OPTIMAL CONVERGENCE OF THE SIMPLIFIED NEWTON ITERATIONS FOR NON STIFF PROBLEMS

Ronald Tshelametse*

ABSTRACT:

The paper deals with the numerical solution of IVP's for systems of stiff ODE's with particular emphasis on implicit linear multistep methods (LMM), particularly the backward differentiation formulae (BDF). In thispaper we intend to find the value of the optimal convergence rate factor that will minimize the computational costs in the context of terminating simplified Newton iterations for non-stiff problems. We refer to this convergence rate factor, η_{ref} as the optimal convergence rate. We conclude that for non-stiff problems the optimal convergence rate is such that $0.3 \leq \eta_{ref} \leq 0.5$. This compares favourably with the optimal convergence rate for non-stiff problems using small stepsizes as obtained in Gustaffsson and Soderlind [1, Fig2, p27]. We further conduct experiments using several values ofη_{ref} in solving the van de Pol problem which uses small stepsizes when solved using ode15s with , $\sigma = 10$ for $t \in [0,30]$. We monitor the number of function calls for various convergence rate. The results obtained confirm our lower bound on η_{ref}.

* Department of Mathematics, University of Botswana, Private Bag 0022, Gaborone, Botswana.

1 Introduction

The paper is concerned with the numerical integration of

$$\frac{du}{dt} = f(t,u),\ \ 0 < t \le T,\qquad u(0) = u_0,\ \ f: R \times R^m \to R^m \qquad (1)$$

In the literature some initial value problems (1) are referred to as stiff. A prominent feature for these problems is that they are extremely difficult tosolve by standard explicit methods. The time integration of stiff systems isusually achieved using implicit methods, and for many codes by linear multistepmethods. A linear multistep method also called linear k-step methods [2], [3], [4], [5], [6] instandard constant stepsize form a linear multistep or k-step method is defined thus:

$$\sum_{i=0}^{k} \alpha_i u_{n-i} = h \sum_{i=0}^{k} \beta_i f_{n-i}\,, \qquad (2)$$

where $\alpha_i$and β_i are constants and α_0 = 1. f_{n-i}denotes $f(t_{n-i}, u_{n-i})$, $t_{n-i} = t_n - ih$, $i = 0,1,\dots,k$ and $h$is the stepsize. The condition that α_0 = 1removes the arbitrariness that arises from the fact that both sides of the IVPcould be multiplied by the same constant without altering the method. Thelinear multistep method (2) is said to be explicit if$\beta_0 = 0$and implicit if$\beta_0 \neq 0$. This leads to

$$F(u_n) \equiv u_n + \varphi_n - h_n\beta_0 f(u_n) = 0, \qquad (4)$$

where $\varphi_n = \sum_{i=1}^{k} \alpha_i\, u_{n-i}$is a known value. [7], [8], [9], [10], [11],[12]. At each integration step $t_n$we must solve the nonlinear equation (4). To solve for $u_n$most codes use the Newton iterative method and its variantsin the following form

$$W_n^{(l)} \varepsilon_n^{(l)} = -F\left(u_n^{(l)}\right), \qquad u_n^{(l+1)} = u_n^{(l)} + \varepsilon_n^{(l)} \qquad l = 0,12,\dots \qquad (5)$$

with the starting value $u_n^{(0)}$known and "fairly" accurate. For the full Newtonmethod

$$W_n^{(l)} = F'\left(u_n^{(l)}\right) = I - h_h\beta_0 f'\left(u_n^{(l)}\right) \qquad (6)$$

The use of the Newton method is due to the stiffness phenomenon. For largeproblems evaluating the Jacobian, $J = f'(u_n^{(l)})$ (and hence the Newton iteration matrix $W_n^{(l)}$) and solving the linear algebraic system are by far the most computationally expensive operations in the integration.

There are various strategiesused in practice to try and minimise the cost of computing the Jacobian and theNewton matrix. These measures are mainly centredonadministering the iteration matrix in (6) leading to simplified Newton methods amongst others [13], [14], [15]. In this paper we focus on optimally terminating the simplified Newton iterations.

2Theory

We intend to find the value of η (set to 0.9 in ode15s) that will minimize the computational costs in the context ofsimplified Newton. We refer to this convergence rate factor, η_{ref} as the optimalconvergence rate. For functional iteration this task is considered in Gustafsson and Soderlind [1], ideal for nonstiff problems, where$\eta_{ref} \approx 0.4$. To simulatethe nonstiff implementation in the context of simplified Newton we consider thecase when the stepsize h_n is small. We emphasise here that the convergencerate of an iterative solver usuallydepends on the stepsize. Thus the step sizeselection strategy is important.

