Energy plus maximum bound preserving Runge–Kutta methods for the Allen–Cahn equation

Zhaohui Fu$^{1,2}$, Tao Tang$^{3,4}$ and Jiang Yang$^{1,4}$

$^1$Department of Mathematics, Southern University of Science and Technology, Shenzhen, 518055, China.
$^2$Department of Mathematics, University of British Columbia, Vancouver, V6T1Z4, Canada.
$^3$Division of Science and Technology, BNU-HKBU United International College, Zhuhai, 519000, China.
$^4$SUSTech International Center for Mathematics, Southern University of Science and Technology, Shenzhen, 518055, China.

Contributing authors: fuzhmath@math.ubc.ca; tangt@sustech.edu.cn; yangj7@sustech.edu.cn;

Abstract

It is difficult to design high order numerical schemes which could preserve both the maximum bound property (MBP) and energy dissipation law for certain phase field equations. Strong stability preserving (SSP) Runge–Kutta methods have been developed for numerical solution of hyperbolic partial differential equations in the past few decades, where strong stability means the non-increasing of the maximum bound of the underlying solutions. However, existing framework of SSP RK methods can not handle nonlinear stabilities like energy dissipation law. The aim of this work is to extend this SSP theory to deal with the nonlinear phase field equation of the Allen–Cahn type which typically satisfies both maximum bound preserving (MBP) and energy dissipation law. More precisely, for Runge–Kutta time discretizations, we first derive a general necessary and sufficient condition under which MBP is satisfied; and we further provide a necessary condition under which the MBP scheme satisfies energy dissipation.
1 Introduction

Strong stability preserving Runge–Kutta (SSP-RK) methods have been developed for numerical solution of *hyperbolic partial differential equations*, starting by Shu [1] it was observed that some Runge–Kutta methods can be decomposed into *convex combinations* of forward Euler steps, and so any convex functional property satisfied by forward Euler will be preserved by these higher-order time discretizations, generally under a different time-step restriction. This approach was used to develop second- and third-order Runge–Kutta methods that preserve the strong stability properties of the spatial discretizations developed in that work. In fact, this approach also guarantees that the intermediate stages in a Runge–Kutta method satisfy the strong stability property as well. More references in this direction can be found in [2–5] and a useful survey article of Gottlieb, Shu and Tadmor [6].

The aim of this work is to extend this SSP theory to deal with the nonlinear phase field equation of the Allen-Cahn type. To this end, we consider the numerical approximation of the Allen–Cahn equation

\[ u_t = \varepsilon \Delta u + \frac{1}{\varepsilon} f(u), \quad x \in \Omega, \ t \in (0, T], \tag{1.1} \]

with initial condition

\[ u(x, 0) = u_0(x), \quad x \in \Omega, \tag{1.2} \]

and the homogeneous Neumann boundary condition or periodic boundary condition, where \( \Omega \) is a bounded domain in \( \mathbb{R}^d \ (d = 1, 2, 3) \). In this paper, we consider the polynomial double-well potential

\[ F(u) = \frac{1}{4} (1 - u^2)^2 \tag{1.3} \]

and correspondingly,

\[ f(u) = -F'(u) = u - u^3. \tag{1.4} \]

The solution \( u(x, t) \) describes the concentration of two crystal orientations of the same material. In this phase model, \( u = 1 \) represents one orientation and \( u = -1 \) represents the other. The parameter \( \varepsilon \) here is the width of the interface between two phases, which is positive and small.
The Allen–Cahn equation can be viewed as the $L^2$ gradient flow of the Ginzburg-Landau free energy

$$\mathcal{E}(u) = \int_{\Omega} \left( \frac{\varepsilon}{2} |\nabla u|^2 + \frac{1}{\varepsilon} F(u) \right) dx. \quad (1.5)$$

The $L^2$ gradient flow structure corresponds to an energy dissipation law. This means that the energy is decreasing as a function of time,

$$\frac{d\mathcal{E}}{dt} = -\int_{\Omega} \left( \varepsilon \Delta u + \frac{1}{\varepsilon} f(u) \right)^2 dx \leq 0. \quad (1.6)$$

Another significant feature of the Allen–Cahn equation is its maximum bound preserving (MBP) property in the sense

$$\|u(\cdot, t)\|_{\infty} \leq 1 \quad (1.7)$$

provided that the initial and boundary values are bounded by 1.

The Allen–Cahn equation was originally introduced by Allen and Cahn in [7] to describe the motion of anti-phase boundaries in crystalline solids. In this context, $u$ represents the concentration of one of the two metallic components of the alloy and the parameter $\varepsilon$ represents the interfacial width, which is small compared to the characteristic length of the laboratory scale. The homogeneous Neumann boundary condition implies that no mass loss occurs across the boundary walls. Since then, the Allen–Cahn equation has been widely applied to many complicated moving interface problems in materials science and fluid dynamics through a phase-field approach. Since essential features of the Allen–Cahn equation are the dissipation law (1.6) and the MBP property (1.7), it is important to design numerical schemes satisfying both of them.

There have been many energy-dissipation studies for various numerical schemes for the Allen–Cahn equation, see, e.g., [8–12], and there have been also many recent works on MBP schemes, see, e.g., [13–16]. It is shown in [14] that the first- and second-order exponential time differencing (ETD) schemes satisfy MBP (1.7) unconditionally, while [13] established an abstract framework on MBP for more general semi-linear parabolic equations. Note that most of the relevant works consider the MBP (1.7) and the energy dissipation law (1.6) separately. The exception includes [17–19], but the schemes under consideration are only of first-order accuracy in time.

