Finite differencing schemes as Padé approximants

Peter Jäckel*

First version: 13th September 2011
This version: 6th May 2013

Abstract

In this note, we discuss Padé schemes [Smi86, chapter 3, pp 116-124] for the numerical integration of spatially discretized parabolic partial integro differential equations.

1 Introduction

There are many numerical schemes for the approximation of solutions to the parabolic partial integro-differential equation of the form

$$u_t + L \cdot u = 0 \tag{1.1}$$

with L being a Sturm-Liouville operator. Among them are, apart from simulation techniques by virtue of the Feynman-Kac theorem's transformation, various finite differencing, as well as finite element methods. In this note, we elaborate on the properties of some temporal finite differencing schemes with a particular emphasis on their categorisation as *Padé approximants* of the analytically exact solution.

2 Finite differencing schemes

Standard finite differencing schemes for equations of the form (1.1) are based on a spatially discretised approximation $\tilde{L}$ of the Sturm-Liouville operator L, and a *time-propagator* $A = A(t, \Delta t, \tilde{L})$ that is applied iteratively to the (terminal, or initial, depending on your application) spatially discretized representation $\tilde{u}$ of u. The standard iteration is simply

$$\tilde{u}(t - \Delta t) = A(t, \Delta t, \tilde{L}) \cdot \tilde{u}(t) \tag{2.1}$$

where we have assumed that the desired direction of solution is backwards in time. Analytically, we know of course that the formal solution to (1.1) is

$$u(t) = \mathrm{e}^{\int_t^T L \,\mathrm{d}s} \cdot u(T) \tag{2.2}$$

*Deputy head of Quantitative Research, VTB Capital

Key words and phrases. finite differencing, numerical solution of parabolic equations.

which translates directly to the spatially discretized version

$$\tilde{u}(t) = \mathrm{e}^{\int_t^T \tilde{L}\mathrm{d}s} \cdot \tilde{u}(T). \tag{2.3}$$

It is clear from here that, for a time-homogeneous setting, a scheme whose time-propagator is

$$A = \mathrm{e}^{\Delta t \cdot \tilde{L}} \tag{2.4}$$

incurs no temporal discretisation error. Unfortunately, the explicit computation of (2.4) requires the full diagonalisation of the operator $\tilde{L}$ which is a task that is numerically much harder than finding a numerical solution to (1.1). Nevertheless, we commonly use the Taylor expansion of any given scheme's time-propagator A to compare with the Taylor expansion of $\mathrm{e}^{\Delta t \cdot \tilde{L}}$ to determine the scheme's *temporal convergence order*.

A first order direct expansion of $\mathrm{e}^{\Delta t \cdot \tilde{L}}$ leads to

$$A_{(1,0)} = 1 + \Delta t \cdot \tilde{L} \tag{2.5}$$

and this is known as the *first order fully explicit scheme* (and we shall explain the subscript in due course).

3 Padé approximants

Clearly, many other approximations for A to $\mathrm{e}^{\Delta t \cdot \tilde{L}}$ are possible. The family of functions of the rational form

$$A(\tilde{L}) \;=\; R_{(m,n)}(\theta)\big|_{\theta=\Delta t\cdot\tilde{L}} \;\equiv\; \frac{P_m(\theta)}{Q_n(\theta)} \;\equiv\; \frac{1+p_1\theta+p_2\theta^2+\cdots+p_m\theta^m}{1+q_1\theta+q_2\theta^2+\cdots+q_n\theta^n} \tag{3.1}$$

are called the *Padé approximants* of order (m,n). The coefficients of $R_{(m,n)}(\theta)$ are selected to match the first $(m+n)$ derivatives of e^θ at zero, in line with the common meaning of Padé approximations. For e^x, they are explicitly known [Pad92, Var61]:

$$\mathrm{e}^x \;=\; \frac{\sum_{j=0}^m c_{m,n,j,m}\cdot(+x)^j}{\sum_{j=0}^n c_{m,n,j,n}\cdot(-x)^j} + \mathcal{O}\left(x^{(m+n+1)}\right) \tag{3.2}$$

with

$$c_{m,n,j,k} \;=\; \frac{(m+n-j)!k!}{(m+n)!j!(k-j)!}\,. \tag{3.3}$$

As a first example, let us consider the (0,1) Padé approximant

$$R_{(0,1)}(\theta) = \frac{1}{1-\theta}\,. \tag{3.4}$$

When we substitute this into the time-propagation rule (2.1), we obtain

$$\tilde{u}(t-\Delta t) = (1-\Delta t\cdot\tilde{L})^{-1}\cdot\tilde{u}(t) \tag{3.5}$$

which, in the context of linear operators, of course, simply means that we have to solve the system

$$(1 - \Delta t \cdot \tilde{L}) \cdot \tilde{u}(t - \Delta t) = \tilde{u}(t) \tag{3.6}$$

for $\tilde{u}(t-\Delta t)$. This scheme is conventionally known as the *first order fully implicit* scheme.

Next, consider the (1,1) Padé approximant

$$R_{(1,1)}(\theta) = \frac{1 + \frac{1}{2}\theta}{1 - \frac{1}{2}\theta} \tag{3.7}$$

which gives us

$$(1 - \tfrac{1}{2}\Delta t \cdot \tilde{L}) \cdot \tilde{u}(t - \Delta t) = (1 + \tfrac{1}{2}\Delta t \cdot \tilde{L})\tilde{u}(t) \tag{3.8}$$

also known as the *Crank-Nicolson* scheme.

Of particular interest in this note is the (0,2) Padé scheme

$$(1 - \Delta t \cdot \tilde{L} + \tfrac{1}{2} \cdot \Delta t^2 \cdot \tilde{L}^2) \cdot \tilde{u}(t - \Delta t) = \tilde{u}(t) \tag{3.9}$$

which we shall refer to as the *second order fully implicit (0,2) Padé* scheme.

Remark 3.1. The literature is unfortunately not unanimous on the notation of the Padé indices (m, n) in the context of finite differencing schemes for parabolic partical differential equations. This seems to depend largely on whether the approximant for e^x or e^{-x} is referred to, and on whether the respective parabolic equation is being solved forward or backward in time. Smith [Smi86, chapter 3, pp 116-124], for instance, refers to the first order explicit scheme as the (0,1) scheme, as does Varga [Var61], and to the (1,0) Padé approximant as the backward implicit scheme, and so on, whereas other authors [TZA+06, WKY+07, Sid10] tend to use the notation that relates the first index m in (m, n) to the explicit part and the second index, n, to the implicit part of the calculation. We use the latter notation throughout.

4 Stability and oscillations

The numerical stability of any time-propagation scheme is governed by the eigenvalues of its time-propagator A. Modes associated with eigenvalues larger than one in absolute value will grow with each iteration of (2.1). Modes with eigenvalues whose real parts are negative respond with a sign reversal under (2.1), making them appear to oscillate, or *ring*, as we iterate. When all eigenvalues are inside the unit circle, or exactly at one on the real axis, the scheme is considered *unconditionally stable*. Schemes with eigenvalues with negative real parts are referred to as *oscillatory*, or *ringing*. When all eigenvalues have non-negative real parts, the scheme is considered *non-oscillatory*, or *quiet*.

For Padé approximants, the spectrum of A can readily be related to the spectrum of $\tilde{L}$ since for every eigenvector $\boldsymbol{x}_i$ of $\tilde{L}$ that satisfies

$$\tilde{L} \cdot \boldsymbol{x}_i = \lambda_i \boldsymbol{x}_i \tag{4.1}$$

we have

$$P_n(\Delta t \cdot \tilde{L}) \cdot \boldsymbol{x}_i = P_n(\Delta t \cdot \lambda_i) \cdot \boldsymbol{x}_i \tag{4.2}$$

$$Q_m(\Delta t \cdot \tilde{L}) \cdot \boldsymbol{x}_i = Q_m(\Delta t \cdot \lambda_i) \cdot \boldsymbol{x}_i \tag{4.3}$$

$$R_{(m,n)}(\Delta t \cdot \lambda_i) \cdot Q_m(\Delta t \cdot \tilde{L}) \cdot \boldsymbol{x}_i = P_n(\Delta t \cdot \tilde{L}) \cdot \boldsymbol{x}_i \tag{4.4}$$

and thus

$$R_{(m,n)}(\Delta t \cdot \tilde{L}) \cdot \boldsymbol{x}_i = R_{(m,n)}(\Delta t \cdot \lambda_i) \cdot \boldsymbol{x}_i \ . \tag{4.5}$$

In other words, the eigenvector $\boldsymbol{x}_i$ of $\tilde{L}$ with eigenvalue λ_i is also eigenvector of $R_{(m,n)}(\Delta t \cdot \tilde{L})$ with eigenvalue $R_{(m,n)}(\Delta t \cdot \lambda_i)$.

