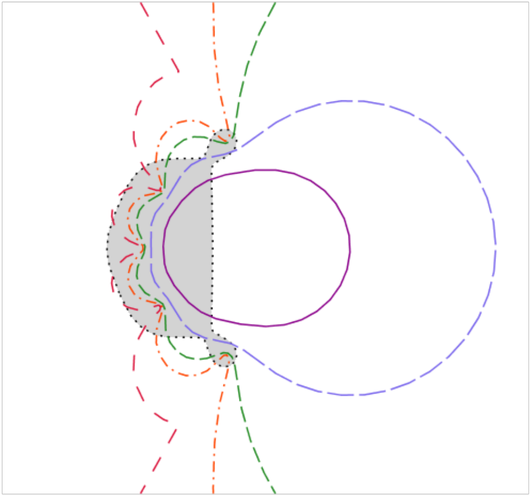
Optimal Backward Error & the Dahlquist Test Problem

Robert M Corless[1]

Joint work with Yalçin Kaya[2] and Robert H C Moir[1]

[1]Department of Applied Mathematics
The University of Western Ontario
[2]Division of Information Technology, Engineering and the Environment
School of Information Technology and Mathematical Sciences
University of South Australia

2015 AMMCS-CAIMS Congress
Modified for McMaster University
October 8, 2015



Suppose that you have an initial-value problem (ODE) to solve:

$$\dot{x} = f(t, x), \quad x(t_0) = x_0, \quad t_0 \leq t \leq t_{final}.$$

- Would you bet \$1000 that your numerical solution to *your* initial-value problem was correct?
- What if someone else wrote the solver?
- What if you had an easy way to test the solution before you bet?
- What does "correct" mean anyway?

- Suppose $x_{ref}(t)$ is the "true" reference solution to the IVP $\dot{x} = f(t,x), \quad x(t_0) = x_0$.

We do not know $x_{ref}(t)$ (else why compute?)

- Suppose $z(t)$ is our computed solution, interpolating the *skeleton* (t_k, x_k) which has mesh widths (step sizes) $h_k = t_{k+1} - t_k$.
- The "forward error" is $z(t) - x_{ref}(t)$.
- The "backward error" (residual) is $\Delta = \dot{z}(t) - f(t, z(t))$.

$\Delta(t)$ is computable, and can often be understood or interpreted as a model perturbation: $\dot{z} = f(t, z(t)) + \Delta(t)$.

- The "local error" needs a new concept:

"local reference solution"

the reference solution to $\dot{x} = f(t, x), \quad x(t_k) = x_k$.

- Call these $x_k(t), \quad t_k \leq t \leq t_{k+1}$.
- The "local errors" are $z(t) - x_k(t)$ on $t_k \leq t \leq t_{k+1}$.
- Typically these are largest at t_{k+1}^-.

- Forward error and residual are related by "conditioning" (sensitivity), *e.g.* by

Gröbner-Alexeev nonlinear variation of constants formula

$$z(t) - x_{ref}(t) = \int_{t_0}^{t} G(t, \tau, z, x_{ref})\Delta(\tau)d\tau.$$

- If $\|\Delta(t)\| = O(h^p)$ as $h = mean\ h_k \to 0$ then $\|z - x_{ref}\| = O(h^p)$ also (we say the numerical method has order p).

- J. Wilkinson first popularized backward error in numerical linear algebra.
- He attributed it to Givens, but von Neumann & Goldstine had the notion of "condition number" ("figure of merit").
- Henrici realized the notion was very general.
- Warming & Hyett and then Griffiths & Sanz-Serna looked at "the method of modified equations".
- Zadunaisky invented "defect correction", an iterated improvement scheme using backward error.
- Stetter proved that, asymptotically as $h \to 0$, $\|\Delta\| \sim (\text{local error})/h$.
- Enright, in the 1980s, showed defect (residual) *control* was a viable strategy for RK solvers; Shampine used it for BVP solvers (esp `bvp4c` in MATLAB).
- The book Corless & Fillion uses backward error throughout.

