

# MULTIRATE EXPONENTIAL ROSENBRACK METHODS \*

VU THAI LUAN<sup>†</sup>, RUJEKO CHINOMONA<sup>‡</sup>, AND DANIEL R. REYNOLDS<sup>‡</sup>

**Key words.** multirate time integration, exponential Rosenbrock methods, convergence analysis

**AMS subject classifications.** 65L05, 65L06, 65M20, 65L20

**Abstract.** In this paper we propose a novel class of methods for high order accurate integration of multirate systems of ordinary differential equation initial-value problems. The proposed methods construct multirate schemes by approximating the action of matrix  $\varphi$ -functions within explicit exponential Rosenbrock (ExpRB) methods, thereby called *Multirate Exponential Rosenbrock* (MERB) methods. They consist of the solution to a sequence of modified “fast” initial-value problems, that may themselves be approximated through subcycling any desired IVP solver. In addition to proving how to construct MERB methods from certain classes of ExpRB methods, we provide rigorous convergence analysis of these methods and derive efficient MERB schemes of orders two through six (the highest order ever constructed infinitesimal multirate methods). We then present numerical simulations to confirm these theoretical convergence rates, and to compare the efficiency of MERB methods against other recently-introduced high order multirate methods.

**1. Introduction.** In this paper, we consider numerical methods to perform highly accurate time integration for multirate systems of ordinary differential equation (ODE) initial-value problems (IVPs). The primary characteristic of these problems is that they are comprised of two or more components that on their own would evolve on significantly different time scales. Such problems may be written in the general additive form

$$(1.1) \quad u'(t) = F(t, u(t)) := F_f(t, u) + F_s(t, u), \quad t \in [t_0, T], \quad u(t_0) = u_0,$$

where  $F_f$  and  $F_s$  contain the “fast” and “slow” operators or variables, respectively. Typically, either due to stability or accuracy limitations the fast processes must be evolved with small step sizes; however the slow processes could allow much larger time steps. Such problems frequently arise in the simulation of “multiphysics” systems, wherein separate models are combined together to simulate complex physical phenomena [7]. While such problems may be treated using explicit, implicit, or mixed implicit-explicit time integration methods that evolve the full problem using a shared time step size, this treatment may prove inefficient, inaccurate or unstable, depending on which time scale is used to dictate this shared step size. Historically, scientific simulations have treated such problems using *ad hoc* operator splitting schemes where faster components are “subcycled” using smaller time steps than slower components. Schemes in this category include Lie–Trotter [20] and Strang–Marchuk [19, 27] techniques, that are first and second-order accurate, respectively. In recent years, however, methods with increasingly high orders of accuracy have been introduced. Our particular interest lies in methods allowing so-called “infinitesimal” formulations, wherein

---

\*Submitted to the editors DATE.

**Funding:** The first author is supported by NSF grant DMS–2012022. The second and third authors were supported in part by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Scientific Discovery through Advanced Computing (SciDAC) Program through the FASTMath Institute, under Lawrence Livermore National Laboratory subcontract B626484 and DOE award DE-SC0021354.

<sup>†</sup>Department of Mathematics and Statistics, Mississippi State University, Mississippi State, MS, 39762 ([luan@math.msstate.edu](mailto:luan@math.msstate.edu))

<sup>‡</sup>Department of Mathematics, Southern Methodist University, Dallas, TX 75275-0156 ([rchinomona@smu.edu](mailto:rchinomona@smu.edu), [reynolds@smu.edu](mailto:reynolds@smu.edu)).

the fast time scale is assumed to be solved exactly, typically through evolution of a sequence of modified fast IVPs,

$$v'(\tau) = F_f(\tau, v) + g(\tau), \quad \tau \in [\tau_0, \tau_f], \quad v(\tau_0) = v_0,$$

and where the forcing function  $g(\tau)$ , time interval  $[\tau_0, \tau_f]$ , and initial condition  $v_0$  are determined by the multirate method to incorporate information from the slow time scale. In practice, however, these fast IVPs are solved using any viable numerical method, typically with smaller step size than is used for the slow dynamics. While both the legacy Lie–Trotter and Strang–Marchuk schemes satisfy this description, each uses  $g(\tau) = 0$ , and only couple the time scales through the initial condition  $v_0$ . The first higher-order infinitesimal multirate methods were the *multirate infinitesimal step* (MIS) methods [25, 29], that allowed up to third order accuracy. These have been extended by numerous authors in recent years to support fourth and fifth orders of accuracy, as well as implicit or even mixed implicit-explicit treatment of the slow time scale [1, 3, 14, 23, 26].

Most higher-order ( $\geq 3$ ) infinitesimal methods, including MIS, relaxed MIS [26], extended MIS [1], multirate infinitesimal GARK [22, 23], and implicit-explicit multirate infinitesimal GARK [3], place no restrictions on the operators  $F_f$  and  $F_s$ . The corresponding order conditions for these methods are rooted in partitioned Runge–Kutta theory, to the end that the number of order conditions grows exponentially with the desired order of accuracy, to the effect that none of these methods have been proposed with order of accuracy greater than four.

In previous work, we presented an alternate approach for deriving infinitesimal multirate methods that was based on exponential Runge–Kutta (ExpRK) theory, named *multirate exponential Runge–Kutta* (MERK) methods [14]. A particular benefit of this theory is that exponential Runge–Kutta methods require fewer order conditions than partitioned Runge–Kutta methods; however, to leverage this theory, MERK methods require that the fast time scale operator is autonomous and that it depends *linearly* on the solution  $u$ , i.e., these consider the IVP

$$(1.2) \quad u'(t) = F(t, u(t)) := \mathcal{L}u + \mathcal{N}(t, u), \quad t \in [t_0, T], \quad u(t_0) = u_0,$$

where the “fast” and “slow” components are  $F_f(t, u) = \mathcal{L}u$  and  $F_s(t, u) = \mathcal{N}(t, u)$ , respectively. With this restriction in place, however, MERK methods have been proposed with orders of accuracy up to five.

In this work, we address the case of a non-autonomous and nonlinear fast time scale operator  $F_f(t, u)$  by proposing to use a dynamic linearization approach that updates the operators  $\mathcal{L}$  and  $\mathcal{N}$  within each time step. We then leverage this dynamic linearization approach through building multirate schemes from exponential Rosenbrock (ExpRB) methods. This new class of multirate schemes, called *Multirate Exponential Rosenbrock* (MERB) methods, approximates the action of matrix  $\varphi$ -functions within explicit ExpRB methods, and consist of solving a sequence of modified linear ODE-IVPs, which can be integrated using any desired ODE solvers. Moreover, we establish an elegant convergence theory for MERB methods, allowing us to determine a minimum order of accuracy for the numerical methods needed for solving the corresponding fast time scale IVPs. In addition to this theory, we generalize the coefficients for a number of high-order ExpRB methods and exploit their parallel stage structure to derive efficient multirate methods of very high order (including the first-ever infinitesimal multirate method of order six), with optimized numbers of modified fast IVPs. Our numerical experiments show that these new proposed

MERB schemes are uniformly the most efficient when considering slow function calls (this is particular of interest for multirate systems where the fast component is much less costly to compute than the slow component), and thus are very competitive in comparison with recently developed high order multirate methods such as MERK and MRI-GARK.

The remainder of this paper is organized as follows. We first present the structure of ExprB methods (Section 2.1). Then in Section 2.2 we interpret the corresponding ExprB internal stages and time step approximations as exact solutions to modified “fast” initial-value problems, thereby deriving MERB methods. In Section 2.3 we present rigorous convergence analysis for this family of newly-proposed methods. Then in Section 2.4 we construct specific multirate methods from this family for practical use, and discuss techniques for their numerical implementation in Section 2.5. In Section 3 we provide detailed numerical results to compare the performance of the proposed methods with the recent MERK methods of orders three through five, as well as with third and fourth order explicit MRI-GARK methods. Finally, we provide concluding remarks and discuss avenues for future research in Section 4.

## 2. Multirate Exponential Rosenbrock Methods.

**2.1. Exponential Rosenbrock schemes.** ExprB methods are constructed by linearizing the vector field  $F(t, u)$  at each step along the numerical solution  $(t_n, u_n)$ ,

$$(2.1) \quad u'(t) = F(t, u(t)) = J_n u(t) + V_n t + N_n(t, u(t))$$

with

$$(2.2) \quad J_n = \frac{\partial F}{\partial u}(t_n, u_n), \quad V_n = \frac{\partial F}{\partial t}(t_n, u_n), \quad N_n(t, u) = F(t, u) - J_n u - V_n t.$$

We note that if (1.1) is in fact autonomous, i.e.,  $u'(t) = F(u(t))$ , then this linearization simplifies since  $V_n = 0$  and  $N_n(t, u) = N_n(u) = F(u) - J_n u$ .

One can represent the exact solution to (2.1) at time  $t_{n+1} = t_n + H$  as in [12] by applying the variation-of-constants formula (a.k.a., Duhamel’s principle),

$$(2.3) \quad \begin{aligned} u(t_{n+1}) &= e^{H J_n} u(t_n) + \int_0^H e^{(H-\tau) J_n} \left( V_n(t_n + \tau) + N_n(t_n + \tau, u(t_n + \tau)) \right) d\tau \\ &= e^{H J_n} u(t_n) + H \varphi_1(H J_n) V_n t_n + H^2 \varphi_2(H J_n) V_n \\ &\quad + \int_0^H e^{(H-\tau) J_n} N_n(t_n + \tau, u(t_n + \tau)) d\tau, \end{aligned}$$

where  $\varphi_k(Z)$  ( $Z = H\mathcal{L}$ ) belong to the family of  $\varphi$ -functions given by

$$(2.4) \quad \varphi_k(Z) = \frac{1}{H^k} \int_0^H e^{(H-\tau) \frac{Z}{H}} \frac{\tau^{k-1}}{(k-1)!} d\tau, \quad k \geq 1.$$

Explicit ExprB methods approximate the integral in (2.3) by using a quadrature rule with nodes  $c_i$  in  $[0, 1]$  ( $i = 1, \dots, s$ ) ( $c_1 = 0$ ). Denoting the resulting approximations  $u_n \approx u(t_n)$  and  $U_{ni} \approx u(t_n + c_i H)$ , ExprB methods may be written as

$$(2.5) \quad \begin{aligned} U_{ni} &= u_n + c_i H \varphi_1(c_i H J_n) F(t_n, u_n) + c_i^2 H^2 \varphi_2(c_i H J_n) V_n + H \sum_{j=2}^{i-1} a_{ij}(H J_n) D_{nj}, \\ u_{n+1} &= u_n + H \varphi_1(H J_n) F(t_n, u_n) + H^2 \varphi_2(H J_n) V_n + H \sum_{i=2}^s b_i(H J_n) D_{ni}, \end{aligned}$$

where

$$(2.6) \quad D_{ni} = N_n(t_n + c_i H, U_{ni}) - N_n(t_n, u_n),$$

( $i = 2, \dots, s$ ) and where  $D_{n1} = 0$  [9, 12]. Here, the weights  $a_{ij}(HJ_n)$  and  $b_i(HJ_n)$  are usually chosen (by construction) as linear combinations of the  $\varphi_k(c_i HJ_n)$  and  $\varphi_k(HJ_n)$  functions given in (2.4), respectively. These unknown functions can be determined by solving order conditions, depending on the required order of accuracy.

*Remark 2.1.* (Order conditions) For later use, in Table 1 we recall the stiff order conditions for ExprB methods up to order 6 from [16]. We note that an ExprB method of order 6 only requires 7 conditions, which is much less than the 36 conditions needed for explicit Runge–Kutta or exponential Runge–Kutta methods of the same order. This is the advantage of the dynamic linearization approach (2.1), and can be understood by observing from (2.2) that

$$(2.7) \quad \frac{\partial N_n}{\partial u}(t_n, u_n) = 0 \quad \text{and} \quad \frac{\partial N_n}{\partial t}(t_n, u_n) = 0.$$

This property significantly simplifies the number of order conditions, particularly for higher-order schemes. A further consequence of (2.7) is that from (2.6) we have  $D_{ni} = \mathcal{O}(H^2)$ , meaning that ExprB methods are at least of order 2.

Table 1: Stiff order conditions for ExprB methods up to order 6 (from [16]). Here  $Z, K$ , and  $M$  denote arbitrary square matrices.

No.	Order condition	Order
1	$\sum_{i=2}^s b_i(Z) c_i^2 = 2\varphi_3(Z)$	3
2	$\sum_{i=2}^s b_i(Z) c_i^3 = 6\varphi_4(Z)$	4
3	$\sum_{i=2}^s b_i(Z) c_i^4 = 24\varphi_5(Z)$	5
4	$\sum_{i=2}^s b_i(Z) c_i K \left( \sum_{k=2}^{i-1} a_{ik}(Z) \frac{c_k^2}{2!} - c_i^3 \varphi_3(c_i Z) \right) = 0$	5
5	$\sum_{i=2}^s b_i(Z) c_i^5 = 120\varphi_6(Z)$	6
6	$\sum_{i=2}^s b_i(Z) c_i^2 M \left( \sum_{k=2}^{i-1} a_{ik}(Z) \frac{c_k^2}{2!} - c_i^3 \varphi_3(c_i Z) \right) = 0$	6
7	$\sum_{i=2}^s b_i(Z) c_i K \left( \sum_{k=2}^{i-1} a_{ik}(Z) \frac{c_k^3}{3!} - c_i^4 \varphi_4(c_i Z) \right) = 0$	6

**2.2. A multirate procedure for ExprB methods.** Inspired by our recent work [14], we now show how ExprB schemes can be interpreted as a class of multirate infinitesimal step-type methods. Namely, we construct modified differential equations whose exact solutions corresponding to the ExprB internal stages  $U_{ni}$  ( $i = 2, \dots, s$ ) and the final stage  $u_{n+1}$ .