5 Pure diffusions

Consider a linear operator $\tilde{L}$ whose matrix representation is given by

$$\tilde{L} = \begin{pmatrix} -2 & 1 & 0 & 0 & \cdots & 0 \\ 1 & -2 & 1 & 0 & & \vdots \\ 0 & 1 & -2 & 1 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ \vdots & & & 1 & -2 & 1 \\ 0 & \cdots & \cdots & 0 & 1 & -2 \end{pmatrix} . \tag{5.1}$$

This form arises from the centered uniform discretisation of pure diffusions in one dimension[1]. The matrix (5.1) is a special variation of the family of Toeplitz matrices. Its eigenvalues are well known [Ray94], [Smi86, chapter 3, page 120, between (3.14) and (3.15)] to be

$$\lambda_k = -2 + 2\cos\left(\tfrac{k \cdot \pi}{N+1}\right) \tag{5.2}$$

with N being the spatial discretization number. Eigenvalues with a higher index k are associated with higher spatial frequencies, which means that a good time-propagator should ideally have eigenvalues monotonically decreasing in absolute value with increasing index k since diffusions dampen out high frequencies. Note that for (5.2), the eigenvalues for low k are near 1, and, as k increases, converge to -4. An example is shown in figure 1.

Given that the eigenvalues of (5.1) are always bounded by $[-4, 0]$, the spectrum of any specific (m, n) Padé approximant time-propagator for a pure one-dimensional diffusion is bounded by the range to which the interval $[-4 \cdot \Delta t, 0]$ is mapped by the respectively associated function $R_{(m,n)}(\cdot)$. We show the stability implications for the first five Padé schemes as a function of the time step size Δt in figure 2. We can see that the standard first

[1] The symmetric form shown in (5.1) does in fact include a Dirichlet boundary condition of the solution being zero on either edge, with the edge nodes being excluded from the discretization. The following discussion is, however, in practice, largely unaffected by the small difference in the spectrum of $\tilde{L}$ that arises from the choice of different boundary conditions.

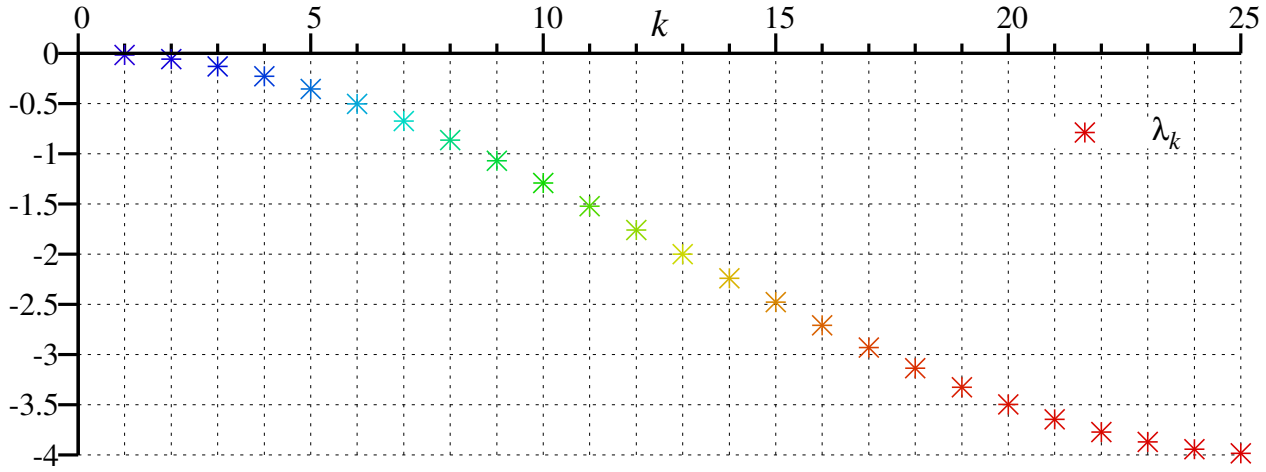


FIGURE 1: The eigenvalues of matrix (5.1) for $N = 25$.

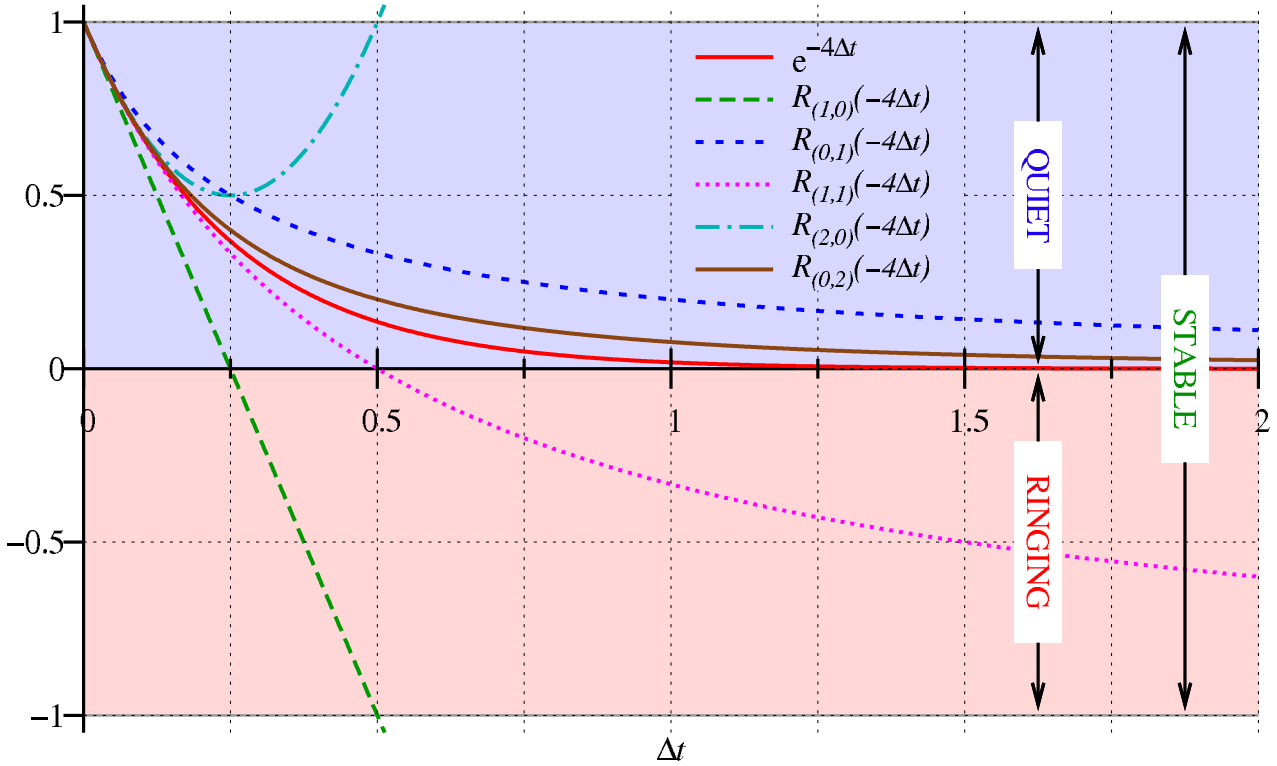


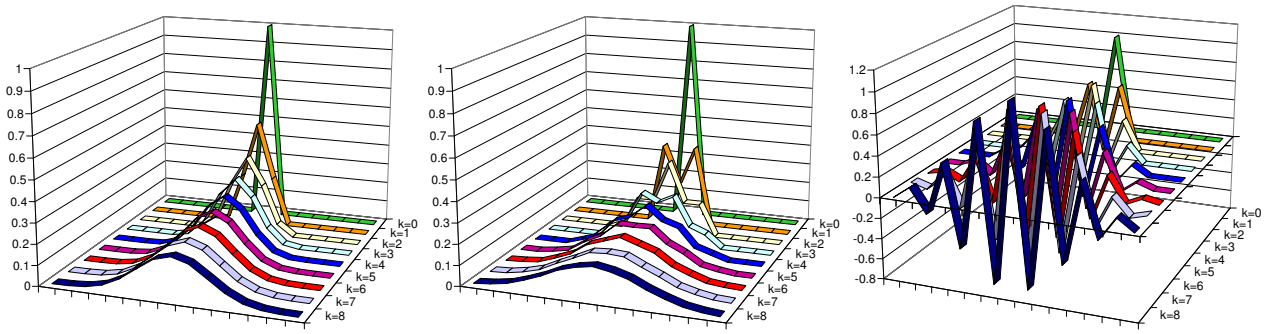
FIGURE 2: The first five Padé approximant schemes' stability zones.

order fully explicit $(1, 0)$ scheme is stable when $\Delta t < 1/2$, and quiet when $\Delta t < 1/4$. The second order $(1, 1)$ Crank-Nicolson scheme starts ringing at $\Delta t \geq 1/2$ but remains stable for all Δt. The second order fully explicit $(2, 0)$ scheme given by $R_{(2,0)}(\lambda) = 1+\lambda+\frac{1}{2}\lambda^2$ has the same stability range as the first order scheme, but never oscillates. In contrast, both the first and the second order fully implicit schemes, $(0, 1)$ and $(0, 2)$ are unconditionally stable and quiet.

The above observations have been made for pure one-dimensional diffusions. It is worth noting, however, that even where analytical proofs have not been obtained, it has been observed empirically that most of the stability properties still apply when multi-dimensional diffusions, or even other types of PIDEs, are discretized, including correlation,

as well as when convection terms are being added.