Remark: backward error is not a panacea — there are problems for which small forward error is possible but not backward error.

- Most codes supply interpolants: for graphical output, for event location, for handling delay DE.
- These interpolants should be $O(h^p)$ accurate, but *sometimes aren't*.

Example

In MATLAB, ode45 uses a fifth-order Runge-Kutta Fehlberg formula, but has only a fourth-order interpolant: so $z'(t)$ will only be third-order accurate.
This sometimes overestimates $\Delta(t) := \dot{z}(t) - f(t, z)$.

- Some years ago RMC proposed finding "optimal" interpolants, that minimized

$$\|\Delta\|_2^2 = \frac{1}{h}\int_{t_n}^{t_{n+1}} \Delta^H(\tau)\Delta(\tau)d\tau.$$

- This leads to the Euler-Lagrange equations

$$\begin{aligned} \dot{z} - f(t,z) &= \Delta \\ \dot{\Delta} + J_f^H\Delta &= 0, \end{aligned}$$

  to be solved as a Boundary-Value problem with $z(t_n) = x_n, \quad z(t_{n+1}) = x_{n+1}.$
- This works, but it's not what we'll talk about today.

- YK suggested we look also at minimizing $\|\Delta\|_\infty$, which leads to optimal control problems: find $u(t)$ such that

$$\dot{z} = f(t, z) + u(t)$$

steers $z(t)$ from $z(t_n) = x_n$ to $z(t_{n+1}) = x_{n+1}$ with minimal $\|u\|_\infty$.
- These turn out to be solvable in some cases using the Pontrjagin maximum principle, and in others by using optimization packages such as `AMPL`.
- We are interested in the *relative* optimality:

$$\dot{z} = f(t, z)(1 + \delta(t)).$$

- We'll just do some "baby" optimal control problems here.

- Suppose $f(t,x)$ is scalar, and separable: $f(x,t) = X(x)T(t)$; so that the equation $\dot{z} = f(t,z)(1+\delta(t))$ is also separable.
- Then,

$$\dot{z} = X(z)T(t)(1+\delta(t)) \Rightarrow \frac{dz}{X(z)} = T(t)dt + T(t)\delta(t)dt.$$

- So,

$$\int_{x_n}^{z(t)} \frac{d\zeta}{X(\zeta)} - \int_{t_n}^{t} T(\tau)d\tau = \int_{t_n}^{t} T(\tau)\delta(\tau)d\tau$$

giving the constraint

$$C = \int_{x_n}^{x_{n+1}} \frac{d\zeta}{X(\zeta)} - \int_{t_n}^{t_n+h} T(\tau)d\tau = \int_{t_n}^{t_n+h} T(\tau)\delta(\tau)d\tau.$$

- C is in principle known.

Optimal Interpolants for Optimal Residual (∞-Norm)

- By the triangle inequality,

$$|C| \leq \int_{t_n}^{t_n+h} |T(\tau)| d\tau \quad \cdot \quad \|\delta(t)\|_\infty .$$

- So no matter what control $\delta(t)$ is chosen,

$$\|\delta\|_\infty \geq \frac{|C|}{\int_{t_n}^{t_n+h} |T(\tau)| d\tau} .$$

- By choosing

$$\delta(\tau) = \overline{\text{signum}(T(\tau))} \cdot \frac{C}{\int_{t_n}^{t_n+h} |T(\tau)| d\tau}$$

(signum$(re^{i\theta}) = e^{i\theta}$) this bound is achieved while satisfying the constraint:

$$\int_{t_n}^{t_n+h} T(\tau)\delta(\tau) d\tau = \int_{t_n}^{t_n+h} T(\tau) \frac{\overline{\text{signum}(T(\tau))} \cdot (C)}{\int_{t_n}^{t_n+h} |T(\tau)| d\tau} d\tau = C.$$

- This explicitly gives us our minimum residual (and the optimal interpolant is $z(t)$).
- Note: more general equations need the full maximum principle.

