Symbolic-numeric Arithmetic with Rational Functions

George Labahn

Symbolic Computation Group
Cheriton School of Computer Science
University of Waterloo

Symbolic-Numeric Seminar, New York, March 10, 2016

Joint work with Bernhard Beckermann and Ana C. Matos (Lille, France)

Report on Papers:

This talk is a report on:

- B. Beckermann, G. Labahn, A. Matos,
  On Rational Functions without Froissart Doublets, (2016)

and on related papers:

- B. Beckermann and A. Matos,
  Algebraic properties of robust Padé approximants,
  Journal of Approximation Theory 190, 91-115 (2015)

- P. Gonnet, S. Güttel and L. N. Trefethen,
  Robust Padé approximation via SVD, SIAM Review, (2013)

Outline

Definition of Padé approximants

Given $f(z) \approx \sum_{j=0}^{\infty} c_j z^j$ its $[m|n]$ Padé approximant $\frac{p}{q}$ satisfies

$$p(z) = \sum_{j=0}^{m} p_j z^j, \quad q(z) = \sum_{j=0}^{n} q_j z^j \not\equiv 0, \quad f(z)q(z) - p(z) = \sum_{j=m+n+1}^{\infty} e_j z^j.$$

Definition of Padé approximants

Given $f(z) \approx \sum_{j=0}^{\infty} c_j z^j$ its $[m|n]$ Padé approximant $\frac{p}{q}$ satisfies

$$p(z) = \sum_{j=0}^{m} p_j z^j, \quad q(z) = \sum_{j=0}^{n} q_j z^j \not\equiv 0, \quad f(z)q(z) - p(z) = \sum_{j=m+n+1}^{\infty} e_j z^j.$$

Existence: Non-trivial solution of linear system $C \cdot \mathrm{vec}(q) = 0$,

$$C = \begin{bmatrix} c_{m+1} & \cdots & c_{m-n+2} & c_{m-n+1} \\ c_{m+2} & \cdots & c_{m-n+3} & c_{m-n+2} \\ \vdots & & \vdots & \vdots \\ c_{m+n} & \cdots & c_{m+1} & c_m \end{bmatrix}, \quad \mathrm{vec}(q) = \begin{bmatrix} q_0 \\ q_1 \\ \vdots \\ q_n \end{bmatrix},$$

with the convention $c_j = 0$ for $j < 0$.

Definition of Padé approximants

Given $f(z) \approx \sum_{j=0}^{\infty} c_j z^j$ its $[m|n]$ Padé approximant $\frac{p}{q}$ satisfies

$$p(z) = \sum_{j=0}^{m} p_j z^j, \quad q(z) = \sum_{j=0}^{n} q_j z^j \not\equiv 0, \quad f(z)q(z) - p(z) = \sum_{j=m+n+1}^{\infty} e_j z^j.$$

Existence: Non-trivial solution of linear system $C \cdot \mathrm{vec}(q) = 0$,

$$C = \begin{bmatrix} c_{m+1} & \cdots & c_{m-n+2} & c_{m-n+1} \\ c_{m+2} & \cdots & c_{m-n+3} & c_{m-n+2} \\ \vdots & & \vdots & \vdots \\ c_{m+n} & \cdots & c_{m+1} & c_m \end{bmatrix}, \quad \mathrm{vec}(q) = \begin{bmatrix} q_0 \\ q_1 \\ \vdots \\ q_n \end{bmatrix},$$

with the convention $c_j = 0$ for $j < 0$.

Uniqueness: Yes for rational function $\frac{p}{q}$.

Scaling : $c_j \mapsto \beta c_j \gamma^j$ for suitable $\beta, \gamma \in \mathbb{C}$.

Why are Padé Approximants useful?

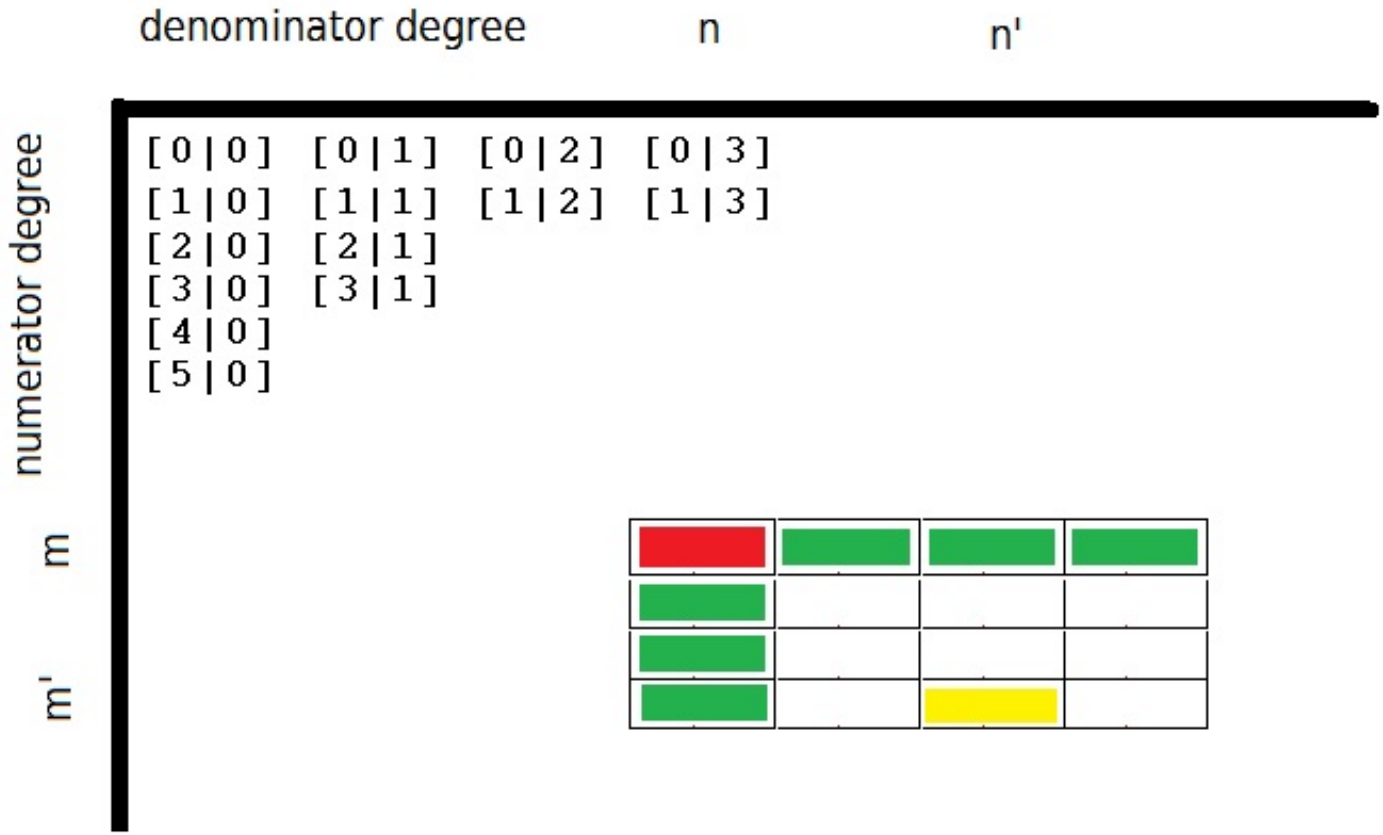
- Approximation of functions via local information
- Approximation of singularities of functions
- Inversion of structured matrices
- Sparse polynomial interpolation (Prony)
- Quadrature formulas

Other types of Padé approximants

- Simultaneous-Padé, Hermite-Padé, . . .
- Vector and matrix versions

Structures in Padé table - degeneracy [Padé, 1892]

Equal entries in Padé table form square, here $[m'|n'] = [m|n]$.



Yellow: desired entry $[m'|n']$,

Red & green: non-degenerate approximants (at least 1 degree exact).

Numerical issues for Padé approximants: coefficients versus values

Stability: small changes in vector of Taylor coefficients c should correspond to small changes in Padé approximant $\frac{p}{q}$.

Numerical issues for Padé approximants: coefficients versus values

Stability: small changes in vector of Taylor coefficients c should correspond to small changes in Padé approximant $\frac{p}{q}$.

Forward conditioning: Does a slightly different $c \approx \overline{c}$ give

- a slightly different vector of coefficients $\begin{bmatrix} vec(p) \\ vec(q) \end{bmatrix}$?
- a slightly different value $\frac{p(z)}{q(z)}$ for a fixed z or slightly different values for all z in the closed unit disk $\mathbb{D}$?

Numerical issues for Padé approximants: coefficients versus values

Stability: small changes in vector of Taylor coefficients c should correspond to small changes in Padé approximant $\frac{p}{q}$.

Forward conditioning: Does a slightly different $c \approx \overline{c}$ give

- a slightly different vector of coefficients $\begin{bmatrix} vec(p) \\ vec(q) \end{bmatrix}$?
- a slightly different value $\frac{p(z)}{q(z)}$ for a fixed z or slightly different values for all z in the closed unit disk $\mathbb{D}$?

Backward conditioning: Does closeby vector of coefficients represent a Padé approx. of closeby vector of Taylor coefficients?

Numerical issues for Padé approximants: coefficients versus values

Stability: small changes in vector of Taylor coefficients c should correspond to small changes in Padé approximant $\frac{p}{q}$.

Forward conditioning: Does a slightly different $c \approx \overline{c}$ give

- a slightly different vector of coefficients $\begin{bmatrix} vec(p) \\ vec(q) \end{bmatrix}$?
- a slightly different value $\frac{p(z)}{q(z)}$ for a fixed z or slightly different values for all z in the closed unit disk $\mathbb{D}$?

Backward conditioning: Does closeby vector of coefficients represent a Padé approx. of closeby vector of Taylor coefficients?

What happens with the Padé poles? $\longrightarrow$ [Beckermann, Golub, & L '07]

- *Here: worst amplification of infinitesimally small relative errors.*

Other numerical issues for Padé approximants: spurious poles

Spurious poles: poles of $\frac{p}{q}$ "far" from singularities of function f

- they occur in exact arithmetic

Other numerical issues for Padé approximants: spurious poles

Spurious poles: poles of $\frac{p}{q}$ "far" from singularities of function f

- they occur in exact arithmetic

Numerical counterpart for such poles in $\mathbb{D} = \{|z| \leq 1\}$ (f analytic)

- **Froissart doublet: (simple) pole with close-by zero**

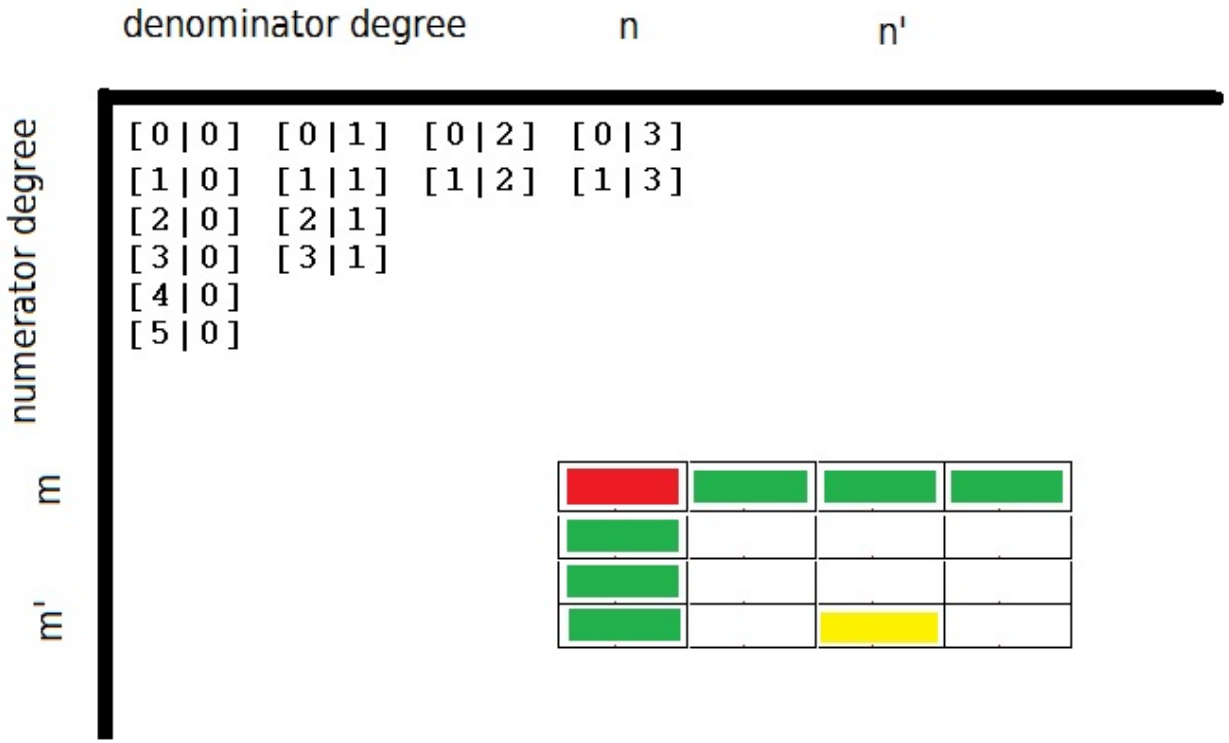
$$Froissart = \min\{\ |\sigma - \tau|\ :\ p(\sigma) = 0, q(\tau) = 0, |\tau| \leq 1\},$$

- **(simple) poles with small residual:**

$$Residual = \min\{\ |\frac{p(\tau)}{q'(\tau)}|\ :\ q(\tau) = 0, |\tau| \leq 1\}.$$

Structures in Padé table - degeneracy [Padé, 1892]

Equal entries in Padé table form square (in exact arithmetic).



In yellow: desired entry $[m'|n']$,
In red: robust approximant $[m|n] = [m'|n']$ (both degrees exact),
In red and green: nondegenerate approximants (one degree exact).

Look-ahead [Cabay Meleshko'93], Look-around [Graves-Morris'97]...

Robust Padé approx [Gonnet, Güttel & Trefethen'12]

Input: $(m', n'), (c_0, ..., c_{m'+n'})^T, tol$

Output: $m \leq m'$, $n \leq n'$, $[m|n]$ Padé approximant p/q s.t.

- m, n as large as possible, p, q coprime and of degree $= m, n$;
- $\frac{1}{\|C^\dagger\|} \geq tol > 0$, i.e., numerical rank$(C) = n$.

Some points:

(1) Computed via use of SVD on C

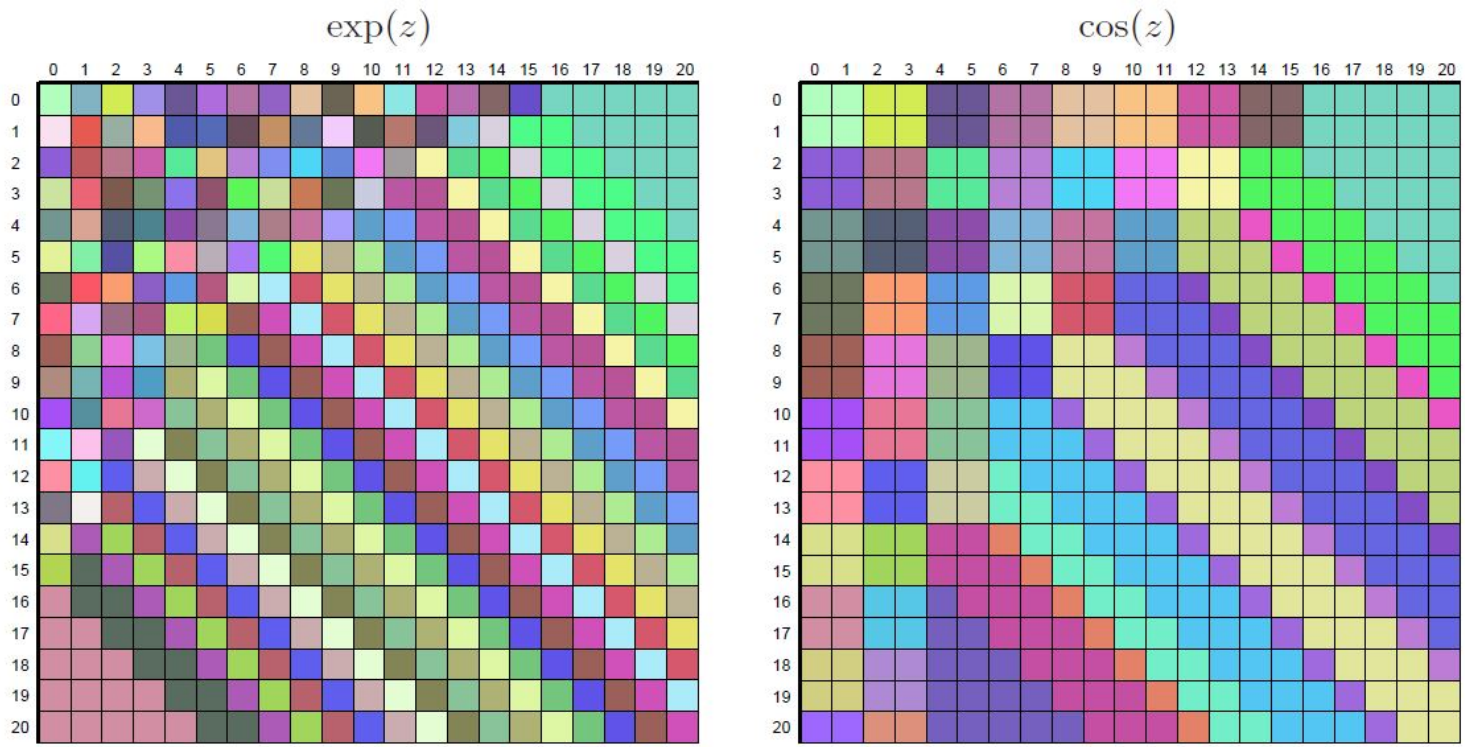
(2) Thus robust Padé approximants are forward stable.

