BIT Numerical Mathematics
https://doi.org/10.1007/s10543-020-00825-0

Error estimation and uncertainty quantification for first time to a threshold value

Jehanzeb H. Chaudhry[1] · **Donald Estep**[3] · **Zachary Stevens**[1] · **Simon J. Tavener**[2]

Received: 11 February 2020 / Accepted: 22 July 2020

Abstract

Classical a posteriori error analysis for differential equations quantifies the error in a Quantity of Interest which is represented as a bounded linear functional of the solution. In this work we consider a posteriori error estimates of a quantity of interest that cannot be represented in this fashion, namely the time at which a threshold is crossed for the first time. We derive two representations for such errors and use an adjoint-based a posteriori approach to estimate unknown terms that appear in our representation. The first representation is based on linearizations using Taylor's Theorem. The second representation is obtained by implementing standard root-finding techniques. We provide several examples which demonstrate the accuracy of the methods. We then embed these error estimates within a framework to provide error bounds on a cumulative distribution function when the parameters of the differential equations are uncertain.

Keywords Initial value problems · Error bounds · Monte Carlo methods · Adjoint based error estimation · Uncertainty quantification

Mathematics Subject Classification 65L05 · 65L70 · 65C05

Communicated by Axel Målqvist.

✉ Jehanzeb H. Chaudhry
jehanzeb@unm.edu

Donald Estep
donald_estep@sfu.ca

Zachary Stevens
zstevens@unm.edu

Simon J. Tavener
tavener@math.colostate.edu

1 The University of New Mexico, Albuquerque, NM 87131, USA

2 Colorado State University, Fort Collins, CO 80523, USA

3 Simon Fraser University, Burnaby, BC V5A 1S6, Canada

1 Introduction

There are many situations in which the purpose of a computation is to determine *when* a functional of the solution to (1) achieves a certain event, for example when a temperature or a species concentration reaches a specified level, the wave height of a tsunami crosses a threshold at a certain location, an orbiting body completes a revolution etc. In this article we perform a posteriori analysis for the error in the computed value and computed probability distribution of the time at which a threshold value is realized for the first time in the context of ordinary differential equations (ODEs). More precisely, consider a system of first order ODEs

$$\dot{y} = f(y, t; \theta),\ t \in (0, T], \quad y(0) = y_0, \tag{1}$$

where $\dot{y} \equiv \frac{dy(t)}{dt}$, $f : \mathbb{R}^m \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}^m$ is a Lipschitz continuous function and θ is a deterministic or random parameter. Let $S(y(t))$ be a linear functional of $y(t)$ and $Q(y)$ be the first time $t \in (0, T]$ at which a threshold $S(y(t)) = R$ is crossed. We assume such a $t < T$ exists. That is,

$$Q(y) := \min_{t \in (0,T]} \arg(S(y(t)) = R). \tag{2}$$

Hence, we refer to this as a non-standard QoI in the context of a posteriori error analysis. An example of this non-standard QoI for the Lorenz system (see Sect. 2.1) is illustrated in Fig. 1a.

Standard adjoint-based a posteriori error analysis seeks to estimate the error in a quantity of interest (QoI) that can be expressed as a bounded functional of the solution and is widely used for a broad range of numerical methods [1,4,6,7,9–13,15–19,22,23,28,30,32,34,35,40]. In these cases, the error analysis involves computable residuals of the numerical solution, the generalized Green's function solving an adjoint problem and variational analysis [1,7,31,34,35]. This work is briefly summarized for initial value problems in Sect. 3.2. It is usually employed within a finite element (variational) solution strategy, but can also be applied to finite difference and finite volume methods by recasting them as equivalent finite element methods [16,20,22–25,29,33,41]. Nonlinear QoIs are treated by first linearizing around a computed solution e.g. see [7,10,21].

The goal of the current work is to derive accurate error estimates for the non-standard QoI given by equation (2) that *cannot* be expressed as a bounded linear functional of the solution y. In addition, we use the result to bound the error in an empirical distribution function for the nonstandard QoI corresponding to a stochastic parameter θ. This is similar to the a posteriori analysis for the error in CDF for standard QoIs for PDEs with random coefficients and random geometries appearing in [14,36,37]. The situation is more complex for a stochastic differential equation when seeking to compute the expected value of a functional of the solution at an "exit time" τ, when the solution first leaves a specified region, since the continuous solution trajectory may leave and re-enter the specified region undetected within a single time-step of the numerical integration scheme. Discussion of this problem appears in, e.g., [8,26,39].

We first perform a a priori error analysis for the non-standard QoI given by equation (2) assuming the initial value problem (1) is solved using a numerical method with $O(h^p)$ convergence rate and show that the error in the non-standard QoI converges at the same asymptotic rate. The a posteriori analysis for the error in the non-standard QoI appears in Sect. 3. The first approach in Sect. 3.3 takes advantage of linearization via Taylor series and employs auxiliary quantities of interest to obtain a formula that directly estimates the error in the QoI. Our second approach, in Sect. 3.4 proceeds by using different root finding methods and again employs auxiliary quantities of interest. Numerical results supporting the accuracy of the error estimates for a deterministic system appear in Sect. 4. Details of the error estimate for the CDF when θ is a random variable are provided in Sect. 3.6 and the bounds are computed for several examples in Sect. 5.

2 A priori analysis

In this section, we present a general a priori result regarding the convergence of the approximate time that a threshold condition is met as the discretization is refined. Assume a continuous numerical solution, $Y(t)$, of order p is computed to approximate the solution to the initial value problem (1). That is,

$$\|y(t) - Y(t)\|_{\mathbb{R}^m} \leq Ch^p, \tag{3}$$

for all $t \in [0, T]$, for some constant $C > 0$. Here $\|\cdot\|_{\mathbb{R}^m}$ denotes the standard Euclidean norm in $\mathbb{R}^m$ and h denotes the step-size used to compute the numerical solution $Y(t)$. For a given value of the threshold R, define t_t and t_c such that

$$S(y(t_t)) = R = S(Y(t_c)), \tag{4}$$

where $t_t = Q(y)$ is the true value of the QoI (2), and $t_c = \min_{t\in(0,T]} \arg(S(Y(t)) = R)$ is a computed approximation to the QoI. Here, we assume that S satisfies the Lipschitz condition in y,

$$|S(y_1(t)) - S(y_2(t))| \leq K\|y_1(t) - y_2(t)\|_{\mathbb{R}^m}, \tag{5}$$

for some constant $K > 0$. Define the true error in the QoI, e_Q, to be

$$e_Q = t_t - t_c. \tag{6}$$

Theorem 1 (Convergence of the non-standard QoI) *Assume there is a numerical approximation to the solution of* (1) *satisfying* (3)*, and the functional $S(y(t))$ is continuously differentiable with respect to t in a neighborhood, B, which contains both the true QoI, t_t, as well as its numerical approximation, t_c. Further assume there exists an $M > 0$ such that*

$$\left|\frac{dS}{dt}(y(t))\right| > M, \tag{7}$$

for all $t \in B$. *Then the error* e_Q *in the computed QoI, defined by* (6)*, satisfies the bound,*

$$e_Q \leq \widehat{C}h^p,$$

for some constant $\widehat{C}$ *which depends on* M, C *and* K.

Proof Given the true solution $y(t)$ to (1), we consider the functional S as an explicit function of t, i.e.,

$$S(y) = S(y(t)) = S(t). \tag{8}$$

Since $S(y(t))$ is continuously differentiable in t, for $t \in B$, by the Inverse Function Theorem (see [42]) we have $t = t(S)$ for S in the image of B, and

$$\frac{dt}{dS}(S(y(t))) = \frac{1}{\frac{dS}{dt}(t(S))}. \tag{9}$$

Applying the Mean-value Theorem (see [2]) we have, for some ξ between $S(y(t_t))$ and $S(y(t_c))$,

$$t_t - t_c = \frac{dt}{dS}(\xi)\left[S(y(t_t)) - S(y(t_c))\right] = \frac{1}{\frac{dS}{dt}(t(\xi))}\left[S(y(t_t)) - S(y(t_c))\right]. \tag{10}$$

Adding and subtracting the term $S(Y(t_c))$ and recalling that $S(y(t_t)) = R = S(Y(t_c))$,

$$\begin{aligned} t_t - t_c &= \frac{1}{\frac{dS}{dt}(t(\xi))}\left[S(y(t_t)) - S(y(t_c)) + S(Y(t_c)) - S(Y(t_c))\right], \\ &= \frac{1}{\frac{dS}{dt}(t(\xi))}\left[S(Y(t_c)) - S(y(t_c))\right]. \end{aligned} \tag{11}$$

Taking norms (absolute values for scalars), and using (7) and (5),

$$\begin{aligned} |t_t - t_c| &= \left|\frac{1}{\frac{dS}{dt}(t(\xi))}\right| |S(Y(t_c)) - S(y(t_c))| \\ &\leq \frac{1}{M}K\,\|y(t_c) - Y(t_c)\|_{\mathbb{R}^m} \leq \frac{1}{M}KCh^p. \end{aligned} \tag{12}$$

Defining $\widehat{C} := \frac{KC}{M}$ gives the desired result. □

Remark 1 Theorem 1 and (12) implies that the approximate QoI t_c converges to t_t with at least the same rate as the numerical solution $Y(t)$. Further, the conclusion that the constant in the final bound (12) is *inversely* proportional to the lower bound on the absolute value of the derivative accords with the intuition developed through standard root-finding techniques such as Newton's method.

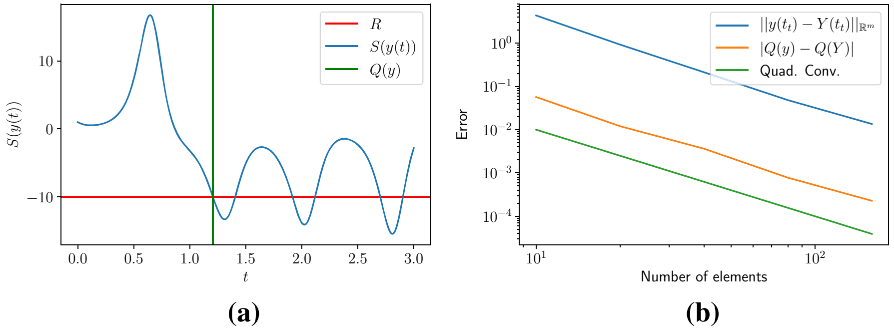


Fig. 1 **a** Reference solution and QoI for the Lorenz system (13). **b** Converge rates of the error in the solution and the error in the QoI. The numerical solution Y and QoI t_c, are computed using the cG(1) method

2.1 Example: Lorenz system

To illustrate the convergence results in Theorem 1, we consider the Lorenz system,

$$\left.\begin{aligned} \dot{y}_1 &= \sigma(y_2 - y_1), \\ \dot{y}_2 &= r y_1 - y_2 - y_1 y_3, \\ \dot{y}_3 &= y_1 y_2 - b y_3, \end{aligned}\right\} \; t \in (0, 3] \quad \text{with} \quad \left\{\begin{aligned} y_1(0) &= 1, \\ y_2(0) &= 0, \\ y_3(0) &= 24, \end{aligned}\right. \tag{13}$$

and set $\sigma = 10$, $r = 28$, and $b = \frac{8}{3}$ (see Sect. 5.2 for more details of this example). We define the functional $S(y(t)) = y_1(t)$ and set the threshold value $R = -10$. Figure 1a illustrates an accurate reference solution as well as the threshold value and the QoI. Figure 1b shows the convergence rates for the error in the solution and the error in the non-standard QoI when using the cG(1) method for computing the numerical solution. The cG(1) method (see Sect. 3.1 for details) has second order accuracy and this convergence rate is observed both for the solution and the non-standard QoI.

