BIT Numer Math
DOI 10.1007/s10543-014-0534-9

BIT

A posteriori error analysis for finite element methods with projection operators as applied to explicit time integration techniques

J. B. Collins · D. Estep · S. Tavener

Received: 3 November 2012 / Accepted: 26 September 2014

Abstract We derive a posteriori error estimates for two classes of explicit finite difference schemes for ordinary differential equations. To facilitate the analysis, we derive a systematic reformulation of the finite difference schemes as finite element methods. The a posteriori error estimates quantify various sources of discretization errors, including effects arising from explicit discretization. This provides a way to judge the relative sizes of the contributions, which in turn can be used to guide the choice of various discretization parameters in order to achieve accuracy in an efficient way. We demonstrate the accuracy of the estimate and the behavior of various error contributions in a set of numerical examples.

Communicated by Ralf Hiptmair.

This research is supported in part by the Defense Threat Reduction Agency (HDTRA1-09-1-0036), Department of Energy (DE-FG02-04ER25620, DE-FG02-05ER25699, DE-FC02-07ER54909, DE-SC0001724, DE-SC0005304, INL00120133), Idaho National Laboratory (00069249, 00115474), Lawrence Livermore National Laboratory (B584647, B590495), National Science Foundation (DMS-0107832, DMS-0715135, DGE-0221595003, MSPA-CSE-0434354, ECCS-0700559, DMS-1065046, DMS-1016268, DMS-FRG-1065046), National Institutes of Health (#R01GM096192).

J. B. Collins
Department of Mathematics, Chemistry, and Physics, Western Texas A&M University, Canyon, TX 79016, USA
e-mail: jcollins@wtamu.edu

D. Estep (✉)
Department of Statistics, Colorado State University, Fort Collins, CO 80523, USA
e-mail: estep@stat.colostate.edu

S. Tavener
Department of Mathematics, Colorado State University, Fort Collins, CO 80523, USA
e-mail: tavener@math.colostate.edu

Published online: 11 December 2014

Keywords A posteriori error estimate · Explicit schemes · Ordinary differential equations

Mathematics Subject Classification 65L02 · 65G02

1 Introduction

We develop and test computational error estimates for two classes of explicit time integration schemes for an ordinary differential equation: Find $y \in C^1([0,T];\mathbb{R}^d)$ satisfying,

$$\begin{cases} \dot{y}(t) = f(y(t),t), & 0 < t \leq T, \\ y(0) = y_0, \end{cases} \tag{1.1}$$

where $\dot{y}(t) = \frac{d}{dt}y(t)$, $y_0 \in \mathbb{R}^d$, $f : \mathbb{R}^d \times [0,T] \to \mathbb{R}^d$ is differentiable, and $C^i([0,T];\mathbb{R}^d)$ denotes the usual space of $\mathbb{R}^d$-valued functions on $[0,T]$ with i continuous derivatives. The computational estimates are based on an a posteriori error analysis that uses variational arguments, adjoint problems, and computable residuals to produce an error estimate for a specified quantity of interest that quantifies the contributions from all sources of discretization error. The application-specific quantity of interest is given by a linear functional $\mathcal{Q} : \mathbb{R}^d \to \mathbb{R}$. Typical examples include weighted averages over particular time intervals, or the value at a point in time. Such "goal oriented" a posteriori error estimates are widely employed for finite element methods [1,2,4,10,14]. We note that this paper is concerned with estimating the total error in a quantity of interest. This aim is significantly different than estimating the truncation error ("local error") of an approximation, e.g. as is the goal for the classic RK45 method.

Prior work [4,6,12,14,15] on a posteriori error estimates for initial value problems has focused on implicit rather than explicit methods. One important reason is that this simplifies the definition of an adjoint problem. In recent years, a posteriori analysis has been extended to include the effects of finite iteration and operator splitting in computation of implicit approximations [7]. See also [5] that addresses stability issues in explicit time stepping. The main goal of this paper is to systematically extend a posteriori error analysis to explicit time stepping methods and thereby quantify the specific consequences of using explicit versus implicit discretization schemes.

The variational analysis is facilitated by adopting a finite element formulation of finite difference methods [3,13,19]. We construct finite element descriptions of several popular explicit time integration schemes that allow for a posteriori error estimation. In order to do so, we introduce special operators in the formulation of the numerical method, and then quantify the effects of these operators on the numerical error.

In Sect. 2, we construct two classes of finite element methods for solving initial value problems (1.1) and demonstrate how particular choices for the approximation space and quadrature produce finite element methods that yield the same solution as specific *implicit* finite difference schemes at a given set of nodes. The analysis

provides the means to distinguish discretization and quadrature error. In Sect. 3, we extend the analysis to construct finite element methods that are nodally equivalent to specific *explicit* finite difference schemes and estimate discretization, quadrature and "explicit" errors for these schemes. In Sect. 4, we report the results for a range of numerical experiments. In Sect. 5, we investigate the difference between the adjoints to the continuous and discretized solution operators and present an alternative error analysis that takes into account this difference.

2 Finite element description of implicit finite difference methods

To construct a finite element description of the finite difference schemes, we employ the following variational formulation of (1.1): Find $y \in C^1([0,T];\mathbb{R}^d)$ such that

$$\begin{cases} \mathcal{N}(y,v) := \langle \dot{y} - f(y,t), v\rangle_{[0,T]} = 0, \quad \forall\, v \in C^0([0,T];\mathbb{R}^d), \\ y(0) = y_0, \end{cases} \tag{2.1}$$

where $\langle\cdot,\cdot\rangle_{[a,b]}$ and $(\cdot,\cdot)$ denote the $L^2([a,b];\mathbb{R}^d)$ and $\mathbb{R}^d$ inner products respectively. The construction is divided into two stages. In the first stage (Sect. 2.1), we approximate the solution space by a space of piecewise polynomial functions. In the second stage (Sect. 2.2), we introduce various approximations to both the integrand and the integral in (2.1).

2.1 Approximation of the solution space

We begin by constructing and analyzing the finite element approximation assuming all integrals in the variational formulation are evaluated exactly. We consider the continuous and discontinuous Galerkin methods [6,13] which produce piecewise polynomial approximations on the domain $[0,T]$ corresponding to a grid,

$$0 = t_0 < t_1 < \cdots < t_{N-1} < t_N = T,$$

with time steps $k_n = t_n - t_{n-1}$ and subintervals $I_n = [t_{n-1}, t_n]$. The space of continuous piecewise polynomials is,

$$\mathcal{C}^q([0,T]) = \{w \in C^0([0,T];\mathbb{R}^d) : w|_{I_n} \in \mathcal{P}^q(I_n),\ 1 \le n \le N\},$$

where $\mathcal{P}^q(I_n)$ is the space of polynomials of degree $\le q$ valued in $\mathbb{R}^d$ on I_n. The space of discontinuous piecewise polynomials is,

$$\mathcal{D}^q([0,T]) = \{w : w|_{I_n} \in \mathcal{P}^q(I_n),\ 1 \le n \le N\}.$$

The continuous Galerkin method of order $q+1$, cG(q), is defined interval-by-interval as: Find $Y \in V = \mathcal{C}^q([0,T])$ such that $Y(0) = y_0$ and for $n = 1, \ldots, N$,

$$\begin{cases} \langle \dot{Y} - f(Y,t), v_k \rangle_{I_n} = 0, \quad \forall\, v_k \in \tilde{V} = \mathcal{P}^{q-1}(I_n), \\ Y(t_{n-1}^+) = Y(t_{n-1}^-). \end{cases} \tag{2.2}$$

The interval-by-interval formulation of the discontinuous Galerkin method of order $q+1$, dG(q), is: Find $Y \in V = \mathcal{D}^q([0,T])$ such that $Y(0^-) = y_0$ and for $n = 1, \ldots, N$,

$$\langle \dot{Y} - f(Y,t), v_k \rangle_{I_n} + ([Y]_{n-1}, v_k(t_{n-1}^+)) = 0, \ \forall\, v_k \in \tilde{V} = \mathcal{P}^q(I_n), \tag{2.3}$$

where $[Y]_n = Y(t_n^+) - Y(t_n^-)$.

The discretizations (2.2) and (2.3) yield a (nonlinear) system of equations for the coefficients of the approximation with respect to the chosen basis of $\mathcal{P}^q(I_n)$ in each interval. For linear constant coefficient problems, these approximations agree with some standard finite difference schemes at node values, e.g., at nodes, dG(0) agrees with the backward Euler, dG(1) agrees with a subdiagonal Pade scheme, and cG(1) with Crank-Nicolson. In general, we say two approximations are *nodally equivalent* if they yield the same approximation values on any given set of nodes $\{t_n\}$ that partition the domain. The dG and cG approximations are *not* nodally equivalent with any commonly encountered finite difference scheme for nonlinear problems in general.

2.1.1 A priori convergence properties

A priori analysis [6] shows that dG(q) is order $q+1$ at every point in time and has superconvergent order $2q+1$ at time nodes t_n under certain conditions. For a linear constant coefficient problem, the extra accuracy obtained at time nodes agrees with the expected accuracy of the nodally equivalent finite difference scheme. Likewise, the cG(q) scheme is order $q+1$ globally with superconvergence $2q$ at time nodes.

2.1.2 A posteriori error analysis

The analysis begins with the definition of a suitable adjoint problem. There are many ways to define an adjoint problem for a nonlinear problem [16]. We use a standard choice for the analysis of implicit methods. We represent the quantity of interest by $\mathcal{Q}(y) = \langle y, \psi \rangle_{[0,T]} + (y(T), \psi_T)$, where $\psi : [0,T] \to \mathbb{R}^d$, $\psi_T \in \mathbb{R}^d$. The choice of ψ and ψ_T is application dependent. Then the adjoint problem is defined as follows. With $\varphi(T) = \psi_T$, for $n = N, N-1, \ldots, 1$, find $\varphi \in C^1(I_n; \mathbb{R}^d)$ such that

$$\begin{cases} \langle v, -\dot{\varphi} - \bar{A}^* \varphi \rangle_{I_n} = \langle v, \psi \rangle_{I_n}, \quad \forall\, v \in C^0(I_n; \mathbb{R}^d), \\ \varphi(t_n^-) = \varphi(t_n^+), \end{cases} \tag{2.4}$$

where

$$\bar{A} = \int_0^1 f'(sy + (1-s)Y, t)ds,$$

$f' = \frac{\partial f}{\partial y}$ and ψ, ψ_T define the quantity of interest. We note that $\varphi \in C^0([0, T]; \mathbb{R}^d)$.

The standard a posteriori error estimate for both cG(q) and dG(q) discretization is provided in Theorem 1 [6,18].

Theorem 1 (General Error Representation Formula) *If $Y(t)$ is an approximation of (2.1) obtained via the cG(q) method, and the error is defined by $e(t) = y(t) - Y(t)$, then the error in the quantity of interest defined by ψ and ψ_T is given by,*

$$\langle e, \psi \rangle_{[0,T]} + (e(T), \psi_T) = \sum_{n=1}^{N} \langle \mathcal{R}(Y), \varphi - \pi_k \varphi \rangle_{I_n}, \tag{2.5}$$

where φ solves the adjoint problem (2.4), π_k is a projection onto $\tilde{V}$, and $\mathcal{R}(Y) = f(Y, t) - \dot{Y}$ is evaluated in the interior of each interval.

If instead, $Y(t)$ is a dG(q) approximation, then

$$\begin{aligned} &\langle e, \psi \rangle_{[0,T]} + (e(T), \psi_T) \\ &= \sum_{n=1}^{N} \left(\langle \mathcal{R}(Y), \varphi - \pi_k \varphi \rangle_{I_n} - \left([Y]_{n-1}, \varphi(t_{n-1}) - \pi_k \varphi(t_{n-1}^+) \right) \right). \end{aligned} \tag{2.6}$$

Remark 2.1 In practice, the exact adjoint solution φ is not available. We first approximate

$$\bar{A} = \int_0^1 f'(sy + (1-s)Y, t)ds \approx f'(Y, t),$$

which is reasonable by a simple Taylor's theorem argument provided $Y \approx y$. Since $Y(0) = y_0$ for cG(q) and $Y(0)^- = y_0$ for dG(q), this is guaranteed to hold for at least an initial transient period by the standard convergence results. Then we compute a numerical solution of the resulting adjoint problem. In order to evaluate the "Galerkin orthogonality" weight $\varphi - \pi_k \varphi$, which amounts to estimating derivatives of the adjoint solution, we either use a higher order method or finer time steps to solve the adjoint problem.

2.1.3 Illustrative example

To illustrate the effects of subsequent stages of discretization, we present results for the very simple linear problem,

$$\dot{y}(t) = \begin{bmatrix} 0 & \mathrm{e}^{0.2t} \\ -\mathrm{e}^{0.2t} & 0 \end{bmatrix} y(t), \quad t \in [0, T], \qquad y(0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{2.7}$$

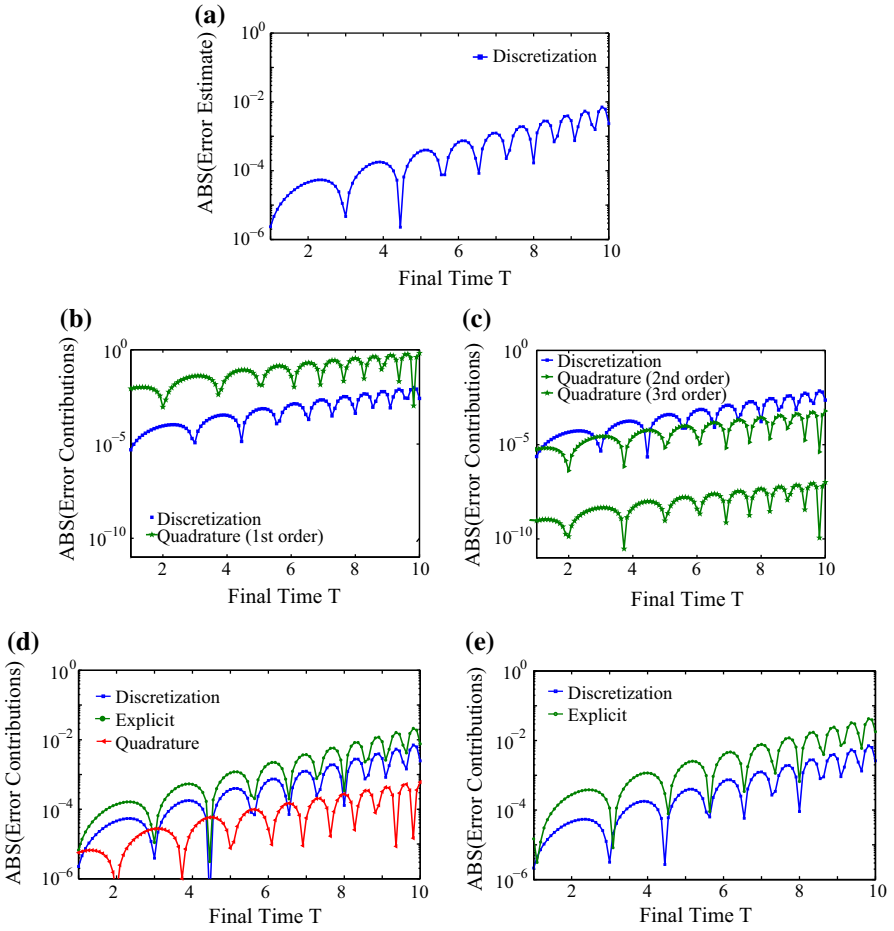

Fig. 1 **a** The a posteriori error estimate (2.5) for the cG(1) approximation with exact quadrature for (2.7) versus final time T. **b** Absolute error contributions versus final time using the cG(1) scheme with first order quadrature; **c** Absolute error contributions versus final time for the cG(1) method using second and third order quadratures. **d** Absolute error contributions using the explicit trapezoid method, refer to Theorem 3; **e** Absolute error contributions using the second-order Adams–Bashforth method, refer to Theorem 4.

The quantity of interest is defined by $\psi = (0, 0)^\top$, $\psi_T = (1, 1)^\top$ giving the error at the final time. The adjoint problem is solved numerically using cG(2), and in each example we use a step size of $k_n = 0.1$. In Fig. 1a, we show the error estimate versus final time for the cG(1) method. There is an overall increasing exponential trend in the errors, yet there is also substantial accumulation and cancellation of errors.

2.2 Approximation of the differential operator using quadrature

Next, we consider approximations of the differential operator based on applying quadrature to the integrals in the weak formulation. The specific quadratures are cho-

sen as part of the effort to produce finite element approximations that are nodally equivalent to specific finite difference schemes.

The cG(q) method with quadrature is written as: Find $Y \in \mathcal{C}^q([0,T])$ such that $Y(0) = y_0$ and for $n = 1, \ldots, N$,

$$\begin{cases} \langle \dot{Y}, v_k \rangle_{I_n} - \langle f(Y,t), v_k \rangle_{I_n,K_n} = 0, \quad \forall\, v_k \in \mathcal{P}^{q-1}(I_n), \\ Y(t_{n-1}^+) = Y(t_{n-1}^-), \end{cases} \tag{2.8}$$

where the nonlinear term uses the approximate inner product

$$\langle g, h \rangle_{I_n,K_n} = \sum_{i=1}^{K_n} g(s_{i,n}) h(s_{i,n}) w_{i,n},$$

defined by nodes $s_{i,n}$ and weights $w_{i,n}$ associated with I_n. A similar change is made to the discontinuous Galerkin method to implement quadrature.

For example, the cG(1) method with the trapezoid rule quadrature gives

$$Y(t_n) = Y(t_{n-1}) + \frac{k_n}{2}(f(Y(t_{n-1}), t_{n-1}) + f(Y(t_n), t_n)),$$

so the cG(1) method with trapezoid rule quadrature is nodally equivalent to the implicit trapezoid method. Similarly, the cG(1) method with midpoint rule quadrature is nodally equivalent to the implicit midpoint method.

2.2.1 A priori convergence results

Following standard finite element convergence analysis, using a quadrature formula of sufficient accuracy preserves the nominal optimal order of convergence of the method. Superconvergence results are more difficult to show.

2.2.2 A posteriori error analysis

We use the standard adjoint problem (2.4) for a posteriori error estimation. This yields the following result [18].