(3) Gonnet, Güttel and Trefethen conjectured :

No Froissart doublets, no small residuals.

Robust Padé Approximants: (Gonnet, Güttel and Trefethen)

Example of Padé table of Robust Padé approximants from GGT:



Questions for rational functions in numerical setting

Q1) How to measure distance between rational functions?

- answer depends on how to represent rational functions...
- should we ask that values are close?
- should we ask that vectors of coefficients are close? Which basis?

Questions for rational functions in numerical setting

Q1) How to measure distance between rational functions?

- answer depends on how to represent rational functions...
- should we ask that values are close?
- should we ask that vectors of coefficients are close? Which basis?

Q2) Is a given rational function "close" to a rational function with lower degree?

Questions for rational functions in numerical setting

Q1) How to measure distance between rational functions?

- answer depends on how to represent rational functions...
- should we ask that values are close?
- should we ask that vectors of coefficients are close? Which basis?

Q2) Is a given rational function "close" to a rational function with lower degree?

Q3) Can we find a simple and reliable indicator for Froissart doublets or/and for small residuals?

Notations and reminders

Fix integers $m, n \geq 0$

- $\mathcal{R}_{m,n}$ set of rational functions $\frac{p}{q}$ with $\partial p \leq m, \quad \partial q \leq n$
- rational function $\frac{p}{q} \in \mathcal{R}_{m,n}$ called *nondegenerate* if
  - p and q are co-prime,
  - *defect*$(p, q) = \min\{m - \deg p, n - \deg q\} = 0$;
- $a_1 \lesssim a_2$ means that there exist modest constants $b, r > 0$ not depending on m, n such that $a_1 \leq b(m + n + 1)^r a_2$.
- $a_1 \sim a_2$ if $a_1 \lesssim a_2$ and $a_2 \lesssim a_1$.
- $\mathbb{D} = \{z \in \mathbb{C} : |z| \leq 1\}$ is the unit disk
- $c(z) = c_0 + c_1 z + \cdots + c_n z^n$ let vec$(c) = (c_0, c_1, \cdots, c_n)^T$

Sylvester type matrix S

Sylvester matrix is denoted $S_* \in \mathbb{C}^{(m+n)\times(m+n)}$

Sylvester type matrix of two polynomials $\frac{p}{q}$

$$S = \left(\begin{array}{ccccc|ccccc} p_0 & & & & & q_0 & & & & \\ p_1 & p_0 & & & & q_1 & q_0 & & & \\ \vdots & & \ddots & & & \vdots & & \ddots & & \\ & p_1 & & & & & q_1 & & & \\ p_m & \vdots & \ddots & p_0 & & q_n & \vdots & \ddots & p_0 & \\ & p_m & & p_1 & & & q_n & & q_1 & \\ & & \ddots & \vdots & & & & \ddots & \vdots & \\ & & & p_m & & & & & q_n & \end{array} \right) \in \mathbb{C}^{(m+n+1)\times(m+n+2)}.$$

$n+1$ columns of p coefficients; $m+1$ columns of q coefficients.

Note:

– $\frac{p}{q}$ nondegenerate iff S full rank (S_* is invertible)

$$(1,\ z, ..., z^{m+n}) \cdot S = (q(z),\ ...,\ q(z)z^m,\ p(z),\ ...,\ p(z)z^n).$$

Measure of coprimeness $\epsilon(p,q)$

Set

$$\begin{aligned}\epsilon(p,q) &= \min\left\{\left\| \begin{array}{c}\mathrm{vec}(p-\widetilde{p})\\ \mathrm{vec}(q-\widetilde{q})\end{array}\right\| : \frac{\widetilde{p}}{\widetilde{q}} \text{ degenerate}\right\}\\ &\sim \{\|S(q,p)-S(\widetilde{p},\widetilde{q})\| : S(\widetilde{p},\widetilde{q}) \text{ not full rank}\}\end{aligned}$$

For a closed $K \subset \overline{\mathbb{C}}$ set

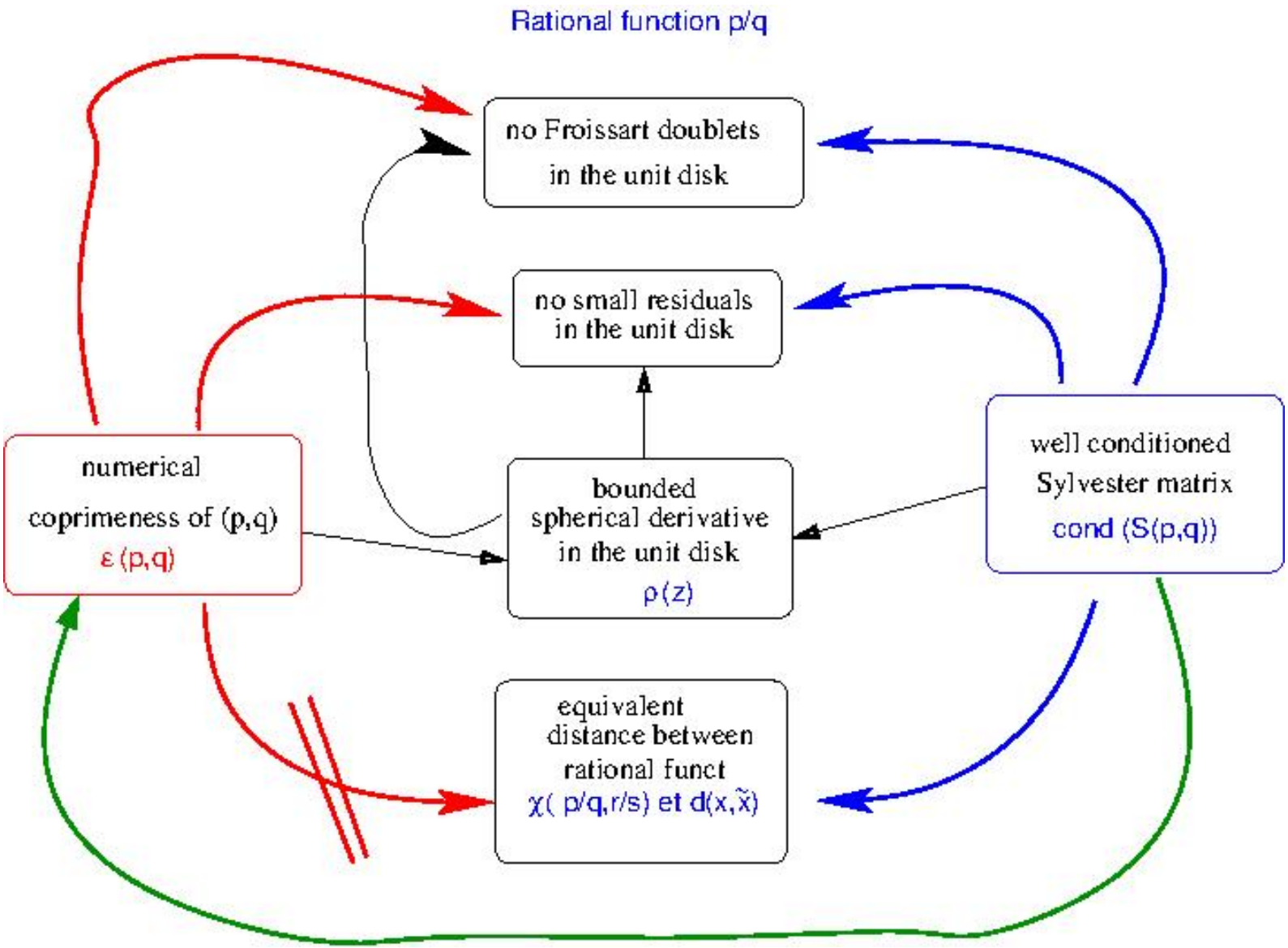
$$\begin{aligned}\epsilon_K(p,q) &= \inf_{z\in K}\left(\frac{|p(z)|^2}{\sum_{j=0}^m |z|^{2j}} + \frac{|q(z)|^2}{\sum_{j=0}^n |z|^{2j}}\right)^{1/2}\\ &\sim \inf_{z\in K}\frac{\|(1,z,...,z^{m+n})S\|}{\|(1,z,...,z^{m+n})\|}\end{aligned}$$

Then [Corless, Gianni, Trager, Watt, 1995], [Beckermann & L'98]

$$\epsilon(p,q) = \epsilon_{\overline{\mathbb{C}}}(p,q) \qquad \text{(computable)}$$

$$\frac{\|(\mathrm{vec}(p)^T,\mathrm{vec}(q)^T)\|}{\epsilon(p,q)} \lesssim cond(S) \quad \text{(structured condition number)}$$

BLM 2016 - Summary of results - BM 2015



How to measure distances in $\mathcal{R}_{m,n}$?

For $r = \frac{p}{q}$, $\widetilde{r} \in \mathcal{R}_{m,n}$ we define the coefficient vector

$$x(r) = \begin{bmatrix} \text{vec}(p) \\ \text{vec}(q) \end{bmatrix}$$

always supposed to be of norm 1.

How to measure distances in $\mathcal{R}_{m,n}$?

For $r = \frac{p}{q}$, $\widetilde{r} \in \mathcal{R}_{m,n}$ we define the coefficient vector

$$x(r) = \begin{bmatrix} \mathrm{vec}(p) \\ \mathrm{vec}(q) \end{bmatrix}$$

always supposed to be of norm 1.

Uniform chordal metric: for closed $K \subset \mathbb{C}$

$$\chi_K(r, \widetilde{r}) = \max_{z \in K} \chi(r(z), \widetilde{r}(z)), \quad \chi(a, b) = \frac{|a - b|}{\sqrt{1 + |a|^2}\sqrt{1 + |b|^2}}$$

How to measure distances in $\mathcal{R}_{m,n}$?

For $r = \frac{p}{q}$, $\widetilde{r} \in \mathcal{R}_{m,n}$ we define the coefficient vector

$$x(r) = \begin{bmatrix} \mathrm{vec}(p) \\ \mathrm{vec}(q) \end{bmatrix}$$

always supposed to be of norm 1.

Uniform chordal metric: for closed $K \subset \mathbb{C}$

$$\chi_K(r, \widetilde{r}) = \max_{z \in K} \chi(r(z), \widetilde{r}(z)), \quad \chi(a, b) = \frac{|a - b|}{\sqrt{1 + |a|^2}\sqrt{1 + |b|^2}}$$

Distance of coeff. vectors of norm 1 with optimal phase:

$$d(r, \widetilde{r}) := \min\{\|x(r) - a x(\widetilde{r})\| : a \in \mathbb{C}, |a| = 1\}.$$

(if $x(r), x(\widetilde{r})$ real then best $a \in \{\pm 1\}$).

How to compare distances in $\mathcal{R}_{m,n}$?

For $r, \widetilde{r} \in \mathcal{R}_{m,n}$ we define the distances

$$\chi_K(r, \widetilde{r}) = \max_{z \in K} \chi(r(z), \widetilde{r}(z)), \quad \chi(a,b) = \frac{|a-b|}{\sqrt{1+|a|^2}\sqrt{1+|b|^2}}$$

$$d(r, \widetilde{r}) := \min\{\|x(r) - a x(\widetilde{r})\| : a \in \mathbb{C}, |a| = 1\}, \quad x(r) = \begin{bmatrix} \mathrm{vec}(p) \\ \mathrm{vec}(q) \end{bmatrix}.$$

Theorem [Matos& Beckermann '15] Let $r = \frac{p}{q}$ be nondegenerate. Then for all $\widetilde{r} \in \mathcal{R}_{m,n}$ and $K \subset \mathbb{D}$

$$\chi_K(r, \widetilde{r}) \lesssim \mathrm{cond}(S)\, d(r, \widetilde{r}),$$

and for all $K \subset \mathbb{C}$ containing $(m+n+1)$th roots of unity

$$\frac{d(r, \widetilde{r})}{\mathrm{cond(S)}} \lesssim \chi_K(r, \widetilde{r}).$$

Distances of same magnitude if cond(S) modest.

Again how to compare distances in $\mathcal{R}_{m,n}$?

For $r, \widetilde{r} \in \mathcal{R}_{m,n}$ we define the distances

$$\chi_K(r, \widetilde{r}) = \max_{z \in K} \chi(r(z), \widetilde{r}(z)), \quad \chi(a,b) = \frac{|a-b|}{\sqrt{1+|a|^2}\sqrt{1+|b|^2}}$$

$$d(r, \widetilde{r}) := \min\{\|x(r) - a x(\widetilde{r})\| : a \in \mathbb{C}, |a| = 1\}, \quad x(r) = \begin{bmatrix} \mathrm{vec}(p) \\ \mathrm{vec}(q) \end{bmatrix}.$$

Theorem [Beckermann & L & Matos '16] : Let $r = \frac{p}{q}$ be nondegenerate. Then for all $\widetilde{r} \in \mathcal{R}_{m,n}$ and $K \subset \mathbb{D}$

$$\chi_K(r, \widetilde{r}) \lesssim \frac{d(r, \widetilde{r})}{\epsilon_K(p,q)}.$$

and

$$\frac{d(r, \widetilde{r})}{\mathrm{cond(S)}} \lesssim \chi_{\mathbb{D}}(r, \widetilde{r}),$$

Cannot be essentially improved: we have example with varying $m = n$ with $\epsilon_{\mathbb{D}}(p,q) \sim 1$, $\epsilon_{\mathbb{C}}(p,q) \sim 3^{-n}$, $\mathrm{cond(S)} \sim 8^n$.

Indicators under perturbations

Our indicators cond(S(q,p)) and $\epsilon_K(p,q)$ do not vary much:

If $2(m+n+1)\text{cond}(S(q,p))\, d(r,\widetilde{r}) \leq 1/3$ then

$$\frac{\text{cond}(S(\widetilde{q},\widetilde{p}))}{\text{cond}(S(q,p))} \in [1/2, 2].$$

Indicators under perturbations

Our indicators cond(S(q,p)) and $\epsilon_K(p,q)$ do not vary much:

If $2(m+n+1)\text{cond}(S(q,p))\, d(r,\widetilde{r}) \leq 1/3$ then

$$\frac{\text{cond}(S(\widetilde{q},\widetilde{p}))}{\text{cond}(S(q,p))} \in [1/2, 2].$$

If $2\, d(r,\widetilde{r}) \leq \epsilon_K(p,q)$ then

$$\frac{\epsilon_K(\widetilde{p},\widetilde{q})}{\epsilon_K(p,q)} \in [1/2, 3/2].$$

Indicators under perturbations

Our indicators cond(S(q,p)) and $\epsilon_K(p,q)$ do not vary much:

If $2(m+n+1)\text{cond}(S(q,p))\, d(r,\widetilde{r}) \leq 1/3$ then

$$\frac{\text{cond}(S(\widetilde{q},\widetilde{p}))}{\text{cond}(S(q,p))} \in [1/2, 2].$$

If $2\, d(r,\widetilde{r}) \leq \epsilon_K(p,q)$ then

$$\frac{\epsilon_K(\widetilde{p},\widetilde{q})}{\epsilon_K(p,q)} \in [1/2, 3/2].$$

By combining with Theorem 2

$$\inf\{\chi_{\mathbb{D}}(r,\widetilde{r}) : \widetilde{r} \in \mathcal{R}_{m-1,n-1}\} \gtrsim \frac{\epsilon(p,q)}{cond(S(q,p))}.$$

Lower bounds for pole-zero distance

- $r = p/q \in \mathcal{R}_{m,n}$ nondegenerate,
- $K \subset \mathbb{C}$ a closed disk or half plane or $K = \mathbb{C}$.
- $\sigma, \tau \in K$ with $p(\sigma) = 0$ (zero) and $q(\tau) = 0$ (pole).

Question: can we give a simple lower bound for $|\sigma - \tau|$?

Lower bounds for pole-zero distance

- $r = p/q \in \mathcal{R}_{m,n}$ nondegenerate,
- $K \subset \mathbb{C}$ a closed disk or half plane or $K = \mathbb{C}$.
- $\sigma, \tau \in K$ with $p(\sigma) = 0$ (zero) and $q(\tau) = 0$ (pole).

