On the Diagonal Approximation of Full Matrices

Walter M. Lioen <walter@cwi.nl>
CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

ABSTRACT
In this paper the construction of diagonal matrices, in some sense approximating the inverse of a given square matrix, is described. The matrices are constructed using the well-known computer algebra system Maple. The techniques we show are applicable to square matrices in general. Results are given for use in Parallel diagonal-implicit Runge-Kutta (PDIRK) methods. For an s-stage Radau IIA corrector we conjecture $s!$ possibilities for the diagonal matrices.

1991 Mathematics Subject Classification: Primary: 65H10, Secondary: 65D99, 65F35
1991 Computing Reviews Classification System: G.1.2, G.1.5, G.1.7, I.1.4
Keywords and Phrases: Parallel diagonal-implicit Runge-Kutta methods, convergence, method parameters, symbolic computation.
Note: The research reported in this paper was supported by the Technology Foundation (STW) in the Netherlands.

1. Introduction

In this paper the construction of diagonal matrices D, approximating the inverse of a given square matrix A, is described. The diagonal matrices are chosen such that the spectral radius $\rho(I-D^{-1}A) = 0$. The diagonal matrices are constructed using the well-known computer algebra system Maple. In this paper we construct diagonal matrices for a special class of square matrices. However, the techniques we use are independent of the choice of the square matrices.

2. The construction of diagonal matrices D using Maple

We are solving $\rho(I - D^{-1}A) = 0$. This is equivalent to $\det(D^{-1}A - \lambda I) = 0$ having s unit zeros:

$$\det(D^{-1}A - \lambda I) = (1-\lambda)^s.$$

Evaluation yields

$$(-\lambda)^s + \sum_{i=1}^{s} \chi_i (-\lambda)^{s-i} = (-\lambda)^s + \sum_{i=1}^{s} \binom{s}{i} (-\lambda)^{s-i}, \tag{2.1}$$

where the χ_i denote the coefficients of the characteristic polynomial of $D^{-1}A$. The χ_i equal the sum of the i-th degree principal minors (symmetrical with respect to the main diagonal) of $D^{-1}A$. In particular, $\chi_1 = \mathrm{trace}(D^{-1}A)$ and $\chi_s = \det(D^{-1}A)$. In (2.1), the left-hand side polynomial equals the right-hand side polynomial if the corresponding polynomial coefficients are equal. This gives us a nonlinear system in s unknowns

$$\chi_i = \binom{s}{i}, \qquad i = 1, \ldots, s, \tag{2.2}$$

where, since A is given, the χ_i are multivariate polynomials in the diagonal matrix elements of $D^{-1} := \mathrm{diag}(d_1, \ldots, d_s)$. Since a χ_i consists of $\binom{s}{i}$ terms (multivariate polynomial coefficients in $d_1, \ldots, d_s$), the nonlinear system (2.2) will have $\sum_{i=1}^{s} \binom{s}{i} = 2^s - 1$ terms in total.

For given s, $\rho(I-D^{-1}A)=0$ means an eigenvalue 0 with algebraic multiplicity s. This eigenproblem is extremely bad conditioned. For the value of a numerically computed $\rho(I-D^{-1}A)$, we expect something of the order $\sqrt[s]{\epsilon}$, where ϵ denotes the machine precision. The Maple engine uses *Digits* decimal digits (user specifiable), so Maple's machine precision is $\epsilon = 10^{1-Digits}$. The extremely bad conditioned eigenproblem is the reason that we need to compute in higher precision, however, for growing s, the number of terms in our nonlinear system grows exponentially: 2^s-1. Using a higher precision, by using multiple machine words to represent a floating point number, implies a quadratic growth (in the number of machine words) of the computational complexity of a multiplication, say. The exponential growth in the number of terms and the quadratic growth in the number of machine words together are the reason we cannot compute the D in arbitrary precision.

