



Simplification and Scaling

Lee A. Segel

SIAM Review, Vol. 14, No. 4. (Oct., 1972), pp. 547-571.

Stable URL:

<http://links.jstor.org/sici?sici=0036-1445%28197210%2914%3A4%3C547%3ASAS%3E2.0.CO%3B2-U>

SIAM Review is currently published by Society for Industrial and Applied Mathematics.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/about/terms.html>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/journals/siam.html>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

SIMPLIFICATION AND SCALING*

LEE A. SEGEL†

Abstract. An attempt is made to clarify two frequently used applied mathematical techniques. Section 1 begins with a description of the *basic simplification procedure* in which a term is neglected under the assumption that it is small, and the consistency of this assumption is later checked. Successful uses of the basic simplification procedure are illustrated. Wretched consistent approximations are presented, showing that the procedure can be misused. The situation is clarified by a discussion of the relation between the size of the residual and the goodness of the approximation in three simple problem classes. Section 2 discusses *scaling*: how to choose dimensionless variables in such a way that the relative size of the various terms in an equation is explicitly indicated by the magnitudes of the dimensionless parameters which precede them. Scaling is illustrated on a simple physical problem and on several known functions. It is pointed out that more than one scale may be necessary, and the connection with singular perturbation theory is established. Advice is given on circumventing the dilemma of choosing scales for the unknown functions which comprise the solution of the very problem whose investigation one is trying to facilitate by appropriate scaling.

Introduction. In his effort to acquire understanding of the subject he is investigating, an applied mathematician cannot restrict himself to rigorously justified operations. Sometimes years of research are necessary before his results can be firmly established, and he cannot wait until this is accomplished. Yet, aside from its obvious merit of providing reliable deductions from assumptions, rigorous reasoning has the further advantage that it proceeds in a carefully defined and well-understood manner. By contrast, reliable as they have proved to be, the freer and more diverse techniques of the applied mathematician are rarely explicitly delineated but rather are transmitted indirectly and informally. The techniques I refer to here are not, say, use of Fourier methods to solve linear problems but the techniques of constructing models and of simplifying equations, etc. I believe that it will be useful to have available detailed explicit expositions of these matters.

The techniques I have in mind comprise the core of the applied mathematician's art (or craft), so their description is necessarily incomplete and somewhat subjective. Attempts to explain part of the artists' methods yield material which probably is not easy to read and (if the present effort is a guide) is definitely not easy to write. Yet it is my experience with unpublished attempts in this direction that the effort is worthwhile. Students are repelled by presentations beginning with incantations like "Many years work by experts indicates . . ." but have their interest aroused and their understanding aided by honest struggles to describe what is being done. Researchers find that their command of the techniques is enhanced by efforts to make them more readily understandable.

This paper comprises an attempt to describe two frequently used applied mathematical methods: the utilization of consistent simplifications of problems wherein a neglected term is verified a posteriori to be small, and the employment of dimensionless variables which are selected so as to cause the appearance in the

* Received by the editors November 11, 1971.

† Department of Mathematics, Rensselaer Polytechnic Institute, Troy, New York 12181. This work was supported in part by the Office of Naval Research (Mechanics Branch). It was completed while the author was on leave at the Weizmann Institute, supported by grants from the Guggenheim Foundation and from Rensselaer.

governing problem of meaningful small parameters (if this is possible). The exposition proceeds to a large extent from consideration of various examples. These examples are very simple, so that the vista of ideas is not blocked by mountains of calculations. (In the training of students, more complicated examples should be referred to as well—whether by further class exposition, homework, or assigned reading—to dispel notions that the ideas are trivial.) In trying to make explicit what is “well-known” I have probably briefly discussed a few matters which really are thoroughly understood, but this does no harm. No attempt has been made to give final and detailed procedural recommendations for specified classes of problems; such recommendations can never be complete, and applied mathematicians must cultivate the ability to extract a widely applicable method from the analysis of a particular problem.

1. The basic simplification procedure. Simplification of a given set of equations may make it possible to avoid large machine calculations or massive analytic work and still to obtain a useful answer. Even if the equations can be solved exactly without undue effort, simplified equations may yield a sufficiently accurate solution whose features are more readily apparent than those of the exact solution. Drastic simplification may allow rapid solution and immediate determination of whether one is on the right track.

The following *basic simplification procedure* is used time and again by applied mathematicians:

- (i) Somehow *identify terms which appear relatively small*.
- (ii) *Delete apparently small terms* and solve the resulting simplified problem.
- (iii) *Check for consistency*. That is, use the approximate solution just obtained to evaluate the neglected terms, and check that they are indeed relatively small. In practice, the smallness of a given term is generally gauged in relation to other terms in the same equation.

We shall now present some examples of the procedure. In doing so we shall use the symbol \sim to mean “seems to be approximately equal to” and \approx to mean “is approximately equal to.”

Example 1.1. Pretend that the problem

$$(1.1a) \quad x + 10y = 21,$$

$$(1.1b) \quad 5x + y = 7$$

is nontrivial. In (1.1a) the coefficient of the x -term is small compared to the coefficient of the y term so it is tempting to assume that the x -term may be neglected to first approximation. Omitting this term we obtain

$$(1.2) \quad y \sim 2.1, \quad x \sim \frac{1}{5}(7 - 2.1) \approx 0.998.$$

Approximating the unknowns by means of the values given in (1.2), we estimate that the ratio of the first term on the left side of (1.1a) to the second has magnitude $0.998/21 \approx 0.05$. This number is small compared to unity so our approximation appears consistent. And our approximate values of x and y are close to the true values $x = 1, y = 2$.

Example 1.2. Consider a body of constant mass m which is radially projected upward from the earth's surface with initial speed V . Let R denote the radius of the earth and let $x^*(t^*)$ denote radial distance from the earth's surface at time t^* . (We use starred variables here so that later we can introduce unstarred dimensionless variables.) Neglecting air resistance, the governing equations and initial conditions are

$$(1.3) \quad \frac{d^2x^*}{dt^{*2}} = -\frac{gR^2}{(x^* + R)^2}, \quad x^*(0) = 0, \quad \frac{dx^*}{dt^*}(0) = V.$$

If V is small in some sense, the displacement x^* should be small compared to R . Instead of (1.3), we can then consider the simplified problem

$$(1.4) \quad \frac{d^2x^*}{dt^{*2}} = -g, \quad x^*(0) = 0, \quad \frac{dx^*}{dt^*}(0) = V,$$

with solution

$$(1.5) \quad \frac{dx^*}{dt^*} = -gt^* + V, \quad x^* = -\frac{1}{2}gt^{*2} + Vt^*.$$

At the time $t^* = Vg^{-1}$, when the speed vanishes, we find from (1.5) that x^* reaches a maximum of $\frac{1}{2}V^2/g$. Thus x^*/R is at most $\frac{1}{2}V^2/(gR)$, and our approximation is consistent when V^2 is small compared to gR .

Example 1.3. (Our discussion will be similar to that found in [4, p. 258].) Consider two-dimensional surface waves in inviscid water. The equation obtained by balancing the product of the density ρ and the vertical acceleration with the vertical pressure gradient and the gravitational body force is

$$(1.6) \quad \rho \frac{Dv}{Dt} = -\frac{\partial p}{\partial y} - g\rho.$$

Let us consider wave motions which are slow in the sense that the vertical acceleration Dv/Dt is negligible compared to the gravitational acceleration g , so that (1.6) can be integrated with respect to y to obtain

$$(1.7) \quad p - p_0 = g\rho(y_0 + \eta - y).$$

This equation states that the difference between the pressure p and the atmospheric pressure p_0 is due simply to the weight of water below the free surface $y(x, t) = y_0 + \eta(x, t)$. A second equation is obtained from the requirement of mass conservation. The resulting problem yields waves whose speed c is roughly $(gh)^{1/2}$, where h is the water depth. (If the problem is linearized, $c = (gh)^{1/2}$ exactly.) If these waves have length $4L$, a particle rises from the mean surface height y_0 to the peak height η in about the time it takes the wave to move a quarter wavelength, namely, L/c . Assuming smooth changes in η , we thus estimate the vertical speed as $c\eta/L$ and the vertical acceleration by $\eta c^2/L^2$. For consistency we must have $\eta c^2/L^2 \ll g$, where $c^2 \approx gh$. As $\eta \leq h$, we certainly have a consistent approximation for waves which are long compared to the water depth in the sense $L^2 \gg h^2$.

There is a feeling of general well-being attached to the property of being consistent, but precisely what does consistency imply in the present context? To generate interest in the answer to this question, let us turn to some examples which illustrate some possibly disturbing aspects of the basic simplification procedure.

