Patankar-type Runge-Kutta schemes for linear PDEs

Sigrun Ortleb and , and Willem Hundsdorfer

Citation: AIP Conference Proceedings **1863**, 320008 (2017); doi: 10.1063/1.4992489
View online: http://dx.doi.org/10.1063/1.4992489
View Table of Contents: http://aip.scitation.org/toc/apc/1863/1
Published by the American Institute of Physics

Patankar-Type Runge-Kutta Schemes for Linear PDEs

Sigrun Ortleb[1,a)] and Willem Hundsdorfer[2,b)]

[1] *University of Kassel, Department of Mathematics, Heinrich-Plett-Straße 40, 34132 Kassel, Germany.*
[2] *CWI, Science Park 123, 1098 XG Amsterdam, The Netherlands.*

[a)] Corresponding author: ortleb@mathematik.uni-kassel.de
[b)] willem.hundsdorfer@cwi.nl

Abstract. We study the local discretization error of Patankar-type Runge-Kutta methods applied to semi-discrete PDEs. For a known two-stage Patankar-type scheme the local error in PDE sense for linear advection or diffusion is shown to be of the maximal order $O(\Delta t^3)$ for sufficiently smooth and positive exact solutions. However, in a test case mimicking a wetting-drying situation as in the context of shallow water flows, this scheme yields large errors in the drying region. A more realistic approximation is obtained by a modification of the Patankar approach incorporating an explicit testing stage into the implicit trapezoidal rule.

Quasi-Linear Equations In Production-Destruction Form Arising From PDEs

In the context of geobiochemical models, so-called production-destruction equations are frequently encountered. These models describe the time-evolution of non-negative quantities and often take into account some type of mass conservation. The underlying ODE systems describing the time-evolution of non-negative quantities $u(t)$ can usually be written in the form $u_i' = \Big(\sum_{j=1}^m p_{ij}(u)\cdot u_j\Big) - q_i(u)\cdot u_i$ for $i = 1, 2, \dots, m$, with production terms $p_{ij}(u)u_j$ and destruction terms $q_i(u)u_i$ such that $p_{ij}(v),\ q_i(v) \ \geq\ 0$ for all $v \in \mathbb{R}^m$ and $i, j = 1, 2, \dots, m$. Usually, we also have $p_{ii}(v) = 0$ for all $v \in \mathbb{R}^m$ and some mass conservation property such as $\sum_{j=1}^m p_{ji}(u) = q_i(u)$. In vector form, we write

$$u' \ =\ P(u)\,u\ -\ Q(u)\,u, \tag{1}$$

with matrix-valued functions $P, Q : \mathbb{R}^m \to \mathbb{R}^{m\times m}$ such that $P(v) = \big(p_{ij}(v)\big) \geq 0,\ Q(v) = \mathrm{diag}\big(q_i(v)\big) \geq 0$ for all $v \in \mathbb{R}^m$. Setting $A(u) = P(u) - Q(u)$, this is written more shortly in the standard quasi-linear form $u' \ =\ A(u)\,u$.

While many interesting biochemical reactions fit into this framework, it also includes certain space-discretized partial differential equations, e.g. the heat equation discretized by second-order differences and the first-order upwind-discretized advection equation. In the context of shallow water flows discretized by the discontinuous Galerkin method, a production-destruction approach as in [1] guarantees non-negativity of the water height for any time step size while still preserving conservativity. In that work, the production-destruction equations where specifically formulated in order to account for the production and destruction terms which influence the cell-wise water volume.

Generally, numerical methods discretizing (1) are supposed to be positivity preserving, conservative and of sufficiently high order. While positivity preservation and conservativity may be directly carried over from the context of ODEs to that of PDEs, the issue of consistency and convergence is more subtle for PDEs. In this work, we hence take a closer look at the local discretization error of Patankar-type methods applied to systems arising from linear PDEs.