The present work seems to be the first effort to study high-order time discretizations aiming to preserve both (1.6) and (1.7). By applying Shu’s SSP-RK theory [1], i.e., using the property of the forward Euler method repetitively, we will first obtain a sufficient condition to verify whether a Runge–Kutta method is MBP, and also give a necessary and sufficient condition for $s$-stage $s$-th order MBP-RK methods. Both results will be established by using the so-called Butcher Tableau so the results are easy to verify. Moreover, we will
provide a necessary condition to judge whether the MBP-RK solutions preserve the energy dissipation law. Finally, we will provide some RK2, RK3 and RK4 methods which satisfy both (1.6) and (1.7). A special RK3 method violating the energy dissipation law will be also reported.

The paper is organized as follows. Section 2 contains some preliminaries and notations. Section 3 analyzes high-order MBP-RK methods to the Allen–Cahn equation by using Shu’s theory. We build up the relationship between the Butcher Tableau and the so-called Shu-Osher form [20]. Section 4 studies how to preserve the energy dissipation law (1.6) for the relevant Runge–Kutta methods. Section 5 applies the theory of Sections 3 and 4 to some typical Runge–Kutta schemes for the Allen–Cahn equation. The final section provides some concluding remarks.

2 Interplay between the Butcher Tableau and Shu-Osher form

The Runge–Kutta methods are a family of implicit and explicit iterative methods used in temporal discretization for the approximate solutions of ordinary differential equations (ODEs). Consider an ODE system in time $u' = G(u)$. An explicit Runge–Kutta method is commonly written in the form:

$$
\begin{align*}
    v_0 &= u_n, \\
    v_i &= u_n + \tau \sum_{j=0}^{i-1} a_{ij} G(v_j), \quad 1 \leq i \leq s-1 \\
    u_{n+1} &= u_n + \tau \sum_{j=0}^{s-1} b_j G(v_j).
\end{align*}
$$

(2.1)

In other words, to specify a particular method, one needs to provide the integer $s$ (the number of stages), and the coefficients $a_{ij}$ (for $1 \leq j < i \leq s$), $b_j$ (for $j = 1, \ldots, s$) and $c_j$ (for $j = 1, \ldots, s-1$). The matrix $(a_{ij})$ is called the Runge–Kutta matrix, while the $b_j$ and $c_j$ are known as the weights and the nodes [21]. These data are usually arranged in a mnemonic device, known as a Butcher tableau (after John C. Butcher):

$$
\begin{array}{c|cccccccc}
0 & 0 \\
\hline
\text{c}_1 & a_{1,0} & 0 \\
\text{c}_2 & a_{2,0} & a_{2,1} & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\text{c}_{s-1} & a_{s-1,0} & a_{s-1,1} & \ldots & a_{s-1,s-2} & 0 \\
\hline
b_0 & b_1 & \ldots & \ldots & b_{s-1}
\end{array}
$$

(2.2)
Energy plus maximum bound preserving RK methods for AC

where

\[ c_i = \sum_{j=0}^{i-1} a_{ij}, \quad i \geq 1. \quad (2.3) \]

If we define \( a_{sj} = b_j \) for all \( j \geq 0 \), then the scheme (2.1) becomes

\[
\begin{align*}
    u_n &= v_0, \\
    v_i &= u_n + \tau \sum_{j=0}^{i-1} a_{ij} G(v_j), \quad 1 \leq i \leq s \\
    u_{n+1} &= v_s.
\end{align*}
\]

We further define a strictly lower-triangular matrix \( A_L \) as

\[
A_L = \begin{bmatrix}
0 & a_{1,0} & 0 \\
& a_{2,0} & a_{2,1} & 0 \\
& & \ddots & \ddots & \ddots \\
& & & a_{s,0} & a_{s,1} & \ldots & a_{s,s-1} & 0
\end{bmatrix}.
\quad (2.5)
\]

On the other hand, the Runge–Kutta method can be written in the Shu–Osher form \([20]\):

\[
\begin{align*}
    v_0 &= u_n, \\
    v_i &= \sum_{k=0}^{i-1} \left( \alpha_{ik} v_k + \tau \beta_{ik} G(v_k) \right), \quad 1 \leq i \leq s \\
    u_{n+1} &= v_s,
\end{align*}
\]

where consistency condition requires

\[
\sum_{k=0}^{i-1} \alpha_{ik} = 1, \quad 1 \leq i \leq s. \quad (2.7)
\]

It is observed in Shu \([1, 20]\) if all coefficients are positive, i.e., \( \alpha_{ik} > 0 \) and \( \beta_{ik} \geq 0 \), then the solution can be viewed as convex combinations of forward Euler solutions. Based on this theory, the consistency condition (2.7) and the positivity conditions \( \alpha_{ik} \geq 0 \) and \( \beta_{ik} \geq 0 \) can ensure the Strong Stability Preserving (SSP) properties.

**Proposition 2.1** ([1, 20]) If the following so-called RK-SSP condition is satisfied

\[
\sum_{k=0}^{i-1} \alpha_{ik} = 1, \quad 1 \leq i \leq s; \quad \alpha_{ik} \geq 0, \quad \beta_{ik} \geq 0, \quad 0 \leq k < i \leq s, \quad (2.8)
\]
(when $\alpha_{ik} = 0$ then $\beta_{ik} = 0$), then the Runge–Kutta type method of type (2.6) satisfies the SSP condition in the sense that

$$\|u_{n+1}\| \leq \|u_n\|,$$

(2.9)

where $\| \cdot \|$ is the maximum norm or in the TV semi-norm.

