SOLVING ODES IN QUASI STEADY STATE

L. F. Shampine
Applied Mathematics Research Department
Sandia National Laboratories
Albuquerque, New Mexico 87185, U.S.A.

MASTER

Abstract

Solving ordinary differential equations (ODEs) with solutions in a quasi steady state has been studied by computational chemists, applied mathematicians, and numerical analysts. Because of this, it is a very appropriate topic for this interdisciplinary workshop.

In this paper we shall first discuss what stiffness is for model problems arising in chemical kinetics. Chemists and applied mathematicians have made use of quasi steady state approximations (singular perturbation theory) to alter the problem so as to avoid stiffness. The approach is described and some difficulties noted. Numerical analysts have developed methods to solve general stiff ODEs. How they relate to the problem at hand is described and some difficulties pointed out. Finally, ideas from both approaches are combined. The new combination deals effectively with stiffness when the quasi steady state hypothesis is valid.

DISCLAIMER

This book was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Distribution of this document is unlimited eb

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

We are concerned with the integration of a system of N first order ordinary differential equations which in vector form are

$$(1) \quad \frac{d\mathbf{y}}{dt} = \dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y}), \mathbf{y}(t_0) \text{ given}$$

We use underlining to denote vectors and dots to denote differentiation by the independent variable t. We shall refer to t as the "time." When no confusion seems likely, we shall suppress t in subsequent expressions. The time behavior of the concentration of chemical species reacting according to the mass action law is an important example we shall refer to continually. As formulated by Edsberg [1], the equations are

$$(2) \quad \dot{\mathbf{y}} = \mathbf{A}\mathbf{p}$$

where A is a rectangular matrix with entries $A_{ij} = q_{ij} - r_{ji}$. Here the q_{ij} and r_{ji} are non-negative integers. The components p_j are the rate functions given by

$$(3) \quad p_j = k_j \prod_i y_i^{r_{ji}}$$

where k_j is a rate constant. For some purposes terms are grouped so that

$$(4) \quad \dot{y}_i = P_i(t, \mathbf{y}) - Q_i(t, \mathbf{y})y_i \quad i = 1, \dots, N.$$

The term P_i is the rate of production of y_i and the other term, the rate of loss. For (2), P_i and Q_i do not depend on t and P_i does not depend on y_i .

From the first numerical solutions of (2) on, puzzling difficulties were observed. It was recognized that the difficulties were often associated with very reactive species for which the production and loss rates were large but the time rates of change of the concentrations of the species were small -- they were nearly in equilibrium. This corresponds, at least for a short time interval, to P_i and Q_i being roughly constant and $Q_i \gg 1$. We shall formulate a hypothesis about the behavior of the coefficients of a system (4).

Quasi steady state hypothesis (QSSH):

For at least one i in (4), P_i and Q_i are roughly constant and $Q_i \gg 1$.

We are led to consider how numerical methods behave when applied to a set of equations

$$(5) \quad \dot{y}_i = P_i - Q_i y_i \quad i = 1, \dots, N$$

with all the P_i and Q_i constant and $Q_i \neq 0$.

The forward Euler method is representative of classical methods like Runge-Kutta and Adams. A Taylor series expansion says that for any smooth function

$$(6) \quad y_i(t_n+h) = y_i(t_n) + h\dot{y}_i(t_n) + \frac{h^2}{2}\ddot{y}_i(t_n) + \dots$$

We wish to proceed from an approximation $y_{i,n}$ of the solution $y_i(t_n)$ to an approximation of $y_i(t_n+h)$, thus advancing the numerical solution one time step of length h . Euler's method uses the linear terms of the Taylor series. The fact that we are talking about a solution of a differential equation (1) appears when we use the relation

$$\dot{y}_i(t_n) = f_i(t_n, y(t_n)).$$

Thus Euler's forward method is

$$(7) \quad y_{i,n+1} = y_{i,n} + hf_i(t_n, y_{i,n}) \quad i = 1, \dots, N$$

It is plausible, and essentially true, that the error of (7) is the first term omitted in (6). It is called the local (truncation) error. The typical code receives from its user a desired accuracy tolerance and it attempts to adjust h so as to keep the magnitude of the local error less than this tolerance.