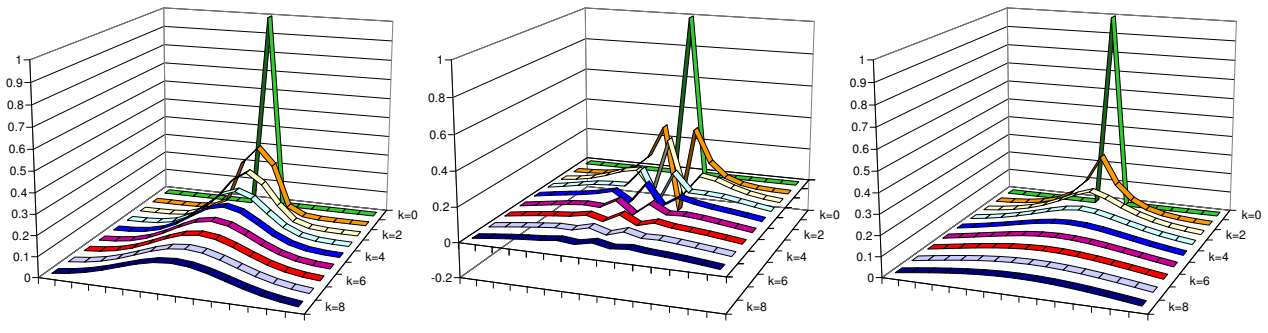
In figure 3, we show the behaviour of the fully explicit (1,0) Padé scheme for three different step sizes: in (a), we have a small step size such that the scheme appears to converge smoothly; in (b), the step size is beyond 1/4 such that we observe the onset of oscillations caused by negative eigenvalues; and in (c), the step size is greater than 1/2 and the scheme has become unstable. In contrast, we show in figure 4 the behaviour of



(a) (b) (c)

FIGURE 3: The (1,0) Padé scheme subject to an input given by a central spike for different time step sizes and up to $k = 8$ iterations. In (a), we have $\Delta t = 0.24$, in (b), $\Delta t = 0.4$, and in (c), we have $\Delta t = 0.564$. Note that the oscillations in (b) die down as the scheme is iterated, meaning that these are the signature of negative eigenvalues and *not* instabilities.

the (1,1) Padé scheme for two different step sizes, and the (0,1) scheme for just one large step size.



(a) (b) (c)

FIGURE 4: The (1,1) and (0,1) Padé schemes subject to an input given by a central spike for different time step sizes and up to $k = 8$ iterations. In (a), we have the (1,1) scheme with $\Delta t = 0.6$, and in (b), the same scheme with $\Delta t = 2.5$. In (c), we have the (0,1) scheme with $\Delta t = 2.5$. Note that the oscillations in (b) die down as the scheme is iterated, meaning that these are the signature of negative eigenvalues and *not* instabilities, and that the fully implicit (0,1) scheme shows no oscillations nor instabilities even for the large step.

6 Direct implementation of the $(0,2)$ Padé scheme

For pure one-dimensional diffusions, an explicit matrix representation of the linear system that needs to be solved for the $(0,2)$ Padé scheme

$$\left[A_{(0,2)}(\Delta t \cdot \tilde{L})\right]^{-1} = \left[1 - \Delta t \cdot \tilde{L} + \tfrac{1}{2}\Delta t^2 \cdot \tilde{L}^2\right] \tag{6.1}$$

results in a pentadiagonal matrix. In comparison, both the first order $(0,1)$ scheme, and the $(1,1)$ scheme give rise to a tridiagonal system, and hence standard procedures of explicit substitution to solve the respective linear systems are, judging by the mere number of required floating point operations, more than twice as fast for the $(0,1)$ and $(1,1)$ schemes. This makes especially the $(1,1)$ Crank-Nicolson scheme often the preferred choice over the $(0,2)$ scheme. An exception is the situation when time steps significantly larger than the (normalized) $\Delta t = 1/2$ threshold are to be used, and, at the same time, the right hand side of (2.1) contains components colinear with the spatial high frequency modes that are associated with the negative eigenvalues of largest absolute value present in this scheme for large Δt. High frequency modes are typically part of the right hand side of (2.1) when initial conditions are not smooth in some form or another, which, in financial modelling applications, alas, is very common. Another exception is when the effective linear system is not solved with explicit substitution schemes, but with iterative methods, as is usually done for multi-dimensional partial integro differential equations. In that case, the advantage of a tridiagonal matrix representation is no longer so obvious, albeit that in practice even with iterative methods the solution of the linear systems for the $(0,1)$ and the $(1,1)$ schemes is usually faster than that for the $(0,2)$ scheme.

For an efficient implementation of $\left[A_{(0,2)}(\Delta t \cdot \tilde{L})\right]^{-1}$ for iterative solvers, we mention that it is advantageous to use its Horner form representation

$$\left[A_{(0,2)}(\Delta t \cdot \tilde{L})\right]^{-1} = \left[1 - \Delta t \cdot \tilde{L} \cdot \left(1 - \tfrac{1}{2}\Delta t \cdot \tilde{L}\right)\right] \ . \tag{6.2}$$

In practical terms, this means whenever within the iterative linear solver routine for the solution of

$$M \cdot x = b \tag{6.3}$$

with

$$M = \left[A_{(0,2)}(\Delta t \cdot \tilde{L})\right]^{-1}, \qquad x = \tilde{u}(t - \Delta t), \qquad b = \tilde{u}(t),$$

the evaluation of $y := M \cdot x$ is required, we execute the following steps:-

$$x' \ = \left(1 - \tfrac{1}{2}\Delta t \cdot \tilde{L}\right) \cdot x \tag{6.4}$$

$$y := x - \Delta t \cdot \tilde{L} \cdot x' \ . \tag{6.5}$$

6.1 Complex time

Another representation of (6.1) is

$$\left[A_{(0,2)}(\Delta t \cdot \tilde{L})\right]^{-1} = \left(1 - \tau_{(+)} \cdot \tilde{L}\right)\left(1 - \tau_{(-)} \cdot \tilde{L}\right) \tag{6.6}$$

with

$$\tau_{(+)/(-)} = \tfrac{1}{2}\,(1 \pm i) \cdot \Delta t \tag{6.7}$$

as suggested by J. Andreasen at the September 2011 Danske Kwant Festival in Copenhagen[2]. This means, instead of directly solving

$$\left[1 - \Delta t \cdot \tilde{L} \cdot \left(1 - \tfrac{1}{2}\Delta t \cdot \tilde{L}\right)\right] \cdot \tilde{u}(t - \Delta t) = \tilde{u}(t)\ , \tag{6.8}$$

we can instead proceed in two *complex time* steps, namely, first solve

$$\left[1 - \tau_{(+)} \cdot \tilde{L}\right] v = \tilde{u}(t) \tag{6.9}$$

for v, then solve

$$\left[1 - \tau_{(-)} \cdot \tilde{L}\right] \tilde{u}(t - \Delta t) = v \tag{6.10}$$

for $\tilde{u}(t - \Delta t)$. The apparent advantage of this is that both of these steps only involve tridiagonal systems when explicit matrix representation linear system solving is used, as opposed to the pentadiagonal system resulting from (6.8). However, it is not clear whether the decomposition (6.9)–(6.10) really does lead to a numerical advantage due to the fact that the total number of floating point operations for (explicitly) solving two tridiagonal systems in complex arithmetic is signficantly higher than the number of floating point operations required for the solution of a single pentadiagonal system[3]. In addition, we refer to section 10 for further reasons as to why taking two complex time steps in sequence is numerically inexpedient.

7 Other implicit second order schemes

Padé approximants are of course not the only possible choice for the approximation of $\mathrm{e}^{\Delta t \cdot \tilde{L}}$. Consider, for example, solving in steps according to

$$(1 - \Delta t \cdot \tilde{L}) \cdot v = \tilde{u}(t) \tag{7.1}$$

$$(1 - \Delta t \cdot \tilde{L}) \cdot w = v \tag{7.2}$$

and setting

$$\tilde{u}(t - \Delta t) := -\tfrac{1}{2}\tilde{u}(t) + 2v - \tfrac{1}{2}w\ . \tag{7.3}$$

[2] It came to light in a subsequent literature research that the method of splitting into two steps with complex stepping coefficients had been known among specialist for some time [GS89b, TZA+06].

[3] Of particular cost are complex divisions, and a host of numerical analysis and engineering literature is dedicated to efficient representation of complex divisions as real-valued operations. Cursory experiments with the highly optimized GNU C++ library implementation of complex operations indicated that a single complex division, on average, may cost as much as 14 real-valued floating point operations comprised by addition, subtraction, multiplication, and division.

This is equivalent to the time-propagator definition

$$A_{uvw}(\Delta t \cdot \tilde{L}) := -\tfrac{1}{2} + 2 \cdot \left(1 - \Delta t \cdot \tilde{L}\right)^{-1} - \tfrac{1}{2} \cdot \left(1 - \Delta t \cdot \tilde{L}\right)^{-2} \tag{7.4}$$

where we have dubbed this scheme the (uvw)-*scheme* for reasons of lack of imagination. Taylor expansion gives

$$A_{uvw}(x) := 1 + x + x^2/2 + \cdots \tag{7.5}$$

which means that this is a second order scheme. Alternatively, consider the Lawson-Morris [LM78] scheme

$$(1 - \tfrac{1}{2}\Delta t \cdot \tilde{L})^2 \cdot v = \tilde{u}(t) \tag{7.6}$$

$$(1 - \Delta t \cdot \tilde{L}) \cdot w = \tilde{u}(t) \tag{7.7}$$

and setting

$$\tilde{u}(t - \Delta t) := 2v - w \; . \tag{7.8}$$

This is equivalent to the time-propagator definition

$$A_{LM}(\Delta t \cdot \tilde{L}) := 2 \cdot \left(1 - \tfrac{1}{2}\Delta t \cdot \tilde{L}\right)^{-2} - \left(1 - \Delta t \cdot \tilde{L}\right)^{-1} \; . \tag{7.9}$$

Its Taylor expansion also confirms this as a second order scheme:

$$A_{LM}(x) := 1 + x + x^2/2 + \cdots \tag{7.10}$$

Unfortunately, as first empirically noticed and reported by J. Andreasen to the author, neither of these schemes preserve positivity for pure diffusions as the time step Δt is increased. We will return to this point in section 9.

