A NOTE ON STABILITY
OF THE DOUGLAS SPLITTING METHOD

WILLEM HUNDSDORFER

ABSTRACT. In this note some stability results are derived for the Douglas splitting method. The relevance of the theoretical results is tested for an advection-reaction equation.

1. PRESENTATION OF THE RESULTS

Consider the initial value problem for a system of ODEs

$$(1.1) \qquad u'(t) = F(t, u(t))$$

with $0 \le t \le T$ and given initial value $u(0)$. We shall consider numerical schemes with step size τ yielding approximations u_n to the exact solution $u(t_n)$ at time levels $t_n = n\tau$ for $n = 0, 1, 2, \cdots$, starting with $u_0 = u(0)$.

For problems that arise by spatial discretization of multi-dimensional PDEs it is often possible to decompose the function F into a number of simpler component functions,

$$(1.2) \qquad F(t, w) = F_1(t, w) + F_2(t, w) + \cdots + F_s(t, w).$$

Splitting methods use this decomposition by treating in each stage at most one of the components implicitly. The best known method of this type is the ADI-Peaceman-Rachford method, but this method can only deal with 2-component splittings, see [5]. In this paper we shall consider the related second-order method of Douglas [1], also known as the method of Stabilizing Corrections [4],

$$(1.3)$$
$$v_0 = u_n + \tau F(t_n, u_n),$$
$$v_i = v_{i-1} + \tfrac{1}{2}\tau\Big(F_i(t_{n+1}, v_i) - F_i(t_n, u_n)\Big) \qquad (i = 1, 2, \cdots, s),$$
$$u_{n+1} = v_s,$$

with internal vectors v_i.

A big advantage of (1.3) over many other splitting methods [4, 5] is that all internal vectors v_i are consistent approximations to the exact solution, namely at time t_{n+1}. This implies that if we are in a steady state $F(u) = 0$, with F independent of t, then this steady state is also a stationary point of the scheme (1.3).

We shall present some stability results for the scalar complex test equation where

(1.4) $$F_j(t,w) = \lambda_j w$$

with $\lambda_j \in \mathbb{C}$. In applications for PDEs the λ_j will represent eigenvalues for the various components, found by inserting Fourier modes. Let $z_j = \tau\lambda_j$. For the test equation the method reduces to

(1.5) $$u_{n+1} = R\,u_n$$

with growth factor

(1.6) $$R = 1 + \Big(\prod_{j=1}^{s}(1 - \tfrac{1}{2}z_j)\Big)^{-1}\sum_{j=1}^{s} z_j.$$

This R corresponds to the stability function for standard one-step methods. Ideally, one would have $|R| \le 1$ for arbitrary λ_j in the left half-plane $\mathbb{C}^-$ without restriction on the time step. As we shall see, for R given by (1.6), this is not true if $s \ge 3$.

It is easy to verify that $|R| \le 1$ when all z_j are real and negative (unconditional stability for purely parabolic equations, see Douglas [1]). On the other hand, it can also be shown that if $s \ge 3$ and all $z_j = iy$, then $|R| > 1$ for any $y \ne 0$ (unconditional instability for purely hyperbolic equations, see Warming and Beam [6] and also Remark 2.1). In this paper we shall present some intermediate results which are applicable to advection-diffusion and advection-reaction equations. It will be assumed that the z_j belong to the wedge $W_\alpha = \{\zeta \in \mathbb{C} : |\arg(-\zeta)| \le \alpha\}$ in the left half-plane. We consider the statement
(A) $\cdots$ $\qquad |R| \le 1$ for all $z_j \in W_\alpha$.

Theorem 1. *Let R be given by (1.6) with $s \ge 2$. We have*

$$\text{(A)} \quad\Longleftrightarrow\quad \alpha \le \frac{1}{s-1}\,\frac{\pi}{2}\,.$$

For $s = 2$ we thus get stability for $\alpha \le \pi/2$, which allows the z_j to range over the whole left half-plane. However, for $s = 3$ we get the condition $\alpha \le \pi/4$, which is already quite restrictive. One may expect the situation to become better if some z_j are real and negative. In the following theorem we assume that there are r such $z_j < 0$. Consider the statement
(B) $\cdots$ $\qquad |R| \le 1$ for arbitrary $z_1, ..., z_{s-r} \in W_\beta$, $z_{s-r+1}, ..., z_s < 0$.