For the model problem (5) it is easy to understand what local error control means. First we notice that the equations have a steady state (equilibrium) solution with $\dot{y}_i \equiv 0$, namely $s_i = P_i/Q_i$. The solution of (5) with $y_i(t_n) = y_{i,n}$ is

$$y_i(t) = s_i + [y_{i,n} - s_i] \exp(-Q_i(t-t_n)) \quad \text{for } t > t_n.$$

For $Q_i > 0$ there is a boundary layer (transition region) of rapid change as the solution approaches the steady state. The step size h is chosen so that

$$|\frac{h^2}{2}\ddot{y}_{i,n}| < \text{given constant}$$

at each step. Here

$$\ddot{y}_{i,n} = [y_{i,n} - s_i] Q_i^2.$$

If the numerical solution $y_{i,n}$ is far from the steady state, we see that

$$(8) \quad h \sim \frac{1}{|Q_i|}.$$

As the steady state is approached, the step size permissible becomes arbitrarily large. In both cases this is just what we would expect from the behavior of the solution.

What about the stability of the method? The formula is

$$y_{i,n+1} = y_{i,n} + h[P_i - Q_i y_{i,n}].$$

A perturbation of $y_{i,n}$ by $\delta_{i,n}$ leads to a perturbation of $y_{i,n+1}$ by $[1 - hQ_i] \delta_{i,n}$. Numerical stability of the formula requires that perturbations not be amplified, hence that $|1 - hQ_i| < 1$. This means a condition like (8) for all t ! This is extremely frustrating because we must use small step sizes for $Q_i \gg 1$ even when the solution is barely changing.

The stability restriction of Euler's forward method is typical of classical numerical methods and is one manifestation of stiffness. If the QSSH is valid, the model analysis says we shall suffer from stiffness. We may suffer from stiffness if the QSSH is not valid, but in the context of chemical kinetics it seems from experience that stiffness is usually accompanied by the QSSH.

One way to avoid stiffness is to change the problem. With the model problem it is pointless to continue integrating a solution component nearly in steady state. If one replaces the differential equation by an algebraic equation expressing equilibrium, he arrives at the QSS approximation (QSSA). In detail, if the QSSH holds for equation i

$$\dot{y}_i = P_i(t, \underline{y}) - Q_i(t, \underline{y})y_i,$$

it is replaced by

$$0 = P_i(t, \underline{y}) - Q_i(t, \underline{y})y_i.$$

The package [2] is an example of this for chemical kinetics. More generally, applied mathematicians have encountered many physical problems for which there is a natural parameter $\epsilon > 0$ and the differential equation (1) can be written in partitioned form as

$$(9) \quad \begin{aligned} \dot{\underline{u}} &= \underline{g}(\underline{u}, \underline{v}), \\ \epsilon \dot{\underline{v}} &= \underline{G}(\underline{u}, \underline{v}), \end{aligned}$$

and the solution $\underline{y}(t; \epsilon) = (\underline{u}(t; \epsilon), \underline{v}(t; \epsilon))^T$. It is natural to consider the "reduced" problem

$$(10) \quad \begin{aligned} \dot{\underline{u}} &= \underline{g}(\underline{u}, \underline{v}), \\ 0 &= \underline{G}(\underline{u}, \underline{v}), \end{aligned}$$

with solution $\underline{y}(t; 0) = (\underline{u}(t; 0), \underline{v}(t; 0))^T$. The relation between the problems for

$\epsilon > 0$ and $\epsilon = 0$ can be quite complex. We shall be interested in those problems for which $\underline{y}(t; \epsilon) \rightarrow \underline{y}(t; 0)$ on some time interval of interest as $\epsilon \rightarrow 0$. The solution $\underline{y}(t; 0)$ is the "outer" approximation for $\underline{y}(t; \epsilon)$ in a singular perturbation analysis. Of course the solution of the reduced problem cannot, in general, satisfy all the initial conditions of the full problem so it cannot be expected to be a good approximation in an initial boundary layer.

