to graduate students, faculty, and practicing scientists and engineers. No prior knowledge is required beyond a good working knowledge of the calculus. The emphasis is on practical results. Most of the illustrative examples are taken from the fields of nonlinear diffusion, heat and mass transfer, applied superconductivity, and helium cryogenics.

PREFACE

In his book *How to Solve It*, George Polya gives a short caricature of the traditional mathematics professor. According to Polya, when faced with a differential equation, the traditional professor says, "In order to solve this differential equation you look at it till a solution occurs to you." This advice is comical, as Polya intends it to be, because it is really no advice at all: it gives no clue as to how to proceed; it applies to everything and solves nothing. In fact, it fits nicely a second dictum of the traditional professor: "This principle is so perfectly general that no particular application of it is possible."

Unfortunately, authors with serious intentions sometimes speak with words close to those of Polya's traditional professor. Consider, for example, the following passage: "It is thus apparent that the first objective in the study of a nonlinear equation is to ascertain whether or not a solution can be obtained either explicitly or implicitly in terms of classical functions. The procedure in such a study is to discover a transformation which will reduce the equation to some type that is known to have a solution of the desired kind. Failing this, one seeks a transformation which will reduce the equation to one that is *asymptotic* to a form solvable by known functions." The author of this says nothing about how to find such transformations, so that this advice is as insubstantial as that of the traditional professor. His illustrative example only deepens the mystery:

"[An] example [of the second procedure] is furnished by the following nonlinear equation:

$$\frac{dy}{dx} = y^2 + x \tag{8}$$

upon which we make the following transformation of both the dependent and the independent variables:

$$x = \left(\frac{3}{2}t\right)^{2/3}, \; y = \sqrt{x}\, w \tag{9}$$

Equation (8) is then reduced to the following:

$$\frac{dw}{dt} + \frac{1}{3}\frac{w}{t} = w^2 + 1 \tag{10}$$

which, as t increases, is asymptotic to the equation:

$$\frac{dw}{dt} = w^2 + 1 \tag{11}$$

"The solution of Eq. (11) is the function $w = \tan(t - t_0)$ and we can infer, therefore, that w, the solution of Eq. (10), is asymptotic to this function." Polya says something in the preface of his book that I am sure expresses the reader's reaction at this juncture: "Yes, the solution seems to work, it appears to be correct; but how is it possible to invent such a solution?" On this point, our author is silent. He makes his magic passes and leaves us convinced but mystified.

Polya again: "A derivation correctly presented in the book or on the blackboard may be inaccessible and uninstructive, if the purpose of the successive steps is

incomprehensible, if the reader or listener cannot understand how it was humanly possible to find such an argument, if he is not able to derive any suggestion from the presentation as to how he could find such an argument by himself." *Accordingly, the first aim of this book is so to present the material that the reader will always feel that the subject is unfolding naturally along a path he himself might easily have followed.*

Another way the traditional mathematics professor hampers understanding is by leaving vital aspects of the problem to be finished by the reader. Our author does this when he says he regards "a linear differential equation as solved, if its solution can be reduced to the quadrature of a known function, even though the quadrature cannot be expressed simply in terms of the classical algebraic or transcendental functions, [and] regard[s] a nonlinear equation as solved, if it can be reduced to the solution of a linear equation, even though the solution is not explicitly reducible to the classical functions." In this book, on the contrary, a problem is not considered solved until the nature of the solution can be seen, in Polya's often-repeated words, "at a glance." *Presenting the material in such a way as to keep it always clear at a glance is by no means easy, but it is a burden I cheerfully accept.*

This brings us to the matter of rigor. Here, too, I take guidance from Polya, who recommends what he calls incomplete proofs "as a sort of mnemotechnic device ... when the aim is tolerable coherence of presentation and not strictly logical consistency." After all, he says, "the facts must be presented in some connection and in some sort of system, since isolated items are laboriously acquired and easily forgotten. Any sort of connection that unites the facts simply, naturally, and durably is welcome here ... proofs may be useful, especially simple proofs." *So I place clarity before rigor and strive for simplicity and directness of proof.*

What about the choice of subject matter? Here the guiding principle has been breadth of application. Accordingly, in the first chapter on first-order ordinary differential equations, I have stressed *analysis of the direction field* because it can be done for any first-order equation. Strangely, one rarely sees this subject dealt with in courses on differential equations, yet in the truest sense it deals directly with the soul of the differential equation (if one may be permitted to speak thus). Perhaps the incapacity of the present generation of technologists to deal with differential equations stems from neglect during their training of such fundamental matters as the direction field in favor of more advanced but less useful knowledge. There is a tendency in teaching these days, which I shall strive to avoid, to despise the elementary.

In the middle chapters of this book, I concentrate on *the Lie theory of differential equations*. As I have said in another book, I believe that because of its broad applicability, this theory should become a practical workhorse for handling nonlinear differential equations. Strangely, too, one rarely sees this subject in courses on differential equations, although it was invented specifically for solving them a century ago by the Norwegian genius Sophus Lie. I never heard any mention of it during my own education and only learned of it later when, by pure chance, I came across Cohen's 1911 book while browsing in Oak Ridge National Laboratory's library. The ideas I found in that old book electrified me and convinced me that Lie's theorems could be applied widely and with tremendous effect (as I hope this

book will show) by reducing second-order differential equations to first order. The latter can then be treated by the graphical means of Chap. 1.

It was fortunate that in applying Lie's theory I ignored Jacobi's usually correct advice always to generalize and instead started off by concentrating on the affine (stretching) groups, which in my experience were the ones that showed up most often in applications. This led me to discover some useful properties of partial differential equations invariant to families of affine groups. The upshot of all this work is to allow calculation of *similarity solutions of a broad class of second-order partial differential equations* by successive reduction, first to second-order ordinary differential equations and then to ordinary differential equations of the first order. This method has been described in detail in an earlier monograph—here it is described fully but with fewer illustrative examples.

The wide applicability of the ideas mentioned above (analysis of the direction field and Lie theory) arises from their being rooted in very general strategies, namely graphical analysis and exploitation of symmetry. Another broad general strategy is to look for information in the form of inequalities when equalities are too difficult to obtain. Certain methods are available for this purpose, and they form the third main division of this book. They deal with *monotone operators, differential inequalities, maximum and minimum principles, and complementary variational principles.*

An early version of this book was used as the text of a graduate course that I gave in the summer of 1987 at the University of Wisconsin in Madison. Great efforts were expended in getting it ready on time by the staff of the Fusion Energy Division of Oak Ridge National Laboratory. I wish to note for special thanks Sandra Vaughan and Kathy Zell, who initially transcribed my handwritten notes; Darcus Johnson and Brenda Smith, who typed the entire text, including the many complicated equations; Jane Parrott and her graphics staff, who drew the figures; and Bonnie Nestor, who edited the text.

Lawrence Dresner
Oak Ridge, Tennessee
November 1987

Chapter 1

ANALYSIS OF THE DIRECTION FIELD OF FIRST-ORDER ORDINARY DIFFERENTIAL EQUATIONS

"It adds a precious seeing to the eye."
—W. Shakespeare
Love's Labour's Lost

1.1 After having criticized another author's treatment of the first-order differential equation

$$\dot{y} = \frac{dy}{dx} = y^2 + x \ , \tag{1}$$

I feel compelled to start by making good my boast that I can present a heuristic treatment that will at every stage be clear "at a glance."

The entire content of a first-order differential equation can be epitomized by its direction field, a drawing in which is plotted at every point (x, y) a short line segment having as its slope the value dy/dx calculated from the differential equation. The integral curves that satisfy the differential equation must be everywhere tangent to these line segments. Figure 1 shows the direction field of Eq. (1). By letting the eye sweep along the line segments in the direction they indicate, it is possible to form an immediate impression of what the integral curves are like.

In these days of powerful computers and computer graphics, it is no trouble to produce a direction field like that of Fig. 1 (which was obtained on a time-share VAX 8600 in a couple of seconds). Since the direction field is a logical equivalent of the differential equation, one might say that the problem of the first-order differential equation is entirely solved and that analytic techniques for the treatment of the direction field are obsolete. There is good deal of truth in this, but, in my opinion, the time has not quite arrived when the analytic techniques are as obsolete as flint knapping. So I shall turn back the clock to 1917 and consider the method described by S. Brodetsky (quoted in *Introduction to Nonlinear Differential and Integral Equations*, Harold T. Davis, Dover, New York, 1962, pp. 26–27) for dealing with the equation $\dot{y} = f(x, y)$:

> "Draw the locus of all points at which the required family of curves are parallel to the axis of x: it is of course $f(x, y) = 0$. Draw the locus of points where they are parallel to the axis of y, i.e. $1/f(x, y) = 0$.
>
> "One or the other or both of these loci may not exist in the finite part of the plane; but in any case we get the plane divided up into a number of compartments: in some the required curves have positive dy/dx, in others negative dy/dx. Now calculate d^2y/dx^2 from the given differential equation. This can always be done. Draw the locus of points of inflection, i.e., $d^2y/dx^2 = 0$. We now have a number of compartments, in some of which the curves are concave upward, viz. d^2y/dx^2 positive, in others [concave] downward, viz. d^2y/dx^2 negative. We have thus divided up the plane into

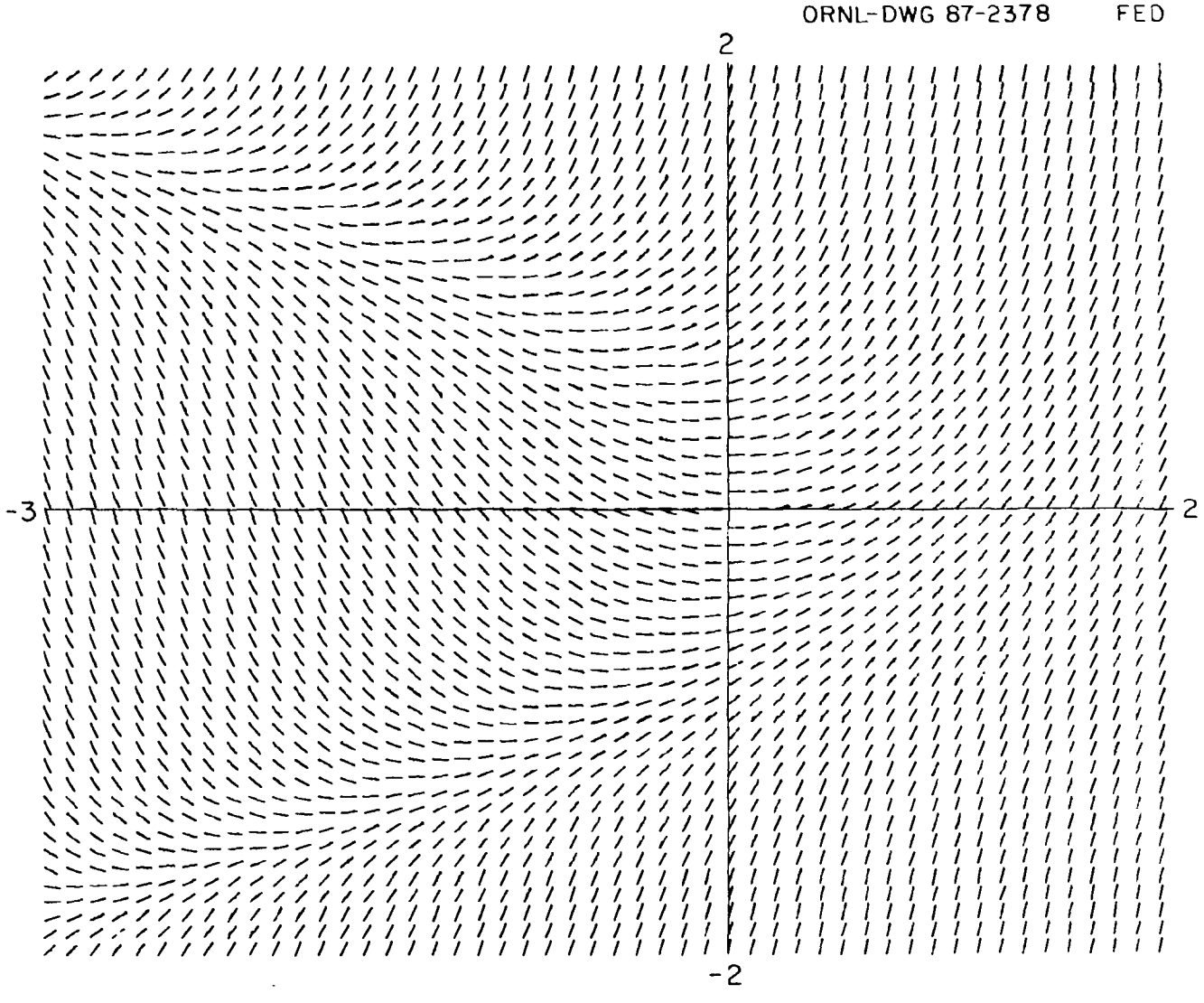

Fig. 1. The direction field of Eq. (1) in the portion of the plane $-3 < x < 2, -2 < y < 2$.

spaces, in each of which the curves satisfying the differential equation have one of the general forms

(1) ⌝ , (2) ⌜ , (3) ⌞ , (4) ⌟ , ...

Now draw a number of short tangents at a convenient number of points, and the geometrical solution of the differential equation is obtained."

Shown in Fig. 2 are the results of carrying out this procedure. The solid curve is the locus C_1 of zero slope ($\dot{y} = 0 : \quad x = -y^2$), and the dashed curve is the locus C_2 of zero curvature ($\ddot{y} = 0 : \; x = -y^2 - 1/2y$). For Eq. (1) it is easy to see that both slope and curvature [$\ddot{y} = 2y(y^2 + x) + 1$] are positive in the first quadrant. The slope changes sign as we cross locus C_1, the curvature as we cross locus C_2. This enables us at once to put the marks (1–4) above in the regions into which the plane is divided by these loci.

If we superimpose the curves C_1 and C_2 on the direction field, we see from this combined drawing (Fig. 3) that there are integral curves like 1 that appear to rise from $-\infty$, cross C_2, and approach $y = +\infty$. We might suspect that this is so from

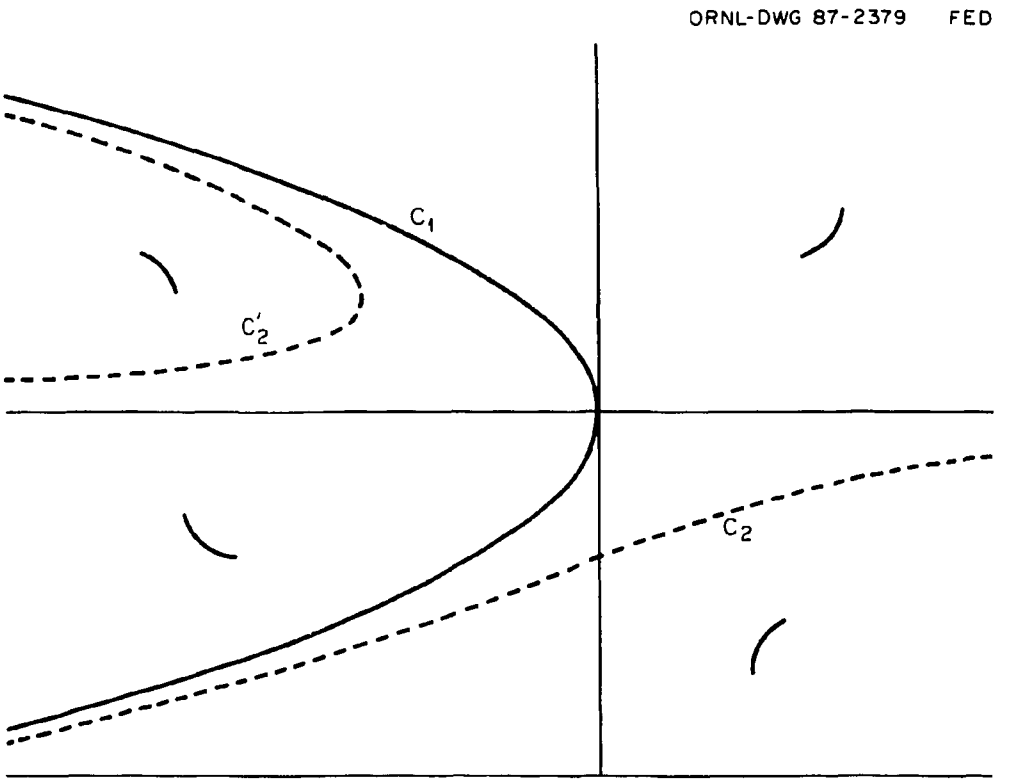

Fig. 2. The locus C_1 of zero slope ($\dot{y} = 0$, solid curve) and the locus C_2, C_2' of zero curvature ($\ddot{y} = 0$, dashed curve).

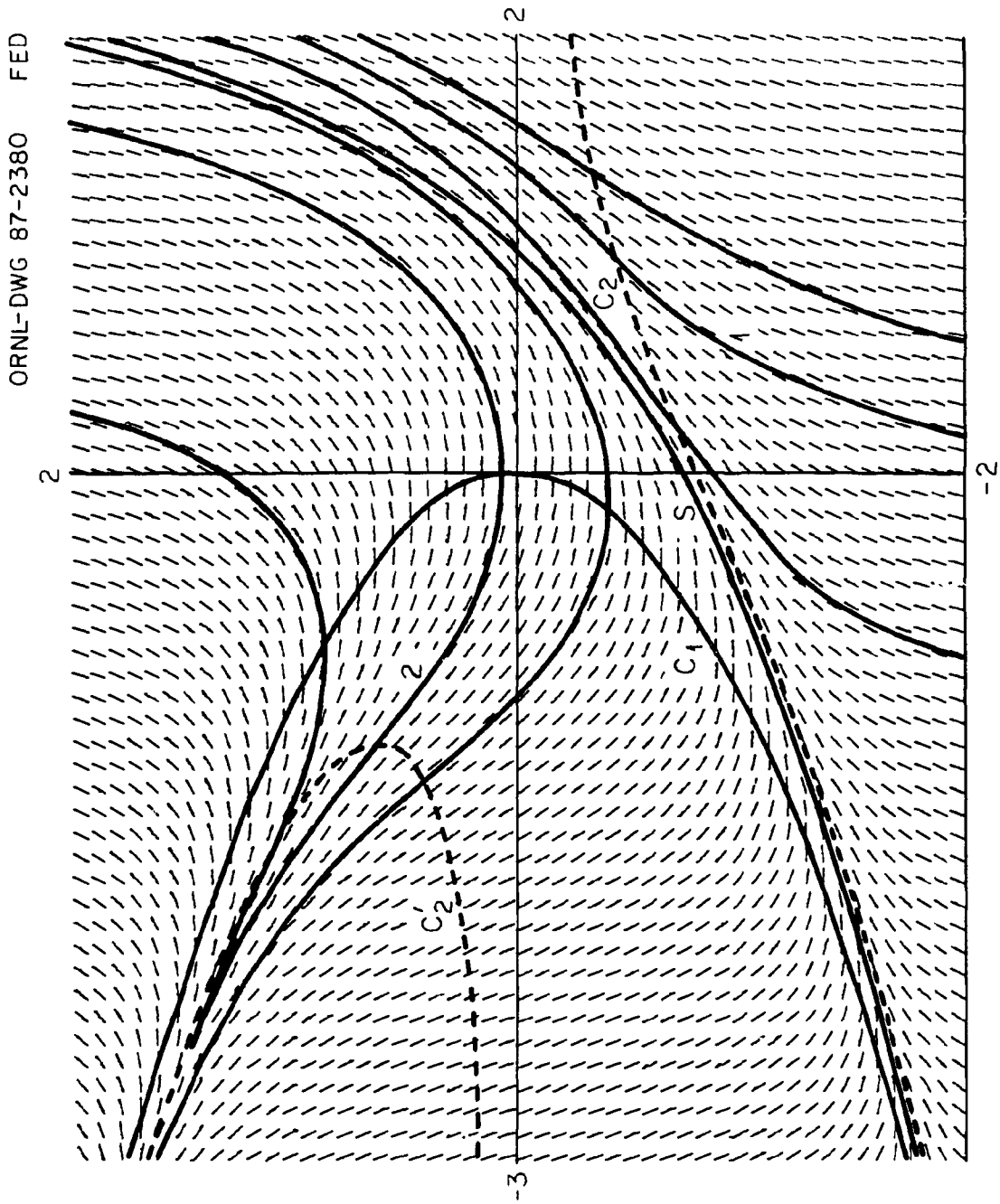

Fig. 3. The curves C_1, C_2, and C_2' of Fig. 2 superimposed on the direction field of Fig. 1.

the curves of Fig. 2 alone, and here is how we can verify our suspicion. First, we answer the question, how do the integral curves cross C_2? The slope of the integral curves is given by Eq. (1). If we evaluate the right-hand side of Eq. (1) on C_2, where $x = -y^2 - 1/2y$, we find

$$\dot{y}_{ic}(C_2) = -\frac{1}{2y} , \tag{2}$$

where the notation on the left-hand side of Eq. (2) means the slope $\dot{y}$ of the integral curves at points (x, y) of C_2. If we differentiate the equation defining C_2 we get

$$1 = -2y\dot{y} + \frac{\dot{y}}{2y^2} \tag{3a}$$

so that

$$\dot{y}_{C_2} = \frac{1}{-2y + 1/2y^2} , \tag{3b}$$

where the left-hand side of Eq. (3b) means the slope $\dot{y}$ of the curve C_2. Then

$$\dot{y}_{ic}(C_2) > \dot{y}_{C_2} > 0 . \tag{4}$$

This means that the integral curves cross C_2 from lower left to upper right.

Now we turn to the question of how integral curves like 1 behave when $|y|$ is large. From inspection of Fig. 1 we might suspect that on each integral curve like 1, y approaches ∞ for a certain value of x and $-\infty$ for a certain smaller value of x. How can we show this without creating the entire direction field? *When $|y|$ is large*, one of the following three mutually exclusive alternatives must hold:

$$|y| \ll \sqrt{|x|} , \quad |y| \sim \sqrt{|x|} , \quad |y| \gg \sqrt{|x|} .$$

If the first of these holds, then the right-hand side of Eq. (1) can be replaced by x and Eq. (1) can be integrated at once: $y = x^2/2 + c$, where c is a constant of integration. For large enough y, the constant becomes negligible, so the first alternative gives $y = x^2/2$. But this contradicts the assumption that $|y| \ll \sqrt{|x|}$, i.e., that $y^2 \ll |x|$. So the first alternative leads to a contradiction and thus cannot hold.

The third alternative means that $\dot{y} = y^2$ to leading order, so that $-1/y = x + c$, where c is a constant of integration. As $y \to +\infty$, $-1/y \to 0$ from below, i.e., $-1/y$ ascends through negative values to zero, so that c must be negative and $y \to \infty$ as $x \to |c|$. In other words, if we replace c by $-b$, where now $b > 0$, $y \sim 1/(b - x)$. Thus each integral curve has a simple pole at which $y \to \infty$ as x approaches the pole from below.

When $y \to -\infty$, $-1/y \to 0$ from above, i.e., $-1/y$ descends through positive values to zero so that c must be positive and $y \to -\infty$ as $x \to -c$. Thus $y \sim -1/(c + x)$, and each integral curve has a second simple pole at which $y \to -\infty$ as x approaches the pole from above.

The second alternative means that $y = A\sqrt{x}$, $x > 0$, and $y = A\sqrt{-x}$, $x < 0$, where A stands for some generic constant of proportionality. When substituted into Eq. (1), this gives, for $x > 0$, $A/2\sqrt{x} = (A^2 + 1)x$, which is self-contradictory no matter what the value of A. However, when $x < 0$, this gives $A/2\sqrt{-x} = (1 - A^2)(-x)$, which can be satisfied, *to leading order* (remember, if $|y|$ is large, so will $|x|$ be), by $A = \pm 1$. So it is possible for y to approach $\pm\infty$ as $x \to -\infty$ according to the asymptotic laws $y \sim \sqrt{-x}$ or $y \sim -\sqrt{-x}$.

These last possibilities do not affect curves like 1, however; these curves therefore stretch from pole to pole in the manner of the tangent curve. They fill part of the plane densely, and the locations of the upper and lower poles vary continuously from curve to curve. These locations could be expressed, for example, as functions of the intersection of each integral curve with the x-axis.

Integral curves like 2 (which cannot cross C_2 because they would cross from upper left to lower right) also cross the x-axis. If we advance along the positive x-axis from the origin we eventually pass from the family of curves like 2 to the family of curves like 1. The locus of the intersections of the curves of the family 2 with the x-axis, being dense on the x-axis and bounded from above, has an upper limit point on the x-axis. This limit point is also the lower limit point of the intersections of the curves of 1 with the x-axis, these intersections being dense on the x-axis and bounded from below. This limit point thus separates the intersections of the two families with the x-axis. There is such a limit point on any line parallel to the x-axis; their locus is a curve S that separates the two families of integral curves. Because S lies infinitely close to integral curves of both families, it must have the slope prescribed by Eq. (1), i.e., it must be a solution of the differential equation. Such a limiting solution that separates two qualitatively different families of integral curves is called a separatrix. Separatrices are important because, as we shall see later, they often turn out to be the thing we must calculate in order to obtain a similarity solution of a partial differential equation.

The curve S lies above the integral curves of family 1; therefore it must lie above curve C_2. Futhermore, it lies below the curves of the family 2; therefore it lies below curve C_1. Consequently, as $x \to -\infty$, $y_s \sim -\sqrt{-x}$, since this is the common asymptote of curves C_1 and C_2. This asymptote can be used to obtain starting values for the numerical computation of S.* Since the value of y_s is known for x large and negative, we integrate numerically in the positive x-direction. This is fortunate because that is the stable direction of integration. By stable we mean

*It is possible to obtain an asymptotic series for S at the cost of some computational labor. If we set $x' = -x$ and $y' = -y$, for convenience, then Eq. (1) becomes $y' = y'^2 - x'$ and S' is asymptotic to $\sqrt{x'}$ for $y' \gg 1$. Using the method of undetermined coefficients, we can obtain the asymptotic series

$$y' = \sqrt{x'} + \frac{1}{4x'} - \frac{5}{32x'^{5/2}} + \frac{15}{64x'^4} - \frac{1105}{2048x'^{11/2}} + \cdots$$

If we again change the sign of x and y we get points on the separatrix S of Fig. 3.

†This differential equation, like most others in this book, is not contrived but arose in the author's study of the expulsion of cold helium from a long, slender, heated tube.

that small errors (e.g., the truncation error of a finite-difference scheme or the error incurred by the finite-decimal representation of numbers in the computer) do not increase without bound in the course of integration. This is because neighboring integral curves converge on S as we advance in the positive x-direction. Small errors such as roundoff and truncation errors heal themselves as we integrate forwards. On the other hand, if we integrate backwards (i.e., in the negative x-direction) we are eventually thrown off either to one side or to the other.

What about the behavior of integral curves like 2? The same reasoning applied to integral curve 1 shows that on curves like 2, y can approach $+\infty$ either by approaching a pole from below or by approaching asymptotically $\sqrt{-x}$ as $x \to -\infty$. On the right they must clearly have a pole. Since they cannot cross C_1 again on the far left, they must always lie below it and so must approach $\sqrt{-x}$ for large enough $-x$. Furthermore, since integral curves cross the upper branch of C_2' from lower left to upper right, curves like 2 approach the common asymptote $\sqrt{-x}$ of C_1 and C_2' from below C_2'.

The diagram in Fig. 3 summarizes what there is to be known about Eq. (1), and it is fair to say that its content can be taken in at a glance. It is my contention that the qualitative nature of the curves of families 1 and 2 could have been deduced from Fig. 2 alone by augmenting Brodetsky's method with the two additional methods used here, namely: (i) the study of how the integral curves cross C_2 and (ii) the study of asymptotic behavior by enumeration of cases.

1.2 The example we have just studied is of a very simple kind in which the slope y is uniquely determined at every point (x,y) of the plane. More complicated cases arise when points (x,y) exist at which $f(x,y)$ is multivalued. The differential equation[†]

$$\dot{y} = \frac{y(2-x)}{3x-y} \tag{5}$$

provides an example of this. At the point O:(0,0) and P:(2,6) the right-hand side of Eq. (5) becomes indeterminate in the manner 0/0. Such points are called singular points of the differential equation. To see what happens at these singular points (as well as everywhere else) we study the direction field of Eq. (5). We shall not actually construct it as we did in Fig. 1 but rather infer its general appearance by following Brodetsky's advice.

The slope $\dot{y}$ vanishes when the numerator of the right-hand side vanishes, i.e., when $y = 0$ or $x = 2$; it is infinite when the denominator vanishes, i.e., when $y = 3x$. These lines are shown in Fig. 4 along with hatch marks to indicate the slope of the integral curves on them. The points O and P, which are the intersections of lines on which $\dot{y} = 0$ with lines on which $\dot{y} = \pm\infty$, are shown as black dots. These lines divide the plane into seven regions, in each of which the slope has a constant sign. The slope changes sign as we cross each line.

If we find the sign of the slope at any convenient point, we can then assign the sign everywhere by simply crossing the lines from region to region. Since the slope on the y-axis ($x = 0$) is -2 (except possibly at 0), the sign of the slope in each

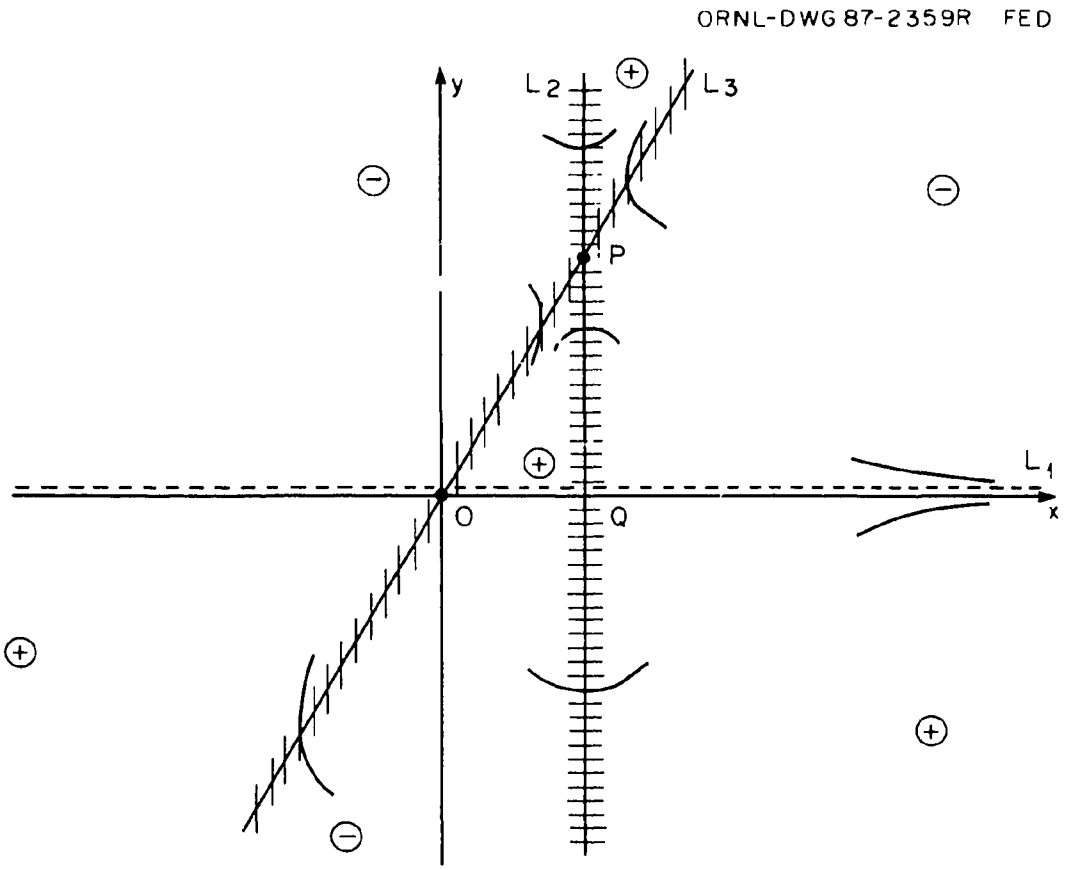

Fig. 4. A partial sketch of the direction field and integral curves of Eq. (5).

region must be as shown. With these assignments fixed, we can begin to sketch in parts of the integral curves. In Fig. 4 are six short arcs showing how the integral curves must cross the lines L_2 and L_3. No integral curve can cross L_1 (except possibly at 0) because the slope $\dot{y}$ of the integral curves is equal to the slope of L_1 itself. [This means, of course, that $L_1 : y = 0$ is an integral curve of Eq. (5).] With the assignments of slope given in Fig. 4, this shows that $y \to 0$ as $x \to \infty$ on any integral curve, as indicated by the two short arcs near line L_1.

What happened to the two integral curves shown intersecting the segments of OP and PQ as $x \to 0$? They cannot escape from the triangle OPQ by crossing either line L_1 or line L_3, so they must pass through the origin O. To study how they might do this, we first note that close to the origin O, the differential equation of Eq. (5) can be replaced by

$$\dot{y} = \frac{2y}{3x - y} \quad (\text{near } O) \tag{6}$$

since close to the origin $|x| \ll 2$. This differential equation can be solved; the solution is $x = y + Cy^{3/2}$, where C is a constant of integration. You can verify this by calculating dx/dy and comparing it with $1/\dot{y}$ calculated from Eq. (6). Later you will learn a straightforward way of finding such a solution. But right now you may not know "how it is possible to invent such a solution." From this solution we see that all integral curves (save the exceptional one $y = 0$) approach the origin along the line $y = x$ (remember, when $y \ll 1$, $y^{3/2} \ll y$). Now I will show you how to obtain that information from Eq. (6) without solving it by studying its limiting behavior by enumeration of cases.

Any curve entering the origin can do so in one of three mutually exclusive ways: $|y| \ll |x|$, $|y| \sim |x|$, and $|y| \gg |x|$. Since the curves we are interested in lie between $L_3 : y = 3x$ and $L_1 : y = 0$, the third alternative is excluded. The first alternative simplifies Eq. (6) to $\dot{y} = 2y/3x$, which can be solved at once to give $y = \text{const}\, x^{2/3}$. No matter what the value of the constant, when $|x|$ is small enough, this contradicts the hypothesis $|y| \ll |x|$. Hence the first alternative is likewise excluded. The second alternative means $y = Ax$ when x is small enough. Inserting this form into Eq. (6), we obtain the algebraic equation $A = 2A/(3 - A)$ for A, which has the solution $A = 0$ and $A = 1$. The first of these contradicts the hypothesis $|y| \sim |x|$, so we are left with the second. Thus, integral curves entering the origin do so along the line $y = x$.

The integral curves in the triangle OPQ are of two types, those that eventually cross the segment of OP of L_3 and those that eventually cross the segment of PQ of L_2. These two families must be separated by a separatrix S that, because it belongs to neither family, must exit through the point P. The point P, lying as it does at the center of four different families of integral curves, must be traversed by two separatrices (see Fig. 5). One of them is S; the other intersects S at an angle. The slopes of these two separatrices at P can be determined by an application of l'Hospital's rule:

$$\dot{y}_P = \frac{2\dot{y}_P - y_P - x_P\dot{y}_P}{3 - \dot{y}_p} = \frac{-6}{3 - \dot{y}_P} , \tag{7a}$$

$$\dot{y}_P = (3 \pm \sqrt{33})/2 . \tag{7b}$$

A singular point like P traversed by two separatrices separating four families of integral curves is called a saddle point.

When $|x|$ is large, Eq. (5) becomes

$$\dot{y} = \frac{-xy}{3x - y} . \tag{8}$$

The integral curves in the first and fourth quadrants must approach L_1 as $x \to \infty$. Therefore, for them $y \ll x$, and Eq. (8) becomes $\dot{y} = -y/3$, which can easily be solved to give $y = \text{const}\exp(-x/3)$. So these integral curves approach L_1 exponentially.

To analyze the asymptotic behavior of the integral curves in the second and third quadrants, i.e., as $x \to -\infty$, we again resort to the enumeration of alternatives. As

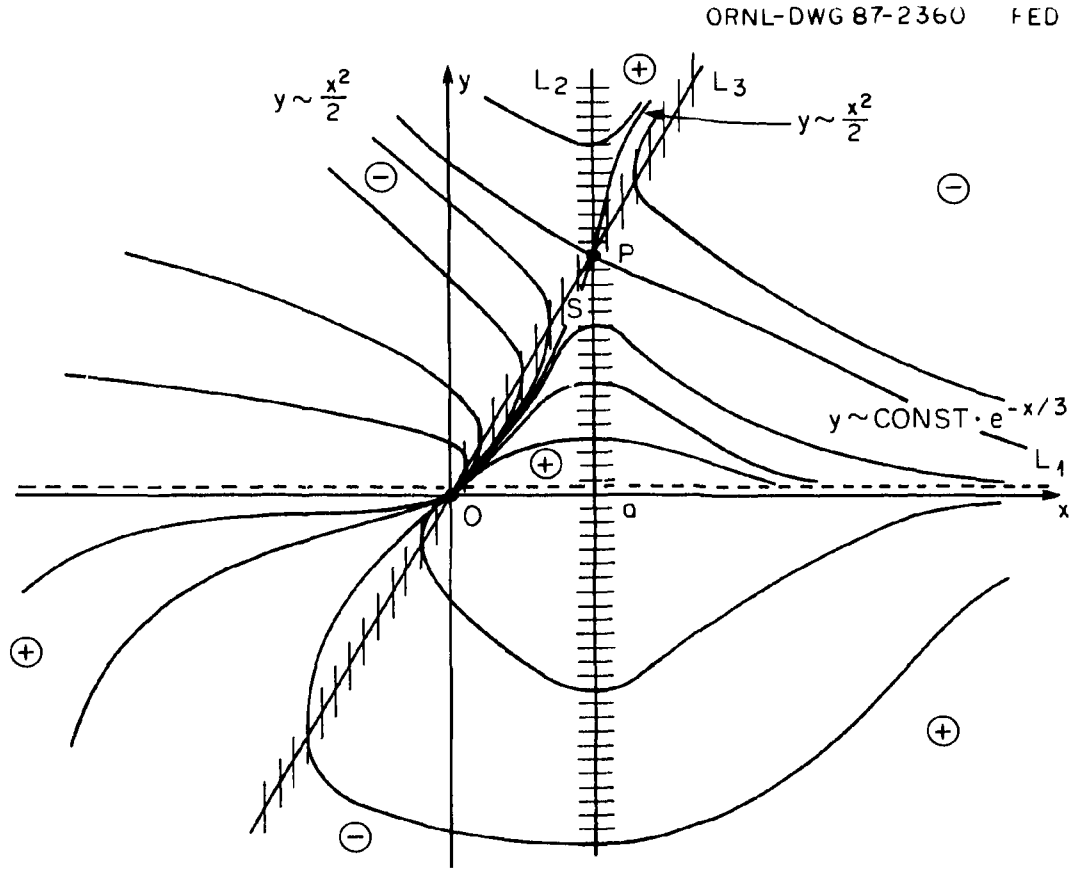

Fig. 5. Second stage in the construction of the integral curves of Eq. (5).

$x \to -\infty$, either $|y| \gg |x|$, $|y| \sim |x|$, or $|y| \ll |x|$. The first alternative leads to $\dot{y} = x$ or $y = x^2/2 + \text{const}$. When $|x|$ is large enough the constant of integration will be negligible so that $y \sim x^2/2$. This is consistent with the hypothesis $|y| \gg |x|$, but only for integral curves in the second quadrant, where $y > 0$. The third alternative, $|y| \ll |x|$, leads to $y = \text{const}\exp(-x/3)$ as before. But now as $x \to -\infty$, it contradicts the hypothesis $|y| \ll |x|$. Finally, $|y| \sim |x|$ also leads to a contradiction because the numerator of Eq. (8) is of order 2 while the denominator is of order 1. Thus the integral curves in the second quadrant stretch toward infinity asymptotically to $y = x^2/2$. None of the alternatives is free of contradiction for integral curves in the third quadrant, so they cannot stretch to infinity. Instead, they must intersect the line L_3 and loop around into the fourth quadrant as shown. Finally, the integral curves between the lines L_2 and L_3 can only fulfill the alternative $|y| \gg |x|$ and so are asymptotic to $y = x^2/2$.

Figure 5 summarizes all the information we have gained and displays the content of the differential equation (5) so it can be comprehended at a glance. It is surprising

how such a simple differential equation can give rise to so complex an array of integral curves. In the practical problem that gave rise to Eq. (5), it was the section of the separatrix S between O and P that was needed. It was calculated numerically by integrating from P to O [the stable direction using the positive slope equation (7b) to obtain starting values close to P].

1.3 The study of asymptotic behavior by enumeration of alternatives, if handled unthinkingly, can lead to unexpected paradoxes. The differential equation $\dot{y} = (x + y)/x$ provides an example. Figure 6 shows the direction field. Only the first and fourth quadrants are shown; the second and third are images of the first and fourth under the transformation $x' = -x, y' = -y$, to which the differential equation is invariant. There must certainly be some integral curves like those shown. How do these integral curves enter the origin? They can do so in one of three mutually exclusive ways, namely, $|y| \ll |x|$, $|y| \sim |x|$, and $|y| \gg |x|$. The first alternative leads to $y = x +$ const, which contradicts the hypothesis $|y| \ll |x|$. The second alternative, which means $y = Ax$ for small enough x and y, leads to $A = A+1$, which has no solution. The third alternative leads to $\dot{y} = y/x$ so that $y =$ const x. This, too, contradicts the hypothesis $|y| \gg |x|$. So none of the three mutually exclusive alternatives appears free of contradiction. The resolution of this paradox is this: the constants of integration denoted above by "const" are not necessarily constants, but may be slowly varying functions of x. Consider again the first alternative $|y| \ll |x|$. If it applies, the differential equation becomes $\dot{y} = 1$ *to leading order.* This differential equation is satisfied, again to leading order, by expressions of the sort $y = x + C(x)$, where $C(x)$ is a sufficiently slowly varying function of x. For then, $\dot{y} = 1 + \dot{C}(x)$, so if $\dot{C}(x) \ll 1$, $y = 1$ to leading order. Even with this enlargement of the meaning of "const," the first alternative leads to a contradiction. So, too, does the second alternative. But the third alternative does not!

When $|y| \gg |x|$, the differential equation becomes $\dot{y} = y/x$ to leading order. Were this exact, it would give $y =$ const x. Try instead $y = C(x)x$, where $C(x)$ is a slowly varying function of x. Differentiating, we find $\dot{y} = C + x\dot{C} = y/x + x\dot{C}$. If $|x\dot{C}| \ll |y/x| = |C|$, the solution of $y = C(x)x$ satisfies the differential equation $y = y/x$ *to leading order.* If $\lim_{x\to 0} |C(x)| = \infty$, it is then possible for $|y| \ll |x|$ for small x. [An example of a function $C(x)$ that satisfies these requirements is $C(x) = \ln x$.] From the relation $\dot{y} = C + x\dot{C}$ we see at once that $|\dot{y}(0)| = \infty$. By differentiating the differential equation we find that $\ddot{y} = x^{-1}$, which is positive in the first and fourth quadrants. So the integral curves are all concave upwards. This precludes the possibility of any integral curves rising vertically from 0 in the positive y-direction, so the integral curves must all look like those shown in Fig. 6.

The general solution of the differential equation $\dot{y} = (x + y)/y$ is $y = x \ln(Ax)$, as the reader may verify by differentiation. Later we shall learn a direct method of solving this differential equation.

1.4 The singular points O and P in Fig. 5 are the intersections of one line on which $\dot{y} = 0$ and another on which $\dot{y} = \pm\infty$. Such intersections are surrounded

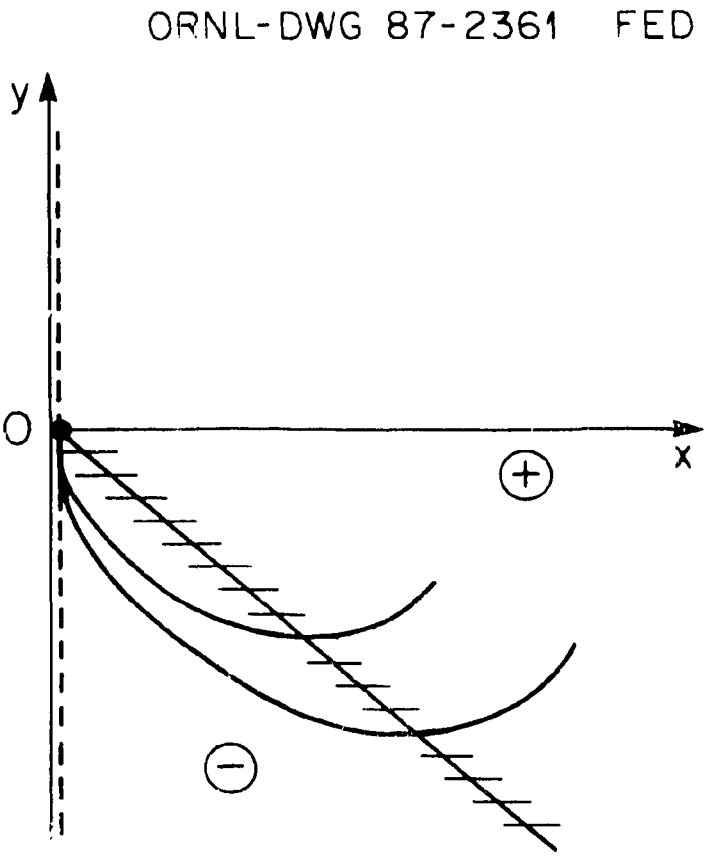

Fig. 6. Part of the direction field of the differential equation $\dot{y} = (x + y)/x$.

by characteristic patterns of integral curves of only a few different types, and we display them below.