Example 1.4. Consider the equations

$$(1.8a) \quad 0.01x + y = 0.1,$$

$$(1.8b) \quad x + 101y = 11.$$

In (1.8a) the coefficient of the x -term is small compared to the coefficient of the y -term. Seeing no reason to believe that x and y differ too much, we neglect the former term to obtain

$$(1.9) \quad y \sim 0.1, \quad x \sim 11 - 101(0.1) = 0.9.$$

Using (1.9) we estimate that the ratio of the first term on the left side of (1.8) to the second has magnitude $(0.01)(0.9)/(0.1) = 0.09$. This number is small compared to unity so our approximation appears to be consistent. But the exact answer is $y = 1$, $x = -90$, so our "approximation" for y is off by a factor of 10, while our "approximation" for x is off by a factor of 100 and has the wrong sign to boot.

To see what went wrong, let us generalize (1.8) and consider the following equations for $x(\varepsilon)$ and $y(\varepsilon)$:

$$(1.10) \quad \varepsilon x + y = 0.1, \quad x + 101y = 11.$$

Equations (1.10) reduce to (1.8) when $\varepsilon = 0.01$. The approximation (1.9) corresponds to taking $\varepsilon = 0$. But is it true that $x(\varepsilon) \approx x(0)$, $y(\varepsilon) \approx y(0)$ when $\varepsilon = 0.01$? As soon as we ask this question it becomes clear that we shall be in trouble if $x(\varepsilon)$ and $y(\varepsilon)$ change rapidly with ε near $\varepsilon = 0$. That this is indeed the case can be seen from the exact solution to (1.10):

$$(1.11) \quad x = \frac{0.9}{1 - 101\varepsilon}, \quad y = \frac{0.1 - 11\varepsilon}{1 - 101\varepsilon}, \quad \varepsilon \neq 1/101.$$

Example 1.4 illustrates the fact that an approximation which appears consistent may not actually be so. In approximating $x(0.01)$ and $y(0.01)$ by $x(0)$ and $y(0)$ we simplified the equations by neglecting the εx (ε) term. This is a consistent procedure if $|0.01 x(0.01)|$ is small compared to $|y(0.01)|$. Pretending that we only knew the approximate solutions $x(0)$ and $y(0)$, we checked for consistency as best we could by examining $|0.01 x(0)/y(0)|$. This ratio had the satisfactorily small value 0.09 so the approximation appeared consistent. We were deceived, however, since

$$|0.01 x(0.01)/y(0.01)| = 9.$$

We thus distinguish *apparent consistency* (the approximation to the neglected term is small) from *genuine consistency* (the term neglected is truly small). The third step in the basic simplification procedure should really be called *checking for apparent consistency*.

It is not surprising that our approximation in Example 1.4 was a poor one, because it was apparently consistent but not genuinely consistent. But it is disheartening that an approximation can be apparently consistent but not genuinely so. Even more disheartening is the fact that an approximation which seems genuinely consistent may be very inaccurate, as is illustrated in the next example.

Example 1.5 [6, p. 41]. The polynomial

$$(1.12) \quad (x - 1)(x - 2)(x - 3) \cdots (x - 20) = x^{20} - 210x^{19} + \cdots$$

has as zeros the first twenty positive integers. If the coefficient of x^{19} is changed by the addition of εx^{19} where $\varepsilon = -2^{-23} \approx -1.19 \times 10^{-8}$, then the smaller zeros are almost unaltered, but the larger zeros are so radically altered that the changed equation has five pairs of complex conjugate roots, given roughly by $10 \pm 0.64i$, $12 \pm 1.7i$, $14 \pm 2.5i$, $17 \pm 2.8i$, $20 \pm 1.9i$. Whether the added term εx^{19} is evaluated using the “approximate” zeros or the exact zeros, it is still small in magnitude compared to the retained term $-210x^{19}$. Both apparent and genuine consistency seem verified, but the approximation is again a poor one.

We have established the existence of *wretched consistent approximations*. It is clearly necessary to examine matters with more care. We shall find that certain general remarks can be made, but that further understanding requires a somewhat detailed analysis of various problem classes.

As was illustrated in Example 1.4, that a problem has been altered slightly is a significant fact only if it is known that small changes in the problem cause small changes in its solution. Numerical analysts call a problem *ill-conditioned* if this is not the case. It will be convenient here also to use the terminology that the *solution* to a problem is *sensitive to neglect of a term T* if such neglect cannot be expected to result in a small change in the solution even when the neglected term is genuinely small. Thus Wilkinson reports in [6] that the smaller zeros of (1.12) are not sensitive to alterations in *any* of the coefficients (these alterations can be regarded as the neglect of certain terms in an obvious way) and even the larger zeros are not sensitive to the neglect of *certain* coefficients. But the larger zeros *are* sensitive to the neglect of some of the coefficients, namely, those multiplying the higher powered terms in the polynomial.

The only blanket statement which I can make concerning the value of consistency checks is that *lack of apparent consistency is almost certainly associated with poor approximation*. To show this, let x symbolize the true solution to some problem and let \tilde{x} denote the approximate solution which one obtains when the term T is neglected. Neglect of T is apparently or genuinely consistent, respectively, if $T(\tilde{x})$ or $T(x)$ is small. (The notation used here is symbolic. x could be a vector function, for example, and T could depend on the components of x and several of their derivatives.) If the solution is sensitive to the neglect of T , then, by definition, such neglect cannot be expected to lead to a good approximation. If the solution is *not* sensitive to the neglect of T , then

$$[T(x) \text{ small}] \text{ implies } [\tilde{x} \approx x] \text{ implies } [T(\tilde{x}) \text{ small}],$$

assuming that T varies continuously. Finding a large apparent error $T(\tilde{x})$ is therefore logically inconsistent with the supposition that the term neglected is genuinely small, and neglect of a genuinely large term cannot be expected to lead to a good approximation.

The imprecise words “small” and “large” appear again and again, but these words can only be given a definite meaning in the context of a particular problem. Thus, the numerical magnitude of an acceleration (say) can be changed by altering the time scale, but this change can have no effect on the question of whether the

acceleration is negligible or not. To mention a slightly different example, if a certain temperature should be 1°C but was calculated as 1.1°C the error is 10%. If the Fahrenheit scale is employed, the percentage error is only about 0.5%, but it is without significance that the latter percentage is much smaller than the former. Whatever temperature scale is used, the decision as to whether an error is acceptable or not rests on scientific grounds and must be independent of the units employed. (*Note*: The difference between the alterations of the time scale and the temperature scale is that the former involves only the choice of a unit, while the latter also involves the choice of a zero.)

To make further progress, it is necessary to consider particular types of problems. We shall discuss linear algebraic equations, the determination of the zeros of a function, and second order ordinary differential equations.

1.1. Linear algebraic equations. Suppose that we are confronted with the equation $Ax = b$, where A is a square matrix and x is a column vector. Instead, we solve a simplified or modified problem in which A is replaced by A_1 . (We assume the existence of A^{-1} and A_1^{-1} .) If we denote the solution to the modified problem by x_1 , we have $A_1x_1 = b$. We shall designate the terms which are apparently neglected in the modified problem as the *residual* r and the terms which are genuinely neglected as the *genuine equation error* g . The vectors g and r are given by

$$(1.13) \quad g = Ax - A_1x = b - A_1x, \quad r = Ax_1 - A_1x_1 = Ax_1 - b.$$

We denote the difference between the answers to the modified and exact problems by h : $h = x_1 - x$. Since $A(x_1 + h) = b$ we find

$$(1.14) \quad Ah = r, \quad h = A^{-1}r,$$

so the error h can be expressed in terms of the residual r and the original matrix A . (We can also express h in terms of the genuine equation error g and the modified matrix A_1 , by $h = A_1^{-1}g$, but this is less useful.) From (1.14),

$$\|A\| \|h\| \geq \|r\| \quad \text{and} \quad \|h\| \leq \|A^{-1}\| \|r\|,$$

by the fundamental inequality connecting a vector norm and its subordinate matrix norm [6, p. 80]. Thus we can bound the error by

$$(1.15) \quad \|A\|^{-1} \|r\| \leq \|h\| \leq \|A^{-1}\| \|r\|.$$

To bound the relative error we use the original equation $Ax = b$ to write

$$\|b\| \|A\|^{-1} \leq \|x\| \quad \text{and} \quad \|x\| \leq \|A^{-1}\| \|b\|.$$

We thereby obtain

$$(1.16) \quad \kappa^{-1}R \leq E \leq \kappa R,$$

where

$$E = \text{relative magnitude of error} = \|h\|/\|x\|,$$

$$R = \text{relative magnitude of residual} = \|r\|/\|b\|,$$

$$\kappa = \text{condition number} = \|A\| \|A^{-1}\|.$$

Hopefully, the residual is close to the genuine equation error. As we have already mentioned, however, the error in the answer depends not only on the error in the equation, but also on how sensitive the equation is to modification. As is frequently the case in matrix problems, in (1.16) the sensitivity or *condition* is quantitatively expressed by the condition number $\|A\| \|A^{-1}\| \equiv \kappa$. We shall not use the Euclidean norm, so that $\|I\| = 1$ and $\kappa \geq \|AA^{-1}\| = 1$. Thus the condition number cannot be less than unity, but (by definition) in well-conditioned problems it will not be too much greater than unity. For well-conditioned problems, then, (1.16) shows that the error will be large if the residual is large, and the error will be small if the residual is small. For ill-conditioned problems, on the other hand, the error can be large in spite of the fact that the residual is small. Further, the error can be small even though the residual is large. This last result demonstrates particularly forcefully that “anything can happen” with an ill-conditioned problem, for a small error may result even if such a problem is drastically altered.

