Journal of Computational and Applied Mathematics 439 (2024) 115629

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

Avoiding order reduction phenomenon for general linear methods when integrating linear problems with time dependent boundary values

I. Alonso-Mallo [a,*,1], N. Reguera [b,1]

[a] IMUVA, Departamento de Matemática Aplicada, Universidad de Valladolid, Paseo de Belén, 7, 47011 Valladolid, Spain
[b] IMUVA, Departamento de Matemáticas y Computación, Escuela Politécnica Superior, Universidad de Burgos, Avda. Cantabria, 09006 Burgos, Spain

ARTICLE INFO

Article history:
Received 7 June 2023
Received in revised form 12 September 2023

Keywords:
Order reduction
General linear methods
Initial boundary value problems

ABSTRACT

When applied to stiff problems, the effective order of convergence of general linear methods is governed by their stage order, which is less than or equal to the classical order of the method. This produces an order reduction phenomenon, present in all general linear methods except those with high stage order, in a manner similar to that observed in other time integrators with internal stages.

In this paper, we investigate the order reduction which arises when general linear methods are used as time integrators when using the method of lines for solving numerically initial boundary value problems with time dependent boundary values.

We propose a technique, based on making an appropriate choice of the boundary values for the internal stages, with which it is possible to recover one unit of order, as we prove in this work. As expected, this implies a considerable improvement for the general linear methods suffering order reduction. Moreover, numerical experiments show that the improvement is not only in these cases, but that, even when the order reduction is not expected, the size of the errors is drastically reduced by using the technique proposed in this paper.

1. Introduction

General linear methods (GLMs), introduced by Burrage and Butcher in [1], are a class of multistep-multistage methods used for the time integration of systems of ordinary differential equations [2–7]. Many of the time integrators most commonly used in practice for numerically solving ordinary differential equations are particular cases of GLM, including Runge–Kutta methods and linear multistep methods (see e.g. [8,9]).

When a GLM is used for the time discretization of a stiff system of ordinary differential equations, the order observed in practice can be lower than the classical order of the GLM. This order reduction phenomenon is ubiquitous, and it has been observed for many schemes with internal stages [10–13], and also for exponential schemes [14–16].

* Corresponding author.
E-mail addresses: isaias@mac.uva.es (I. Alonso-Mallo), nreguera@ubu.es (N. Reguera).

[1] This research has been supported by project PGC2018-101443-B-100, Ministerio de Ciencia, Innovación y Universidades (Spain) and by Junta de Castilla y León (Grant numbers VA169P20 and VA193P20).

https://doi.org/10.1016/j.cam.2023.115629

The order reduction suffered by GLMs has been studied in [9] for the particular case when they are applied to Prothero–Robinson test problems [17]

$$\begin{aligned} u'(t) &= \lambda(u(t)-\phi(t))+\phi'(t), \qquad t\in[t_0,T],\\ u(t_0) &= \phi(t_0). \end{aligned}$$

More precisely, in [9], it is shown that the effective global order of a GLM is equal to its stage order q and, for GLMs with Runge–Kutta stability which are A(0)-stable, it is equal to $q+1$ when the stability function $R(z)$ of the underlying Runge–Kutta method is such that $R(+\infty)\neq 1$. Therefore, for GLMs with these stability properties, the order reduction phenomenon is observed when $q+1$ is less than the classical order p, and, for general GLMs, order reduction is observed if $q<p$. It follows that only with GLMs with high stage order, i.e., $q=p$ or $q=p-1$, the order reduction phenomenon will not appear. Consequently, recent literature on general linear methods applied to stiff problems has dealt with methods that have high stage order, such as Diagonally Implicit Multistage Integration methods (DIMSIMs) [18–20] and also implicit-explicit methods (IMEX) [21].

An important case of very stiff high dimensional systems of ordinary differential equations arises when a time evolution partial differential equation is discretized in space with the idea of using the method of lines to obtain a full discretization (see e.g. [22,23], where GLMs are used for the time integration). Although the order reduction is not observed in the special case when the solution and its derivatives vanish sufficiently at the boundary of the spatial domain where the solutions are defined (see [24] for the case in which Runge–Kutta methods are used as integrators in time), order reduction arises in the more general and interesting case of time dependent boundary conditions. Due to the importance of this phenomenon, we can find in the literature the proposal of different techniques that allow to avoid the order reduction, when a initial boundary value problem (IBVP) is full discretized in time and space, for the case of some schemes with internal stages, such as Runge–Kutta [10,12], Rosenbrock methods [11], Spectral-fractional step Runge–Kutta [12], as well as for exponential splitting schemes [14–16].

The aim of this paper is, on the one hand, to study the order reduction that GLMs suffer when they are used to solve well posed linear abstract IBVP (see [24,25] for a detailed description of this type of problems). Since it is assumed that the differential operator of the problem is the infinitesimal generator of a C_0-semigroup when the boundary values vanish, both hyperbolic and parabolic cases are included. We prove that the effective local order is $q+1$; then the effective global order is also $q+1$ when the stability function $R(z)$ of the underlying Runge–Kutta method is such that $R(+\infty)\neq 1$, similarly to the conclusions in [9] for Prothero–Robinson test problems. The numerical experiments that we show in this paper confirm this results.

On the other hand, we also prove that, in the general case of time dependent boundary values, a unit of order can be recovered by an appropriate choice of the boundary values of the GLM stages. In this way, with the technique we propose, we prove that the local order is $\min(q+2,p+1)$ instead of order $q+1$ that would be obtained with a standard implementation. Then, for GLMs that present order reduction, one unit of order is gained, with the improvement that this entails. But, as we will shown in the numerical experiments, even in the particular case of GLMs for which no order reduction is expected, the size of the errors is much smaller with our technique, also providing an improvement to the standard implementation. Furthermore, we note that the computational cost of the technique we propose is negligible compared to that of the method itself.

The organization of the paper is as follows. Section 2 is devoted to present in an abstract way the IBVP which we want to solve numerically, and the GLMs that we will use for this. In Section 3, we prove the order reduction that GLMs suffer when using a standard implementation We also propose the boundary values for the stages that must be used to avoid the order reduction at least in one unit. The numerical experiments, which confirm the previous results, are included in Section 4.

