An Improved Version of the Reduction
to Scalar CDS Method for the Numerical Solution
of Separably Stiff Initial Value Problems

By Peter Alfeld

Abstract. In [1] the Reduction to Scalar CDS method for the solution of separably stiff initial value problems is proposed. In this paper an improved version is given that is equivalent for linear problems but considerably superior for nonlinear problems. A naturally arising numerical example is given, for which the old version fails, yet the new version yields very good results. The disadvantage of the new version is that in the case of several dominant eigenvalues $s > 1$, say, a system of s nonlinear equations has to be solved, whereas the old version gives rise to s uncoupled nonlinear equations.

1. **Introduction.** In order to avoid repetition the reader is assumed to be familiar with reference [1], particularly with Sections 1, 2, 3-IV, and 5-IV. There the Reduction to Scalar (RS) method has been derived by constructing a set of s scalar initial value problems whose solution evaluated at x_{n+1} provides an approximation to the components of $y(x_{n+1})$ corresponding to dominant eigenvalues. The scalar initial value problems are then solved using the trapezoidal rule.

In this paper the reverse procedure is used. The trapezoidal rule is applied to the full initial value problem, then the numerical problem is scalarized by forming the scalar product with the left eigenvectors corresponding to the dominant eigenvalues. In Section 2 the new RS version is introduced, and it is shown that both versions are equivalent for linear initial value problems. In Section 3, the application of both versions to nonlinear problems is described and compared. In Section 4 a naturally arising separably stiff problem is given for which the old version fails, and the new one produces very good results.

2. **The New RS Version Applied to Linear Problems.** In this section we consider the linear initial value problem

$$(2.1) \qquad y' = f(x, y) = A(x)y + g(x), \qquad y(a) = \eta, \qquad y, g \in \mathbf{R}^m,$$

where $A(x)$ is an $m \times m$ separably stiff matrix. (In the next section we will consider the case that $f(x, y)$ is nonlinear.) To (2.1) we apply the CDS scheme

(2.2(i)) $$\tilde{y}_{n+1} = By_n,$$

(2.2(ii)) $$y_{n+1} = \tilde{y}_{n+1} + \sum_{i=1}^{s} \xi_{n+1}^{(i)} c_{n+1}^{(i)};$$

see [1].

In [1] the following choice of correction factors is suggested:

$$\xi_{n+1}^{(i)} = -\langle d_{n+1}^{(i)}, \tilde{y}_{n+1} \rangle + \kappa_{n+1}^{(i)}, \qquad i = 1, 2, \ldots, s,$$

where

$$\kappa_{n+1}^{(i)} = \langle d_{n+1}^{(i)}, y_n \rangle$$

(2.3) $$+ \frac{h}{2} [\langle d_{n+1}^{(i)}, f_n \rangle + \langle d_{n+1}^{(i)}, f(x_{n+1}, y_n + (\kappa_{n+1}^{(i)} - \langle d_{n+1}^{(i)}, y_n \rangle) c_{n+1}^{(i)}) \rangle],$$

$$i = 1, 2, \ldots, s.$$

This system of equations is obtained by setting up s scalar initial value problems whose exact solutions are approximations to the components of the exact solution corresponding to the dominant eigenvalues, and then solving these scalar problems using the trapezoidal rule.

Since (2.1) is stiff it would be advantageous if we could apply the trapezoidal rule

(2.4) $$y_{n+1} - y_n = \frac{h}{2} (f(x_{n+1}, y_{n+1}) - f_n)$$

to the full problem (2.1), but this would lead to a system of m nonlinear equations if (2.1) were nonlinear. However, forming the inner product with the ith left dominant eigenvector, $d_{n+k}^{(i)}$, in (2.4), and substituting (2.2(ii)) into (2.4) yields

$$\left\langle d_{n+1}^{(i)}, \tilde{y}_{n+1} + \sum_{j=1}^{s} \xi_{n+1}^{(i)} c_{n+1}^{(i)} - y_n \right\rangle$$

(2.5) $$= \frac{h}{2} \left\langle d_{n+1}^{(i)}, f_n + f\left(x_{n+1}, \tilde{y}_{n+1} + \sum_{j=1}^{s} \xi_{n+1}^{(j)} c_{n+1}^{(j)}\right) \right\rangle,$$

$$i = 1, 2, \ldots, s.$$

This is a system of s equations for the s correction factors $\xi_{n+1}^{(i)}$, $i = 1, 2, \ldots, s$. Making use of the linearity of (2.1), and the biorthonormality of the eigensystem (see [1]), and proceeding similarly as in [1], (2.5) can be solved exactly, giving

(2.6) $$\xi_{n+1}^{(i)} = -\langle d_{n+1}^{(i)}, \tilde{y}_{n+1} \rangle + \frac{\left\langle d_{n+1}^{(i)}, y_n + \frac{h}{2}(f_n + g(x_{n+1})) \right\rangle}{1 - \frac{h}{2} \lambda_{n+1}^{(i)}},$$

which is the same correction factor as obtained in [1] as the solution of the system (2.3).

