Average energy dissipation rates of explicit exponential Runge-Kutta methods for gradient flow problems

Hong-lin Liao * Xuping Wang†

Abstract

We propose a unified theoretical framework to examine the energy dissipation properties at all stages of explicit exponential Runge-Kutta (EERK) methods for gradient flow problems. The main part of the novel framework is to construct the differential form of EERK method by using the difference coefficients of method and the so-called discrete orthogonal convolution kernels. As the main result, we prove that an EERK method can preserve the original energy dissipation law unconditionally if the associated differentiation matrix is positive semi-definite. A simple indicator, namely average dissipation rate, is also introduced for these multi-stage methods to evaluate the overall energy dissipation rate of an EERK method such that one can choose proper parameters in some parameterized EERK methods or compare different kinds of EERK methods. Some existing EERK methods in the literature are evaluated from the perspective of preserving the original energy dissipation law and the energy dissipation rate. Some numerical examples are also included to support our theory.

KEYWORDS: gradient flow problem, explicit exponential Runge-Kutta method, discrete orthogonal convolution kernels, stage energy dissipation laws, average dissipation rate

AMS subject classifications: 65L20, 65M06, 65M12

1 Introduction

We propose a unified theoretical framework to examine the energy dissipation properties at all stages of explicit exponential Runge-Kutta (EERK) methods for solving the following semi-discrete semi-linear parabolic problem

$$u_h'(t) + L_h u_h(t) = g_h(u_h(t)), \quad u_h(t_0) = u_h^0, \tag{1.1}$$

where L_h is a symmetric, positive definite matrix resulting from certain spatial discretization of stiff term, typically the Laplacian operator $-\Delta$ with periodic boundary conditions, and g_h represents a nonlinear but non-stiff term. Without losing the generality, the finite difference method is assumed to approximate spatial operators and we define the discrete L^2 inner product $\langle u, v\rangle := v^T u$ and the L^2 norm $\|v\| := \sqrt{\langle v, v\rangle}$. Assume that there exists a non-negative Lyapunov function G_h such that $g_h(v) = -\frac{\delta}{\delta v} G_h(v)$. Then the problem (1.1) can be formulated into a gradient system

$$\frac{\mathrm{d}u_h}{\mathrm{d}t} = -\frac{\delta E}{\delta u_h} \quad \text{with} \quad E[v_h] := \frac{1}{2}\langle v_h, L_h v_h\rangle + \langle G_h(v_h), 1\rangle. \tag{1.2}$$

*ORCID 0000-0003-0777-6832. School of Mathematics, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China; Key Laboratory of Mathematical Modeling and High Performance Computing of Air Vehicles (NUAA), MIIT, Nanjing 211106, China. Emails: liaohl@nuaa.edu.cn and liaohl@csrc.ac.cn. This author's work is supported by NSF of China under grant number 12071216.

†School of Mathematics, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China. Email: wangxp@nuaa.edu.cn

The dynamics approaching the steady-state solution u_h^*, that is $L_h u_h^* = g_h(u_h^*)$, of this dissipative system (1.1) satisfies the following *original* energy dissipation law

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \Big\langle \frac{\delta E}{\delta u_h}, \frac{\mathrm{d}u_h}{\mathrm{d}t} \Big\rangle = -\Big\langle \frac{\mathrm{d}u_h}{\mathrm{d}t}, \frac{\mathrm{d}u_h}{\mathrm{d}t} \Big\rangle \leq 0. \tag{1.3}$$

In simulating the semilinear parabolic problems (1.1) and related gradient flow problems (1.2), explicit exponential (including exponential integrating factor and exponential time differencing) integrators turned out to be very competitive, see [1–6, 14–19, 33]. For a detailed overview of such integrators and their implementation, we refer to [6, 8, 14, 16]. The main idea behind these methods is to treat the linear part of problem exactly and the nonlinearity in an explicit way and dates back to the 1960s, see [2,3,12,20,29,31,32]. For stiff problems, Hochbruck and Ostermann [13] constructed explicit exponential Runge-Kutta (also called exponential time differencing Runge-Kutta, ETDRK) methods with stiff orders up to four and established the convergence in an abstract Banach space framework of sectorial operators and locally Lipschitz continuous nonlinearities. Luan and Ostermann [26] showed that there does not exist an EERK method of order five with less than or equal to six stages and constructed a fifth-order method with eight stages for semilinear parabolic problems. For the stability properties of EERK methods, Maset and Zennaro [28] derived sufficient conditions of unconditional contractivity and unconditional asymptotic stability and investigated some popular EERK methods with respect to the two stability properties.

In the past decade, the explicit ETDRK methods [2, 3, 14] became popular in simulating gradient flow problems, see [5,6,15–17,21,25,33,34], in the context of partial differential equations. One of main concerns is whether these ETD type methods can preserve the decaying of original energy $E[u_h(t)]$. Although the first-order ETD1 method has been proven in [5, 6, 15] to preserve the energy decaying, the energy dissipation property of high-order EERK methods seems theoretically challenging due to their multi-stage nature. Very recently, the second-order ETD2RK method in [3] has been shown to preserve the original energy decaying of the scalar gradient system [9] and the matrix gradient system [25]. These works are theoretically interesting, while their analysis may be limited since the proofs for the energy decaying heavily rely on technical skills and would be difficult to extend for other ETD methods or general situations, such as the parameterized EERK methods constructed by Hochbruck and Ostermann [13, 14].

In this article, we will focus on whether and to what extent the multi-stage EERK methods preserve the original energy dissipation law (1.3). In the next section, a unified theoretical framework for the stage energy dissipation property of EERK methods is established by constructing the differential forms of EERK methods and a new concept, namely average dissipation rate, is introduced for these multi-stage methods to evaluate the overall energy dissipation rate of an EERK method such that one can choose proper parameters in some parameterized EERK methods or compare different kinds of EERK methods. Our main results are stated in Theorem 2.1 and Lemma 2.2.

As applications of our theory, three parameterized second-order EERK methods, including the widespread ETD2RK scheme [3] and the three-stage method by Strehmel and Weiner [31], are discussed in Section 3. Some popular methods are evaluated and suggested for practical numerical simulations, see Table 1, in which the abscissa choices for the contractivity and the energy stability of three second-order EERK methods are summarized. Section 4 addresses four third-order EERK methods, including the ETD3RK [3], ETD2CF3 [2] and two parameterized methods developed by Hochbruck and Ostermann [13]. Table 2 collects some abscissa choices for the energy stability of four third-order EERK methods. Numerical experiments are presented in Section 5 to support our theory. Short comments on four fourth-order EERK methods from [3, 13, 19, 31] and some concluding remarks on the new theory are presented in the last section.

2 Stage energy laws of EERK methods

2.1 General class of EERK methods

Let u_h^k be the numerical approximation of $u_h(t_k)$ at the grid point t_k for $0 \le k \le N$. To integrate the semilinear parabolic problem (1.1) from the discrete time t_{n-1} ($n \ge 1$) to the next grid point $t_n = t_{n-1}+\tau$, the construction of one-step EERK methods (typically, τ also represents a variable-step size) starts from the following variation-of-constants formula

$$u_h(t_{n-1}+\tau) = e^{-\tau L_h} u_h(t_{n-1}) + \int_0^\tau e^{-(\tau-\sigma)L_h} g_h\left[u_h(t_{n-1}+\sigma)\right] \mathrm{d}\sigma.$$

Let $U^{n,i}$ be the approximation of $u_h(t_{n-1}+c_i\tau)$ at the abscissas $c_1 := 0$, $c_i \in (0,1]$ for $2 \le i \le s$, and $c_{s+1} := 1$. By replacing τ by $c_i\tau$ to define the internal stages $t_{n-1}+c_i\tau$, one can construct the following general class of EERK methods:

$$U^{n,1} = u_h^{n-1}, \tag{2.1a}$$

$$U^{n,i+1} = \chi_{i+1}(-\tau L_h)U^{n,1} + \tau \sum_{j=1}^{i} a_{i+1,j}(-\tau L_h) g_h(U^{n,j}), \quad 1 \le i \le s-1, \tag{2.1b}$$

$$U^{n,s+1} = \chi(-\tau L_h)U^{n,1} + \tau \sum_{j=1}^{s} b_j(-\tau L_h) g_h(U^{n,j}), \tag{2.1c}$$

$$u_h^n = U^{n,s+1}. \tag{2.1d}$$

The method coefficients χ_i, χ, a_{ij} and b_j are constructed from linear combinations of the entire functions $\varphi_j(z)$ and scaled versions thereof. These functions are given by

$$\varphi_0(z) = e^z \quad \text{and} \quad \varphi_j(z) := \int_0^1 e^{(1-s)z} \frac{s^{j-1}}{(j-1)!}\,\mathrm{d}s \quad \text{for } z \in \mathbb{C} \text{ and } j \ge 1, \tag{2.2}$$

which satisfy the recursion formula

$$\varphi_{k+1}(z) = \frac{\varphi_k(z) - 1/k!}{z} \quad \text{for } k \ge 0. \tag{2.3}$$

Here the involved matrix functions $\varphi_j(-\tau L_h)$ are defined on the spectrum of $-\tau L_h$, that is, the values $\{\varphi_j(\lambda_k) : 1 \le k \le \dim(-\tau L_h)\}$ exist, where λ_k are the eigenvalues of $-\tau L_h$ and thus $\varphi_j(\lambda_k)$ are the eigenvalues of $\varphi_j(-\tau L_h)$. More properties on the matrix functions can be found in [11, Theorem 1.13], and, typically in this article, $f(-\tau L_h)$ is a positive definite operator if the given entire function f is positive.

Always we assume that $\chi_i(0) = 1$ and $\chi(0) = 1$ for consistency. This scheme (2.1) reduces to an explicit Runge-Kutta method with coefficients $a_{ij} := a_{ij}(0)$ and $b_j := b_j(0)$ if we put $L_h = 0$. The latter method will be called the *underlying explicit Runge-Kutta method* henceforth. We suppose throughout the paper that the underlying Runge-Kutta method satisfies

$$\sum_{j=1}^{s} b_j(0) = 1 \quad \text{and} \quad \sum_{j=1}^{i-1} a_{ij}(0) = c_i \quad \text{for } i = 1, 2, \cdots, s,$$

which makes it invariant under the transformation of (2.1) to the non-autonomous system. A desirable property of numerical methods is that they preserve equilibria u_h^* of (1.2). Requiring

$U^{n,i} = u_h^n = u_h^*$ for all i and $n \geq 0$ immediately yields the necessary and sufficient conditions. It turns out that the method coefficients have to satisfy

$$\sum_{j=1}^{s} b_j(z) = \frac{\chi(z)-1}{z} \quad \text{and} \quad \sum_{j=1}^{i-1} a_{ij}(z) = \frac{\chi_i(z)-1}{z} \quad \text{for } i = 1, 2, \cdots, s. \tag{2.4}$$

Without further mention, we consider the methods with $\chi(z) = e^z$ and $\chi_i(z) = e^{c_i z}$ for $1 \leq i \leq s$. With the help of (2.4), the functions χ_i and χ can be eliminated in (2.1). The numerical scheme (2.1) then takes the form

$$U^{n,i+1} = U^{n,1} + \tau \sum_{j=1}^{i} a_{i+1,j}(-\tau L_h)\left[g_h(U^{n,j}) - L_h U^{n,1}\right] \quad \text{for } 1 \leq i \leq s-1, \tag{2.5a}$$

$$U^{n,s+1} = U^{n,1} + \tau \sum_{j=1}^{s} b_j(-\tau L_h)\left[g_h(U^{n,j}) - L_h U^{n,1}\right]. \tag{2.5b}$$

To simplify our notations, define

$$a_{s+1,j}(z) := b_j(z), \quad 1 \leq j \leq s. \tag{2.6}$$

Then the EERK method (2.5) applying to (1.1) reads

$$U^{n,i+1} = U^{n,1} + \tau \sum_{j=1}^{i} a_{i+1,j}(-\tau L_h)\left[g_h(U^{n,j}) - L_h U^{n,1}\right] \quad \text{for } 1 \leq i \leq s. \tag{2.7}$$

Always, we assume that $a_{k+1,k}(z) \neq 0$ for any $1 \leq k \leq s$. The associated Butcher tableau reads, where we use the abbreviations $a_{ij} := a_{ij}(-\tau L_h)$,

$$\begin{array}{c|ccccc} c_1 & 0 & & & & \\ c_2 & a_{21} & 0 & & & \\ c_3 & a_{31} & a_{32} & 0 & & \\ \vdots & \vdots & \vdots & \ddots & \ddots & \\ c_s & a_{s,1} & a_{s,2} & \cdots & a_{s,s-1} & 0 \\ \hline & a_{s+1,1} & a_{s+1,2} & \cdots & a_{s+1,s-1} & a_{s+1,s} \end{array}.$$

2.2 Our theoretical framework

Motivated by Du et al. [5,6], we introduce the stabilized operators with a parameter $\kappa \geq 0$,

$$L_\kappa := L_h + \kappa I \quad \text{and} \quad g_\kappa(u) := g_h(u) + \kappa u, \tag{2.8}$$

such that the problem (1.1) becomes the stabilized version

$$u_h'(t) = -L_\kappa u_h(t) + g_\kappa(u_h), \quad u_h(t_0) = u_h^0. \tag{2.9}$$

Thus, applying (2.7) to (2.9), we have the following EERK method

$$U^{n,i+1} = U^{n,1} + \sum_{j=1}^{i} a_{i+1,j}(-\tau L_\kappa)\left[\tau g_\kappa(U^{n,j}) - \tau L_\kappa U^{n,1}\right] \quad \text{for } 1 \leq i \leq s. \tag{2.10}$$

To make our idea more concise, we assume further that the nonlinear function g_h is Lipschitz continuous with a constant $\ell_g > 0$, cf. [30] or the recent discussions in [9, subsection 2.2]. In theoretical manner, the stabilization parameter κ in (2.8) is chosen properly large $\kappa \geq 2\ell_g$, see Remark 1 for an alternative choice, to enhance the dissipation of linear part so that the nonlinear growth of g_h can be formally controlled in the numerical analysis. In this sense, if an EERK method is proven to maintain the original energy dissipation law (1.3) unconditionally, we mean that this EERK method can be stabilized by setting a properly large parameter κ (which may not be necessary in actual calculations). To derive the energy dissipation law of the general EERK method (2.10), we need the following result. The proof is standard and we include it for completeness.

Lemma 2.1. *If g_h is Lipschitz-continuous with a constant $\ell_g > 0$ and $\kappa \geq 2\ell_g$, then*

$$\big\langle u - v, g_\kappa(v) - \frac{1}{2}L_\kappa(u+v)\big\rangle \leq E[v] - E[u],$$

where the energy E is defined in (1.2).

Proof. Since g_h is Lipschitz continuous, [30, Lemma 2.8.20] gives

$$\big\langle u - v, g_h(v)\big\rangle \leq \big\langle G_h(v) - G_h(u), 1\big\rangle + \ell_g\big\|u - v\big\|^2. \tag{2.11}$$

It follows that

$$\begin{aligned}&\big\langle u - v, g_\kappa(v) - \frac{\kappa}{2}(u+v)\big\rangle = \big\langle u - v, g_h(v) - \frac{\kappa}{2}(u-v)\big\rangle\\ =&\big\langle u - v, g_h(v)\big\rangle - \frac{\kappa}{2}\big\|u - v\big\|^2 \leq \big\langle G_h(v) - G_h(u), 1\big\rangle - \frac{1}{2}(\kappa - 2\ell_g)\big\|u - v\big\|^2.\end{aligned}$$

Also, it is easy to know that

$$\big\langle u - v, \frac{\kappa}{2}(u+v) - \frac{1}{2}L_\kappa(u+v)\big\rangle = \frac{1}{2}\big\langle v - u, L_h(u+v)\big\rangle = \frac{1}{2}\langle v, L_h v\rangle - \frac{1}{2}\langle u, L_h u\rangle.$$

Adding up the above two results yields the claimed inequality and completes the proof. □

Our theoretical framework contains three main steps:

(1) Compute difference coefficients: we introduce a class of difference coefficients, for $i = 1, 2, \cdots, s$,

$$\underline{a}_{i+1,i}(z) := a_{i+1,i}(z) \quad \text{and} \quad \underline{a}_{i+1,j}(z) := a_{i+1,j}(z) - a_{i,j}(z) \quad \text{for } 1 \leq j \leq i-1. \tag{2.12}$$

It is not difficult to derive from (2.10) that

$$\delta_\tau U^{n,i+1} = \sum_{j=1}^{i} \underline{a}_{i+1,j}(-\tau L_\kappa)\left[\tau g_\kappa(U^{n,j}) - \tau L_\kappa U^{n,1}\right] \quad \text{for } 1 \leq i \leq s, \tag{2.13}$$

where the (stage) time difference $\delta_\tau U^{n,i+1} := U^{n,i+1} - U^{n,i}$ for $1 \leq i \leq s$. The associated Butcher difference (Butcher-Diff) tableau reads

Butcher-Diff tableau:

$$\begin{array}{c|ccccc} c_1 & 0 & & & & \\ c_2 & \underline{a}_{21} & 0 & & & \\ c_3 & \underline{a}_{31} & \underline{a}_{32} & 0 & & \\ \vdots & \vdots & \vdots & \ddots & \ddots & \\ c_s & \underline{a}_{s,1} & \underline{a}_{s,2} & \cdots & \underline{a}_{s,s-1} & 0 \\ \hline & \underline{a}_{s+1,1} & \underline{a}_{s+1,2} & \cdots & \underline{a}_{s+1,s-1} & \underline{a}_{s+1,s} \end{array} .$$

(2) Determine DOC kernels and differential form: we introduce the so-called discrete orthogonal convolution (DOC) kernels $\underline{\theta}_{k,j}(z)$ with respect to the coefficient $\underline{a}_{ij}$, cf. [22–24],

$$\underline{\theta}_{k,k}(z) := \frac{1}{\underline{a}_{k+1,k}(z)} \quad \text{and} \quad \underline{\theta}_{k,j}(z) := -\sum_{\ell=j+1}^{k} \underline{\theta}_{k,\ell}(z)\frac{\underline{a}_{\ell+1,j}(z)}{\underline{a}_{j+1,j}(z)} \quad \text{for } 1 \le j \le k-1. \tag{2.14}$$

It is easy to check the following discrete orthogonal identity,

$$\sum_{\ell=j}^{m} \underline{\theta}_{m,\ell}(z)\underline{a}_{\ell+1,j}(z) \equiv \delta_{mj} \quad \text{for } 1 \le j \le m \le s, \tag{2.15}$$

where δ_{mj} is the Kronecker delta symbol with $\delta_{mj} = 0$ if $j \neq m$. Multiplying the above equation (2.13) by the DOC kernels (matrices) $\underline{\theta}_{k,i}(-\tau L_\kappa)$, and summing i from 1 to k, one can apply the discrete orthogonal identity (2.15) to find that

$$\begin{aligned}
\sum_{i=1}^{k} \underline{\theta}_{k,i}(-\tau L_\kappa)\delta_\tau U^{n,i+1} &= \sum_{i=1}^{k} \underline{\theta}_{k,i}(-\tau L_\kappa) \sum_{j=1}^{i} \underline{a}_{i+1,j}(-\tau L_\kappa)\left[\tau g_\kappa(U^{n,j}) - \tau L_\kappa U^{n,1}\right] \\
&= \sum_{j=1}^{k}\sum_{i=j}^{k} \underline{\theta}_{k,i}(-\tau L_\kappa)\underline{a}_{i+1,j}(-\tau L_\kappa)\left[\tau g_\kappa(U^{n,j}) - \tau L_\kappa U^{n,1}\right] \\
&= \tau g_\kappa(U^{n,k}) - \tau L_\kappa U^{n,1} \\
&= \tau g_\kappa(U^{n,k}) - \tau L_\kappa U^{n,k+1} + \tau L_\kappa \sum_{\ell=1}^{k} \delta_\tau U^{n,\ell+1} \quad \text{for } 1 \le k \le s.
\end{aligned}$$

Thus we have an equivalent form (differential form) of the EERK method (2.10)

$$\sum_{\ell=1}^{k} d_{k\ell}(-\tau L_\kappa)\delta_\tau U^{n,\ell+1} = \tau g_\kappa(U^{n,k}) - \frac{\tau}{2} L_\kappa(U^{n,k+1} + U^{n,k}) \quad \text{for } 1 \le k \le s, \tag{2.16}$$

where the functions $d_{k\ell}$ are defined by

$$d_{k\ell}(z) := \underline{\theta}_{k\ell}(z) + \frac{z}{2}\big(2 - \delta_{k\ell}\big) \quad \text{for } 1 \le \ell \le k \le s \quad \text{and} \quad d_{k\ell}(z) := 0 \quad \text{for } \ell > k. \tag{2.17}$$

The associated lower triangular matrix $D := (d_{k\ell})_{s\times s}$ is called the differentiation matrix. Always, we denote the symmetric part $\mathcal{S}(D;z) := \frac{1}{2}[D(z) + D(z)^T]$.

