INTERNATIONAL JOURNAL FOR NUMERICAL METHODS IN ENGINEERING
Int. J. Numer. Meth. Engng 2004; **60**:2031–2043 (DOI: 10.1002/nme.1035)

Truncation error and stability analysis of iterative and non-iterative Thomas–Gladwell methods for first-order non-linear differential equations

Dmitri Kavetski, Philip Binning*, † and Scott W. Sloan

School of Engineering, University of Newcastle Callaghan, NSW 2308, Australia

SUMMARY

The consistency and stability of a Thomas–Gladwell family of multistage time-stepping schemes for the solution of first-order non-linear differential equations are examined. It is shown that the consistency and stability conditions are less stringent than those derived for second-order governing equations. Second-order accuracy is achieved by approximating the solution and its derivative at the same location within the time step. Useful flexibility is available in the evaluation of the non-linear coefficients and is exploited to develop a new non-iterative modification of the Thomas–Gladwell method that is second-order accurate and unconditionally stable. A case study from applied hydrogeology using the non-linear Richards equation confirms the analytic convergence assessment and demonstrates the efficiency of the non-iterative formulation. Copyright © 2004 John Wiley & Sons, Ltd.

KEY WORDS: Thomas–Gladwell methods; non-iterative linearization; non-linear differential equations; Richards equation

INTRODUCTION

The numerical solution of ordinary differential equations (ODEs) and, more generally, differential-algebraic equations (DAEs) has been the subject of considerable research. The approximation of low-order ODE systems that arise in many areas of physical sciences and engineering is of particular importance. Higher-order ODEs can be converted to equivalent first-order ODE systems using state augmentation.

Thomas and Gladwell [1] presented a family of multistage time-stepping schemes for the solution of second-order ODE systems

$$C\frac{\mathrm{d}^2\mathbf{U}(t)}{\mathrm{d}t^2} + \mathbf{M}(\mathbf{U}, t)\frac{\mathrm{d}\mathbf{U}(t)}{\mathrm{d}t} + \mathbf{K}(\mathbf{U}, t)\mathbf{U}(t) = \mathbf{F}(\mathbf{U}, t) \tag{1}$$

*Correspondence to: Philip Binning, Institute of Environment and Resources, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark.
†E-mail: pjb@er.dtu.dk

Received 9 December 2002
Revised 11 November 2003
Accepted 28 November 2003

Published online 24 May 2004

The schemes are defined by the weighting parameters φ_1, φ_2 and φ_3 and are given by

$$[C + \varphi_2 \Delta t \mathbf{m} + \varphi_3 \Delta t^2 \mathbf{k}]\ddot{\mathbf{u}}^n = -\mathbf{m}\dot{\mathbf{u}}^n - \mathbf{k}(\mathbf{u}^n + \varphi_1 \Delta t \dot{\mathbf{u}}^n) + \mathbf{f}^{n+\varphi_1} \tag{2}$$

$$\mathbf{u}^{n+1} = \mathbf{u}^n + \Delta t \dot{\mathbf{u}}^n + 1/2 \Delta t^2 \ddot{\mathbf{u}}^n \tag{3}$$

$$\dot{\mathbf{u}}^{n+1} = \dot{\mathbf{u}}^n + \Delta t \ddot{\mathbf{u}}^n \tag{4}$$

where the lower-case symbols denote the numerical approximations to the analytical (upper-case) terms and $\Delta t = t^{n+1} - t^n$ is a finite increment (time step), not necessarily constant throughout the integration. Updates (3) and (4) can be viewed as a truncated Taylor series, where $\dot{\mathbf{u}}$ approximates $\mathrm{d}\mathbf{U}/\mathrm{d}t$ and $\ddot{\mathbf{u}} \approx \mathrm{d}^2\mathbf{U}/\mathrm{d}t^2$.

In the original implementation by Thomas and Gladwell [1], when the ODEs (1) are non-linear, the coefficients are evaluated as follows:

$$\mathbf{x} = \mathbf{X}(\mathbf{u}^n + \varphi_1 \Delta t \dot{\mathbf{u}}^n + \varphi_3 \Delta t^2 \ddot{\mathbf{u}}^n) \tag{5}$$

where $\mathbf{X} = \mathbf{M}, \mathbf{K}$ and $\mathbf{F}$ and $\mathbf{x}$ is the value used in the algorithm in Equation (2).

Thomas and Gladwell [1] state that, when schemes (2)–(4) are applied to (1), they are at least $O(\Delta t)$ accurate and when the parameters take the values

$$\varphi_1 = \varphi_2 = 1/2 \tag{6}$$

the approximations become $O(\Delta t^2)$ accurate.

The stability constraints for the Thomas–Gladwell integration family applied to (1) are summarized by Thomas and Gladwell [1] in their Table I. The methods are stable provided

$$2\varphi_3 \geqslant \varphi_1 \geqslant 1/2 \quad \text{and} \quad \varphi_2 \geqslant 1/2 \tag{7}$$

Although originally derived for second-order ODEs, the Thomas–Gladwell scheme can be applied, as a special case, to first-order DAE systems by setting $C \equiv 0$

$$\mathbf{M}(\mathbf{U}, t)\frac{\mathrm{d}\mathbf{U}(t)}{\mathrm{d}t} + \mathbf{K}(\mathbf{U}, t)\mathbf{U}(t) = \mathbf{F}(\mathbf{U}, t) \tag{8}$$

Equations of this structure often arise following finite element and finite difference semi-discretization of partial differential equations (PDEs). Such PDEs form the basis of physical modelling in many areas of science and engineering. In these applications, t typically denotes time, while the vector $\mathbf{U}$ contains quantities of interest such as soil displacements, water pressures, etc. In most cases, the non-linearity of the constitutive functions $\mathbf{M}$ and $\mathbf{K}$, as well as non-trivial forcing conditions $\mathbf{F}$, require that numerical approximations be employed. Although DAEs (8) can often be converted to ODEs by pre-multiplying all terms by $\mathbf{M}^{-1}(\mathbf{U}, t)$, this exacerbates numerical ill-conditioning when $\mathbf{M}$ is near-singular or poorly conditioned [2]; moreover, $\mathbf{M}$ can be singular in some applications.