3 A posteriori error analysis

The aim is to derive an accurate a posteriori error estimate $\eta \approx e_Q$. The accuracy of the error estimate is quantified by the effectivity ratio,

$$\rho_{\text{eff}} = \frac{\eta}{e_Q}. \tag{14}$$

An effectivity ratio close to one indicates an accurate error estimate. We let ϵ denote the error in the solution to (1), i.e.,

$$\epsilon(t) = y(t) - Y(t). \tag{15}$$

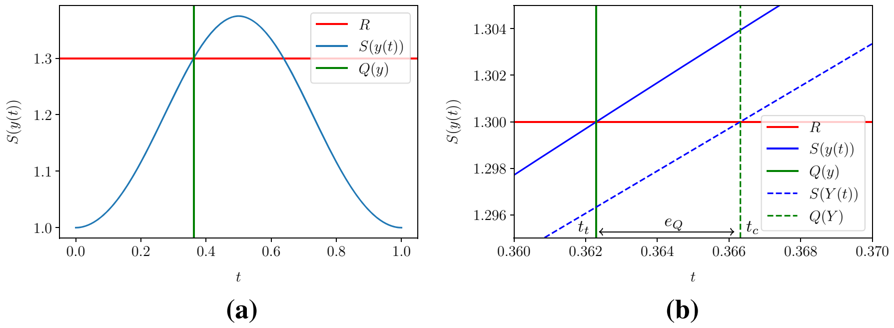


Fig. 2 **a** Graph showing true functional $S(y(t))$, chosen value of R, and true value of QoI for the example in Sect. 4.1. **b** Close up of true QoI and numerical QoI for the example in Sect. 4.1, solved using the Crank–Nicolson scheme

3.1 Integration schemes

For simplicity, we consider a continuous FEM approximation $Y(t)$, $t \in [0, T]$, with approximate functional $S(Y(t))$ as illustrated in Fig. 2. For each problem the linear functional $S(y(t))$ and the value of R are specified, and the problems are solved using two different numerical schemes: (i) a variational cG(1) finite element scheme using 40 equally-sized elements and high-order Gaussian quadrature, and (ii) a Crank–Nicolson finite difference scheme with 21 equally-spaced nodes. However, we stress that the analysis can be extended to a wide variety of numerical methods for which equivalence to a finite element method can be established, as discussed in Sect. 1.

Given the partition $\Lambda = \{0 = t_0, t_1, \ldots, t_N = T\}$ define the space,

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

where $\mathcal{P}^q(I_n)$ is the space of all polynomials of degree q or less on $I_n := [t_n, t_{n+1}]$. The continuous Galerkin finite element method of order $q + 1$, denoted cG(q), for solving (1) is defined interval-wise by: Find $Y \in \mathcal{V}^q$ such that

$$\int_{t_n}^{t_{n+1}} \dot{Y}(t) \cdot v(t)\, \mathrm{d}t = \int_{t_n}^{t_{n+1}} f(Y, t) \cdot v(t) dt, \quad \forall v \in \mathcal{P}^{q-1}(I_n), \tag{16}$$

for $n = 0, 1, 2, \ldots, N - 1$. The cG($q$) schemes are variational and hence well suited for adjoint based analysis. However, the Crank–Nicolson is also nodally equivalent to a variational scheme, see Theorem 5 in "Appendix A".

3.2 Adjoint-based a posteriori error analysis for standard QoIs

We derive error estimates for the nonstandard QoI in terms of expressions involving errors in linear functionals of the numerical solution. This section presents a standard a posteriori error estimate for a linear functional of a solution. Let $(\cdot, \cdot)$ denote the inner-product pairing in $\mathbb{R}^m$.

Theorem 2 (Adjoint-based a posteriori error analysis for IVPs) *Given a finite element solution Y(t) of* (1) *and* $\psi \in \mathbb{R}^m$*, the error* $(\psi, \epsilon(\hat{t}))$ *at* $\hat{t} \in (0, T]$ *is represented as*

$$(\psi, \epsilon(\hat{t})) = (\psi, y(\hat{t})) - (\psi, Y(\hat{t})) = \int_0^{\hat{t}} (\phi, [f(Y, t) - \dot{Y}])\, dt, \tag{17}$$

where ϕ *is the solution to the adjoint equation*

$$\begin{cases} -\dot{\phi} = \overline{f_{y,Y}(t)}^T \phi, \ t \in [0, \hat{t}), \\ \phi(\hat{t}) = \psi, \end{cases} \tag{18}$$

with

$$\overline{f_{y,Y}(t)} = \int_0^1 \frac{df}{dz}(z, t)\mathrm{d}s \tag{19}$$

and $z = sy + (1 - s)Y$*,*

Proof The proof is standard see [28]. □

Note that the adjoint equation (18) is solved backward in time from $\hat{t}$ to 0.

3.3 A posteriori analysis for the non-standard QoI based on Taylor series

We denote the error in the non-standard QoI as $e_Q = t_t - t_c$.

Theorem 3 *For an approximate solution* $Y(t)$ *to* (1) *and a bounded linear functional* $S(y(t))$ *on* $(H^1((0, T]))^m$*, if the function* $f(y, t)$ *is continuously differentiable in* t*, then the error in the QoI* (2) *is given by*

$$e_Q = \frac{S(Y(t_c)) - S(y(t_c)) - \mathcal{R}_1(t_c, t_t)}{\nabla_y S(Y(t_c)) \cdot f(Y(t_c), t_c) + \nabla_y[\nabla_y S(Y(t_c)) \cdot f(Y(t_c), t_c)] \cdot (y(t_c) - Y(t_c)) + \mathcal{R}_2(Y(t_c))}, \tag{20}$$

where the remainder terms $\mathcal{R}_1(t_t, t_c)$ *and* $\mathcal{R}_2(t_c)$ *satisfy*

$$\mathcal{R}_1(t_t, t_c) = \frac{1}{2}\frac{d^2 S}{dt^2}(y(\xi))(t_t - t_c)^2, \tag{21}$$

for some ξ *between* t_t *and* t_c *and*

$$\mathcal{R}_2(Y(t_c)) = ||y(t_c) - Y(t_c)||\mathcal{H}_2(Y(t_c)), \ \textit{for } \mathcal{H}_2 \textit{ with } \lim_{Y(t_c) \to y(t_c)} \mathcal{H}_2(Y(t_c)) = 0,$$

and $||\cdot||$ *denotes the Euclidean norm on* $\mathbb{R}^m$*.*

Proof From the definition of the functional $S(y(t))$ and R,

$$S(Y(t_c)) = R = S(y(t_t)). \tag{22}$$

Expanding $S(y(t_t))$ using Taylor's Theorem with remainder centered at t_c (e.g. see [2]) in (22),

$$S(Y(t_c)) = S(y(t_c)) + \frac{dS}{dt}(y(t_c))(t_t - t_c) + \mathcal{R}_1(t_c, t_t), \tag{23}$$

Applying the chain-rule to the derivative in (23) and using (1) gives

$$S(Y(t_c)) = S(y(t_c)) + \left[\nabla_y S(y(t_c)) \cdot f(y(t_c), t_c)\right](t_t - t_c) + \mathcal{R}_1(t_c, t_t). \tag{24}$$

Adding and subtracting the term $\nabla_y S(Y(t_c)) \cdot f(Y(t_c), t_c)$ inside the square brackets gives

$$\begin{aligned} S(Y(t_c)) = S(y(t_c)) & \\ & + \left[\nabla_y S(Y(t_c)) \cdot f(Y(t_c), t_c) + \left(\nabla_y S(y(t_c)) \cdot f(y(t_c), t_c)\right.\right. \\ & \left.\left. -\nabla_y S(Y(t_c)) \cdot f(Y(t_c), t_c)\right)\right](t_t - t_c) \\ & + \mathcal{R}_1(t_c, t_t). \end{aligned} \tag{25}$$

Using the multi-variable Taylor's Theorem with remainder centered at Y (e.g. see [3]) gives

$$\begin{aligned} & \nabla_y S(y(t_c)) \cdot f(y(t_c), t_c) - \nabla_y S(Y(t_c)) \cdot f(Y(t_c), t_c) \\ & \quad = \nabla_y [\nabla_y S(Y(t_c)) \cdot f(Y(t_c), t_c)] \cdot (y(t_c) - Y(t_c)) + \mathcal{R}_2(Y(t_c)), \end{aligned} \tag{26}$$

where the remainder is of the form

$$\mathcal{R}_2(Y(t_c)) = \frac{1}{2}(Y(t_c) - y(t_c))^\top \mathbf{H}_y(\nabla_y S(\xi) \cdot f(\xi, t_c))(Y(t_c) - y(t_c)), \tag{27}$$

for some ξ between $y(t_c)$ and $Y(t_c)$, and where $\mathbf{H}_y$ is the Hessian

$$(\mathbf{H}_y)_{i,j} = \frac{\partial^2}{\partial y_i \partial y_j}.$$

Substituting (26) in to (25) and rearranging to isolate the error of the QoI, results in

$$(t_t - t_c) = \frac{S(Y(t_c)) - S(y(t_c)) - \mathcal{R}_1(t_c, t_t)}{\nabla_y S(Y(t_c)) \cdot f(Y(t_c), t_c) + \nabla_y [\nabla_y S(Y(t_c)) \cdot f(Y(t_c), t_c)] \cdot (y(t_c) - Y(t_c)) + \mathcal{R}_2(Y(t_c))}. \tag{28}$$

□

Corollary 1 *For functionals $S(y(t), t)$ that are explicitly dependent on t,*

$$e_Q = \frac{S(Y(t_c), t_c) - S(y(t_c), t_c) - \mathcal{R}_1(t_c, t_t)}{\frac{\partial S}{\partial t}(y(t_c), t_c) + \nabla_y S(Y(t_c), t_c) \cdot f(Y(t_c), t_c) + \nabla_y[\nabla_y S(Y(t_c), t_c) \cdot f(Y(t_c), t_c)] \cdot (y(t_c) - Y(t_c)) + \mathcal{R}_2(Y(t_c)))}. \tag{29}$$

Where the partial derivative of S with respect to t appears from the chain-rule applied to (23).

Proof If S depends explicitly on t, then (24) becomes

$$S(Y(t_c), t_c) = S(y(t_c), t_c) + \left[\nabla_y S(y(t_c), t_c) \cdot f(y(t_c), t_c) + \frac{\partial S}{\partial t}(y(t_c), t_c)\right](t_t - t_c) + \mathcal{R}_1(t_c, t_t).$$

The remainder of the proof mimics the proof of Theorem 3 retaining this additional partial derivative. □

Note that functionals $S(y(t), t)$ that depend directly on t require special treatment of the term $\frac{\partial S}{\partial t}(y(t_c), t_c)$ in (29). More precisely, one can use another application of Taylor's Theorem centered at $Y(t_c)$ in order to make this term computable.

