Numerical Questions in ODE Boundary Value Problems

M.R.Osborne *

Contents

Abstract

Classical considerations of stability in ODE initial and boundary problems have been mirrored by corresponding properties (stiff stability, di-stability) in problem discretizations. However, computational categories are not precise, and qualitative descriptors such as "of moderate size" cannot be avoided where size varies with the sensitivity of the Newton iteration in nonlinear problems for example. Sensitive Newton iterations require close enough initial estimates. The main tool for providing this in boundary value problems is continuation with respect to a parameter. If stable discretizations are not available then adaptive meshing is needed to follow rapidly changing solutions. Use of such tools can be necessary in stable nonlinear situations not covered by classical considerations. Implications for the estimation

*Mathematical Sciences Institute, Australian National University, ACT 0200, Australia. mailto:Mike.Osborne@maths.anu.edu.au

problem are sketched. It is shown how to choose appropriate boundary conditions for the embedding method. The simultaneous method is formulated as a constrained optimization problem. It appears distinctly promising in avoiding explicit ODE characterization. However, its properties are not yet completely understood.

1 Introduction

The stability of the solution of the initial value problem (IVP) for systems of ordinary differential equations has been studied extensively. There is a corresponding theory for numerical schemes for estimating solutions of these problems but it possesses some important differences. These arise through the requirement to produce numerical results for problems which cannot fit the classical framework, and this leads to differences both in emphasis and requirement. Boundary value problems (BVP) involve a global statement which makes corresponding results more elusive, but problems which are similar in kind occur. The estimation problems is an inverse problem which arises in trying to quantify system properties using information obtained by observing solution trajectories, typically in the presence of noise. Important in applications, the requirement here is to clarify the role of intrinsic properties of the differential equation in the well determinedness or otherwise of the estimation problem solution.

The differential equation is written

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}\,(t, \mathbf{x}) \tag{1}$$

where $\mathbf{x} \in R^m$, $\mathbf{f} \in R^1 \times R^m \to R^m$, and the forcing function $\mathbf{f}$ is assumed to have any required degree of smoothness. Boundary conditions have the form

$$B_0\mathbf{x}(0) + B_1\mathbf{x}(1) = \mathbf{b} \tag{2}$$

where $\mathbf{b} \in R^m$ and B_0, $B_1 \in R^m \to R^m$. Rank conditions are necessary on B_0, B_1 for the existence of a unique solution (see (6)). Boundary conditions are separated if no row of (2) couples solution values at both boundary points. This formulation contains the IVP as the special separated case $B_0 = I$, $B_1 = 0$. The IVP is distinguished because relatively weak conditions are sufficient to guarantee a local solution. The general boundary formulation also suffices for multipoint problems which can be reduced to BVP form by mapping each of the subintervals into $[0, 1]$.

Conditions for the existence of solutions of the BVP can be obtained by first considering the linear differential equation

$$\frac{d\mathbf{x}}{dt} = A(t)\mathbf{x} + \mathbf{q}(t). \tag{3}$$

Associated with this equation are IVP's for the fundamental matrix $X(t, \xi)$ satisfying

$$\frac{dX}{dt} = A(t)X,\ X(\boldsymbol{\xi}, \boldsymbol{\xi}) = I, \tag{4}$$

and the particular integral $\mathbf{w}(t, \xi)$ satisfying

$$\frac{d\mathbf{w}}{dt} = A(t)\mathbf{w} + \mathbf{q}(t), \quad \mathbf{w}(\xi, \xi) = 0. \tag{5}$$

The solution of the BVP can now be written by supposition

$$\mathbf{x} = X(t, 0)\mathbf{x}(0) + \mathbf{w}(t, 0)$$

where $\mathbf{x}(0)$ must be chosen to satisfy the boundary value equations

$$(B_0 + B_1 X(1, 0))\,\mathbf{x}(0) = \mathbf{b} - B_1\mathbf{w}(1, 0). \tag{6}$$

This can be satisfied provided $B_0 + B_1X(1, 0)$ has a bounded inverse. Solution conditions in the nonlinear case can now be obtained by linearising around an assumed solution and applying the Newton-Kantorovich theory [6].

In the linear case knowledge of the fundamental matrix permits explicit solution representations to be written down. In particular, the Green's matrix is given by

$$\begin{aligned} G(t, s) &= X(t)\left[B_0X(0) + B_1X(1)\right]^{-1} B_0X(0)X^{-1}(s),\ t > s, \\ &= -X(t)\left[B_0X(0) + B_1X(1)\right]^{-1} B_1X(1)X^{-1}(s),\ t < s. \end{aligned}$$

Note that G is independent of the form of initial condition on the fundamental matrix provided only it has full rank. The size of the Green's matrix governs the sensitivity of the BVP solution to perturbations in $\mathbf{q}(t)$. This is clearly important in stability considerations. The quantity

$$\alpha = \max_{0 \le t,s \le 1} \|G(t, s\|_2 \tag{7}$$

is called the *stability constant*.