Wood [3] discusses various approximations to second-order ODEs (1) and states that the Thomas–Gladwell scheme is equivalent to an SS22 algorithm of Zienkiewicz *et al.* [4] provided $\varphi_1 = \varphi_2$. The notation SSab denotes a single-step $O(\Delta t^a)$ approximation to an ODE of order b. Setting $C = 0$ converts an SS$a2$ method to an SS$a1$ scheme. In addition, there are other $O(\Delta t^2)$ approximations to (8), e.g. the Crank–Nicolson scheme. Although the simplification of

the governing DE from second- to first-order affects the convergence properties of the Thomas–Gladwell scheme, this issue has not been specifically addressed in the literature, despite the importance of the methods as a generalized family of approximations to low-order ODE systems.

Reliable and robust ODE software is now widely available [5], typically employing variable-order variable-step Runge–Kutta, Adams and Gear schemes. Thomas–Gladwell methods have also been used in variable-order variable-step ODE codes [6] and are of practical significance since they include many common schemes as special cases [1].

Conversely, current PDE codes typically employ robust lower-order algorithms, since (i) semi-discrete PDEs are often very stiff; (ii) using high-order schemes presupposes well behaved high-order solution derivatives that may not even exist, particularly for non-smooth forcing functions and constitutive relationships; (iii) temporal accuracy must be balanced with accuracy in other dimensions (e.g. space); and (iv) extremely high accuracy is rarely needed in engineering applications [3]. For example, an adaptive time-stepping scheme used for PDEs in geomechanics [7] and hydrogeology [8, 9] monitors its local truncation error by the difference between an $O(\Delta t)$ backward Euler approximation and a Thomas–Gladwell scheme. In order to provide an asymptotic truncation error measure, the Thomas–Gladwell scheme must be $O(\Delta t^2)$ accurate.

The first objective of this paper is therefore to analyse the consistency and stability of Thomas–Gladwell approximations to first-order DEs (8) and determine any parameter constraints on its convergence. We seek to verify the order of accuracy and stability for fully non-linear DEs, as opposed to linear or quasi-linear forms. It is found that both constraints (6) and (7) can be weakened, indicating that Thomas–Gladwell methods are applicable to a wider range of problems than originally intended.

Another issue of practical importance is the approximation of the non-linear coefficients $\mathbf{m}$, $\mathbf{k}$ and $\mathbf{f}$. Whilst Thomas and Gladwell [1] employ (5), it may be advantageous to use different approximations, either from the point of view of truncation error or due to software implementation aspects. For example, non-iterative time-stepping schemes for the non-linear Richards equation in hydrogeology exploit linearizations of the non-linear coefficients to increase efficiency [10, 11].

The second aim of the paper is hence to demonstrate, both analytically and empirically, that valuable freedom is available in the selection of the arguments of the non-linear terms in the time-stepping scheme. In particular, a compact non-iterative formulation is derived that maintains second-order accuracy and stability without iterated inversion of non-linear systems. Empirical assessment illustrates the advantages of the non-iterative scheme compared to an analogous time-stepping algorithm with an iterative solver.

TRUNCATION ERROR ANALYSIS

The truncation error analysis is presented for a scalar ODE. The results generalize to ODE systems since the same approximation scheme is applied to each term in the system. The approximation of a system of m ODEs such as (8) then becomes a sum of m nth-order approximations and remains nth-order accurate.

The Thomas–Gladwell equations can be formulated as a weighted difference approximation to ODE (8). It is also convenient to eliminate $\ddot{u}^n$, giving a 2-stage form as

$$m[\varphi_2 \dot{u}^{n+1} + (1-\varphi_2)\dot{u}^n] + k[u^n + \Delta t\{\varphi_3 \dot{u}^{n+1} + (\varphi_1 - \varphi_3)\dot{u}^n\}] = f^{n+\varphi_1} \tag{9}$$

$$u^{n+1} = u^n + 1/2\Delta t(\dot{u}^n + \dot{u}^{n+1}) \tag{10}$$

This form of the Thomas–Gladwell methods shows that the derivative $\mathrm{d}U/\mathrm{d}t$ in the governing ODE is approximated at a different location within the step than U and F; this has implications for the order of accuracy of the entire approximation.

To obtain the local truncation error, it is assumed that the initial conditions u^n and $\dot{u}^n$ are exact, i.e.

$$u^n = U^n \quad \text{and} \quad \dot{u}^n = \dot{U}^n \tag{11}$$

Rearrangement and Taylor series expansion of Equation (10) then yields

$$\dot{u}^{n+1} = \frac{\mathrm{d}U^n}{\mathrm{d}t} + \Delta t \frac{\mathrm{d}^2U^n}{\mathrm{d}t^2} + \frac{\Delta t^2}{3}\frac{\mathrm{d}^3U^n}{\mathrm{d}t^3} + O(\Delta t^3) = \frac{\mathrm{d}U^n}{\mathrm{d}t} + O(\Delta t) \tag{12}$$

The expansion for the second derivative $\ddot{u}^n$ can be derived from (12) and (4), showing that $\ddot{u}^n$ is an $O(\Delta t)$ approximation to $\mathrm{d}^2U(t^n)/\mathrm{d}t$

$$\ddot{u}^n = \frac{\dot{u}^{n+1} - \dot{u}^n}{\Delta t} = \frac{\mathrm{d}^2U^n}{\mathrm{d}t^2} + \frac{\Delta t}{3}\frac{\mathrm{d}^3U^n}{\mathrm{d}t^3} + O(\Delta t^2) = \frac{\mathrm{d}^2U^n}{\mathrm{d}t^2} + O(\Delta t) \tag{13}$$

It can also be shown that $\ddot{u}^n$ is an $O(\Delta t^2)$ approximation to $\mathrm{d}^2U(t^n + 1/3\Delta t)/\mathrm{d}t$, with the truncation error given by $1/6\Delta t^2\mathrm{d}^4U(t^n)/\mathrm{d}t^4 + O(\Delta t^3)$.