The Patankar-Euler Method And Its Modification

The forward Euler method applied to (1) will obviously be positivity preserving if we have $I - \Delta t\, Q(u^n) \geq 0$, but this requires a very severe time step restriction on Δt for stiff systems. To avoid this, a variant was proposed by Patankar [2], originally in the context of source terms in heat transfer. This method given by

$$u^{n+1} \ =\ u^n + \Delta t\, P(u^n)u^n - \Delta t\, Q(u^n)u^{n+1} \tag{2}$$

International Conference of Numerical Analysis and Applied Mathematics (ICNAAM 2016)
AIP Conf. Proc. 1863, 320008-1–320008-4; doi: 10.1063/1.4992489
Published by AIP Publishing. 978-0-7354-1538-6/$30.00

is unconditionally positivity preserving but not mass conserving. In addition, while (2) is of order one in the ODE sense, consistency is lost for stiff problems such as the discretized heat equation. In fact, the semi-discrete 1D heat equation $u_i'(t) = \frac{1}{\Delta x^2}\Big(u_{i-1}(t) - 2u_i(t) + u_{i+1}(t)\Big)$ for $i = 1, 2, \ldots, m$, with spatial periodicity, i.e. $u_0(t) = u_m(t)$ and $u_{m+1}(t) = u_1(t)$ fits in the form (1) with diagonal destruction matrix $Q(u) = 2\Delta x^{-2} I$. The Patankar-Euler scheme (2), written out per component, now reads $u_i^{n+1} = u_i^n + \frac{\Delta t}{\Delta x^2}\Big(u_{i-1}^n - 2u_i^{n+1} + u_{i+1}^n\Big)$. This scheme is unconditionally positivity preserving as well as unconditionally contractive in the maximum norm. However, inserting exact PDE solution values in the scheme, we obtain $u(x_i, t^{n+1}) = u(x_i, t^n) + \frac{\Delta t}{\Delta x^2}\Big(u(x_{i-1}, t^n) - 2u(x_i, t^{n+1}) + u(x_{i+1}, t^n)\Big) + \Delta t \rho_i^n$. Hence, Taylor development shows that for small Δt and Δx the leading term in these local truncation errors is given by $\rho_i^n = \frac{2\Delta t}{\Delta x^2} u_t(x_i, t^n) + O\Big(\frac{\Delta t^2}{\Delta x^2}\Big)$. It follows that the scheme will only be convergent in case of a very severe time step restriction of $\Delta t / \Delta x^2 \to 0$.

In order to obtain an unconditionally positive and additionally mass conservative scheme for production-destruction equations, the modification $u^{n+1} = u^n + \Delta t\, P(u^n) u^{n+1} - \Delta t\, Q(u^n) u^{n+1} = u^n + \Delta t\, A(u^n) u^{n+1}$ has been proposed in [3]. For linear problems $u' = Au$ with constant matrix A, such as the linear heat equation, this modified method now reduces to the implicit Euler method, so consistency in PDE sense is not a problem there. In [3], also a second-order method has been proposed. We will refer to this method as mPaRK2. This scheme does not fit directly in the vector production-loss formulation and thus has to be written per component, starting with the quasi-linear form $u_i' = \sum_{j=1}^m a_{ij}(u) u_j, i = 1, 2, \ldots, m$. The mPaRK2 method is then based on the trapezoidal rule with an Euler-type prediction to provide the internal stage value $v_i^{n+1} \approx u_i(t^{n+1})$ and reads

$$v_i^{n+1} = u_i^n + \Delta t \sum_j a_{ij}(u^n) v_j^{n+1}, \qquad u_i^{n+1} = u_i^n + \frac{1}{2}\Delta t \sum_j \Big(a_{ij}(u^n) \frac{u_j^n}{v_j^{n+1}} u_j^{n+1} + a_{ij}(v^{n+1}) u_j^{n+1}\Big). \tag{3}$$

As shown in [3], this scheme is unconditionally positivity preserving and mass conserving, and the order is two in the ODE sense. However, it is unknown whether there will be order reduction for stiff problems, in particular for semi-discrete problems obtained from PDEs after space discretization. Regarding the local discretization error, consistency of $O(\Delta t^3)$ can be proven for sufficiently smooth exact solutions. This is dealt with in the next section.

Error Recursions And Numerical Results

We will study error recursions for the mPaRK2 method applied to linear problems with constant coefficients. These are naturally non-linear for this method, even for linear equations. For a linear problem $u'(t) = Au(t)$ with $A = (a_{ij}) \in \mathbb{R}^{m \times m}$, we will first write (3) in vector form by introducing the diagonal matrix $W^n = \mathrm{diag}(u_i^n / v_i^{n+1})$. Then (3) can be written compactly as

$$v^{n+1} = u^n + \Delta t\, A v^{n+1}, \qquad u^{n+1} = u^n + \frac{1}{2}\Delta t\, A (W^n + I) u^{n+1}. \tag{4}$$