Corollary 2 *For functionals of the form $S(y(t)) = v \cdot y(t)$, for some $v \in \mathbb{R}^m$, $\nabla_y S(y(t)) = v$, and (20) becomes*

$$\begin{aligned} e_Q &= \frac{-v \cdot (y(t_c) - Y(t_c)) - \mathcal{R}_1(t_c, t_t)}{v \cdot f(Y(t_c), t_c) + v^\top \nabla_y f(Y(t_c), t_c) \cdot (y(t_c) - Y(t_c)) + \mathcal{R}_2(Y(t_c))} \\ &= \frac{-v \cdot \epsilon(t_c) - \mathcal{R}_1(t_c, t_t)}{v \cdot f(Y(t_c), t_c) + v^\top \nabla_y f(Y(t_c), t_c) \cdot \epsilon(t_c) + \mathcal{R}_2(Y(t_c))} \end{aligned} \tag{30}$$

Obtaining a computable error estimate. Taylor's Theorem gives that the two functions $\mathcal{R}_1$ and $\mathcal{R}_2$ in equations (20) and (30) decay to zero super-linearly as $t_c \to t_t$ and $Y(t_c) \to y(t_c)$, respectively. Provided the numerical solution $Y(t)$ is fairly accurate, $\mathcal{R}_1$ will be small compared to the other terms in (23) and $\mathcal{R}_2$ will be small compared to the terms in (26). This leads to the first approximation of the error,

$$\eta(Y) = \frac{-v \cdot \epsilon(t_c)}{v \cdot f(Y(t_c), t_c) + (v^\top \nabla_y f(Y(t_c), t_c)) \cdot \epsilon(t_c)}. \tag{31}$$

Remark 2 Note that the functional S may achieve the value R at multiple times. Assume there exists a time $\tilde{t} > t_t$ such that $S(y(\tilde{t})) = R$. Equation (22) is then valid at time $\tilde{t}$, i.e., $S(Y(t_c)) = R = S(y(\tilde{t}))$ and (20) follows with t_t replaced by $\tilde{t}$ and ξ replaced by $\tilde{\xi}$. In the estimate (31) we approximate the term $\mathcal{R}_1(t_c, \cdot)$ by zero. If the numerical solution is sufficiently accurate, then $|t_t - t_c| < |\tilde{t} - t_c|$ and $0 \approx \mathcal{R}_1(t_c, t_t) \ll \mathcal{R}_1(t_c, \tilde{t})$. However, if the numerical solution is inaccurate, we may have the reverse situation, where $|t_t - t_c| > |\tilde{t} - t_c|$, in which case the error estimate will be inaccurate or worse, $\mathcal{R}_1(t_c, \tilde{t}) \approx 0$ and the estimate may indicate the value of $\tilde{t} - t_c$ rather than $t_t - t_c$. We observe this phenomenon in Sect. 4.4.1 and is illustrated by Table 10 and Fig. 6b.

The estimate (31) contains two terms that are linear functionals of the error at time t_c. These can both be approximated by the standard techniques in Sect. 3.2 as is discussed next.

First adjoint problem In order to estimate $-v \cdot \epsilon(t_c)$, we solve (18) with adjoint data $\psi = \psi_1 = -v$ and $\hat{t} = t_c$, then substitute the solution ϕ_1 in (17) to provide the estimate

$$\mathcal{E}_1(Y, \phi_1) \approx \psi_1 \cdot \epsilon(t_c) = -v \cdot \epsilon(t_c). \tag{32}$$

Second adjoint problem In order to estimate $v^T \nabla_y f(Y(t_c), t_c) \cdot \epsilon(t_c)$, we solve (18) with adjoint data $\psi = \psi_2 = v^T \nabla_y f(Y(t_c), t_c)$ and $\hat{t} = t_c$, then substitute the solution ϕ_2 in (17) to provide the estimate

$$\mathcal{E}_2(Y, \phi_2) \approx \psi_2 \cdot \epsilon(t_c) = v^\top \nabla_y f(Y(t_c), t_c) \cdot \epsilon(t_c). \tag{33}$$

Computable error based on Taylor series and adjoint techniques. For an approximate solution $Y(t)$ to (1) and a linear functional $S(Y(t)) = v \cdot Y(t)$, a computable estimate of the error in the QoI (2) is obtained by substituting (32) and (33) in (31),

$$\eta(Y, \phi_1, \phi_2) = \frac{\mathcal{E}_1(Y, \phi_1)}{v \cdot f(Y(t_c), t_c) + \mathcal{E}_2(Y, \phi_2)}. \tag{34}$$

3.4 Error in non-standard QoI based on iterative techniques

Given an approximate solution $Y(t)$ to (1) with numerical QoI t_c, define $g(t)$ as

$$\begin{aligned} g(t) &= S(y(t)) - R, \\ &= S(Y(t)) + \big(S(y(t)) - S(Y(t))\big) - R, \end{aligned} \tag{35}$$

so

$$g(t_t) = 0. \tag{36}$$

In the case where $S(t)$ is a linear functional of $y(t)$, i.e., $S(y(t)) = v \cdot y(t)$, then

$$g(t) = S(Y(t)) + v \cdot \epsilon(t) - R.$$

At $t = \hat{t}$ we estimate $v \cdot \epsilon(\hat{t})$ by solving (18) with adjoint data $\psi = \psi_3 = v^\top$ and substituting the solution ϕ_3 in to (17) to provide the estimate

$$\mathcal{E}_3(Y, \phi_3; \hat{t}) \approx v^\top \cdot (y(\hat{t}) - Y(\hat{t})), \tag{37}$$

hence

$$g(\hat{t}) = S(Y(\hat{t})) + \mathcal{E}_3(Y, \phi_3; \hat{t}) - R.$$

We find t^* such that $g(t^*) \approx 0$ via a standard root finding procedure, then

$$\eta(Y) = t^* - t_c. \tag{38}$$

There are many options for root finding methods for computing η. In this article, we use two of the basic root finding methods: the secant method and the inverse quadratic method.

3.4.1 Error estimate based on the secant method

Given initial values $x_0,\ x_1$, the method is defined by the recurrence

$$x_n = \frac{x_{n-2} * g(x_{n-1}) - x_{n-1} * g(x_{n-2})}{g(x_{n-1}) - g(x_{n-2})} \quad n = 2, 3, \ldots \tag{39}$$

(See [38]). For the initial guesses the examples presented choose $x_0 < t_c < x_1$. These choices are made precise in the numerical examples in Sect. 4.

3.4.2 Error estimate based on inverse quadratic interpolation

Given initial values $x_0,\ x_1,\ x_2$, the method is defined by the recurrence

$$\begin{aligned} x_n = {} & \frac{x_{n-3}\ g_{n-2}\ g_{n-1}}{(g_{n-3} - g_{n-2})(g_{n-3} - g_{n-1})} + \frac{x_{n-2}\ g_{n-3}\ g_{n-1}}{(g_{n-2} - g_{n-3})(g_{n-2} - g_{n-1})} \\ & + \frac{x_{n-1}\ g_{n-2}\ g_{n-3}}{(g_{n-1} - g_{n-2})(g_{n-1} - g_{n-3})}. \\ & n = 3, 4, \ldots. \end{aligned} \tag{40}$$

(See [27]). The choice of the initial guesses is made precise in the numerical examples in Sect. 4.

3.5 Comparison of the two error estimation methods

The method based on Taylor series always requires fewer adjoint problems to be solved than using one of the iterative methods. However, the estimate (31) neglects certain terms compared to the error representation (30). If any of the neglected terms are large, the error estimate may be inaccurate even though an accurate numerical solution is used. The iterative methods only rely on the initial guesses and pointwise error computation, which is computed accurately. The initial guesses defined in Sect. 4 bracket the computed QoI, and provided the computed solution is sufficiently accurate and the initial bracket contains only a single value t such that $S(y(t)) = R$, the iterative methods will be accurate. Numerical comparisons of the two methods, as well as limitations of both are discussed throughout Sect. 4.

3.6 Error in a cumulative density function

If the differential equation (1) depends on a random parameter θ, then the solution $y(t;\theta)$ and the QoI, $Q(y;\theta)$, are random variables. As a random variable, $Q(y;\theta)$ has a corresponding cumulative distribution function (CDF),

$$F(t) = P(\{\theta : Q(y;\theta) \le t\}) = P(Q \le t).$$

An approximation to the CDF is computed using the Monte Carlo method with a finite number of numerically computed sample values $\{\hat{Q}(Y^{[n]}, \theta^{[n]}) = \hat{Q}^{[n]}\}_{n=1}^{N}$,

$$\hat{F}_N(t) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}(\hat{Q}^{[n]} \le t), \tag{41}$$

where $\mathbb{1}$ is the indicator function. A nominal sample distribution is computed using exact values of the QoI,

$$F_N(t) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}(Q^{[n]} \le t). \tag{42}$$

An estimate of the error in an approximate distribution of the QoI (2) is computed for two examples in Sect. 5. The estimate takes into account error contributions due to finite sampling and errors arising from the discretization of the ODE. The expressions (41) and (42) decompose the error in to sampling and discretization contributions,

$$F(t) - \hat{F}_N(t) = (F(t) - F_N(t)) + (F_N(t) - \hat{F}_N(t)).$$

This decomposition is used to derive the following error bound.

Theorem 4 *For* $0 < \varepsilon < 1$,

$$\begin{aligned} \left|F(t) - \hat{F}_N(t)\right| \le &\left(\frac{\hat{F}_N(t)\left(1-\hat{F}_N(t)\right)}{N\varepsilon}\right)^{1/2} \\ &+ \left(\frac{1}{N} + \frac{1}{N\varepsilon^{1/2}}\right)\left|\sum_{n=1}^{N}\left(\mathbb{1}(\hat{Q}^{[n]} - \left|e_Q^{[n]}\right| \le t \le \hat{Q}^{[n]} + \left|e_Q^{[n]}\right|)\right)\right| \\ &+ \frac{2}{(2N\varepsilon)^{3/4}} \end{aligned} \tag{43}$$

with probability greater than or equal to $1 - 2\varepsilon + \varepsilon^2$, *where* $e_Q^{[n]} = Q^{[n]} - \hat{Q}^{[n]}$ *is the error in a numerically computed sample of the QoI.*

Proof We decompose the error as

$$\left|F(t) - \hat{F}_N(t)\right| \le |F(t) - F_N(t)| + \left|F_N(t) - \hat{F}_N(t)\right| = I + II. \tag{44}$$

Focusing on the term $II = \left|F_N(t) - \hat{F}_N(t)\right| = \left|\hat{F}_N(t) - F_N(t)\right|$,

$$\begin{aligned}
II &= \left|\frac{1}{N}\sum_{n=1}^{N}\left(\mathbb{1}(\hat{Q}^{[n]} \le t) - \mathbb{1}(Q^{[n]} \le t)\right)\right| \\
&= \left|\frac{1}{N}\sum_{n=1}^{N}\left(\mathbb{1}(\hat{Q}^{[n]} \le t) - \mathbb{1}(\hat{Q}^{[n]} + e_Q^{[n]} \le t)\right)\right|, \\
&= \left|\frac{1}{N}\sum_{\substack{n=1\\ e_Q^{[n]} \le 0}}^{N}\left(\mathbb{1}(\hat{Q}^{[n]} - \left|e_Q^{[n]}\right| \le t \le \hat{Q}^{[n]})\right) + \frac{1}{N}\sum_{\substack{n=1\\ e_Q^{[n]} > 0}}^{N}\left(\mathbb{1}(\hat{Q}^{[n]} \le t \le \hat{Q}^{[n]} + \left|e_Q^{[n]}\right|)\right)\right|, \\
&\le \left|\frac{1}{N}\sum_{n=1}^{N}\left(\mathbb{1}(\hat{Q}^{[n]} - \left|e_Q^{[n]}\right| \le t \le \hat{Q}^{[n]})\right) + \frac{1}{N}\sum_{n=1}^{N}\left(\mathbb{1}(\hat{Q}^{[n]} \le t \le \hat{Q}^{[n]} + \left|e_Q^{[n]}\right|)\right)\right|, \\
&= \left|\frac{1}{N}\sum_{n=1}^{N}\left(\mathbb{1}(\hat{Q}^{[n]} - \left|e_Q^{[n]}\right| \le t \le \hat{Q}^{[n]} + \left|e_Q^{[n]}\right|)\right)\right|,
\end{aligned} \tag{45}$$

Now consider the term $I = |F(t) - F_N(t)|$. We start with the Chebyshev Inequality:

$$P\left(|F(t) - F_N(t)| \ge ks\right) \le \frac{1}{k^2}$$

for any real number k, where s^2 is the variance of F_N given by [36,43],

$$s^2 = \frac{F(t)\left(1 - F(t)\right)}{N}.$$

Choosing $\varepsilon = \frac{1}{k^2}$ leads to

$$I = |F(t) - F_N(t)| \le \left(\frac{F(t)\left(1 - F(t)\right)}{N\varepsilon}\right)^{1/2}, \tag{46}$$

with a probability greater than $1 - \varepsilon$. Now,

$$F(t)\left(1 - F(t)\right) = F_N(t)\left(1 - F_N(t)\right) + \left(F(t) - F_N(t)\right)\left(1 - F(t) - F_N(t)\right). \tag{47}$$

Taking absolute values in (47), dividing by $N\varepsilon$, taking the square root, and using $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for any $a, b \ge 0$,

$$\left|\frac{F(t)\,(1-F(t))}{N\varepsilon}\right|^{1/2} \le \left|\frac{F_N(t)\,(1-F_N(t))}{N\varepsilon}\right|^{1/2} + \left|\frac{(F(t)-F_N(t))\,(1-F(t)-F_N(t))}{N\varepsilon}\right|^{1/2}. \tag{48}$$

Multiplying and dividing the second term on the right-hand side of (48) by $\sqrt{2}\delta$ and using the fact that $ab \le \frac{a^2}{2} + \frac{b^2}{2}$,

$$\begin{aligned}\left|\frac{(F(t)-F_N(t))\,(1-F(t)-F_N(t))}{N\varepsilon}\right|^{1/2} &\le \left|\delta^2\,(F(t)-F_N(t))^2 + \frac{(1-F(t)-F_N(t))^2}{4\delta^2 N^2 \varepsilon^2}\right|^{1/2} \\ &\le \delta\,|F(t)-F_N(t)| + \frac{1}{2\delta N\varepsilon},\end{aligned}$$

where we obtain the final line by observing that $(1-F(t)-F_N(t))^2 \le 1$.