The convergence properties of the approximations to the non-linear coefficients, m, k and f are examined using Taylor-series expansions about u^n, e.g.

$$m = M(u^n + \Delta u) = M^n + \Delta u\frac{\mathrm{d}M^n}{\mathrm{d}U} + \frac{\Delta u^2}{2}\frac{\mathrm{d}^2M^n}{\mathrm{d}U^2} + O(\Delta u^3) \tag{14}$$

For the methods originally proposed by Thomas and Gladwell [1]

$$\Delta u = \varphi_1\Delta t\dot{u}^n + \varphi_3\Delta t^2\ddot{u}^n \tag{15}$$

and the full expansion is then given by

$$\begin{aligned} m &= M(u^n + [\varphi_1\Delta t\dot{u}^n + \varphi_3\Delta t^2\ddot{u}^n]) \\ &= M^n + [\varphi_1\Delta t\dot{u}^n + \varphi_3\Delta t^2\ddot{u}^n]\frac{\mathrm{d}M^n}{\mathrm{d}U} + \frac{[\varphi_1\Delta t\dot{u}^n + \varphi_3\Delta t^2\ddot{u}^n]^2}{2}\frac{\mathrm{d}^2M^n}{\mathrm{d}U^2} \\ &\quad + O([\varphi_1\Delta t\dot{u}^n + \varphi_3\Delta t^2\ddot{u}^n]^3) \end{aligned} \tag{16}$$

Re-arranging terms, substituting (13) and assuming exact initial conditions (11) yields

$$m = M + \varphi_1\Delta t\frac{\mathrm{d}U}{\mathrm{d}t}\frac{\mathrm{d}M}{\mathrm{d}U} + \varphi_3\Delta t^2\frac{\mathrm{d}^2U}{\mathrm{d}t^2}\frac{\mathrm{d}M}{\mathrm{d}U} + \frac{\varphi_1^2}{2}\Delta t^2\left(\frac{\mathrm{d}U}{\mathrm{d}t}\right)^2\frac{\mathrm{d}^2M}{\mathrm{d}U^2} + O(\Delta t^3) \tag{17}$$

where all terms on the RHS are evaluated at t^n. The coefficient k has an analogous expansion.

When the forcing function f is dependent on U, (i.e. $f = f(U)$), it can be treated as

$$f = F + \varphi_1\Delta t\frac{\mathrm{d}U}{\mathrm{d}t}\frac{\mathrm{d}F}{\mathrm{d}U} + \varphi_3\Delta t^2\frac{\mathrm{d}^2U}{\mathrm{d}t^2}\frac{\mathrm{d}F}{\mathrm{d}U} + \frac{\varphi_1^2}{2}\Delta t^2\left(\frac{\mathrm{d}U}{\mathrm{d}t}\right)^2\frac{\mathrm{d}^2F}{\mathrm{d}U^2} + O(\Delta t^3) \tag{18}$$

where the RHS is evaluated at t^n.

To obtain the truncation error of the Thomas–Gladwell methods, (10) is solved for $\dot{u}^{n+1}$ and substituted into (9). All terms are then expanded in Taylor series about t^n. Since (12) is already an expansion of (10), it can be substituted directly into (9). The non-linear coefficients are expanded as in (14). The resulting expression is then compared with the governing ODE (8). The difference between them constitutes the local truncation error T_L

$$T_L = \Delta t \left[\varphi_2 M \frac{d^2U}{dt^2} + \varphi_1 \frac{dM}{dU} \left(\frac{dU}{dt} \right)^2 + \varphi_1 K \frac{dU}{dt} + \varphi_1 \frac{dK}{dU} \frac{dU}{dt} U - \varphi_1 \frac{dF}{dU} \frac{dU}{dt} \right]$$

$$+ \Delta t^2 \left[\begin{array}{l} \varphi_3 \dfrac{dM}{dU} \dfrac{dU}{dt} \dfrac{d^2U}{dt^2} + \dfrac{\varphi_1^2}{2} \dfrac{d^2M}{dU^2} \left(\dfrac{dU}{dt} \right)^3 + \varphi_1 \varphi_2 \dfrac{dM}{dU} \dfrac{dU}{dt} \dfrac{d^2U}{dt^2} + \dfrac{\varphi_2}{3} M \dfrac{d^3U}{dt^3} + \varphi_3 \dfrac{dK}{dU} \dfrac{d^2U}{dt^2} U \\ + \dfrac{\varphi_1^2}{2} \dfrac{d^2K}{dU^2} \left(\dfrac{dU}{dt} \right)^2 U + \varphi_1^2 \dfrac{dK}{dU} \left(\dfrac{dU}{dt} \right)^2 + \varphi_3 K \dfrac{d^2U}{dt^2} - \varphi_3 \dfrac{dF}{dU} \dfrac{d^2U}{dt^2} - \dfrac{\varphi_1^2}{2} \dfrac{d^2F}{dU^2} \left(\dfrac{dU}{dt} \right)^2 \end{array} \right]$$

$$+ O(\Delta t^3) \tag{19}$$

The truncation error is hence $O(\Delta t)$ and contains advective and diffusive terms.