To understand the features of these schemes, we note that they can be represented as rational functions in standard form:-

$$\begin{aligned} A_{uvw}(x) &= \frac{2 - 2x - x^2}{2 - 4x + 2x^2} \\ A_{LM}(x) &= \frac{-4 + 4x + x^2}{-4 + 8x - 5x^2 + x^3} \end{aligned} \tag{7.11}$$

For pure diffusions, we can apply the same eigenvalue analysis we previously did for Padé approximants. We show the highest spatial frequency eigenvalue bound as a function of Δt in figure 5. We can see from this that for $\Delta t < 1/2$ both the Lawson-Morris and the (uvw)-scheme are better approximations for the exponential function than the $(0, 2)$ Padé approximant. We can also see by further consultation of equations (7.11) that they are both unconditionally stable. Alas, however, it emerges that neither of those schemes are without oscillations. In fact, we can determine that the Lawson-Morris scheme has negative eigenvalues when $\Delta t > (1+\sqrt{2})/2 \approx 1.21$, albeit that negative eigenvalues never exceed ≈ -0.04. The same applies to the (uvw)-scheme when $\Delta t > (1 + \sqrt{3})/4 \approx 0.68$, with negative eigenvalues never exceeding $-1/2$.

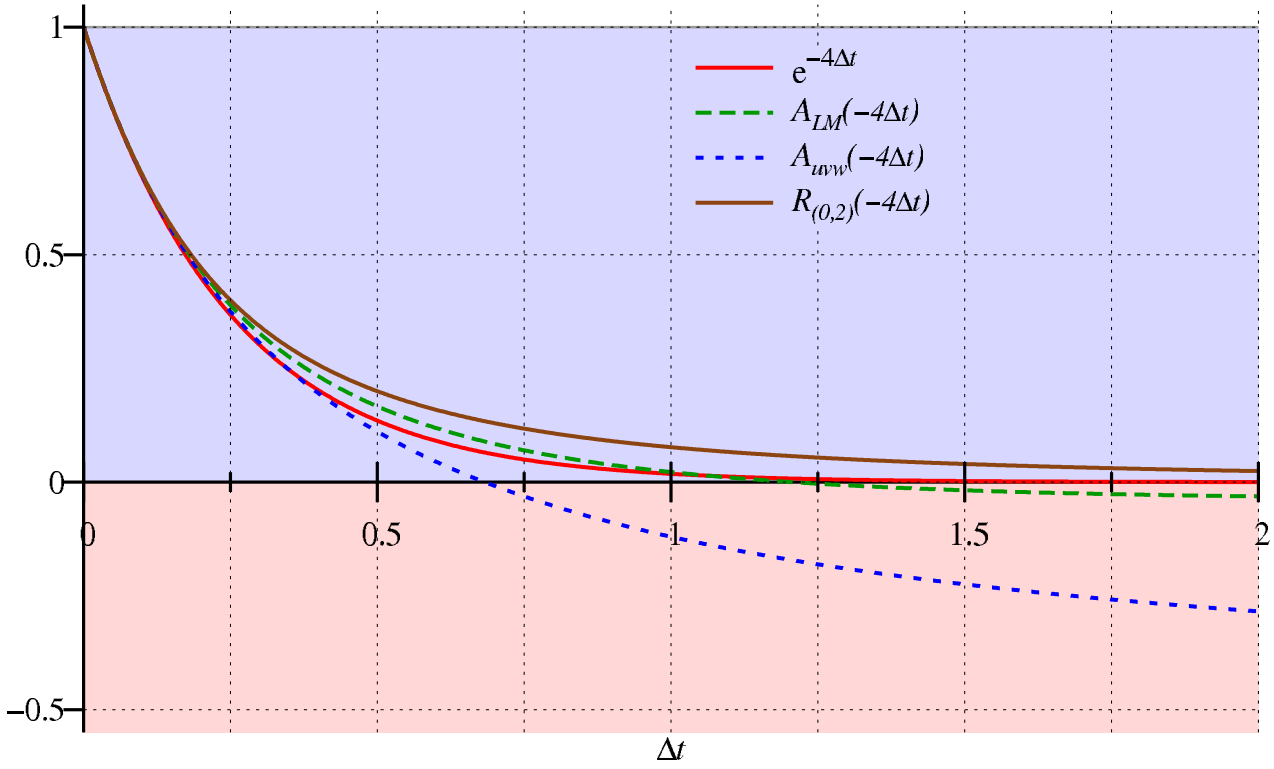


FIGURE 5: Highest spatial frequency eigenvalue bound of the Lawson-Morris and the (uvw)-scheme.

The crux of the above mentioned mixing schemes is that, whilst they are both the result of linear combinations of sub-step schemes that *individually* are unconditionally stable and quiet, they are linear combinations involving *negative weights*, as can be seen in the coefficients on the right hand side of equations (7.3) and (7.8). Linear combinations with negative weights, unlike *convex combinations*, do not preserve any positivity guarantees. This phenomenon is not unique to finite differencing schemes: it is exactly the same that gives rise to stable quadrature schemes only ever involving positive coefficients as, e.g., in Gauss-Hermite, Gauss-Legendre, Clenshaw-Curtis, Kronrod, and so many other quadrature rules.

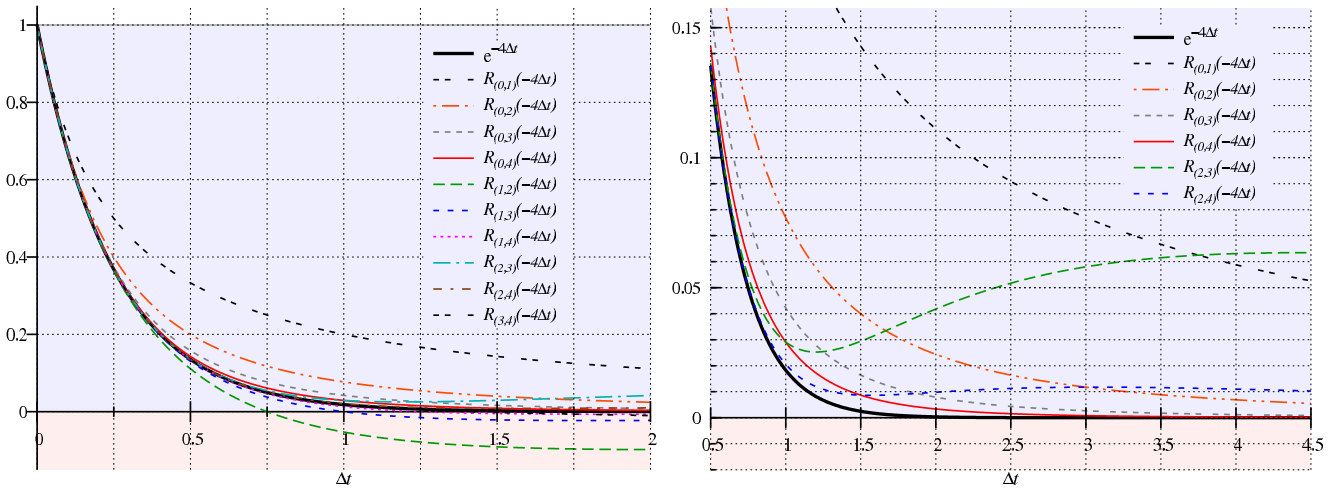
8 Higher order Padé schemes

It is clear from the rational form (3.1) that all $R_{(m,n)}(x)$ with $m > n$ grow unbounded for $x \to -\infty$, and hence associated schemes are not unconditionally stable. Also, for the so-called *diagonal* approximants given by $R_{(m,m)}(x)$, we can deduce from the coefficients given in (3.2) that

$$\lim_{x\to-\infty} R_{(m,m)}(x) = (-1)^m \,. \tag{8.1}$$

This means that high frequency modes are in each finite difference scheme iteration step attenuated by a factor that, in absolute value, approaches 1 in the limit of long step sizes and refined spatial discretization. Since diffusions in the exact analytical limit dampen higher frequencies faster, this is an undesirable behaviour for any scheme we wish to use for all step sizes, including very large steps.

