Working with rational functions in a numeric environment - some contributions

Ana C. Matos

Laboratoire Paul Painlevé
Université de Lille
(joint work with B. Beckermann and G. Labahn)

Journées Matrices Structurées
Université de Limoges
23 may 2019

Rational functions I

- Let $\mathbb{C}_n[z]$ be the space of polynomials of degree at most n with complex coefficients,

$$\mathbb{C}_{m,n}[z] = \left\{ \frac{p}{q}, \quad p \in \mathbb{C}_m[z], q \in \mathbb{C}_n[z], \; q \neq 0 \right\}$$

the set of rational functions.

- important role in applied mathematics: approximation (Padé approximants), analytic continuation , determining singularities, extracting information from noisy signals, sparse interpolation, exponential analysis, modelling ...
- with rational functions we solve different problems:
  - rational interpolation [Trefethen, Berrut, Cuyt, ...]
  - best uniform rational approximation [Stahl, Varga, Petrushev, ...]
  - Padé approximation [Baker, Graves-Morris, Brezinski, ...]

Rational functions I

- Let $\mathbb{C}_n[z]$ be the space of polynomials of degree at most n with complex coefficients,

$$\mathbb{C}_{m,n}[z] = \left\{ \frac{p}{q}, \quad p \in \mathbb{C}_m[z], q \in \mathbb{C}_n[z], \; q \neq 0 \right\}$$

the set of rational functions.

- important role in applied mathematics: approximation (Padé approximants), analytic continuation , determining singularities, extracting information from noisy signals, sparse interpolation, exponential analysis, modelling ...
- with rational functions we solve different problems:
  - rational interpolation [Trefethen, Berrut, Cuyt, ...]
  - best uniform rational approximation [Stahl, Varga, Petrushev, ...]
  - Padé approximation [Baker, Graves-Morris, Brezinski, ...]

Rational functions II

Some difficulties:

- How to choose the degrees for a given type of rational approximation (Padé, interpolation,...) ?
  Overshooting the degree leads to: spurious poles, Froissart doublets, poles with small residues...$\Rightarrow$ numerical instabilities
- we fix the degrees $n \in \mathbb{N}$ of numerator and $m \in \mathbb{N}$ of denominator of the rational r function we want to use for modelling / approximating. In order to have "good" numerical properties the chosen rational function
  - $r = p/q$ must be nondegenerate i.e., the polynomials p and q are co-prime, and the defect $(\min\{m - deg\, p, n - deg\, q\})$ is equal to zero;
  - r sufficiently "far" from $\mathbb{C}_{m-1,n-1}[z]$

How to ensure these properties?

Outline of the talk

Working with rational functions in a numerical environment give rise to numerical instabilities. How to prevent them?

1. Padé approximation
   - definitions, theoretical properties
   - stability issues: condition number of the Sylvester matrix to control
     - conditioning of the Padé map
     - Froissart doublets
2. Rational functions - numerical issues
   - Froissart doublets and small residues
   - How to control their existence? Give lower bounds on the distance pole-zero based on 3 different quantities
     - condition number of a Sylvester type matrix
     - numerical coprimeness of numerator and denominator polynomials
     - spherical derivative
3. Future work ⇒ towards the construction of rational functions with good numerical properties

Outline of the talk

Working with rational functions in a numerical environment give rise to numerical instabilities. How to prevent them?

1. Padé approximation
   - definitions, theoretical properties
   - stability issues: condition number of the Sylvester matrix to control
     - conditioning of the Padé map
     - Froissart doublets
2. Rational functions - numerical issues
   - Froissart doublets and small residues
   - How to control their existence? Give lower bounds on the distance pole-zero based on 3 different quantities
     - condition number of a Sylvester type matrix
     - numerical coprimeness of numerator and denominator polynomials
     - spherical derivative
3. Future work $\Rightarrow$ towards the construction of rational functions with good numerical properties

Outline of the talk

Working with rational functions in a numerical environment give rise to numerical instabilities. How to prevent them?

1. Padé approximation
   - definitions, theoretical properties
   - stability issues: condition number of the Sylvester matrix to control
     - conditioning of the Padé map
     - Froissart doublets
2. Rational functions - numerical issues
   - Froissart doublets and small residues
   - How to control their existence? Give lower bounds on the distance pole-zero based on 3 different quantities
     - condition number of a Sylvester type matrix
     - numerical coprimeness of numerator and denominator polynomials
     - spherical derivative
3. Future work ⇒ towards the construction of rational functions with good numerical properties

Outline of the talk

Working with rational functions in a numerical environment give rise to numerical instabilities. How to prevent them?