An important related problem is the *estimation problem*. Here the argument of the forcing function $\mathbf{f}$ in (1) contains also a vector of parameters $\boldsymbol{\beta} \in R^p$,

$$\mathbf{f} \leftarrow \mathbf{f}(t, \mathbf{x}, \boldsymbol{\beta}).$$

This vector of parameters is to be estimated using information gained by observing a solution trajectory in the presence of noise at a set of points $T_n = \{t_1, t_2, \cdots, t_n\}$. This information is assumed to have the form

$$\mathbf{y}_i = H\mathbf{x}(t_i, \boldsymbol{\beta}) + \boldsymbol{\varepsilon}_i,\ t_i \in T_n, \tag{8}$$

where $\mathbf{y}_i \in R^k$, $k \leq m$, $H : R^m \to R^k$ has rank k, and $\boldsymbol{\varepsilon}_i \sim N(0, \sigma^2 I)$ are independent samples from a random process. Differences with the BVP arise not only from the presence of the noise process, but also from a requirement that a sufficiently rich set of observations be available. A minimum condition assumed is that $nk > m$ so the problem is formally strictly over-determined. This means that the best that can be done in general is to seek a solution that minimizes a goodness of fit criterion with respect to the observed data. Here this criterion is assumed to be

$$F(\boldsymbol{\beta}) = \sum_{i=1}^{n} \|\mathbf{y}_i - H\mathbf{x}(t_i, \boldsymbol{\beta})\|_2^2 . \tag{9}$$

It can be interpreted either in a least squares or maximum likelihood sense. The resulting optimization problem can have two forms depending on the manner of generating the comparison functions $\mathbf{x}(t, \boldsymbol{\beta})$.

Embedding : Here boundary matrices B_0, B_1 are selected in order to embed the comparison function in a family of BVP's. Now the appropriate right hand side vector $\mathbf{b}$ becomes a vector of auxiliary parameters to be determined as part of the estimation problem. The selection of B_0, B_1 has to be specified a priori in a manner compatible with stability constraints on the problem, and this approach has the further disadvantage that a (in general nonlinear) BVP must be solved at each Newton iteration.

Simultaneous : The idea is to impose a discretized form of the differential equation as a set of equality constraints on the problem of minimizing F. Estimates of $\boldsymbol{\beta}$ and the state variables $\mathbf{x}(t_i, \boldsymbol{\beta})$ are then generated simultaneously by solving the resulting constrained optimization problem. Here it simplifies discussion to combine the state vector with the parameter vector $\mathbf{x}^T \leftarrow \left[\ \mathbf{x}^T \quad \boldsymbol{\beta}^T\ \right]$ and to augment the differential equation system with the additional equations $\frac{d\boldsymbol{\beta}}{dt} = 0$.

An important aspect of the estimation problem is the selection of the points $t_i \in T_n$. Here it is necessary to distinguish two classes of experiment.

1. One in which observations are not available outside a finite interval which is here assumed to be $[0,1]$. Full information requires that T_n can be generated for arbitrarily large n. These experiments are called *planned* if the sets T_n satisfy the condition

$$\frac{1}{n}\sum_{i=1}^{n} v(t_i) \to \int_0^1 v(t)d\rho(t),\ n \to \infty$$

for all $v(t) \in C[0,1]$. This requirement just reflects the requirement that the non negative density function ρ is associated with the mechanism which must be set in place to generate an unbounded number of observations.

2. The alternative situation is one in which observations on a trajectory for arbitrarily large time contain parametric information. An important class of these problems is the class of stationary processes. One problem here which can be posed as an estimation problem is that of determining frequencies. These are functions of the coefficients in an ODE with constant coefficients.

The stochastic aspects of the estimation problem can have an important bearing on the choice of T_n.

- The asymptotic analysis of the effects of noisy data on maximum likelihood estimates of the parameters shows that this gets small no faster than $O\left(n^{-1/2}\right)$ under planned experiment conditions. A higher rate $(O\left(n^{-3/2}\right))$ is theoretically possible in maximum likelihood estimates in the frequency estimation problem but direct maximization is not the way to obtain these coefficients [12].

- It is not difficult to obtain ODE discretizations that give errors at most $O\left(n^{-2}\right)$.

This suggests that the trapezoidal rule provides an adequate discretisation. Here it has the form:

$$\mathbf{c}_i(\mathbf{x}_c) = \mathbf{x}_{i+1} - \mathbf{x}_i - \frac{h}{2}\left(\mathbf{f}_{i+1} + \mathbf{f}_i\right) = 0, \quad i = 1, 2, \cdots, n-1, \qquad (10)$$

with $\mathbf{x}_i = \mathbf{x}(t_i, \boldsymbol{\beta})$, $\mathbf{x}_c$ the composite vector with sub-vector components $\mathbf{x}_i$, and h the discretization mesh spacing. It is known to be endowed with attractive properties [4].

2 ODE stability

2.1 Initial value problem stability

Consider first the stability of the IVP

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}\left(t, \mathbf{x}\right), \quad \mathbf{x}(0) = \mathbf{b}.$$

Stability here means that solutions with close initial conditions $\mathbf{x}_1(0),\ \mathbf{x}_2(0)$ remain close in an appropriate sense.

- Let $\|\mathbf{x}_1(t) - \mathbf{x}_2(t)\|$ remain bounded ($\rightarrow 0$) as $t \rightarrow \infty$. This corresponds to weak (strong) IVS. Note that systems with bounded oscillatory solutions may well be weakly stable but the classification may not be particularly useful.

- Here computation introduces the idea of stiff discretizations which preserve the stability characteristics of the original equation in the sense that decaying solutions of the differential equation are mapped onto decaying solutions of the computational problem. The advantage is that the computation does not have to follow rapidly decaying solutions in detail. This is one area where there are genuine nonlinear results - for example, Butcher's work on BN stability of Runge-Kutta methods [2].

However, not all relevant IVPs are stable. The classical BVP solution method of multiple shooting provides an example [10]. This requires computation of the multiple shooting matrix of the linearized equation:

$$\begin{bmatrix} -X(t_2,t_1) & I & & & \\ & -X(t_3,t_2) & I & & \\ & & \cdots & \cdots & \\ B_0 & & & & B_1 \end{bmatrix}.$$

The problem is that the IVP for computing $X(t_{i+1}, t_i)$ could well be unstable in both forward and backward directions when the BVP has a well determined solution. This does not mean progress cannot be made. This is a consequence of Dahlquist's famous "consistency + stability implies convergence as $h \rightarrow 0$" theorem [3] which does not require IVP stability. However, it's setting implies exact arithmetic. The problem for practical computation is a form of numerical instability. This occurs in trying to follow a decreasing solution in the presence of rapidly increasing solutions. Rounding error introduces components of these fast solutions and these will eventually swamp that required. Compromise is necessary. Here this takes the form of restrictions

on the length of the interval of integration. This control in multiple shooting is provided by the choice of the $\{t_i\}$. Multiple shooting in this form appears to require accurate computation of all solutions and this is potentially a weakness.

