Field of value error bounds for approximating matrix functions via rational Arnoldi

Bernhard Beckermann

http://math.univ-lille1.fr/~bbecker

Laboratoire Paul Painlevé UMR 8524 (équipe ANO-EDP)
UFR Mathématiques, Université de Lille 1

Structured Matrices, Cortona, September 2008

Joint work with
Lothar Reichel, Kent State University

Outline

The problem: approximating $g(A)b$ for large sparse $A \in \mathbb{R}^{d\times d}$, $b \in \mathbb{R}^d$,
and, e.g., $g(z) = \exp(\tau z)$, $g(z) = \sqrt{z}$, $g(z) = \log(z)$

The method: (rational) Arnoldi

The aim: "simple" a priori error estimates in terms of
field of values $W(A) = \{y^* Ay : \|y\| = 1\}$

The question: link with best polynomial/rational approx. of g on $W(A)$?

The aim: Simple sharp explicit upper bounds

Outline

The problem: approximating $g(A)b$ for large sparse $A \in \mathbb{R}^{d\times d}$, $b \in \mathbb{R}^d$,
and, e.g., $g(z) = \exp(\tau z)$, $g(z) = \sqrt{z}$, $g(z) = \log(z)$

The method: (rational) Arnoldi

The aim: "simple" a priori error estimates in terms of
field of values $W(A) = \{y^* A y : \|y\| = 1\}$

The question: link with best polynomial/rational approx. of g on $W(A)$?

The aim: Simple sharp explicit upper bounds

The problem and some applications

How to approximately compute $g(A)b$, where

$$\|b\| = 1, \quad A \in \mathbb{R}^{d\times d} \text{ large, sparse, non-symmetric...?}$$

Applications for $g(z) = e^z$, $g(z) = \cos(z)$, $g(z) = \sin(z)$: semi-discretized PDEs or ODEs.

Applications for $g(z) = 1/\sqrt{z}$: splitting techniques in implicit schemes, stochastic differential equations.

Applications for $g(z) = \log(z)$, $g(z) = \tanh(z)$, ...

...recent SIAM book of Nick Higham.

Approximating via the Arnoldi method

We compute ONB $V_m = (v_1, ..., v_m) \in \mathbb{C}^{d\times m}$ of $\text{span}(b, Ab, ..., A^{m-1}b)$ via

$$v_1 = b, \qquad h_{j+1,j}v_{j+1} = Av_j - h_{1,j}v_1 - ... - h_{j,j}v_j,$$

leading to Arnoldi decomposition

$$AV_m = V_m H_m + (0, ..., 0, h_{m+1,m}v_{m+1}), \quad H_m = V_m^* A V_m \text{ upper Hessenberg.}$$

Approximation via Arnoldi:

- compute Arnoldi decomposition V_m, H_m for "small" m
- compute exactly $g(H_m)$
- approach $g(A)b$ by $V_m g(H_m) e_1$.

Error estimate? For each polynomial p of degree $< m$ we have $p(A)b = p(A)V_m e_1 = V_m p(H_m) e_1$, and thus

$$\begin{aligned} \|g(A)b - V_m g(H_m)e_1\| &= \|(g-p)(A)b - V_m(g-p)(H_m)e_1\| \\ &\leq \|(g-p)(A)\| + \|(g-p)(H_m)\|. \end{aligned}$$

Approximating via rational Arnoldi

Consider some fixed denominator polynomial $q(z) = (1 - z/z_1)...(1 - z/z_m)$.

Rough idea (following Ruhe):

we compute ONB $V_{m+1} = (v_1, ..., v_{m+1})$ of

$$q(A)^{-1}\text{span}(b, Ab, ..., A^m b) = \text{span}(b, (z_1 I - A)^{-1} b, ..., (z_m I - A)^{-1} b)$$

and then project:

$$A_{m+1} = V_{m+1}^* A V_{m+1}.$$

The rational Arnoldi approximation with fixed denominator q of $g(A)b$

$$V_{m+1} g(A_{m+1}) V_{m+1}^* b = V_{m+1} g(A_{m+1}) e_1$$

gives the exact answer for any rational function $g = p/q$ with $\deg p \leq m$.
Also, we recover ordinary Arnoldi with $A_{m+1} = H_{m+1}$ for
$z_1 = ... = z_m = \infty$.