Theorem 2 (Quadrature Error Representation Formula) *If $Y(t)$ is an approximation of (2.1) obtained via the cG(q) method with a quadrature defined by the inner product $\langle \cdot, \cdot \rangle_{I_n,K_n}$, then the error in the quantity of interest defined by ψ and ψ_T is given by,*

$$\langle e, \psi \rangle_{[0,T]} + (e(T), \psi_T) = \sum_{n=1}^{N} \Bigg(\underbrace{\langle \mathcal{R}(Y), \varphi - \pi_k \varphi \rangle_{I_n}}_{\text{Discretization Contribution}} + \underbrace{\langle f(Y,t), \pi_k \varphi \rangle_{I_n} - \langle f(Y,t), \pi_k \varphi \rangle_{I_n,K_n}}_{\text{Quadrature Contribution}} \Bigg), \tag{2.9}$$

where $\mathcal{R}(Y), \varphi$ and π_k are defined as above. For the dG(q) approximation, the discretization contribution is replaced by,

$$\langle \mathcal{R}(Y), \varphi - \pi_k \varphi \rangle_{I_n} - ([Y]_{n-1}, \varphi(t_{n-1}) - \pi_k \varphi(t_{n-1}^+)).$$

The estimates distinguish contributions from discretization of the solution space from the effects of quadrature. The discretization contribution and the quadrature contribution use different adjoint weights, $\varphi - \pi_k \varphi$ and $\pi_k \varphi$ respectively, reflecting the fact that these two sources of discretization error accumulate, propagate, and cancel in different ways in general. One practical consequence, is that we can use the relative size of the contributions to identify the dominant source of error. In the case that quadrature contribution dominates, we can decrease the error very efficiently by simply using a higher order quadrature formula. We illustrate below.

2.2.3 Illustrative example

For the linear system (2.7), we solve (2.2) with cG(1) and with quadratures of three different orders. The discretization contributions are roughly the same regardless of the quadrature. However, the first order quadrature contribution dominates the discretization contribution to the error, while for third order quadrature, the discretization contribution dominates. The error contributions are roughly the same when second order quadrature is used. Generally, using a higher order quadrature to evaluate the integrals is less computationally expensive than increasing the order of approximation or using a smaller time step for the finite element discretization.

3 Finite element description of explicit finite difference methods

We next consider the explicit discretizations forward Euler, explicit trapezoid, Runge–Kutta 4, and the family of Adams–Bashforth methods. Extending the approach to other explicit schemes is possible on a case-by-case basis. We treat explicit schemes in the finite element framework by using *extrapolation* of some approximation of f in the variational integrals. The extrapolation is formulated in terms of a certain operator inserted in the variational equation. We discuss two approaches that yield different families of schemes. For the purpose of intuition, it is useful to consider the "modified equation" that results from building the extrapolation operators into the original differential equation.

The first extrapolation operator is based on a local series expansion, and yields a family of one-step explicit methods such as Runge–Kutta. Often these expansions are approximations of a truncated Taylor series, though they can be more general. Formally, Eq. (2.1) is modified to become

$$\begin{cases} \sum_{n=1}^{N} \langle \dot{\tilde{y}} - \sum_{i=1}^{L} \alpha_i f(P_n^i \tilde{y}, t), v_k \rangle_{I_n}, \quad \forall\, v_k \in \tilde{V} = \mathcal{P}^{q-1}(I_n), \\ \tilde{y}(0) = y_0, \end{cases} \tag{3.1}$$

where $\sum_{i=1}^{L} \alpha_i = 1$ and the operators P_n^i are described below. Similarly for a discontinuous Galerkin method.

The second extrapolation operator replaces f by an extrapolation of a polynomial interpolant computed from previous time nodes. This approach yields a family of methods that include the multi-step explicit methods such as Adams-Bashforth. Formally, Eq. (2.1) is modified to become

$$\begin{cases} \sum_{n=1}^{\ell-1} \langle \dot{\tilde{y}} - f(\tilde{y}, t), v \rangle_{I_n} + \sum_{n=\ell}^{N} \langle \dot{\tilde{y}} - Q_n^{\ell} f(\tilde{y}, t), v_k \rangle_{I_n} = 0, \quad \forall\, v_k \in \tilde{V} = \mathcal{P}^{q-1}(I_n), \\ \tilde{y}(0) = y_0, \end{cases} \tag{3.2}$$

where ℓ and the operators Q_n^{ℓ} are described in Sect. 3.2.

3.1 Taylor series-like approximation

We express the approximation as the result of an operator P applied to piecewise polynomials. The operator P is the composition of two operators $P = TS$. At a given node t_{n-1}, the operator P maps a $BV([0, T]; \mathbb{R}^d)$ function to a polynomial defined on $[t_{n-2}, t_n]$. The first operator S projects a function with limited regularity into a space of functions with sufficient regularity for a series constructed using derivatives to be defined. This is needed since we apply P to finite element functions that have discontinuities in value and/or derivative at time nodes. Given n and $v \in BV([0, T]; \mathbb{R}^d)$ on an interval containing time nodes $\{t_i,\ i \in \mathcal{I}_n\}$, we define S as the polynomial that interpolates v with values $\{v(t_i^-), i \in \mathcal{I}_n\}$. Typically, $\mathcal{I}_n$ includes $n, n-1, n-2, \ldots$ for the number of nodes equal to the order of the series expansion.

The second operator T maps a function $v \in C^k([t_{n-2}, t_n]; \mathbb{R}^d)$ to the series,

$$Tv = v(t_{n-1}) + \sum_{i=1}^{k} \frac{d^{(i-1)}}{dt^{(i-1)}} f(v(t_{n-1}), t_{n-1}) \frac{(t - t_{n-1})^i}{i!},$$

where the time derivatives of f are computed using the chain rule. Note that if v is a solution of the differential equation, Tv is the Taylor polynomial of order k.

We denote the restriction of P to $[t_{n-1}, t_n]$ by P_n. The cG(q) method incorporating a Taylor-series like approximation is: Find $Y \in \mathcal{C}^q(I_n)$ such that $Y(0) = y_0$ and for $n = 1, \ldots, N$,

$$\begin{cases} \langle \dot{Y}, v_k \rangle_{I_n} = \langle f(P_n Y, t), v_k \rangle_{I_n, K_n}, \quad \forall\, v_k \in \mathcal{P}^{q-1}(I_n), \\ Y(t_{n-1}^+) = Y(t_{n-1}^-). \end{cases} \tag{3.3}$$

The definition of the dG(q) method is analogous.

We note that these cG and dG approximations can be obtained by applying the finite element discretizations including quadrature, to the nominal modified problem (3.1).

3.1.1 Examples

As a simple example, we consider $L = 1$ (and therefore $\alpha_1 = 1$) and a series of order zero $P_n Y(t) = Y(t_{n-1})$. Using dG(0) and left hand quadrature, the modified problem becomes,

$$Y(t_n^-) = Y(t_{n-1}^-) + k_n f(Y(t_{n-1}^-), t_{n-1}^-), \quad n = 1, \ldots, N,$$

which is the update formula for the forward Euler method.

Next, we consider $L = 1$ (and therefore $\alpha_1 = 1$) and a series of order one,

$$P_n Y = Y(t_{n-1}) + f(Y(t_{n-1}), t_{n-1})(t - t_{n-1}).$$

The cG(1) approximation is determined by,

$$Y(t_n) = Y(t_{n-1}) + \langle f\big(Y(t_{n-1}) + f(Y(t_{n-1}), t_{n-1})(t - t_{n-1})\big), 1\rangle_{I_n, K_n}, \quad n = 1, \ldots, N.$$

By varying the quadrature, we can obtain different finite difference schemes. For example, the midpoint quadrature rule yields the explicit midpoint method [17],

$$\begin{aligned} \hat{Y}_n &= Y(t_{n-1}) + \frac{k_n}{2} f(Y(t_{n-1}), t_{n-1}) \\ Y(t_n) &= Y(t_{n-1}) + k_n f(\hat{Y}_n, t_{n-1/2}), \end{aligned}$$

with $t_{n-1/2} = t_n - \frac{k_n}{2}$. If we use the trapezoid rule for quadrature, we obtain the explicit trapezoid, Heun's or RK2 method [17],

$$\begin{aligned} \hat{Y}_n &= Y(t_{n-1}) + k_n f(Y(t_{n-1}), t_{n-1}) \\ Y(t_n) &= Y(t_{n-1}) + \frac{k_n}{2}\big(f(Y(t_{n-1}), t_{n-1}) + f(\hat{Y}_n, t_n)\big). \end{aligned}$$

As mentioned, changing the quadrature formula can be an efficient way to improve accuracy. For example, replacing the trapezoid rule in the RK2 method with Simpson's rule gives a method we call RK2/4,

$$\begin{aligned} Y_1 = Y(t_{n-1}), \quad Y_2 = Y_1 + \frac{k_n}{2} f(Y_1, t_{n-1}), \quad Y_3 = Y_1 + k_n f(Y_1, t_{n-1}) \\ Y(t_n) = Y(t_{n-1}) + \frac{k_n}{6}\big(f(Y_1, t_{n-1}) + 4 f(Y_2, t_{n-1/2}) + f(Y_3, t_n)\big). \end{aligned}$$

Note that while we use a fourth order quadrature in RK2/4 with the cost of a single additional function evaluation, the method is still second order overall. Nonetheless, this can lead to significant improvement in accuracy when the quadrature error contribution in the RK2 approximation is dominant, as we illustrate below.