This discussion is readily illustrated in the constant coefficient case. Consider the ODE

$$\mathbf{f}(t, \mathbf{x}) = A\mathbf{x} - \mathbf{q}$$

If A is non-defective then weak IVS requires the eigenvalues $\lambda_i(A)$ to satisfy $Re\{\lambda_i\} \leq 0$ while this inequality must be strict for strong IVS.
A one-step discretization of the ODE (ignoring the $\mathbf{q}$ contribution) can be written

$$\mathbf{x}_{i+1} = T_h(A)\,\mathbf{x}_i.$$

where $T_h(A)$ is the amplification matrix. A stiff discretization requires the stability inequalities to map into the condition $|\lambda_i(T_h)| \leq 1$. For the trapezoidal rule

$$\begin{aligned} |\lambda_i(T_h)| &= \left|\frac{1 + h\lambda_i(A)/2}{1 - h\lambda_i(A)/2}\right|, \\ &\leq 1 \text{ if } Re\{\lambda_i(A)\} \leq 0. \end{aligned}$$

2.2 Boundary value problem stability

The generalisation of IVS that is appropriate for linear differential equations is provided by the property of *dichotomy*: The key paper discussing the computational context is de Hoog and Mattheij [5]. Here only a weak form of dichotomy is considered. It requires that there exists a projection P depending on the choice of X such that, given

$$S_1 \leftarrow \{XP\mathbf{w},\ \mathbf{w} \in R^m\}, \quad S_2 \leftarrow \{X(I - P)\mathbf{w},\ \mathbf{w} \in R^m\},$$

then for all s, t

$$\begin{aligned} \phi \in S_1 &\Rightarrow \frac{|\phi(t)|}{|\phi(s)|} \leq \kappa, \quad t \geq s, \\ \phi \in S_2 &\Rightarrow \frac{|\phi(t)|}{|\phi(s)|} \leq \kappa, \quad t \leq s. \end{aligned}$$

This is the structural property that connects linear BVP stability with the detailed behaviour of the range of possible solutions. However, the BVP is specified on a finite interval. This means that on that interval, provided the fundamental matrix is bounded, there is always a bounded κ. The additional

feature in the computational context is that a modest κ is required for $t, s \in [0,1]$. The key result is that if X satisfies $B_0X(0) + B_1X(1) = I$ then $P = B_0X(0)$ is a suitable projection in sense that for separated boundary conditions the choice $\kappa = \alpha$ is allowed where α is the stability constant. There is an intimate connection between stability and dichotomy. Dichotomy permits a form of generalisation of A stability to the BVP case.

- The dichotomy projection separates increasing and decreasing solutions of the differential equation. *Compatible* boundary conditions pin down rapidly decreasing solutions at 0, and rapidly increasing solutions at 1.
- The discretization needs similar property in order that the given boundary conditions exercise the same control on the discretized system.
- This requires solutions of the ODE which are rapidly increasing (decreasing) in magnitude be mapped into solutions of the discretization which are rapidly increasing (decreasing) in magnitude.

This property is called *di-stability* in [7]. They show that the trapezoidal rule is di-stable in the constant coefficient case. The argument is straight forward:

$$\lambda(A) > 0 \Rightarrow \left| \frac{1 + h\lambda(A)/2}{1 - h\lambda(A)/2} \right| > 1. \tag{11}$$

Example 1 *Mattheij suggested a problem which provides an interesting test of discretization methods. Consider the differential system defined by*

$$A(t) = \begin{bmatrix} 1 - 19\cos 2t & 0 & 1 + 19\sin 2t \\ 0 & 19 & 0 \\ -1 + 19\sin 2t & 0 & 1 + 19\cos 2t \end{bmatrix},$$

$$\mathbf{q}(t) = \begin{bmatrix} e^t\left(-1 + 19\left(\cos 2t - \sin 2t\right)\right) \\ -18e^t \\ e^t\left(1 - 19\left(\cos 2t + \sin 2t\right)\right) \end{bmatrix}.$$

Here the right hand side is chosen so that $\mathbf{z}(t) = e^t\mathbf{e}$ *satisfies the differential equation. The fundamental matrix displays two fast and one slow solution showing that this system exhibits strong dichotomy:*

$$X(t,0) = \begin{bmatrix} e^{-18t}\cos t & 0 & e^{20t}\sin t \\ 0 & e^{19t} & 0 \\ -e^{-18t}\sin t & 0 & e^{20t}\cos t \end{bmatrix}.$$

	$h = .1$			$h = .01$		
$\mathbf{x}(0)$	1.0000	.9999	.9999	1.0000	1.0000	1.0000
$\mathbf{x}(1)$	2.7183	2.7183	2.7183	2.7183	2.7183	2.7183

Table 1: Boundary point values - stable computation

	$h = .1$			$h = .01$		
$\mathbf{x}(0)$	1.0000	.9999	1.0000	1.0000	1.0000	1.0000
$\mathbf{x}(1)$	-7.9+11	2.7183	-4.7+11	2.03+2	2.7183	1.31+2

Table 2: Boundary point values - unstable computation

For boundary data with two terminal conditions, one initial condition, and right hand side chosen to match the exponential solution :

$$B_0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \; B_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \; \mathbf{b} = \begin{bmatrix} e \\ e \\ 1 \end{bmatrix},$$

the trapezoidal rule discretization scheme gives the results in Table 1. These computations are apparently satisfactory.