Along with this, we also consider the scheme with the exact solution inserted,

$$\bar{v}^{n+1} = u(t^n) + \Delta t\, A \bar{v}^{n+1}, \qquad u(t^{n+1}) = u(t^n) + \frac{1}{2}\Delta t\, A (\bar{W}^n + I) u(t^{n+1}) + \rho^n, \tag{5}$$

where $\bar{W}^n = \mathrm{diag}\Big(u_i(t^n)/\bar{v}_i^{n+1}\Big)$ and $\rho^n = (\rho_i^n) \in \mathbb{R}^m$. Subtraction of (4) from (5) gives a recursion for the global discretization errors $e^n = u(t^n) - u^n$ of the form $e^{n+1} = R^n e^n + d^n$, with amplification matrix and local errors given by

$$R^n = \Big(I - \frac{1}{2}\Delta t A\Big(\bar{W}^n + I\Big)\Big)^{-1}\Big(I + \frac{1}{2}\Delta t A G^n\Big), \quad d^n = \Big(I - \frac{1}{2}\Delta t A\Big(\bar{W}^n + I\Big)\Big)^{-1} \rho^n,$$

with the matrix $G^n \in \mathbb{R}^{m \times m}$ given by $G^n = \mathrm{diag}(u_i^{n+1}/\bar{v}_i^{n+1}) - \mathrm{diag}((u_i^n u_i^{n+1})/(\bar{v}_i^{n+1} v_i^{n+1}))(I - \Delta t A)^{-1}$. The difference between ρ^n and its counterpart resulting from the implicit trapezoidal rule can be determined from

$$\rho^n = u(t^{n+1}) - u(t^n) - \tfrac{1}{2}\Delta t A \left(u(t^n) + u(t^{n+1})\right) + \tfrac{1}{2}\Delta t A \left(u(t^n) - \bar{W}^n u(t^{n+1})\right). \tag{6}$$

Thus, the term $\tilde{\rho}^n = \frac{1}{2}\Delta t A \left(u(t^n) - \bar{W}^n u(t^{n+1})\right) = \frac{1}{2}\Delta t A\, \mathrm{diag}\left(\bar{v}_i^{n+1} - u_i(t^{n+1})\right)\left(\mathrm{diag}\left(\bar{v}_i^{n+1}\right)\right)^{-1} u(t^n) = \frac{1}{2}\Delta t A\, D_1\, D_2\, u(t^n)$ represents the difference in local errors between the implicit trapezoidal rule and the mPaRK2 scheme. For the diagonal matrices, we have $D_1 = \mathrm{diag}\left(\left((I - \Delta t A)^{-1} - e^{\Delta t A}\right) u(t^n)\right) = O(\Delta t^2)$ and $D_2 = \left(\mathrm{diag}\left((I - \Delta t A)^{-1} u(t^n)\right)\right)^{-1}$. In addition, if we assume $u(t^n) > 0$ then D_2 is bounded for $\Delta t \to 0$, i.e. $D_2 = O(1)$.

Semi-discrete linear advection and linear diffusion A reasonable assumption in the case of the semi-discrete linear advection with $A = \frac{1}{\Delta x}$ tridiag$[1\ -1\ \ 0]$ is a time-step choice such that $\Delta tA = O(1)$. Then we have

$$\begin{aligned}\tilde{\rho}^n &= \tfrac{1}{2}\underbrace{\Delta tA}_{O(1)}\left(\underbrace{\left((I-\Delta tA)^{-1}-e^{\Delta tA}\right)}_{O(\Delta t^2)}+\underbrace{\operatorname{diag}\left(\frac{u_i(t^n)-\bar{v}_i^{n+1}}{\bar{v}_i^{n+1}}\right)}_{O(\Delta t)}\underbrace{\left((I-\Delta tA)^{-1}-e^{\Delta tA}\right)}_{O(\Delta t^2)}\right)u(t^n) \\ &= \tfrac{1}{2}\Delta tA\left((I-\Delta tA)^{-1}-e^{\Delta tA}\right)u(t^n)+O(\Delta t^3).\end{aligned} \tag{7}$$

Due to the smoothness of the implicit Euler scheme, it holds that $A\left((I-\Delta tA)^{-1}-e^{\Delta tA}\right)u(t^n) = O(\Delta t^2)$. Hence, the local discretization error of the mPaRK2 method applied to the semi-discrete linear advection equation with positive initial data is bounded by $\rho^n = O(\Delta t^3)$.