Below we explore the relationship between the original form (2.1) and the Shu–Osher form (2.6). We rewrite the Butcher form with the help of the consistency condition (2.7):

$$v_i = v_0 + \tau \sum_{j=0}^{i-1} a_{ij} G(v_j) = \alpha_{i0} v_0 + \sum_{j=1}^{i-1} \alpha_{ij} v_0 + \tau \sum_{j=0}^{i-1} a_{ij} G(v_j).$$

(2.10)

We further use (2.1) for the above result to obtain

$$v_i = \alpha_{i0} v_0 + \sum_{j=1}^{i-1} \alpha_{ij} \left( v_j - \tau \sum_{k=0}^{j-1} a_{jk} G(v_k) \right) + \tau \sum_{k=0}^{i-1} a_{ik} G(v_k)$$

$$= \sum_{k=0}^{i-1} \left[ \alpha_{ik} v_k + \tau \left( a_{ik} - \sum_{j=k+1}^{i-1} \alpha_{ij} a_{jk} \right) G(v_k) \right], \quad 1 \leq i \leq s.$$ (2.11)

By defining

$$\beta_{ik} = a_{ik} - \sum_{j=k+1}^{i-1} \alpha_{ij} a_{jk}, \quad 0 \leq k \leq i - 1,$$

(2.12)

the relationship between the original form (2.1) and the Shu–Osher form (2.6) is established.

**Theorem 2.2** If all elements in the strictly lower-triangular matrix $A_L$ in (2.5) are positive, i.e. $a_{ik} > 0$ for all $0 \leq k < i \leq s$, then there exist coefficients $\alpha_{ij}, \beta_{ij} \geq 0$ such that the corresponding explicit Runge–Kutta scheme (2.1) satisfies the RK-SSP condition.

**Proof** We need to use the given positive elements $a_{ij}$ ($0 \leq j < i \leq s$) to construct positive coefficient pairs $(\alpha_{ik}, \beta_{ik})$. Let

$$\delta = \min_{0 \leq k < i \leq s} \frac{a_{ik}}{\sum_{j=k+1}^{i-1} a_{jk}},$$

(2.13)

and let

$$\alpha_{ij} = \min \left\{ \frac{\delta}{2}, \frac{1}{2(i-1)} \right\}, \quad \forall 1 \leq i \leq s, 1 \leq j < i,$$

$$\alpha_{i0} = 1 - (i-1) \cdot a_{i1}, \quad 1 \leq i \leq s.$$ (2.14)
It is easy to check that $\alpha_{ik} > 0$ for all $0 \leq k < i \leq s$. Using the relation (2.12) and the fact $\alpha_{ij} < \delta$ gives

$$\beta_{ik} = a_{ik} - \sum_{j=k+1}^{i-1} \alpha_{ij}a_{jk}$$

$$> a_{ik} - \sum_{j=k+1}^{i-1} \delta a_{jk}$$

$$= \sum_{j=k+1}^{i-1} a_{jk} \left( \frac{a_{ik}}{\sum_{j=k+1}^{i-1} a_{jk}} - \delta \right) \geq 0. \quad (2.15)$$

Finally, it is easy to observe that

$$\sum_{k=0}^{i-1} \alpha_{ik} = \alpha_{i0} + \sum_{k=1}^{i-1} \alpha_{i1} = 1 - (i-1) \cdot \alpha_{i1} + (i-1) \cdot \alpha_{i1} = 1. \quad (2.16)$$

This completes the proof of the theorem. □

The above theorem gives a simple sufficient condition which can convert a Runge–Kutta method to be of Shu–Osher type satisfying the RK-SSP condition (2.1). Below we derive a sufficient and necessary condition for a wide class of Runge–Kutta method.

**Theorem 2.3** An explicit Runge–Kutta method with non-zero sub-diagonal elements satisfies the RK-SSP condition (2.8) if and only if all elements in the strictly lower-triangular part of $A_L$ in (2.5) are positive.

**Proof** The sufficient condition is proved in Theorem 2.2. We now prove the necessary condition. In this case, the explicit Runge–Kutta method with non-zero sub-diagonal elements satisfies the RK-SSP condition (2.8). Define $\text{order}(a_{ik}) = (i-1)s + k$, $0 \leq k < i \leq s$. If there exist non-positive elements in $A$, we take the first of them in the sense of order, $a_{pq} \leq 0$. Since the RK scheme satisfies (2.8), we have

$$\alpha_{ik} \geq 0, \quad \beta_{ik} = a_{ik} - \sum_{j=k+1}^{i-1} \alpha_{ij}a_{jk} \geq 0, \quad 0 \leq k \leq i-1. \quad (2.17)$$

In particular, we have

$$a_{pq} - \sum_{j=q+1}^{p-1} \alpha_{pj}a_{jq} \geq 0. \quad (2.18)$$

As $a_{pq}$ is the first non-positive element in the sense of order, all $a_{jq}$ in the summation above are all positive. We then have two cases.

- If $a_{pq} < 0$, then it is easy to see a contradiction to (2.18).
- If $a_{pq} = 0$, as the other $a_{jq}$ is positive then it follows from (2.18) that all $\alpha_{pj}$ in (2.18) are 0, and in particular, $\alpha_{p,p-1} = 0$, which leads to $\beta_{p,p-1} = 0$. Note that by (2.12) we have $\beta_{p,p-1} = a_{p,p-1}$. Consequently, we have $a_{p,p-1} = 0$ which contradicts the non-zero sub-diagonal element assumption.
This completes the proof of the theorem. □

One direct result is the following proposition.

**Proposition 2.4** An $s$-stage $s$th-order explicit Runge–Kutta method satisfies the RK-SSP condition (2.8) if and only if all elements in the strictly lower-triangular part of $A_L$ are positive.

**Proof** For the $s$-order RK scheme, in order to match the highest order term in the Taylor expansion, we must have

$$\frac{1}{s!} = a_{1,0}a_{2,1} \cdots a_{s,s-1}$$

which guarantees all sub-diagonal elements $a_{i,i-1}$ are non-zero. □

The following proposition is given in [4, 21], while Theorem 2.3 provides a different perspective.

**Proposition 2.5** There does not exist any 4-stage 4th-order explicit Runge–Kutta method satisfying the RK-SSP condition (2.8).

**Proof** The only 4th-order RK whose coefficients are all non-negative is the classic RK4 [21], whose Butcher tableau reads

```
0 | 0
----|----
1/2 | 1/2
1/2 | 0 1/2
1/2 | 0 0 1/2
```

Note that $a_{21} = a_{30} = a_{31} = 0$, i.e., they are not positive. Consequently, the classical RK4 does not satisfy the RK-SSP condition (2.8). □

**Proposition 2.6** Any irreducible RK method whose elements in the strictly lower-triangular part are all positive can not have order greater than 4.