To illustrate (1.16), consider Example 1.1. There,

$$A = \begin{bmatrix} 1 & 10 \\ 5 & 1 \end{bmatrix}, \quad A^{-1} = \begin{bmatrix} -1/49 & 10/49 \\ 5/49 & -1/49 \end{bmatrix}, \quad r = \begin{bmatrix} 0.998 \\ 0 \end{bmatrix}, \quad b = \begin{bmatrix} 21 \\ 7 \end{bmatrix}.$$

We shall employ the sup norm so that

$$\|x\| = \max_i |x_i| \quad \text{and} \quad \|A\| = \max_i \sum_j |a_{ij}|,$$

where the x_i and a_{ij} are the elements of x and A respectively. Thus

$$\kappa = \|A\| \|A^{-1}\| = 11(11/49) \approx 2.4,$$

$$R = \|r\|/\|b\| = 0.998/21 \approx 0.05,$$

and (1.16) gives, approximately,

$$(1.17) \quad 0.02 \leq E \leq 0.12.$$

Actually,

$$x = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad h = \begin{bmatrix} 0.002 \\ -0.1 \end{bmatrix}, \quad E = \frac{\|h\|}{\|x\|} \approx 0.05.$$

Example 1.4 shows what happens in an ill-conditioned problem. We have

$$(1.18) \quad \|A\| = 102, \quad \|A^{-1}\| = 10,200, \quad \kappa \approx 10^6, \quad R \approx 10^{-3},$$

so roughly, $10^{-9} \leq E \leq 10^3$. Actually, $E \approx 1$. The error bounds are of little help, but the vast size of the condition number tells us all we need to know—that our “approximation” is almost certainly useless.

The estimate (1.16) does not give any justification for comparing the residual with “a term retained,” but rather shows that comparison with the right side of the equation is informative. I have been unable to find any evidence for the usefulness of the statement *the residual is small compared to a term retained* except the following very crude remark: If the term retained is “typical,” and if there is no near-cancellation of typical terms which results in a decrease in magnitude of the left side of the equation, then the italicized statement is equivalent to a comparison with the right side of the equation. Rather than contemplate this string of “ifs,” why not compare r with b ?¹

¹ In Example 1.1, the apparent value of the retained term $10y$ equals 21 so that $|x| \ll |10y|$ implies the desired fact that $|x| \ll \max [21, 7]$.

The estimates given by (1.16) are of a general type which is common in the numerical analysis literature, but I was unaware of explicit mention of them. E. Rogers has pointed out to me, however, that (1.16) is found in L. Collatz [7, p. 100] and in E. Isaacson and H. Keller [8, p. 48].

Successive reduction of residuals by an iteration process forms the basis of many techniques of numerical analysis. For example, this is the explicit strategy of relaxation methods. In the present case, the equation $Ah = r$ can be used to approximate the error h , and a better approximation can generally be obtained by repeating the process. (For details, see [6, p. 121 ff].) But we are primarily interested in situations where it is not feasible to calculate higher approximations, so to gain further understanding of the meaning of a small residual we turn from linear systems to another class of problems.

1.2. Zeros of a function. Let a zero of $f(x, \varepsilon)$ be $x(\varepsilon)$, so

$$(1.19) \quad f[x(\varepsilon), \varepsilon] \equiv 0.$$

Suppose ε is a small parameter, and let an approximation $x^{(0)}$ to the zero $x(\varepsilon)$ be found by setting $\varepsilon = 0$, so that $f[x^{(0)}, 0] = 0$. (We shall consistently use a superscript zero to indicate that ε has been set equal to zero.) The equation error in equating $f(x, 0)$ to zero rather than $f(x, \varepsilon)$ is $f(x, \varepsilon) - f(x, 0)$. The error is apparent or genuine depending on whether this expression is evaluated when $x = x^{(0)}$ or $x = x(\varepsilon)$. Thus for the genuine equation error g we have

$$g \equiv f[x(\varepsilon), \varepsilon] - f[x(\varepsilon), 0] = -\varepsilon f_x^{(0)} x'(0) + O(\varepsilon^2),$$

while the residual satisfies

$$r \equiv f[x^{(0)}, \varepsilon] - f[x^{(0)}, 0] = f[x^{(0)}, \varepsilon] = \varepsilon f_\varepsilon^{(0)} + O(\varepsilon^2).$$

We wish to relate r to $h(\varepsilon) \equiv x(\varepsilon) - x^{(0)}$, the error in the solution. We shall retain only lowest powers of ε . To lowest order,

$$h(\varepsilon) = \varepsilon x'(0).$$

But

$$0 \equiv f[x(\varepsilon), \varepsilon] - f[x_0, 0] = \varepsilon \{ f_x^{(0)} x'(0) + f_\varepsilon^{(0)} \} + O(\varepsilon^2).$$

The vanishing of the expression within the curly brackets implies two things.

- (i) The genuine error and the apparent error are equal to first approximation.
- (ii) To first order, the error is given by

$$(1.20) \quad h = -r/f_x^{(0)}.$$

The error again depends not only on the residual r but on a measure of the condition of the problem, in this case $f_x^{(0)}$. Here, the influence of condition is particularly clear. If $f_x^{(0)}$ is small, then a small change in f is associated with an order one change in x ; in other words, a small change in the equation is associated with a non-small change in the solution. When this is the case, one expects a large error in spite of a small residual.

Leaving conditioning aside, when relative error is to be estimated it is the relative residual r/x which is relevant, where x is the solution. Barring cases of

ill-conditioning, r can be compared to the approximate solution. Consequently, in Example 1.5 one should not examine the ratio of the neglected term to a term retained, but the ratio of the neglected term to the estimated solution.² The latter ratio is $\varepsilon \bar{x}^{19}/\bar{x}$. For $\varepsilon = -2^{-23}$ this ratio is small for \bar{x} close to unity but becomes large for \bar{x} much bigger than 2. There is a clear warning of difficulty for the larger roots. Actually the warning is somewhat premature, for here the conditioning factor acts to keep the errors reasonable for zeros up to about 10. (See [6] for a full discussion of sensitivity in this problem.) "Normally" the various terms in a polynomial might be expected to be of the same size with coefficients of unit magnitude; in such a case comparison with any term retained will give the magnitude of the desired ratio between the residual and the solution. The present polynomial is not "normal," however. Many of its coefficients are very large, but the terms nearly cancel because of sign alternations. "Abnormal" equations with such large but nearly canceling terms are not a rarity; this must be kept in mind when comparing a term neglected with an allegedly "typical" term retained.

1.3. A differential equation. Consider the following initial value problem for $x(t, \varepsilon)$:

$$(1.21) \quad \begin{aligned} f(t, x, \dot{x}, \ddot{x}, \varepsilon) &= 0; & x(0) &= A, \dot{x}(0) = B; \\ | \varepsilon | &\ll 1; & \dot{} &= d/dt. \end{aligned}$$

The following reasoning closely parallels that used in the discussion of (19).