Figure 7a shows one possible configuration. Because the sign of $\dot{y}$ changes as we move across either of the two lines, the sign alternates from quadrant to quadrant as we circulate around the singularity P. The array of four families of integral curves separated by two separatrices characterizes the *saddle point.*

If we keep the same configuration of lines but change the sign of y by multiplying the right-hand side of the differential equation by -1, we get the configuration shown in Fig. 7b. The integral curves can either spiral into P (in which case P is called a focus) or surround P as closed curves (in which case P is called a center or a vortex point).

A new behavior occurs in the degenerate case in which the locus of zero (infinite) slope is itself a line of zero (infinite) slope. Again, two assignments of sign are possible. One (Fig. 7c) leads again to a saddle point, the other (Fig. 7d) to integral curves radiating from P like the spokes of a wheel—it is called a node.

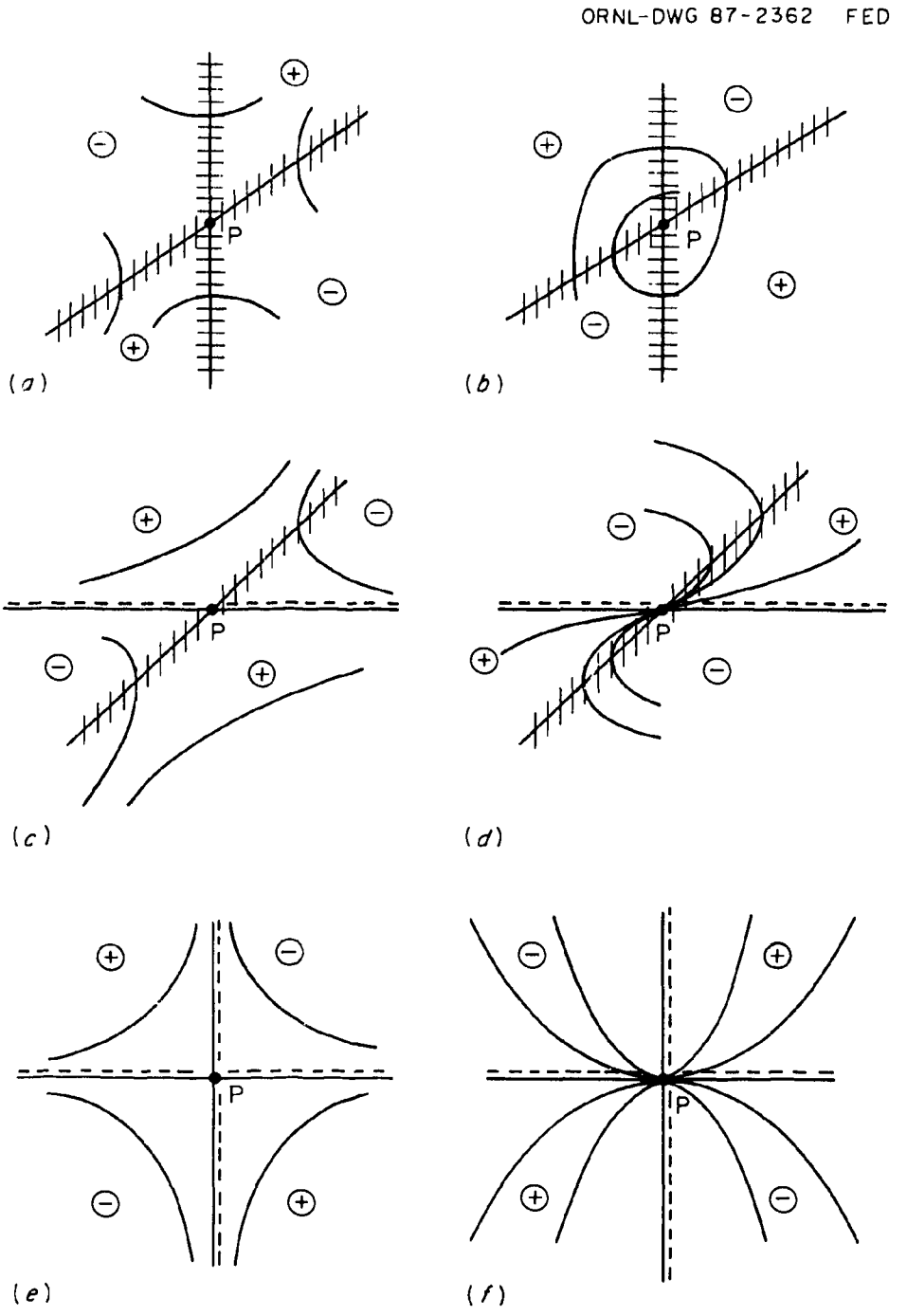

Fig. 7. (a) A saddle point, (b) a focus or a center (vortex point), (c) a saddle point, (d) a node, (e) a saddle point, and (f) a node.

If the locus of zero slope is a line of zero slope and the locus of infinite slope is simultaneously a line of infinite slope (a kind of double degenerate case) we get the configurations in Figs. 7e and 7f, giving, respectively, a saddle point and a node.

The integral curves entering the node of Fig. 7d might be considered as degenerate spirals that are prevented from making more than a half-turn around P by the line of zero slope, which they cannot cross. Some justification for this viewpoint can be found in a topological characterization of the direction field, the Poincaré index. The Poincaré index is a topological invariant of a continuous vector field. A vector field is a diagram in which a small vector is plotted at every point (x, y) according to some given prescription. It differs from a direction field only in that the little hatch marks of the direction field have been supplied with arrow heads showing in which direction they point. A continuous vector field is one in which the directions of the two vectors at two neighboring points are close to one another.

Suppose we draw a closed curve in such a vector field. As we advance along this curve in the positive (counterclockwise) direction, the local vector of the vector field will continuously change its direction. When we return to our starting point it will have returned to its original direction. In doing so, it may have executed several complete revolutions—the number of such revolutions (counted positive when executed counterclockwise and negative when executed clockwise) is the Poincaré index.

To convert a direction field to a vector field, we start by putting an arrowhead on any arbitrary hatch mark. The arrow direction everywhere else is determined by the requirement that the vector field be continuous. Figure 8 shows the results of such a construction at (a) an ordinary point of the vector field, (b) a node, (c) a center, (d) a saddle, and (e) a focus, together with the Poincaré index I of the curve C.

If the curves C of Figs. 8b–8e are imagined to shrink down continuously around the points S inside them, their index I will remain unchanged. For the index can only change by an integer, something that cannot happen continuously. The index can only change when the curve C crosses a singularity. So the index of any curve surrounding a singularity is the same, and we can therefore call its value the index of the point. Saddles have index -1; nodes, centers, and focuses have index $+1$; and ordinary points have index 0.

Among the most useful facts about the index are these. The index of a closed *integral* curve is 1. Consequently, such an integral curve must surround some singular point. Furthermore, the index of any closed curve C is the sum of the indexes of the singular points it contains. [To see this, surround each singularity with an infinitesimal circle and join these circles to the curve C by cuts that will be traversed twice in opposite directions (Fig. 9). The index of the entire cut curve is zero since it contains no singularity (case 8a). Since the cuts contribute nothing to the overall vector rotation because they are traversed alternately in opposite directions, $I_C - I_{C1} - I_{C2} = 0$, as was to be proved.] As an example of how this last theorem can be applied, imagine a large contour in Fig. 5 surrounding both the singularities O and P. It is easy to see that the index of the large contour is zero, so $I_O + I_P = 0$.

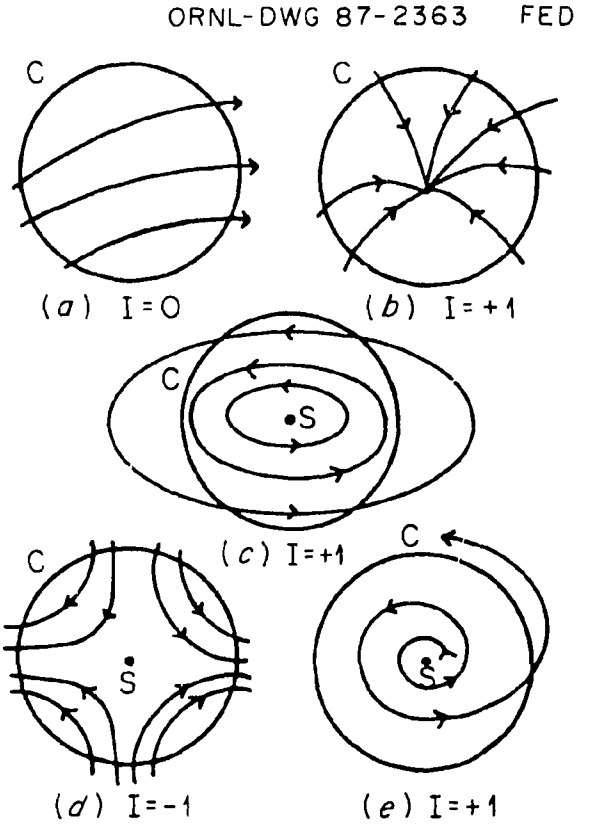

Fig. 8. A vector field at (a) an ordinary point, (b) a node, (c) a center, (d) a saddle, and (e) a focus. I is the Poincaré index of the curve C.

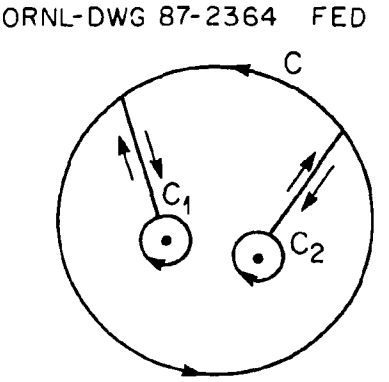

Fig. 9. Sketch to aid in the calculation of the index of a curve surrounding two singularities.

Clearly, then, one of the singularities must be a saddle, while the other must be a center, a node, or a focus (it is in fact a node).

The singularities dealt with so far are particularly simple. More complex singularities can arise from the confluence of several simple singularities. For example, the differential equation

$$\dot{y} = \frac{x^2}{x+y} \tag{9}$$

has a single singularity at the origin (see Fig. 10). This singularity has a Poincaré index of zero. The reason for this peculiar behavior is that $x = 0$ is a double root of $x^2 = 0$, the equation we obtain when we set the numerator equal to zero. Alternatively, we may note that the sign of y does not change as we cross the locus of $\dot{y} = 0$.

Equation (9) may be considered as the limit of the differential equation

$$\dot{y} = \frac{x(x-\epsilon)}{x+y} \tag{10}$$

as $\epsilon \to 0$. Equation (10) has two singularities, a focus at the origin and a saddle at the point $(\epsilon, -\epsilon)$. These two merge as $\epsilon \to 0$, giving a compound singularity whose Poincaré index is zero. Two separatrices emerge from the singularity.

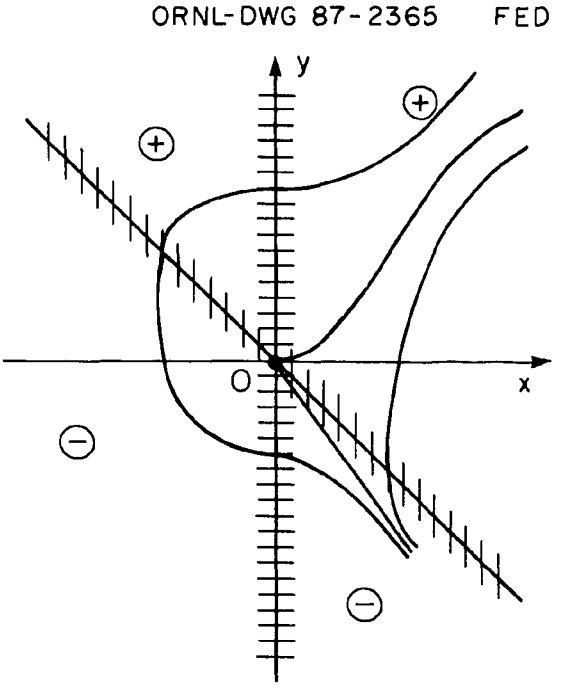

Fig. 10. The direction field of Eq. (9).

The differential equation

$$\dot{y} = -\frac{x^3}{x+y}\,, \tag{11}$$

which is similar to Eq. (9), is a reduced form of the second-order Emden-Fowler equation. (The Emden-Fowler equation arises in the study of the equilibrium mass distribution of a cloud of gas held together by gravity. We shall study the method of reducing it to a first-order equation in later chapters.) Its one singularity, located at the origin, has a Poincaré index of 1. Figure 11 shows its direction field. At large enough radii, the integral curves spiral around the origin, but once within a critical radius they approach the origin, drawing ever closer to the line $y = -x$ as they do so. Curves intersecting the line $y = -x$ at abscissas whose absolute values are greater than some value x_0 make another half-circuit counterclockwise, whereas curves intersecting $y = -x$ at abscissas whose absolute values are less than x_0 approach the origin along the line $y = -x$.

What about the exceptional integral curve that intersects the line $y = -x$ at $x = \pm x_0$? It approaches the origin along the x-axis, i.e., with zero slope, which

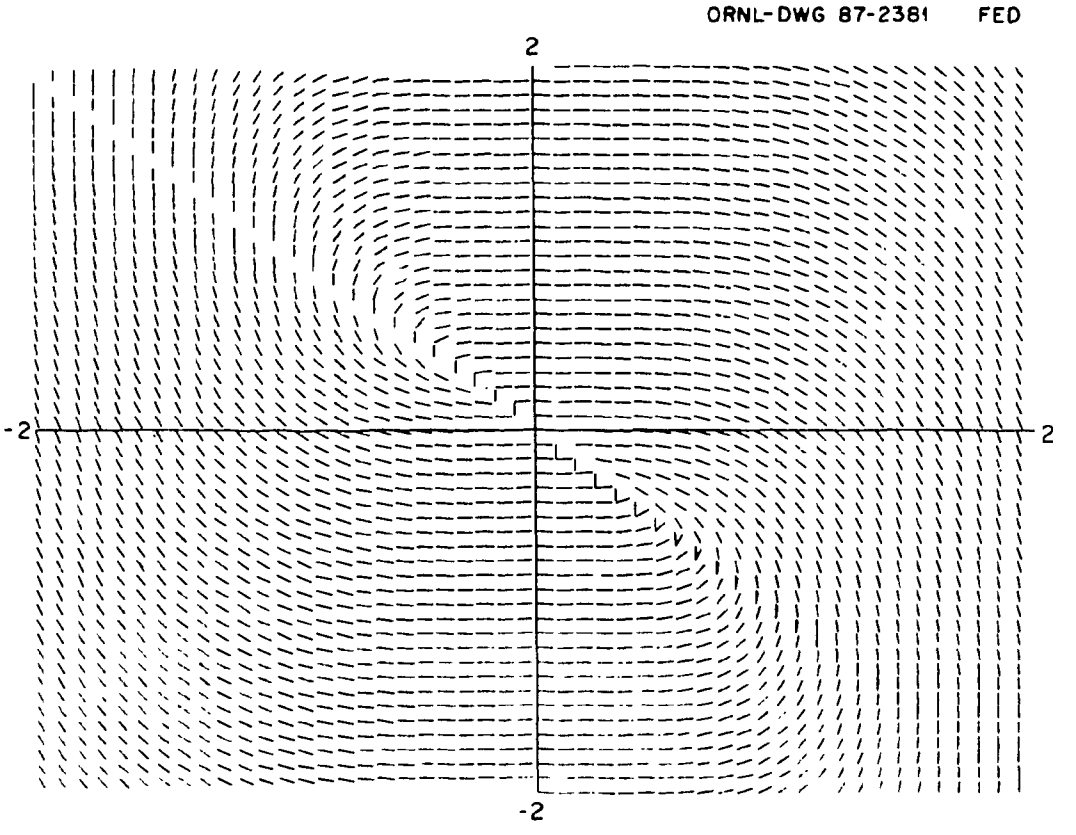

Fig. 11. The direction field of Eq. (10).

means $|y| \ll |x|$. It follows, then, from Eq. (11), that to leading order S is given by $y = -x^3/3$. In fact, if we set

$$y = -\frac{x^3}{3} + Ax^5 + Bx^7 + Cx^9 + \ldots \tag{12}$$

we find

$$\dot{y} = -x^2 + 5Ax^4 + 7Bx^6 + 9Cx^8 + \ldots$$

$$x + y = x - \frac{x^3}{3} + Ax^5 + Bx^7 + Cx^9 + \ldots$$

and

$$-x^3 = (x+y)\dot{y}$$

$$= -x^3 + \left(\frac{1}{3} + 5A\right)x^5 + \left(7B - \frac{8A}{3}\right)x^7 + \left(9C - \frac{10B}{3} + 5A^2\right)x^9 + \ldots$$

so that

$$A = -\frac{1}{15}\ , B = -\frac{8}{315}\ , C = -\frac{101}{8505}\ ,\ \text{etc.}$$

Thus the integral curve S is given by the series

$$y = -\frac{x^3}{3} - \frac{x^5}{15} - \frac{8x^7}{315} - \frac{101x^9}{8505} - \ldots\ ,$$

which represents it close to the origin. It is this integral curve that interests us in astrophysical applications.

No integral curve can enter the origin in such a manner that $|y|$ is always $\gg |x|$. For then, Eq. (11) would become $\dot{y} = -x^3/y$ so that $y^2/2 + x^4/4 = \text{const}$. If such a curve passes through the origin, the constant must vanish, and then so must x and y, a contradiction. But integral curves can enter the origin in such a manner that $|y| \sim |x|$. If we set $y = ax$, we find $a(a+1)x = -x^3$, which can be satisfied to leading order if $a = -1$. In fact, if we set y equal to a power series in the odd powers of x, and proceed as we did above, we find the series

$$y = -x + x^3 + 3x^5 + 24x^7 + 289x^9 + \ldots \tag{13}$$

The series of Eq. (13) is a formal solution of Eq. (11). If it converged, then within its radius of convergence all integral curves that approach the origin along $y = -x$ would have to be identical with it. This is not the case, as one can see from Fig. 11, where infinitely many different integral curves approach 0 and along $y = -x$. So Eq. (13) never converges, no matter how small x is. We might have suspected this from the rapidity with which the coefficients grow. Equation (12), on the other hand, representing a particular special integral curve, probably converges, as we might suspect from the decreasing of its coefficients. Neither assertion about convergence has been proved here.

1.5 Besides having isolated singular points at which many slopes are possible, differential equations may have more than one slope at every point! To see how this can happen, let us begin by considering the one-parameter family of parabolas

$$y = (x - a)^2 + a \ , \qquad a = \text{a parameter} \ . \tag{14}$$

These parabolas have minima $y = a$ at $x = a$, and so a sketch of the family looks like Fig. 12. Two parabolas pass through each point (x, y) of the plane, one with a positive slope and one with a negative slope.

We can find these two slopes by converting Eq. (14) into a differential equation. We do this by first differentiating to get $\dot{y} = 2(x - a)$ and then eliminating a in favor of y:

$$(\dot{y} - 1)^2 = 4y - 4x + 1 \ . \tag{15}$$

The two slopes arise from the two signs of the square root that are possible. This differential equation has the family Eq. (14) as the family of its integral curves.

The family Eq. (14) has an envelope E given $f(x, y, a) = 0$ and $f_a(x, y, a) = 0$ where $f(x, y, a) = y - (x - a)^2 - a$. The envelope is the straight line $y = x - 1/4$. This straight line, because it is everywhere tangent to a curve of the family Eq. (14), must also be a solution of the differential equation (15). Substitution shows this to be so. Such a solution is called a singular solution.

The usual situation is to be given the differential equation, not the family of integral curves. It turns out that we can find the singular solution (if one exists, of course) from the differential equation even if we cannot integrate the differential equation to find the family of integral curves. Here is how we proceed. Suppose the differential equation can be written as $f(x, y, \dot{y}) = 0$. From the sketch in Fig. 12 we can see that on the singular solution the two roots for $\dot{y}$ collapse to one double root. When the function f has a double root, then $f_{\dot{y}} = 0$ at the double root, too. So we find the singular solution by eliminating $\dot{y}$ from the equations $f(x, y, \dot{y}) = 0$ and $f_{\dot{y}}(x, y, \dot{y}) = 0$. Applying this to Eq. (15), we obtain at once $\dot{y} = 1$, $y = x - 1/4$.

The procedure outlined above can produce loci that are not solutions of the differential equation at all. Figure 13a shows one way in which this can happen. The integral curves again have two branches, which this time meet at a cusp. At the cusp, the slopes of the two branches become equal. Solution of the equations of $f = 0$ and $f_{\dot{y}} = 0$ will yield the locus L of the cusps. But L is clearly not a singular solution of the differential equation because it nowhere has the slope of the integral curves.

It is possible, however, for a cusp locus to be a singular solution, and Fig. 13b shows how this can happen. An analytic criterion that distinguishes case (a) from case (b) can be found as follows: at neighboring points (x, y) and $(x + dx,\ y + dy)$ on the locus L, we have $f(x, y, \dot{y}) = 0$ and $f(x + dx,\ y + dy,\ \dot{y} + d\dot{y}) = 0$. Subtracting these two equations, we obtain $f_x dx + f_y dy + f_{\dot{y}} d\dot{y} = 0$. Now on L we must also have $f_{\dot{y}} = 0$. Thus, $f_x dx + f_y dy = 0$. Now dy/dx is the slope of L, and if L is to be a singular solution this slope must equal a value of $\dot{y}$ obtained from the differential equation. So if $f_x + f_y \dot{y} = 0$, L is a singular solution.

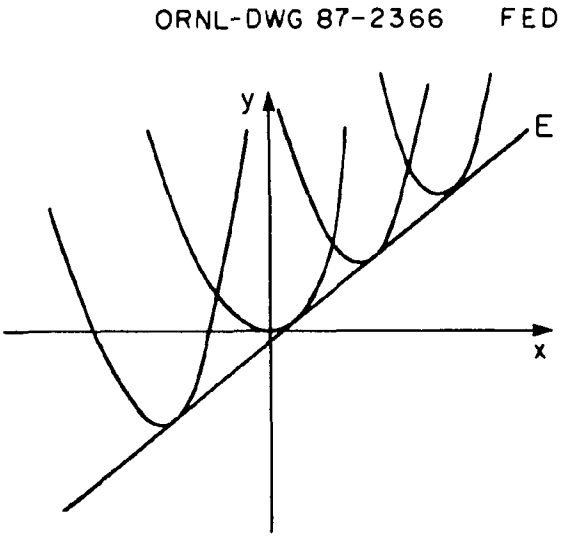

Fig. 12. Sketch of the family of parabolas $y = (x - a)^2 + a$.

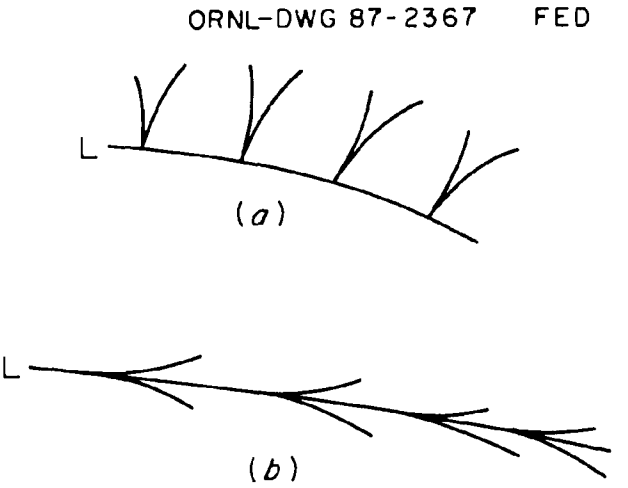

Fig. 13. Sketches showing cusp loci L which (a) are not and (b) are singular solutions.

The reader should realize that another way to test whether a locus like L is a singular solution is to substitute it into the differential equation. Such a test is unimpeachable.

1.6 Singular solutions have the attractive property that we can obtain them without integrating the differential equation to display explicitly the entire family of integral curves. Separatrices have the same attractive property, although to find them we need more information than just the function f but still less than the full, explicit form of the integral curves. We need a quantity called the integrating factor, which is defined below.

In discussing the integrating factor it is convenient to write the differential equation $\dot{y} = f(x,y)$ in the form

$$M(x,y)dx + N(x,y)dy = 0 \ , \tag{16}$$

where $f(x,y) = -M(x,y)/N(x,y)$. The general solution of a first-order differential equation like Eq. (16) is a one-parameter family of curves, the parameter being essentially a constant of integration. We represent the family of integral curves as $\phi(x,y) = C$, where C is the parameter that labels the curves. If we differentiate along an integral curve, C may be treated as a constant, and we have

$$\phi_x dx + \phi_y dy = 0 \ . \tag{17}$$

Since the incremental vector (dx, dy) lies along an integral curve, it satisfies Eq. (16) as well; the two equations, (16) and (17), have a nontrivial solution if and only if

$$\frac{\phi_x}{M} = \frac{\phi_y}{N} \ . \tag{18}$$

The two sides of Eq. (18) represent a function of x and y; denote it by $\mu(x,y)$. It is called an integrating factor because if we multiply the differential equation (16) by it, the differential equation takes the form [Eq. (17)] of a perfect differential.

Since $\phi_x = \mu M$ and $\phi_y = \mu N$, equality of the cross derivatives $\phi_{xy} = \phi_{yx}$ gives the condition $(\mu M)_y = (\mu N)_x$, which μ must satisfy. This condition is equivalent to the partial differential equation

$$N\mu_x - M\mu_y = \mu(M_y - N_x) \ . \tag{19}$$

Any particular solution of Eq. (19) is a suitable integrating factor. It is not necessary to find the general solution of Eq. (19).

Suppose we know two different integrating factors, $\mu(x,y)$ and $\nu(x,y)$. Multiplying Eq. (16) by them converts Eq. (16) into two different perfect differentials which upon integration give $\phi(x,y) = a$ and $\psi(x,y) = b$ (because $\mu M = \phi_x$, $\mu N = \phi_y$, $\nu M = \psi_x$, and $\nu N = \psi_y$). Here a and b are constants. Both of these equations represent the same integral curves, each curve labeled with a particular value of a [if we are representing them by $\phi(x,y) = a$] or with a value of b [if we are representing them by $\psi(x,y) = b$]. A value of a determines a particular curve and thus a particular value of b, which means b is a function of a: $b = F(a)$.

Consequently, $\psi(x,y) = b = F(a) = F[\phi(x,y)]$ and the functions ψ and ϕ are functionally dependent. Differentiating this last equation partially with respect to, say, x gives $\psi_x = \dot{F}(\phi)\phi_x$ or, since $\psi_x = \nu M$ and $\phi_x = \mu M$, $\nu = \dot{F}(\phi)\mu$. Since ϕ is a constant on any integral curve, the integral curves are given by the condition $\nu = C\mu$, where $C = \dot{F}(a)$ is a constant labeling the different integral curves. Conversely, any function of the form $\dot{F}(\phi)\mu$, where $\dot{F}(\phi)$ is any function of ϕ, is an integrating factor, converting Eq. (16) into the perfect differential form $\dot{F}(\phi)\phi_x dx + \dot{F}(\phi)\phi_y dy = \dot{F}(\phi)d\phi = dF(\phi) = 0$. So the most general form of the integrating factor is $\mu G(\phi)$, where G is any function of ϕ.

The differential equation

$$(y^2 - 2xy)dx + x^2 dy = 0 \tag{20}$$

furnishes an illustrative example of these ideas. Here $M = y^2 - 2xy$, $N = x^2$, $M_y = 2(y - x)$, and $N_x = 2x$. Since $M_y \neq N_x$, Eq. (20) is not yet in the form of a perfect differential and needs to be multiplied by an integrating factor. If Eq. (20) were already a perfect differential, $\mu = 1$ would be an integrating factor, and when $\mu = 1$, Eq. (19) becomes $M_y - N_x = 0$. Equation (19) is now

$$x^2\mu_x - y(y - 2x)\mu_y = 2(y - 2x)\mu \ . \tag{21}$$

The first term will vanish if a particular solution for μ is sought that is only a function of y. The factor $y - 2x$ cancels from the remaining two terms, so that we have $-y(d\mu/dy) = 2\mu$, which gives $\mu = \text{const}\ y^{-2}$. The value of the constant is irrelevant (as long as it is not zero) so we take for our integrating factor $\mu = y^{-2}$.

If we multiply M and N by μ we find

$$\begin{aligned} \phi_x &= \mu M = 1 - 2x/y \ , \\ \phi_y &= \mu N = x^2/y^2 \ . \end{aligned} \tag{22}$$

We can integrate the first of these equations if we treat y as a constant, which we must do since the derivative ϕ_x is a partial derivative. We get $\phi = x - x^2/y + H(y)$, where $H(y)$ is the "constant" of integration. We determine $H(y)$ by differentiating partially with respect to y and comparing with the second part of Eq. (22). We find $\phi_y = x^2/y^2 + \dot{H}$, so that $\dot{H} = 0$ and H is at most a constant. Since $\phi = \text{const}$ labels the integral curves, we can incorporate H in ϕ and obtain for the integral curves

$$y = \frac{x^2}{x + C} \ , \tag{23}$$

where $C = H - \phi$ is a constant labeling the various curves. The most general integrating factor then has the form

$$\nu = \frac{G(C)}{y^2} = \frac{1}{y^2} G\left[\frac{x(x - y)}{y}\right] \ , \tag{24}$$

where $G(z)$ is any function. For example, if $G(z) = z^{-2}$, $\nu = x^{-2}(x - y)^{-2}$ is also an integrating factor.

The equation $\mu^{-1} = 0$ may specify one or more separatrices. To see this, consider the one-parameter family of integral curves $\phi(x,y) = C$ sketched in Fig. 14. The family consists of two qualitatively different parts separated by a separatrix S corresponding to the value C_0 of the parameter C.

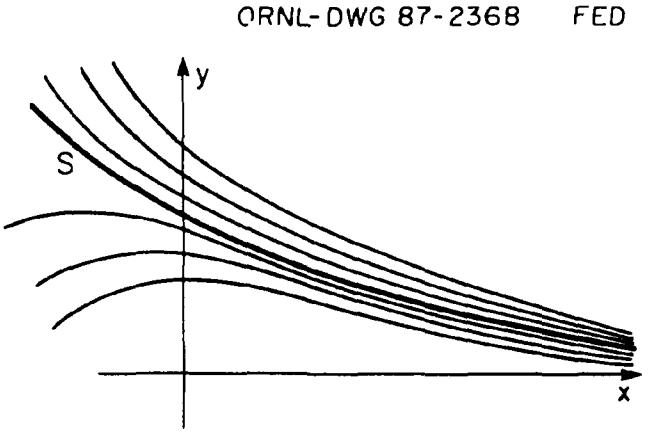

Fig. 14. A one-parameter family of curves $\phi(x,y) = C$ having a separatrix S corresponding to $C = C_0$.

If (x,y) and $(x + dx, y + dy)$ are two neighboring points on the same integral curve, then $\phi(x,y) = C$ and $\phi(x + dy, y + dy) = C$. Thus $\phi_x dx + \phi_y dy = 0$, which means the vector (ϕ_x, ϕ_y) is perpendicular to the tangent vector (dx, dy). Accordingly, the unit normal to the curves $\phi(x,y) = C$ is the vector $(\phi_x, \phi_y)/(\phi_x^2 + \phi_y^2)^{1/2}$. By similar reasoning, we find that if (x,y) and $(x + dx, y + dy)$ are points on two neighboring curves having parameters C and $C + dC$, respectively, then $\phi_x dx + \phi_y dy = dC$. Now if (dx, dy) is perpendicular to $\phi(x,y) = C$, then $dx = ds\phi_x/(\phi_x^2 + \phi_y^2)^{1/2}$ and $dy = ds\phi_y/(\phi_x^2 + \phi_y^2)^{1/2}$, where ds is the normal distance between curves C and $C + dC$ at (x,y). Substituting these values for dx and dy into the expression for dC, we finally obtain $ds(\phi_x^2 + \phi_y^2)^{1/2} = dC$.

At a separatrix, $ds/dC = 0$. This is because curves corresponding to a finite interval of dC are packed into an infinitesimally small normal distance from the separatrix. Said another way, at a separatrix, the density dC/ds of integral curves is infinite. Now $ds/dC = (\phi_x^2 + \phi_y^2)^{-1/2} = \mu^{-1}(M^2 + N^2)^{-1/2}$. If, as in Eq. (20), neither M nor N is ever infinite, ds/dC can only vanish if $\mu^{-1} = 0$. So $\mu^{-1} = 0$ may specify separatrices.

We can check this with the example begun with Eq. (20). Figure 15 shows a plot of the family of curves given by Eq. (23). From the diagram, we can see that there are three separatrices that divide the plane into six parts. The separatrices are the lines $y = 0$, $x = 0$, and $y = x$. The integrating factor $\mu = y^{-2}$ gives the

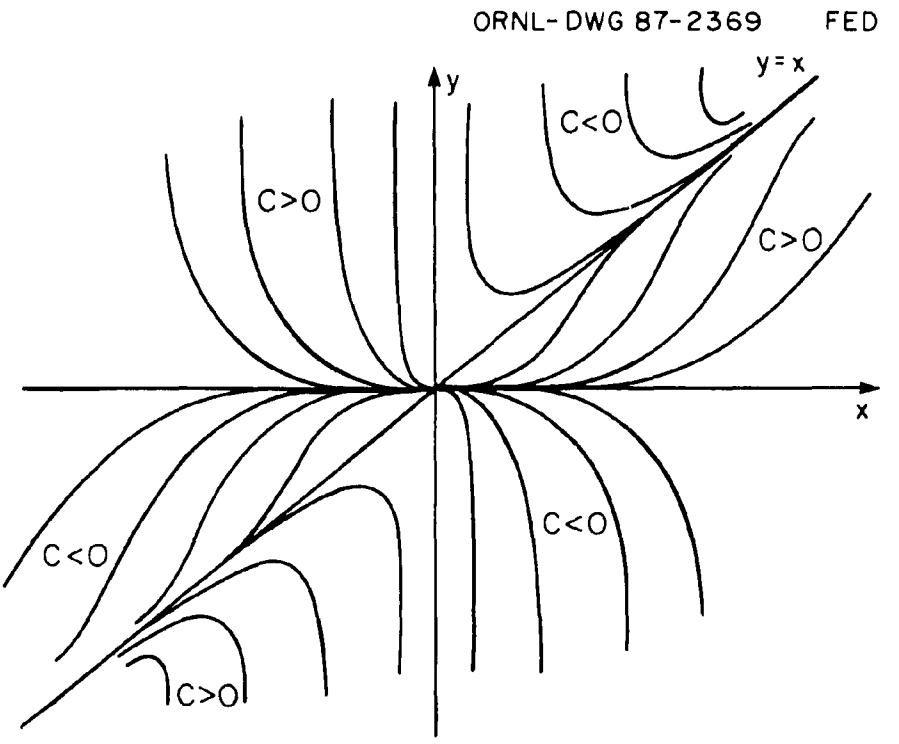

Fig. 15. The family of curves $y = x^2/(x + C)$, $-\infty < C < \infty$.

separatrix $y = 0$. The integrating factor $\nu = x^{-2}(x - y)^{-2}$ gives the separatrices $x = 0$ and $y = x$. From this example, we see that knowing one integrating factor may not be enough to find all the separatrices without integration, though we may find some. If we know two integrating factors, of course, we can find all the integral curves without integration.

Chapter 2
THE LIE THEORY OF DIFFERENTIAL EQUATIONS

"Plus ça change, plus c'est la même chose."
—Alphonse Karr
Les Guêpes

2.1 Lie has given a method of finding an integrating factor if the differential equation is invariant to a one-parameter group of transformations. What this last phrase means is best made clear by means of an example. If we change the variables in the differential equation (1.20)* to x' and y', where

$$\begin{aligned} x' &= \lambda x \\ y' &= \lambda y \end{aligned} \quad , \qquad 0 < \lambda < \infty \tag{1}$$

and λ is some positive constant, then the resulting differential equation in the primed variables is identical to the original differential equation (1.20) in the unprimed variables. This is true no matter what the value of λ is, as long as it is not zero. The differential equation (1.20) is said then to be invariant to the transformations of Eq. (1).

The transformations of Eq. (1) are said to be a group because they obey the three group postulates, namely: (i) Two transformations carried out in succession are equivalent to some other single transformation of the group. Thus, if $x' = \lambda_1 x$, $y' = \lambda_1 y$, and $x'' = \lambda_2 x'$, $y'' = \lambda_2 y'$, then $x'' = \lambda_1\lambda_2 x$, $y'' = \lambda_1\lambda_2 y$. (ii) There is an identity transformation, i.e., one that leaves the variables x, y unchanged. For the transformations of Eq. (1), the identity transformation is the one for which $\lambda = 1$. (iii) For every transformation, there is an inverse, i.e., a second transformation that undoes the effect of the first. For the transformations of Eq. (1), the inverse transformation has $\lambda_2 = 1/\lambda_1$. (Thus $x'' = x$ and $y'' = y$.)

A first-order differential equation is logically equivalent to a one-parameter family of integral curves, and so the family, too, must be transformed into itself by a group under which the differential equation is invariant. In general, each curve of the family has as its image under transformation some other curve of the family, and only certain exceptional curves transform into themselves. For example, the integral curve $y = x^2/(x + C)$ transforms under the transformations of Eq. (1) into the integral curve $y' = x'^2/(x' + \lambda C)$. So the integral curve belonging to label C has as its image the integral curve with label λC. Only the curves for $C = 0$ and $C = \infty$ transform into themselves.

Lie's method of constructing an integrating factor is based on the observation that the image of an integral curve is another integral curve. Represent the family of integral curves as $\phi(x, y) = C$ and focus attention on the curve Q for which $C = C_0$. Transform each point (x, y) of Q into its image (x', y'); denote the locus of these images as curve Q'. The curve Q' is also an integral curve belonging to a

*That is, Eq. (20) in Chap. 1.

label C that depends on λ and C_0. So we can write $\phi(x',y') = C(\lambda, C_0)$ or, using Eq. (1) to replace primed variables by unprimed variables,

$$\phi(\lambda x, \lambda y) = C(\lambda, C_0) = C[\lambda, \phi(x,y)] \quad . \tag{2}$$

Now we differentiate with respect to λ and then set $\lambda = 1$:

$$x\phi_x + y\phi_y = \frac{\partial C(\lambda,\phi)}{\partial \lambda}\Big|_{\lambda=1} \equiv F(\phi) \quad . \tag{3}$$

Since $\phi_x = \mu M$ and $\phi_y = \mu N$, Eq. (3) can be written

$$\mu = \frac{F(\phi)}{xM + yN} \quad . \tag{4}$$

As we saw in Chap. 1, if we multiply an integrating factor by any function of ϕ we get another integrating factor. Therefore, $\mu = (xM + yN)^{-1}$ must be an integrating factor. Since $M = y^2 - 2xy$ and $N = x^2$ for the differential equation (1.20), $\mu^{-1} = xM + yN = xy(y-x)$, which satisfies Eq. (1.21), as it should. Interestingly, this integrating factor yields all three separatrices $x = 0$, $y = 0$, and $y = x$ when μ^{-1} is set to zero.

2.2 Lie considered groups more general than the simple stretching group of Eq. (1). We can write the most general one-parameter family of transformations of x and y in the form

$$x' = X(x,y,\lambda) \quad , \tag{5a}$$

$$y' = Y(x,y,\lambda) \quad . \tag{5b}$$

The functions of X and Y cannot be chosen arbitrarily because of the requirement that they conform to the group property that two such transformations executed in succession are equivalent to a certain other transformation. The restrictions on X and Y may be found, as Lie has proposed, by composing finite transformations out of a succession of infinitesimal transformations. This means the following.

Suppose $\lambda = \lambda_0$ corresponds to the identity transformation. When $\lambda - \lambda_0$ is very small, i.e., when λ is close to λ_0, Eqs. (5a,b) can be replaced by the linear terms in their Taylor series around $\lambda = \lambda_0$:

$$x' = x + \xi(x,y)(\lambda - \lambda_0) + \ldots \quad , \tag{6a}$$

$$y' = y + \eta(x,y)(\lambda - \lambda_0) + \ldots \quad , \tag{6b}$$

where

$$\xi(x,y) \equiv \frac{\partial X(x,y,\lambda)}{\partial \lambda}\Big|_{\lambda=\lambda_0} \quad \text{and} \quad \eta(x,y) = \frac{\partial Y(x,y,\lambda)}{\partial \lambda}\Big|_{\lambda=\lambda_0} \quad . \tag{6c}$$

The meaning of Eqs. (6a,b) is that nearby images of the point (x,y) lie on a small line segment through (x,y) having the slope $(y'-y)/(x'-x) = \eta(x,y)/\xi(x,y)$. The

transformations of Eqs. (6a,b) are infinitesimal transformations. The geometric interpretation of composing a finite transformation out of a succession of infinitesimal transformations is that we reach a remote image of (x,y) by stepping successively along a series of neighboring points, each of which is a nearby image of its predecessor. That all these points are images of one another follows from the group property. The locus traced out by this series of steps has the slope η/ξ everywhere and hence is an integral curve of the differential equation

$$\frac{dy}{dx} = \frac{\eta(x,y)}{\xi(x,y)} \quad . \tag{7}$$

These integral curves are called the orbits of the group. If we parameterize the points of an orbit by setting $d\lambda = dx/\xi = dy/\eta$, we obtain the functions X and Y by integrating these differential equations. Replacing $d\lambda$ by $F(\lambda)d\lambda$, where F is any function of λ, just corresponds to a different parameterization of the points of the orbit. *The group is thus entirely characterized by the two functions* $\xi(x,y)$ *and* $\eta(x,y)$.

For the simple stretching group of Eq. (1), $\xi \equiv (\partial x'/\partial\lambda)_{\lambda=1} = x$ and $\eta \equiv (\partial y'/\partial\lambda)_{\lambda=1} = y$. The orbits are then straight lines through the origin. If we parameterize the orbits according to $d\lambda = dx/\xi = dy/\eta$, we obtain by integration $x = x_0 e^{\lambda-\lambda_0}$ and $y = y_0 e^{\lambda-\lambda_0}$, which has the same form as Eq. (1) if we identify $e^{\lambda-\lambda_0}$ here with λ there. If we parameterize the orbits according to $d\lambda/\lambda = dx/\xi = dy/\eta$, we obtain by integration $x = x_0(\lambda/\lambda_0)$, $y = y_0(\lambda/\lambda_0)$, which is the same as Eq. (1) if we choose λ_0, the parameter corresponding to the identity transformation, to be 1.

The orbits, being composed of points which transform into one another, are invariant curves, i.e., they transform into themselves. They are moreover the only invariant curves. Separatrices are invariant curves because they separate two invariant families of curves. They are also integral curves of the differential equation. So they must simultaneously satisfy the differential equations $dy/dx = \eta/\xi$ and $dy/dx = -M/N$. Equating these slopes we get the *algebraic* equation $\xi M + \eta N = 0$ for invariant integral curves. This equation must include all the separatrices.

Now we can find Lie's general expression for an integrating factor in terms of the components ξ and η of the infinitesimal transformation. We start again with the relation

$$\phi(x',y') = C(\lambda, C_0) = C[\lambda, \phi(x,y)] \tag{8}$$

and again differentiate with respect to λ and then set $\lambda = \lambda_0$. We get

$$\xi\phi_x + \eta\phi_y = F(\phi) \quad , \tag{9}$$

since $(\partial x'/\partial\lambda)_{\lambda=\lambda_0} \equiv \xi$ and $(\partial y'/\partial\lambda)_{\lambda=\lambda_0} \equiv \eta$. Proceeding now exactly as before, we find that

$$\mu = (\xi M + \eta N)^{-1} \tag{10}$$

is an integrating factor. The algebraic equation $\xi M + \eta N = 0$ derived in the last paragraph for the separatrices is thus the same as the earlier result $\mu^{-1} = 0$.

Furthermore, we know now that when μ is Lie's integrating factor, $\mu^{-1} = 0$ gives all the separatrices.