Substituting back into (48) and combining with (46),

$$I \le \left(\frac{F_N(t)\,(1-F_N(t))}{N\varepsilon}\right)^{1/2} + \delta\,|F(t)-F_N(t)| + \frac{1}{2\delta N\varepsilon}. \tag{49}$$

From [43], for any $\varepsilon > 0$ we have with a probability greater than $1-\varepsilon$,

$$I \le \left(\frac{\log(\varepsilon^{-1})}{2N}\right)^{1/2} \le \left(\frac{1}{2N\varepsilon}\right)^{1/2}, \tag{50}$$

where we also used that $\log(x) \le x$ for all $x > 0$. Substituting this into the right-hand side of (49),

$$I \le \left(\frac{F_N(t)\,(1-F_N(t))}{N\varepsilon}\right)^{1/2} + \delta\left(\frac{1}{2N\varepsilon}\right)^{1/2} + \frac{1}{2\delta N\varepsilon}. \tag{51}$$

Consider the function

$$D(\delta) = \frac{\delta}{(2N\varepsilon)^{1/2}} + \frac{1}{\delta(2N\varepsilon)},$$

Elementary calculus shows that the minimum of $D(\delta)$, for $\delta > 0$, occurs at $\delta = \left(\frac{1}{2N\varepsilon}\right)^{1/4}$.

With this choice of δ, (51) becomes

$$I \le \left(\frac{F_N(t)\,(1-F_N(t))}{N\varepsilon}\right)^{1/2} + \frac{2}{(2N\varepsilon)^{3/4}}. \tag{52}$$

The numerator of the first term in (52) is expanded as

$$|F_N(t)\,(1-F_N(t))| = \left|\hat{F}_N(t)\left(1-\hat{F}_N(t)\right) + \left(F_N(t)-\hat{F}_N(t)\right)\left(1-F_N(t)-\hat{F}_N(t)\right)\right|$$

$$\leq \left|\hat{F}_N(t)\left(1-\hat{F}_N(t)\right)\right| + \left|\left(F_N(t)-\hat{F}_N(t)\right)\left(1-F_N(t)-\hat{F}_N(t)\right)\right|. \tag{53}$$

Using $\left|1-F_N(t)-\hat{F}_N(t)\right| \leq 1$ in (53) together with (45) and (52),

$$\begin{aligned} I &\leq \left(\frac{\hat{F}_N(t)\left(1-\hat{F}_N(t)\right)}{N\varepsilon}\right)^{1/2} \\ &\quad + \frac{1}{N\varepsilon^{1/2}}\left(\left|\sum_{n=1}^{N}\left(\mathbb{1}(\hat{Q}^{[n]}-\left|e_Q^{[n]}\right| \leq t \leq \hat{Q}^{[n]}+\left|e_Q^{[n]}\right|)\right)\right|\right)^{1/2} + \frac{2}{(2N\varepsilon)^{3/4}}, \\ &\leq \left(\frac{\hat{F}_N(t)\left(1-\hat{F}_N(t)\right)}{N\varepsilon}\right)^{1/2} \\ &\quad + \frac{1}{N\varepsilon^{1/2}}\left(\left|\sum_{n=1}^{N}\left(\mathbb{1}(\hat{Q}^{[n]}-\left|e_Q^{[n]}\right| \leq t \leq \hat{Q}^{[n]}+\left|e_Q^{[n]}\right|)\right)\right|\right) + \frac{2}{(2N\varepsilon)^{3/4}}, \end{aligned} \tag{54}$$

where we also used $\sqrt{x} \leq x$ if $x = 0$ or $x \geq 1$.

Since (54) relies on both (46) and (50), this bound occurs with a probability of at least $(1-\varepsilon)^2 = 1-2\varepsilon+\varepsilon^2$. Combining (45) and (54) with (44) completes the proof. □

The estimate (34) is used to approximate $\eta^{[n]} \approx e_Q^{[n]}$. The first term on the right-hand side of the bound (43) quantifies the error contribution from finite sampling, while the second term represents error due to discretization.

4 Numerical examples

This section considers a wide range of types of linear and nonlinear ODEs in order to explore the accuracy of the estimates.

Since the Crank–Nicolson finite difference scheme is nodally equivalent to the cG(1) finite element method with a trapezoidal rule quadrature, given $t_i < t_c < t_{i+1}$, the numerical QoI may be computed by using linear interpolation as,

$$t_c = \frac{R(t_i - t_{i+1})}{Y(t_i) - Y(t_{i+1})} - \frac{t_i Y(t_i) - t_{i+1} Y(t_{i+1})}{Y(t_i) - Y(t_{i+1})}.$$

When implementing the secant method (39), the two grid-points closest to the QoI are used as initial guesses:

$$x_0 = t_L \text{ and } x_1 = t_R, \tag{55}$$

where $t_L < t_c < t_R$, with no other grid-points in between. For the inverse quadratic interpolation scheme (40), the initial guesses are the two closest grid-points to the left of the QoI and one to the right:

$$x_0 = t_{LL},\ x_1 = t_L \text{ and } x_2 = t_R, \tag{56}$$

where $t_{LL} < t_L < t_c < t_R$, with no other grid-points in between. For most examples the adjoint solutions, needed for the estimates (32), (33) and (34), are computed using the cG(3) method with 100 finite elements, with the exceptions of Sect. 4.5 where cG(3) is used with 40 elements and Sect. 5.2 where cG(2) with 100 elements is used. For all methods define n_{adj} to be the number of adjoint solutions required to compute the error in the QoI. This number can be seen as the relative cost of implementing the different methods.

4.1 Linear problem

We consider the initial value problem

$$\dot{y} = \sin(2\pi t)y, \quad t \in (0, 1], \quad y(0) = 1,$$

with analytic solution

$$y(t) = \exp\left(\frac{1}{2\pi}(1 - \cos(2\pi t))\right).$$

Let $R = 1.3$ and $S(y(t)) = y(t)$. The true QoI is given by

$$t_t = Q(y) = \min_{t\in(0,1]} \arg(y(t) = 1.3) = \frac{1}{2\pi}(\arccos(-2\pi \ln(1.3) + 1)).$$

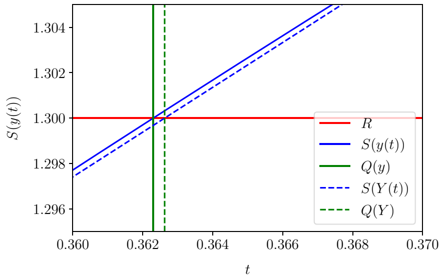
For this problem, the terms in (31) are

$$v = 1,\ f(y, t) = \sin(2\pi t)y,\ \nabla_y f(y, t) = \sin(2\pi t),$$

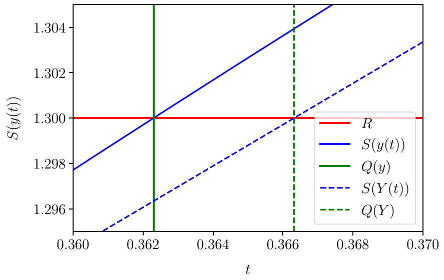
hence, for (32), (33), and (37) the values needed are

$$\psi_1 = -1,\ \psi_2 = \sin(2\pi t_c),\ \psi_3 = 1.$$

The true solution and QoI are shown in Fig. 3. This graph includes a horizontal line at $S(y(t)) = R$, to indicate the threshold value of interest, as well as a vertical line denoting the true value of the QoI, i.e. the first time the threshold is crossed. Figure 3 compares the numerical QoI to the true value for both the numerical schemes. True errors, error estimates and effectivity ratios are provided in Tables 1 and 2. All methods provide excellent effectivity ratios, but the iterative methods require many more applications of Theorem 2 and hence require solving more adjoint problems of the form (18), as shown by the values of η_{adj}.



(a)



(b)

Fig. 3 **a** Comparing cG(1) solution and computed QoI(2) to the true values for example in Sect. 4.1. **b** Comparing Crank–Nicolson solution and computed QoI (2) to the true values for example in Sect. 4.1

Table 1 Results of the different methods on the example in Sect. 4.1 using cG(1) with 40 elements

Method	t_c	t_{LL}	t_L	t_R	e_Q	η	ρ_{eff}	n_{adj}
Taylor series	0.3626	–	–	–	$-3.267\text{e}{-04}$	$-3.269\text{e}{-04}$	1.000	2
Secant	0.3626	–	0.35	0.375	$-3.267\text{e}{-04}$	$-3.267\text{e}{-04}$	1.000	6
Inverse quad.	0.3626	0.325	0.35	0.375	$-3.267\text{e}{-04}$	$-3.267\text{e}{-04}$	1.000	7

Table 2 Results of the different methods on the example in Sect. 4.1 using Crank–Nicolson with 21 nodes

Method	t_c	t_{LL}	t_L	t_R	e_Q	η	ρ_{eff}	n_{adj}
Taylor series	0.3663	–	–	–	$-4.017\text{e}{-03}$	$-4.056\text{e}{-03}$	1.010	2
Secant	0.3663	–	0.35	0.4	$-4.017\text{e}{-03}$	$-4.017\text{e}{-03}$	1.000	7
Inverse quad.	0.3663	0.3	0.35	0.4	$-4.017\text{e}{-03}$	$-4.017\text{e}{-03}$	1.000	7

4.2 Nonlinear problem

Next we consider the nonlinear initial value problem

$$\dot{y}(t) = \sin(2\pi y(t)),\ t \in (0, 1], \quad y(0) = \frac{1}{4}.$$

The analytic solution to this problem is

$$y(t) = \frac{1}{\pi} \arctan(e^{2\pi t}).$$

Let $R = 0.4$ and $S(y(t)) = y(t)$. The true QoI is

$$t_t = Q(y) = \min_{t\in[0,1]} \arg(y(t) = 0.4) = \frac{\ln(\tan(0.4\pi))}{2\pi}.$$

Here, the terms in (31) are

$$v = 1, \ \ f(y,t) = \sin(2\pi y), \ \ \nabla_y f(y,t) = 2\pi \cos(2\pi y),$$

so the data needed for (32), (33), and (37) are

$$\psi_1 = -1, \ \ \psi_2 = 2\pi \cos(2\pi R), \ \ \psi_3 = 1.$$

Figure 4a shows the true values of the linear functional $S(y(t))$ as well as the event in question and the true QoI. The values in Tables 3 and 4 indicate that all three methods are fairly accurate. The two iterative methods again require more adjoint equations to be solved.