In contrast, posing two initial and one terminal condition:

$$B_0 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \; B_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \; \mathbf{b} = \begin{bmatrix} 1 \\ e \\ 1 \end{bmatrix}$$

gives the results in Table 2 The effects of instability are seen clearly in the first and third solution components.

The feature of this example is the role of di-stability. Consider the trapezoidal rule denominator in (11). This suggests large and spurious amplification is likely in case $h = .1$. However, this is not evident in the stable computation. However, the unstable case does show more influence of instability than the case $h = .01$. The small denominator in (11) suggests the likely explanation.

2.3 Nonlinear stability

In nonlinear problems stability becomes a property of the linear problem governing the behaviour of perturbations about a current trajectory. In this sense it is a local property. Stable nonlinear problems are associated with relatively slow perturbation growth. Such problems can be expected to have the property that Newton's method applied to solve the discretized problem

will have a reasonable domain of convergence. The linear IVP/BVP stability requirements are inflexible in the sense that solutions must not depart from the classification as increasing/decreasing. Important conflicting examples occur in the linearised equations associated with dynamical systems. These include solutions which

- can have a stable character - for example, limiting trajectories which attract neighboring orbits;
- and clearly switch between the increasing and decreasing modes of the linearised system in a manner characteristic of oscillatory behaviour and so cannot satisfy the linear IVP/BVP stability requirements.

Limit cycle behavior provides a familiar example that is of this type. Intriguingly it can share some of the properties of stationary processes in the sense that observations contain trajectory information for all t.

Example 2 *To exhibit limit cycle behaviour consider the FitzHugh-Nagumo equations:*

$$\frac{dV}{dt} = \gamma\left(V - \frac{V^3}{3} + R\right),$$
$$\frac{dR}{dt} = -\frac{1}{\gamma}\left(V - \alpha - \beta R\right).$$

Solution components for $\alpha = .2$, $\beta = .2$, $\gamma = 1$ are illustrated in Figure 1. Note that the positive and negative components of the individual cycles are not exact opposites.

Example 3 *The Van der Pol equation:*

$$\frac{d^2x}{dt^2} - \lambda\left(1 - x^2\right)\frac{dx}{dt} + x = 0.$$

This provides a difficult ODE example with difficulty increasing with λ. The solutions here are exactly periodic. Stability is illustrated by convergence of trajectories from nearby initial points to the limit cycle. In Figure 2 rapid convergence to the limit cycle for $\lambda = 1, 10$ computed using standard Scilab code is illustrated. Computational problems occur because of the need to follow rapidly changing trajectories in detail. The clustering of integration points near the approximately vertical section of each trajectory shows the importance of adaptive mesh selection. Note the rapid convergence to the limiting trajectory showing this is certainly a stable situation.

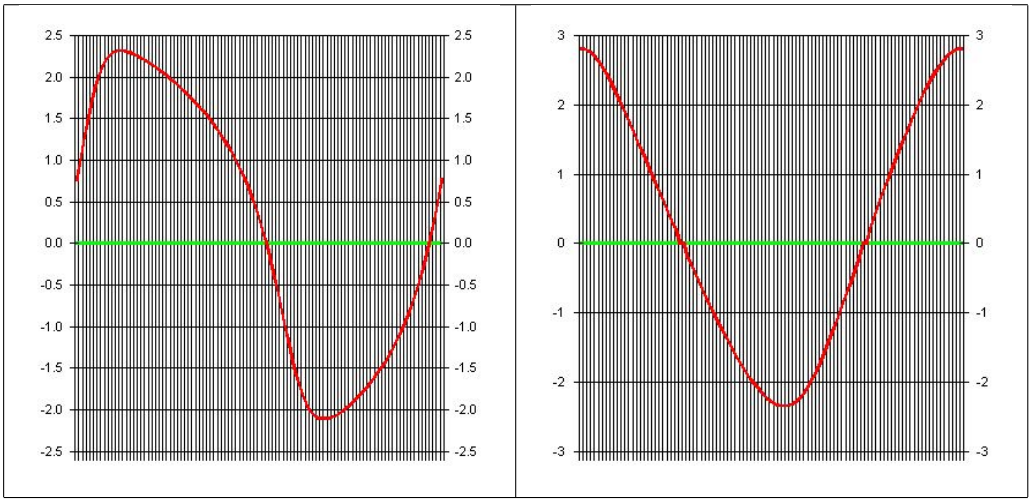

Figure 1: FitzHugh-Nagumo BVP solutions V, R (one cycle)

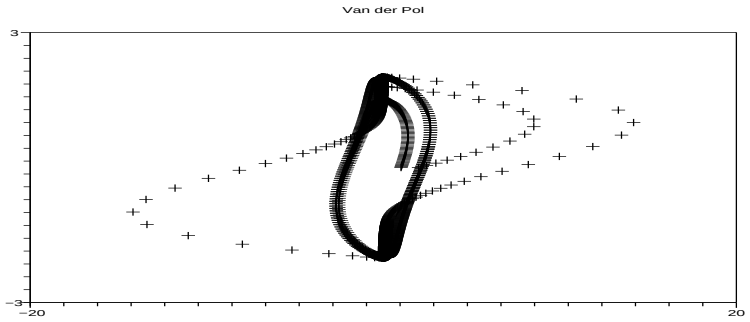

Figure 2: Scilab plot of Van der Pol trajectories for $\lambda = 1$, 10

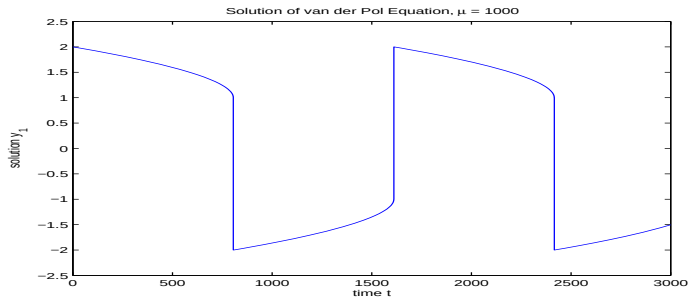

Figure 3: Matlab plot of state variable for $\lambda = 1000$, $x_1(0) = 2$

Matlab also uses this example in demonstration software but the output is less useful as it gives state information but not the derivative values for the case $\lambda = 1000$ *(Figure 3). This plot of a difficult case implies an excellent IVP solver. The starting values* $(2, 0)$ *used are rather special as:*