An approximate solution $x^{(0)}(t)$ is found by setting $\varepsilon = 0$:

$$f(t, x^{(0)}, \dot{x}^{(0)}, \ddot{x}^{(0)}, 0) = 0.$$

The residual r satisfies

$$r \equiv f(t, x^{(0)}, \dot{x}^{(0)}, \ddot{x}^{(0)}, \varepsilon) = \varepsilon [f_5]_{\varepsilon=0} + O(\varepsilon^2),$$

while the genuine equation error g satisfies

$$g = -f(t, x, \dot{x}, \ddot{x}, 0) = -\varepsilon [f_2 x_\varepsilon + f_3 \dot{x}_\varepsilon + f_4 \ddot{x}_\varepsilon]_{\varepsilon=0} + O(\varepsilon^2).$$

(Subscripts denote partial derivatives.) But

$$\begin{aligned} 0 &= f(t, x, \dot{x}, \ddot{x}, \varepsilon) - f(t, x^{(0)}, \dot{x}^{(0)}, \ddot{x}^{(0)}, 0) \\ &= \varepsilon [f_2 x_\varepsilon + f_3 \dot{x}_\varepsilon + f_4 \ddot{x}_\varepsilon + f_5]_{\varepsilon=0} + O(\varepsilon^2). \end{aligned}$$

Thus, as before, using a superscript to remind us that all terms are evaluated at $\varepsilon = 0$, we have

$$(1.22) \quad f_4^{(0)} \ddot{x}_\varepsilon^{(0)} + f_3^{(0)} \dot{x}_\varepsilon^{(0)} + f_2^{(0)} x_\varepsilon^{(0)} = -f_5^{(0)}.$$

We conclude from (1.22) that the residual and the genuine equation error are the same to lowest order. Also, we can now obtain an expression for the error in terms of the residual. To do this we make the definitions

$$(1.23) \quad \begin{aligned} \mathbf{F}(t) &\equiv (f_2^{(0)}, f_3^{(0)}, f_4^{(0)}), \\ \mathbf{h}(t) &\equiv (x - x^{(0)}, \dot{x} - \dot{x}^{(0)}, \ddot{x} - \ddot{x}^{(0)}). \end{aligned}$$

² Infatuates of our terminology could say that Example 1.5 provides an example of "apparently genuine consistency."

We note that

$$\mathbf{h}(t) = \varepsilon(x_\varepsilon^{(0)}, \dot{x}_\varepsilon^{(0)}, \ddot{x}_\varepsilon^{(0)}) + O(\varepsilon^2).$$

Thus to lowest order, (1.22) implies

$$(1.24) \quad \mathbf{F} \cdot \mathbf{h} = -r.$$

Since $|r| \leq |\mathbf{F}| |\mathbf{h}|$, we have

$$|\mathbf{h}| \geq \frac{|r|}{|\mathbf{F}|} \geq \frac{|r|}{\min |\mathbf{F}|}.$$

Here is the presently appropriate version of our now-familiar result that a large residual precludes a small error unless the problem is ill-conditioned. If a single condition number is required, that number is the smallest value of $|\mathbf{F}|$. Note that both the error vector \mathbf{h} and the condition vector \mathbf{F} involve not only the behavior of the solution but also of its first two derivatives. Thus, even if a small change in the equation is associated with a small change in the solution itself, ill-conditioning arises if, for example, the solution's second derivative is markedly affected.

It is difficult to make any general statements from (1.24) about whether a small residual implies a small error. All one can say is that, as usual, ill-conditioning must be excluded. Further, one must somehow satisfy oneself that the smallness of r does not arise from cancellation of non-small errors in $x_\varepsilon^{(0)}$, $\dot{x}_\varepsilon^{(0)}$ and $\ddot{x}_\varepsilon^{(0)}$. In particular cases one can often make stronger statements. To illustrate this, consider

$$(1.25) \quad \ddot{x} + (1 + \varepsilon x)^{-2} = 0; \quad x(0) = 0, \quad \dot{x}(0) = 1.$$

As we shall demonstrate below, (1.25) is the governing equation of the appropriately nondimensionalized version of the projectile problem of Example 1.2 when the parameter $\varepsilon \equiv V^2/Rg$ is small. In this case, $\mathbf{F} = (0, 0, 1)$ and $r = \varepsilon x^{(0)}(t)$ so one can immediately deduce from (1.24) that

$$(1.26) \quad |\varepsilon \ddot{x}_\varepsilon^{(0)}| \leq \varepsilon \max x^{(0)}(t).$$

To first order in ε , the above relation shows that the error in the acceleration is small. Two integrations permit estimates of the error in x itself.

An important point can be introduced via consideration of small vibrations of a simple nonlinear pendulum. It can be shown that the proper formulation of this problem is

$$(1.27) \quad \ddot{\theta} + \varepsilon^{-1/2} \sin(\varepsilon^{1/2}\theta) = 0; \quad \theta(0) = 1, \quad \dot{\theta}(0) = 0; \quad 0 < \varepsilon \ll 1.$$

(It is not necessary here to identify the variables.) The zeroth approximation is $\theta^{(0)} = \cos t$. Using (1.24) to express the error, we find

$$(1.28) \quad \varepsilon[\ddot{\theta}_\varepsilon^{(0)} + \theta_\varepsilon^{(0)}] = r = \frac{1}{6}\varepsilon[\theta^{(0)}]^3.$$

We see illustrated here why it is difficult to bound the error from a knowledge of the residual, for (1.28) is nothing less than a differential equation for $\theta_\varepsilon^{(0)}$. We are faced with the problem of obtaining an estimate of the solution to a linear differential equation from a knowledge of the magnitude of its forcing term. There is no

universal formula here, but the following simple calculation is typical of what can be done. Using the method of variation of parameters and observing that the initial conditions $\theta_\varepsilon^{(0)}(0) = \dot{\theta}_\varepsilon^{(0)}(0) = 0$ hold, we write

$$\varepsilon\theta_\varepsilon^{(0)} = \int_0^t \sin(t - \zeta)r(\zeta) d\zeta.$$

Since $|r| \leq \varepsilon/6$,

$$|\varepsilon\theta_\varepsilon^{(0)}(t)| \leq \varepsilon T/6 \quad \text{for } 0 \leq t \leq T.$$

As is well known, there is indeed an increase of error with T .

From another point of view, the current example illustrates the fact that perhaps the best way to obtain information about the validity of the first approximation is to compute the second approximation. After all, $\theta_\varepsilon^{(0)}$ is nothing more than the $O(\varepsilon)$ coefficient in the power series expansion of $\theta(t, \varepsilon)$. Equation (1.28) which governs $\theta_\varepsilon^{(0)}$ was obtained by what can be regarded as the parametric differentiation approach to perturbation theory.

This is where we terminate our analysis of the basic simplification procedure. We have begun a probe into the relation between the residual and the error. We gave some examples of approximations which were poor in spite of small residuals. (See [5, p. 198 ff. and p. 276 ff.] for other examples.) In trying to pinpoint what went wrong we considered some classical problems but only scratched the surface of the available body of knowledge concerning the use of residuals to obtain good approximate solutions. For such problems, the basic simplification procedure is only the beginning of a proper approach. (Those interested in passing beyond the beginner level are urged to consult the slim but extremely helpful book of Wilkinson [6] which we have repeatedly cited.)

But time and time again the applied mathematician is faced with a problem which is so formidable that he can barely solve a highly simplified version of it. We have discussed classical examples as a means for acquiring a general understanding which will be useful in attacking these more formidable problems. Our discussion leads us to the following *recommendations*:

(i) Use the basic simplification procedure in difficult problems. Not much extra work is required to estimate the magnitude of neglected terms, and you will at least learn where your simplification is almost certainly invalid (when the residual is large).

(ii) Although it is difficult to bound the error associated with a given residual, it may be helpful to regard the residual as an extraneous forcing and to use physical intuition to estimate its effect. Another possibility is to replace the residual by its mean or maximum value and to evaluate the effect of this constant forcing. To spot hidden ill-conditioning, make small random modifications in the problem and solve again [5, p. 170].

(iii) Beware of comparing a term neglected with a term retained. To estimate relative error, compare the neglected term with your approximate solution.

(iv) An applied mathematician studies simplified models to gain an understanding of complicated situations. In this spirit, regard a deep study of the simpler classical problems of numerical analysis and perturbation theory as not only of value in itself but also of value in its indication of what to expect in more complicated problems.

2. Scaling. In much of our previous discussion, we took for granted that a first approximation was obtained to the solution of some problem by letting the small parameter ε tend to zero with the independent variable fixed. For this to be the case, the variables must be properly chosen. As an illustration, consider

$$(2.1a) \quad u(x, \varepsilon) \equiv x + e^{-x/\varepsilon} \quad \text{for } 0 \leq x \leq 1, \quad \varepsilon > 0,$$

where

$$(2.1b) \quad \lim_{\varepsilon \rightarrow 0} u(x, \varepsilon) = x.$$

If we switch to the new independent variable $\xi = x/\varepsilon$, we have

$$(2.2a) \quad v(\xi, \varepsilon) \equiv u(\varepsilon\xi, \varepsilon) = \varepsilon\xi + e^{-\xi},$$

$$(2.2b) \quad \lim_{\varepsilon \rightarrow 0} v(\xi, \varepsilon) = e^{-\xi}.$$

On the other hand, if we introduce $\eta = x/\varepsilon^2$, we obtain

$$(2.3a) \quad w(\eta, \varepsilon) = u(\varepsilon^2\eta, \varepsilon) = \varepsilon^2\eta + e^{-\varepsilon\eta},$$

$$(2.3b) \quad \lim_{\varepsilon \rightarrow 0} w(\eta, \varepsilon) = 1.$$

Which of (2.1b), (2.2b), and (2.3b)—if any—gives “the correct first approximation” to (2.1a)? This is not a difficult question, but let us proceed slowly.