Thus, for the linear problem (2.1), the two versions of the reduction to scalar method are identical, and we have as an immediate consequence

THEOREM. *The CDS scheme defined by* (2.5) *is dominantly stable.*

3. The Two RS Versions Applied to Nonlinear Problems. We now consider the fully nonlinear separably stiff initial value problem

$$(3.1) \qquad y' = f(x, y), \qquad y(a) = \eta, \qquad y, f \in \mathbf{R}^m.$$

Both the systems (2.3) and (2.5) are now nonlinear, but (2.3) is a set of s uncoupled scalar equations, whereas the unknowns $\xi_{n+1}^{(i)}$ in (2.5) are not independent of each other.

The reason why (2.3) is uncoupled is that the scalar initial value problems giving rise to (2.3) are defined in terms of functions whose components corresponding to eigenvalues other than the ith are independent of y_{n+1}. This is an advantage computationally, but may lead to problems if (3.1) is so badly nonlinear, that the components corresponding to different eigenvalues cannot be considered independent even locally,

The system (2.5), on the other hand, represents the trapezoidal rule, as it is applied to the full problem (3.1), and then reduced to the dominant components. Thus, we can expect that the effect of the interdependence of the components corresponding to different eigenvalues is represented in the numerical solution of (3.1) by a CDS scheme based on (2.5). To solve (2.5) numerically we first consider the case $s = 1$, and then generalize to the case $s > 1$.

If $s = 1$, (2.5) can be solved by Newton's method, giving

$$[\xi_{n+1}^{(1)}]^{[t+1]} = [\xi_{n+1}^{(1)}]^{[t]}$$

$$(3.2) \qquad - \frac{\left\langle \tilde{d}_{n+1}^{(1)}, \tilde{y}_{n+1} + [\xi_{n+1}^{(1)}]^{[t]} \tilde{c}_{n+1}^{(1)} - y_n - \frac{h}{2}(f_n + f(x_{n+1}, \tilde{y}_{n+1} + [\xi_{n+1}^{(1)}]^{[t]} \tilde{c}_{n+1}^{(1)})) \right\rangle}{1 - \frac{h}{2} \left\langle \tilde{d}_{n+1}^{(1)}, \frac{\partial f}{\partial y}(x_{n+1}, \tilde{y}_{n+1} + [\xi_{n+1}^{(1)}]^{[t]} \tilde{c}_{n+1}^{(1)}) \tilde{c}_{n+1}^{(1)} \right\rangle},$$

$$t = 0, 1, 2, \ldots .$$

Approximating the denominator by $1 - h\lambda_{n+1}^{(1)}/2$ and defining

$$y_{n+1}^{[t]} = \tilde{y}_{n+1} + [\xi_{n+1}^{(1)}]^{[t]} \tilde{c}_{n+1}^{(1)},$$

(3.2) becomes

$$(3.3) \qquad [\xi_{n+1}^{(1)}]^{[t+1]} = [\xi_{n+1}^{(1)}]^{[t]} - \frac{\left\langle \tilde{d}_{n+1}^{(1)}, y_{n+1}^{[t]} - y_n - \frac{h}{2}(f_n + f(x_{n+1}, y_{n+1}^{[t]})) \right\rangle}{1 - h\tilde{\lambda}_{n+1}^{(1)}/2},$$

$$t = 0, 1, 2, \ldots .$$

If $s > 1$, s iterations of the type (3.3) can be linked. This, and defining initial approximations for the correction factors, gives

$$(3.4(i)) \qquad [\xi_{n+1}^{(i)}]^{[0]} = \xi_n^{(i)} \qquad (= 0 \text{ if } \xi_n^{(i)} \text{ is not defined}),$$

$$[\xi_{n+1}^{(i)}]^{[t+1]} = [\xi_{n+1}^{(i)}]^{[t]} - \frac{\left\langle \widetilde{d}_{n+1}^{(1)}, y_{n+1}^{[t]} - y_n - \frac{h}{2}(f_n + f(x_{n+1}, y_{n+1}^{[t]})) \right\rangle}{1 - h\widetilde{\lambda}_{n+1}^{(i)}/2},$$

(3.4(ii))

$$i = 1, 2, \ldots, s, \ t = 0, 1, 2, \ldots,$$

where

(3.4(iii)) $$y_{n+1}^{[t]} = \widetilde{y}_{n+1} + \sum_{i=1}^{s} [\xi_{n+1}^{(i)}]^{[t]} \widetilde{c}_{n+1}^{(i)}.$$

It is straightforward (but somewhat tedious) to see that for the linear problem (2.1) the iteration (3.4) terminates after one step and reduces to (2.6).