Rational Arnoldi: some more details

For fixed denominator polynomial $q(z) = (1 - z/z_1)...(1 - z/z_m)$, take z_0 far from the other z_j, and put $z_{m+1} = \infty$.

We compute ONB $V_{m+1} = (v_1, ..., v_{m+1})$ of $q(A)^{-1}\text{span}(b, Ab, ..., A^m b)$ via

$$v_1 = b, \qquad h_{j+1,j}v_{j+1} = (z_j I - A)^{-1}(z_j - z_0)(A - z_0 I)v_j - h_{1,j}v_1 - ... - h_{j,j}v_j,$$

leading to a more complicated Arnoldi decomposition: with $D_{m+1} = \text{diag}\,(\frac{1}{z_1 - z_0}, ..., \frac{1}{z_{m+1} - z_0})$, and $H_{m+1} = (h_{j,k})_{j,k=1,...,m+1}$ upper Hessenberg, we get after some computations

$$(A - z_0 I)V_{m+1}(I + H_{m+1}D_{m+1}) = V_{m+1}H_{m+1} + (0, ..., 0, h_{m+2,m+1}v_{m+2}),$$

implying that $A_{m+1} := V_{m+1}^* A V_{m+1} = z_0 I + H_{m+1}(I + H_{m+1}D_{m+1})^{-1}$.

Structured matrices are helpful for solving shifted systems.
If no structure, take repeated poles ($\Longrightarrow$ few LU decompositions)

How to get error estimates?

In both cases the error is governed by $\|g(A) - p(A)\|$ or $\|g(A) - \frac{p}{q}(A)\|$ for arbitrary polynomials p of degree $\leq m$.

Link with

$$\eta_m^q(g, \mathbb{E}) = \min_{\deg p \leq m} \|g - \frac{p}{q}\|_{L_\infty(\mathbb{E})} \quad ???$$

For normal matrices: take the maximum on the (convex hull of) the spectrum, but in the general case?

Crouzeix 2006: There exists a universal constant $C \in [2, 11.5]$ such that for any matrix $B \in \mathbb{C}^{d \times d}$ and for any function f analytic in the field of values

$$W(B) = \{y^* B y : y \in \mathbb{C}^d, \|y\| = 1\},$$

there holds

$$\|f(B)\| \leq C \, \|f\|_{L_\infty(W(B))}.$$

In what follows let $\mathbb{E} \subset \mathbb{C}$ be some convex and compact set symmetric with respect to the real axis and containing the field of values.

How to get error estimates?

In both cases the error is governed by $\|g(A) - p(A)\|$ or $\|g(A) - \frac{p}{q}(A)\|$ for arbitrary polynomials p of degree $\leq m$.

Link with

$$\eta_m^q(g, \mathbb{E}) = \min_{\deg p \leq m} \|g - \frac{p}{q}\|_{L_\infty(\mathbb{E})} \quad ???$$

For normal matrices: take the maximum on the (convex hull of) the spectrum, but in the general case?

Crouzeix 2006: There exists a universal constant $C \in [2, 11.5]$ such that for any matrix $B \in \mathbb{C}^{d \times d}$ and for any function f analytic in the field of values

$$W(B) = \{y^* B y : y \in \mathbb{C}^d, \|y\| = 1\},$$

there holds

$$\|f(B)\| \leq C \, \|f\|_{L_\infty(W(B))}.$$

In what follows let $\mathbb{E} \subset \mathbb{C}$ be some convex and compact set symmetric with respect to the real axis and containing the field of values.

How to get error estimates?

In both cases the error is governed by $\|g(A) - p(A)\|$ or $\|g(A) - \frac{p}{q}(A)\|$ for arbitrary polynomials p of degree $\leq m$.

Link with

$$\eta_m^q(g, \mathbb{E}) = \min_{\deg p \leq m} \|g - \frac{p}{q}\|_{L_\infty(\mathbb{E})} \quad ???$$

For normal matrices: take the maximum on the (convex hull of) the spectrum, but in the general case?