2.1. Dimensional analysis. It is appropriate to begin by reviewing the process of introducing dimensionless variables [1], [2]. We shall illustrate the ideas on the projectile problem of Example 1.2. We repeat the equations of (1.3):

$$(2.4) \quad \frac{d^2x^*}{dt^{*2}} = -\frac{gR^2}{(x^* + R)^2}, \quad x^*(0) = 0, \quad \frac{dx^*}{dt^*}(0) = V.$$

In (2.4) the parameters and their dimensions in terms of length (L) and time (T) are:

<i>Parameter</i>	<i>Dimension</i>
Earth radius R	L
Gravitational acceleration g	LT^{-2}
Initial speed V	LT^{-1}

The quantities R and R/V are appropriate choices for an intrinsic reference length and an intrinsic reference time, where *intrinsic reference quantities* are defined to be *standards of measurement formed from the parameters of a given problem*. We measure the dependent variable x^* and the independent variable t^* using the above-named reference quantities as units. This is equivalent to introducing the variables

$$(2.5) \quad y = x^*/R, \quad \tau = t^*/(RV^{-1}),$$

with which (2.4) becomes

$$(2.6a) \quad \varepsilon \frac{d^2 y}{d\tau^2} = -\frac{1}{(y+1)^2},$$

$$(2.6b) \quad y(0) = 0,$$

$$(2.6c) \quad \frac{dy}{d\tau}(0) = 1,$$

where

$$(2.7) \quad \varepsilon = V^2/(gR).$$

The variables y and τ in (2.6) are *dimensionless* because their values are independent of the units used in the measurement process. The parameter ε is also dimensionless.

If we explicitly note parameter dependence, (2.4) appears to have a solution of the form

$$(2.8) \quad x^* = x^*(t^*; g, R, V),$$

while the solution to (2.6) can be written

$$(2.9) \quad y = y(\tau; \varepsilon).$$

The solution (2.8) depends on the three parameters g , R , and V , but in (2.9) only the single dimensionless parameter ε appears.

As an alternative to (2.5) we could base our time scale on the acceleration g , writing

$$(2.10) \quad z = x^*/R, \quad \tau_1 = t^*/\sqrt{(Rg^{-1})}$$

with which (2.4) becomes

$$(2.11) \quad \frac{d^2 z}{d\tau_1^2} = -\frac{1}{(z+1)^2}, \quad z(0) = 0, \quad \frac{dz}{d\tau_1}(0) = \varepsilon^{1/2}.$$

Now we expect a solution of the form

$$(2.12) \quad z = z(\tau_1, \varepsilon).$$

There are innumerable reference times, for $h(\varepsilon)RV^{-1}$ will serve, for any function h . Whichever of these reference times is employed, it will still be the case, as in (2.9) and (2.12), that the dimensionless distance from the earth depends only on a dimensionless time and the single dimensionless parameter ε .

A priori it would seem advantageous in a given problem to compare lengths with an intrinsic reference length rather than with an arbitrary length such as the distance between two scratches on a certain metal bar. As our example illustrates, this is the case—and the advantage is that *the number of parameters which appear in a problem is reduced when dimensionless variables are employed.* (To be more precise, no parameters disappear, but they are seen to occur only in certain dimensionless combinations called *dimensionless groups*.)

Our example illustrates the fact that in all but the simplest problems, *non-dimensionalization can be carried out in a variety of ways*, each of which confers the same reduction in the number of parameters which appear. Is there a preferred set of nondimensional variables? The ensuing remarks will show that often there is.

We turn now to the matter of how to take advantage of the knowledge that a certain parameter is small. That a *dimensional* parameter is “small” is virtually without meaning, for the size of the parameter depends on the units of measurement. We are thus concerned with the presence of a small parameter in a set of dimensionless equations. It is natural to derive a simplified problem by passing to the limit when the small parameter is zero. This is not a straightforward matter, however, because in obtaining an appropriate simplified equation, *a term cannot be neglected merely because it is preceded by a small dimensionless parameter*. If this were *not* the case, then in the trajectory problem we would deduce from (2.6) that the lowest order approximation y_0 satisfies

$$(2.13) \quad -(y_0 + 1)^{-2} = 0, \quad y_0(0) = 0, \quad \frac{dy_0}{d\tau}(0) = 1.$$

But this problem has no solution. Using (2.11) on the other hand, the lowest order equation would be

$$(2.14) \quad \frac{d^2 z_0}{d\tau_1^2} = -\frac{1}{(z_0 + 1)^2}, \quad z_0(0) = 0, \quad \frac{dz_0}{d\tau_1}(0) = 0.$$

The solution here is negative, and can therefore only apply beneath the surface of the earth!

2.2. Definition of scaling. Since dimensionless variables can be chosen in a variety of ways, one cannot expect that the appearance of a small dimensionless parameter will inevitably signify the presence of a relatively small term. *Scaling* amounts to nondimensionalizing so that the relative magnitude of each term is indicated by a dimensionless factor preceding that term. More formally, *in the process of scaling one attempts to select intrinsic reference variables so that each term in the dimensional equations transforms into the product of a constant dimensional factor which closely estimates the term's order of magnitude and a dimensionless factor of unit order of magnitude*. (For the time being we shall use the phrase “order of magnitude” in the sense of “approximate size.” We shall shortly specify the meaning of this phrase more precisely.) Intrinsic reference variables which are selected by this process are called *scales*. Generally, scales differ for different parameter ranges. Also, we shall see that for a given range of parameters it may be necessary to choose different scales for different ranges of independent variables.

2.3. Scaling the trajectory problem. To illustrate the scaling procedure, let us consider the projectile problem in situations where the projectile's distance from the earth's surface x^* is always small compared to the earth's radius R . Such limited motion occurs only when the initial speed V is “sufficiently small.” On dimensional grounds we can assert that V must be small compared to some multiple of $(Rg)^{1/2}$, the only combination of parameters other than V with dimension length/time.

When $x^* \ll R$, it is clear that the acceleration has order of magnitude g , the gravitational acceleration at the earth's surface. Now if a projectile is launched with initial speed V and is then acted on by a *uniform* deceleration of magnitude g , the projectile will momentarily come to rest (at its maximum elevation) in time V/g . Taking the average of its initial and final speeds, we estimate that in time V/g it will move a distance equal to $(\frac{1}{2}V)(V/g) = \frac{1}{2}V^2/g$. To keep the factor $\frac{1}{2}$ in this expression might imply more accuracy than we can legitimately claim—remember that we have replaced the continuously changing speed of the particle by the average of its initial and final values and we have ignored the change of the force of gravity with distance. We thus take V^2/g as our estimate of the order of magnitude of x^* .

Remark. Our initial assumption that the parameters are such that x^* is always small compared to R is now seen to require that V^2/g be small compared to R . Because we have thought more deeply about the problem, it is not surprising that this is a more precise statement than the requirement “ V must be small compared to some multiple of $(Rg)^{1/2}$,” which is all one can deduce on dimensional grounds.

We have completed the most difficult part of the scaling procedure, estimation of the size of various terms in the special circumstances under consideration. We now show how to take formal advantage of the above estimates which show (i) that the displacement x^* has order of magnitude V^2/g and (ii) that the acceleration d^2x^*/dt^{*2} has order of magnitude g .

Using (i), the dimensionless displacement x should be defined by

$$(2.15) \quad x = x^*/(V^2/g^{-1}).$$

With this change of variable x^* will be replaced according to the equation

$$(2.16) \quad x^* = (V^2/g)x.$$

The first factor on the right side of (2.16), namely, (V^2/g) , correctly and explicitly shows the order of magnitude of x^* . As x^* is of order of magnitude V^2/g , (2.15) shows that the dimensionless factor x must have order of magnitude unity, as required by the scaling procedure.

From estimate (ii) we must choose a *time scale* T such that if the dimensionless time t , defined by

$$(2.17) \quad t = t^*/T,$$

is introduced into (2.4), then the term d^2x^*/dt^{*2} is transformed into the product of a dimensionless term and the constant g . But, from (2.15) and (2.17),

$$\frac{d^2x^*}{dt^{*2}} = \frac{V^2}{gT^2} \frac{d^2x}{dt^2}.$$

The requirement $V^2/(gT^2) = g$ gives $T = V/g$ as an equation defining the time scale.

