A new backward error analysis for the matrix exponential based on pseudo-spectra

Marco Caliari*

University of Verona, Italy

Integrating the Integrators for Nonlinear Evolution Equations
2–7 December 2018, Banff (Alberta, CANADA)

*joint work with Franco Zivcovich

Exponential integrators

We consider the simplest exponential integrator for

$$u'(t) = Au(t) + g(u(t)), \quad u(0) = u_0$$

that is exponential Euler

$$u_{n+1} = u_n + h\varphi_1(hA)(Au_n + g(u_n))$$

where h is the time step and φ_1 is the entire function

$$\varphi_1(z) = \frac{\mathrm{e}^z - 1}{z}.$$

Exponential integrators

We consider the simplest exponential integrator for

$$u'(t) = Au(t) + g(u(t)), \quad u(0) = u_0$$

that is exponential Euler

$$u_{n+1} = u_n + h\varphi_1(hA)(Au_n + g(u_n))$$

where h is the time step and φ_1 is the entire function

$$\varphi_1(z) = \frac{\mathrm{e}^z - 1}{z}.$$

Given the augmented matrix

$$\tilde{A} = \begin{bmatrix} A & v \\ 0 & 0 \end{bmatrix}, \quad v = Au_n + g(u_n)$$

we have

$$\exp(h\tilde{A}) \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} h\varphi_1(hA)v \\ 1 \end{bmatrix}$$

Power series expansion of the backward error for $\exp(A)$

We formally approximate $\exp(A)$ as

$$p(s^{-1}A)^s = \exp(A + \Delta A) = \exp(A + sh(s^{-1}A))$$

where $p(z)$ is a polynomial of degree m (with $p(0) = 1$) and $h(z)$ has a power series expansion

$$h(z) = \log(\mathrm{e}^{-z}p(z)) = \sum_{k=\ell+1}^{\infty} c_k z^k$$

where ℓ is the largest integer such that $p^{(j)}(0) = 1$, $j = 0, 1, \ldots, \ell$.

Power series expansion of the backward error for $\exp(A)$

We formally approximate $\exp(A)$ as

$$p(s^{-1}A)^s = \exp(A + \Delta A) = \exp(A + sh(s^{-1}A))$$

where $p(z)$ is a polynomial of degree m (with $p(0) = 1$) and $h(z)$ has a power series expansion

$$h(z) = \log(\mathrm{e}^{-z} p(z)) = \sum_{k=\ell+1}^{\infty} c_k z^k$$

where ℓ is the largest integer such that $p^{(j)}(0) = 1$, $j = 0, 1, \ldots, \ell$. Therefore, $\|\Delta A\| \leq \mathrm{tol} \cdot \|A\|$ if

$$\frac{\|\Delta A\|}{\|A\|} = \frac{\|h(s^{-1}A)\|}{\|s^{-1}A\|} \leq \frac{\tilde{h}(s^{-1}\|A\|)}{s^{-1}\|A\|} \leq \mathrm{tol}$$

where $\tilde{h}(z) = \sum_{k=\ell+1}^{\infty} |c_k| z^k$.

Precomputation of the threshold

We can precompute in high precision the threshold θ such that

$$\frac{\tilde{h}(\theta)}{\theta} = \mathrm{tol}.$$

Then

$$\|\Delta A\| \leq \mathrm{tol} \cdot \|A\| \quad \text{if } s^{-1}\|A\| \leq \theta.$$

Precomputation of the threshold

We can precompute in high precision the threshold θ such that

$$\frac{\tilde{h}(\theta)}{\theta} = \mathrm{tol}.$$

Then

$$\|\Delta A\| \le \mathrm{tol} \cdot \|A\| \quad \text{if } s^{-1}\|A\| \le \theta.$$

Given

$$\alpha_q(A) = \max\{\|A^q\|^{1/q}, \|A^{q+1}\|^{1/(q+1)}\}$$

then

$$\|\Delta A\| \le \mathrm{tol} \cdot \|A\| \quad \text{if } s^{-1}\alpha_q(A) \le \theta \text{ and } q(q-1) \le \ell + 1$$

The sequence $\{\alpha_q(A)\}_q$ usually decreases for nonnormal matrices.