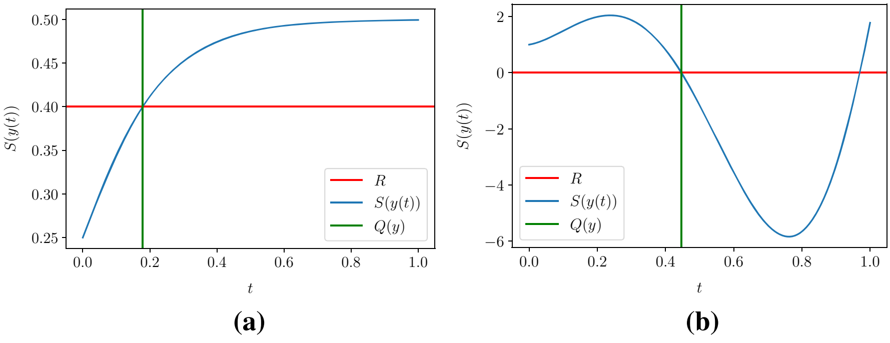


Fig. 4 **a** Chosen value of R, true data $S(y(t))$, and true QoI for example in Sect. 4.2. **b** Chosen value of R, true data $S(y(t))$, and true QoI for example in Sect. 4.3

Table 3 Results for Sect. 4.2 using the different methods on cG(1) solution with 40 elements

Method	t_c	t_{LL}	t_L	t_R	e_Q	η	ρ_{eff}	n_{adj}
Taylor series	0.1790	–	–	–	− 1.087e−04	− 1.086e−04	1.000	2
Secant	0.1790	–	0.175	0.2	− 1.087e−04	− 1.087e−04	1.000	6
Inverse quad.	0.1790	0.15	0.175	0.2	− 1.087e−04	− 1.087e−04	1.000	6

Table 4 Results for Sect. 4.2 using the different methods on Crank–Nicolson solution with 21 nodes

Method	t_c	t_{LL}	t_L	t_R	e_Q	η	ρ_{eff}	n_{adj}
Taylor series	0.1810	–	–	–	− 2.156e−03	− 2.141e−03	1.007	2
Secant	0.1810	–	0.15	0.2	− 2.156e−03	− 2.144e−03	1.001	7
Inverse quad.	0.1810	0.1	0.15	0.2	− 2.156e−03	− 2.144e−03	1.001	7

4.3 Linear system

We consider the two dimensional system $\dot{y} + A(t)y = 0$,

$$\begin{pmatrix} \dot{y}_1(t) \\ \dot{y}_2(t) \end{pmatrix} + \begin{pmatrix} 1 + 9\cos^2(6t) - 6\sin(12t) & -12\cos^2(6t) - 9/2\sin(12t) \\ 12\sin^2(6t) - 9/2\sin(12t) & 1 + 9\sin^2(6t) + 6\sin(12t) \end{pmatrix} \begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \ t \in (0, 1],$$

with initial conditions $y_1(0) = y_2(0) = 1$. The analytic solution to this problem is

$$\begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix} = \begin{pmatrix} 3/5\exp(2t)(\cos(6t) + 2\sin(6t)) - 1/5\exp(-13t)(\sin(6t) - 2\cos(6t)) \\ 3/5\exp(2t)(2\cos(6t) - \sin(6t)) - 1/5\exp(-13t)(\cos(6t) + 2\sin(6t)) \end{pmatrix}.$$

Set $R = 0$ and $S(y(t)) = y_1(t)$ in order to analyze the first component. The true quantity of interest is

$$t_t := Q(y) = 0.446255366908554$$

The parameters needed for (31) are

$$v = (1, 0)^\top, \ f(y, t) = -A(t)y, \ \nabla_y f(y, t) = -A(t).$$

For (32), (33), and (37) the values needed are

$$\psi_1 = -(1, 0)^\top, \quad \psi_3 = (1, 0)^\top,$$
$$\psi_2 = (1 + 9\cos^2(6t_c) - 6\sin(12t_c) - 12\cos^2(6t_c) - \frac{9}{2}\sin(12t_c))^\top.$$

The true solution and QoI are shown in Fig. 4b. Tables 5 and 6 show the results for cG(1) and Crank–Nicolson respectively. Again, all methods are accurate using either numerical method. The two iterative methods require many more adjoint problems to be solved than the Taylor series method without any increase in accuracy.

Table 5 Results of the different methods on example in Sect. 4.3 using cG(1) with 40 elements

Method	t_c	t_{LL}	t_L	t_R	e_Q	η	ρ_{eff}	n_{adj}
Taylor series	0.4463	–	–	–	− 1.323e−04	− 1.322e−04	0.999	2
Secant method	0.4463	–	0.425	0.45	− 1.323e−04	− 1.323e−04	1.000	6
Inverse quad.	0.4463	0.4	0.425	0.45	− 1.323e−04	− 1.323e−04	1.000	8

Table 6 Results of the different methods on example in Sect. 4.3 using Crank–Nicolson with 21 nodes

Method	t_c	t_{LL}	t_L	t_R	e_Q	η	ρ_{eff}	n_{adj}
Taylor series	0.4462	–	–	–	2.675e−05	2.675e−05	1.000	2
Secant	0.4462	–	0.4	0.45	2.675e−05	2.675e−05	1.000	6
Inverse quad.	0.4462	0.35	0.4	0.45	2.675e−05	2.675e−05	1.000	8

4.4 Harmonic oscillator

Consider the harmonic oscillator

$$\ddot{\omega} = -\frac{k}{m}\omega - \frac{c}{m}\dot{\omega} + \frac{F_0}{m}\cos(\gamma t + \theta_d), \;\; t \in (0, 2], \quad \omega(0) = 5, \;\; \dot{\omega}(0) = 0.$$

with

$$k = 50, \;\; m = 0.25, \;\; c = 1, \;\; F_0 = 50, \;\; \theta_d = 0, \;\; \gamma = 10.$$

Rewriting as a system of first-order ODEs, $\dot{y} + Ay = h(t)$, gives

$$\begin{pmatrix} \dot{y_1}(t) \\ \dot{y_2}(t) \end{pmatrix} + \begin{pmatrix} 0 & -1 \\ 200 & 4 \end{pmatrix} \begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix} = \begin{pmatrix} 0 \\ 200\cos(10t) \end{pmatrix}.$$

Set $R = 0$ and $S(y(t)) = y_1(t)$ in order to observe when the oscillator first reaches the origin. The true solution in [5] is used to determine

$$t_t := Q(\omega) = 0.14034864129073557.$$

Here for (31), the values needed are

$$v = (1, 0)^\top, \;\; f(y, t) = -Ay + h(t), \;\; \nabla_y f(y, t) = -A.$$

To compute (32), (33), and (37), let

$$\psi_1 = -(1, 0)^\top, \;\; \psi_2 = (0, 1)^\top, \;\; \psi_3 = (1, 0)^\top.$$

The true data $S(y(t))$ and QoI are given in Fig. 5a and the results using cG(1) and Crank–Nicolson method are provided in Tables 7 and 8 respectively. All methods using either numerical method give effectivity ratios close to one. The two iterative methods require more adjoint problems to be solved than the Taylor series estimate, but they do lead to a slightly more accurate error estimate.

4.4.1 Harmonic oscillator: effect of the choice of interval

We consider the same equation and function as in Sect. 4.4, except over the time interval $t \in (0.2, 2]$ and with $R = 1.8$.

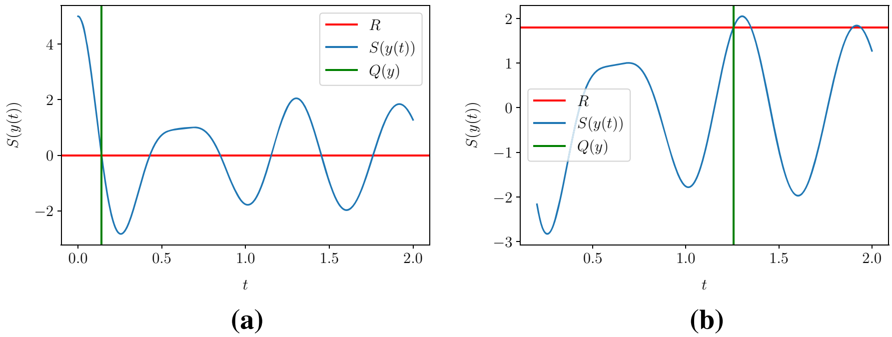


Fig. 5 **a** Chosen value of R, true data $S(y(t)$, and true QoI $Q(y)$ for example 4.4. **b** Chosen value of R, true data $S(y(t))$, and true QoI $Q(y)$ for example in Sect. 4.4.1

Table 7 Results of the different methods on the example in Sect. 4.4 using cG(1) with 40 elements

Method	t_c	t_{LL}	t_L	t_R	e_Q	η	ρ_{eff}	n_{adj}
Taylor series	0.1447	–	–	–	− 4.440e−03	− 4.449e−03	1.011	2
Secant method	0.1447	–	0.1	0.15	− 4.440e−03	− 4.440e−03	1.000	7
Inverse quad.	0.1447	0.05	0.1	0.15	− 4.440e−03	− 4.440e−03	1.000	8

Table 8 Results of the different methods on the example in Sect. 4.4 using Crank–Nicolson with 21 nodes

Method	t_c	t_{LL}	t_L	t_R	e_Q	η	ρ_{eff}	n_{adj}
Taylor series	0.1575	–	–	–	− 1.715-02	− 1.816e−02	1.059	2
Secant method	0.1575	–	0.1	0.2	− 1.715-02	− 1.715e−02	0.999	8
Inverse quad.	0.1575	0.0	0.1	0.2	− 1.715-02	− 1.715e−02	0.999	10

Applying the secant method to the true solution results in the true QoI,

$$t_t = 1.2558594599461572.$$

Since this problem has the same ODE and functional S as in Sect. 4.4, the parameters and steps laid out in that section can be used to obtain the error estimates.

The true functional and QoI are shown in Fig. 5b and the results when using the different methods in Tables 9 and 10. The Taylor series method is slightly less accurate compared to the iterative methods when using the cG(1) method. This is due to the size of the second derivative of the functional near the event, leading to a larger absolute value of the remainder in (26). Since the error estimate (31) neglects this remainder, if its absolute value is too large the estimate will not be accurate. Examples in Sect. 4.4.2 take a further look into this effect.

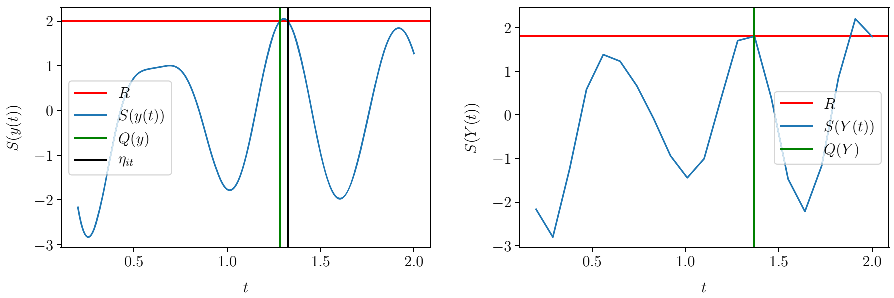
Both the Taylor series and iterative methods are poor for the Crank–Nicolson method. This is due to the low accuracy of the numerical solution as illustrated in Fig. 6b. The potential inaccuracy of the Taylor series estimate under these circum-

Table 9 Results of the different methods on example in Sect. 4.4.1 using cG(1) with 40 elements

Method	t_c	t_{LL}	t_L	t_R	e_Q	η	ρ_{eff}	n_{adj}
Taylor series	1.2637	–	–	–	−7.887e−03	−8.623e−03	1.093	2
Secant	1.2637	–	1.235	1.37	−7.887e−03	−7.887e−03	0.999	8
Inverse quad.	1.2637	1.19	1.235	1.37	−7.887e−03	−7.887e−03	0.999	9

Table 10 Results of the different methods on example in Sect. 4.4.1 using Crank–Nicolson with 21 nodes

Method	t_c	t_{LL}	t_L	t_R	e_Q	η	ρ_{eff}	n_{adj}
Taylor series	1.3674	–	–	–	−1.116e−01	−1.542e−02	0.138	2
Secant method	1.3674	–	1.28	1.37	−1.116e−01	−1.746e−02	0.156	8
Inverse quad.	1.3674	1.19	1.28	1.37	−1.116e−01	−1.746e−02	0.156	10



(a) **(b)**

Fig. 6 **a** Figure detailing issue with iterative methods in Sect. 4.4.1 and Sect. 4.4.2 when the numerical solution is not accurate near the event. The iterative methods result in $t^* = \eta_{it}$, which is the second occurrence of the event rather than the first. This figure specifically details the case when $R = 2$. **b** Numerical values for example in Sect. 4.4.1 when using Crank–Nicolson method with 21 nodes

stances is discussed in Remark 2. The root-finding methods are converging to the *second* time the event occurs (which is 1.3237), rather than the first. Because of the small difference in time between the locations of the two roots (see Fig. 6a), the proximity of the second root to the numerical QoI, and the size of the numerical time step, both roots are contained within the initial interval over which the iterative methods are applied. It is therefore possible for the iterative methods to converge to the larger of the two roots.

4.4.2 Harmonic oscillator: effect of the choice of *R*

Again consider the harmonic oscillator of Sect. 4.4.1, and estimate the error of the QoI (2) with several different values of R, increasing R until it is very close to the maximum of the true data. The maximum value of the true data is approximately 2.05015 (see Fig. 7). Results are provided in Tables 11, 12 and 13 for increasingly fine

Table 11 Effectivity ratio for the different methods for varying values of R on example in Sect. 4.4 using cG(1) with 40 elements

Method	R = 1.95	R = 2.0	R = 2.01	R = 2.02	R = 2.03	R = 2.04	R = 2.05
Taylor series	1.061	1.095	1.251	1.603	3.470	– 1.137	0.427
Secant method	0.999	– 11.305	– 4.952	– 2.650	– 1.405	1.000	Fail
Inverse quad.	0.999	– 11.305	– 4.952	– 2.650	– 1.405	Fail	Fail

Table 12 Effectivity ratio for the different methods for varying values of R on example in Sect. 4.4 using cG(1) with 60 elements

Method	R = 1.95	R = 2.0	R = 2.01	R = 2.02	R = 2.03	R = 2.04	R = 2.05
Taylor series	1.033	0.999	1.043	1.100	1.179	1.283	0.758
Secant	1.000	0.999	0.999	0.999	– 6.545	– 4.520	3.133
Inverse quad.	1.000	0.999	0.999	0.999	– 6.545	– 4.520	3.133

Table 13 Effectivity ratio for the different methods for varying values of R on example in Sect. 4.4 using cG(1) with 100 elements

Method	R = 1.95	R = 2.0	R = 2.01	R = 2.02	R = 2.03	R = 2.04	R = 2.05
Taylor series	1.017	1.001	1.019	1.100	1.039	0.998	0.588
Secant	0.999	0.999	0.999	1.000	0.999	0.999	0.999
Inverse quad.	0.999	0.999	0.999	1.000	0.999	0.999	0.999

finite element meshes. The tables contain the effectivity ratios, ρ_{eff}, for each method and each value of R.

Notice that the iterative methods appear to be more sensitive to the accuracy of the numerical solution than the Taylor series method. In extreme cases, the iterative methods fail to converge. This occurs when a root-finding iteration falls outside of the domain of the IVP (1), i.e., if x_n the approximation to the root at the nth iteration, $x_n < 0$ or $x_n > T$. As the number of finite elements used to solve the ODE increases, the two iterative methods eventually recover their accuracy even when the threshold value is very close to an extremum. For the cases where the iterative methods are inaccurate, note that the root-finding schemes do *not* converge to the true QoI. Instead, the convergence is to the *second* occurrence of the event rather than the first (see Fig. 6a).

The estimate derived from Taylor's theorem is generally more accurate for the less accurate numerical solutions, However, even when using an accurate numerical solution, the Taylor series approach becomes inaccurate when the curvature of S as a function of t is large near the threshold value. The remainder $\mathcal{R}_1(t_t, t_c)$, given by (21), is one half of the second derivative of S with respect to t at some point between t_t and t_c. As the threshold value R moves closer to the local maximum, this derivative grows and the assumption that $\mathcal{R}_1(t_t, t_c)$ is small is no longer valid, resulting in an inaccurate estimate. The iterative methods do not depend on the values of the second derivative of

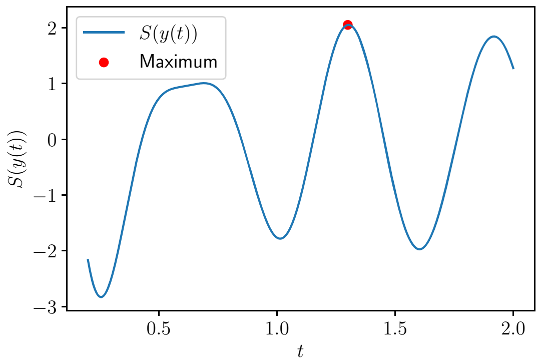


Fig. 7 True data for example in Sect. 4.4.1, showing max value of ≈ 2.05015

the solution and those methods are able to produce accurate error estimates provided the numerical solution is sufficiently accurate near the event.

4.5 One dimensional heat equation

We consider the one dimensional heat equation with boundary and initial conditions

$$\begin{aligned} u_t(x,t) &= u_{xx}(x,t) + 3e^t \sin(\pi x), \quad (x,t) \in (0,1) \times (0,1], \\ u(x,0) &= 0, \quad x \in (0,1), \\ u(0,t) &= 0,\ u(1,t) = 0, \quad t \in (0,1]. \end{aligned}$$

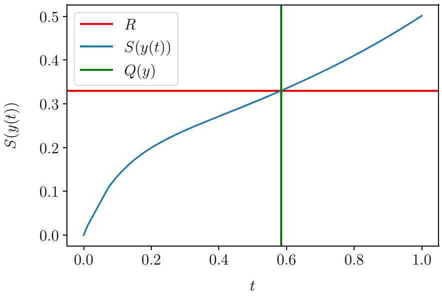
This section analyzes the system of ordinary differential equations that arises from a spatial discretization of (4.5) using a central-difference method. In particular using a uniform partition of the spatial interval [0, 1] with 22 nodes:

$$\{0 = x_0 < x_1 < \cdots < x_{21} = 1\}.$$

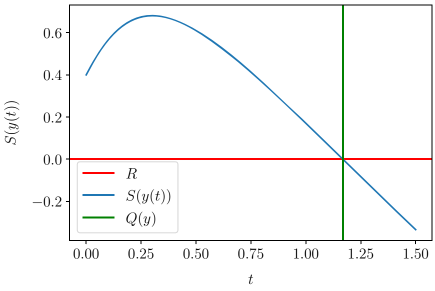
Since boundary values are specified, this semi-discretization leads to a system of 20 first-order ODEs of the form $\dot{y}(t) = Ay(t) + k(t)$, where $h = \frac{1}{21}$ and

$$A = \frac{1}{h^2}\begin{pmatrix} -2 & 1 & 0 & \cdots & \cdots & 0 \\ 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & -2 & 1 \\ 0 & \cdots & \cdots & 0 & 1 & -2 \end{pmatrix}, \quad k(t) = \begin{pmatrix} 3e^t \sin(\pi x_1) \\ 3e^t \sin(\pi x_2) \\ 3e^t \sin(\pi x_3) \\ \vdots \\ 3e^t \sin(\pi x_{19}) \\ 3e^t \sin(\pi x_{20}) \end{pmatrix}$$

Since this problem will only analyze the semi-discrete system and not the full PDE, a reference solution is obtained using an accurate time-integrator (SciPy's solve_ivp) using an absolute tolerance of 10^{-15}. Let $R = 0.33$ and $S(y(t)) = \frac{1}{20}\sum_{i=1}^{20} y_i(t)$ in order to analyze the discrete average of the solution over the spatial domain at a time



(a)



(b)

Fig. 8 **a** Chosen value of R, true data $S(y(t))$, and true QoI for example in Sect. 4.5. **b** Chosen value of R, true data $S(y(t))$, and true QoI for example in Sect. 4.6

Table 14 Results of the different methods on the example in Sect. 4.5 using cG(1) with 40 elements

Method	t_c	t_{LL}	t_L	t_R	e_Q	η	ρ_{eff}	n_{adj}
Taylor series	0.5834	–	–	–	6.157e−05	6.151e−05	0.999	2
Secant	0.5834	–	0.575	0.6	6.157e−05	6.150e−05	0.999	6
Inverse quad.	0.5834	0.55	0.575	0.6	6.157e−05	6.150e−05	0.999	7

t. This library function also has the capability of tracking when specified events occur, which is used to obtain a reference for the true QoI,

$$t_t = 0.5834435609935992.$$

For this problem, the parameters in (31) are

$$v = \frac{1}{20}(1, 1, \dots, 1)^{\top}, \;\; f(y, t) = Ay + k(t), \;\; \nabla_y f(y, t) = A.$$

For (32), (33), and (37), set

$$\psi_1 = -\frac{1}{20}(1, 1, \dots, 1)^{\top}, \;\; \psi_2 = \frac{1}{20h^2}(-1, 0, \dots, 0, -1)^{\top}, \;\; \psi_3 = \frac{1}{20}(1, 1, \dots, 1)^{\top}.$$

The true solution and QoI are shown in Fig. 8a and the results when using cG(1) or Crank–Nicolson methods are shown in Tables 14 and 15 respectively. All methods are accurate using either numerical method. The two iterative methods require more adjoint problems to be solved than the Taylor series estimate without any noticeable increase in accuracy.

Table 15 Results of the different methods on the example in Sect. 4.5 using Crank–Nicolson with 21 nodes

Method	t_c	t_{LL}	t_L	t_R	e_Q	η	ρ_{eff}	n_{adj}
Taylor series	0.5830	–	–	–	4.457e−04	4.457e−04	1.000	2
Secant	0.5830	–	0.55	0.6	4.457e−04	4.456e−04	0.999	6
Inverse quad.	0.5830	0.5	0.55	0.6	4.457e−04	4.456e−04	0.999	7

4.6 Two body problem

We consider the two body problem

$$\left.\begin{aligned} \dot{y}_1 &= y_3, \\ \dot{y}_2 &= y_4, \\ \dot{y}_3 &= \frac{-y_1}{(y_1^2+y_2^2)^{3/2}}, \\ \dot{y}_4 &= \frac{-y_2}{(y_1^2+y_2^2)^{3/2}}, \end{aligned}\right\} \quad t \in (0, 1.5], \quad y(0) = (0.4, 0, 0, 2.0)^\top,$$

which models a small body orbiting a much larger body in two dimensions. Here y_1, y_2 are the spatial coordinates of the orbiting body relative to the larger body, and y_3, y_4 are the respective velocities. The initial conditions are chosen so that the analytic solution is [22]

$$y = \left(\cos(\tau) - 0.6, 0.8\sin(\tau), \frac{-\sin(\tau)}{1-0.6\cos(\tau)}, \frac{0.8\cos(\tau)}{1-0.6\cos(\tau)}\right)^\top,$$

where τ solves $\tau - 0.6\sin(\tau) = t$. Let $R = 0$ and $S(y(t)) = y_1(t) + y_2(t)$. The true QoI can be found exactly:

$$t_t = Q(y) = \cos^{-1}((15 - 16\sqrt{2})/41) - 0.6\sin\left(\cos^{-1}((15 - 16\sqrt{2})/41)\right).$$

The values needed to compute (31) are

$$v = (1, 1, 0, 0)^\top, \; f(y, t) = \left(y_3, y_4, \frac{-y_1}{(y_1^2+y_2^2)^{3/2}}, \frac{-y_2}{(y_1^2+y_2^2)^{3/2}}\right)^\top,$$

and

$$\nabla_y f(y, t) = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \frac{2y_1^2 - y_2}{(y_1^2+y_2^2)^{5/2}} & \frac{3y_1 y_2}{(y_1^2+y_2^2)^{5/2}} & 0 & 0 \\ \frac{3y_1 y_2}{(y_1^2+y_2^2)^{5/2}} & \frac{2y_1^2 - y_2}{(y_1^2+y_2^2)^{5/2}} & 0 & 0 \end{pmatrix}.$$

Table 16 Results of the different methods on the example in Sect. 4.6 using cG(1) with 40 elements

Method	t_c	t_{LL}	t_L	t_R	e_Q	η	ρ_{eff}	n_{adj}
Taylor series	1.1601	–	–	–	8.262e−03	8.287e−03	1.003	2
Secant method	1.1601	–	1.125	1.1625	8.262e−03	8.287e−03	1.003	5
Inverse quad.	1.1601	1.0875	1.125	1.1625	8.262e−03	8.287e−03	1.003	6

Table 17 Results of the different methods on the example in Sect. 4.6 using Crank–Nicolson with 21 nodes

Method	t_c	t_{LL}	t_L	t_R	e_Q	η	ρ_{eff}	n_{adj}
Taylor series	1.2091	–	–	–	−4.068e−02	−4.078e−02	1.002	2
Secant	1.2091	–	1.2	1.275	−4.068e−02	−4.077e−02	1.002	5
Inverse quad.	1.2091	1.125	1.2	1.275	−4.068e−02	−4.077e−02	1.002	6

For (32), (33), and (37), the data needed are

$$\psi_1 = (-1, -1, 0, 0)^\top, \ \ \psi_2 = (0, 0, 1, 1)^\top \ \ \psi_3 = (1, 1, 0, 0)^\top.$$

The true data $S(y(t))$ and QoI are shown in Fig. 8b and the results using the cG(1) and Crank–Nicolson method appear in Tables 16 and 17 respectively. All methods have larger error than in other examples so far, probably as a result of the non-linear nature of (4.6). However the error estimates are still accurate using either numerical method; each with an effectivity ratio close to one.

4.7 Logistic equation

Consider the Logistic equation

$$\dot{y} = ky\left(1 - \frac{y}{K}\right), \ t \in (0, 20], \ \ y(0) = \frac{1}{2}, \tag{57}$$

where $k = 0.25$ and $K = 1$. The analytic solution is,

$$y(t) = \frac{K\ y(0)}{y(0) + (K - y(0))e^{-kt}} = \frac{1}{1 + e^{-0.25t}}. \tag{58}$$

Let $S(y(t)) = y(t)$ and consider several threshold values, $R \in \{0.55, 0.8, 0.9, 0.94, 0.98, 0.99, 0.995\}$. The values needed for (31) are

$$v = 1, \ \ f(y, t) = ky(1 - \frac{y}{k}), \ \ \nabla_y f(y, t) = k - \frac{2k}{K}y,$$

Fig. 9 True values of functional and QoI for example in Sect. 4.7, when $R = 0.94$

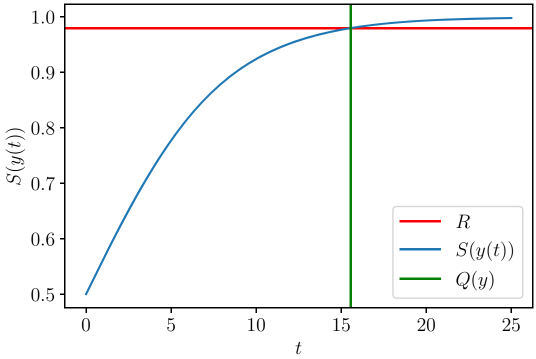


Table 18 Error in QoI and effectivity ratio of the different methods for varying values of R on example in Sect. 4.7 using cG(1) with 5 elements

	R = 0.55	R = 0.8	R = 0.9	R = 0.94	R = 0.98	R = 0.99	R = 0.995
e_Q	−0.090	−0.117	−0.166	0.194	0.829	0.610	1.513
Taylor series	1.001	1.021	1.041	0.957	0.902	0.919	0.830
Secant	0.999	0.987	0.977	1.023	1.007	1.011	1.005
Inverse quad.	0.999	0.987	0.977	1.023	1.007	1.011	1.005

so the data needed for (32), (33), and (37) are

$$\psi_1 = -1, \ \psi_2 = k - \frac{2k}{K} R, \ \psi_3 = 1.$$

The numerical solution is computed using the cG(1) method with five elements. Figure 9 shows the true functional and QoI for a chosen threshold value. Table 18 shows the true error in the QoI and the effectivity ratio for each method as the threshold value increases. As the error in the QoI increases, the Taylor series method loses accuracy, presumably since the remainder terms are no longer negligible, despite the fact the second derivatives with respect to t are small. However, the iterative methods are accurate even when the true error is large.

4.8 Conclusions for deterministic examples

Both the Taylor series and the root-finding approaches provide accurate error estimates in most cases. Some limitations of these methods were revealed in Sects. 4.4.1 and 4.4.2. The poor results in Sect. 4.4.1 are caused by the use of a low accuracy solution and the fact that computed QoI was closer to the second time the threshold value was crossed than the first. In Sect. 4.4.2, specifically Tables 11, 12 and 13, we observed that the issue that arose in Sect. 4.4.1 can be remedied by using a numerical solution that is more accurate near the QoI. Although another issue is revealed in the final column of Table 13, where the Taylor series approach gives poor results even though the numerical solution is quite accurate. In that instance the poor result is due to the assumption that terms involving the second derivative of $S(t)$ with respect to t can be neglected. The example in Sect. 4.7 shows that the Taylor series approach may not be accurate if the error in the QoI is large, but the iterative methods are accurate provided the root finding technique finds the correct root.

5 Numerical examples for error in the CDF of the non-standard QoI

The techniques outlined in Sect. 3.6 are applied to some examples below. The error bound (43) relies on accurate error estimates for the non-standard QoI. In the numerical examples, the estimates for the error in each sample value had an effectivity ratio close to one.

5.1 Harmonic oscillator

Reconsider the harmonic oscillator from Sect. 4.4 this time with parameters k and m as random variables:

$$\begin{pmatrix} \dot{y_1}(t) \\ \dot{y_2}(t) \end{pmatrix} + \begin{pmatrix} 0 & -1 \\ k/m & 1/m \end{pmatrix} \begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix} = \begin{pmatrix} 0 \\ 50/m * \cos(10t) \end{pmatrix}, \; t \in (0, 2],$$

with initial conditions $(y_1(0), y_2(0)) = (5, 0)$. Let k have a normal distribution with mean 50 and a standard deviation of 5 and m be uniformly distributed over $[.125, .325]$. For the QoI, set $R = -1$ and $S(y(t)) = y_1(t)$. With $\varepsilon = 0.05$ in (43), the nominal CDF (42) is computed using the true solution given in [5] with 1000 samples. The numerical solution is obtained using cG(1) with 40 elements and the approximate CDF (41) is computed with $N = 100$ samples (see Fig. 10a). Both sources contribute to the error, with the sampling error being slightly more dominant (Fig. 11). The computed bound is indeed larger than the actual error in the distribution. Both the bound and the error peak near the inflection point of the CDF, with the error bound being about six times larger than the true error.

5.2 Lorenz system

Consider the Lorenz system (13), where we let one of the initial conditions be a random variable. More precisely, $y_1(0) = \theta$ is uniformly distributed over the interval $(0, 2]$. Again let $\sigma = 10$, $r = 28$, and $b = \frac{8}{3}$. For the QoI (2), set $R = 3$ and $S(y(t)) = y_1(t)$. A reference solution and QoI are obtained using an accurate time-integrator (SciPy's solve_ivp with event tracker) with an absolute tolerance of 10^{-15} and a relative tolerance of 10^{-8}. This time, the numerical solution is computed using the cG(1) method with 30 elements.

The values needed for Eq. (31) are

$$v = (1, 0, 0)^\top, \; f(y, t) = (\sigma(y_2 - y_1), \; ry_1 - y_2 - y_1 y_3, \; y_1 y_2 - b y_3)^\top,$$

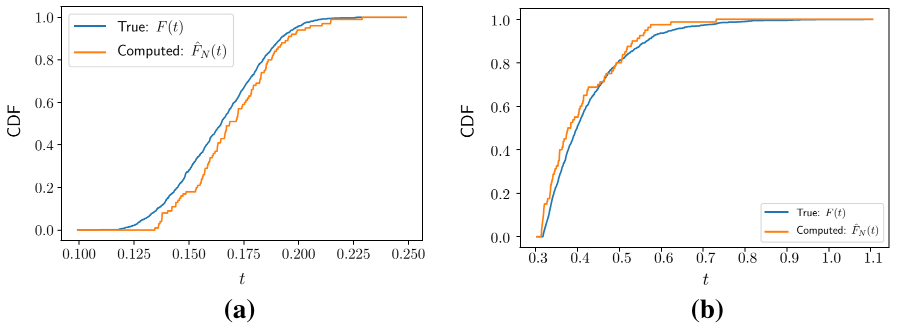


Fig. 10 **a** Nominal CDF using 1000 samples and computed CDF using 100 samples for example in Sect. 5.1. **b** Comparing nominal CDF using 1000 samples to computed CDF using 80 samples for example in Sect. 5.2

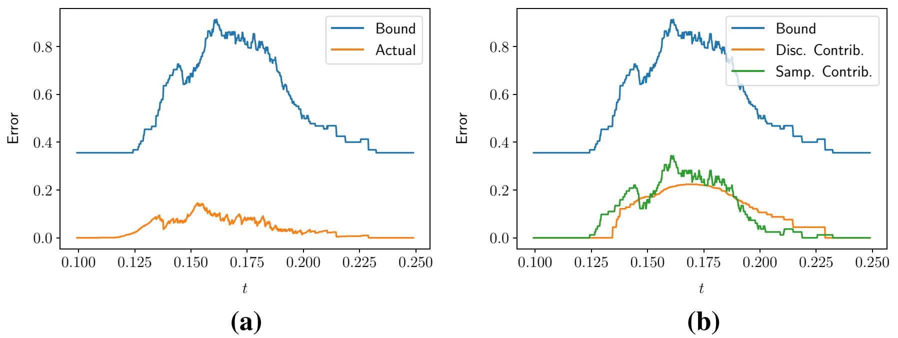


Fig. 11 Error bound for example in Sect. 5.1. **a** Comparing computed error bound (43) to true error for the problem in Sect. 5.1 when using 1000 samples for the nominal CDF and 100 samples for the numerical CDF. **b** Breaking the error bound into sampling and discretization contributions for the problem in Sect. 5.1 when using 100 samples. The sampling and discretization contributions are computed as the first and second terms of (43), respectively

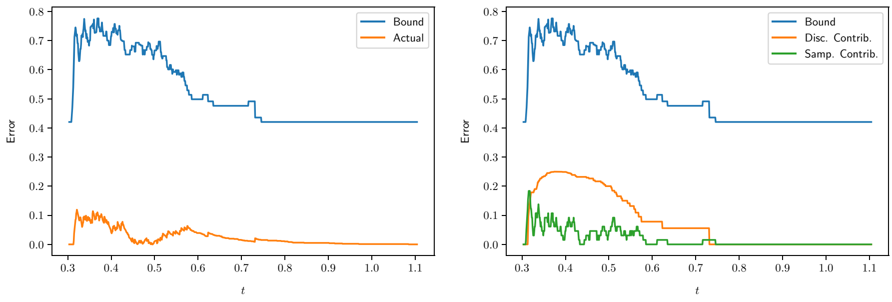
and

$$\nabla_y f(y, t) = \begin{pmatrix} \sigma & -\sigma & 0 \\ r - y_3 & -1 & -y_1 \\ -y_2 & y_1 & -b \end{pmatrix}.$$

hence, for (32), (33), and (37) the data are

$$\psi_1 = (-1, 0, 0)^\top, \psi_2 = (-\sigma, \sigma, 0)^\top, \psi_3 = (1, 0, 0)^\top.$$

The bound (43) is computed with $\varepsilon = 0.05$. The Fig. 10b compares the numerical CDF computed using 80 samples to the nominal CDF using 1000 samples. Figure 12 shows the discretization and sampling contributions to the calculated error bound. For this example, the discretization is the larger contributor to the error in the CDF, which is likely due to the chaotic nature of the system. As in Sect. 5.1 the error bound is roughly six times the true error at its peak.