(3) Establish stage energy dissipation law: this process is standard and we have the following result, which simulates the original energy dissipation law (1.3) at all stages.

Theorem 2.1. *If $\mathcal{S}(D;z)$ is positive (semi-)definite, the EERK method* (2.10) *preserves the original energy dissipation law* (1.3) *at all stages without any time-step constraints,*

$$E[U^{n,j+1}] - E[U^{n,1}] \le -\frac{1}{\tau}\sum_{k=1}^{j} \Big\langle \delta_\tau U^{n,k+1}, \sum_{\ell=1}^{k} d_{k\ell}(-\tau L_\kappa)\delta_\tau U^{n,\ell+1} \Big\rangle \quad \textit{for } 1 \le j \le s, \tag{2.18}$$

and in particular, by taking $j := s$,

$$E[u_h^n] - E[u_h^{n-1}] \le -\frac{1}{\tau}\sum_{k=1}^{s} \Big\langle \delta_\tau U^{n,k+1}, \sum_{\ell=1}^{k} d_{k\ell}(-\tau L_\kappa)\delta_\tau U^{n,\ell+1} \Big\rangle \quad \textit{for } n \ge 1.$$

Proof. Making the inner product of the equivalent form (2.16) with $\frac{1}{\tau}\delta_\tau U^{n,k+1}$ and summing k from $k=1$ to j, one can find that

$$\frac{1}{\tau}\sum_{k=1}^{j}\Big\langle \delta_\tau U^{n,k+1}, \sum_{\ell=1}^{k} d_{k\ell}(-\tau L_\kappa)\delta_\tau U^{n,\ell+1}\Big\rangle = \sum_{k=1}^{j}\Big\langle \delta_\tau U^{n,k+1}, g_\kappa(U^{n,k}) - \frac{1}{2}L_\kappa(U^{n,k+1}+U^{n,k})\Big\rangle$$

for $1\le j\le s$. Lemma 2.1 yields the following energy dissipation law at each stage

$$E[U^{n,j+1}] - E[U^{n,1}] + \frac{1}{\tau}\sum_{k=1}^{j}\Big\langle \delta_\tau U^{n,k+1}, \sum_{\ell=1}^{k} d_{k\ell}(-\tau L_\kappa)\delta_\tau U^{n,\ell+1}\Big\rangle \le 0$$

for $1\le j\le s$. It completes the proof. □

For $1\le j\le s$, let $D_j := D[1:j,1:j]$ be the j-th sequential sub-matrix of the matrix D. We denote further that $\delta_\tau\vec{U}_{n,j+1} := (\delta_\tau U^{n,2}, \delta_\tau U^{n,3}, \cdots, \delta_\tau U^{n,j+1})^T$. The above stage energy dissipation law (2.18) can be formulated as

$$E[U^{n,j+1}] - E[U^{n,1}] \le -\frac{1}{\tau}\Big\langle \delta_\tau\vec{U}_{n,j+1}, D_j(-\tau L_\kappa)\delta_\tau\vec{U}_{n,j+1}\Big\rangle \quad \text{for } 1\le j\le s. \tag{2.19}$$

After the completion of this article, we are informed that, by computing the original energy difference $E[u_h^n] - E[u_h^{n-1}]$ with a key inequality, Fu, Shen and Yang independently derived the same sufficient condition of Theorem 2.1, cf. [10, Theorem 2.1], to ensure that the EERK method (2.10) maintains the decreasing of original energy, that is, $E[u_h^n]\le E[u_h^{n-1}]$. In the following subsection, we will introduce a simple indicator for evaluating to what extent the multi-stage EERK method (2.10) preserves the original energy dissipation law (1.3).

2.3 Averaged dissipation rate

Theorem 2.1 shows that the EERK method (2.10) is unconditionally energy stable if the differentiation matrix $D(z)$ is semi-positive definite, that is, all eigenvalues $\lambda_i(z)$ $(i=1,2,\cdots,s)$ of the symmetric part $\mathcal{S}(D;z)$ are nonnegative. A necessary condition is that the average eigenvalue is nonnegative,

$$\overline{\lambda}(z) := \frac{1}{s}\sum_{i=1}^{s}\lambda_i(z) = \frac{1}{s}\mathrm{tr}\big(D(z)\big) \ge 0.$$

If $\lambda_{\min}\le\lambda_i(z)\le\lambda_{\max}$ $(i=1,2,\cdots,s)$ for any $z\le 0$, one has

$$\lambda_{\min}\langle\vec{v},\vec{v}\rangle \le \langle\vec{v}, D(-\tau L_\kappa)\vec{v}\rangle \le \lambda_{\max}\langle\vec{v},\vec{v}\rangle.$$

Then, according to (2.19), the overall energy dissipation rate of the energy $E[u_h^n]$ could be roughly estimated by the average eigenvalue $\overline{\lambda}(z)$ of $\mathcal{S}(D;z)$. If $\overline{\lambda}(z)\ge 0$, one could use the following *average dissipation rate*

$$\mathcal{R}(z) := \frac{1}{s}\mathrm{tr}\big(D(z)\big) \quad \text{for } z\le 0, \tag{2.20}$$

to examine the energy dissipation behaviors among different methods, see detailed arguments for second-order EERK methods in the next section. By using the definitions (2.17) and (2.14) to compute the diagonal elements $d_{kk}(z)$ for $1\le k\le s$, it is not difficult to obtain the following result.

Lemma 2.2. *If the EERK method* (2.10) *preserves the original energy dissipation law* (1.3) *unconditionally, then the average dissipation rate is nonnegative, that is,*

$$\mathcal{R}(z) = \frac{z}{2} + \frac{1}{s}\sum_{i=1}^{s} \frac{1}{a_{i+1,i}(z)} \geq 0 \quad \textit{for } z \leq 0.$$

Typically, if $\mathcal{R}(z) > 1$, the discrete energy $E[u_h^n]$ decays faster than the continuous counterpart $E[u_h(t_n)]$ and the dynamics approaching the steady-state solution appears a time "ahead" effect. If $0 < \mathcal{R}(z) < 1$, the discrete energy $E[u_h^n]$ may decay slower and the dynamics appears a time "delay" effect. In general, a time-stepping method is a "good" candidate to preserve the original energy dissipation law (1.3) unconditionally if the average dissipation rate $\mathcal{R}(z)$ is nonnegative for $z \leq 0$ and is as close to 1 as possible within properly large range of z. Lemma 2.2 provides us a simple criterion to evaluate the overall energy dissipation rate of an EERK method and then choose proper parameters in some parameterized EERK methods or compare different EERK methods.

Remark 1. *The differential form* (2.16) *and the associated differentiation matrix* $D(z)$ *of the EERK method* (2.10) *would be "optimal" to evaluate the energy dissipation property although they are not unique. A direct choice is to retain only the pure implicit form of stiff term, that is,*

$$\sum_{\ell=1}^{k} \tilde{d}_{k\ell}(-\tau L_\kappa)\delta_\tau U^{n,\ell+1} = \tau g_\kappa(U^{n,k}) - \tau L_\kappa U^{n,k+1} \quad \textit{for } 1 \leq k \leq s, \tag{2.21}$$

where the elements of differentiation matrix $\widetilde{D} := (\tilde{d}_{k\ell})_{s\times s}$ *are defined by*

$$\tilde{d}_{k\ell}(z) := \underline{\theta}_{k\ell}(z) + z \quad \textit{for } 1 \leq \ell \leq k \leq s \quad \textit{and} \quad \tilde{d}_{k\ell}(z) := 0 \quad \textit{for } \ell > k. \tag{2.22}$$

If $\mathcal{S}(\widetilde{D}; z)$ *is positive (semi-)definite, one can follow the proof of Theorem 2.1 to get*

$$\begin{aligned} E[U^{n,j+1}] - E[U^{n,1}] \leq & -\frac{1}{\tau}\sum_{k=1}^{j}\Big\langle \delta_\tau U^{n,k+1}, \sum_{\ell=1}^{k} \tilde{d}_{k\ell}(-\tau L_\kappa)\delta_\tau U^{n,\ell+1}\Big\rangle \\ & -\frac{1}{\tau}\sum_{k=1}^{j}\Big\langle \delta_\tau U^{n,k+1}, \frac{1}{2}\tau L_\kappa \delta_\tau U^{n,k+1}\Big\rangle \quad \textit{for } 1 \leq j \leq s, \end{aligned} \tag{2.23}$$

in which the following result similar to Lemma 2.1 has been used,

$$\langle u - v, g_\kappa(v) - L_\kappa u\rangle \leq E[v] - E[u] - \frac{1}{2}\langle u - v, L_\kappa(u - v)\rangle \quad \textit{for } \kappa \geq \ell_g.$$

It is easy to see that the positive (semi-)definiteness of $\mathcal{S}(\widetilde{D}; z)$ *is much severer than the condition of Theorem 2.1 because the energy dissipation estimate* (2.23) *ignores the dissipation effect of the last term compared with* (2.18)*. Correspondingly, the overall dissipation rate will be also underestimated via the average dissipation rate* $\widetilde{\mathcal{R}}(z)$*, that is,*

$$\widetilde{\mathcal{R}}(z) := \frac{1}{s}\mathrm{tr}\big(\widetilde{D}(z)\big) = z + \frac{1}{s}\sum_{i=1}^{s} \frac{1}{a_{i+1,i}(z)} < \mathcal{R}(z) \quad \textit{for } z < 0.$$

In this situation, one may make misjudgment on the energy dissipation property of EERK methods. For example, consider the one-parameter EERK2 method (3.2) *described below with*

$$\widetilde{\mathcal{R}}(c_2, z) := z + \frac{1}{2c_2\varphi_1(c_2 z)} + \frac{c_2}{2\varphi_2(z)}.$$

It is easy to know that $\lim_{z\to-\infty}\widetilde{\mathcal{R}}(c_2,z)=-\infty$ *if* $c_2\in(0,1)$*, while* $\lim_{z\to-\infty}\widetilde{\mathcal{R}}(1,z)=\frac{1}{2}$*. This directly leads to incorrect conclusion that the EERK2 method with* $c_2=1$ *is the only possible case to preserve the energy dissipation law* (1.3)*; In contrast, Corollary 3.1 says that the EERK2 method preserves the energy dissipation law* (1.3) *unconditionally for* $c_2\in[\frac{1}{2},1]$*. In summary, the condition of Theorem 2.1 is nearly "optimal" although we can not claim that the positive semi-definiteness of differentiation matrix* $D(z)$ *is also necessary to the energy stability of the EERK method (the only loss of dissipation rate comes from the inequality* (2.11) *for controlling the nonlinear growth).*

2.4 Simple case: ETD1

To end this section, we consider a simple case with $s=1$. The only reasonable choice is the exponential forward Euler [13] or ETD1 [3,5] method with stiff order one. Applied to (2.9), it is

$$\delta_\tau u^{n,2}=\varphi_1(-\tau L_\kappa)\left[\tau g_\kappa(u^{n,1})-\tau L_\kappa u^{n,1}\right] \tag{2.24}$$

or, recalling the recursive formula (2.3),

$$u^{n,2}=\varphi_0(-\tau L_\kappa)u^{n,1}+\tau\varphi_1(-\tau L_\kappa)g_\kappa(u^{n,1}). \tag{2.25}$$

The associated Butcher and Butcher-Diff tableaux are the same, that is,

$$\text{ETD1 Butcher or Butcher-Diff:}\quad \begin{array}{c|c} 0 & 0 \\ \hline & \varphi_1 \end{array}.$$

The definition (2.17) gives

$$D^{(1)}=(d_{11}^{(1)})\quad\text{with}\quad d_{11}^{(1)}(z)=\frac{z}{2}+\frac{1}{\varphi_1(z)}=\frac{z(1+e^{-z})}{2(1-e^{-z})}\geq 1\quad\text{for } z\leq 0.$$

Here and hereafter, the superscript (p) is always used to indicate the order of the method, that is to say, $D^{(p)}$ and $\mathcal{R}^{(p)}$ denote the associated differential matrix and the average dissipation rate, respectively, of a formal p-th order EERK method. Obviously, Theorem 2.1 yields

Corollary 2.1. *The exponential forward Euler* (2.24) *preserves the energy dissipation law* (1.3)*,*

$$E[u_h^n]-E[u_h^{n-1}]\leq-\frac{1}{\tau}\big\langle\delta_\tau u^n,D^{(1)}(-\tau L_\kappa)\big(\delta_\tau u^n\big)\big\rangle\quad\textit{for } n\geq 1.$$

By the definition (2.20), one has $\mathcal{R}^{(1)}(z):=d_{11}^{(1)}(z)$ such that $\mathcal{R}^{(1)}(z)\geq 1$ for any $z<0$ and $\lim_{z\to-\infty}\mathcal{R}^{(1)}(z)=+\infty$. It means that the dissipation rate of the discrete energy $E[u_h^n]$ approaches the original rate as the step size $\tau\to 0$; while the exponential forward Euler (2.24) always generates a time "ahead" (compared with the continuous counterpart $E[u_h(t_n)]$) for any time-step sizes.

By the form (2.25), it is easy to find that the ETD1 method is unconditionally contractive, also see [28]. The contractivity of EERK methods is essential to preserve the maximum bound principle of semilinear parabolic problems, cf. [5,6,21,25,34] and references therein; while detailed discussions are out of our current scope in this article.

3 Discrete energy laws of second-order methods

Second-order methods require two internal stages, $s=2$, at least. Hochbruck and Ostermann [13] derived the following stiff order conditions (the stiff order describes the behavior of the local error

independently of the norm of the matrix L_κ) with a parameter c_2 ($0 < c_2 \leq 1$)

$$a_{31}(-\tau L_\kappa) + a_{32}(-\tau L_\kappa) = \varphi_1(-\tau L_\kappa), \tag{3.1a}$$
$$a_{32}(-\tau L_\kappa)c_2 = \varphi_2(-\tau L_\kappa), \tag{3.1b}$$
$$a_{21}(-\tau L_\kappa) = c_2\varphi_1(-c_2\tau L_\kappa). \tag{3.1c}$$

They lead to the following one-parameter family of second-order EERK (EERK2) method with the following Butcher tableau

$$\begin{array}{c|cc} 0 & & \\ c_2 & c_2\varphi_{1,2} & \\ \hline & \varphi_1 - \frac{1}{c_2}\varphi_2 & \frac{1}{c_2}\varphi_2 \end{array}, \tag{3.2}$$

where the notations $\varphi_{i,j}$ are defined by

$$\varphi_{i,j} := \varphi_{i,j}(-\tau L_\kappa) = \varphi_i(-c_j\tau L_\kappa), \quad i \geq 0,\ 1 \leq j \leq s+1. \tag{3.3}$$

Note that, these abbreviations will be also used in the Butcher tableaus below. This EERK2 method (3.2) fulfills all conditions in (3.1) and thus is stiff order two. If the abscissa $c_2 = 1$, it reduces to the so-called ETD2RK [3, 5, 9] with the following form

$$U^{n,2} = \varphi_0(-\tau L_\kappa)U^{n,1} + \tau\varphi_1(-\tau L_\kappa)g_\kappa(U^{n,1}), \tag{3.4a}$$
$$U^{n,3} = U^{n,2} + \tau\varphi_2(-\tau L_\kappa)\left[g_\kappa(U^{n,2}) - g_\kappa(U^{n,1})\right]. \tag{3.4b}$$

This case is also the only scenario to ensure the unconditional contractivity of EERK2 method (3.2), cf. [28], due to the fact $\varphi_1(z) \geq \varphi_2(z)$ for $z \leq 0$.

By weakening the condition (3.1b) to $a_{32}(0)c_2 = \varphi_2(0) = \frac{1}{2}$, one has a one-parameter weak variant (called EERK2-w in short)

$$\begin{array}{c|cc} 0 & & \\ c_2 & c_2\varphi_{1,2} & \\ \hline & (1 - \frac{1}{2c_2})\varphi_1 & \frac{1}{2c_2}\varphi_1 \end{array}. \tag{3.5}$$

Although the EERK2-w method (3.5) does not have stiff order two, it achieves stiff convergence order two [13, Section 5.1] under certain requirements (boundedness) on the discrete operator τL_κ. In the following, we consider their stage energy dissipation properties in simulating (2.9).