2. Discretization of an IBVP by means of a GLM

Let X and Y be two Banach spaces, $D(A)\subset X$ a dense subspace of X, and let $A: D(A)\subset X\to X$, $B: D(A)\subset X\to Y$, be two linear operators.

We consider an abstract non homogeneous linear IBVP

$$\begin{aligned} u'(t) &= Au(t)+f(t),\ t\in[0,T],\\ u(0) &= u_0,\\ Bu(t) &= g(t)\in Y. \end{aligned} \tag{1}$$

where A and B are operators for which we assume, as in [24,25], the following hypothesis:

(A1) The operator B is onto.
(A2) $\mathrm{Ker}(B)=D(A_0)$ is a dense subspace of X and $A_0=A|_{D(A_0)}$ is the infinitesimal generator of a C_0 semigroup $\{S(t)\}_{t\geq 0}$ in X with type ω.
(A3) Let If $z>\omega$, the steady state problem

$$Ax = zx,$$

$$Bx = v \in Y,$$

possesses a unique solution denoted by $x = K(z)v$ and there exists a constant C such that the linear operator $K(z) : Y \to D(A)$ satisfies, for any $z \geq \omega_0 > \omega$,

$$\|K(z)v\|_X \leq C\|v\|_Y.$$

(A4) The solution u in (1) satisfies $u(t) \in D(A)$ for $t \in [0, T]$ and it is smooth enough in time.

We will consider a GLM for the numerical time integration of the IBVP (1). For this, let N be a natural number and $k = T/N$ will denote the time step and $t_n = nk$, $n = 0, \ldots, N$. A GLM of s stages and r steps applied to (1) is given by the equations

$$U_i^{[n]} = k\sum_{j=1}^{s} a_{ij}[AU_j^{[n]} + f(t_n + c_j k)] + \sum_{j=1}^{r} u_{ij}u_j^{[n]}, \qquad i = 1, 2, \ldots, s, \tag{2}$$

$$u_i^{[n+1]} = k\sum_{j=1}^{s} b_{ij}[AU_j^{[n]} + f(t_n + c_j k)] + \sum_{j=1}^{r} v_{ij}u_j^{[n]}, \qquad i = 1, 2, \ldots, r, \tag{3}$$

where the stages $U_i^{[n]}$ are approximations to the values $u(t_n + c_i k)$ of order q of the solution,

$$U_i^{[n]} = u(t_n + c_i k) + O(k^{q+1}), \qquad i = 1, 2, \ldots, s$$

and $u_i^{[n]}$ are approximations of order p to certain linear combinations of derivatives of the solution at the point t_n,

$$u_i^{[n]} = \sum_{l=0}^{p} q_{il}k^l u^{(l)}(t_n) + O(k^{p+1}), \qquad i = 1, 2, \ldots, r. \tag{4}$$

The values p and q, which are the classical order and the stage order (see Definition 2.1) of the GLM, will be important throughout this work.

These equations of a GLM can be rewritten by means of the abscissa vector $\mathbf{c} = [c_1, \ldots, c_s]^T$, the vectors $\mathbf{q}_0 = [q_{1,0}, \ldots, q_{r,0}]^T$, $\mathbf{q}_1 = [q_{1,1}, \ldots, q_{r,1}]^T$, ..., $\mathbf{q}_p = [q_{1,p}, \ldots, q_{r,p}]^T$, and the matrices $\mathbf{A} = [a_{ij}] \in \mathbb{R}^{s\times s}$, $\mathbf{U} = [u_{ij}] \in \mathbb{R}^{s\times r}$, $\mathbf{B} = [b_{ij}] \in \mathbb{R}^{r\times s}$, $\mathbf{V} = [v_{ij}] \in \mathbb{R}^{r\times r}$ and $\mathbf{C} = \mathrm{diag}(c_1, \ldots, c_s) \in \mathbb{R}^{s\times s}$.

Next, we are going to mention some definitions and properties of GLMs that will be of interest for the study we will carry out next.

The GLM is strictly zero-stable when all the eigenvalues of the coefficient matrix $\mathbf{V}$ are inside of the unit circle an it has a unique eigenvalue on the unit circle. As a consequence of the strict zero-stability, we can assure that there exists

$$\lim_{n\to\infty} \mathbf{V}^n = \widetilde{\mathbf{V}} = \mathbf{u}\mathbf{v}^T$$

where $\mathbf{u}, \mathbf{v} \in \mathbb{R}^r$ are vectors satisfying $\widetilde{\mathbf{V}}\mathbf{u} = \mathbf{u}$, $\mathbf{v}^T\widetilde{\mathbf{V}} = \mathbf{v}^T$, and $\mathbf{v}^T\mathbf{u} = 1$ (cf. [8]).

We use the notation $\mathbf{e} = [1, \ldots, 1]^T$ and $\mathbf{c}^l = [c_1^l, \ldots, c_s^l]^T$ and we consider the following values:

$$\left.\begin{array}{rcl} \gamma_0 &=& \mathbf{e} - \mathbf{U}\mathbf{q}_0 \\ \gamma_k &=& \dfrac{\mathbf{c}^k}{k!} - \dfrac{\mathbf{A}\mathbf{c}^{k-1}}{(k-1)!} - \mathbf{U}\mathbf{q}_k,\ k = 1, \ldots, p. \end{array}\right\}$$

Then, a way to define the stage order is as follows.

Definition 2.1. The maximum number $q \leq p$ such that $\gamma_k = 0$, $k = 0, 1, \ldots, q$, is the stage order of the GLM method.

In the rest of this work, we assume the stage consistency conditions

$$\left.\begin{array}{lll} \gamma_0 = 0 & \Leftrightarrow \mathbf{U}\mathbf{q}_0 = \mathbf{e}, \\ \gamma_1 = 0 & \Leftrightarrow \mathbf{A}\mathbf{e} + \mathbf{U}\mathbf{q}_1 = \mathbf{c}, \end{array}\right\} \tag{5}$$

or, equivalently, we assume that $q \geq 1$.