The similarity between the $O(\Delta t)$ terms in the truncation error (19) and the first derivative of the governing ODE (obtained by differentiating each term in (8) with respect to t) can be exploited by setting

$$\varphi_1 = \varphi_2 \tag{20}$$

This choice eliminates the first-order terms, raising the convergence of the schemes to second-order. Hence, the reduction of the governing ODE from second- to first-order relaxes the parameter condition for $O(\Delta t^2)$ accuracy from (6) to (20).

The intuitive meaning of (20) is that all dependent variables in the ODE (both the solution $U(t)$ and the derivative dU/dt) become approximated at the same location within the time step, namely $t^n + \varphi_1 \Delta t = t^n + \varphi_2 \Delta t + O(\Delta t^2)$. Indeed, (19) shows that it is precisely the approximation of $U(t)$ and dU/dt at different t-locations that introduces $O(\Delta t)$ errors and degrades the accuracy of the approximation.

If $\varphi_1 = \varphi_2$ and the governing ODE is linear or at least quasi-linear (e.g. with sufficiently slowly varying coefficients, so that $|\partial X/\partial U| \ll |X|$ for $X = M, K, F$), the truncation error is further reduced to

$$T_L = \Delta t^2 \left[\frac{\varphi_1}{3} M \frac{d^3U}{dt^3} + \varphi_3 K \frac{d^2U}{dt^2} - \frac{\varphi_1^2}{2} \frac{d^2F}{dt^2} \right] + O(\Delta t^3) \tag{21}$$

Alternative evaluation of non-linear coefficients

Useful flexibility is available in the evaluation of the coefficients m, k and f when the governing ODEs are non-linear. For example, instead of using (5), which does not correspond to any particular t-level, it is possible to use $m^{n+\varphi_1}$, $k^{n+\varphi_1}$ and $f^{n+\varphi_1}$ where, for example

$$m^{n+\varphi_1} = M(u(t^{n+\varphi_1})) \tag{22}$$

To evaluate these coefficients, a simple interpolation can be undertaken based on $O(\Delta t^2)$ forward Taylor series consistent with (3), e.g.:

$$m^{n+\varphi_1} = m(u^n + [\varphi_1 \Delta t]\dot{u}^n + 1/2[\varphi_1 \Delta t]^2 \ddot{u}^n) \tag{23}$$

To obtain the truncation error associated with (23), the non-linear coefficients can be expanded using Taylor series as

$$m^{n+\varphi_1} = M(u^n + [\varphi_1 \Delta t \dot{u}^n + 1/2[\varphi_1 \Delta t]^2 \ddot{u}^n]) = M^n + [\varphi_1 \Delta t \dot{u}^n + 1/2[\varphi_1 \Delta t]^2 \ddot{u}^n] \frac{\mathrm{d}M^n}{\mathrm{d}U}$$

$$+1/2[\varphi_1 \Delta t \dot{u}^n + 1/2[\varphi_1 \Delta t]^2 \ddot{u}^n]^2 \frac{\mathrm{d}^2 M^n}{\mathrm{d}U^2} + O([\varphi_1 \Delta t \dot{u}^n + 1/2[\varphi_1 \Delta t]^2 \ddot{u}^n]^3) \tag{24}$$

Expanding (24) and collecting first- and second-order terms leads to

$$m^{n+\varphi_1} = M(u^n + [\varphi_1 \Delta t \dot{u}^n + [\varphi_1 \Delta t]^2 \ddot{u}^n]) = M^n + \varphi_1 \Delta t \frac{\mathrm{d}U}{\mathrm{d}t}^n \frac{\mathrm{d}M^n}{\mathrm{d}U} + O(\Delta t^2) \tag{25}$$

Comparing (25) and (17), it can be seen that the precise manner in which the second derivative is incorporated affects only the coefficient of the $O(\Delta t^2)$ term. That is, if (5) is replaced by (23), the truncation error of the approximation will remain $O(\Delta t^2)$ accurate, although the coefficients of error terms will change.

A minor disadvantage of using (16) or (23) is that they require the evaluation and storage of the second derivatives $\ddot{\mathbf{u}}^n$. In practical software, it may be preferable to rearrange the Thomas–Gladwell schemes to be solved for $\dot{\mathbf{u}}^{n+1}$, i.e.

$$[\varphi_2 \mathbf{m} + \varphi_3 \Delta t \mathbf{k}]\dot{\mathbf{u}}^{n+1} = -(1 - \varphi_2)\mathbf{m}\dot{\mathbf{u}}^n - \mathbf{k}[\mathbf{u}^n + (\varphi_1 - \varphi_3)\Delta t \dot{\mathbf{u}}^n] + \mathbf{f}^{n+\varphi_1} \tag{26}$$

and then updating the solution according to (10). In addition, the argument of the non-linear coefficients $\mathbf{m}$, $\mathbf{k}$ and $\mathbf{f}$ can be set to

$$\mathbf{u}^{n+\varphi_1} = \mathbf{u}^n + \varphi_1 \Delta t \dot{\mathbf{u}}^{n+1} \tag{27}$$

This modification eliminates $\ddot{\mathbf{u}}^n$ and maintains second-order accuracy.