Note that in the above approach we are not restricted to using the trapezoidal rule (2.4). Any (implicit) linear multistep method with suitable stability properties could be used instead, before scalarizing.

4. Numerical Example. Enright et al. in [3] give the following separably stiff initial value problem, arising in insulator physics:

(4.1(i)) $$^{1}y' = -^{1}y + 10^8 * {}^{3}y(1 - {}^{1}y), \qquad {}^{1}y(0) = 1,$$

(4.1(ii)) $$^{2}y' = 10 * {}^{2}y + 3 * 10^7 * {}^{3}y(1 - {}^{2}y), \qquad {}^{2}y(0) = 0,$$

(4.1(iii)) $$^{3}y' = -^{1}y' - {}^{2}y', \qquad {}^{3}y(0) = 0,$$

$$x \in [0, 1].$$

The Jacobian $\partial f(x, y)/\partial x$ of this system is singular for all (x, y) (this is implied by (4.1(iii))); thus, one of the eigenvalues is zero. On the exact solution curve the second eigenvalue decreases from -1 to -8.6, the third and dominant one decreases from $-3 * 10^7$ to approximately $-4 * 10^7$, as x increases from 0 to 1.

(4.1) was tackled by both versions of the RS method, using the 4th order Adams-Bashforth method as basic method, a step-length $h = 0.01$ and "exact" starting values $y(x_1), y(x_2), y(x_3), y(x_4)$, computed by a Runge-Kutta method with sufficiently small step-length. The initial value $y(x_0) = \eta$ was not used because it lies in the transient phase.

The first version of the reduction to scalar method, as it is described in [1], failed after a few steps, because the iteration for the correction factor did not converge in 20 steps. If the accuracy requirements on the correction factor were decreased, the iteration did converge, but then the numerical solution was sufficiently inaccurate for the Jacobian no longer to be separably stiff (i.e. $\lambda^{(2)}$ became much larger), and the power method did not converge. The failure occurred at $x_{n+1} = 0.07$ or 0.08 depending on the convergence parameters in the power method, and the routine used to compute the correction factor. With our new version, the computation was successful.

For comparison purposes, an "exact" solution was computed using the standard 4th order Runge-Kutta method with the very small step-length $h = 5 * 10^{-8}$, which was required by stability.

Measures of the maximum correction factor, and the maximum dominant and subdominant error, respectively, are given by

$$MC = \max_{5 \leqslant n \leqslant 100} |\xi_n^{(1)}|,$$

$$MD = \max_{5 \leqslant n \leqslant 100} |\langle \widetilde{d}_n^{(1)}, y(x_n) - y_n \rangle|,$$

$$MS = \max_{5 \leqslant n \leqslant 100} \|y(x_n) - y_n - \langle \widetilde{d}_n^{(1)}, y(x_n) - y_n \rangle \widetilde{c}_n^{(1)}\|_\infty.$$

Thus, MD gives the maximum error in the dominant component of the numerical solution; and MS gives the infinity norm of the error after the dominant component has been subtracted.

The computations were carried out in DOUBLE PRECISION, carrying 19 digits. The results for our new version are:

$$MC = 9.63 * 10^{-12}, \quad MD = 3.08 * 10^{-17}, \quad MS = 1.60 * 10^{-6}.$$

The numerical solution is stable, and the results are accurate. The subdominant accuracy can be improved by decreasing the step-size.

For a comparison of the above results with those obtained by a gradient prediction CDS method; see [2].

Conclusions. We have seen that the first version of the Reduction to Scalar method, advocated in [1] as the best CDS method available, may fail for severely nonlinear naturally arising problems. The improved version, proposed in this paper, is more robust with respect to nonlinearity. Its disadvantage is that it is computationally more expensive than the first version if $s > 1$.

Department of Mathematics
University of Utah
Salt Lake City, Utah 84112

1. P. ALFELD & J. D. LAMBERT, "Correction in the dominant space: A numerical technique for a certain class of stiff initial value problems," *Math. Comp.*, v. 31, 1977, pp. 922−938.

2. P. ALFELD, "Inverse linear multistep methods for the numerical solution of initial value problems of ordinary differential equations," *Math. Comp.*, v. 33, 1979, pp. 111−124.

3. W. E. ENRIGHT, T. E. HULL & B. LINDBERG, "Comparing numerical methods for stiff systems of o.d.e.s," *BIT*, v. 15, 1975, pp. 10−48.