**Proof** It is known that there is no irreducible SSP-RK method which has order greater than 4 [4, 5]. If an explicit irreducible Runge–Kutta method has positive strictly lower-triangular part, then based on Theorem 2.2 it must satisfy the RK-SSP condition, which contradicts the existing theory of [4, 5]. □

**Remark 1** Theorems in this section could also be derived by the contractivity theory [4, 22], although the approaches and illustrations are different.
3 MBP-RK methods for the Allen–Cahn equation

We will use the central finite difference discretization to the Allen–Cahn equation in space. Without loss of generality, we consider the computational domain $[0, 2\pi]$ with the periodic boundary condition and let the space mesh size $h = 2\pi/N$. Denote the grid points as $\{x_j = jh, j = 0, 1, \ldots, N - 1\}$ and the forward finite difference matrix of $\partial_x$ by $D_1$

$$D_1 = \frac{1}{h} \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ -1 & 1 & -1 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}_{N \times N}. \tag{3.1}$$

Thus we have the central difference discretization operator $D = -D_1^T D_1$ for the Laplacian $\Delta$. It is well-known that the discrete operator $D$ is of second-order accuracy to approximate the Laplacian operator.

**Lemma 3.7** Given any vector $v$ and scalar $\alpha > 2$, the following inequality holds:

$$\| (I + \frac{1}{\alpha} h^2 D) v \|_\infty \leq \| v \|_\infty. \tag{3.2}$$

**Proof** When $\alpha > 2$, $\alpha I + D$ is a tri-diagonal matrix whose elements are all positive. Besides, Note that the sum of each row of $D$ is zero. Consequently, the sum of every row of $\alpha I + h^2 D$ equals to constant $\alpha$. Observe that

$$\| (\alpha I + h^2 D) v \|_\infty = \max_j \|a_j v_{j-1} + b_j v_j + c_j v_{j+1}\|_\infty \leq (|a_j| + |b_j| + |c_j|) \max_j \|v_j\|_\infty = \alpha \|v\|_\infty. \tag{3.3}$$

This completes the proof. □

**Lemma 3.8** ([23]) Denote the discrete Fourier transform as $F_N$ and the conjugate transpose as $(\cdot)^H$, then it holds that

$$D = F_N^H \Lambda F_N, \quad \Lambda = \text{diag}(\lambda_0, \ldots, \lambda_{N-1}), \tag{3.4}$$

where $\lambda_j = -(2 - 2 \cos(jh))/h^2$ are eigenvalues of $D$.

One direct result of this lemma is the following inverse inequality: Given any vector $u$, it holds that

$$0 \leq -u^T D u \leq \frac{4}{h^2} u^T u. \tag{3.5}$$

Note that the above property holds for more general boundary conditions and domains, and in these situations the coefficient 4 in (3.5) will be replaced by a constant $C$ depending only on the boundary conditions and the domain.
For simplicity and for ease of demonstrating the main ideas, in this paper we only consider the 1D case. For multi-dimension cases, by using the tensor product for the discrete Laplacian operator in 2D and 3D, results similar to the 1D case can be obtained.

The semi-discrete finite difference discretization of the Allen–Cahn equation reads

$$\frac{d}{dt} u = \varepsilon D_1 u + \frac{1}{\varepsilon} f(u) =: G(u). \quad (3.6)$$

We list following two properties for system (3.6) resulting from the so-called method-of-line approach.

- If the initial value satisfies $\|u_0\|_{\infty} \leq 1$, then the solution $u(t)$ given by (3.6) satisfies the maximum bound preserving (MBP) property:

$$\|u(t)\|_{\infty} \leq 1, \quad \forall t \geq 0. \quad (3.7)$$

- Let

$$E_h(u) = \varepsilon \|D_1 u\|_{l_2}^2 + \frac{1}{4\varepsilon} \|1 - u^2\|_{l_2}^2. \quad (3.8)$$

Then the solutions of system (3.6) satisfy the semi-discrete energy dissipation law

$$\frac{d}{dt} E_h = - \left\| \frac{du}{dt} \right\|_{l_2}^2 \leq 0. \quad (3.9)$$

Note that the first result can be found in, e.g., [24], and the second result can be obtained by taking the $L^2$ inner product of (3.6) with $\frac{du}{dt}$.

In this section, we are concerned with MBP Runge–Kutta method for the Allen–Cahn equation. The main strategy is to extend the Shu-Osher theory for the hyperbolic conservation laws to deal with the Allen–Cahn solutions.

### 3.1 Forward Euler solution

In this section we discretize the semi-discrete system in the time direction by applying forward Euler method.

Before providing a useful theorem, we need following simple results, which can be obtained by an elementary proof.

**Lemma 3.9** For any positive number $a$, if $-4a \leq c \leq a/2$, then the function $g(x) = ax + c(x - x^3)$ satisfies

$$|g(x)| \leq a, \quad \forall x \in [-1, 1]. \quad (3.10)$$

The following theorem characterizes the Euler property for the system (3.6).