(a) (b)

Fig. 12 **a** Comparing error bound (43) to true error for example in Sect. 5.2 when using 1000 samples for the nominal CDF and 80 samples for the numerical CDF. **b** Showing the sampling and discretization contributions to the error bound for the example in Sect. 5.2 when using 100 samples. The sampling contribution is computed as the first term of (43), while the second term gives the discretization contribution

6 Conclusions

We develop two different classes of accurate a posteriori error estimates for a QoI that cannot be expressed as a bounded functional of the solution, namely the first time when a given functional S of the solution achieves a specific value. The first method is based on Taylor's Theorem and is accurate whenever the numerical solution is sufficiently accurate and the curvature of the functional S is not too large. Moreover this method is cost effective, requiring the solution of only two adjoint problems. The second class of methods are based on standard root-finding techniques and are accurate provided the numerical solution is sufficiently accurate near the event of interest. These estimates however are more costly, requiring an adjoint solution per iteration of the root-finding algorithm. Both methods can be used as a basis for determining the discretization contribution to an error bound on a CDF of the functional when one or more of the parameters governing the system of differential equations are random variables.

Acknowledgements J. Chaudhry's and Z. Stevens's work is supported by the NSF-DMS 1720402. S. Tavener's and D. Estep's work is supported by NSF-DMS 1720473.

A Theorem 5

Theorem 5 *Numerical solutions obtained via the Crank–Nicolson finite difference scheme are nodally equivalent to solutions obtained using a cG(1) finite element method in which the integrals are evaluated with the trapezoidal rule.*

Proof The cG(1) formulation over a sub-interval (t_n, t_{n+1}), with the constant test function $v(t) = 1$ is

$$\int_{t_n}^{t_{n+1}} \dot{y}\, \mathrm{d}t = \int_{t_n}^{t_{n+1}} f(y, t)\, \mathrm{d}t. \tag{59}$$

Where by the fundamental theorem of calculus

$$\int_{t_n}^{t_{n+1}} \dot{y}\ \mathrm{d}t = y(t_{n+1}) - y(t_n). \tag{60}$$

Using the trapezoidal quadrature rule, we obtain

$$\int_{t_n}^{t_{n+1}} f(y,t)\ \mathrm{d}t \approx \frac{t_{n+1}-t_n}{2}\big(f(y(t_{n+1}),t_{n+1}) + f(y(t_n),t_n)\big). \tag{61}$$

Substituting (60) and (61) into (59) results in the Crank–Nicolson scheme. □

References

1. Ainsworth, M., Oden, T.: A Posteriori Error Estimation in Finite Element Analysis. John Wiley-Teubner, Chichester (2000)
2. Apostol Tom, M.: Calculus, vol. 1, 2nd edn. Wiley, Hoboken (1967)
3. Apostol Tom, M.: Calculus, vol. 2, 2nd edn. Wiley, Hoboken (1969)
4. Bangerth, W., Rannacher, R.: Adaptive Finite Element Methods for Differential Equations. Birkhauser Verlag, Basel (2003)
5. Barger, V., Olsson, M.: Classical Mechanics. A Modern Perspective. McGraw-Hill, New York (1973)
6. Barth, T.J.: A Posteriori Error Estimation and Mesh Adaptivity for Finite Volume and Finite Element Methods. Lecture Notes in Computational Science and Engineering, vol. 41. Springer, New York (2004)
7. Becker, R., Rannacher, R.: An optimal control approach to a posteriori error estimation in finite element methods. Acta Numer. **10**, 1–102 (2001)
8. Bouchard, B., Geiss, S., Gobet, E., et al.: First time to exit of a continuous Itô process: general moment estimates and L_1-convergence rate for discrete time approximations. Bernoulli **23**(3), 1631–1662 (2017)
9. Cao, Y., Petzold, L.: A posteriori error estimation and global error control for ordinary differential equations by the adjoint method. SIAM J. Sci. Comput. **26**(2), 359–374 (2004)
10. Carey, V., Estep, D., Tavener, S.: A posteriori analysis and adaptive error control for multiscale operator decomposition solution of elliptic systems I: triangular systems. SIAM J. Numer. Anal. **47**(1), 740–761 (2008)
11. Chaudhry, J.H., Estep, D., Ginting, V., Tavener, S.J.: A posteriori analysis for iterative solvers for nonautonomous evolution problems. SIAM/ASA J. Uncertain. Quantif. **3**(1), 434–459 (2015)
12. Chaudhry, J., Estep, D., Tavener, S.: A posteriori error analysis for Schwarz overlapping domain decomposition methods, July (2019). arXiv e-prints, arXiv:1907.01139
13. Chaudhry, J.H.: A posteriori analysis and efficient refinement strategies for the Poisson–Boltzmann equation. SIAM J. Sci. Comput. **40**(4), A2519–A2542 (2018)
14. Chaudhry, J.H., Burch, N., Estep, D.: Efficient distribution estimation and uncertainty quantification for elliptic problems on domains with stochastic boundaries. SIAM/ASA J. Uncertain. Quantif. **6**(3), 1127–1150 (2018)
15. Chaudhry, J.H., Collins, J.B., Shadid, J.N.: A posteriori error estimation for multi-stage Runge–Kutta IMEX schemes. Appl. Numer. Math. **117**, 36–49 (2017)
16. Chaudhry, J.H., Estep, D., Ginting, V., Shadid, J.N., Tavener, S.: A posteriori error analysis of IMEX multi-step time integration methods for advection–diffusion–reaction equations. Comput. Methods Appl. Mech. Eng. **285**, 730–751 (2015)
17. Chaudhry, J.H., Estep, D., Tavener, S., Carey, V., Sandelin, J.: A posteriori error analysis of two-stage computation methods with application to efficient discretization and the parareal algorithm. SIAM J. Numer. Anal. **54**(5), 2974–3002 (2016)

18. Chaudhry, J.H., Estep, D., Ginting, V., Tavener, S.J.: A posteriori analysis of an iterative multi-discretization method for reaction–diffusion systems. Comput. Methods Appl. Mech. Eng. **267**, 1–22 (2013)
19. Chaudhry, J.H., Shadid, J.N., Wildey, T.: A posteriori analysis of an IMEX entropy-viscosity formulation for hyperbolic conservation laws with dissipation. Appl. Numer. Math. **135**, 129–142 (2019)
20. Chaudry, J.H., Estep, D., Ginting, V., Tavener, S.: A posteriori analysis for iterative solvers for non-autonomous evolution problems. SIAM J. Uncertain. Quantif. **3**, 434–459 (2015)
21. Cliffe, K.A., Collis, J., Houston, P.: Goal-oriented a posteriori error estimation for the travel time functional in porous media flows. SIAM J. Sci. Comput. **37**(2), B127–B152 (2015)
22. Collins, J.B., Estep, D., Tavener, S.: A posteriori error analysis for finite element methods with projection operators as applied to explicit time integration techniques. BIT Numer. Math. **55**(4), 1017–1042 (2015)
23. Collins, J.B., Estep, D., Tavener, S.: A posteriori error estimation for the Lax–Wendroff finite difference scheme. J. Comput. Appl. Math. **263**, 299–311 (2014)
24. Delfour, M., Hager, W., Trochu, F.: Discontinuous Galerkin methods for ordinary differential equations. Math. Comput. **36**(154), 455–473 (1981)
25. Delfour, M.C., Dubeau, F.: Discontinuous polynomial approximations in the theory of one-step, hybrid and multistep methods for nonlinear ordinary differential equations. Math. Comput. **47**(175), 169–189 (1986)
26. Dzougoutov, A., Moon, K.-S., von Schwerin, E., Szepessy, A., Tempone, R.: Adaptive Monte Carlo algorithms for stopped diffusion. In: Engquist, B., Runborg, O., Lötstedt, P. (eds.) Multiscale Methods in Science and Engineering, pp. 59–88. Springer, Berlin (2005)
27. Epperson, J.F.: An Introduction to Numerical Methods and Analysis. Wiley, Hoboken (2007)
28. Eriksson, K., Estep, D., Hansbo, P., Johnson, C.: Introduction to adaptive methods for differential equations. Acta Numerica. **3**, 105–158 (1995)
29. Eriksson, K., Johnson, C., Logg, A.: Explicit time-stepping for stiff ODEs. SIAM J. Sci. Comput. **25**(4), 1142–1157 (2004)
30. Estep, D.: A posteriori error bounds and global error control for approximation of ordinary differential equations. SIAM J. Numer. Anal. **32**(1), 1–48 (1995)
31. Estep, D.: A short course on duality, adjoint operators, Green's functions, and a posteriori error analysis. Unpublished (2004)
32. Estep, D.: Error estimates for multiscale operator decomposition for multiphysics models. In: Fish, J. (ed.) Multiscale Methods: Bridging the Scales in Science and Engineering. Oxford University Press, Oxford (2009)
33. 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)
34. 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)
35. Estep, D., Larson, M., Williams, R.: Estimating the error of numerical solutions of systems of reaction–diffusion equations, p. 07. Memoirs of the American Mathematical Society, Providence (2000)
36. Estep, D., Målqvist, A., Tavener, S.: Nonparametric density estimation for randomly perturbed elliptic problems I: computational methods, a posteriori analysis, and adaptive error control. SIAM J. Sci. Comput. **31**(4), 2935–2959 (2009)
37. Estep, D., Målqvist, A., Tavener, S.: Nonparametric density estimation for randomly perturbed elliptic problems II: applications and adaptive modeling. Int. J. Numer. Methods Eng. **80**(6–7), 846–867 (2009)
38. Gautschi, W.: Numerical Analysis. Birkhäuser, Basel (2011)
39. Gobet, E.: Euler schemes and half-space approximation for the simulation of diffusion in a domain. ESAIM: Probab. Stat. **5**, 261–297 (2001)
40. Johansson, A., Chaudhry, J.H., Carey, V., Estep, D., Ginting, V., Larson, M., Tavener, S.J.: Adaptive finite element solution of multiscale PDE–ODE systems. Comput. Methods Appl. Mech. Eng. **287**, 150–171 (2015)
41. Logg, A.: Multi-adaptive time integration. Appl. Numer. Math. **48**(3–4), 339–354 (2004)
42. Rudin, W.: Principles of Mathematical Analysis, 3rd edn. McGraw-Hill, New York (1976)
43. Serfling, R.J.: Approximation Theorems of Mathematical Statistics. Wiley, Hoboken (1980)