2.3 If explicit expressions for X and Y are available, it is relatively easy to decide whether a given differential equation is invariant to a given group. But if no explicit expressions are available, i.e., if Eq. (7) cannot be integrated explicitly, how can we answer this question? To do so, we need the transformation law for the derivative $\dot{y}$, which, as we shall now see, is entirely determined by the transformation laws for x and y. Suppose we consider two neighboring points $P_1 : (x,y)$ and $P_2 : (x+dx, y+dy)$ joined by a short line segment whose slope is $\dot{y} = dy/dx$. Under the infinitesimal transformation with parameter $d\lambda = \lambda - \lambda_0$, P_1 goes into the point $P_1' : (x', y')$ and P_2 into the point $P_2' : (x'+dx', y'+dy')$, where

$$\begin{aligned} x' &= x + \xi(x,y)d\lambda \quad , \\ y' &= y + \eta(x,y)d\lambda \quad , \end{aligned} \tag{11}$$

$$\begin{aligned} x' + dx' &= x + dx + \xi(x+dx, y+dy)d\lambda \quad , \\ y' + dy' &= y + dy + \eta(x+dx, y+dy)d\lambda \quad . \end{aligned} \tag{12}$$

The slope $\dot{y}' = dy'/dx'$ of the segment $P_1'\ P_2'$ is thus completely determined by the transformation laws for x and y:

$$\dot{y}' = \frac{dy'}{dx'} = \frac{dy + [\eta(x+dx, y+dy) - \eta(x,y)]d\lambda}{dx + [\xi(x+dx, y+dy) - \xi(x,y)]d\lambda} \quad . \tag{13}$$

If we expand the square brackets to first order in dy and dx and divide the numerator and denominator of the right-hand side by dx, Eq. (13) becomes

$$\dot{y}' = \frac{\dot{y} + (\eta_x + \eta_y \dot{y})d\lambda}{1 + (\xi_x + \xi_y \dot{y})d\lambda} \tag{14a}$$

$$= \dot{y} + (\eta_x + \eta_y \dot{y} - \xi_x \dot{y} - \xi_y \dot{y}^2)d\lambda \tag{14b}$$

$$= \dot{y} + \left(\frac{d\eta}{dx} - \dot{y}\frac{d\xi}{dx}\right) d\lambda \quad , \tag{14c}$$

where d/dx applied to a function of (x,y) means $\partial/\partial x + \dot{y}\partial/\partial y$. The quantity

$$\eta_d = \frac{d\eta}{dx} - \dot{y}\frac{d\xi}{dx} \tag{15}$$

is Lie's expression for the component of the extended infinitesimal transformation belonging to $\dot{y}$.

A first-order differential equation is a functional relation connecting x, y, and $\dot{y}$:

$$g(x, y, \dot{y}) = 0 \quad . \tag{16}$$

If it is invariant to the extended infinitesimal transformation with components $\xi(x,y)$, $\eta(x,y)$, and $\eta_d(x,y,\dot{y})$, it will be invariant to the entire group equation (5)

(since the transformation of the group can be composed of a succession of infinitesimal transformations). Invariance means that $g(x', y', \dot{y}') = 0$, where x', y', and $\dot{y}'$ are the images of x, y, and $\dot{y}$. Thus

$$g(x + \xi d\lambda, y + \eta d\lambda, \dot{y} + \eta_d d\lambda) = 0 \quad . \tag{17}$$

From Eqs. (16) and (17) follows the condition

$$\xi g_x + \eta g_y + \eta_d g_{\dot{y}} = 0 \tag{18}$$

that the differential equation (16) is invariant to the group equation (5).

The condition equation (17) can be looked upon as a first-order *linear* partial differential equation for g if we imagine that ξ and η are known. Its general solution therefore supplies the answer to the question, "What is the most general first-order ordinary differential equation invariant to the group whose infinitesimal transformation has the components ξ and η?" The general solution for a first-order linear partial differential equation like Eq. (18) can be obtained by integrating the characteristic equations

$$\frac{dx}{\xi} = \frac{dy}{\eta} = \frac{d\dot{y}}{\eta_d} \quad . \tag{19}$$

If we can find two independent integrals* $u(x, y, \dot{y})$ and $v(x, y, \dot{y})$ of Eq. (19), the general solution can be obtained by setting $v = F(u)$, where F is any arbitrary function. The equation $v = F(u)$ then gives the most general functional relation between x, y, and $\dot{y}$ that is invariant to the group with infinitesimal components ξ, η. If we seek explicit representation of the most general differential equation, we shall have to have explicit representations of both u and v. Eliminating $\dot{y}$ between these two integrals of Eq. (19) gives an integral of the first part of Eq. (19), $dx/\xi = dy/\eta$. Such an integral is an explicit representation of the orbits of the group ξ, η. So we shall be able to attain an explicit representation for the most general differential equation at best for all groups for which an explicit representation of the orbits is also possible.

Any one-parameter family of curves can serve as the orbits of a group; for example, the family

$$y = \frac{x^2}{x + u} \qquad (u = \text{parameter}) \quad . \tag{20}$$

This family has as its differential equation Eq. (1.20), which, when written in the form

$$\frac{dx}{x^2} = \frac{dy}{2xy - y^2} \quad , \tag{21}$$

allows us to identify the infinitesimal components of the group:

$$\xi = x^2 \quad , \quad \eta = 2xy - y^2 \quad . \tag{22}$$

*An integral is a function of x, y, and $\dot{y}$ whose value remains constant as we move along a curve in x, y, $\dot{y}$ space whose direction is given by Eq. (19).

Then, Eq. (15) gives

$$\eta_d = 2y(1-\dot{y}) \quad . \tag{23}$$

The characteristic equations, Eq. (19), are then

$$\frac{dx}{x^2} = \frac{dy}{2xy-y^2} = \frac{d\dot{y}}{2y(1-\dot{y})} \quad . \tag{24}$$

From Eq. (20) it follows that $u = x^2/y - x$ is one integral of Eq. (24). We can find a second integral by substituting for y from Eq. (20) in the last term of Eq. (24). Then

$$\frac{dx}{x^2} = \frac{d\dot{y}}{2[x^2/(x+u)](1-\dot{y})} \tag{25a}$$

or

$$\frac{d\dot{y}}{1-\dot{y}} = \frac{x+u}{2}dx \quad . \tag{25b}$$

Integrating Eq. (25b), we get

$$-\ln(1-\dot{y}) = \frac{x^2}{4} + \frac{xu}{2} + F(u) \quad , \tag{26}$$

where v, the constant of integration on the right-hand side, has been set equal to $F(u)$, an arbitrary function of u. Since $u = x^2/y - x$, Eq. (26) can be written finally as

$$\dot{y} = 1 - \exp\left(\frac{x^2}{4} - \frac{x^3}{2y}\right) G\left(\frac{x^2}{y} - x\right) \quad , \tag{27}$$

where $G = e^{-F}$ is also an arbitrary function of its argument. Equation (27) is the most general first-order differential equation invariant to the group whose infinitesimal components are given in Eq. (22).

The infinitesimal components in Eq. (22) are not the only ones that reduce the equation of the orbits, Eq. (7), to Eq. (21). Components obtained by multiplying Eq. (22) by a common factor will work just as well. Thus, the orbits do not uniquely determine the group, and different groups may have the same orbits. This is made clear by an example simpler than the foregoing one. Suppose the orbits are the lines that radiate from the origin, $y = ux$. Then their differential equation is $dy/y = dx/x$. If we choose $\xi = x$ and $\eta = y$, we are led to the most general differential equation $\dot{y} = F(y/x)$, where F can be any function. If, on the other hand, we choose $\xi = x^{a+1}$, $\eta = yx^a$, we are led to the most general differential equation $\dot{y} = y/x + x^{-a}F(y/x)$. If we choose $\xi = x^2y$, $\eta = xy^2$, which also leads to the orbits $y = ux$, we find the most general differential equation is $\dot{y} = (y/x)[x^2F(y/x) - 1]/[x^2F(y/x) + 1]$.

In the manner just outlined, we can construct tables of first-order differential equations for which groups and therefore integrating factors are known. Cohen gives such a table (A. Cohen, *An Introduction to the Lie Theory of One-Parameter Groups*, G. E. Stechert and Co., New York, 1931).

2.4 An alternative to using an integrating factor to solve a first-order differential equation is to separate variables. Lie has shown how to find, by means of the group, new variables in which the differential equation is separable. According to Cohen, this method antedates Lie's discovery of the integrating factor by five years, having been discovered in 1869.

Suppose we change variables from x, y to new coordinates x_1, y_1, where x_1 and y_1 are prescribed functions of x and y. To each point P in the plane belong a pair of values (x,y) and another pair (x_1,y_1) calculable from (x,y). Under the transformation with parameter λ the point $P:(x,y)$ is transformed into its image point $P':(x',y')$, where x' and y' are calculable from Eqs. (5a) and (5b). From (x',y') we can calculate x_1' and y_1', the new coordinates of P'. This procedure implicitly defines a pair of functions X_1 and Y_1 such that $x_1' = X_1(x_1,y_1,\lambda)$ and $y_1' = Y_1(x_1,y_1,\lambda)$.

Now

$$\xi_1 = \left(\frac{\partial x_1}{\partial \lambda}\right)_{\lambda=\lambda_0} = \frac{\partial x_1}{\partial x}\left(\frac{\partial x}{\partial \lambda}\right)_{\lambda=\lambda_0} + \left(\frac{\partial x_1}{\partial y}\right)\left(\frac{\partial y}{\partial \lambda}\right)_{\lambda=\lambda_0}$$
$$= \xi\frac{\partial x_1}{\partial x} + \eta\frac{\partial x_1}{\partial y} \quad , \tag{28a}$$

and similarly

$$\eta_1 = \left(\frac{\partial y_1}{\partial \lambda}\right)_{\lambda=\lambda_0} = \frac{\partial y_1}{\partial x}\left(\frac{\partial x}{\partial \lambda}\right)_{\lambda=\lambda_0} + \frac{\partial y_1}{\partial y}\left(\frac{\partial y}{\partial \lambda}\right)_{\lambda=\lambda_0}$$
$$= \xi\frac{\partial y_1}{\partial x} + \eta\frac{\partial y_1}{\partial y} \quad . \tag{28b}$$

Lie has chosen as canonical variables x_1, y_1 those for which $\xi_1 = 0$ and $\eta_1 = 1$. The functional dependence of these canonical variables on the original variables x, y may then be found by solving the pair of first-order partial differential equations

$$\xi\frac{\partial x_1}{\partial x} + \eta\frac{\partial x_1}{\partial y} = 0 \quad , \tag{29a}$$

$$\xi\frac{\partial y_1}{\partial x} + \eta\frac{\partial y_1}{\partial y} = 1 \quad . \tag{29b}$$

Any particular pair of solutions x_1, y_1 of Eqs. (29a) and (29b) will provide satisfactory canonical coordinates for which $\xi_1 = 0$ and $\eta_1 = 1$.

The characteristic equations for the linear partial differential equations (29a) and (29b) are

$$\frac{dx}{\xi} = \frac{dy}{\eta} \tag{30a}$$

and

$$\frac{dx}{\xi} = \frac{dy}{\eta} = dy_1 \quad . \tag{30b}$$

Since Eq. (30a) is the same as Eq. (7), its integral gives the equations of the orbits. If we have an explicit representation of the orbits, we already have an integral of the first equation of Eq. (30b), so finding the second integral involves only two quadratures.

When $\xi_1 = 0$ and $\eta_1 = 1$, $\eta_{1d} = 0$ according to Eq. (15). Equation (19) then becomes $dx_1/0 = dy_1/1 = d\dot{y}_1/0$, for which two integrals are $u = x_1$ and $v = \dot{y}_1$. So the most general differential equation invariant to the group $\xi_1 = 0$, $\eta_1 = 1$ is $\dot{y}_1 = F(x_1)$, *which is separable.*

As an illustration, let us pursue the last example in Para. 2.3 in which $\xi = x^2y$ and $\eta = xy^2$. An integral of Eq. (30a) is y/x, so we can take x_1 to be any function of y/x. The simplest choice is $x_1 = y/x$. This function is also an integral of Eq. (30b). If we substitute it in the expression for η, the last equality of Eq. (30b) becomes $dy_1 = dx/x_1x^3$, which is satisfied by $y_1 = -1/2x^2x_1 = -1/2xy$. Thus $y_1 = -1/2xy$, $x_1 = y/x$ are a suitable pair of canonical coordinates. If we use them in the most general differential equation $\dot{y} = (y/x)[x^2F(y/x) - 1]/[x^2F(y/x) + 1]$, it becomes the separable equation $y_1 = -F(x_1)/2x_1^2$.

In the important special case that the group is an affine (stretching) group, replacement of the dependent variable y by a group invariant causes the differential equation to separate. [A group invariant is a function $u(x,y)$, which transforms into itself under the action of the group.] The most general stretching group in two variables is

$$y' = \lambda^\beta y \quad , \tag{31}$$
$$x' = \lambda x \quad ,$$

where β is a constant. We lose no generality by making the exponent of the multiplier of x equal to 1. The transformation law for y is then

$$\dot{y}' = \frac{dy'}{dx'} = \frac{\lambda^\beta dy}{\lambda\, dx} = \lambda^{\beta-1}\dot{y} \quad . \tag{32}$$

We write the differential equation in the form $\dot{y} = f(x,y)$. If this differential equation is to be invariant to Eqs. (31) and (32), it must have the same form in the primed variables, namely, $\dot{y}' = f(x',y')$ or

$$\lambda^{\beta-1}\dot{y} = f(\lambda x, \lambda^\beta y) \tag{33a}$$

or

$$\lambda^{\beta-1}f(x,y) = f(\lambda x, \lambda^\beta y) \quad . \tag{33b}$$

Differentiating with respect to λ and setting $\lambda = 1$, we obtain

$$(\beta - 1)f = xf_x + \beta y f_y \quad , \tag{34}$$

a linear partial differential equation for f. The characteristic equations are

$$\frac{dx}{x} = \frac{dy}{\beta y} = \frac{df}{(\beta - 1)f} \quad . \tag{35}$$

Two integrals of these characteristic equations are y/x^β and $f/x^{\beta-1}$, so the general solution of Eq. (35) is

$$\frac{f}{x^{\beta-1}} = F\left(\frac{y}{x^\beta}\right) \quad , \tag{36}$$

where F is an arbitrary function. This equation expresses the restriction of the form of f imposed by the condition of invariance of the differential equation to the stretching group equation (31).

An invariant of the group is the function $u = y/x^\beta$. If we replace y by u we shall get a separable differential equation. For

$$\begin{aligned}\frac{du}{dx} &= \frac{y}{x^\beta} - \frac{\beta y}{x^{\beta+1}} = \frac{1}{x}\left(\frac{f}{x^{\beta-1}} - \beta\frac{y}{x^\beta}\right) \\ &= \frac{1}{x}\left[F\left(\frac{y}{x^\beta}\right) - \beta\frac{y}{x^\beta}\right] = \frac{1}{x}[F(u) - \beta u] \quad ,\end{aligned} \tag{37a}$$

so that

$$\frac{dx}{x} = \frac{du}{F(u) - \beta u} \quad . \tag{37b}$$

2.5 Lie also considered second-order differential equations. Such equations have the general form $g(x, y, \dot{y}, \ddot{y}) = 0$. To test whether such an equation is invariant to the group with infinitesimal components ξ, η we must calculate the transformation law for the second derivative. A computation following the line from Eq. (12) to Eq. (15) gives

$$\eta_{dd} = \frac{d\eta_d}{dx} - \ddot{y}\frac{d\xi}{dx} \equiv (\eta_d)_x + \dot{y}(\eta_d)_y + \ddot{y}(\eta_d)_{\dot{y}} - \ddot{y}\xi_x - \ddot{y}\dot{y}\xi_y \tag{38}$$

for the component of the extended infinitesimal transformation belonging to $\ddot{y}$. (Remember η_d is a function of $x, y,$ *and* $\dot{y}$!) The invariance of $g(x, y, \dot{y}, \ddot{y}) = 0$ means that

$$\xi g_x + \eta g_y + \eta_d g_{\dot{y}} + \eta_{dd} g_{\ddot{y}} = 0 \quad , \tag{39}$$

which is derived exactly as Eq. (18) was.

Suppose now we imagine the second-order differential equation solved for $\ddot{y}$:$\ddot{y} = f(x, y, \dot{y})$. Introduce the new variable $z = \dot{y}$. Then the second-order differential equation becomes a pair of coupled first-order differential equations,

$$\begin{aligned}\dot{z} &= f(x, y, z) \quad , \\ \dot{y} &= z \quad ,\end{aligned} \tag{40}$$

which can be written in the form

$$\frac{dz}{f(x, y, z)} = \frac{dy}{z} = dx \quad , \tag{41}$$

which is slightly more transparent than Eq. (40) for the purposes of this discussion.

Equation (44) determines a line element at every point (x, y, z) in three-dimensional space. The totality of these line elements comprises the direction field of the second-order differential equation. In these days of powerful computer graphics it is not overly ambitious to aspire to plot this direction field, but even if we could, comprehending its content at a glance probably would tax our skills beyond their limits. But the concept of a three-dimensional direction field is not without its use. The direction field determines a two-parameter family of integral curves that fill all of space. (That two parameters are involved can be seen by noting that the intersection of a curve with some fiducial plane is specified by two coordinates that can serve to identify the curve.) If the differential equation is invariant to the group (ξ, η), this family of integral curves must be transformed into itself by the group since it is logically equivalent to the differential equation. (In transforming the curves, of course, η_d is used as the component of the infinitesimal transformation belonging to z.)

The image of an integral curve of the family is another integral curve of the family. Since the group is a *one-parameter* group of transformations, a curve and all its images form a one-parameter family of curves in space, i.e., a surface. This surface, by the manner of its construction, is furthermore invariant to the group, i.e., it transforms into itself.

An invariant surface $h(x, y, z) = 0$ in three-dimensional space must satisfy the relation

$$\xi h_x + \eta h_y + \eta_d h_z = 0 \tag{42}$$

[since $h(x + \xi d\lambda,\ y + \eta d\lambda,\ z + \eta_d d\lambda)$ also equals 0]. The characteristic equations of (42) are

$$\frac{dx}{\xi} = \frac{dy}{\eta} = \frac{dz}{\eta_d} \ . \tag{43}$$

If we know two integrals of Eq. (43), $u(x, y)$ and $v(x, y, z)$, the most general solution of Eq. (42) is $h(x, y, z) = F(u, v) = 0$, where F is an arbitrary function. This is the most general form of surface invariant to the group (ξ, η, η_d).

The one-parameter family of invariant surfaces into which the integral curves can be grouped thus takes the form $F(u, v, C) = 0$, where C is the parameter labeling the individual surfaces. But such a form corresponds to a one-parameter family of curves in the (u, v) plane. Such a one-parameter family of curves is logically identical to a first-order differential equation in u and v. So introduction of the new variables $u(x, y)$ and $v(x, y, \dot{y})$ into the second-order differential equation reduces it to a first-order differential equation.

Because u and v are integrals of Eq. (43), they are invariant to transformations of the group, i.e., they are group invariants. The invariant v, because it involves $\dot{y}$ as well as x and y, is called a first differential invariant. So we may state Lie's very important theorem about second-order ordinary differential equations as follows: if we introduce as new variables an invariant and a first differential invariant of a group leaving a second-order ordinary differential equation invariant, the differential equation reduces to first order. The importance of this theorem is that we can comprehend the contents of the first-order ordinary differential equation "at a glance."

As an illustration, let us choose the Emden-Fowler equation, which, as tioned in Chap. 1, arises in the study of the equilibrium mass distribu cloud of gas held together by gravity. We specialize first to a gas with a specific heats (adiabatic exponent) of 4/3. The Emden-Fowler equation the form

$$\ddot{y} + \frac{2\dot{y}}{x} + y^3 = 0 \quad . \tag{44}$$

This differential equation is invariant to the affine (stretching) group

$$\begin{aligned} y' &= \lambda^{-1} y \quad , \\ x' &= \lambda x \quad . \end{aligned} \tag{45}$$

To see that this is true we calculate the transformation laws for $\dot{y}$ and $\ddot{y}$:

$$\dot{y}' = \lambda^{-2}\dot{y} \quad \text{and} \quad \ddot{y}' = \lambda^{-3}\ddot{y} \quad . \tag{46}$$

So if we imagine the differential equation (44) written in the primed form and use Eqs. (45) and (46) to transform to the unprimed form, each term in Eq. (44) individually is multiplied by the factor λ^{-3}. This common factor can be cancelled, so that in the unprimed form Eq. (44) has precisely the same form as in the primed form.

Because a multiplicative group like Eq. (45) will cause each term in Eq. (44) to be multiplied by a power of λ, the computations outlined above can be done in one's head. When we see an equation like Eq. (44) whose terms are products of powers of $x, y, \dot{y}$, and $\ddot{y}$, we should at once test to see if it is invariant to a stretching group. Since the most general stretching group in two variables has the form

$$\begin{aligned} y' &= \lambda^{\beta} y \quad , \\ x' &= \lambda x \quad , \end{aligned} \tag{47}$$

the transformation laws for $\dot{y}$ and $\ddot{y}$ are

$$\begin{aligned} \dot{y}' &= \lambda^{\beta-1}\dot{y} \quad , \\ \ddot{y}' &= \lambda^{\beta-2}\ddot{y} \quad . \end{aligned} \tag{48}$$

If we imagine Eq. (44) to be written in the primed form and transform to the unprimed form, the terms in Eq. (44) are multiplied by the factors $\lambda^{\beta-2}$, $\lambda^{\beta-2}$, and $\lambda^{3\beta}$, respectively. In order for these terms to be equal (so we can cancel them as a common factor), $\beta - 2$ must equal 3β, i.e., β must be -1.

We can now write down an invariant u and a first differential invariant v for the group equation (45) at once:

$$u = xy, \quad v = x^2\dot{y} \quad . \tag{49}$$

The choices of Eq. (49) are not the only possible ones [$u = xy$, $v = \dot{y}/y^2$ or $u = x^2y^2$, $v = (\dot{y}/y^2)\exp(xy)$ are also possible]. However, Eq. (49) is a suitable choice. Then

$$\frac{dv}{dx} = 2x\dot{y} + x^2\ddot{y} = 2x\dot{y} + x^2\left(-\frac{2\dot{y}}{x} - y^3\right) = -x^2y^3 = -u^3/x \quad , \tag{50a}$$

$$\frac{du}{dx} = y + x\dot{y} = (u+v)/x \quad . \tag{50b}$$

Here we have eliminated $\ddot{y}$ using the differential equation (44) and then eliminated y and $\dot{y}$ in favor of u and v. Upon dividing, we get the first-order differential equation

$$\frac{dv}{du} = -\frac{u^3}{u+v} \quad . \tag{50c}$$

This differential equation was studied in Chap. 1, where it was given as Eq. (1.20).

The Emden-Fowler equation for a gas of adiabatic exponent 6/5 is

$$\ddot{y} + \frac{2}{x}\dot{y} + y^5 = 0 \quad , \tag{51}$$

and this can easily be shown to be invariant to the group

$$\begin{aligned} y' &= \lambda^{-1/2} y \quad , \\ x' &= \lambda x \quad . \end{aligned} \tag{52}$$

An invariant and a first differential invariant are $u = y\sqrt{x}$ and $v = \dot{y}x^{3/2}$. Differentiating them with respect to x, we find

$$\frac{dv}{du} = -\frac{v + 2u^5}{2v + u} \quad , \tag{53}$$

which can be integrated explicitly! Writing Eq. (53) as

$$2v\ dv + u\ dv + v\ du + 2u^5 du = 0 \quad , \tag{54a}$$

we see that it is already in the form of a perfect differential. Thus

$$3v^2 + 3uv + u^6 = \text{const} \quad . \tag{54b}$$

If we now replace u and v by their equivalents in terms of x, y, and $\dot{y}$, we find that Eq. (54b) is equivalent to

$$3x^3\dot{y}^2 + 3y\dot{y}x^2 + x^3y^6 = \text{const} \quad . \tag{54c}$$

So we are faced with the task of integrating another first-order differential equation. But because Eq. (54c) is equivalent to Eq. (54b), and because Eq. (54b) is invariant to Eq. (52) (it is composed of invariants!), Eq. (54c) must be invariant to Eq. (52). This means that, for example, we can separate variables by introducing an invariant in place of y [remember, Eq. (52) is a stretching group]. A convenient choice is $w = u^2 = xy^2$, which causes Eq. (54c) to separate:

$$\frac{dx}{x} = \frac{\sqrt{3}}{2}\frac{dw}{(3w^2/4 - w^4 + \text{const} \cdot w)^{1/2}} \quad . \tag{55}$$

Because the interpretation of y is a gravitational potential, the physically interesting solution of Eq. (51) is the one for which y is finite at the origin and has zero derivative

there. For that solution, the constant in Eqs. (54c) and (55) must be zero. Then Eq. (55) can be integrated by setting $w = (\sqrt{3}/2)\sin(\theta/2)$. We find after some tedious computation

$$w = \frac{3ax}{x^2 + 3a^2}, \quad y = \left(\frac{3a}{x^2 + 3a^2}\right)^{1/2}, \quad a = \text{constant of integration} \ . \tag{56}$$

2.6 The Emden-Fowler equation (51) could be solved analytically after its reduction to the first-order differential equation (53). This must be counted as good fortune and is generally not the case. How then do we proceed? The answer to this question is best given by means of an example, the solution of the Thomas-Fermi equation. This nonlinear second-order equation arises in the determination of the screening of the Coulomb potential of a nucleus by the electron cloud surrounding it. It has the form

$$x^{1/2}\ddot{y} = y^{3/2} \ , \tag{57}$$

where x is the radial coordinate (in suitable atomic units) and y is a multiplicative correction factor to the unshielded nuclear Coulomb potential. The integral curve of Eq. (57) we seek is one for which

$$y(0) = 1 \ \text{ and } \ y(\infty) = 0 \ . \tag{58}$$

Since Eq. (57) is composed of products of powers of x, y, and $\ddot{y}$, we try the stretching group equation (47). Substituting Eqs. (47) and (48) into the primed form of Eq. (57), we find $\beta - 3/2 = (3/2)\beta$ as the condition for invariance. Thus $\beta = -3$, and Eq. (57) is invariant to the stretching group

$$\begin{aligned} x' &= \lambda x \ , \\ y' &= \lambda^{-3} y \ . \end{aligned} \tag{59}$$

If we use $u = x^3 y$ and $v = x^4\dot{y}$ as an invariant and a first differential invariant, we find

$$x\frac{dv}{dx} = 4x^4\dot{y} + x^5\ddot{y} = 4x^4\dot{y} + x^5(x^{-1/2}y^{3/2}) = 4v + u^{3/2} \ , \tag{60a}$$

$$x\frac{du}{dx} = 3x^3 y + x^4\dot{y} = 3u + v \ , \tag{60b}$$

so that

$$\frac{dv}{du} = \frac{4v + u^{3/2}}{3u + v} \ . \tag{60c}$$

Equation (60c) is not explicitly integrable in terms of elementary functions, so we shall turn to an analysis of its direction field to help us to solve Eqs. (57) and (58). Now since x is positive (being a radius) and y varies between 0 and 1, we may guess y is positive and $\dot{y}$ negative. Therefore $u = x^3 y > 0$ and $v = x^4\dot{y} < 0$, so we shall only be interested in the fourth quadrant of the (u, v) plane. Figure 1 shows the direction field of Eq. (60c) in this quadrant. The curve of zero slope is

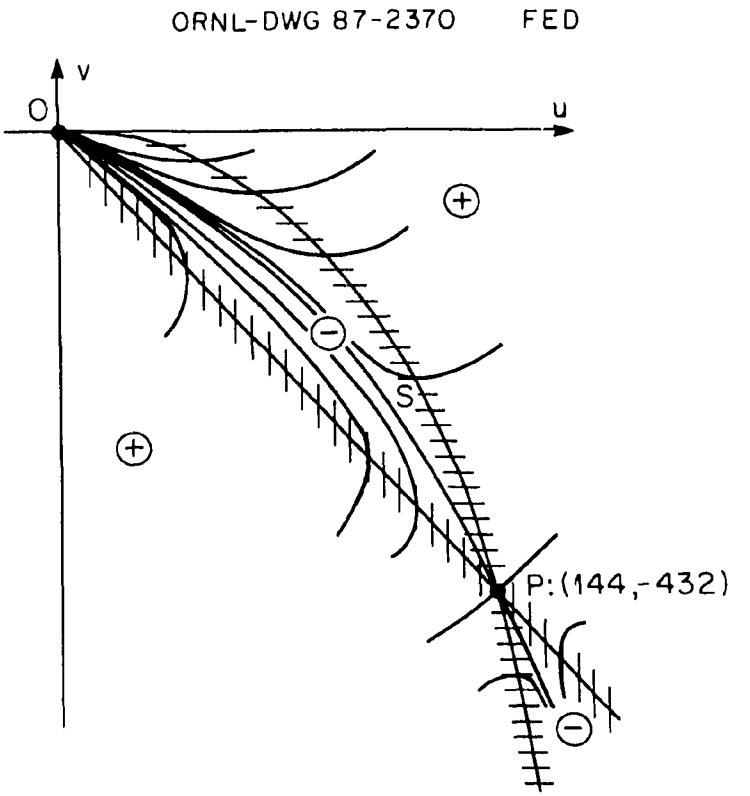

Fig. 1. The fourth quadrant of the direction field of Eq. (60c).

$v = -u^{3/2}/4$; the curve of infinite slope is $v = -3u$. These curves intersect at two singularities, the origin O and the point P: (144, −432).

The signs of the slope dv/du being as shown in the figure, the origin must be a node and the point P a saddle. One of the integral curves in Fig. 1 corresponds to the solution of Eqs. (57) and (58) that we seek. How shall we find out which one? In the first place, when $x = 0$, $y = 1$, so $u = x^3y = 0$. If $\dot{y}(0)$ is finite, then $v = x^4\dot{y} = 0$ as well. So the integral curve we seek passes through the origin O, which corresponds to $x = 0$. All the curves that emanate from the origin except the separatrix S eventually leave the fourth quadrant. So our attention is naturally focused on S.

The point P corresponds to the limit $x = \infty$. This we can see as follows. The slopes of the two separatrices through P can be calculated from Eq. (60c) using l'Hospital's rule; they are $(1 \pm \sqrt{73})/2$. So if near P we write $u = 144 + \Delta u$ and $v = -432 + \Delta v$, then $\Delta v/\Delta u = (1 - \sqrt{73})/2$ on S. Then, near P, Eq. (60b) becomes

$$\frac{dx}{x} = \frac{du}{3u + v} = \frac{2du}{(7 - \sqrt{73})\Delta u} = \frac{2du}{(7 - \sqrt{73})(u - 144)} \quad . \tag{61}$$

Thus, as $u \to 144$ from below, $x \to +\infty$. It follows, furthermore, from the definition of u that, when x is large,

$$y \sim \frac{u_p}{x^3} = \frac{144}{x^3} \quad . \tag{62}$$

So without yet having solved any differential equations we already have the asymptotic form of the solution we seek.

We can find additional useful information by studying the behavior of the separatrix S near the origin. Since the separatrix lies between the curves $v = -u^{3/2}/4$ and $v = -u$, it can approach the origin in one of three mutually exclusive ways, namely, (i) $-v \sim u$, (ii) $u \gg -v \gg u^{3/2}$, and (iii) $-v \sim u^{3/2}$. The first alternative means $v = au$ near O. Then Eq. (60c) reduces in leading order to $a = 4a/(a+3)$, so $a = 1$. This curve does not lie in the fourth quadrant and so cannot represent S. The third alternative means $v = au^{3/2}$, which converts Eq. (60c) into $3/2a = (4a+1)/3$ in leading order, so $a = 2$. Again, the curve does not lie in the fourth quadrant. The second alternative converts Eq. (60c) into $dv/du = 4v/3u$ in leading order, which implies $v = -au^{4/3}$, where a is some positive constant. So S and indeed all integral curves entering the origin through the fourth quadrant do so along curves of the form $v = -au^{4/3}$, different curves being labeled by different values of a.

The value of a for the separatrix has an interesting and useful interpretation. Since $v = -au^{4/3}$ near O, we have $x^4\dot{y} = -a(x^3y)^{4/3}$ near $x = 0$. The powers of x cancel, and since $y(0) = 1$, we thus have $\dot{y}(0) = -a$. We can find the value of a by numerically integrating along S from P to O, using the slope $(1 - \sqrt{73})/2$ to obtain starting values close to P. Once we have done so, we have starting values for the integration of Eq. (57) at $x = 0$. Thus, at the cost of a single numerical integration of a first-order differential equation, we have converted the two-point boundary value problem expressed by the conditions of Eq. (58) into an initial-value problem.

This program of calculation is not so easily carried out. To see why, note that if a is of the order of unity, $v = au^{4/3}$ will not be greater than $u^{3/2}$ until $u^{-1/6} \gg 1$. If we want the ratio $v/u^{3/2}$ to be, say, 1000, u will have to be smaller than 10^{-18}! So we will not be able to obtain a simply as $\lim_{u\to 0}(v/u^{4/3})$. We can circumvent this difficulty by constructing a power series for v that starts with the leading term $au^{4/3}$. A tedious calculation gives

$$v = az^8 + 2z^9 - \frac{4}{3}a^2z^{10} - \frac{34}{9}az^{11} + (2a^3 - 3)z^{12} + \frac{1102}{135}a^2z^{13} + \left(\frac{950a - 260a^4}{81}\right)z^{14} + \ldots \; ; \quad z = u^{1/6} \quad . \tag{63}$$

A numerical integration (fourth-order Runge-Kutta) from P toward O gives $v(10^{-6}) = -0.141663 \times 10^{-7}$. The value of a calculated from Eq. (63) is then $a = -1.58806$, within 3 parts in 10^4 of Baker's value of -1.588588.

Once we have the value of a, we can find starting values of y and $\dot{y}$ near the origin using the following power series given by Baker:

$$y = 1 - ax + x^3/3 - 2ax^4/5 + \ldots + x^{3/2}\left[\frac{4}{3} - 2ax/5 + 3a^2x^2/70 + \frac{4}{63}\left(\frac{2}{3} + \frac{a^3}{16}\right)x^3 + \ldots\right] \quad . \quad (64)$$

Shown in Fig. 2 are two sets of points calculated by forward integration of Eq. (57) (fourth-order Runge-Kutta) for $a = -1.588$ and $a = -1.588588$. The two sets of points coincide well for $x \lesssim 3$, but beyond $x = 3$, they diverge from one another. This is because a forward integration is equivalent to an integration in Fig. 1 along the separatrix in the direction $O \rightarrow P$. This is the unstable direction, and sooner or later a numerical calculation will be thrown off the separatrix to one side or the other. We could graphically join the points at small x to the asymptote $144/x^3$ with a curve like the solid one in Fig. 1. Such an interpolation gives a reasonable depiction of the solution, but not a highly accurate one because of the uncertainty of the graphical interpolation.

There is another way to calculate the curve of $y(x)$ without resorting to graphical interpolation. The procedure is this. First we find by numerical integration of Eq. (60c) some convenient point (u, v) on the separatrix S near P. From u and v we calculate values of y and $\dot{y}$ according to $y = u/x^3$, $\dot{y} = v/x^4$; the value of x we

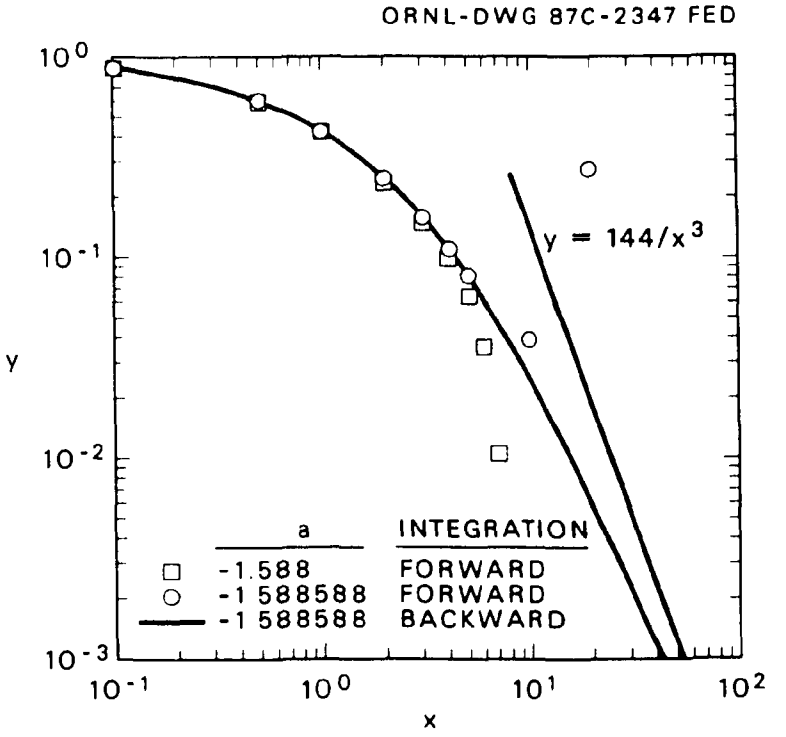

Fig. 2. The solution $y(x)$ of the Thomas-Fermi equation for which $y(0) = 1$.

choose arbitrarily. Then using $x, y, \dot{y}$ as initial data we integrate *backward* toward $x = 0$ (i.e., in the stable direction). In general, this integration will produce a $y(0) \neq 1$. Choose $\lambda = [y(0)]^{1/3}$ and calculate new starting values $x', y', \dot{y}'$ from the old starting values $x, y, \dot{y}$ according to $x' = \lambda x$, $y' = \lambda^{-3} y$, $\dot{y}' = \lambda^{-4}\dot{y}$. The primed starting values, when integrated backward, will lead to the value of $y'(0) = 1$ as required and hence define the solution curve we are seeking. Furthermore, since the backward direction of integration of Eq. (57) corresponds to motion along the separatrix in the direction $P \rightarrow O$, it is the stable direction of integration. The solid curve in Fig. 2 was produced in this way.

The reason that this works can be understood as follows. Suppose we denote the solution of Eq. (57) that obeys the boundary conditions of Eq. (58) by $y_*(x)$. If $y'(x')$ is any image of $y_*(x)$ under the transformation equation (59), then

$$y'(0) = \lambda^{-3} y_*(0) = \lambda^{-3} \quad , \tag{65a}$$

since $x = 0$ transforms into $x' = 0$, and

$$y'(x') \sim \lambda^{-3} \frac{144}{x^3} = \frac{144}{x'^3} \rightarrow 0 \quad \text{as} \quad x' \rightarrow \infty \quad . \tag{65b}$$

From this we can see at once that y_* and its one-parameter family of images look like Fig. 3 when plotted in the x, y plane.

As we have seen, when u and v are calculated from x, y_*, and $\dot{y}_*$, their locus in the u, v plane is the separatrix S. Any image point of x, y_*, and $\dot{y}_*$ will lead to the same values of u and v because u and v are invariants of the transformations. Hence y_* and its one-parameter family of images all map into the separatrix S.

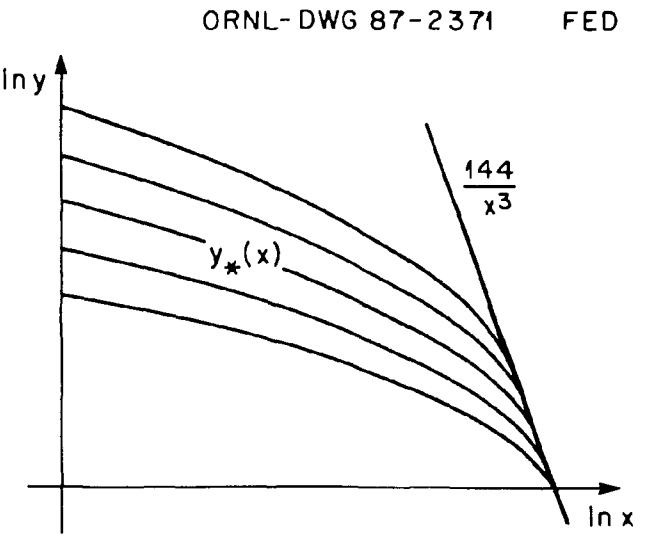

Fig. 3. The solution $y_*(x)$ for which $y_*(0) = 1$ and its one-parameter family of images.

The image of $x, y_*, \dot{y}_*$ is $\lambda x, \lambda^{-3} y_*, \lambda^{-4} \dot{y}_*$. So to any values u and v, *any* value of x can correspond, depending on the value of λ. If we know a point u, v on S and choose a value of x, we have implicitly chosen a value of λ, i.e., a particular curve of the family. Using the values of y and $\dot{y}$ corresponding to the chosen value of x, we can integrate backward to find $y(0)$. Then we determine λ using Eq. (65a). Having determined λ, we can scale the curve $y(x)$ that we just calculated to $y_*(x_*)$ according to $x = \lambda^{-1} x$, $y_* = \lambda^3 y$.

The reasons for the elaborate procedures just outlined are twofold, namely, that the boundary conditions (58) are two-point boundary conditions and that numerical integration of Eq. (57) in the forward x-direction is unstable. Because of these reasons, straightforward trial-and-error solution of Eqs. (57) and (58) is unrewarding, tedious, and inaccurate. The methods given here circumvent trial-and-error and are, moreover, capable of high accuracy.

2.7 Another second-order equation whose associated first-order equation cannot be solved in simple terms is van der Pol's equation,

$$\ddot{y} - \epsilon(1 - y^2)\dot{y} + y = 0 \ , \quad \epsilon > 0 \ . \tag{66}$$

This equation can be considered as the equation of harmonic motion ($\ddot{y} + y = 0$) with a term added which dampens the motion for large amplitudes and supports it for small motions. Because x, the independent variable, does not appear explicitly, Eq. (66) is invariant to the translation group

$$\begin{aligned} y' &= y \ , \\ x' &= x + \lambda \ . \end{aligned} \tag{67}$$

The dependent variable y is an invariant u of the group equation (67) and the derivative $\dot{y}$ is a first differential invariant v. (These simple choices are not the only ones possible: any function of y is an invariant, and any function of y and $\dot{y}$ is a first differential invariant!) Substituting $u = y$ and $v = \dot{y}$ in Eq. (66), we find the associated first-order equation

$$\frac{dv}{du} = \frac{\epsilon(1 - u^2)v - u}{v} \ . \tag{68}$$

Figure 4, the direction field of Eq. (68), shows the loci of zero and infinite slope. If we focus our attention on the region of the u-axis far to the right of the origin, we can see that there are two families of curves there, those that cross the u-axis and those that cross the locus of zero slope. These two families must be separated by a separatrix in the fourth quadrant, shown as curve S. A second separatrix S', the image of S under reflection in the origin, emerges in the portion of the second quadrant near the u-axis far to the left of the origin. Because these curves cannot cross, they both must wind inward as we traverse them in the clockwise direction.

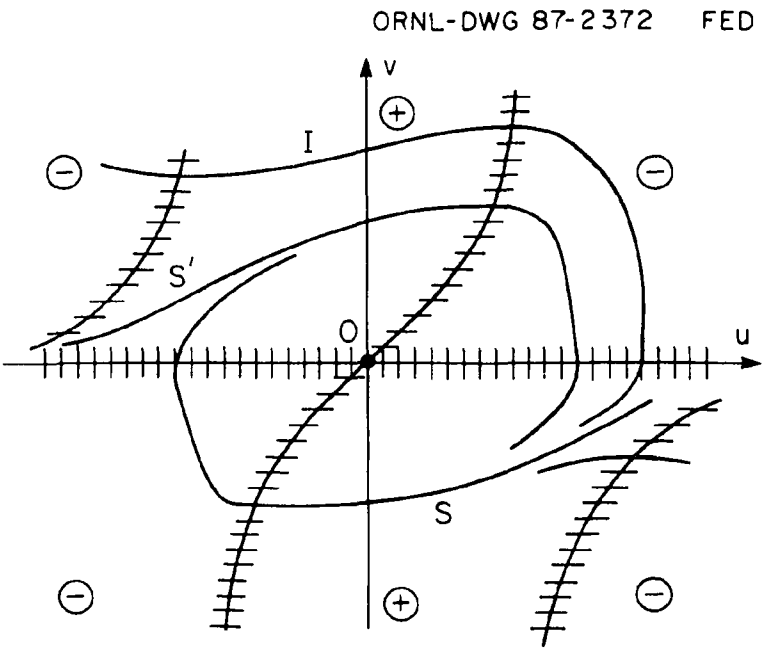

Fig. 4. Sketch of the direction field of Eq. (68), which is associated with van der Pol's equation (66).

The separatrices occur as a symmetrical pair because Eq. (68) is invariant to the single transformation $u' = -u$, $v' = -v$ that represents inversion in the origin. By extension, all integral curves occur in symmetrical pairs.

Shown in Fig. 4 is a typical integral curve I lying above S' in the second quadrant. As we proceed along it in the clockwise direction it, too, winds inward. Does it wind inward to the origin, or does it finally approach some limiting orbit that encircles the origin and closes upon itself? Such a closed trajectory, if it exists, would correspond to a periodic solution of Eq. (66). To see whether a closed trajectory exists, let us examine how the integral curves behave in the neighborhood of the origin. If they spiral out as we advance clockwise, there will have to be at least one closed orbit.

Near the origin $u^2 \ll 1$, so Eq. (68) becomes

$$\frac{dv}{du} = \frac{\epsilon v - u}{v} \quad . \tag{69}$$

Equation (69) is invariant to the group $v' = \lambda v$, $u' = \lambda u$, so we can integrate it explicitly; however, instead of plunging directly ahead, we employ an idea of Liénard's that will help us determine with only a little computational labor whether the spiral integral curves wind in or out. Write Eq. (69) as

$$d\left(\frac{v^2}{2} + \frac{u^2}{2}\right) = v\,dv + u\,du = \epsilon v\,du \quad . \tag{70}$$

Let us now integrate Eq. (70) clockwise over the upper half of the trajectory shown in Fig. 5:

$$\frac{1}{2}(u_2^2 - u_1^2) = \epsilon \int_1^2 v\, du > 0 \quad , \tag{71a}$$

since $v > 0$ on the upper half of the orbit. Thus $u_2^2 > u_1^2$ or $u_2 > |u_1|$. If we integrate Eq. (70) clockwise over the bottom half of the trajectory, we get

$$\frac{1}{2}(u_3^2 - u_2^2) = \epsilon \int_2^3 v\, du > 0 \quad , \tag{71b}$$

since $v < 0$, but we are integrating in the negative u-direction. Thus, $u_3^2 > u_2^2$ or $|u_3| > u_2$. Since $|u_3| > u_2 > |u_1|$, the integral curves near the origin must spiral outward in the clockwise direction.