Also, we will make the hypothesis that, for the GLM that is being used, the stage order $q < p$, so that the order reduction phenomenon will appear for the local error (see Theorem 3.1).

On the other hand, we also consider the values

$$\left.\begin{array}{rcl} \widehat{\gamma}_0 &=& \mathbf{q}_0 - \mathbf{V}\mathbf{q}_0 \\ \widehat{\gamma}_k &=& \displaystyle\sum_{l=0}^{k} \frac{\mathbf{q}_l}{(k-l)!} - \frac{\mathbf{B}\mathbf{c}^{k-1}}{(k-1)!} - \mathbf{V}\mathbf{q}_k,\ k = 1, \ldots, p. \end{array}\right\}$$

Table 1
Order conditions for GLMs of order $p \leq 4$.

Order p	order condition
0	$\widehat{\gamma}_0 = 0, \gamma_0 = 0$
1	$\widehat{\gamma}_1 = 0, \gamma_1 = 0$
2	$\widehat{\gamma}_2 = 0$
3	$\widehat{\gamma}_3 = 0$
	$\widetilde{\mathbf{V}}\mathbf{B}\gamma_2 = 0$ or $\mathbf{B}\gamma_2 = 0$
4	$\widehat{\gamma}_4 = 0$
	$\widetilde{\mathbf{V}}\mathbf{B}\gamma_3 = 0$
	$\widetilde{\mathbf{V}}\mathbf{B}\mathbf{A}\gamma_2 = 0$
	$\widetilde{\mathbf{V}}\mathbf{B}\mathbf{C}\gamma_2 = 0$

Then, the conditions

$$\widehat{\gamma}_k = 0,\ k = 0, 1, \ldots, p, \tag{6}$$

are necessary (but not sufficient) for the GLM to have order p. In the rest of this work, we also assume the consistency conditions

$$\widehat{\gamma}_0 = 0 \Leftrightarrow \mathbf{V}\mathbf{q}_0 = \mathbf{q}_0,$$
$$\widehat{\gamma}_1 = 0 \Leftrightarrow \mathbf{B}\mathbf{e} + \mathbf{V}\mathbf{q}_1 = \mathbf{q}_0 + \mathbf{q}_1,$$

which imply, along with (5), that the GLM has classical order $p \geq 1$.

Order conditions can be deduced in order to obtain a GLM of any desired order. See for example [8], where they are displayed up to order $p = 6$. We have listed in Table 1 these conditions up to the order $p = 4$.

To approximate the solution of (1) by means of a GLM, the first step is to solve the equation for the internal stages (2), which can be rewritten as

$$(\mathbf{I} \otimes I - \mathbf{A} \otimes kA)U^{[n]} = (\mathbf{U} \otimes I)u^{[n]} + k(\mathbf{A} \otimes I)F^{[n+c]} \tag{7}$$

where $U^{[n]} = [U_1^{[n]}, \ldots, U_s^{[n]}]^T$ is the vector of internal stages, $u^{[n]} = [u_1^{[n]}, \ldots, u_r^{[n]}]^T$ is the approximation at $t = t_n$ and $F^{[n+c]} = [f(t_n + c_1 k), \ldots, f(t_n + c_s k)]^T$.

When the internal stages have been computed, the approximation $u^{[n+1]}$ is obtained with (3), which can be rewritten as

$$u^{[n+1]} = (\mathbf{B} \otimes kA)U^{[n]} + k(\mathbf{B} \otimes I)F^{[n+c]} + (\mathbf{V} \otimes I)u^{[n]}. \tag{8}$$

To obtain a unique solution of the Eqs. (7), it is necessary to incorporate the boundary values of the internal stages. If we choose the boundary values $G^{[n]} = [G_1^{[n]}, \ldots, G_s^{[n]}]^T$, we obtain the following equations for the stages,

$$\left.\begin{array}{rcl} (\mathbf{I} \otimes I - \mathbf{A} \otimes kA)U^{[n]} & = & (\mathbf{U} \otimes I)u^{[n]} + k(\mathbf{A} \otimes I)F^{[n+c]}, \\ BU^{[n]} & = & G^{[n]}. \end{array}\right\} \tag{9}$$

In order to see that (9) has a unique solution, it suffices to consider the problems

$$\left.\begin{array}{rcl} (\mathbf{I} \otimes I - \mathbf{A} \otimes kA_0)U^{[n,0]} & = & (\mathbf{U} \otimes I)u^{[n]} + k(\mathbf{A} \otimes I)F^{[n+c]}, \\ BU^{[n,0]} & = & 0, \end{array}\right\} \tag{10}$$

and

$$\left.\begin{array}{rcl} (\mathbf{I} \otimes I - \mathbf{A} \otimes kA)U^{[n,b]} & = & 0, \\ BU^{[n,b]} & = & G^{[n]}. \end{array}\right\} \tag{11}$$

We endow the space X^s with the usual norm product. Then, the solvability of (10) and (11) comes from the following two lemmas (see [10]).

Lemma 2.2. *There exists a constant $C > 0$ such that the operator $(\mathbf{I} \otimes I - \mathbf{A} \otimes kA_0)^{-1} : X^s \to X^s$ is boundedly invertible and*

$$\|(\mathbf{I} \otimes I - \mathbf{A} \otimes kA_0) : D(A_0^s) \subset X^s \to X^s\| \leq C$$

for $k > 0$ small enough.

Lemma 2.3. *For $k > 0$ small enough and $W = [W_1, \ldots, W_s]^T \in Y^s$, the problem,*

$$\left.\begin{array}{rcl} (\mathbf{I} \otimes I - \mathbf{A} \otimes kA)V & = & 0, \\ BV & = & W. \end{array}\right\}$$

possesses a unique solution $V := K((kA)^{-1})W \in X^s$. *Moreover, there exists a constant* $C > 0$ *such that,*

$$\|K((kA)^{-1})W\| \leq C\|W\|.$$

3. Local error analysis of the time semidiscrete problem

In order to define the time semidiscrete local error, we consider $U^{[n+c]} = [u(t_n + c_1k), \ldots, u(t_n + c_sk)]^T$, and we use the following notation

$$\begin{aligned} G^{[n+c]} &= BU^{[n+c]} = [Bu(t_n + c_1k), \ldots, Bu(t_n + c_sk)]^T \\ &= [g(t_n + c_1k), \ldots, g(t_n + c_sk)]^T. \end{aligned} \tag{12}$$

These boundary values are the ones used for the stages in the standard method of lines.