The QSSA is an outer approximation where we can identify ϵ as a characteristic magnitude for those Q_i with equation i satisfying the QSSH. Robertson [3] argues that the chemical kinetic equations (2) have the form (9) where the variables $\underline{y}$ are those species participating in "fast" reactions.

In the package [2] the user designates which species react "fast," and they are assumed in equilibrium for all time.

The QSSA has been widely used, but a number of serious objections have been raised. For one thing, how is the partition of fast and slow variables to be made? The reaction rates may span a wide range, but not fall into two distinct groups. A partition may change for nonlinear problems. Consider the famous van der Pol equation

$$\ddot{x} - \mu(1-x^2) \dot{x} + x = 0.$$

Written as a system with $y_1 = x$, $y_2 = \dot{x}$ we have

$$\dot{y}_1 = y_2$$

$$\dot{y}_2 = -y_1 + \mu(1-y_1^2)y_2$$

We are interested in relaxation oscillations with the parameter $\mu \gg 1$. Plots of the solution with $\mu = 1000$ can be found in [4,p.326]. This problem has the form (4), but the sign of Q_2 is wrong when $1-y_1^2 < 0$. All solution curves approach rapidly a steady oscillation which has a maximum of about 2. The QSSH is valid and the solution y_1 slowly decreases to about 1, at which point the character changes and there is an almost discontinuous drop to about -2. Once again the QSSH is valid and there is a slow increase of y_1 to about -1, followed there by an almost discontinuous jump to about +2. It is clear that application of the QSSA is not straightforward.

It is not apparent how to estimate the error of the QSSA even when it is applicable. It is still less clear how to handle boundary layers. The chemical kinetic

equations allow linear conservation laws, e.g. mass balance. These are preserved exactly by the typical numerical methods but not by the QSSA.

Because of all these factors, using the QSSA sometimes leads to results which are qualitatively incorrect. Workers have devised variants designed to respond to some of the difficulties and research continues. As we shall see, numerical analysts in the meantime produced a reliable alternative.

Stiff differential equations arise in the most varied contexts, see e.g. the survey [5]. Numerical analysts have tried to respond to the need to solve such problems by developing formulas which do not suffer the stability limitations of the forward Euler. The backward Euler formula is an example. It is derived similar to the forward formula with the derivative taken at the other end of the step:

$$(11) \quad y_{i,n+1} = y_{i,n} + hf_i(y_{i,n+1}).$$

If $y_{i,n}$ is perturbed by $\delta_{i,n}$, the result $y_{i,n+1}$ for the model problem (5) is perturbed to

$$y_{i,n+1} + \frac{\delta_{i,n}}{1 + h Q_i}.$$

Thus the formula is stable for all $Q_i > 0$, and there is no restriction on the size of h .

The backward Euler and similar formulas do away with the stability limitation, more or less successfully, but a price must be paid. As (11) exemplifies, they are all implicit, i.e. a system of algebraic equations must be solved at each step for the new approximation $y_{i,n+1}$. The generic form is

$$(12) \quad \underline{z} = \underline{\psi} + h\gamma \underline{f}(\underline{z}) \text{ or } \underline{F}(\underline{z}) = \underline{z} - h\gamma \underline{f}(\underline{z}) - \underline{\psi} = \underline{0}.$$

Here $\underline{z}$ represents the new solution approximation, $\underline{\psi}$ comes from previously computed quantities, γ is a constant determined by the formula, and h and $\underline{f}$ have their usual meanings. Let $\underline{z}^*$ denote a solution of the algebraic equations (12).