3.1.2 Runge–Kutta 4

The fourth order Runge–Kutta method (RK4),

$$\begin{aligned} Y^1 &= Y(t_{n-1}),\, Y^2 = Y^1 + \frac{k_n}{2} f(Y^1, t_{n-1}) \\ Y^3 &= Y^1 + \frac{k_n}{2} f(Y^2, t_{n-1/2}),\;\; Y^4 = Y^1 + k_n f(Y^3, t_{n-1/2}) \\ Y(t_n) &= Y(t_{n-1}) + \frac{k_n}{6}\big(f(Y^1, t_{n-1}) + 2f(Y^2, t_{n-1/2}) \\ &\quad + 2f(Y^3, t_{n-1/2}) + f(Y^4, t_n)\big), \end{aligned}$$

is a popular method since it has fairly high order with relatively few function evaluations.

We use four series approximations to construct the equivalent finite element method. The first mapping is $P_n^1 y = y(t_{n-1})$ while the second mapping is

$$P_n^2 y = y(t_{n-1}) + f(y(t_{n-1}), t_{n-1})(t - t_{n-1}).$$

The third and fourth mappings are more complicated. We introduce the projection $\hat{P}_n y = y(t_{n-1/2})$. Expanding y in its Taylor series, we have,

$$\begin{aligned} y(t) &= y(t_{n-1}) + \int_{t_{n-1}}^{t} f(y(s), s)\, ds \approx y(t_{n-1}) + \hat{P}_n f(y, t)(t - t_{n-1}) \\ &\approx y(t_{n-1}) + \hat{P}_n f(P_n^2 y, t)(t - t_{n-1}). \end{aligned}$$

We define the third mapping as,

$$P_n^3 Y = Y(t_{n-1}) + \hat{P}_n f(P_n^2 Y, t)(t - t_{n-1}).$$

With similar motivation using P_n^3, we define the fourth mapping,

$$P_n^4 Y = Y(t_{n-1}) + \hat{P}_n f(P_n^3 Y, t)(t - t_{n-1}).$$

Next, we apply the cG(3) method with approximate inner products $\langle g, h\rangle_{I_n,L}$, $\langle g, h\rangle_{I_n,M}$ and $\langle g, h\rangle_{I_n,R}$ defined to be the left hand, midpoint and right hand quadrature rules respectively. The resulting finite element method is: Find $Y \in \mathcal{C}^3(I_n)$ such that $Y(0) = y_0$ and for $n = 1, \ldots, N$,

$$\begin{cases} \langle \dot{Y}, v_k\rangle_{I_n} &= \frac{1}{6}\big[\langle f(P_n^1 Y, t), v_k\rangle_{I_n,L} + 2\langle f(P_n^2 Y, t) + f(P_n^3 Y, t), v_k\rangle_{I_n,M} \\ &\qquad + \langle f(P_n^4 Y, t), v_k\rangle_{I_n,R}\big] \quad \forall\, v_k \in \mathcal{P}^2(I_n), \\ Y(t_{n-1}^+) &= Y(t_{n-1}^-). \end{cases} \tag{3.4}$$

Nodal equivalence to RK4 is easily proven, as the piecewise constant function is a test function for cG(3). It is also straightforward to show that $P_n^1 Y(t_{n-1}) = Y^1$, $P_n^2 Y(t_{n-1/2}) = Y^2$, $P_n^3 Y(t_{n-1/2}) = Y^3$, and $P_n^4 Y(t_n) = Y^4$.

The additional degrees of freedom of the cG(3) finite element approximation within I_n can be determined explicitly from the values of the approximation at t_{n-1} and t_n using (3.4).

While the finite difference scheme is fourth order, the approximations to the Taylor series and the quadrature used in the finite element formulation are less than fourth order. This confirms that the RK4 scheme achieves its order through a special cancelation of error contributions.

3.1.3 A posteriori error analysis

Because truncated Taylor series become exact in the limit of increasing order, the sequence of "explicit" modified problems (3.1) approaches the true problem in the limit of increasing order. This suggests that using the same adjoint problem for error analysis as used to treat implicit methods is reasonable, at least in the limit of increasing order.

Theorem 3 (Taylor Series Error Representation Formula) *If $Y(t)$ is an approximation of (3.1) obtained via the cG(q) method with quadrature defined by the inner product $\langle \cdot, \cdot \rangle_{I_n, K_n^i}$, then the error in the quantity of interest defined by ψ and ψ_T is given by,*

$$\begin{aligned}\langle e, \psi \rangle_{[0,T]} + (e(T), \psi_T) = \sum_{n=1}^{N} \Bigg(& \underbrace{\langle \mathcal{R}_P(Y), \varphi - \pi_k \varphi \rangle_{I_n}}_{\textit{Discretization Contribution}} \\ & \underbrace{+ \langle f(Y,t) - \sum_{i=1}^{L} \alpha_i f(P_n^i Y, t), \varphi \rangle_{I_n}}_{\textit{Explicit Contribution}} \\ & \underbrace{+ \sum_{i=1}^{L} \alpha_i \langle f(P_n^i Y, t), \pi_k \varphi \rangle_{I_n} - \sum_{i=1}^{L} \alpha_i \langle f(P_n^i Y, t), \pi_k \varphi \rangle_{I_n, K_n^i}}_{\textit{Quadrature Contribution}} \Bigg),\end{aligned}$$

where φ and π_k are defined as above, and $\mathcal{R}_P(Z) = \sum_{i=1}^{L} \alpha_i f(P^i Z, t) - \dot{Z}$ is the modified residual.

For the dG(q) approximation, the discretization contribution is replaced by,

$$\langle \mathcal{R}_P(Y), \varphi - \pi_k \varphi \rangle_{I_n} - ([Y]_{n-1}, \varphi(t_{n-1}) - \pi_k \varphi(t_{n-1}^+)).$$

Proof We begin by defining the following nonlinear form,

$$\mathcal{N}_P(Z, v) = \sum_{n=1}^{N} \langle \dot{Z} - \sum_{i=1}^{L} \alpha_i f(P_n^i Z, t), v \rangle_{I_n}, \tag{3.5}$$

for $Z \in H^1([0,T];\mathbb{R}^d)$ and $v \in L^2([0,T];\mathbb{R}^d)$. We evaluate (2.1) and (3.5) at y and Y respectively and subtract to obtain,

$$\begin{aligned}\sum_{n=1}^{N}\langle \mathcal{R}_P(Y), v\rangle_{I_n} &= \mathcal{N}(y,v) - \mathcal{N}_P(Y,v)\\ &= (\mathcal{N}(y,v) - \mathcal{N}(Y,v)) + (\mathcal{N}(Y,v) - \mathcal{N}_P(Y,v))\,.\end{aligned}$$

This yields a relation between the error and the residual,

$$\sum_{n=1}^{N}\langle \mathcal{R}_P(Y), v\rangle_{I_n} = \sum_{n=1}^{N}\left(\langle \dot{e} - \bar{A}e, v\rangle_{I_n} - \langle f(Y,t) - \sum_{i=1}^{L}\alpha_i f(P_n^i Y, t), v\rangle_{I_n}\right).$$

Using the definition of φ in Eq. (2.4) and integrating by parts, we obtain,

$$\begin{aligned}\langle e,\psi\rangle_{[0,T]} + (e(T),\psi_T) &= \sum_{n=1}^{N}\left(\langle \dot{e} - \bar{A}e, \varphi\rangle_{I_n}\right)\\ &= \sum_{n=1}^{N}\Big(\langle \mathcal{R}_P(Y), \varphi\rangle_{I_n}\\ &\qquad + \langle f(Y,t) - \sum_{i=1}^{L}\alpha_i f(P_n^i Y, t), \varphi\rangle_{I_n}\Big).\end{aligned}$$

Next we use Galerkin orthogonality defined in (3.3) to obtain the error representation formula. The argument for the dG method is similar. □

3.1.4 Illustrative example

For the linear system (2.7), we plot the discretization, quadrature, and a new "explicit" contributions for the explicit trapezoid method in Fig. 1d. The explicit contribution is larger than the discretization error while the quadrature contribution is significantly smaller overall. However, the quadrature contribution is "out of phase" with the other two contributions.

3.2 Polynomial extrapolation

We use an operator $Q_n^\ell : BV([0,T];\mathbb{R}^d) \to \mathcal{P}^{\ell-1}(I_n)$ that produces a polynomial $Q_n^\ell f$ interpolating the function f at the nodes $t_{n-\ell},\ldots,t_{n-1}$. We define the operator Q^ℓ so that $Q^\ell|_{I_n} = Q_n^\ell$ on $[t_{\ell-1},T]$. We define the approximation differently on the "initial" interval $[0,t_{\ell-1}]$, but make sure to preserve the order of the general method. There are various ways to do this. To simplify the analysis, we provide the analysis for the implicit cG($\ell-1$) method, which has the same order as the corresponding explicit method, i.e. letting $\ell = q+1$.

The cG(q) method incorporating polynomial extrapolation is: Find $Y \in \mathcal{C}^q([0,T])$ such that $Y(0) = y_0$, for $n = 1, \ldots, \ell - 1$,

$$\begin{cases} \langle \dot{Y}, v_k \rangle_{I_n} - \langle f(Y,t), v_k \rangle_{I_n} = 0 \quad \forall\, v_k \in \mathcal{P}^{q-1}(I_n), \\ Y(t_{n-1}^+) = Y(t_{n-1}^-), \end{cases}$$

and for $n = \ell, \ldots, N$,

$$\begin{cases} \langle \dot{Y}, v_k \rangle_{I_n} - \langle Q_n^\ell f(Y,t), v_k \rangle_{I_n} = 0 \quad \forall\, v_k \in \mathcal{P}^{q-1}(I_n), \\ Y(t_{n-1}^+) = Y(t_{n-1}^-). \end{cases} \tag{3.6}$$

We can view (3.6) as applying the cG(q) method to the modified problem (3.2).

3.2.1 Examples

The finite element approximations are nodally equivalent to the Adams-Bashforth multi-step finite difference methods. For example, applying cG(1) to (3.2) with $\ell = 2$ yields,

$$Y(t_n) = Y(t_{n-1}) + \frac{3}{2} k_n f(Y(t_{n-1}), t_{n-1}) - \frac{1}{2} k_n f(Y(t_{n-2}), t_{n-2}),$$

which is the update formula for the second order Adams-Bashforth method. For a general projection Q_n^ℓ the finite element approximation of (3.2) is nodally equivalent to the ℓth order Adams–Bashforth method.

3.2.2 A posteriori error analysis

To obtain the following error estimate, we again employ the classic adjoint (2.4) used for implicit discretizations.

Theorem 4 (Extrapolation Error Representation Formula) *If $Y(t)$ is an approximation of (3.2) obtained via the cG(q) method, then the error in the quantity of interest defined by ψ and ψ_T is given by,*

$$\begin{aligned} \langle e, \psi \rangle_{[0,T]} + (e(T), \psi_T) = &\sum_{n=1}^{\ell-1} \underbrace{\langle \mathcal{R}(Y), \varphi - \pi_k \varphi \rangle_{I_n}}_{\textit{Initial Contr.}} \\ &+ \sum_{n=\ell}^{N} \Bigg(\underbrace{\langle \mathcal{R}_Q^\ell(Y), \varphi - \pi_k \varphi \rangle_{I_n}}_{\textit{Discretization Contr.}} + \underbrace{\langle f(Y,t) - Q_n^\ell f(Y,t), \varphi \rangle_{I_n}}_{\textit{Explicit Contr.}} \Bigg) \end{aligned}$$

where $\mathcal{R}$, φ and π_k are defined as above, and $\mathcal{R}_Q^\ell(Z) = Q^\ell f(Z,t) - \dot{Z}$ is the modified residual.

For the dG(q) approximation, the discretization contribution is replaced by

$$\langle \mathcal{R}_Q^\ell(Y), \varphi - \pi_k\varphi\rangle_{I_n} - ([Y]_{n-1}, \varphi(t_{n-1}) - \pi_k\varphi(t_{n-1}^+)).$$

Note that the initial contribution is the same as provided by Theorem 1. The quadrature error is zero because $\langle Q_n^\ell f(Y,t), v\rangle_{I_n}$ can be integrated exactly.

Proof The proof follows the argument used for Theorem 3 except that we apply the appropriate form of Galerkin orthogonality to the initial interval and the remaining intervals independently.

3.2.3 Illustrative example

For the linear system (2.7), we show two error contributions, i.e., the discretization and "explicit" contributions in Fig. 1e. The discretization contribution is almost identical to the discretization contributions for previous methods. The explicit error contribution clearly dominates.

4 Numerical experiments

We explore various aspects of the a posteriori error estimates using several examples chosen to stress particular characteristics. We are particularly focussed on the relative sizes of error contributions and the overall accuracy of the estimates. We note that the level of accuracy required for an a posteriori error estimate depends strongly on the needs of the application, and can range from being roughly of the correct order of magnitude to being correct to several digits. As a way of portraying estimate accuracy, we use the effectivity ratio,

$$\mathcal{E} = \left| \frac{\text{Estimated Error}}{\text{Exact Error}} \right|.$$

In these examples, the effectivity ratio is close to unity, and we use the discrepancy defined as the distance of $\mathcal{E}$ from one,

$$\eta = |\mathcal{E} - 1|.$$

In some cases, we construct the problem to have a known solution so the error is computable. In other cases, we solve the problem using a higher order method with very fine time steps to get a much more accurate solution which is used to approximate the true error.

In the following examples, we consider the explicit trapezoid (RK2), RK4/2, and second order Adams-Bashforth (AB2) methods. We solve the adjoint problems using the third order cG(2) method. We also compare to the numerical solution obtained by applying the fully implicit cG(1) method to the original equation.

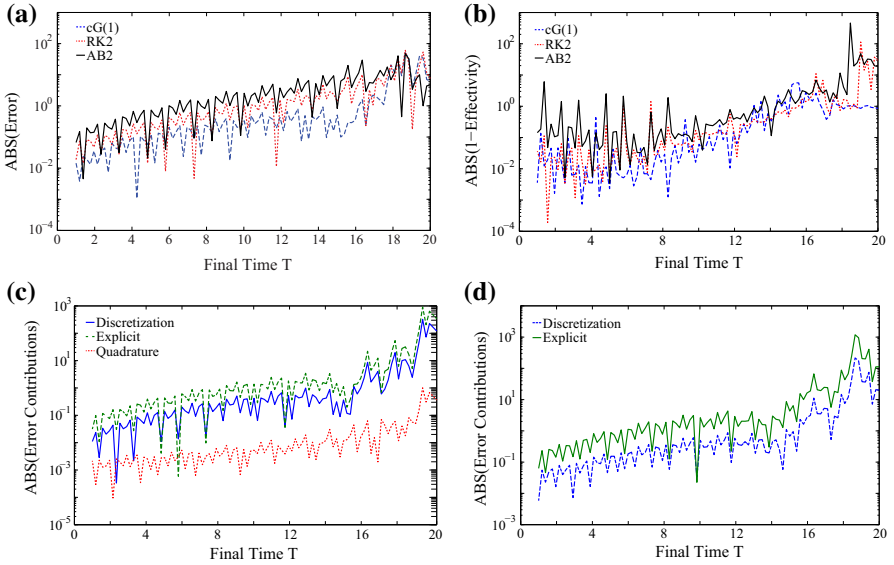

Fig. 2 Results for the Lorenz problem (4.1). **a** Error for the cG(1), RK2 and AB2 methods. **b** Discrepancy in the effectivity ratio for for the cG(1), RK2 and AB2 methods. **c** Error contributions for RK2. **d** Error contributions for AB2

4.1 The Lorenz equation

We solve the Lorenz system,

$$\begin{cases} \dot{y}_1 = -10.1y_1 + 10.1y_2 \\ \dot{y}_2 = 28.0y_1 - y_2 - y_1y_3 \\ \dot{y}_3 = -8/3y_3 + y_1y_2 \\ y(0) = [-9.408, -9.096, 28.581]^T. \end{cases} \tag{4.1}$$

This is a well-known chaotic system [11,18]. In general, a posteriori error estimates for implicit methods are accurate up to a critical time, when the numerical error becomes quite large, and the estimates are inaccurate after that.

In the following examples, we use a uniform time step of $k_n = .01$ and set $\psi = (0, 0, 0)^\top$, $\psi_T = (1, 1, 1)^\top$ for a sequence of final times T between 1 and 20. Fig. 2a shows the value of $\mathcal{E}$ and Fig. 2b shows the value of η for the cG(1), RK2, and AB2 methods. The error in the explicit methods grows more rapidly. On the other hand, the difference in the effectivity ratios is not significant, and estimates for the explicit methods are reasonably accurate.

In Fig. 2c, d, we plot the absolute error contributions for the RK2 and AB2 methods respectively. In both cases, the explicit contribution dominates, and it dominates more in the extrapolation method. This is common to most examples we tested. We also see that the quadrature error is negligible. This is because the Lorenz system is almost linear, containing only two bilinear terms.

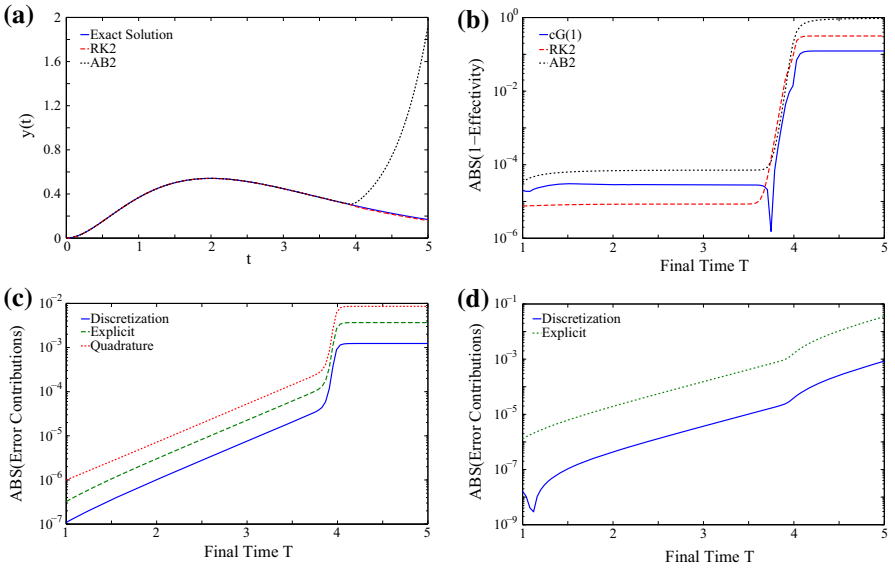

Fig. 3 Results for the nonlinear problem (4.2). **a** Exact and approximate RK2 and AB2 solutions. **b** Discrepancy in the effectivity ratios for the cG(1), RK2 and AB2 methods. **c** Error contributions for RK2. **d** Error contributions for AB2

4.2 A highly nonlinear example

The next example is

$$\begin{cases} \dot{y} = y(1 + \tanh(\alpha(y - .3))) - te^{-t}\left(-2 + 2\,t + t\tanh\left(\alpha\left(t^2e^{-t} - .3\right)\right)\right), \\ y(0) = 0, \end{cases} \tag{4.2}$$

where the forcing term is constructed to give the exact solution $y(t) = t^2e^{-t}$. The hyperbolic tangent, with $\alpha = 100$, in the nonlinearity leads to a sudden change in value of the right hand side as the solution changes, providing a challenge for extrapolation methods. We again compare the cG(1), RK2, and AB2 methods. We set $\psi = 0$, $\psi_T = 1$ for a sequence of final times T. The modified equations in the explicit methods are solved with a uniform time step of $k_n = .001$.

Figure 3a compares the exact solution and approximate solutions obtained via the RK2 and AB2 methods and Fig. 3b shows the value of η for cG(1), RK2 and AB2 methods. Near $t = 4$, there is a large jump η for all methods. As can be seen from Fig. 3c, d in which we plot the error contributions for the RK2 and AB2 methods, respectively, the jump in the discrepancy η for RK2 and AB2 occurs simultaneously with a sudden loss of accuracy. For the RK2 method, the quadrature error dominates all contributions while the explicit contribution dominates the discretization contribution. For AB2, the explicit contribution is dominant.

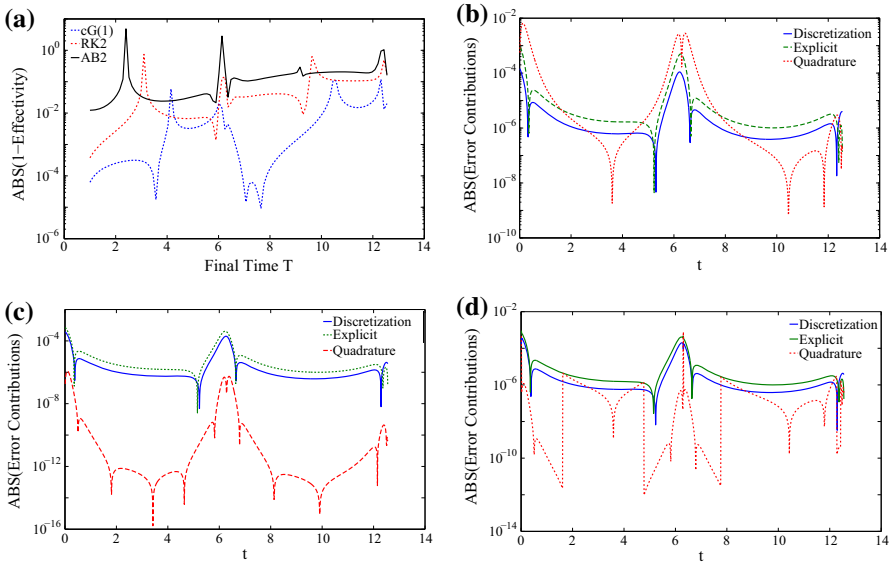

Fig. 4 Results for the two body problem. **a** Discrepancy in the effectivity ratio for cG(1), RK2, and AB2. **b** Absolute error contributions from each time interval for RK2. **c** Absolute error contributions from each time interval for RK2/4. **d** Absolute error contributions from each time interval for the adaptive trapezoid method

4.3 The two body problem

We next consider the well known two body problem,

$$\begin{cases} \dot{y}_1 = y_3, \quad \dot{y}_3 = \dfrac{-y_1}{(y_1^2 + y_2^2)^{3/2}}, \\ \dot{y}_2 = y_4, \quad \dot{y}_4 = \dfrac{-y_2}{(y_1^2 + y_2^2)^{3/2}}, \\ y(0) = [0.4, 0, 0, 2.0]^\top. \end{cases}$$

The two body problem is a Hamiltonian system with a complicated dynamic structure. For the specified choice of initial conditions, there is an exact analytic periodic solution determined by the equation,

$$y = \left[\cos(\tau) - .6, .8\sin(\tau), \frac{-\sin(\tau)}{1 - .6\cos(\tau)}, \frac{.8\cos(\tau)}{1 - .6\cos(\tau)}\right]^\top,$$

where τ solves $\tau - .6\sin(\tau) = t$.

Figure 4a shows the value of η for the cG(1), RK2, and AB2 methods. We see that the error estimate becomes inaccurate around specific times during the first part of the solution and the inaccuracy gradually increases as time passes.

Next, we examine the absolute error contributions from each time interval during one computation. We use $\psi = (1, 1, 1, 1)^\top$ and $\psi_T = (0, 0, 0, 0)^\top$. We solve up to time $T = 12.55$ using a uniform time step of $k_n = .01$. In Fig. 4b, we plot the error

Table 1 Errors in various quantities of interest for the explicit trapezoid method and it's variations

	$\psi = (1,1,1,1)^\top$ $\psi_T = (0,0,0,0)^\top$	$\psi = (0,0,0,0)^\top$ $\psi_T = (1,1,1,1)^\top$	$\psi = (0,1,0,0)^\top$ $\psi_T = (0,0,0,0)^\top$
RK2	1.34e−1	−1.06e−1	−3.16e−2
RK2/4	−1.19e−2	−1.49e−2	−2.95e−2
Adaptive Quad.	−1.08e−2 (37.5 %)	−1.58e−2 (45.3 %)	−3.60e−2 (28.9 %)
RK2 with $\frac{k_n}{2}$	3.11e−2	−3.04e−2	−7.68e−3

contributions for RK2. We note that the quadrature contribution dominates during periods of the solution, so we also present results for the RK2/4 method in Fig. 4c. As expected, the discretization and explicit contributions are significantly larger than the quadrature contribution in RK2/4 because we are computing integrals in the variational formulation more accurately.

Since the quadrature error only dominates over part of the domain, while the RK2/4 method costs more per time step than RK2, this suggest use of an adaptive quadrature approach in which the higher order quadrature is only used when the quadrature error is the largest component of the total error. In Fig. 4d we show results for such an approach and see the quadrature error has been reduced so that it no longer dominates.

In Table 1, we give errors for various quantities of interest for RK2, RK2/4, an adaptive quadrature method (where we indicate the percentage of the time domain in which the RK2/4 method is implemented), and RK2 with the time step cut in half. The We see that for certain quantities of interest, an adaptive quadrature scheme can give improved accuracy for less cost than halving the time step.

4.4 The Bistable problem

We next consider solution of a large dimension system obtained by a method of lines discretization in space of the well known bistable, or Allen–Cahn, parabolic problem,

$$\begin{cases} u_t = u - u^3 + \epsilon\, u_{xx}, & 0 < x < 1,\ 0 < t, \\ \frac{\partial u}{\partial x}(0,t) = \frac{\partial u}{\partial x}(1,t) = 0, & 0 < t, \\ u(x,0) = u_0(x), & 0 < x < 1. \end{cases}$$

For small ϵ, the solution of this problem exhibits "metastability", that is long periods of quasi-steady state behavior punctuated by rapid transients [12]. We consider initial data that gives two metastable periods over [0, 150],

$$u_0(x) = \begin{cases} \tanh((.2-x)/(2\sqrt{\epsilon})), & 0 \le x < .28, \\ \tanh((x-.36)/(2\sqrt{\epsilon})) & .28 \le x < .4865, \\ \tanh((.613-x)/(2\sqrt{\epsilon})) & .4865 \le x < .7065, \\ \tanh((x-.8)/(2\sqrt{\epsilon})) & .7065 \le x < 1, \end{cases}$$

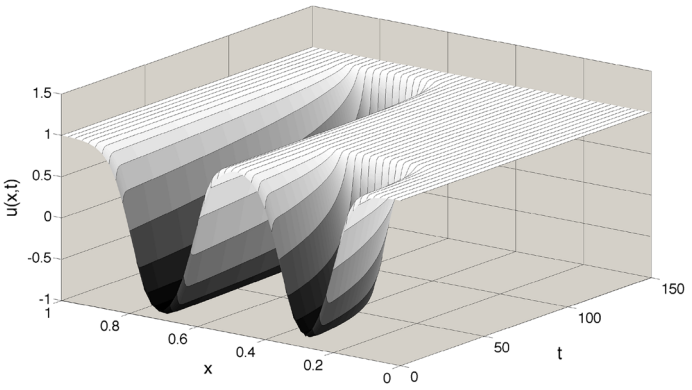

Fig. 5 Exact solution of the Bistable problem with given initial data. One well collapses at $t \approx 41$ while the other collapses at $t \approx 141$

We show the numerical solution with $\epsilon = .0009$ in Fig. 5. The solution begins with two "wells" and at $t \approx 41$ and $t \approx 141$, the wells sharply collapse.

We discretize the spatial variable with a standard cG(1) finite element method using a uniform mesh size of $h = 0.02$ to obtain,

$$\begin{cases} \dot{u} = u - u^3 + \dfrac{\epsilon}{h^2} Au, \quad 0 < t, \\ u(0) = u_0(x), \end{cases}, \qquad A = \begin{bmatrix} -2 & 2 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 2 & -2 \end{bmatrix},$$

where x is the discretized spatial variable. We present results for the RK2 and AB2 methods as well as the fourth order Runge–Kutta method (RK4) and fourth order Adams–Bashforth method (AB4), all using the uniform time step of $k_n = .01$. We solve the adjoint problems with the cG(4) method. For a quantity of interest, we consider the average of the error over both the temporal and spatial domain. We solve up to time $T = 150$ in order to include the collapses of both wells.

We plot the absolute error contributions from each time interval for one solution for all four methods in Fig. 6. In Fig. 6a and in Fig. 6b we see two sharp jumps in the error contributions arising from rapid transitions of the solution. It is also interesting to observe the difference between the explicit contribution and the discretization contribution. For RK2 they are fairly close, while for AB2 the contributions are about twice as far apart. This suggests that the extrapolation approximation is less accurate than the Taylor series approximation. For the higher order methods, RK4 in Fig. 6c and AB4 in Fig. 6d respectively, we see that the explicit contribution is significantly larger than the discretization contribution (The explicit error and quadrature error lie on top of one another in Fig. 6c). This suggest that explicit contribution is more significant for higher order methods.

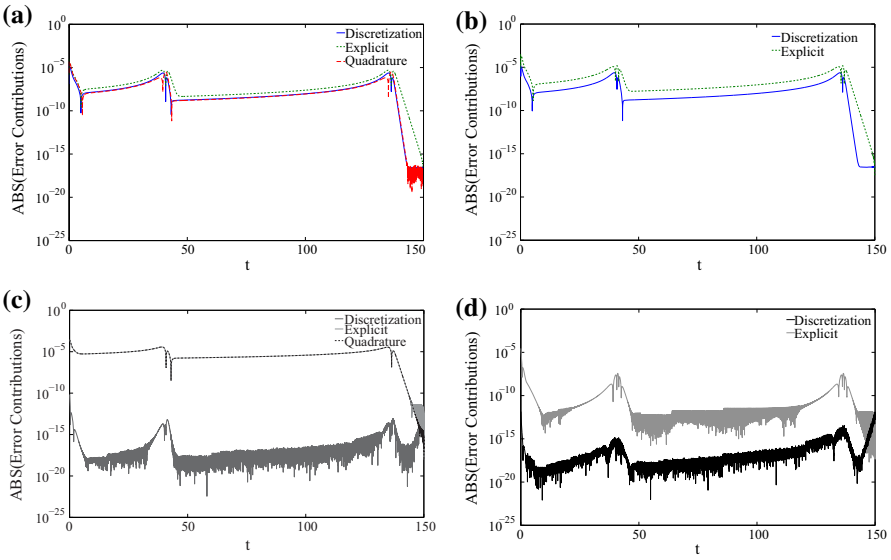

Fig. 6 Results for the discretized Bistable problem. **a** Absolute error contributions from each time interval for RK2. **b** Absolute error contributions from each time interval for AB2. **c** Absolute error contributions from each time interval for RK4. **d** Absolute error contributions from each time interval for AB4

5 Error analysis involving the adjoint of the discretized problem

In the analysis so far, we have used the standard definition of an adjoint operator for a nonlinear operator that is based on linearization of the continuous (nonlinear) operator [7,16]. However, in situations in which the numerical solution has significantly different stability properties from the true solution, it may be necessary to account for the fact that the continuous and discretized solution operators have different adjoint operators [7]. We briefly describe how this can be carried out for an explicit discretization methods using a Taylor series approximation.

The idea is to separately linearize the continuous and discrete problems around a given solution that is common to both problems, e.g., a steady-state solution. We define adjoint problems to these two linearized problems and then derive an estimate.

We linearize $\mathcal{N}(y, v)$ [Eq. (2.1)] and $\mathcal{N}_P(Y, v)$ [Eq. (3.5)] about a common solution $w(t)$ to obtain,

$$0 = \langle A_w w - f(w,t), v\rangle_{[0,T]} + \langle \dot{y} - A_w y, v\rangle_{[0,T]},$$

and

$$-\sum_{n=1}^{N}\langle \mathcal{R}_P(Y), v\rangle = \sum_{n=1}^{N}\Big(\langle A_{P_n,w}B_{n,w}w - f(P_n w,t), v\rangle_{I_n} + \langle \dot{Y} - A_{P_n,w}B_{n,w}Y, v\rangle_{I_n}\Big),$$

where,

$$A_w := \int_0^1 f'(sy + (1-s)w, t)\, ds,$$
$$A_{P_n,w} := \sum_{i=1}^{L} \left(\alpha_i \int_0^1 f'(s P_n^i Y + (1-s) P_n w, t)\, ds \right),$$
$$B_{n,w} := \int_0^1 P_n'(sY + (1-s)w)\, ds,$$

and P_n' denotes the Fréchet derivative of the approximation series operator. This yields the linear differential equations,

$$\langle \dot{y} - A_w y, v \rangle_{[0,T]} = \langle f(w,t) - A_w w, v \rangle_{[0,T]}, \tag{5.1}$$
$$\sum_{n=1}^{N} \langle \dot{Y} - A_{P_n,w} B_{n,w} Y, v \rangle_{I_n} = \sum_{n=1}^{N} \langle f(P_n w, t) - A_{P_n,w} B_{n,w} w - \mathcal{R}_P(Y), v \rangle_{I_n}. \tag{5.2}$$

We now define adjoint problems for these linear problems. For (5.1): With $\varphi_w(T) = \psi_T$, for $n = N, N-1, \ldots, 1$, find $\varphi_w \in C^1(I_n; \mathbb{R}^d)$ such that

$$\begin{cases} \langle v, -\dot{\varphi}_w - A_w^* \varphi_w \rangle_{I_n} = \langle v, \psi \rangle_{I_n}, \ \forall\, v \in C^0(I_n; \mathbb{R}^d), \\ \varphi_w(t_n^-) = \varphi_w(t_n^+). \end{cases} \tag{5.3}$$

For (5.2): With $\varphi_{P_w} = \psi_T$, for $n = N, N-1, \ldots, 1$, find $\varphi_{P_w} \in C^1(I_n; \mathbb{R}^d)$ such that

$$\begin{cases} \langle v, -\dot{\varphi}_{P_w} \rangle_{I_n} - \langle B_{n,w} v, A_{P_n,w}^* \varphi_{P_w} \rangle_{I_n} = \langle v, \psi \rangle_{I_n}, \ \forall\, v \in C^0(I_n; \mathbb{R}^d), \\ \varphi_{P_w}(t_n^-) = \varphi_{P_w}(t_n^+). \end{cases} \tag{5.4}$$

Theorem 5 (Two Adjoint Error Representation Formula) *If $Y(t)$ is an approximation of the modified Eq. (3.1) obtained via the cG(q) method with quadrature defined by the inner product $\langle \cdot, \cdot \rangle_{I_n, K_n}$, and $w \in L^2([0,T]; \mathbb{R}^d)$, then the error in the quantity of interest defined by ψ and ψ_T is given by,*

$$\begin{aligned} &\langle e, \psi \rangle_{[0,T]} + (e(T), \psi_T) \\ &= \sum_{n=1}^{N} \Big(\underbrace{\langle \mathcal{R}_P(Y), \varphi_{P_w} - \pi_k \varphi_{P_w} \rangle_{I_n}}_{\textit{Discretization Contribution}} \Big) \\ &+ \sum_{n=1}^{N} \Big(\underbrace{\langle f(P_n Y, t), \pi_k \varphi_{P_w} \rangle_{I_n} - \langle f(P_n Y, t), \pi_k \varphi_{P_w} \rangle_{I_n, K_n}}_{\textit{Quadrature Contribution}} \Big) \end{aligned}$$

$$+\underbrace{\big(y_0, \varphi_w(0) - \varphi_{P_w}(0)\big)}_{\textit{Explicit Contribution}}$$
$$+\sum_{n=1}^{N}\Big(\underbrace{\langle f(w,t) - A_w w, \varphi_w\rangle_{I_n} - \langle f(P_n w, t) - A_{P_n,w}B_{n,w}w, \varphi_{P_w}\rangle_{I_n}}_{\textit{Difference in Linearization Contribution}}\Big) \tag{5.5}$$

where φ_w and φ_{P_w} are solutions of (5.3) and (5.4) respectively, and π_k and $\mathcal{R}_P(Y)$ are as above.

For the dG(q) approximation, the discretization contribution is replaced by,

$$\langle \mathcal{R}_P(Y), \varphi_{P_w} - \pi_k\varphi_{P_w}\rangle_{I_n} - \big([Y]_{n-1}, \varphi_{P_w}(t_{n-1}) - \pi_k\varphi_{P_w}(t_{n-1}^+)\big).$$

Proof We begin by splitting the error,

$$\langle e, \psi\rangle_{[0,T]} = \langle y, \psi\rangle_{[0,T]} - \langle Y, \psi\rangle_{[0,T]}. \tag{5.6}$$

We deal with each of these terms separately. For the first term, we use y as the test function in (5.3) and integrate by parts and use (5.1) to obtain,

$$\begin{aligned}\langle y, \psi\rangle_{[0,T]} &= \langle \dot{y} - A_w y, \varphi_w\rangle_{[0,T]} + (y_0, \varphi_w(0)) - (y(T), \psi_T)\\ &= \langle f(w,t) - A_w w, \varphi_w\rangle_{[0,T]} + (y_0, \varphi_w(0)) - (y(T), \psi_T).\end{aligned}$$

For the second term, we use Y as the test function in (5.4) and integrate by parts and use (5.2) to obtain,

$$\begin{aligned}\langle Y, \psi\rangle_{[0,T]} &= \sum_{n=1}^{N}\Big(\langle \dot{Y} - A_{P_n,w}B_{n,w}Y, \varphi_{P_w}\rangle_{I_n} + \big([Y]_{n-1}, \varphi_{P_w}(t_{n-1})\big)\Big)\\ &\quad + (y_0, \varphi_{P_w}(0)) - (Y(T), \psi_T)\\ &= \sum_{n=1}^{N}\Big(\langle f(P_n w, t) - A_{P_n,w}B_{n,w}w - \mathcal{R}_P(Y), \varphi_{P_w}\rangle_{I_n}\\ &\quad + \big([Y]_{n-1}, \varphi_{P_w}(t_{n-1})\big)\Big) + (y_0, \varphi_{P_w}(0)) - (Y(T), \psi_T).\end{aligned}$$

Then Eq. (5.6) becomes,

$$\begin{aligned}\langle e, \psi\rangle_{[0,T]} = \sum_{n=1}^{N}\{&\langle \mathcal{R}_P(Y), \varphi_{P_w}\rangle_{I_n}\\ &+ \langle f(w,t) - A_w w, \varphi_w\rangle_{I_n} - \langle f(P_n w, t) - A_{P_n,w}B_{n,w}w, \varphi_{P_w}\rangle_{I_n}\}\\ &+ (y_0, \varphi_w(0) - \varphi_{P_w}(0)) - (e(T), \psi_T).\end{aligned}$$

We can then use Galerkin orthogonality to complete the proof. The proof for the dG methods is analogous.

An error estimate can also be derived for multistep methods that incorporate extrapolation and use the modified Eq. (3.2). These a posteriori results are complicated to use because of the need to consider the solution of two adjoint problems and because the term labelled "Difference in Linearization Contribution" involves a common solution. In practice, these estimates are often manipulated further to obtain expressions more amenable to computation plus additional terms that cannot be estimated but are provably higher order, see [8,9].

5.1 A numerical example

We consider the solution of the system arising from a method of lines discretization of the linear heat equation,

$$\begin{cases} u_t = u_{xx}, & 0 < x < 3\pi,\ t > 0, \\ u(0,t) = u(1,t) = 0, & t > 0, \\ u(x,0) = u_0(x), & 0 < x < 3\pi, \end{cases}$$

where the initial data is given by $u_0(x) = \sin(x)$. Using a standard cG(1) method in space on a uniform mesh with spacing $h \approx .304$ gives,

$$\begin{cases} \dot{u} = \frac{1}{h^2} A u, & t > 0, \\ u(0) = u_0(x), \end{cases} \quad , \quad A = \begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix}. \tag{5.7}$$

We solve (5.7) using the forward Euler method. The dG(0) finite element method is used to solve the corresponding modified equation. From the point of view of stability for numerical solution of the full partial differential equation, the forward Euler scheme with a uniform time step k_n must satisfy,

$$\alpha := \frac{k_n}{h^2} \leq \frac{1}{2}.$$

Since instability does not occur in the true solution, the solution of the adjoint problem (2.4) does not indicate instability. However, the adjoint to the discretization (5.4) should reflect instability in the numerical solution. We plot relative error between the modified adjoint φ_{P_w} and the continuous adjoint φ_w, i.e.

$$RE(t) := \frac{\|\varphi_w(t) - \varphi_{P_w}(t)\|_\infty}{\|\varphi_w(t)\|_\infty} \tag{5.8}$$

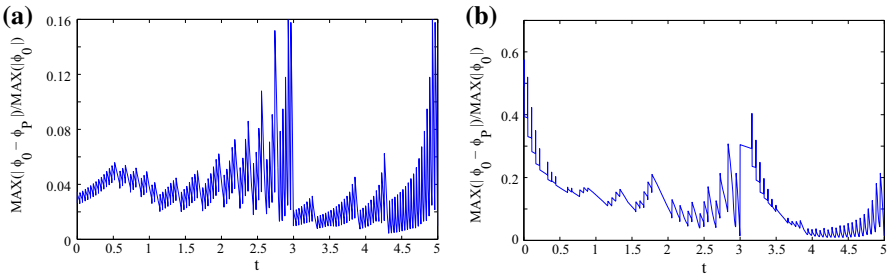

Fig. 7 Results for the discretized heat equation. Relative difference in the adjoint solutions when the quantity of interest is the average value over the first half of the domain. **a** $\alpha = 0.4$. **b** $\alpha = 0.6$

where the norms are in $\mathbb{R}^N$, for two different values of α. We consider the quantity of interest as the average value over the entire spatio-temporal domain with final time $T = 5$.

We show the results in Fig. 7. When $\alpha = 0.4$ in Fig. 7a, the stability condition is satisfied and there is relatively little difference between the adjoint solutions for either quantity of interest. However, the CFL stability condition is violated when $\alpha = 0.6$ in Fig. 7b and we can see the large difference in the two adjoint solutions.

6 Conclusions

We present an a posteriori error analysis for approximate solutions of nonlinear ordinary differential equations solved with explicit finite difference methods. To obtain this analysis, we represent two classes of explicit finite difference methods as finite element methods, whose solutions, while defined over the entire domain, are equal to the finite difference solutions at nodes. In particular, we distinguish between error contributions from restricting the solution to lie in a particular piecewise polynomial approximation space, the use of quadrature to evaluate the integrals in the finite element formulation and the approximation of the differential operator which we modify via the use of projection operators. Splitting the contributions to the error in this manner has the potential to enable us to determine the best method of adaptation to reduce the error in our numerical solution. Finally, we give an adjoint problem for the explicit method, and show how it can be used to determine numerical stability when compared with the adjoint of the ODE.

References

1. Ainsworth, M., Oden, J.T.: A posteriori error estimation in finite element analysis. Comput. Methods Appl. Mech. Eng. **142**, 1–88 (1997)
2. Becker, R., Rannacher, R.: An optimal control approach to a posteriori error estimation in finite element methods. Acta Numer. **10**, 1–102 (2001)
3. Delfour, M., Trochu, F., Hager, W.: Discontinuous Galerkin methods for ordinary differential equations. Math. Comput. **36**, 455–473 (1981)
4. Eriksson, K., Estep, D., Hansbo, P., Johnson, C.: Introduction to adaptive methods for differential equations. Acta Numer. **4**, 105–158 (1995)

5. Eriksson, K., Johnson, C., Logg, A.: Explicit time-stepping for stiff ODE's. SIAM J. Sci. Comput. **25**, 1142–1157 (2003)
6. Estep, D.: A posteriori error bounds and global error control for approximation of ordinary differential equations. SIAM J. Numer. Anal. **32**, 1–48 (1995)
7. Estep, D.: Error estimates for multiscale operator decomposition for multiphysics models. In: Fish, J. (ed.) Multiscale Methods: Bridging the Scales in Science and Engineering, chap. 11. Oxford University Press, New York (2009)
8. Estep, D., Ginting, V., Ropp, D., Shadid, J.N., Tavener, S.: An a posteriori-a priori analysis of multiscale operator splitting. SIAM J Numer Anal **46**, 1116–1146 (2008)
9. Estep, D., Ginting, V., Tavener, S.: A posteriori analysis of a multirate numerical method for ordinary differential equations. Comput. Methods Appl. Mech. Eng. **223**, 10–27 (2012)
10. Estep, D., Holst, M., Mikulencak, D.: Accounting for stability: a posteriori error estimates based on residuals and variational analysis. Commun. Numer. Methods Eng. **18**, 15–30 (2002)
11. Estep, D., Johnson, C.: The pointwise computability of the Lorenz system. Math. Models Methods Appl. Sci. **8**, 1277–1306 (1998)
12. Estep, D.J., Larson, M.G., Williams, R.D.: Estimating the Error of Numerical Solutions of Systems of Reaction-Diffusion Equations. American Mathematical Society, Providence (2000)
13. Estep, D.J., Stuart, A.M.: The dynamical behavior of the discontinuous Galerkin method and related difference schemes. Math. Comput. **71**, 1075–1104 (2002)
14. Giles, M., Süli, E.: Adjoint methods for PDEs: a posteriori error analysis and postprocessing by duality. Acta Numer. **11**, 145–236 (2002)
15. Houston, P., Rannacher, R., Süli, E.: A posteriori error analysis for stabilised finite element approximations of transport problems. Comput. Methods Appl. Mech. Eng. **190**, 1483–1508 (2000)
16. Marchuk, G.I., Agoshkov, V.I., Shutiaev, V.P.: Adjoint Equations and Perturbation Algorithms in Nonlinear Problems. CRC Press, Boca Raton (1996)
17. Petzold, L.R., Ascher, U.M.: Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations. SIAM, Philadelphia (1998)
18. Sandelin, J.D.: Global estimate and control of model, numerical, and parameter error. Ph.D. thesis, Colorado State University (2007)
19. Zhao, S., Wei, G.W.: A unified discontinuous Galerkin framework for time integration. Math. Methods Appl. Sci. **37**, 1042–1071 (2014)