3.1 EERK2 method

To establish the stage energy laws, we present the Butcher-Diff tableau

$$\text{EERK2 Butcher-Diff:} \quad \begin{array}{c|cc} 0 & & \\ c_2 & c_2\varphi_{1,2} & \\ \hline & \varphi_1 - \frac{1}{c_2}\varphi_2 - c_2\varphi_{1,2} & \frac{1}{c_2}\varphi_2 \end{array}. \tag{3.6}$$

By the procedure (2.14), one can compute the associated DOC kernels

$$\underline{\theta}_{11}(z) = \frac{1}{c_2\varphi_1(c_2z)}, \quad \underline{\theta}_{22}(z) = \frac{c_2}{\varphi_2(z)} \quad \text{and} \quad \underline{\theta}_{21}(z) = \frac{c_2\varphi_1(c_2z) - \varphi_1(z) + \frac{1}{c_2}\varphi_2(z)}{\varphi_2(z)\varphi_1(c_2z)}.$$

The definition (2.17) gives the following differentiation matrix

$$D^{(2)}(c_2,z) := \begin{pmatrix} \frac{1}{c_2\varphi_1(c_2z)} + \frac{z}{2} & 0 \\ \frac{c_2\varphi_1(c_2z)-\varphi_1(z)+\frac{1}{c_2}\varphi_2(z)}{\varphi_2(z)\varphi_1(c_2z)} + z & \frac{c_2}{\varphi_2(z)} + \frac{z}{2} \end{pmatrix}.$$

Now we consider the matrix $\mathcal{S}(D^{(2)};c_2,z)$, the symmetric part of $D^{(2)}(c_2,z)$. The first leading principal minor reads

$$\mathrm{Det}\big[\mathcal{S}(D_1^{(2)};c_2,z)\big] = d_{11}^{(2)}(c_2,z) = \frac{z(e^{c_2z}+1)}{2(e^{c_2z}-1)} \geq \frac{1}{c_2} \quad \text{for } c_2 \in (0,1] \text{ and } z \leq 0.$$

The second leading principal minor (determinant) of $\mathcal{S}(D^{(2)};c_2,z)$ is given by

$$\mathrm{Det}\big[\mathcal{S}(D^{(2)};c_2,z)\big] = \frac{(e^{c_2z}-1)^{-2}z^2}{4(z-e^z+1)^2} g_{21}(c_2,z) \quad \text{for } z<0,$$

where the auxiliary function g_{21} is defined by

$$\begin{aligned} g_{21}(c_2,z) :=& 2c_2ze^{(c_2+2)z} - c_2^2z^2e^{2c_2z} - 2c_2ze^{c_2z+z}(1-(c_2-1)z) - e^{2z}(c_2^2z^2+1) \\ &+ 2e^z(1-(c_2-1)z) + (z+1)((2c_2-1)z-1) \quad \text{for } z<0. \end{aligned} \tag{3.7}$$

Now we develop a technique of comparison function to handle the function g_{21}.

Proposition 3.1. *The function g_{21} in (3.7) is positive for $c_2 \in [\frac{1}{2},1]$ and $z<0$.*

Proof. The condition $c_2 \in [\frac{1}{2},1]$ comes from the simple fact $\lim_{z\to-\infty} g_{21}(c_2,z)/z^2 = 2c_2-1 \geq 0$. To handle g_{21}, we consider a comparison function (by setting $c_2 := \frac{1}{2}$ in all exponents of g_{21})

$$\begin{aligned} g_{21}^*(c_2,z) :=& 2c_2ze^{\frac{5z}{2}} - c_2^2z^2e^z - 2c_2ze^{\frac{3z}{2}}(1-(c_2-1)z) - e^{2z}(c_2^2z^2+1) \\ &+ 2e^z(1-(c_2-1)z) + (z+1)((2c_2-1)z-1) \quad \text{for } z<0. \end{aligned}$$

It is easy to check that $g_{21}(c_2,z) \geq g_{21}^*(c_2,z)$ for $c_2 \in [\frac{1}{2},1]$ and $z<0$. Actually,

$$\begin{aligned} g_{21}(c_2,z) - g_{21}^*(c_2,z) &= c_2e^z(e^{z/2}-e^{c_2z})\big[c_2z^2(e^{-z/2}+e^{c_2z-z}-2) - 2z(e^z-1-z)\big] \\ &\geq c_2e^z(e^{z/2}-e^{c_2z})\big[c_2z^2(e^{-z/2}-1) - 2z(e^z-z-1)\big] \geq 0 \end{aligned}$$

due to the facts $e^{-z/2}-1>0$ and $e^z-z-1>0$ for $z<0$.

Note that, g_{21}^* is a concave, quadratic polynomial with respect to c_2, that is,

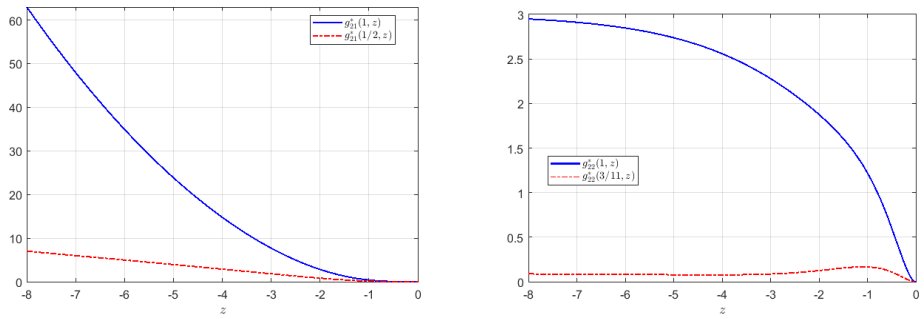
$$g_{21}^*(c_2,z) = -e^z(e^{z/2}-1)^2z^2c_2^2 + 2z(1-e^{\frac{3z}{2}})(z+1-e^z)c_2 - (z+1-e^z)^2 \quad \text{for } z<0.$$

Moreover, one can check that (technical details are omitted here), cf. Figure 1(a),

$$g_{21}^*(1,z) > 0 \quad \text{and} \quad g_{21}^*(\tfrac{1}{2},z) > 0 \quad \text{for } z<0.$$

They imply that $g_{21}^*(c_2,z)>0$ and then $g_{21}(c_2,z)>0$ for $c_2 \in [\frac{1}{2},1]$ and $z<0$. □

Proposition 3.1 shows that $\mathrm{Det}\big[\mathcal{S}(D^{(2)};c_2,z)\big] > 0$ for $c_2 \in [\frac{1}{2},1]$ and $z<0$. Thus, the sequential principal minors of $\mathcal{S}(D^{(2)};c_2,z)$ are positive and then the differentiation matrix $D^{(2)}(c_2,z)$ is positive definite. Theorem 2.1 gives the following result.



(a) $g_{21}^*(c_2,z)$ (b) $g_{22}^*(c_2,z)$

Figure 1: Curves of comparison functions g_{21}^* and g_{22}^*.

Corollary 3.1. *The EERK2 method* (3.2) *with* $c_2 \in [\frac{1}{2}, 1]$ *preserves the energy dissipation law* (1.3) *unconditionally at all stages in the sense that*

$$\begin{aligned} E[U^{n,j+1}] - E[U^{n,1}] \le & -\frac{1}{\tau}\sum_{k=1}^{j}\Big\langle \delta_\tau U^{n,k+1}, \sum_{\ell=1}^{k} d_{k\ell}^{(2)}(c_2, -\tau L_\kappa)\delta_\tau U^{n,\ell+1}\Big\rangle \\ = & -\frac{1}{\tau}\Big\langle \delta_\tau \vec{U}_{n,j+1}, D_j^{(2)}(c_2, -\tau L_\kappa)\delta_\tau \vec{U}_{n,j+1}\Big\rangle \quad \textit{for } 1 \le j \le 2. \end{aligned}$$

According to Lemma 2.2, the EERK2 method (3.2) has the average dissipation rate

$$\mathcal{R}^{(2)}(c_2, z) := \frac{z}{2} + \frac{1}{2c_2\varphi_1(c_2 z)} + \frac{c_2}{2\varphi_2(z)} \quad \text{for } c_2 \in (0,1] \text{ and } z \le 0. \tag{3.8}$$

It is different for different choices of c_2. One has

$$\lim_{z\to 0}\mathcal{R}^{(2)}(c_2,z) = \frac{1}{2c_2} + c_2 \quad \text{and} \quad \lim_{z\to-\infty}\mathcal{R}^{(2)}(c_2,z) = +\infty \quad \text{for } c_2 \in (0,1].$$

For properly large time-step sizes, the ETD2RK method (3.4) with the case $c_2 = 1$ has the largest dissipation rate, see Figure 2 (a), while the average dissipation rate of the case $c_2 = \frac{1}{2}$ is much more closer to 1. The former is preferred to preserve the unconditional contractivity, while the latter would be preferred to preserve the energy dissipation law (1.3) unconditionally. As seen, $\mathcal{R}^{(2)}(c_2, z) > 1$ for $c_2 \in [\frac{1}{2}, 1]$ and $z \le 0$, so that the EERK2 method (3.2) always generates a time "ahead" in simulating the gradient system (2.9).

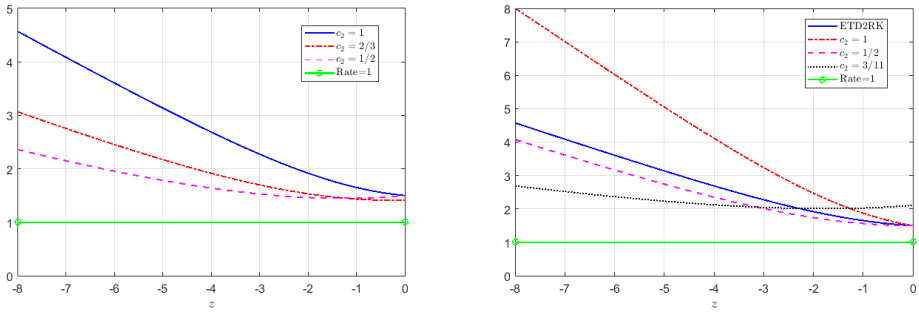
3.2 Weak variants of EERK2 method

For the EERK2-w methods with tableau (3.5), the associated Butcher-Diff tableau

$$\text{EERK2-w Butcher-Diff:} \quad \begin{array}{c|cc} 0 & & \\ c_2 & c_2\varphi_{1,2} & \\ \hline & (1-\frac{1}{2c_2})\varphi_1 - c_2\varphi_{1,2} & \frac{1}{2c_2}\varphi_1 \end{array} .$$

By the procedure (2.14), one can compute the associated DOC kernels

$$\underline{\theta}_{11}^{(2,w)} = \frac{1}{c_2\varphi_1(c_2 z)}, \quad \underline{\theta}_{22}^{(2,w)} = \frac{2c_2}{\varphi_1(z)} \quad \text{and} \quad \underline{\theta}_{21}^{(2,w)} = \frac{2c_2}{\varphi_1(z)} + \frac{(1-2c_2)}{c_2\varphi_1(c_2 z)}.$$



(a) EERK2 (b) EERK2-w

Figure 2: Averaged dissipation rates $\mathcal{R}^{(2)}(c_2, z)$ and $\mathcal{R}^{(2,w)}(c_2, z)$ for different abscissas c_2.

The definition (2.17) gives the following one-parameter differentiation matrix

$$D^{(2,w)}(c_2, z) := \begin{pmatrix} \frac{1}{c_2\varphi_1(c_2 z)} + \frac{z}{2} & 0 \\ \frac{2c_2}{\varphi_1(z)} + \frac{(1-2c_2)}{c_2\varphi_1(c_2 z)} + z & \frac{2c_2}{\varphi_1(z)} + \frac{z}{2} \end{pmatrix}.$$

It is not difficult to check that

$$\mathrm{Det}\big[\mathcal{S}(D_1^{(2,w)}; c_2, z)\big] = \frac{z(e^{c_2 z}+1)}{2(e^{c_2 z}-1)} \geq \frac{1}{c_2} \quad \text{for } c_2 \in (0,1] \text{ and } z \leq 0.$$

To handle the second leading principal minor, we need the following result.

Proposition 3.2. *For the abscissa $c_2 \in [\frac{3}{11}, 1]$ and $z < 0$, it holds that*

$$g_{22}(c_2, z) := -4c_2^2(e^{c_2 z} - e^z)^2 + 4c_2(1-e^z)(1-e^{c_2 z + z}) - (1-e^z)^2 > 0. \tag{3.9}$$

Proof. We consider the following comparison function

$$g_{22}^*(c_2, z) := -4c_2^2\big(e^{\frac{3z}{11}} - e^z\big)^2 + 4c_2(1-e^z)\big(1-e^{\frac{24z}{11}}\big) - (1-e^z)^2 \quad \text{for } z < 0.$$

For $c_2 \in [\frac{3}{11}, 1]$ and $z < 0$, it is obvious that

$$g_{22}(c_2, z) - g_{22}^*(c_2, z) = 4c_2\big(e^{\frac{3z}{11}} - e^{c_2 z}\big)e^z\Big[c_2\big(e^{-\frac{8z}{11}} + e^{(c_2-1)z} - 2\big) + (1-e^z)\Big] \geq 0.$$

Note that, g_{22}^* is a concave, quadratic polynomial with respect to c_2 due to $\partial_{c_2}^2 g_{22}^* < 0$. Reminding that $\lim\limits_{z\to-\infty} g_{22}^*(c_2, z) = 4c_2 - 1 > 0$ for $c_2 \in [\frac{3}{11}, 1]$, it is not difficult to check that, cf. Figure 1(b),

$$g_{22}^*(1, z) > 0 \quad \text{and} \quad g_{22}^*(\tfrac{3}{11}, z) > 0 \quad \text{for } z < 0.$$

They imply that $g_{22}^*(c_2, z) > 0$ and then $g_{22}(c_2, z) \geq g_{22}^*(c_2, z) > 0$ for $c_2 \in [\frac{3}{11}, 1]$ and $z < 0$. □

Reminding the auxiliary function (3.9) and Proposition 3.2, we know that the second leading principal minor of $\mathcal{S}(D^{(2,w)}; c_2, z)$ is positive, that is,

$$\mathrm{Det}\big[\mathcal{S}\big(D^{(2,w)}; c_2, z\big)\big] = \frac{z^2 g_{22}(c_2, z)}{4(e^z-1)^2(e^{c_2 z}-1)^2} > 0 \quad \text{for } c_2 \in [\tfrac{3}{11}, 1] \text{ and } z < 0.$$

Thus the sequential principal minors of $\mathcal{S}(D^{(2,w)}; c_2, z)$ are positive and then the differentiation matrix $D^{(2,w)}(c_2, z)$ is positive definite. Theorem 2.1 gives the following result.

Corollary 3.2. *The EERK2-w method* (3.5) *with* $c_2 \in [\frac{3}{11}, 1]$ *preserves the original energy dissipation law* (1.3) *at all stages in the sense that*

$$E[U^{n,j+1}] - E[U^{n,1}] \le -\frac{1}{\tau}\sum_{k=1}^{j}\Big\langle \delta_\tau U^{n,k+1}, \sum_{\ell=1}^{k} d_{k\ell}^{(2,w)}(c_2, -\tau L_\kappa)\delta_\tau U^{n,\ell+1}\Big\rangle \quad \textit{for } 1 \le j \le 2.$$

It is worth mentioning that the choice $c_2 \in [\frac{1}{2}, 1]$ makes the EERK2-w method (3.5) unconditionally contractive [28]. Similar to the ETD2RK scheme (3.4), the EERK2-w method (3.5) arrives at the following weak variant of ETD2RK scheme for $c_2 \in [\frac{1}{2}, 1]$,

$$U^{n,2} = \varphi_0(-c_2\tau L_\kappa)U^{n,1} + \tau c_2\varphi_1(-c_2\tau L_\kappa)g_\kappa(U^{n,1}), \tag{3.10a}$$

$$U^{n,3} = \varphi_0(-\tau L_\kappa)U^{n,1} + (1 - \frac{1}{2c_2})\tau\varphi_1(-\tau L_\kappa)g_\kappa(U^{n,1}) + \frac{\tau}{2c_2}\varphi_1(-\tau L_\kappa)g_\kappa(U^{n,2}). \tag{3.10b}$$

As an advantage over the ETD2RK scheme (3.4), the weak variant (3.10) provides more choice of the abscissa c_2 to preserve both the contractivity and energy dissipation law unconditionally.

For the one-parameter EERK2-w method (3.5), the average dissipation rate

$$\mathcal{R}^{(2,w)}(c_2, z) := \frac{z}{2} + \frac{1}{2c_2\varphi_1(c_2 z)} + \frac{c_2}{\varphi_1(z)} \quad \text{for } c_2 \in (0, 1] \text{ and } z \le 0. \tag{3.11}$$

It is easy to find that

$$\lim_{z\to 0}\mathcal{R}^{(2,w)}(c_2, z) = \frac{1}{2c_2} + c_2 \quad \text{and} \quad \lim_{z\to -\infty}\mathcal{R}^{(2,w)}(c_2, z) = +\infty \quad \text{for } c_2 \in (0, 1].$$

For properly large time-step sizes, the case $c_2 = 1$ has the largest dissipation rate, see Figure 2 (b), while the case $c_2 = \frac{1}{2}$ has the smallest rate near $z = 0$ and the case $c_2 = \frac{3}{11}$ has the smallest rate for $z < -3$. More interestingly, the case $c_2 = \frac{1}{2}$ seems superior to the ETD2RK method (3.4) since the dissipation rate of the former is much closer to 1. For all cases $c_2 \in [\frac{3}{11}, 1]$, the average rate $\mathcal{R}^{(2,w)}(c_2, z) > 1$ and the EERK2-w method (3.5) always generates a time "ahead" for the gradient flow system (2.9).

3.3 Remarks for the three-stage EERK method

We present some remarks for the one-parameter family of 3-stage EERK (called EERK2-S in short) method proposed by Strehmel and Weiner [31] with second-order B-consistency,

$$\begin{array}{c|ccc} 0 & & & \\ c_2 & c_2\varphi_{1,2} & & \\ 1 & \varphi_{1,3} - \frac{1}{c_2}\varphi_{2,3} & \frac{1}{c_2}\varphi_{2,3} & \\ \hline & \varphi_1 - \varphi_2 & 0 & \varphi_2 \end{array}. \tag{3.12}$$

These methods satisfy all conditions up to stiff order two, see (4.1a)-(4.1d), described in Section 4. By following the arguments in [28], we know that the EERK2-S method (3.12) with the abscissa $c_2 = 1$ is also unconditionally contractive since $\varphi_1(z) \ge \varphi_2(z)$ for $z \le 0$.

To establish the stage energy laws, we present the Butcher-Diff tableau

$$\text{EERK2-S Butcher-Diff:} \quad \begin{array}{c|ccc} 0 & & & \\ c_2 & c_2\varphi_{1,2} & & \\ 1 & \varphi_1 - \frac{1}{c_2}\varphi_2 - c_2\varphi_{1,2} & \frac{1}{c_2}\varphi_2 & \\ \hline & \frac{1-c_2}{c_2}\varphi_2 & -\frac{1}{c_2}\varphi_2 & \varphi_2 \end{array}.$$

Note that the first three lines of this Butcher-Diff tableau are the same to the Butcher-Diff tableau (3.6) of the EERK2 method (3.2). The stage energies $E[U^{n,j}]$ $(j = 2, 3)$ of the EERK2-S method (3.12) have the same dissipation rates to those of the two-stage EERK2 method. Thus, to preserve the energy dissipation law (1.3) at all stages, the condition $c_2 \in [\frac{1}{2}, 1]$ is also necessary for the EERK2-S method (3.12).