At first implicit formulas were evaluated by simple or functional iteration

$$(13) \quad \underline{z}^{(m+1)} = \underline{\psi} + h\gamma \underline{f}(\underline{z}^{(m)}).$$

It is possible using past data to predict efficiently a starting approximation $\underline{z}^{(0)}$. For non-stiff problems the iteration is very efficient. If we apply it to the model problem (5), we easily find that

$$(z_i^{(m+1)} - z_i^*) = -h\gamma Q_i(z_i^{(m)} - z_i^*).$$

This means that we must have $|h\gamma Q_i| < 1$ to get convergence and the rate $R_1 = |h\gamma Q_i|$. We have not gained a thing, because once again we have a restriction on h like (8). The answer is to go to a Newton iteration scheme. If we linearize $\underline{f}(\underline{z})$ about some iterate $\underline{z}^{(m)}$, we are led to Newton's iteration

$$\underline{z}^{(m+1)} = \underline{\psi} + h\gamma \underline{f}(\underline{z}^{(m)}) + \mathcal{J}(\underline{z}^{(m)})(\underline{z}^{(m+1)} - \underline{z}^{(m)}).$$

Here $\mathcal{J}$ is the Jacobian matrix $\left(\frac{\partial f_i}{\partial y_j}\right)$. It is too expensive to reevaluate $\mathcal{J}$ at each iterate, so codes form an approximation J and use it as long as convergence is adequate. The resulting iteration is actually carried out in the form

$$(14) \quad M(\underline{z}^{(m+1)} - \underline{z}^{(m)}) = -\underline{F}(\underline{z}^{(m)})$$

where

$$(15) \quad M = I - h\gamma J.$$

Let

$$(16) \quad H(\underline{z}^*) = I - M^{-1}(I - h\gamma \mathcal{J}(\underline{z}^*)).$$

With some smoothness assumptions, if $\|H(\underline{z}^*)\|_\infty < r < 1$, then the iteration (14) converges for all $\underline{z}^{(0)}$ sufficiently close to $\underline{z}^*$ and the (linear) rate of convergence is at least r . Here

$$\|H\|_\infty = \max_i \sum_{j=1}^N |H_{ij}|.$$

Notice that if J is sufficiently close to $\mathcal{J}(\underline{z}^*)$ and (15) is used, the iteration converges and does so rapidly.

The procedure described is closely related to the singular perturbation approach. The algebraic equations (12) for the differential equations (9) assume the partitioned form

$$\underline{F}(\underline{z}) = \begin{pmatrix} \underline{z}_1 - h\gamma \underline{g}(\underline{z}_1, \underline{z}_2) - \underline{\psi}_1 \\ \underline{z}_2 - \frac{h\gamma}{\epsilon} \underline{G}(\underline{z}_1, \underline{z}_2) - \underline{\psi}_2 \end{pmatrix} = \underline{0}$$

Scaling the second set of equations by $\epsilon/h\gamma$ does not affect the solution, but reveals what is going on:

$$\begin{pmatrix} \underline{z}_1 - h\gamma \underline{g}(\underline{z}_1, \underline{z}_2) - \underline{\psi}_1 \\ \frac{\epsilon}{h\gamma} \underline{z}_2 - \underline{G}(\underline{z}_1, \underline{z}_2) - \frac{\epsilon}{h\gamma} \underline{\psi}_2 \end{pmatrix} = \underline{0}$$

If ϵ were 0, these would be the algebraic equations arising from application of the integration formula to the reduced problem (10). Thus the general approach does much the same thing as the singular perturbation approach when the latter is valid, but it is generally applicable and reliable.

The general approach to stiff ODEs described is quite effective, but the task is far more difficult and expensive than the solution of non-stiff problems. For one thing, it is necessary to approximate $\mathcal{J}(\underline{z}^*)$. This is often very expensive and/or a lot of trouble. For chemical kinetics it is neither. Recall that $\dot{\underline{y}} = A\underline{y}$ where A is constant. Because

$$\frac{\partial p_i}{\partial y_j} = r_{ij} \frac{p_i}{y_i}$$

it is easy and inexpensive to form the Jacobian. The storage required for the solution of the differential equation goes from a multiple of N for non-stiff problems to a multiple of N^2 because of the Jacobian. This can be very serious, even prohibitive.

The costs of the linear algebra, which have no analog for non-stiff problems, can be quite important. The main cost is that of decomposing M. Repeated solution of the linear systems in (14) can also be a significant amount of work.