Question: can we give a simple lower bound for $|\sigma - \tau|$?

Theorem [Beckermann & Matos '15]: if $\sigma, \tau \in \mathbb{D}$ then

$$|\sigma - \tau| \gtrsim \frac{1}{\mathrm{cond(S)}}.$$

The same estimate remains true for σ (and τ) a root (a pole) of some meromorphic function f with $\chi_{\mathbb{D}}(f, r) \leq 1/3$.

Lower bounds for pole-zero distance

- $r = p/q \in \mathcal{R}_{m,n}$ nondegenerate,
- $K \subset \mathbb{C}$ a closed disk or half plane or $K = \mathbb{C}$.
- $\sigma, \tau \in K$ with $p(\sigma) = 0$ (zero) and $q(\tau) = 0$ (pole).

Question: can we give a simple lower bound for $|\sigma - \tau|$?

Theorem [Beckermann & Matos '15]: if $\sigma, \tau \in \mathbb{D}$ then

$$|\sigma - \tau| \gtrsim \frac{1}{\mathrm{cond(S)}}.$$

The same estimate remains true for σ (and τ) a root (a pole) of some meromorphic function f with $\chi_{\mathbb{D}}(f, r) \leq 1/3$.

Froissart doublet should be linked with coprimeness : for all K

$$\sigma, \tau \in K \quad \Longrightarrow \quad \chi(\sigma, \tau) \geq \frac{1}{2} \frac{\epsilon_K(p, q)}{m\|\mathrm{vec}(p)\| + n\|\mathrm{vec}(q)\|}$$

A closer look at Froissart doublets

... in fact a Froissart doublet means that for a small change of arguments we have a big change of values, i.e., a very big Lipschitz constant

$$L_K(r) := \sup\left\{\frac{\chi(r(z), r(w))}{\chi(z, w)} : z, w \in K\right\}.$$

A closer look at Froissart doublets

... in fact a Froissart doublet means that for a small change of arguments we have a big change of values, i.e., a very big Lipschitz constant

$$L_K(r) := \sup\left\{\frac{\chi(r(z), r(w))}{\chi(z, w)} : z, w \in K\right\}.$$

$w \to z$ shows that

$$L_K(r) \geq \nu_K(r) := \sup_{z \in K}(1 + |z|^2)\rho(r)(z), \quad \rho(r)(z) = \frac{|r'(z)|}{1 + |r(z)|^2}$$

(spherical derivative), but also $L_K(r) \leq \frac{\pi}{2}\nu_K(r)$.

A closer look at Froissart doublets

... in fact a Froissart doublet means that for a small change of arguments we have a big change of values, i.e., a very big Lipschitz constant

$$L_K(r) := \sup\left\{\frac{\chi(r(z), r(w))}{\chi(z, w)} : z, w \in K\right\}.$$

$w \to z$ shows that

$$L_K(r) \geq \nu_K(r) := \sup_{z \in K}(1 + |z|^2)\rho(r)(z), \quad \rho(r)(z) = \frac{|r'(z)|}{1 + |r(z)|^2}$$

(spherical derivative), but also $L_K(r) \leq \frac{\pi}{2}\nu_K(r)$.

Theorem [Beckermann & L & Matos '16]: if f is meromorphic in K with $\chi_K(f, r) \leq \frac{1}{3}$ then

$$\chi(\sigma, \tau) \geq \frac{1}{3L_K(f)} \geq \frac{2}{3\pi}\frac{1}{\nu_K(f)}.$$

Moreover, if τ is a simple pole of r

$$|\text{Residual}(r)(\tau)| = \frac{1}{\rho(r)(\tau)} \geq \frac{1}{\nu_K(r)}.$$

Interlude: why $L_K(r) \leq \frac{\pi}{2}\nu_K(r)$

Recall: with stereographic projection T from $\mathbb{C}$ onto the sphere of diameter 1:

$$\chi(w,z) = \|T(w) - T(z)\|$$

and spherical distance $\widetilde{\chi}(w,z)$ is length of shortest path on sphere joining $T(z)$ and $T(w)$. Hence for suitable path $\Gamma \subset \mathbb{C}$ joining z and w

$$\begin{aligned}
\chi(r(z),r(w)) \leq \widetilde{\chi}(r(z),r(w)) &\leq \int_{r(\Gamma)} \frac{|du|}{1+|u|^2} \\
&\leq \int_\Gamma \rho(r)(v)|dv| \leq \nu_K(r) \int_\Gamma \frac{|dv|}{1+|v|^2} \\
&= \nu_K(r)\widetilde{\chi}(z,w) \leq \frac{\pi}{2}\nu_K(r)\chi(z,w).
\end{aligned}$$

Is this new estimate better?

Theorem [Beckermann & L & Matos'16]: if f is meromorphic in K with $\chi_K(f,r) \le \frac{1}{3}$ then

$$\chi(\sigma,\tau) \ge \frac{1}{3L_K(f)} \ge \frac{2}{3\pi}\frac{1}{\nu_K(f)}.$$

Moreover, if τ is a simple pole of r

$$|\text{Residual}(r)(\tau)| = \frac{1}{\rho(r)(\tau)} \ge \frac{1}{\nu_K(r)}.$$

Last theorem can be much sharper since

$$\nu_{1/K}(\widetilde{r}) = \nu_K(r) \quad \text{for} \quad \widetilde{r}(z) = r(1/z),$$

$$\frac{\epsilon_K(p,q)}{m\|\text{vec}(p)\| + n\|\text{vec}(q)\|} \lesssim \frac{1}{\nu_K(r)} \quad \text{for } K \subset \mathbb{D} \text{ or } m = n$$

$$\epsilon_K(p^m,q^m) = \epsilon_K(p,q)^m, \quad \nu_K(r^m) \le m\,\nu_K(r).$$

Is this new estimate better?

Theorem [Beckermann & L & Matos'16]: if f is meromorphic in K with $\chi_K(f,r) \le \frac{1}{3}$ then

$$\chi(\sigma,\tau) \ge \frac{1}{3L_K(f)} \ge \frac{2}{3\pi}\frac{1}{\nu_K(f)}.$$

Moreover, if τ is a simple pole of r

$$|\text{Residual}(r)(\tau)| = \frac{1}{\rho(r)(\tau)} \ge \frac{1}{\nu_K(r)}.$$

Last theorem can be much sharper since

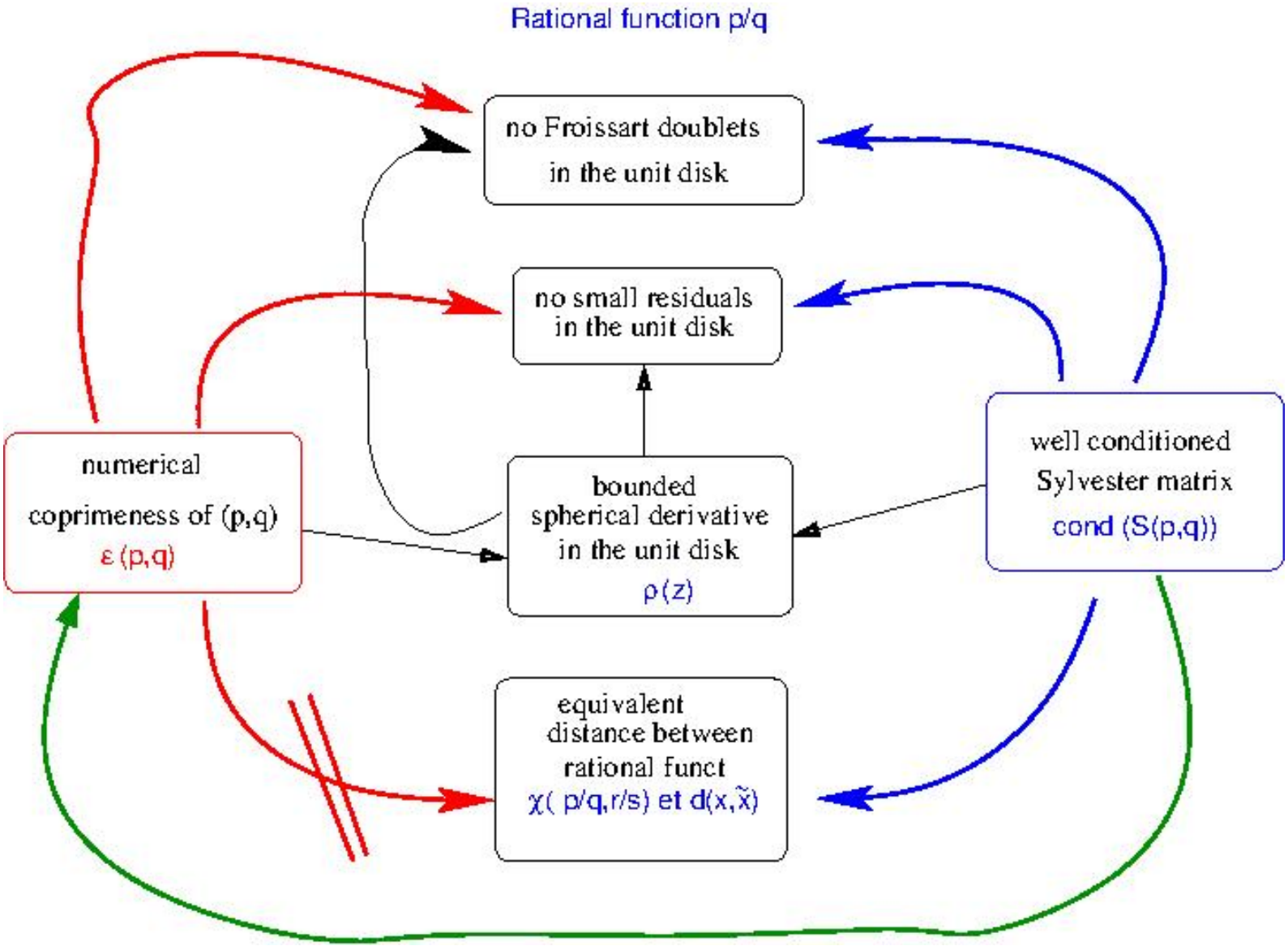
$$\nu_{1/K}(\widetilde{r}) = \nu_K(r) \quad \text{for} \quad \widetilde{r}(z) = r(1/z),$$

$$\frac{\epsilon_K(p,q)}{m\|\text{vec}(p)\| + n\|\text{vec}(q)\|} \lesssim \frac{1}{\nu_K(r)} \quad \text{for } K \subset \mathbb{D} \text{ or } m = n$$

$$\epsilon_K(p^m,q^m) = \epsilon_K(p,q)^m, \quad \nu_K(r^m) \le m\,\nu_K(r).$$

In any case, asking modest Lipschitz constant is reasonable.

Summary of results



The (real) Padé coefficient map

$$\begin{array}{rcl} F : \mathbb{R}^{m+n+1} & \mapsto & \mathbb{R}^{m+n+2} \\ (c_0, \ldots, c_{m+n})^T & \mapsto & \begin{bmatrix} \text{vec}(p) \\ \text{vec}(q) \end{bmatrix} \end{array}$$

- non-trivial gcd of $[m|n]$ Padé approximant $\frac{p}{q}$ cancelled ($\Longrightarrow q(0) \neq 0$) and
- normalization: $\|F(c)\|^2 = \|\text{vec}(p)\|^2 + \|\text{vec}(q)\|^2 = 1,$
  phase: $q(0) > 0.$

Theorem [≈ Werner & Wuytack, '83]:
F is continuous in neighborhood $\mathcal{N}$ of $\overline{c} \Longleftrightarrow$
$F(\overline{c})$ is nondegenerate (i.e., $defect = \min(m - \deg p, n - \deg q) = 0$).

Jacobian of Padé map

Beckermann and Matos make use of the matrices:

$$T \in \mathbb{C}^{(m+n+1)\times(m+n+2)} \quad \text{and} \quad Q \in \mathbb{C}^{(m+n+1)\times(m+n+1)}:$$

$$\begin{bmatrix} 1 & 0 & \cdots & 0 & -c_0 & 0 & \cdots & 0 \\ & & \ddots & \vdots & & & \ddots & \vdots \\ 0 & 1 & \ddots & \vdots & -c_1 & -c_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots & \ddots & \ddots & 0 \\ & & & & & & \ddots & \\ 0 & \cdots & 0 & 1 & -c_m & & \ddots & -c_0 \\ 0 & \cdots & \cdots & 0 & -c_{m+1} & \cdots & \cdots & -c_1 \\ \vdots & & & \vdots & \vdots & & & \vdots \\ 0 & \cdots & \cdots & 0 & -c_{m+n} & \cdots & \cdots & -c_m \end{bmatrix},$$

$$\begin{bmatrix} q_0 & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & & & \vdots \\ q_n & & q_0 & 0 & \cdots & 0 \\ & \ddots & & \ddots & \ddots & \vdots \\ 0 & \ddots & & \ddots & \ddots & \\ \vdots & \ddots & \ddots & & \ddots & 0 \\ 0 & \cdots & 0 & q_n & \cdots & q_0 \end{bmatrix}.$$

Jacobian of Padé map

Beckermann and Matos make use of the matrices:

$$T \in \mathbb{C}^{(m+n+1)\times(m+n+2)} \quad \text{and} \quad Q \in \mathbb{C}^{(m+n+1)\times(m+n+1)}:$$

$$\begin{bmatrix} 1 & 0 & \cdots & 0 & -c_0 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \vdots & -c_1 & -c_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & -c_m & & \ddots & -c_0 \\ 0 & \cdots & \cdots & 0 & -c_{m+1} & \cdots & \cdots & -c_1 \\ \vdots & & & \vdots & \vdots & & & \vdots \\ 0 & \cdots & \cdots & 0 & -c_{m+n} & \cdots & \cdots & -c_m \end{bmatrix},$$

$$\begin{bmatrix} q_0 & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & & & \vdots \\ q_n & & q_0 & 0 & \cdots & 0 \\ & \ddots & & \ddots & \ddots & \vdots \\ 0 & \ddots & & \ddots & \ddots & \\ \vdots & \ddots & \ddots & & \ddots & 0 \\ 0 & \cdots & 0 & q_n & \cdots & q_0 \end{bmatrix}.$$

Notice that

$$Q \cdot c = \begin{bmatrix} vec(p) \\ 0 \end{bmatrix} \quad \text{and} \quad T \begin{bmatrix} vec(p) \\ vec(q) \end{bmatrix} = 0, \quad \text{rank}(T) = m + n + 1.$$

Forward/backward stability

Theorem [Matos & Beckermann '15]: let $F(\overline{c})$ be nondegenerate, $\mathcal{N}$ neighborhood as before. Assume $||\overline{c}|| = 1$. Then:

$$cond_{forward}(F) = \limsup_{c \to \overline{c}} \frac{||F(c) - F(\overline{c})||/||F(\overline{c})||}{||c - \overline{c}||/||\overline{c}||} = ||J_F(\overline{c})||.$$

$$cond_{backward}(F) = \limsup_{F(c) \to F(\overline{c})} \frac{||c - \overline{c}||/||\overline{c}||}{||F(c) - F(\overline{c})||/||F(\overline{c})||} = ||J_F(\overline{c})^\dagger||$$

$$= \limsup_{\substack{y \to F(\overline{c}) \\ ||F(c)-y|| = \text{dist}(y, F(\mathcal{N}))}} \frac{||c - \overline{c}||/||\overline{c}||}{||y - F(\overline{c})||/||F(\overline{c})||}.$$

Here $J_F(c) = T^\dagger \cdot Q$ and $J_F(c)^\dagger = Q^{-1} \cdot T$

Jacobian of Padé map and a Sylvester type matrix

We have

$$J_F(c) = T^\dagger \cdot Q = S^\dagger \cdot Q^2, \quad J_F(c)^\dagger = Q^{-1} \cdot T = Q^{-2} \cdot S,$$

with the

$$\begin{bmatrix} q_0 & 0 & \cdots & 0 & -p_0 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots & \vdots & \ddots & \ddots & \vdots \\ q_n & & \ddots & 0 & -p_m & & \ddots & 0 \\ 0 & \ddots & & q_0 & 0 & \ddots & & -p_0 \\ \vdots & \ddots & \ddots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & q_n & 0 & \cdots & 0 & -p_m \end{bmatrix},$$

$$\begin{bmatrix} q_0 & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & & & \vdots \\ q_n & & q_0 & 0 & \cdots & 0 \\ 0 & \ddots & & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & & \ddots & 0 \\ 0 & \cdots & 0 & q_n & \cdots & q_0 \end{bmatrix}$$

being $S = Q \cdot T \in \mathbb{C}^{(m+n+1)\times(m+n+2)}$ and $Q \in \mathbb{C}^{(m+n+1)\times(m+n+1)}$.