It can be shown that with $e_l = u_n^{(l)} - u_n$,

$$e_{l+1} = W^{-1}(h_n\beta_0\tilde{J} - h_0\beta_0 J)e_l, \qquad (7)$$

where$W = I - h_0\beta_0 J$ and $\tilde{J}$ is the mean value Jacobian at $t_n$such that

$$\tilde{j}e_l = f\left(u_n^{(l)}\right) - f(u_n). \qquad (8)$$

If the iteration is convergent, there follows

$$\|e_{l+1}\| \leq \eta\|e_l\|, \qquad (9)$$

where $\left\|W^{-1}(h_n\beta_0\tilde{J} - h_0\beta_0 J)e_l\right\| \leq \eta$. Now since we assume that $\eta < 1$,

$$\|e_l\| \leq \left\|u_n^{(l)} - u_n^{(l-1)}\right\| + \|e_{l+1}\| \leq \eta\|e_l\| + \left\|u_n^{(l)} - u_n^{(l-1)}\right\|, \qquad (10)$$

yielding in (9)

$$\|e_{l+1}\| \leq \frac{\eta}{1-\eta}\left\|u_n^{(l)} - u_n^{(l-1)}\right\|, \qquad (11)$$

In practice η is replaced with the estimate of the convergence rate factor to getthe practical convergence test of the form

$$\|e_{l+1}\|\sim \leq \frac{\eta}{1-\eta}\left\|u_n^{(l)}-u_n^{(l-1)}\right\| \leq \tau, \qquad (12)$$

for some appropriate τ . We now follow Gustafsson and Soderlind [1]. Let l bethe number of iterations needed to satisfy (12). Then

$$\tau \geq \frac{\eta}{1-\eta}\left\|u_n^{(l)}-u_n^{(l-1)}\right\| \geq \sim\|e_l\| \approx \eta^l\|e_0\|, \qquad (13)$$

wheree_0 is the error in the initial $u_n^{(0)}$. Thus

$$l\sim \geq \frac{log\tau - log\|e_0\|}{log\eta}, \qquad (\eta<1) \qquad (14)$$

implying that, the number of function evaluations per unit step (i.e. per unittime of integration) is proportional to $\bar{l}$, where approximately

$$\bar{l} = \frac{l}{h_n} = \frac{log\tau - log\|e_0\|}{h_n log\eta}, \qquad (15)$$

For modelling the relationship between η and h_n we have from (9) the approximation

$$\eta \approx \left\|W^{-1}(h_n\beta_0\tilde{J} - h_0\beta_0 J)\right\|. \qquad (16)$$

Hence if $h_n = h_0$

$$\eta \approx h_n\beta_0\left\|(I-h_n\beta_0 J)^{-1}(\tilde{J}-J)\right\|. \qquad (17)$$

For small h_n,

$$(I - h_n\beta_0 J)^{-1} \approx I. \qquad (18)$$

We may therefore use the model

$$\eta \approx \mathrm{K}h_n \qquad (19)$$

for some constant K. In ode15s the iteration matrix is updated and refactored every time the step size, h is changed, see Shampine and Reichelt [16], so that for every successful time step the condition$h_n = h_0$ holds.

This model is also used by Gustafsson and Soderlind [1] who analyse thecase for functional iteration, namely,$J = 0$. Therefore

$$\bar{l} \approx \frac{log\tau - log\|e_0\|}{\eta log\eta}, \qquad (20)$$

The starting value u_n^0 is usually constructed using an interpolation formula,which makes the initial error e_0 depend on h_n(and henceη). Assuming thatthe dependence is usually notstrong, $log\tau - log\|e_0\|$is assumed to be aconstant, so that we have the proportionality relationship

$$\bar{l} \approx -\frac{1}{\eta log\eta}, \qquad (21)$$

where the minus sign is a result of assuming $e_0 > \tau$. The expression (21) hasa minimum at$\eta = e^{-1} \approx 0.36788$ which indicates the appropriate selection of the convergence rate as, say $\eta_{ref} \approx 0.4$. The function (21) is fairly flat around the minimum, any value $0.3 \leq \eta_{ref} \leq 0.5$ would probably be acceptable. See Figure 1 below.

It should be noted that this model, for both the functional iteration [1] and our analysis for small h_n, relies on the assumption that the number of iterations l, is not limited and $l$is significantly large. From [1, Fig2, p.27] for functional iteration for a non stiff problem van de Pol problem

with$\sigma = 10$, if the maximum allowed iterations is 10 then any value$0.2 \leq \eta_{ref} \leq 0.95$ is acceptable. Their theoretical expectations are approximately met when $l = 100$.We conduct experiments to investigate the range of our η_{ref}in solving the same problem.

3Numerical Experiments

In this experiment we aim to obtain a plot similar to the plots obtained byGustafsson and Soderlind [1, Fig.2,p.27] for the simplified Newton iteration for small h_n. That is, we intend to obtain a plot of the total work (numberof function calls) as a function of the convergence rate when integrating thenonstiff van de Pol problem with $\sigma = 10$,for$t \in$ [0, 30] withode15s. The code ode15s uses small stepsizes(maximum stepsize is about 0.9) throughout the integration when solving this problem. This will thensimulate the above small h_n theory for the simplified Newton method. We havea nonstiff problem solved by the stiff methods of ode15s verifying that the theory forsmall,$h_n$corresponds to using functional iteration in the case of Gustafssonand Soderlind [1].

In ode15s the acceptable convergence rate is $\eta_{ref} \leq 0.9$ and the maximumallowed number of iterations is 4. We use the weighted infinity norm and experiment with various convergence rates η_{ref}in the range[0.01, 0.97]. For the default ode15s settings discussed above the plot of functioncalls against convergence rate is plotted as rings (o) in Figure 1. We furtherincrease the maximum number of allowed iterations in ode15s to 100 and investigate the behaviour of function evaluations versus convergence rate. Thebehaviour is plotted as crosses (+) in Figure 1. To determine whether ode15sconforms to the theory we also include the plot of the function (21) for theoptimal convergence rate as a continuous line in Figure 1. For the plots to fiton the same scale the expression (11) was normalised with a constant $k$thatmakes its minimum equal to the minimum number of function callsexperienced in practice, that is,

$$k = \frac{448}{\left[\frac{1}{-0.36788 \log(0.36788)}\right]} = 164.81$$

We obtain **Figure 1.**

4Results and Conclusions

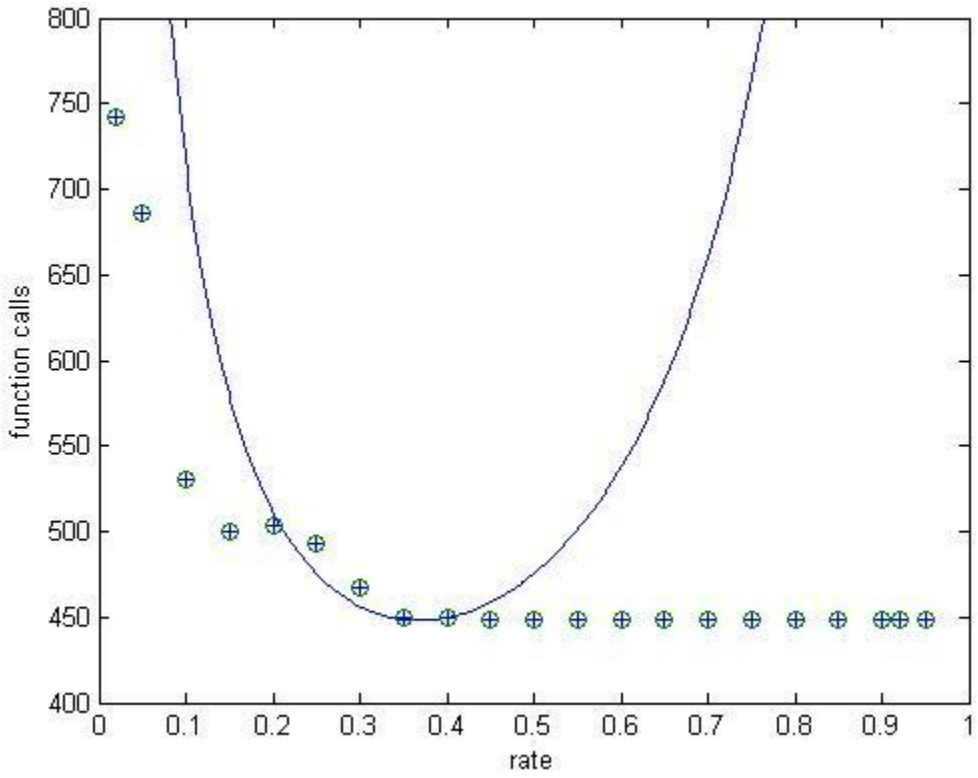

Figure1: Total work (number of function calls) as a function of the convergence rate set-point when integrating the van de Pol problem with $\sigma = 10$.The equation solver includes a limit on the number of allowed iterations. The crosses and the rings correspond to setting this limit as 100 and 4 respectively. The solid line is the graph of the function (21). We use default tolerances.