1. Padé approximation
   - definitions, theoretical properties
   - stability issues: condition number of the Sylvester matrix to control
     - conditioning of the Padé map
     - Froissart doublets
2. Rational functions - numerical issues
   - Froissart doublets and small residues
   - How to control their existence? Give lower bounds on the distance pole-zero based on 3 different quantities
     - condition number of a Sylvester type matrix
     - numerical coprimeness of numerator and denominator polynomials
     - spherical derivative
3. Future work ⇒ towards the construction of rational functions with good numerical properties

Definition: Padé approximants I

Initial data: $(c_i)_{i=0}^{m+n}$ coefficients of $f(x) \approx \sum_{i=0}^{\infty} c_i x^i, \quad (c_0 \neq 0)$
The Padé approximant of type (m, n) of f is rational function defined by

$$[m/n]_f(x) = \frac{p(x)}{q(x)} \text{ with}$$

- $p(x) = p_0 + p_1 x + \cdots + p_m x^m$, $q(x) = q_0 + q_1 x + \cdots + q_n x^n$, $q(x) \neq 0$
- $q(x)f(x) - p(x) = \mathcal{O}(x^{m+n+1})(x \to 0)$

We set the Toeplitz matrix ($c_i = 0$ if $i < 0$)

$$C = \begin{pmatrix} c_{m+1} & c_m & \cdots & c_{m+1-n} \\ c_{m+2} & c_{m+1} & \cdots & c_{m+2-n} \\ \vdots & \vdots & \vdots & \vdots \\ c_{m+n} & c_{m+n-1} & \cdots & c_m \end{pmatrix} \in \mathbb{C}^{n\times(n+1)}$$

Definitions: Padé approximants II

The coefficients of $p(x)$ and $q(x)$ are solution of a linear system:

- Denominator coefficients $\vec{q}$: solution of an homogeneous $n \times (n+1)$ system

$$C\vec{q} = 0, \quad \vec{q} = (q_0, q_1, \cdots, q_n)^T$$

- Numerator coefficients $\vec{p}$:

$$p_k = \sum_{i=0}^{m} c_{k-i} q_i \text{ for } k = 0, 1, \cdots, m$$

So there are infinitely many solutions but the rational function p/q is unique. To define uniquely polynomials p and q we impose:

- p and q are coprime;
- normalisation: $\|\vec{p}\|^2 + \|\vec{q}\|^2 = 1, \quad q(0) > 0.$

Definitions: Padé approximants III

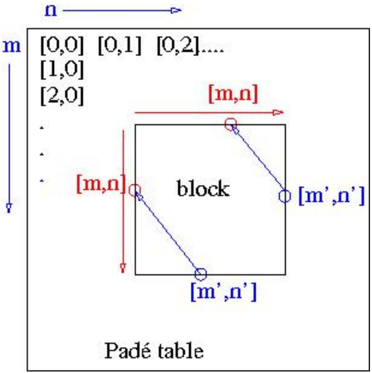
In a matrix form we can write

$$T \begin{bmatrix} \vec{p} \\ \vec{q} \end{bmatrix} = 0,$$

$$T = \begin{bmatrix} 1 & 0 & \cdots & 0 & -c_0 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \vdots & -c_1 & -c_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & -c_m & & \ddots & -c_0 \\ 0 & \cdots & \cdots & 0 & -c_{m+1} & \cdots & \cdots & -c_1 \\ \vdots & & & \vdots & \vdots & & & \vdots \\ 0 & \cdots & \cdots & 0 & -c_{m+n} & \cdots & \cdots & -c_m \end{bmatrix} \in \mathbb{C}^{(m+n+1)\times(m+n+2)},$$

Padé table

- we dispose the approximants in a double entry table
- degeneracies may occur: block structure of the table (in each block the approximants are identical) depending on the rank of C
-



- good convergence properties:
- diagonal sequences $([n|n])_n, ([n-1|n])_n$ (Stieltjes functions $f(z) = \int \frac{d\mu(x)}{x-z}$)
- columns $([m|n])_m$ (n fixed) (Montessus de Ballore for meromorphic functions)
- ray sequences $([\gamma n|n])_n$