Additional assumptions on the smoothness of the initial data are necessary for the semi-discrete linear diffusion equation with $A = \frac{1}{\Delta x^2}$ tridiag$[1\ -2\ \ 1]$ as shown in Fig. 1 and Table 1. As the diagonal matrix D_1 in the definition of $\tilde{\rho}^n$ can be bounded by $D_1 = \operatorname{diag}\left(\frac{1}{2}(\Delta tA)^2u(t^n)\right)+O((\Delta tA)^3)$, a smoothness condition on the exact solution of the type

$$S := A\operatorname{diag}\left((A^2u(t^n))_i\right)\left(\operatorname{diag}((I-\Delta tA)^{-1}u(t^n))\right)^{-1}u(t^n) = O(1) \tag{8}$$

guarantees $\tilde{\rho}^n = O(\Delta t^3)$ and hence a local error of third order. The left part of Fig. 1 depicts the situation for an initial solution $u_0 = 0.1+\sin^2(2\pi x)$ which satisfy the smoothness condition. Here, the quantities u_i/v_i approximate 1 for $\Delta t \to 0$ with $\Delta t = O(\Delta x)$. Consequently, the difference between the quantities $A^2u_0 \approx u_{0,xxxx}$ and $\operatorname{diag}(u_i/v_i)\,A^2u_0$ is small leading to basically constant smoothness indicators $\|S\|_2$ as shown in Table 1. This Table also lists the local error in the first mPaRK2 step. In accordance with the designed order of convergence, this local error behaves as $O(\Delta t^3)$. On the other hand, for an initial solution of $u_0 = \sin^2(2\pi x)$, Table 1 shows an order reduction to about $O(\Delta t^{2.3})$ in addition to an increasing value of the indicator $\|S\|_2$. As depicted in Fig. 1, this behavior is due the fact that for $u_i = 0$, we also have $u_i/v_i = 0$. Values at nearby grid points will tend to 1 for $\Delta t \to 0$ while zeros of u_i/v_i remain unchanged. This leads to the boundary layer effect for $u_i/v_i = 0$ visible in the right part of Fig. 1 as well as a locally large difference in curvature between A^2u_0 and $\operatorname{diag}(u_i/v_i)\,A^2u_0$.

TABLE 1. Effect of the smoothness condition (8) on the error of consistency for the mPaRK2 scheme.

	$u_0 = 0.1+\sin^2(2\pi x)$			$u_0 = \sin^2(2\pi x)$		
	L^2 loc. error	L^2 consistency	Indicator $\|S\|_2$	L^2 loc. error	L^2 consistency	Indicator $\|S\|_2$
$m = 40$	0.00177		1.70e+06	0.00218		2.55e+06
$m = 80$	0.00036	2.31	1.65e+06	0.00054	2.02	3.40e+06
$m = 160$	5.74e-05	2.64	1.53e+06	0.00012	2.20	4.89e+06
$m = 320$	8.13e-06	2.82	1.44e+06	2.45e-05	2.25	7.52e+06
$m = 640$	1.08e-06	2.91	1.39e+06	5.12e-06	2.26	1.21e+07
$m = 1280$	1.40e-07	2.95	1.38e+06	1.07e-06	2.26	1.98e+07
$m = 2560$	1.78e-08	2.97	1.81e+06	2.24e-07	2.25	3.30e+07

So far, these investigations show a local discretization error of order $O(\Delta t^3)$ for sufficiently smooth and positive solutions. This positivity requirement includes thin-layer approaches for the shallow water equations, where a thin film of water is retained also in regions marked as dry. However, we should remark that for a full convergence analysis, stability has to be proven as well. This necessitates boundedness of products of amplification matrices R^n which seems to be quite difficult to prove due to the non-linearity of the method.

Finally, we also consider a modification to the mPaRK2 scheme which follows more closely the approach in [1]. This modification is based on a direct correction of the explicit part of the implicit trapezoidal rule and reads as

$$v^{n+1/2} = u^n + \frac{\Delta t}{2}Au^n\,, \qquad u^{n+1/2} = u^n + \frac{1}{2}\Delta tA\tilde{W}u^{n+1/2}\,, \qquad u^{n+1} = u^{n+1/2} + \frac{1}{2}\Delta t\,Au^{n+1}\,, \tag{9}$$

with $\tilde{W} = \operatorname{diag}\left(u_i^n/\tilde{v}_i^{n+1/2}\right)$ determined by the correction $\tilde{v}^{n+1/2}$ to the quantity $v^{n+1/2}$ which may have negative components. More precisely, $\tilde{v}^{n+1/2}$ is given by $\tilde{v}_i^{n+1/2} = v_i^{n+1/2}$ if $v_i^{n+1/2} > 0$ and $\tilde{v}_i^{n+1/2} = u_i^n$ otherwise. We will denote this