The expression (21) hasa minimum at$\eta = e^{-1} \approx 0.36788$ which indicates the appropriate selection of the convergence rate as, say $\eta_{ref} \approx 0.4$. The function (21) is fairly flat around the minimum, any value $0.3 \leq \eta_{ref} \leq 0.5$would probably be acceptable. See Figure 1 below.This is in agreement with$\eta_{ref} \approx 0.4$ as obtained by Gustafssonand Soderlind [1].

Regarding our experiments, From Figure 1, for the simplified Newton iteration with small $h_n$it is clear thatfor the nonstiff problem, van de Pol problem if the maximumnumber of allowed iterations is 4 or 100 then any value $0.2 \leq \eta_{ref} \leq 0.95$ is acceptable. The largest observed

number of iterations performed throughout the entire integration was 3. As seen in (13), the key requirement is that $\|e_l\| \approx \eta^l \|e_0\|$, which is more meaningful if l is large. Moreover the result (21) is only valid when the dependence of the initial error e_0 on the step size is not strong. We note that in ode15s, $u_n^{(0)}$ is constructed using an interpolation polynomial, so that the initial error $e_0$depends onh_n. This in part explains the difference between our plots and the plots in Gustafssonand Soderlind [1] for the van de Pol problem using a functional iteration based code.The convergence rate $\eta_{ref} \leq 0.9$ will (for the non-stiff van de Pol problem) yield the minimum number of function evaluations. The same conclusions can be drawn even when the maximum number of allowed iterations is increased to 100, see Figure 1.

Meanwhile, we conclude that the optimal convergence rate for non-stiff problems in the solution of odes using BDFs in the simplified Newton context is such that:

$$0.3 \leq \eta_{ref} \leq 0.5$$

To be specific$\eta_{ref} = e^{-1} = 0.36788$.

References

[1] Kjell Gustafsson, and Gustaf Soderlind. Control strategies for the iterative solution of nonlinear equations in ODE solver. *SIAM J. Sci. Stat. Comput.*, 18, No. 1:23–40, January 1997.

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

[3] W. H. Enright, T. E. Hull, and B. Linberg. Comparing numerical methodsfor stiff systems of ODEs. *BIT*, 15:10–48, 1975.

[4] E. Hairer. Backward error analysis for linear multistep methods. *Numer. Math*., 84:2:199–232, 1999.

[5] E. Hairer. Conjugate-symplecticity of linear multistep methods. *J. Computational Mathematics*, 26:5:657–659, 2008.

[6] E. Hairer and C. Lubich. Symmetric multistep methods over long times.*Numer. Math*., 97:4:699–723, 2004.

[7] J. D. Lambert. Numerical Methods for Ordinary Differential Systems. *John Wiley and Sons*, 1991.

[8] Peter N. Brown, George D. Byrne, and Alan C. Hindmarsh. VODE: a variable-coefficient ODE solver. *SIAM J. Sci. Stat. Comput.*, 10, No. 5:1038–1051, September 1989.

[9] C. T. Kelley. Iterative methods for linear and nonlinear equations. *Society for Industrial and Applied Mathematics, Philadelphia, PA, USA*, 1995.

[10] Kenneth R. Jackson. The numerical solution of stiff IVPs for ODEs. *J. Applied Numerical Mathematics*, 1995.

[11] L. F. Shampine and P. Bogacki. The effect of changing the step size inthe linear multistep codes. *SIAM J. Sci. Stat. Comput*., 10:1010–1023,September 1989.

[12] R. Tshelametse. Terminating simplified newton iterations: A modifiedstrategy.*International Journal of Management, IT and Engineering*, 4,Issue 4:pp. 246–266, 2014.

[13] G. D. Byrne and A. C. Hindmarsh. A polyalgorithm for the numericalsolution of ordinary differential equation. *Comm. ACM*, 1, No. 1:71–96,March 1975.

[14] G. Gheri and P.Marzulli. Parallel shooting with error estimate for increasing the accuracy. *J. Comput. and Appl. Math*., 115, Issues 1-2:213–227,March 2000.

[15] Lawrence F. Shampine. Numerical solution of ordinary differential equations. *Chapman and Hall*, 1994.

[16] Lawrence F. Shampine and Mark W. Reichelt. The MATLAB ODE suite. *SIAM J. Sci. Stat. Comput*., 18, No 1:103–118,September 1989.