**Theorem 3.10** Consider the ODE system (3.6). If $\tau < \tau_0 := \min\{4h^2/\varepsilon, \varepsilon/4\}$, then for any vector $u$ satisfying $\|u\|_{\infty} \leq 1$, we have

$$\|u + \tau G(u)\|_{\infty} \leq 1. \quad (3.11)$$
Energy plus maximum bound preserving RK methods for AC

Proof Note that
\[
\|u + \tau G(u)\|_\infty = \|u + \tau \left(\varepsilon Du + \frac{1}{\varepsilon} f(u)\right)\|_\infty \\
= \left\|\left(\frac{1}{2}u + \tau \varepsilon Du\right) + \left(\frac{1}{2}u + \frac{\tau}{\varepsilon} f(u)\right)\right\|_\infty \\
\leq \left\|\frac{1}{2}u + \tau \varepsilon Du\right\|_\infty + \left\|\frac{1}{2}u + \frac{\tau}{\varepsilon} f(u)\right\|_\infty.
\] (3.12)

Using Lemma 3.7 and the assumption \(\tau < 4h^2/\varepsilon\) gives
\[
\left\|\frac{1}{2}u + \tau \varepsilon Du\right\|_\infty \leq \frac{1}{2}.
\]

Using (1.4), Lemma 3.9 and the assumption \(\tau < \varepsilon/4\) yields
\[
\left\|\frac{1}{2}u + \frac{\tau}{\varepsilon} f(u)\right\|_\infty \leq \frac{1}{2}.
\]

Combining the above three results gives the desired result. \(\square\)

Remark 2 The relationship between the time step and \(\varepsilon\) comes from here and is inevitable if one wants to solve with explicit methods directly.

3.2 MBP-RK methods

**Theorem 3.11** Consider the Runge-Kutta scheme (2.6) with \(G\) defined by (3.6). If the SSP-RK property (2.8) is satisfied, then
\[
\|u^n\|_\infty \leq 1 \implies \|u^{n+1}\|_\infty \leq 1
\] (3.13)
under the time-step restriction
\[
\tau \leq \tau_{SSP} := \min_{0 \leq k < i \leq s} \frac{\alpha_{ik}}{\beta_{ik}} \cdot \tau_0, \quad \text{with} \quad \tau_0 = \min \left\{ \frac{4h^2}{\varepsilon}, \frac{\varepsilon}{4} \right\}.
\] (3.14)

Note that the ratio above is understood as infinity whenever \(\beta_{ik} = 0\).

Proof The proof is based on the original SSP machinery, see, e.g., [3, 6]. In particular, note that for each \(i\), we have
\[
\|v_i\|_\infty = \left\|\sum_{k=0}^{i-1} \left(\alpha_{ik} v_k + \tau \beta_{ik} G(v_k)\right)\right\|_\infty \leq \left\|\sum_{k=0}^{i-1} \alpha_{ik} \left(v_k + \tau \frac{\beta_{ik}}{\alpha_{ik}} G(v_k)\right)\right\|_\infty.
\] (3.15)

Under the assumption (3.14), we have \(\tau \beta_{ik}/\alpha_{ik} \leq \tau_0\). Then using Theorem 3.10 gives
\[
\|v_i\|_\infty \leq \sum_{k=0}^{i-1} \alpha_{ik} \left\|v_k + \tau \frac{\beta_{ik}}{\alpha_{ik}} G(v_k)\right\|_\infty \leq \sum_{k=0}^{i-1} \alpha_{ik} \cdot 1 = 1.
\] (3.16)

This yields the desired result (3.13). \(\square\)
4 The discrete energy dissipation law

The discrete energy is defined as follows

\[ E(u) = -\frac{\varepsilon}{2} u^T D u + \frac{1}{\varepsilon} \sum_{j=1}^{N} F(u_j). \]  

(4.1)

For ease of our derivation, we will consider a special class of Runge–Kutta schemes.

**Lemma 4.12** Given a Butcher tableau (2.2) and the corresponding RK scheme (2.4). By suitably arranging coefficients \( \{c_{ik}\} \), we can obtain a class of RK scheme of the following form:

\[ v_i = \sum_{k=0}^{i-1} p_{ik} v_k + d_i \tau G(v_{i-1}), \quad 1 \leq i \leq s. \]  

(4.2)

**Proof** We wish to convert the RK formula (2.4) into the Shu-Osher format. It follows from (2.12) that

\[
\begin{align*}
v_i &= \sum_{k=0}^{i-1} \left[ \alpha_{ik} v_k + \left( a_{ik} - \sum_{l=k+1}^{i-1} a_{lk} \alpha_{il} \right) \tau G(v_k) \right] \\
&= \sum_{k=0}^{i-1} \alpha_{ik} v_k + \sum_{k=0}^{i-2} \left( a_{ik} - \sum_{l=k+1}^{i-1} a_{lk} \alpha_{il} \right) \tau G(v_k) + a_{i,i-1} \tau G(v_{i-1}).
\end{align*}

(4.3)

By forcing the second last term in (4.3) to 0, a set of values of \( \{\alpha_{ik}\} \) can be determined by \( \{a_{ik}\} \). This will leave only the last G-term in (4.3). Therefore, we derive \( p_{ik} = \alpha_{ik} \) and \( d_i = a_{i,i-1} \) and thus the scheme has the unique form (4.2). \( \square \)

Note that the consistency condition requires \( \sum_{k=0}^{i-1} p_{ik} = 1 \), but now the coefficients in (4.2) may be negative.

Before we present the main result of this section, we state the following result whose proof is quite straightforward:

\[
\frac{1}{4} [(a^2 - 1)^2 - (b^2 - 1)^2] \leq (b^3 - b)(a - b) + (a - b)^2, \quad \forall a, b \in [-1, 1].
\]  

(4.4)

**Theorem 4.13** For a given SSP-RK solution which has the form (4.2), we define a upper triangular matrix \( \Phi \) given by

\[
\Phi_{ij} = \sum_{k=0}^{i-1} \frac{p_{jk}}{d_j}, \quad i \leq j
\]  

(4.5)

and the energy discriminant

\[
\Delta E = \frac{1}{2} (\Phi + \Phi^T).
\]  

(4.6)
If $\Delta E$ is positive-definite, then the energy is non-increasing under the time step restriction

$$\tau \leq \min \left\{ \frac{\lambda}{\varepsilon} + \frac{2\varepsilon}{h^2}, \tau_{SSP} \right\},$$

(4.7)

where $\tau_{SSP}$ is the SSP-RK time-restriction given by (3.14), and $\lambda$ is the smallest eigenvalue of $\Delta E$.