- analytic continuation (ex: $\log(1+z)$)
- approximation of singularities

Drawbacks: Spurious poles in Padé approximants

- these approximants can have poles that don't correspond to singularities of the function called "spurious poles" (H. Stahl): z_n pole of $[n|n]$ and $\lim_{n\to\infty} z_n = z_0$ with f analytic in z_0 (asymptotic definition)
- these poles prevent uniform convergence and can be dense in $\mathbb{C}$
  $f(z) = \int_0^1 \frac{1}{1-zx}\frac{dx}{\sqrt{1-x^2}}, \quad g(z) = \int_0^1 \frac{(x-\cos(\alpha_1))(x-\cos(\alpha_2))}{1-zx}\frac{dx}{\sqrt{1-x^2}}$
  - f et g are analytic in $\Omega = \mathbb{C}\backslash[1,+\infty[$
  - the approximants $[n-1|n]$ of f converge locally uniformly to f in Ω;
  - the poles of $[n-1|n]$ of g are dense in $\mathbb{C}$
- it is important to be able to eliminate spurious poles: Gonchar's lemma: convergence in capacity + absence of poles $\Rightarrow$ uniform convergence
- for meromorphic functions we can show that to each "spurious pole" z_n of $[n|n]$ correspond a zero ξ_n such that $\lim_{n\to\infty} |z_n - \xi_n| = 0$

Drawbacks: Spurious poles in Padé approximants

- these approximants can have poles that don't correspond to singularities of the function called "spurious poles" (H. Stahl): z_n pole of $[n|n]$ and $\lim_{n\to\infty} z_n = z_0$ with f analytic in z_0 (asymptotic definition)
- these poles prevent uniform convergence and can be dense in $\mathbb{C}$
  $f(z) = \int_0^1 \frac{1}{1-zx}\frac{dx}{\sqrt{1-x^2}}, \quad g(z) = \int_0^1 \frac{(x-\cos(\alpha_1))(x-\cos(\alpha_2))}{1-zx}\frac{dx}{\sqrt{1-x^2}}$
  - f et g are analytic in $\Omega = \mathbb{C}\backslash[1,+\infty[$
  - the approximants $[n-1|n]$ of f converge locally uniformly to f in Ω;
  - the poles of $[n-1|n]$ of g are dense in $\mathbb{C}$
- it is important to be able to eliminate spurious poles: Gonchar's lemma: convergence in capacity + absence of poles $\Rightarrow$ uniform convergence
- for meromorphic functions we can show that to each "spurious pole" z_n of $[n|n]$ correspond a zero ξ_n such that $\lim_{n\to\infty} |z_n - \xi_n| = 0$

Drawbacks: Spurious poles in Padé approximants

- these approximants can have poles that don't correspond to singularities of the function called "spurious poles" (H. Stahl): z_n pole of $[n|n]$ and $\lim_{n\to\infty} z_n = z_0$ with f analytic in z_0 (asymptotic definition)
- these poles prevent uniform convergence and can be dense in $\mathbb{C}$
  $f(z) = \int_0^1 \frac{1}{1-zx}\frac{dx}{\sqrt{1-x^2}}, \quad g(z) = \int_0^1 \frac{(x-\cos(\alpha_1))(x-\cos(\alpha_2))}{1-zx}\frac{dx}{\sqrt{1-x^2}}$
  - f et g are analytic in $\Omega = \mathbb{C}\backslash[1,+\infty[$
  - the approximants $[n-1|n]$ of f converge locally uniformly to f in Ω;
  - the poles of $[n-1|n]$ of g are dense in $\mathbb{C}$
- it is important to be able to eliminate spurious poles: Gonchar's lemma: convergence in capacity + absence of poles $\Rightarrow$ uniform convergence
- for meromorphic functions we can show that to each "spurious pole" z_n of $[n|n]$ correspond a zero ξ_n such that $\lim_{n\to\infty} |z_n - \xi_n| = 0$