With the appropriate scaled dimensionless variables

$$(2.18) \quad x = x^*/(V^2 g^{-1}) \quad \text{and} \quad t = t^*/(V g^{-1}),$$

(2.4) becomes

$$(2.19a) \quad g \frac{d^2 x}{dt^2} = -\frac{R^2}{(xV^2 g^{-1} + R)^2},$$

$$(2.19b) \quad x(0) = 0,$$

$$(2.19c) \quad \frac{dx}{dt}(0) = 1.$$

To simplify we divide both sides of (2.19a) by g , and divide the numerator and denominator of its right side by R^2 . This gives

$$(2.20) \quad \ddot{x} = -(1 + \varepsilon x)^{-2}; \quad x(0) = 0, \quad \dot{x}(0) = 1.$$

(Note that because of this division, it is *relative* orders of magnitude which are now explicitly displayed. In particular, the factor ε gives the order of magnitude of x^* divided by R .)

Since (2.4) has been nondimensionalized in the particular way called for by the scaling procedure, we are confident that when ε is small compared to unity the term εx is small compared to 1. If the lowest approximation to $x(t)$ is denoted by $x^{(0)}(t)$, we have from (2.20):

$$(2.21) \quad \ddot{x}^{(0)} = -1, \quad x^{(0)}(0) = 0, \quad \dot{x}^{(0)}(0) = 1.$$

Thus

$$x^{(0)} = t - \frac{1}{2}t^2 \quad \text{and} \quad 0 \leq x^{(0)} \leq \frac{1}{2} \quad \text{for} \quad 0 \leq t \leq 2,$$

so from (1.26) we find the error estimate

$$|x(t, \varepsilon) - x^{(0)}(t)| \leq \varepsilon + O(\varepsilon^2) \quad \text{for} \quad 0 \leq t \leq 2 + O(\varepsilon).$$

The inappropriateness of the previously obtained “approximate equations” (2.13) and (2.14) when $\varepsilon \ll 1$ can now be ascribed to the fact that in both (2.6) and (2.11) the nondimensionalization was not in accordance with the scaling procedure. That these incorrect “approximate” problems do not have a sensible solution is comforting, typical, but unfortunately not inevitable. (Recall Examples 1.4 and 1.5.)

We must emphasize that *the choice of scales depends on the parameter range under consideration*. As an illustration of this, observe that we have scaled the projectile problem when $\varepsilon \ll 1$, but if ε^{-1} is a small parameter, then new scaling is required. To mention one facet of the new situation, when ε^{-1} is small the initial speed is so large that the particle soon passes many earth radii from the earth’s surface and it becomes wrong to estimate that $d^2 x^*/dt^{*2}$ has magnitude g .

2.4. Order of magnitude. Further work will be facilitated if we use the phrase “order of magnitude” in a precisely defined way. A number A will be said to have *order of magnitude* 10^n , n an integer, if

$$5 \cdot 10^{n-1} < |A| \leq 5 \cdot 10^n.$$

By the *order of magnitude of a function* f defined over a certain region we mean the order of magnitude of the number M , where M is the maximum (or perhaps the least upper bound) of $|f|$ over the given region.

In some situations, the degree of accuracy required could make it expedient to base one's definitions on powers of three or on powers of 100 rather than on powers of 10. Also, note the distinction between the *order* of a function defined over a certain domain, denoted by the O symbol, and the numerical *order of magnitude* of a function as defined above.

2.5. Scaling known functions. Scaling of *known* functions is now a straightforward matter. It will be helpful to consider a few examples.

For definiteness let us at first consider a phenomenon which is governed by a single first order ordinary differential equation in which the independent variable x^* does not explicitly appear, say,

$$(2.22) \quad F(u^*, du^*/dx^*) = 0.$$

Suppose that x^* is restricted to an interval I^* (which may be infinite). It will be convenient to refer to u^* as a velocity and to imagine that x^* is a spatial variable, although our discussion applies to any dimensional dependent and independent variables u^* and x^* .

Let L be the *length scale* and U the *velocity scale*. Introduce the dimensionless velocity $u(x)$ by

$$(2.23) \quad x = x^*/L, \quad u = u^*/U.$$

We have

$$(2.24) \quad u^*(x^*) = Uu(x^*/L), \quad \frac{du^*}{dx^*} = \frac{U}{L} \frac{du(x)}{dx} \Big|_{x=x^*/L}.$$

If U and L are indeed appropriate scales, then the combinations of U and L which appear on the right side of the equations in (2.24) must be reasonable estimates for the maximum absolute values of the terms on the left. Now ordinary scales can usefully be regarded as estimates of *exact scales* in which U and U/L actually equal these maximum absolute values. For exact scales,

$$(2.25) \quad U = \max_{x^* \in I^*} |u^*(x^*)|$$

and

$$(2.26) \quad \frac{U}{L} = \max_{x^* \in I^*} \left| \frac{du^*}{dx^*} \right|.$$

From (2.26), using (2.25), we have

$$(2.27) \quad L \equiv \frac{|u^*|_{\max}}{|du^*/dx^*|_{\max}}.$$

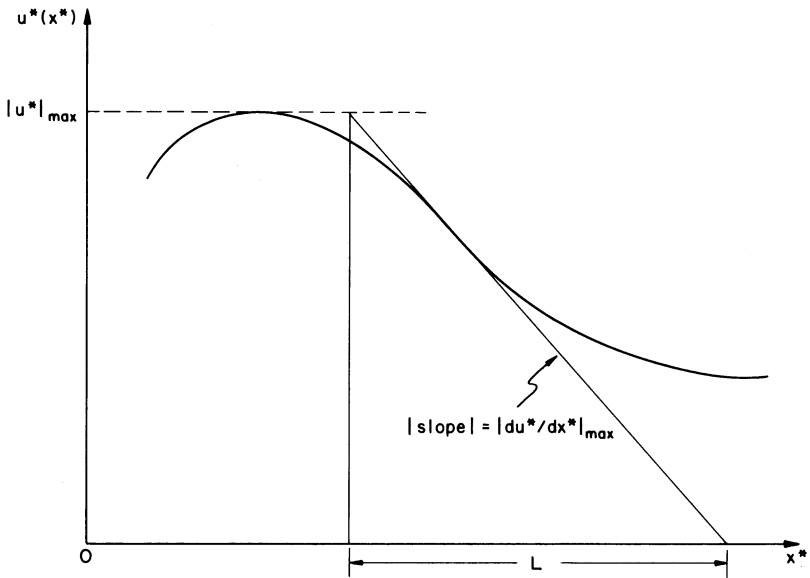


FIG. 1. The length scale L according to formula (2.27)

Equations (2.25) and (2.27) give explicit expressions for the exact velocity and length scales in problems governed by (2.22). Note from Fig. 1 that L can be interpreted as the base of a right triangle whose altitude is $|u^*|_{\max}$ and whose hypotenuse has the slope $|du^*/dx^*|_{\max}$.

Example 2.1. $u^*(x^*) = A \sin \lambda x^*$, $-\infty < x^* < \infty$; A and λ positive constants. Obviously $U = A$. Since

$$|du^*/dx^*|_{\max} = |A\lambda \cos \lambda x^*|_{\max} = A\lambda,$$

(2.27) implies $L = \lambda^{-1}$.

Example 2.2. $u^*(x^*) = A[x^* + \exp(-x^*/\varepsilon)]$, $0 \leq x^* \leq 1$; A and ε positive constants, $\varepsilon \ll 1$. Here

$$|u^*|_{\max} \approx A, \quad |du^*/dx^*|_{\max} = |A + A\varepsilon^{-1} \exp(-x^*/\varepsilon)|_{\max} \approx A\varepsilon^{-1},$$

so $L = \varepsilon$. See Fig. 2.