- We've solved a number of examples this way, including $\dot{x} = x^2$, a non-compact example, and $\dot{x} = -\sqrt{x}$ (Torricelli's law, which is not Lipschitz), and several systems.
- We note that this method of assessment is *independent* of the numerical method used: all we need is $(t_n, x_n), \quad (t_n + h, x_{n+1}),$ and the original equation.
- We have examined several such methods.
- Today we'll look at perhaps the simplest interesting problem:

$$\dot{x} = \lambda x, \qquad x(t_n) = x_n.$$

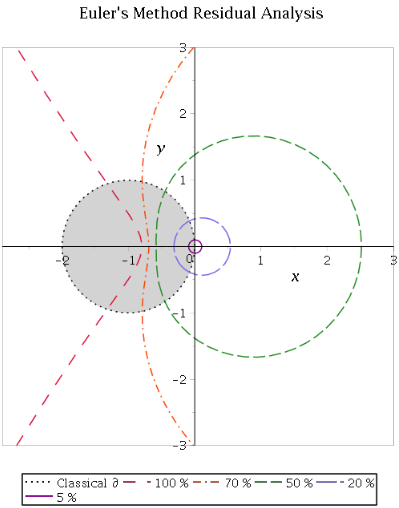
- Without loss of generality we can shift the origin to t_n

$$\dot{x} = \lambda x, \quad x(0) = x_n \quad \text{on } 0 \leq t \leq h.$$

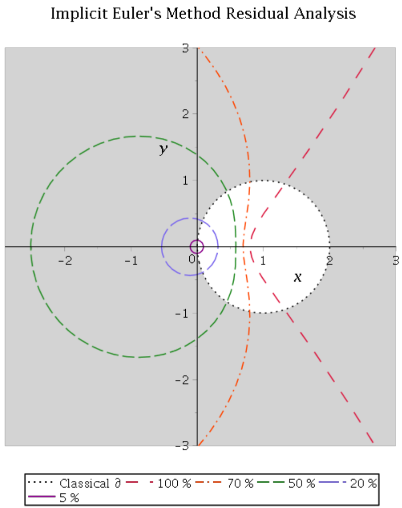
- This is the Dahlquist test problem.
- For $\mathrm{Re}(\lambda) \ll 0$ it is the simplest example of a “stiff” problem.
- It arises also on linearization of nonlinear problems and doing eigenvalue analysis.
- It's more important to numerical analysis than one might think.

- Considering the parameter $\mu = \lambda h$ and the function $R(\mu)$, which is the approximation of the exponential function provided by a given numerical method (the exponential being the solution to the reference problem), we obtain a classical measure of the stability of the numerical method.
- For $|R(\mu)| < 1$, the numerical solution of the Dahlquist test problem is uniformly bounded in $n \geq 0$.
- This determines the classical stability region in the complex plane of the given numerical method.
- These regions give rise to a variety of "familiar diagrams" for familiar numerical methods.

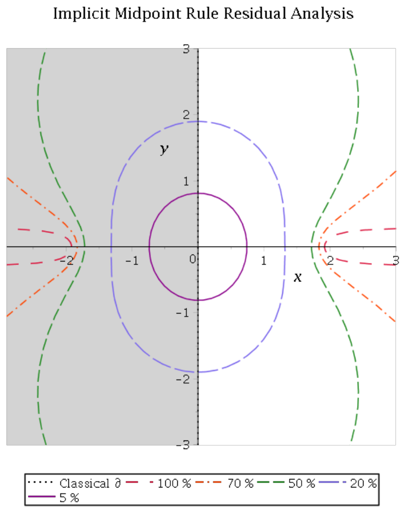
Classical Stability: Euler Method



Classical Stability: Implicit Euler Method

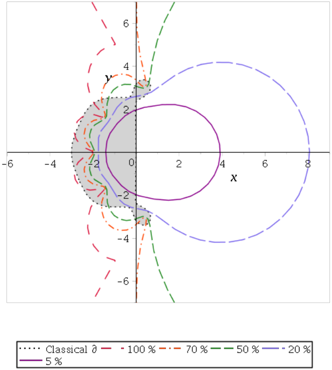


Classical Stability: Implicit Midpoint Rule (Or Any Exactly A-Stable Method)