Theorem 2. *Let R be given by (1.6), and let $1 \le r \le s - 1$. We have*

$$\text{(B)} \quad\Longleftrightarrow\quad \beta \le \frac{1}{s-r}\,\frac{\pi}{2}\,.$$

It is somewhat surprising that for $r = 1$ we get the same condition as in Theorem 1. So again, already for $s = 3$ we may get a quite restrictive condition, unless there are two z_j that are real and negative. If $s = 3$ with arbitrary $z_1, z_2 \in \mathbb{C}^-$, then we have stability if $z_3 = 0$, but letting $z_3 < 0$ may destroy this stability.

The proof of these results will be given in the next section. In Section 3 some numerical results will be presented for an advection-reaction equation.

Remark. For linear problems, where $F_j(t,w) = A_j w^i + g_j(t)$, the stability results can be applied provided the matrices A_j are normal and commuting. Some results for non-commuting matrices were given by Douglas and Gunn [2] under very strict conditions on the step size.

2. Proofs

2.1. Proof of Theorem 1.

In the following, all summations will be from 1 to s, unless indicated otherwise. Let

$$\xi = R - 1 \quad \text{and} \quad \eta = -|\sum_j z_j|^2 \left(\frac{1}{\xi} + \frac{1}{2}\right).$$

Clearly $|R| \leq 1$ is equivalent with the following

$$|1 + \xi| \leq 1 \quad \Longleftrightarrow \quad \text{Re}\,\frac{1}{\xi} \leq -\frac{1}{2} \quad \Longleftrightarrow \quad \text{Re}\,\eta \geq 0.$$

The last criterion will be used in this proof. By some calculations it is seen that

(2.1)
$$\eta = -\left(\sum_j \bar{z}_j\right)\left(1 + (-\tfrac{1}{2})^2 \sum_{j<k} z_j z_k + (-\tfrac{1}{2})^3 \sum_{j<k<l} z_j z_k z_l + \cdots + (-\tfrac{1}{2})^s z_1 z_2 ... z_s\right).$$

To verify the statement of the theorem it is, according to the maximum modulus theorem, sufficient to consider z_j on the boundary of W_α. In the following, let $t_j \geq 0$ be arbitrary, $0 \leq q \leq \frac{1}{2}s$ and

$$z_j = -e^{i\alpha}t_j \quad (1 \leq j \leq s - q), \qquad z_j = -e^{-i\alpha}t_j \quad (s - q < j \leq s).$$

First, consider $q = 0$. Then we obtain

$$\eta = \left(e^{-i\alpha}\sum_j t_j\right)\left(1 + (\tfrac{1}{2})^2 e^{2i\alpha}\sum_{j<k} t_j t_k + \cdots + (\tfrac{1}{2})^s e^{si\alpha}t_1 t_2 ... t_s\right),$$

$$\text{Re}\,\eta = \left(\sum_j t_j\right)\left(\cos(\alpha) + (\tfrac{1}{2})^2 \cos(\alpha)\sum_{j<k} t_j t_k + \cdots + (\tfrac{1}{2})^s \cos((s-1)\alpha)t_1 t_2 ... t_s\right).$$

It follows that $\text{Re}\,\eta \geq 0$ for all $t_j \geq 0$ if and only if $\cos(k\alpha) \geq 0$ $(k = 1, 2, ..., s-1)$, that is,

(2.2)
$$(s - 1)\alpha \leq \frac{\pi}{2}.$$

Next, suppose that $q = 1$. Then (2.1) gives

$$\eta = \left(\sum_{j<s} e^{-i\alpha}t_j + e^{i\alpha}t_s\right)\left(\sum_{k=0}^{s-1} p_k e^{ik\alpha}\right)$$

with $p_k \geq 0$ depending on the t_j. The actual expressions easily follow from (2.1). In particular we find

$$p_{s-1} = (\tfrac{1}{2})^{s-1} t_1 t_2 ... t_{s-1},$$

$$p_{s-3} \geq (\tfrac{1}{2})^{s-1} t_1 t_2 ... t_{s-1}\left(\frac{1}{t_1} + \frac{1}{t_2} + \cdots + \frac{1}{t_{s-1}}\right)t_s.$$

Assuming (2.2), it follows that

$$\text{Re}\,\eta \geq (\tfrac{1}{2})^{s-1} \cos(s\alpha)t_1 t_2 ... t_s + (\tfrac{1}{2})^{s-1}(s - 1)\cos((s - 4)\alpha)t_1 t_2 ... t_s.$$

Further, (2.2) implies $\cos(s\alpha) + (s - 1)\cos((s - 4)\alpha) \geq 0$. Hence, also for this case $q = 1$, we see that (2.2) implies $|R| \leq 1$.