By an elaboration of the above argument, Liénard proved not only that the van der Pol equation had closed trajectories, but also that there was exactly one such closed trajectory. Now since $u = y$ and $v = \dot{y}$, as x increases we traverse integral curves in the first and second quadrants ($v > 0$) in the direction of increasing u ($du = dy = \dot{y}\, dx = v\, dx > 0$). Similarly, we traverse integral curves in the third and fourth quadrants in the direction of decreasing u. Clearly, then, as x increases, we spiral clockwise around the origin in the (u, v) plane. This means that as x increases, we spiral toward the fixed trajectory in the (u, v) plane. This

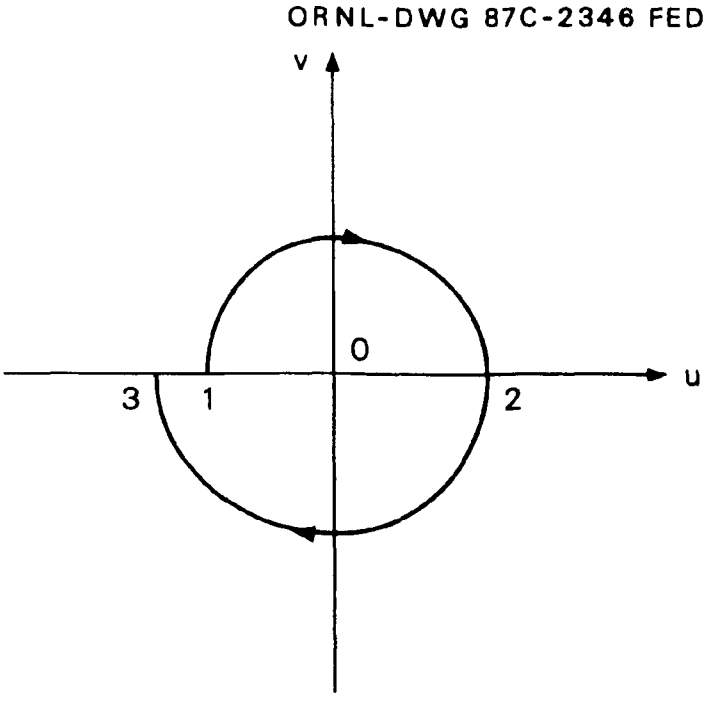

Fig. 5. Part of a spiral trajectory near the origin O.

closed trajectory represents a *stable, periodic* limit in the (x, y) plane to which every solution therefore tends as x increases. It is called a limit cycle. Shown in Fig. 6 is the solution of van der Pol's equation for $\epsilon = 5$, for which $y(0) = 0$ and $\dot{y}(0) = 0.01$. These initial conditions are quite distant from those that describe the limit cycle, for which $\dot{y} = 4.3752$ when $y = 0$. Nevertheless, the solution becomes virtually indistinguishable from the stable limit cycle after only one oscillation.

2.8 The use of Liénard's simple argument is not a conceit but in fact is probably the simplest and most straightforward way of determining whether the integral curves near the origin spiral inward or outward as we circulate clockwise. If we had plunged straight ahead instead of using Liénard's argument and solved Eq. (69) directly, we should have found, after some tedious calculation,

$$\epsilon < 2 : \frac{1}{2}\ln(v^2 - \epsilon uv + u^2) + \frac{\epsilon}{\sqrt{4-\epsilon^2}}\tan^{-1}\left(\frac{2v - \epsilon u}{u\sqrt{4-\epsilon^2}}\right) = \text{const} \quad , \tag{72a}$$

$$\epsilon > 2 : \frac{1}{2}\ln(v^2 - \epsilon uv + u^2) - \frac{\epsilon}{\sqrt{4-\epsilon^2}}\tanh^{-1}\left(\frac{2v - \epsilon u}{u\sqrt{\epsilon^2-4}}\right) = \text{const} \quad . \tag{72b}$$

These expressions are far from illuminating, and it is by no means clear at a glance that the integral curves they describe spiral outward in the clockwise direction.

A better alternative to solving Eq. (69) directly is based on the linearity of both the numerator and denominator of the right-hand side. Let us introduce a new parameter t by writing Eq. (69) as the coupled pair of linear equations

$$\frac{dv}{dt} = \epsilon v - u \quad , \tag{73a}$$

$$\frac{du}{dt} = v \quad . \tag{73b}$$

We can write the general solution of these as a sum of exponentials in t. If we set $v = Ae^{\lambda t}$ and $u = Be^{\lambda t}$, Eqs. (73a) and (73b) become

$$A\lambda = \epsilon A - B, \quad B\lambda = A \tag{74a}$$

or

$$\lambda^2 - \epsilon\lambda + 1 = 0 \quad . \tag{74b}$$

Thus,

$$\lambda = \frac{1}{2}(\epsilon + \sqrt{\epsilon^2 - 4}) \quad . \tag{74c}$$

When $\epsilon > 2$, the two roots given by Eq. (74c) are positive; when $\epsilon < 2$, the two roots are complex conjugates whose real part is positive. So in either case, as t grows larger (corresponding to clockwise circulation about the origin), u and v move away from the origin. This method of analysis can be used for any singular point at which the leading term in both numerator and denominator is linear.

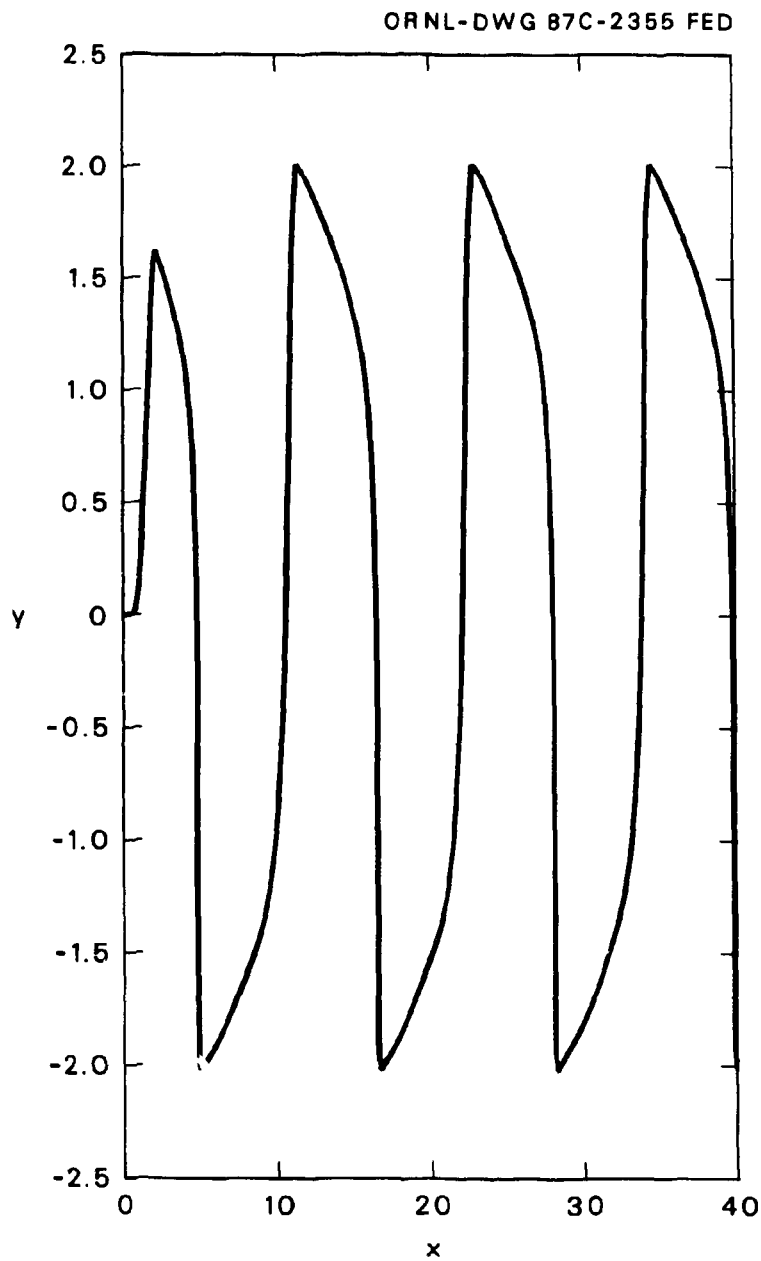

Fig. 6. Solution of van der Pol's equation for $\epsilon = 5$, for which $y(0) = 0$ and $\dot{y}(0) = 0.01$.

Chapter 3

SIMILARITY SOLUTIONS OF SECOND-ORDER PARTIAL DIFFERENTIAL EQUATIONS

"I have multiplied visions, and used similitudes."
—Hosea 12:10

3.1 The heart and soul of this chapter is based on an idea first proposed and exploited by Birkhoff, who considered partial differential equations with one dependent and two independent variables (for the sake of concreteness, call them c, z, and t, respectively). Many partial differential equations of physics and engineering are of this type: a good example for the reader to keep in mind for the moment is the ordinary diffusion equation $c_t = c_{zz}$. Quite often such partial differential equations are invariant to one or more one-parameter groups of transformations. For example, the diffusion equation is invariant to the affine group $c' = \lambda^\alpha c$, $z' = \lambda z$, $t' = \lambda^2 t$, where, owing to the linearity of the ordinary diffusion equation, α can be any fixed number.

When the partial differential equation is invariant to a group, every transformation of the group carries a solution into another solution. Among the very wide manifold of solutions usual for a partial differential equation there may be some that transform into themselves, i.e., are invariant to the group. The condition of group invariance restricts the form of such solutions. In the example we have been pursuing of the diffusion equation, solutions invariant to the affine group must have the form $c = t^{\alpha/2} y(z/t^{1/2})$, where y is an arbitrary function of the argument $x = z/t^{1/2}$. (We shall see presently why this is so.) Solutions invariant to affine groups are called similarity solutions.

Birkhoff realized that, because the unknown function y is a function of one variable only, when the invariant form is substituted into the diffusion equation, the result is an *ordinary* differential equation for y in terms of x. The calculation of this ordinary differential equation is instructive. If

$$c = t^{\alpha/2} y(z/t^{1/2}) \,, \tag{1a}$$

$$c_t = t^{\alpha/2-1}\left(\frac{\alpha y}{2} - \frac{z\dot{y}}{2t^{1/2}}\right) = t^{\alpha/2-1}\left(\frac{\alpha y}{2} - \frac{x\dot{y}}{2}\right) \,, \tag{1b}$$

$$c_z = t^{(\alpha-1)/2}\dot{y} \,, \tag{1c}$$

$$c_{zz} = t^{\alpha/2-1}\ddot{y} \,. \tag{1d}$$

Equating the right-hand side of Eq. (1b) and the right-hand side of Eq. (1d) we find, after cancelling the common factor $t^{\alpha/2-1}$, the second-order ordinary differential equation

$$\ddot{y} = \frac{\alpha}{2} y - \frac{1}{2} x\dot{y} \,. \tag{2}$$

Any solution of Eq. (2) will furnish a solution $c(z,t)$ of the diffusion equation through the connection equation (1a).

In obtaining the rightmost form in Eq. (1b), we have combined powers of z and t to obtain powers of x. Some power of t was left over, namely, $t^{\alpha/2-1}$. The same power appears on the right-hand side of Eq. (1d), so that it can be cancelled in obtaining Eq. (2). If we had chosen for the argument of the function y a combination of z and t other than $z/t^{1/2}$, e.g., z/t, this would not have been true. When we had eliminated all explicit appearance of z from the ordinary differential equation, there would not have remained a cancellable common power of t. (Try it!) So the group invariance helps us to find the right combination of z and t to use as the argument of y.

3.2 Different values of the constant α distinguish different solutions of the partial differential equation. Now, different solutions of a partial differential equation satisfy different boundary and initial conditions, so we may expect that α is somehow determined by the boundary and initial conditions. To simplify discussion of the boundary and initial conditions let us use the language of heat diffusion, so that the dependent variable c can be called temperature and its negative derivative $-c_z$ can be called heat flux (or just flux).

Consider now what I call the problem of clamped temperature in a half-space. Imagine the half-space $z > 0$ initially held at zero temperature to have its front face ($z = 0$) suddenly raised to unit temperature, e.g., by being brought into contact with a heat bath. How does the temperature rise in the half-space as a function of time? The mathematical representation of the boundary and initial conditions of this problem is

$$c(z,0) = 0 \quad (z > 0)\ , \tag{3a}$$

$$c(0,t) = 1 \quad (t > 0)\ , \tag{3b}$$

$$c(\infty,t) = 0 \quad (t > 0)\ , \tag{3c}$$

where Eq. (3c) expresses the implied condition far from the heated boundary. Let us rewrite these conditions using the invariant form of Eq. (1a). It is convenient to start with Eq. (3b); the reader will see why in a moment. According to Eq. (3b),

$$1 = t^{\alpha/2} y(0)\ , \tag{4a}$$

which can only be satisfied if

$$\alpha = 0 \quad \text{and} \quad y(0) = 1\ . \tag{4b}$$

If α had any other value than zero, the right-hand side of Eq. (4a) could not be held constant as the time t changed. When $\alpha = 0$, Eq. (1a) takes the form $c = y(z/t^{1/2})$; then Eq. (3a) and Eq. (3c) *both* become

$$y(\infty) = 0\ . \tag{4c}$$

Thus the *three* boundary and initial conditions, Eqs. (3a)–(3c), for the partial differential equation collapse to *two* boundary conditions, Eqs. (4b) and (4c), for the ordinary differential equation (2). This collapse of the boundary conditions from

three to two is essential to the success of the method of similarity solutions because in general three conditions overdetermine the solution of a second-order ordinary differential equation.

Since $\alpha = 0$ for the clamped-temperature problem, Eq. (2) takes the form $\ddot{y} = -(1/2)x\dot{y}$, which can be integrated at once to give $\dot{y} = -C\exp(-x^2/4)$, where C is a (positive) constant of integration. A second integration gives $y = C\int_x^\infty \exp(-u^2/4)\,du$, which already obeys the boundary conditions of Eq. (4c). To satisfy Eq. (4c), C must equal $1/\sqrt{\pi}$. Then $y = \mathrm{erfc}(x/2)$, where erfc is the complementary error function. Rewritten in terms of c, this solution takes the well-known form

$$c = \mathrm{erfc}\left(\frac{z}{2\sqrt{t}}\right) . \tag{5}$$

3.3 Since Eq. (5) is invariant to the affine group $c' = c$, $z' = \lambda z$, $t' = \lambda^2 t$ (remember $\alpha = 0$ for the clamped-temperature problem!), so must be the boundary and initial conditions, Eqs. (3a)–(3c), that determine it. If they were not, then the boundary and initial conditions in the primed variables would be different from those in the unprimed variables. These different sets of boundary and initial conditions would therefore determine different solutions of the partial differential equation, which, being images of one another, could not be their own images, i.e., could not be invariant.

When Eqs. (3a)–(3c) are written in terms of the primed variables they become

$$c'\left(\frac{z'}{\lambda},\ 0\right) = 0 , \tag{6a}$$

$$c'\left(0,\ \frac{t'}{\lambda^2}\right) = 1 , \tag{6b}$$

$$c'\left(\infty,\ \frac{t'}{\lambda^2}\right) = 0 , \tag{6c}$$

since when $z = 0$, $z' = 0$, when $z = \infty$, $z' = \infty$ and when $t = 0$, $t' = 0$. Because $z = z'/\lambda$ and $t = t'/\lambda^2$ can have any value, Eqs. (6a) and (6b) can only be satisfied if

$$c'(z', 0) = 0 , \tag{7a}$$

$$c'(0, t') = 1 , \tag{7b}$$

$$c'(\infty,\ t') = 0 , \tag{7c}$$

for all z' and t'. Equations (7a)–(7c) are the same as Eqs. (3a)–(3c) except that they refer to the primed variables.

If the clamped-temperature problem referred not to the half-space $z > 0$ but to the finite slab $0 < z < L$, the boundary condition (3c) would have to be replaced by the condition $c(L, t) = 0$, which, upon transformation to the primed variables, becomes $c'(L/\lambda,\ t'/\lambda^2) = 0$. Since $t = t'/\lambda^2$ can have any value, this boundary condition is equivalent to $c'(L/\lambda\,,\ t') = 0$ for all t'. Now this is not the same as

$c'(L,\ t') = 0$ because in general $\lambda \neq 1$. This means that the clamped-temperature problem in a finite slab cannot be solved in terms of a similarity solution of the partial differential equation but instead requires a solution of a more complicated kind.

At an early epoch, however, when $t \ll L^2/4$, the diffusing heat is not yet affected by the presence of the cold boundary at $z = L$. The temperature distribution does not yet "know" about the cold boundary and "thinks" the heat is diffusing in a semi-infinite half-space. So for short times, at least, the similarity solution gives a good approximation to the temperature distribution. For long times, $t \gg L^2/4$, the steady-state solution $c = 1 - z/L$ is a good approximation to the temperature distribution. Knowing these two limiting temperature distributions often enables us to estimate quantities of interest. Suppose, for example, we wanted to know the heat flux $-c_z(0,t)$ through the slab as a function of time. Then

$$-c_z(0,t) = 1/\sqrt{\pi t}\ , \quad t \ll L^2/4\ , \quad \text{similarity solution,} \tag{8a}$$

$$= \frac{1}{L}\ , \qquad t \gg L^2/4\ , \quad \text{steady-state solution.} \tag{8b}$$

A simple graphical interpolation between these limits may well provide a sufficient estimate for practical purposes.

3.4 Next we consider what I call the pulsed-source problem in an infinite medium. At $t = 0$, an amount of heat Q per unit area is instantaneously introduced in the plane $z = 0$ and subsequently spreads out toward $z = \pm\infty$ by diffusion. The boundary and initial conditions for this problem are

$$c(z,0) = 0\ , \tag{9a}$$

$$\int_{-\infty}^{+\infty} c(z,t)\ dz = Q\ , \tag{9b}$$

$$c(\pm\infty,t) = 0\ . \tag{9c}$$

Equation (9b) expresses conservation of the heat injected by the initial pulse. If we substitute Eq. (1a) into Eq. (9b), the latter becomes

$$Q = \int_{-\infty}^{+\infty} t^{\alpha/2} y(z/t^{1/2})\ dz = t^{(\alpha+1)/2} \int_{-\infty}^{+\infty} y(x)\ dx\ , \tag{10}$$

where, as before, $x = z/t^{1/2}$. The integral on the right-hand side of Eq. (10) is a pure number, so for Eq. (10) to be satisfied for all t, α must equal -1. Then

$$c(z,t) = t^{-1/2} y(x)\ , \quad x = z/t^{1/2}\ . \tag{11a}$$

Since $y(x)$ must be symmetric, i.e., since $y(x) = y(-x)$, the boundary and initial conditions, Eqs. (9a)-(9c), collapse to

$$\int_0^{\infty} y(x)\ dx = Q/2 \tag{11b}$$

and

$$y(\infty) = 0 \ . \tag{11c}$$

[The boundary condition (11c) is sufficient as it stands to satisfy Eq. (9c). To satisfy Eq. (9a), y must approach zero sufficiently rapidly as x approaches infinity so that $\lim_{t\to 0} t^{-1/2} y(z/t^{1/2}) = 0$. Whether this requirement is fulfilled can only be tested a posteriori once we have solved for $y(x)$. If the requirement is met, the similarity solution is the solution to the stated problem. If not, the solution to the stated problem is not a similarity solution but a solution of some other kind.]

When $\alpha = -1$, the ordinary differential equation (2) is again easily solvable, for its right-hand side is just the perfect differential $-d/dx[(1/2)xy]$. So, integrating once, we get

$$\dot{y} = -\frac{1}{2} xy \ . \tag{12}$$

The constant of integration vanishes since, by symmetry, $\dot{y}(0) = 0$. Integrating again, we find

$$y = C \exp(-x^2/4) \ , \tag{13a}$$

which obeys Eq. (11c). From Eq. (11b) it follows that $C = Q/\sqrt{4\pi}$, so that

$$C = Q \frac{\exp(-z^2/4t)}{(4\pi t)^{1/2}} \ , \tag{14}$$

another well-known solution. [Now we can verify that the initial condition, Eq. (9a), is satisfied, i.e., that $\lim_{t\to 0} c(z,t) = 0$, because the exponential term overpowers the factor $t^{-1/2}$.]

It is perhaps worthwhile to note that Eqs. (5) and (14) are solutions of different ordinary differential equations because they satisfy different versions of the generalized ordinary differential equation (2) corresponding to different values of α.

3.5 What we have done so far has been based on the form of Eq. (1a) for an invariant solution. We can certainly see at once that Eq. (1a) is invariant to the affine group $c' = \lambda^\alpha c$, $t' = \lambda^2 t$, and $z' = \lambda z$, and everything we have done so far could have been based on looking for special solutions of the form of Eq. (1a). As it happens, Eq. (1a) is the most general form for a relation among c, z, and t that is invariant to the affine group. We can prove this easily by methods introduced in Chap. 2, and we can generalize at no extra cost of labor to the affine group

$$c' = \lambda^\alpha c \ , \tag{15a}$$

$$t' = \lambda^\beta t \ , \tag{15b}$$

$$z' = \lambda z \ , \tag{15c}$$

where the exponents α and β are particular prescribed constants. [Note that no generality is lost by taking the exponents of λ in Eq. (15c) equal to 1.]

A relation $c = f(z,t)$ among c, z, and t can be visualized as a surface S in three-dimensional space. If this surface is to be invariant to the group equation (15), the image $(c',\ z',\ t')$ of any point (c,z,t) on S must also lie on S. This means $c' = f(z',t')$ or, what is the same thing, $\lambda^\alpha c = f(\lambda z, \lambda^\beta t)$. If we differentiate this last equation with respect to λ and set $\lambda = 1$ (the value of λ for the identity transformation), we get the first-order linear partial differential equation

$$\alpha f = z f_z + \beta t f_t \ , \tag{16}$$

whose characteristic equations are

$$\frac{df}{\alpha f} = \frac{dz}{z} = \frac{dt}{\beta t} \ . \tag{17}$$

Two integrals of Eq. (17) are $z/t^{1/\beta}$ and $f/t^{\alpha/\beta}$. The most general solution of Eq. (16) is obtained by equating one of these integrals to an arbitrary function y of the other:

$$\frac{c}{t^{\alpha/\beta}} = \frac{f}{t^{\alpha/\beta}} = y\left(\frac{z}{t^{1/\beta}}\right) \ . \tag{18}$$

When $\beta = 2$, this form reduces to Eq. (1a).

3.6 So far we have applied Birkhoff's idea of seeking invariant solutions to the *linear* diffusion equation, for which there are excellent alternative methods of solution based on the principle of superposition, e.g., Fourier series and Laplace transformation. Now let us turn our attention to a nonlinear diffusion equation for which Birkhoff's method seems to me to be the only one available.

At low temperatures, the thermal conductivities of metals (e.g., copper or aluminum) are directly proportional to temperature, and their specific heats are proportional to the cube of the temperature. So the ordinary one-dimensional heat diffusion equation for such a material becomes

$$ST^3 \frac{\partial T}{\partial t} = \frac{\partial}{\partial z}\left(kT \frac{\partial T}{\partial z}\right) \ , \tag{19}$$

where S is a constant having the dimensions of $\mathrm{J{\cdot}m^{-3}{\cdot}K^{-4}}$ and k is a constant having the dimensions of $\mathrm{W{\cdot}m^{-1}{\cdot}K^{-2}}$. ($ST^3$ is the heat capacity per unit volume and kT is the thermal conductivity.) If we set $c = T^2$, then Eq. (19) becomes

$$c \frac{\partial c}{\partial t} = \frac{k}{S} \frac{\partial^2 c}{\partial z^2} \ . \tag{20}$$

Suppose now we consider what I call the clamped-flux problem in a semi-infinite half-space. At $t = 0$, a heater covering the front face $z = 0$ of the cold half-space $z > 0$ is suddenly energized and begins producing a steady heat flux q (dimensions: $\mathrm{W{\cdot}m^{-2}}$) into the half-space. How does the temperature in the half-space rise as

a function of time? The boundary and initial conditions corresponding to this problem are

$$\left.\begin{aligned} &T(z,0) = 0 \\ &-(kT)T_z|_{z=0} = q \\ &T(\infty,t) = 0 \end{aligned}\right\} \qquad \text{or} \qquad \begin{cases} c(z,0) = 0 & \text{(21a)} \\ -kc_z(0,t) = 2q & \text{(21b)} \\ c(\infty,t) = 0 & \text{(21c)} \end{cases}$$

if we assume that the half-space is initially at zero temperature. We can eliminate dimensional quantities k, S, and q by choosing to work in a special system of units in which the constants k, S, and $2q$ all have the numerical value 1. Then Eqs. (20) and (21) become

$$cc_t = c_{zz} \tag{22a}$$

and

$$c(z,0) = 0 \ , \tag{22b}$$

$$c_z(0,t) = -1 \ , \tag{22c}$$

$$c(\infty,t) = 0 \ . \tag{22d}$$

Now we test Eq. (22a) for invariance to the affine group (15): a short computation shows that it will be invariant only if the constants α and β obey the linear constraint

$$\alpha - \beta = -2 \ . \tag{23}$$

[The easiest way to see this is to imagine Eq. (22a) written in the primed form and then replace the primed variables by their equivalents expressed in terms of the unprimed variables according to Eq. (15). Then we get Eq. (22a) in the unprimed variables with the left-hand side multiplied by the factor $\lambda^{2\alpha-\beta}$ and the right-hand side multiplied by the factor $\lambda^{\alpha-2}$. If these two factors are equal, they may be cancelled. Then Eq. (22a) in the primed form implies Eq. (22a) in the unprimed form, and Eq. (22a) is invariant to Eq. (15). Thus, the exponents of λ in the two factors must be equal, from which Eq. (23) follows at once.]

The boundary condition, Eq. (22c), will be invariant if and only if $\alpha - 1 = 0$, i.e., $\alpha = 1$. Then from Eq. (23), it follows that $\beta = 3$. So according to Eq. (18) we should take the form

$$c = t^{1/3} y(z/t^{1/3}) \tag{24}$$

for the invariant solution of Eqs. (22a)–(22d) that we seek. Differentiating Eq. (24), we obtain

$$c_t = t^{-2/3}\left(\frac{y}{3} - \frac{1}{3}x\dot{y}\right) \ , \tag{25a}$$

$$c_z = \dot{y} \ , \tag{25b}$$

$$c_{zz} = t^{-1/3}\ddot{y} \ , \tag{25c}$$

so that Eq. (22a) becomes, after some slight rearrangement,

$$3\ddot{y} + xy\dot{y} - y^2 = 0 \ . \tag{26}$$

The boundary and initial conditions, Eqs. (22b) and (22c), collapse to the two conditions

$$\dot{y}(0) = -1 \ , \tag{27a}$$

$$y(\infty) = 0 \ . \tag{27b}$$

Equations (26), (27a), and (27b) together make up a two-point boundary value problem. Since Eq. (26) is not solvable in terms of tabulated functions, we shall have to solve it numerically.* In order to start the numerical solution of a second-order ordinary differential equation we need two initial conditions, a value and a slope. We, therefore, have to guess the value at the origin, integrate forward, and test whether $y(\infty) = 0$. As it turns out, if we guess $y(0)$ too high, the curve $y(x)$ we get has a positive minimum and thereafter approaches ∞ asymptotically with a constant slope. As we lower $y(0)$ the minimum moves down and to the right. If we guess $y(0)$ too low, the $y(x)$ we obtain plunges toward $-\infty$ at some finite value of x. As we raise $y(0)$, this singularity moves to the right. (The reader is urged to try out some numerical integrations if he can.) It is possible, then, to improve our guesses of $y(0)$. But the trial-and-error process outlined here is very laborious and converges rather slowly. Moreover, it is inelegant, although that may not really matter.

There is a less laborious and much more elegant way of dealing with this two-point boundary value problem based on the invariance of the ordinary differential equation (26) to the affine group

$$\begin{aligned} y' &= \mu^{-2} y \ , \\ x' &= \mu x \ , \end{aligned} \qquad 0 < \mu < \infty \ . \tag{28}$$

[For the moment, the existence of this group seems to be a piece of luck. Later we shall see that the invariance of Eq. (26) to Eq. (28) could have been foretold from the invariance of the partial differential equation (22a) to the one-parameter family of groups given by Eqs. (15a)–(15c) and (23).] If we introduce the invariant $u = x^2 y$ and the first differential invariant $v = x^3 \dot{y}$ as new variables, the second-order ordinary differential equation (26) reduces to the first-order ordinary differential equation

$$\frac{dv}{du} = \frac{9v - uv + u^2}{3(2u + v)} \ . \tag{29}$$

Now we examine the direction field of Eq. (29). Since we expect $\dot{y}$ to be negative and y to be positive, we expect $u > 0$ and $v < 0$. Thus, we want only the fourth quadrant of the direction field. Figure 1 shows a sketch of this quadrant. The curve of zero slope is $v = u^2/(u - 9)$; the curve of infinite slope is $v = -2u$. These two curves intersect in two singularities, the origin O and the point P : (6,−12). The signs of the slope dv/du being as shown, the origin must be a node and the point P a saddle. The direction field is quantitatively the same as that shown in Fig. 2.1,* and the procedure we follow is similar to that which we followed for the Thomas-Fermi equation.

*That is, Fig. 1 of Chap. 2.

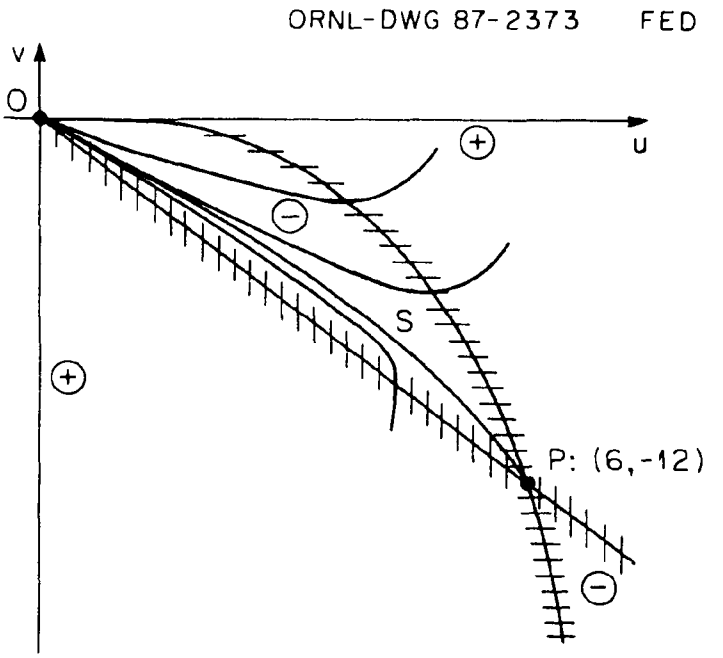

Fig. 1. A sketch of the fourth quadrant of the direction field of Eq. (29).

When $x = 0$, both u and v are zero, so the origin O in the (u,v) plane corresponds to the initial value $x = 0$. As in the case of the Thomas-Fermi equation, the point P corresponds to $x = \infty$. Also as before, two separatrices pass through P, one having the positive slope $(\sqrt{33} - 1)/2$, the other the negative slope $-(\sqrt{33} + 1)/2$. The one with the negative slope, S, passes through the origin and is the integral curve of Eq. (29) that we want. In a manner similar to that which we used to obtain Eq. (2.61) we now find that, near P, $dx/x = 2\ du/(\sqrt{33} - 3)(u_P - u)$, so that $x \to \infty$ as $u \to u_P$ along S. When x is large and u is very near $u_P = 6$, $y = u/x^2 \sim u_P/x^2 = 6/x^2$, which fulfills boundary condition (27b).

How does the separatrix S behave near the origin O? Since it lies between the locus of zero slope and the locus of infinite slope, $2u \geq |v| \geq u^2/9$ near the origin. Now, close to the origin, $|u|$ and $|v|$ are $\ll 1$, so Eq. (29) becomes

$$\frac{dv}{du} = \frac{9v + u^2}{3(2u + v)} \tag{30}$$

because $9|v| \gg |uv|$. (Note that we cannot say that the first-order term $9v$ greatly exceeds the quadratic term u^2 because we do not know the relative magnitudes of u and v.) Three possibilities exist: $|v| \sim u$, $u \gg |v| \gg u^2$, and $|v| \sim u^2$. The first leads to $v = u$, which does not lie in the fourth quadrant. The second leads to $v = Cu^{3/2}$, where C is a constant of integration. The third leads to $v = u^2/3$, which also does not lie in the fourth quadrant. Only the second alternative yields an allowable result; we expect $C < 0$.

If we substitute for u and v their definitions in terms of x and y, the relation $v = Cu^{3/2}$ becomes

$$C = \dot{y}(0)/y^{3/2}(0) \tag{31}$$

if we remember that the point $x = 0$ corresponds to the origin $u = 0$, $v = 0$ in the (u,v) plane. To find C we can integrate Eq. (29) numerically from P to O. To start, we step away from P using the slope $-(\sqrt{33}+1)/2$ obtained from Eq. (29) with l'Hospital's rule. Then we integrate toward O, decreasing the interval of integration as we approach O, until the ratio $v/u^{3/2}$ becomes constant to the desired number of figures. This procedure, which requires one integration only, gives $C = -0.5383$ to four significant figures. Armed with this value of C, we can find the hitherto unknown value of $y(0)$ corresponding to the slope $\dot{y}(0) = -1$, namely, $y(0) = 1.511$.

Figure 2 shows a curve obtained by forward integration of Eq. (26) with the initial conditions $y(0) = 1.511$, $\dot{y}(0) = -1$ [curve (a)]. As we might have expected from the divergence of the integral curves in Fig. 1 near the saddle point P, forward integration ($0 \rightarrow P$ in Fig. 1) is unstable. That is the reason that beyond about $x = 5$, curve (a) progressively diverges more and more from the asymptotic limit $6/x^2$ that it should approach. For practical purposes it may be satisfactory to join the points of the numerically calculated curve for $x < 5$ graphically to the asymptote $6/x^2$. If higher accuracy is desired, we can calculate $y(x)$ numerically by backward integration as described in the last paragraphs of Sect. 2.6 [curve (b)].

According to Eq. (24), $c(0,t) = y(0)t^{1/3}$ so that $T(0,t) = [y(0)]^{1/2}t^{1/6} = 1.229\ t^{1/6}$. This formula is written in the system of special units. To convert it

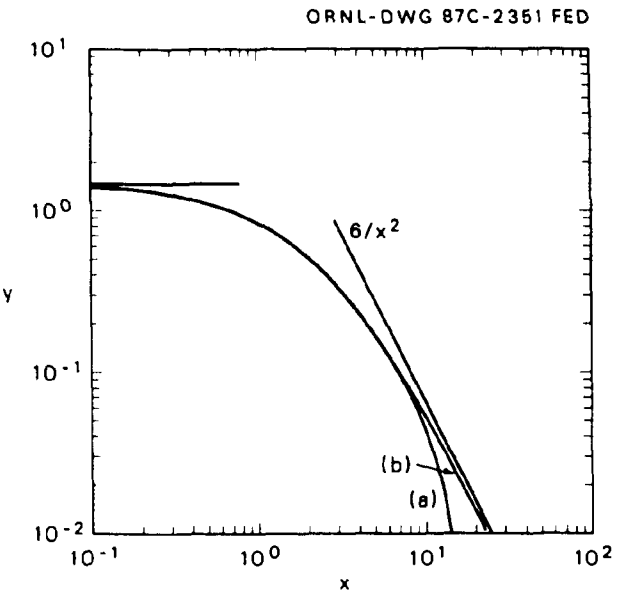

Fig. 2. Solution of the ordinary differential equation (26) and the boundary condition (27). Curve (a) was obtained by a forward integration that eventually becomes unstable. Curve (b) was obtained by a backward integration that is always stable.

into a form that is correct in any set of units, we make it dimensionally homogeneous by multiplying with suitable powers of k, S, and $2q$. Because the latter quantities are all numerically equal to 1 in special units, multiplying the terms of an equation by powers of them changes nothing. Once the equation is dimensionally homogeneous, it is then correct in any set of units. Thus $T(0,t) = 1.548(q^2t/kS)^{1/6}$, which gives the temperature at the front face of the half-space.

3.7 The turning point in Sect. 3.6 was the recognition of the invariance of the ordinary differential equation (26) to the affine group (28). The existence of such an associated affine group for the ordinary differential equation is a consequence of the partial differential equation's invariance to a *one-parameter family* of affine groups of the type

$$c' = \lambda^{\alpha} c \ , \tag{32a}$$

$$t' = \lambda^{\beta} t \ , \tag{32b}$$

$$z' = \lambda z \ , \tag{32c}$$

where α and β fulfill the linear constraint

$$M\alpha + N\beta = L \tag{32d}$$

and M, N, and L are fixed coefficients determined by the structure of the partial differential equation [cf. Eq. (23)]. The parameter λ labels the individual transformations of a group; the parameter α labels the groups of the family. If the partial differential equation is invariant to such a one-parameter (α) family of one-parameter (λ) groups, the ordinary differential equation that gives its similarity solutions is invariant to the associated affine group,

$$y' = \mu^{L/M} y \ , \tag{33a}$$

$$x' = \mu x \tag{33b}$$

(here μ is the group parameter of the associated group).

To see why this is so, we begin by noting that functions $c(z,t)$ invariant to a group of the family of Eq. (32a), say the group corresponding to the parameters α_0, β_0, must have the form of Eq. (18), namely,

$$c = t^{\alpha_0/\beta_0} y\left(\frac{z}{t^{1/\beta_0}}\right) = t^{\alpha_0/\beta_0} y(x) \ , \quad x \equiv \frac{z}{t^{1/\beta_0}} \ . \tag{34}$$

The parameters α_0, β_0, which obey the constraint (32d), are determined by the boundary and initial conditions that specify the particular problem we are dealing with.

If we transform Eq. (34) (imagined written in the primed form) by Eqs. (32a)–(32c) with $\alpha = \alpha_0$ and $\beta = \beta_0$, we recover Eq. (34) itself in the unprimed form. What happens if we transform Eq. (34), written in the primed form, by a group of the family for which $\alpha \neq \alpha_0$, $\beta \neq \beta_0$? We shall certainly get another solution of the

partial differential equation, for the image of any solution is another solution. This new solution is given by

$$c = \lambda^{(\alpha_0\beta-\alpha\beta_0)/\beta_0} t^{\alpha_0/\beta_0} y\left(\lambda^{1-\beta/\beta_0}\frac{z}{t^{1/\beta_0}}\right) \tag{35a}$$

$$= \mu^{(\alpha_0\beta-\alpha\beta_0)/(\beta_0-\beta)} t^{\alpha_0/\beta_0} y(\mu x) \ , \tag{35b}$$

where $\mu = \lambda^{1-\beta/\beta_0}$. Because the pairs α, β and α_0, β_0 separately obey the linear constraint (32d), it follows that

$$\frac{\alpha_0\beta - \alpha\beta_0}{\beta - \beta_0} = \frac{L}{M} \ , \tag{36}$$

so that the new solution of the partial differential equation is given by

$$c = t^{\alpha_0/\beta_0}\mu^{-L/M} y(\mu x) \ . \tag{37}$$

Equation (37) has the same form as Eq. (34), i.e., t^{α_0/β_0} times a function of $x = z/t^{1/\beta_0}$, which means that $y(x)$ and $\mu^{-L/M}y(\mu x)$ must satisfy the same ordinary differential equation. Now the one-parameter family of functions $\mu^{-L/M}y(\mu x)$, $0 < \mu < \infty$, is the same as the one-parameter family of images of $y(x)$ under the group of transformations

$$y' = \eta^{L/M} y \ , \tag{38a}$$

$$x' = \eta x \ , \qquad 0 < \eta < \infty \ . \tag{38b}$$

In fact, the function $\mu^{-L/M}y(\mu x)$ is the image of $y(x)$ for the transformation of the group (38) for which $\eta = \mu^{-1}$. Seeing this last assertion has troubled some of my students, so I give below two proofs of it, a short one and a long one; the long one has the virtue (I hope) of complete transparency.

The short proof is embodied in the line of equalities

$$y'(x') = \eta^{L/M} y(x) = \eta^{L/M} y\left(\frac{x'}{\eta}\right) = \mu^{-L/M} y(\mu x') \ . \tag{39}$$

The first equality comes from Eq. (38a), which says that the value of y' at the image point x' is $\eta^{L/M}$ times the value of y at the source point x. The second equality follows from Eq. (38b). The third equality follows from taking $\eta = \mu^{-1}$. The interpretation of Eq. (39) is this: the image function $y'(\ldots)$ is the same as the function $\mu^{-L/M}y(\mu\ldots)$, where the three dots signify the place at which the argument (the same for both functions) must be inserted.

The longer proof makes use of the three diagrams shown in Figs. 3(a)–3(c). Figure 3(a) shows curve $y = f(x)$, which will be transformed in Fig. 3(b) to a new curve $y = \mu^{-a}f(\mu x)$ and in Fig. 3(c) into the image under Eqs. (38a) and (38b) for which $\eta = \mu^{-1}$ ($a = L/M$). Shown again for reference in Fig. 3(b) is the curve $y = f(x)$. Let us choose an abscissa x and calculate graphically the value of $y = \mu^{-a}f(\mu x)$. Suppose the abscissa x lies at point A. Then μx would be at point B. The height of point C gives the magnitude of $f(\mu x)$ and the height of point D the

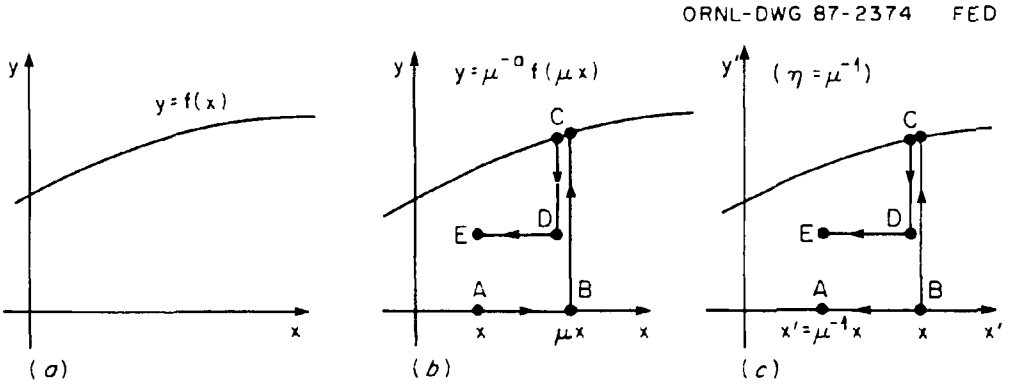

Fig. 3. Auxiliary sketches for use in the proof that $\mu^{-L/M}y(\mu x)$ is an image of $y(x)$ under a transformation of the group (38a) and (38b).

magnitude of $\mu^{-a}f(\mu x)$. When this last height is plotted over the abscissa A, we have a point E belonging to the curve $y = \mu^{-a}f(\mu x)$.

This point also lies on the image curve of $y = f(x)$ under Eqs. (38a) and (38b) with $\eta = \mu^{-1}$. This time we start with the abscissa x lying at point B. Then point C gives $f(x)$ and point D gives $y' = \eta^a f(x) = \mu^{-a}f(x)$. The abscissa $x' = \eta x = \mu^{-1}x$ must then be at point A. The point $(x',\ y')$ thus lies at point E. By this construction we see that any point on one curve lies on the other, and conversely. So the two curves are the same, which is what we wanted to prove.

What we have proved so far is that every image under the associated group, Eqs. (38a) and (38b), of a solution of the ordinary differential equation for $y(x)$ is also a solution. So the total manifold of solutions of this ordinary differential equation must be carried into itself by the transformations of this group, that is, must be invariant to this group. Now since a differential equation and its manifold of solutions are logically identical, the differential equation itself must be invariant to the associated group (38a) and (38b).

My earlier book, *Similarity Solutions of Nonlinear Partial Differential Equations* (Research Notes in Mathematics 88, Pitman Advanced Publishing Program, Pitman Publishing Inc., 1020 Plain Street, Marshfield, Massachusetts 02050), is devoted to the exploitation of the invariance of the ordinary differential equation for $y(x)$ to the associated group (38a) and (38b). Among the partial differential equations treated there are (1) $C_t = (C^n C_z)_z$, which occurs in soil mechanics and boundary-layer flow; (2) $C_t = (C_z^{1/3})_z$, which occurs in the theory of counterflow heat transport in superfluid helium; and (3) $C_{tt} = C_{zz}\int_0^1 C_z^2 dz$, which occurs in the theory of motion of a shock-loaded membrane. Since that book is an ample reference for the interested reader, I close this chapter here.