Having excluded all (m, n) schemes with $m \geq n$, we show the Padé approximation of e^{-x} of all remaining schemes for all $m, n \leq 4$ in figure 6. To the left, in (a), we can see



(a) (b)

FIGURE 6: Padé approximants of e^{-x} for $m < n$, $m, n \leq 4$.

that various schemes allow for negative eigenvalues. Further excluding those, we show the behaviour of the remaining six schemes, for a wider range of Δt, in (b). We can see that out of those remaining six schemes, only those with $m = 0$ are monotonic functions of Δt. Since we consider it desirable to preserve monotonicity in order to avoid high frequency modes being less dampened in *longer* time steps, we exclude the $(2, 3)$ and $(2, 4)$ schemes, too. This leaves us with the fully implicit schemes $(0, n)$ with $n = 1, 2, 3, 4$ as candidates for high accuracy finite differencing schemes. Finally, we point out that we can see in figure 6(b) that the $(0, n)$ schemes, with increasing n, provide a formidable approximation to e^{-x} indeed.

Remark 8.1. The first two properties we required in the above selection of finite differencing schemes, namely, positivity of the Padé approximant $R_{(m,n)}(x)$, and the limiting behaviour $\lim_{x\to-\infty} R_{(m,n)}(x) = 0$, form what is known as *L-stability* in numerical analysis. In addition to L-stability, we also demanded monotonicity of $R_{(m,n)}(-x)$ for $x > 0$.

9 Positivity

A rarely discussed aspect of finite differencing scheme is the concept of *positivity preservation*. We consider a scheme *positivity preserving* if for any vector $\tilde{u}$ that is non-negative in all elements, the iteration result

$$A(t, \Delta t, \tilde{L}) \cdot \tilde{u}$$

remains non-negative in all elements. For this to hold for all possible non-negative choices of the vector $\tilde{u}$, the time-propagator A must be non-negative in all its elements. When the diffusion generator $\tilde{L}$ is symmetric and tridiagonal in the standard form (5.1), then

$$A^{-1}_{(0,1)} \;=\; 1 - \Delta t \cdot \tilde{L}$$

forms a *Stieltjes matrix*[4]. Since any Stieltjes matrix is invertible to a matrix with non-negative entries, this means that the $(0,1)$ scheme preserves positivity for any step size Δt. Equally, we can see that the $(1,1)$ scheme preserves positivity when $\Delta t \leq 1$ by inspecting

$$A_{(1,1)} \;=\; \left(1 - \tfrac{1}{2}\Delta t \cdot \tilde{L}\right)^{-1} \cdot \left(1 + \tfrac{1}{2}\Delta t \cdot \tilde{L}\right) \;.$$

The first term on the right hand side is always positivity preserving as already discussed. The second term is non-negative in all entries when $\Delta t \leq 1$ as can be seen from the elments in $\tilde{L}$ as shown in equation (5.1), and the product of two positivity preserving operators does of course also preserve positivity. Unfortunately, however, for other schemes, and in particular for the higher order fully implicit $(0,n)$ schemes with $n = 2, 3, \ldots$ that are unconditionally stable and non-oscillatory, as well as approximating e^{-x} very accurately, positivity is in general not preserved. We demonstrate this visually in figures 7 to 10 where we display the leading digits and the exponents of the propagators $A_{(0,1)}$, $A_{(0,2)}$, $A_{(0,3)}$, and $A_{(0,4)}$, for a discretisation with 35 spatial levels with negative entries in red colour. We can clearly see a structure of negative bands for $A_{(0,n)}$ for $n > 1$, albeit that

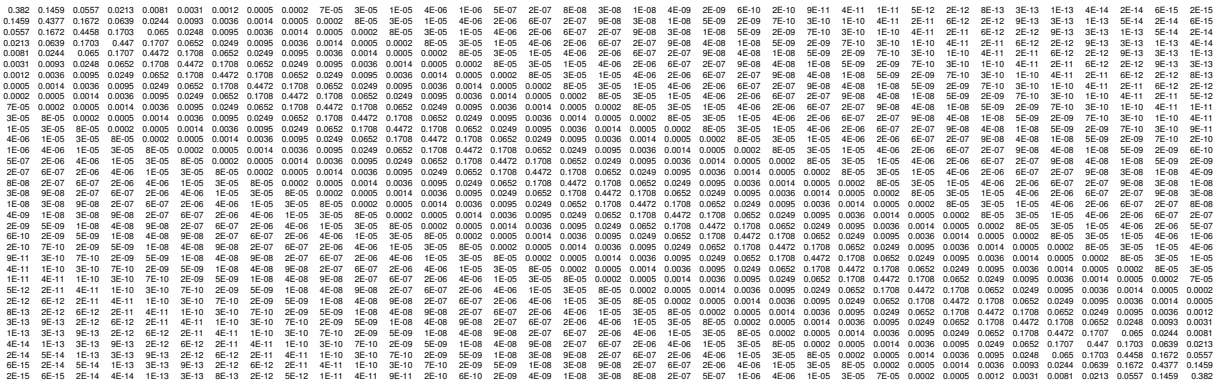


FIGURE 7: $A^{35\times35}_{(0,1)}$ for $\Delta t = 1$.

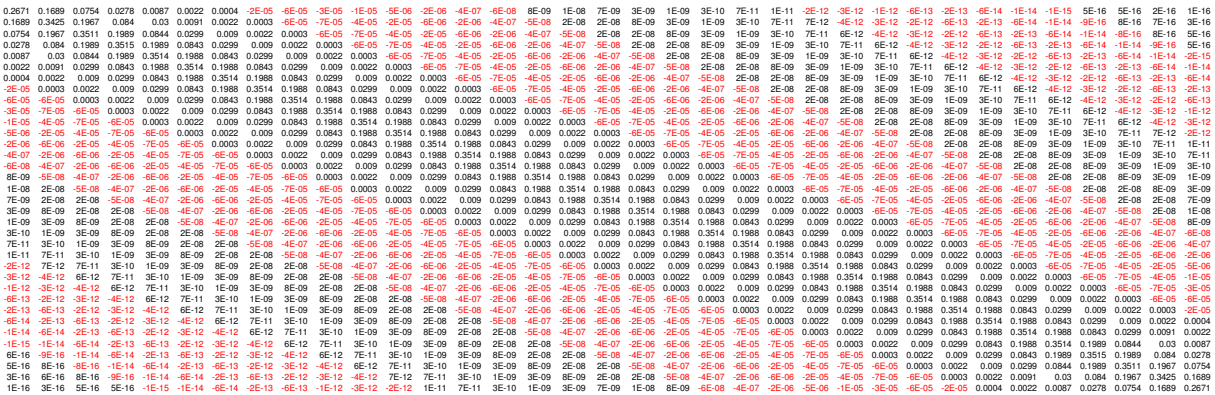


FIGURE 8: $A^{35\times35}_{(0,2)}$ for $\Delta t = 1$.

the negative entries are small in absolute value. An even better visualization is given by the display of those matrices as surface plots since this facilitates an intuitive relative comparison of the extent to which $A_{(0,n)}$ for $n > 1$ neglects positivity. Since the absolute

[4] A real symmetric positive definite matrix with nonpositive off-diagonal entries.

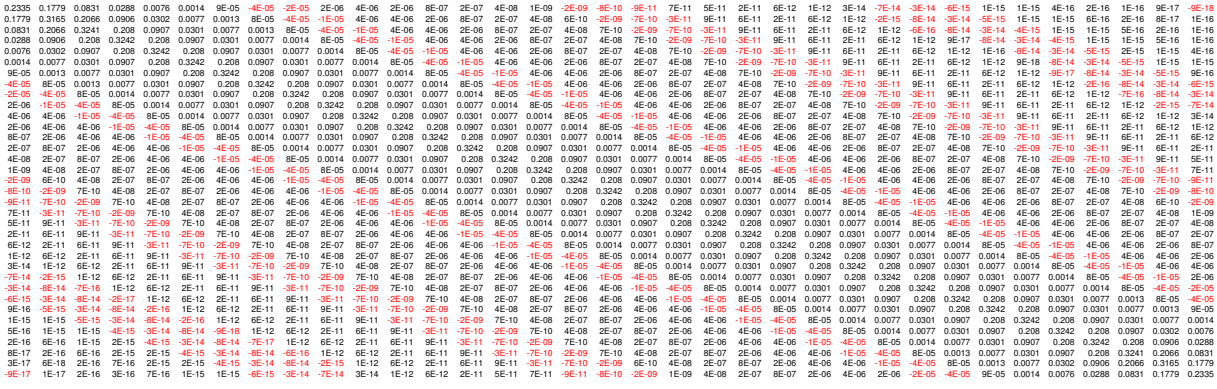


FIGURE 9: $A^{35\times35}_{(0,3)}$ for $\Delta t = 1$.