Drawbacks: Spurious poles in Padé approximants

- these approximants can have poles that don't correspond to singularities of the function called "spurious poles" (H. Stahl): z_n pole of $[n|n]$ and $\lim_{n\to\infty} z_n = z_0$ with f analytic in z_0 (asymptotic definition)
- these poles prevent uniform convergence and can be dense in $\mathbb{C}$
  $f(z) = \int_0^1 \frac{1}{1-zx} \frac{dx}{\sqrt{1-x^2}}, \quad g(z) = \int_0^1 \frac{(x-\cos(\alpha_1))(x-\cos(\alpha_2))}{1-zx} \frac{dx}{\sqrt{1-x^2}}$
  - f et g are analytic in $\Omega = \mathbb{C}\backslash[1, +\infty[$
  - the approximants $[n-1|n]$ of f converge locally uniformly to f in Ω;
  - the poles of $[n-1|n]$ of g are dense in $\mathbb{C}$
- it is important to be able to eliminate spurious poles: Gonchar's lemma: convergence in capacity + absence of poles $\Rightarrow$ uniform convergence
- for meromorphic functions we can show that to each "spurious pole" z_n of $[n|n]$ correspond a zero ξ_n such that $\lim_{n\to\infty} |z_n - \xi_n| = 0$

From theory to numerical analysis

- interested in computing
  - the Padé approximants (or rational functions) coefficients
  - the values of a rational function in a point
- in numerical computations:

  finite precision arithmetic + noise in the coefficients

  can amplify these phenomena $\Rightarrow$ numerical instabilities
- AIM:
  - identify the principal sources of numerical problems
  - propose some indicators of the good numerical properties of a rational function
- in a numerical setting we need to define a metric in the set of rational functions
  - should we ask that values are closed?
  - should we ask that coefficients are close? in which basis?

From theory to numerical analysis

- interested in computing
  - the Padé approximants (or rational functions) coefficients
  - the values of a rational function in a point
- in numerical computations:

  finite precision arithmetic + noise in the coefficients

  can amplify these phenomena ⇒ numerical instabilities
- AIM:
  - identify the principal sources of numerical problems
  - propose some indicators of the good numerical properties of a rational function
- in a numerical setting we need to define a metric in the set of rational functions
  - should we ask that values are closed?
  - should we ask that coefficients are close? in which basis?

From theory to numerical analysis

- interested in computing
  - the Padé approximants (or rational functions) coefficients
  - the values of a rational function in a point
- in numerical computations:

  finite precision arithmetic + noise in the coefficients

  can amplify these phenomena $\Rightarrow$ numerical instabilities
- AIM:
  - identify the principal sources of numerical problems
  - propose some indicators of the good numerical properties of a rational function
- in a numerical setting we need to define a metric in the set of rational functions
  - should we ask that values are closed?
  - should we ask that coefficients are close? in which basis?

How to measure distances in $\mathbb{C}_{m,n}[z]$?

For $r = p/q, \widetilde{r} \in \mathbb{C}_{m,n}$ we define the coefficient vector

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

always supposed to be of norm 1.

(a) Distance of coefficient vectors of norm 1 with optimal phase:

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

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

(b) 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 $\mathbb{C}_{m,n}[z]$?

For $r = p/q, \widetilde{r} \in \mathbb{C}_{m,n}$ we define the coefficient vector

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

always supposed to be of norm 1.

(a) Distance of coefficient vectors of norm 1 with optimal phase:

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

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

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

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

Definition of the Sylvester type matrix S

We fix $n, m \in \mathbb{N}$

$$S = \left[\begin{array}{cccccccc} q_0 & 0 & \cdots & 0 & p_0 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots & \vdots & \ddots & \ddots & \vdots \\ q_n & & \ddots & 0 & p_m & & \ddots & 0 \\ 0 & \ddots & & q_0 & 0 & \ddots & & p_0 \\ \vdots & \ddots & \ddots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & q_n & 0 & \cdots & 0 & p_m \end{array}\right], \quad S = S(q,p) \in \mathbb{C}^{(m+n+1)\times(m+n+2)}$$

Sylvester type matrix of two polynomials built from the coefficients of the numerator and denominator of the rational function.
The Sylvester matrix is denoted $S_*(q,p) \in \mathbb{C}^{(m+n)\times(m+n)}$

Remark

S has full rank iff p and q are coprime and $p_m \neq 0$ or $q_n \neq 0$

Equivalence of the two distances

Theorem

Let $r = p/q$ be nondegenerate, then for all $\widetilde{r} \in \mathcal{C}_{m,n}[z]$

$$\frac{1}{\mathrm{cond}(S)}\, d(r, \widetilde{r}) \lesssim \chi_{\mathbb{D}}(r, \widetilde{r}) \lesssim \mathrm{cond}(S)\, d(r, \widetilde{r})$$

Notation: for simplicity we set $a_1 \lesssim a_2$ meaning that there exist modest constants $b, r > 0$ not depending on m, n such that $a_1 \leq b(m+n+1)^r\, a_2$.