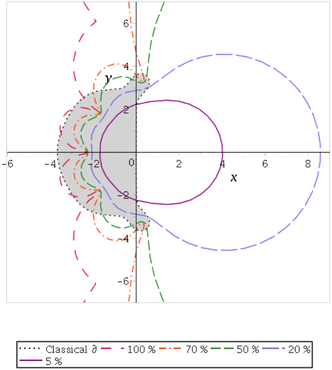


Classical Stability: RKF45

Fourth Order Part RKF45 Method Residual Analysis

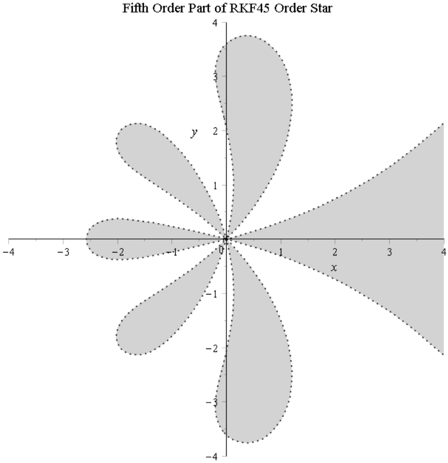


Fifth Order Part RKF45 Method Residual Analysis



RKF45 Order Star

- By considering not $|R(\mu)| < 1$ but the relative forward error $|R(\mu)e^{-\mu}|$ we are led to the theory of "order stars", which considers the regions $A_+ : |R(\mu)e^{-\mu}| > 1, A_0 : |R(\mu)e^{-\mu}| = 1, A_- : |R(\mu)e^{-\mu}| < 1.$



- But we want the relative backward error:
- For $f(x) = \lambda x$, we get $\frac{dz}{z} = \lambda dt + \delta(t)dt$, so

$$\int_{x_n}^{x_{n+1}} \frac{dz}{z} - \lambda \int_{t_n}^{t_n+h} d\tau = \lambda \int_{t_n}^{t_n+h} \delta(\tau)d\tau,$$

or

$$\left|\ln_k\left(\frac{x_{n+1}}{x_n}\right) - \lambda h\right| = \left|\lambda \int_{t_n}^{t_n+h} \delta(\tau)d\tau\right| \leq |\lambda h| \, \|\delta\|_\infty.$$

- So, $\|\delta\|_\infty \geq \left|\frac{1}{\mu}\ln_k(R(\mu)) - 1\right|$, and equality is obtained if

$$\delta(\mu) = \frac{1}{\mu}\ln_k(R(\mu)) - 1.$$

- N.B. $\ln_k a := \ln a + 2\pi ik$

- Indeed this is the exact solution to the *same kind* of problem

$$x_n = R(\mu)^n x_0 = e^{\frac{\ln_k R(\mu)}{h} nh} x_0 .$$

- So,

$$\Lambda = \frac{\ln_k R(\mu)}{h} = \lambda \frac{\ln_k R(\mu)}{\mu}, \quad \dot{y} = \Lambda y!$$

$$y_{n+1} = R(\mu)y_n \qquad \text{e.g. } R(\mu) = 1+\mu \quad \text{Euler}$$
$$\text{or } R(\mu) = (1-\mu)^{-1} \quad \text{Implicit Euler}$$

Then $y_n = R^n(\mu)y_0$ by induction

$$= \mathrm{e}^{n \ln R(\mu)} y_0 \qquad = \mathrm{e}^{n(\ln R(\mu)+2\pi i k)} y_0$$
$$= \mathrm{e}^{n \ln_k R(\mu)} y_0 \qquad = \mathrm{e}^{nh\mu \frac{\ln_k R(\mu)}{\mu}} y_0$$
$$= \mathrm{e}^{\lambda t_n \cdot (1+\delta)} y_0 \quad \text{if } 1+\delta = \frac{\ln_k R(\mu)}{\mu}$$

$\therefore$ interpolant $z(t) = \mathrm{e}^{\lambda(1+\delta)t} y_0$ satisfies

$$\dot{z}(t) = \lambda(1+\delta)z(0) = y_0 \,.$$

Define:

$$K_R(\mu) = \mathrm{round}\left(\frac{\mathrm{Im}(\mu - \ln R(\mu))}{2\pi}\right)$$

(a kind of unwinding number, cf $K(z) = \frac{z - \ln e^z}{2\pi i} = \lceil \frac{\mathrm{Im}(z) - \pi}{2\pi} \rceil$)

Theorem:

$$K_R(\mu) = \arg_k \min |\delta|$$

Proof:

$$\begin{aligned}
\arg_k \min |\delta| &= \frac{\arg_k \min |\ln_k(R\mu)) - \mu|}{\mu} \\
&= \arg_k \min \left| \ln_k(\rho e^{i\theta}) - (\sigma + i\tau) \right| \\
&= \arg_k \min |\ln \rho - \sigma + i(\theta + 2\pi i k - \tau)|
\end{aligned}$$

k cannot affect the real part, and minimizes the imaginary part exactly when k is the nearest integer to $\frac{\tau-\theta}{2\pi}$. QED.

N.B. there may be more than one such k but they give the same δ .

- This will give us a *quantitative* assessment of the quality of the method $x_{n+1} = R(\mu)x_n$.

Example: Euler Method

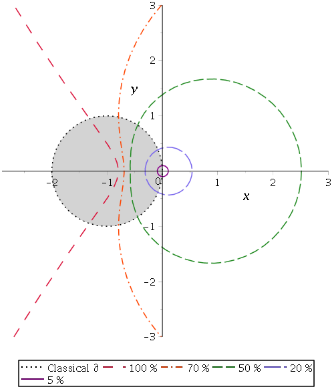
$$x_{n+1} = x_n + h\lambda x_n = (1 + h\lambda)x_n = (1 + \mu)x_n.$$

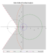
- Now, $R(\mu) \approx e^{\mu}$ so $\delta = \frac{\ln_k R(\mu)}{\mu} - 1 \approx 0$, but the *size* of δ tells us the relative backward error
- Note: In general $|\delta| = O(h^p)$ as $h \to 0$, so this is a nonlinear pseudospectral problem (plot contours of $|\delta|$ in the complex μ plane).

- If $|\delta| > 1$ then the problem we've solved is more than 100% different to the problem we wanted to solve.
- The curve $|\delta| = 1$ gives a qualitative upper limit: μ outside that region means the solution (decaying or not) is pretty lousy (probably worthless).
- If $|\delta| < 0.05$ then we've solved a problem within 5% of the one we wanted to (analogous to the 95% confidence limit!)
- If $|\delta| < \varepsilon$ (user's tolerance) then the solver has done its job [you win your bet!]