Chapter 4
MAXIMUM PRINCIPLES AND DIFFERENTIAL INEQUALITIES

"To compare great things with small."
—John Milton
Paradise Lost

4.1 In the preface to their book on maximum principles, Protter and Weinberger introduce the subject with the following words: "[A maximum] principle is a generalization of the elementary fact of calculus that any function $f(x)$ which satisfies the inequality $f'' > 0$ on an interval $[a,b]$ achieves its maximum value at one of the endpoints of the interval. We say that solutions of the inequality $f'' > 0$ satisfy a *maximum principle*. More generally, functions which satisfy a differential inequality in a domain D and, because of it, achieve their maxima on the boundary of D are said to possess a maximum principle."

The chief use of maximum principles is to provide bounds for solutions of differential equations. We begin our discussion with the linear homogeneous, second-order ordinary differential equation

$$\ddot{y} + g(x)\dot{y} + h(x)y = 0 \quad ; \quad h(x) < 0 \; . \tag{1}$$

Can the function $y(x)$ have a positive maximum on any interval $[a,b]$? At a positive maximum, $y > 0$, $\dot{y} = 0$, and $\ddot{y} < 0$. These conditions are inconsistent with Eq. (1), for then the first and third terms will be negative while the second will vanish; the three terms on the left-hand side cannot then sum to zero. So if $y(a)$ and $y(b)$ are both positive, the larger of the two must be the maximum value of y on the interval $[a,b]$. By a similar argument, we find that y cannot have a negative minimum. Now if $y(a)$ and $y(b)$ are both positive, y cannot become negative anywhere on the interval $[a,b]$. For if it did, it would have to possess a negative minimum, which it cannot. So with the meagerest of hypotheses we have proved that $0 < y \leq \max[y(a), y(b)]$ if $y(a)$ and $y(b)$ are positive.

The same style of reasoning we have just used can be employed to find bounds to solutions of Eq. (1). Suppose we know a function $u(x)$ that, while not satisfying the ordinary differential equation (1), does satisfy the differential inequality

$$\ddot{u} + g\dot{u} + hu > 0 \tag{2}$$

with the boundary inequalities

$$u(a) < y(a) \; , \tag{3a}$$

$$u(b) < y(b) \; . \tag{3b}$$

If we subtract Eq. (2) from Eq. (1) and write $w = y - u$ we get

$$\ddot{w} + g\dot{w} + hw < 0 \; . \tag{4}$$

Furthermore, from Eq. (3) we get

$$w(a)\ ,\ w(b) > 0\ . \tag{5}$$

The function $w(x)$ cannot have a negative minimum, for at a negative minimum $w < 0, \dot{w} = 0, \ddot{w} > 0$, which cannot satisfy Eq. (4). But then w can never dip below zero in the interval $[a,b]$. So $w > 0$ or, what is the same thing,

$$y \succ u\ . \tag{6}$$

The assertions made above hold if the direction of all inequalities is reversed. The assertions are true as well if on the right-hand side of Eqs. (1) and (2) zero is replaced by a function $f(x)$.

4.2 The restriction $h < 0$ plays an essential role in the foregoing arguments, which collapse completely without it. But even if h is not everywhere negative in $[a,b]$, if it is possible to find a function $t(x)$ *positive in* $[a,b]$ and such that

$$\ddot{t} + g\dot{t} + ht < 0\ , \tag{7}$$

then the above theorems can be rescued. To see how this works, let us start again with Eqs. (1), (2), and (3) and proceed exactly as before to Eqs. (4) and (5). If a positive function t obeying Eq. (7) can be found, then we set $w = st$. A short computation shows that

$$\ddot{s} + \left(g + 2\frac{\dot{t}}{t}\right)\dot{s} + \left(\frac{\ddot{t} + g\dot{t} + ht}{t}\right)s < 0 \tag{8}$$

while

$$s(a)\ ,\ s(b) > 0\ . \tag{9}$$

In view of Eq. (7), Eq. (8) is covered by the $h < 0$ case. Therefore, as in Sect. 4.1, $s > 0$. Since $t > 0$, this means $w > 0$ and $y > u$.

As an example of the use of these techniques, we take the following problem of Collatz: given

$$\ddot{y} + (1 + x^2)y + 1 = 0 \tag{10a}$$

with

$$y(\pm 1) = 0\ , \tag{10b}$$

estimate $y(0)$. To get a lower limit we need a function $u(x)$ that will make the left-hand side of Eq. (10a) greater than zero in the interval $(-1,\ +1)$. If $\ddot{u} + u + 1 = 0$ and if $u > 0$ in $(-1,\ +1)$ then $\ddot{u} + (1 + x^2)u + 1 > 0$ since $x^2u > 0$. If we take $u(\pm 1) = 0$, too, we then find $u = \sec 1 \cdot \cos x - 1$. If, as before, $w = y - u$, we find $\ddot{w} + (1 + x^2)w < 0$, $w(\pm 1) = 0$. Here $g = 0$ and $h = (1 + x^2) > 0$, so we must look for a function t satisfying Eq. (7) on the interval $(-1,\ +1)$. The function $t = 1 - x^2$ suffices, for $\ddot{t} + (1 + x^2)t = -2 + (1 - x^4) = -1 - x^4 < 0$. Then, as above at the beginning of this section, $w > 0$ or $y > u$. Thus, $y(0) > u(0) = \sec 1 - 1 = 0.8508$.

To get an upper limit we need a function $v(x)$ that will make the left-hand side of Eq. (10a) less than zero in the interval $(-1, +1)$. We try $v = a(1 - x^2)$, where a is a constant yet to be determined. Then $\ddot{v} + (1 + x^2)v + 1 = 1 - a(1 + x^4)$. For the right-hand side to be <0, we must have $a > (1 + x^4)^{-1}$. The largest value of the right-hand side of this last equation occurs where $x = 0$. Thus we must have $a > 1$. If $w = y - v$, then $\ddot{w} + (1 + x^2)w > 0$, $w(\pm 1) = 0$. Then, using the same kind of reasoning as at the beginning of this section, we find $w < 0$ or $y < v$. Then $y(0) < v(0) = a$, which we can take to be as low as but not lower than 1. So finally $0.8505 < y(0) < 1$. The geometric mean of these values, 0.9224, has the smallest maximum error, namely 8.4%.

Because Eq. (10a) is linear it can easily be solved by the sum of the solution of the inhomogeneous equation and a multiple of the solution of the homogeneous equation for both of which $\dot{y}(0) = 0$ and $y(0) = 1$, say. Both of these solutions are easily calculated numerically. The multiple of the solution of the homogeneous equation must be chosen to make $y(1) = 0$ for the sum. In this way, we find $y(0) = 0.932054$. The *closeness of the geometric mean to the exact value is pure coincidence*!

The same kind of logic as applied above to the two-point boundary value problem can be applied to the initial value problem, i.e., to the differential equation (1) with the values of $y(a)$ and $\dot{y}(a)$ specified. Suppose we have a function obeying the differential inequality of Eq. (2). As before, we find that the difference $w = y - u$ cannot have a negative minimum. If $w(0) < 0$ and $\dot{w}(0) < 0$, then w must be <0 everywhere. So if $u(a) > y(a)$ and $\dot{u}(a) > \dot{y}(a)$, then $u > y$ everywhere. If h is not <0, we can again rescue the various theorems if we can find a t satisfying Eq. (7).

4.3 The subject of this book is nonlinear differential equations, and the foregoing discussion of linear differential equations has been used only to illustrate the central idea of this chapter, namely, *that the differential equation or differential inequality restricts the kind of extrema the solutions may have.* Let us now turn our attention to a nonlinear two-point boundary value problem of the type we encountered in Sect. 3.6:

$$\ddot{y} + \dot{y}^2 - y^2 = 0 \,, \tag{11a}$$

$$y(0) = 1, \; y(\infty) = 0 \,. \tag{11b}$$

Equation (11a) has been chosen specifically because it is not invariant to an affine group.*

In order to solve the problem just posed, we need to learn how the integral curves through the point (0,1) behave. Maximum principles alone will not tell us everything we want to know, and their proper use, as we shall see in the examples below, is as an adjunct to other, more direct methods of analysis. A first cursory glance tells us that the integral curves of Eq. (11a) can never have maxima because at an extremum (if one exists at all!) $\ddot{y} = y^2 > 0$. A corollary is that integral curves emanating from the point (0,1) with non-negative slopes are monotone increasing. A slightly less obvious conclusion is this: two integral curves that emanate from the

*It is, however, invariant to the translation group $y' = y$, $x' = x + \lambda$.

point (0,1) with different slopes and always remain positive never intersect a second time. To see this, call the two solutions y_1 and y_2 and suppose that $\dot{y}_1(0) > \dot{y}_2(0)$ while $y_1(0) = y_2(0) = 1$. If we subtract Eq. (11a) written for y_2 from Eq. (11a) written for y_1, we find that $w = y_1 - y_2$ obeys the ordinary differential equation $\ddot{w} + (\dot{y}_1 + \dot{y}_2)\dot{w} - (y_1 + y_2)w = 0$ and the boundary and initial conditions $w(0) = 0$, $\dot{w}(0) > 0$. Since by hypothesis y_1 and y_2 are positive, w can never have a positive maximum (at which $\ddot{w} < 0$, $\dot{w} = 0$, $w > 0$). Therefore $w > 0$ everywhere, which means $y_1 > y_2$.

In order to find out more about the integral curves through (0,1), we study their asymptotic behavior for large x. This is easier than studying their behavior in general because we need keep only the dominant terms. In the extreme of large x, we expect one of the three terms in Eq. (11a) to become negligible with respect to the other two [which remain comparable, since they must cancel according to Eq. (11a)]. (1) Suppose the first two terms are comparable and the last negligible. Then $\ddot{y} = -\dot{y}^2$, which can be integrated at once to give $y = \ln(Ax + B)$. But then $\dot{y} = A/Ax + B \ll y$ when x is large, contrary to hypothesis. So this supposition is wrong. (2) Suppose the middle term is negligible compared with the other two. Then $\ddot{y} = y^2$, which can be integrated once to give $3\dot{y}^2 = 2y^3 + A$. Now if y gets large as $x \to \infty$, eventually A becomes negligible. If y gets small as $x \to \infty$, so must $\dot{y}$, in which case A must equal zero. So in either case, we continue by integrating a second time the differential equation $3\dot{y}^2 = 2y^3$ to obtain $y = 6/(x + B)^2 \sim 6/x^2$. The neglected middle term $\dot{y}^2 \sim 144/x^6$ is truly small compared with the first or third terms, $36/x^4$, when x is large enough. So $6/x^2$ is a consistent asymptotic behavior. (3) Suppose the first term can be neglected compared with the other two. Then $\dot{y}^2 = y^2$ so that $y \sim Ae^{\pm x}$. If $y = Ae^x$, $\ddot{y} = Ae^x \ll y^2 = A^2e^{2x}$, so Ae^x is a consistent asymptotic behavior. If $y = Ae^{-x}$, $|\ddot{y}| = Ae^{-x} \gg y^2 = A^2e^{-2x}$, so Ae^{-x} is not a consistent asymptotic behavior. All three terms cannot be asymptotically comparable because no two pairs lead to the same asymptotic behavior. The upshot of this line of argument is that the only possible asymptotic behaviors for the curves through (0,1) are Ae^x and $6/x^2$.

In view of these findings, the integral curves through the point (0,1) behave as sketched in Fig. 1. The upper curves behave asymptotically as Ae^x with positive A, the lower curves as Ae^x with negative A, and the separatrix between them as $6/x^2$. It is the separatrix that we want. From the sketch we see that numerical integration in the forward direction will be unstable. So to calculate the separatrix we shall have to integrate backward.

To get a pair of consistent boundary values $y(x)$ and $\dot{y}(x)$ with which to start the backward integration, we calculate an asymptotic series for the separatrix:

$$y = \frac{6}{x^2} + \frac{A}{x^3} + \frac{B}{x^4} + \frac{C}{x^5} + \frac{D}{x^6} + \cdots , \tag{12a}$$

where

$$B = (A^2 - 144)/8 , \tag{12b}$$

$$C = A(A^2 - 432)/72 , \tag{12c}$$

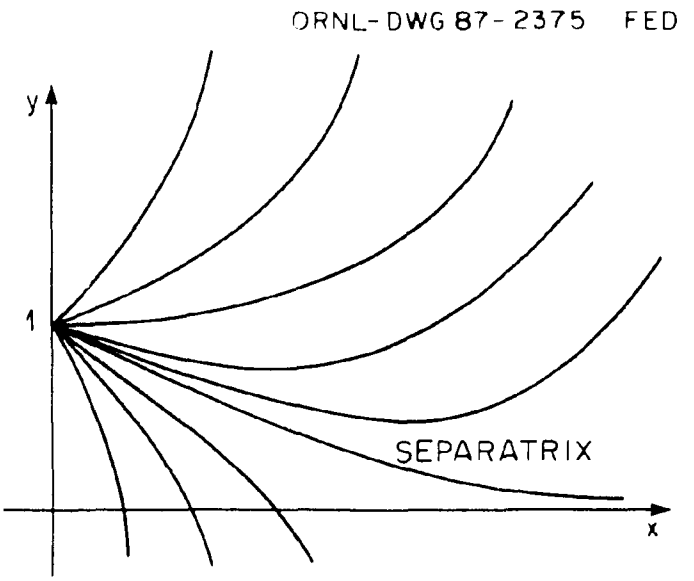

Fig. 1. A sketch of the integral curves of Eq. (11a) emanating from the point (0,1).

and

$$D = 5A^4/3456 - 5A^2/4 + 342/5 \,. \tag{12d}$$

Equations (12b)-(12d) have been obtained by inserting Eq. (12a) into the differential equation (11a), collecting terms, and equating the coefficient of each power of x individually to zero. Each value of A corresponds to a particular value of $y(0)$. By trial and error, aided in the last stages by interpolation to get the next guess, we find that for $A = -22.12$, $y(0) = 0.999957$ and $\dot{y}(0) = -0.657483$. The curve of $y(x)$ obtained by backward integration is drawn in Fig. 2. Drawn also is another curve obtained by forward integration using the above values of $y(0)$ and $\dot{y}(0)$; it shows clearly the instability caused by the divergence at large x of the integral curves in Fig. 1.

We can get quite satisfactory upper and lower bounds for $y(x)$ by using the maximum principle and taking as our family of comparison functions the family

$$u = \frac{6}{x^2 + ax + b} \,, \tag{13}$$

for which $u(0) = 6/b = y(0)$, $\dot{u}(0) = -6a/b^2$, and $u \sim 6/x^2$ for large x. A tedious but straightforward calculation shows that

$$\ddot{u} + \dot{u}^2 - u^2 = 36(x^2 + ax + b)^{-4} \left\{ \left[4 - \frac{1}{3}(a^2 - 4b) \right] x^2 + ax + b \right] + a^2 - 4b \right\} \,. \tag{14}$$

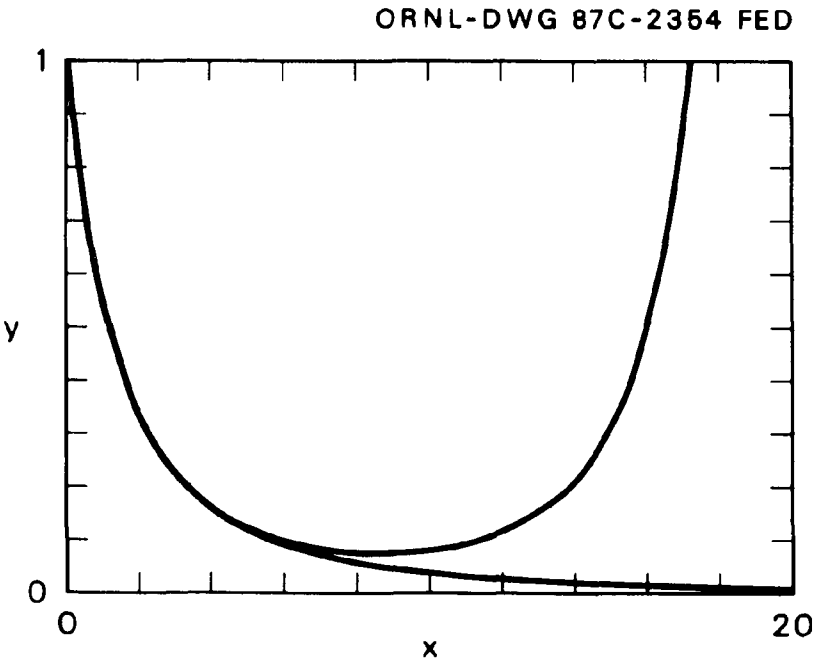

Fig. 2. The solution of Eqs. (11a) and (11b). The upper curve was obtained by a forward, unstable integration; the lower curve by a backward, stable integration.

To get a lower limit we want the right-hand side of Eq. (14) to be >0.* Now since $b > 0$ and $a > 0$, $x^2 + ax + b$ is positive and monotone increasing for $x > 0$. (Its minimum occurs at $x = -a/2 < 0$; at $x = 0$ it equals $b > 0$.) If the right-hand side of Eq. (14) is to be positive, then $4 + (1/3)(a^2 - 4b)$ must be positive, for if it were negative, then for large enough x the right-hand side of Eq. (14) would be negative. The smallest positive contribution the product of the square brackets makes occurs when $x = 0$. For the right-hand side of Eq. (14) still to be >0 when $x = 0$, we must have

$$\left[4 + \frac{1}{3}(a^2 - 4b)\right] b + a^2 - 4b > 0 \tag{15a}$$

or

$$a^2 > \frac{4b^2}{b+3} \,. \tag{15b}$$

The best lower limit of the family of Eq. (13) will have the smallest allowable value of a, namely,

$$a = \frac{2b}{\sqrt{b+3}} \,. \tag{15c}$$

When $b = 6$ $[y(0) = 1]$, $a = 4$.

*This is proved below.

To get an upper limit we want the right-hand side of Eq. (14) to be <0.* Then $4 + (1/3)(a^2 - 4b)$ must be negative. Then

$$a^2 < 4b - 12 \ . \tag{16a}$$

[This incidentally requires $b > 3$ since $a^2 > 0$. Thus, the family (13) will give an upper limit only if $y(0) = u(0) < 2$.] The best upper limit of the family (13) will have the largest allowable value of a, namely,

$$a = 2\sqrt{b-3} \ . \tag{16b}$$

When $b = 6$, $a = 2\sqrt{3}$. Thus

$$\frac{6}{x^2 + 4x + 6} < y < \frac{6}{x^2 + 2\sqrt{3}x + 6} \ . \tag{17}$$

To prove the first inequality rigorously, we start with the case of Eq. (15c) for which $\ddot{u} + \dot{u}^2 - u^2 > 0$, $u(0) = y(0) = 1$, and $u(x) \sim y(x) \sim 6/x^2$. Then $\ddot{w} + (\dot{y} + \dot{u})\dot{w} - (y + u)w < 0$, where $w = y - u$. Furthermore, $w(0) = 0$ and $w \to 0$ faster than $6/x^2$ as $x \to \infty$. Since $y + u > 0$, w can have no negative minimum. Then w must be >0 everywhere, so that $y > u$. The second inequality is proved in an entirely analogous manner: when $\ddot{u} + \dot{u}^2 - u^2 < 0$, w can have no positive maximum and so must be <0. Therefore, $y < u$.

Shown in Table 1 are values of $y(x)$ calculated by backward numerical integration with $A = -22.12$ and values given by the upper and lower limits in Eq. (17). The geometric mean of the two bounds differs fractionally from either bound by less than 3% so that, for practical purposes, it may be a satisfactory estimate.

In our brief study of linear equations in Sects. 4.1 and 4.2, we achieved some generality, but any such generality in the study of nonlinear equations hardly seems possible because of their wide variety of form.

4.4 Not only do ordinary differential equations have maximum principles, but so do partial differential equations. The best and simplest examples are Laplace's and Poisson's equations and the ordinary diffusion equation. We begin with them, and after making the principles clear, we move on to some nonlinear partial differential equations.

*This, too, is proved below.

Table 1. Exact values of $y(x)$, the solution of Eqs. (11a) and (11b), and the upper and lower bounds of Eq. (17)

x	Lower bound	Exact value	Upper bound
0.0	1.0000	1.0000	1.0000
0.5	0.7273	0.7318	0.7517
1.0	0.5455	0.5523	0.5734
2.0	0.3333	0.3402	0.3544
3.0	0.2222	0.2276	0.2363
4.0	0.1579	0.1618	0.1673
5.0	0.1176	0.1205	0.1242
6.0	0.09091	0.09305	0.09556
7.0	0.07229	0.07391	0.07571
8.0	0.05882	0.06008	0.06140
9.0	0.04878	0.04977	0.05077
10.0	0.04110	0.04189	0.04266
12.0	0.03030	0.03083	0.03132
14.0	0.02326	0.02362	0.02395
16.0	0.01840	0.01867	0.01890
18.0	0.01493	0.01512	0.01529
20.0	0.01235	0.01250	0.01262

Solutions of Laplace's equation, $\nabla^2\phi = 0$, always have their largest and smallest values on the boundary B on any closed region R and not in the interior. For, at a relative maximum $\phi_{xx} < 0$ and $\phi_{yy} < 0$, whereas at a relative minimum $\phi_{xx} > 0$ and $\phi_{yy} > 0$. Both of these necessary conditions are incompatible with Laplace's equation $\nabla^2\phi = \phi_{xx} + \phi_{yy} = 0$. So ϕ cannot have a relative maximum or a relative minimum in R. Its largest and smallest values therefore lie on B, the boundary of R. For such functions ϕ it is possible at every point to find at least one direction in which ϕ increases and at least one direction in which ϕ decreases.

This property of Laplace's equation enables us to get bounds on the solutions to problems involving Laplace's and Poisson's equations. Consider, for example, the following problem: $\nabla^2\phi = -1$; $\phi = 0$ on the perimeter P of a square S of side 2; find ϕ at the center of the square. One way to get an estimate of $\phi(0,0)$ is to construct a function ψ such that $\nabla^2\psi = -1$; in general this function will not vanish on P. The difference $\psi - \phi$ satisfies Laplace's equation $\nabla^2(\psi - \phi) = 0$ in S, so its maximum and minimum values must lie on P, where $\phi = 0$. Therefore, everywhere

in S, $\psi_{\min}(P) < \psi - \phi < \psi_{\max}(P)$. The art in this method is to try to make $\psi_{\min}(P)$ and $\psi_{\max}(P)$ close together.

The linearity of Laplace's and Poisson's equations enables us to form solutions by superposition. The function $\psi = -(1/4)(x^2 + y^2)$ satisfies $\nabla^2\psi = -1$. Its maximum and minimum on the perimeter of the square (sides $x = \pm 1$, $y = \pm 1$) are $-1/4$ and $-1/2$, respectively. If we add a constant A to the ψ we have a ψ that still obeys $\nabla^2\psi = -1$, but whose maximum and minimum values on the perimeter are now $A - 1/4$ and $A - 1/2$. To minimize the absolute deviation of ϕ from ψ we choose A to make $A - 1/4$ and $A - 1/2$ equal but opposite in sign, i.e., we choose $A = 3/8$. Since $\psi(0,0) = A = 3/8$, we have at last $-1/8 < 3/8 - \phi(0,0) < 1/8$ or $1/4 < \phi(0,0) < 1/2$. The estimate $\phi(0,0) = 3/8$ is thus correct within a maximum possible error of 33%.

To improve our estimate we must add to our trial function additional solutions of Laplace's equation. Since we are working in two dimensions, we can find such functions by taking the real and imaginary parts of any analytic function of the complex variable $x + iy$. The necessary symmetry of ϕ [$\phi(-x,y) = \phi(x,y) = \phi(x,-y)$] requires us to take the real part. Let us try adding a multiple of $\mathrm{Re}(x + iy)^4 = x^4 - 6x^2y^2 + y^4$, i.e., let us take

$$\psi = A - \frac{1}{4}(x^2 + y^2) + a(x^4 - 6x^2y^2 + y^4) \; . \tag{18}$$

Because of the symmetry of ψ we need consider only the line segment $x = 1$, $0 < y < 1$ in determining the largest and smallest values of ψ on P:

$$\psi(1,y) = A - \frac{1}{4}(1 + y^2) + a(1 - 6y^2 + y^4) \; . \tag{19}$$

Our task is now to choose A and a to make the difference between $\psi_{\max}$ and $\psi_{\min}$ calculated from Eq. (19) as small as possible. A moment's thought should make it clear that this task can be accomplished by choosing a to make the difference of the largest and smallest values of

$$f(y) = -\frac{1}{4}(1 + y^2) + a(1 - 6y^2 + y^4) \tag{20}$$

on the interval $0 < y < 1$ as close together as possible.

This task is more challenging than it might appear at first glance. The first thing we do is to find out whether $f(y)$ has an extremum on the interval $(0,1)$ and what kind it is. The extremum ($\dot{f} = 0$) lies at values of y satisfying $y^2 = 3 + 1/8a$. So there will be an extremum in the y interval $(0,1)$ only if $-1/16 < a < -1/24$.

Outside this range of a, the largest and smallest values of $f(y)$ occur at $y = 0$ and $y = 1$; their absolute difference equals $|5a + (1/4)|$. The smallest value this absolute difference has outside the interval $-1/16 < a < -1/24$ occurs for $a = -1/24$ and equals $1/24$.

When $-1/16 < a < -1/24$, $f(y)$ has an extremum for $y^2 = 3 + 1/8a$. At this extremum, $\ddot{f} = 24a + 1 < 0$, so the extremum is a maximum. A short computation shows $f_{max} = -(1 + 8a + 1/64a)$. The minimum value of f occurs at either $y = 0$ or $y = 1$; it is the smaller of $a - (1/4)$ and $-4a - (1/2)$. For $a < -1/20$, $f_{min} = a - (1/4)$; for $a > -1/20$, $f_{min} = -4a - (1/2)$. Thus, for $-1/16 < a < -1/20$, $\Delta f = f_{max} - f_{min} = -3/4 - 9a - 1/64a$, which is monotonic decreasing in that interval. For $-1/20 < a < -1/24$, $\Delta f = -1/2 - 4a - 1/64a$, which is monotonic increasing in that interval. Clearly, then, the best value of $a = 1/20$, for which $\Delta f = 1/80$; then $f_{max} = -23/80$, $f_{min} = -24/80$. If we choose $A = 47/160$, we find $-1/160 < A - \phi(0,0) < 1/160$ or $23/80 < \phi(0,0) < 24/80$. Thus, the estimate $\phi(0,0) = 47/160$ has a maximum possible error of $\pm 1/160$ ($\pm 2.1\%$).

Our estimates so far have been based on a solution of the partial differential equation that does not satisfy the boundary conditions. We can also get estimates from functions that do satisfy the boundary conditions but do not satisfy the partial differential equation. Suppose, for example, we have a function ψ that vanishes on the perimeter P of the square S but satisfies only the differential inequality $\nabla^2\psi > -1$. The difference $\psi - \phi$ vanishes on P and satisfies the differential inequality $\nabla^2(\psi - \phi) > 0$. Therefore $\psi - \phi$ cannot have a relative maximum anywhere (although now a relative minimum is possible). Its largest value occurs on the perimeter P—this value, of course, is zero. So inside S, $\psi - \phi < 0$ or $\psi < \phi$. In particular, $\psi(0,0) < \phi(0,0)$. A similar result holds when the sense of all the inequalities is reversed.

Let us choose for ψ the function

$$\psi = (1 - x^2)(1 - y^2)[a + b(x^2 + y^2)] , \tag{21}$$

where a and b are constants yet to be determined. (This function has been chosen in the following way. The first two factors have been chosen to ensure that $\psi = 0$ when $x = \pm 1$ or $y = \pm 1$. The squares are used to give ψ even symmetry under the transformation $x' = -x$ and $y' = -y$, to which the partial differential equation and boundary conditions are invariant. Similarly, ψ has been made symmetric under interchange of x and y just as the partial differential equation and boundary conditions are.) A short computation shows that

$$\nabla^2\psi = 4(b - a) - (16b - 2a)(x^2 + y^2) + 24bx^2y^2 + 2b(x^4 + y^4) . \tag{22}$$

In the corners of the square ($x = \pm 1$, $y = \pm 1$), $\nabla^2\psi = 0$ no matter what the values of a and b. So the trial function of Eq. (21) can at most satisfy the inequality $\nabla^2\psi > -1$ and therefore can only provide us with a *lower* limit to ϕ. Our task is to find the largest possible value of $a = \psi(0,0)$ consistent with the inequality

$$G(x,y,a,b) = 4(b - a) - (16a - 2a)(x^2 + y^2) + 24bx^2y^2 + 2b(x^4 + y^4) > -1 . \tag{23}$$

We begin by determining if G has any extrema in the square S. Differentiating, we find

$$G_x = -4(8b - a)x + 48bxy^2 + 8bx^3 \ , \tag{24a}$$

$$G_{xx} = -4(8b - a) + 48by^2 + 24bx^2 \ , \tag{24b}$$

and corresponding expressions for G_y and G_{yy} in which x and y are interchanged. Furthermore,

$$G_{xy} = 96bxy \ . \tag{24c}$$

Relative maxima and minima can occur only where $G_x = G_y = 0$. These points are

$$O: \quad x = 0, \quad y = 0 \ , \tag{25a}$$

$$Q: \quad x = 0, \quad y^2 = (8b - a)/2b \ \text{ and } \ y = 0, \quad x^2 = (8b - a)/2b \ , \tag{25b}$$

$$R: \quad x^2 = y^2 = (8b - a)/14b \ . \tag{25c}$$

The origin O is a relative minimum when $8b < a$ and a relative maximum when $8b > a$. The points Q are relative minima when $8b > a$ and relative maxima when $8b < a$. The points R are saddle points ($G_{xx}G_{yy} < G_{xy}^2$).

When $8b < a$, the minimum of G is at the origin and equals $G_{\min} = 4(b - a) > -1$. To find the largest value of a consistent with these inequalities, we plot the lines $8b = a$ and $4(b - a) = -1$ (see Fig. 3). The only admissible values of a and b correspond to points below the first line and above the second (hatched area). The largest possible value of a is that corresponding to the intersection $a = 2/7$, $b = 1/28$.

When $8b > a$, the minimum value of G occurs at the points Q, where

$$G_{\min} = 4(b - a) - \frac{(8b - a)^2}{2b} > -1 \ . \tag{26}$$

To find the largest value of a consistent with the inequality (26) and the inequality $8b > a$, we again plot them as equalities (see Fig. 4). The admissible values of a are in the hatched area. The largest value of a corresponds to the intersection R : $a = 2/7$, $b = 1/28$. [The maximum M of the curve lies at $b = (7 + \sqrt{14})/280 > 1/28$ and so cannot fulfill the requirement that $8b > a$.]

The lower limit $a = 2/7$, which we obtain in both cases, is close to the lower limit $23/80$ that we obtained earlier and is slightly inferior to it.

The restriction that ψ obey the boundary condition $\psi = 0$ on P is more stringent than we need, and $\psi(P) < 0$ is enough to prove that $\psi < \phi$ everywhere in S. For then $\nabla^2(\psi - \phi) > 0$ and $(\psi - \phi)_P > 0$. Since the maximum value of $\psi - \phi$ occurs on the boundary P, $\psi - \phi < (\psi - \phi)_{\max} < 0$, so that $\psi < \phi$ everywhere.

Many different combinations of differential and boundary inequalities are possible and have been discussed exhaustively by Protter and Weinberger. Always at the root of the discussion lie restrictions placed by the differential equation or inequality on the kind of extrema the solution may have.

4.5 A nonlinear analogue of Laplace's equation arises when we attempt to calculate steady temperature distributions in superfluid helium (He-II). Superfluid helium

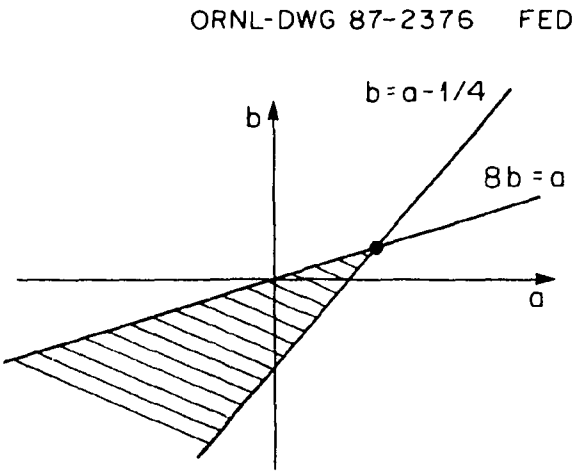

Fig. 3. Graphical determination of the maximum possible value of a when $8b < a$.

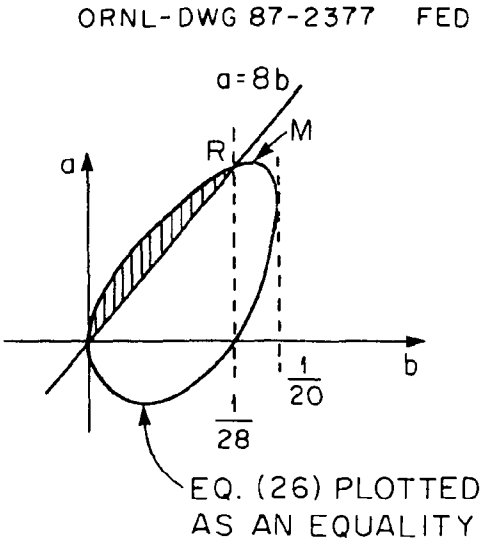

Fig. 4. Graphical determination of the maximum possible value of a when $8b > a$.

(He-II) is a very-low-temperature phase of helium ($T < 2.2$ K) that has some unusual physical properties. One of these, which interests us here, is that for heat fluxes in the practical range ($\gtrsim 0.1$ W/cm^2), the heat flux $\vec{Q}$ is proportional not to the temperature gradient T, but to its cube root:

$$\vec{Q} = -K(\nabla T)^{1/3} , \tag{27}$$

where K is a constant of proportionality taken to be independent of temperature. In steady heat flow, $\nabla \cdot \vec{Q} = 0$, so that

$$\nabla \cdot \left[K(\nabla T)^{1/3}\right] = 0 . \tag{28}$$

Equation (28) has a maximum principle, i.e., the largest and smallest temperatures lie on the boundary B of any region R. To see this, suppose that T has a relative maximum at some point P in the interior of R. In the neighborhood of P, the level surfaces of T are closed surfaces enclosing P. The vector $-\nabla T$ is the outward normal to these surfaces. Now $\nabla T = -Q^2\vec{Q}/K^3$, so $\vec{Q}\cdot(-\nabla T) = Q^4/K^3 > 0$, which means that the vector $\vec{Q}$ makes an acute angle with $-\nabla T$, the outward normal to the level surfaces of T. Hence $\int\int \vec{Q}\cdot\vec{ds} > 0$ when taken over a level surface of T. But since $\nabla \cdot \vec{Q} = 0$ everywhere, this integral must vanish. This is a contradiction, so our original supposition that T had a relative maximum must be false. A similar argument applies to relative minima.

In the case of a linear equation, the difference of two solutions, being a solution itself, has a maximum and a minimum principle. However, this simple argument does not suffice for Eq. (28) because it is nonlinear. Nevertheless, even though the difference of two solutions is *not* necessarily a solution, the difference obeys a maximum and a minimum principle. Suppose the two solutions are T_1 and T_2. Then

$$\begin{aligned} -K^3\nabla(T_1 - T_2)\cdot(\vec{Q}_1 - \vec{Q}_2) &= (Q_1^2\vec{Q}_1 - Q_2^2\vec{Q}_2)\cdot(\vec{Q}_1 - \vec{Q}_2) \\ &= Q_1^4 - (Q_1^2 + Q_2^2)\vec{Q}_1\cdot\vec{Q}_2 + Q_2^4 \\ &> Q_1^4 - (Q_1^2 + Q_2^2)Q_1Q_2 + Q_2^4 \\ &= (Q_1^3 - Q_2^3)(Q_1 - Q_2) \\ &= (Q_1^2 + Q_1Q_2 + Q_2^2)(Q_1 - Q_2)^2 > 0 . \end{aligned} \tag{29}$$

Thus $\vec{Q}_1 - \vec{Q}_2$ makes an acute angle with the normal $-\nabla(T_1 - T_2)$ to the level surfaces of $T_1 - T_2$. Since $\nabla\cdot(\vec{Q}_1 - \vec{Q}_2) = 0$, these level surfaces cannot be closed, i.e., $T_1 - T_2$ cannot have either a relative maximum or a relative minimum in the interior of any region R.

This argument can be extended to functions T_1 obeying differential and boundary inequalities. Suppose, for example, we have a function T_1 for which $\nabla\cdot\left[K(\nabla T_1)^{1/3}\right] > 0$ and for which $T_1(B) < T_2(B)$, where T_2 is a solution of Eq. (28). Then $\nabla\cdot\vec{Q}_1 < 0$ and so $\nabla\cdot(\vec{Q}_1 - \vec{Q}_2) < 0$. Thus $T_1 - T_2$ cannot have a relative maximum in B. For then, $\int\int(\vec{Q}_1 - \vec{Q}_2)\cdot d\vec{S}$ must be >0 when taken over

a closed level surface around the maximum. This contradicts $\nabla \cdot (\vec{Q}_1 - \vec{Q}_2) < 0$. Therefore, the largest value of $T_1 - T_2$ lies on B. Then $T_1 - T_2 < (T_1 - T_2)_{max} < 0$ since $T_1(B) < T_2(B)$, and thus $T_1 < T_2$ everywhere in R. The same argument applies when the inequalities are reversed and the words "largest" and "maximum" are replaced by the words "smallest" and "minimum," respectively.

As a numerical example let us choose the analogous problem to that considered in Sect. 4.4, namely $\nabla \cdot (\nabla T)^{1/3} = -1$; $T = 0$ on the perimeter P of a square S of side 2; find T at the center of the square. (For convenience we work in special units in which $K = 1$.) The function $T_1 = (R^4 - r^4)/32$ is a solution of the partial differential equation $\nabla \cdot (\nabla T)^{1/3} = -1$ written in cylindrical coordinates, $(1/r)(d/dr)[r(dT/dr)^{1/3}] = -1$. Here R^4 is a constant of integration yet to be chosen. The difference of the solution T we seek and the solution T_1 has its maximum on the perimeter P of the square S. Since $T(P) = 0$, we have in the interior of S

$$\min[T_1(P)] < T_1 - T < \max[T_1(P)] \; . \tag{30}$$

Owing to the geometric symmetry of the problem, we need only consider the values of $T_1(P)$ on the interval $x = 1$, $0 \leq y \leq 1$, where $T_1(P) = [R^4 - (1 + y^2)^2]/32$. Then we see at once that $\min[T_1(P)] = (R^4 - 4)/32$ and $\max[T_1(P)] = (R^4 - 1)/32$. Since $T_1(0,0) = R^4/32$, it follows from Eq. (30) that $1/32 < T(0,0) < 1/8$. The geometric mean of these extremes, $1/16$, is then correct to within a factor of 2.

4.6 The ordinary diffusion equation $C_t = C_{zz}$ has maximum and minimum principles. Consider the following typical boundary–initial value problem: $C(a,t)$, $C(b,t)$, and $C(z,0)$ are specified; what is the value of C at any point z, $a \leq z \leq b$, at any time $t > 0$? (See Fig. 5.) The solution C cannot have either a minimum or a maximum in the interior R of any finite region $a \leq z \leq b$, $0 \leq t < \infty$. For, at a maximum, $C_t = C_z = 0$ and $C_{zz} < 0$, which contradicts the equality $C_t = C_{zz}$, and similarly at a minimum. Hence the largest and smallest values of C lie on the boundary. Furthermore, they cannot lie on the segment AB, for if the maximum of C lay on the interior of segment AB, then there $C_t = C_{zz}$ would be <0. But then larger values of C would lie at smaller t and the same z, i.e., inside the region R. A similar argument holds for the minimum of C. So the largest and smallest values of C are determined by the boundary and initial conditions.

Since the ordinary diffusion equation is linear, the difference w of the two solutions C_1 and C_2 is also a solution. If $C_1(a,t) > C_2(a,t)$, $C_1(b,t) > C_2(b,t)$, and $C_1(z,0) > C_2(z,0)$, then $C_1 > C_2$ everywhere in R, for the smallest value of $w = C_1 - C_2$ must be on the boundary. But there $w > 0$. So $w > 0$ in R, i.e., $C_1 - C_2 > 0$ in R.

Solutions of the ordinary diffusion equation with a linear source term, $C_t = C_{zz} + h(z,t)C$, can similarly be compared when $h < 0$. If $C_1 > C_2$ on $z = a$, $z = b$, and $t = 0$, then $w = C_1 - C_2$ obeys $w_t = w_{zz} + hw$ and $w > 0$ on $z = a$, $z = b$, and $t = 0$. Can w have a minimum in R? If it does, then at the minimum $w_t = 0$ and $w_{zz} > 0$. Therefore $w > 0$, too. So if w has a minimum in R it must be positive, and therefore $w > 0$ everywhere in R. If w does not have a minimum in R, its

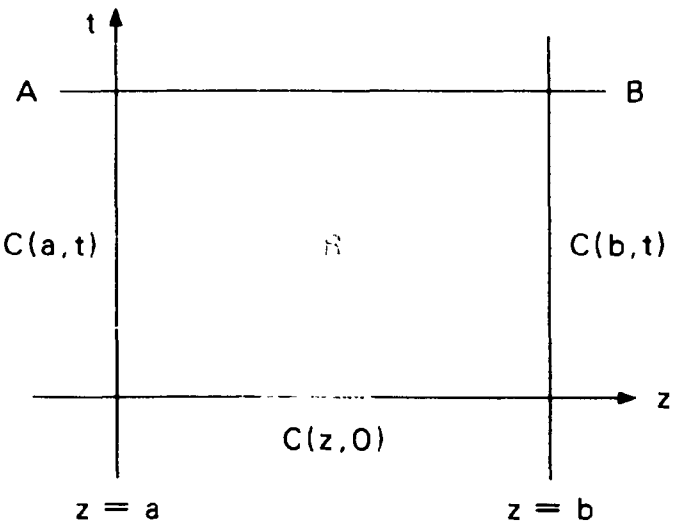

Fig. 5. A sketch of the boundary and initial conditions for the diffusion equation.

smallest value lies on the boundary where $w > 0$. Thus the smallest value of w is always positive, so $C_1 - C_2 = w > 0$ everywhere in R.

If h is not always negative, but is bounded in the interval $a \leq z \leq b$, we can rescue the result of the preceding paragraph by considering the function $g(z,t)$ defined by $C = ge^{\lambda t}$. Substitution into the partial differential equation for C shows that g obeys the partial differential equation $g_t = g_{zz} + (h - \lambda)g$. If we choose $\lambda > \max_{a \leq z \leq b}(h)$, then we can apply the reasoning of the foregoing paragraph to g and, because $e^{\lambda t} > 0$, ultimately to C.

An application of these ideas arises in a problem drawn from the domain of applied superconductivity. Shorn of its physical derivation, the mathematical problem comes down to this: the temperature C in a certain kind of superconducting magnet obeys to a good approximation the diffusion equation with source

$$C_t = C_{zz} + G(C) \,, \tag{31a}$$

where

$$G(C) = 0, \quad C < a \,, \tag{31b}$$

$$G(C) = b(C - a), \quad C > a \,. \tag{31c}$$

At time $t = 0$, a sudden heat pulse of strength q is introduced at the origin; that is to say, for $t = 0+$, the initial temperature distribution is taken to be

$$C = q\frac{\exp(-z^2/4t)}{(4\pi t)^{1/2}} \,. \tag{32}$$

The reader may recognize the second factor on the right-hand side of Eq. (2) as the pulsed-source solution of Eq. (3.14) of Sect. 3.4. If q is small enough, the temperature C everywhere eventually approaches zero (called recovery). If q is large enough, the temperature C everywhere eventually grows without bound (called quenching). We seek the value of q that divides these two kinds of behavior.

The key to solving this rather formidable nonlinear eigenvalue problem is to consider functions (trial solutions) of the form*

$$C_1 = h(t)\frac{\exp(-z^2/4t)}{(4\pi t)^{1/2}} = h(t)S(z,t) \ , \tag{33}$$

where S is an abbreviation for the pulsed-source solution (3.14). If we substitute Eq. (33) into Eq. (31a), we get

$$(C_1)_{zz} + G(C_1) - (C_1)_t = hS_{zz} + G(hS) - \dot{h}S - hS_t \tag{34a}$$

$$= G(hS) - \dot{h}S \ . \tag{34b}$$

We choose h to make the right-hand side of Eq. (34b) vanish when $z = 0$, i.e., to satisfy

$$\dot{h} = (4\pi t)^{1/2} G\left[\frac{h}{(4\pi t)^{1/2}}\right] \ . \tag{35}$$

With this value of h, the right-hand side of Eq. (34b) becomes

$$G(hS) - \dot{h}S = G\left[\frac{h}{(4\pi t)^{1/2}} e^{-z^2/4t}\right] - e^{-z^2/4t} G\left[\frac{h}{(4\pi t)^{1/2}}\right] \ . \tag{36}$$

Because $G(C)$ is *concave upward*, it has the property that

$$G[\theta C_1 + (1-\theta)C_2] \le \theta G(C_1) + (1-\theta)G(C_2), \quad 0 \le \theta \le 1 \tag{37a}$$

(see Fig. 6). This means $G(hS) - \dot{h}S \le 0$ with the equality occurring only for $z = 0$. Therefore $(C_1)_{zz} + G(C_1) - (C_1)_t \le 0$ with the equality occurring only at $z = 0$.

Now let us consider the difference w between C and C_1: $w = C - C_1$. It must satisfy

$$w_{zz} + G(C) - G(C_1) - w_t \ge 0 \tag{37b}$$

or

$$w_{zz} + \dot{G}[\theta C + (1-\theta)C_1]w - w_t \ge 0 \tag{37c}$$

if we use the law of the mean. Again, there is equality only if $z = 0$. If we choose the same initial values of Eq. (32) for C_1 as for C, then $w(z,0) = 0$. Furthermore, since $C(\pm\infty,\ t) = C_1(\pm\infty,\ t) = 0$, $w(\pm\infty,\ t) = 0$. By symmetry, $w_z(0,t) = 0$. These boundary and initial conditions are summarized in Fig. 7.

---

*While this trial solution may look like a *Deus ex machina* of the type I promised not to introduce, a little experimentation will show the reader that there is hardly any place else to begin—at least, I have not found any.

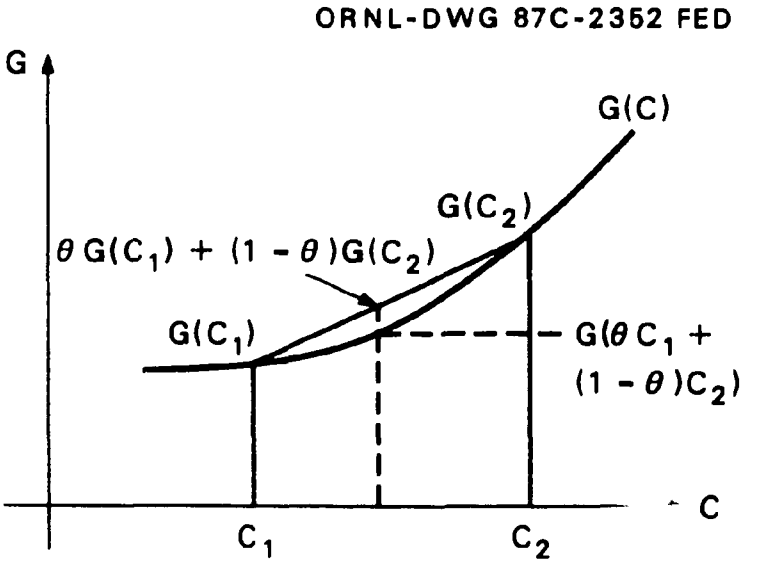

Fig. 6. Sketch illustrating property (37a) of functions that are concave upward.

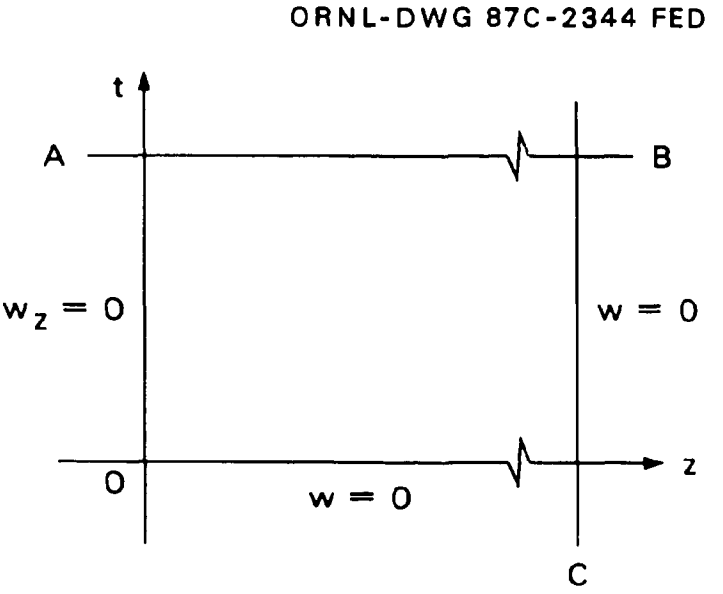

Fig. 7. Sketch showing the boundary and initial conditions w obeys. The line BC lies at very large z.

The function w cannot be positive in the interior of the rectangle $OABC$. To prove this we need to consider not w but the related function v, defined by $w = ve^{\lambda t}$; this is because $\dot{G} \geq 0$. The function v obeys the differential inequality

$$v_{zz} + (\dot{G} - \lambda)v - v_t \geq 0 \tag{38}$$

with equality only if $z = 0$. As a consequence of Eq. (38), $v \leq 0$ in the rectangle $OABC$. We prove this most easily by *reductio ad absurdum.*

Assume $v > 0$ somewhere in $OABC$. Then v can have no maximum in the interior of $OABC$. For if it did, then at the maximum, $v_{zz} < 0$, $v_t = 0$, which, together with $v > 0$, contradict Eq. (38) [if $\lambda > \max(\dot{G})$]. The largest value of v must then lie on the boundary of $OABC$. It cannot be on OC or CB, for then the largest value of v would be zero. If the maximum of v lay on OA, then there $v > 0$ and $v_t = 0$, so that, from Eq. (38), $v_{zz} > 0$. But since $v_z = 0$ on OA, $v_{zz} > 0$ means there are larger values of v just inside $OABC$ than on OA, so the largest value of v cannot be on OA. It cannot be on AB either, because if it were, v_{zz} would be <0 and v would be >0, so that from Eq. (38), v_t would be <0. Then there would be larger values of v just inside $OABC$, again a contradiction. Thus we are always led to a contradiction. So we must reject the hypothesis $v > 0$ somewhere and therefore must have $v < 0$ in $OABC$. But since $e^{\lambda t} > 0$, $w \leq 0$, or $C \leq C_1$.

This inequality means that if we choose $h(0)$ so that C_1 recovers, so must C. The solution C will surely recover for any smaller value of q, so $h(0)$ will be a *lower limit to the limiting value of* q. It remains only to calculate the largest value of $h(0)$ for which C_1 recovers.

We are interested only in temperature distributions for which $C_1(0,t) > a$. But then, since $h/(4\pi t)^{1/2} = C_1(0,t)$, G on the right-hand side of Eq. (35) is given by Eq. (31c):

$$\dot{h} = (4\pi t)^{1/2} b \left[\frac{h}{(4\pi t)^{1/2}} - a \right] \tag{39a}$$

$$= bh - ab(4\pi t)^{1/2} \ . \tag{39b}$$

The solution of Eq. (39b) is

$$h = \left[h(0) - ab \int_0^t (4\pi t)^{1/2} e^{-bt} dt \right] \cdot e^{bt} \ . \tag{40}$$

When $t \to \infty$, the second term in the square brackets approaches the value $\pi a/\sqrt{b}$. If $h(0) > \pi a/\sqrt{b}$, $h \to \infty$ exponentially (quench). If $h(0) < \pi a/\sqrt{b}$, $h \to 0$ exponentially. [In fact, it does not, for once $h/(4\pi t)^{1/2}$ drops below a, G must be replaced by zero, i.e., Eq. (39) no longer applies.] This corresponds to recovery. If $h(0) = \pi a/\sqrt{b}$,

$$C_1(0,t) = \frac{h}{(4\pi t)^{1/2}} = \frac{ae^{bt}}{(bt)^{1/2}} \int_{bt}^{\infty} e^{-u} u^{1/2} du \tag{41a}$$

$$\to a \text{ as } t \to \infty \ . \tag{41b}$$

Clearly, then, $h(0) = \pi a/\sqrt{b}$ is the limiting value of $h(0)$ we have been seeking. So finally, then, if $q < \pi a/\sqrt{b}$, C must recover; therefore the limiting value of q is $\geq \pi a/\sqrt{b}$.

4.7 The problem of the previous paragraph has an interesting group-theoretic property: it is invariant to the one-parameter family of groups of transformations

$$\begin{aligned} C' &= \lambda^{\alpha} C \qquad b' = \lambda^{-2} b \\ t' &= \lambda^2 t \qquad a' = \lambda^{\alpha} a \qquad 0 < \lambda < \infty \ , \\ z' &= \lambda z \qquad q' = \lambda^{\alpha+1} q \end{aligned} \tag{42}$$

where α is arbitrary. Now the limiting value of q can only be a function of a and b: $q = F(a,b)$. Moreover, this function relationship must hold unchanged for the primed values since, they, too, satisfy the stated problem. Thus

$$q' = F(a',b') \ , \tag{43}$$

$$\lambda^{\alpha+1} F = \lambda^{\alpha+1} q = F(\lambda^{\alpha} a, \lambda^{-2} b) \ . \tag{44}$$

If we differentiate with respect to λ and set $\lambda = 1$, we find

$$(\alpha + 1)F = \alpha a F_a - 2bF_b \ . \tag{45}$$

The characteristic equations are

$$\frac{da}{\alpha a} = \frac{db}{-2b} = \frac{dF}{(\alpha+1)F} \tag{46}$$

so that, most generally,

$$F(a,b) = a^{(\alpha+1)/\alpha} H(a^2 b^{\alpha}) \ , \tag{47}$$

where H is an arbitrary function.

Suppose we consider Eq. (47) written in terms of the primed variables for a particular value of α, namely, α_0:

$$F' = \alpha'^{(\alpha_0+1)/\alpha_0} H(a'^2 b'^{\alpha_0}) \ . \tag{48}$$

Let us now replace the primed variables by the unprimed variables according to Eq. (42):

$$\lambda^{\alpha+1} F = \lambda^{\alpha(\alpha_0+1)/\alpha_0} a^{(\alpha_0+1)/\alpha_0} H(\lambda^{2(\alpha-\alpha_0)} a^2 b^{\alpha_0}) \ . \tag{49}$$

If we introduce the abbreviations $\mu = \lambda^{2(\alpha-\alpha_0)}$ and $x = a^2 b^{\alpha_0}$ and substitute for F from Eq. (47) on the left-hand side of Eq. (49) we get, after some rearrangement,

$$H(\mu x) = \mu^{-1/2\alpha_0} H(x) \ . \tag{50}$$

We can determine the form of $H(x)$ by differentiating Eq. (50) with respect to μ and then setting $\mu = 1$:

$$x\dot{H} = -\frac{1}{2\alpha_0}H \ . \tag{51}$$

Equation (51) has the solution

$$H(x) = \text{const}\ x^{-(1/2\alpha_0)} \ . \tag{52}$$

Then

$$q = F(a,b) = a^{(\alpha_0+1)/\alpha_0} \cdot \text{const} \cdot (a^2 b^{\alpha_0})^{-1/2\alpha_0} = \text{const} \cdot ab^{-1/2} \ . \tag{53}$$

This is precisely the form derived at the end of Sect. 4.8, where the lower limit π was obtained for the constant.

Having discovered the form in Eq. (53), we now need only to find q numerically for a single choice of a and b in order to know it for all a and b. We can do this by repeatedly solving Eq. (31a) for various q, thereby bracketing the sought-for limiting value. The constant in Eq. (53) turns out to be 3.88 to three figures, about 24% larger than the lower bound π.

Chapter 5

MONOTONE OPERATORS AND ITERATION

"'Does the road wind up-hill all the way?'
'Yes, to the very end.'"
—Christina Rossetti
"Up-Hill"

5.1 Iteration is a very old technique for getting solutions of all kinds of equations—algebraic, transcendental, ordinary and partial differential, etc. Two problems beset its use. The first is the increasing complexity of computation required to evaluate higher iterates. The second, an issue of principle, is whether or not the sequence of iterates converges. Collatz has identified a broad class of iteration problems for which the question of convergence can be answered, namely, those based on the iteration of monotone operators.

An operator T is *monotone* if $w \geq v$ implies $Tw \geq Tv$. An operator T is *antitone* if $w > v$ implies $Tw \leq Tv$. Monotone and antitone operators and operators that can be written as the sum of a monotone and an antitone operator can all be made the basis of convergent iteration schemes. How to do this is the subject of this chapter.

Suppose we begin with the simple case of a pure monotone operator T, and suppose we can find an *upper* solution u_0 and a *lower* solution v_0, i.e., functions u_0 and v_0 that obey the following conditions:

$$u_0 \geq Tu_0 \equiv u_1 \quad , \tag{1a}$$

$$v_0 \leq Tv_0 \equiv v_1 \quad , \tag{1b}$$

$$v_0 \leq u_0 \quad . \tag{1c}$$

If we then create two iterative sequences, $u_{n+1} = Tu_n$ starting with u_0 and $v_{n+1} = Tv_n$ starting with v_0, we can show by induction that the sequence of u-iterates decreases, the sequence of v-iterates increases, and the nth u-iterate is greater than the nth v-iterate. The induction proceeds straightforwardly as follows. If $u_n \leq u_{n-1}$, then $u_{n+1} = Tu_n \leq Tu_{n-1} = u_n$; furthermore the inductive hypothesis holds for $n = 0$. Similarly, if $v_n \geq v_{n-1}$, $v_{n+1} = Tv_n \geq Tv_{n-1} = v_n$; the inductive hypothesis again holds for $n = 1$. Finally, if $v_n \leq u_n$, $v_{n+1} = Tv_n \leq Tu_n = u_{n+1}$, and the inductive hypothesis holds for $n = 0$.

The sequence of u-iterates decreases and is bounded from below; the sequence of v-iterates increases and is bounded from above. The sequences therefore have limit points u and v which obey $u = Tu$ and $v = Tv$. Frequently these limit points will be the same. So the iterates give upper and lower bounds to the solutions of $u = Tu$.

To see how we can apply this scheme to the approximation of solutions of differential equations, let us begin with an example of Collatz's, namely the first-order ordinary differential equation and boundary conditions

$$\dot{y} = (1 - x)y^2 \quad , \tag{2a}$$

$$y(0) = 1 \quad . \tag{2b}$$

Equations (2a) and (2b) can be written as

$$y = 1 + \int_0^x (1-t)y^2(t)\, dt \quad , \tag{3}$$

where t is a dummy variable of integration. We take the right-hand side of Eq. (3) to be the operator T:

$$Ty = 1 + \int_0^x (1-t)y^2(t)\, dt \quad . \tag{4}$$

Is T monotone? From Eq. (4) we see that

$$Tu - Tv = \int_0^x (1-t)(u^2 - v^2)dt = \int_0^x (1-t)(u+v)(u-v)dt \quad . \tag{5}$$

So long as $x \leq 1$ and u and v are positive, T is monotone.

Since y is monotone increasing for $x \leq 1$ [see Eq. (2a)], a possible value for $v_0 = 1$. Then,

$$v_1 = 1 + x - x^2/2 \quad , \tag{6a}$$

$$v_2 = 1 + x + x^2/2 - 2x^3/3 - x^4/4 + x^5/4 - x^6/24 \quad . \tag{6b}$$

For $x \leq 1$, $v_1 > v_0$ as desired. For u_0 we try the form

$$u_0 = 1 + x + ax^2 \quad , \tag{7a}$$

with a as yet undetermined. Then

$$u_1 = 1 + x + x^2/2 + (2a-1)x^3/3 - x^4/4 + (a^2 - 2a)x^5/5 - a^2x^6/6 \quad . \tag{7b}$$

By comparing Eqs. (7a) and (7b) we can see that u_1 will be $\leq u_0$ if $a = 1/2$. Then

$$u_1 = 1 + x + x^2/2 - x^4/4 - 3x^5/20 - x^6/24 \quad . \tag{7c}$$

Further iteration is extremely laborious so we stop here, noting that $v_2 < y < u_2 \leq u_1$, so that u_1 is an upper limit to y and v_2 is a lower limit. Equation (2a) was picked deliberately because it is solvable in terms of simple functions: $y = (1 - x + x^2/2)^{-1}$. Shown in Table s a comparison of v_2, y, and u_1.

Table 1. A comparison of v_2, y, and u_1 for $x = 1$

x	v_2	y	u_1
0.0	1.0000	1.0000	1.0000
0.2	1.2143	1.2195	1.2195
0.4	1.4333	1.4706	1.4719
0.6	1.6211	1.7241	1.7340
0.8	1.7473	1.9231	1.9575
1.0	1.7917	2.0000	2.0583

5.2 Our next example is the Poisson-Boltzmann equation $\nabla^2 y = e^y$ that comes up in certain problems of ionic distribution in strong ion exchangers. Suppose we look for the regular solution inside a region R that vanishes on the boundary B of R. The operator T we identify with $(\nabla^2)_{\text{inverse}} e^{(\ )}$, that is to say, we define Ty as the solution of

$$\nabla^2(Ty) = e^y \quad , \tag{8a}$$

$$Ty(B) = 0 \quad . \tag{8b}$$

ntitone, as we prove next.

. have two functions u_0 and v_0 such that $u_0 > v_0$. If we define $u_1 = Tu_0$ and $v_1 = Tv_0$, we have

$$\nabla^2 u_1 = e^{u_0} \ , \quad u_1(B) = 0 \ , \tag{9a}$$

$$\nabla^2 v_1 = e^{v_0} \ , \quad v_1(B) = 0 \ . \tag{9b}$$

If we subtract Eq. (9b) from Eq. (9a), we get

$$\nabla^2(u_1 - v_1) = e^{u_0} - e^{v_0} > 0 \ . \tag{10}$$

Thus, $u_1 - v_1$ cannot have a maximum in the interior of R [for at a maximum $(u_1 - v_1)_{xx} < 0$, $(u_1 - v_1)_{yy} < 0$, and $(u_1 - v_1)_{zz} < 0$]. The largest value of $u_1 - v_1$ must thus occur on the boundary. But the boundary value is zero. Hence in R, $u_1 - v_1 < 0$ or $u_1 < v_1$. So T is antitone.

We start again with upper and lower solutions u_0 and v_0, now defined such that

$$v_0 \leq u_0 \ , \tag{11a}$$

$$v_1 = Tu_0 \geq v_0 \ , \tag{11b}$$

$$u_1 = Tv_0 \leq u_0 \ , \tag{11c}$$

and create the iterative sequence $v_{n+1} = Tu_n$ and $u_{n+1} = Tv_n$. As before, we prove by induction the assertions $v_n \leq v_{n+1}$, $u_n \geq u_{n+1}$, and $v_n \leq u_n$. Thus, as before, the two sequences provide upper and lower bounds to stationary solutions $y = Ty$ confined between them.

Let us take for the region R a cylinder of radius 1. The reason for this choice is that this problem has an analytic solution that we can use to compare with the limits we calculate by iteration. For u_0 we choose $u_0 = 0$. The rationale behind this choice is the following. Since $\nabla^2 y = e^y > 0$, y cannot have a maximum inside the region R. Since $y(B) = 0$, and the largest value of y occurs on B, $y < 0$ in R. So $u_0 = 0$ is a simple convenient upper limit. For v_0 we take $v_0 = -b$, where b is an as yet undetermined positive constant, thus satisfying Eq. (11a). Then

$$\frac{1}{r}\frac{d}{dr} r \frac{dv_1}{dr} = e^{u_0} = 1 \ , \quad v_1(1) = 0 \ , \tag{12a}$$

$$\frac{1}{r}\frac{d}{dr} r \frac{du_1}{dr} = e^{v_0} = e^{-b} \ , \quad u_1(r) = 0 \ , \tag{12b}$$

so that

$$v_1 = (r^2 - 1)/4 \quad , \tag{13a}$$

$$u_1 = e^{-b}(r^2 - 1)/4 \quad . \tag{13b}$$

Since $r \leq 1$, $(r^2 - 1)/4 \leq 0$ and u_1 and u_0 satisfy Eq. (11c). In order to satisfy Eq. (11b), we must have $v_1 = (r^2 - 1)/4 \geq -b = v_0$ or $b \geq (1 - r^2)/4$. Thus $b \geq 1$, and

$$e^{-1/4}(r^2 - 1)/4 \geq y \geq (r^2 - 1)/4 \quad . \tag{14}$$

The Poisson-Boltzmann equation $\nabla^2 y = e^y$ is solvable in cylindrical coordinates. The most direct approach is to make use of the invariance to the group $y' = y - 2 \ln \lambda$, $r' = \lambda r$ and apply the method of Sect. 2.5. The computations are tedious and will not be repeated here—they are summarized in my paper in *J. Math. Phys.* **12** (7), 1339 (1971). The result, which can be verified by substitution, is

$$y = -\ln\left[\frac{a}{8}\left(1 - \frac{r^2}{a}\right)^2\right] \quad , \quad a = 5 + \sqrt{24} = 9.8990 \quad . \tag{15}$$

A comparison of the limits, Eq. (14), and the exact solution, Eq. (15), is shown in Table 2. The geometric mean of the limits has the smallest maximum possible error, namely 13%. Because of the exponential on the right-hand side of the Poisson-Boltzmann equation, further iteration is extremely difficult.

Table 2. A comparison of the limits, Eq. (14), and the exact solution, Eq. (15)

r	y_{lower}	y_{exact}	y_{upper}
0.0	−0.2500	−0.2130	−0.1947
0.1	−0.2475	−0.2110	−0.1928
0.2	−0.2400	−0.2049	−0.1869
0.3	−0.2275	−0.1947	−0.1772
0.4	−0.2100	−0.1804	−0.1635
0.5	−0.1875	−0.1618	−0.1460
0.6	−0.1600	−0.1389	−0.1246
0.7	−0.1275	−0.1115	−0.0993
0.8	−0.0900	−0.0793	−0.0701
0.9	−0.0475	−0.0423	−0.0370
1.0	0	0	0

This iterative technique can be extended to a region R of any shape. If we take $u_0 = 0$ and $v_0 = b$, we find that $\phi \leq y \leq \phi \exp(\phi_{\min})$, where ϕ is the solution of the *linear* problem $\nabla^2 \phi = 1$, $\phi(B) = 0$, and $\phi_{\min}$ is its minimum value in R. So for a sphere of unit radius, for example, $(r^2 - 1)/6 \leq y \leq e^{-1/6}(r^2 - 1)/6$. No exact solution is available for this case.

5.3 Another equation to which Collatz's method of monotone operators might be applied by way of example is the equation of D. Anderson and M. Lisak, which they

obtained from a similarity treatment of a problem in plasma physics [*IEEE Trans. Plasma Sci.* **PS-9** (2), 73–75 (1981)]:

$$\ddot{y} + x\dot{y}e^{-y} = 0 \quad , \tag{16a}$$

$$y(0) = a \quad , \tag{16b}$$

$$y(\infty) = 0 \quad . \tag{16c}$$

By dividing Eq. (16a) by $\dot{y}$ and integrating twice with respect to x, we obtain

$$y = a - b\int_0^x \exp\left(-\int_0^{x'} x''e^{-y(x'')}dx''\right)dx' \quad , \tag{17}$$

where a and b are positive numbers equal to $y(0)$ and $-\dot{y}(0)$, respectively. The right-hand side of Eq. (17) is defined as the operator T acting on y. If y increases, the inner exponential decreases, the outer exponential increases, and the right-hand side decreases. Thus if $u > v$, $Tu < Tv$, and T is antitone. So we look for upper and lower solutions u_0 and v_0 such that (i) $u_0 \geq v_0$, (ii) $Tv_0 \equiv u_1 \leq u_0$, and (iii) $Tu_0 \equiv v_1 \geq v_0$ to start our iterative sequence $u_{n+1} = Tv_n$ and $v_{n+1} = Tu_n$. We choose $u_0 = a$ and $v_0 = 0$. Then

$$u_1 = a - b\sqrt{\frac{\pi}{2}}\mathrm{erf}\left(\frac{x}{\sqrt{2}}\right) \quad , \tag{18a}$$

$$v_1 = a - b\sqrt{\frac{\pi}{2}}e^{a/2}\mathrm{erf}\left(\frac{x}{\sqrt{2}}e^{-a/2}\right) \quad . \tag{18b}$$

Conditions (i) and (ii) are satisfied by these functions no matter what the (positive) values of a and b. What about condition (iii)? Since the error function is <1 (and approaches 1 as $x \to \infty$), $v_1 \geq 0$ requires

$$a \geq b\sqrt{\frac{\pi}{2}}e^{a/2} \quad . \tag{19}$$

When Eq. (19) is satisfied, then $u_1 \geq y \geq v_1$, i.e.,

$$a - b\sqrt{\frac{\pi}{2}}\mathrm{erf}\left(\frac{x}{\sqrt{2}}\right) \geq y \geq a - b\sqrt{\frac{\pi}{2}}e^{a/2}\mathrm{erf}\left(\frac{x}{\sqrt{2}}e^{-a/2}\right) \quad . \tag{20}$$

According to Eq. (20), when $x \to \infty$,

$$a - b\sqrt{\frac{\pi}{2}} \geq y(\infty) \geq a - b\sqrt{\frac{\pi}{2}}e^{a/2} \quad , \tag{21}$$

the last inequality following from Eq. (19).

When a and b obey the strict equality (19), i.e., when the left-hand side is greater than the right-hand side, then it follows from Eq. (21) that $y(\infty) > 0$. Thus the solution that the limits of Eq. (20) enclose cannot be the one we seek [remember

Eq. (16c)!]. Only if both sides of Eq. (19) are equal is there even a chance for $y(\infty)$ to be zero. But numerical calculations show that even then $y(\infty) > 0$. So Collatz's iteration method tells us nothing about Anderson and Lisak's problem propounded in Eq. (16), although a certain amount of analysis is required to determine this.

In spite of this disappointment, the monotonicity of operator on the right-hand side of Eq. (17) can be of use of us. For if y is the exact solution of Eq. (16), then Eq. (17) gives

$$\frac{a}{b} = \int_0^\infty \exp\left(-\int_0^x x' e^{-y(x')} dx'\right) dx \quad . \tag{22}$$

The right-hand of Eq. (22) is a monotone operator acting on y. Now y itself is monotonic decreasing, as we can see from Eq. (16). For if it were not, it would have to possess an extremum at which $\dot{y} = 0$ and $\ddot{y} = 0$. But the only solution for which $\dot{y}$ and $\ddot{y}$ vanish simultaneously is a constant, which cannot fulfill Eqs. (16b) and (16c) at the same time. Thus $a = y(0) > y > 0 = y(\infty)$. Using $y = a$ and $y = 0$ in Eq. (22), we obt:

$$\sqrt{\frac{\pi}{2}} e^{a/2} \geq \frac{a}{b} \geq \sqrt{\frac{\pi}{2}} \tag{23a}$$

so that

$$\sqrt{\frac{2}{\pi}} a \geq b \geq \sqrt{\frac{2}{\pi}} a e^{-a/2} \quad . \tag{23b}$$

Shown in Fig. 1 is a curve of b vs a calculated numerically, as described below, and the limits shown in Eq. (23b).

Another procedure exactly like the one just carried out begins by integrating the differential equation (16a) from zero to x:

$$\dot{y} + b = -\int_0^x x\dot{y} e^{-y} dx = x e^{-y} - \int_0^x e^{-y} dx \quad . \tag{24}$$

Since y is monotonic decreasing and positive, $\ddot{y} = -\dot{y} e^{-y} > 0$. Thus y is also concave upward. But then $y \geq a - bx$, $0 < x < a/b$. If we choose $x > a/b$ and replace y in the integral by the comparison function

$$u = \begin{matrix} a - bx \ , & a \leq x \leq a/b \\ 0 \ , & a/b < x \end{matrix} \quad , \tag{25}$$

we find

$$\dot{y} + b > x(e^{-y} - 1) + (a - 1 + e^{-a})/b \quad . \tag{26}$$

In passing from Eq. (24) to Eq. (26) we have used the fact that the operator $Ty \equiv \int_0^x e^{-y} dx$ is antitone. If we now let $x \to \infty$, then $\dot{y} \to 0$ and so does $x(e^{-y} - 1)$. Thus Eq. (26) becomes

$$b > (a - 1 + e^{-a})^{1/2} \quad . \tag{27}$$

Figure 1 shows the limits (23b) and (27) as well as a curve calculated numerically. The numerical calculations were carried out with the aid of the invariance of the differential equation (16a) to the mixed translation-stretching group $x' = \lambda x$,

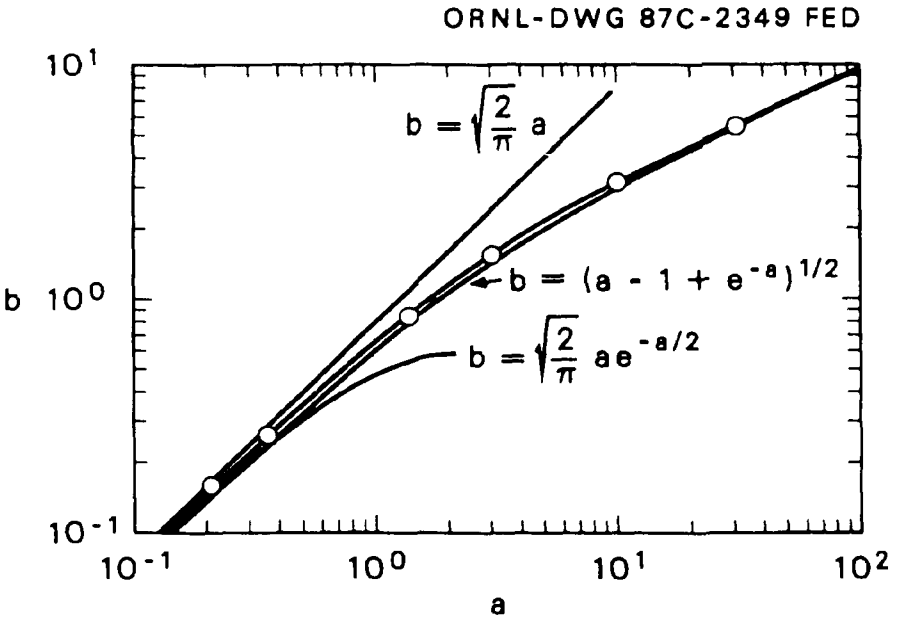

Fig. 1. The limits (23b) and (27) and a curve calculated numerically.

$y' = y + 2 \ln \lambda$. We proceed by picking $y(0)$ and $\dot{y}(0)$ arbitrarily and finding $y(\infty)$ by numerical integration (the integral curves all approach constants for large x). Then we transform the integral curve to get an image with $y'(\infty) = 0$. In this way, we find one point on the curve b vs a with each numerical integration.

It is clear from Eq. (23b) that $b \to \sqrt{2/\pi a}$ as $a \to 0$. It also happens that b approaches the limit (27) for large a, and this should not surprise us because the comparison function (25) becomes a closer and closer lower limit to the true solution, the larger a is.

5.4 Occasionally, one meets with operators that are neither monotone nor antitone, but which can be written as the sum of a monotone operator T_1 and an antitone operator T_2. To solve the problem $u = Tu + r$, Collatz sets up the iterative scheme

$$v_{n+1} = T_1 v_n + T_2 u_n + r \quad , \tag{28a}$$

$$u_{n+1} = T_1 u_n + T_2 v_n + r \quad , \tag{28b}$$

with starting values that obey the inequalities

$$v_0 \leq v_1 \leq u_1 \leq u_0 \quad . \tag{28c}$$

The success of the whole method depends on finding a u_0 and a v_0 that fulfill Eq. (28c). If we can do so, then

$$v_{n-1} \leq v_n \leq u_n \leq u_{n-1} \quad . \tag{29}$$

We prove Eq. (29) straightforwardly by induction:

$$v_{n+1} = T_1 v_n + T_2 u_n + r \geq T_1 v_{n-1} + T_2 u_{n-1} + r = v_n \quad , \tag{30a}$$

$$u_{n+1} = T_1 u_n + T_2 v_n + r \leq T_1 u_{n-1} + T_2 v_{n-1} + r = u_n \quad , \tag{30b}$$

$$v_{n+1} = T_1 v_n + T_2 u_n + r \leq T_1 u_n + T_2 v_n + r = u_{n+1} \quad . \tag{30c}$$

Chapter 6
COMPLEMENTARY VARIATIONAL PRINCIPLES

"Searcher for the fecund minimum."
—Wallace Stevens
"The Comedian as the Letter C"

6.1 The variational technique for solving differential equations is based on the connection between the extrema (maxima or minima) of a functional and the solution of a related differential equation. (A functional is a function of a function: you put in a function as the independent variable and get back a number. For example, $\int_0^1 y(x)\,dx$ is a functional of y.) The connection between functionals and differential equations is explored thoroughly in the calculus of variations, where the functionals are chosen because of their intrinsic interest. For example, in the classical brachistochrone problem the functional is the time it takes for a bead to slide down a wire connecting two points. Desired is the shape of the wire to make the time of transit a minimum. The wire shape is calculated by solving a related differential equation calculable from the particular functional. How to obtain this differential equation from the functional is part of the lore of the calculus of variations.

The process can be inverted. Given a particular differential equation, we may sometimes be able to find a functional that is minimized or maximized by solutions of the differential equation. Then we can choose a family of trial functions containing one or more undetermined parameters, evaluate the functional, and choose the parameters to make the functional an extremum. Used in this way, the functional provides a criterion of best fit. But it is not the only criterion of best fit. Indeed, it is not always even the most convenient. Its real power shines when the functional represents a quantity in which we may have some interest. Then, because the functional is an extremum for the solutions of the differential equation, when the error ϵ in the trial function is small, the error in the value of the functional is of order ϵ^2. Roughly speaking, then, a 10% trial function will provide a 1% estimate of the functional. If the latter is something we should like to know, we shall have gotten something for nothing.

The variational method has been used for a long time in the manner just described, and variational estimates have been obtained for myriad quantities of interest in science and technology. But all of these estimates suffered the peculiar defect that, while they were felt to be accurate, no rigorous measure of their error was available.

About 20 years ago, B. Noble remedied this defect for a wide class of differential equations. He showed that it was possible to find two variational principles, called complementary, one of which attained a maximum and the other an equal minimum for exact solutions of the differential equation. In such a case, trial functions provide two estimates of the desired quantity of second-order accuracy, and furthermore one necessarily is a lower limit and the other an upper limit. So Noble's method provides us with close upper and lower bounds to the desired quantity. Noble's method has been elaborated in a very fine monograph by A. M. Arthurs.

6.2 The key to Noble's method is the formulation of the problem in the Hamiltonian form. To understand the Hamiltonian form, we must first understand the Euler-Lagrange form. Suppose we start with an ordinary second-order differential equation, which is the so-called Euler-Lagrange equation of a Lagrangean $L(q, \dot{q})$:

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}\right) - \frac{\partial L}{\partial q} = 0 \quad . \tag{1}$$

For example, the differential equation $\ddot{q} - q = 0$ (whose solutions are $e^{\pm t}$) is the Euler-Lagrange equation of the Lagrangean $L = \dot{q}^2/2 + q^2/2$. The importance of the Lagrangean is this. Among all the functions $q(t)$ for which $q(a) = q_1$ and $q(b) = q_2$, the solution $q_*(t)$ of the Euler-Lagrange differential equation (1) that fulfills the boundary conditions $q_*(a) = q_1$ and $q_*(b) = q_2$ makes the functional

$$A = \int_a^b L(q, \dot{q})\, dt \tag{2}$$

an extremum (in the example being discussed, a minimum).

To see the meaning of this last statement in some detail and to set the stage for further developments, let us consider the problem of finding the solution of $\ddot{q} - q = 0$ and its associated value of A when $q(0) = 0$ and $q(1) = 1$. The exact solution is $q = \sinh t / \sinh 1 \equiv q_*(t)$. The value of A corresponding to it is $\sinh 2/4 \sinh^2 1 = 0.656518$. Another function of t, not a solution of the differential equation $\ddot{q} - q = 0$, but obeying the boundary conditions $q(0) = 0$ and $q(1) = 1$, is $q = t$. For it, the value of A is $2/3$, a slight overestimate of the correct value by about 1.5%.

It is easy to see from the differential equation that $\ddot{q}_* \neq 0$ in general. In fact, q_* must be concave upward. The trial function $q = t$, on the other hand, has no curvature. We can try to improve our trial function by including some curvature. So, for example, we can take as our trial function $q = at + (1 - a)t^2$, where a is some number not yet specified. For this trial function, $A = (5a^2 - 8a + 19)/24$. In order to make A an extremum (in this case, a minimum, as we shall see below) we set $dA/da = 0$. Then we find at once that $a = 4/5$. The corresponding value of A is $79/120$, which overestimates the correct value by a scant 0.28%. Shown below in Table 1 are the values of q_* and the two trial functions t and $t(4 + t)/5$. The trial function t is larger than q_* by as much as 17% in places, but the corresponding value of A is only 1.5% larger than A_*. The trial function $t(4 + t)/5$ sometimes exceeds q_* and sometimes is exceeded by it, but the percentage difference between them is at most about 4% and is usually less. The corresponding value of A exceeds A_* by only 0.28%. This example shows clearly how much better an estimate A is of A_* than q is of q_*.

Table 1. The exact solution q_* and two trial functions

t	q_*	t	$t(4+t)/5$
0.00	0	0	0
0.10	0.085234	0.10	0.082000
0.20	0.171320	0.20	0.168000
0.30	0.259122	0.30	0.258000
0.40	0.349517	0.40	0.352000
0.50	0.443409	0.50	0.450000
0.60	0.541740	0.60	0.552000
0.70	0.645493	0.70	0.658000
0.80	0.755705	0.80	0.768000
0.90	0.873482	0.90	0.882000
1.00	1.000000	1.00	1.000000

I have said above that in the example being discussed the functional A is a minimum when $q = q_*$, and now is the time to show it. Suppose we choose as trial functions the family of functions $q = q_* + \eta$, where η is an arbitrary function of t except that $\eta(a) = \eta(b) = 0$. Thus $q(a) = q_*(a) = q_1$ and $q(b) = q_*(b) = q_2$, i.e., q obeys the same boundary conditions as q_*. Then

$$A = \frac{1}{2}\int_a^b (\dot{q}_*^2 + q_*^2)\, dt + \int_a^b (\dot{\eta}\dot{q}_* + \eta q_*)\, dt + \frac{1}{2}\int_a^b (\dot{\eta}^2 + \eta^2)\, dt \quad . \tag{3}$$

The first term on the right-hand side of Eq. (3) is A_*. The second term we treat by integration by parts:

$$\int_a^b (\dot{\eta}\dot{q}_* + \eta q_*)\, dt = \eta\dot{q}_*\Big|_a^b + \int_a^b \eta(q_* - \ddot{q}_*)\, dt = 0 \quad . \tag{4}$$

The integrated term vanishes because $\eta(a) = \eta(b) = 0$. The integral on the right-hand side vanishes because q_* obeys the differential equation $\ddot{q}_* - q_* = 0$. Thus,

$$A = A_* + \frac{1}{2}\int_a^b (\dot{\eta}^2 + \eta^2)\, dt \quad . \tag{5}$$

Since the integral on the right-hand side is always positive, $A > A_*$, with equality being achieved if and only if $\eta = 0$, i.e., $q = q_*$. Thus A has as its minimum value A_*, which is attained only for the solution of the differential equation. Furthermore, the integral on the right-hand side is of second order in η, so if η is small, the estimate that Eq. (5) provides of A_* is much better than the estimate that q provides of q_*.

6.3 The reasoning just applied to the functional A given in Eq. (3) can be extended to the general functional A given in Eq. (2). Thereby we shall show that the solution of the Euler-Lagrange equation, Eq. (1), makes A an extremum, and we shall find conditions that will tell us whether the extremum is a minimum, a maximum, or

neither. Suppose $q = q_* + \eta$, $\dot{q} = \dot{q}_* + \dot{\eta}$, where q_* is the solution of the Euler-Lagrange equation obeying the boundary conditions $q_*(a) = q_1$ and $q_*(b) = q_2$ and where q is a trial function obeying the same boundary conditions. Then, to second order in η we have

$$A = \int_a^b dt \left[L(q_*, \dot{q}_*) + \frac{\partial L}{\partial q_*}\eta + \frac{\partial L}{\partial \dot{q}_*}\dot{\eta} + \frac{1}{2}\frac{\partial^2 L}{\partial q_*^2}\eta^2 + \frac{\partial^2 L}{\partial q_* \partial \dot{q}_*}\eta\dot{\eta} + \frac{1}{2}\frac{\partial^2 L}{\partial \dot{q}_*^2}\dot{\eta}^2 \right] \quad (6a)$$

$$= A_* + \int_a^b \left(\frac{\partial L}{\partial q_*}\eta + \frac{\partial L}{\partial \dot{q}_*}\dot{\eta} \right) dt$$

$$+ \frac{1}{2}\int_a^b \left(\frac{\partial^2 L}{\partial q_*^2}\eta^2 + 2\frac{\partial^2 L}{\partial q_* \partial \dot{q}_*}\eta\dot{\eta} + \frac{\partial^2 L}{\partial \dot{q}_*^2}\dot{\eta}^2 \right) dt \ . \quad (6b)$$

In order for A to differ from A_* in second order, the first-order term, which is the first integral on the right-hand side of Eq. (6b), must vanish for any arbitrary η for which $\eta(a) = \eta(b) = 0$ [remember, $\eta(a) = q(a) - q_*(a) = 0$]. A possible and convenient choice for η is a sharply peaked function centered on some point $t = t_0$ in the interval $a < t < b$ (see Fig. 1a). The first term in the first integral on the right-hand side of Eq. (6b) is then $(\partial L/\partial q_*)_{t=t_0} \int_a^b \eta \ dt$. We lose no generality by taking the area under the sharp peak to be unity, so that $\int_a^b \eta \ dt = 1$. Then the first term in the first integral is just $(\partial L/\partial q_*)_{t=t_0}$.

We can use the same trick on the second term with one slight addition of complexity. Because the derivative $\dot{\eta}$ does not have a single sharp peak (see Fig. 1b), it does not simply pick out the value of its coefficient at $t = t_0$. But an integration by parts is all we need to complete our calculation:

$$\int_a^b \left(\frac{\partial L}{\partial \dot{q}_*} \right) \dot{\eta} dt = \left(\frac{\partial L}{\partial \dot{q}_*} \right) \eta \Big|_a^b - \int_a^b \eta \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_*} \right) dt \quad (7a)$$

$$= -\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_*} \right)_{t=t_0} \ . \quad (7b)$$

The integrated term vanishes because $\eta(a) = \eta(b) = 0$. Adding the two terms, we find for the first integral on the right-hand side of Eq. (6b)

$$\left[\frac{\partial L}{\partial q_*} - \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_*} \right) \right]_{t=t_0} = 0 \ . \quad (7c)$$

If this first integral is to vanish, the quantity in brackets in Eq. (7c) must vanish. Since the choice of t_0 on which to center the sharply peaked function was arbitrary, Eq. (7c) must vanish for all t_0 in the interval (a, b). But this means that then q_* satisfies the Euler-Lagrange equation, Eq. (1).

Next we determine whether the extreme value A_* of the functional A is a minimum, a maximum, or neither. We can write the second integral as

$$\frac{1}{2}\int_a^b \eta^2 \left[L_{q_* q_*} + 2L_{q_* \dot{q}_*}\left(\frac{\dot{\eta}}{\eta} \right) + L_{\dot{q}_* \dot{q}_*}\left(\frac{\dot{\eta}}{\eta} \right)^2 \right] dt \ . \quad (8a)$$

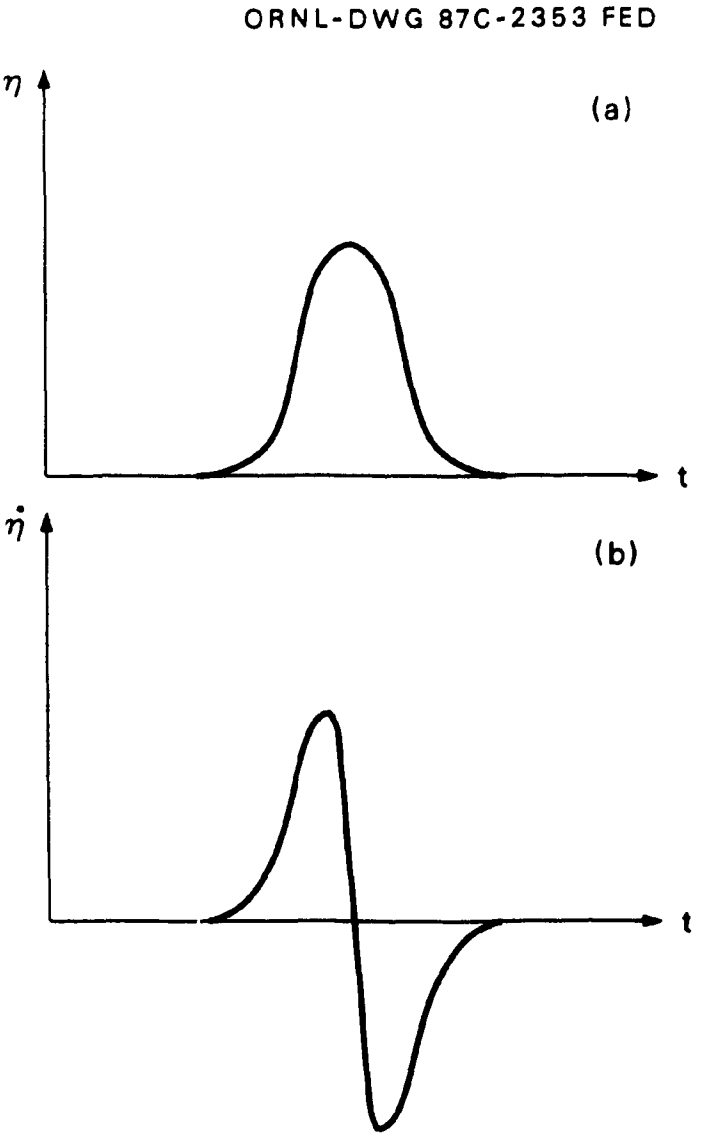

Fig. 1. The sharply peaked function $\eta(t)$ and its derivative $\dot{\eta}(t)$.