Conclusion on stability of Padé coefficient map

$$J_F(c) = T^\dagger Q = S^\dagger Q^2, \quad J_F(c)^\dagger = Q^{-1} T = Q^{-2} S,$$

With normalizations $\|c\| = 1, \|F(c)\| = 1$:

$$\|Q\| \sim 1,\ \|C\| \leq \|T\| \sim 1,\ \|S\| \sim 1,$$
$$\|T^\dagger\| \sim \|C^\dagger\|,\ \max(\|Q^{-1}\|, \|T^\dagger\|) \lesssim \|S^\dagger\|.$$

and $\|J_F(\overline{c})^\dagger\| \sim \|Q^{-1}\|$.

Conclusion on stability of Padé coefficient map

$$J_F(c) = T^\dagger Q = S^\dagger Q^2, \quad J_F(c)^\dagger = Q^{-1} T = Q^{-2} S,$$

With normalizations $\|c\| = 1, \|F(c)\| = 1$:

$$\|Q\| \sim 1,\ \|C\| \leq \|T\| \sim 1,\ \|S\| \sim 1,$$
$$\|T^\dagger\| \sim \|C^\dagger\|,\ \max(\|Q^{-1}\|, \|T^\dagger\|) \lesssim \|S^\dagger\|.$$

and $\|J_F(\overline{c})^\dagger\| \sim \|Q^{-1}\|$.

- $cond(T) \sim \|C^\dagger\|$ modest $\Longrightarrow$ forward stable,
- $cond(Q)$ modest $\Longleftrightarrow$ backward stable.

In both cases sufficient condition: $cond(S)$ modest.

Stability of Padé value map

Theorem [Matos & Beckermann '15]: for $m \geq n-1$ and for any $K \subset \mathbb{D}$ provided that $[m|n]_{\overline{c}} = p_{\overline{c}}/q_{\overline{c}}$ is nondegenerate

$$\limsup_{c \mapsto \overline{c}} \max_{z \in K} \frac{\chi([m|n]_c(z), [m|n]_{\overline{c}}(z))}{\|c-\overline{c}\|/\|\overline{c}\|} \lesssim \gamma(K)+1, \quad \gtrsim \gamma(K)-1,$$

$$\gamma(K) = \max_{z \in K} \frac{|z^{m+n+1}| \overbrace{\|\mathsf{vec}(q_{\overline{c}})\|^2}^{\sim 1}}{|p_{\overline{c}}(z)|^2 + |q_{\overline{c}}(z)|^2}.$$

Stability of Padé value map

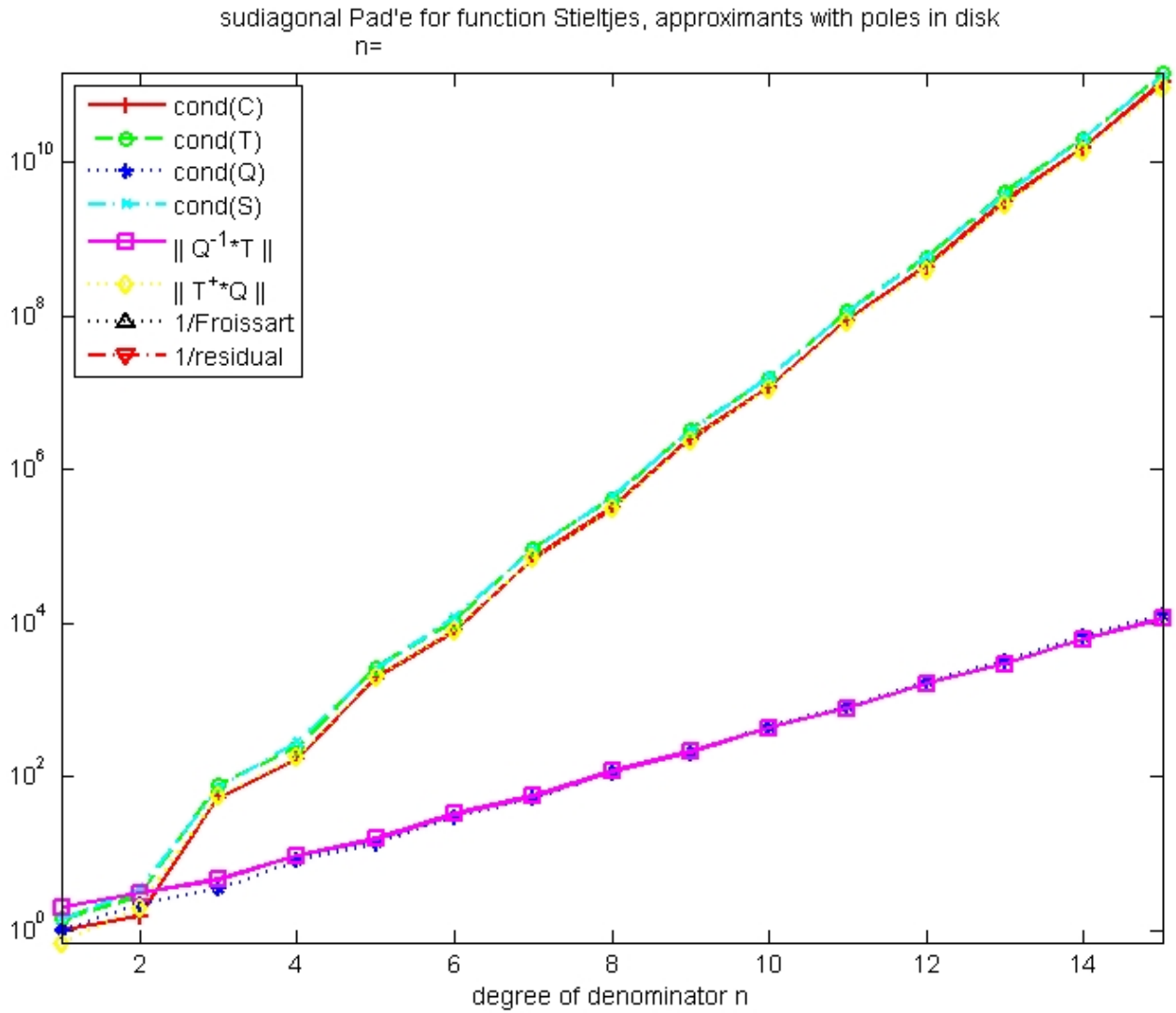
Theorem [Matos & Beckermann '15]: for $m \geq n-1$ and for any $K \subset \mathbb{D}$ provided that $[m|n]_{\overline{c}} = p_{\overline{c}}/q_{\overline{c}}$ is nondegenerate

$$\limsup_{c \mapsto \overline{c}} \max_{z \in K} \frac{\chi([m|n]_c(z), [m|n]_{\overline{c}}(z))}{\|c - \overline{c}\|/\|\overline{c}\|} \lesssim \gamma(K) + 1, \quad \gtrsim \gamma(K) - 1,$$

$$\gamma(K) = \max_{z \in K} \frac{|z^{m+n+1}| \overbrace{\|\mathrm{vec}(q_{\overline{c}})\|^2}^{\sim 1}}{|p_{\overline{c}}(z)|^2 + |q_{\overline{c}}(z)|^2}.$$