Using the boundary values (12), we define $\overline{U}^{[n+c]} = [\overline{U}_1^{[n+c]}, \ldots, \overline{U}_s^{[n+c]}]^T$ as the solution of

$$\left.\begin{aligned} (\mathbf{I} \otimes I - \mathbf{A} \otimes kA)\overline{U}^{[n+c]} &= (\mathbf{U} \otimes I)\widetilde{u}(t_n) + k(\mathbf{A} \otimes I)F^{[n+c]}, \\ B\overline{U}^{[n+c]} &= G^{[n+c]}, \end{aligned}\right\} \tag{13}$$

where $\widetilde{u}(t_n) = [\widetilde{u}_1(t_n), \ldots, \widetilde{u}_r(t_n)]^T$, is given taking into account (4) with exact values of the solution and its derivatives, that is

$$\widetilde{u}_i(t_n) = \sum_{l=0}^{p} q_{il}k^l u^{(l)}(t_n),\ i = 1, \ldots, r. \tag{14}$$

Notice that we can consider $\widetilde{u}(t_n)$ as the optimal approximation of the exact solution at $t = t_n$ that we expect to obtain with the GLM.

Then, let $\overline{u}^{[n+1]}$ be given by

$$\overline{u}^{[n+1]} = (\mathbf{B} \otimes kA)\overline{U}^{[n+c]} + k(\mathbf{B} \otimes I)F^{[n+c]} + (\mathbf{V} \otimes I)\widetilde{u}(t_n), \tag{15}$$

Then, we define the time semidiscrete local error at t_{n+1} as follows.

$$\rho_{n+1} = \widetilde{u}(t_{n+1}) - \overline{u}^{[n+1]}.$$

Theorem 3.1. *Let* u *be the solution of* (1) *and let us assume the hypotheses (A1)–(A4). We consider the GLM given by the stages Eq.* (13)*, with the boundary values* $G^{[n+c]}$ *defined in* (12)*, and the final step* (15)*. We assume that* q*, the stage order of the GLM, is smaller that* p*, the classical order. Then the time semidiscrete local errors* $\rho_n = \widetilde{u}(t_n) - \overline{u}^{[n]}$, $1 \leq n \leq N$, *satisfy*

$$\|\rho_n\| \leq Ck^{q+1}, \quad \text{for } k > 0, \tag{16}$$

where the constant C *depends only on the derivatives of* u*, the GLM method and the differential operator* A*.*

Proof. Let $\varepsilon_{n,k}$ be the value satisfying

$$(\mathbf{I} \otimes I - \mathbf{A} \otimes kA)U^{[n+c]} = (\mathbf{U} \otimes I)\widetilde{u}(t_n) + k(\mathbf{A} \otimes I)F^{[n+c]} + \varepsilon_{n,k}. \tag{17}$$

Then,

$$\begin{aligned} \varepsilon_{n,k} &= U^{[n+c]} - (\mathbf{A} \otimes kI)(AU^{[n+c]} + F^{[n+c]}) - (\mathbf{U} \otimes I)\widetilde{u}(t_n) \\ &= U^{[n+c]} - (\mathbf{A} \otimes kI)U'^{[n+c]} - (\mathbf{U} \otimes I)\widetilde{u}(t_n) \end{aligned}$$

where we have used the notation $U'^{[n+c]} = [u'(t_n + c_1k), \ldots, u'(t_n + c_sk)]^T$.

Expanding $U^{[n+c]}$ and $U'^{[n+c]}$ into Taylor's series around t_n, we obtain

$$\begin{aligned} \varepsilon_{n,k} &= \gamma_0 + \sum_{l=0}^{p} \gamma_l k^l u^{(l)}(t_n) + O(k^{p+1}) \\ &= \sum_{l=q+1}^{p} \gamma_l k^l u^{(l)}(t_n) + O(k^{p+1}), \end{aligned} \tag{18}$$

where q is the stage order of the GLM and we have assumed that $q < p$.

Denoting $E_{n,k} = U^{[n+c]} - \overline{U}^{[n+c]}$, we subtract (17) from (13) and, taking into account that $U^{[n+c]}$ and $\overline{U}^{[n+c]}$ share the same boundary values, we deduce that

$$\left.\begin{aligned} (\mathbf{I} \otimes I - \mathbf{A} \otimes kA)E_{n,k} &= \varepsilon_{n,k} \\ BE_{n,k} &= 0 \end{aligned}\right\}$$

and, using Lemma 2.2,

$$E_{n,k} = (\mathbf{I} \otimes I - \mathbf{A} \otimes kA_0)^{-1} \varepsilon_{n,k}.$$

On the other hand,

$$\begin{aligned} u^{[n+1]} &= (\mathbf{V} \otimes I)\widetilde{u}(t_n) + (\mathbf{B} \otimes kA)U^{[n+c]} + k(\mathbf{B} \otimes I)F^{[n+c]} \\ &= (\mathbf{V} \otimes I)\widetilde{u}(t_n) + (\mathbf{B} \otimes kA)U'^{[n+c]} \\ &= \widetilde{u}(t_{n+1}) - \widetilde{u}(t_{n+1}) + (\mathbf{V} \otimes I)\widetilde{u}(t_n) + (\mathbf{B} \otimes kA)U'^{[n+c]} \\ &= \widetilde{u}(t_{n+1}) - \widehat{\gamma}_0 - \sum_{l=0}^{p} \widehat{\gamma}_l k^l u^{(l)}(t_n) + O(k^{p+1}) \\ &= \widetilde{u}(t_{n+1}) + O(k^{p+1}), \end{aligned} \tag{19}$$

where, as for the stages equation, we have expanded $\widetilde{u}(t_{n+1})$ and $U'^{[n+c]}$ into Taylor's series around t_n and we have used the order conditions (6).

Finally, we subtract (15) from (19),

$$\begin{aligned} \rho_{n+1} &= \widetilde{u}(t_{n+1}) - \overline{u}^{[n+1]} \\ &= (\mathbf{B} \otimes kA_0)(\mathbf{I} \otimes I - \mathbf{A} \otimes kA_0)^{-1} \varepsilon_{n,k} \end{aligned}$$

and, from (18) and using that $(\mathbf{B} \otimes kA_0)(\mathbf{I} \otimes I - \mathbf{A} \otimes kA_0)^{-1}$ are well-defined and bounded for $k > 0$, we deduce that

$$\rho_{n+1} = O(k^{q+1}).$$

We will now propose a technique to increase the order by one unit in cases where order reduction is present. The key is to choose properly the boundary values for the stages. For this, we consider the following vector

$$\begin{aligned} U^{[n+c,1]} &= (\mathbf{U} \otimes I)\widetilde{u}(t_n) + (\mathbf{A} \otimes kA)U^{[n+c]} + k(\mathbf{A} \otimes I)F^{[n+c]} \\ &= (\mathbf{U} \otimes I)\widetilde{u}(t_n) + (\mathbf{A} \otimes kA)U'^{[n+c]} \end{aligned} \tag{20}$$

and we denote

$$\begin{aligned} G^{[n+c,1]} = BU^{[n+c,1]} &= (\mathbf{U} \otimes I)B\widetilde{u}(t_n) + (\mathbf{A} \otimes kA)BU'^{[n+c]} \\ &= (\mathbf{U} \otimes I)\widetilde{G}(t_n) + (\mathbf{A} \otimes kA)G'^{[n+c]}, \end{aligned} \tag{21}$$

where

$$G'^{[n+c]} = [g'(t_n + c_1 k), \ldots, g'(t_n + c_s k)]^T,$$

and

$$\widetilde{G}(t_n) = [B\widetilde{u}_1(t_n), \ldots, B\widetilde{u}_r(t_n)]^T,$$

with

$$B\widetilde{u}_1(t_n) = \sum_{l=0}^{p} q_{il} k^l g^{(l)}(t_n),\ i = 1, \ldots, r.$$

Using these boundary values, we define the s-dimensional vector $\overline{U}^{[n+c,1]}$ as the solution of

$$\left.\begin{aligned} (\mathbf{I} \otimes I - \mathbf{A} \otimes kA)\overline{U}^{[n+c,1]} &= (\mathbf{U} \otimes I)\widetilde{u}(t_n) + k(\mathbf{A} \otimes I)F^{[n+c]} \\ B\overline{U}^{[n+c,1]} &= G^{[n+c,1]} \end{aligned}\right\} \tag{22}$$

and the second step of the GLM is given by

$$\overline{u}^{[n+1,1]} = (\mathbf{V} \otimes I)\widetilde{u}(t_n) + (\mathbf{B} \otimes kA)\overline{U}^{[n+c,1]} + k(\mathbf{B} \otimes I)F^{[n+c]}. \tag{23}$$

Then, we define the time semidiscrete local error at t_{n+1}

$$\rho_{n+1,1} = \widetilde{u}(t_{n+1}) - \overline{u}^{[n+1,1]}.$$

Theorem 3.2. *Let u be the solution of (1) and let us assume the hypotheses (A1)–(A4). We use (13) with the boundary values $G^{[n+c,1]}$ defined in (21) and the final step (23). Then the time semidiscrete local error $\rho_{n,1} = \widetilde{u}(t_n) - \overline{u}^{[n,1]}$, $1 \le n \le N$, satisfy*

$$\|\rho_{n,1}\| \le Ck^{\min(q+2,p+1)}, \quad \text{for } k > 0, \tag{24}$$

where q and p are, respectively, the stage order and the classical order of the GLM, and the constant C depends only on the derivatives of u, the GLM method and the differential operator A.

Proof. We subtract (20) from (17) and we deduce that

$$\varepsilon_{n,k} = U^{[n+c]} - U^{[n+c,1]}. \tag{25}$$

Let $\varepsilon_{n,k,1}$ be the value satisfying

$$U^{[n+c,1]} = (\mathbf{U}\otimes I)\widetilde{u}(t_n) + (\mathbf{A}\otimes kA)U^{[n+c,1]} + k(\mathbf{A}\otimes I)F^{[n+c]} + \varepsilon_{n,k,1}. \tag{26}$$

Now, we subtract (20) from (26) and we have

$$0 = (\mathbf{A}\otimes kA)(U^{[n+c,1]} - U^{[n+c]}) + \varepsilon_{n,k,1} \tag{27}$$

and, using (18),

$$\begin{aligned}\varepsilon_{n,k,1} &= (\mathbf{A}\otimes kA)(U^{[n+c]} - U^{[n+c,1]})\\ &= (\mathbf{A}\otimes kA)\varepsilon_{n,k}\\ &= \sum_{l=q+1}^{p} \mathbf{A}\gamma_l k^{l+1}Au^{(l)}(t_n) + O(k^{p+1})\\ &= O(k^{q+2}) + O(k^{p+1}).\end{aligned}$$

On the order hand, we define

$$\begin{aligned}u^{[n+1,1]} &= (\mathbf{V}\otimes I)\widetilde{u}(t_n) + (\mathbf{B}\otimes kA)U^{[n+c,1]} + k(\mathbf{B}\otimes I)F^{[n+c]}\\ &= (\mathbf{V}\otimes I)\widetilde{u}(t_n) + (\mathbf{B}\otimes kA)U^{[n+c]} + k(\mathbf{B}\otimes I)F^{[n+c]}\\ &\quad + (\mathbf{B}\otimes kA)(U^{[n+c,1]} - U^{[n+c]})\\ &= u^{[n+1]} + (\mathbf{B}\otimes kA)\varepsilon_{n,k} = \widetilde{u}(t_{n+1}) + O(k^{p+1}) + O(k^{q+2}).\end{aligned} \tag{28}$$

where we have used (19).

Now, we use the notation $E_{n,k,1} = U^{[n+c,1]} - \overline{U}^{[n+c,1]}$. Then, from (20), (25) and (26),

$$E_{n,k,1} = (\mathbf{I}\otimes I - \mathbf{A}\otimes kA_0)^{-1}\varepsilon_{n,k,1}.$$

Finally, from (28)

$$\begin{aligned}\rho_{n+1,1} &= \widetilde{u}(t_{n+1}) - u^{[n+1,1]} + u^{[n+1,1]} - \overline{u}^{[n+1,1]}\\ &= O(k^{p+1}) + O(k^{q+2}) + (\mathbf{B}\otimes kA)(U^{[n+c,1]} - \overline{U}^{[n+c,1]})\\ &= O(k^{p+1}) + O(k^{q+2}) + (\mathbf{B}\otimes kA_0)(\mathbf{I}\otimes I - \mathbf{A}\otimes kA_0)^{-1}\varepsilon_{n,k,1}\\ &= O(k^{p+1}) + O(k^{q+2}).\end{aligned}$$

4. Numerical experiments

In the previous section, Theorem 3.1 and 3.2 show that the technique we propose, based on using appropriate boundary conditions for the stages (20), allows us to increase the local order by one unit in the general case in which there is order reduction.

In this section we show numerical experiments that support these theoretical results. Moreover, we show that also in the case of GLMs for which order reduction is not expected, our technique provides an improvement because, also in this case, it allows to obtain much smaller errors than with a standard implementation.

In this section we are going to solve numerically the initial boundary value problem

$$\begin{aligned}u_t(x,t) &= u_{xx}(x,t) + f(x,t), \quad x\in[0,1]\\ u(x,0) &= u_0(x)\\ u(0,t) &= g_0(t), \quad u(1,t) = g_1(t)\end{aligned} \tag{29}$$

with $f(x,t) = e^t(x^2-1)$, $u_0(x) = x^2+1$, $g_0(t) = e^t$ and $g_1(t) = 2e^t$, whose exact solution is $u(x,t) = e^t(x^2+1)$.

This problem fits into the abstract IBVPs theory developed in [25], taking the spaces $X = L^2(0,1)$, $Y = \mathbb{R}^2$ and defining, for $u\in D(A) = H^2(0,1)$, $Au = u_{xx}$ and $Bu = [g_0(0), g_1(0)]^T$.

For the spatial discretization of (29) we will consider the second-order symmetric difference scheme. We take $J\in\mathbb{N}$, $J\geq 3$, and $h = 1/J$ and we consider the nodes $x_j = jh$, $j = 0,\dots,J-1$. We denote

$$A_{h,0} = \frac{1}{h^2}tridiag(1,-2,1), \qquad C_h g(t) = \frac{1}{h^2}[g_0(t),0,\dots,0,g_1(t)]^T.$$

Notice that, since the solution of (29) is a second degree polynomial in space, the error in space vanishes since we are using second-order symmetric differences. In this way, we can better study the error in time, which is what we are interested in.

Table 2
Local and global error when integrating problem (29) with GLM31 with a standard implementation. For the discretization in space $h = 4 \times 10^{-3}$ has been considered.

	$k = 4 \times 10^{-2}$	$k = 2 \times 10^{-2}$	$k = 10^{-2}$	$k = 5 \times 10^{-3}$	$k = 2.5 \times 10^{-3}$
L^2-local error	1.2931e−01	2.7007e−02	5.6031e−3	1.1553e−3	2.3611e−4
Order		2.26	2.27	2.28	2.29
L-global error	5.6650e−02	1.2217e−02	2.5743e−3	5.3406e−4	1.0924e−4
Order		2.21	2.25	2.27	2.29

Table 3
Local and global error when integrating problem (29) with GLM31 avoiding order reduction. For the discretization in space $h = 4 \times 10^{-3}$ has been considered.

	$k = 4 \times 10^{-2}$	$k = 2 \times 10^{-2}$	$k = 10^{-2}$	$k = 5 \times 10^{-3}$	$k = 2.5 \times 10^{-3}$
L^2-local error	1.5281e−03	1.5434e−04	1.6616e−5	1.9643e−6	2.5371e−7
Order		3.31	3.22	3.10	3.00
L^2-global error	9.8820e−04	1.2439e−04	1.5806e−5	2.0039e−6	2.5335e−7
Order		2.99	2.98	2.98	2.98

After applying this spatial discretization, we obtain the semidiscrete problem

$$U_h'(t) = A_{h,0}U_h(t) + f_h(t) + C_h g(t) \tag{30}$$

where $U_h(t) = [U_1(t), \ldots, U_{J-1}(t)]^T$ and $f_h(t) = [f(x_1, t), \ldots, f(x_{J-1}, t)]^T$.

We are going to carry out the numerical integration of (30) with two different GLM studying in each case the order in time that is observed.

First, we use the method GLM31, deduced in [9] that has classical order $p = 3$ and stage order $q = 1$. For this method, $\mathbf{V}$ is a rank one matrix such that $\mathbf{V} = \mathbf{e}_1\mathbf{v}^T$, where $e_1 = [1, 0, \ldots, 0]^T$ is the first vector of the canonical basis of $\mathbb{R}^{p+1}$ and $\mathbf{v}^T\mathbf{e}_1 = 1$. In this way, $\mathbf{V}^n = \mathbf{e}_1\mathbf{v}^T = \mathbf{V}$ and we deduce that $\widetilde{\mathbf{V}} = \mathbf{V}$. Then, the condition for order three is given by $\widetilde{\mathbf{V}}\mathbf{B}\gamma_2 = \mathbf{V}\mathbf{B}\gamma_2 = 0$. We include here the coefficients of the GLM because there is a small error in [9]:

$$\mathbf{A} = \begin{bmatrix} 0.5 & 0 & 0 \\ 0.6114715765267838 & 0.5 & 0 \\ 34.95200172589030 & -1.545721169870557 & 0.5 \end{bmatrix},$$

$$\mathbf{U} = \begin{bmatrix} 1 & -0.5 & -0.2509325024749301 & 0 \\ 1 & -0.6114715765267838 & -\frac{1}{4} & \frac{1}{3} \\ 1 & -32.90628055601974 & \frac{1}{2} & -\frac{1}{8} \end{bmatrix},$$

$$\mathbf{B} = \begin{bmatrix} 36.20771429329408 & -2.724800143643149 & 1.340965575511875 \\ 0 & 0 & 1 \\ 1 & -4 & 3 \\ 4 & -8 & 4 \end{bmatrix},$$

$$\mathbf{v} = \begin{bmatrix} 1 & -33.82387972516280 & 0.5214344963096989 & -0.1632161031338774 \end{bmatrix}^T$$

We can observe in Table 2 that, when we solve numerically the problem (29) with the spatial discretization previously explained and the scheme GLM31 for the time integration, with its standard implementation, local and global orders in time $q+1 = 2$ are obtained. This local order is expected from Theorem 3.1. Although in the general case the global order is expected to be 1, we get again the value 2 due to the summation by parts.

The numerical results observed in Table 2 show that order reduction is taking place when we use the standard implementation of GLM31. In order to avoid this order reduction effect, now we are going to carry out the same numerical experiment, but using the implementation (22) that we propose in this work for avoiding the order reduction phenomenon. In Table 3, we display the local and global orders in time, observing that now, both of them are $q+1+1 = 3$, so the order reduction is completely avoided. These numerical results are in agreement with Theorem 3.2. Note that, when we use the boundary values proposed in Theorem 3.2 to avoid order reduction, we get the added advantage that the errors are much smaller (compare the errors in Table 3 with those in Table 2). This is clearly seen in Fig. 1.

Now, we are going to consider the method GLM32 deduced in [9] with $p = 3$ and $q = 2$. The coefficients of such a method, are as follows

$$\mathbf{A} = \begin{bmatrix} 0.5 & 0 & 0 \\ 19.03844038247158 & 0.5 & 0 \\ -3.961809715877235 & -0.1911525347850310 & 0.5 \end{bmatrix},$$

$$\mathbf{U} = \begin{bmatrix} 1 & -0.5 & 0 & 0 \\ 1 & -19.03844038247158 & -0.125 & 0.333333333333333 \\ 1 & 4.652962250662266 & 0.09557626739251550 & -0.125 \end{bmatrix},$$

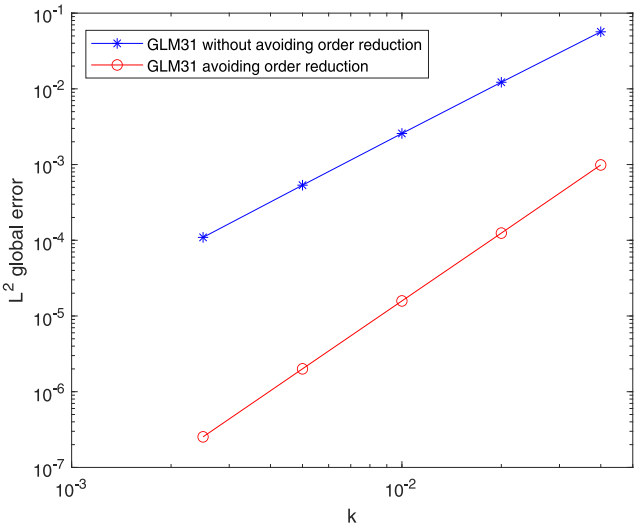

Fig. 1. Global errors when integrating problem (29) with GLM31.

Table 4
Local and global error when integrating problem (29) with GLM32 with a standard implementation. For the discretization in space $h = 4 \times 10^{-3}$ has been considered.

	$k = 4 \times 10^{-2}$	$k = 2 \times 10^{-2}$	$k = 10^{-2}$	$k = 5 \times 10^{-3}$	$k = 2.5 \times 10^{-3}$
L^2-local error	2.3028e−04	2.43442e−05	2.5611e−06	2.6760e−07	2.7773e−08
Order		3.24	3.25	3.26	3.27
L^2-global error	7.0327e−04	7.6149e−05	8.0456e−06	8.3339e−07	8.4674e−08
Order		3.21	3.24	3.27	3.30

Table 5
Local and global error when integrating problem (29) with GLM32 avoiding order reduction. For the discretization in space $h = 4 \times 10^{-3}$ has been considered.

	$k = 4 \times 10^{-2}$	$k = 2 \times 10^{-2}$	$k = 10^{-2}$	$k = 5 \times 10^{-3}$	$k = 2.5 \times 10^{-3}$
L^2-local error	1.5405e−05	1.0631e−06	7.0663e−08	4.5948e−09	2.9492e−10
Order		3.86	3.91	3.94	3.96
L^2-global error	1.4321e−04	7.9134e−06	4.3307e−07	2.3743e−08	1.3305e−09
Order		4.18	4.19	4.19	4.16

$$\mathbf{B} = \begin{bmatrix} -1.624407520530026 & -0.1715153809360207 & 0.5228985152929018 \\ 0 & 0 & 1 \\ 1 & -4 & 3 \\ 4 & -8 & 4 \end{bmatrix},$$

$$\mathbf{v} = \begin{bmatrix} 1 & 2.273024386173144 & 0.06285917517510861 & -0.07334316836278161 \end{bmatrix}^T$$

We can observe in Table 4 that when we integrate problem (29) considering the method GLM32 for the time integration, with its standard implementation, local and global order $q + 1 = 3$ are obtained. The global order is 3, and it seems that there is no order reduction. However, if there were no order reduction, local order $p + 1 = 4$ would be expected and this is not the case in Table 4. Therefore, we have used the implementation of the GLM32 that we propose in order to avoid order reduction and that is based in the boundary values of Theorem 3.2. The local and global errors displayed in Table 5 show that the local order predicted by Theorem 3.2, $q + 2 = 4$, is reached. In addition, the global error sizes are much smaller than when using the standard implementation, which justifies the use of the new boundary values even when q+1=p, in which case no order reduction is expected (see Fig. 2).

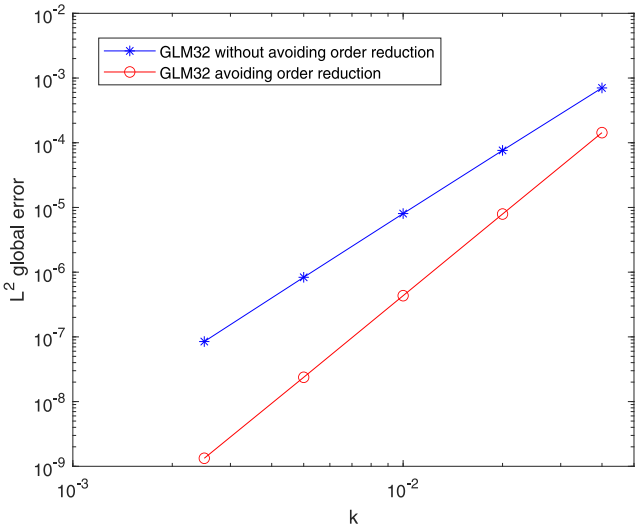

Fig. 2. Global errors when integrating problem (29) with GLM32.

Data availability

No data was used for the research described in the article.

References

[1] K. Burrage, J.C. Butcher, Nonlinear stability of a general class of differential equation methods, BIT 20 (1980) 185–203.
[2] K. Burrage, Parallel and Sequential Methods for Ordinary Differential Equations, The Clarendon Press, Oxford University Press, 1995.
[3] J.C. Butcher, The Numerical Analysis of Ordinary Differential Equations: Runge–Kutta and General Linear Methods, A Wiley–Interscience Publication, John Wiley & Sons Ltd., Chichester, 1987.
[4] J.C. Butcher, Numerical Methods for Ordinary Differential Equations, second ed., John Wiley & Sons Ltd., Chichester, 2008.
[5] J.C. Butcher, General linear methods, Acta Numer. 15 (2006) 157–256.
[6] E. Hairer, G. Wanner, Solving Ordinary Differential Equations. II. Stiff and Differential–Algebraic Problems, second ed., in: Springer Series in Computational Mathematics, vol. 14, Springer-Verlag, Berlin, 1996.
[7] Z. Jackiewicz, General Linear Methods for Ordinary Differential Equations, John Wiley & Sons, Inc., Hoboken, NJ, 2009.
[8] A. Cardone, Z. Jackiewicz, J.H. Verner, B. Welfert, Order conditions for general linear methods, J. Comput. App. Math. 290 (2015) 44–64.
[9] M. Braš, A. Cardone, Z. Jackiewicz, B. Welfert, Order reduction phenomenon for general linear methods, Appl. Numer. Math. 119 (2017) 94–114.
[10] I. Alonso-Mallo, Runge–Kutta methods without order reduction for linear initial boundary value problems, Numer. Math. 91 (2002) 577–603.
[11] I. Alonso-Mallo, B. Cano, Spectral/Rosenbrock discretizations without order reduction for linear parabolic problems appl, Numer. Math. 41 (2002) 247–268.
[12] I. Alonso-Mallo, B. Cano, J.C. Jorge, Spectral-fractional step Runge–Kutta discretizations for initial boundary value problems with time dependent boundary conditions, Math. Comput. 73 (2004) 1801–1825.
[13] B. Cano, N. Reguera, How to avoid order reduction when Lawson methods integrate nonlinear initial boundary value problems, BIT 62 (2022) 431–463.
[14] I. Alonso-Mallo, B. Cano, N. Reguera, Avoiding order reduction when integrating reaction–diffusion boundary value problems with exponential splitting methods, J. Comput. Appl. Math. 357 (2019) 228–250.
[15] L. Einkemmer, A. Ostermann, Overcoming order reduction in diffusion-reaction splitting. Part 1: Dirichlet boundary conditions, SIAM J. Sci. Comput. 37 (2015) 1577–1592.
[16] L. Einkemmer, A. Ostermann, Overcoming order reduction in diffusion-reaction splitting. Part 2: Oblique boundary conditions, SIAM J. Sci. Comput. 38 (2016) 3741–3757.
[17] A. Prothero, A. Robinson, On the stability and accuracy of one-step methods for solving stiff systems of ordinary differential equations, Math. Comp. 28 (1974) 145–162.
[18] J.C. Butcher, Diagonally-implicit multi-stage integration methods, Appl. Numer. Math. 11 (1993) 347–363.
[19] J.C. Butcher, Z. Jackiewicz, Diagonally implicit general linear methods for ordinary differential equations, BIT 33 (1993) 452–472.
[20] I. Th Famelis, Z. Jackiewicz, A new approach to the construction of DIMSIMs of high order and stage order, Appl. Numer. Math. 119 (2017) 79–93.
[21] Z. Jackiewicz, H. Mittelmann, Construction of IMEX DIMSIMs of high order and stage order, Appl. Numer. Math. 121 (2017) 234–248.
[22] A. Jaust, J. Schütz, General Linear Methods for Time-Dependent PDEs. Theory, Numerics and Applications of Hyperbolic Problems. II, in: Springer Proc. Math. Stat., vol. 237, Springer, Cham, 2018, pp. 59–70.
[23] P.E.J. Vos, C. Eskilsson, A. Bolis, S. Chun, R.M. Kirby, S.J. Sherwin, A generic framework for time-stepping partial differential equations (PDEs): general linear methods, object-oriented implementation and application to fluid problems, Int. J. Comput. Fluid Dyn. 25 (2011) 107–125.
[24] I. Alonso-Mallo, C. Palencia, Optimal orders of convergence for Runge–Kutta methods and linear, initial boundary value problems, Appl. Numer. Math. 44 (2003) 1–19.
[25] C. Palencia, I. Alonso-Mallo, Abstract initial–boundary value problems, Proc. Roy. Soc. Edinburgh Sect. A 124 (1994) 879–908.