3. The construction of D matrices for PDIRK methods

Van der Houwen and Sommeijer introduced the Parallel Diagonal-Implicit Iteration of Runge-Kutta (PDIRK) methods [2] for the solution of stiff initial value problems:

$$y'(t)=f(y(t)), \qquad y(t_0)=y_0, \qquad y,f\in\mathbf{R}^d.$$

The PDIRK method can be presented as follows

for $n=1,\ldots,N$
  $Y_n^{(0)}=P(Y_{n-1}^{(m)},y_{n-1})$
  for $j=1,\ldots,m$
    $Y_n^{(j)}-h_n(D\otimes I_d)F(Y_n^{(j)})=e\otimes y_{n-1}+h_n((A-D)\otimes I_d)F(Y_n^{(j-1)})$
  $y_n=y_{n-1}+h_n(b^T\otimes I_d)F(Y_n^{(m)})$

where s denotes the number of stages of the Runge-Kutta method described by c, A, and b. The vector Y_n consists of the s stacked stage vectors $y_{n,1},\ldots,y_{n,s}$, each being an approximation to the solution at the intermediate time points $t_{n-1}+c_ih_n$, and $F(Y_n)$ consists of the stacked vectors $f(y_{n,1}),\ldots,f(y_{n,s})$. I_d denotes the d-dimensional identity matrix and $e=(1,\ldots,1)^T\in\mathbf{R}^s$. P denotes a predictor, N is the number of time-steps, and m is the number of iterations needed to solve the corrector equation to some prescribed precision. By requiring D being a diagonal matrix, the s stage vectors in $Y_n^{(j)}$ can be computed in parallel. Otherwise, D is still free and can be used to obtain good convergence.

Let $Z(z)$ denote the iteration matrix of the PDIRK method with $z=\lambda h$, h being the stepsize and with λ running through the spectrum $\Lambda(J)$ of the Jacobian J. In [2] it was shown that $Z(z)=z(I-zD)^{-1}(A-D)$. It is desirable to have $\rho(Z(z))$ small in the closed left halfplane. Because $\rho(Z(z))$ is an analytic function in the closed left halfplane, its maximum in $\Re(z)\le 0$ is assumed on the boundary, i.e., on the imaginary axis.

For $|z|\to 0$, corresponding with the nonstiff error components, $Z(z)\to z(A-D)$. For $|z|\to\infty$, corresponding with the stiff error components, $Z(z)\to I-D^{-1}A$, which is the matrix in the equation $\rho(I-D^{-1}A)=0$ we try to solve.

In their paper, Van der Houwen and Sommeijer [2] discuss several possibilities for the choice of the matrices D for several types of Runge-Kutta correctors. The approach for the choice of the D matrix adopted in this paper is based on the minimization of the spectral radius of the matrix $I-D^{-1}A$, resulting in a strong damping of the stiff error components. For the two-stage Radau IIA corrector, the matrix D was computed straightforwardly in such a way that the spectral radius $\rho(I-D^{-1}A)=0$ (giving two solutions). For $s>2$ the D matrices were computed using a numerical minimization routine on $\rho(I-D^{-1}A)$. (The bad condition of the eigenproblem is one of the reasons, the minimization routine used in [2] had a hard time searching for solutions.)

In 't Hout [1] suggested not just minimizing $\rho(I-D^{-1}A)$ for $s>2$, but merely solving $\rho(I-D^{-1}A)=0$. This way, Sommeijer [3] was able to construct D matrices for three-stage Radau IIA and Gauss-Legendre correctors. In both cases, there are exactly four real solutions. In this paper, we shall extend this approach for $s>3$.