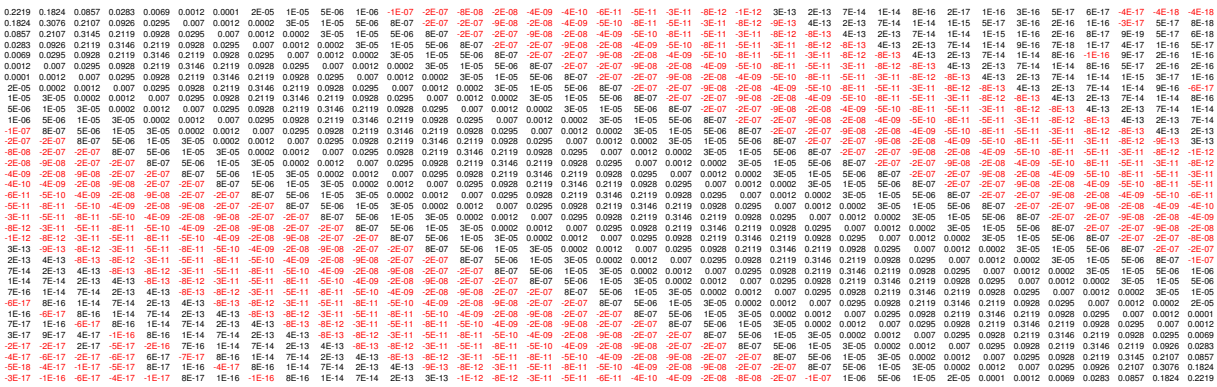


FIGURE 10: $A^{35\times35}_{(0,4)}$ for $\Delta t = 1$.

numbers are small, a direct plot simply results in surfaces with a dominant ridge along the diagonal. Hence, we use a non-linear scaling to enhance the visual magnitude of the small off-diagonal values, and show in figure 11 the values of $A^{1/5}_{(0,1)}$, $A^{1/5}_{(0,2)}$, $A^{1/5}_{(0,3)}$, and $A^{1/5}_{(0,4)}$, with the power coefficient $(1/5)$ having been chosen more or less arbitrarily to emphasize the visibility of the bands of non-negligible negative values. We can clearly see that whilst $A_{(0,1)}$ is positive in all entries, the higher order fully implicit propagators display negative bands (marked in red where non-negligible). They are most pronounced for $A_{(0,2)}$, and become smaller in absolute magnitude as n is increased in $A_{(0,n)}$.

The occurrence of negative entries in the time propagators $A_{(0,n)}$ for $n > 1$ is not independent from the choice of step size Δt. We show in figure 12 a study of the magnitude of the most negative entry in the time propagators for a spatial discretization with 101 levels. We can see that for moderately small step sizes the most negative entries in $A_{(0,4)}$ are in absolute value orders of magnitude smaller than those for $A_{(0,3)}$ and $A_{(0,2)}$. We also observe that for $A_{(0,3)}$ there appears to be a critical threshold below which the matrix systematically preserves positivity. We have already seen, albeit from analytical considerations, the existence of a critical threshold for the preservation of positivity with the $(1,1)$ scheme and thus need not be too surprised about this phenomenon. Whilst the reasons for it to happen for the $(0,3)$ are not as straightforwardly obvious as for the $(1,1)$, we are content with the numerical evidence for it, and don't pursue this issue any further in this context.

Remark 9.1. A matrix that is non-negative in all its elements is often called simply a

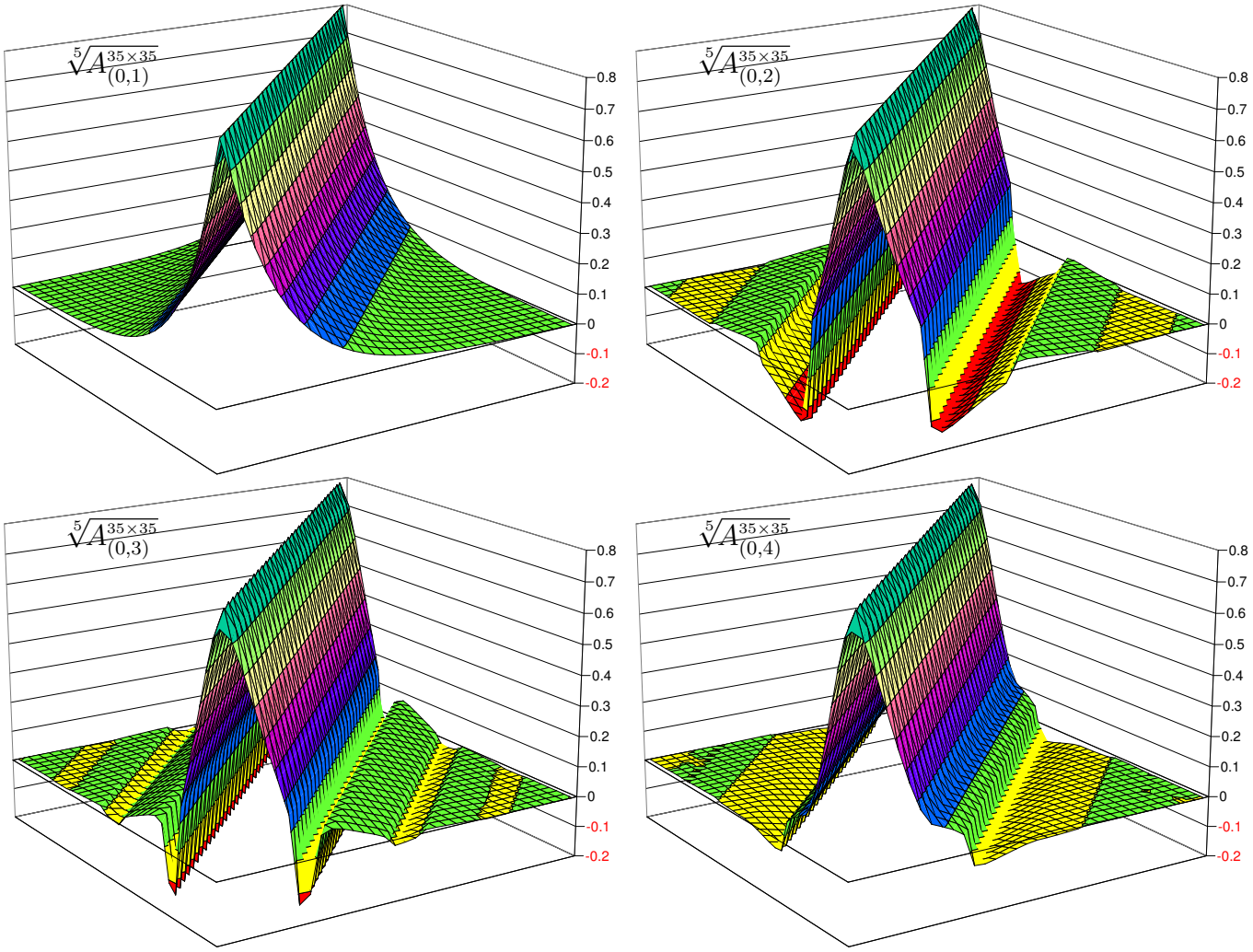


FIGURE 11: Fully implicit propagator matrices for 35 spatial levels and $\Delta t = 1$, arbitrarily scaled by a power of (1/5) to aid visualization of negative regions of small but non-negligible absolute value.

non-negative matrix in the literature. Here, however, we do not use this terminology to avoid the confusion arising from the fact that some authors define *non-negative* to mean having only real-valued non-negative eigenvalues. Unfortunately, the ambiguity in the terminology in the literature does not end there: some authors [Sid10] refer to Padé schemes for parabolic equations as *positivity preserving* when they have no negative eigenvalues. The latter definition is used to refer to whether the Padé approximant $R_{(m,n)}(-x)$ preserves the sign of the function e^{-x} that it is to approximate, and this translates to the absence of negative eigenvalues. For the finite differencing scheme to preserve positivity of the solution vector under iteration, however, this is not sufficient.

10 Partial Fraction Decomposition

We have seen that higher order fully implicit Padé finite difference integration schemes can be of significant benefit since they permit comparatively large step sizes whilst retaining good accuracy. A consideration of practical concern, particularly for large numbers of spatial discretization levels, is the question as to the chosen method for the solution of

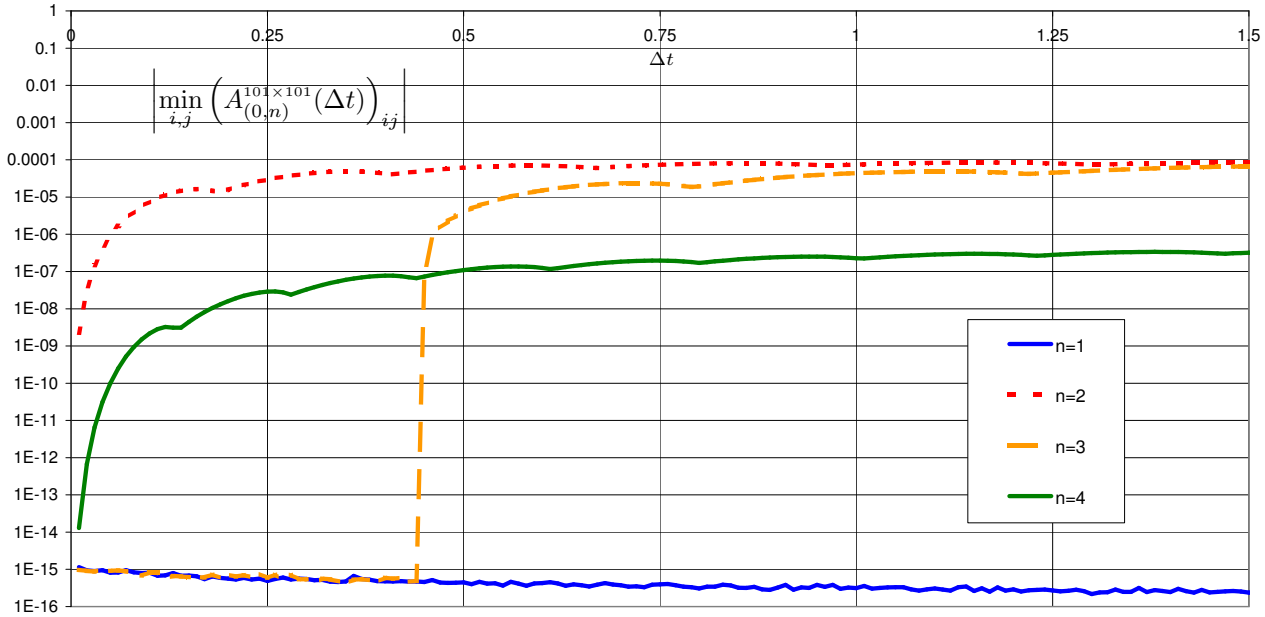


FIGURE 12: The magnitude of the most negative entry in the time propagator matrix $A^{101\times101}_{(0,n)}$ as a function of step size Δt for $n = 1, 2, 3, 4$. Note that the line for $A^{101\times101}_{(0,1)}$ is at the level of numerical resolution and thus represents the base line that ought to be considered numerically equivalent to zero.