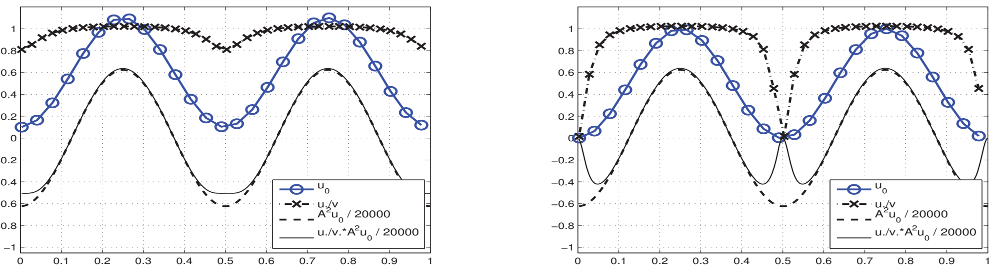

FIGURE 1. Initial conditions for the semi-discrete linear heat equation which satisfy (left) or do not satisfy (right) condition (8).

scheme by mPaRK2ex. Due to this switch in case of vanishing components, we cannot expect an overall second order of convergence as the update reduces to two steps of the implicit Euler scheme if $v^{n+1/2} = 0$. However, for a test case of an advected wave mimicking wetting and drying, i.e. advection of the initial condition $u_0 = 0.01 + \sin^4(\pi x)$, this method behaves much better than mPaRK2 as shown in Fig. 2.

A comparison of the Patankar-type schemes is carried out for the upwind-discretized linear advection on 160 grid points using spatial periodicity up to a final time of $T = 2$. As shown on the left of Fig. 2, using a time step of $\Delta t = 0.025$ corresponding to a Courant number of 4 does not exhibit significant differences of the schemes mPaRK2 and mPaRK2ex, also in comparison to the implicit trapezoidal rule. However, a larger time step of $\Delta t = 0.0625$ corresponding to a Courant number of 10 shows the drawback of mPaRK2 on the right part of Fig. 2. While mPaRK2 does not account for the vanishing solution in the interval $[0, 0.15]$ and the implicit trapezoidal rule clearly yields negative values, the modified scheme mPaRK2ex seems to combine the best features of both methods. The solution is non-negative in the whole computational domain and very accurate in the almost dry regions. Hence, this method seems quite promising and should be further investigated, in particular with respect to its stability.

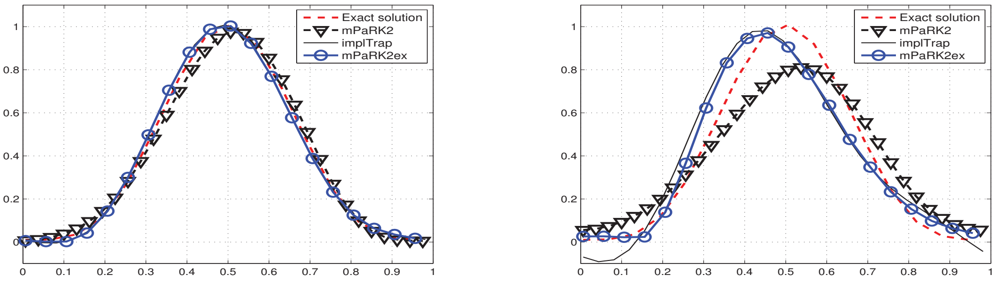

FIGURE 2. Linear transport of a wave using a Courant number of 4 (left) and 10 (right). Comparison of mPaRK2 to mPaRK2ex.

ACKNOWLEDGMENTS

S. Ortleb gratefully acknowledges the hospitality of the Centrum Wiskunde & Informatica where this research was partly carried out during her visit in the period of February/March 2016.

REFERENCES

[1] A. Meister and S. Ortleb, International Journal for Numerical Methods in Fluids **76**, 69–94 (2014).

[2] S. V. Patankar, *Numerical heat transfer and fluid flow*, Series in computational methods in mechanics and thermal sciences (Hemisphere Pub. Corp. New York, Washington, 1980).

[3] H. Burchard, E. Deleersnijder, and A. Meister, Appl. Numer. Math. **47**, 1–30 (2003).