$$x_1(0) = 2 + \frac{1}{3}\alpha\lambda^{-4/3} - \frac{16}{27}\lambda^{-2}\ln(\lambda) + O(\lambda^{-2})$$

where $\alpha = 2.33811....$

Example 4 *The Van der Pol equation is exactly cyclic so the problem of computing a half cycle can be cast in BVP form on the interval* $[0, 2]$ *by making the transformation* $s = 4t/T$*. The unknown interval length can be treated as an additional variable by setting* $x_3 = T/4$*. The resulting ODE system becomes*

$$\begin{aligned}
\frac{dx_1}{ds} &= x_2, \\
\frac{dx_2}{ds} &= \lambda\left(1 - x_1^2\right)x_2x_3 - x_1x_3^2, \\
\frac{dx_3}{ds} &= 0.
\end{aligned}$$

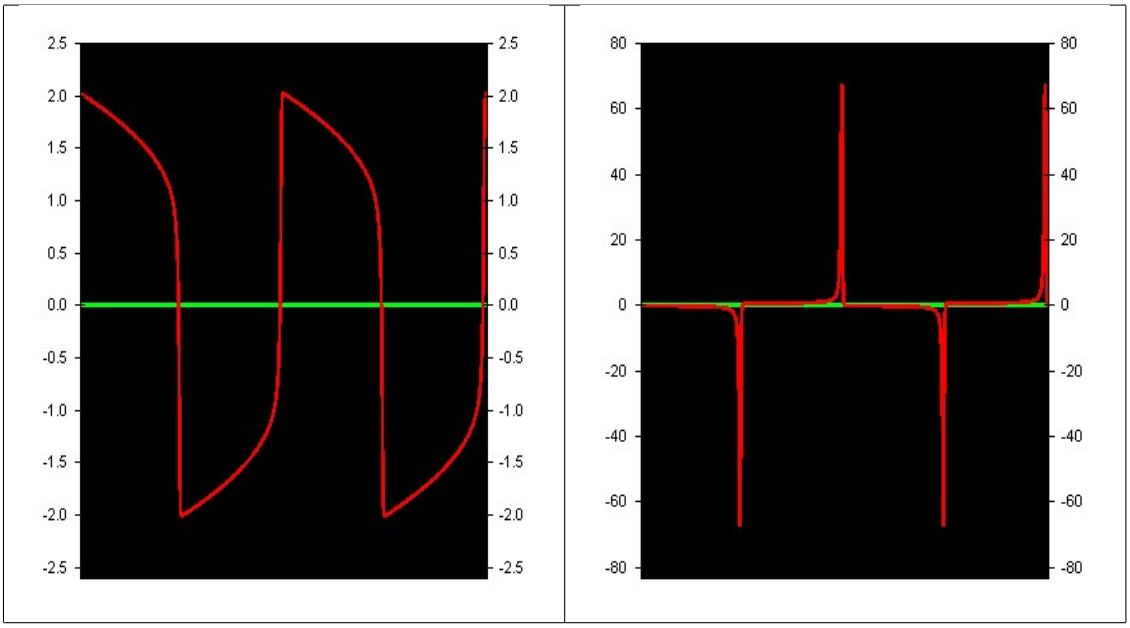

Figure 4: Van der Pol solution x_1, x_2 for λ=10

The appropriate boundary data is

$$B_0 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \; B_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \; \mathbf{b} = 0.$$

The nonlinear system has the trivial solution $\mathbf{x} = 0$ *so it is necessary to choose appropriate nonzero initial estimates. Here this has been done by taking the periodic solution for* $\lambda = 0$ *with* $x_1(0) = 2$, $x_2(0) = 0$ *as the initial estimate. This gives convergence for the Newton iteration for* $\lambda = 1$ *and continues to work for* $\lambda \leq 5$. *Continuation with* $\Delta\lambda = 1$ *is used for higher values. The fixed discretizations exemplified are* $h = 1/100, 1/1000$. *This is not ideal for this problem as the IVP computations have illustrated the importance of adaptive meshing. The BVP results for* $\lambda = 10$, $h = 1/1000$ *are given in Figure 4. These reinforce the need for the use of appropriately graded mesh selection.*

2.4 Stability consequences

The ODE stability conditions provide sharp distinctions - in part because they are specifying global properties. Computational requirements force compromise. In the IVP this is provided by various control devices: for example,

automatic step length control. There are two classes of computational stability problem:

- The difference approximation is unstable and does not satisfy the Dahlquist root condition $\rho(1) = 0$; $\rho(t) = 0$, $t \neq 1$, $\Rightarrow |t| < 1$. In this case errors grow with n and so are unbounded as $h \to 0$. This occurs whether or not the original problem is unstable.

- In unstable IVP's a computed slow solution will be swamped eventually as a result of the growth of rounding error induced perturbations which can grow like the Gronwall Lemma bound $\gamma \exp(Kt)$ in a worst case. This is the problem which multiple shooting seeks to control. This device appears not be necessary in stable BVP's if di-stable discretizations are used.