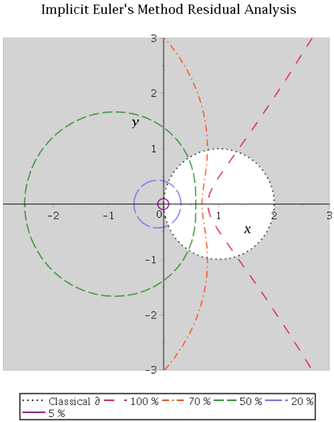
Euler Method

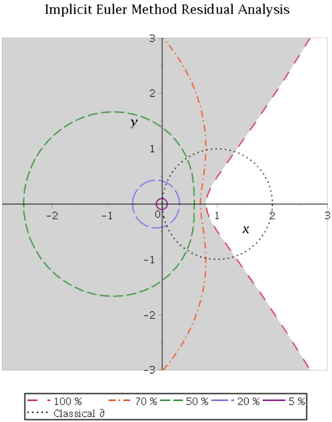
Euler's Method Residual Analysis



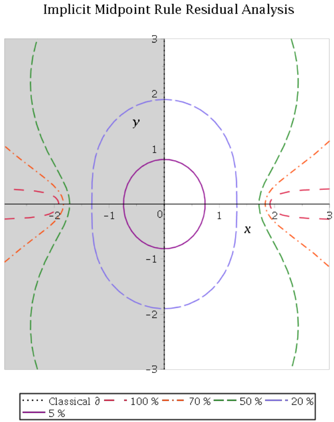


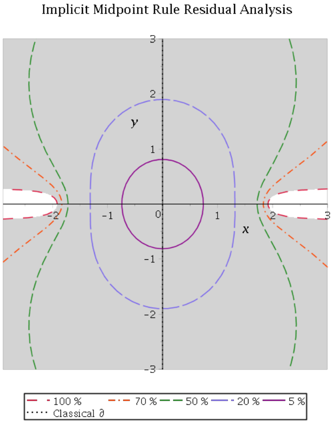
Implicit Euler Method





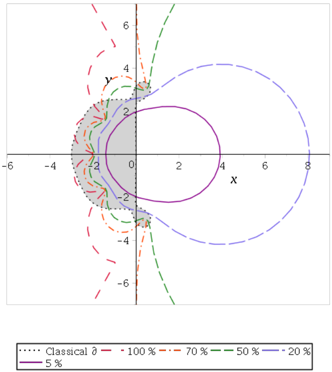
Implicit Midpoint Rule



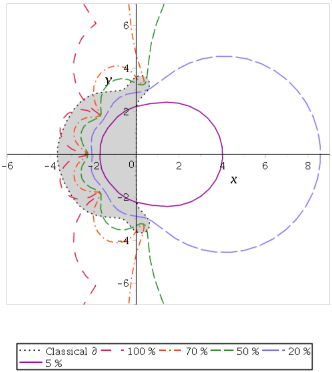


RKF45

Fourth Order Part RKF45 Method Residual Analysis

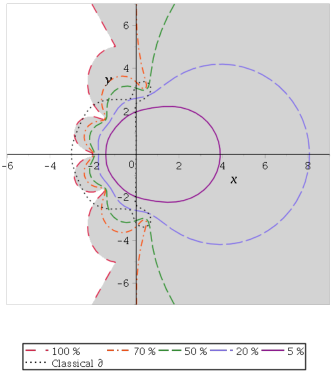


Fifth Order Part RKF45 Method Residual Analysis

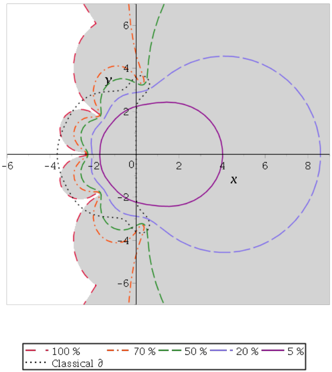


RKF45

Fourth Order Part RKF45 Method Residual Analysis

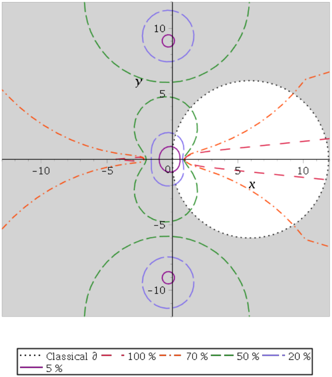


Fifth Order Part RKF45 Method Residual Analysis

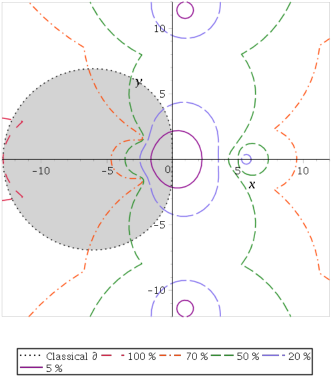


SDIRK 2-Stage 3rd Order

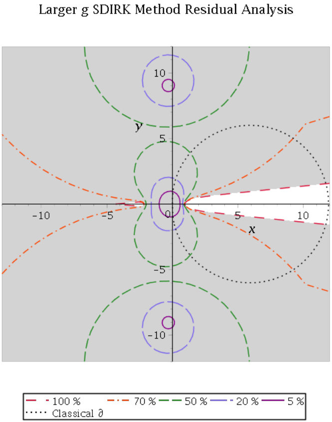
Larger g SDIRK Method Residual Analysis

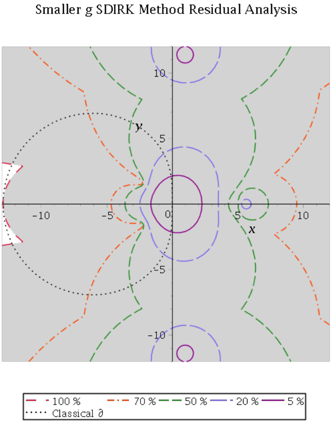


Smaller g SDIRK Method Residual Analysis

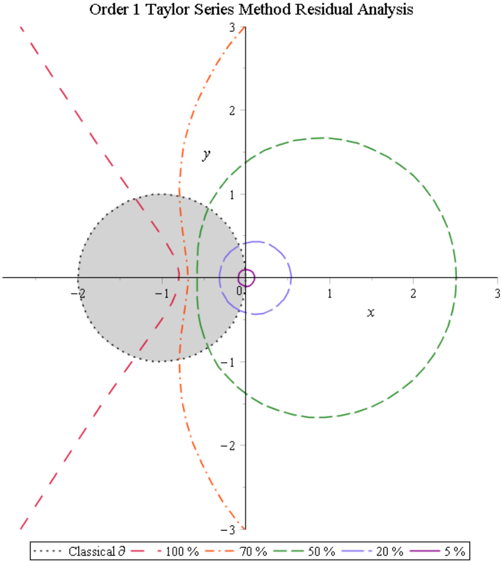


SDIRK 2-Stage 3rd Order

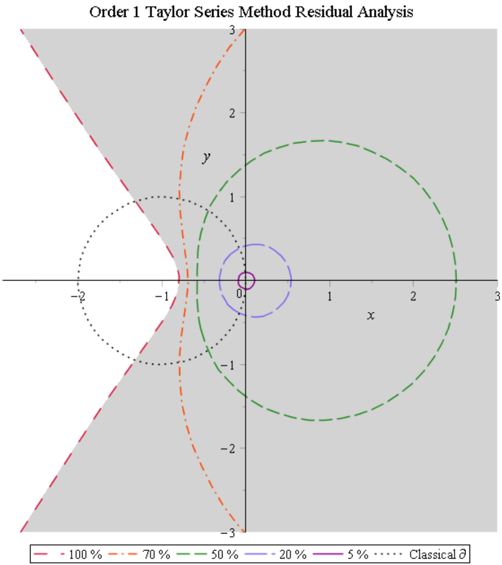




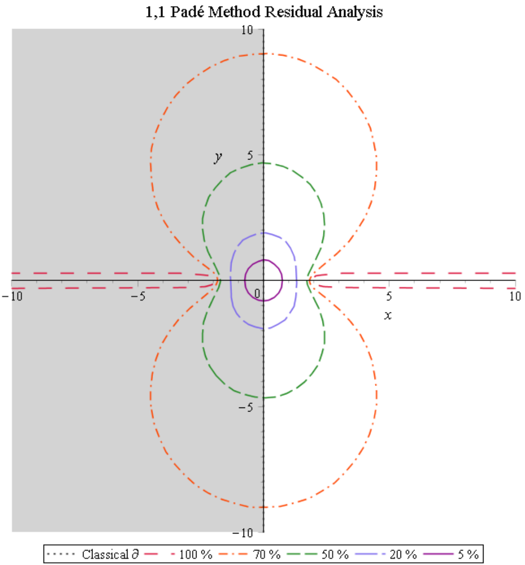
High Order Talyor Series Methods



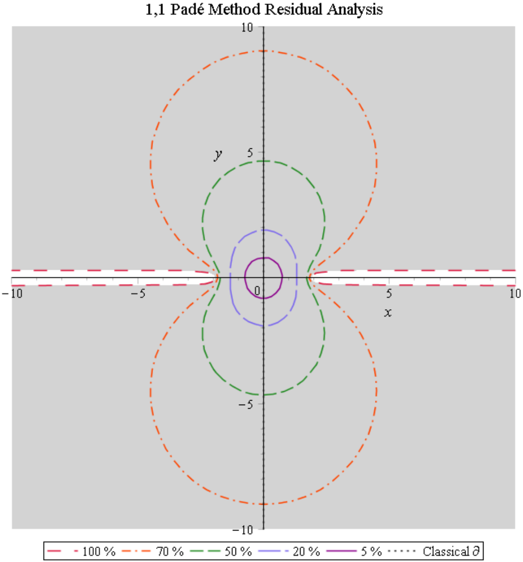