Crouzeix 2006: There exists a universal constant $C \in [2, 11.5]$ such that for any matrix $B \in \mathbb{C}^{d\times d}$ and for any function f analytic in the field of values

$$W(B) = \{y^* B y : y \in \mathbb{C}^d, \|y\| = 1\},$$

there holds

$$\|f(B)\| \leq C \, \|f\|_{L_\infty(W(B))}.$$

In what follows let $\mathbb{E} \subset \mathbb{C}$ be some convex and compact set symmetric with respect to the real axis and containing the field of values.

Riemann maps and Faber operators

Let $\mathbb{E}$ convex, compact as before, $\mathbb{D}$ closed unit disc, then there exists unique conformal map $\phi : \mathbb{E}^c \mapsto \mathbb{D}^c$ with $\phi(\infty) = \infty$, $\phi'(\infty) > 0$, $\psi := \phi^{-1}$.

Faber operator: bijection between G analytic in $\mathbb{D}$ and g analytic in $\mathbb{E}$

$$z \in \mathrm{Int}(\mathbb{E}) : \qquad g(z) = \mathcal{F}(G)(z) = \frac{1}{2\pi i} \int_{|w|=1} \frac{\psi'(w)}{\psi(w) - z} G(w)\, dw,$$

$$w \in \mathrm{Int}(\mathbb{D}) : \qquad G(w) = \mathcal{F}^{-1}(g)(w) = \frac{1}{2\pi i} \int_{|\zeta|=1} \frac{g(\psi(\zeta))}{\zeta - w}\, d\zeta.$$

Faber polynomial: $F_n(z) = \mathcal{F}(w^n)(z)$ polynomial of degree n, namely F_n polynomial part of ϕ^n.

$G(w) = \frac{a}{w - w_1}$ with $|w_1| > 1$, $\mathcal{F}(G)(z) = \frac{a\psi'(w_1)}{z - \psi(w_1)}$, similar for multiple poles.

Hence with $Q(w) = \prod_j (w - w_j)$, $q(z) = \prod_j (z - z_j)$, $z_j = \psi(w_j)$ (Ellacott '83):

$$g = \mathcal{F}(G) \quad \Longrightarrow \quad \frac{1}{\|\mathcal{F}^{-1}\|}\, \eta_m^Q(G, \mathbb{D}) \le \eta_m^q(g, \mathbb{E}) \le 2\, \eta_m^Q(G, \mathbb{D}).$$

Our a priori bound

THEOREM 1: Let $\mathbb{E}$ as before containing the field of values $W(A)$ and let $g = \mathcal{F}(G)$ be analytic on $\mathbb{E}$, then for the rational q-Arnoldi method

$$\|g(A)b - V_{m+1}g(A_{m+1})e_1\| \leq 4\,\eta_m^Q(G, \mathbb{D})$$

(put $q = Q = 1$ for classical Arnoldi).

Our proof is inspired from Crouzeix, Delyon, Badea, BB 02-07, in particular the CRAS '05 of BB: $\|F_n(A)\| \leq 2$.

Also, we use $W(A_{m+1}) \subset W(A)$.

Previous work of Eiermann (1993), Greenbaum (1997), ...

Idea of proof of Theorem 1

It is sufficient to show

$$\|h(A)\| \leq 2\|H\|_{L_\infty(\mathbb{D})}, \quad h = \mathcal{F}(H) + H(0).$$

Here $W(A) \subset \text{Int}(\mathbb{E})$ for simplicity. We have

$$\mathcal{F}(w^m)(A) = \frac{1}{2\pi i}\int_{|w|=1} w^m \psi'(w)(\psi(w) - A)^{-1} dw = \begin{cases} F_m(A) & \text{if } m = 0, 1, 2, ..., \\ 0 & \text{if } m = -1, -2, ... \end{cases}$$

Hence

$$h(A) = \frac{1}{2\pi}\int_{|w|=1} H(w) \left(\underbrace{\Big(w\psi'(w)(\psi(w) - A)^{-1}\Big) + \Big(w\psi'(w)(\psi(w) - A)^{-1}\Big)^*}_{\text{positive definite}} \right) \frac{dw}{iw}.$$

Arnoldi $\Longrightarrow$ best polynomial approximation on $\mathbb{D}$

The Faber coefficients are given by

$$g_j = \frac{1}{2\pi i} \int_{|w|=1} \frac{g(\psi(w))}{w^{j+1}} \, dw \quad \Longrightarrow \quad g = \mathcal{F}(G), \quad G(w) = \sum_{j=0}^{\infty} g_j w^j,$$

In the polynomial case $q = Q = 1$, we have

LEMMA 2:

$$|g_m| \le \eta^1_{m-1}(G, \mathbb{D}) \le \sum_{j=0}^{\infty} |g_{m+j}|.$$

Knizhnerman '91 gave a similar upper bound with additional powers of $m + j$

Hochbruck & Lubich '97 gave a more complicated bound, weaker up to factor 0.75.

Example: $g(z) = \exp(\tau z)$, $\tau > 0$: if $m \ge \tau \text{cap}(\mathbb{E})$ then both $|g_m|$ and $\eta_m(G, \mathbb{D})$ can be bounded above and below by constants times $(\tau \text{cap}(\mathbb{E}))^m/(m!)$ (as in the disk case).

In general, if G is analytic in $|w| \le R$ then $|g_m| \le R^{-m} \|G\|_{L_\infty(\{|w| \le R\})}$.

Arnoldi $\Longrightarrow$ best polynomial approximation on $\mathbb{D}$

The Faber coefficients are given by

$$g_j = \frac{1}{2\pi i} \int_{|w|=1} \frac{g(\psi(w))}{w^{j+1}}\, dw \quad \Longrightarrow \quad g = \mathcal{F}(G), \quad G(w) = \sum_{j=0}^{\infty} g_j w^j,$$

In the polynomial case $q = Q = 1$, we have

LEMMA 2:

$$|g_m| \le \eta^1_{m-1}(G, \mathbb{D}) \le \sum_{j=0}^{\infty} |g_{m+j}|.$$

Knizhnerman '91 gave a similar upper bound with additional powers of $m + j$

Hochbruck & Lubich '97 gave a more complicated bound, weaker up to factor 0.75.

Example: $g(z) = \exp(\tau z)$, $\tau > 0$: if $m \ge \tau \text{cap}(\mathbb{E})$ then both $|g_m|$ and $\eta_m(G, \mathbb{D})$ can be bounded above and below by constants times $(\tau \text{cap}(\mathbb{E}))^m/(m!)$ (as in the disk case).

In general, if G is analytic in $|w| \le R$ then $|g_m| \le R^{-m} \|G\|_{L_\infty(\{|w| \le R\})}$.

Arnoldi $\Longrightarrow$ best polynomial approximation on $\mathbb{D}$

The Faber coefficients are given by

$$g_j = \frac{1}{2\pi i}\int_{|w|=1} \frac{g(\psi(w))}{w^{j+1}}\,dw \quad \Longrightarrow \quad g = \mathcal{F}(G), \quad G(w) = \sum_{j=0}^{\infty} g_j w^j,$$

In the polynomial case $q = Q = 1$, we have

LEMMA 2:

$$|g_m| \leq \eta^1_{m-1}(G,\mathbb{D}) \leq \sum_{j=0}^{\infty} |g_{m+j}|.$$

Knizhnerman '91 gave a similar upper bound with additional powers of $m + j$

Hochbruck & Lubich '97 gave a more complicated bound, weaker up to factor 0.75.

Example: $g(z) = \exp(\tau z)$, $\tau > 0$: if $m \geq \tau\text{cap}(\mathbb{E})$ then both $|g_m|$ and $\eta_m(G,\mathbb{D})$ can be bounded above and below by constants times $(\tau\text{cap}(\mathbb{E}))^m/(m!)$ (as in the disk case).

In general, if G is analytic in $|w| \leq R$ then $|g_m| \leq R^{-m}\,\|G\|_{L_\infty(\{|w|\leq R\})}$.

Arnoldi $\Longrightarrow$ best polynomial approximation on $\mathbb{D}$

The Faber coefficients are given by

$$g_j = \frac{1}{2\pi i} \int_{|w|=1} \frac{g(\psi(w))}{w^{j+1}} \, dw \quad \Longrightarrow \quad g = \mathcal{F}(G), \quad G(w) = \sum_{j=0}^{\infty} g_j w^j,$$

In the polynomial case $q = Q = 1$, we have

LEMMA 2:

$$|g_m| \leq \eta^1_{m-1}(G, \mathbb{D}) \leq \sum_{j=0}^{\infty} |g_{m+j}|.$$

Knizhnerman '91 gave a similar upper bound with additional powers of $m + j$

Hochbruck & Lubich '97 gave a more complicated bound, weaker up to factor 0.75.

Example: $g(z) = \exp(\tau z)$, $\tau > 0$: if $m \geq \tau \mathrm{cap}(\mathbb{E})$ then both $|g_m|$ and $\eta_m(G, \mathbb{D})$ can be bounded above and below by constants times $(\tau \mathrm{cap}(\mathbb{E}))^m/(m!)$ (as in the disk case).

In general, if G is analytic in $|w| \leq R$ then $|g_m| \leq R^{-m} \|G\|_{L_\infty(\{|w| \leq R\})}$.

Rational Arnoldi $\Longrightarrow$ best rational approximation on $\mathbb{D}$ with prescribed poles

Rational approximation (with fixed denominator) allows for a better rate of convergence for functions analytic in parts of the plane like

$$\frac{1}{\sqrt{z}} = \int_{-\infty}^{0} \frac{1}{z-t} \frac{dx}{\pi\sqrt{|x|}}, \qquad \frac{\log(z)}{z-1} = \int_{-\infty}^{0} \frac{1}{z-x} \frac{dx}{1+|x|},$$

$$z^{\kappa} = \frac{\sin(\pi|\kappa|)}{\pi} \int_{-\infty}^{0} \frac{|x|^{\kappa}}{z-x} dx, \qquad \kappa \in (-1, 0),$$

or more generally **MARKOV FUNCTIONS**

$$g(z) = \int_{\alpha}^{\beta} \frac{d\mu(x)}{z-x}, \quad \alpha < \beta < \gamma = \min\{\mathrm{Re}(z) : z \in \mathbb{E}\}, \; \mu \geq 0.$$

Notice: $G = \mathcal{F}^{-1} g$ is also Markov function with support $[\phi(\alpha), \phi(\beta)] \subset [-\infty, -1)$.

The special case of Markov functions

Suppose in what follows that w_j occur in conjugate pairs, and $w_j = \phi(z_j) \in (\phi(\alpha), \phi(\beta))$ have even multiplicities. Thus $B(w) := \prod_{j=1}^{m} \frac{1 - w\overline{w_j}}{w - w_j}$ is of unique sign and of modulus ≥ 1 in $[\phi(\alpha), \phi(\beta)]$.

THEOREM 3: Under the above assumptions for $\mathbb{E}$, w_j, g

$$\int \frac{\phi'(x)}{|\phi(x)|^2 - 1/|B(\phi(x))|} \frac{d\mu(x)}{|B(\phi(x))|} \leq \eta_m^Q(G, \mathbb{D})$$

$$\leq \int \frac{\phi'(x)}{|\phi(x)|^2 - 1} \frac{d\mu(x)}{|B(\phi(x))|} \leq \|g\|_{L_\infty(\mathbb{E})} \max_{y \in [\phi(\alpha), \phi(\beta)]} \frac{1}{|yB(y)|}.$$

NB1: lower/upper bound differ by constant $1 - |\phi(\beta)|^{-2}$.

NB2: upper bound remains valid without assumptions on multiplicities.

Ideas of proof: for upper bound construct modified interpolant at $1/\overline{w}_j$ and 0. For lower bound compute explicitly best rational approximant on discrete subset of $\partial\mathbb{D}$.

How to choose the poles?

We have so far for Markov functions g

$$\|g(A)b - V_{m+1}g(A_{m+1})e_1\| \leq \frac{4\|g\|_{L_\infty(\mathbb{E})}}{|\phi(\beta)|} \max_{w\in[\phi(\alpha),\phi(\beta)]} \frac{1}{|B(w)|}, \qquad B(w) := \prod_{j=1}^{m} \frac{1 - w\overline{w_j}}{w - w_j}.$$

Choice of poles $z_j = \psi(w_j)$ by minimizing $1/B$ on $[\phi(\alpha), \phi(\beta)]$?