Finally, suppose that $q \geq 2$. Then

$$\eta = \Big(\sum_{j \leq s-q} e^{-i\alpha} t_j + \sum_{j > s-q} e^{i\alpha} t_j \Big) \Big(\sum_{k=-q}^{s-q} q_k e^{ik\alpha} \Big)$$

with $q_k \geq 0$. Therefore, $\mathrm{Re}\,\eta$ is a sum of $\cos(k\alpha)$ terms with $-(q+1) \leq k \leq s-q+1$ and nonnegative coefficients, and again it follows that (2.2) is sufficient to have $|R| \leq 1$.

Remark 2.1. If we have $s = 3$ and $z_j = iy$ for $j = 1, 2, 3$, then $\mathrm{Re}\,\eta = -\frac{3}{8} y^4 < 0$ for any $y \neq 0$. Hence, for any $C > 0$, we have $\max\{|R| : z_j = i\,y_j,\ |y_j| \leq C\} > 1$. This instability result was already obtained by Warming and Beam [6] for a class of multistep splitting methods, containing the Douglas method as a special case.

2.2. Proof of Theorem 2. First we consider the case where one of the z_k is real and negative, say $z_s < 0$. The other z_j are assumed to lie in the wedge W_β. It will be shown that

$$(2.3) \qquad (s-1)\beta \leq \frac{\pi}{2}$$

is necessary to guarantee that $|R| \leq 1$ if $z_s \to -\infty$. By Theorem 1 we already know that this is a sufficient condition for arbitrary $z_s < 0$.

In the limit $z_s \to -\infty$ we have $R \to S$ with

$$S = 1 - 2 \Big(\prod_{j<s} (1 - \tfrac{1}{2} z_j) \Big)^{-1}.$$

It is easily seen that $|S| \leq 1$ is equivalent with

$$\mathrm{Re} \prod_{j<s} (1 - \tfrac{1}{2} z_j) \geq 1.$$

Take $z_j = -e^{i\beta} t_j$ $(1 \leq j \leq s-1)$ with $t_j > 0$. Then

$$\mathrm{Re} \prod_{j<s} (1 - \tfrac{1}{2} z_j)$$

$$= \mathrm{Re}\Big(1 + (-\tfrac{1}{2}) \sum_{j<s} z_j + (-\tfrac{1}{2})^2 \sum_{j<k<s} z_j z_k + \cdots + (-\tfrac{1}{2})^{s-1} z_1 z_2 \ldots z_{s-1} \Big)$$

$$= 1 + \tfrac{1}{2} \cos(\beta) \sum_{j<s} t_j$$

$$+ (\tfrac{1}{2})^2 \cos(2\beta) \sum_{j<k<s} t_j t_k + \cdots + (\tfrac{1}{2})^{s-1} \cos((s-1)\beta) t_1 t_2 \ldots t_{s-1}.$$

Thus we see that (2.3) is necessary if $t_1, \ldots, t_{s-1}$ are sufficiently large.

Next we consider the general situation $z_1, \ldots, z_{s-r} \in W_\beta$, $z_{s-r+1}, \ldots, z_s < 0$ with $1 \leq r \leq s-1$. We now have to show that

$$(2.4) \qquad (s-r)\beta \leq \frac{\pi}{2}$$

is necessary and sufficient for $|R| \leq 1$.