Table 1: D matrices and some statistics for Radau IIA, $s = 2$

D	$\rho(A-D)$	$z, \max(\rho(Z(z)))$	$\rho(I-D^{-1}A)$
$\mathrm{diag}(\frac{2}{3}-\frac{1}{6}\sqrt{6}, \frac{2}{5}+\frac{1}{10}\sqrt{6})$	$\frac{2}{5}-\frac{1}{15}\sqrt{6}$	$\sqrt{6}i, \frac{1}{250}(6-\sqrt{6})(16+\sqrt{6})$	0
$\mathrm{diag}(\frac{2}{3}+\frac{1}{6}\sqrt{6}, \frac{2}{5}-\frac{1}{10}\sqrt{6})$	$\frac{2}{5}+\frac{1}{15}\sqrt{6}$	$\sqrt{6}i, \frac{1}{250}(6+\sqrt{6})(16-\sqrt{6})$	0

Using Maple, being a computer algebra system for doing symbolic computations, suggests we are going to use symbolic computations throughout. However, for Radau IIA and for many other Runge-Kutta methods the entries of c, that in turn define the matrix A, are the roots of an s-degree polynomial $P(x)$, say. In general, explicit roots in terms of radicals for polynomials of degree greater than 4 do not exist. Since for Radau IIA, $P(x)/(x-1)$ is irreducible for at least $s = 2, \ldots, 8$ we cannot compute the c and therefore A and b symbolically for $s > 5$. Although we can compute A for $s = 4, 5$ symbolically, we refrain from doing so for reasons we will explain below. Because we are still using a computer algebra system, we can, once we found an approximate solution, refine this solution to any given accuracy. This can be done using Maple's fsolve, in this case a numerical nonlinear system solver, with a more accurate approximation of the nonlinear system and a small interval containing the approximation found thus far.

In the next subsections, we derive the matrices $D^{-1} = \mathrm{diag}(d_1, \ldots, d_s)$ for the Radau IIA correctors with $s = 2, \ldots, 8$. If there are multiple solutions to $\rho(I - D^{-1}A) = 0$ we choose the D that minimizes $\max(\rho(Z(z)))$ over the closed left halfplane.

3.1 Construction of diagonal matrix D for Radau IIA, $s = 2$

The Radau IIA matrix for $s = 2$ is given by

$$A = \begin{pmatrix} \frac{5}{12} & -\frac{1}{12} \\ \frac{3}{4} & \frac{1}{4} \end{pmatrix}, \qquad D^{-1}A = \begin{pmatrix} \frac{5}{12}d_1 & -\frac{1}{12}d_1 \\ \frac{3}{4}d_2 & \frac{1}{4}d_2 \end{pmatrix}.$$

Equating the polynomial coefficients, as described in the previous section, gives rise to the following nonlinear system

$$\begin{cases} \mathrm{trace}(D^{-1}A) = \binom{2}{1} \\ \det(D^{-1}A) = \binom{2}{2} \end{cases} \quad \Leftrightarrow \quad \begin{cases} \frac{5}{12}d_1 + \frac{1}{4}d_2 = 2 \\ \frac{1}{6}d_1 d_2 = 1 \end{cases}.$$

Solution of this system is straightforward and gives the following two solutions

$$d_1 = \frac{12}{5} \mp \frac{3}{5}\sqrt{6}, \quad d_2 = 4 \pm \sqrt{6}.$$

This leads to the figures listed in Table 1. Clearly, the first solution in this table gives the best overall convergence because its $\max(\rho(Z(z)))$ is smallest.

3.2 Construction of diagonal matrix D for Radau IIA, $s = 3$

For $s = 3$ we still use the symbolically computed A matrix. However, the solution of the nonlinear system involves the roots of an irreducible 6-th degree polynomial. The d_2 in the solutions are the roots of

$$\begin{aligned} &\left(19772 + 6483\sqrt{6}\right)x^6 + \left(-425520 - 125280\sqrt{6}\right)x^5 + \left(3502800 + 767700\sqrt{6}\right)x^4 \\ &\quad + \left(-14308800 - 1126800\sqrt{6}\right)x^3 + \left(30844800 - 3337200\sqrt{6}\right)x^2 \\ &\quad + \left(-34041600 + 10022400\sqrt{6}\right)x + 16358400 - 6177600\sqrt{6}, \end{aligned}$$