Suppose that the problem is governed by

$$(2.28) \quad F(u^*, du^*/dx^*, \dots, d^N u^*/dx^{*N}) = 0,$$

generalizing (2.18). Equation (2.25) is still an appropriate definition of the velocity scale U , but in choosing the length scale L we have to consider each of the derivatives

$$\frac{d^i u^*(x^*)}{dx^{*i}} = \frac{U}{L^i} \left[\frac{d^i u(x)}{dx^i} \right]_{x=x^*/L}, \quad i = 1, 2, \dots, N.$$

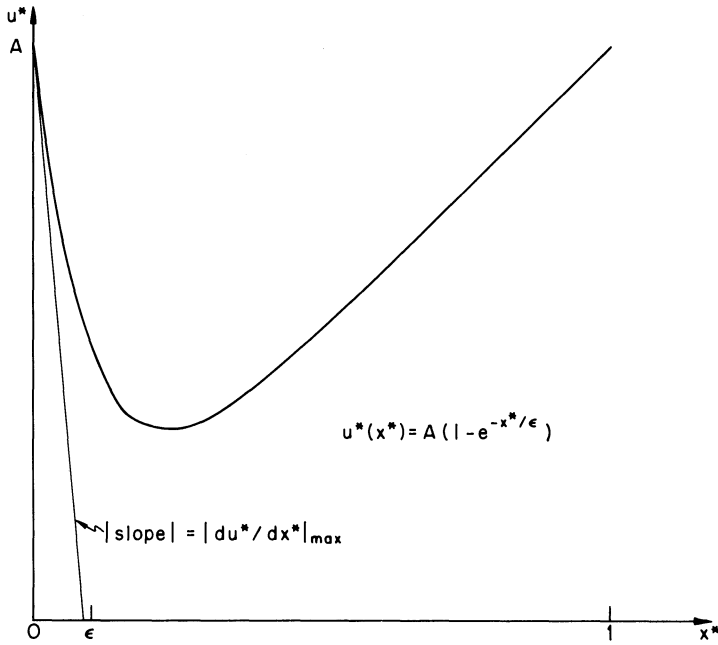


FIG. 2. The length scale L for the function u^* , where $u^*(x^*) = A[x^* + \exp(-x^*/\epsilon)]$; A and ϵ positive constants, $0 \leq x^* \leq 1$, $\epsilon \ll 1$.

We must choose L so that

$$(2.29) \quad \left| \frac{d^i u^*}{dx^{*i}} \right| \leq \frac{U}{L^i}, \quad i = 1, 2, \dots, N.$$

To make our estimate as sharp as possible we must select the largest value of L such that (2.29) is satisfied. It follows that

$$L \text{ is the largest constant such that } L \leq \left[\frac{U}{|d^i u^*/dx^{*i}|} \right]^{1/i}$$

for $i = 1, 2, \dots, N$; or

$$L = \text{smallest of } \frac{U}{|du^*/dx^*|_{\max}}, \left[\frac{U}{|d^2 u^*/dx^{*2}|_{\max}} \right]^{1/2}, \dots, \left[\frac{U}{|d^N u^*/dx^{*N}|_{\max}} \right]^{1/N}.$$

Example 2.3. For the sinusoid of Example 2.1,

$$(2.30) \quad \left[\frac{U}{|d^i u^*/dx^{*i}|_{\max}} \right]^{1/i} = \left[\frac{A}{A\lambda^i} \right]^{1/i} = \lambda^{-1},$$

so $L = \lambda^{-1}$ regardless of how many derivatives are involved in the governing equation (2.28).

The absence of x^* in (2.28) is characteristic of a wide class of *spatially homogeneous* problems. For such problems, "one place is just as good as another." Formally, spatial homogeneity manifests itself in the fact that the governing

equation, having no explicit appearance of the independent variable x^* , retains its form when subject to the axis translation $x^* \rightarrow x^* + \text{const.}$ As would be expected, length scales for problems *not* characterized by spatial homogeneity must take into account the variations of the inhomogeneity. For homogeneous problems, however, we can conclude that the length scale (λ^{-1}) of a sinusoidal function of zero mean value is approximately 1/6 of its period ($2\pi\lambda^{-1}$).

Example 2.4. The function considered in Example 2.3 has a length scale which is independent of N , the number of derivatives which must be taken into account. An example of a function whose length scale depends upon N is

$$u^*(x^*) = M + A \sin \lambda x^*, \quad -\infty < x^* < \infty,$$

M , A , and λ positive constants. Here

$$U = M + A, \quad \left[\frac{U}{|d^i u^*/dx^{*i}|_{\max}} \right]^{1/i} = \frac{1}{\lambda} \left[1 + \frac{M}{A} \right]^i, \quad \text{so } L = \frac{1}{\lambda} \left[1 + \frac{M}{A} \right]^{1/N}.$$

The dependence on N is weak providing M/A is not large compared to unity. When M is large, the length scale is large.

Our discussion may be extended to problems where there is more than one dependent variable or more than one independent variable. In the latter case, scales are assigned to independent variables by treating each separately. In the former case, the scale assigned to a given independent variable must take into account derivatives of every dependent variable.

2.6. Orthodoxy. Two matters of concern remain even after the process of scaling enables one to put a problem into a form which explicitly reveals the presence of terms (if any) having relatively small orders of magnitude. The first matter is that neglect of relatively small terms may have a large effect. This was discussed in § 1. In what follows we shall assume that relatively small terms are negligible.

The second matter of concern stems from the fact that the order of magnitude of a term estimates that term's *maximum* magnitude. If the magnitudes of the various terms in an equation stray too far from their maximum values, then the order of magnitude estimates may give misleading impressions of their relative sizes over much of their domain of definition. We say that a term in an equation satisfies the *orthodoxy requirement* in a given domain if the term's absolute value does *not* differ drastically from its maximum absolute value except perhaps for a negligible portion of the domain in question. Suppose that in a certain equation, the order of magnitude of a term T_1 is much greater than that of a second term T_2 . If the first term fails to satisfy the orthodoxy requirement, then the first term may be smaller in absolute value than the second for much of the domain in question, even though the maximum absolute value of the first term is much greater than the maximum absolute value of the second term.

Difficulties due to unorthodoxy are less widespread than one might think. Unorthodoxy might seem almost inevitable, for example, when the same length scale must serve to characterize the change of several variables; but this is not the case. The reason seems to reside in the fact that the dependent variables are bound to combine so that certain differential equations are satisfied. It is difficult to

imagine how such variables can differ widely in behavior. Presumably this is why it commonly occurs in practice that a length scale selected by concentrating on just one of several dependent variables frequently serves well for all of them.

“Harmless” unorthodox functions occur in the study of small amplitude water waves [4]. According to linear theory, the velocity components, the (dynamic) pressure and their derivatives are unorthodox functions because they decay rapidly with depth. But all decay at exactly the same exponential rate, so the relative order of magnitude of the various terms is correctly estimated by examining what goes on near the water surface. Such behavior is to be expected whenever the problem can be regarded as governed by a system of linear equations with constant coefficients.

Another common situation where a degree of unorthodoxy is tolerable involves oscillatory terms of relatively large maximum amplitude. Thus, one would say that terms of unit order of magnitude in an equation can probably be neglected to first approximation if other terms are known to behave like sinusoids of large amplitude. This is so in spite of the fact that the sinusoidal terms are small near their zeros. Because there is almost no room for interesting behavior in a function which must pass smoothly from values specified on one side of a narrow region to *nearby* values which are specified on the other side, it would appear that these small regions of anomalous behavior can be ignored. Orthodoxy is present not merely because the region of anomalous behavior is relatively small but because disregard of the anomaly has a negligible effect.

An important failure of orthodoxy often occurs when dependent variables behave like the function

$$(2.31) \quad u^*(x^*) = A[x^* + \exp(-x^*/\varepsilon)], \quad 0 \leq x^* \leq 1,$$

whose graph appears in Fig. 2. Suppose that the term du^*/dx^* must be assessed. The rapid change of the exponential in (2.31) means that an estimate of $|du^*/dx^*|$ in an interval containing points near $x^* = 0$ is a gross overestimate for an interval not containing such points. For example, when $3\varepsilon \leq x^* \leq 1$,

$$|du^*/dx^*|_{\max} = A\varepsilon^{-1}e^{-3}$$

so $L = e^3\varepsilon$, a length scale more than ten times as large as ε , the length scale appropriate for $0 \leq x^* \leq 1$. As in the situation discussed in the previous paragraph, our present example contains only a narrow region of anomalous behavior. But here the anomaly cannot be disregarded, because there is a sufficiently rapid rate of change to give an appreciable effect even though the change is confined to a narrow interval.