Non-iterative linearizations

A more fundamental limitation of using (5), (23) and (26) for non-linear DEs is that they require iterated inversion of non-linear systems at each time step. This is a major computational expense that can be avoided by approximating $u^{n+\varphi_1}$ in the non-linear coefficients at $t^n + \varphi_1 \Delta t$ using a truncated Taylor series

$$\tilde{u}^{n+\varphi_1} = u^n + \varphi_1 \Delta t \dot{u}^n \tag{28}$$

All coefficients are then approximated with $O(\Delta t^2)$ accuracy, e.g.

$$m^{n+\varphi_1} = M(\tilde{u}^{n+\varphi_1}) = M(u^n + \varphi_1 \Delta t \dot{u}^n) = M^n + \varphi_1 \Delta t \dot{u}^n \frac{\mathrm{d}M^n}{\mathrm{d}U} + O(\Delta t^2) \tag{29}$$

Linearization (29) preserves the second-order accuracy of the Thomas–Gladwell scheme (provided $\varphi_1 = \varphi_2$ as derived above) and yields non-iterative $O(\Delta t^2)$ methods for (8) that may have significant efficiency advantages over traditional iterative methods.

STABILITY ANALYSIS

In addition to consistency, stability is necessary to guarantee convergence. Several definitions of stability have been used in the literature [3]. In this study, the behaviour of the numerical schemes is examined for the model equation

$$M\frac{\mathrm{d}U}{\mathrm{d}t} + KU = 0 \tag{30}$$

which, provided the (possibly) non-linear terms $M(\cdot)$ and $K(\cdot)$ are positive, has a bounded decaying solution. Model (30) arises in eigenvalue decompositions of linear and non-linear ODE systems in finite difference and finite element applications [3].

The stability at infinity of the numerical schemes requires that the amplification factor $\xi = u^{n+1}/u^n$ satisfy $|\xi| \leqslant 1$ as $K\Delta t \to \infty$. Note that $K \to \infty$ corresponds to increasing the stiffness of the ODE. Provided the modulus of the amplification factor is below unity, the numerical solution is bounded and decays as required by model (30). The analysis parallels the original stability assessment of Thomas and Gladwell [1], who also examined the behaviour of their methods at infinity.

The multi-step analogue of (9)–(10) can be obtained from the multi-step analogue of (2)–(4) provided by Thomas and Gladwell [1] in their Equation (5) by setting $C = 0$:

$$\begin{aligned}[\varphi_2 M + \varphi_3 \Delta t K]u^{n+2} + [(1 - 2\varphi_2)M + (1/2 + \varphi_1 - 2\varphi_3)\Delta t K]u^{n+1} \\ + [(\varphi_2 - 1)M + (\varphi_3 - \varphi_1 + 1/2)\Delta t K]u^n = 0\end{aligned} \tag{31}$$

Dividing through by u^n and letting $K\Delta t \to \infty$ yields the stability polynomial of the Thomas–Gladwell scheme

$$2\varphi_3\xi^2 + (1 + 2\varphi_1 - 4\varphi_3)\xi + (1 - 2\varphi_1 + 2\varphi_3) = 0 \tag{32}$$

where $\xi^2 = u^{n+2}/u^n$. The non-linearity of $M(\cdot)$ and $K(\cdot)$ does not affect (32), which in turn implies that linearizations of the coefficients do not affect the stability polynomial. The absence of φ_2 in the stability polynomial is also noted, implying that the stability constraint (7) is relaxed when the governing equation is first-order.

The roots of quadratic (32) are given by

$$\xi_{1,2} = \frac{-1 - 2\varphi_1 + 4\varphi_3 \pm \sqrt{1 + 4\varphi_1 - 16\varphi_3 + 4\varphi_1^2}}{4\varphi_3} \tag{33}$$

The solution is bounded provided the amplification factors satisfy $|\xi_{1,2}(\varphi_1, \varphi_3)| \leqslant 1$. It can be shown that

$$|\xi_{1,2}(\varphi_1, \varphi_3)| \leqslant 1 \quad \text{iff } 2\varphi_3 \geqslant \varphi_1 \geqslant 1/2 \tag{34}$$

where $|\cdot|$ denotes the complex modulus whenever ξ is complex.

Although (34) does not constrain φ_2 for stability, in practice we require $\varphi_1 = \varphi_2$ for second-order accuracy. This is an important consideration in the design of adaptive integrators with

embedded error control. In addition, we expect numerical ill-conditioning for $\varphi_2 \to 0$, since the time stepping matrix $[\varphi_2\mathbf{m} + \varphi_3\Delta t\mathbf{k}]$ that must be factorized in (26) becomes progressively dominated by $\mathbf{k}$. In practice, this term often represents the stiffness matrix in finite element formulations and has a condition number $\kappa \propto 1/l^2$ where l is the spatial element size. Moreover, while $\mathbf{m}$ is often diagonal or at least diagonally dominant, $\mathbf{k}$ can be indefinite for certain boundary conditions.

Consistency, stability, conditioning and efficiency considerations favour the following particularly useful $O(\Delta t^2)$ non-iterative scheme:

$$\varphi_1 = \varphi_2 = \varphi_3 = 1 \tag{35}$$

This choice leads to the compact algorithm

$$[\mathbf{m} + \Delta t\mathbf{k}]\dot{\mathbf{u}}^{n+1} = -\mathbf{k}\mathbf{u}^n + \mathbf{f}^{n+1} \tag{36}$$

$$\mathbf{u}^{n+1} = \mathbf{u}^n + 1/2\Delta t(\dot{\mathbf{u}}^n + \dot{\mathbf{u}}^{n+1}) \tag{37}$$

with the coefficients evaluated as

$$\mathbf{x} = \mathbf{X}(\mathbf{u}^n + \Delta t\dot{\mathbf{u}}^n) \tag{38}$$

where $\mathbf{x}$ denotes the functions $\mathbf{m}$, $\mathbf{k}$ and $\mathbf{f}$.

The parameter set (35) cancels several terms in the approximation and, more importantly, allows the forcing term $\mathbf{F}(\cdot)$ to be evaluated with the same arguments as the other non-linear terms. When the non-linearities are expensive to evaluate and there are common terms in $\mathbf{M}$, $\mathbf{K}$ and $\mathbf{F}$ (as is often the case in practice), the computational cost of a time step is significantly reduced. In addition, the derivative estimate (36) is analogous to the strongly stable backward Euler method useful in simple adaptive algorithms [7, 9].