Precomputation of the threshold

We can precompute in high precision the threshold θ such that

$$\frac{\tilde{h}(\theta)}{\theta} = \mathrm{tol}.$$

Then

$$\|\Delta A\| \leq \mathrm{tol} \cdot \|A\| \quad \text{if } s^{-1}\|A\| \leq \theta.$$

Given

$$\alpha_q(A) = \max\{\|A^q\|^{1/q}, \|A^{q+1}\|^{1/(q+1)}\}$$

then

$$\|\Delta A\| \leq \mathrm{tol} \cdot \|A\| \quad \text{if } s^{-1}\alpha_q(A) \leq \theta \text{ and } q(q-1) \leq \ell + 1$$

The sequence $\{\alpha_q(A)\}_q$ usually decreases for nonnormal matrices.
Usually we work with shifted matrices $B = A - \mu I$.

Families of polynomial approximations

Instead of a single polynomial of degree m, we can consider sequences $\{p_m\}_m$. For instance

- truncated Taylor series $p_m(z) = \sum_{i=0}^{m} z^i/i!$ [Al-Mohy–Higham, 2011]

Families of polynomial approximations

Instead of a single polynomial of degree m, we can consider sequences $\{p_m\}_m$. For instance

- truncated Taylor series $p_m(z) = \sum_{i=0}^{m} z^i/i!$ [Al-Mohy–Higham, 2011]
- Interpolation $p_m(z) = \sum_{i=0}^{m} \mathrm{e}^{[z_0,z_1,\ldots,z_i]} \prod_{j=0}^{i-1}(z - z_j)$ at Leja–Hermite points [C., Kandolf, Ostermann, Rainer, Zivcovich 2016–2018]

$$z_0 = z_1 = \ldots = z_\ell = 0,$$

$$z_{i+1} \in \arg \max_{x \in [-c,c]} \prod_{j=0}^{i} |x - z_j| \quad i = \ell, \ell+1, \ldots, m-1$$

For each m, c can be chosen in order to maximize θ.

More information from the spectrum of A

The field of values $\mathcal{W}(A)$ satisfies

$$\mathcal{W}(A) = \mathcal{W}(A_{\mathrm{H}} + A_{\mathrm{SH}}) \subseteq \mathcal{W}(A_{\mathrm{H}}) + \mathcal{W}(A_{\mathrm{SH}}) = \\ \mathrm{conv}(\sigma(A_{\mathrm{H}})) + \mathrm{conv}(\sigma(A_{\mathrm{SH}})) \subseteq [\alpha, \nu] + \mathrm{i}[\eta, \beta]$$

We use Gershgorin's disks to obtain the rectangle $[\alpha, \nu] + \mathrm{i}[\eta, \beta]$.

More information from the spectrum of A

The field of values $\mathcal{W}(A)$ satisfies

$$\begin{aligned}\mathcal{W}(A) = \mathcal{W}(A_{\mathrm{H}} + A_{\mathrm{SH}}) \subseteq \mathcal{W}(A_{\mathrm{H}}) + \mathcal{W}(A_{\mathrm{SH}}) = \\ \mathrm{conv}(\sigma(A_{\mathrm{H}})) + \mathrm{conv}(\sigma(A_{\mathrm{SH}})) \subseteq [\alpha, \nu] + \mathrm{i}[\eta, \beta]\end{aligned}$$

We use Gershgorin's disks to obtain the rectangle $[\alpha, \nu] + \mathrm{i}[\eta, \beta]$. After applying the obvious shift μ, with abuse of notation, we get

$$\mathcal{W}(A) \subseteq R(A) = [-\nu, \nu] + \mathrm{i}[-\beta, \beta]$$

More information from the spectrum of A

The field of values $\mathcal{W}(A)$ satisfies

$$\mathcal{W}(A) = \mathcal{W}(A_{\mathrm{H}} + A_{\mathrm{SH}}) \subseteq \mathcal{W}(A_{\mathrm{H}}) + \mathcal{W}(A_{\mathrm{SH}}) = \\ \mathrm{conv}(\sigma(A_{\mathrm{H}})) + \mathrm{conv}(\sigma(A_{\mathrm{SH}})) \subseteq [\alpha, \nu] + \mathrm{i}[\eta, \beta]$$