$\Rightarrow$ for a modest value of $\mathrm{cond}(S)$ the two distances are equivalent

Sources of numerical instabilities

1. block structure of the Padé table: degeneracy (rank deficient matrices)
   - near singular systems, numerical rank of the matrix
   - small perturbations can fracture the block - ill posed problems

   $\Rightarrow$ study the stability of the Padé maps

2. Froissart doublets in rational functions: pair zero-pole (z_p, z_q) "sufficiently" close and such that z_q doesn't correspond to a singularity of the function that r represents or approach (numerical counterpart of the spurious poles)

3. small residues in the partial fraction decomposition

Sources of numerical instabilities

1. block structure of the Padé table: degeneracy (rank deficient matrices)
   - near singular systems, numerical rank of the matrix
   - small perturbations can fracture the block - ill posed problems

   $\Rightarrow$ study the stability of the Padé maps
2. Froissart doublets in rational functions: pair zero-pole (z_p, z_q) "sufficiently" close and such that z_q doesn't correspond to a singularity of the function that r represents or approach (numerical counterpart of the spurious poles)
3. small residues in the partial fraction decomposition

Next

In order to prevent numerical instabilities , some numerical analysis to help in choosing approximants with good numerical properties:

1. Stability issues for Padé approximation
    1. non degenerate approximants
    2. conditioning of the Padé maps
    3. robust Padé approximants
2. Froissart doublets and small residues
    1. why we want to eliminate them?
    2. how to control their presence?
3. how to use these results?

Next

In order to prevent numerical instabilities , some numerical analysis to help in choosing approximants with good numerical properties:

1. Stability issues for Padé approximation
    1. non degenerate approximants
    2. conditioning of the Padé maps
    3. robust Padé approximants
2. Froissart doublets and small residues
    1. why we want to eliminate them?
    2. how to control their presence?
3. how to use these results?

1 - Stability Issues for Padé approximation

The Padé map

$$F : \mathbb{C}^{m+n+1} \ni c = (c_0, \ldots, c_{m+n})^T \mapsto y = \begin{bmatrix} \vec{p} \\ \vec{q} \end{bmatrix} \in \mathbb{C}^{m+n+2}$$

mapping the vector of $(m+n+1)$ Taylor coefficients to the coefficient vector in the basis of monomials of the numerator and denominator of an $[m|n]$ Padé approximant p/q

Uniqueness obtained by:

- p and q have no common divisor;
- normalization:

$$\|F(c)\|^2 = \|\vec{p}\|^2 + \|\vec{q}\|^2 = 1, \quad q(0) > 0.$$

Theorem [Werner, Wuytack '83]

F is continuous in a neighborhood of c if and only if its $[m|n]$ Padé approximant $F(c)$ is nondegenerate i.e.,

$$defect = \min(m - deg\, p, n - deg\, q) = 0.$$

The Padé map

$$F : \mathbb{C}^{m+n+1} \ni c = (c_0, \dots, c_{m+n})^T \mapsto y = \begin{bmatrix} \vec{p} \\ \vec{q} \end{bmatrix} \in \mathbb{C}^{m+n+2}$$

mapping the vector of $(m+n+1)$ Taylor coefficients to the coefficient vector in the basis of monomials of the numerator and denominator of an $[m|n]$ Padé approximant p/q

Uniqueness obtained by:

- p and q have no common divisor;
- normalization:

$$\|F(c)\|^2 = \|\vec{p}\|^2 + \|\vec{q}\|^2 = 1, \quad q(0) > 0.$$

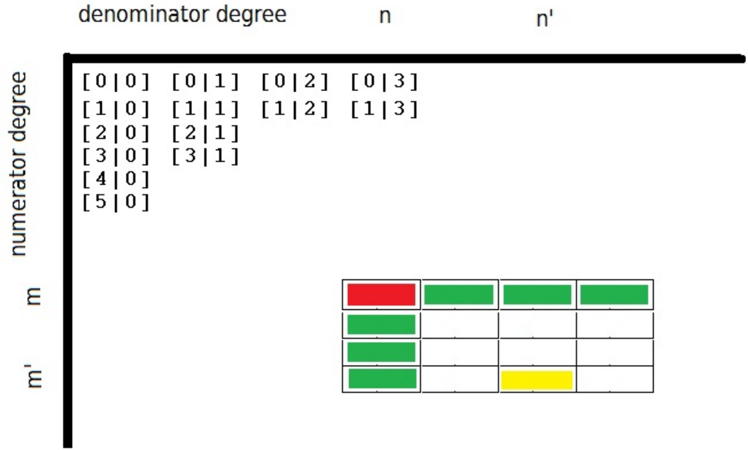
Theorem [Werner, Wuytack '83]

F is continuous in a neighborhood of c if and only if its $[m|n]$ Padé approximant $F(c)$ is nondegenerate i.e.,

$$\mathit{defect} = \min(m - \mathit{deg}\, p, n - \mathit{deg}\, q) = 0.$$

Structures in Padé table - degeneracy

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



In red and green: nondegenerate approximants (at least one degree exact).

The Padé map: conditioning

Hypothesis: *real Padé map*: $F(c) \in \mathbb{R}^{m+n+2}$, p/q is nondegenerate.
We want to study for $y = F(c)$ the perturbed equation

$$\widetilde{y} = F(\widetilde{c}) + \eta, \quad \|\eta\| = \mathrm{dist}(\widetilde{y}, F(\mathbb{R}^{m+n+1})).$$

1. Forward conditioning $\kappa_{for}(F)$: 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/for all z in the closed unit disk $\mathbb{D}$?
2. Backward conditioning $\kappa_{back}(F)$: Does a closeby vector of coefficients represent a Padé approximant of a closeby vector of Taylor coefficients?

The Padé map: conditioning

Hypothesis: *real Padé map*: $F(c) \in \mathbb{R}^{m+n+2}$, p/q is nondegenerate.
We want to study for $y = F(c)$ the perturbed equation

$$\widetilde{y} = F(\widetilde{c}) + \eta, \quad \|\eta\| = \mathrm{dist}(\widetilde{y}, F(\mathbb{R}^{m+n+1})).$$

1. Forward conditioning $\kappa_{for}(F)$: 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/for all z in the closed unit disk $\mathbb{D}$?
2. Backward conditioning $\kappa_{back}(F)$: Does a closeby vector of coefficients represent a Padé approximant of a closeby vector of Taylor coefficients?

Some related matrices

Definition

$$Q = \begin{bmatrix} q_0 & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & & & \vdots \\ q_n & & q_0 & 0 & \cdots & 0 \\ 0 & \ddots & & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & & \ddots & 0 \\ 0 & \cdots & 0 & q_n & \cdots & q_0 \end{bmatrix}, \quad Q \in \mathbb{C}^{(m+n+1)\times(m+n+1)}$$

lower triangular matrix

q_i are the coefficients of the denominator polynomial.

The Padé map: conditioning

Theorem [Beckermann, AM '13]

Suppose that F is continuous in a neighborhood of $c \in \mathbb{R}^{m+n+1}$. Then we have the following amplification of real errors

$$\kappa_{for}(F)(\overline{c}) := \limsup_{c \to \overline{c}} \frac{\|F(c) - F(\overline{c})\|}{\|F(\overline{c})\|} \Big/ \frac{\|c - \overline{c}\|}{\|c\|} = \|T^{\dagger}Q\| = \|S^{\dagger}Q^{2}\|.$$

$$\kappa_{back}(F)(\overline{c}) := \limsup_{F(c) \to F(\overline{c})} \frac{\|c - \overline{c}\|/\|\overline{c}\|}{\|F(c) - F(\overline{c})\|/\|F(\overline{c})\|} = \|Q^{-1}T\| = \|Q^{-2}S\|.$$

Consequences:

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

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

- Thus
  - $cond(T) \sim \|C^\dagger\|$ modest $\Longrightarrow$ the real Padé map is forward well-conditioned
  - $cond(Q)$ modest $\Longleftrightarrow$ backward well-conditioned.