Unconditional stability of method (35) for any positive stepsize Δt can be shown by checking that $|\xi^*_{1,2}| \leqslant 1 \ \forall\{\Delta t > 0\}$, where

$$\xi^*_{1,2} = \frac{2M + \Delta t K \pm \sqrt{4M^2 - 4M\Delta t K - 7\Delta t^2 K^2}}{4\Delta t K + 4M} \tag{39}$$

To non-dimensionalize the problem, define $\lambda = \Delta t K/M$, yielding

$$\xi^*_{1,2} = \frac{2 + \lambda \pm \sqrt{4 - 4\lambda - 7\lambda^2}}{4\lambda + 4} \tag{40}$$

It can then be seen that the stability condition is satisfied for any positive interval I_λ, e.g. by plotting $|\xi^*_{1,2}(\lambda)|$ over some large interval, say $\lambda \in I_\lambda = [10^{-20}, 10^{20}]$. At the upper limit $\lambda \to \infty$ the stability at infinity analysis holds (where the method was shown to be stable, with $\lim_{\lambda\to\infty} |\xi^*_{1,2}| = 0.71$), whereas at the lower limit one finds $\lim_{\lambda\to 0} \xi^*_1 = 1$ and $\lim_{\lambda\to 0} \xi^*_2 = 0$.

The proposed non-iterative scheme is similar to the implicit factored method and other non-iterative schemes used for the non-linear Richards equation by Paniconi *et al.* [10]. They found that non-iterative schemes based on Newton and Picard linearizations of the Crank–Nicolson scheme were in principle competitive with the iterative schemes for low accuracy and did not

suffer stability restrictions on step size, but were generally inefficient due to a degradation of convergence from second to first-order. This degradation arises due to the truncation of the $O(\Delta t)$ term in the coefficient linearization used by Paniconi *et al.* [10] in the Crank–Nicolson scheme, yielding only first-order approximations of the non-linear coefficients. The derivation of the second-order convergent non-iterative linearization in this paper addresses this deficiency, yielding more accurate non-iterative methods. In addition, the use of the multistage Thomas–Gladwell method facilitates the design of adaptive truncation error control for the non-iterative time-stepping formulation.

EMPIRICAL CONVERGENCE ANALYSIS

An empirical study verifies the convergence of the schemes and illustrates the behaviour of the non-iterative formulation. The test problem consists of solving the spatially discrete Richards equation, a highly non-linear equation describing unsaturated flow through porous media and widely used in subsurface hydrogeology [12, 13]. The moisture form of Richards equation is given by

$$\frac{\partial \theta}{\partial t} - \nabla \cdot D(\theta)\nabla\theta + \frac{\partial K(\theta)}{\partial z} = 0 \tag{41}$$

where θ is the volumetric moisture content [dimensionless], z is the (positive downward) depth [L], t is time [T], $K(\theta)$ is the hydraulic conductivity [L/T] and $D(\theta)$ is the diffusivity [L^2/T].

The spatial approximation is accomplished using finite elements, yielding the semi-discrete moisture-based Richards equation

$$\mathbf{M}\frac{\mathrm{d}\boldsymbol{\theta}(t)}{\mathrm{d}t} + \mathbf{K}(\boldsymbol{\theta})\boldsymbol{\theta}(t) = \mathbf{F}(\boldsymbol{\theta}) \tag{42}$$

where $\boldsymbol{\theta}(t)$ is the moisture content at the finite element nodes at time t and $\mathbf{M}$, $\mathbf{K}$ and $\mathbf{F}$ are the global finite element matrices [9]. The vector $\mathbf{F}$ contains the gravity drainage term and the forcing functions (including the boundary conditions).

Adaptive time stepping is employed based on an assessment of the truncation error of the solution [9]. At the $(n+1)$th time step the error is given by the difference between the second-order update $\boldsymbol{\theta}^{n+1}$ (37) and a first-order estimate $\boldsymbol{\theta}_{(1)}^{n+1} = \boldsymbol{\theta}^n + \Delta t\dot{\boldsymbol{\theta}}^{n+1}$:

$$\mathbf{e}^{n+1} = \boldsymbol{\theta}_{(1)}^{n+1} - \boldsymbol{\theta}^{n+1} \tag{43}$$

The time step is accepted if the error is suitably small, that is, if

$$\|\mathbf{e}^{n+1}\|_r \equiv \max_i \left|\frac{e_i^{n+1}}{\theta_i^{n+1}}\right| \leqslant \tau \tag{44}$$

where τ is a user-prescribed tolerance [dimensionless]. A first estimate of the $(n+1)$th time step Δt_1^{n+1} in terms of (44) is given by

$$\Delta t_1^{n+1} = s\Delta t^n \sqrt{\frac{\tau}{\|\mathbf{e}^n\|_r}} \tag{45}$$

where $s \sim 0.8$ is a safety factor to limit the number of steps that just fail to meet the error requirement. If a step fails to satisfy (44), it is re-attempted with a smaller step size

$$\Delta t_{j+1}^{n+1} = s\Delta t_j^{n+1}\sqrt{\frac{\tau}{\|\mathbf{e}_j^{n+1}\|_r}} \tag{46}$$

where the subscript j indexes the consecutive time-step estimates. Additionally, the step size variation factor is constrained between $r_{\min} \sim 0.1$ and $r_{\max} \sim 4$ to prevent excessive variation if the asymptotic behaviour of the error control breaks down. A complete algorithmic description of the adaptive algorithm is available [8, 9].