In the BVP stability discussion the dichotomy considerations are restricted to a finite interval on which we ask for "moderate" κ, arguing that it can be related to a bound for the Green's matrix and so directly relates to problem stability. Here the individual terms in the inverse of the multiple shooting matrix can then be interpreted using the Green's matrix. If κ is large then the BVP will be associated with a sensitive Newton iteration because the inverse Jacobian matrices must contain terms of $O(\kappa)$. Available tools for overcoming this problem include:

- use of adaptive mesh control in positioning discretization points - but this may be difficult if good initial estimates are not available;

- adaptive continuation with respect to a parameter in order to move from a known to a required BVP solution in a sequence of steps in which the current solution provides a good enough initial estimate for convergence of the Newton iteration at the next continuation increment.

3 The estimation problem

3.1 Estimation via embedding

The embedding form of the estimation problem (1) leads to a nonlinear least squares problem to minimize (9) for the unknown parameters $\boldsymbol{\beta}$, $\mathbf{b}$. This can be solved by an application of the Gauss-Newton method [9] once the boundary conditions needed to specify the embedding have been specified. This can be done by noting that a good choice should lead to a relatively

well conditioned linear system in setting up the linear least squares problem for the Gauss-Newton correction. To see what is involved note that the trapezoidal rule discretization of the differential equation (10) can be written in the form

$$\mathbf{c}_i\left(\mathbf{x}_i, \mathbf{x}_{i+1}\right) = \mathbf{c}_{ii}(\mathbf{x}_i) + \mathbf{c}_{i(i+1)}(\mathbf{x}_{i+1}).$$

Here the state variables enter additively. As a consequence the gradient system has the block bi-diagonal matrix C given by

$$C = \begin{bmatrix} C_{11} & C_{12} & & & \\ & C_{22} & C_{23} & & \\ & & & \cdots & \\ & & & \cdots & C_{(n-1)n} \end{bmatrix}. \tag{12}$$

Consider the orthogonal factorization of this system with the first column permuted to the last place:

$$\left[\begin{array}{cccc|c} C_{12} & & & & C_{11} \\ C_{22} & C_{23} & & & \\ \hline & & & C_{(n-1)(n-1)} & C_{(n-1)n} \end{array}\right] \rightarrow Q \left[\begin{array}{ccc|c} U & & & Z \\ \hline 0 & \cdots & D & G \end{array}\right]$$

This step is independent of the boundary conditions. It permits a solution representation of the form

$$\mathbf{x}_i = V_i \mathbf{x}_1 + W_i \mathbf{x}_n + \mathbf{w}_i,\ i = 2, \cdots, n-1. \tag{13}$$

with unknowns $\mathbf{x}_1$, $\mathbf{x}_n$. Factorization by orthogonal cyclic reduction associates $\{V_i,\ W_i,\ \mathbf{w}_i\}$ with solutions of the ODE system of twice the order [11]:

$$\left\{\frac{d}{dt} + A^T(t)\right\}\left\{\frac{d}{dt} - A(t)\right\}\mathbf{z} = \cdots$$

where A is the coefficient matrix of the linear ODE. Does a dichotomy result for this system follow from dichotomy for the original? The result is true for systems with constant coefficients. The orthogonal factorization suggested here is not the same as cyclic reduction but the performance appears similar.

If boundary conditions are inserted at this point there results a system for $\mathbf{x}_1$, $\mathbf{x}_n$ with matrix $\begin{bmatrix} D & G \\ B_1 & B_0 \end{bmatrix}$. Orthogonal factorization again provides a useful strategy.

$$\begin{bmatrix} D & G \end{bmatrix} = \begin{bmatrix} L & 0 \end{bmatrix} \begin{bmatrix} S_1^T \\ S_2^T \end{bmatrix}$$

It follows that the system determining $\mathbf{x}_1,\ \mathbf{x}_n$ is best conditioned by choosing

$$\left[\begin{array}{cc} B_1 & B_0 \end{array}\right] = S_2^T.$$

This choice of B_0, B_1 depends only on the ODE, but it does depend on $\boldsymbol{\beta}$ in the estimation problem as a consequence. It's use is illustrated in Example 5.

To set up the Gauss-Newton iteration let $\nabla_{(\beta,b)}\mathbf{x} = \left[\frac{\partial \mathbf{x}}{\partial \boldsymbol{\beta}}, \frac{\partial \mathbf{x}}{\partial \mathbf{b}}\right]$, $\mathbf{r}_i = \mathbf{y}_i - H\mathbf{x}\left(t_i, \boldsymbol{\beta}, \mathbf{b}\right)$. Then the gradient of F is

$$\nabla_{(\beta,b)} F = -2\sum_{i=1}^{n} \mathbf{r}_i^T H \nabla_{(\beta,b)} \mathbf{x}_i.$$

The gradient terms with respect to $\boldsymbol{\beta}$ are found by solving the BVP's

$$B_0 \frac{\partial \mathbf{x}}{\partial \boldsymbol{\beta}}(0) + B_1 \frac{\partial \mathbf{x}}{\partial \boldsymbol{\beta}}(1) = 0,$$
$$\frac{d}{dt}\frac{\partial \mathbf{x}}{\partial \boldsymbol{\beta}} = \nabla_x \mathbf{f} \frac{\partial \mathbf{x}}{\partial \boldsymbol{\beta}} + \nabla_\beta \mathbf{f},$$

while the corresponding terms with respect to $\mathbf{b}$ satisfy the BVP's

$$B_0 \frac{\partial \mathbf{x}}{\partial \mathbf{b}}(0) + B_1 \frac{\partial \mathbf{x}}{\partial \mathbf{b}}(1) = I,$$
$$\frac{d}{dt}\frac{\partial \mathbf{x}}{\partial \mathbf{b}} = \nabla_x \mathbf{f} \frac{\partial \mathbf{x}}{\partial \mathbf{b}}.$$