Note that R is fractional linear in all z_j with denominator $1 - \tfrac{1}{2} z_j$. Considering fixed $z_1, \ldots, z_{s-r}$, it follows that we have

$$|R| \leq 1 \quad \text{for all} \quad z_{s-r+1}, \ldots, z_s < 0$$

iff this holds for $z_{s-r+1}, ..., z_s$ equal to 0 or ∞. This amounts to verification of the two inequalities :

$$\left| 1 + \left(\prod_{j \leq s-r} (1 - \tfrac{1}{2}z_j) \right)^{-1} \sum_{j \leq s-r} z_j \right| \leq 1, \qquad \left| 1 - 2 \left(\prod_{j \leq s-r} (1 - \tfrac{1}{2}z_j) \right)^{-1} \right| \leq 1.$$

From the above results, with $s - r$ replacing s and $s - 1$, respectively, it follows that this will be satisfied for arbitrary $z_1, ..., z_{s-r} \in W_\beta$ if and only if (2.4) holds.

3. EXAMPLE

In this section we shall present some numerical tests for an advection-reaction equation. For comparison we also consider the following method,

$$\begin{aligned}
(3.1) \quad & v_0 = u_n, \\
& v_i = v_{i-1} + \tfrac{1}{2}\tau F_i(t_n, v_{i-1}) \qquad (i = 1, 2, \cdots, s), \\
& v_{s+i} = v_{s+i-1} + \tfrac{1}{2}\tau F_{s+1-i}(t_{n+1}, v_{s+i}) \qquad (i = 1, 2, \cdots, s), \\
& u_{n+1} = v_{2s}.
\end{aligned}$$

This method has been tested in [3], where it was called the *trapezoidal splitting* method. It is also a second-order method but the internal vectors v_j are not consistent approximations to the exact solution. It is more stable than the Douglas scheme, however. If we apply (3.1) to the scalar test equation (1.4) we get the growth factor

$$(3.2) \qquad R = \prod_{j=1}^{s} (1 - \tfrac{1}{2}z_j)^{-1}(1 + \tfrac{1}{2}z_j),$$

and thus with this method we have $|R| \leq 1$ for arbitrary z_j in the left half-plane $\mathbb{C}^-$, irrespective of s.

To verify the relevance of the results for the scalar test equation (1.4), we consider the following advection equation with a linear reaction term,

$$\mathbf{u}_t = a\mathbf{u}_x + b\mathbf{u}_y + G\mathbf{u} \qquad \text{on } \Omega = [0,1]^2.$$

The velocities are given by $a(x,y,t) = 2\pi(y - \tfrac{1}{2})$, $b(x,y,t) = 2\pi(\tfrac{1}{2} - x)$. Further,

$$\mathbf{u}(x,y,t) = \left(\begin{array}{c} \mathbf{u}_1(x,y,t) \\ \mathbf{u}_2(x,y,t) \end{array} \right), \qquad G = \left(\begin{array}{cc} -k_1 & k_2 \\ k_1 & -k_2 \end{array} \right).$$

We take $k_1 = 1$. The second reaction constant k_2 will be used to vary the stiffness of this reaction term. Note that the matrix G has eigenvalues 0 and $-(k_1 + k_2)$. We have a chemical equilibrium if $\mathbf{u}_1/\mathbf{u}_2 = k_2/k_1$.

The initial condition is chosen as

$$\mathbf{u}_1(x,y,0) = c, \quad \mathbf{u}_2(x,y,0) = (1 - c) + \mu \exp(-80(x - \tfrac{1}{2})^2 - 80(y - \tfrac{3}{4})^2)$$

with $c = k_2/(k_1 + k_2)$. To avoid a strong transient phase, we take $\mu = 100/k_2$. So, if k_2 increases we start closer to the chemical equilibrium to maintain some smoothness.

The exact solution is given by

$$\begin{aligned}
\mathbf{u}_1(x,y,t) &= c\left(e^{-(k_1+k_2)t} + (1 - e^{-(k_1+k_2)t})d(x,y,t) \right), \\
\mathbf{u}_2(x,y,t) &= d(x,y,t) - \mathbf{u}_1(x,y,t),
\end{aligned}$$

with

$$d(x, y, t) = 1 + \mu \, \exp(-80\xi^2 - 80(\eta - \tfrac{1}{4})^2),$$

$$\xi = \cos(2\pi t)(x - \tfrac{1}{2}) - \sin(2\pi t)(y - \tfrac{1}{2}), \quad \eta = \sin(2\pi t)(x - \tfrac{1}{2}) + \cos(2\pi t)(y - \tfrac{1}{2}).$$

After a mild transient phase this is purely an advection problem, and the velocity field gives a rotation around the center of the domain. At $t = 1$ one rotation is completed.