the generated linear systems. For multi-dimensional diffusion problems, almost invariably, one will wish to resort to iterative methods such as the BiCGStab method [dV92]. For one-dimensional diffusions, however, for schemes that are of first order in their implicit part, e.g., the $(1, 1)$ Crank-Nicolson or the $(0, 1)$ fully implicit scheme, the resulting linear system is tri-diagonal in its matrix representation, and amenable to highly efficient direct forward- and back-substitution. Higher order implicit schemes when represented in their inverse matrix form, are no longer tri-diagonal. In fact, the matrix representation of $A^{-1}_{(0,n)}$ has a total of $2n + 1$ diagonals that are non-zero, meaning that $A^{-1}_{(0,2)}$ is pentadiagonal, $A^{-1}_{(0,3)}$ is hepta-diagonal, and $A^{-1}_{(0,4)}$ is nona-diagonal. This does not only increase the numerical effort, which is ultimately inevitable, but it also increases the risk of the linear system ending up numerically ill-conditioned, especially when no pivoting is employed in aid of maximum speed. One way to overcome this [Gal89, GS89a, GS89b], is to decompose the denominator of $R_{(0,n)}$ as given in (3.1) into its roots

$$Q_n(x) \equiv q_n \cdot \prod_{j=1}^{n} (\rho_j - x) \tag{10.1}$$

and to solve sequentially

$$q_n \cdot (\rho_1 - \Delta t \cdot \tilde{L}) \cdot \ldots \cdot (\rho_n - \Delta t \cdot \tilde{L}) \cdot \tilde{u}(t - \Delta t) = \tilde{u}(t) \tag{10.2}$$

one linear system composed by $(\rho_j - \Delta t \cdot \tilde{L})$ after another, with $q_n = \frac{1}{n!}$ as obvious from (3.2) and (3.3). Note that for the fully implicit $(0, n)$ schemes this means that we need to solve complex linear systems for all roots that form complex conjugate pairs. This is of course the generalization of what was referred to as *complex time* in section 6.1.

An ingenious alternative to the sequence of complex linear systems is to use the *partial*

fraction decomposition

$$\frac{n!}{\prod_{j=1}^{n}(\rho_j - x)} \;=\; \sum_{j=1}^{n} \frac{\alpha_j}{\rho_j - x} \tag{10.3}$$

with

$$\alpha_j \;=\; \frac{n!}{\prod_{k\neq j}(\rho_k - \rho_j)} \;. \tag{10.4}$$

This was first demonstrated in [Gal89, GS89a, GS89b] specifically for the purpose of parallelization of the solution of the respective linear systems in (10.2). This strategy transforms the task of solving the one linear system

$$Q_n(\Delta t \cdot \tilde{L}) \cdot \tilde{u}(t - \Delta t) \;=\; \tilde{u}(t) \tag{10.5}$$

whose matrix form is $(2n+1)$-diagonal, into n individual linear systems

$$(\rho_j - \tilde{L}) \cdot v_j \;=\; \tilde{u}(t) \tag{10.6}$$

whose solutions comprise the finite differencing scheme by virtue of

$$\tilde{u}(t - \Delta t) \;=\; \sum_{k} \alpha_j v_j \;. \tag{10.7}$$

A further simplification is yet possible [GS89b]. For all roots that form complex conjugate pairs, i.e, when $\rho_{j+1} = \rho_j^*$, we observe that $\alpha_{j+1} = \alpha_j^*$, and hence

$$\frac{\alpha_j}{\rho_j - x} + \frac{\alpha_{j+1}}{\rho_{j+1} - x} \;=\; \frac{\alpha_j}{\rho_j - x} + \frac{\alpha_j^*}{\rho_j^* - x} \;=\; 2 \cdot \Re\left[\frac{\alpha_j}{\rho_j - x}\right] \tag{10.8}$$

since $x \in \mathbb{R}$. This in turn means that the $A_{(0,2)}$ scheme can in fact be implemented as the solution of a *single* complex-valued linear system. For example, since

$$2! \cdot \left(1 - x + x^2/2\right) \;=\; [(1+i) - x] \cdot [(1-i) - x] \;, \tag{10.9}$$

which obviously can be identified as the scheme discussed earlier in section 6.1, we have

$$\begin{aligned} \alpha_1 \;&=\; \frac{2!}{(1-i) - (1+i)} \\ &=\; i \end{aligned} \tag{10.10}$$

and thus

$$\begin{aligned} u_{(0,2)}(t - \Delta t) \;&=\; 2 \cdot \Re\left[i \cdot \left((1+i) - \Delta t \cdot \tilde{L}\right)^{-1} \cdot \tilde{u}(t)\right] \\ &=\; -2 \cdot \Im\left[\left((1+i) - \Delta t \cdot \tilde{L}\right)^{-1} \cdot \tilde{u}(t)\right] \end{aligned} \tag{10.11}$$

for the $(0,2)$ scheme. We see here that it is in fact not necessary to take two complex time steps since they can be collapsed into one.

For the $(0,3)$ scheme, we obtain

$$\begin{aligned} u_{(0,3)}(t-\Delta t) \; &= \alpha_{3,1} \cdot \left(\rho_{3,1} - \Delta t \cdot \tilde{L}\right)^{-1} \cdot \tilde{u}(t) \\ &\quad + 2 \cdot \Re\left[\alpha_{3,2} \cdot \left(\rho_{3,2} - \Delta t \cdot \tilde{L}\right)^{-1} \cdot \tilde{u}(t)\right] \end{aligned} \tag{10.12}$$

with

$$\begin{aligned} \alpha_{3,1} &= 1.47568651779572090 \\ \rho_{3,1} &= 1.59607163798332152 \\ \alpha_{3,2} &= -0.73784325889786049 + 0.36501784080102848 \cdot i \\ \rho_{3,2} &= 0.70196418100833929 + 1.80733949445202179 \cdot i \; . \end{aligned} \tag{10.13}$$

And finally, for the $(0,4)$ scheme, we end up with

$$\begin{aligned} u_{(0,4)}(t-\Delta t) \; = \; 2 \cdot \Re\Big[\;\; &\alpha_{4,1} \cdot \left(\rho_{4,1} - \Delta t \cdot \tilde{L}\right)^{-1} \cdot \tilde{u}(t) \\ &+ \alpha_{4,2} \cdot \left(\rho_{4,2} - \Delta t \cdot \tilde{L}\right)^{-1} \cdot \tilde{u}(t) \Big] \end{aligned} \tag{10.14}$$

with

$$\begin{aligned} \alpha_{4,1} &= 0.54141334842915765 + 1.58885918222327870 \cdot i \\ \rho_{4,1} &= 1.72944423106770540 + 0.88897437612186581 \cdot i \\ \alpha_{4,2} &= -0.54141334842915765 - 0.24856252086611912 \cdot i \\ \rho_{4,2} &= 0.27055576893229461 + 2.50477590436243425 \cdot i \; . \end{aligned} \tag{10.15}$$

10.1 Solving the complex linear system

The partial fraction decomposition approach requires the solution of one or more linear systems of the form

$$(\rho \cdot \mathbf{1}_n - G) \cdot x \; = \; b \tag{10.16}$$

with ρ complex, G being a real linear operator given by the discretized diffusion generator, b being a real vector, and n being the number of spatial discretization nodes. When the underlying dynamics represent a one-dimensional diffusion, the generator G is tri-diagonal in a matrix representation, and the system (10.16) is readily amenable to direct solution with standard tri-diagonal forward an backward substitution on the space of complex numbers.