Example 5 *Consider the modification of the Mattheij problem with parameters $\beta_1^* = \gamma$, and $\beta_2^* = 2$ corresponding to the solution $\mathbf{x}\left(t, \boldsymbol{\beta}^*\right) = e^t \mathbf{e}$:*

$$A(t) = \left[\begin{array}{ccc} 1 - \beta_1 \cos \beta_2 t & 0 & 1 + \beta_1 \sin \beta_2 t \\ 0 & \beta_1 & 0 \\ -1 + \beta_1 \sin \beta_2 t & 0 & 1 + \beta_1 \cos \beta_2 t \end{array}\right],$$
$$\mathbf{q}(t) = \left[\begin{array}{c} e^t\left(-1 + \gamma\left(\cos 2t - \sin 2t\right)\right) \\ -(\gamma - 1)e^t \\ e^t\left(1 - \gamma\left(\cos 2t + \sin 2t\right)\right) \end{array}\right].$$

In the numerical experiments optimal boundary conditions are set at the first iteration. The aim is to recover estimates of $\boldsymbol{\beta}^*, \mathbf{b}^*$ from simulated data $e^{t_i} H\mathbf{e} + \boldsymbol{\varepsilon}_i$, $\boldsymbol{\varepsilon}_i \sim N(0, .01I)$ using Gauss-Newton, stopping when $\nabla F \mathbf{h} < 10^{-8}$. Results are given in Table 3.

Here the effect of varying $\boldsymbol{\beta}$, $\mathbf{b}$ proves negligible. The angle between the initial conditions and the optimal conditions for the subsequent values of $\boldsymbol{\beta}$, $\mathbf{b}$ is determined by $\| \left[\begin{array}{cc} B_1 & B_2 \end{array}\right]_1 \left[\begin{array}{cc} B_1 & B_2 \end{array}\right]_k^T - I\|_F < 10^{-3}$, $k > 1$. This example possesses a dichotomy so these results confirm the efficacy of the embedding method for stable problems.

$H = \begin{bmatrix} 1/3 & 1/3 & 1/3 \end{bmatrix}$	$H = \begin{bmatrix} .5 & 0 & .5 \\ 0 & 1 & 0 \end{bmatrix}$
$n = 51,\ \gamma = 10,\ \sigma = .1$	$n = 51,\ \gamma = 10,\ \sigma = .1$
14 iterations	5 iterations
$n = 51,\ \gamma = 20,\ \sigma = .1$	$n = 51,\ \gamma = 20,\ \sigma = .1$
11 iterations	9 iterations
$n = 251,\ \gamma = 10,\ \sigma = .1$	$n = 251,\ \gamma = 10,\ \sigma = .1$
9 iterations	4 iterations
$n = 251,\ \gamma = 20,\ \sigma = .1$	$n = 251,\ \gamma = 20,\ \sigma = .1$
8 iterations	5 iterations

Table 3: Embedding method: Gauss-Newton results for the Mattheij problem

3.2 Simultaneous estimation

The simultaneous method 1 leads to the optimization problem:

$$\min_{\mathbf{x}_c} F(\mathbf{x}_c); \quad \mathbf{c}_i(\mathbf{x}_c) = 0,\ i = 1, 2, \cdots, n-1, \tag{14}$$

where $\mathbf{x}_c \in R^{nm}$ is the composite vector with block sub-vectors $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$, and where the individual state and parameter vectors are bundled together to form composite state sub-vectors. Introducing the Lagrangian function

$$\mathcal{L} = F(\mathbf{x}_c) + \sum_{i=1}^{n-1} \boldsymbol{\lambda}_i^T \mathbf{c}_i.$$

permits the necessary conditions to be written:

$$\nabla_{\mathbf{x}_i} \mathcal{L} = 0,\ i = 1, 2, \cdots, n, \quad \mathbf{c}(\mathbf{x}_c) = 0.$$

The basic algorithmic approach involves the use of Newton's method or one of its variants to solve this nonlinear system. The resulting system determining corrections $\mathbf{dx}_c, \mathbf{d}\boldsymbol{\lambda}_c$ is:

$$\nabla^2_{\mathbf{xx}} \mathcal{L} \mathbf{dx}_c + \nabla^2_{\mathbf{x}\boldsymbol{\lambda}} \mathcal{L} \mathbf{d}\boldsymbol{\lambda}_c = -\nabla_{\mathbf{x}} \mathcal{L}^T, \tag{15}$$

$$\nabla_{\mathbf{x}} \mathbf{c}(\mathbf{x}_c)\, \mathbf{dx}_c = C \mathbf{dx}_c = -\mathbf{c}(\mathbf{x}_c), \tag{16}$$

where the block bidiagonal matrix C is defined in equation (12) The sparsity is a consequence of the trapezoidal rule. Here $\nabla^2_{\mathbf{xx}} \mathcal{L}$ is block diagonal while $\nabla^2_{\mathbf{x}\boldsymbol{\lambda}} \mathcal{L} = C^T$ is block bidiagonal. In [9] these equations are connected to necessary conditions for the solution of a quadratic program. This leads to consideration of two main solution approaches.

Elimination The constraint equations (16) can be solved for $\mathbf{dx}_i$, $i = 2, \cdots, n-1$ in terms of $\mathbf{dx}_1$ and $\mathbf{dx}_n$ as in equation (13). This permits the quadratic program to be reduced to a problem in just these variables with the constraint determined by the last row of the factored matrix. Second order sufficiency conditions must still be satisfied for this reduced problem. This is discussed in [8] and the references cited there. This approach has been tested for boundary value stable problems. Simpler elimination schemes are possible, but these correspond essentially to simple shooting [1] and cannot be boundary value stable.