Table 2: D matrices and some statistics for Radau IIA, $s = 3$

D	$z, \max(\rho(Z(z)))$	$\rho(I - D^{-1}A)$
diag(0.3203827776857808, 0.1399668046773267, 0.3716674595229115)	$5.86i, 0.401$	$0.226 \cdot 10^{-18}$
diag(0.5587475041000601, 0.1365365144606963, 0.2184662437345018)	$4.17i, 0.466$	$0.743 \cdot 10^{-18}$
diag(0.1040499402500167, 0.3328127454285067, 0.4812901402100924)	$5.39i, 0.472$	$0.257 \cdot 10^{-18}$
diag(0.7476215577338980, 0.4040614917633793, 0.0551720944386193 4)	$8.72i, 0.658$	$0.149 \cdot 10^{-18}$

the d_1 and d_3 belonging to the root chosen for d_2 are both algebraic expressions containing this root. Although we still can compute the D elements in an arbitrary precision, we no longer can compute them symbolically. This is where we switch to numerical computations. By using 60 digits (so our Maple machine precision $\epsilon = 10^{-59}$), we may expect and also get a $\rho(I - D^{-1}A)$ value in the order of magnitude of 10^{-20}. We give the D matrix elements in 16 digits accuracy, enough for 64 bit IEEE floating point arithmetic used by most today workstations. Since the 6-th degree polynomial only has 4 real-valued roots, we only get 4 solutions in Table 2. Clearly, the first solution in this table gives the best overall convergence because its $\max(\rho(Z(z)))$ is smallest.

3.3 Construction of diagonal matrix D for Radau IIA, $s = 4$

The nonlinear system for $s = 4$ (here presented with truncated precision floating point values for the multivariate polynomial coefficients) looks as follows.

$$\begin{aligned} 0.113d_1 + 0.207d_2 + 0.189d_3 + 0.0625d_4 &= 4 \\ 0.0328d_1d_2 + 0.0158d_1d_3 + 0.00925d_1d_4 + 0.0585d_2d_3 + 0.00670d_2d_4 + 0.0198d_3d_4 &= 6 \\ 0.0101d_1d_2d_3 + 0.000756d_1d_2d_4 + 0.00145d_1d_3d_4 + 0.00672d_2d_3d_4 &= 4 \\ 0.00119d_1d_2d_3d_4 &= 1 \end{aligned}$$

If we compute the nonlinear system for $s = 4$ symbolically, every attempt solving this system ended with Maple's error message 'Error, (in expand/bigprod) object too large'. As we will explain in the next subsection, for $s > 4$, we can not determine all solutions anymore if we switch to using floating-point arithmetic. However, we are still able to approximate all solutions by first solving a rational approximation of this nonlinear system. We do this by taking a rational approximation for the multivariate polynomial coefficients. This way, the Maple object for the nonlinear system becomes much smaller than the symbolically determined nonlinear system. Even following this approach, Maple needs some help: first we solve $\{d_1, d_2, d_3\}$ from equations 1, 2 and 4 yielding d_1, d_2 and d_3 as expressions containing the roots of a 6-th degree polynomial with d_4 in its coefficients. Substituting these solutions in equation 3 yields d_4 as roots of a 24-th degree polynomial and roots of a 4-th degree polynomial. All roots of the 4-th degree polynomial appear to be spurious solutions either introduced by the substitution or introduced by Maple's nonlinear system solver. Every d_4, root of the 24-th degree polynomial, corresponds with exactly 1 root of the 6-th degree polynomial in the expressions for d_1, d_2 and d_3. However, if we substitute a d_4 in the d_1, d_2 and d_3 expressions containing the roots of a 6-th degree polynomial with d_4 in its coefficients, we find 6 solutions. By looking at the residues we are able to select the correct root. Thus we find 24 solutions for the D matrices. Of these 24 solutions only 8 are real-valued (the complex solutions all have a significant imaginary part). Using the previously found 8 real-valued approximations as starting values for Maple's numerical nonlinear system solver, we refined the solutions. The results are presented in Table 3. Clearly, the first solution in this table gives the best overall convergence because its $\max(\rho(Z(z)))$ is smallest. Note that the best solution is only marginally better than the third solution in this table, which corresponds with the D and refines the D presented in [2].

3.4 Construction of diagonal matrices D for Radau IIA, $s = 5, \ldots, 8$

For $s = 2$ we found 2 solutions and for $s = 3$ we found 6 solutions (two complex conjugated roots, leaving only 4 interesting ones). For $s = 4$ we found 24 solutions, 8 of them being real-valued.

Table 3: D matrices and some statistics for Radau IIA, $s = 4$

D	$z, \max(\rho(Z(z)))$
diag(0.1527853137467750, 0.08774983992555644, 0.2636113044230077, 0.3368439415344046)	$7.35i, 0.516$
diag(0.2486697765179635, 0.08780166100259280, 0.1959784536013074, 0.2782189143400108)	$9.55i, 0.527$
diag(0.3192979656769842, 0.08871403314492813, 0.1809065091618870, 0.2323154243215252)	$8.83i, 0.528$
diag(0.3813455220335083, 0.2236996990743941, 0.07777745058977572, 0.1794250018603748)	$10.7i, 0.575$
diag(0.3247262709818679, 0.2121007378979642, 0.07907332247161736, 0.2185904187299887)	$10.5i, 0.592$
diag(0.05363587665020470, 0.1829772752695088, 0.3149333835926415, 0.3851673585460386)	$10.7i, 0.626$
diag(0.4736365821033994, 0.3205127563668334, 0.07509019550034840, 0.1044352044734334)	$8.01i, 0.627$
diag(0.5589306159049463, 0.4027756536996861, 0.1964584550831966, 0.02691713787967921)	$23.1i, 0.828$

If we look carefully at the structure of the nonlinear system and look what happens if we try the eliminations for $s = 4$ by hand, we see that eliminating the first equation gives rise to equations of degree 2. Eliminating an equation of degree 2 we have 2 solutions and the other equations become of degree 3. Eliminating an equation of degree 3 we have 3 solutions and the other equations become of degree 4, and so on. Thus, we are tempted to formulate the following conjecture.

Conjecture 1 *For Radau IIA matrices, the nonlinear system has $s!$ solutions.*

This conjecture is true for $s \leq 3$, since we found $s!$ solutions analytically. For $s = 4$ we did not formally prove $s!$ solutions, however, using interval arithmetic, the $s!$ numerical solutions found, can almost probably be proven to be solutions.

However, not all solutions have to be useful. The solutions not only have to give rise to positive real numbers, but also $\max(\rho(Z(z))) < 1$ should hold on the imaginary axis. (In fact, for $s = 8$ we found a solution with the property $\max(\rho(Z(z))) > 1$.)

For $s > 4$ we could not find an attack for finding all solutions. Using a rational approximation of the nonlinear system, as we did for $s = 4$, and trying to follow the heuristic reasoning leading to the conjecture above, does not work: inevitably, we will end up with some d_i being the roots of a too high order polynomial containing d_j-s, $j \neq i$, for which we cannot solve the roots explicitly.

For $s > 4$ we solve a floating-point approximation of the nonlinear system using Maple's fsolve (solve using floating-point arithmetic) in this case a numerical nonlinear system solver. Since fsolve attempts to compute a single real root, we have to specify search intervals for finding different real roots. Finally, for $s = 9$, even using 1 word per multivariate polynomial coefficient, the nonlinear system became too big for Maple to solve: 'Error, (in expand/bigprod) object too large'.

In Table 4–5 we only present the solutions found with the smallest $\max(\rho(Z(z)))$ for $s = 2, 3, 4$. For $s = 5, \ldots, 8$ we found multiple solutions—definitely not all—and for given s, we present the one with minimal $\max(\rho(Z(z)))$. Consequently, we do not claim the tabulated D to be optimal in the sense that $\max(\rho(Z(z)))$ is minimal, for $s > 4$. An impression on the behavior of $\rho(Z(z))$, can be found in Figure 1–2.

4. Conclusions

We were able to construct D matrices for the PDIRK method with the Radau IIA corrector for $s = 2, \ldots, 8$. There is no reason why we could not use the same method for other correctors. The machine readable coefficients can be obtained from the author.

Acknowledgements

I gratefully acknowledge Dr. K.J. in 't Hout and Dr. B.P. Sommeijer for introducing me to the problem of the construction of diagonal matrices. I also want to acknowledge Dr. B.P. Sommeijer for his valuable comments during the preparation of this paper.

Table 4: D matrices for Radau IIA, $s = 2, \ldots, 8$

s	D
2	$\mathrm{diag}(\frac{2}{3} - \frac{1}{6}\sqrt{6}, \quad \frac{2}{5} + \frac{1}{10}\sqrt{6})$
3	$\mathrm{diag}(0.3203827776857808, \ 0.1399668046773267, \ 0.3716674595229115)$
4	$\mathrm{diag}(0.1527853137467750, \ 0.0877498399255644, 0.2636113044230077, \ 0.3368439415344046)$
5	$\mathrm{diag}(0.2030587241544029, \ 0.1359509620271233, \ 0.0634672671977216 1, 0.1739534363905672, 0.2170008621974159)$
6	$\mathrm{diag}(0.2821041897853654, \ 0.2357235133749619, \ 0.1599420020814069, \ 0.0448827902540156 1, 0.0725183723002662 8, 0.0868405694117026 5)$
7	$\mathrm{diag}(0.2532796578435856, \ 0.2241262251569997, \ 0.1760533860274691, \ 0.1147022645744301, 0.0327728917938160 0, 0.0509280510349084 9, 0.0604325607811130 5)$
8	$\mathrm{diag}(0.2276829014827774, \ 0.2081000197483005, \ 0.1752690782773552, \ 0.1328007951785409, 0.0844317036804653 4, 0.0246390605306660 5, 0.0376039194943007 6, 0.0446744893554889 5)$

Table 5: Some statistics of the D matrices presented in Table 4

s	$\rho(A - D)$	$z, \max(\rho(Z(z)))$	$\rho(I - D^{-1}A)$
2	$\frac{2}{5} - \frac{1}{15}\sqrt{6}$	$\sqrt{6}i, \frac{1}{250}(6 - \sqrt{6})(16 + \sqrt{6})$	0
3	0.155	$5.86i, 0.401$	$0.226 \cdot 10^{-18}$
4	0.199	$7.35i, 0.516$	$0.129 \cdot 10^{-14}$
5	0.113	$14.6i, 0.622$	$0.816 \cdot 10^{-12}$
6	0.191	$16.5i, 0.720$	$0.149 \cdot 10^{-9}$
7	0.181	$23.1i, 0.813$	$0.478 \cdot 10^{-8}$
8	0.168	$32.1i, 0.898$	$0.571 \cdot 10^{-7}$