- one pole of multiplicity m.
- two poles of multiplicity $m/2$.
- k poles of multiplicity p, $m = pk$.

How to choose a single pole?

Let m be even. The problem

$$\min_{w_1 \in \mathbb{C}} \max_{[\phi(\alpha),\phi(\beta)]} \left| \frac{1 - \overline{w}_1 w}{w - w_1} \right|$$

can be solved explicitly with optimal pole

$$w_1 = w_{opt} = \frac{\widetilde{\phi}(\beta)\widetilde{\phi}(\alpha) + 1}{\widetilde{\phi}(\beta)\widetilde{\phi}(\alpha) - 1}, \qquad \widetilde{\phi}(w) = \sqrt{\frac{|\phi(w)| - 1}{|\phi(w)| + 1}} \in (0,1),$$

leading to the convergence rate

$$\max_{w \in [\phi(\alpha),\phi(\beta)]} \left| \frac{1 - \overline{w}_1 w}{w - w_1} \right|^m = \left(\frac{1 - \widetilde{\phi}(\beta)/\widetilde{\phi}(\alpha)}{1 + \widetilde{\phi}(\beta)/\widetilde{\phi}(\alpha)} \right)^m .$$

Example: if $[\alpha,\beta] = [-\infty, 0], \mathbb{E} = [\lambda_{\min}, \lambda_{\max}]$ then rate $\approx e^{-2m/\sqrt[4]{\kappa}}$, $\kappa = \frac{\lambda_{\max}}{\lambda_{\min}}$.

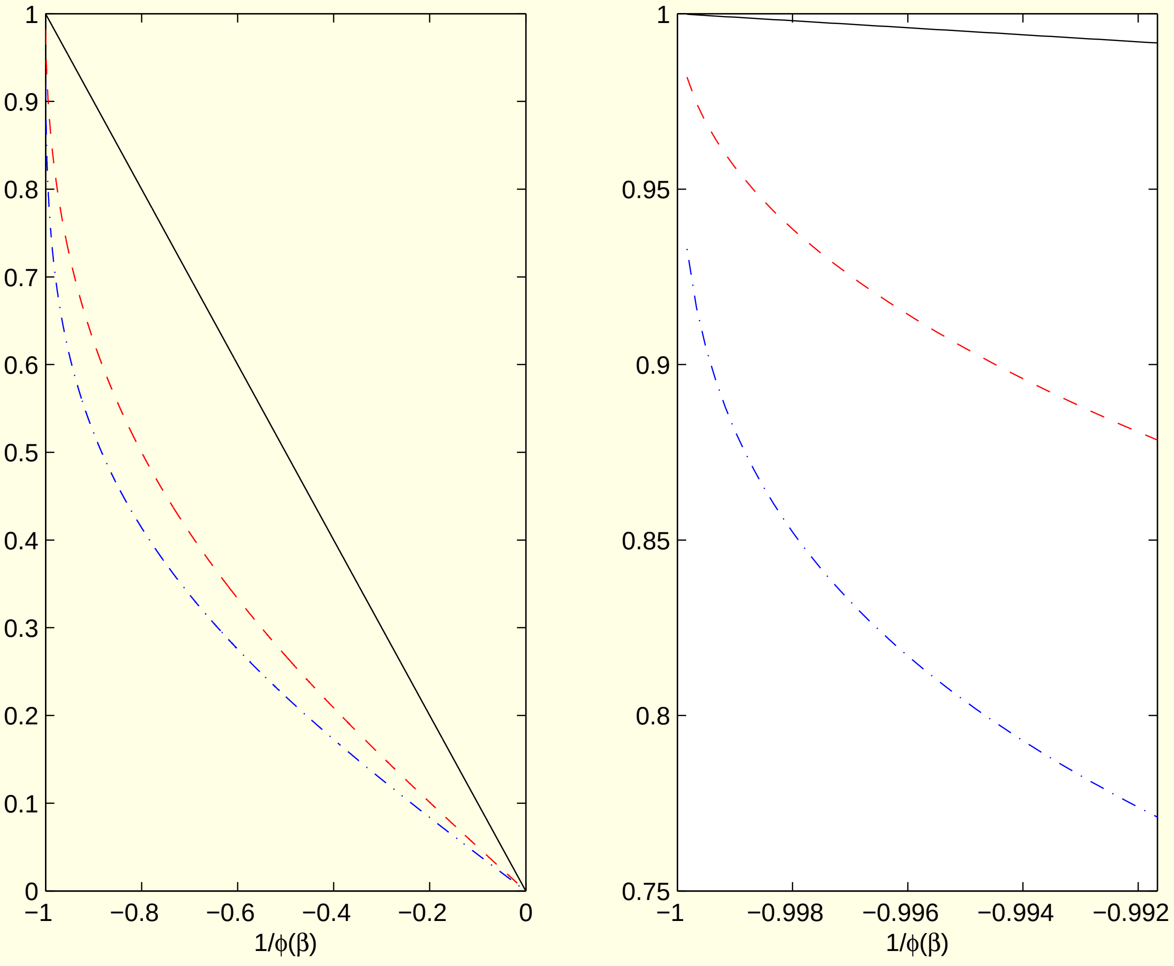
How to choose one finite pole?