To satisfy the orthodoxy requirement for an equation involving the function u^* of (2.31) and its first derivative, we must split $[0, 1]$ into two parts and choose a different length scale in each part. In the *outer region*, more than a few multiples of ε from $x^* = 0$, we have

$$|u^*|_{\max} \approx A \quad \text{and} \quad |du^*/dx^*|_{\max} \approx A,$$

so $U = A$ and $L = 1$. Introducing these scales we obtain

$$u = u^*/A, \quad x = x^*,$$

so

$$u(x, \varepsilon) \equiv A^{-1}u^*(x, \varepsilon) = x + e^{-x/\varepsilon}.$$

In the *inner region*, within a few multiples of ε from $x^* = 0$, we find

$$|u^*|_{\max} \approx A \quad \text{and} \quad |du^*/dx^*|_{\max} \approx A\varepsilon^{-1},$$

so $U = A$ and $L = \varepsilon^{-1}$. Using ξ and v for scaled variables in the inner region, we have

$$v = u^*/A, \quad \xi = x^*/\varepsilon,$$

so

$$v(\xi, \varepsilon) = A^{-1}u^*(\varepsilon\xi, \varepsilon) = \varepsilon\xi + e^{-\xi}.$$

To obtain a first approximation in the two regions we let $\varepsilon \rightarrow 0$, keeping the respective independent variables fixed. We obtain in the outer region,

$$(2.32) \quad u(x, \varepsilon) \approx x,$$

and in the inner region,

$$(2.33) \quad v(\xi, \varepsilon) \approx e^{-\xi}, \quad \text{so } u(x, \varepsilon) \approx e^{-x/\varepsilon}.$$

Our simple example illustrates that lack of orthodoxy in a given domain can be remedied by introducing subdomains in each of which orthodoxy is present. Different scales will be required in the different subdomains. Moreover, *in the presence of unorthodoxy one should not seek a single first approximation but rather should seek different first approximations in different subdomains.*

By our last point, as illustrated in (2.32) and (2.33), we have essentially answered the question under (2.3). It remains only to state that the approximation $u \approx 1$, which is equivalent to (2.3b), is only valid very near $x = 0$. This is to be expected, for lengths are measured with respect to ε^2 , a quantity which is small even compared to the width of the inner region. The limit $\varepsilon \rightarrow 0$, $\eta = \varepsilon^{-2}x$ fixed, gives an approximation which is valid only within a few multiples of ε^2 of $x = 0$; $u \approx 1$ is clearly the appropriate form for such an approximation. Nothing of interest is gained here by considering an "inner-inner" region of width $O(\varepsilon^2)$.

In the present instance we have advocated the introduction of different scales in different regions to overcome unorthodoxy, while in our earlier discussion of large amplitude sinusoids we suggested that the unorthodoxy be ignored. The region of unorthodoxy is narrow in both cases, but only in the present case does the function change rapidly and so accumulate considerable alteration.

An example of a situation in which more than one scale must be introduced is provided by the projectile problem when ε^{-1} is large. The scaling of (14) is still appropriate when the projectile is near the earth, but another scaling is required when the projectile is far from the earth. As above there is an inner region (thickness $\approx V^2/g$) and a much larger outer region (particle more than a few radii from the earth's surface).

It can happen that two scales coexist, as in $(5 + e^{-x}) \sin ex$. Under some circumstances it is profitable to recognize the simultaneous existence of more than one scale, e.g., in the slow change in the amplitude and frequency of lightly damped

oscillations. Under other circumstances it may be best to ignore the rapid variation or to average it out, even though this means that large errors will be made in approximating the derivatives of the answer. Blurring small scale irregularities in order to reveal major trends is the strategy of many successful phenomenological theories. For a recent example, see Drew's [3] use of averaging methods to obtain field equations for two-phase problems in mechanics.

2.7. Perturbation theory. Once a problem has been correctly scaled, one can in principle derive arbitrarily accurate approximations by systematic exploitation, via perturbation theory, of the presence in the equations of a small parameter. From the present point of view, regular perturbation theory deals with orthodox situations where one set of scales suffices. Then, in particular, one first approximation holds throughout the domain of interest. Singular perturbation theory has been developed to handle unorthodox situations where more than one scale must be introduced in order to obtain an approximation which is uniformly valid throughout the domain.

It is not appropriate to enter into an extended discussion of perturbation theory, but a few remarks are in order. First, experience indicates that familiarity with scaling, as we have presented it, makes it relatively easy to assimilate the basic ideas of perturbation theory. The introductory remarks of the previous paragraph can serve as a sample of the sort of explanation which is readily comprehensible.

A second point stems from the observation that the choice of boundary layer scales (or, equivalently, the choice of stretching transformation) has been raised to a fine art by practitioners of singular perturbation theory. The art does not seem to proceed from contemplation of the physical problem, but rather from relatively abstract considerations of what is necessary to construct a uniformly valid solution. (We can speak of *abstract scaling* as opposed to *physical scaling*.) What then is the place of a physically-based attempt to select appropriate scales? One answer is that abstract expertise on choosing boundary layer scales has been acquired by generalizing singular perturbation solutions which *have* been obtained with the help of physical reasoning and of experimental knowledge of the phenomenon.

For the simple projectile problem, abstract scaling easily provides an exit from the paradoxes we presented. Faced with the failure of (2.6) to furnish a meaningful approximation for small ε , for example, a trained analyst would immediately introduce new variables such as $y = \varepsilon^a x$, $\tau = \varepsilon^b t$, with which (2.6) becomes

$$\varepsilon^{1+a-2b} \frac{d^2 x}{dt^2} = -\frac{1}{(1 + \varepsilon^a x)^2};$$

$$x(0) = 0, \quad \varepsilon^{a-b} \frac{dx}{dt}(0) = 1.$$

The second derivative term must be retained if the two initial conditions are to be satisfied. Thus $1 + a - 2b = 0$. Initially at least, the speed must certainly be of the same magnitude as the initial speed, so $a - b = 0$. Thus $a = b = 1$ —and we have arrived at the correctly scaled equation (2.18) by an essentially abstract approach.

It is no surprise that a routine application of abstract scaling techniques will provide a quick solution to a simple problem. For complicated perturbation problems at the frontiers of our technical ability, however, it may not be feasible to examine all possible distinguished limiting equations, to select appropriate matching conditions, and thereby to piece together a uniformly valid solution. For such problems, physical scaling provides an extra weapon which can turn failure into success.

2.8. Scaling unknown functions. Order of magnitude estimation and scaling require knowledge of the main features of the solution to the very problem one is trying to solve.

How can one obtain the information necessary for the required order of magnitude estimates? We list six possibilities.

(i) Utilize experimental or observational evidence concerning the phenomena in question.

(ii) Obtain hints from experience of related problems.

(iii) Solve highly simplified versions of the given problem. (This was illustrated in our discussion of the projectile problem where we used knowledge of the answer obtained when variation of gravitational force with distance is neglected. Knowing the zeroth approximation, we employed scaling to adapt the problem for efficient determination of higher approximations.)

(iv) (Inverse procedure). Make certain order of magnitude assumptions merely because the concomitant neglect of terms renders the problem tractable. (For example, assume that nonlinear terms are negligible.) Evaluate the neglected terms using the approximate solution. If possible, select parameter ranges so that these terms do indeed appear to be negligible. Hope that these parameter ranges are characteristic of interesting phenomena, and that the demonstrated lack of inconsistency signals justified neglect of small terms.

(v) Use a trial and error approach. Assume a certain scaling, solve the resulting simplified dimensionless problem, and then check to see whether the dimensionless terms are of unit order of magnitude.

(vi) Employ the results of numerical calculations for particular but representative values of the various parameters involved in the problem.

The last item forms the basis for a joint numerical-analytical attack on difficult problems. Such problems may involve many dimensionless parameters and two or three spatial dimensions. To follow the temporal evolution of the solution even once by numerical methods may be extremely demanding of computer time. There may thus be no possibility of thoroughly understanding the effect of parameter variation by using a direct numerical attack. But if a few numerical solutions are available, order of magnitude estimates can often be made which provide the key to the simplifications required in an analytic investigation. If successful, such an investigation will exhibit the dependence of the solution on the various parameters and will also reveal the general features of the solution which form the basis for physical understanding.

Acknowledgments. I have benefited from discussions with G. Habetler, E. F. and J. B. Keller, C. C. Lin, L. Rubinfeld, M. Shimshoni, and W. Siegmann.

REFERENCES

- [1] G. BIRKHOFF, *Hydrodynamics*, Dover, New York, 1955.
- [2] P. BRIDGMAN, *Dimensional Analysis*, Yale University Press, New Haven, Conn., 1931.
- [3] D. DREW, *Averaged field equations for two-phase media*, *Studies in Appl. Math.*, 50 (1971), pp. 133–166.
- [4] H. LAMB, *Hydrodynamics*, 6th ed., Dover, New York, 1945.
- [5] C. LANCZOS, *Applied Analysis*, Pitman, London, 1957.
- [6] J. WILKINSON, *Rounding Errors in Algebraic Processes*, Her Majesty's Stationery Office, London, 1963.
- [7] L. COLLATZ, *Functional Analysis and Numerical Methods*, Academic Press, New York, 1966.
- [8] E. ISAACSON AND H. KELLER, *Analysis of Numerical Methods*, John Wiley, New York, 1966.