Non-linear iteration is required for the solution of the standard Thomas–Gladwell formulation. In this study, Picard iteration is used, where the following linearized system is inverted at each iteration

$$[\mathbf{m} + \Delta t\mathbf{k}^{n+1,m}]\dot{\boldsymbol{\theta}}^{n+1,m+1} = -\mathbf{k}^{n+1,m}\boldsymbol{\theta}^n + \mathbf{f}^{n+1,m} \tag{47}$$

where m is the iteration counter. Iterations are terminated when a relative convergence test of the following form is satisfied:

$$\max_i \left| \frac{\boldsymbol{\theta}_i^{n+1,m+1} - \boldsymbol{\theta}_i^{n+1,m}}{\boldsymbol{\theta}_i^{n+1,m+1}} \right| \leqslant \tau_{\mathrm{PI}} \tag{48}$$

where τ_{PI} is the Picard iteration tolerance [dimensionless]. Here we set $\tau_{\mathrm{PI}} = 0.1\tau$.

A test study, adopted from Celia *et al.* [13] and Rathfelder and Abriola [14], verifies the second-order convergence of the Thomas–Gladwell schemes and the advantages of the non-iterative formulation. Moisture fluxes through a 60-cm vertical column of New Mexico soil with initially low moisture content are modelled. The non-linear van Genuchten constitutive relationships are used to describe the soil hydraulic properties:

$$K(\theta) = K_s\theta_e^{1/2}\{1 - (1 - \theta_e^{1/m})^m\}^2 \tag{49}$$

$$D(\theta) = \frac{(1-m)K_s}{\alpha m(\theta_s - \theta_r)}\,\theta_e^{(m-2)/2m}\left\{\frac{1}{(1-\theta_e^{1/m})^m} + (1-\theta_e^{1/m})^m - 2\right\} \tag{50}$$

where $\theta_e = (\theta - \theta_r)/(\theta_s - \theta_r)$, $\alpha = 0.0335$, $\theta_s = 0.368$, $\theta_r = 0.102$, $m = 0.5$ and $K_s = 0.00922\,\mathrm{cm/s}$.

Boundary moisture contents of $\theta(z = 0, t) = 0.2004$ and $\theta(z = 60, t) = 0.11$ are imposed, the initial conditions are defined as

$$\theta(z, t = 0) = \begin{cases} 0.11, & z \geqslant 0.6 \\ 0.2004 - \dfrac{0.2004 - 0.11}{0.6}\,z & 0 \leqslant z < 0.6 \end{cases} \tag{51}$$

These conditions induce a strong variation of temporal gradients in the solution and are well suited for the analysis of temporal accuracy.

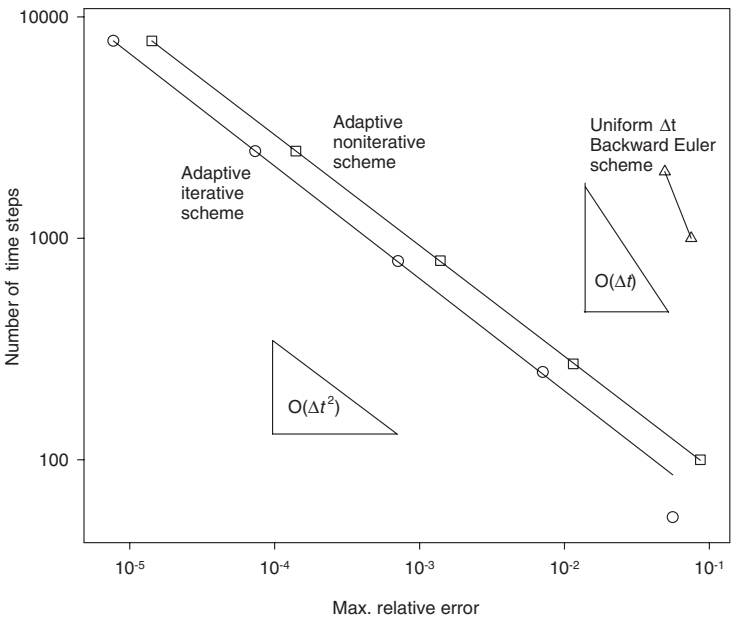

Figure 1. Convergence of iterative and non-iterative Thomas–Gladwell schemes. Adaptive time stepping used in both cases. For comparison, the uniform time step backward Euler solution used by Kavetski *et al.* [9] is also shown.

All solutions are obtained using identical spatial grids comprising 100 linear elements of uniform size. Identical spatial approximations ensure that the error measure reflects only temporal errors since, as the error tolerance $\tau \to 0$, the numerical solutions converge to the exact solution of (42) regardless of element size and type.

In the absence of an analytical solution, a surrogate 'exact' solution is required for algorithm benchmarking. The reference solution is approximated numerically by the iterative Thomas–Gladwell scheme with adaptive time stepping and very tight truncation error and iteration tolerances ($\tau = 10^{-7}$ and $\tau_{\mathrm{PI}} = 10^{-9}$). This ensures that any differences between the 'exact' and 'approximate' solutions are dominated by the truncation error of the 'approximate' solution. Since a relative error test is used in the truncation error controller, the actual errors are also expressed in a relative form,

$$\varepsilon(t^n) = \max_i \left| \frac{\theta_i^n - \bar{\theta}_i^n}{\bar{\theta}_i^n} \right| \tag{52}$$

where $\bar{\theta}_i^n$ is the 'exact' solution at the ith node.

Figure 1 and Table I show the number of times steps required to solve the problem as the error tolerance of the adaptive truncation error control is reduced. The table also shows the number of iterations and the number of failed steps for each solution. The second-order convergence of the iterative Thomas–Gladwell and the non-iterative scheme is evident, confirming that condition (6) is weakened to (20) when the order of the governing equation is reduced from second to first.