The costs mentioned lead to a less satisfactory adaptation of step size h method (as reflected in γ) than in the non-stiff case. This is because changes of h and/or γ may force a factorization of a new M and possibly the formation of a new J.

For the reasons sketched the methods for stiff problems are much more expensive per step than the methods for non-stiff problems. Despite this, fantastic savings are possible for stiff problems because the step size can be so much larger. Naturally a great deal of research is being devoted to reducing still more the costs of these procedures.

The general approach furnishes a reliable, effective solution procedure, but takes no advantage of special structure. It is quite plausible that with equations in the production-loss form (4), one can do a much better job when the QSSH is valid. We shall suggest here ways to do this.

There are several ways to motivate the first two variations on a theme that we propose. For brevity we take a rather formal approach. First we note that the iteration matrix M does not have to arise directly from an approximation to $I - h_Y J$. It is only necessary that it be a reasonably good approximation. Second we note that most of the disadvantages listed for the general approach would disappear if M were a diagonal matrix. Thus formally we suggest a variant which uses the diagonal of the matrix arising from Newton's method:

V1: Take M in (14) to be $\text{diag}\{1 - h_Y \frac{\partial f_i}{\partial y_i}(\underline{z}^{(m)})\}$.

The iteration for solving (12) is then

$$(1 - h_Y \frac{\partial f_i}{\partial y_i}(\underline{z}^{(m)}))(z_i^{(m+1)} - z_i^{(m)}) = -F_i(\underline{z}^{(m)}).$$

This form is to be used mainly for theoretical purposes. A more practical variant is:

V2: Take M in (14) to be $\text{diag}\{1 - h_Y J_{ii}\}$.

Thus we are simply using the diagonal of the usual iteration matrix (15).

We shall state a convergence result, but first let us consider why it might suffice to use a diagonal matrix. We are interested in

$$(4) \quad \dot{y}_i = P_i(t, \underline{y}) - Q_i(t, \underline{y})y_i = f_i(t, \underline{y}) \quad i = 1, \dots, N.$$

The QSSH is that for some i , P_i and Q_i are roughly constant, which we take to mean that their partial derivatives are all "small." Also $Q_i \gg 1$. This means that

$$\text{QSSH:} \quad \frac{\partial f_i}{\partial y_i} \ll -1, \quad \left| \frac{\partial f_i}{\partial y_i} \right| \gg \left| \frac{\partial f_i}{\partial y_j} \right| \quad j \neq i.$$

That is, the diagonal element of the Jacobian dominates strongly the off-diagonal elements in a row corresponding to an equation for which the QSSH is valid.

The form (4) suggests another iteration which has been noticed by a number of authors. Namely, to solve (12), which is for (4)

$$F_i(\underline{z}) = z_i - \psi_i - h_Y [P_i(t, \underline{z}) - Q_i(t, \underline{z})z_i] = 0,$$

use the iteration

$$\text{V3:} \quad (1 + h_Y Q_i(t, \underline{z}^{(m)}))(z_i^{(m+1)} - z_i^{(m)}) = -F_i(\underline{z}^{(m)}).$$

If the QSSH is valid, V3 is essentially the same as V1, but perhaps more convenient.

The relationship is especially close for the chemical kinetic equations (2). In general

$$\frac{\partial f_i}{\partial y_i} = -\frac{\partial P_i}{\partial y_i} + \frac{\partial Q_i}{\partial y_i} y_i + Q_i.$$

For (2) we group so that $\partial P_i / \partial y_i = 0$. The QSSH assumes that $\partial Q_i / \partial y_i$ is "small." Moreover Robertson [3] has argued that the very reactive species for which the QSSH should be valid should also have y_i "small."