We use Gershgorin's disks to obtain the rectangle $[\alpha, \nu] + \mathrm{i}[\eta, \beta]$.
After applying the obvious shift μ, with abuse of notation, we get

$$\mathcal{W}(A) \subseteq R(A) = [-\nu, \nu] + \mathrm{i}[-\beta, \beta]$$

and

$$\Lambda_\varepsilon(A) \subseteq \mathcal{W}(A) + \Delta_\varepsilon \subseteq R(A) + \Delta_\varepsilon$$

where $\Lambda_\varepsilon(A) = \{z \in \mathbb{C} \colon \|(zI - A)^{-1}\|_2 \geq \varepsilon^{-1}\}$ is the ε-pseudo-spectrum of A and $\Delta_\varepsilon = \{z \in \mathbb{C} \colon |z| \leq \varepsilon\}$.

Contour integral expansion of the backward error

$\Lambda_\varepsilon(A)$ does not scale with A: we consider instead

$$\Lambda_{\delta\|tA\|_2}(tA) \subseteq \mathcal{W}(tA) + \Delta_{\delta\|tA\|_2} \subseteq R(tA) + \Delta_{\delta\|tA\|_2} = t(R(A) + \Delta_{\delta\|A\|_2}) \subseteq tR_\delta(A)$$

where $R_\delta(A)$ is the extented rectangle

$$R_\delta(A) = [-\nu - \delta\|A\|_2, \nu + \delta\|A\|_2] + \mathrm{i}[-\beta - \delta\|A\|_2, \beta + \delta\|A\|_2].$$

Contour integral expansion of the backward error

$\Lambda_\varepsilon(A)$ does not scale with A: we consider instead

$$\Lambda_{\delta\|tA\|_2}(tA) \subseteq \mathcal{W}(tA) + \Delta_{\delta\|tA\|_2} \subseteq R(tA) + \Delta_{\delta\|tA\|_2} = t(R(A) + \Delta_{\delta\|A\|_2}) \subseteq tR_\delta(A)$$

where $R_\delta(A)$ is the extented rectangle

$$R_\delta(A) = [-\nu - \delta\|A\|_2, \nu + \delta\|A\|_2] + \mathrm{i}[-\beta - \delta\|A\|_2, \beta + \delta\|A\|_2].$$

Then

$$\frac{\|\Delta A\|_2}{\|A\|_2} = \frac{\|h(s^{-1}A)\|_2}{\|s^{-1}A\|_2} \leq \|s^{-1}A\|_2 \|g(s^{-1}A)\|_2 = \|s^{-1}A\|_2 \left\| \frac{1}{2\pi\mathrm{i}} \int_\Gamma g(z)(zI - s^{-1}A)^{-1} \mathrm{d}z \right\|_2 \leq \frac{\mathcal{L}(\Gamma)}{2\pi\delta} \|g\|_\Gamma,$$

if $h(z) = z^2 g(z)$ ($\ell \geq 1$) and $\Gamma = \partial K$ encloses $\Lambda_{\delta\|s^{-1}A\|_2}(s^{-1}A)$.