**Proof** Rewrite (4.2) by using the form of $G$ (for simplicity we drop the $\varepsilon$ scale in this proof and notice that here $v_i$ are vectors) and the consistency condition:

$$f(v_i) = \frac{1}{d_i+1\tau} \left( v_{i+1} - \sum_{k=0}^{i} p_i+1,k v_k \right) - Dv_i$$

$$= \frac{v_{i+1} - v_i}{d_i+1\tau} + \frac{1}{d_i+1\tau} \sum_{k=0}^{i-1} p_{i+1,k}(v_i - v_k) - Dv_i.$$  

(4.8)

By using the definition of the potential $F$ and by using (4.4), we obtain

$$\sum_{j=1}^{N} F((u_{n+1})_j) - F((u_n)_j) = \sum_{i=0}^{s-1} \sum_{j=1}^{N} F((v_{i+1})_j) - F((v_{i})_j)$$

$$\leq \sum_{i=0}^{s-1} -(v_{i+1} - v_i)^T f(v_i) + (v_{i+1} - v_i)^2 =: J_1 + J_2 + J_3,$$

(4.9)

where

$$J_1 = \sum_{i=0}^{s-1} \left( 1 - \frac{1}{d_i+1\tau} \right) (v_{i+1} - v_i)^2, \quad J_2 = \sum_{i=0}^{s-1} (v_{i+1} - v_i)^T Dv_i,$$

$$J_3 = -\sum_{i=0}^{s-1} \frac{1}{d_i+1\tau} \sum_{k=0}^{i-1} p_{i+1,k}(v_i - v_k)^T (v_{i+1} - v_i).$$

It can be easily seen that $J_1$ is simply quadratic, which will be negative for sufficiently small $\tau$. By denoting $w_i = v_i - v_{i-1}$, we have

$$J_2 = \sum_{i=0}^{s-1} (v_{i+1} - v_i)^T D \left( \frac{v_{i+1} + v_i}{2} - \frac{v_{i+1} - v_i}{2} \right)$$

$$= \sum_{i=0}^{s-1} \frac{1}{2} \left( v_{i+1}^T Dv_{i+1} - v_i^T Dv_i - w_{i+1}^T Dw_{i+1} \right)$$

$$= \frac{1}{2} (u_{n+1}^T Du_{n+1} - u_n^T Du_n) - \frac{1}{2} \sum_{i=1}^{s} w_i^T Dw_i;$$

(4.10)

and also

$$J_3 = \sum_{i=0}^{s-1} \frac{1}{d_i+1\tau} \sum_{k=0}^{i-1} p_{i+1,k} \left( \sum_{m=k+1}^{i} w_m^T \right) w_{i+1}$$.  


\[ E_{n+1} - E_n = \sum_{j=1}^{N} F((u_{n+1})_j) - F((u_n)_j) - \frac{1}{2} (u_{n+1}^T Du_{n+1} - u_n^T Du_n) \]

\[ \leq \sum_{i=1}^{s} \left( 1 - \frac{1}{d_i \tau} \right) w_i^2 - \sum_{i=1}^{s} \sum_{m=1}^{i-1} \sum_{k=0}^{m-1} p_{ik} w_m^T w_i - \frac{1}{2} \sum_{i=1}^{s} w_i^T D w_i \]

\[ = \sum_{i=1}^{s} w_i^2 - \frac{1}{\tau} \sum_{m,i=1}^{s} w_m^T \Phi_{mi} w_i - \frac{1}{2} \sum_{i=1}^{s} w_i^T D w_i, \quad (4.12) \]

where we have defined an upper triangle matrix \( \Phi \) by (notice that \( \sum_{k=0}^{i-1} p_{ik} = 1 \))

\[ \Phi_{ij} = \sum_{k=0}^{i-1} p_{jk}/d_j, \quad i \leq j. \quad (4.13) \]

Consider the energy discriminant \( \Delta_E \) defined by (4.6). Recall that we dropped the \( \varepsilon \) scale in the very beginning. If we keep \( \varepsilon \) in the derivation, the change of the energy (4.12) becomes

\[ E_{n+1} - E_n \leq \sum_{i=1}^{s} w_i^2 - \frac{\varepsilon}{\tau} \sum_{i,j=1}^{s} w_i^T \Phi_{ij} w_j - \frac{\varepsilon^2}{2} \sum_{i=1}^{s} w_i^T D w_i. \quad (4.14) \]

If \( \Delta_E \) is positive-definite and \( \lambda \) is the smallest eigenvalue of \( \Delta_E \), then we have

\[ \sum_{i,j=1}^{s} w_i^T \Phi_{ij} w_j = \sum_{i,j=1}^{s} w_i^T \Delta_{E,ij} w_j \geq \lambda \sum_{i=1}^{s} w_i^2. \quad (4.15) \]

It follows from (3.5) that

\[ -\sum_{i=1}^{s} w_i^T D w_i \leq 4h^{-2} \sum_{i=1}^{s} w_i^2. \quad (4.16) \]

Thus, by (4.14), to make sure the energy dissipation we only need

\[ 1 - \frac{\varepsilon \lambda}{\tau} + \frac{2\varepsilon^2}{h^2} \leq 0, \quad (4.17) \]

which is true under the assumption (4.7). \( \square \)

5 Some energy plus MBP RK methods

In this section we present some RK2, RK3 and RK4 methods which is maximum bound preserving and energy dissipation. We will also give an RK3 method which is MBP but does not satisfy our condition for the energy dissipation law. All examples in this section are existing schemes, and the special 5-stage 4th-order example comes from [3].
5.1 An RK2 satisfying energy-dissipation and MBP

The Butcher Tableau is as follows:

\[
\begin{array}{c|cc}
0 & 1 & 0 \\
1/2 & 1/2 & 1/2 \\
\end{array}
\]

The corresponding (4.2) form is given by

\[
\begin{align*}
v_0 &= u_n, \\
v_1 &= v_0 + \tau G(v_0), \\
v_2 &= v_0 + \frac{\tau}{2} G(v_0) + \frac{\tau}{2} G(v_1) = \frac{1}{2} v_0 + \frac{1}{2} v_1 + \frac{\tau}{2} G(v_1). \quad (5.1)
\end{align*}
\]

It follows from the theory in Section 3 the scheme (5.1) is MBP.

The energy form coincides with (5.1) and the energy discriminant is

\[
\Phi = \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}, \quad \Delta E = \frac{1}{2} (\Phi + \Phi^T) = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 2 \end{pmatrix}. \quad (5.2)
\]

Note that \(\Delta E\) is positive-definite and the smallest eigenvalue of \(\Delta E\) is \(\frac{1}{2} (3 - \sqrt{2})\). Hence with suitably small time-step, this MBP-RK2 scheme preserves both maximum bound and the energy dissipation law.

5.2 An RK3 satisfying energy-dissipation and MBP

Consider the Butcher tableau:

\[
\begin{array}{c|cccc}
0 & 1 & 1 & 1 \\
1/2 & 1/2 & 1/2 & 1/2 \\
3/4 & 3/4 & 3/4 & 3/4 \\
\end{array}
\]

The corresponding (4.2) form is given by

\[
\begin{align*}
v_0 &= u_n, \\
v_1 &= v_0 + \tau G(v_0), \\
v_2 &= \frac{3}{4} v_0 + \frac{1}{4} v_1 + \frac{\tau}{4} G(v_1), \\
v_3 &= \frac{1}{3} v_0 + \frac{2}{3} v_2 + \frac{2\tau}{3} G(v_2). \quad (5.3)
\end{align*}
\]
The energy discriminant reads

$$\Phi = \begin{pmatrix} 1 & 3 & \frac{1}{2} \\ 0 & 4 & \frac{1}{2} \\ 0 & 0 & 2 \end{pmatrix}, \quad \Delta_E = \frac{1}{2} (\Phi + \Phi^T) = \begin{pmatrix} 1 & \frac{3}{2} & \frac{1}{2} \\ \frac{3}{2} & 4 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 2 \end{pmatrix}. \quad (5.4)$$

Note that $\Delta_E$ is positive-definite and the smallest eigenvalue of $\Delta_E$ is about 0.362228. Again based on the theory in Section 4, with sufficiently small time-step, this MBP-RK3 scheme preserves both maximum bound and the energy dissipation law.

### 5.3 An RK3 satisfying MBP but not sure energy-dissipation

Consider the Butcher tableau:

\[
\begin{array}{c|ccc}
 & 0 & 1 & 2 \\
\hline
0 & \frac{5}{6} & \frac{1}{6} \\
1 & 1 & 1 & 1 \\
\end{array}
\]

The corresponding (4.2) form is given by

- $v_0 = u_n,$
- $v_1 = v_0 + \tau G(v_0),$
- $v_2 = v_1 + \tau G(v_1),$
- $v_3 = \frac{1}{3}v_0 + \frac{1}{2}v_1 + \frac{1}{6}v_2 + \frac{\tau}{6}G(v_2). \quad (5.5)$

However, It can be shown that the energy discriminant is not positive-definite in this case. Note

$$\Phi = \begin{pmatrix} 1 & 0 & 2 \\ 0 & 1 & 5 \\ 0 & 0 & 6 \end{pmatrix}, \quad \Delta_E = \frac{1}{2} (\Phi + \Phi^T) = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & \frac{5}{2} \\ 1 & \frac{5}{2} & 6 \end{pmatrix}. \quad (5.6)$$

The smallest eigenvalue of $\Delta_E$ is $\frac{1}{2}(7 - 3\sqrt{6}) \approx -0.174.$ Thus, this scheme is not guaranteed to decrease the energy by our approach.

### 5.4 An 5-stage RK4 satisfying MBP and energy-dissipation

It is well known that there is no 4-stage 4th-order RK4. We then just consider the following 5-stage RK scheme:

$$v_0 = u_n,$$
\[v_1 = v_0 + d_1 \tau G(v_0),\]
\[v_2 = p_{20} v_0 + p_{21} v_1 + d_2 \tau G(v_1),\]
\[v_3 = p_{30} v_0 + p_{32} v_2 + d_3 \tau G(v_2),\]
\[v_4 = p_{40} v_0 + p_{43} v_3 + d_4 \tau G(v_3),\]
\[v_5 = p_{52} v_2 + p_{53} v_3 + d_5 \tau G(v_3) + p_{54} v_4 + d_5 \tau G(v_4),\]

where

\[d_1 = 0.391752226571890, \quad p_{20} = 0.444370493651235,\]
\[p_{21} = 0.555629506348765, \quad d_2 = 0.368410593050371,\]
\[p_{30} = 0.620101851488403, \quad p_{32} = 0.379898148511597,\]
\[d_3 = 0.251891774271694, \quad p_{40} = 0.178079954393132,\]
\[p_{43} = 0.821920045606868, \quad d_4 = 0.544974750228521,\]
\[p_{52} = 0.517231671970585, \quad p_{53} = 0.09609710526147,\]
\[d_5 = 0.063692468666290, \quad p_{54} = 0.386708617503269,\]
\[d_5 = 0.226007483236906.\]

Using the theory of Section 3, it is known that the scheme (5.7) satisfies MBP. In order to obtain the energy form, we rewrite the last line as

\[v_5 = p_{52} v_2 + p_{53} v_3 + d_5 \frac{v_4 - p_{40} v_0 - p_{43} v_3}{d_4} + p_{54} v_4 + d_5 \tau G(v_4)\]

\[= -d_5 \frac{p_{40}}{d_4} v_0 + p_{52} v_2 + \left( p_{53} - \frac{d_5 p_{43}}{d_4} \right) v_3 + \left( p_{54} + \frac{d_5}{d_4} \right) v_4 + d_5 \tau G(v_5)\]

Thus the energy discriminant is

\[
\Phi = \begin{pmatrix}
\frac{1}{d_1} & \frac{p_{20}}{d_2} & \frac{p_{30}}{d_3} & \frac{p_{40}}{d_4} & -\frac{d_5 p_{40}}{d_4} \\
0 & \frac{1}{d_2} & \frac{p_{21}}{d_3} & \frac{p_{32}}{d_4} & -\frac{d_5 p_{43}}{d_4} \\
0 & 0 & \frac{1}{d_3} & \frac{p_{30}}{d_4} & p_{52} - \frac{d_5 p_{50}}{d_4} \\
0 & 0 & 0 & \frac{1}{d_4} & p_{52} + p_{53} - \frac{d_5}{d_4} \\
0 & 0 & 0 & 0 & \frac{1}{d_5}
\end{pmatrix}, \quad \Delta_E = \frac{1}{2} \left( \Phi + \Phi^T \right). \quad (5.9)
\]

The smallest eigenvalue of \(\Delta_E\) is about 1.706. Hence, based on the theory of Section 4, then with sufficiently small time-step this 5-stage RK4 preserves both maximum bound and energy dissipation law.

**Acknowledges**

The work of J. Yang is supported by National Natural Science Foundation of China (NSFC) Grant No. 11871264, Natural Science Foundation of Guangdong Province (2018A0303130123), and NSFC/Hong Kong RRC Joint
Research Scheme (NFSC/RGC 11961160718), and the research of Z. Yuan and Z. Zhou is partially supported by Hong Kong RGC grant (No. 15304420).

References

[1] Shu, C.-W.: Total-variation-diminishing time discretizations. SIAM Journal on Scientific and Statistical Computing 9, 1073–1084 (1988)

[2] Gottlieb, S., Ketcheson, D.I., Shu, C.-W.: Strong stability preserving Runge–Kutta and multistep time discretizations. World Scientific Press (2011)

[3] Gottlieb, S., Shu, C.-W.: Total variation diminishing Runge-Kutta schemes. Mathematics of Computation 67, 73–85 (1998)

[4] Kraaijevanger, J.F.B.M.: Contractivity of Runge-Kutta methods. BIT 31, 482–528 (1991)

[5] Ruuth, S.J., Spiteri, R.J.: Two barriers on strong-stability-preserving time discretization methods. J. Sci. Comput. 17, 211–220 (2002)

[6] Gottlieb, S., Shu, C.-W., Tadmor, E.: Strong stability-preserving high-order time discretization methods. SIAM Rev. 43, 89–112 (2001)

[7] Allen, S.M., Cahn, J.W.: A microscopic theory for anti-phase boundary motion and its application to anti-phase domain coarsening. Acta Metall 27, 1085–1095 (1979)

[8] Eyre, D.J.: An unconditionally stable one-step scheme for gradient systems. unpublished, see http://www.math.utah.edu/eyre/research/methods/stable.ps.

[9] Gomez, H., Hughes, T.: Provably unconditionally stable, second-order time-accurate, mixed variational methods for phase-field models. J. Comput. Phys. 230, 5310–5327 (2011)

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

[11] Shen, J., Xu, J., Yang, J.: A new class of efficient and robust energy stable schemes for gradient flows. SIAM Rev. 61, 474–506 (2019)

[12] Xu, C., Tang, T.: Stability analysis of large time-stepping methods for epitaxial growth models. SIAM J. Numer. Anal. 44, 1759–1779 (2006)

[13] Du, Q., Ju, L., Li, X., Qiao, Z.: Maximum bound principles for a class of semilinear parabolic equations and exponential time differencing schemes.
arXiv preprint: 2005.11465, to appear in SIAM Review (2020)

[14] Du, Q., Ju, L., Li, X., Qiao, Z.: Maximum principle preserving exponential time differencing schemes for the nonlocal Allen–Cahn equations. SIAM J. Numer. Anal. 57, 875–898 (2019)

[15] Li, B., Yang, J., Zhou, Z.: Arbitrarily high-order exponential cut-off methods for preserving maximum principle of parabolic equations. SIAM J. Sci. Comput. 42(6), 3957–3978 (2020)

[16] Stehlík, P., Volek, J.: Maximum principles for discrete and semidiscrete reaction-diffusion equation. Discrete Dyn. Nat. Soc., 1–13 (2015)

[17] Shen, J., Tang, T., Jiang, Y.: On the maximum principle preserving schemes for the generalized Allen–Cahn equation. Comm. Math. Sci. 14(6), 1517–1534 (2016)

[18] Tang, T.: Revisit of semi-implicit schemes for phase-field equations. Anal. Theory Appl. 36(3), 235–242 (2020)

[19] Tang, T., Yang, J.: Implicit-explicit scheme for the Allen–Cahn equation preserves the maximum principle. J. Comput. Math. 34(5), 471–481 (2016)

[20] Shu, C.-W., Osher, S.: Efficient implementation of essentially non-oscillatory shock-capturing schemes. Journal of Computational Physics (77), 439–471 (1988)

[21] Butcher, J.C.: Numerical Methods for Ordinary Differential Equations. Wiley (2003)

[22] Ferracina, L., Spijker, M.N.: Stepsize restrictions for the total-variation-diminishing property in general runge–kutta methods. SIAM Journal on Numerical Analysis 42(3), 1073–1093 (2004)

[23] Gong, Y., Wang, Q., Wang, Y., Cai, J.: A conservative Fourier pseudospectral method for the nonlinear Schrödinger equation. Journal of Computational Physics 328, 354–370 (2017)

[24] Yang, J., Du, Q., Zhang, W.: Uniform $l^p$ -bound of the Allen–Cahn equation and its numerical discretization. International Journal of Numerical Analysis and Modeling 15 (1-2), 213–227 (2018)