Dirichlet conditions are prescribed at the inflow boundaries. At the outflow boundaries we shall use an upwind discretization in space, in the interior second-order central differences are used. We consider splitting with F_1, F_2 the finite difference operators for advection in the x and y direction, respectively, and with F_3 for the linear reaction term. The test has been performed on a fixed 80×80 grid, and with $\tau = 1/80$ and $1/160$. The spatial difference operators will have eigenvalues close to the imaginary axis.

Although we are not in a model situation with commuting, normal operators, application of a standard von Neumann analysis (ignoring boundary conditions and freezing the coefficients) yields local growth factors R with z_1, z_2 on the imaginary axis between $-iC$ and iC with local Courant numbers C, which are maximally $80\pi\tau$, and with $z_3 = 0$ or $-\tau(k_1 + k_2)$.

On the basis of Theorem 2 we expect the Douglas scheme to be stable only if k_2 is not large. The following table shows that this scheme becomes indeed unstable for large k_2, whereas the trapezoidal splitting remains stable. Note however that the transition from stable to unstable is very hesitant. Several rotations are sometimes needed to give a significant instability.

TABLE 3.1. Maximum errors for the Douglas method (1 rotation and 4 rotations) and trapezoidal splitting (4 rotations). The entry *** denotes overflow.

		Douglas, 1 rot.	Douglas, 4 rot.	TrapSplit, 4 rot.
$k_2 = 500$	$\tau = 1/80$	$4.5 \ 10^{-2}$	$1.0 \ 10^{-1}$	$1.0 \ 10^{-1}$
	$\tau = 1/160$	$2.9 \ 10^{-2}$	$8.3 \ 10^{-2}$	$8.3 \ 10^{-2}$
$k_2 = 1000$	$\tau = 1/80$	$2.2 \ 10^{-2}$	$8.1 \ 10^{+8}$	$5.0 \ 10^{-2}$
	$\tau = 1/160$	$1.4 \ 10^{-2}$	$4.1 \ 10^{-2}$	$4.1 \ 10^{-2}$
$k_2 = 2000$	$\tau = 1/80$	8.4	$7.6 \ 10^{+22}$	$2.5 \ 10^{-2}$
	$\tau = 1/160$	$7.2 \ 10^{-3}$	7.5	$2.1 \ 10^{-2}$
$k_2 = 4000$	$\tau = 1/80$	$1.9 \ 10^{+3}$	$5.6 \ 10^{+30}$	$1.2 \ 10^{-2}$
	$\tau = 1/160$	$1.4 \ 10^{+3}$	***	$1.0 \ 10^{-2}$

The reason for the fact that the transition from stability to instability is not clear-cut with the Douglas method lies in the fact that the growth factors do not become large. In the following figure the modulus $|R|$ is plotted for z_1, z_2 fixed on the imaginary axis and $z_3 < 0$ varying. Although we have $|R| > 1$ if $z_3 << 0$,
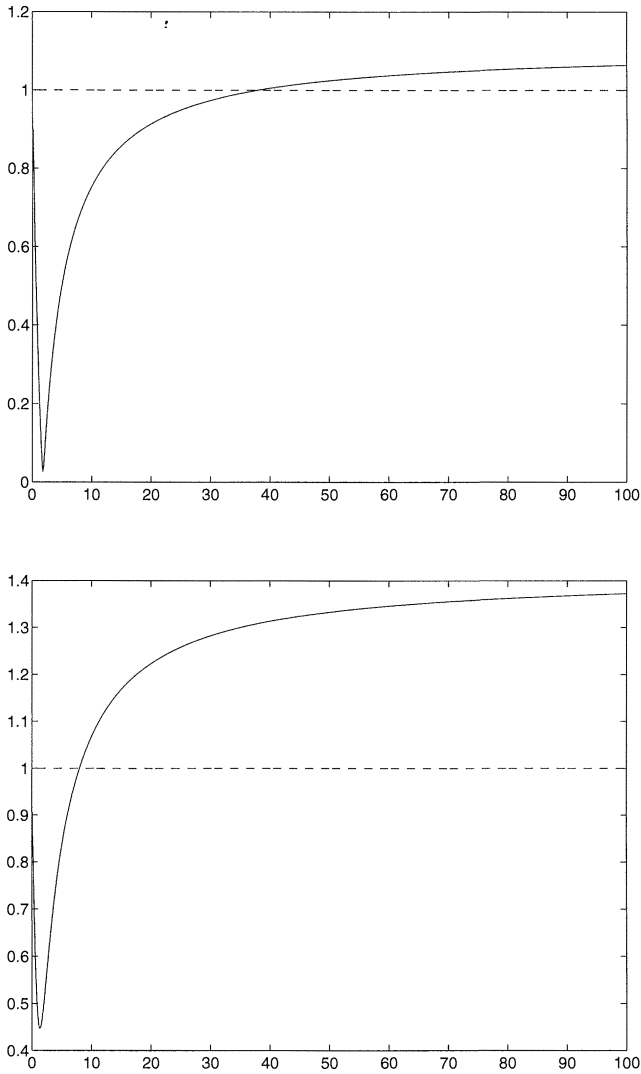
FIGURE 1. Modulus $|R|$ versus $x \in [0, 100]$. Top picture with $z_1 = z_2 = \frac{1}{2}i, z_3 = -x$. Bottom picture with $z_1 = z_2 = 2i, z_3 = -x$.