High Order Talyor Series Methods



High Order Diagonal Padé Methods



High Order Diagonal Padé Methods



Future work on this project will include:

- look at more methods, more pictures, more systems
- try to test the "preference change" predictions
- look at symplectic methods
- prove some things about the pictures

Calvo, MP, Murua, A, & Sanz-Serna, JM. 1994. Modified equations for ODEs. *Contemporary Mathematics*, **172**, 63–63.

Corless, Robert M. 1994. Error backward. *Contemporary Mathematics*, **172**, 31–31.

Enright, Wayne H. 1989. A new error-control for initial value solvers. *Applied Mathematics and Computation*, **31**, 288–301.

Enright, Wayne H, & Hayes, Wayne B. 2007. Robust and reliable defect control for Runge-Kutta methods. *ACM Transactions on Mathematical Software (TOMS)*, **33**(1), 1.

Grcar, Joseph F. 2011. John von Neumann's analysis of Gaussian elimination and the origins of modern Numerical Analysis. *SIAM review*, **53**(4), 607–682.

Griffiths, DF, & Sanz-Serna, JM. 1986. On the scope of the method of modified equations. *SIAM Journal on Scientific and Statistical Computing*, **7**(3), 994–1008.

Henrici, Peter. 1964. *Elements of numerical analysis*. Wiley.

Turing, Alan M. 1948. Rounding-off errors in matrix processes. *The Quarterly Journal of Mechanics and Applied Mathematics*, **1**(1), 287–308.

Von Neumann, John, & Goldstine, Herman H. 1947. Numerical inverting of matrices of high order. *Bulletin of the American Mathematical Society*, **53**(11), 1021–1099.

Warming, RF, & Hyett, BJ. 1974. The modified equation approach to the stability and accuracy analysis of finite-difference methods. *Journal of computational physics*, **14**(2), 159–179.

Wilkinson, James Hardy. 1963. *Rounding errors in algebraic processes*. Prentice-Hall Series in Automatic Computaton. Prentice-Hall.

Thank You!

Optimal Backward Error & the Dahlquist Test Problem

Robert M Corless[1]

Joint work with Yalçin Kaya[2] and Robert H C Moir[1]

[1]Department of Applied Mathematics
The University of Western Ontario
[2]Division of Information Technology, Engineering and the Environment
School of Information Technology and Mathematical Sciences
University of South Australia

2015 AMMCS-CAIMS Congress
Modified for McMaster University
October 8, 2015