10.2 Solving iteratively with real-valued sparse systems

In higher dimensions, or for non-diffusive dynamics, one may wish to use iterative algorithms for the solution of (10.16). In that case, one may use an iterative algorithm that is designed for complex systems. Since the generic iterative solution of complex linear systems can be significantly more difficult than that of benign real-valued linear systems, one may, instead, want to use one of the well tried and tested algorithms such as the

BiCGStab method [dV92], by the aid of the following reformulation of (10.16). Consider that

$$\rho = \rho_r + i \cdot \rho_i \tag{10.17}$$

$$x = x_r + i \cdot x_i \tag{10.18}$$

with $\rho_r, \rho_i \in \mathbb{R}$, $x_r, x_i \in \mathbb{R}^{\kappa}$. Then, the system (10.16) can be represented in block form as

$$\begin{pmatrix} (\rho \cdot \mathbf{1}_n - G) & -\rho_i \cdot \mathbf{1}_n \\ \rho_i \cdot \mathbf{1}_n & (\rho \cdot \mathbf{1}_n - G) \end{pmatrix} \cdot \begin{pmatrix} x_r \\ x_i \end{pmatrix} = \begin{pmatrix} b \\ 0_n \end{pmatrix} \tag{10.19}$$

with 0_n standing for a vector of n entries that are all zero. As for an initial guess in the iterative scheme, we may take guidance by what tends to work well in the conventional first order fully implicit scheme. For that scheme, we have $\rho = 1$ and $x^{(0)} = b$ is a tried and tested, even if possibly not optimal, initial guess. In other words, to lowest order, the generator G, is considered to have only a slight effect when going from $b = \tilde{u}(t)$ to $x = \tilde{u}(t - \Delta t)$, and in the setting of the initial guess $x^{(0)}$, the generator G is simply dropped from (10.16). In this context, this translates to

$$x^{(0)} = \frac{1}{\rho} \cdot b \tag{10.20}$$

which means

$$x_r^{(0)} = \frac{\rho_r}{|\rho_i|^2} \cdot b \tag{10.21}$$

$$x_i^{(0)} = -\frac{\rho_i}{|\rho_i|^2} \cdot b \,. \tag{10.22}$$

10.3 Sequentially solving real-valued pentadiagonal systems

In the one-dimensional case for $n > 1$, an alternative to iterative solutions of real-valued sparse systems is to factorize the denominator of $R_{(0,n)}$ only to quadratic terms. Then, the implicit solution can be obtained by sequential direct solution of pentadiagonal systems. Specifically, we have

$$R_{(0,1)}(x)^{-1} = (1 - x) \tag{10.23}$$

$$R_{(0,2)}(x)^{-1} = \left(1 - x + \tfrac{1}{2}x^2\right) \tag{10.24}$$

$$\begin{aligned} R_{(0,3)}(x)^{-1} &= \left(1 - x + \tfrac{1}{2}x^2 - \tfrac{1}{6}x^3\right) \\ &= (1 - 0.62653829327079973114 \cdot x) \\ &\quad \cdot (1 - 0.37346170672920026886 \cdot x + 0.26601193966388692052 \cdot x^2) \end{aligned} \tag{10.25}$$

$$\begin{aligned} R_{(0,4)}(x)^{-1} &= \left(1 - x + \tfrac{1}{2}x^2 - \tfrac{1}{6}x^3 + \tfrac{1}{24}x^4\right) \\ &= (1 - 0.91474668699459514598 \cdot x + 0.26446261479904986361 \cdot x^2) \\ &\quad \cdot (1 - 0.08525331300540485402 \cdot x + 0.15755219957394281479 \cdot x^2) \end{aligned} \tag{10.26}$$

Substituting $x \equiv \Delta t \cdot \tilde{L}$, we then obtain the implicit linear system from the spatially discrete operator $\tilde{L}$.

11 Conclusion

We have reviewed a range of time stepping schemes for finite differencing methods for the solution of partial integro differential equations. We focussed on the two time level family of Padé approximants as taken from [Smi86, chapter 3], but also included two alternative second order implicit schemes. For pure one-dimensional diffusions, we have explained the mechanisms that give rise to stability, or instability, and negative responses when spatial high frequency modes are present in the initial conditions, as is so often the case in financial derivatives mathematics applications. We have identified that the $R_{(0,n)}(\Delta t \cdot \tilde{L})$ Padé schemes with $n = 1, 2, 3, 4$ given by

$$R_{(0,1)}(x) = (1 - x)^{-1} \tag{11.1}$$

$$R_{(0,2)}(x) = \left(1 - x + \tfrac{1}{2}x^2\right)^{-1} \tag{11.2}$$

$$R_{(0,3)}(x) = \left(1 - x + \tfrac{1}{2}x^2 - \tfrac{1}{6}x^3\right)^{-1} \tag{11.3}$$

$$R_{(0,4)}(x) = \left(1 - x + \tfrac{1}{2}x^2 - \tfrac{1}{6}x^3 + \tfrac{1}{24}x^4\right)^{-1} \tag{11.4}$$

are all L-stable, i.e., *unconditionally stable* and *non-oscillatory*, as well as monotonic in their damping factor as a function of step size. We have given some numerical evidence that the numerical magnitude of the violation of strict positivity preservation for $n > 1$ seems to decrease as n increases. Only the $(0, 1)$ scheme, however, is strictly positivity-preserving within numerical round-off.

As for a general summary, alas, we can only mention that it is useful to recall that a great variety of finite differencing schemes exist, even in just a one-dimensional setting, some more and some less known, with different levels of convergence order, complexity of implementation, implications for positivity, and so on. Far be it from us, however, to attempt giving any wholesale advice as to when to use which method within this short note. Suffice it to say that the choice of finite differencing techniques is definitely the domain of *horses for courses*.

Acknowledgement

The purpose of this note is not the presentation of any new results but primarily an elaboration of Padé approximant schemes which are documented in the excellent reference [Smi86, chapter 3]. The idea of complex time steps is due to J. Andreasen to whom the author is indebted for the useful discussions held at the September 2011 Danske Kwant Festival in Copenhagen. The Lawson-Morris scheme was also mentioned to this author by J. Andreasen without reference and subsequent research revealed that it was first published in 1977 as a technical report by Lawson and Morris [LM78]. The (uvw)-scheme is likely to be already in the literature under a different name unbeknownst to this author who derived it by simple coefficient matching, though, it is clearly rendered redundant by the superior properties of the $(0, 2)$ Padé scheme and its higher order cousins. Whilst most of this note is a mere presentation of other researchers' results, all errors remain of course the responsibility of the author.

The author is grateful to Charles-Henri Roubinet, Head of Quantitative Research at VTB Capital, for authorizing the release of this note (originally from September 2011) into the public domain.

Finally, the author thanks David Shelton for his feedback and pointing out typographical errors.

References

[dV92] H. Van der Vorst. A fast and smoothly converging variant of BiCG for the solution of nonsysmmetric linear systems. *SIAM J. Sci. Statist. Comput.*, 13:631–644, 1992.

[Gal89] E. Gallopoulos. A partial fraction decomposition approach to improved efficiency of some parabolic solvers. Technical report, Technical Report 874, Center for Supercomputing Reseach and Development, 1989.

[GS89a] E. Gallopoulos and Y. Saad. A parallel block cyclic reduction algorithm for the fast solution of elliptic equations. *Parallel Computing*, 10(2):143–159, 1989. scgroup.hpclab.ceid.upatras.gr/faculty/stratis/Papers/GallopoulosSaadBCR.pdf.

[GS89b] E. Gallopoulos and Y. Saad. On the Parallel Solution of Parabolic Equations. Technical report, CSRD Report 854, Univ. of Illinois, Urbana-Champaigne, 1989. ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19900014693_1990014693.pdf.

[LM78] J.D. Lawson and J.Ll. Morris. The Extrapolation of First Order Methods for Parabolic Partial Differential Equations I. *SIAM Journal of Numerical Analysis*, 15:1212–1224, 1978. www.cs.uwaterloo.ca/research/tr/1977/CS-77-20.pdf.

[Pad92] H. Padé. *Sur la répresentation approchée d'une fonction par des fractions rationelles*. PhD thesis, Ann. École Nor. (3), 9, 1892.

[Ray94] J.W.S. Rayleigh. *The Theory of Sound*. Various (re-)publishers, 1894.

[Sid10] M. Siddique. Fourth Order Positively Smoothed Pad Schemes for Parabolic Partial Differential Equations with Nonlocal Boundary Conditions. *Journal of Computational and Applied Mathematics*, 4(42):2065–2080, 2010. www.m-hikari.com/ams/ams-2010/ams-41-44-2010/siddiqueAMS41-44-2010.pdf.

[Smi86] G. Smith. *Numerical Solution of Partial Differential Equations: Finite Difference Methods*. Oxford University Press, January 1986.

[TZA+06] R.K. Thulasiram, C. Zhen, Chhabra A, P. Thulasiraman, and A.B. Gumel. A second order L0 stable algorithm for evaluating European options. *International Journal of High Performance Computing and Networking*, 4(5/6):311–320, 2006. ftp://ftp.cs.umanitoba.ca/pub/IJHPCN/FINAL/029-Fin-Tulsi/Journal.pdf.

[Var61] R.S. Varga. On higher order stable implicit methods for solving parabolic partial differential equations. *Journal of Mathematics and Physics*, 40(3):220–231, 1961. www.math.kent.edu/~varga/pub/paper_19.pdf.

[WKY+07] B.A. Wade, A.Q.M. Khaliq, M. Yousuf, J. Vigo-Aguiard, and R. Deiningere. On smoothing of the CrankNicolson scheme and higher order schemes for pricing barrier options. *Journal of Computational and Applied Mathematics*, 204(1):144–158, 2007. https://pantherfile.uwm.edu/wade/www/index_files/wade_barriers_jcam.pdf.