Contour integral expansion of the backward error

$\Lambda_\varepsilon(A)$ does not scale with A: we consider instead

$$\Lambda_{\delta\|tA\|_2}(tA) \subseteq \mathcal{W}(tA) + \Delta_{\delta\|tA\|_2} \subseteq R(tA) + \Delta_{\delta\|tA\|_2} = t(R(A) + \Delta_{\delta\|A\|_2}) \subseteq tR_\delta(A)$$

where $R_\delta(A)$ is the extented rectangle

$$R_\delta(A) = [-\nu - \delta\|A\|_2, \nu + \delta\|A\|_2] + \mathrm{i}[-\beta - \delta\|A\|_2, \beta + \delta\|A\|_2].$$

Then

$$\frac{\|\Delta A\|_2}{\|A\|_2} = \frac{\|h(s^{-1}A)\|_2}{\|s^{-1}A\|_2} \leq \|s^{-1}A\|_2 \|g(s^{-1}A)\|_2 = \|s^{-1}A\|_2 \left\| \frac{1}{2\pi\mathrm{i}} \int_\Gamma g(z)(zI - s^{-1}A)^{-1} \mathrm{d}z \right\|_2 \leq \frac{\mathcal{L}(\Gamma)}{2\pi\delta} \|g\|_\Gamma,$$

if $h(z) = z^2 g(z)$ ($\ell \geq 1$) and $\Gamma = \partial K$ encloses $\Lambda_{\delta\|s^{-1}A\|_2}(s^{-1}A)$.
This is true if $s^{-1}R_\delta(A) \subseteq K$.

Choice of K

For given c and δ, we consider the ellipse Γ_γ of foci $(\pm c, 0)$ and capacity (half sum of the semi-axes) γ. We look for γ_δ such that

$$\frac{\|\Delta A\|_2}{\|A\|_2} \leq \ldots \leq \frac{\mathcal{L}(\Gamma_{\gamma_\delta})}{2\pi\delta}\|g\|_{\Gamma_{\gamma_\delta}} = \mathrm{tol}$$

where g is associated to a given polynomial $p_m\colon [-c, c] \to \mathbb{R}$.

Choice of K

For given c and δ, we consider the ellipse Γ_γ of foci $(\pm c, 0)$ and capacity (half sum of the semi-axes) γ. We look for γ_δ such that

$$\frac{\|\Delta A\|_2}{\|A\|_2} \leq \ldots \leq \frac{\mathcal{L}(\Gamma_{\gamma_\delta})}{2\pi\delta}\|g\|_{\Gamma_{\gamma_\delta}} = \mathrm{tol}$$

where g is associated to a given polynomial $p_m \colon [-c, c] \to \mathbb{R}$.

- For a given (shifted) matrix A, compute the rectangle $R_\delta(A) = [-\nu - \delta\|A\|_2, \nu + \delta\|A\|_2] + \mathrm{i}[-\beta - \delta\|A\|_2, \beta + \delta\|A\|_2]$

Choice of K

For given c and δ, we consider the ellipse Γ_γ of foci $(\pm c, 0)$ and capacity (half sum of the semi-axes) γ. We look for γ_δ such that

$$\frac{\|\Delta A\|_2}{\|A\|_2} \leq \ldots \leq \frac{\mathcal{L}(\Gamma_{\gamma_\delta})}{2\pi\delta}\|g\|_{\Gamma_{\gamma_\delta}} = \mathrm{tol}$$

where g is associated to a given polynomial $p_m \colon [-c, c] \to \mathbb{R}$.

- For a given (shifted) matrix A, compute the rectangle $R_\delta(A) = [-\nu - \delta\|A\|_2, \nu + \delta\|A\|_2] + \mathrm{i}[-\beta - \delta\|A\|_2, \beta + \delta\|A\|_2]$
- compute s as the smallest integer such that $s^{-1}R_\delta(A) \subseteq K_{\gamma_\delta}$

$$\frac{(\nu + \delta\|A\|_2)^2}{s^2 a_\delta^2} + \frac{(\beta + \delta\|A\|_2)^2}{s^2 b_\delta^2} \leq 1$$

Choice of K

For given c and δ, we consider the ellipse Γ_γ of foci $(\pm c, 0)$ and capacity (half sum of the semi-axes) γ. We look for γ_δ such that

$$\frac{\|\Delta A\|_2}{\|A\|_2} \leq \ldots \leq \frac{\mathcal{L}(\Gamma_{\gamma_\delta})}{2\pi\delta}\|g\|_{\Gamma_{\gamma_\delta}} = \mathrm{tol}$$

where g is associated to a given polynomial $p_m \colon [-c, c] \to \mathbb{R}$.