If we apply the convergence result already stated to the M of V1, we have the sufficient condition for convergence at rate r

$$\|H\|_{\infty} = \max_i \frac{\sum_{j \neq i} h \gamma |J_{ij}|}{|1 - h \gamma J_{ii}|} < r,$$

where we write

$$J_{ij} = \frac{\partial f_i(z^*)}{\partial y_j}$$

Simpler sufficient conditions are that for all $i = 1, \dots, N$

$$(17a) \quad \text{if } J_{ii} > 0, h \gamma \sum_j |J_{ij}| < r$$

$$(17b) \quad \text{if } J_{ii} < 0, \sum_{j \neq i} |J_{ij}| < r |J_{ii}|.$$

The first condition is the same kind of condition as arises from simple iteration, but the second is quite noteworthy because it does not restrict h at all. If the QSSH is valid for equation i , condition (17b) obviously holds, and the step size is not restricted. Notice that the magnitudes of the entries of the Jacobian in such a row do not matter, just that the diagonal entry be negative and dominate. In particular, for a singularly perturbed equation

$$\epsilon \dot{y}_i = f_i(t, \underline{y}),$$

the size of ϵ does not matter. If (17b) holds, $\epsilon \rightarrow 0$ does not make the algebraic problem harder to solve by the iteration V1. It is easy to establish similar results for V2 and V3.

Two different ways of using these results suggest themselves. Suppose we are using a general purpose code with $M = I - h \gamma J$. We can test (17) to see if the diagonal of M would suffice to yield the desired rate of convergence r with the step size h . If it will, we use V2. This avoids a matrix decomposition and the cost of repeatedly solving linear systems, and we can adjust the step size and formula as we please. Thus if the QSSH should at any time be valid, we can take

full advantage of it. If the QSSH is not valid, or is not responsible for the stiffness, the problem is solved in the usual, reliable way.

Alternatively one could ask for the problem in the form (4) and always use the iteration V3 with a suitable formula. The author is pursuing this matter with a code based on both explicit and implicit Runge-Kutta methods. For simplicity suppose we have an efficient, highly stable, implicit Runge-Kutta formula. Because V3 is essentially simple iteration for non-stiff problems, we can expect to do about as well with it as is possible for non-stiff problems. In addition, we solve efficiently a class of stiff problems. It is important to realize that it is the nature of the stiffness and not how stiff a problem is that counts in such a code. We regard such a code as an extension of classical Runge-Kutta codes. Young and Boris [6,7] have been concerned with reactive flow problems. Storage is so critical there that they believe the general purpose codes cannot be used. Furthermore, based on a combination of physical insight and experience, they believe that for many important problems, the QSSH is valid and that it is the source of the stiffness. They use a refined QSSA which overcomes some of the objections raised, but not all. The code outlined here removes all the objections Young and Boris raise in connection with general purpose methods and, of course, does not experience the difficulties of the QSSA. For the problems treated by Young and Boris, the approach presented here appears to be extremely attractive.

References

- [1] Edsberg, L., Integration package for chemical kinetics, in R. A. Willoughby, ed., Stiff Differential Systems, Plenum Press, New York, 1974.
- [2] Clark, R. L. and G. F. Groner, A CSMP/360 precompiler for kinetic chemical equations, Simulation, 19(1972) 127-132.
- [3] Robertson, H. H., Numerical integration of systems of stiff ordinary differential equations with special structure, in G. Hall and J. M. Watt, eds., Modern Numerical Methods for Ordinary Differential Equations, Clarendon Press, Oxford, 1976.
- [4] Villadsen, J., and M. L. Michelsen, Solution of Differential Equation Models by Polynomial Approximation, Prentice-Hall, Englewood Cliffs, NJ, 1978.
- [5] Bjurel, G., et al., Survey of stiff ordinary differential equations, Rept. NA 70.11, Dept. Inf. Proc. Comp. Sci., Royal Inst. Tech., Stockholm, 1970.
- [6] Young, T. R., and J. P. Boris, A numerical technique for solving stiff ordinary differential equations associated with the chemical kinetics of reactive-flow problems, J. Phys. Chem., 81 (1977) 2424-2427.
- [7] Young, T. R., CHEMEQ - a subroutine for solving stiff ordinary differential equations, Rept. NRL 4091, Naval Res. Lab., Washington, D.C., 1980.