We obtain exactly the same rate for a single pole as if we take the two poles $w_1 = \phi(\alpha)$ and $w_2 = \phi(\beta)$ with multiplicities $m/2$!

Special case $\alpha = \infty = w_1$: see Druskin & Knizhnerman '98 for symmetric A and Knizhnerman & Simoncini '08 for general A.

But we get a better rate c^m for $w_1 = \infty$ if we put the finite pole at $w_2 = -\frac{c^2+1}{2c}$, where c unique solution in $(0, 1/|\phi(\beta)|)$ of

$$-\sqrt{\frac{2c^2}{1+c^4}} = \frac{1/\phi(\beta)+c}{1+c/\phi(\beta)}.$$



How to choose real poles?

Let $m = pk$ for some integers $p, k \geq 1$.

This pole placement problem is reduced to the (classical) third Zolotarev problem for minimal Blaschke products on intervals $I \subset \mathbb{R} \setminus \mathbb{D}$:

find $Z_{k,I}$ the minimum $L_\infty(I)$ norm of a Blaschke product of order k,

the required poles $w_1, ..., w_k$ (repeated periodically) being the zeros of such a minimal Blaschke product $B_{k,I}$.

It is known that $R(I, \mathbb{D})^{-k} \leq Z_{k,I} \leq 2R(I, \mathbb{D})^{-k}$, and that for $j = 1, 2, ..., k$

$$w_j = \chi_{I,\mathbb{D}}\Big(\exp(2\pi i \frac{2j-1}{4k})\Big)$$

where $R(I, \mathbb{D})$ is the ring modulus and $\chi_{I,\mathbb{D}}$ the conformal map from $1 < |\zeta| < R$ onto the doubly connected set $\overline{C} \setminus (\mathbb{D} \cup I)$....

Convergence rate for k LU decompositions:

$$\max_{w \in [\phi(\alpha), \phi(\beta)]} 1/|B(w)| \leq 2^p R([\phi(\alpha), \phi(\beta)], \mathbb{D})^{-m}.$$

... some final comments ...

... and what to do with

$$\log(z) = (z-1)\int_{-\infty}^{0} \frac{1}{z-t}\frac{dt}{1-t},$$

$$z^{7/2} = z^4 z^{-1/2} = z^4 \int_{-\infty}^{0} \frac{1}{z-t}\frac{dt}{\pi\sqrt{|t|}},$$

... we can go back to the proof of THM 1: if $\widetilde{g}(z) = p_1(z) + p_2(z)g(z)$ with $\deg p_2 = s, \deg p_1 \leq m+s$ and $G = \mathcal{F}(g)$ then with $z_{m+1} = ... = z_{m+s+1} = \infty$

$$\|\widetilde{g}(A)b - V_{m+s+1}\widetilde{g}(A_{m+s+1})e_1\| \leq 4\,\|p_2(A)b\|\,\eta_m^Q(G,\mathbb{D}).$$