- For a given (shifted) matrix A, compute the rectangle $R_\delta(A) = [-\nu - \delta\|A\|_2, \nu + \delta\|A\|_2] + \mathrm{i}[-\beta - \delta\|A\|_2, \beta + \delta\|A\|_2]$
- compute s as the smallest integer such that $s^{-1}R_\delta(A) \subseteq K_{\gamma_\delta}$

$$\frac{(\nu + \delta\|A\|_2)^2}{s^2 a_\delta^2} + \frac{(\beta + \delta\|A\|_2)^2}{s^2 b_\delta^2} \leq 1$$

- approximate $\exp(A)v$ as $\underbrace{p_m(s^{-1}A)(\ldots(p_m(s^{-1}A)}_{s \text{ times}}\, v)\ldots)$

Details

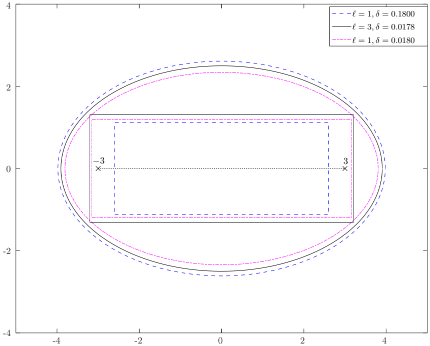
- We used Leja–Hermite interpolation polynomials with $\ell \geq 1$

Details

- We used Leja–Hermite interpolation polynomials with $\ell \geq 1$
- for given intervals $[-c, c]$ and $\mathrm{i}[-c, c]$ and given degrees m up to 55, we computed the corresponding ellipses Γ_{γ_δ}

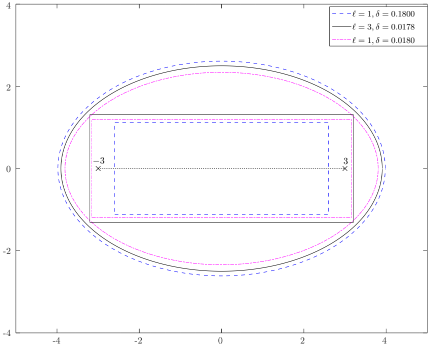
Details

- We used Leja–Hermite interpolation polynomials with $\ell \geq 1$
- for given intervals $[-c, c]$ and $\mathrm{i}[-c, c]$ and given degrees m up to 55, we computed the corresponding ellipses Γ_{γ_δ}
- we optimized over δ and ℓ



Details

- We used Leja–Hermite interpolation polynomials with $\ell \geq 1$
- for given intervals $[-c, c]$ and $\mathrm{i}[-c, c]$ and given degrees m up to 55, we computed the corresponding ellipses Γ_{γ_δ}
- we optimized over δ and ℓ



- given the matrix A, we minimize $s \cdot m$ (matrix-vector cost)

Numerical results: 1

A is a 2D diffusion matrix, size 2041×2041, $\|A^q\|_1^{1/q} = 100$

Method	s	m	c	θ or γ	ℓ	$s \cdot m$	act. its.	rel. err.
Taylor	11	53	0	9.3	53	583	495	4.4e-14
L–H p.s.	10	55	4.8	1.0e1	0	550	460	3.3e-14
L–H c.i.	8	51	1.3e1	7.1	15	408	268	1.3e-14

The number of actual iterations is smaller than $s \cdot m$ because of an early termination criterion

$$\text{if } \left\| \frac{A^k}{k!} v^{(l)} \right\| \leq \mathrm{tol} \cdot \left\| \sum_{i=0}^{k} \frac{A^i}{i!} v^{(l)} \right\| \quad \text{for } k < m \text{ and } 0 \leq l \leq s-1$$

then stop substep l

Numerical results: 2

A is a 1D Schrödinger matrix, size 69×69, $\|A^q\|_1^{1/q} = 2450$

Method	s	m	c	θ or γ	ℓ	$s \cdot m$	act. its.	rel. err.
Taylor	249	55	0	9.9	55	13695	13197	7.3e-11
L–H p.s.	292	55	8.4	8.4	1	16060	10220	2.7e-13
L–H c.i.	186	54	1.3e1	7.9	42	10044	9858	1.7e-13