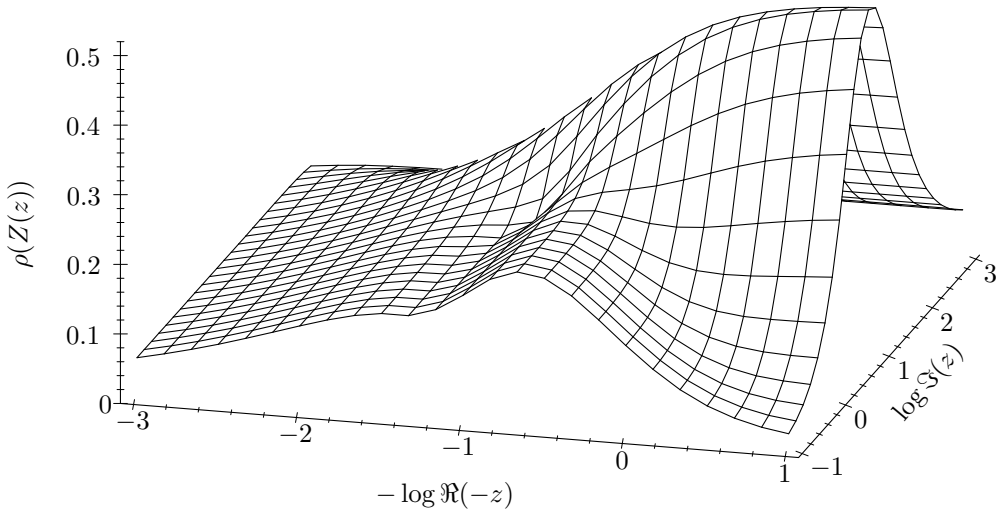

Figure 1: $\rho(Z(z))$ in the second quadrant for Radau IIA, $s = 4$

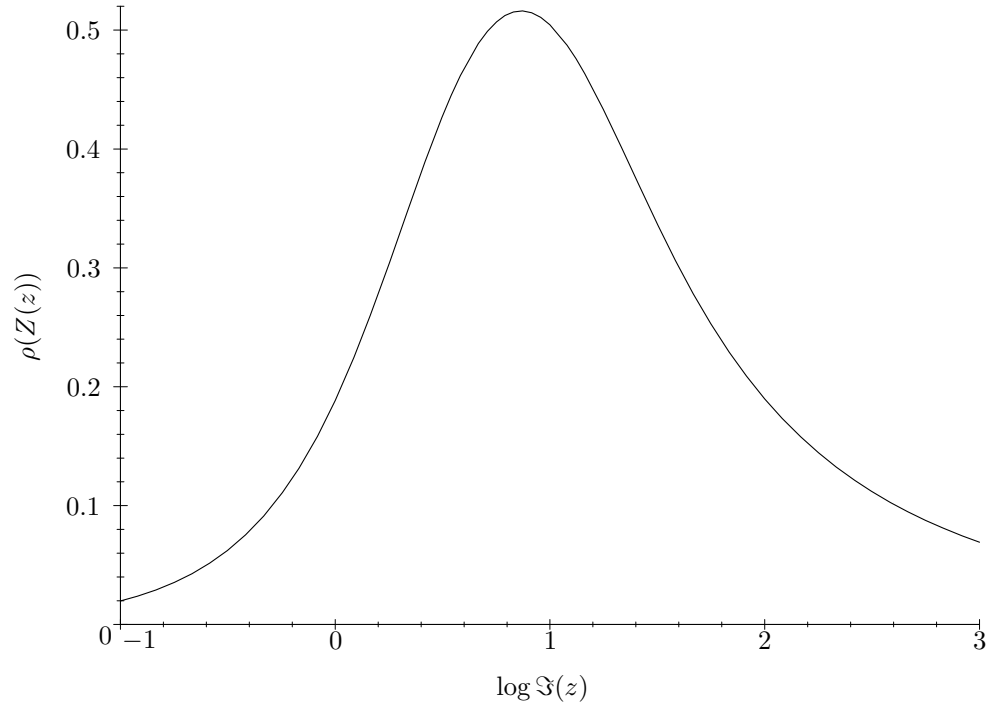

Figure 2: $\rho(Z(z))$ along the positive imaginary axis for Radau IIA, $s = 4$

References

1. K.J. in 't Hout. Private communication, April 1994.
2. P.J. van der Houwen and B.P. Sommeijer. Iterated Runge-Kutta methods on parallel computers. *SIAM J. Sci. Stat. Comput.*, 12(5):1000–1028, 1991.
3. B.P. Sommeijer. Private communication, April 1994.