LEMMA 2.2. *Consider an explicit exponential Rosenbrock scheme (2.5) where the weights  $a_{ij}(HJ_n)$  and  $b_i(HJ_n)$  can be written as linear combinations of  $\varphi_k$  functions,*

$$(2.8) \quad a_{ij}(HJ_n) = \sum_{k=1}^{\ell_{ij}} \alpha_{ij}^{(k)} \varphi_k(c_i HJ_n), \quad b_i(HJ_n) = \sum_{k=1}^{m_i} \beta_i^{(k)} \varphi_k(HJ_n),$$

and where  $\ell_{ij}$  and  $m_i$  are some positive integers. Then,  $U_{ni}$  and  $u_{n+1}$  are the exact solutions of the following (linear) modified differential equations

$$(2.9a) \quad v'_{ni}(\tau) = J_n v_{ni}(\tau) + p_{ni}(\tau), \quad v_{ni}(0) = u_n, \quad i = 2, \dots, s,$$

$$(2.9b) \quad v'_{n+1}(\tau) = J_n v_{n+1}(\tau) + q_n(\tau), \quad v_{n+1}(0) = u_n$$

at the times  $\tau = c_i H$  and  $\tau = H$ , respectively. Here  $p_{ni}(\tau)$  and  $q_n(\tau)$  are polynomials in  $\tau$  given by

$$(2.10a) \quad p_{ni}(\tau) = N_n(t_n, u_n) + (t_n + \tau)V_n + \sum_{j=2}^{i-1} \left( \sum_{k=1}^{\ell_{ij}} \frac{\alpha_{ij}^{(k)}}{c_i^k H^{k-1} (k-1)!} \tau^{k-1} \right) D_{nj},$$

$$(2.10b) \quad q_n(\tau) = N_n(t_n, u_n) + (t_n + \tau)V_n + \sum_{i=2}^s \left( \sum_{k=1}^{m_i} \frac{\beta_i^{(k)}}{H^{k-1} (k-1)!} \tau^{k-1} \right) D_{ni}.$$

*Proof.* The proof can be carried out in a very similar manner as in [14, Theorem 3.1]. Here, we only sketch the main idea. First, we insert the  $\varphi_k$  functions from (2.4) into (2.8) to get the integral representations of  $a_{ij}(HJ_n)$  and  $b_i(HJ_n)$ :

$$(2.11a) \quad a_{ij}(HJ_n) = \int_0^{c_i H} e^{(c_i H - \tau)J_n} \sum_{k=1}^{\ell_{ij}} \frac{\alpha_{ij}^{(k)}}{(c_i H)^k (k-1)!} \tau^{k-1} d\tau,$$

$$(2.11b) \quad b_i(HJ_n) = \int_0^H e^{(H - \tau)J_n} \sum_{k=1}^{m_i} \frac{\beta_i^{(k)}}{H^k (k-1)!} \tau^{k-1} d\tau.$$

Inserting these into (2.5) shows that the ExpRB stages and time step update may be written as

$$(2.12a) \quad U_{ni} = e^{c_i H J_n} u_n + \int_0^{c_i H} e^{(c_i H - \tau)J_n} p_{ni}(\tau) d\tau, \quad i = 2, \dots, s,$$

$$(2.12b) \quad u_{n+1} = e^{H J_n} u_n + \int_0^H e^{(H - \tau)J_n} q_n(\tau) d\tau,$$

which clearly show that  $U_{ni} = v_{ni}(c_i H)$  and  $u_{n+1} = v_{n+1}(H)$  by means of the variation-of-constants formula applied to (2.9a) and (2.9b), respectively.  $\square$

**MERB methods.** Starting from the initial value  $u_0 = u(t_0)$ , equations (2.9) from Lemma 2.2 suggest a multirate procedure to approximate the numerical solutions  $u_{n+1}$  ( $n = 0, 1, 2, \dots$ ) obtained by ExpRB methods. Specifically, one may integrate the slow process ( $V_n t + N_n(t, u)$ ) using a macro time step  $H$ , and integrate the fast process ( $J_n u$ ) using a micro time step  $h = H/m$  (where  $m > 1$  is an integer representing the time scale separation factor) via solving the “fast” ODEs (2.9a) on  $[0, c_i H]$  and (2.9b) on  $[0, H]$ . Let us denote the corresponding numerical solutions of these ODEs as  $\hat{U}_{ni}$  ( $\approx v_{ni}(c_i H) = U_{ni}$ ) and  $\hat{u}_{n+1}$  ( $\approx v_{n+1}(H) = u_{n+1}$ ). Then this multirate procedure consists in each step of solving (2.9)–(2.10) with the initial value  $\hat{u}_n$  ( $\hat{u}_0 = u_0$ ). Since we must linearize each step around the approximate solution  $\hat{u}_n$  instead of the true value  $u_n$ , we denote the approximations of  $J_n, V_n, N_n(t, u)$ , and  $D_{nj}$  appearing in polynomials (2.10) as

$$(2.13a) \quad \hat{J}_n = \frac{\partial F}{\partial u}(t_n, \hat{u}_n), \quad \hat{V}_n = \frac{\partial F}{\partial t}(t_n, \hat{u}_n), \quad \hat{N}_n(t, u) = F(t, u) - \hat{J}_n u - \hat{V}_n t,$$

$$(2.13b) \quad \hat{D}_{nj} = \hat{N}_n(t_n + c_j H, \hat{U}_{nj}) - \hat{N}_n(t_n, \hat{u}_n).$$

Thus, starting with  $\hat{u}_0 = u_0$ , for each time step  $t_n \rightarrow t_{n+1}$  we solve perturbed linear ODEs for  $i = 2, \dots, s$ :

$$(2.14) \quad y'_{ni}(\tau) = \hat{J}_n y_{ni}(\tau) + \hat{p}_{ni}(\tau), \quad \tau \in [0, c_i H], \quad y_{ni}(0) = \hat{u}_n,$$

with

$$(2.15) \quad \hat{p}_{ni}(\tau) = \hat{N}_n(t_n, \hat{u}_n) + (t_n + \tau)\hat{V}_n + \sum_{j=2}^{i-1} \left( \sum_{k=1}^{\ell_{ij}} \frac{\alpha_{ij}^{(k)}}{c_i^k H^{k-1} (k-1)!} \tau^{k-1} \right) \hat{D}_{nj},$$

to obtain

$$\hat{U}_{ni} \approx y_{ni}(c_i H) \approx v_{ni}(c_i H) = U_{ni}.$$

Then, using these approximations, we find

$$(2.16) \quad \hat{q}_n(\tau) = \hat{N}_n(t_n, \hat{u}_n) + (t_n + \tau)\hat{V}_n + \sum_{i=2}^s \left( \sum_{k=1}^{m_i} \frac{\beta_i^{(k)}}{H^{k-1} (k-1)!} \tau^{k-1} \right) \hat{D}_{ni}$$

and solve one additional linear ODE

$$(2.17) \quad y'_{n+1}(\tau) = \hat{J}_n y_{n+1}(\tau) + \hat{q}_n(\tau), \quad \tau \in [0, H], \quad y_{n+1}(0) = \hat{u}_n$$

to obtain the update

$$\hat{u}_{n+1} \approx y_{n+1}(H) \approx v_{n+1}(H) = u_{n+1}.$$

Since this process can be derived from ExprB schemes satisfying (2.8), we call the resulting methods (2.14)–(2.17) *Multirate Exponential Rosenbrock (MERB)* methods. Note that since  $\hat{U}_{n1}$  and  $y_{n1}(0)$  do not enter the MERB scheme, for the sake of completeness, one can define  $\hat{U}_{n1} = y_{n1}(0) = \hat{u}_n$ .

*Remark 2.3.* (A comparison with MERK methods). Based on their structure in (2.14)–(2.17), MERB methods have similar structure to MERK methods. Hence, they can retain MERK's interesting features, including very few evaluations of the costly slow components, and they do not require computing matrix functions as ExprB methods do. The main difference is that at each integration step MERB methods must update the linearization components  $\hat{J}_n$ ,  $\hat{V}_n$ ,  $\hat{N}_n$  and  $\hat{D}_{nj}$ . However, this increased cost may be balanced by the fact that, due to the property (2.7), high order MERB methods should require considerably fewer modified ODEs than MERK methods of the same order (see Section 2.4).

### 2.3. Convergence analysis of MERB methods.

**2.3.1. Analytical framework.** To analyze the convergence of MERB methods, we employ the abstract framework of strongly continuous semigroups (see, e.g., [5, 21]) on a Banach space  $X$ . Throughout this paper, we denote the norm in  $X$  by  $\|\cdot\|$ . Let

$$(2.18) \quad J = \frac{\partial F}{\partial u}(t, u)$$

be the Fréchet partial derivative of  $F$ . We make use of the following assumptions.

*Assumption 1.* The Jacobian (2.18) is the generator of a strongly continuous semigroup  $e^{tJ}$  in  $X$ . This implies that there exist constants  $C$  and  $\omega$  such that

$$(2.19) \quad \|e^{tJ}\| \leq C e^{\omega t}, \quad t \geq 0,$$

and consequently  $\varphi_k(HJ)$ ,  $a_{ij}(HJ)$  and  $b_i(HJ)$  are bounded operators.

*Assumption 2.* The solution  $u : [t_0, T] \rightarrow X$  of (1.1) is sufficiently smooth with derivatives in  $X$ , and  $F : [t_0, T] \times X \rightarrow X$  is sufficiently Fréchet differentiable in a strip along the exact solution to (1.1). All derivatives occurring are assumed to be uniformly bounded.

*Stability bound.* Since  $\hat{J}_n = \frac{\partial F}{\partial u}(t_n, \hat{u}_n)$  arising in MERB methods changes at every step, and  $\hat{J}_n \approx J_n$ , we also employ the following stability bound (for the discrete evolution operators on  $X$ ) of exponential Rosenbrock methods (see [9, Sect. 3.3]) to have

$$(2.20) \quad \left\| \prod_{j=0}^{n-k} e^{H\hat{J}_{n-j}} \right\| \leq C_S, \quad t_0 \leq t_k \leq t_n \leq T.$$

The importance of this bound is that the constant  $C_S$  is uniform in  $k$  and  $n$ , despite the fact that  $J_n$  varies from step to step.

**2.3.2. A global error representation of MERB methods.** Since MERB methods (2.14)–(2.17) result in a numerical solution  $\hat{u}_{n+1}$  which approximates the numerical solution  $u_{n+1}$  of ExprB methods (as denoted above) at time  $t_{n+1}$ , we will employ the local errors of ExprB methods to analyze the global error of MERB methods. Throughout the paper the following error notations will be used.

- *Global error notation for MERB methods.* We denote the global error at time  $t_{n+1}$  of a MERB method as

$$(2.21) \quad \hat{e}_{n+1} = \hat{u}_{n+1} - u(t_{n+1}).$$

- *Local error notation for ExprB methods.* We denote the local error at  $t_{n+1}$  of the base ExprB method as

$$(2.22) \quad \tilde{e}_{n+1} = \tilde{u}_{n+1} - u(t_{n+1})$$

Here,  $\tilde{u}_{n+1}$  is the numerical solution of the base ExprB method obtained after carrying out one step of (2.5) starting from the exact solution  $u(t_n)$  as the initial value, i.e.,

$$(2.23a) \quad \tilde{u}_{n+1} = e^{H\tilde{J}_n} u(t_n) + H\varphi_1(H\tilde{J}_n)\tilde{V}_n t_n + H^2\varphi_2(H\tilde{J}_n)\tilde{V}_n \\ + H \sum_{i=1}^s b_i(H\tilde{J}_n)\tilde{N}_n(t_n + c_i H, \tilde{U}_{ni}),$$

$$(2.23b) \quad \tilde{U}_{ni} = e^{c_i H\tilde{J}_n} u(t_n) + c_i H\varphi_1(c_i H\tilde{J}_n)\tilde{V}_n t_n + c_i^2 H^2\varphi_2(c_i H\tilde{J}_n)\tilde{V}_n \\ + H \sum_{j=1}^{i-1} a_{ij}(H\tilde{J}_n)\tilde{N}_n(t_n + c_j H, \tilde{U}_{nj}),$$

where

$$(2.24) \quad \tilde{J}_n = \frac{\partial F}{\partial u}(t_n, u(t_n)), \quad \tilde{V}_n = \frac{\partial F}{\partial t}(t_n, u(t_n)), \quad \tilde{N}_n(t, u) = F(t, u) - \tilde{J}_n u - \tilde{V}_n t.$$

Note that from Lemma 2.2, (2.23) is equivalent to one step of the MERB scheme starting from the exact initial value  $y_{n+1}(0) = u(t_n)$  (for which the solution of the

IVP (2.17) on  $[0, H]$  is “known” to be  $y_{n+1}(H) = \tilde{u}_{n+1}$ . Therefore, one can consider that  $\tilde{e}_{n+1}$  is also the local error of MERB methods.

• *Global error notation for approximation of the IVP (2.17)*. As  $\hat{u}_{n+1} \approx y_{n+1}(H)$  (the true solution of the ODE (2.17)), we denote the global error of an ODE solver used for integrating (2.17) on  $[0, H]$  as

$$(2.25) \quad \hat{\varepsilon}_{n+1} = \hat{u}_{n+1} - y_{n+1}(H).$$

• *Global error notation for approximation of the IVP (2.14)*. Similarly, since  $\hat{U}_{ni}$  is the numerical solution of (2.14) on  $[0, c_i H]$  obtained by an ODE solver, let us denote the global error of this approximation as

$$(2.26) \quad \hat{\varepsilon}_{ni} = \hat{U}_{ni} - y_{ni}(c_i H).$$

Note that by applying the variation-of-constants formula to (2.17) and using (2.11b),  $y_{n+1}(H)$  can be represented as

$$(2.27) \quad \begin{aligned} y_{n+1}(H) = & e^{H\hat{J}_n} \hat{u}_n + H\varphi_1(H\hat{J}_n)\hat{V}_n t_n + H^2\varphi_2(H\hat{J}_n)\hat{V}_n \\ & + H \sum_{i=1}^s b_i(H\hat{J}_n)\hat{N}_n(t_n + c_i H, \hat{U}_{ni}). \end{aligned}$$

In view of (2.21), (2.22), and (2.25), we can write

$$(2.28) \quad \hat{e}_{n+1} = \hat{u}_{n+1} - \tilde{u}_{n+1} + \tilde{e}_{n+1} = \hat{\varepsilon}_{n+1} + (y_{n+1}(H) - \tilde{u}_{n+1}) + \tilde{e}_{n+1},$$

i.e., the global error arising from the MERB scheme can be written as the sum of the global error of the ODE solver used for the IVP (2.17), the difference between the true solution to the IVP (2.17) and the numerical solution obtained by the base ExprB method (2.23), and the local error arising from this ExprB-based method.

To keep our presentation in a compact form, we introduce

$$(2.29a) \quad t_{ni} = t_n + c_i H,$$

$$(2.29b) \quad \hat{B}_n = \varphi_1(H\hat{J}_n)\hat{V}_n t_n + H\varphi_2(H\hat{J}_n)\hat{V}_n + \sum_{i=1}^s b_i(H\hat{J}_n)\hat{N}_n(t_{ni}, \hat{U}_{ni}),$$

$$(2.29c) \quad \tilde{B}_n = \varphi_1(H\tilde{J}_n)\tilde{V}_n t_n + H\varphi_2(H\tilde{J}_n)\tilde{V}_n + \sum_{i=1}^s b_i(H\tilde{J}_n)\tilde{N}_n(t_{ni}, \tilde{U}_{ni}).$$

Using (2.29), we now derive a full expansion of (2.28), which later tells us how the global error of MERB methods can be estimated by the sum of the propagated local errors of ExprB methods and the global errors of the ODE solvers used for integrating (2.14) and (2.17).

**THEOREM 2.4.** *The global error  $\hat{e}_{n+1}$  of MERB methods (2.14)–(2.17) at time  $t_{n+1}$  can be expressed as*

$$(2.30) \quad \begin{aligned} \hat{e}_{n+1} = & \underbrace{\left( \prod_{j=0}^n e^{H\hat{J}_{n-j}} - \prod_{j=0}^n e^{H\tilde{J}_{n-j}} \right) u_0}_{Error1} + \underbrace{\sum_{k=0}^n \left( \prod_{j=0}^{n-k-1} e^{H\hat{J}_{n-j}} \right) \tilde{e}_{k+1}}_{Error2} \\ & + \underbrace{\sum_{k=0}^n \left( \prod_{j=0}^{n-k-1} e^{H\hat{J}_{n-j}} \right) \hat{\varepsilon}_{k+1}}_{Error3} + \underbrace{H \sum_{k=0}^n \left[ \left( \prod_{j=0}^{n-k-1} e^{H\hat{J}_{n-j}} \right) \hat{B}_k - \left( \prod_{j=0}^{n-k-1} e^{H\tilde{J}_{n-j}} \right) \tilde{B}_k \right]}_{Error4}. \end{aligned}$$

*Proof.* In view of (2.28), we first study the difference  $(y_{n+1}(H) - \tilde{u}_{n+1})$ . Using (2.29b) and (2.25) (which implies  $\hat{u}_n = y_n(H) + \hat{\varepsilon}_n$ ), we have

$$(2.31) \quad y_{n+1}(H) = e^{H\hat{J}_n}\hat{u}_n + H\hat{B}_n = e^{H\hat{J}_n}y_n(H) + e^{H\hat{J}_n}\hat{\varepsilon}_n + H\hat{B}_n.$$

Solving this recurrence relation (with note that  $y_0(H) = u(t_0) = u_0$ ) gives

$$(2.32) \quad y_{n+1}(H) = \left( \prod_{j=0}^n e^{H\hat{J}_{n-j}} \right) u_0 + \sum_{k=0}^{n-1} \left( \prod_{j=0}^{n-k-1} e^{H\hat{J}_{n-j}} \right) \hat{\varepsilon}_{k+1} + H \sum_{k=0}^n \left( \prod_{j=0}^{n-k-1} e^{H\hat{J}_{n-j}} \right) \hat{B}_k.$$

Similarly, using (2.29c) and (2.22) (which implies  $u(t_n) = \tilde{u}_n - \tilde{\varepsilon}_n$ ), we can write  $\tilde{u}_{n+1}$  in (2.23a) as

$$(2.33) \quad \tilde{u}_{n+1} = e^{H\tilde{J}_n}u(t_n) + H\tilde{B}_n = e^{H\tilde{J}_n}\tilde{u}_n - e^{H\tilde{J}_n}\tilde{\varepsilon}_n + H\tilde{B}_n.$$

After solving this recurrence, we end up with

$$(2.34) \quad \tilde{u}_{n+1} = \left( \prod_{j=0}^n e^{H\tilde{J}_{n-j}} \right) u_0 - \sum_{k=0}^{n-1} \left( \prod_{j=0}^{n-k-1} e^{H\tilde{J}_{n-j}} \right) \tilde{\varepsilon}_{k+1} + H \sum_{k=0}^n \left( \prod_{j=0}^{n-k-1} e^{H\tilde{J}_{n-j}} \right) \tilde{B}_k.$$

Subtracting (2.34) from (2.32) gives  $(y_{n+1}(H) - \tilde{u}_{n+1})$  and inserting this into (2.28) proves (2.30).  $\square$

Next, in order to estimate the global error  $\hat{\varepsilon}_{n+1}$ , we prove some preliminary results.

### 2.3.3. Preliminary results and error bounds.

LEMMA 2.5. *The term Error4 in (2.30) can be further expressed as*

$$(2.35) \quad \text{Error4} = H \sum_{k=0}^n \left[ \left( \prod_{j=0}^{n-k-1} e^{H\hat{J}_{n-j}} - \prod_{j=0}^{n-k-1} e^{H\tilde{J}_{n-j}} \right) \hat{B}_k + \left( \prod_{j=0}^{n-k-1} e^{H\tilde{J}_{n-j}} \right) (\hat{B}_k - \tilde{B}_k) \right],$$

where

$$(2.36) \quad \begin{aligned} \hat{B}_k - \tilde{B}_k &= \sum_{j=1}^2 [(\varphi_j(H\hat{J}_k) - \varphi_j(H\tilde{J}_k))\tilde{V}_k + \varphi_j(H\hat{J}_k)(\hat{V}_k - \tilde{V}_k)] t_k^{2-j} H^{j-1} \\ &+ \sum_{i=1}^s (b_i(H\hat{J}_k) - b_i(H\tilde{J}_k)) \tilde{N}_k(t_{ki}, \tilde{U}_{ki}) \\ &+ \sum_{i=1}^s b_i(H\hat{J}_k) (\hat{N}_k(t_{ki}, \hat{U}_{ki}) - \tilde{N}_k(t_{ki}, \tilde{U}_{ki})). \end{aligned}$$

*Proof.* The derivation of (2.35) is straightforward by subtracting and adding the same term  $\prod_{j=0}^{n-k-1} e^{H\tilde{J}_{n-j}} \hat{B}_k$  within the sum  $\sum_{k=0}^n [\cdot]$  in Error4. Also, by subtracting (2.29c) from (2.29b), one can easily obtain (2.36).  $\square$

To estimate the difference in the nonlinear terms at each internal MERB and ExpRB stage,  $(\hat{N}_k(t_{ki}, \hat{U}_{ki}) - \tilde{N}_k(t_{ki}, \tilde{U}_{ki}))$  in (2.36), we first study the difference

$$(2.37) \quad \hat{E}_{ni} = \hat{U}_{ni} - \tilde{U}_{ni}.$$

Denoting

$$(2.38a) \quad \hat{A}_{ni} = c_i \varphi_1(c_i H \hat{J}_n) \hat{V}_n t_n + c_i^2 H \varphi_2(c_i H J_n) \hat{V}_n + \sum_{j=1}^{i-1} a_{ij}(H \hat{J}_n) \hat{N}_n(t_{nj}, \hat{U}_{nj}),$$

$$(2.38b) \quad \tilde{A}_{ni} = c_i \varphi_1(c_i H \tilde{J}_n) \tilde{V}_n t_n + c_i^2 H \varphi_2(c_i H J_n) \tilde{V}_n + \sum_{j=1}^{i-1} a_{ij}(H \tilde{J}_n) \tilde{N}_n(t_{nj}, \tilde{U}_{nj}),$$

we obtain the following result.

LEMMA 2.6. *The difference between  $\hat{U}_{ni}$  and  $\tilde{U}_{ni}$  can be expressed as*

$$(2.39) \quad \hat{E}_{ni} = \hat{\varepsilon}_{ni} + e^{c_i H \hat{J}_n} \hat{\varepsilon}_n + (e^{c_i H \hat{J}_n} - e^{c_i H \tilde{J}_n}) u(t_n) + H(\hat{A}_{ni} - \tilde{A}_{ni})$$

with

$$(2.40) \quad \begin{aligned} \hat{A}_{ni} - \tilde{A}_{ni} &= \sum_{\ell=1}^2 [(\varphi_\ell(c_i H \hat{J}_n) - \varphi_\ell(c_i H \tilde{J}_n)) \tilde{V}_n + \varphi_\ell(c_i H \hat{J}_n) (\hat{V}_n - \tilde{V}_n)] c_i^\ell t_n^{2-\ell} H^{\ell-1} \\ &+ \sum_{j=1}^{i-1} (a_{ij}(H \hat{J}_n) - a_{ij}(H \tilde{J}_n)) \tilde{N}_n(t_{nj}, \tilde{U}_{nj}) \\ &+ \sum_{j=1}^{i-1} a_{ij}(H \hat{J}_n) (\hat{N}_n(t_{nj}, \hat{U}_{nj}) - \tilde{N}_n(t_{nj}, \tilde{U}_{nj})). \end{aligned}$$

Here,  $\hat{\varepsilon}_{n1} = \hat{U}_{n1} - y_{n1}(c_1 H) = \hat{u}_n - y_{n1}(0) = 0$  (due to  $c_1 = 0$ ) and thus  $\hat{E}_{n1} = \hat{\varepsilon}_n$ .

*Proof.* From (2.37) and (2.26), we have

$$(2.41) \quad \hat{E}_{ni} = \hat{\varepsilon}_{ni} + y_{ni}(c_i H) - \tilde{U}_{ni}.$$

Using (2.38b), one can write  $\tilde{U}_{ni}$  given in (2.23b) as

$$(2.42) \quad \tilde{U}_{ni} = e^{c_i H \tilde{J}_n} u(t_n) + H \tilde{A}_{ni}.$$

By applying the variation-of-constants formula to (2.14) and using (2.11a),

$$(2.43) \quad y_{ni}(c_i H) = e^{c_i H \hat{J}_n} \hat{u}_n + H \hat{A}_{ni} = e^{c_i H \tilde{J}_n} (\hat{\varepsilon}_n + u(t_n)) + H \hat{A}_{ni},$$

where  $\hat{A}_{ni}$  is given in (2.38a). Inserting (2.42) and (2.43) into (2.41) gives (2.39). Similarly to (2.36), the expression (2.40) can be verified by subtracting (2.38b) from (2.38a) first and then adding and subtracting to the result the same terms  $c_i \varphi_1(c_i H \hat{J}_n) \tilde{V}_n t_n$ ,  $c_i^2 H \varphi_2(c_i H \hat{J}_n) \tilde{V}_n$ , and  $\sum_{j=1}^{i-1} a_{ij}(H \hat{J}_n) \tilde{N}_n(t_n + c_j H, \tilde{U}_{nj})$ .  $\square$

Next, we prove several bounds needed to estimate the terms in (2.36) and (2.40). To simplify our presentation within both this and the following subsections, we will use  $C$  as a generic constant that may have different values at each occurrence.

LEMMA 2.7. *Under Assumption 2, the bound*

$$(2.44) \quad \|\hat{N}_n(t_{ni}, \hat{U}_{ni}) - \tilde{N}_n(t_{ni}, \tilde{U}_{ni})\| \leq C \|\hat{E}_{ni}\| + C \|\hat{E}_{ni}\|^2 + C \|\hat{J}_n - \tilde{J}_n\| + C \|\hat{V}_n - \tilde{V}_n\|$$

holds for all  $n$  and  $i$  as long as  $\hat{E}_{ni}$  remains in a sufficiently small neighborhood of 0.

*Proof.* First, we split

$$\hat{N}_n(t_{ni}, \hat{U}_{ni}) - \tilde{N}_n(t_{ni}, \tilde{U}_{ni}) = \underbrace{\hat{N}_n(t_{ni}, \hat{U}_{ni}) - \hat{N}_n(t_{ni}, \tilde{U}_{ni})}_{N_{split1}} + \underbrace{\hat{N}_n(t_{ni}, \tilde{U}_{ni}) - \tilde{N}_n(t_{ni}, \tilde{U}_{ni})}_{N_{split2}}.$$

Using (2.13a) and (2.24), we write the term  $N_{split2}$  as

$$(2.45) \quad \begin{aligned} N_{split2} &= (F(t_{ni}, \tilde{U}_{ni}) - \hat{J}_n \tilde{U}_{ni} - \hat{V}_n t_{ni}) - (F(t_{ni}, \tilde{U}_{ni}) - \tilde{J}_n \tilde{U}_{ni} - \tilde{V}_n t_{ni}) \\ &= (\tilde{J}_n - \hat{J}_n) \tilde{U}_{ni} + (\tilde{V}_n - \hat{V}_n) t_{ni}. \end{aligned}$$

Expanding  $\hat{N}_n(t_{ni}, \hat{U}_{ni})$  into a Taylor series expansion around  $(t_{ni}, \tilde{U}_{ni})$  gives

$$(2.46) \quad N_{split1} = \frac{\partial \hat{N}_n}{\partial u}(t_{ni}, \tilde{U}_{ni}) \hat{E}_{ni} + \int_0^1 (1-\theta) \frac{\partial^2 \hat{N}_n}{\partial u^2}(t_{ni}, \tilde{U}_{ni} + \theta \hat{E}_{ni}) (\hat{E}_{ni}, \hat{E}_{ni}) d\theta.$$

Under Assumption 2, (2.44) follows by bounding  $\|N_{split1}\| + \|N_{split2}\|$ .  $\square$

LEMMA 2.8. *Under Assumptions 1 and 2, the bounds*

$$(2.47a) \quad \|\hat{J}_n - \tilde{J}_n\| \leq C \|\hat{e}_n\| + C \|\hat{e}_n\|^2,$$

$$(2.47b) \quad \|\hat{V}_n - \tilde{V}_n\| \leq C \|\hat{e}_n\| + C \|\hat{e}_n\|^2,$$

$$(2.47c) \quad \|e^{t\hat{J}_n} - e^{t\tilde{J}_n}\| \leq Ct \|\hat{e}_n\|, \quad t \geq 0$$

$$(2.47d) \quad \|\varphi_\ell(t\hat{J}_n) - \varphi_\ell(t\tilde{J}_n)\| \leq Ct \|\hat{e}_n\|, \quad t \geq 0$$

$$(2.47e) \quad \|b_i(H\hat{J}_n) - b_i(H\tilde{J}_n)\| \leq CH \|\hat{e}_n\|,$$

$$(2.47f) \quad \|a_{ij}(H\hat{J}_n) - a_{ij}(H\tilde{J}_n)\| \leq CH \|\hat{e}_n\|$$

hold for all  $n, \ell, i$  and  $j$ , as long as the global errors  $\hat{e}_n$  remain in a sufficiently small neighborhood of 0.

*Proof.* By definition,  $\hat{J}_n - \tilde{J}_n = \frac{\partial F}{\partial u}(t_n, \hat{u}_n) - \frac{\partial F}{\partial u}(t_n, u(t_n))$ . Using Assumption 2, one can expand  $G(t, u) := \frac{\partial F}{\partial u}(t, u)$  in a Taylor series around  $(t_n, u(t_n))$  to get

$$\hat{J}_n - \tilde{J}_n = \frac{\partial G}{\partial u}(t_n, u(t_n)) \hat{e}_n + \mathcal{O}(\|\hat{e}_n\|^2),$$

which shows (2.47a). Similarly (2.47b) may be verified by expanding  $\frac{\partial F}{\partial t}(t, u)$  in a Taylor series around  $(t_n, u(t_n))$ .

Next, we estimate the difference between the two semigroups  $e^{t\hat{J}_n}$  and  $e^{t\tilde{J}_n}$  in a similar manner as in [17, Lemma 4.2]. Observing that  $e^{t\tilde{J}_n}$  is the solution of the IVP

$$w'(t) = \hat{J}_n w(t) = \tilde{J}_n w(t) + (\hat{J}_n - \tilde{J}_n) w(t), \quad w(0) = I,$$

We apply the variation-of-constants formula to this IVP to obtain

$$e^{t\hat{J}_n} - e^{t\tilde{J}_n} = t \int_0^1 e^{(1-\theta)t\tilde{J}_n} (\hat{J}_n - \tilde{J}_n) e^{\theta t\tilde{J}_n} d\theta.$$

Therefore, (2.47c) follows directly from (2.19) and (2.47a). Using this, the bounds (2.47d)–(2.47f) follow from using (2.4) and (2.8) (see also [17, Lemma 4.3]).  $\square$

Using the results from Lemmas 2.6, 2.7, and 2.8, we obtain the following result.

COROLLARY 2.9. *Under Assumptions 1 and 2, the estimate*

$$(2.48) \quad \|\hat{B}_k - \tilde{B}_k\| \leq \sum_{j=1}^i C\|\hat{e}_{kj}\| + C\|\hat{e}_k\| + C\|\hat{e}_k\|^2$$

holds for all  $k$ , as long as  $\hat{E}_{ki}$  and the global errors  $\hat{e}_k$  remain in a sufficiently small neighborhood of 0.

*Proof.* Using Lemmas 2.8 and 2.7, one can bound (2.36) as

$$(2.49) \quad \|\hat{B}_k - \tilde{B}_k\| \leq CH\|\hat{e}_k\| + C\|\hat{e}_k\| + C\|\hat{e}_k\|^2 + C\|\hat{E}_{ki}\| + C\|\hat{E}_{ki}\|^2.$$

Next, we apply Lemma 2.6 (with  $n = k$ ) to get  $\hat{E}_{ki}$  and then estimate it by using (2.19) and Lemma 2.8 (the bound (2.47c)):

$$(2.50) \quad \|\hat{E}_{ki}\| \leq \|\hat{e}_{ki}\| + C\|\hat{e}_k\| + CH\|\hat{e}_k\| + H\|\hat{A}_{ki} - \tilde{A}_{ki}\|.$$

Again using Lemmas 2.8 and 2.7, the bound on  $\|\hat{A}_{ki} - \tilde{A}_{ki}\|$  from (2.40) is similar to (2.49). Inserting this into (2.50) we have

$$(2.51) \quad \|\hat{E}_{ki}\| \leq \|\hat{e}_{ki}\| + CH\|\hat{e}_k\| + C\|\hat{e}_k\| + C\|\hat{e}_k\|^2 + \sum_{j=1}^{i-1} C\|\hat{E}_{kj}\|.$$

Since  $\hat{E}_{k1} = \hat{e}_k$  (see Lemma 2.6), this relation finally shows that

$$(2.52) \quad \|\hat{E}_{ki}\| \leq \|\hat{e}_{ki}\| + CH\|\hat{e}_k\| + C\|\hat{e}_k\| + C\|\hat{e}_k\|^2 + \sum_{j=1}^{i-1} C\|\hat{e}_{kj}\|.$$

Now using the fact that  $CH\|\hat{e}_k\| + C\|\hat{e}_k\| = (CH + C)\|\hat{e}_k\| \leq C\|\hat{e}_k\|$ , one can easily show (2.48) from (2.49) and (2.52).  $\square$

Finally, we give a technical lemma, which can be later used to estimate the term *Error1* appearing in (2.30).

LEMMA 2.10. *Let  $\{Z_j\}_{j=0}^n$  and  $\{Y_j\}_{j=0}^n$  be two sequences (of operators) in  $X$ . We have*

$$(2.53) \quad \prod_{j=0}^n Z_{n-j} - \prod_{j=0}^n Y_{n-j} = \sum_{k=0}^n \left( \prod_{j=0}^{n-k-1} Z_{n-j} \right) (Z_k - Y_k) \left( \prod_{j=n-k+1}^n Y_{n-j} \right).$$

*Proof.* By adding and subtracting  $\prod_{j=0}^{n-1} Z_{n-j} Y_0$  and then  $\prod_{j=0}^{n-2} Z_{n-j} Y_0 Y_1$ , the left hand side of (2.53) can be written as

$$\begin{aligned} & Z_n Z_{n-1} \dots Z_1 (Z_0 - Y_0) + Z_n Z_{n-1} \dots Z_2 (Z_1 - Y_1) Y_0 + (Z_n Z_{n-1} \dots Z_2 - Y_n Y_{n-1} \dots Y_2) Y_1 Y_0 \\ &= \left( \prod_{j=0}^{n-1} Z_{n-j} \right) (Z_0 - Y_0) + \left( \prod_{j=0}^{n-2} Z_{n-j} \right) (Z_1 - Y_1) Y_0 + (Z_n Z_{n-1} \dots Z_2 - Y_n Y_{n-1} \dots Y_2) Y_1 Y_0. \end{aligned}$$

We continue adding and subtracting  $\left( \prod_{j=0}^{n-k-1} Z_{n-j} \right) \left( \prod_{j=n-k+1}^n Y_{n-j} \right)$  in this manner until  $k = n$  to obtain the right hand side (2.53).  $\square$

COROLLARY 2.11. *Under Assumptions 1 and 2, the estimate*

$$(2.54) \quad \left\| \prod_{j=0}^n e^{H\hat{J}_{n-j}} - \prod_{j=0}^n e^{H\tilde{J}_{n-j}} \right\| \leq H \sum_{k=0}^n C \|\hat{e}_k\|.$$

holds for all  $n$  as long as the global errors  $\hat{e}_k$  remain sufficiently small.

*Proof.* This follows by applying Lemma 2.10 to  $Z_{n-j} = e^{H\hat{J}_{n-j}}$  and  $Y_{n-j} = e^{H\tilde{J}_{n-j}}$ , and using the stability bound (2.19) and the bound (2.47c) from Lemma 2.8.  $\square$

**2.3.4. MERB convergence.** With the above preparation in hand, we are now ready to prove convergence of our MERB methods.

THEOREM 2.12. *Let the initial value problem (1.1) satisfy Assumptions 1–2. Consider for its numerical solution a MERB method (2.14)–(2.17) that is constructed from an ExprB method of global order  $p$  using with macro time step  $H$ . Let  $m$  denote the number of fast steps per slow step. If the fast ODEs (2.14) and (2.17) associated with the MERB method are integrated with micro time step  $h = H/m$  by using ODE solvers that have global order of convergence  $q$  and  $r$ , respectively, then the MERB method is convergent with the error bound*

$$(2.55) \quad \|\hat{u}_n - u(t_n)\| \leq CH^p + CHh^q + Ch^r$$

on compact time intervals  $t_0 \leq t_n = t_0 + nH \leq T$ . Here, while the first error constant depends on  $T - t_0$  (but is independent of  $n$  and  $H$ ), the second and third error constants also depend on the error constants of the chosen ODE solvers.

*Proof.* We begin with the global error expansion given in Theorem 2.4, and estimate each of the terms in (2.30). First, from Corollary 2.11 it is obvious that  $\|\text{Error1}\| \leq H \sum_{k=0}^n C \|\hat{e}_k\|$ . Then the stability bound (2.19) tells us that  $\|\text{Error2}\| \leq \sum_{k=0}^n C \|\hat{e}_{k+1}\|$  and  $\|\text{Error3}\| \leq \sum_{k=0}^n C \|\hat{e}_{k+1}\|$ . Next, examining the expression (2.35) we employ Corollaries 2.9 and 2.11, along with the stability bound (2.19), to obtain  $\|\text{Error4}\| \leq H \sum_{k=0}^n [CH \|\hat{e}_k\| + \sum_{j=1}^i C \|\hat{e}_{kj}\| + C \|\hat{e}_k\| + C \|\hat{e}_k\|^2]$ . Therefore, we derive from (2.30) that

$$(2.56) \quad \|\hat{e}_{n+1}\| \leq H \sum_{k=0}^n C \|\hat{e}_k\| + \sum_{k=0}^n C \|\tilde{e}_{k+1}\| + \sum_{k=0}^n C \|\hat{e}_{k+1}\| + H \sum_{k=0}^n \left( \sum_{j=1}^i C \|\hat{e}_{kj}\| \right).$$

From our assumption that the base ExprB method has global order  $p$ , its local error satisfies  $\|\tilde{e}_{k+1}\| \leq CH^{p+1}$ .

As for the global errors  $\hat{e}_{k+1}$  and  $\hat{e}_{kj}$  obtained by solving the fast ODEs (2.17) and (2.14) on  $[0, H]$  and  $[0, c_i H]$ , respectively (using micro time step  $h$ ), the global error analysis from [8, Theorem 3.4] guarantees that

$$(2.57a) \quad \|\hat{e}_{k+1}\| \leq h^r \frac{C}{L} (e^{LH} - 1) = Ch^r H \frac{e^{LH} - 1}{LH} = Ch^r H \varphi_1(LH) \leq Ch^r H,$$

$$(2.57b) \quad \|\hat{e}_{kj}\| \leq h^q \frac{C}{L} (e^{Lc_i H} - 1) = Ch^q c_i H \frac{e^{Lc_i H} - 1}{Lc_i H} \leq Ch^q H \varphi_1(Lc_i H) \leq Ch^q H.$$

These bounds require that the Jacobian  $\hat{J}_k$  of the right hand sides of both ODEs satisfies  $\|\hat{J}_k\| \leq L$ . This follows from  $\|\frac{\partial F}{\partial u}(t, u)\| \leq L$ , which easily follows from Assumption 2. Combining these bounds and shifting the index  $n$  in (2.56) to  $n - 1$ , we obtain

$$(2.58) \quad \|\hat{e}_n\| \leq H \sum_{k=0}^{n-1} C \|\hat{e}_k\| + \sum_{k=0}^{n-1} (CH^{p+1} + Ch^r H + Ch^q H^2).$$

The error bound (2.55) results from applying a discrete Gronwall lemma to (2.58).  $\square$

*Remark 2.13.* Since  $h = H/m$ , one can write the error bound (2.55) as  $\|\hat{u}_n - u(t_n)\| \leq CH^p + \frac{C}{m^q}H^{q+1} + \frac{C}{m^r}H^r$ . Thus for a fixed  $m$ , a MERB method (2.14)–(2.17) will converge with order  $p$  provided that the inner ODE solvers for (2.14) and (2.17) have orders  $q \geq p - 1$  and  $r \geq p$ , respectively. We note that this is an improvement compared to MRI-GARK methods [23], that in fact require both  $q \geq p$  and  $r \geq p$  for a method of order  $p$ . It is also worth mentioning that the error bound (2.55) for MERB methods is similar to the one obtained with MERK methods [14].

**2.4. Construction of specific MERB methods.** Guided by Theorem 2.12, in order to derive MERB methods it is important to begin with base ExprB methods that satisfy Lemma 2.2. Fortunately, such ExprB methods are available up to order 6 in the literature, see [9, 17, 18]. In this subsection, we extend some of these methods to write their coefficients more generally, and then derive MERB methods of orders 2 through 6 from these schemes. Note that since a MERB method (2.14)–(2.17) is uniquely characterized by its polynomials  $\hat{p}_{ni}(\tau)$  and  $\hat{q}_n(\tau)$ , we only provide those polynomials here. In particular, we note that these MERB methods require fewer modified ODEs to be solved per slow time step than comparable order MRI-GARK [23] and MERK methods [14].

**2.4.1. Second-order methods.** First, consider the second-order ExprB-Euler scheme (see [9], and [12, Sect. 1.2.2] for non-autonomous problems)

$$u_{n+1} = u_n + H\varphi_1(HJ_n)F(t_n, u_n) + H^2\varphi_2(HJ_n)V_n.$$

Using Lemma 2.2 we immediately derive from this the second-order MERB2 method:

$$(2.59) \quad \hat{q}_n(\tau) = \hat{N}_n(t_n, \hat{u}_n) + (t_n + \tau)\hat{V}_n, \quad \tau \in [0, H].$$

This only requires the solution of one modified ODE. We note that since second order multirate methods have been available for some time, we do not include MERB2 in our numerical results, and instead focus on higher order multirate methods.

**2.4.2. Third-order methods.** In [9], a 2-stage 3rd-order ExprB method called `exprb32` was constructed (using  $c_2 = 1$ ) for autonomous problems. Extending this to non-autonomous problems and writing this for general  $c_2$ , we solve condition 1 of Table 1 directly (with  $s = 2$ ) to give a general family of third-order methods:

$$(2.60) \quad \begin{aligned} U_{n2} &= u_n + c_2H\varphi_1(c_2HJ_n)F(t_n, u_n) + c_2^2H^2\varphi_2(c_2HJ_n)V_n, \\ u_{n+1} &= u_n + H\varphi_1(HJ_n)F(t_n, u_n) + H^2\varphi_2(HJ_n)V_n + H\frac{2}{c_2^2}\varphi_3(HJ_n)D_{n2}. \end{aligned}$$

From this we construct the MERB3 family of third-order methods:

$$(2.61) \quad \begin{aligned} \hat{p}_{n2}(\tau) &= \hat{N}_n(t_n, \hat{u}_n) + (t_n + \tau)\hat{V}_n, & \tau \in [0, c_2H], \\ \hat{q}_n(\tau) &= \hat{N}_n(t_n, \hat{u}_n) + (t_n + \tau)\hat{V}_n + \frac{\tau^2}{c_2^2H^2}\hat{D}_{n2}, & \tau \in [0, H]. \end{aligned}$$

Clearly, this requires the solution of 2 modified ODEs per slow time step (whereas third-order MERK and MRI-GARK methods require solving 3 modified ODEs per step). In our numerical experiments we take  $c_2 = \frac{1}{2}$ , which gives rise to a total fast time step traversal for MERB3 of  $(1 + c_2)H = 1.5H$ .

**2.4.3. Fourth-order method.** There exist several 4th-order ExprB schemes [9, 17, 18, 13, 15] with coefficients fulfilling Lemma 2.2. However, we chose a 2-stage 4th-order ExprB method called `exprb42` which was constructed for autonomous problems in [13]. Transforming this to non-autonomous form, we have

$$(2.62) \quad \begin{aligned} U_{n2} &= u_n + \frac{3}{4}H\varphi_1\left(\frac{3}{4}HJ_n\right)F(t_n, u_n) + \frac{9}{16}H^2\varphi_2\left(\frac{3}{4}HJ_n\right)V_n, \\ u_{n+1} &= u_n + H\varphi_1(HJ_n)F(t_n, u_n) + H^2\varphi_2(HJ_n)V_n + H\frac{16}{9}\varphi_3(HJ_n)D_{n2}. \end{aligned}$$

We then apply Lemma 2.2 to construct the 4th-order MERB4 method:

$$(2.63) \quad \begin{aligned} \hat{p}_{n2}(\tau) &= \hat{N}_n(t_n, \hat{u}_n) + (t_n + \tau)\hat{V}_n, & \tau \in [0, \frac{3}{4}H] \\ \hat{q}_n(\tau) &= \hat{N}_n(t_n, \hat{u}_n) + (t_n + \tau)\hat{V}_n + \frac{16}{9}\frac{\tau^2}{H^2}\hat{D}_{n2}, & \tau \in [0, H]. \end{aligned}$$

MERB4 only requires solving 2 modified ODEs per slow time step, whereas 4th-order MRI-GARK and MERK methods require 5 and 4 modified ODEs in each step, respectively. We further note that (2.63) has a total fast traversal time of  $\frac{7}{4}H = 1.75H$ .

**2.4.4. Fifth-order methods.** ExprB methods of order 5 can be found in [17, 18]. Here, for efficiency purposes, we consider a parallel scheme called `pexprb54s4`, whose coefficients (with fixed nodes  $c_i$ ) satisfy Lemma 2.2. It uses  $s = 4$  stages and is embedded with a fourth-order scheme (for stepsize adaptivity) but can be implemented as a 3-stage method. A detailed derivation of `pexprb54s4` is given in [18] (solving conditions 1–4 of Table 1 with the choices  $b_2(Z) = 0$ ,  $a_{43}(Z) = 0$ ,  $a_{32}(Z) = \frac{2c_3^3}{c_2^3}\varphi_3(c_3Z)$ , and  $a_{42} = \frac{2c_4^3}{c_2^3}\varphi_3(c_4Z)$ ). Following that derivation, we present here a family of fifth-order ExprB methods (depending on parameters  $c_2, c_3, c_4$ ) for non-autonomous problems:

$$(2.64) \quad \begin{aligned} U_{n2} &= u_n + H(c_2\varphi_1(c_2HJ_n)F(t_n, u_n) + c_2^2H\varphi_2(c_2HJ_n)V_n), \\ U_{n3} &= u_n + H\left(c_3\varphi_1(c_3HJ_n)F(t_n, u_n) + c_3^2H\varphi_2(c_3HJ_n)V_n + \frac{2c_3^3}{c_2^3}\varphi_3(c_3HJ_n)D_{n2}\right), \\ U_{n4} &= u_n + H\left(c_4\varphi_1(c_4HJ_n)F(t_n, u_n) + c_4^2H\varphi_2(c_4HJ_n)V_n + \frac{2c_4^3}{c_2^3}\varphi_3(c_4HJ_n)D_{n2}\right), \\ u_{n+1} &= u_n + H(\varphi_1(HJ_n)F(t_n, u_n) + H\varphi_2(HJ_n)V_n + b_3(HJ_n)D_{n3} + b_4(HJ_n)D_{n4}) \end{aligned}$$

with

$$\begin{aligned} b_3(HJ_n) &= \frac{1}{c_3^2(c_4 - c_3)}(c_4\varphi_3(HJ_n) - 6\varphi_4(HJ_n)), \\ b_4(HJ_n) &= \frac{1}{c_4^2(c_3 - c_4)}(2c_3\varphi_3(HJ_n) - 6\varphi_4(HJ_n)), \\ c_4 &= \frac{3(5c_3 - 4)}{5(4c_3 - 3)}. \end{aligned}$$

We note that the two internal stages  $\{U_{n3}, U_{n4}\}$  are independent of one another and thus can be computed simultaneously. They also have the same format, in that they have the same formula but only act on different inputs  $c_3$  and  $c_4$ , which we exploit below to give the same polynomial for their corresponding modified ODEs.

Applying Lemma 2.2 to (2.64) results in the fifth-order family of MERB5 methods:

$$(2.65) \quad \begin{aligned} \hat{p}_{n2}(\tau) &= \hat{N}_n(t_n, \hat{u}_n) + (t_n + \tau)\hat{V}_n, & \tau \in [0, c_2H] \\ \hat{p}_{n3}(\tau) &\equiv \hat{p}_{n4}(\tau) = \hat{N}_n(t_n, \hat{u}_n) + (t_n + \tau)\hat{V}_n + \left(\frac{\tau}{c_2H}\right)^2 \hat{D}_{n2}, & \tau \in [0, c_3H] \\ \hat{q}_n(\tau) &= \hat{N}_n(t_n, \hat{u}_n) + (t_n + \tau)\hat{V}_n + \frac{\tau^2}{H^2} \left( \frac{c_4}{c_3^2(c_4 - c_3)} \hat{D}_{n3} + \frac{c_3}{c_4^2(c_3 - c_4)} \hat{D}_{n4} \right) \\ &\quad - \frac{\tau^3}{H^3} \left( \frac{1}{c_3^2(c_4 - c_3)} \hat{D}_{n3} + \frac{1}{c_4^2(c_3 - c_4)} \hat{D}_{n4} \right), & \tau \in [0, H]. \end{aligned}$$

This only requires solving 3 modified ODEs per slow step (the only existing fifth-order multirate method, MERK5, requires 5). In our experiments we choose  $c_2 = c_4 = \frac{1}{4} < c_3 = \frac{33}{40}$ , so we can solve the modified ODE (2.14) using the polynomial  $\hat{p}_{n3}(\tau)$  on  $[0, c_3H]$  to obtain *both*  $\hat{U}_{n3} \approx U_{n3}$  and  $\hat{U}_{n4} \approx U_{n4}$  (since  $c_4 < c_3$ ), without solving an additional fast ODE on  $[0, c_4H]$ . Using this strategy, the total fast traversal time for MERB5 is  $(1 + c_2 + c_3)H = \frac{83}{40}H = 2.075H$ .

**2.4.5. Sixth-order methods.** To the best of our knowledge, the only existing ExprB method of order 6, named `pexprb65s7`, is given in [18]. It uses  $s = 7$  stages and is embedded with a fifth-order method. As with (2.64), this method consists of multiple independent internal stages (namely the stages in two groups  $\{U_{n2}, U_{n3}\}$  and  $\{U_{n4}, U_{n5}, U_{n6}, U_{n7}\}$ ) that can be computed simultaneously, which we exploit to implement like a 3-stage method. While `pexprb65s7` is constructed for autonomous problems and uses a set of fixed nodes  $c_i$ , we extend the derivation from [18] to construct a family of 7-stage sixth-order methods for non-autonomous problems:

$$(2.66) \quad \begin{aligned} U_{nk} &= u_n + c_k H \varphi_1(c_k H J_n) F(t_n, u_n) + (c_k H)^2 \varphi_2(c_k H J_n) V_n, \quad k = 2, 3 \\ U_{ni} &= u_n + c_i H \varphi_1(c_i H J_n) F(t_n, u_n) + (c_i H)^2 \varphi_2(c_i H J_n) V_n, \\ &\quad + H a_{i2}(H J_n) D_{n2} + H a_{i3}(H J_n) D_{n3}, \quad i = 4, 5, 6, 7 \\ u_{n+1} &= u_n + H \varphi_1(H J_n) F(t_n, u_n) + H^2 \varphi_2(H J_n) V_n + H \sum_{i=4}^7 b_i(H J_n) D_{ni}, \end{aligned}$$

where

$$\begin{aligned} a_{i2}(H J_n) &= \frac{1}{c_2^2(c_3 - c_2)} (2c_i^3 c_3 \varphi_3(c_i H J_n) - 6c_i^4 \varphi_4(c_i H J_n)), \\ a_{i3}(H J_n) &= \frac{1}{c_3^2(c_2 - c_3)} (2c_i^3 c_2 \varphi_3(c_i H J_n) - 6c_i^4 \varphi_4(c_i H J_n)), \\ b_i(H J_n) &= -2\hat{\alpha}_i \varphi_3(H J_n) + 6\hat{\eta}_i \varphi_4(H J_n) - 24\hat{\beta}_i \varphi_5(H J_n) + 120\hat{\gamma}_i \varphi_6(H J_n), \\ \hat{\gamma}_i &= \frac{1}{c_i^2(c_i - c_k)(c_i - c_l)(c_i - c_m)}, \quad \hat{\alpha}_i = c_k c_l c_m \hat{\gamma}_i, \\ \hat{\beta}_i &= (c_k + c_l + c_m) \hat{\gamma}_i, \quad \hat{\eta}_i = (c_k c_l + c_l c_m + c_k c_m) \hat{\gamma}_i. \end{aligned}$$

Here  $i, k, l, m \in \{4, 5, 6, 7\}$  are distinct indices and  $c_i, c_k, c_l, c_m$  are distinct (positive) nodes. Applying Lemma 2.2 we obtain the first-ever sixth-order infinitesimal multirate

method, MERB6:

$$\begin{aligned}
\hat{p}_{n2}(\tau) &\equiv \hat{p}_{n3}(\tau) = \hat{N}_n(t_n, \hat{u}_n) + (t_n + \tau)\hat{V}_n, & \tau \in [0, c_2H] \\
\hat{p}_{n4}(\tau) &\equiv \hat{p}_{n5}(\tau) \equiv \hat{p}_{n6}(\tau) \equiv \hat{p}_{n7}(\tau) = \hat{N}_n(t_n, \hat{u}_n) + (t_n + \tau)\hat{V}_n \\
&\quad + \frac{\tau^2}{(c_3 - c_2)H^2} \left( \frac{c_3}{c_2^2} \hat{D}_{n2} - \frac{c_2}{c_3^2} \hat{D}_{n3} \right) - \frac{\tau^3}{(c_3 - c_2)H^3} \left( \frac{1}{c_2^2} \hat{D}_{n2} - \frac{1}{c_3^2} \hat{D}_{n3} \right), & \tau \in [0, c_4H] \\
\hat{q}_n(\tau) &= \hat{N}_n(t_n, \hat{u}_n) + (t_n + \tau)\hat{V}_n - \frac{\tau^2}{H^2} \sum_{i=4}^7 \hat{\alpha}_i \hat{D}_{ni} + \frac{\tau^3}{H^3} \sum_{i=4}^7 \hat{\eta}_i \hat{D}_{ni} \\
&\quad - \frac{\tau^4}{H^4} \sum_{i=4}^7 \hat{\beta}_i \hat{D}_{ni} + \frac{\tau^5}{H^5} \sum_{i=4}^7 \hat{\gamma}_i \hat{D}_{ni}. & \tau \in [0, H]
\end{aligned}$$

As seen, MERB6 requires only 3 modified ODEs per each slow time step like MERB5, reflecting the fact that its base 6th-order ExpRB method (2.66) has the structure of a 3-stage method. MERB6 can be also implemented in an efficient way by choosing  $c_3 < c_2$  and  $c_5, c_6, c_7 < c_4$ . With these choices, we can solve the modified ODE (2.14) using  $\hat{p}_{n2}(\tau)$  on  $[0, c_2H]$  to obtain both  $\hat{U}_{n2} \approx U_{n2}$  and  $\hat{U}_{n3} \approx U_{n3}$  without solving an additional fast ODE on  $[0, c_3H]$ . Similarly, we can solve (2.14) using  $\hat{p}_{n4}(\tau)$  on  $[0, c_4H]$  to get *all four approximations*  $\hat{U}_{ni} \approx U_{ni}$  ( $i = 4, 5, 6, 7$ ) without solving 3 additional ODEs on  $[0, c_5H]$ ,  $[0, c_6H]$ , and  $[0, c_7H]$ . In our numerical experiments, we take  $c_3 = c_5 = \frac{1}{10} < c_2 = c_6 = \frac{1}{9} < c_7 = \frac{1}{8} < c_4 = \frac{1}{7}$ . This gives a total fast traversal time of  $(1 + c_2 + c_4)H = \frac{79}{63}H \approx 1.253H$ .

**2.5. MERB method implementation.** In Algorithm 2.1 we provide a precise description of the MERB algorithm. We note that in our implementations of MERB

---

**Algorithm 2.1** MERB method

---

- **Input:**  $F; J; V; t_0; u_0; s; c_i$  ( $i = 1, \dots, s$ );  $H$
  - **Initialization:** Set  $n = 0; \hat{u}_n = u_0$ .  
While  $t_n < T$ 
    1. Set  $\hat{U}_{n1} = \hat{u}_n$ .
    2. Compute  $\hat{J}_n = J(t_n, \hat{u}_n)$  and  $\hat{V}_n = V(t_n, \hat{u}_n)$
    3. For  $i = 2, \dots, s$  do
      - (a) Find  $\hat{p}_{ni}(\tau)$  as in (2.15).
      - (b) Solve (2.14) on  $[0, c_iH]$  to obtain  $\hat{U}_{ni} \approx y_{ni}(c_iH)$ .
    4. Find  $\hat{q}_n(\tau)$  as in (2.16)
    5. Solve (2.17) on  $[0, H]$  to get  $\hat{u}_{n+1} \approx y_{n+1}(H)$ .
    6. Update  $t_{n+1} := t_n + H, n := n + 1$ .
  - **Output:** Approximate values  $\hat{u}_n \approx u_n, n = 1, 2, \dots$  (where  $u_n$  is the numerical solution at time  $t_n$  obtained by an ExpRB method).
- 

methods, we found it beneficial to include formulas for  $\hat{N}_n(t, u)$  and  $\hat{D}_{ni}(t, u)$  as additional inputs to the algorithm (provided they can be pre-computed) for use in equations (2.15) and (2.16) to avoid floating-point cancellation errors when seeking very accurate solutions. On the other hand, we note that within the MERB algorithm, both the products  $Jw$  and  $V\tau$  can be approximated from  $F$  using finite differences,

$$\begin{aligned}
J(t, u)w &= \frac{1}{\sigma} (F(t, u + \sigma w) - F(t, u)) + \mathcal{O}(\sigma), \quad \text{and} \\
V(t, u)\tau &= \frac{1}{\sigma} (F(t + \sigma\tau, u) - F(t, u)) + \mathcal{O}(\sigma),
\end{aligned}$$

instead of  $J$  and  $V$  being provided analytically; however, when seeking high accuracy then such approximations can cause excessive floating-point cancellation error.

**3. Numerical Experiments.** In this section, we implement MERB methods on select multirate test problems to demonstrate accuracy and efficiency. We first discuss choices for the inner fast integrators, fast-slow splitting, optimal time scale separation factors, and give a general description of how the error and efficiency are measured. We then present numerical results for a reaction-diffusion problem and a semi-linear nonautonomous system with coupling between the fast and slow variables. For each problem, we compare the proposed MERB3, MERB4, MERB5, and MERB6 methods with other recently developed multirate methods that treat the slow time scale explicitly, namely MERK3, MERK4, and MERK5 from [14], plus MRI-GARK-ERK33a and MRI-GARK-ERK45a from [23], written here in short form as MRI-GARK33a and MRI-GARK45a. MATLAB implementations of all tests are provided on Github [4].

**3.1. Choice of inner integrators.** For uniformity in our implementations of MERB, MERK, and MRI-GARK methods of the same order, we use the same explicit fast integrators for solving all modified ODEs. Third-order methods use a 3-stage explicit third-order method from equation (233f) of [2], fourth-order methods use a 4-stage explicit fourth-order method commonly known as ‘‘RK4’’ from [11], fifth-order methods use an 8-stage fifth-order method which is the explicit part of ARK5(4)8L[2]SA from [10], while the sixth-order method uses an 8-stage explicit sixth-order method based on the 8,5(6) procedure of [28]. We note that although both MERK and MERB methods could compute the internal stages using a lower order integrator, for the sake of simplicity that approach is not used here.

**3.2. Fast-slow splitting.** The splitting of an IVP into fast and slow components,  $u'(t) = F(t, u) = F_f(t, u) + F_s(t, u)$ , for MERB methods is dictated by the dynamic linearization process at each time step,

$$(3.1) \quad \hat{u}'(t) = F(t, \hat{u}(t)) = \hat{J}_n \hat{u}(t) + \hat{V}_n t + \hat{N}_n(t, \hat{u}(t)),$$

where the multirate splitting becomes  $F_f(t, u) = \hat{J}_n u$  and  $F_s(t, u) = \hat{V}_n t + \hat{N}_n(t, u)$ . This brings interesting questions when comparing against MERK and MRI-GARK methods that do not require dynamic linearization. MERK methods require that  $F_f(t, u) = \mathcal{L}u$ , but MRI-GARK methods have no constraints on  $F_f$  or  $F_s$ . Thus to provide a more thorough picture in the following comparisons of MERB, MERK and MRI-GARK methods, we consider two separate fast-slow splittings for each problem. The first is the dynamic linearization (3.1), that can place more of a problem’s dynamics at the fast time scale than other fixed multirate splittings; this offers a potential for greater multirate accuracy but at the expense constructing the dynamic linearization at each slow step. Our second splitting defines a fixed  $F_f(t, u) = \mathcal{L}u$ , leaving  $F_s(t, u) = F(t, u) - \mathcal{L}u$ ; in the ensuing results we call this the ‘fixed linearization’. Though the motivation for this splitting arises from the MERK requirement on  $F_f$ , we also apply this splitting to MRI-GARK methods. We note that other fixed splittings which can offer different accuracy and efficiency insights on multirate methods are possible, however we only focus on one fixed splitting for each test problem. Methods that use fixed linearization are denoted with an asterisk in our results, for instance, MERK3\* uses a fixed linearization while MERK3 uses dynamic linearization.

**3.3. Optimal time scale separation.** In order to compare methods at their peak performance, we strive to determine an optimal time scale separation factor for each multirate method on each test problem. The optimal time scale separation factor

$m = H/h$  is the integer ratio between the slow and fast time step sizes that results in maximal efficiency. We follow the approach from [14] for determining this value experimentally, by comparing efficiency in terms of slow-only function evaluations and total (slow+fast) function evaluations for several different values of  $m$  and  $H$ .

Method	Slow stages	Modified ODEs	Fast time traversal of $[0, H]$	React.-diffusion optimal $m$		Bidirect. coupling optimal $m$	
				Dynamic	Fixed	Dynamic	Fixed
MERB3	2	2	1.5	10		80	
MERK3	3	3	2.166	20	10	80	10
MRI-GARK33a	3	3	1	20	5	80	10
MERB4	2	2	1.75	10		40	
MERK4	6	4	2.833	20	10	40	10
MRI-GARK45a	5	5	1	10	1	40	1
MERB5	4	3	2.075	5		10	
MERK5	10	5	3.2	5	5	10	10
MERB6	7	3	1.253	5		5	

Table 2: Multirate method properties: number of slow internal stages and modified ODEs, total step traversal times, and optimal  $m$  factors for each problem and splitting.

Table 2 presents the optimal  $m$  values for each method and each test problem splitting. A trend emerges among MERK and MRI-GARK methods that use both dynamic and fixed linearization: dynamic linearization almost exclusively results in larger optimal  $m$  values than fixed linearization, supporting our earlier hypothesis that dynamic linearization includes more of the problem within the fast dynamics, thereby requiring a larger  $m$  value to resolve. We also note that for the fixed linearization, both MRI-GARK methods have smaller  $m$  values than other methods, resulting in less computational work at the fast time scale for a given  $H$  value.

**3.4. Presentation of results.** For each test problem we sort our results into 3 groups:  $\mathcal{O}(H^3)$  methods,  $\mathcal{O}(H^4)$  methods, and  $\mathcal{O}(H^5)$  with  $\mathcal{O}(H^6)$  methods. In each group we provide four kinds of “log-log” plots: one convergence plot (error versus  $H$ ) and three efficiency plots that measure cost through slow function calls, total function calls, and MATLAB runtimes, respectively. Solution error is computed as the maximum absolute error over all spatial grid points and time outputs, as measured against either an analytical solution or highly accurate reference solution. We also compute convergence rates using a linear least squares fit of the  $\log(\text{error})$  versus  $\log(H)$  data, neglecting points at the reference solution floor. Each of our efficiency measurements tells a different story. First, slow function calls illustrate the cost of a multirate method when applied to IVP systems with expense dominated by the slow components  $F_s(t, u)$ . Second, total function calls capture the cost of  $F_f(t, u)$ , and highlight properties of methods related to their total fast traversal times. Lastly, even though MATLAB runtimes are a poor proxy for runtimes on HPC applications, we use them here to capture the costs associated with dynamic linearization, and to measure how these costs affect efficiency.

**3.5. Reaction-diffusion.** From Savcenca et al.[24], we consider the reaction-diffusion equation:

$$(3.2) \quad u_t = \epsilon u_{xx} + \gamma u^2(1-u), \quad 0 < x < 5, \quad 0 < t \leq 5.$$

The initial and boundary conditions are  $u(x, 0) = (1 + \exp(\lambda(x-1)))^{-1}$  and  $u_x(0, t) = u_x(5, t) = 0$  respectively, where  $\lambda = \frac{1}{2}\sqrt{2\gamma/\epsilon}$ . Multiple combinations of  $\gamma$  and  $\epsilon$  are possible, here we choose  $\gamma = 0.1$  and  $\epsilon = 0.01$  that lead to an optimal  $m > 1$  when using dynamic linearization. We use a second-order centered finite difference scheme with 101 spatial grid points to discretize the diffusion term. In addition to dynamic linearization, MERK and MRI-GARK methods also use a fixed splitting where  $F_f(t, u) = \epsilon u_{xx}$  and  $F_s(t, u) = \gamma u^2(1-u)$ . The numerical solution is considered at 10 evenly spaced points within the time interval, and all methods are tested with slow time steps  $H = 0.5 \times 2^{-k}$ , for  $k = 0, \dots, 6$ . We compute error by comparing against a reference solution obtained using MATLAB's `ode15s` with relative tolerance  $10^{-13}$  and absolute tolerance  $10^{-14}$ .

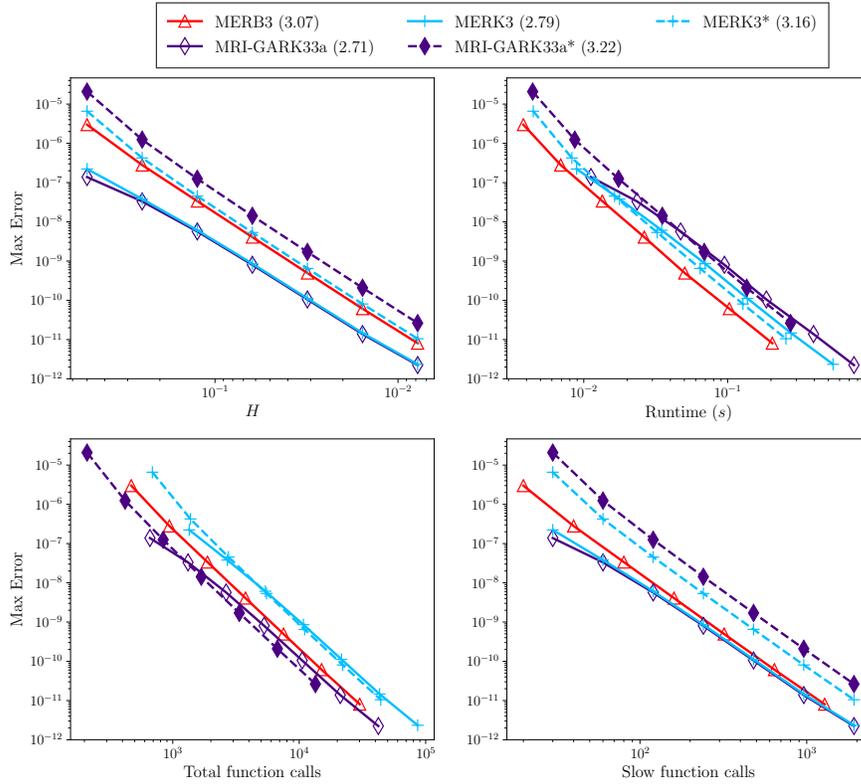


Fig. 1: Convergence (top-left) and efficiency (top-right, bottom) for  $\mathcal{O}(H^3)$  methods on the reaction-diffusion problem of Section 3.5. The legend displays the measured convergence rates for each method in parentheses.

Figures 1-3 show accuracy and efficiency results for this problem. Examining the top-left of Figure 1 and the legend, we see that each third-order method attains the expected order of convergence. The observed errors for the dynamic linearization

approach on all methods are less than for fixed linearization. This can be attributed to inclusion of more of the problem at the fast time scale in the case of dynamic linearization, which results in higher optimal time scale separation factors (as shown in Table 2) and lower errors. Among the methods that apply dynamic linearization, **MERK3** and **MRI-GARK33a** have almost identical errors that are lower than those for **MERB3**, which uses an  $m$  two times smaller. **MERK3\*** and **MRI-GARK33a\*** have the largest errors on this test problem.

Turning to the efficiency results at the top-right and bottom of Figure 1, the most efficient methods in each of these plots are closest to the bottom left corner. For our MATLAB implementations, **MERB3** has an obvious advantage in terms of runtime, while both **MRI-GARK33a** and **MRI-GARK33a\*** have the least efficient implementation. Taking into account only **MERK** and **MRI-GARK** methods, there is not a significant runtime difference between the dynamic fixed linearization approaches, although the fixed linearization is very slightly faster at tighter accuracies. When looking at total function calls, both **MRI-GARK33a** and **MRI-GARK33a\*** are the most efficient of the group, largely owing to their shorter fast traversal time of  $1.0H$ , while **MERB3** and **MERK3** have traversal times of  $1.5H$  and  $2.166H$ , respectively. The slow function call efficiency is closely aligned with the convergence behavior: at large values of  $H$ , **MERK3** and **MRI-GARK33a** are the most efficient, but **MERB3** is just as efficient as **MERK3** and **MRI-GARK33a** at tighter accuracies.

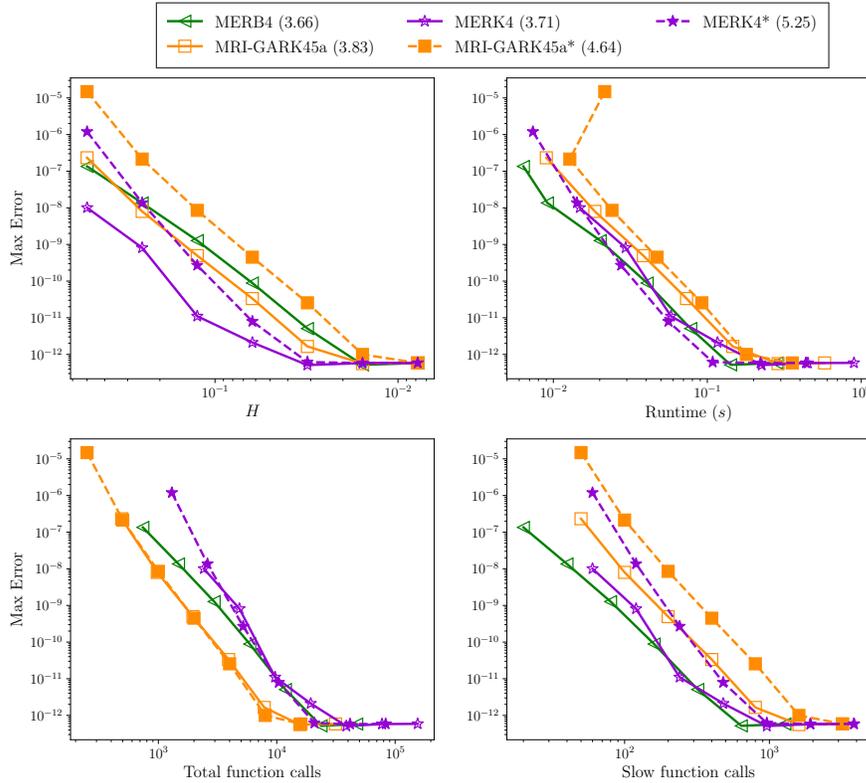


Fig. 2: Convergence (top-left) and efficiency (top-right, bottom) for  $\mathcal{O}(H^4)$  methods on the reaction-diffusion problem of Section 3.5.

In Figure 2 we present the corresponding plots for the fourth-order methods. Here, all methods approximately reach their expected order of convergence, with  $\text{MERK4}^*$  and  $\text{MRI-GARK45a}^*$  outperforming their expectations.  $\text{MERK4}$  has the smallest error, but also uses an  $m$  value that is two times greater than other fourth-order methods on this test problem (see Table 2).  $\text{MERK4}^*$  starts off with larger errors than  $\text{MERB4}$  and  $\text{MRI-GARK45a}$ , but because it converges at fifth-order for this test problem, its errors quickly drop below those for  $\text{MERB4}$  and  $\text{MRI-GARK45a}$ .  $\text{MRI-GARK45a}^*$  has an  $m = 1$  which seemingly puts it at a disadvantage when comparing accuracy with other methods, however, larger values of  $m$  only lead to more total function evaluations with no reduction in error. Focusing on runtime efficiency,  $\text{MERB4}$  is more efficient at larger error values, but  $\text{MERK4}$  is eventually the most efficient at smaller error values. Total function call efficiency repeats the previous pattern from third-order methods:  $\text{MRI-GARK45a}$  and  $\text{MRI-GARK45a}^*$  are the most efficient and closely line up,  $\text{MERB4}$  performs better than  $\text{MERK4}$  and  $\text{MERK4}^*$  due to its shorter total traversal time of  $1.75H$  versus  $2.833H$ . Finally, when comparing slow function calls  $\text{MERB4}$  is the most efficient. This is expected since  $\text{MERB4}$  only has 2 slow stages, compared with 6 for  $\text{MERK4}$  and 5 for  $\text{MRI-GARK45a}$ .

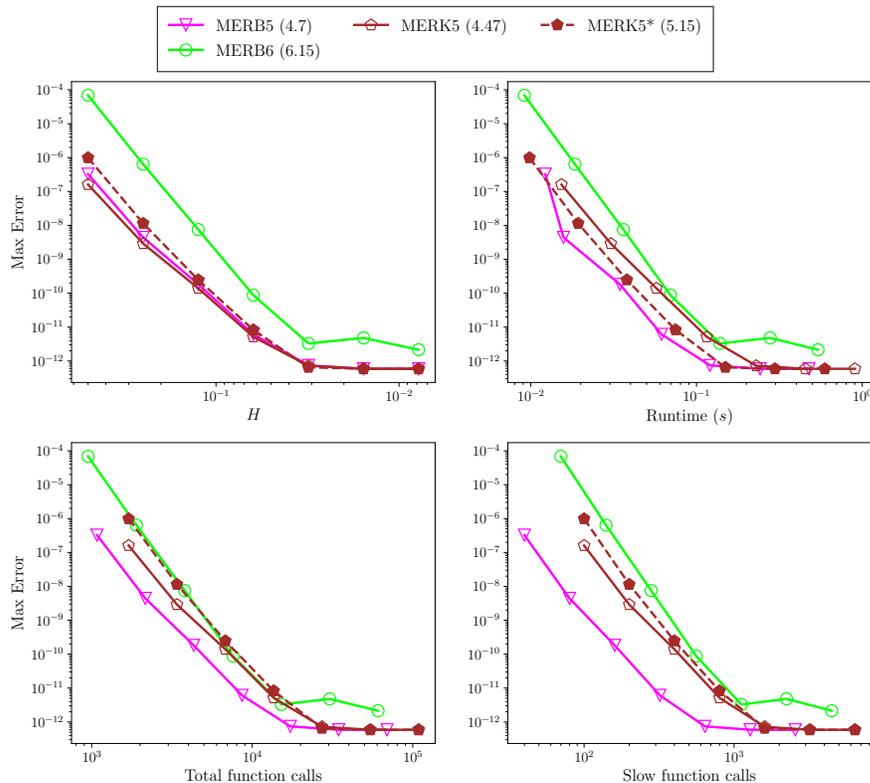


Fig. 3: Convergence (top-left) and efficiency (top-right, bottom) of  $\mathcal{O}(H^5)$  and  $\mathcal{O}(H^6)$  methods on the reaction-diffusion problem of Section 3.5.

The first thing to note discussing the fifth and sixth-order methods is that they all use the same  $m = 5$  for this problem (Table 2). Their convergence and efficiency plots

are provided in Figure 3. On this problem, all methods converge at their expected rates, although **MERB6** starts out with larger error values than the fifth-order methods, that all cluster around similar error values, although the dynamic linearization used by **MERK5** results in slightly less error than the fixed linearization used in **MERK5\***. In all three measures of efficiency **MERB5** is the most efficient. Looking at total function calls efficiency, **MERB5** has a total traversal time of  $2.075H$  compared to  $3.2H$  for **MERK5** and  $1.253H$  for **MERB6** (though we barely get to see advantages of this due to its larger error on this problem). When it comes to slow function calls, **MERB5**'s 4 slow stages is much lower than the 10 stages for **MERK5** and 7 stages for **MERB6**. Combining the merits of **MERB5** from total function calls and slow function calls explains its runtime efficiency performance.

**3.6. Bidirectional coupling system.** Inspired by [6, Sect. 5.1], we propose the following semi-linear, nonautonomous bidirectional coupling problem on  $0 < t \leq 1$ :

$$(3.3a) \quad u' = \sigma v - w - \beta t,$$

$$(3.3b) \quad v' = -\sigma u,$$

$$(3.3c) \quad w' = -\lambda(w + \beta t) - \beta \left( u - \frac{a(w + \beta t)}{a\lambda + b\sigma} \right)^2 - \beta \left( v - \frac{b(w + \beta t)}{a\lambda + b\sigma} \right)^2,$$

with exact solution  $u(t) = \cos(\sigma t) + ae^{-\lambda t}$ ,  $v(t) = -\sin(\sigma t) + be^{-\lambda t}$ , and  $w(t) = (a\lambda + b\sigma)e^{-\lambda t} - \beta t$ . This problem features linear coupling from slow to fast time scales through the equation (3.3a), and nonlinear coupling from fast to slow time scales through the equation for (3.3c). In addition, it includes tunable parameters  $\{a, b, \beta, \lambda, \sigma\}$  taken here to be  $\{1, 20, 0.01, 5, 100\}$ , with  $a\sigma = b\lambda$ ;  $\sigma$  determines the frequency of the fast time scale and  $\beta$  controls the strength of the nonlinearity. In the case of dynamic linearization, smaller values of  $\beta$  correspond with weaker nonlinearity, resulting in higher values of the optimal time scale separation factor  $m$ . While there are various possible fixed splittings, we chose the most natural splitting into fast variables and slow variables informed by the exact solution:

$$F_f(t, \mathbf{u}) = \begin{bmatrix} \sigma v \\ -\sigma u \\ 0 \end{bmatrix}, \quad F_s(t, \mathbf{u}) = \begin{bmatrix} -w - \beta t \\ 0 \\ -\lambda(w + \beta t) - \beta \left( u - \frac{a(w - \beta t)}{a\lambda + b\sigma} \right)^2 - \beta \left( v - \frac{b(w - \beta t)}{a\lambda + b\sigma} \right)^2 \end{bmatrix}$$

We assess error at 20 equally spaced points within the time interval and consider time steps  $H = 0.05 \times 2^{-k}$  for integers  $k = 0, 1, \dots, 7$ .

Accuracy and efficiency plots for this problem are shown in Figures 4-6. Starting with third-order methods in Figure 4, all methods incorporating dynamic linearization have similar errors, coinciding with their uniform time scale separation factor of  $m = 80$ . Similarly, the methods using fixed linearization **MERK3\*** and **MRI-GARK33a\*** have the same  $m = 10$ , leading to comparable errors. As before, dynamic linearization leads to lower errors than fixed linearization (here the difference in errors for the same  $H$  is up to  $10^3$ ). The previous efficiency observations are repeated here as well: **MERB3** is the most efficient in runtime and slow function evaluations, while **MRI-GARK33a** is slightly more efficient in total function evaluations.

Results for fourth-order methods are plotted in Figure 5. Like with the third-order methods, we use the same  $m$  for dynamic linearization methods, but here there is slightly more variation in errors, with **MERK4** being slightly more accurate than the

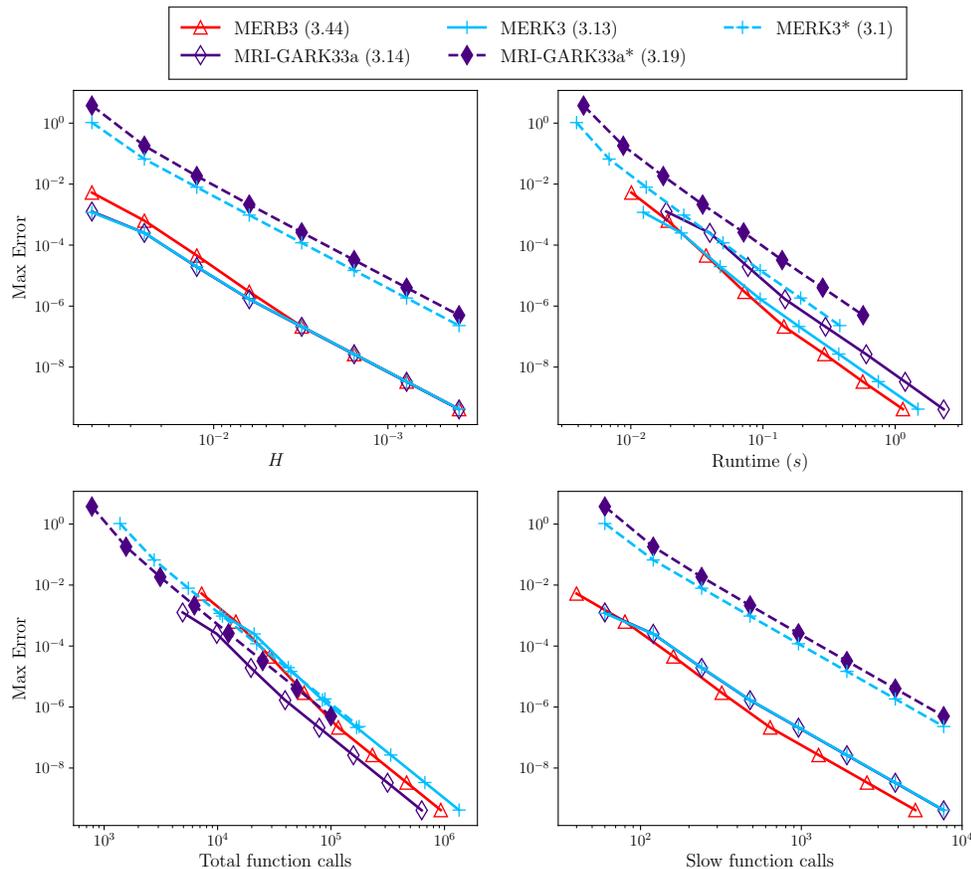


Fig. 4: Convergence (top-left) and efficiency (top-right, bottom) of  $\mathcal{O}(H^3)$  methods on the bidirectional coupling problem of Section 3.6.

others in this group. Both MERB4 and MERK4 show optimal runtime efficiency, the MRI-GARK methods are the most efficient in total function calls, and MERB4 is again the most efficient in slow function calls.

Finally, the performance of fifth and sixth-order methods on the bidirectional coupling problem is illustrated in Figure 6. The accuracy of these methods is almost identical on this test problem with MERB6 demonstrating a slightly steeper line, so we focus on the efficiency comparisons. Both of our new MERB methods are the most competitive for this test problem. We observe that MERB5 is slightly more efficient in terms of runtime at larger error values, but at smaller errors MERB6 becomes more efficient due to its higher order of accuracy. MERB6 is also the most efficient in total function calls followed by MERB5, due to their smaller total traversal times in comparison with MERK5. The small number of stages for MERB5 makes it clearly more efficient in terms of slow function calls.

**4. Conclusions.** We have introduced a new approach for multirate integration of initial-value problems that evolve on multiple time scales. Employing an MIS-like approach wherein the couplings between slow and fast time scales occurs through

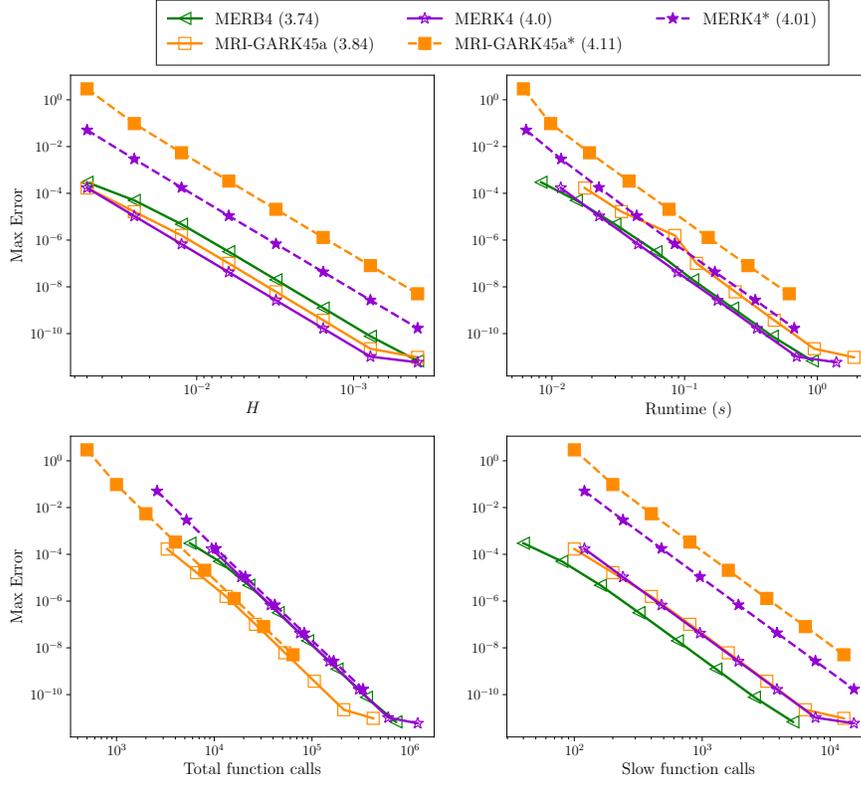


Fig. 5: Convergence (top-left) and efficiency (top-right, bottom) of  $\mathcal{O}(H^4)$  methods on the bidirectional coupling problem of Section 3.6.

defining a sequence of modified IVPs at the fast time scale, and built off of existing ExprB methods, the proposed MERB methods allow creation of multirate methods with very high order of accuracy, and minimize the amount of costly processing of the slow time scale operator. In addition to deriving a clear mechanism for constructing these from certain classes of ExprB schemes, we provide rigorous convergence analysis for MERB methods. We note that the style of this analysis is much more elegant than our approach for MERK methods [14], in that we analyze the overall MERB error by separately quantifying the error between the MERB approximation of the underlying ExprB method, and the error in the ExprB approximation of the original IVP. With this theory in hand, we propose a suite of MERB methods with orders 2 through 6, where in the cases of orders 3–6, we additionally provide generalizations of the base ExprB methods and extend these to non-autonomous problems.

We examine the performance of the proposed MERB methods of orders 3 through 6, comparing these against existing MERK and explicit MRI-GARK methods on two test problems, and where the MERK and MRI-GARK methods are tested with two potential multirate splittings on each problem. While all MERB, MERK and MRI-GARK methods exhibited their theoretical convergence rates on these problems and splittings, their efficiency differed considerably. In order to provide results that potentially apply to a broad range of multirate applications, we investigate efficiency using three separate measurements of cost: MATLAB runtime, total function calls

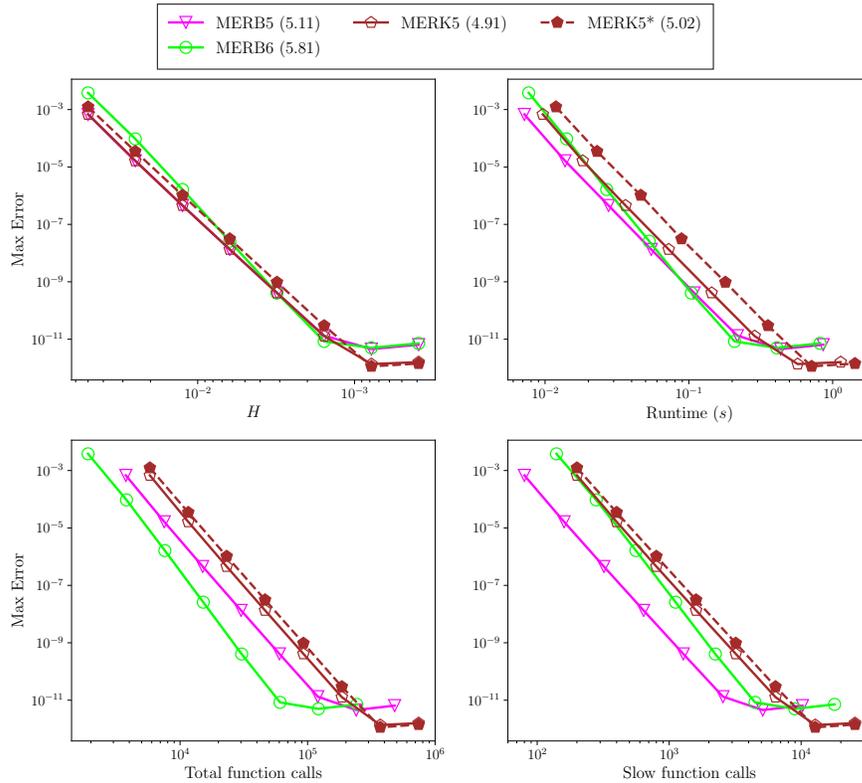


Fig. 6: Convergence (top-left) and efficiency (top-right, bottom) of  $\mathcal{O}(H^5)$  and  $\mathcal{O}(H^6)$  methods on the bidirectional coupling problem of Section 3.6.

(both fast and slow), and slow function calls only. Within these metrics, some general patterns emerge. First, most of the methods exhibited optimal efficiency at higher  $m = H/h$  values when using multirate splittings based on dynamic linearization as opposed to fixed splittings. Second, the proposed MERB methods show the best runtime efficiency of all methods and splittings, although in some cases the equivalent order MERK method with dynamic splitting is competitive. Third, due to their total fast time scale traversal times of  $1.0H$ , the MRI-GARK methods always exhibit the best total function call efficiency. Lastly, due to their low number of slow stages, the proposed MERB methods are uniformly the most efficient when considering slow function calls (only in a few instances MERK with dynamic splitting was competitive). This is particular of interest for multirate problems where the fast component is much less costly to compute than the slow component.

Based on these results, we find that the newly proposed MERB methods provide a unique avenue to construction of high order MIS-like multirate methods, and that they are very competitive in comparison with other recently-developed high order MIS-like multirate schemes. More work remains, however. An obvious extension is to include embeddings to enable low-cost temporal error estimation, as well as to investigate robust techniques for error-based multirate time step adaptivity. A further extension of MERB methods could focus on applications that require implicit or mixed implicit-explicit treatment of processes at the slow time scale.

## REFERENCES

- [1] T. P. BAUER AND O. KNOTH, *Extended multirate infinitesimal step methods: Derivation of order conditions*, Journal of Computational and Applied Mathematics, (2019), p. 112541, <https://doi.org/https://doi.org/10.1016/j.cam.2019.112541>.
- [2] J. C. BUTCHER, *Numerical Methods for Ordinary Differential Equations*, John Wiley & Sons, Apr. 2008.
- [3] R. CHINOMONA AND D. R. REYNOLDS, *Implicit-explicit multirate infinitesimal GARK methods*, SIAM J. Sci. Comput., (accepted, 2021).
- [4] R. CHINOMONA, D. R. REYNOLDS, AND V. T. LUAN, *Multirate exponential Rosenbrock methods (MERB)*. <https://github.com/rujekoc/merbrepo>, 2021.
- [5] K. J. ENGEL AND R. NAGEL, *One-parameter semigroups for linear evolution equations*, Springer, New York, 2000.
- [6] D. ESTEP, V. GINTING, AND S. TAVENER, *A posteriori analysis of a multirate numerical method for ordinary differential equations*, Computer Methods in Applied Mechanics and Engineering, 223–224 (2012), p. 10–27.
- [7] D. E. K. ET AL., *Multiphysics simulations: Challenges and opportunities*, The International Journal of High Performance Computing Applications, 27 (2013), pp. 4–83.
- [8] E. HAIRER, S. P. NØRSETT, AND G. WANNER, *Solving Ordinary Differential Equations I (2nd Revised. Ed.): Nonstiff Problems*, Springer-Verlag, Berlin, Heidelberg, 1993.
- [9] M. HOCHBRUCK, A. OSTERMANN, AND J. SCHWEITZER, *Exponential Rosenbrock-type methods*, SIAM J. Numer. Anal., 47 (2009), pp. 786–803.
- [10] C. A. KENNEDY AND M. H. CARPENTER, *Additive Runge–Kutta schemes for convection–diffusion–reaction equations*, Applied Numerical Mathematics, 44 (2003), pp. 139–181, [https://doi.org/10.1016/S0168-9274\(02\)00138-1](https://doi.org/10.1016/S0168-9274(02)00138-1).
- [11] W. KUTTA, *Beitrag zur näherungsweise integration totaler differentialgleichungen*, Zeitschrift für Math. u. Phys., 46 (1901), pp. 435–453.
- [12] V. T. LUAN, *High-order exponential integrators*, PhD thesis, University of Innsbruck, 2014.
- [13] V. T. LUAN, *Fourth-order two-stage explicit exponential integrators for time-dependent PDEs*, Applied Numerical Mathematics, 112 (2017), pp. 91–103.
- [14] V. T. LUAN, R. CHINOMONA, AND D. R. REYNOLDS, *A New Class of High-Order Methods for Multirate Differential Equations*, SIAM Journal on Scientific Computing, 42 (2020), pp. A1245–A1268, <https://doi.org/10.1137/19M125621X>.
- [15] V. T. LUAN AND D. L. MICHELS, *Efficient exponential time integration for simulating nonlinear coupled oscillators*, Journal of Computational and Applied Mathematics, 391 (2021), p. 113429.
- [16] V. T. LUAN AND A. OSTERMANN, *Exponential B-series: The stiff case*, SIAM Journal on Numerical Analysis, 51 (2013), pp. 3431–3445.
- [17] V. T. LUAN AND A. OSTERMANN, *Exponential Rosenbrock methods of order five–construction, analysis and numerical comparisons*, Journal of Computational and Applied Mathematics, 255 (2014), pp. 417–431.
- [18] V. T. LUAN AND A. OSTERMANN, *Parallel exponential Rosenbrock methods*, Computers & Mathematics with Applications, 71 (2016), pp. 1137–1150.
- [19] G. I. MARCHUK, *Some application of splitting-up methods to the solution of mathematical physics problems*, Aplikace Matematiky, 13 (1968), pp. 103–132, <http://eudml.org/doc/14518>.
- [20] R. I. MCLACHLAN AND G. R. W. QUISPTEL, *Splitting methods*, Acta Numerica, 11 (2002), pp. 341–434, <https://doi.org/10.1017/S0962492902000053>.
- [21] A. PAZY, *Semigroups of linear operators and applications to partial differential equations*, Springer, New York, (1983).
- [22] S. ROBERTS, A. SARSHAR, AND A. SANDU, *Coupled multirate infinitesimal GARK schemes for stiff systems with multiple time scales*, SIAM Journal on Scientific Computing, 42 (2020), pp. A1609–A1638, <https://doi.org/10.1137/19M1266952>.
- [23] A. SANDU, *A class of multirate infinitesimal gark methods*, SIAM Journal on Numerical Analysis, 57 (2019), pp. 2300–2327, <https://doi.org/10.1137/18M1205492>.
- [24] V. SAVCENCO, W. HUNSDORFER, AND J. G. VERWER, *A multirate time stepping strategy for stiff ordinary differential equations*, BIT, 47 (2007), pp. 137–155.
- [25] M. SCHLEGEL, O. KNOTH, M. ARNOLD, AND R. WOLKE, *Multirate Runge–Kutta schemes for advection equations*, Journal of Computational and Applied Mathematics, 226 (2009), pp. 345–357, <https://doi.org/10.1016/j.cam.2008.08.009>.
- [26] J. M. SEXTON AND D. R. REYNOLDS, *Relaxed multirate infinitesimal step methods for initial-value problems*, arXiv:1808.03718 [cs, math], (2018), <https://arxiv.org/abs/1808.03718>.

- [27] G. STRANG, *On the construction and comparison of difference schemes*, SIAM J. Numer. Anal., 5 (1968), pp. 506–517, <https://doi.org/10.1137/0705041>.
- [28] J. H. VERNER, *Explicit Runge–Kutta methods with estimates of the local truncation error*, SIAM Journal on Numerical Analysis, 15 (1978), pp. 772–790.
- [29] J. WENSCH, O. KNOTH, AND A. GALANT, *Multirate infinitesimal step methods for atmospheric flow simulation*, BIT Numer. Math., 49 (2009), pp. 449–473, <https://doi.org/10.1007/s10543-009-0222-3>.