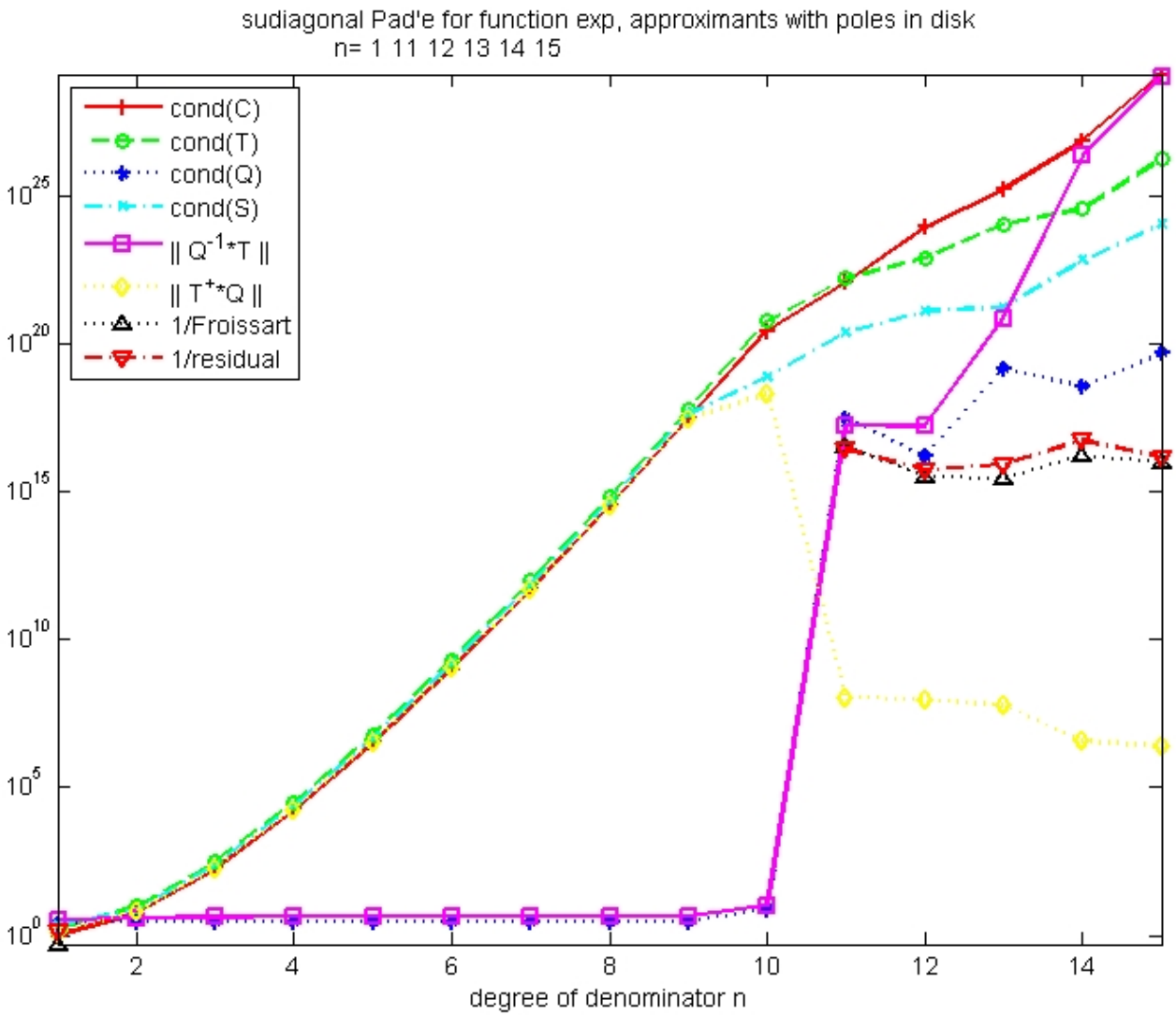
- Either modest condition number or behavior like $\gamma(K)$.
- Link between $\frac{1}{\sqrt{\gamma(K)}}$ and numerical coprimeness [L & B'98]?

Example 1: Stieltjes function $f(z) = \int_{-1}^{1} \frac{1}{1-zx} \frac{dx}{\sqrt{1-x^2}}$



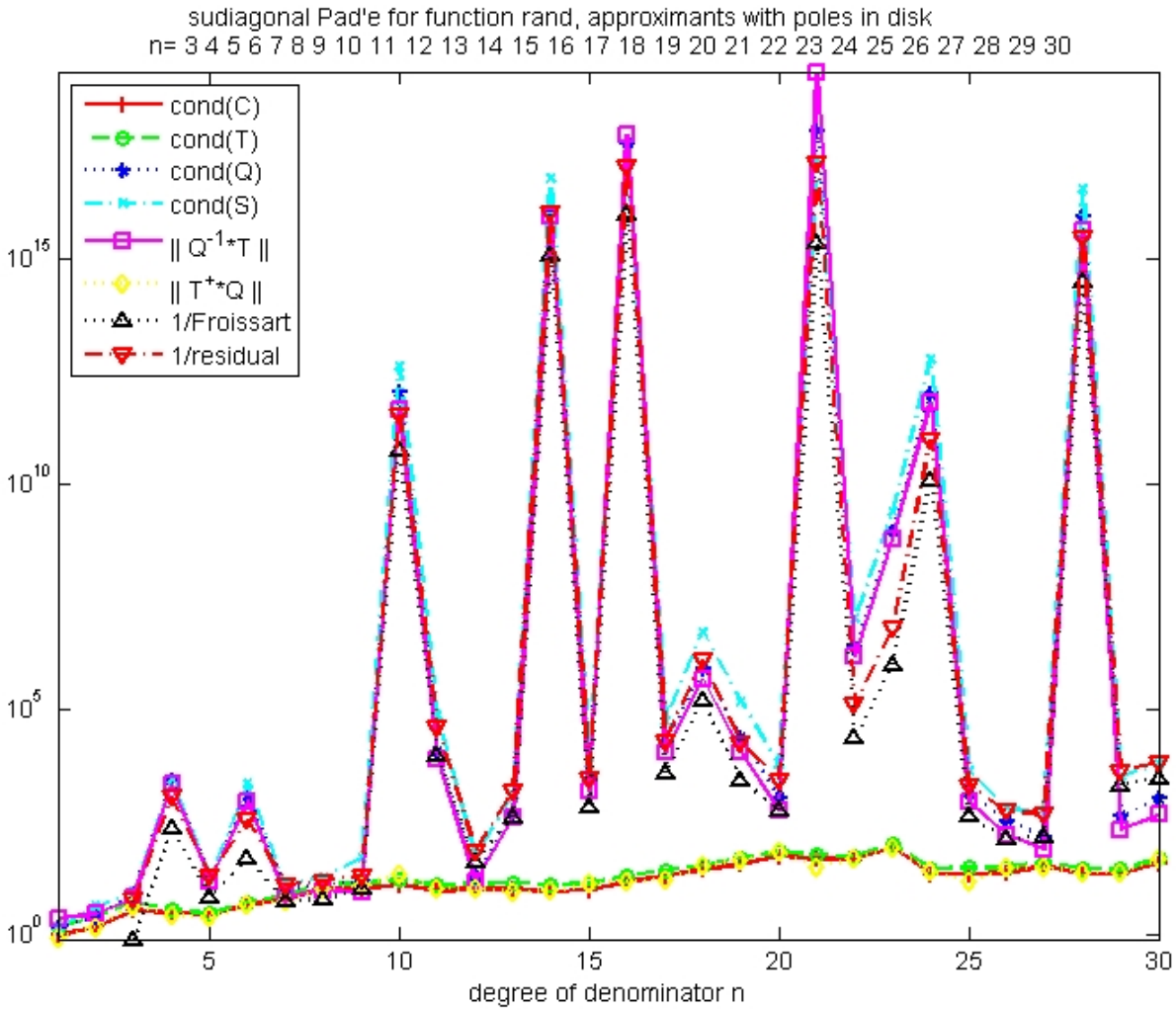
$cond(C)$ increases exponentially, but no spurious poles.

Example 2: exponential $f(z) = exp(z)$



reach machine precision for $n \geq 9$.

Example 3: random $f(z) = \sum_j c_j z^j, c_j = randn(1,1)$.



$cond(C)$ always modest, but there are spurious poles,
$1/Residual \approx 1/Froissart \approx cond(S)$

Example 4: modification of Gammel's example

[Mascarenhas'13]

For arbitrary $z_k \in \mathbb{C}, |z_k| \geq 1/3$, suitable integers $n_0 < m_0 \leq n_1 < m_1 \leq ...$, and suitable $\alpha_k \in \mathbb{C}$

$$f(z) = \sum_{k=0}^{\infty} \alpha_k \sum_{j=n_k}^{m_k-1} (z/z_k)^j = \sum_{k=0}^{\infty} \alpha_k \frac{(z/z_k)^{n_k} - (z/z_k)^{m_k}}{1 - (z/z_k)}$$

analytic in $\mathbb{D}$.

For all $m = n = n_k$ we have that $cond(C) \leq 6$, thus $[n_k|n_k] = [n_k|1]$ is robust and has pole z_k with $Residual = |z_k \alpha_k|$.

Example 4: modification of Gammel's example

[Mascarenhas'13]

For arbitrary $z_k \in \mathbb{C}, |z_k| \geq 1/3$, suitable integers $n_0 < m_0 \leq n_1 < m_1 \leq ...$, and suitable $\alpha_k \in \mathbb{C}$

$$f(z) = \sum_{k=0}^{\infty} \alpha_k \sum_{j=n_k}^{m_k-1} (z/z_k)^j = \sum_{k=0}^{\infty} \alpha_k \frac{(z/z_k)^{n_k} - (z/z_k)^{m_k}}{1 - (z/z_k)}$$

analytic in $\mathbb{D}$.

For all $m = n = n_k$ we have that $cond(C) \leq 6$, thus $[n_k|n_k] = [n_k|1]$ is robust and has pole z_k with $Residual = |z_k \alpha_k|$.

Gives negative answer to question of [GGT'12]: robust Padé approximants do not always eliminate "spurious" poles.

Summary : Beckermann and Matos (2015)

Modest condition number of Sylvester-type matrix S implies

- forward stability of Padé map
- backward stability of Padé map
- absence of Froissart doublets and small residuals ?