Null space An alternative approach which does not depend on a boundary value formulation can be based on the factorization

$$C^T = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} U \\ 0 \end{bmatrix}.$$

Then the Newton equations can be written

$$\begin{bmatrix} Q^T \nabla^2_{\mathbf{xx}} \mathcal{L} Q & \begin{bmatrix} U \\ 0 \end{bmatrix} \\ \begin{bmatrix} U^T & 0 \end{bmatrix} & 0 \end{bmatrix} \begin{bmatrix} Q^T \mathbf{dx}_c \\ \mathbf{d\lambda}_c \end{bmatrix} = - \begin{bmatrix} Q^T \nabla_{\mathbf{x}} \mathcal{L}^T \\ \mathbf{c} \end{bmatrix}.$$

They can be solved in the sequence

$$U^T Q_1^T \mathbf{dx}_c = -\mathbf{c}, \tag{17}$$

$$Q_2^T \nabla^2_{\mathbf{xx}} \mathcal{L} Q_2 Q_2^T \mathbf{dx}_c = -Q_2^T \nabla^2_{\mathbf{xx}} \mathcal{L} Q_1 Q_1^T \mathbf{dx}_c - Q_2^T \nabla_{\mathbf{x}} \mathcal{L}^T, \tag{18}$$

$$U \mathbf{d\lambda}_c = -Q_1^T \nabla^2_{\mathbf{xx}} \mathcal{L} \mathbf{dx}_c - Q_1^T \nabla_{\mathbf{x}} \mathcal{L}^T. \tag{19}$$

Sufficient conditions are just the second order sufficiency conditions

1. The matrix C has full row rank so the linearised constraints are linearly independent.
2. The matrix $Q_2^T \nabla^2_{\mathbf{xx}} \mathcal{L} Q_2$ is nonsingular.

Remark 6 *The null space method does not depend explicitly on techniques associated with boundary value problem solution methods. Thus it is of interest to ask if it possesses wider stability tolerances. Discussion of the method's properties is complicated by the presence of the Lagrange multipliers for which initial estimates have to be provided. Typically this is done by computing the generalised inverse solution to the necessary condition*

$$C^T \boldsymbol{\lambda}_c + \nabla_{\mathbf{x}} F^T = 0$$

at the initial point. Note this equation has formal similarity to the discretization of the adjoint differential equation and so could connect the null space method back to stability questions.

test results $n = 11$	particular integral $Q_1^T x$
.87665 -.97130 -1.0001	.87660 -.97134 -1.0001
.74089 -1.0987 -1.3432	.74083 -1.0988 -1.3432
.47327 -1.2149 -1.6230	.47321 -1.2150 -1.6231
.11498 -1.3427 -1.8611	.11491 -1.3428 -1.8612
-.32987 -1.4839 -2.0366	-.32994 -1.4840 -2.0367
-.85368 -1.6400 -2.1250	-.85376 -1.6401 -2.1250
-1.4428 -1.8125 -2.1018	-1.4429 -1.8125 -2.1019
-2.0773 -2.0031 -1.9444	-2.0774 -2.0032 -1.9444
-2.7309 -2.2137 -1.6330	-2.7310 -2.2138 -1.6331
-3.3719 -2.4466 -1.1526	-3.3720 -2.4467 -1.1527

Table 4: Stability test: comparison of exact and computed values

Remark 7 *If the null space method is applied to the Mattheij problem with initial estimate* $\mathbf{x}_c = 0$ *then the first step solves* $C\mathbf{dx}_c = \mathbf{q}_c$. *It follows that* $Q_1^T \mathbf{dx}_c = U^{-T}\mathbf{q}_c$ *should estimate* $Q_1^T \mathrm{vec}\left\{ \exp(t_i) \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right\}$. *Computed and exact results are displayed in Table 4 in the interesting case* $h = .1$. *The results suggest that the null space method can exploit di-stability.*

References

[1] U.M. Ascher, R.M.M. Mattheij, and R.D. Russell, *Numerical solution of boundary value problems for ordinary differential equations*, SIAM, Philadelphia, 1995.

[2] J.C. Butcher, *Numerical methods for ordinary differential equations*, John Wiley and Sons, 2003.

[3] G. Dahlquist, *Convergence and stability in the numerical integration of ordinary differential equations*, Math. Scand. **4** (1956), 33–53.

[4] ———, *A special stability problem for linear multistep methods*, BIT **3** (1963), 27–43.

[5] F.R. de Hoog and R.M.M. Mattheij, *On dichotomy and well-conditioning in BVP*, SIAM J. Numer. Anal. **24** (1987), 89–105.

[6] P. Deuflhard, *Newton methods for nonlinear problems*, Springer-Verlag, Berlin Heidelberg, 2004.

[7] R. England and R.M.M. Mattheij, *Boundary value problems and dichotomic stability*, SIAM J. Numer. Anal. **25** (1988), 1037–1054.

[8] Z. Li, M.R. Osborne, and T. Prvan, *Parameter estimation of ordinary differential equations*, IMA J. Numer. Anal. **25** (2005), 264–285.

[9] J. Nocedal and S.J. Wright, *Numerical optimization*, Springer Verlag, 1999.

[10] M. R. Osborne, *On shooting methods for boundary value problems*, J. Math. Analysis and Applic. **27** (1969), 417–433.

[11] M.R. Osborne, *Cyclic reduction, dichotomy, and the estimation of differential equations*, J. Comp. and Appl. Math. **86** (1997), 271–286.

[12] B.J. Quinn and E.J. Hannan, *The estimation and tracking of frequency*, Cambridge University Press, Cambridge, United Kingdom, 2001.