There is a hump phenomenon for Taylor series approximation. We mean that

$$\left\| \sum_{i=0}^{k} \frac{A^i}{i!} v^{(l)} \right\| \gg \left\| \sum_{i=0}^{m} \frac{A^i}{i!} v^{(l)} \right\| \quad \text{for } k < m \text{ and } 0 \leq l \leq s-1$$

and cancellation takes place.

Numerical results: 3

A is `triu(-4*ones(20),1)` (nilpotent), v is `cos((1:20)')`,
$\|A\|_1 = 76$, $\alpha_8(A) = 16.29$, $\lim_{q\to\infty} \alpha_q(A) = \rho(A) = 0$

Method	s	m	c	θ or γ	ℓ	$s \cdot m$	act. its.	rel. err.
Taylor	2	54	0	9.6	54	108	42	3.2e-14
L–H p.s.	2	53	6.7	9.6	41	106	42	4.2e-14
L–H c.i.	6	55	5.5	9.2	2	330	186	2.0e-14

Since it is not possible to use the values $\alpha_q(A)$ for L–H c.i., there is overscaling.

Numerical results: 3

A is `triu(-4*ones(20),1)` (nilpotent), v is `cos((1:20)')`,
$\|A\|_1 = 76$, $\alpha_8(A) = 16.29$, $\lim_{q\to\infty} \alpha_q(A) = \rho(A) = 0$

Method	s	m	c	θ or γ	ℓ	$s \cdot m$	act. its.	rel. err.
Taylor	2	54	0	9.6	54	108	42	3.2e-14
L–H p.s.	2	53	6.7	9.6	41	106	42	4.2e-14
L–H c.i.	6	55	5.5	9.2	2	330	186	2.0e-14

Since it is not possible to use the values $\alpha_q(A)$ for L–H c.i., there is overscaling.
This is the famous `triw` example by [Al-Mohy–Higham, 2011] for which Krylov and rational methods may suffer of loss of accuracy.

Numerical results: 4

A is `triu(-4*ones(110),1)` (nilpotent), v is `ones(110,1)`, $\|A\|_1 = 436$, $\alpha_8(A) = 112.08$

Method	s	m	c	θ or γ	ℓ	$s \cdot m$	act. its.	rel. err.
Taylor	12	55	0	9.9	55	660	313	3.2e-12
L–H c.i.	34	55	0	9.2	55	1870	635	2.2e-14

In this case, we have that L–H c.i. is Taylor, but the abuse of the values $\alpha_q(A)$ makes Taylor to underscale.

Conclusions

- A mixture of power series and contour integral expansions would probably be optimal

Conclusions

- A mixture of power series and contour integral expansions would probably be optimal
- The backward error analysis can be applied to any polynomial method (Taylor truncated series, interpolation, Chebyshev series, . . . , Krylov)

Conclusions

- A mixture of power series and contour integral expansions would probably be optimal
- The backward error analysis can be applied to any polynomial method (Taylor truncated series, interpolation, Chebyshev series, . . . , Krylov)
- It should be possible to perform the backward error analysis on-the-fly [C. and Zivcovich, 2018]

Conclusions

- A mixture of power series and contour integral expansions would probably be optimal
- The backward error analysis can be applied to any polynomial method (Taylor truncated series, interpolation, Chebyshev series, . . . , Krylov)
- It should be possible to perform the backward error analysis on-the-fly [C. and Zivcovich, 2018]
- matrix-free?

Conclusions

- A mixture of power series and contour integral expansions would probably be optimal
- The backward error analysis can be applied to any polynomial method (Taylor truncated series, interpolation, Chebyshev series, ..., Krylov)
- It should be possible to perform the backward error analysis on-the-fly [C. and Zivcovich, 2018]
- matrix-free?
- Thanks for your attention