(For convenience, I abbreviate $L_{q_*q_*} = \partial^2 L/\partial q_*^2$, etc.) The quantity in the parentheses is a quadratic expression in the variable $\dot{\eta}/\eta$. If it has a positive discriminant,

$$(L_{q_*\dot{q}_*})^2 - L_{q_*q_*}L_{\dot{q}_*\dot{q}_*} \quad , \tag{8b}$$

it has real roots, i.e., there are values of $\dot{\eta}/\eta$ for which it vanishes, and it is positive for some values of $\dot{\eta}/\eta$ and negative for others. Hence A can be either greater than or smaller than A_*, although it always differs from A_* in second order. On the other hand, if the discriminant is negative, there are no roots, and the expression in the parentheses must always have the same sign. It will always be positive if $L_{q_*q_*}$ and $L_{\dot{q}_*\dot{q}_*}$ are both positive [if the discriminant (8b) is negative, $L_{q_*q_*}$ and $L_{\dot{q}_*\dot{q}_*}$ must have the same sign]. Then A is always greater than A_*, and A_* is a minimum. If $L_{q_*q_*}$ and $L_{\dot{q}_*\dot{q}_*}$ are negative, A will always be less than A_*, and A_* is a maximum.

6.4 As a simple example illustrating the application of the foregoing ideas, let us consider a problem suggested by Collatz (L. Collatz, *Differentialgleichungen*, B. G. Teubner, Stuttgart, 1967, pp. 172–5), namely, the linear eigenvalue problem

$$\ddot{y} + \lambda x y = 0 \quad , \tag{9a}$$

$$\dot{y}(0) = 0 \ , \quad y(1) = 0 \quad , \tag{9b}$$

which arises in the calculation of the mechanical stability of a vertical rod supporting its own weight. (The lowest eigenvalue λ gives the critical value of $\mu g l^3/B$ at which the rod buckles under its own weight. Here μ is the mass of the rod per unit length, g the acceleration of gravity, l the length of the rod, and B its flexural rigidity.) The Lagrangean for the differential equation (9a) is

$$L = \frac{1}{2}\dot{y}^2 - \frac{1}{2}\lambda x y^2 \quad , \tag{10a}$$

and the functional A (in mechanics called the action) corresponding to it is

$$A = \frac{1}{2}\int_0^1 (\dot{y}^2 - \lambda x y^2)\, dx \quad . \tag{10b}$$

What is the value of the action when $y = y_*$, the solution of the eigenvalue problem (9a,b)? If we integrate the first term on the right in Eq. (10b) by parts, we get

$$2A_* = y_*\dot{y}_*\Big|_0^1 - \int_0^1 (y_*\ddot{y}_* + \lambda x y_*^2)\, dx = 0 \tag{10c}$$

because $y\dot{y} = 0$ when $x = 0$ [$\dot{y}(0) = 0$] or 1 [$y(1) = 0$] and, according to Eq. (9a), $y_*\ddot{y}_* = -\lambda x y_*^2$. What this means is that, when y is a trial function that obeys the boundary conditions of Eq. (9b) and differs from y_* by an error of order ϵ,

$$A[y] = \frac{1}{2}\int_0^1 (\dot{y}^2 - \lambda x y^2)\, dx = A_* + O(\epsilon^2) = O(\epsilon^2) \quad . \tag{10d}$$

Thus Eq. (10d) provides the estimate for the eigenvalue,

$$\lambda = \frac{\int_0^1 \dot{y}^2 \, dx}{\int_0^1 xy^2 \, dx} \quad , \tag{11a}$$

the error in which is of order ϵ^2.

A simple trial function that obeys the boundary conditions of Eq. (9b) is $y = 1 - x^n$, where n is a parameter yet to be determined. A short calculation then shows that, according to Eq. (11a),

$$\lambda = \frac{2(n+1)(n+2)}{2n-1} \quad . \tag{11b}$$

The best value of n is that which makes the right-hand side of Eq. (11b) an extremum. (To see this, note that two trial functions, with neighboring values of n near the best value, must each lead to trial values of λ that differ from the correct value in second order. Hence the trial values of λ must differ from one another in second order.) The extremum of the right-hand side occurs when $n = (\sqrt{15}+1)/2$ and is equal to $\sqrt{15}+4 = 7.872983$.

Collatz's problem is soluble in terms of Bessel functions of order $1/3$, and using the properties of these functions, Collatz has obtained the value $\lambda = 7.83735$, from which our variational estimate differs by only 0.45%.

6.5 The Euler-Lagrange equation is a second-order equation. Hamilton's equations are an equivalent set of two coupled first-order equations. To derive them, Hamilton employed the so-called Legendre transformation that is used in thermodynamics to change independent variables. [A simple example of the Legendre transformation is the passage from the internal energy U to the Helmholtz free energy F. According to the two laws of thermodynamics, $dU = T\,dS - P\,dV$; thus U may conveniently be considered a function of the entropy S and the volume V. If we subtract $d(TS)$ from both sides, we obtain $d(U-TS) = T\,dS - P\,dV - S\,dT - T\,dS = -S\,dT - P\,dV$. The new function $F = U - TS$, called the Helmholtz free energy, is most conveniently considered a function of T and V.]

To reduce the Lagrange equation to a pair of first-order equations, Hamilton introduced the new variable $p = \partial L/\partial \dot{q}$. In terms of it, the Euler-Lagrange equation, Eq. (1), becomes $\dot{p} \equiv dp/dt = \partial L/\partial q$. Hamilton then introduced, in place of the Lagrangean, a related function H that could be considered a function of p and q; this he did by means of the Legendre transformation

$$H = p\dot{q} - L \quad . \tag{12a}$$

Then

$$\begin{aligned} dH &= \dot{q}\,dp + p\,d\dot{q} - L_q\,dq - L_{\dot{q}}\,d\dot{q} \\ &= \dot{q}\,dp + p\,d\dot{q} - \dot{p}\,dq - p\,d\dot{q} \\ &= \dot{q}\,dp - \dot{p}\,dq \end{aligned} \tag{12b}$$

because $L_{\dot{q}} = p$ and $L_q = d/dt(L_{\dot{q}}) = \dot{p}$. Thus

$$H_p \equiv \frac{\partial H}{\partial p} = \dot{q} \ , \tag{13a}$$

$$H_q \equiv \frac{\partial H}{\partial q} = -\dot{p} \ . \tag{13b}$$

Equations (13a) and (13b) are the coupled first-order equations of Hamilton. The function H is called, appropriately enough, the Hamiltonian.

6.6 In terms of the Hamiltonian, the functional A has the form

$$A = \int_a^b (p\dot{q} - H)\, dt \ . \tag{14}$$

Noble's idea is to study the behavior of A when p and q are trial functions that (i) are close to the exact solutions p_* and q_* of Eqs. (13a) and (13b) and (ii) obey *either* Eq. (13a) *or* Eq. (13b). If the Hamiltonian is of a certain type, then one of these families of trial functions will give an upper limit to A and the other will give a lower limit. Thus we shall be able to bracket the true value.

Suppose

$$p = p_* + \zeta \ , \tag{15a}$$

$$q = q_* + \eta \ . \tag{15b}$$

Then to terms of second order in ζ and η,

$$\begin{aligned} A &= \int_a^b [(p_* + \zeta)(\dot{q}_* + \dot{\eta}) - H(p_* + \zeta\ , q_* + \eta)]\, dt \\ &= \int_a^b \Big(p_*\dot{q}_* + \zeta\dot{q}_* + \dot{\eta}p_* + \zeta\dot{\eta} - H_* - \zeta H_{p_*} - \eta H_{q_*} \\ &\qquad - \frac{\zeta^2}{2} H_{p_*p_*} - \eta\zeta H_{p_*q_*} - \frac{\eta^2}{2} H_{q_*q_*} \Big) dt \\ &= A_* + \int_a^b (\zeta\dot{q}_* + \dot{\eta}p_* - \zeta H_{p_*} - \eta H_{q_*})\, dt \\ &\qquad + \int_a^b \left(\dot{\eta}\zeta - \frac{\zeta^2}{2} H_{p_*p_*} - \eta\zeta H_{p_*q_*} - \frac{\eta^2}{2} H_{q_*q_*} \right) dt \ . \end{aligned} \tag{16}$$

If we integrate the term $\dot{\eta}p_*$ in the first integral by parts we get

$$\int_a^b \dot{\eta}p_*\, dt = \eta p_* \Big|_a^b - \int_a^b \eta\dot{p}_*\, dt \ . \tag{17}$$

The integrated term vanishes if $\eta(a) = \eta(b) = 0$, i.e., if q obeys the same boundary conditions as q_*. Using the Hamilton equations, (13a) and (13b), we see then that the first-order term (first integral on the right) vanishes.

Suppose now we consider trial functions p and q that obey Eq. (13a). Then

$$\dot{q} = \dot{q}_* + \dot{\eta} = H_p(p_* + \zeta\ , q_* + \eta) = H_{p_*} + H_{p_*p_*}\zeta + H_{p_*q_*}\eta \qquad (18a)$$

or

$$\dot{\eta} = H_{p_*p_*}\zeta + H_{p_*q_*}\eta\ . \qquad (18b)$$

Substituting Eq. (18b) into the second-order term (second integral on the right), we find

$$A = A_* + \frac{1}{2}\int_a^b (\zeta^2 H_{p_*p_*} - \eta^2 H_{q_*q_*})dt \quad [p, q \text{ obey (13a)}]\ . \qquad (19)$$

If p and q instead obey Eq. (13b), then

$$\dot{p} = \dot{p}_* + \dot{\zeta} = -H_q(p_* + \zeta\ , q_* + \eta) = -H_{q_*} - H_{q_*p_*}\zeta - H_{q_*q_*}\eta \qquad (20a)$$

or

$$\dot{\zeta} = -H_{q_*p_*}\zeta - H_{q_*q_*}\eta\ . \qquad (20b)$$

Since $\int_a^b \dot{\eta}\zeta\ dt = \eta\zeta|_a^b - \int_a^b \eta\dot{\zeta}\ dt = -\int_a^b \eta\dot{\zeta}\ dt$ (because $\eta(a) = \eta(b) = 0$—remember, q and q_* obey the same boundary conditions),

$$A = A_* + \frac{1}{2}\int_a^b (\eta^2 H_{q_*q_*} - \zeta^2 H_{p_*p_*})\ dt \quad [p, q \text{ obey (13b)}]\ . \qquad (21)$$

[N.B.: The symbols η and ζ appearing in Eq. (21) are not numerically the same as those appearing in Eq. (19)!]

If $H_{q_*q_*}$ and $H_{p_*p_*}$ have *opposite* signs, or if one of them is zero, then the second-order terms in Eqs. (19) and (21) will have opposite signs. Thus, one of these equations will give an upper limit to A_* and the other a lower limit.

As an illustrative example, let us take the problem dealt with in Sect. 6.2, namely, $\ddot{q} - q = 0$, $q(0) = 0$, $q(1) = 1$. Then $H = (p^2 - q^2)/2$ so that Hamilton's equations are $\dot{p} = q$ and $\dot{q} = p$, which are clearly the equivalent of the second-order equation. The functional A is then given by

$$A = \int_0^1 \left(p\dot{q} - \frac{1}{2}p^2 + \frac{1}{2}q^2\right) dt\ . \qquad (22)$$

If $p = p_*$ and $q = q_*$, then $p_* = \dot{q}_*$ and

$$A_* = \frac{1}{2}\int_0^1 (\dot{q}_*^2 + q_*^2)\ dt = \frac{1}{2}\left[q_*\dot{q}_*\Big|_0^1 - \int_0^1 q_*(\ddot{q}_* - q_*)\ dt\right] = \frac{1}{2}\dot{q}_*(1)\ . \qquad (23)$$

So the limits we shall get will provide upper and lower bounds of second-order accuracy on the slope $\dot{q}_*$ at $t = 1$.

When p and q obey Eq. (13a): $\dot{q} = p$, we get an upper limit to A_* (since $H_{pp} = 1$ and $H_{qq} = -1$). Then A given by Eq. (22) becomes

$$A = \frac{1}{2}\int_0^1 (\dot{q}^2 + q^2)\ dt\ , \quad q(0) = 0\ , \quad q(1) = 1\ . \qquad (24)$$

This is what we had earlier in Sect. 6.2. There the choice $q = t$ gave $A = 2/3$. When p and q obey Eq. (13b): $\dot{p} = q$, we get a lower limit to A_*. If we take $q = t$ to satisfy the requirement that q and q_* obey the same boundary conditions, then we find $p = a + t^2/2$, where a is an as yet undetermined constant of integration. Then, substituting into Eq. (22), we find

$$A = \int_0^1 \left[\left(a + \frac{t^2}{2} \right) - \frac{1}{2} \left(a + \frac{t^2}{2} \right)^2 + \frac{1}{2} t^2 \right] = \frac{37}{120} + \frac{5}{6} a - \frac{1}{2} a^2 \quad . \tag{25}$$

Since Eq. (25) provides a lower limit, its maximum value of 59/90, which occurs when $a = 5/6$, is the best such lower limit. So we find then that $59/90 < A_* < 2/3$. The geometric mean of these limits, 0.661088, cannot be in error by more than 0.84% (its error is in fact 0.70%).

In Collatz's example (Sect. 6.4), the differential equation $\ddot{y} + \lambda x y = 0$ has the Hamiltonian $H = p^2/2 + \lambda x y^2/2$. Then $H_{pp} = 1$ and $H_{yy} = \lambda x$, and both are positive. Thus, the conditions for applying Noble's idea are not fulfilled, and although both Eqs. (19) and (21) provide second-order estimates of A, we have no guarantee that one is always an upper and the other always a lower bound.

6.7 The work up to now has dwelt on solutions $q(t)$ of ordinary differential equations. Now we turn to solutions $q(x, y, z)$ of partial differential equations. If such solutions make a Lagrangean of the form $L(q, q_x, q_y, q_z)$ an extremum, what is the form of the Euler-Lagrange differential equation? To answer this question, we proceed just as we did in Sect. 6.3 and set $q = q_* + \eta$:

$$A = \iiint_R L \, dx \, dy \, dz = A_* + \iiint_R (L_q \eta + L_{q_x} \eta_x + L_{q_y} \eta_y + L_{q_z} \eta_z) dx \, dy \, dz + \cdots \quad , \tag{26}$$

where the derivatives are to be evaluated for $q = q_*$. This can be written conveniently in vector notation if we define the vector $\vec{L}_{\nabla q}$ to be the vector with components L_{q_x}, L_{q_y}, and L_{q_z}. Then Eq. (26) becomes

$$A = A_* + \iiint_R (L_q \eta + \vec{L}_{\nabla q} \cdot \nabla \eta) dx \, dy \, dz + \cdots \tag{27a}$$

$$= A_* + \iint_C \eta \vec{L}_{\nabla q} \cdot d\vec{S} + \iiint_R (L_q - \nabla \cdot \vec{L}_{\nabla q}) \eta \, dx \, dy \, dz + \cdots \quad . \tag{27b}$$

Here C is the bounding surface of the region R and $d\vec{S}$ is its outward normal. The passage from Eq. (27a) to Eq. (27b) is by means of the vector identity $\nabla \cdot (s\vec{v}) = \vec{v} \cdot \nabla s + s \nabla \cdot \vec{v}$ and the divergence theorem.

If $\eta(C) = 0$ but η is otherwise arbitrary, we find the Euler-Lagrange equation

$$\nabla \cdot \vec{L}_{\nabla q} - L_q = 0 \tag{28}$$

for the exact solution q_*. So if we confine our trial functions q to those that obey the same boundary conditions as q_*, namely, $q(C) = q_*(C)$, then A differs from A_* in second order.

The partial differential equation $\nabla^2 q = -1$ furnishes an example of these considerations. This partial differential equation occurs in many applications with the boundary conditions $q(C) = 0$. Among those known to me are eddy current generation in noncircular plates by ramped fields, torsion of noncircular bars, and laminar flow of viscous fluids through noncircular pipes. According to Eq. (28), the Lagrangian L is

$$L = \frac{1}{2}(\nabla q)^2 - q \quad . \tag{29}$$

The functional A, *when evaluated for* $q = q_*$, the exact solution of the partial differential equation and boundary conditions, is

$$A_* = -\frac{1}{2}\iiint_R (\nabla q_*)^2 dx\ dy\ dz = -\frac{1}{2}\iiint_R q_*\ dx\ dy\ dz \quad . \tag{30}$$

In the three problems mentioned above, the value of A_* is directly related to the total eddy power dissipation in the plate, the torsional rigidity of the bar, and the total flow in the pipe, quantities of incontestable physical interest.

6.8 By recasting these equations in the Hamiltonian form, we can obtain upper and lower limits to the extreme value A_*. If we set

$$\vec{p} = \vec{L}_{\nabla q} \tag{31a}$$

and

$$H = \vec{p} \cdot \nabla q - L \quad , \tag{31b}$$

we find

$$\begin{aligned} dH &= d\vec{p} \cdot \nabla q + \vec{p} \cdot d(\nabla q) - \vec{L}_{\nabla q} \cdot d(\nabla q) - L_q\ dq \\ &= d\vec{p} \cdot \nabla q - \nabla \cdot \vec{p}\ dq \quad . \end{aligned} \tag{31c}$$

Thus the Hamilton equations become

$$\nabla q = H_{\vec{p}} \quad , \tag{31d}$$

$$-\nabla \cdot \vec{p} = H_q \quad , \tag{31e}$$

where $H_{\vec{p}}$ is the vector with components $\partial H/\partial p_x$, $\partial H/\partial p_y$, and $\partial H/\partial p_z$. In terms of H, A becomes

$$A \equiv \iiint_R [\vec{p} \cdot \nabla q - H(\vec{p}, q)]\ dx\ dy\ dz \quad . \tag{32}$$

Suppose now we substitute into Eq. (32) trial functions $\vec{p}$ and q that differ slightly from the true solutions $\vec{p}_*$ and q_*:

$$\vec{p} = \vec{p}_* + \vec{\zeta} \quad , \tag{33a}$$

$$q = q_* + \eta \quad . \tag{33b}$$

Then, to terms of second order in η and ζ,

$$A \equiv \iiint_R \Big[(\vec{p}_* + \vec{\zeta}) \cdot (\nabla q + \nabla \eta) - H_{\vec{p}} \cdot \vec{\zeta} - H_q \eta - \frac{1}{2} \vec{\zeta} \cdot H_{\vec{p}\vec{p}} \cdot \vec{\zeta}$$

$$- \eta H_{q\vec{p}} \cdot \vec{\zeta} - \frac{1}{2} H_{qq} \eta^2 \Big] dx\ dy\ dz \quad , \tag{34}$$

where all derivatives are to be evaluated for $q = q_*$ and $\vec{p} = \vec{p}_*$. Here $H_{\vec{p}\vec{p}}$ is the symmetric tensor whose components are $H_{p_x p_x}$, $H_{p_x p_y}$, etc. The first-order term vanishes if $\vec{p}$ and q obey appropriate boundary conditions on C, as we now see:

$$\iiint_R [\vec{p}_* \cdot \nabla \eta + \vec{\zeta} \cdot \nabla q_* - H_{\vec{p}} \cdot \vec{\zeta} - H_q \eta] dx\ dy\ dz$$

$$= \iint_C \eta \vec{p}_* \cdot d\vec{S} + \iiint_R [-\eta \nabla \cdot \vec{p}_*$$

$$+ \vec{\zeta} \cdot \nabla q_* - H_{\vec{p}} \cdot \vec{\zeta} - H_q \eta] dx\ dy\ dz \quad .$$

This transformation has been achieved using the vector identity $\nabla \cdot (\vec{p}_* \eta) = \eta \nabla \cdot \vec{p}_* + \vec{p}_* \cdot \nabla \eta$ and the divergence theorem. Because $\vec{p}_*$ and q_* obey the Hamilton equations [(31d) and (31e)] the terms in the last integral cancel in pairs (first and fourth, second and third). So if either (i) $\eta(C) = 0$ or (ii) $\vec{p}_* \cdot d\vec{S} = 0$, i.e., $\vec{p}_*$ is tangential to C, the surface integral vanishes and so does the first-order term. The first of these conditions means that the trial function q must obey the same boundary condition on C as does the exact solution. The second boundary condition depends on the problem we are solving and may or may not be fulfilled. Thus, to terms of second order,

$$A = A_* + \frac{1}{2} \iiint_R [2\vec{\zeta} \cdot \nabla \eta - \vec{\zeta} \cdot H_{\vec{p}\vec{p}} \cdot \zeta - H_{qq} \eta^2 - 2\eta H_{q\vec{p}} \cdot \vec{\zeta}] dx\ dy\ dz \quad . \tag{35}$$

If $\vec{p}$ and q obey the first Hamilton equation, Eq. (31d), then

$$\nabla q_* + \nabla \eta = H_{\vec{p}}(\vec{p}_* + \zeta\ , q_* + \eta) = H_{\vec{p}} + H_{\vec{p}\vec{p}} \cdot \vec{\zeta} + H_{\vec{p}q} \eta \quad . \tag{36}$$

Since q_* and p_* obey the Hamilton equations, the first terms on the left-hand side and right-hand side of Eq. (36) cancel. Substituting from Eq. (36) for $\nabla \eta$ into Eq. (35), we find

$$A = A_* + \frac{1}{2} \iiint_R [\vec{\zeta} \cdot H_{\vec{p}\vec{p}} \cdot \vec{\zeta} - H_{qq} \eta^2] dx\ dy\ dz \quad . \tag{37}$$

If, on the other hand, $\vec{p}$ and q obey the second Hamilton equation, Eq. (31e), then

$$-\nabla \cdot p_* - \nabla \cdot \vec{\zeta} = H_q(\vec{p}_* + \vec{\zeta}\ , q_* + \eta) = H_q + H_{q\vec{p}} \cdot \vec{\zeta} + H_{qq} \eta \quad . \tag{38}$$

Again, the first terms on the left-hand side and the right-hand side cancel because $\vec{p}_*$ and q_* obey the Hamilton equations. We shall use Eq. (38) to obtain the term $2\vec{\zeta}\cdot\nabla\eta$ in Eq. (35) by means of an integration by parts:

$$\iiint_R \vec{\zeta}\cdot\nabla\eta\ dx\ dy\ dz = \iint_C \eta\vec{\zeta}\cdot d\vec{S} - \iiint_R \eta\nabla\cdot\vec{\zeta}\ dx\ dy\ dz$$

$$= \iiint_R (\eta H_{q\vec{p}}\cdot\vec{\zeta} + H_{qq}\eta^2)dx\ dy\ dz \quad . \tag{39}$$

The surface integral vanishes if either (i) $\eta(C) = 0$ or (ii) $\vec{\zeta}\cdot dS = 0$. Substituting from Eq. (39) into Eq. (35), we find

$$A = A_* + \frac{1}{2}\iiint_R (H_{qq}\eta^2 - \vec{\zeta}\cdot H_{\vec{p}\vec{p}}\cdot\vec{\zeta})dx\ dy\ dz \quad . \tag{40}$$

If the tensor $H_{\vec{p}\vec{p}}$ is positive or negative definite and if H_{qq} has the opposite sign to it, then Eqs. (37) and (40) give an upper and a lower limit to A_*. Another condition under which this would be true would be if, say, H_{qq} were zero and $H_{\vec{p}\vec{p}}$ were either positive or negative definite. Acceptable boundary conditions for q and $\vec{p}$ are these: (1) either $q(C) = q_*(C)$ or $\vec{p}_*\cdot d\vec{S} = 0$ on C for the trial functions obeying the first Hamilton equation, Eq. (31d), $\nabla q = H_{\vec{p}}$, and (2) either $q(C) = q_*(C)$ or $\vec{p}_*\cdot d\vec{S} = \vec{p}\cdot d\vec{S} = 0$ on C for the trial functions obeying the second Hamilton equation, Eq. (31e), $-\nabla\cdot\vec{p} = H_q$.

Let us now return to the example we pursued in Sect. 6.7, namely, $\nabla^2 q = -1$, $q(C) = 0$. The Lagrangean is given in Eq. (29). According to Eqs. (31a) and (31b), $\vec{p} = \nabla q$ and $H = p^2/2 + q$. Thus the tensor $H_{\vec{p}\vec{p}}$ has 1 for its diagonal elements and zero for all others; it is therefore positive definite. Furthermore, $H_{qq} = 0$. So we expect the two estimates of A obtained from Eq. (32) by choosing trial values of $\vec{p}$ and q that satisfy one or the other of Hamilton's equations to be upper and lower bounds. Hamilton's equations are

$$\nabla q = H_{\vec{p}} = \vec{p} \quad , \tag{41a}$$

$$-\nabla\cdot\vec{p} = H_q = 1 \quad . \tag{41b}$$

If $\vec{p}$ and q obey Eq. (41a), then

$$A = \iiint_R \left[\frac{1}{2}(\nabla q)^2 - q\right] dx\ dy\ dz \ , \quad q(C) = 0 \quad . \tag{42}$$

Except for the boundary conditions $q(C) = 0$, q is completely arbitrary. If $\vec{p}$ obeys Eq. (41b), no restriction is placed on q. If we choose $q = 0$ so as to satisfy the requirement that $q(C) = q_*(C)$, then Eq. (32) becomes

$$A = -\frac{1}{2}\iiint p^2 dx\ dy\ dz \ , \quad \nabla\cdot\vec{p} = -1 \quad . \tag{43}$$

The same result can be obtained by choosing $q = q_*$; since q_* does not appear in Eq. (43) we do not actually have to know it to imagine $q = q_*$. Combining Eqs. (43), (42), and (30), we get

$$\iiint_R p^2 dx\ dy\ dz > \iiint_R (\nabla q_*)^2 dx\ dy\ dz > \iiint_R [2q - (\nabla q)^2] dx\ dy\ dz \quad , \qquad (44)$$

where $\nabla \cdot \vec{p} = -1$ and $q(C) = 0$ but $\vec{p}$ and q are otherwise arbitrary.

Suppose now that R is a thin square disk with corners $(\pm 1, \pm 1)$. A convenient trial function for q is $a(1-x^2)(1-y^2)$, where a is a constant yet to be determined. A short computation shows that the right-hand side of Eq. (44) is $(160a - 256a^2)/45$. The maximum value of this expression occurs when $a = 5/16$ and equals $5/9$, which is the best lower limit attainable with the family of trial functions chosen for q. A suitable trial function for p is the vector $(-x/2, -y/2)$, whose divergence is -1. A short computation then shows that the left-hand side of Eq. (44) is $2/3$, which is an upper limit. The geometric mean of these limits, $10/27 = 0.6086$, has a percentage difference from the exact value of no more than 9.5%. The exact value, 0.5623, can be calculated from a series given by Sikora.

The inequalities of Eq. (44) can be made the basis of a number of formulas for estimating $\iiint_R (\nabla q_*)^2\ dx\ dy\ dz$ for a variety of irregularly shaped two-dimensional disks. [See, for example, my paper "Eddy Current Heating of Irregularly Shaped Plates by Slow Ramped Fields," p. 89 in *Proceedings of the Eighth Symposium on Engineering Problems of Fusion Research, San Francisco, California, November 13–16, 1979*, IEEE, New York, 1979, and the references contained therein. This paper deals largely with means of choosing suitable trial functions and evaluating the multiple integrals on the left-hand and right-hand sides of Eq. (44).]

6.9 The foregoing section was devoted to an important but linear problem. This section is devoted to the nonlinear problem of Sect. 4.5, namely, steady heat flow in superfluid helium [see Eqs. (4.27) and (4.28)]. A Lagrangean for Eq. (4.28) is

$$L = \frac{3}{4}|\nabla T|^{4/3} \quad . \qquad (45)$$

(We assume here, as before, that K is independent of temperature. The factor $3/4$ has been inserted for convenience.) According to Eq. (31),

$$\vec{p} = (\nabla T)^{1/3} \equiv \nabla T / |\nabla T|^{2/3} \qquad (46)$$

and

$$H = \frac{1}{4}p^4 \quad . \qquad (47)$$

The Hamiltonian equations are then

$$\nabla T = \vec{p}^3 \ (\equiv p^2 \vec{p}) \quad , \qquad (48a)$$

$$\nabla \cdot \vec{p} = 0 \quad . \tag{48b}$$

The functional A is given by

$$A = \iiint_R \left(\vec{p} \cdot \nabla T - \frac{p^4}{4} \right) dx\ dy\ dz \quad . \tag{49}$$

In order to see if the method of complementary variational principles is of any use in this problem, we must identify the meaning of A_*, the exact value of A. Now

$$\begin{aligned} A_* &= \iiint_R (\vec{p}_* \cdot \nabla T_* - p_*^4/4)\ dx\ dy\ dz \\ &= \frac{3}{4} \iiint_R \vec{p}_* \cdot \nabla T_*\ dx\ dy\ dz \quad \text{[remember } \vec{p}_* \text{ and } T_* \text{ obey both Eqs. (48a) and (48b)]} \\ &= \frac{3}{4} \iiint_R \nabla \cdot (\vec{p}_* T_*)\ dx\ dy\ dz \qquad (\nabla \cdot \vec{p}_* = 0) \\ &= \frac{3}{4} \iint_C T_* \vec{p}_* \cdot d\vec{S} \quad . \end{aligned} \tag{50}$$

Suppose we now take R to be a duct with two plane parallel isothermal surfaces and two irregular adiabatic surfaces (see Fig. 2). From Eq. (46) or Eq. (48a) we see that $\vec{p}_*$ is parallel to the heat flux vector and is therefore parallel to the adiabatic surfaces. Therefore, on the adiabatic surfaces $\vec{p}_* \cdot d\vec{S} = 0$. Since $T_* = 0$ on the isothermal surface BD (T is the temperature *rise*),

$$A_* = \frac{3}{4}(\Delta T) \iint_{AC} (\nabla T_*)^{1/3} \cdot d\vec{S} = \frac{3}{4} \frac{\Delta T}{K} Q \quad , \tag{51}$$

where Q is the total heat flow into the face AC of the duct [N.B.: $(\nabla T_*)^{1/3}$ and $d\vec{S}$ are oppositely directed on AC.] So our variational principles will give us accurate bounds on the total heat flow through the duct, a quite useful quantity to have.

First, let us choose trial functions $\vec{T}$ and $\vec{p}$ obeying Eq. (48a). Then Eq. (49) becomes

$$A = \frac{3}{4} \iiint_R |\nabla T|^{4/3}\ dx\ dy\ dz \quad , \tag{52}$$

where the trial function T must obey the boundary conditions $T = \Delta T$ at $x = 0$ and $T = 0$ at $x = L$. It is easy to see that Eq. (52) will give the upper limit to A_*: since T is arbitrary except on the isothermal surfaces, we can add to it a high-frequency flutter that can make ∇T as large as we please.

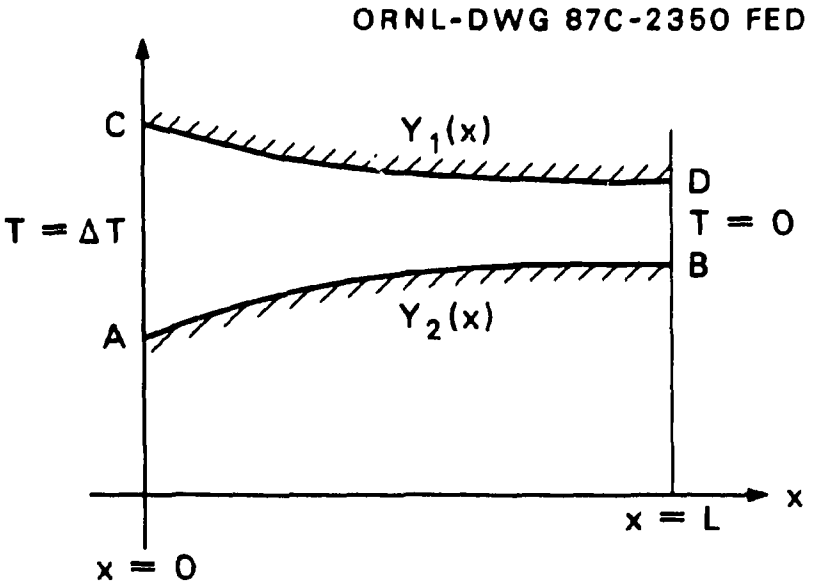

Fig. 2. The He-II-filled duct with isothermal surfaces $x = 0$ and $x = L$ and adiabatic surfaces $y = Y_1(x)$ and $y = Y_2(x)$.

Second, let us choose trial functions $\vec{p}$ and $\vec{T}$ obeying Eq. (48b). Since $\nabla \cdot \vec{p} = 0$, the first term in Eq. (49) can be converted to a surface integral, and A becomes

$$A = \iint_C T_* \vec{p} \cdot d\vec{S} - \frac{1}{4} \iiint_R p^4 \, dx \, dy \, dz \tag{53}$$

if we choose $T = T_*$. So finally we have

$$\iiint_R |\nabla T|^{4/3} \, dx \, dy \, dz \geq \frac{Q\Delta T}{K} \geq \frac{4}{3} \iint_C T_* \vec{p} \cdot d\vec{S} - \frac{1}{3} \iiint_R p^4 \, dx \, dy \, dz \ . \tag{54}$$

6.10 In this section and the next, we shall undertake the evaluation of the left- and right-hand sides of Eq. (54). Let us begin with the left-hand side, which is the easier of the two. Suppose we consider a unit width of duct in the z-direction and take the isothermal surfaces to be planes parallel to the end planes. This means we take $T = T(x)$. Then

$$\iiint_R |\nabla T|^{4/3} \, dx \, dy \, dz = \int_0^L \left(\frac{dT}{dx} \right)^{4/3} (Y_1 - Y_2) \, dx \ . \tag{55}$$

We choose the function $T(x)$ so as to minimize the right-hand side of Eq. (55). A straightforward variational calculation will give us the minimizing function; call it $T_0(x)$. If we set $T(x) = T_0(x) + \eta(x)$, then the *first-order* term in the expansion of Eq. (55) in powers of η is

$$\frac{4}{3}\int_0^L (Y_1 - Y_2)\left(\frac{dT_0}{dx}\right)^{1/3}\frac{d\eta}{dx}\,dx \quad . \tag{56}$$

If T_0 is to minimize the right-hand side of Eq. (55), Eq. (56) must vanish for all η. If we integrate by parts, Eq. (56) becomes

$$\frac{4}{3}(Y_1 - Y_2)\left(\frac{dT_0}{dx}\right)^{1/3}\eta\Big|_0^L - \frac{4}{3}\int_0^L \frac{d}{dx}\left[(Y_1 - Y_2)\left(\frac{dT_0}{dx}\right)^{1/3}\right]\eta\,dx \quad . \tag{57}$$

Now T, and perforce T_0, must obey the boundary conditions $T(0) = \Delta T$, $T(L) = 0$, so $\eta(0) = \eta(L) = 0$. Therefore, the integrated term vanishes. From the second term we see that T_0 must obey the Euler-Lagrange differential equation

$$\frac{d}{dx}\left[(Y_1 - Y_2)\left(\frac{dT_0}{dx}\right)^{1/3}\right] = 0 \quad . \tag{58}$$

Thus

$$\left(\frac{dT_0}{dx}\right)^{1/3} = \frac{B}{Y_1 - Y_2} \quad , \tag{59a}$$

where B is a constant of integration determined by

$$\Delta T = B^3 \int_0^L \frac{dx}{(Y_1 - Y_2)^3} \quad . \tag{59b}$$

Substituting Eqs. (59a) and (59b) into Eq. (55), we find for Eq. (55) the result

$$(\Delta T)^{4/3}\left[\int_0^L \frac{dx}{(Y_1 - Y_2)^3}\right]^{-1/3} \quad , \tag{60a}$$

so that from Eq. (54) we have

$$Q \le \frac{K(\Delta T)^{1/3}}{[\int_0^L dx/(Y_1 - Y_2)^3]^{1/3}} \quad . \tag{60b}$$

There is a "simple" derivation of Eq. (60b) that proceeds from the assumption that at every abscissa the temperature gradient is given by

$$\left(\frac{dT}{dx}\right)^{1/3} = \frac{Q}{K(Y_1 - Y_2)} \quad . \tag{61}$$

But this simple derivation does not show that the value of Q in Eq. (61) is an upper limit of variational accuracy, two things that are worth knowing.

6.11 Now we turn to the evaluation of the right-hand side of Eq. (54). Since $\vec{p}$ must be a divergenceless vector, let us set

$$p_x = \frac{\partial \psi}{\partial y} , \quad p_y = -\frac{\partial \psi}{\partial x} \quad . \tag{62}$$

In order to evaluate the first integral (the surface integral) we must know T_* on C. We only know it on the bounding isotherms, not on the lateral adiabats, i.e., not on $y = Y_1$ and $y = Y_2$. But if we take $y = Y_1$ and $y = Y_2$ to be level surfaces of ψ, then on them $\vec{p} \cdot d\vec{S}$ will be zero. Then

$$\frac{4}{3} \iint_C T_* \vec{p} \cdot dS = -\frac{4}{3}(\Delta T) \int_{Y_2}^{Y_1} \left(\frac{\partial \psi}{\partial y} \right) dy = -\frac{4}{3}(\Delta T)(\psi_1 - \psi_2) \quad . \tag{63}$$

The minus sign occurs in the last two terms because the outward normal to R on the end surface AC points in the negative x-direction; thus $dS_x = -dy$. So

$$\frac{Q \Delta T}{K} \geq -\frac{4}{3}(\Delta T)(\psi_1 - \psi_2) - \frac{1}{3} \iiint_R (\nabla \psi)^4 \, dx \, dy \, dz \quad . \tag{64}$$

Now in order that our trial functions may include the exact solution, we take

$$\psi_1 - \psi_2 = -\int_{\substack{Y_2 \\ (AC)}}^{Y_1} \vec{p}_* \cdot d\vec{S}_* = -\int_{AC} (\nabla T_*)^{1/3} \cdot d\vec{S}_* = -\frac{Q}{K} \quad . \tag{65}$$

Combining Eqs. (64) and (65), we get

$$\frac{Q \Delta T}{K} \leq \iiint_R (\nabla \psi)^4 \, dx \, dy \, dz \quad . \tag{66}$$

In spite of the direction of the inequality in Eq. (66), we shall ultimately get a lower limit to Q. This is because ψ also involves Q.

We choose as level surfaces* for the trial function ψ the surfaces

$$y = \lambda Y_1(x) + (1 - \lambda) Y_2(x) , \quad 0 \leq \lambda \leq 1 \quad . \tag{67}$$

*This procedure is called by Pólya and Szegö the method of assigned level surfaces.

The most convenient way to evaluate the integral in Eq. (66) is to introduce the new coordinates λ, x. Since the new coordinates are not Cartesian, we employ tensor formalism for the calculations:

$$(dx)^2 + (dy)^2 = (dx)^2 + \{(Y_1 - Y_2)d\lambda + [\lambda\dot{Y}_1 + (1-\lambda)\dot{Y}_2]dx\}^2 \quad ,$$

$$g_{xx} = 1 + [\lambda\dot{Y}_1 + (1-\lambda)\dot{Y}_2]^2 \quad ,$$

$$g_{x\lambda} = g_{\lambda x} = [\lambda\dot{Y}_1 + (1-\lambda)\dot{Y}_2](Y_1 - Y_2) \quad ,$$

$$g_{\lambda\lambda} = (Y_1 - Y_2)^2 \quad ,$$

$$g = \det(g_{ij}) = (Y_1 - Y_2)^2 \quad ,$$

$$g^{\lambda\lambda} = \frac{g_{xx}}{g} = \frac{1 + [\lambda\dot{Y}_1 + (1-\lambda)\dot{Y}_2]^2}{(Y_1 - Y_2)^2} \quad .$$

If ψ is a function only of λ,

$$(\nabla\psi)^4 = \left[g^{\lambda\lambda}\left(\frac{d\psi}{d\lambda}\right)^2\right]^2 \quad .$$

Since

$$\iint_R (\nabla\psi)^4 dx\ dy = \iint_R (\nabla\psi)^4 \sqrt{g}\ d\lambda\ dx \quad ,$$

we finally have

$$\iint_R (\nabla\psi)^4 dx\ dy = \int_0^1 d\lambda \left(\frac{d\psi}{d\lambda}\right)^4 \int_0^L dx\ \frac{\{1 + [\lambda\dot{Y}_1 + (1-\lambda)\dot{Y}_2]^2\}^2}{(Y_1 - Y_2)^3} \quad . \tag{68}$$

Equation (68) has the form

$$\int_0^1 d\lambda \left(\frac{d\psi}{d\lambda}\right)^4 G(\lambda) \quad , \tag{69}$$

where $G(\lambda)$ is $\int_0^L \cdots dx$. We shall choose ψ so as the maximize Eq. (69). A short variational calculation shows that ψ must obey the Euler-Lagrange differential equation

$$\frac{d}{d\lambda}\left[G(\lambda)\left(\frac{d\psi}{d\lambda}\right)^3\right] = 0 \quad . \tag{70}$$

The solution that obeys the boundary conditions $\psi_1 = -Q/K$, $\psi_2 = 0$ [see Eq. (65)] is

$$\psi = -\frac{Q}{K}\frac{\int_0^\lambda G^{-1/3} d\lambda}{\int_0^1 G^{-1/3} d\lambda} \quad . \tag{71}$$

Substituting Eq. (71) into Eq. (69), we find that Eq. (66) takes the form

$$\frac{Q\Delta T}{K} \leq \left(\frac{Q}{K}\right)^4 \left(\int_0^1 G^{-1/3} d\lambda\right)^{-3} \tag{72a}$$

or

$$Q \geq K(\Delta T)^{1/3} \int_0^1 G^{-1/3} d\lambda \ , \tag{72b}$$

where

$$G(\lambda) = \int_0^L \frac{\{1 + [\lambda \dot{Y}_1 + (1-\lambda)\dot{Y}_2]^2\}^2}{(Y_1 - Y_2)^3}\, dx \ . \tag{72c}$$

The function G is simple to evaluate when the adiabatic surfaces are straight lines, i.e., when R is a trapezoid. By way of example, consider the trapezoid shown in Fig. 3, for which $\dot{Y}_1 = -a$ and $\dot{Y}_2 = 0$. In this case, Eq. (72b) becomes

$$Q \geq \frac{K(\Delta T)^{1/3}}{[\int_0^L dx/(Y_1 - Y_2)^3]^{1/3}} \cdot \int_0^1 (1 + \lambda^2 a^2)^{-2/3}\, d\lambda \ . \tag{73}$$

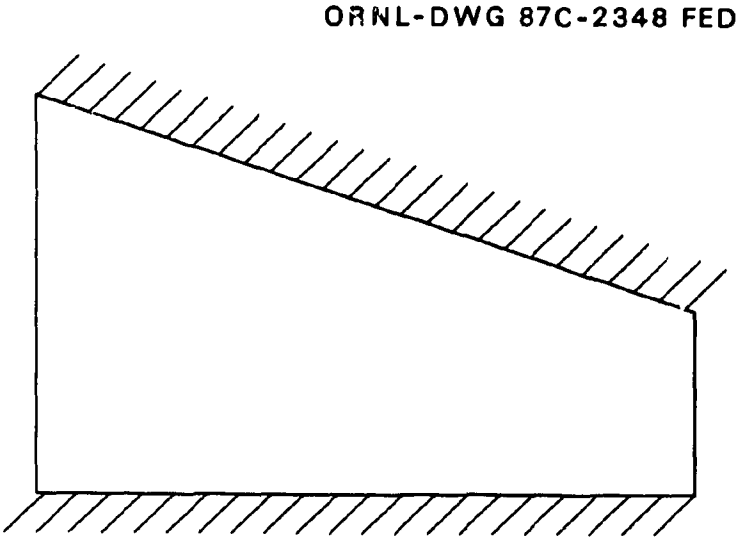

Fig. 3. A trapezoidal duct. The hatched surfaces are adiabatic.

Comparing Eq. (73) with Eq. (60b), we see that the λ-integral in Eq. (73) gives the ratio of the upper and lower variational estimates of Q. The λ-integral is easy to evaluate either by series or with Simpson's rule. A few values are given in Table 2.

Table 2. Values of the integral $\int_0^1 (1+\lambda^2 a^2)^{-2/3}\, d\lambda$

a	$\int_0^1 (1+\lambda^2 a^2)^{-2/3}\, d\lambda$
0.0	1.000000
0.1	0.997789
0.3	0.980852
0.5	0.950452
0.7	0.911607
1.0	0.847138
1.5	0.743754
2.0	0.656516

These numerical values show that, even for substantial slopes, the two bracketing estimates are quite close together.

6.12 The variational method is not without use even when the desired quantity is not the one represented by the functional A. Consider, for example, the problem dealt with at the end of Sect. 4.5, namely: $\nabla\cdot(\nabla T)^{1/3} = -1$; $T = 0$ on the perimeter P of a square S of side 2; find T at the center of the square. The Lagrangean for this problem is

$$L = \frac{3}{4}|\nabla T|^{4/3} - T \quad , \tag{74}$$

and the extreme value of the functional A is

$$A_* = \iint_S \left(\frac{3}{4}|\nabla T_*|^{4/3} - T_*\right) dx\, dy \tag{75a}$$

$$= \iint_S \left\{\frac{3}{4}\nabla\cdot[T_*(\nabla T_*)^{1/3}] - \frac{1}{4}T_*\right\} dx\, dy \tag{75b}$$

$$= \frac{3}{4}\int_P T_*(\nabla T_*)^{1/3}\cdot d\vec{S} - \frac{1}{4}\iint_S T_*\, dx\, dy \tag{75c}$$

$$= -\frac{1}{4}\iint_S T_*\, dx\, dy \tag{75d}$$

since $T_* = 0$ on the perimeter P. {It also follows that $A_* = -\frac{1}{4}\iint_S |\nabla T_*|^{4/3} dx\, dy$ [equate Eqs. (75a) and (75d)].} So if we wanted the average value of T in S, the variational method could give us a second-order estimate of it. But we are interested in T at the center of the square, for which no such second-order estimate is possible.

The variational method can help us to find the "best" trial function of a chosen family. The word "best" is in quotes because the trial function is best only in the sense of making A as close to its extreme value A_s as possible, but in actual fact this is achieved by making the trial function resemble the exact solution as much as possible. In the problem being considered we can again make use of Pólya and Szegö's method of assigned level surfaces. Let us choose the center of the square as the origin of polar coordinates (r, θ) in terms of which the perimeter of the square is given by $r = R(\theta)$. Let us choose as the level surfaces of T the surfaces $r = \lambda R(\theta)$, $0 < \lambda < 1$, that are geometrically similar to P. Since we are taking $T = T(\lambda)$ only, it is convenient to introduce the nonorthogonal coordinates (λ, θ) in place of the coordinates (r, θ). Proceeding as before, we evaluate A:

$$dr^2 + r^2 d\theta^2 = (R\, d\lambda + \lambda \dot{R}\, d\theta)^2 + \lambda^2 R^2\, d\theta^2$$

$$g_{\lambda\lambda} = R^2 \qquad g = \det(g_{ij}) = \lambda^2 R^4$$

$$g_{\lambda\theta} = g_{\theta\lambda} = \lambda R \dot{R} \qquad g^{\lambda\lambda} = \frac{g_{\theta\theta}}{g} = \frac{R^2 + \dot{R}^2}{R^4}$$

$$g_{\theta\theta} = \lambda^2 (R^2 + \dot{R}^2) \qquad (\nabla T)^2 = g^{\lambda\lambda} \left(\frac{dT}{d\lambda}\right)^2 = \frac{R^2 + \dot{R}^2}{R^4} \left(\frac{dT}{d\lambda}\right)^2$$

Thus

$$A = \iint_S \left[\frac{3}{4}\left(\frac{R^2 + \dot{R}^2}{R^4}\right)^{2/3} \left(\frac{dT}{d\lambda}\right)^{4/3} - T\right] \lambda R^2\, d\lambda\, d\theta$$

$$= \int_0^1 \lambda\, d\lambda \int_0^{2\pi} d\theta \left[\frac{3}{4}\left(\frac{R^2 + \dot{R}^2}{R}\right)^{2/3} \left(\frac{dT}{d\lambda}\right)^{4/3} - R^2 T\right]$$

$$= \int_0^1 \lambda\, d\lambda \left[\frac{3}{4} a \left(\frac{dT}{d\lambda}\right)^{4/3} - 2A_s T\right] , \tag{76}$$

where

$$a = \int_0^{2\pi} \left(\frac{R^2 + \dot{R}^2}{R}\right)^{2/3} d\theta \tag{77a}$$

and

$$A_s = \frac{1}{2} \int_0^{2\pi} R^2\, d\theta \tag{77b}$$

is the area of the square S.

We choose $T(\lambda)$ to minimize Eq. (76). A short variational calculation shows that T must satisfy the Euler-Lagrange equation,

$$a\frac{d}{d\lambda}\left[\lambda\left(\frac{dT}{d\lambda}\right)^{1/3}\right]+2A_s\lambda=0\ . \tag{78}$$

The solution (78) that is regular at $\lambda = 0$ and obeys the boundary condition $T(1) = 0$ at $\lambda = 1$ is

$$T=\frac{1}{4}\left(\frac{A_s}{a}\right)^3(1-\lambda^4)\ . \tag{79}$$

From Eq. (79) it follows that

$$T(0)=\frac{1}{4}\left(\frac{A_s}{a}\right)^3 \tag{80a}$$

and

$$\langle T\rangle=\frac{1}{A_s}\iint_S T\,dx\,dy=\frac{1}{6}\left(\frac{A_s}{a}\right)^3\ . \tag{80b}$$

Equation (80b) is accurate to second order; Eq. (80a) is not. The results [Eqs. (80a) and (80b)] apply to any geometric figure. For the square of side 2 an easy calculation shows that $a = 8$ and $A_s = 4$. Thus, we estimate that $T(0) = 1/32$ and $\langle T\rangle = 1/48$.

6.13 When the quantity we are interested in is represented by the functional A and the Hamiltonian has certain properties, we can get rigorous upper and lower bounds for A_*. We are dealing with exact mathematics and we know by how much at most our estimates can be wrong. But when the quantity we are interested in is not the one represented by the functional A, as in the previous section, we can rigorously say little that is useful about our estimate of it. So although we may feel that the estimate $T(0) = 1/32$ is reasonably accurate, there is nothing in the analysis that led to it that can help us quantify this feeling.

Nevertheless, the approach used in Sect. 6.12 and extensions of it mentioned below can be used to get estimates of various quantities that, though they might not satisfy a mathematician, might well satisfy an engineer. I call these methods curve-fitting methods and their common features are these: a family of curves of some generality is chosen to represent the solution of the differential equation, and then the best curve is picked out according to some criterion of best fit. In Sect. 6.12 the criterion of best fit was the variational criterion—the best curve is the one that makes the functional A an extremum. But other criteria are possible. Some deal with the residual, the amount by which the sum of all the terms in the differential equation misses being zero. In the method of collocation, the parameters of the best member of a multiparameter family of trial functions are determined by requiring the residual to vanish at several discrete points. In the method of least squared residual, the parameters are chosen to minimize the integrated squared residual, possibly multiplied by some weighting function. In the Galerkin method, used mainly for linear problems, the trial solution is written as a finite sum of orthogonal functions and the expansion coefficients are chosen to make the residual orthogonal

to all those functions used in the sum (thus we solve the differential equation in the subspace spanned by the trial functions). A method that I personally like is the integral method, in which integral relations constraining the solution are obtained by multiplying the differential equation by various functions and integrating it over the interval of interest. The success of these methods depends more on the choice of trial family than on the criterion of best fit, and generally the latter can be chosen to minimize the labor of calculation.

As an example of these ideas, let us consider the differential equation (3.26) and the boundary conditions (3.27a) and (3.27b):

$$3\ddot{y} + xy\dot{y} - y^2 = 0 \quad , \tag{3.26}$$

$$\dot{y}(0) = -1 \ , \quad y(\infty) = 0 \quad . \tag{3.27}$$

This example is especially interesting because it cannot be written in the Lagrangean form. Now, we know from the analysis of Sect. 3.6 that for large x, $y \sim 6/x^2$. A simple, one-parameter trial family that has this behavior and for which $\dot{y}(0) = -1$ is

$$y = (a + a^2 x + x^2/6)^{-1} \quad . \tag{81}$$

If we insert Eq. (81) into Eq. (3.26), we find the residual

$$\frac{6a^4 + a^2 x - 2a}{(a + a^2 x + x^2/6)^3} \quad . \tag{82}$$

If we require the residual to vanish at $x = 0$, we find $a = 3^{-1/3} = 0.6934$. On the other hand, if we require it to vanish when $x = 1$, we find $a = 0.6136$. If we require a to minimize the integrated squared residual (this integral was done numerically), we find $a = 0.6738$.

The integral method can be applied to Eq. (3.26) by integrating it over the entire interval from zero to infinity. After integrating the middle term by parts we find

$$\int_0^\infty y^2 \, dx + 2\dot{y}(0) = 0 \quad . \tag{83}$$

Substituting Eq. (81) into Eq. (82) and carrying out the indicated integration, we find the following equation for a:

$$\frac{2}{3} \frac{\arccos\sqrt{3a^3/2}}{(2a/3 - a^4)^{3/2}} - \frac{a}{(2a/3 - a^4)} - 2 = 0 \quad . \tag{84}$$

Equation (84) can be solved without too much effort by the Newton-Raphson method and yields $a = 0.6468$.

The result obtained in Sect. 3.6 by numerical integration of the differential equation was $a = (1.511)^{-1} = 0.6618$. The results obtained above are all within 8% of the correct value. Had we not done the numerical integration of the differential equation, these results would suggest to us that the correct value is likely to lie between 0.6 and 0.7. But it should constantly be borne in mind that these results have no rigorous significance. Regarding nonvariational curve-fitting methods, all I can say is let the user beware.

Chapter 7
STABILITY OF NUMERICAL METHODS

"There is nothing stable in the world; uproar's your only music."
—John Keats
Letters

7.1 We have already seen in earlier chapters of this book how a brute-force numerical approach to certain problems involving ordinary differential equations runs into difficulties because of numerical instability (see e.g., Sects. 2.6, 3.6, and 4.3). By instability, I mean runaway departure of the numerically calculated values from the correct solution. The cause of instability in all of these cases was the divergence of neighboring integral curves from one another and from the one we were trying to calculate (see Fig. 4.1). When we tried to advance in the direction of the divergence, the unavoidable small errors of truncation in the numerical procedure threw us off the curve we were trying to calculate onto a near neighbor. Because the integral curves diverge, the numerical solution departed by ever greater amounts from the solution we were trying to calculate, and the numerical solution eventually became worthless. Figures 2.2 and 3.2 show this clearly. A similar thing occurs in the development of "chaos," about which much has been written lately; there, as here, the problem is caused by a very sensitive dependence of the asymptotic behavior on the initial conditions.

When the cause of instability is seen clearly, one realizes that there is no way of finessing a solution marching in the direction of divergence. But, as we have already seen in the examples cited above, numerical integration in the opposite direction is quite successful. All of those examples were two-point boundary-value problems on a semi-infinite interval. In all of them, an asymptotic limit was used to find consistent values of y and $\dot{y}$ for some large value of x that then served as starting values for a stable integration in the backward x-direction. In the examples of Sects. 2.6 and 3.6, we made explicit use of the affine group invariance of the differential equation, but in the example of Sect. 4.3, we deliberately considered a differential equation (4.11a) not invariant to an affine group. There, we postulated the asymptotic series (4.12a) and determined the coefficients A, B, C, etc., by substituting into the differential equation and equating the coefficients of individual powers of x to zero.

In general, the last approach will prove satisfactory, but it must be handled with some delicacy, as the following illustration based on Eq. (3.26) shows. Suppose we want to handle the two-point boundary-value problem of Eqs. (3.26), (3.27a), and (3.27b) without invoking invariance to an affine group. A little numerical trial and error convinces us that forward integration is unstable (try it!). So we look for an asymptotic series with which to start a backward integration. Substitution of the trial form $y \sim A/x^m$ into Eq. (3.26) gives

$$\frac{3m(m+1)A}{x^{m+2}} - \frac{mA^2}{x^{2m}} - \frac{A^2}{x^{2m}} = 0 \quad , \tag{1}$$

which can be satisfied if $2m = m + 2$ and $A = 3m$, i.e., if $m = 2$ and $A = 6$. Thus we find the special solution $6/x^2$. It is tempting at this point to again postulate the asymptotic form, Eq. (4.12a), but a quick calculation shows that all the coefficients of the higher powers, A, B, C, etc., must vanish. This leaves us in a quandary.

The reason for our difficulty is that the form in Eq. (4.12a) assumes too much about the solution to Eq. (3.26). If we assume less at the outset, we fare better. Suppose we assume

$$y \sim \frac{6}{x^2} + \frac{A}{x^m} + \frac{B}{x^n} + \cdots \quad , \tag{2}$$

where $2 < m < n$. Substitution of Eq. (2) into Eq. (3.26) gives

$$(3m^2 - 3m - 24)Ax^{-(m+2)} + (3n^2 - 3n - 24)Bx^{-(n+2)} - (m+1)A^2x^{-2m} - (n+m+2)ABx^{-(m+n)} - (n+1)Bx^{-2n} + \cdots = 0 \quad . \tag{3}$$

Since we do not want A to vanish, we must choose m to be the positive root of $3m^2 - 3m - 24 = 0$, namely, $m = (\sqrt{33}+1)/2 = 3.372$. Since $n > m$, $3n^2 - 3n - 24 \neq 0$; thus B must vanish, unless $n + 2 = 2m$, in which case

$$(3n^2 - 3n - 24)B = (m+1)A^2 \quad . \tag{4}$$

If we add additional terms to Eq. (2) at the start, we can continue in this way, but the calculations are tedious. What we have is sufficient, namely,

$$y \sim \frac{6}{x^2} + \frac{A}{x^m} + \frac{B}{x^{2m-2}} + \cdots \quad , \tag{5a}$$

$$m = (\sqrt{33}+1)/2 \quad , \tag{5b}$$

$$B = \frac{\sqrt{33}+3}{6(27 - 3\sqrt{33})}A^2 \quad . \tag{5c}$$

We expect that different values of A will correspond to different slopes at the origin. Equation (3.27) directs our attention to the curve for which $\dot{y}(0) = -1$. To find the corresponding value of A we use trial and error, improving our guesses with the Newton-Raphson method. If we define $1 + \dot{y}(0) = f(A)$, then

$$A' = A - \frac{hf(A)}{f(A+h) - f(A)} \tag{6}$$

is the Newton-Raphson iterative procedure for finding the root of $f(A) = 0$. Table 1 shows the actual work. The first four trials were guesswork to locate the root approximately. Thereafter, the Newton-Raphson method was used to accelerate convergence. Two initial values of x were used to demonstrate that the final result did not depend on its particular value (as long as it was large enough). Each line represents a numerical integration, carried out by the fourth-order Runge-Kutta method on a time-share VAX 8600 in a couple of seconds. The final result, $y(0) =$

1.511, is the same as that obtained in Sect. 3.6 at the cost of a single numerical integration.

Table 1. Trial and error solution for A using the Newton-Raphson method

x	A	$f(A)$	h	$f(A+h)$	$y(0)$
10.0	−1.0	−1490			
	−0.10	−228952			
	−10.0	−8.850			
	−30.0	−0.021214	−0.01	−0.020585	
	−30.3373	$-0.0^3383800$	−0.0001	$-0.0^3377737$	
	−30.343630	-0.0^6149			1.511171
20.0	−30.0	0.103469	−1.0	0.164363	
	−28.30	−0.016270	−0.01	−0.015498	
	−28.5108	$-0.0^3176209$	−0.0001	$-0.0^3168664$	
	−28.513135	-0.0^7626			1.511176

7.2 Instability of another kind sometimes occurs when we try to solve partial differential equations. Consider, for a moment, the pulsed-source problem in an infinite medium for the ordinary diffusion equation: $C_t = C_{zz}$; $C(\pm\infty, t) = 0$; $C(z, 0) = 0$; $\int_{-\infty}^{\infty} C dz = 1$, $t > 0$. This problem is useful for discussion because it has the known solution $C(z,t) = \exp(-z^2/4t)/(4\pi t)^{1/2}$. To solve it numerically we might use the finite-difference representation

$$\frac{C(z,t+k) - C(z,t)}{k} = \frac{C(z-h,t) - 2C(z,t) + C(z+h,t)}{h^2} \tag{7a}$$

or

$$\frac{C_{n,m+1} - C_{n,m}}{k} = \frac{C_{n-1,m} - 2C_{n,m} + C_{n+1,m}}{h^2} \ , \tag{7b}$$

where $C_{n,m}$ is an abbreviation for $C(z = nh, t = mk)$. Equation (7b) can easily be solved for $C_{n,m+1}$:

$$C_{n,m+1} = C_{n,m} + \frac{k}{h^2}(C_{n-1,m} - 2C_{n,m} + C_{n+1,m}) \ . \tag{7c}$$

It is easy to see that Eq. (7c) allows computation of the C values at the next time step, $C_{n,m+1}$, from the C values at the present time step $C_{n,m}$. Shown in Figs. 1a–1d are the results of such a computation. Figure 1a shows the initial condition calculated from the known analytic solution for $t = 0.25$. The space step h has been taken to be 0.01, the time step k to be 10^{-4}. Figures 1b, 1c, and 1d show the calculated profiles of C after 33, 35, and 37 time steps, respectively. As the reader can see, an oscillatory disturbance appears and grows rapidly, eventually destroying any information we hope to gain from the numerical integration.

Armed with the results of this numerical experiment, we might now guess that the difference equation (7c) has a solution of the form

$$C_{n,m} = (-)^n e_m \ . \tag{8}$$

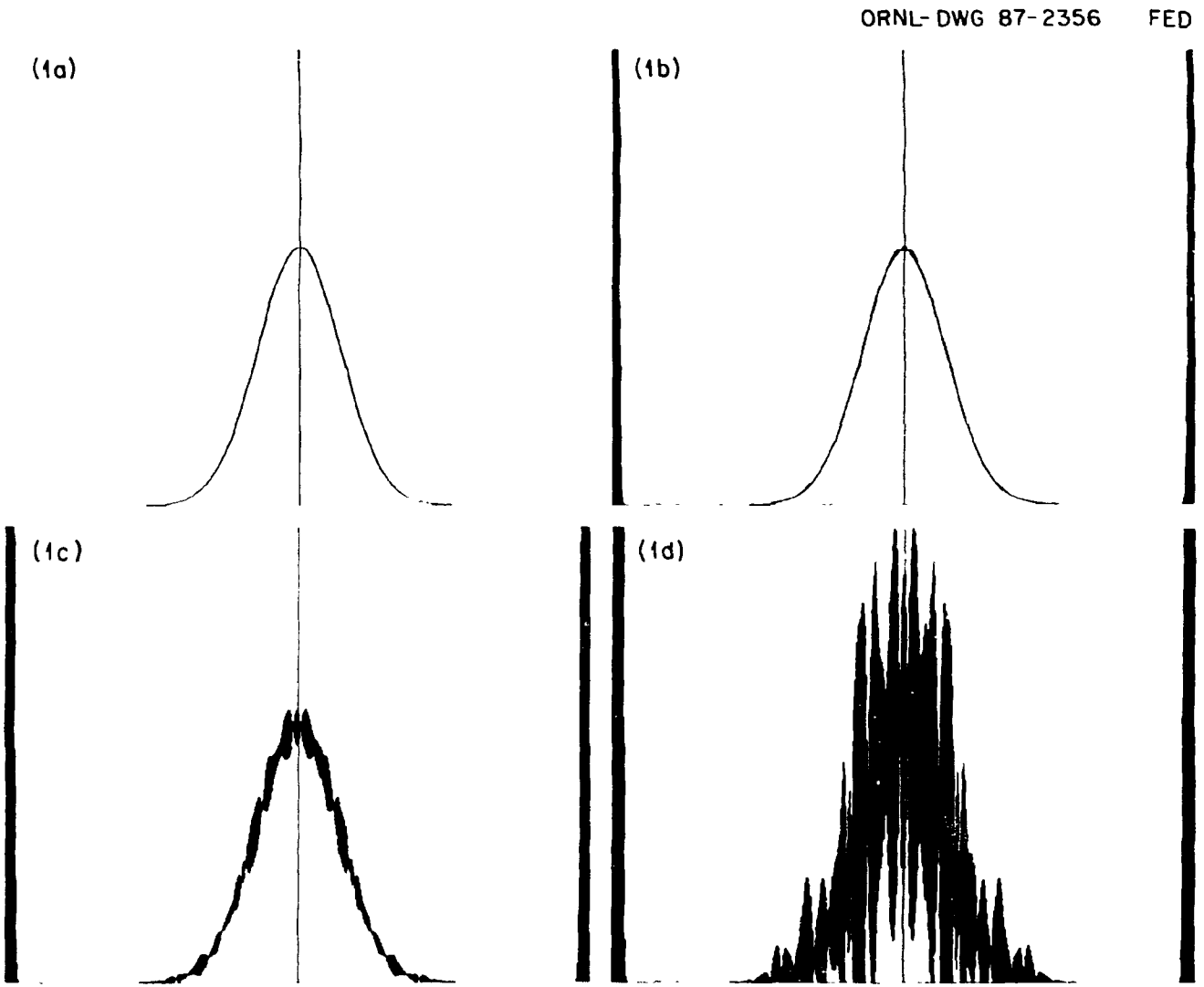

Fig. 1. Numerical integration of the ordinary diffuson equation $C_t = C_{zz}$. (a) Initial condition at $t = 0.25$; (b) after 33 steps with $h = 0.01$ and $k = 10^{-4}$; (c) after 35 steps; (d) after 37 steps. The z-axis contains 1000 space points.

Here the factor $(-)^n$ provides the rapid fluctuation from point to point that is evident in the numerical calculations. Substituting Eq. (8) into Eq. (7c), we find at once that

$$e_{m+1} = e_m \left(1 - \frac{4k}{h^2}\right) \quad . \tag{9}$$

From Eq. (9) we see that if $|1 - (4k/h^2)| > 1$, the e_m will diverge exponentially, whereas if $|1 - (4k/h^2)| < 1$, they will tend toward zero. Thus, if

$$1 > 1 - \frac{4k}{h^2} > -1 \quad , \tag{10a}$$

the e_m tend to zero, whereas otherwise they diverge exponentially. Now Eq. (10a) is equivalent to

$$\frac{h^2}{2} > k > 0 \quad . \tag{10b}$$

The integration that led to Figs. 1a–1d had $k > h^2/2$, so it is now understandable that it became unstable. Reducing the time step by a factor of $1/\sqrt{2}$ or more cures the instability, i.e., prevents the appearance of *unbounded* fluctuations. Bounded fluctuations still occur. They originate from the inadequacy of the finite-difference scheme accurately to represent the solution of the partial differential equation (truncation error) and from the finite-decimal representation of numbers in the computer. If the time step is chosen to satisfy Eq. (10b), and if the errors just mentioned are initially small, they will remain small and not trouble us.

The restriction of the time step expressed by Eq. (10b) is inconvenient because it demands small time steps, and small time steps mean long computing times. This restriction was quite serious in the distant past, when the calculations were done by hand, or even in the recent past, when computers were slow. But with today's fast mainframe computers, the restriction of the time step is not so important. There are finite-difference methods, the so-called implicit methods, that are stable for all values of k/h^2. However, they involve the solution of large (but sparse) matrices, which complicates their programming and slows down their running. They are nonetheless worth a moment's consideration.

Suppose on the right-hand side of Eq. (7) we estimate the second space derivative C_{zz} using the values of C at time $t + k$. Then Eq. (7c) would become

$$C_{n,m+1} = C_{n,m} + \frac{k}{h^2}(C_{n-1,m+1} - 2C_{n,m+1} + C_{n+1,m+1}) \quad . \tag{11}$$

If we now substitute the trial solution, Eq. (8), into Eq. (11), we find

$$e_{m+1} = e_m \Big/ \left(1 + \frac{4k}{h^2}\right) \quad . \tag{12}$$

Thus, no matter what the value of k/h^2, the e_m *never* become unbounded. The equations (11) are a coupled set of linear equations for the $C_{n,m+1}$ which require some small labor to solve. They are linear because the underlying partial differential equation is linear. When the underlying partial differential equation is nonlinear,

the labor of solving these equations can be immense and the implicit method may lose its utility.

7.3 The stability condition, Eq. (10b), between k and h applies for the ordinary diffusion equation, and other conditions may apply for other partial differential equations. Consider, for example, the wave equation $C_{xx} = C_{tt}$. A simple finite-difference approximation to it is

$$\frac{C_{n,m+1} - 2C_{n,m} + C_{n,m-1}}{k^2} = \frac{C_{n+1,m} - 2C_{n,m} + C_{n-1,m}}{h^2} \tag{13a}$$

or

$$C_{n,m+1} = 2C_{n,m} - C_{n,m-1} + \frac{k^2}{h^2}(C_{n+1,m} - 2C_{n,m} + C_{n-1,m}) \ . \tag{13b}$$

If we substitute Eq. (8) into Eq. (13b), we get

$$e_{m+1} = \left(2 - \frac{4k^2}{h^2}\right) e_m - e_{m-1} \tag{14a}$$

$$= Ae_m - e_{m-1} \ , \quad \text{where} \quad A = \left(2 - \frac{4k^2}{h^2}\right) \ . \tag{14b}$$

Equation (14) is a linear difference equation of the second order and has therefore two linearly independent solutions of the form $e_m = Be^{km}$, where k is a root of the equation

$$e^k = A - e^{-k} \quad \text{or} \quad (e^k)^2 - Ae^k + 1 = 0 \ . \tag{15}$$

Now A cannot exceed 2. If $2 > A > -2$, then $e^k = (A \pm i\sqrt{4 - A^2})/2$, the modulus of which is unity. Thus k is pure imaginary and equals $i\theta$, where $\theta = \cos^{-1}(A/2)$. Then

$$e_m = \mathrm{Re}(B_+ e^{im\theta} + B_- e^{-im\theta}) \tag{16}$$

is the general solution of Eq. (14). The modulus of e_m never becomes different in order of magnitude from that of e_0. The e_m do not become unbounded and we have stability.

If $A < -2$, $e^k = (A \pm \sqrt{A^2 - 4})/2$. The root with the minus sign has a modulus larger than 1; the root with the plus sign has a modulus smaller than 1. For large m, the larger root dominates, so that eventually

$$\frac{e_{m+1}}{e_m} \sim \frac{A - \sqrt{A^2 - 4}}{2} \ , \quad A < -2 \ . \tag{17}$$

The right-hand side of Eq. (17) is negative and has a modulus > 1, so the e_m fluctuate in sign and grow in magnitude without bound. This means there is instability for $A < -2$.

When $2 > A > -2$, $0 < k^2/h^2 < 1$, so the condition for stability for the wave equation is $k < h$, which allows much more generous time steps than the ordinary diffusion equation.

The stability criteria derived in this section and the last are necessary criteria. They are also sufficient, but this is more difficult to prove. They refer, of course, to particular finite-difference representations of the underlying partial differential equation.

7.4 The partial differential equation

$$C_t = (C_z^{1/3})_z \tag{18}$$

arises in the study of transient heat transfer in superfluid helium (He-II); see Sect. 4.5. A simple finite-difference representation of Eq. (18) is

$$\frac{C_{n,m+1} - C_{n,m}}{k} = \frac{1}{h^{4/3}}[(C_{n+1,m} - C_{n,m})^{1/3} - (C_{n,m} - C_{n-1,m})^{1/3}] \tag{19a}$$

or

$$C_{n,m+1} = C_{n,m} + \frac{k}{h^{4/3}}[(C_{n+1,m} - C_{n,m})^{1/3} - (C_{n,m} - C_{n-1,m})^{1/3}] \ . \tag{19b}$$

If we substitute Eq. (8) into Eq. (19b) we get

$$e_{m+1} = e_m - \frac{2^{4/3}k}{h^{4/3}} e_m^{1/3} \ . \tag{20}$$

A little numerical experimentation with Eq. (19) shows that it has as a solution a two-cycle, which turns out to be given by

$$e_m = \pm(-)^m \frac{\sqrt{2}k^{3/2}}{h^2} \tag{21}$$

for all values of $k/h^{4/3}$. The reader can verify Eq. (21) by substitution into Eq. (20). From this we might expect that solutions of Eq. (18) will be perturbed by high-frequency fluctuations of the constant amplitude given by Eq. (21).

To test this I performed calculations of the infinite-medium, pulsed-source problem for the partial differential equation (18). I chose this problem because it has the known exact solution

$$C = t^{-3/2} y(x) \ , \tag{22a}$$

$$x = z/t^{3/2} \ , \tag{22b}$$

$$y = \frac{4/3\sqrt{3}}{(x^4 + b^4)^{1/2}} \ , \tag{22c}$$

$$b = \frac{2\left[\Gamma\left(\frac{1}{4}\right)\right]^2}{3\sqrt{3\pi}} = 2.854535\ldots \ . \tag{22d}$$

(This similarity solution was obtained using the techniques of Chap. 3. The details can be found in the author's book mentioned in Sect. 3.8.) Shown in Fig. 2 is the

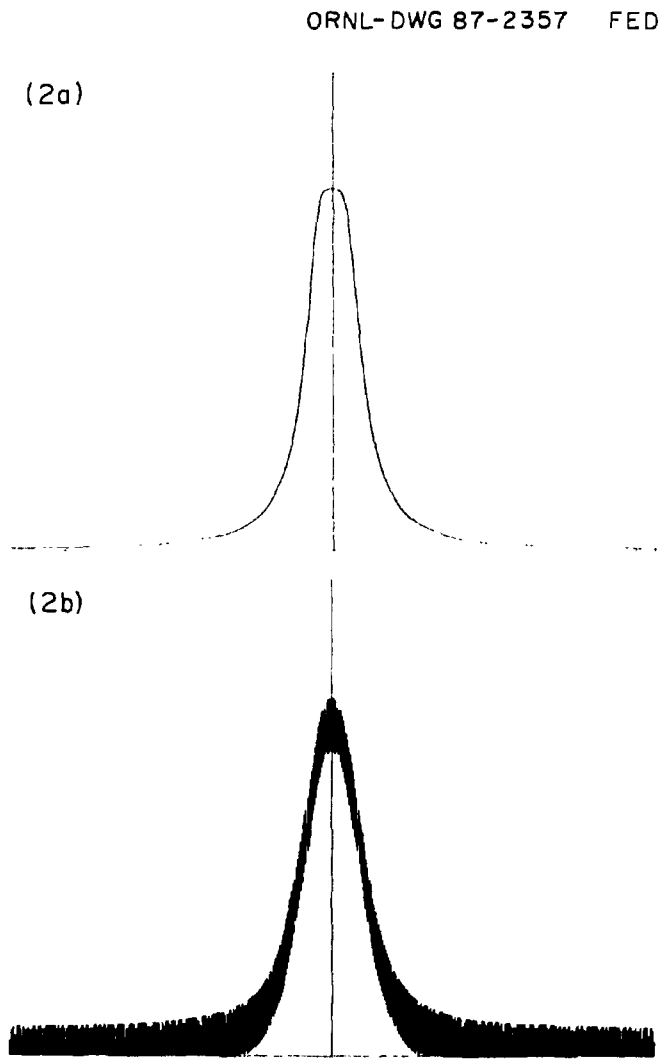

Fig. 2. Numerical integration of Eq. (18). (a) Initial condition at $t = 0.25$; (b) after 1000 steps with $k = 2.5 \times 10^{-4}$. Here $h = 0.01$ and the space axis has 1000 points stretching from $z = -5$ to $z = +5$.

initial condition (22) for $t = 0.25$ and a numerically calculated value for $t = 0.50$ (determined by 1000 time steps with $k = 2.5 \times 10^{-4}$; here $h = 0.01$ and the space axis contains 1000 points stretching from $z = -5$ to $z = +5$). The amplitude of the oscillations agrees perfectly with the value ± 0.05590 given in Eq. (21). But the curve itself does not agree at all with what we expect from Eq. (22). For example, $C(0,t)$ at $t = 0.5$ should be 0.267210, which is less than half the value given by curve (b).

It appears, then, that the oscillations destroy the utility of the numerical integration. It is not hard to see why. If we add a fluctuating quantity to the C's in the differences on the right-hand side of Eq. (19), we can seriously distort the value of the differences, especially when the fluctuating quantity is large compared with the true value of the difference. This reasoning implies, on the other hand, that if we make e_m small enough, by making the time step small enough, the numerical scheme should give the right answer.

To test this last supposition, I performed a second set of calculations going from $t = 0.25$ to $t = 0.50$ but this time with $h = 0.025$, $k = 2.5 \times 10^{-5}$, and 10^4 time steps (now z stretches from -12.5 to $+12.5$). Now the amplitude of fluctuations is only $\pm 2.828 \times 10^{-4}$, 200 times smaller than in the first case. The results are shown in Fig. 3. Included in Fig. 3 is the exact result for $t = 0.50$ calculated from Eq. (22). The agreement between the numerically calculated result (b) and the exact result (c) is very good, but some small discrepancies persist, e.g., the flattening in the wings of the numerically calculated curve.

In practical computations, in which no exact solution is available for comparison, one should calculate over and over again with smaller and smaller time steps until good convergence is achieved.

Finally, a stable, implicit finite difference scheme for integrating Eq. (18) can be based on its representation in the form $C_{zz} = 3q^2 C_t$, $q = C_z^{1/3}$, namely

$$\frac{C_{n+1,m+1} - 2C_{n,m+1} + C_{n-1,m+1}}{h^2} = 3q_{n,m}^2 \frac{C_{n,m+1} - C_{n,m}}{k} \quad , \tag{23a}$$

$$q_{n,m} = \left(\frac{C_{n+1,m} - C_{n-1,m}}{2h} \right)^{1/3} \quad . \tag{23b}$$

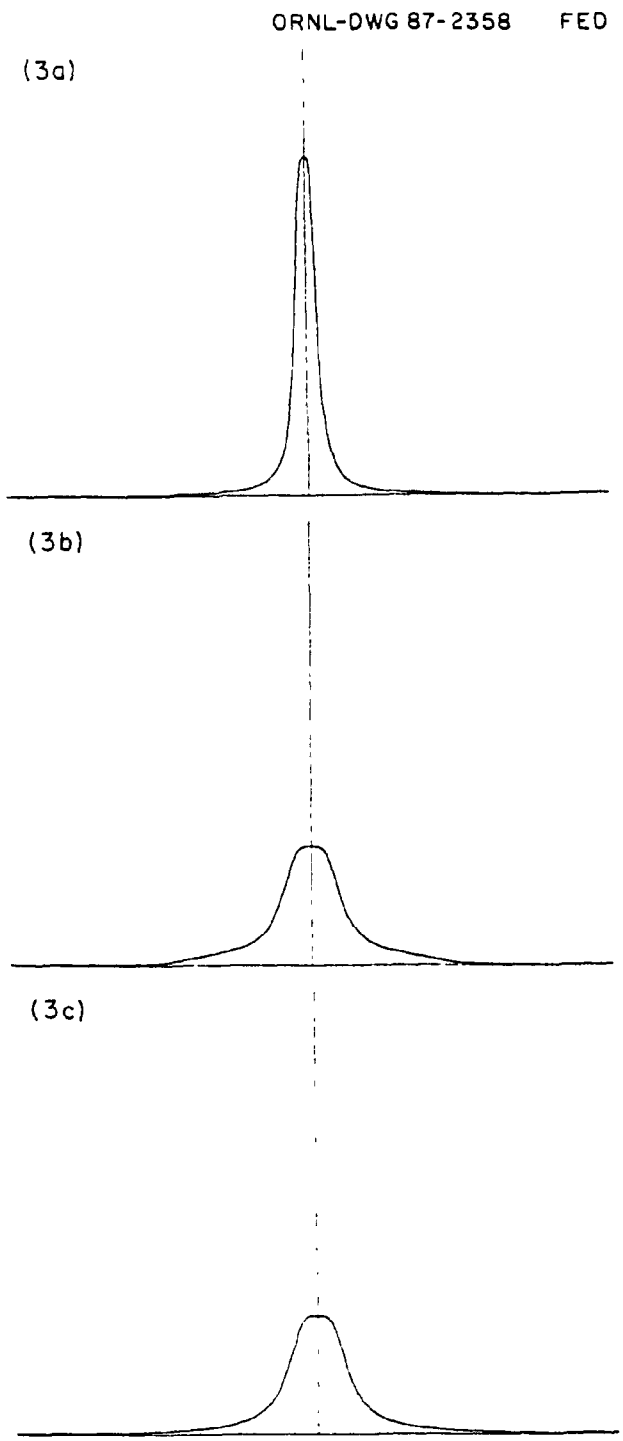

Fig. 3. Numerical integration of Eq. (18). (a) Initial condition at $t = 0.25$. (b) After 10^4 time steps with $k = 2.5 \times 10^{-5}$. Here $h = 0.025$ and the space axis has 1000 points stretching from $z = -12.5$ to $z = +12.5$. (c) Exact result for $t = 0.5$ calculated from Eq. (22).

BIBLIOGRAPHY

A good general reference to the subject of nonlinear differential equations is

H. T. Davis, *Introduction to Nonlinear Differential and Integral Equations,* U.S. Government Printing Office, Washington, D.C., 1960; Dover Publications, New York, 1962.

Two old but quite excellent references to the subject of Lie groups and differential equations are

A. Cohen, *An Introduction to the Lie Theory of One-Parameter Groups,* 2nd. ed., G. E. Stechert, New York, 1931.

L. E. Dickson, "Differential Equations from the Group Standpoint," *Annals of Mathematics, Ser. 2,* **25**, 287–378 (1923).

Many books have already been written about the similarity solutions of partial differential equations. Six good references are:

W. F. Ames, pp. 87–145 in *Nonlinear Partial Differential Equations in Engineering,* vol. II, chap. 2, Academic Press, New York, 1972.

G. I. Barenblatt, *Similarity, Self-Similarity, and Intermediate Asymptotics,* N. Stein, trans., M. van Dyke, trans. ed., Consultants Bureau, New York, 1979.

G. W. Bluman and J. D. Cole, *Similarity Methods for Differential Equations,* Springer-Verlag, New York, 1974.

L. Dresner, *Similarity Solutions of Nonlinear Partial Differential Equations,* Pitman Publishing, Marshfield, Mass., 1983.

A. G. Hansen, *Similarity Analysis of Boundary-Value Problems in Engineering,* Prentice-Hall, Englewood Cliffs, N.J., 1964.

L. I. Sedow, *Similarity and Dimensional Methods in Mechanics,* Academic Press, New York, 1959.

An excellent reference to the subject of maximum principles and differential inequalities is:

M. H. Protter and H. F. Weinberger, *Maximum Principles in Differential Equations,* Prentice-Hall, Englewood Cliffs, N.J., 1967.

A foundational discussion of monotone operators and iteration is contained in

L. Collatz, pp. 275 ff., *Funktionalanalysis und Numerische Mathematik,* chap. III, Springer-Verlag, Berlin, 1964.

An excellent exposition of complementary variational principles can be found in

A. M. Arthurs, *Complementary Variational Principles,* Clarendon Press, Oxford, 1970.

A classic reference to numerical methods for partial differential equations is

G. E. Forsythe and W. R. Wasow, *Finite-Difference Methods for Partial Differential Equations,* Wiley, New York, 1960.

ORNL/TM-10655
Dist. Category UC-20

INTERNAL DISTRIBUTION

1. D. B. Batchelor
2. J. B. Drake
3–17. L. Dresner
18. R. A. Langley
19. M. S. Lubell
20. O. B. Morgan
21. C. W. Nestor, Jr.
22. M. W. Rosenthal
23. J. Sheffield
24. G. Wilson
25–26. Laboratory Records Department
27. Laboratory Records, ORNL-RC
28. Document Reference Section
29. Central Research Library
30. Fusion Energy Division Library
31–32. Fusion Energy Division Publications Office
33. ORNL Patent Office

EXTERNAL DISTRIBUTION

34. Office of the Assistant Manager for Energy Research and Development, U.S. Department of Energy, Oak Ridge Operations Office, P.O. Box E, Oak Ridge, TN 37831
35. J. D. Callen, Department of Nuclear Engineering, University of Wisconsin, Madison, WI 53706-1687
36. J. F. Clarke, Director, Office of Fusion Energy, Office of Energy Research, ER-50 Germantown, U.S. Department of Energy, Washington, DC 20545
37. R. W. Conn, Department of Chemical, Nuclear, and Thermal Engineering, University of California, Los Angeles, CA 90024
38. S. O. Dean, Fusion Power Associates, Inc., 2 Professional Drive, Suite 248, Gaithersburg, MD 20879
39. H. K. Forsen, Bechtel Group, Inc., Research Engineering, P.O. Box 3965, San Francisco, CA 94105
40. J. R. Gilleland, L-644, Lawrence Livermore National Laboratory, P.O. Box 5511, Livermore, CA 94550
41. R. W. Gould, Department of Applied Physics, California Institute of Technology, Pasadena, CA 91125
42. R. A. Gross, Plasma Research Laboratory, Columbia University, New York, NY 10027
43. D. M. Meade, Princeton Plasma Physics Laboratory, P.O. Box 451, Princeton, NJ 08544

44. M. Roberts, International Programs, Office of Fusion Energy, Office of Energy Research, ER-52 Germantown, U.S. Department of Energy, Washington, DC 20545
45. W. M. Stacey, School of Nuclear Engineering and Health Physics, Georgia Institute of Technology, Atlanta, GA 30332
46. D. Steiner, Nuclear Engineering Department, NES Building, Tibbetts Avenue, Rensselaer Polytechnic Institute, Troy, NY 12181
47. R. Varma, Physical Research Laboratory, Navrangpura, Ahmedabad 380009, India
48. Bibliothek, Max-Planck Institut für Plasmaphysik, Boltzmannstrasse 2, D-8046 Garching, Federal Republic of Germany
49. Bibliothek, Institut für Plasmaphysik, KFA Jülich GmbH, Postfach 1913, D-5170 Jülich, Federal Republic of Germany
50. Bibliothek, KfK Karlsruhe GmbH, Postfach 3640, D-7548 Karlsruhe 1, Federal Republic of Germany
51. Bibliotheque, Centre de Récherches en Physique des Plasmas, Ecole Polytechnique Federale de Lausanne, 21 Avenue des Bains, CH-1007 Lausanne, Switzerland
52. F. Prévot, CEN/Cadarache, Departement de Récherches sur la Fusion Contrôlée, F-13108 Saint-Paul-lez-Durance Cedex, France
53. Bibliothèque, CEN/Cadarache, F-13108 Saint-Paul-lez-Durance Cedex, France
54. Documentation S.I.G.N., Departement de la Physique du Plasma et de la Fusion Contrôlée, Association EURATOM-CEA, Centre d'Etudes Nucléaires, B.P. 85, Centre du Tri, F-38041 Grenoble, France
55. Library, Culham Laboratory, UKAEA, Abingdon, Oxfordshire, OX14 3DB, England
56. Library, JET Joint Undertaking, Abingdon, Oxfordshire OX14 3EA, England
57. Library, FOM-Instituut voor Plasmafysica, Rijnhuizen, Edisonbaan 14, 3439 MN Nieuwegein, The Netherlands
58. Library, Institute of Plasma Physics, Nagoya University, Chikusa-ku, Nagoya 464, Japan
59. Library, International Centre for Theoretical Physics, P.O. Box 586, I-34100 Trieste, Italy
60. Library, Centro Richerche Energia Frascati, C.P. 65, I-00044 Frascati (Roma), Italy
61. Library, Plasma Physics Laboratory, Kyoto University, Gokasho, Uji, Kyoto, Japan
62. Plasma Research Laboratory, Australian National University, P.O. Box 4, Canberra, A.C.T. 2601, Australia
63. Library, Japan Atomic Energy Research Institute, Tokai Research Establishment, Tokai, Naka-gun, Ibaraki-ken 311-02, Japan

64. Library, Japan Atomic Energy Research Institute, Naka Research Establishment, Naka-machi, Naka-gun, Ibaraki-ken, Japan
65. G. A. Eliseev, I. V. Kurchatov Institute of Atomic Energy, P.O. Box 3402, 123182 Moscow, U.S.S.R.
66. V. A. Glukhikh, Scientific-Research Institute of Electro-Physical Apparatus, 188631 Leningrad, U.S.S.R.
67. I. Shpigel, Institute of General Physics, U.S.S.R. Academy of Sciences, Ulitsa Vavilova 38, Moscow, U.S.S.R.
68. D. D. Ryutov, Institute of Nuclear Physics, Siberian Branch of the Academy of Sciences of the U.S.S.R., Sovetskaya St. 5, 630090 Novosibirsk, U.S.S.R.
69. V. T. Tolok, Kharkov Physical-Technical Institute, Academical St. 1, 310108 Kharkov, U.S.S.R.
70. Library, Academia Sinica, P.O. Box 3908, Beijing, China (PRC)
71. P. Komarek, KfK Karlsruhe GmbH, Postfach 3640, D-7548 Karlsruhe 1, Federal Republic of Germany
72. S. Shimamoto, Japan Atomic Energy Research Establishment, Tokai Research Establishment, Tokai, Naka-gun, Ibaraki-ken 311-02, Japan

    Department of Mathematics, The University of Tennessee, Knoxville, TN 37916

73. V. Alexiades
74. J. S. Bradley
75. S. Lenhart

    Department of Nuclear Engineering and Engineering Physics, University of Wisconsin, Madison, WI 53706

76. M. M. Abdelsalam
77. R. Boom
78. M. Carbon
79. M. L. Corradini
80. Y. M. Eyssa
81. G. L. Kulcinski
82. D. C. Larbalestier
83. G. A. Moses
84. I. N. Sviatoslavsky
85. S. W. van Sciver
86. P. L. Walstrom

    Swiss Institute for Nuclear Research, CH-5234 Villigen, Switzerland

87. G. Vecsey
88. J. A. Zichy

89–166. Given distribution as shown in TIC-4500, Magnetic Fusion Energy