the value does not become large (it can be shown that for z_1, z_2 on the imaginary axis and $z_3 < 0$, the case $z_1 = z_2 = 2i, z_3 \to \infty$ gives the maximal growth factor, namely $\sqrt{2}$). Therefore, it takes some time for the instability to become visible.

It is also clear from the above table and figures that a mild stiffness in the reaction term is allowed. This is quantified in the following theorem. We consider the statement

(C) $\cdots$ $\quad |R| \leq 1$ for all $z_j = iy_j \, (j = 1, 2), z_3 = -x$ with $|y_j| \leq \gamma \, (j = 1, 2), 0 \leq x \leq \delta$.

Theorem 3. *Let R be given by (1.6) with $s = 3$. We have*

$$(C) \quad \Longleftrightarrow \quad \delta \leq 6 + 8\gamma^{-2}.$$

Proof. Consider η defined by (2.1). Here we have

$$\mathrm{Re}\,\eta = (1 - \tfrac{1}{4}y_1 y_2 - \tfrac{1}{8}y_1 y_2 x)x + \tfrac{1}{4}(y_1 + y_2)^2 x.$$

So, for $x > 0$ we have $\mathrm{Re}\,\eta \geq 0$ iff

$$x \leq \frac{8 + 2(y_1^2 + y_1 y_2 + y_2^2)}{y_1 y_2}$$

in case $y_1 y_2 > 0$, whereas there is no restriction for $y_1 y_2 \leq 0$. By some straightforward analysis the result follows. $\square$

In conclusion it can be said that the Douglas method seems only suited for multi-dimensional PDEs if either

- advection dominates only in one direction, or
- advection dominates in two directions but the other components are nonstiff.

On the other hand, in situations where the method is stable, it is in general more accurate than a method like (3.1), due to the fact that the internal stages are all consistent approximations to the exact solution.

It is an open question whether multi-component splitting methods exist for $s \geq 3$ which are internally consistent and stable for all $z_j \in \mathbb{C}^-$.

REFERENCES

1. J. Douglas, *Alternating direction method for three space variables*. Numer. Math. 4, pp. 41-63 (1962). MR **24**:B2122
2. J. Douglas, J.E. Gunn, *A general formulation of alternating direction methods*. Numer. Math. 6, pp. 428-453 (1964). MR **31**:894
3. W. Hundsdorfer, *Trapezoidal and midpoint splittings for initial-boundary value problems*. CWI Report, 1996.
4. G.I. Marchuk, *Splitting and alternating direction methods*. Handbook of Numerical Analysis 1 (P.G. Ciarlet. J.L. Lions, eds.), North-Holland, Amsterdam, pp. 197-462, 1990. CMP 90:08
5. A.R. Mitchell, D.F. Griffiths, *The Finite Difference Method in Partial Differential Equations*. John Wiley & Sons, Chichester, 1980. MR **82a**:65002
6. R.F. Warming, R.M. Beam, *An extension of A-stability to alternating direction methods*. BIT 19, pp. 395-417 (1979). MR **80h**:65072

CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands