Commun. Comput. Phys.
doi: 10.4208/cicp.OA-2017-0015

Vol. **24**, No. 1, pp. 268-285
July 2018

On Dissipation and Dispersion Errors Optimization, A-Stability and SSP Properties

Arman Rokhzadi[1,*] and Abdolmajid Mohammadian[1]

[1] *Department of Civil Engineering, University of Ottawa, 161 Louis Pasteur, Ottawa, ON, Canada.*

Received 20 January 2017; Accepted (in revised version) 24 October 2017

Abstract. In a recent paper (Du and Ekaterinaris, 2016) optimization of dissipation and dispersion errors was investigated. A Diagonally Implicit Runge-Kutta (DIRK) scheme was developed by using the relative stability concept, i.e. the ratio of absolute numerical stability function to analytical one. They indicated that their new scheme has many similarities to one of the optimized Strong Stability Preserving (SSP) schemes. They concluded that, for steady state simulations, time integration schemes should have high dissipation and low dispersion. In this note, dissipation and dispersion errors for DIRK schemes are studied further. It is shown that relative stability is not an appropriate criterion for numerical stability analyses. Moreover, within absolute stability analysis, it is shown that there are two important concerns, accuracy and stability limits. It is proved that both A-stability and SSP properties aim at minimizing the dissipation and dispersion errors. While A-stability property attempts to increase the stability limit for large time step sizes and by bounding the error propagations via minimizing the numerical dispersion relation, SSP optimized method aims at increasing the accuracy limits by minimizing the difference between analytical and numerical dispersion relations. Hence, it can be concluded that A-stability property is necessary for calculations under large time-step sizes and, more specifically, for calculation of high diffusion terms. Furthermore, it is shown that the oscillatory behavior, reported by Du and Ekaterinaris (2016), is due to Newton method and the tolerances they set and it is not related to the employed temporal schemes.

AMS subject classifications: 65N12

Key words: Diagonally Implicit Runge-Kutta methods, dissipation and dispersion, optimization, numerical stability, steady state.

*Corresponding author. *Email addresses:* arokh061@uottawa.ca (A. Rokhzadi), amohamma@uottawa.ca (A. Mohammadian)

1 Introduction

Steady state solutions could be sought as the long-time mean solutions of the unsteady problems (Du and Ekaterinaris, 2016). However, dissipation and dispersion errors are the main obstacles to achieve long time-step sizes and accordingly to decrease the calculation time.

Among different temporal integrator schemes, Runge-Kutta methods have attracted attentions, as they are single-step methods and have free parameters which could lead to optimization of dissipation and dispersion errors. Excessive research in the literature has been made to control and bound dissipation and dispersion errors to increase the range of stability and accuracy. Consequently, numerous stability properties have been introduced including A-stability property. The reader is referred to ODEs books for more details (e.g. Hairer and Wanner, 1996).

The Strong Stability Preserving (SSP) Runge-Kutta methods are well-known due to their non-oscillatory behavior in shock and discontinuity problems. These methods were designed as convex combinations of Forward Euler (FE) method within limited radius of absolute monotonicity. This class of methods was further developed by Gottlieb and Shu (1998). Then, Ketcheson (2009) developed optimal implicit SSP RungeKutta methods up to order six with eleven stages.

The objective of this paper is to further investigate the stability analysis within studying dissipation and dispersion errors, in order to discuss the conclusions of Du and Ekaterinaris (2016) and to discuss their proposed DIRK scheme.

Du and Ekaterinaris (2016) indicated that this new scheme has many similarities with the three-stage fourth order SSP optimized DIRK scheme. They also described their proposed scheme, so called DIRK-D, as a more accurate model for low wavenumber components than other schemes they employed. However, as will be discussed, in stability analyses of temporal schemes, the main attention is on time step sizes. The wavenumber is assumed as a fixed variable, which basically would be the highest one. It will be shown that the source of instability, imposed by grid mesh, is due to high wavenumbers.

Relative stability analysis, i.e. the ratio of absolute numerical stability function to analytical one, was introduced by Hairer and Wanner (1996) within the concept of Order Star. Du and Ekaterinaris (2016) used relative stability function to design the optimized three-stage fourth order DIRK scheme, DIRK-D, and to examine its performances. However, the Order Star is mainly useful in proving relation between stability and achievable order of accuracy and this idea is not useful for judging the stability. Indeed, absolute stability is the more practical one (Leveque, 2007).

In Section 2, the relative stability analysis is studied further in order to show that this concept is not useful for optimization of the dissipation and dispersion errors. Meanwhile, as indicated by Leveque (2007) and shown in the present paper, this concept just shows the order of accuracy and truncation error. Du and Ekaterinaris (2016) indicated that, for advection-diffusion system, the amplification factor needs to include contribution of physical and numerical diffusion. In contrast, it will be shown that this contri-

bution is not required in minimizing the dissipation and dispersion errors (Section 3). Moreover this section proves that both A-stability and SSP properties cope with minimizing the dissipation and dispersion errors but for different applications. In Section 4, the non-linear viscous Burgers equation is examined with high order WENO-5 spatial scheme, which is very similar to WENO-5M used by Du and Ekaterinaris (2016). They used Newton iteration method, while, in this paper, the Gauss Seidel approach is served. It will be shown that the oscillatory behavior reported by Du and Ekaterinaris (2016) is not caused by temporal schemes and it is due to Newton method and the tolerances they set. Section 6 presents some concluding remarks.

2 Relative stability function

The advection-diffusion model is described by:

$$\frac{\partial u}{\partial t}+c\frac{\partial u}{\partial x}=\nu\frac{\partial^2 u}{\partial x^2}. \tag{2.1}$$

Applying the Fourier transform and assuming the solutions as $u(x,t)=u(t)e^{ikx}$, this equation is simplified as:

$$\frac{\partial \widehat{u}}{\partial t}=-(ick+k^2)\widehat{u}(t). \tag{2.2}$$

It could be found that the analytical solution is an exponential function. Consequently the analytical stability function, R_a, defined as the ratio of two successive time steps solutions, forms as:

$$\frac{\widehat{u}^{n+1}}{\widehat{u}^n}=R_a=e^z, \tag{2.3}$$

in which,

$$z=-(ick+k^2\nu)\Delta t. \tag{2.4}$$

As it is clear, $z\in C^-$, where C^- includes all the complex variables with negative real part. Note that c represents the advection speed which is assumed to obtain negative value as well.

Applying the Fourier transform to time dependent function as $\widehat{u}(t_n)=e^{-i\omega t_n}$, results in the analytical dispersion relation:

$$e^{-i\omega\Delta t}=e^z \tag{2.5}$$

in which ω is analytical frequency. Therefore, Eq. (2.3) is rearranged as:

$$e^{-i\omega\Delta t}=e^{-k^2\nu\Delta t}e^{-ick\Delta t}. \tag{2.6}$$

As can be seen, the diffusion contribution to the equation is in the form of descending exponential function with maximum value equals one. This fact will be discussed in the optimization process in Section 3.2.

In order to analyze the performance of Runge-Kutta scheme, this time integration method would be applied to (2.2) using the following equations:

$$\begin{aligned} u_{n+1} &= u_n + z\mathbf{b}^T\mathbf{Y}, \\ \mathbf{Y} &= u_n\mathbf{e} + z\mathbf{A}\mathbf{Y}. \end{aligned} \tag{2.7}$$

u_n represents the numerical solution at time step t_n, while Y represents the vector of solutions in internal stages. e is vector of ones of size $s\times 1$, and s is the number of internal stages. For DIRK scheme, matrix A and vector b are shown in Butcher tableau as Table 1.

Table 1: Butcher Tableau for s stage DIRK scheme.

$\mathbf{c}$	$\mathbf{A}$
	$\mathbf{b}^T$

c_1	a_{11}	0	0	0
c_2	a_{21}	a_{22}	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
c_s	a_{s1}	a_{s2}	a_{sl}	a_{ss}
	b_1	b_2	$\cdots$	b_s

Similar to the definition of the analytical stability function, the absolute numerical stability function, which is the ratio of the numerical solutions of two successive time steps, is in the form:

$$\frac{u_{n+1}}{u_n} = R_n = 1 + z\mathbf{b}^T(\mathbf{I} - z\mathbf{A})^{-1}\mathbf{e}, \tag{2.8}$$

in which I is $s\times s$ unit matrix.

It should be noted that the DIRKs coefficients which are specified by matrix A and vectors b and c need to satisfy the order conditions. The reader is referred to the literature for more details on order conditions, e.g. Nazari et al. (2015).

Du and Ekaterinaris (2016) described the relative stability function, called E, as the ratio of absolute numerical stability function to analytical one:

$$E = \frac{R_n}{R_a} = \frac{1 + z\mathbf{b}^T(\mathbf{I} - z\mathbf{A})^{-}1\mathbf{e}}{e^z}. \tag{2.9}$$

Taylor series expansion of the numerator reforms this equation as:

$$E = \frac{1 + z\mathbf{b}^T\mathbf{e} + z^2\mathbf{b}^T\mathbf{A}\mathbf{e} + z^3\mathbf{b}^T\mathbf{A}^2\mathbf{e} + \ldots + z^p\mathbf{b}^T\mathbf{A}^{p-1}\mathbf{e} + \sum_{n=p+1}^{\infty} z^n\mathbf{b}^T\mathbf{A}^{n-1}\mathbf{e}}{e^z}. \tag{2.10}$$

This equation could be rearranged as:

$$E=\frac{1+z\mathbf{b}^T\mathbf{e}+z^2\mathbf{b}^T\mathbf{A}\mathbf{e}+z^3\mathbf{b}^T\mathbf{A}^2\mathbf{e}+...+z^p\mathbf{b}^T\mathbf{A}^{p-1}\mathbf{e}}{e^z}+\frac{\sum_{n=p+1}^{\infty}z^n\mathbf{b}^T\mathbf{A}^{n-1}\mathbf{e}}{e^z}. \tag{2.11}$$

Obviously, in the first term of Right Hand Side (RHS), the polynomials coefficients in numerator are part of the order conditions. Depending on the desired order of accuracy, p, imposed on DIRK scheme, this polynomial is an approximation of the Taylor series of exponential function up to order p. Hence, (2.11) could be rearranged as follows:

$$E=1-\frac{\sum_{n=p+1}^{\infty}\frac{z^n}{n!}}{e^z}+\frac{\sum_{n=p+1}^{\infty}z^n\mathbf{b}^T\mathbf{A}^{n-1}\mathbf{e}}{e^z}. \tag{2.12}$$

Comparing Eqs. (2.11) and (2.12) implies that the Runge-Kutta methods try to revive the exponential function depending on the order of accuracy, p.

In stability analysis context, it is tried to design new schemes bearing large time step sizes, which is included in the variable z, (2.4). Note that decay rate of exponential function, denominator in RHS of (2.11) and (2.12), is faster than any power law function. Hence, it is clear that the magnitude of relative stability function, i.e. $|E|$, diverges once the variable z, obtains extremely large negative values in the domain of interest, $z\in C^-$. Moreover, to minimize the difference between the relative stability function and one, it is required to minimize the summation in numerator of the last term in RHS, (2.11) and (2.12), which is the Truncation Error (TE) term of the absolute numerical stability function, comparing (2.8) and numerator of (2.10). Therefore, it would be appropriate to consider the absolute numerical stability function in order to optimize the performance of Runge-Kutta schemes.

Leveque (2007) indicated that the relative stability function is just useful in studying the relation between stability and accuracy. In order to show this relation, Fig. 1 illustrates the relative stability region, i.e. $|E|\leq 1$, with blue color for some optimized SSP schemes. The coefficients of the examined SSP schemes are provided in Tables 2-5 in Butcher tableau form. The first row compares the stable region of the optimized SSP Runge-Kutta schemes with second order of accuracy achieved in two and three stages, so called SSP(2,2) and SSP(3,2) respectively. It is clear that more internal stages, involved in the simulation, results in larger stable region. However, it is worth mentioning that these two schemes preserve the A-stability property, which means that their absolute numerical stability function remain less than one in the whole domain of interest, i.e. $z\in C^-$. However, the relative stability function does not show this property.

Table 2: Butcher Tableau for SSP(2,2) scheme.

0.25	0.25	0
0.75	0.5	0.25
	0.5	0.5

Table 3: Butcher Tableau for SSP(3,2) scheme.

0.333333333333333	0.166666666666667	0	0
0.5	0.333333333333333	0.166666666666667	0
0.833333333333333	0.333333333333333	0.333333333333333	0.166666666666667
	0.333333333333333	0.333333333333333	0.333333333333333

Table 4: Butcher Tableau for SSP(3,3) scheme.

0.146446609406726	0.146446609406726	0	0
0.5	0.353553390593275	0.146446609406726	0
0.853553390593272	0.353553390593273	0.353553390593273	0.146446609406726
	0.333333333333333	0.333333333333333	0.333333333333333

Table 5: Butcher Tableau for SSP(3,4) scheme.

0.128886400515720	0.128886400515720	0	0
0.5	0.371113599484280	0.128886400515720	0
0.871113599484281	0.257772801031442	0.484454397937119	0.128886400515720
	0.302534578182651	0.394930843634698	0.302534578182651

To compare the relative stability region for third and fourth order of accuracy, SSP optimized Runge-Kutta schemes with three stages, so called SSP(3,3) and SSP(3,4) respectively, are illustrated in the second row of Fig. 1. The blue fingers include the stability functions roots (Hairer and Wanner, 1996). As can be seen, the higher order of accuracy results in more fingers and different region of stability. Further details of Order Star could be found in the literature (Hairer and Wanner, 1996; Wanner et al., 1978; Norsett and Wanner, 1979; Norsett and Trickett, 1984; Iserles and Norsett, 1991 and Butcher, 2009).

Consequently, one may notice that the relative stability function is not an appropriate concept to provide the details of Runge-Kutta schemes. Moreover, as shown, it is just informative in studying the relation between order of accuracy and stability. This fact was also indicated by Leveque (2007).

Therefore, it is not more helpful than absolute stability function to minimize the dissipation and dispersion errors.

3 Dissipation and dispersion errors optimization

In this section, two issues, accuracy and stability limits, associated with the exact and numerical dispersion relations will be discussed. Afterward, the optimization of dissipation and dispersion errors and contribution of diffusion terms will be investigated.

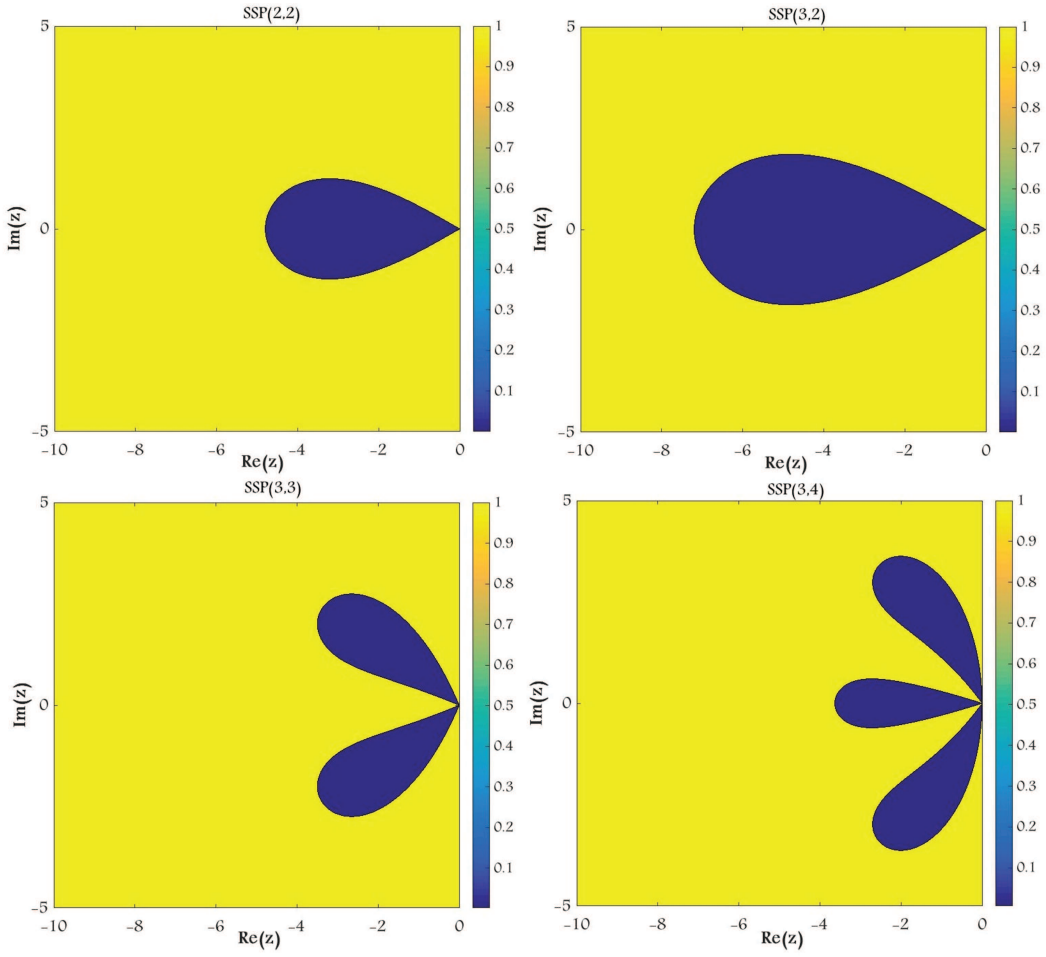

Figure 1: Relative stability analysis within domain of interest, $z \in C^-$, for SSP(2,2) and SSP(3,2) (first row) and SSP(3,3) and SSP(3,4) (second row), stable region, $|E| \leq 1$, shown in blue color.

Finally, the A-stability and SSP properties will be described as the dissipation and dispersion optimization problems.

3.1 Accuracy and stability

Similar to the exact dispersion relation, (2.5), the numerical dispersion relation could be easily found by applying the Fourier transform to temporal term, i.e. $\widehat{u}(t_n) = e^{-i\omega^* t_n}$ and substitution into (2.8):

$$e^{-i\omega^* \Delta t} = 1 + z\mathbf{b}^T(\mathbf{I} - z\mathbf{A})^{-1}\mathbf{e}, \tag{3.1}$$

in which $z = -(ick + k^2\nu)\Delta t$ and ω^* is called numerical frequency.

Furthermore, in stability analysis context, it is common to linearize the nonlinear governing equation by using perturbation technique. Hence, referring to (2.1), the variable is decomposed as $u=\overline{u}+u'$, in which $\overline{u}$ and u' represent the mean and perturbed components. Therefore, the distribution of perturbed component, which represents the propagation of the errors as well, follows the original governing equation (with some approximation). Hence, one may expect that the absolute numerical stability function, (2.8), rules the error propagations as well. This means that Eq. (2.8), representing the ratio of errors in two successive time steps and the associated dispersion relation, (3.1), need to be less than one in order to have a stable scheme. Clearly, large values of the attributed variable, z, which includes time step and wavenumber, may result in large errors. Therefore, the main concern with instability, associated with spatial discretization, is the highest wavenumber corresponding to the finest grid size, i.e. $k_{\max}=2\pi/\lambda_{\min}$, in which $\lambda_{\min}=2\Delta x$. Using (2.9) and (2.12) and the exact dispersion relation, (2.5), the following equation could be obtained:

$$e^{-i\omega^*\Delta t}=e^{-i\omega\Delta t}-\sum_{n=p+1}^{\infty}\frac{z^n}{n!}+\sum_{n=p+1}^{\infty}z^n\mathbf{b}^T\mathbf{A}^{n-1}\mathbf{e}. \tag{3.2}$$

The last RHS summation in (3.2) presents the dissipation and dispersion error. It could be realized that, in order to have an accurate time integrator, the difference between numerical and exact dispersion relations, (3.2), needs to be minimized.

There are two important issues related to dispersion relations; the first one is related to accuracy and it is associated with (3.2). It means that minimizing the difference between exact and numerical dispersion relations in (3.2) increases the accuracy. The other issue is stability concern and corresponds to numerical dispersion relation in (3.1), which is necessary to be less than one. These issues were also indicated by Hu et al. (1996) within the concept of accuracy and stability limits.

In order to show these two issues, Fig. 2 shows the magnitude of numerical dispersion relation, (3.1), in which the variable z just obtains the imaginary component. Referring to (2.5), the exact dispersion relation obtains the magnitude value equal to one. Hence, for accuracy reason, it is expected that time integrator schemes approximate the exact value, which is one, for large interval of the variable. In addition, due to the above discussion, the stable scheme needs to approximate the magnitude of numerical dispersion relation, (3.1), less than or equal one for large interval as well, which is due to error propagation. Both schemes employed in Fig. 2 are three-stage fourth order SDIRK schemes. However, the purple one is SSP optimized, so called SSP(3,4), which could be shown that it has less TE than the other one which is the unique three-stage fourth order A-stable scheme, developed by Crouzeix (1975).

It is clear from left column that SSP(3,4) could remain more accurate due to less TE. However, as the absolute numerical stability function with A-stability property remains less or equal one for extremely large variable, it shows more stable behavior (right column). Therefore, it could be concluded that although SSP optimized schemes were de-

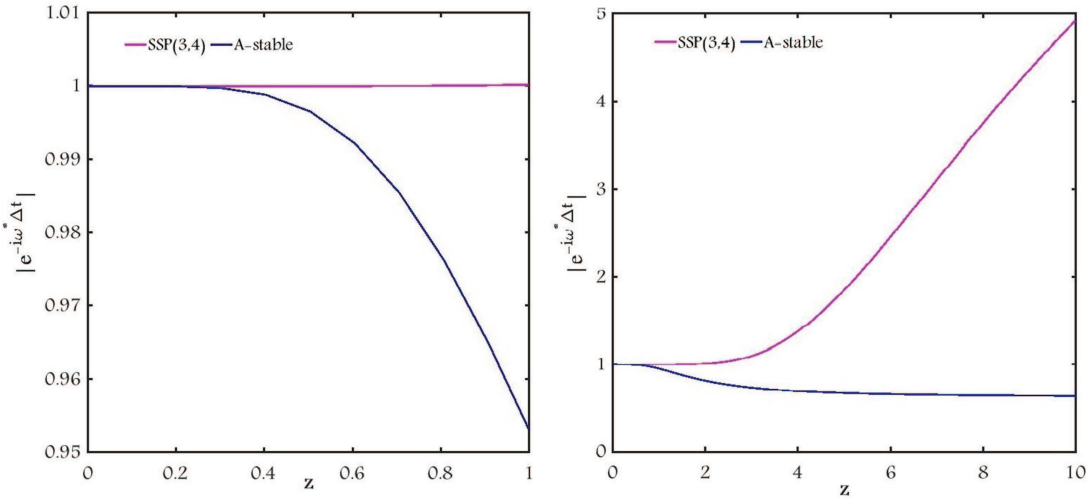

Figure 2: Comparison of absolute numerical stability functions magnitude, SSP(3,4) (purple), three stage fourth order A-stable (blue), in terms of accuracy (left column) and stability (right column).

veloped from different concept and for different purposes, they try to minimize the dissipation and dispersion error to increase the accuracy limit, (3.2). Meanwhile, A-stability property tries to minimize these errors to increase the stability limit, (3.1).

3.2 Dissipation and dispersion optimization

In order to minimize the combination of dissipation and dispersion errors, Hu et al. (1996) introduced an integral function, which is a summation of the difference of exact and numerical phase relations in (3.2) for a specified interval. Note that the difference of exact and numerical dispersion relations is equivalent to the difference of exact and absolute numerical stability functions.

Du and Ekaterinaris (2016) optimized the dissipation and dispersion errors by maximizing the so called acceptable amplification (RAA) and acceptable phase shift (RAP). They indicated that the nominated errors need to include the contribution of physical and numerical diffusion terms. The purpose of the following discussion is to show that the objective function proposed by Hu et al. (1996) does not require including the diffusion terms.

Hu et al. (1996) tried to minimize the following integral as the objective function, while they proved it as a combination of dissipation and dispersion errors.

$$E_{rr} = \min \int_0^{\Gamma} |R_n - R_a|^2 dz. \tag{3.3}$$

As they studied the hyperbolic compressible Euler system, the variable z in (2.3) and (2.8) just obtains imaginary component, i.e. $z=-ick\Delta t$, and consequently, a proposed objective

function, (3.3), they employed was:

$$E_{rr}=\min\int_0^{\Gamma}|R_n-R_a|^2d\sigma, \tag{3.4}$$

in which $R_a=e^{-i\sigma}$ and $\sigma=-ck\Delta t$.

Hu et al. (1996) minimized the objective function in (3.4) in the interval $[0,\Gamma]$ while changing Γ manually. In the present paper, E_{rr} is defined as:

$$E_{rr}=\min\int_0^{z_m}|R_n-R_a|^2dz, \tag{3.5}$$

a more general form of the error function in (3.4). It is the deviation of absolute numerical stability function from analytic stability function, or equivalently the difference of exact and numerical dispersion relations, in which variable z includes the real negative component, a representative of diffusion terms as well.

This integral would be calculated as the l_2-norm of the function R_n-R_a in a discretized grid mesh. For any vector X, the l_2-norm is bounded by l_∞-norm as follows (Boyd and Vandenberghe, 2004):

$$\|X\|_\infty\leq\|X\|_2\leq\sqrt{m}\,\|X\|_\infty \tag{3.6}$$

in which m is dimension of the vector X. In this paper, it depends on the grid mesh size employed to calculate the integral (3.5). It should be mentioned that the l_∞-norm is the largest component of vector X, which, hereafter, means the largest component of function R_n-R_a.

As the l_2-norm is lower bounded by l_∞-norm, it can be shown, for the present case, that maximizing the l_∞-norm results in the minimum value of l_2-norm. Consequently, the objective function in (3.5) could be replaced as follows:

$$E_{rr}=\max\|R_n-R_a\|_\infty. \tag{3.7}$$

From triangular inequality, one obtains:

$$\|R_n\|_\infty-\|R_a\|_\infty\leq\|R_n-R_a\|_\infty\leq\|R_n\|_\infty+\|R_a\|_\infty. \tag{3.8}$$

Referring to (2.3) and considering the fact that the desired domain is $z\in C^-$, the analytical stability function is found as descending function with maximum value of one, i.e. $\|R_n\|_\infty=1$. Furthermore, one can find that "max" is a convex function (Boyd and Vandenberghe, 2004) and it preserves the inequality direction. Consequently, the following inequalities hold:

$$\max\|R_n\|_\infty-1\leq\max\|R_n-R_a\|_\infty\leq\max\|R_n\|_\infty+1. \tag{3.9}$$

Hence, it is clear that maximizing the absolute numerical stability function results in maximum value for the objective function in (3.5), which is equivalent to minimizing the

specified integral in (3.4). Therefore, in minimizing the summation of dissipation and dispersion errors, it is not required to include the contributions of diffusion terms and using the objective function proposed by Hu et al. (1996), Eq. (3.4), results in minimizing the dissipation and dispersion errors. In this regard, the constraint $\| R_n \|_\infty \leq 1$ is necessary to impose on the optimization algorithm, as shown in Section 2.

In conclusion, in order to optimize the dissipation and dispersion errors using (3.4), the above discussions let us consider the analytical stability function without diffusion coefficient. Therefore, the objective function, (3.4), may be considered as:

$$E_{rr}^1 = \min \int_0^\Gamma |R_n - 1|^2 d\sigma. \tag{3.10}$$

Moreover, by using (3.7), one may consider the objective function as:

$$E_{rr}^2 = \max \| R_n - 1 \|_\infty . \tag{3.11}$$

Any above objective functions should be minimized subject to the constraint:

$$\| R_n \|_\infty \leq 1. \tag{3.12}$$

Substituting the absolute numerical stability function, (2.8), into Eqs. (3.11) and (3.12), the new forms could be found as:

$$E_{rr}^2 = \max \| z\mathbf{b}^T (\mathbf{I} - z\mathbf{A})^{-1} \mathbf{e} \|_\infty \leq 1, \tag{3.13}$$

subject to:

$$\| 1 + z\mathbf{b}^T (\mathbf{I} - z\mathbf{A})^{-1} \mathbf{e} \|_\infty \leq 1. \tag{3.14}$$

Eq. (3.13) may be rearranged as:

$$E_{rr}^2 = \max \| \mathbf{b}^T (z^{-1}\mathbf{I} - \mathbf{A})^{-1} \mathbf{e} \|_\infty \leq 1, \tag{3.15}$$

and the constraint may be reformed as:

$$-2 \leq z\mathbf{b}^T (\mathbf{I} - z\mathbf{A})^{-1} \mathbf{e} \leq 0. \tag{3.16}$$

As the desired domain in stability analysis is $z \in C^-$, and z appears in the denominator in Eq. (3.15), it is clear that maximizing $\| z \|$ results in the maximum value for RHS of Eq. (3.15). Hence, the new objective function could be:

$$E_{rr}^3 = \max \| z \|_\infty, \tag{3.17}$$

subject to the constraint in (3.16).

3.3 A-stability property

According to Hairer and Wanner (1996), an implicit Runge-Kutta method is A-stable if and only if $R_n(z)$ is analytic for $Re(z)<0$ and $|R_n(iy)|\leq 1$ for all real y values.

Therefore it could be interpreted that A-stability condition implies bounding the magnitude of absolute numerical stability function in the whole left half plane, i.e. $z\in C^-$ such that $\| R_n \|_\infty \leq 1$.

One may find A-stability property as an optimization problem with objective function as (3.17) subject to the constraint in (3.16). Eq. (3.16) could be rearranged in an appropriate way for the discussion as:

$$-2\leq \mathbf{b}^T(z^{-1}\mathbf{I}-\mathbf{A})^{-1}\mathbf{e}\leq 0. \tag{3.18}$$

The maximum value for $\| z \|_\infty$, which tends to infinity, lets the A-stability conditions become:

$$0\leq \mathbf{b}^T\mathbf{A}^{-1}\mathbf{e}\leq 2. \tag{3.19}$$

Therefore, A-stability property aims at minimizing the dissipation and dispersion errors, which can be obtained by using the proposed integral in (3.4).

In addition, Eq. (2.4) shows that the real component of variable z includes the time step, Δt, the diffusion term, represented by ν, and wavenumber, k. It is clear that large values of time step and/or diffusion terms, in both physical and numerical forms, may move the numerical instability region farther along the negative real axis. This includes the maximum resolvable wavenumber, $k_{\max}=\frac{2\pi}{\lambda_{\min}}$, which represents the finest grid size as $\lambda_{\min}=2\Delta x$.

Consequently, A-stability property is necessary for calculation of large time steps and large diffusion terms. Although Du and Ekaterinaris (2016) correctly indicated that high dissipation is suitable for decreasing the calculation cost toward steady state solutions, their conclusion is inconsistent with their statement that A-stable schemes may not remain stable.

3.4 SSP property

The SSP DIRK schemes are convex combinations of Forward Euler (FE) method (Ketcheson et al., 2009). Hence, they preserve the monotonic behavior of FE method, but for limited range of time-step sizes, specified with absolute radius of monotonicity. This property prevents the non-oscillatory behavior in solving discontinuity and shock capturing. Ketcheson et al. (2009) calculated the optimized radius of monotonicity for SSP DIRK schemes up to order six with eleven stages. They described that the maximum radius of absolute monotonicity of the Runge-Kutta scheme is the largest $r\geq 0$ such that $(\mathbf{I}+r\mathbf{A})^{-1}$ exists and

$$\begin{aligned}&\mathbf{K}(I+r\mathbf{A})^{-1}\geq 0,\\ &r\mathbf{K}(I+r\mathbf{A})^{-1}\mathbf{e}_s\leq \mathbf{e}_{s+1},\end{aligned} \tag{3.20}$$

in which $\mathbf{K}=\begin{bmatrix}\mathbf{A}\\ \mathbf{b}^T\end{bmatrix}$ and $\mathbf{e}_s$ is $s\times 1$ vector of ones.

Clearly the objective function in SSP optimization problem is the same as (3.17) but with some additional constraints.

As the absolute numerical stability function is the ratio of solutions of two successive time steps, one may find that a requirement for non-oscillatory behavior is that the absolute numerical stability function needs to preserve the positive sign, which results in more restricted inequalities than (3.16) as:

$$-1\leq z\mathbf{b}^T(\mathbf{I}-z\mathbf{A})^{-1}\mathbf{e}\leq 0. \tag{3.21}$$

Clearly, as the SSP optimization seeks positive value for radius of absolute monotonicity, r, simply replacing the variable z, in the interval $z\in C^-$, with r gives the following condition:

$$0\leq r\mathbf{b}^T(\mathbf{I}+r\mathbf{A})^{-1}\mathbf{e}\leq 1. \tag{3.22}$$

It is clear that equation set (3.20), i.e. constraints in SSP optimization problem, covers (3.22), the constraints in minimizing dissipation and dispersion errors problem. Consequently, SSP optimization aims at minimizing the dissipation and dispersion errors as well. Additionally, as it is clear from (3.20), SSP optimization imposes non negativity of all coefficients in DIRK scheme in order to guarantee the convex combination of FE (Ketcheson et al., 2009).

Therefore, It can be concluded that both A-stability and optimized SSP properties aim at minimizing the dissipation and dispersion errors. However, referring to the discussion in Section 3.1, A-stability tries to extend the range of stability while SSP optimization tries to extend the range of accuracy.

4 Non-linear viscous Burgers equation

Du and Ekaterinaris (2016), through calculation of non-linear viscous Burgers equation and using Newton method, compared their proposed scheme, DIRK-D, to the only A-stable three-stage fourth order SDIRK scheme (Crouzeix, 1975), so called DIRK-B, and reported some oscillatory behaviors. In this section, the non-linear viscous Burgers equation is examined to show the ability of A-stable schemes in obtaining the steady state solutions against DIRK-D which does not hold A-stability property.

The spatial terms are discretized with WENO-5 (Wang and Spiteri, 2007), which is not much different from WENO-5M employed by Du and Ekaterinaris (2016). The only difference is that they employed Newton iteration method to find the solutions, while here the Gauss Seidel approach is used.

The following equation represents the one-dimensional non-linear viscous Burgers

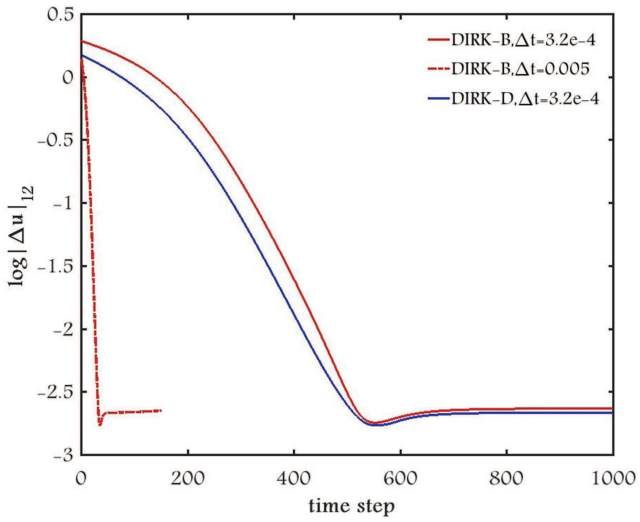

Figure 3: l_2-norm of difference between numerical solution and analytical steady state solution, $u-u_{exact}$, against number of time steps, DIRK-B (red) and DIRK-D (blue), $\Delta t=3.2e-4$ (solid), $\Delta t=5.0e-3$ (dash).

equation:

$$\frac{\partial u}{\partial t}+u\frac{\partial u}{\partial x}=\nu\frac{\partial^2 u}{\partial x^2}. \tag{4.1}$$

Neumann boundary conditions and initial condition are assumed as follows:

$$\begin{aligned} &u'(0,t)=u'(1,t)=0,\\ &u(x,0)=R\cos(\pi x), \quad 0\leq x\leq 1, \end{aligned} \tag{4.2}$$

where $R=5$, and $\nu=0.1$. The governing equation is calculated within domain $x\in[0,1]$, divided uniformly into 200 cells, and boundary conditions are set as symmetric. The analytical steady state solution for non-linear viscous equation was derived by Burns et al. (1998) as follows:

$$u(x)=4.9799\tanh\left[4.9799\frac{0.5-x}{2\nu}\right]. \tag{4.3}$$

In Fig. 3, the l_2-norm of difference between numerical solutions in each time step and exact steady state solution, i.e. $\| u-u_{exact} \|_2$ is illustrated. Both schemes, DIRK-B and DIRK-D, have the same behavior with the same time-step size, $\Delta t=3.2e-4$. It is clear that the steady state solution is obtained by both schemes within 600 time-stepping. DIRK-B, thanks to A-stability property, tolerates larger time step, $\Delta t=5.0e-3$, but DIRK-D diverges under this time step. Therefore, DIRK-B reaches the steady state solution faster than DIRK-D, within less than 200 steps.

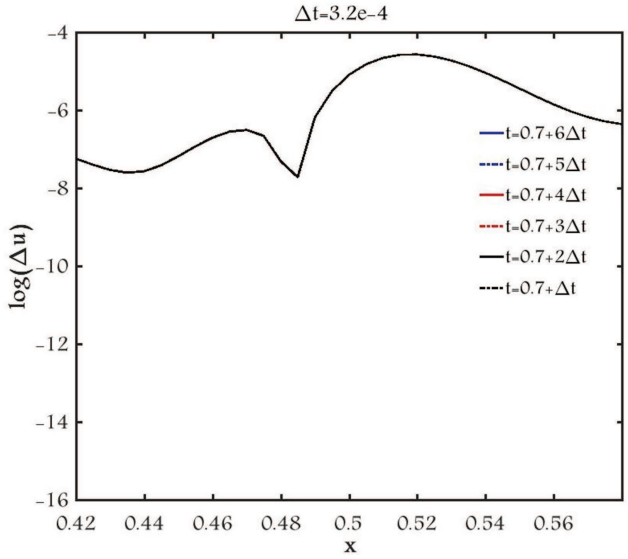

Figure 4: Distribution of errors for six successive time step, DIRK-B and DIRK-D (both identical), $\Delta t = 3.2e-4$s.

Following Du and Ekaterinaris (2016), the distribution of errors, defined as $\Delta u = u - u_{exact}$, at six successive time intervals, is illustrated in Fig. 4 with time-step $\Delta t = 3.2e-4s$. Both DIRK-B and DIRK-D calculate identical error distributions. As mentioned, the Gauss Seidel approach was chosen for iteration. Comparing to the results presented by Du and Ekaterinaris (2016), which was calculated by Newton method, it is realized that Newton method predicts the results more accurately, as the level of errors is lower than those provided by Gauss Seidel.

However, it is clear that no oscillations appear for any of schemes using Gauss Seidel method. Consequently, one may find that the oscillation reported by Du and Ekaterinaris (2016) is related to Newton method and not to temporal integration methods. Moreover, Fig. 5 shows the same distribution of errors generated by DIRK-B with larger time step, $\Delta t = 5.0e-3s$. This figure can demonstrate that oscillatory behavior, reported by Du and Ekaterinaris (2016), is due to Newton method and not due to the temporal integrators.

Fig. 6 shows the CPU time distribution for both DIRK-B and DIRK-D schemes in solving the non-linear viscous Burger equation until $t = 0.76s$. DIRK-D, developed by Du and Ekaterinaris, enjoys the SSP property. This property let the time integrator prevent oscillatory behavior and therefore, it help to reduce the CPU time. As shown, although the stable time step is small for DIRK-D compared to DIRK-B, the CPU time is competitive.

5 Conclusion

In this note, the conclusions of Du and Ekaterinaris (2016) were further investigated. It was shown that the relative stability function, which is the ratio of absolute numerical

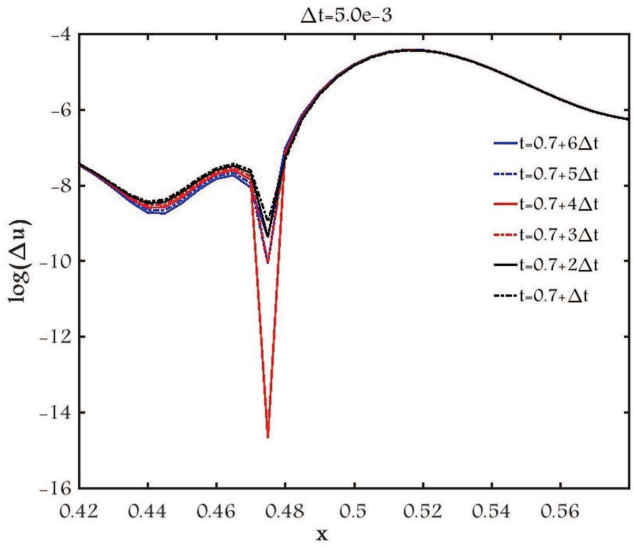

Figure 5: Distribution of errors for six successive time step, DIRK-B, $\Delta t = 5.0e-3$s.

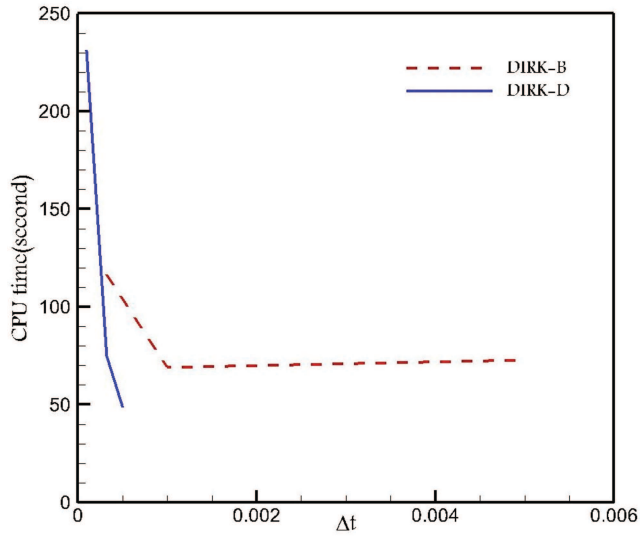

Figure 6: CPU time distribution in solving Burger equation till $t = 0.76$ s, DIRK-B (red) and DIRK-D (blue).

stability function to analytical one, is not informative for optimization of dissipation and dispersion errors. It is just useful to study the relation between stability and order of accuracy. It was discussed that the high wavenumber is the main concern as a source of instability. In the literature, the optimization of dissipation and dispersion errors were studied by Hu et al. (1996) using an integral function as a combination of dissipation and dispersion errors. Du and Ekaterinaris (2016) included contribution of diffusion terms in

their optimization procedure. It was shown that the proposed integral by Hu et al. (1996) is sufficient for optimization purposes and it does not require including the diffusion terms.

Moreover, referring to the analytical dispersion relation, (2.6), it is clear that the diffusion term appears as descending exponential function with maximum value equal one. This idea could be interpreted for numerical dispersion relation as well. Therefore, it is clear that the contribution of diffusion term does not require in the optimization.

Furthermore, it was shown that the proposed integral by Hu et al. (1996) covers the objectives of A-stability and SSP optimized properties. It was shown that A-stability property tries to extend the limit of stability by bounding the error propagation, while SSP optimized scheme tries to increase the limit of accuracy by decreasing the TE, which is commonly called error constant as well.

Calculation of non-linear viscous equation showed that A-stability property is necessary for stability under large time step sizes and large diffusion terms, and consequently, this property can accelerate achieving the steady state solutions. It was also shown that the oscillatory behavior reported by Du and Ekaterinaris (2016) is due to Newton method implementation and probably it was caused by the tolerances they set, as the Gauss Seidel method approaches the solution with no fluctuations.

Acknowledgments

The authors wish to acknowledge financial support from NSERC.

References

[1] S. Boyd and L. Vandenberghe, Convex Optimization, Cambridge University Press, ISBN 0 521 83378 7, 2004.
[2] J. Burns, A. Balogh, D. S. Gilliam, V. I. Shubov, Numerical stationary solutions for a viscous Burgers's equation, Journal of Mathematical Systems, Estimation, and Control, 8 (2) (1998) 116.
[3] J.C. Butcher, Order and stability of generalized Pad approximations, Appl. Numer. Math., 59(3-4) (2009) 558-567.
[4] M. Crouzeix, Ph.D. thesis, Paris: Universite Paris VI; 1975. Sur l'approximation des equations differentielles operationelles lineaires par des methodes de Runge Kutta.
[5] Y. Du and J. A. Ekaterinaris, On the stability and CPU time of the implicit Runge-Kutta schemes for steady state simulations, Commun. Comput. Phys., 20(2) (2016), 486-511.
[6] S. Gottlieb and C.-W. Shu, Total variation diminishing Runge-Kutta schemes, Math. Comp., 67 (1998), 7385.
[7] E. Hairer and G. Wanner, Solving Ordinary Differential Equations II, Springer-Verlag, ISBN 978-3-642-05221-7, 1996.
[8] F. Hu, M. Hussaini and J. Manthey, Low-dissipation and low-dispersion RungeKutta schemes for computational acoustics, J. Comput. Phys., 124 (1996) 177191.
[9] A. Iserles and S.P. Nørsett, Order Stars, Chapman and Hall, New York (1991).

[10] D. Ketcheson, C. Macdonald, S. Gottlieb, Optimal implicit strong stability preserving Runge-Kutta methods, Appl. Numer. Math., 59 (2009) 373392.
[11] R. J. LeVeque, Finite Difference Methods for Ordinary and Partial Differential Equations, SIAM, ISBN 978-0-89871-629-0, 2007.
[12] F. Nazari, A. Mohammadian, M. Charron, High-order low-dissipation low-dispersion diagonally implicit Runge-Kutta schemes, J. Comput. Phys., 286 (2015) 3848.
[13] S.P. Norsett and S.R. Trickett, Exponential fitting of restricted rational approximations to the exponential function, In: Rational Approximation and Interpolation, Lecture Notes in Mathematics, Vol. 1105, Eds: P. Russell Graves-Morris, E.B. Saff and R.S. Varga, Springer-Verlag, Berlin (1984) 466-476.
[14] S.P. Norsett and G. Wanner, The real-pole sandwich for rational approximations and oscillation equations, BIT, 19(1) (1979) 79-94.
[15] R. Wang and R. J. Spiteri, Linear instability of the fifth-order WENO method, SIAM J. Numer. Anal., 45(5) (2007), 18711901.
[16] G. Wanner, E. Hairer and S. P. Nrsett, Order stars and stability theorems, BIT, 18(4) (1978a) 475-489.