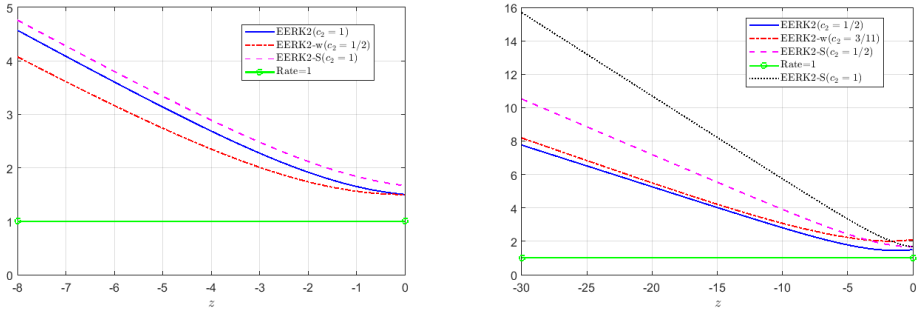
However, the second-order EERK2-S method (3.12) would be not competitive for solving the gradient system (2.9). Actually, one has the average dissipation rate

$$\mathcal{R}^{(2,S)}(c_2, z) = \frac{z}{2} + \frac{1}{3c_2\varphi_1(c_2 z)} + \frac{c_2}{3\varphi_2(z)} + \frac{1}{3\varphi_2(z)} \quad \text{for } z \le 0.$$

It is easy to obtain that

$$\lim_{z\to 0} \mathcal{R}^{(2,S)}(c_2, z) = \frac{2}{3} + \frac{2}{3}\big(c_2 + \frac{1}{2c_2}\big) \quad \text{and} \quad \lim_{z\to -\infty} \mathcal{R}^{(2,S)}(c_2, z) = +\infty \quad \text{for } c_2 \in (0, 1].$$

In Figure 3, we compare the average dissipation rates of the EERK2, EERK2-w and EERK2-S methods. Taking into the contractivity account, we find that the EERK2-w method (3.5) with $c_2 = \frac{1}{2}$ generates the minimum time "ahead" effect since the average dissipation rate $\mathcal{R}^{(2,w)}(\frac{1}{2}, z)$ has the smallest value for any time-step sizes, cf. Figure 3 (a), while the EERK2-S method with $c_2 = 1$ produces the maximum time "ahead" effect. If the contractivity is not considered, Figure 3 (b) suggests that the EERK2 method (3.2) with $c_2 = \frac{1}{2}$ produces the minimum time "ahead" effect among the three methods since the average dissipation rate $\mathcal{R}^{(2)}(\frac{1}{2}, z)$ has the smallest value.



(a) With contractivity (b) General case

Figure 3: Dissipation rate comparisons of EERK2, EERK2-w and EERK2-S methods.

Therefore, from the perspective of preserving the original energy dissipation rate, the EERK2-S methods (3.12) with $c_2 \in [\frac{1}{2}, 1]$ would be not competitive among second-order EERK methods for the gradient system (2.9). As noted in [13, subsection 5.2], maybe more importantly, the three-stage the EERK2-S method (3.12) is not preferred for semilinear parabolic problems because they are more computationally expensive than the two-stage methods in previous subsections.

As the end of this section, Table 1 lists the abscissa choices for the contractivity and the energy stability of three second-order EERK methods. Observations indicate that the abscissa condition maintaining the energy stability is often different from that preserving the contractivity, with the former being weaker than the latter. For the gradient flow system (1.2), certain time-stepping method without the contractivity does not necessarily violate the energy dissipation law (1.3). Actually, next subsection shows that there are many third-order EERK methods preserving the energy dissipation law (1.3); however, as pointed out by [28, Section 4], scholars have not found that any EERK methods of stiff convergence order greater than two can maintain the contractivity.

Table 1: Choice of the abscissa c_2 in second-order EERK methods.

Method	Contractivity	Energy law preserving	Best dissipation rate
EERK2 (3.2)	$c_2 = 1$	$c_2 \in [\frac{1}{2}, 1]$	$c_2 = \frac{1}{2}$
EERK2-w (3.5)	$c_2 \in [\frac{1}{2}, 1]$	$c_2 \in [\frac{3}{11}, 1]$	$c_2 = \frac{3}{11}$ or $\frac{1}{2}$
EERK2-S (3.12)	$c_2 = 1$	at least $c_2 \in [\frac{1}{2}, 1]$	$c_2 = \frac{1}{2}$

4 Discrete energy laws of third-order methods

Third-order methods require three internal stages, $s = 3$, at least. The order conditions for three-stage methods (2.7) are given by [13]

$$\begin{aligned}
a_{41}(-\tau L_\kappa) + a_{42}(-\tau L_\kappa) + a_{43}(-\tau L_\kappa) &= \varphi_1(-\tau L_\kappa), && \text{(4.1a)}\\
a_{42}(-\tau L_\kappa)c_2 + a_{43}(-\tau L_\kappa)c_3 &= \varphi_2(-\tau L_\kappa), && \text{(4.1b)}\\
a_{21}(-\tau L_\kappa) &= c_2\varphi_1(-c_2\tau L_\kappa), && \text{(4.1c)}\\
a_{31}(-\tau L_\kappa) + a_{32}(-\tau L_\kappa) &= c_3\varphi_1(-c_3\tau L_\kappa), && \text{(4.1d)}\\
a_{42}(-\tau L_\kappa)c_2^2 + a_{43}(-\tau L_\kappa)c_3^2 &= 2\varphi_3(-\tau L_\kappa), && \text{(4.1e)}\\
a_{42}(-\tau L_\kappa)Jc_2^2\varphi_2(-c_2\tau L_\kappa) + a_{43}(-\tau L_\kappa)J\psi_{2,3} &= 0, && \text{(4.1f)}
\end{aligned}$$

where J denotes arbitrary bounded operator and

$$\psi_{2,3} := c_3^2\varphi_2(-c_3\tau L_\kappa) - c_2 a_{32}(-\tau L_\kappa).$$

As pointed out in [13, subsection 5.2], condition (4.1f) can be fulfilled by setting (I) $a_{42} = 0$ and $\psi_{2,3} = 0$; or (II) $a_{42} = \gamma a_{43}$ and $c_2^2\varphi_2 + \gamma\psi_{2,3} = 0$.

The choice (I) leads to the following one-parameter family of method (called EERK3-1 in short)

$$\begin{array}{c|ccc}
0 & & & \\
c_2 & c_2\varphi_{1,2} & & \\
\frac{2}{3} & \frac{2}{3}\varphi_{1,3} - \frac{4}{9c_2}\varphi_{2,3} & \frac{4}{9c_2}\varphi_{2,3} & \\
\hline
 & \varphi_1 - \frac{3}{2}\varphi_2 & 0 & \frac{3}{2}\varphi_2
\end{array} . \tag{4.2}$$

The other choice (II) leads to the two-parameter family of method (called EERK3-2)

$$\begin{array}{c|ccc}
0 & & & \\
c_2 & c_2\varphi_{1,2} & & \\
c_3 & c_3\varphi_{1,3} - a_{32} & \gamma c_2\varphi_{2,2} + \frac{c_3^2}{c_2}\varphi_{2,3} & \\
\hline
 & \varphi_1 - a_{42} - a_{43} & \frac{\gamma}{\gamma c_2 + c_3}\varphi_2 & \frac{1}{\gamma c_2 + c_3}\varphi_2
\end{array} , \tag{4.3}$$

where the parameter $\gamma := \frac{(3c_3-2)c_3}{(2-3c_2)c_2}$ for $c_2 \neq \frac{2}{3}$ and $c_2 \neq c_3$ (to ensure $a_{32} \neq 0$). Also, it is to set $c_3 \neq \frac{2}{3}$ since it degrades into the EERK3-1 method (4.2) if $c_3 = \frac{2}{3}$.

In the literature, there are some related three-stage methods that involve the function φ_3. Cox

and Matthews [3] constructed the ETD3RK method with Butcher tableau

$$\begin{array}{c|ccc} 0 & & & \\ \frac{1}{2} & \frac{1}{2}\varphi_{1,2} & & \\ 1 & -\varphi_{1,3} & 2\varphi_{1,3} & \\ \hline & 4\varphi_3-3\varphi_2+\varphi_1 & -8\varphi_3+4\varphi_2 & 4\varphi_3-\varphi_2 \end{array}. \tag{4.4}$$

This method satisfies the conditions (4.1a)-(4.1d), while the conditions (4.1e)-(4.1f) are satisfied only in a very weak form (setting $L_\kappa = 0$). As a variant of the commutator-free Lie group CF3 method due to Celledoni, Marthinsen, and Owren [2], the so-called ETD2CF3 method is given by

$$\begin{array}{c|ccc} 0 & & & \\ \frac{1}{3} & \frac{1}{3}\varphi_{1,2} & & \\ \frac{2}{3} & \frac{2}{3}\varphi_{1,3}-\frac{4}{3}\varphi_{2,3} & \frac{4}{3}\varphi_{2,3} & \\ \hline & \varphi_1-\frac{9}{2}\varphi_2+9\varphi_3 & 6\varphi_2-18\varphi_3 & -\frac{3}{2}\varphi_2+9\varphi_3 \end{array}. \tag{4.5}$$

The ETD2CF3 method satisfies the conditions (4.1a)-(4.1c), while conditions (4.1d) and (4.1f) are satisfied in the weak form (setting $L_\kappa = 0$).

4.1 Simplified procedure and two simple cases

As seen, the method coefficients (and the corresponding differentiation matrix as well) of high-order EERK methods are always more complex than those of lower order methods. Therefore certain symbolic computation system, such as the *Wolfram Mathematica*, will be employed to assist our theoretical derivations for stage energy dissipation laws. To do that, we summarize our theoretical framework in the subsections 2.1 and 2.2 as follows:

Step1. Compute the differentiation matrix $D(z)$ defined via (2.17), or

$$D(z) = \left(E_s^{-1}A(z)\right)^{-1} + zE_s - \frac{z}{2}I = A(z)^{-1}E_s + zE_s - \frac{z}{2}I,$$

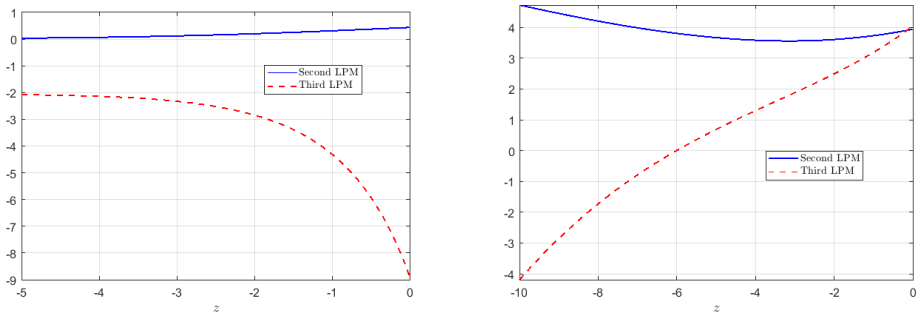
where $A(z) := (\mathrm{a}_{ij})_{s\times s}$ is the coefficient matrix with $\mathrm{a}_{ij} := a_{i+1,j}(z)$ for $1 \le i, j \le s$ and $E_s := (1_{i\ge j})_{s\times s}$ is the lower triangular matrix full of element 1.

Step2. Compute the j-th leading principal minors $\mathrm{Det}\left[\mathcal{S}(D_j; z)\right]$ for $1 \le j \le s$ and check the positive definiteness of the symmetric matrix $\mathcal{S}(D; z)$ for $z \le 0$.

Step3. Establish the stage energy dissipation laws if $\mathcal{S}(D; z)$ is positive definite and compute the average dissipation rate $\mathcal{R}(z)$ using the coefficients $a_{i+1,i}$ $(1 \le i \le s)$ of the EERK method.

In the following, we will apply the procedure (**Step1**)-(**Step3**) to examine the above third-order EERK methods and pick out those preserving the energy dissipation law (1.3) unconditionally.

Before searching for some third-order EERK methods that preserve the energy dissipation law (1.3) unconditionally, we first examine the well-known ETD3RK (4.4) and ETD2CF3 (4.5) methods by computing the associated differential matrices $D^{(3,e)}(z)$ and $D^{(3,f)}(z)$. Figure 4 shows that the determinant of $\mathcal{S}(D^{(3,e)}; z)$ is always negative for $z < 0$ and the differential matrix $D^{(3,e)}(z)$ is not positive definite for any $z < 0$. The determinant of $\mathcal{S}(D^{(3,f)}; z)$ is always negative and the differential matrix $D^{(3,f)}(z)$ is not positive definite for $z < -6$. That is to say, when applied to the gradient system (1.2), both methods may destroy the energy dissipation law (1.3) (especially for large time-step sizes) no matter how large the stabilization parameter κ we set in (2.8).



(a) ETD3RK [3] (b) ETD2CF3 [2]

Figure 4: Leading principal minors (LPM) of associated differential matrices.

4.2 One-parameter EERK3 methods

For the c_2-parameterized EERK3-1 method (4.2), one has the following differentiation matrix

$$D^{(3,1)}(c_2,z) := \begin{pmatrix} \frac{1}{c_2\varphi_1(c_2z)} + \frac{z}{2} & 0 & 0 \\ \frac{9c_2}{4\varphi_2(\frac{2z}{3})} + \frac{1}{c_2\varphi_1(c_2z)} - \frac{3\varphi_1(\frac{2z}{3})}{2\varphi_2(\frac{2z}{3})\varphi_1(c_2z)} + z & \frac{9c_2}{4\varphi_2(\frac{2z}{3})} + \frac{z}{2} & 0 \\ \frac{2c_2\varphi_1(c_2z)-2\varphi_1(z)+3\varphi_2(z)}{3c_2\varphi_1(c_2z)\varphi_2(z)} + z & \frac{2}{3\varphi_2(z)} + z & \frac{2}{3\varphi_2(z)} + \frac{z}{2} \end{pmatrix}. \tag{4.6}$$

It is not difficult to check that

$$\mathrm{Det}\big[\mathcal{S}(D_1^{(3,1)};c_2,z)\big] = \frac{z(e^{c_2z}+1)}{2(e^{c_2z}-1)} \geq \frac{1}{c_2} \quad \text{for } c_2 \in (0,1] \text{ and } z \leq 0.$$

By using the auxiliary function g_{31} in (A.5) and Proposition A.1, one has

$$\mathrm{Det}\big[\mathcal{S}\big(D_2^{(3,1)};c_2,z\big)\big] = \frac{(e^{c_2z}-1)^{-2}z^2}{4(2z-3e^{\frac{2z}{3}}+3)^2} g_{31}(c_2,c_2,z) > 0 \quad \text{for } c_2 \in [\tfrac{4}{9},1] \text{ and } z < 0.$$

Also, the third leading principal minor of $\mathcal{S}(D^{(3,1)};c_2,z)$ is given by

$$\mathrm{Det}\big[\mathcal{S}\big(D^{(3,1)};c_2,z\big)\big] = \frac{z^4(e^{c_2z}-1)^{-2}g_{32}(c_2,c_2,z)}{72(z-e^z+1)^2(2z-3e^{\frac{2z}{3}}+3)^2} > 0 \quad \text{for } c_2 \in [\tfrac{4}{9},1] \text{ and } z < 0,$$

where g_{32} is defined by (A.5) and Proposition A.2 has been used. Then the sequential principal minors of $\mathcal{S}(D^{(3,1)};c_2,z)$ are positive, and then the differentiation matrix $D^{(3,1)}(c_2,z)$ is positive definite. Theorem 2.1 gives the following result.

Corollary 4.1. *The one-parameter EERK3-1 method* (4.2) *with* $c_2 \in [\frac{4}{9},1]$ *preserves the energy dissipation law* (1.3) *at all stages in the sense that*

$$E[U^{n,j+1}] - E[U^{n,1}] \leq -\frac{1}{\tau}\sum_{k=1}^{j}\Big\langle \delta_\tau U^{n,k+1}, \sum_{\ell=1}^{k} d_{k\ell}^{(3,1)}(c_2,-\tau L_\kappa)\delta_\tau U^{n,\ell+1}\Big\rangle \quad \textit{for } 1 \leq j \leq 3.$$

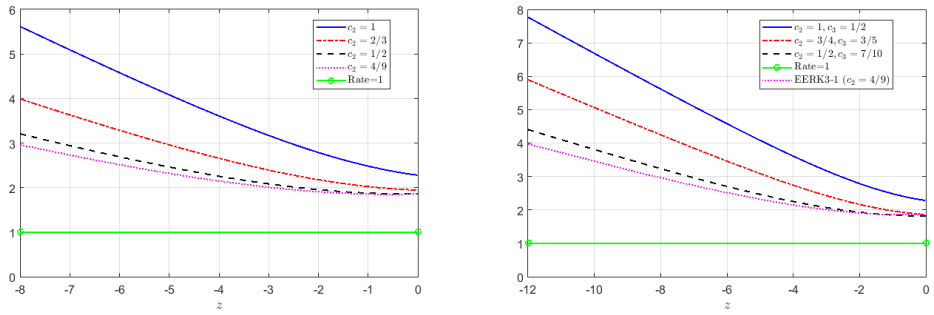
For the one-parameter EERK3-1 method (4.2), the average dissipation rate

$$\mathcal{R}^{(3,1)}(c_2,z) = \frac{z}{2} + \frac{1}{3c_2\varphi_1(c_2z)} + \frac{3c_2}{4\varphi_2(\frac{2z}{3})} + \frac{2}{9\varphi_2(z)}.$$

It is easy to find that

$$\lim_{z\to 0}\mathcal{R}^{(3,1)}(c_2,z)=\frac{3c_2}{2}+\frac{1}{3c_2}+\frac{4}{9}\quad\text{and}\quad\lim_{z\to-\infty}\mathcal{R}^{(3,1)}(c_2,z)=+\infty\quad\text{for } c_2\in[\tfrac{4}{9},1].$$

As seen in Figure 5 (a), the case $c_2=1$ has the largest dissipation rate, while the case $c_2=\frac{4}{9}$ has the smallest rate. For all cases $c_2\in[\frac{4}{9},1]$, the average dissipation rate $\mathcal{R}^{(3,1)}(c_2,z)>1$ and the EERK3-1 method (4.2) always generates a time "ahead" for the gradient system (2.9).



(a) $\mathcal{R}^{(3,1)}(c_2,z)$ (b) $\mathcal{R}^{(3,2)}(c_2,c_3,z)$

Figure 5: Averaged dissipation rates of EERK3-1 and EERK3-2 methods.

4.3 Two-parameter EERK3 methods

For the two-parameter EERK3-2 method (4.3), it is reasonable to properly confine ourselves possible choices of the abscissas c_2 and c_3. According to Lemma 2.2, we consider the following average dissipation rate

$$\mathcal{R}^{(3,2)}(c_2,c_3,z)=z+\frac{1}{3c_2\varphi_1(c_2z)}+\frac{1}{3\gamma c_2\varphi_2(c_2z)+\frac{3c_3^2}{c_2}\varphi_2(c_3z)}+\frac{\gamma c_2+c_3}{3\varphi_2(z)},\tag{4.7}$$

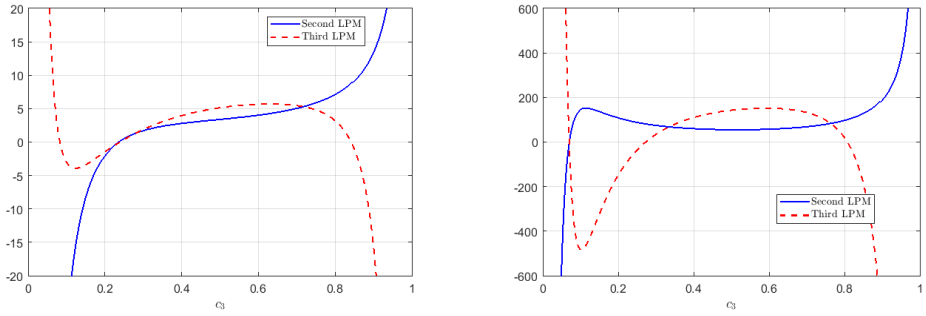
where $\gamma:=\frac{(3c_3-2)c_3}{(2-3c_2)c_2}$ for $c_2\neq\frac{2}{3}$, $c_2\neq c_3$ and $c_3\neq\frac{2}{3}$ (the choice $c_3:=\frac{2}{3}$ gives $\gamma=0$ and the method reduces to the EERK3-1 method (4.2)). It is not difficult to check that the following abscissa (necessary) condition

$$\frac{6c_3(c_2-c_3)}{(3c_2-2)}-1+\frac{2c_2(3c_2-2)}{3c_3(c_2-c_3)}\geq 0\quad\text{for } c_2\neq\frac{2}{3}\text{ and } c_3\neq c_2\tag{4.8}$$

is sufficient to ensure that $\lim_{z\to-\infty}\mathcal{R}^{(3,2)}(c_2,c_3,z)\geq 0$.

In general, the quartic inequality (4.8) gives rise to some theoretical trouble in proving the positive definiteness of the associated differentiation matrix $D^{(3,2)}(c_2,c_3,z)$. Actually, to prove the positivity of the second leading principal minor $\mathrm{Det}\big[\mathcal{S}(D_2^{(3,2)};c_2,c_3,z)\big]$, one has to handle sixth degree polynomials with respect to c_2 or c_3 (while, in the previous subsection 3.1, only second degree polynomials, see Propositions A.1 and A.2, should be handled to determine the positive definiteness of the differentiation matrix $D^{(3,1)}(c_2,z)$ of the EERK3-1 method); and the third leading principal minor $\mathrm{Det}\big[\mathcal{S}(D_3^{(3,2)};c_2,c_3,z)\big]$ involves eighth degree polynomials with respect to c_2 or c_3.

Since we are not able to present a complete discussion on the choices of c_2 and c_3 for the energy stability of EERK3-2 method (4.3), this subsection examines some of concrete examples falling into three cases listed as follows: (a) $c_2=1$ with $c_3\in(0,1]$, (b) $c_2\in(\frac{2}{3},1)$ with $c_3\in(0,c_2)$, and (c) $c_2\in(0,\frac{2}{3})$ with $c_3\in(c_2,1]$. The settings in all cases are necessary for the condition (4.8).



(a) $z_0 = -1$ (b) $z_0 = -10$

Figure 6: Leading principal minors $\mathrm{Det}[\mathcal{S}(D_j^{(3,2)}; 1, c_3, z_0)]$ for $j = 2, 3$.

4.3.1 The case $c_2 = 1$ with $c_3 \in (0, 1]$

At first glance, the condition (4.8) always holds if $c_2 = 1$. We are to choose $c_3 = \frac{1}{2}$ according to some numerical tests, cf. Figure 6 (a)-(b), where the second and third leading principal minors of $\mathcal{S}(D_2^{(3,2)}; 1, c_3, z)$ are depicted for $z_0 = -1$ and $z_0 = -10$, respectively.

It is not difficult to check that

$$\mathrm{Det}\big[\mathcal{S}(D_1^{(3,2)}; 1, \tfrac{1}{2}, z)\big] = \frac{z(e^{z/2}+1)}{2(e^{z/2}-1)} > 0 \quad \text{for } z \le 0.$$

Reminding the auxiliary function g_{41} and g_{42} defined by (B.7)-(B.8), we know that the second and third leading principal minors of $\mathcal{S}(D^{(3,2)}; 1, \frac{1}{2}, z)$ are positive, that is,

$$\mathrm{Det}\big[\mathcal{S}\big(D_2^{(3,2)}; 1, \tfrac{1}{2}, z\big)\big] = \frac{(e^{z/2}-1)^{-2}(e^{z/2}+1)^{-2}z^2}{4(-3z+4e^{z/2}+e^z-5)^2} g_{41}(z) > 0 \quad \text{for } z < 0,$$

$$\mathrm{Det}\big[\mathcal{S}\big(D^{(3,2)}; 1, \tfrac{1}{2}, z\big)\big] = \frac{(e^{z}-1)^{-2}(z-e^z+1)^{-2}z^4}{128(-3z+4e^{z/2}+e^z-5)^2} g_{42}(z) > 0 \quad \text{for } z < 0,$$

where Proposition B.1 has been used. Then the sequential principal minors of $\mathcal{S}(D^{(3,2)}; 1, \frac{1}{2}, z)$ are positive and then the differentiation matrix $D^{(3,2)}(1, \frac{1}{2}, z)$ is positive definite. Theorem 2.1 gives the following result.

Corollary 4.2. *The two-parameter EERK3-2 method* (4.3) *with* $c_2 = 1$ *and* $c_3 = \frac{1}{2}$ *preserves the energy dissipation law* (1.3) *at all stages in the sense that*

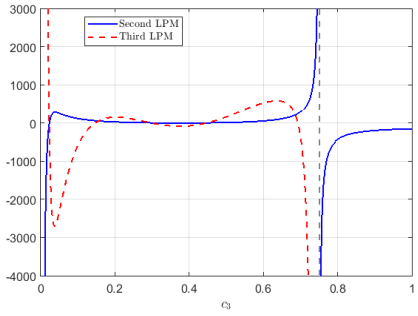
$$E[U^{n,j+1}] - E[U^{n,1}] \le -\frac{1}{\tau}\sum_{k=1}^{j}\Big\langle \delta_\tau U^{n,k+1}, \sum_{\ell=1}^{k} d_{k\ell}^{(3,2)}(1, \tfrac{1}{2}, -\tau L_\kappa)\delta_\tau U^{n,\ell+1}\Big\rangle \quad \textit{for } 1 \le j \le 3.$$

4.3.2 The case $c_2 \in (\frac{2}{3}, 1)$ with $c_3 \in (0, c_2)$

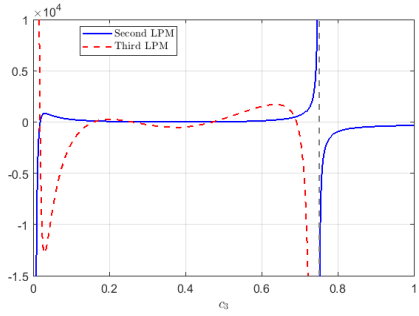
We consider $c_2 = \frac{3}{4}$ and choose $c_3 = \frac{3}{5}$ according to some numerical tests, cf. Figure 7 (a)-(b), where the second and third leading principal minors of $\mathcal{S}(D_2^{(3,2)}; \frac{3}{4}, c_3, z)$ are depicted for $z_0 = -20$ and $z_0 = -30$, respectively.

It is not difficult to check that

$$\mathrm{Det}\big[\mathcal{S}(D_1^{(3,2)}; \tfrac{3}{4}, \tfrac{3}{5}, z)\big] = \frac{z(e^{3z/4}+1)}{2(e^{3z/4}-1)} > 0 \quad \text{for } z \le 0.$$



(a) $z_0 = -20$



(b) $z_0 = -30$

Figure 7: Leading principal minors $\mathrm{Det}[\mathcal{S}(D_j^{(3,2)}; \frac{3}{4}, c_3, z_0)]$ for $j = 2, 3$.

Reminding the auxiliary function g_{51} and g_{52} defined by (B.9)-(B.10), we know that the second and third leading principal minors of $\mathcal{S}(D^{(3,2)}; \frac{3}{4}, \frac{3}{5}, z)$ are positive, that is,

$$\mathrm{Det}\big[\mathcal{S}\big(D_2^{(3,2)}; \tfrac{3}{4}, \tfrac{3}{5}, z\big)\big] = \frac{100(e^{\frac{3z}{4}} - 1)^{-2} z^2}{(27z - 16e^{\frac{3z}{4}} - 25e^{\frac{3z}{5}} + 41)^2} g_{51}(z) > 0 \quad \text{for } z < 0,$$

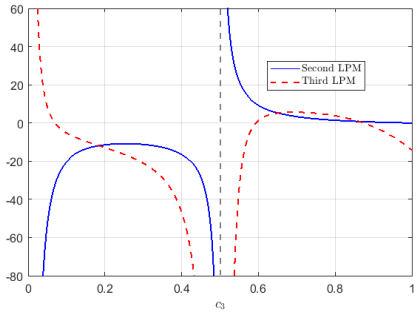
$$\mathrm{Det}\big[\mathcal{S}\big(D^{(3,2)}; \tfrac{3}{4}, \tfrac{3}{5}, z\big)\big] = \frac{300(e^{\frac{3z}{4}} - 1)^{-2}(z - e^z + 1)^{-2} z^4}{(27z - 16e^{\frac{3z}{4}} - 25e^{\frac{3z}{5}} + 41)^2} g_{52}(z) > 0 \quad \text{for } z < 0,$$

where Proposition B.2 has been used. Then the sequential principal minors of $\mathcal{S}(D^{(3,2)}; \frac{3}{4}, \frac{3}{5}, z)$ are positive and then the differentiation matrix $D^{(3,2)}(\frac{3}{4}, \frac{3}{5}, z)$ is positive definite. Theorem 2.1 gives the following result.

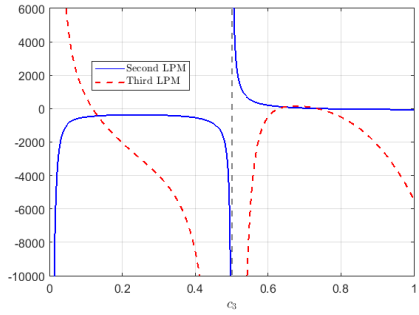
Corollary 4.3. *The two-parameter EERK3-2 method* (4.3) *with* $c_2 = \frac{3}{4}$ *and* $c_3 = \frac{3}{5}$ *preserves the energy dissipation law* (1.3) *at all stages in the sense that*

$$E[U^{n,j+1}] - E[U^{n,1}] \le -\frac{1}{\tau} \sum_{k=1}^{j} \Big\langle \delta_\tau U^{n,k+1}, \sum_{\ell=1}^{k} d_{k\ell}^{(3,2)}(\tfrac{3}{4}, \tfrac{3}{5}, -\tau L_\kappa) \delta_\tau U^{n,\ell+1} \Big\rangle \quad \textit{for } 1 \le j \le 3.$$

4.3.3 The case $c_2 \in (0, \frac{2}{3})$ with $c_3 \in (c_2, 1]$



(a) $z_0 = -1$



(b) $z_0 = -20$

Figure 8: Leading principal minors $\mathrm{Det}[\mathcal{S}(D_j^{(3,2)}; \frac{1}{2}, c_3, z_0)]$ for $j = 2, 3$.

We consider $c_2 = \frac{1}{2}$ and choose the abscissa $c_3 = \frac{7}{10}$ according to some numerical tests, cf. Figure 8 (a)-(b), where the second and third leading principal minors of $\mathcal{S}(D_2^{(3,2)}; \frac{1}{2}, c_3, z)$ are depicted for $z_0 = -1$ and $z_0 = -20$, respectively.

It is not difficult to check that

$$\mathrm{Det}\big[\mathcal{S}(D_1^{(3,2)}; \tfrac{1}{2}, \tfrac{7}{10}, z)\big] = \frac{z(e^{z/2}+1)}{2(e^{z/2}-1)} > 0 \quad \text{for } z \le 0.$$

Reminding the auxiliary function g_{61} and g_{62} defined by (B.11)-(B.12), we know that the second and third leading principal minors of $\mathcal{S}(D^{(3,2)}; \frac{1}{2}, \frac{7}{10}, z)$ are positive, that is,

$$\mathrm{Det}\big[\mathcal{S}(D_2^{(3,2)}; \tfrac{1}{2}, \tfrac{7}{10}, z)\big] = \frac{(e^{z/2}-1)^{-2}z^2}{(21z - 7e^{z/2} - 25e^{\frac{7z}{10}} + 32)^2} g_{61}(z) > 0 \quad \text{for } z < 0,$$

$$\mathrm{Det}\big[\mathcal{S}(D^{(3,2)}; \tfrac{1}{2}, \tfrac{7}{10}, z)\big] = \frac{100(e^{z/2}-1)^{-2}(z-e^z+1)^{-2}z^4}{(21z - 7e^{z/2} - 25e^{\frac{7z}{10}} + 32)^2} g_{62}(z) > 0 \quad \text{for } z < 0,$$

where Proposition B.3 has been used. Then the sequential principal minors of $\mathcal{S}(D^{(3,2)}; \frac{1}{2}, \frac{7}{10}, z)$ are positive and then the differentiation matrix $D^{(3,2)}(\frac{1}{2}, \frac{7}{10}, z)$ is positive definite. Theorem 2.1 gives the following result.

Corollary 4.4. *The two-parameter EERK3-2 method* (4.3) *with* $c_2 = \frac{1}{2}$ *and* $c_3 = \frac{7}{10}$ *preserves the energy dissipation law* (1.3) *at all stages in the sense that*

$$E[U^{n,j+1}] - E[U^{n,1}] \le -\frac{1}{\tau}\sum_{k=1}^{j}\Big\langle \delta_\tau U^{n,k+1}, \sum_{\ell=1}^{k} d_{k\ell}^{(3,2)}(\tfrac{1}{2}, \tfrac{7}{10}, -\tau L_\kappa)\delta_\tau U^{n,\ell+1}\Big\rangle \quad \textit{for } 1 \le j \le 3.$$

By the formula (4.7), we compute the average dissipation rates $\mathcal{R}^{(3,2)}(c_2, c_3, z)$ for the above three examples in Corollaries 4.2, 4.3 and 4.4. For the above three cases of EERK3-2 method (4.3), see Figure 5 (b), the case $c_2 = 1$ with $c_3 = \frac{1}{2}$ has the largest dissipation rate, while the case $c_2 = \frac{1}{2}$ with $c_3 = \frac{7}{10}$ has the smallest dissipation rate. Also, Figure 3 (b) suggests that the EERK3-1 method (3.2) with $c_2 = \frac{4}{9}$ produces the minimum time "ahead" effect among the third-order EERK methods in this section since the average dissipation rate $\mathcal{R}^{(3,1)}(\frac{4}{9}, z)$ has the smallest value. As the end of this section, Table 2 summarizes the abscissa choices for the energy stability of some third-order EERK methods.

Table 2: The parameter choices in third-order EERK methods.

Method	Unconditional energy law preserving	Best dissipation rate
EERK3-1 (4.2)	$c_2 \in [\frac{4}{9}, 1]$	$c_2 = \frac{4}{9}$
EERK3-2 (4.3)	$c_2 = 1$, $c_3 = \frac{1}{2}$ $c_2 = \frac{3}{4}$, $c_3 = \frac{3}{5}$ $c_2 = \frac{1}{2}$, $c_3 = \frac{7}{10}$	$c_2 = \frac{1}{2}$, $c_3 = \frac{7}{10}$
ETD3RK (4.4)	NPD*	–
ETD2CF3 (4.5)	NPD	–

* NPD means that there exists a $z_0 < 0$ such that the associated differential matrix $D(z_0)$ is not positive semi-definite.

5 Numerical experiments

For the sake of generality, we use the Cahn-Hilliard model, $\partial_t u + \epsilon^2\Delta^2 u = \Delta(u^3 - u)$, to perform some numerical tests because this paper mainly focuses on the original energy dissipation properties of various EERK methods (2.7). Let L_h be the discrete matrix of the Laplacian operator $-\Delta$. In such situation, the results of Theorem 2.1 are valid by setting $L_\kappa := \epsilon^2 L_h^2 + \kappa L_h$ and replacing the L^2 inner product $\langle u, v\rangle$ by the H^{-1} inner product $\langle u, v\rangle_{-1} := \langle u, L_h^{-1} v\rangle$ on the zero-mean function space $\{v|\langle v, 1\rangle = 0\}$. That is, we have the following discrete energy law

$$E[U^{n,j+1}] - E[U^{n,1}] \le -\frac{1}{\tau}\sum_{k=1}^{j}\Big\langle \delta_\tau U^{n,k+1}, \sum_{\ell=1}^{k} d_{k\ell}\big(-\epsilon^2\tau L_h^2 - \kappa\tau L_h\big)\delta_\tau U^{n,\ell+1}\Big\rangle_{-1} \quad \text{for } 1 \le j \le s.$$

In the first example, we examine the convergence by considering the EERK2-w methods (3.5) and the EERK3-1 methods (4.2) for different abscissas c_2. In the second example, the energy dissipation rates are demonstrated for different choices of the method parameters including the abscissa c_2, the stabilized parameter κ and the time-step size τ.

5.1 Convergence tests

Example 1. *Consider the Cahn-Hilliard model on $\Omega = (0, 2\pi)$ with $\epsilon = 0.2$ subject to the initial data $u_0 = 0.5\sin(x)$ and Dirichlet boundary condition. Always, we use the center difference approximation with the spacing $h = \pi/320$ for spatial discretization.*

Table 3: EERK2-w errors with different abscissas c_2 for Example 1.

$\tau = 0.01$	$c_2 = 1$		$c_2 = \frac{3}{4}$		$c_2 = \frac{1}{2}$		$c_2 = \frac{3}{11}$	
	$e(\tau)$	Order	$e(\tau)$	Order	$e(\tau)$	Order	$e(\tau)$	Order
τ	6.106e-03	-	5.149e-03	-	4.122e-03	-	3.119e-03	-
$\tau/2$	1.750e-03	1.80	1.462e-03	1.82	1.161e-03	1.83	8.756e-04	1.83
$\tau/4$	4.744e-04	1.88	3.932e-04	1.89	3.098e-04	1.91	2.323e-04	1.91
$\tau/8$	1.220e-04	1.96	1.001e-04	1.97	7.798e-05	1.99	5.756e-05	2.01

We always choose the final time $T = 8$ and the stabilized parameter $\kappa = 2$ in our tests. We run the EERK2-w methods (3.5) with four different abscissas $c_2 = 1$, $\frac{3}{4}$, $\frac{1}{2}$ and $\frac{3}{11}$ for a small time step $\tau = 0.001$. The four schemes work well and the corresponding solution and energy curves (omitted here) are hard to distinguish from each other. The numerical solution of the EERK2-w method with $c_2 = \frac{3}{11}$ and $\tau = 0.01/32$ is taken as the reference solution $u_h^\star$ in the convergence tests. The solution errors recorded in Table 3 are obtained on halving time steps $\tau = 0.01/2^k$ for $k = 0, 1, \cdots, 3$ and the convergence order is computed by Order $\approx \log_2\left(e(\tau)/e(\tau/2)\right)$ where $e(\tau)$ is the L^∞ norm error defined by $e(\tau) := \max_{1\le n\le N}\|u_h^n - u_h^\star\|_\infty$. The numerical results in Table 3 confirm the second-order time accuracy of the EERK2-w methods (3.5).

It is interesting to note that, for each time step, the EERK2-w solutions with different c_2 have slight differences in precision, and the method with $c_2 = \frac{3}{11}$ generates a bit more accurate solution than other cases. This interesting phenomenon is also observed from the solutions (omitted here) of EERK2 methods (3.2): the method with $c_2 = \frac{1}{2}$ generates a bit more accurate solution than other

Table 4: EERK3-1 errors with different abscissas c_2 for Example 1.

$\tau = 0.01$	$c_2 = 1$		$c_2 = \frac{2}{3}$		$c_2 = \frac{1}{2}$		$c_2 = \frac{4}{9}$	
	$e(\tau)$	Order	$e(\tau)$	Order	$e(\tau)$	Order	$e(\tau)$	Order
τ	6.369e-4	-	4.670e-4	-	3.708e-4	-	3.368e-4	-
$\tau/2$	1.107e-4	2.52	7.922e-5	2.56	6.202e-5	2.58	5.604e-5	2.59
$\tau/4$	1.737e-5	2.67	1.218e-5	2.70	9.425e-6	2.72	8.479e-6	2.72
$\tau/8$	2.511e-6	2.79	1.729e-6	2.82	1.321e-6	2.83	1.183e-6	2.84

choices $c_2 > \frac{1}{2}$ including the widespread ETDRK method [3] with $c_2 = 1$, at least for Example 1. Coincidentally, the minimum abscissa choice preserving the energy dissipation law (1.3) are $c_2 = \frac{1}{2}$ for the EERK2 methods (3.2) and $c_2 = \frac{3}{11}$ for the EERK2-w methods (3.5), respectively, see Corollaries 3.1 and 3.2.

The numerical results of EERK3-1 methods (4.2), listed in Table 4, are obtained in similar to those in Table 3. As seen, the EERK3-1 methods always generate third-order solutions, at least for the smooth initial data. Also, we observe that the methods with $c_2 > \frac{4}{9}$ generate a bit less accurate solution than the choice $c_2 = \frac{4}{9}$, which is the minimum abscissa preserving the energy dissipation law (1.3) for the EERK3-1 methods, see Corollary 4.1.

5.2 Energy dissipation property

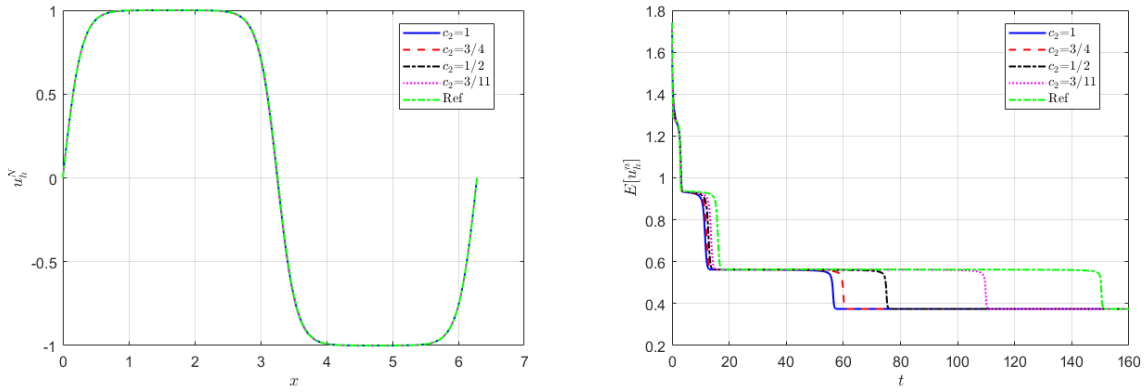
Example 2. *Consider the Cahn-Hilliard model on $\Omega = (0, 2\pi)$ with $\epsilon = 0.2$ and zero-valued Dirichlet boundary condition subject to the following initial data, see [7, guide19],*

$$u_0 = \frac{1}{3}\tanh(2\sin x) - e^{-23.5(x-\frac{\pi}{2})^2} + e^{-27(x-4.2)^2} + e^{-38(x-5.4)^2}.$$

We use the center difference approximation with the spacing $h = \pi/320$ for spatial discretization.

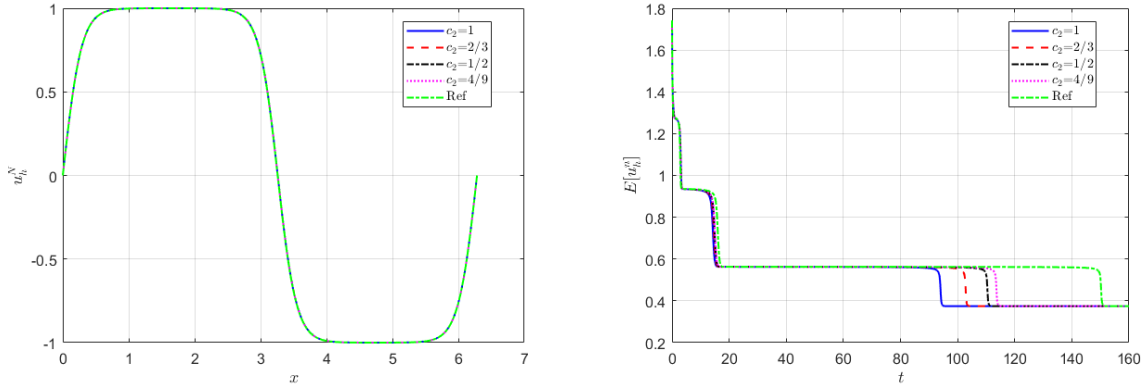
Taking the parameter $\kappa = 2$ and the time-step $\tau = 0.1$, we run the EERK2-w methods (3.5) with four different abscissas $c_2 = 1, \frac{3}{4}, \frac{1}{2}, \frac{3}{11}$, and the EERK3-1 methods (4.2) with four different abscissas $c_2 = 1, \frac{2}{3}, \frac{1}{2}, \frac{4}{9}$ to the final time $T = 160$. The corresponding numerical solution and discrete energy are depicted in Figures 9-10, respectively. Taking a small step size $\tau = 0.001$, we compute the reference solutions and energies (marked by "Ref" here and hereafter) by using the EERK2-w method with $c_2 = \frac{3}{11}$ and the EERK3-1 method with $c_2 = \frac{4}{9}$, respectively. As predicted by Corollaries 3.2 and 4.1, the original energies $E[u_h^n]$ generated by the two methods always decay over the time.

As seen in Figure 9(b), there are some obvious differences in energy dissipation rates for different abscissas c_2. It is not mysterious, according to the discrete energy law in Corollary 3.2, because the EERK2-w methods with different abscissas c_2 have different differentiation matrices $D^{(2,w)}(c_2, z)$. For this example, the discrete energy produced by the case $c_2 = 1$ decays fastest, while that generated by the case $c_2 = \frac{3}{11}$ decays slowest. Qualitatively, they may be explained by the average dissipation rate $\mathcal{R}^{(2,w)}(c_2, z)$ in Figure 2(b), in which we see that $\mathcal{R}^{(2,w)}(1, z)$ has the largest value and $\mathcal{R}^{(2,w)}(\frac{3}{11}, z)$ has the smallest one for properly large $|z| \geq 4$. Similarly, the differences in energy dissipation rates for the EERK3-1 methods with different abscissas c_2, see Figure 10(b), can be attributed to the differences of differentiation matrices $D^{(3,1)}(c_2, z)$ defined in (4.6). Also, they may be qualitatively explained by the average dissipation rate $\mathcal{R}^{(3,1)}(c_2, z)$ in Figure 5(a), in which we see that $\mathcal{R}^{(3,1)}(1, z)$ has the largest value and $\mathcal{R}^{(3,1)}(\frac{4}{9}, z)$ has the smallest one for any $z < 0$.



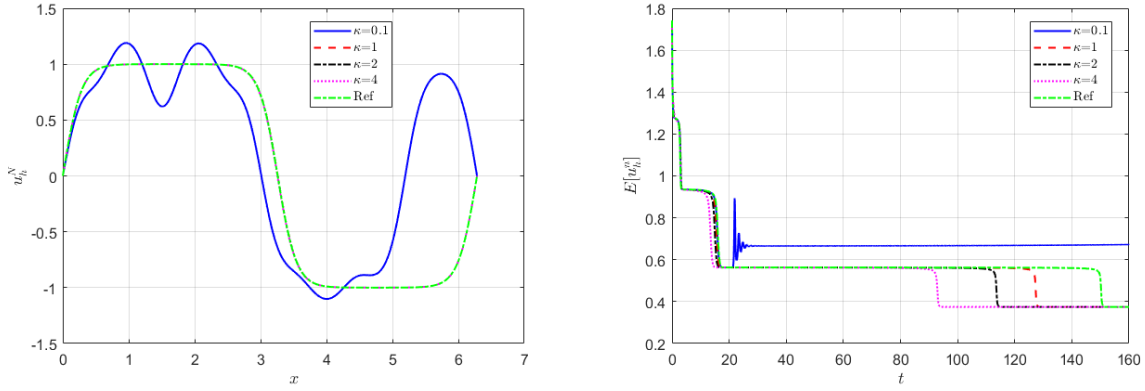
(a) final solution u_h^N (b) energy $E[u_h^n]$

Figure 9: Energy dissipation of the EERK2-w methods (3.5) with $\kappa = 2$ and $\tau = 0.1$.



(a) final solution u_h^N (b) energy $E[u_h^n]$

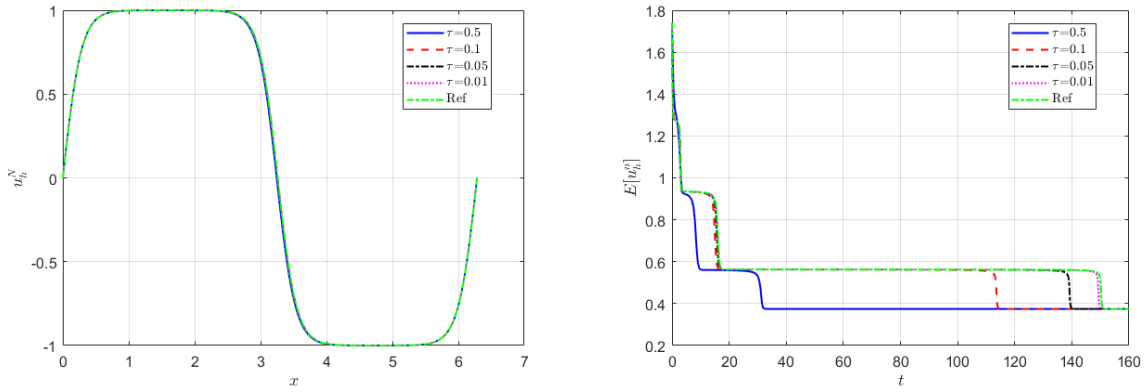
Figure 10: Energy dissipation of the EERK3-1 methods (4.2) with $\kappa = 2$ and $\tau = 0.1$.



(a) final solution u_h^N (b) energy $E[u_h^n]$

Figure 11: Energy dissipation of the EERK3-1 method (4.2) with $c_2 = \frac{4}{9}$ and $\tau = 0.1$.

Obviously, in addition to the different dissipation rates brought by the different choices of c_2, the time step size τ and stabilized parameter κ also have some significant impacts on the discrete energy dissipation property. To explore the influence of stabilized parameter κ, we take a fixed time-step



(a) final solution u_h^N (b) energy $E[u_h^n]$

Figure 12: Energy dissipation of the EERK3-1 method (4.2) with $c_2 = \frac{4}{9}$ and $\kappa = 2$.

$\tau = 0.1$ and run the EERK3-1 method with $c_2 = \frac{4}{9}$ to the final time $T = 160$ for four different parameters $\kappa = 0.1, 1, 2$ and 4, cf. Figure 11, where the reference solution is computed with $\kappa = 4$ and $\tau = 0.001$. The discrete energy for $\kappa = 0.1$ appears non-physical oscillations since the nonlinear stability could not be controlled by the small stabilized parameter. In practice, a properly large κ is always necessary to maintain the stability especially when some large time step τ is employed. With the increase of κ, the energy curve appears some "ahead" effect, that is, the discrete energy dissipates faster as the stabilization parameter κ becomes larger.

Now we fix the stabilized parameter $\kappa = 2$ and run the EERK3-1 method with $c_2 = \frac{4}{9}$ for four different time steps $\tau = 0.5, 0.1, 0.05$ and 0.01, cf. Figure 12, in which the reference solution is obtained with $\tau = 0.001$. We see that, with the increase of time-step size, the energy curve shows some "ahead" effect, that is, the discrete energy dissipate faster as the step size τ becomes larger. Note that, the numerical behaviors in Figure 12(b) and 11(b) would be predictable by the average dissipation rate $\mathcal{R}^{(3,1)}(\frac{4}{9}, z)$, see Figure 5(a), since it is increasing with respect to $|z|$. Actually, we also run the EERK2-w method with $c_2 = \frac{1}{2}$ for Example 2 and find similar behaviors (omitted here) of the discrete energy curves for different time steps τ and different stabilized parameters κ.

6 Fourth-order EERK methods and concluding remarks

We consider firstly three four-stage fourth-order EERK methods from [3, 19, 31]. As noted in [13], these methods do not have the stiff order four although they show a higher order of convergence (generically up to order four) under favorable circumstances. The first one is the following exponential variant of the classical Runge-Kutta method developed by Cox and Matthews [3]

$$\begin{array}{c|cccc} 0 & & & & \\ \frac{1}{2} & \frac{1}{2}\varphi_{1,2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2}\varphi_{1,3} & & \\ 1 & \frac{1}{2}\varphi_{1,3}(\varphi_{0,3}-1) & 0 & \varphi_{1,3} & \\ \hline & \varphi_1 - 3\varphi_2 + 4\varphi_3 & 2\varphi_2 - 4\varphi_3 & 2\varphi_2 - 4\varphi_3 & 4\varphi_3 - \varphi_2 \end{array} . \tag{6.1}$$

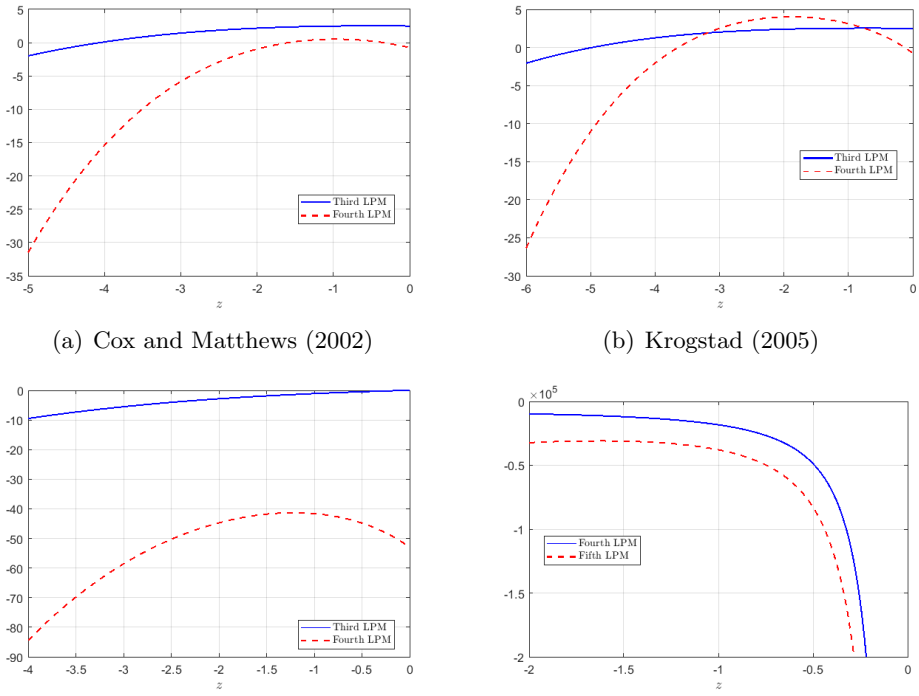
The second one is the Krogstad's method [19] given by

$$\begin{array}{c|cccc} 0 & & & & \\ \frac{1}{2} & \frac{1}{2}\varphi_{1,2} & & & \\ \frac{1}{2} & \frac{1}{2}\varphi_{1,3}-\varphi_{2,3} & \varphi_{2,3} & & \\ 1 & \varphi_{1,4}-2\varphi_{2,4} & 0 & 2\varphi_{2,4} & \\ \hline & \varphi_1-3\varphi_2+4\varphi_3 & 2\varphi_2-4\varphi_3 & 2\varphi_2-4\varphi_3 & -\varphi_2+4\varphi_3 \end{array} . \tag{6.2}$$

The last is the following method from Strehmel and Weiner [31, Example 4.5.5],

$$\begin{array}{c|cccc} 0 & & & & \\ \frac{1}{2} & \frac{1}{2}\varphi_{1,2} & & & \\ \frac{1}{2} & \frac{1}{2}\varphi_{1,3}-\frac{1}{2}\varphi_{2,3} & \frac{1}{2}\varphi_{2,3} & & \\ 1 & \varphi_{1,4}-2\varphi_{2,4} & -2\varphi_{2,4} & 4\varphi_{2,4} & \\ \hline & \varphi_1-3\varphi_2+4\varphi_3 & 0 & 4\varphi_2-8\varphi_3 & -\varphi_2+4\varphi_3 \end{array} . \tag{6.3}$$

We compute the associated differential matrices $D^{(4,C)}(z)$, $D^{(4,K)}(z)$ and $D^{(4,S)}(z)$ of the above three methods (6.1)-(6.3). Numerical results in Figure 13 (a)-(c) show that the third and fourth leading principal minors are not always positive for $z < 0$. That is, the differential matrices $D^{(4,C)}(z)$, $D^{(4,K)}(z)$ and $D^{(4,S)}(z)$ are not positive (semi-)definite. It seems that these EERK methods would not be stabilized to preserve the energy dissipation law (1.3) no matter how large the stabilization parameter κ we set in (2.8).



(a) Cox and Matthews (2002)

(b) Krogstad (2005)

(c) Strehmel and Weine (1992)

(d) Hochbruck and Ostermann (2005)

Figure 13: Some leading principal minors (LPM) of associated differential matrices generated by existing fourth-order EERK methods in [3, 13, 19, 31].

Hochbruck and Ostermann [13] constructed the following five-stage EERK method which has been proved to have the stiff order four,

$$\begin{array}{c|ccccc}
0 & & & & & \\
\frac{1}{2} & \frac{1}{2}\varphi_{1,2} & & & & \\
\frac{1}{2} & \frac{1}{2}\varphi_{1,3}-\varphi_{2,3} & \varphi_{2,3} & & & \\
1 & \varphi_{1,4}-2\varphi_{2,4} & \varphi_{2,4} & \varphi_{2,4} & & \\
\frac{1}{2} & \frac{1}{2}\varphi_{1,5}-2a_{5,2}-a_{5,4} & a_{5,2} & a_{5,2} & \frac{1}{4}\varphi_{2,5}-a_{5,2} & \\
\hline
 & \varphi_1-3\varphi_2+4\varphi_3 & 0 & 0 & -\varphi_2+4\varphi_3 & 4\varphi_2-8\varphi_3
\end{array}, \tag{6.4}$$

with $a_{5,2}=\frac{1}{2}\varphi_{2,5}-\varphi_{3,4}+\frac{1}{4}\varphi_{2,4}-\frac{1}{2}\varphi_{3,5}$. Although this five-stage method is fourth-order accurate for semilinear parabolic problems, it may be not a good candidate for solving the gradient system (2.9). Actually, it would not be stabilized to preserve the energy dissipation law (1.3) unconditionally because the associated differentiation matrix $D^{(4,H)}(z)$ is not positive definite, see Figure 13(d), in which the curves of fourth and fifth leading principal minors of $\mathcal{S}(D^{(4,H)};z)$ are depicted.

Up to now, we are not able to find a fourth-order EERK method that preserves the energy dissipation law (1.3) unconditionally. Nonetheless, this issue would be theoretically interesting and practically important in simulating the gradient system (1.2).

To end this article, we summarize our results in the following. With a unified theoretical framework and a new indicator, namely average dissipation rate, for the energy dissipation properties of EERK methods, we examine some of popular methods and find:

(i) Among second-order EERK methods, the average dissipation rate of the EERK2 method (3.2) with $c_2=\frac{1}{2}$ is the closest to the continuous one so that it preserves the energy dissipation law (1.3) best although the ETD2RK method (3.4), corresponding to the EERK2 method (3.2) with $c_2=1$, seems the most popular for gradient flows, see [3, 5, 6, 9, 15, 25, 34]. If taking into the contractivity account, the EERK2-w method (3.5) with $c_2=\frac{1}{2}$ generates less time "ahead" effect than the well-known ETD2RK method.

(ii) Among third-order EERK methods, the popular ETD3RK and ETD2CF3 methods may destroy the energy dissipation law (1.3), especially for large time-step sizes. For the EERK3-1 (4.2) and EERK3-2 (4.3) methods, one can choose proper parameters (abscissas) to ensure the preserving of original dissipation law, while the EERK3-1 method (4.2) with $c_2=\frac{4}{9}$ produces the minimum time "ahead" effect among the considered third-order EERK methods.

At the same time, our theory is far away from complete. There are many interesting issues that we have not yet addressed. Some of them are listed as follows:

(a) As mentioned, we are not able to find (or prove the non-existence of) a fourth-order EERK method that preserves the energy dissipation law (1.3) unconditionally.

(b) It is noticed that the average dissipation rates $\mathcal{R}(z)$ of the mentioned EERK methods preserving the energy dissipation law (1.3) are greater than 1 and unbounded, that is, $\mathcal{R}(z)\to+\infty$ as $z\to-\infty$. The method with a bounded average dissipation rate would be significantly preferred in the long-time adaptive simulation approaching the steady state. Is there such an EERK method or how do we construct it?

(c) At least, is there a second-order EERK method that has a better dissipation rate than the EERK2 method (3.2) with $c_2=\frac{1}{2}$? Is there a third-order EERK method that has a better dissipation rate than the EERK3-1 method (4.2) with $c_2=\frac{4}{9}$?

Acknowledgements

The authors would like to thank Dr. Cao Wen for his sincere help.

A Auxiliary functions for EERK3-1 methods (4.2)

To examine the second and third leading principal minors of $\mathcal{S}(D^{(3,1)}; c_2, z)$ with the differentiation matrix (4.6), we introduce two auxiliary functions g_{31} and g_{32} as follows,

$$\begin{aligned} g_{31}(c_2,\sigma,z) :=& -9c_2^2z^2e^{2\sigma z} + 18c_2ze^{(\sigma+\frac{4}{3})z} + 6c_2ze^{(\sigma+\frac{2}{3})z}((3c_2-2)z-3) - 9e^{\frac{4z}{3}}(c_2^2z^2+1) \\ & - 6e^{\frac{2z}{3}}((3c_2-2)z-3) + (2z+3)(2(3c_2-1)z-3), \qquad (A.5) \\ g_{32}(c_2,\sigma,z) :=& -27c_2^2z(-z^2+2z+3e^{2z}-2e^z(z+3)+3)(e^{\frac{2z}{3}}-e^{\sigma z})^2 - 4(e^{2z}-1)z(2z-3e^{\frac{2z}{3}}+3)^2 \\ & + 6c_2\left[-12z^2e^{(\sigma+\frac{4}{3})z} + 4(2z+3)z^2e^{(\sigma+\frac{2}{3})z} + 2(2z^2+9z+9)ze^{\sigma z+z} + 18ze^{(\sigma+\frac{8}{3})z} - 6(z+3)ze^{(\sigma+\frac{5}{3})z}\right] \\ & + 6c_2\left[-6(2z+3)ze^{(\sigma+2)z} + 9e^{\frac{2z}{3}}(z^2-2z-3) - 6e^z(2z^2+9z+9) - 3e^{\frac{8z}{3}}(8z^2+6z+9) - 6e^{\frac{7z}{3}}(z+3)z\right] \\ & + 6c_2\left[+2e^{\frac{5z}{3}}(2z^3+9z^2+18z+27) + e^{2z}(8z^3+12z^2+18z+27) + 3(-2z^3+z^2+12z+9) + 18e^{\frac{10z}{3}}z\right]. \qquad (A.6) \end{aligned}$$

For any constants $p_2 > 0$ and $p_1 \geq 0$, one has $\lim_{z\to-\infty} z^{p_1}e^{p_2z} \to 0$. The dominant parts of g_{31} and g_{32} are simple although the expressions of them seem rather complex. Thus the computer-aided proof is always applied for simplicity of presentation. We will prove the following results by applying the technique of comparison function developed in Propositions 3.1 and 3.2.

Proposition A.1. *For the function g_{31} in* (A.5)*, $g_{31}(c_2, c_2, z) > 0$ if $c_2 \in [\frac{4}{9}, 1]$ and $z < 0$.*

Proof. For the function g_{31} in (A.5), we consider a comparison function $g_{31}^*(c_2, z) = g_{31}(c_2, \frac{4}{9}, z)$ such that the difference

$$\begin{aligned} g_{31}(c_2,c_2,z) - g_{31}^*(c_2,z) &= 3c_2(e^{\frac{4z}{9}} - e^{c_2z})e^{\frac{4z}{9}}\left(3c_2z^2 + 3c_2z^2e^{(c_2-\frac{4}{9})z} - 2e^{\frac{2z}{9}}((3c_2-2)z^2-3z) - 6ze^{\frac{8z}{9}}\right) \\ &\geq -3c_2^2z(e^{\frac{4z}{9}} - e^{c_2z})e^{\frac{4z}{9}}\left(-3z(1+e^{\frac{5z}{9}} - 2e^{\frac{2z}{9}}) + 2c_2^{-1}e^{\frac{2z}{9}}(3e^{\frac{2z}{3}} - 3 - 2z)\right) \\ &\geq -3c_2^2z(e^{\frac{4z}{9}} - e^{c_2z})e^{\frac{4z}{9}}r_{31}(z) \geq 0 \quad \text{for } c_2 \in [\tfrac{4}{9}, 1] \text{ and } z < 0, \end{aligned}$$

where the auxiliary function $r_{31}(z) := -3z(1 + e^{\frac{5z}{9}} - 2e^{\frac{2z}{9}}) + 2e^{\frac{2z}{9}}(3e^{\frac{2z}{3}} - 3 - 2z)$ is decreasing and positive for $z < 0$, cf. Figure 14 (a). Note that, $g_{31}^*(c_2, z)$ is a concave, quadratic polynomial with respect to c_2 because

$$\partial_{c_2}^2 g_{31}^*(c_2, z) = -9e^{\frac{8z}{9}}(e^{\frac{2z}{9}} - 1)^2z^2 < 0.$$

Through lengthy and simple calculations, it is not difficult to check that, cf. Figure 14 (b),

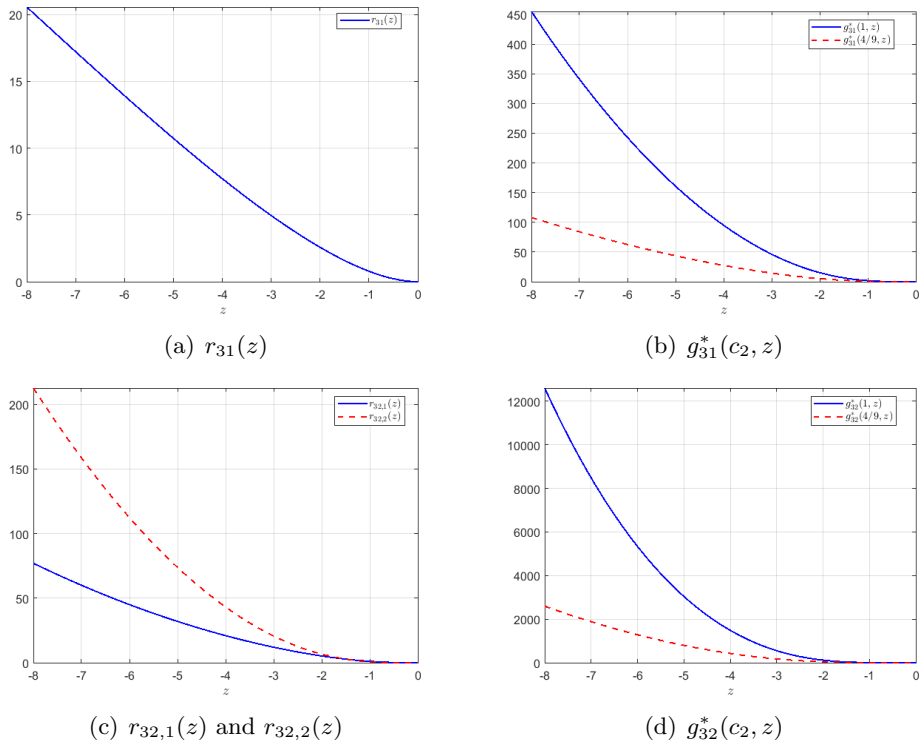
$$g_{31}^*(1, z) > 0 \quad \text{and} \quad g_{31}^*(\tfrac{4}{9}, z) > 0 \quad \text{for } z < 0.$$

They imply that $g_{31}^*(c_2, z) > 0$ and then $g_{31}(c_2, c_2, z) > 0$ for $c_2 \in [\frac{4}{9}, 1]$ and $z < 0$. □

Proposition A.2. *For the function g_{32} in* (A.6)*, $g_{32}(c_2, c_2, z) > 0$ if $c_2 \in [\frac{4}{9}, 1]$ and $z < 0$.*

Proof. For the function g_{32} in (A.6), we consider a comparison function $g_{32}^*(c_2, z) := g_{32}(c_2, \frac{4}{9}, z)$ such that the difference

$$\begin{aligned} g_{32}(c_2,c_2,z) - g_{32}^*(c_2,z) &= -3c_2^2z(e^{\frac{4z}{9}} - e^{c_2z})e^{\frac{2z}{3}}\left[9\big(e^{(c_2-\frac{2}{3})z} + e^{-\frac{2z}{9}} - 2\big)r_{32,1}(z) + 4c_2^{-1}r_{32,2}(z)\right] \\ &\geq -3c_2^2z(e^{\frac{4z}{9}} - e^{c_2z})e^{\frac{2z}{3}}\left[9(e^{\frac{z}{3}} + e^{-\frac{2z}{9}} - 2)r_{32,1}(z) + 4r_{32,2}(z)\right], \end{aligned}$$



(a) $r_{31}(z)$ (b) $g^*_{31}(c_2, z)$

(c) $r_{32,1}(z)$ and $r_{32,2}(z)$ (d) $g^*_{32}(c_2, z)$

Figure 14: Auxiliary functions r_{31}, g^*_{31}, $r_{32,1}$, $r_{32,2}$ and g^*_{32}.

where the two auxiliary functions $r_{32,1}$ and $r_{32,2}$ defined by

$$r_{32,1}(z) := z^2 - 2z - 3e^{2z} + 2e^z(z+3) - 3,$$
$$r_{32,2}(z) := 4z^2 + 6z + e^{\frac{z}{3}}(2z^2 + 9z + 9) - 6e^{\frac{2z}{3}}z + 9e^{2z} - 3e^z(z+3) - 3e^{\frac{4z}{3}}(2z+3).$$

Since the functions $r_{32,1}$ and $r_{32,2}$ are decreasing and positive for $z < 0$, cf. Figure 14 (c), we see that $g_{32}(c_2, c_2, z) \geq g^*_{32}(c_2, z)$ for $c_2 \in [\frac{4}{9}, 1]$ and $z < 0$.

Note that, $g^*_{32}(c_2, z)$ is a concave, quadratic polynomial with respect to c_2 due to

$$\partial^2_{c_2} g^*_{32} = 27ze^{\frac{8z}{9}}(e^{\frac{2z}{9}} - 1)^2 r_{32,1}(z) < 0 \quad \text{for } z < 0.$$

By simple but lengthy calculations, it is not difficult to check that, cf. Figure 14 (d),

$$g^*_{32}(1, z) > 0 \quad \text{and} \quad g^*_{32}(\tfrac{4}{9}, z) > 0 \quad \text{for } z < 0.$$

They imply that $g^*_{32}(c_2, z) > 0$ and then $g_{32}(c_2, c_2, z) > 0$ for $c_2 \in [\frac{4}{9}, 1]$ and $z < 0$. □

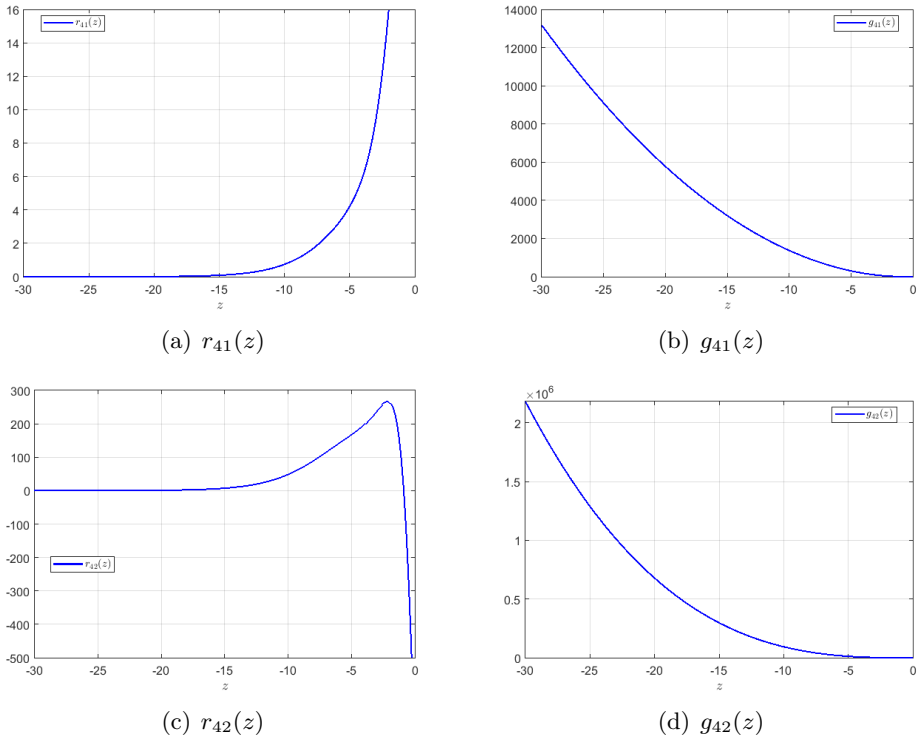
B Auxiliary functions for EERK3-2 methods (4.3)

To examine the second and third leading principal minors of $\mathcal{S}(D^{(3,2)}; 1, \frac{1}{2}, z)$, we introduce two auxiliary functions g_{41} and g_{42} as follows,

$$\begin{aligned} g_{41}(z) :=& 5(3z^2 + 2z - 5) + 8e^{\frac{3z}{2}}(z^2 - 5z - 1) - 2e^z(8z^2 + z + 3) \\ &+ e^{2z}(-16z^2 + 32z - 1) - 8e^{z/2}(z - 5) + 8e^{\frac{5z}{2}}z. \end{aligned} \tag{B.7}$$

$$\begin{aligned} g_{42}(z) :=& 800 + 1241z + 334z^2 - 71z^3 + 8e^{z/2}(z^2 - 97z - 80) - 8e^{\frac{5z}{2}}(121z^2 + 253z + 80) \\ &- 2e^{3z}(149z^2 + 19z + 80) + 8e^{\frac{3z}{2}}(5z^3 + 160z^2 + 216z + 160) + 2e^z(40z^3 - 331z^2 - 981z - 880) \\ &+ e^{2z}(431z^3 + 306z^2 + 928z + 1120) + 1072e^{\frac{7z}{2}}z - 169e^{4z}z. \end{aligned} \tag{B.8}$$

For simplicity of presentation, we always use the computer-aided proof to prove the positivity of the involved auxiliary functions.



(a) $r_{41}(z)$ (b) $g_{41}(z)$

(c) $r_{42}(z)$ (d) $g_{42}(z)$

Figure 15: Curves of the functions $g_{41}(z)$ and $g_{42}(z)$.

Proposition B.1. *The functions* g_{41} *and* g_{42} *in* (B.7)-(B.8) *are positive for* $z < 0$.

Proof. Note that, the quadratic polynomial part $\bar{g}_{41}(z) := 5(3z^2+2z-5)$ of $g_{41}(z)$ is decreasing with respect to $z \in (-1/3, 0)$ and $\lim_{z\to-\infty} \bar{g}_{41}(z) = +\infty$. The remaining part $r_{41}(z) := g_{41}(z) - \bar{g}_{41}(z)$ approaches zero when $|z|$ is properly large such as $z \le z_0 := -30$, see Figure 15(a). Actually, $r_{41}(z_0) \approx -8.6 \times 10^{-5}$. That is to say, $\bar{g}_{41}$ is dominant for $z \in (-\infty, z_0)$. As seen in Figure 15(b), g_{41} is decreasing and positive inside the finite interval $(z_0, 0)$. They lead to $g_{41}(z) > 0$ for $z < 0$.

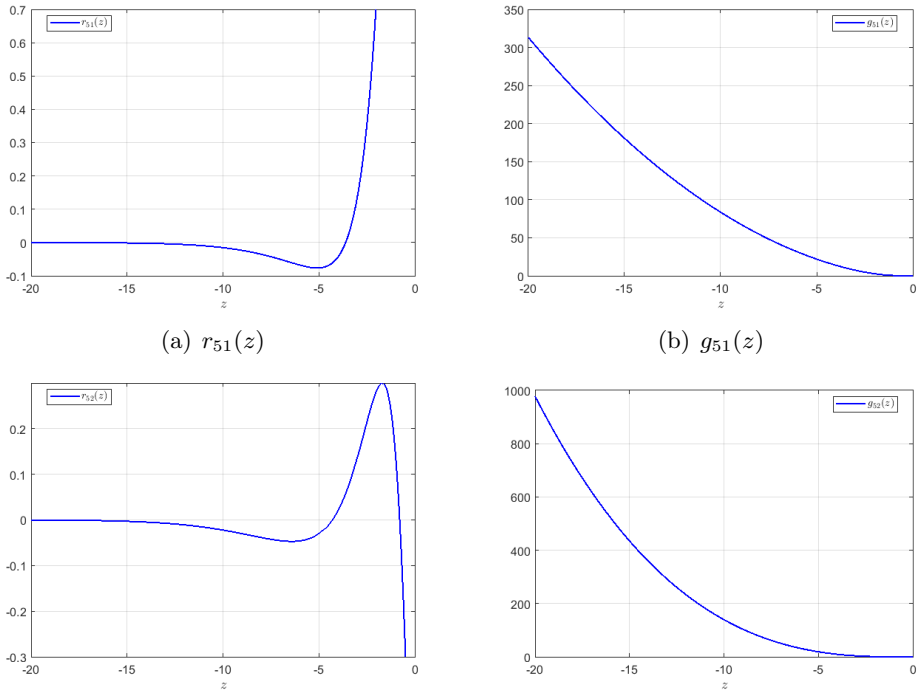
Similarly, the cubic polynomial part $\bar{g}_{42}(z) := 800+1241z+334z^2-71z^3$ of $g_{42}(z)$ is decreasing for $z \in (-3/2, 0)$ and $\lim_{z\to-\infty} \bar{g}_{42}(z) = +\infty$. The remaining part $r_{42}(z) := g_{42}(z) - \bar{g}_{42}(z)$ approaches zero when $|z|$ is properly large such as $z \le z_0 := -30$, see Figure 15(c). Actually, $r_{42}(z_0) \approx -9.1 \times 10^{-3}$. That is to say, $\bar{g}_{42}$ is dominant for $z \in (-\infty, z_0)$. As seen in Figure 15(d), g_{42} is decreasing and positive inside $(z_0, 0)$. They imply that $g_{42}(z) > 0$ for $z < 0$. □

To examine the second and third leading principal minors of $\mathcal{S}(D^{(3,2)}; \frac{3}{4}, \frac{3}{5}, z)$, we define two

auxiliary functions g_{51} and g_{52} as follows,

$$g_{51}(z) := \frac{1}{6400}\Big[8(567z^2-1353z-3362)-625e^{\frac{6z}{5}}(9z^2+16)-50e^{\frac{27z}{20}}(99z^2+492z+256)$$
$$-e^{\frac{3z}{2}}(5625z^2+4096)+9600e^{\frac{21z}{10}}z+15000e^{\frac{39z}{20}}z+128e^{\frac{3z}{4}}(33z+164)+200e^{\frac{3z}{5}}(33z+164)\Big], \tag{B.9}$$

$$g_{52}(z) := \frac{1}{8\times 10^6}\Big[8(-96681z^3+600979z^2+1843541z+1050625)+1875e^{\frac{6z}{5}}(533z^2-1148z-385)z-2332800e^{\frac{21z}{10}}z^2$$
$$+150e^{\frac{27z}{20}}(7319z^2+46552z+46361)z+3e^{\frac{3z}{2}}(453125z^2+62500z-58849)z-1000e^{\frac{13z}{5}}(2754z^2+1107z+5125)$$
$$-400e^{z}(6018z^2+35569z+42025)-128e^{\frac{3z}{4}}(7319z^2+46552z+25625)-200e^{\frac{3z}{5}}(7319z^2+46552z+25625)$$
$$-8e^{\frac{11z}{4}}(529821z^2+54243z+410000)+200e^{\frac{8z}{5}}(2106z^3+37287z^2+52087z+51250)-3645000e^{\frac{39z}{20}}z^2$$
$$+200e^{\frac{7z}{4}}(6642z^3+24111z^2+31963z+32800)+8e^{2z}(137781z^3-328779z^2-65091z+1050625)-7203e^{\frac{7z}{2}}z$$
$$-46875e^{\frac{16z}{5}}z+7650750e^{\frac{67z}{20}}z-7350e^{\frac{5z}{2}}(2z-25)z+18750e^{\frac{11z}{5}}(14z+41)z-300e^{\frac{47z}{20}}(6482z+48683)z\Big]. \tag{B.10}$$



(a) $r_{51}(z)$ (b) $g_{51}(z)$

(c) $r_{52}(z)$ (d) $g_{52}(z)$

Figure 16: Curves of the functions $g_{51}(z)$ and $g_{52}(z)$.

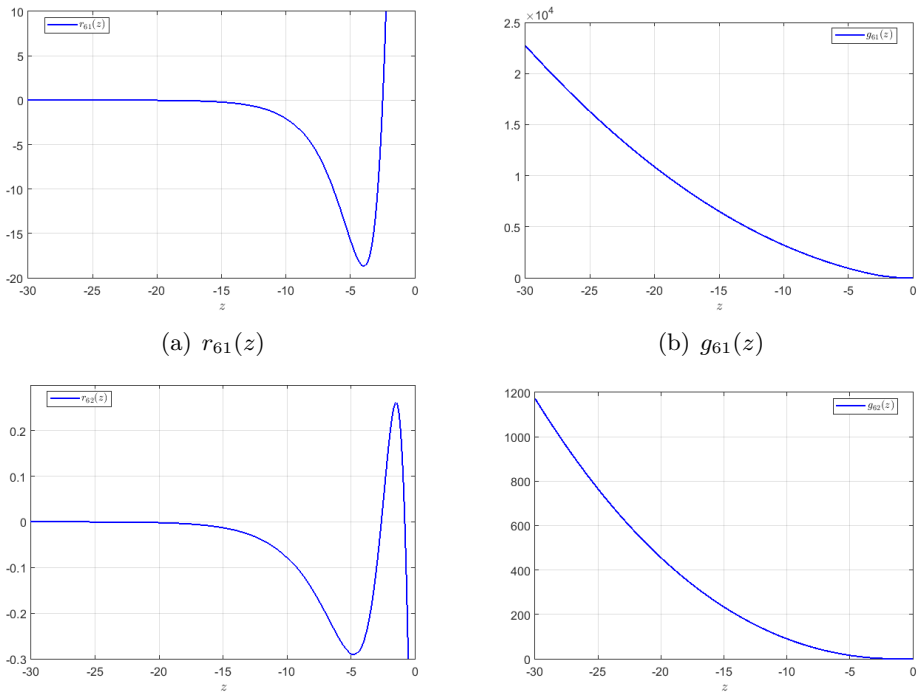
Proposition B.2. *The functions g_{51} and g_{52} in* (B.9)-(B.10) *are positive for $z<0$.*

Proof. The quadratic polynomial part $\bar{g}_{51}(z) := \frac{1}{800}(567z^2-1353z-3362)$ of $g_{51}(z)$ is decreasing with respect to $z\in(-\infty,0)$ and $\lim_{z\to-\infty}\bar{g}_{51}(z)=+\infty$. The remaining part $r_{51}(z) := g_{51}(z)-\bar{g}_{51}(z)$ approaches zero when $|z|$ is properly large such as $z\le z_0 := -20$, see Figure 16(a). Actually, $r_{51}(z_0)\approx -9.8\times 10^{-5}$. That is to say, $\bar{g}_{51}$ is dominant for $z\in(-\infty,z_0)$. As seen in Figure 16(b), g_{51} is decreasing and positive inside $(z_0,0)$. It is easy to conclude that $g_{51}(z)>0$ for $z<0$.

Similarly, the cubic polynomial part $\bar{g}_{52}(z) := \frac{1}{10^6}(-96681z^3+600979z^2+1843541z+1050625)$ of $g_{52}(z)$ is decreasing for $z\in(-3/2,0)$ and $\lim_{z\to-\infty}\bar{g}_{52}(z)=+\infty$. The remaining part $r_{52}(z) := g_{52}(z)-\bar{g}_{52}(z)$ approaches zero when $|z|$ is properly large such as $z\le z_0 := -20$, see Figure 16(c). Actually, $r_{52}(z_0)\approx -3.2\times 10^{-4}$. That is to say, $\bar{g}_{52}$ is dominant for $z\in(-\infty,z_0)$. As seen in Figure 16(d), g_{52} is decreasing and positive inside $(z_0,0)$. They imply that $g_{52}(z)>0$ for $z<0$. □

To examine the second and third leading principal minors of $\mathcal{S}(D^{(3,2)};\frac{1}{2},7/10,z)$, we define two auxiliary functions g_{61} and g_{62} as follows,

$$
\begin{aligned}
g_{61}(z) :=& \frac{1}{16}\Big[-625e^{\frac{7z}{5}}(z^2+4)-50e^{\frac{6z}{5}}(17z^2+64z+28)+16(21z^2-136z-256)\\
&-e^z(625z^2+196)+700e^{\frac{17z}{10}}z+2500e^{\frac{19z}{10}}z+28e^{z/2}(17z+64)+100e^{\frac{7z}{10}}(17z+64)\Big], \qquad \text{(B.11)}\\
g_{62}(z) :=& \frac{1}{2000000}\Big[2560000+4806561z+2007966z^2-25151z^3-1102500e^{\frac{19z}{10}}z^2+2500e^{\frac{7z}{5}}\left(80z^2-176z+185\right)z\\
&+50e^{\frac{6z}{5}}\left(8597z^2+38299z+32348\right)z-250e^{\frac{27z}{10}}\left(1911z^2-2688z+8000\right)-28e^{z/2}\left(8597z^2+38299z+20000\right)\\
&-100e^{\frac{7z}{10}}\left(8597z^2+38299z+20000\right)-14e^{\frac{5z}{2}}\left(76881z^2+63552z+40000\right)\\
&+350e^{\frac{3z}{2}}\left(1029z^3+4443z^2+5606z+3200\right)+50e^{\frac{17z}{10}}\left(6027z^3+33159z^2+63158z+80000\right)\\
&+e^{2z}\left(148176z^3-893416z^2+318039z+2560000\right)+e^z\left(265625z^3-1032550z^2-5015039z-5120000\right)\\
&+2165500e^{\frac{16z}{5}}z-62500e^{\frac{17z}{5}}z-109561e^{3z}z-12500e^{\frac{12z}{5}}(11z+32)z-50e^{\frac{11z}{5}}(17509z+75658)z\Big]. \qquad \text{(B.12)}
\end{aligned}
$$



(a) $r_{61}(z)$ (b) $g_{61}(z)$

(c) $r_{62}(z)$ (d) $g_{62}(z)$

Figure 17: Curves of the functions $g_{61}(z)$ and $g_{62}(z)$.

Proposition B.3. *The functions g_{61} and g_{62} in* (B.11)-(B.12) *are positive for $z<0$.*

Proof. The quadratic polynomial part $\bar{g}_{61}(z) := 21z^2-136z-256$ of $g_{61}(z)$ is decreasing with respect to $z\in(-\infty,0)$ and $\lim_{z\to-\infty}\bar{g}_{61}(z)=+\infty$. The remaining part $r_{61}(z):=g_{61}(z)-\bar{g}_{61}(z)$ approaches zero when $|z|$ is properly large such as $z\le z_0:=-30$, see Figure 17(a). Actually, $r_{61}(z_0)\approx -2.4\times10^{-4}$. That is to say, $\bar{g}_{61}$ is dominant for $z\in(-\infty,z_0)$. As seen in Figure 17(b), g_{61} is decreasing and positive inside $(z_0,0)$. It is easy to conclude that $g_{61}(z)>0$ for $z<0$.

Similarly, the cubic polynomial part $\bar{g}_{62}(z):=\frac{1}{2000000}(2560000+4806561z+2007966z^2-25151z^3)$ of $g_{62}(z)$ is decreasing for $z\in(-3/2,0)$ and $\lim_{z\to-\infty}\bar{g}_{62}(z)=+\infty$. The remaining part $r_{62}(z):=g_{62}(z)-\bar{g}_{62}(z)$ approaches zero when $|z|$ is properly large such as $z\le z_0:=-30$, see Figure 17(c). Actually, $r_{62}(z_0)\approx-2.9\times10^{-5}$. That is to say, $\bar{g}_{62}$ is dominant for $z\in(-\infty,z_0)$. As seen in Figure 17(d), g_{62} is decreasing and positive inside $(z_0,0)$. They imply that $g_{62}(z)>0$ for $z<0$. □

References

[1] B. Cano and M. J. Moreta. Solving reaction-diffusion problems with explicit Runge-Kutta exponential methods without order reduction. *ESAIM: M2AN*, 2024, doi: 10.1051/m2an/2024011.

[2] E. Celledoni, A. Marthinsen and B. Owren. Commutator-free Lie group methods. *Future Gener. Comp. Sys.*, 19: 341-352, 2003.

[3] S. M. Cox and P. C. Matthews. Exponential time differencing for stiff systems. *J. Comput. Phys.*, 176: 430-455, 2002.

[4] G. Dimarco and L. Pareschi. Exponential Runge-Kutta methods for stiff kinetic equations. *SIAM J. Numer. Anal.*, 49: 2057-207, 2011.

[5] Q. Du, L. Ju, X. Li and Z. Qiao. Maximum principle preserving exponential time differencing schemes for the nonlocal Allen-Cahn equation. *SIAM J. Numer. Anal.*, 57(2): 875-898, 2019.

[6] Q. Du, L. Ju, X. Li and Z. Qiao. Maximum bound principles for a class of semilinear parabolic equations and exponential time-differencing schemes. *SIAM Rev.*, 63: 317-359, 2021.

[7] T. A. Driscoll, N. Hale, and L. N. Trefethen, *Chebfun Guide*, Pafnuty Publications, Oxford, 2014, Online version: https://www.chebfun.org/docs/guide/guide19.

[8] M. Fasi, S. Gaudreault, K. Lund and M. Schweitzer. Challenges in computing matrix functions. *arXiv:2401.16132v1*, 2024.

[9] Z. Fu and J. Yang. Energy-decreasing exponential time differencing Runge-Kutta methods for phase-field models. *J. Comput. Phys.*, 454: 110943, 2022.

[10] Z. Fu, J. Shen and J. Yang. Higher-order energy-decreasing exponential time differencing Runge-Kutta methods for gradient flows. arXiv:2402.15142v1, 2024.

[11] N. J. Higham. Functions of Matrices: Theory and Computation. SIAM, Philadelphia, 2008.

[12] M. Hochbruck, C. Lubich and H. Selhofe. Exponential integrators for large systems of differential equations. *SIAM J. Sci. Comput.*, 19: 1552-1574, 1998.

[13] M. Hochbruck and A. Ostermann. Explicit exponential Runge-Kutta methods for semilinear parabolic problems. *SIAM J. Numer. Anal.*, 43: 1069-1090, 2005.

[14] M. Hochbruck and A. Ostermann. Exponential integrators. *Acta Numerica*, 19: 209-286, 2010.

[15] L. Ju, X. Li, Z. Qiao and H. Zhang. Energy stability and error estimates of exponential time differencing schemes for the epitaxial growth model without slope selection. *Math. Comput.*, 87(312): 1859-1885, 2018.

[16] L. Ju, J. Zhang, L. Zhu and Q. Du. Fast explicit integration factor methods for semilinear parabolic equations. *J. Sci. Comput.*, 62: 431-455, 2015.

[17] L. Ju, X. Li and Z. Qiao. Generalized SAV-exponential integrator schemes for Allen-Cahn type gradient flows. *SIAM J. Numer. Anal.*, 60: 1905-1931, 2022.

[18] A.-K. Kassam and L. N. Trefethen. Fourth-order time stepping for stiff PDEs. *SIAM J. Sci. Comput.*, 26: 1214-1233, 2005.

[19] S. Krogstad. Generalized integrating factor methods for stiff PDEs. *J. Comput. Phys.*, 203: 72-78, 2005.

[20] J. D. Lawson. Generalized Runge-Kutta processes for stable systems with large Lipschitz constants. *SIAM J. Numer. Anal.*, 4: 372-380, 1967.

[21] J. Li, X. Li, L. Ju and X. Feng. Stabilized integrating factor Runge-Kutta method and unconditional preservation of maximum bound principle. *SIAM J. Sci. Comput.*, 43: A1780-A1802, 2021.

[22] Z. Li and H.-L. Liao. Stability of variable-step BDF2 and BDF3 methods. *SIAM J. Numer. Anal.*, 60(4): 2253-2272, 2022.

[23] H.-L. Liao, T. Tang and T. Zhou. Positive definiteness of real quadratic forms resulting from variable-step L1-type approximations of convolution operators. *Sci. China. Math.*, 67(2): 237-252, 2024.

[24] H.-L. Liao and Z. Zhang. Analysis of adaptive BDF2 scheme for diffusion equations. *Math. Comput.*, 90: 1207-1226, 2021.

[25] Y. Liu, C. Quan and D. Wang. Maximum bound principle preserving and energy decreasing exponential time differencing schemes for the matrix-valued Allen-Cahn equation. *arXiv: 2312.15613v1*, 2023.

[26] V. T. Luan and A. Ostermann. Explicit exponential Runge-Kutta methods of high order for parabolic problems. *J. Comput. Appl. Math.*, 2014, doi: 10.1016/j.cam.2013.07.027.

[27] C. Moler and C. Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Rev.*, 45 (1): 3-49, 2003.

[28] S. Maset and M. Zennaro. Unconditional stability of explicit exponential Runge-Kutta methods for semi-linear ordinary differential equations. *Math. Comput.*, 78 (266): 957-967, 2009.

[29] D. Pop. An exponential method of numerical integration of ordinary differential equations. *Comm. ACM*, 6: 491-493, 1963.

[30] A. M. Stuart and A. R. Humphries. Dynamical systems and numerical analysis. Cambridge University Press, New York, 1998.

[31] K. Strehmel and R. Weiner. Linear-implizite Runge-Kutta Methoden und ihre Anwendungen. Teubner, Leipzig, 1992.

[32] J. Verwer. On generalized linear multistep methods with zero-parasitic roots and an adaptive principal root. *Numer. Math.*, 27: 143-155, 1977.

[33] X. Wang, L. Ju and Q. Du. Efficient and stable exponential time differencing Runge-Kutta methods for phase field elastic bending energy models. J. Comput. Phys., 316: 21-38, 2016.

[34] H. Zhang, L. Liu, X. Qian and S. Song. Quantifying and eliminating the time delay in stabilization exponential time differencing Runge-Kutta schemes for the Allen-Cahn equation. *ESAIM: M2AN*, 2023, doi: 10.1051/m2an/2023101.