Table I. Computational efficiency of the iterative and non-iterative Thomas–Gladwell schemes. The number of iterations reported is the total of good (accepted) and failed time steps. The total computational effort can be assessed by comparing the number of iterations required for the iterative scheme with the total number of time steps required for the non-iterative scheme.

	Iterative Thomas–Gladwell scheme			Noniterative scheme	
Tolerance τ	Maximum error	Good/failed steps	Iterations	Maximum error	Good/failed steps
10^{-1}	5.58×10^{-2}	55/0	354	8.66×10^{-2}	100/0
10^{-2}	7.08×10^{-3}	249/3	1050	1.15×10^{-2}	271/33
10^{-3}	7.10×10^{-4}	788/1	2352	1.39×10^{-3}	792/8
10^{-4}	7.34×10^{-5}	2472/1	6917	1.40×10^{-4}	2474/1
10^{-5}	7.70×10^{-6}	7782/0	15759	1.42×10^{-5}	7783/0

The empirical results are also consistent with the analytic stability assessment, which predicted no stability restrictions on the step size for both the iterative and non-iterative algorithms.

The compatibility of the approximation scheme and adaptive time stepping is of practical significance. Figure 1 shows that step size variation does not lead to order reduction for both the iterative and non-iterative methods. In addition, Table I shows that in nine out of 10 runs, the failure rate was $\leqslant 1\%$, i.e. only 1% of the attempted steps failed to meet satisfy (44) and had to be repeated with a smaller step size. In one test run, the adaptive non-iterative scheme had a higher failure rate ($\sim 10\%$), which was caused by steps that just failed to meet the error requirement. In these cases, the subsequent automatic time-step attempts were always successful. For the non-iterative scheme, failed steps are very cheap since a single iteration is performed for all time steps. Conversely, a time-step failure of an iterative scheme implies that several iterations are wasted and is hence more costly.

The computational effort required for each solution can be established by examining the number of iterations required to obtain each solution. The non-iterative formulation is about 2–3 times more efficient than the iterative scheme. However, the truncation error of the non-iterative scheme is slightly larger than that of the original iterative scheme for similar time-step sizes due to the additional linearization error in the non-linear coefficients. However, the non-iterative scheme maintains second-order accuracy and the lack of iterations makes it very competitive with analogous time approximations that employ iterative solvers. Further application of the new non-iterative scheme to other forms of the non-linear Richards equation can be found in Reference [11].

CONCLUSIONS

A comprehensive consistency and stability analysis of the two-stage Thomas–Gladwell family of approximations to first-order non-linear DE systems is presented. The reduction of the governing DEs from second to first-order weakens the parameter constraints of the Thomas–Gladwell methods required for second-order accuracy and stability. The method used to evaluate the non-linear coefficients is also discussed, demonstrating that, from the point of view of

maintaining the numerical accuracy of the approximation and increasing efficiency, valuable flexibility is available in the evaluation of these terms. Non-iterative analogues of the standard Thomas–Gladwell methods are discussed. A compact non-iterative linearization is recommended, which preserves second-order accuracy and stability of the time-stepping scheme and improves the computational efficiency relative to iterative implementations.

REFERENCES

1. Thomas RM, Gladwell I. Variable-order variable-step algorithms for second-order systems. Part 1: The methods. *International Journal for Numerical Methods in Engineering* 1988; **26**:39–53.
2. Tocci MD, Kelly CT, Miller CT. Accurate and economical solution of the pressure-head form of Richards' equation by the method of lines. *Advances in Water Resources* 1997; **20**:1–14.
3. Wood WL. *Practical Time Stepping Schemes*. Oxford University Press: Oxford, 1990.
4. Zienkiewicz OC, Wood WL, Hine NL, Taylor RL. A unified set of single step algorithms, Part 1: General formulation and application. *International Journal for Numerical Methods in Engineering* 1984; **23**:1343–1353.
5. Kahaner D, Moler C, Nash S. *Numerical Methods and Software*. Prentice-Hall: Englewood Cliffs, NJ, 1989.
6. Gladwell I, Thomas RM. Variable-order variable-step algorithms for second-order systems. Part 2: The codes. *International Journal for Numerical Methods in Engineering* 1988; **26**:55–80.
7. Sloan SW, Abbo AJ. Biot consolidation analysis with automatic time stepping and error control. Part 1: Theory and implementation. *International Journal for Numerical and Analytical Methods in Geomechanics* 1999; **23**:467–492.
8. Kavetski D, Binning P, Sloan SW. Adaptive time stepping and error control in a mass conservative numerical solution of the mixed form of Richards equation. *Advances in Water Resources* 2001; **24**:595–605.
9. Kavetski D, Binning P, Sloan SW. Adaptive backward Euler time stepping with truncation error control for numerical modelling of unsaturated fluid flow. *International Journal for Numerical Methods in Engineering* 2002; **53**:1301–1322.
10. Paniconi C, Aldama AA, Wood EF. Numerical evaluation of iterative and non-iterative methods for the solution of the non-linear Richard's equation. *Water Resources Research* 1991; **27**:1147–1163.
11. Kavetski D, Binning P, Sloan SW. Noniterative time stepping schemes with adaptive truncation error control for the solution of Richards equation. *Water Resources Research* 2002; **38**:29.1–29.10.
12. Huyakorn PS, Pinder GF. *Computational Methods in Subsurface Flow*. Academic Press Inc.: San Diego, 1985.
13. Celia MA, Bouloutas ET, Zarba RL. A general mass-conservative numerical solution for the unsaturated flow equation. *Water Resources Research* 1990; **26**:1483–1496.
14. Rathfelder K, Abriola LM. Mass conservative numerical solutions of the head-based Richards equation. *Water Resources Research* 1994; **30**:2579–2586.