Sufficient condition:
$cond(S)$ of modest size $\Rightarrow$ Padé map forward and backward well-conditioned

Consequences in Padé approximation

Theorem

Let $r = p/q \in \mathbb{C}_{m,n}[z]$ be nondegenerate and $\widetilde{r} = \widetilde{p}/\widetilde{q} \in \mathbb{C}_{m-1,n-1}[z]$. Then

$$2\,\chi_{\mathbb{D}}(r, \widetilde{r})\, cond(S)^2 \geq (m+n+1)^{-2}$$

- suppose f can be well approximated by some element $\widetilde{r}$ of $\mathbb{C}_{m-1,n-1}[z]$ with respect to the uniform chordal metric in the unit disk : $\chi_{\mathbb{D}}(f, \widetilde{r})$ is small;
- its $[m|n]$ Padé approximant r either does not have a small approximation error $\chi_{\mathbb{D}}(f, r)$, or otherwise cond(S) is necessarily "large".
- this can lead to an early stopping criterion in Padé approximants if we want only to compute well-conditioned rational functions $\Rightarrow$ stability issues

Consequences in Padé approximation

Theorem

Let $r = p/q \in \mathbb{C}_{m,n}[z]$ be nondegenerate and $\widetilde{r} = \widetilde{p}/\widetilde{q} \in \mathbb{C}_{m-1,n-1}[z]$. Then

$$2\,\chi_{\mathbb{D}}(r, \widetilde{r})\mathit{cond}(S)^2 \geq (m+n+1)^{-2}$$

- suppose f can be well approximated by some element $\widetilde{r}$ of $\mathbb{C}_{m-1,n-1}[z]$ with respect to the uniform chordal metric in the unit disk : $\chi_{\mathbb{D}}(f, \widetilde{r})$ is small;
- its $[m|n]$ Padé approximant r either does not have a small approximation error $\chi_{\mathbb{D}}(f, r)$, or otherwise cond(S) is necessarily "large".
- this can lead to an early stopping criterion in Padé approximants if we want only to compute well-conditioned rational functions $\Rightarrow$ stability issues

Well conditioned rational functions

cond(S) has an important role

Definition

a rational function $r = p/q$ is well-conditioned if the corresponding matrix $S(p, q)$ has a modest condition number.

Well conditioned rational functions

$$\Downarrow$$

- forward and backward stability of Padé map
- equivalence of the distances $d(r, \tilde{r})$ and $\chi_K(r, \tilde{r})$
- stop criterium in the computation of sequence of Padé approximants
- no presence of Froissart doublets and small residuals ?

Well conditioned rational functions

cond(S) has an important role

Definition

a rational function $r = p/q$ is well-conditioned if the corresponding matrix $S(p,q)$ has a modest condition number.

Well conditioned rational functions

$\Downarrow$

- forward and backward stability of Padé map
- equivalence of the distances $d(r,\tilde{r})$ and $\chi_K(r,\tilde{r})$
- stop criterium in the computation of sequence of Padé approximants
- no presence of Froissart doublets and small residuals ?

Froissart doublets

2 - Controlling Froissart doublets and small residues

What are Froissart doublets? [Gilewicz & Pindor '97-99, Bessis '96]

Definition

Let $r = p/q \in \mathbb{C}_{m,n}[z]$. A Froissart doublet is a pair zero-pole (z_p, z_q) "sufficiently" close and such that z_q doesn't correspond to a singularity of the function that r represents or approach.

- associated with the occurence of small residuals a_k corresponding to terms $\frac{a_k}{z-z_k}$ in partial fraction decomposition of the approximant$\Rightarrow$ problems in computing the values of the function near z_k.

Why do we want to eliminate them?

- theoretical issues: uniform convergence
- practical issues - modelling noise: if $f \in \mathbb{C}_{n-1,n}(z)$, in presence of noise $\Rightarrow f(z) + \epsilon(z) = \sum_{j=0}^{\infty}(f_j + \epsilon_j)z^j$
  From theoretical results on the convergence of Padé approximants

  $$[m-1/m] \rightarrow_{m\rightarrow\infty} f(z) + \epsilon(z)$$

  $\Rightarrow$ noise can be modelled by the $(m-n)$ spurious poles which come along with $(m-n)$ close zeros $\Rightarrow$ we can filter the noise by identifying and eliminating the Froissart doublets (unstable poles) (A.Cuyt & al.)
- numerical instabilities in the computation of the value of the function: small change in arguments give rise to large variation of the function values - large Lipschitz constant

Why do we want to eliminate them?

- theoretical issues: uniform convergence
- practical issues - modelling noise: if $f \in \mathbb{C}_{n-1,n}(z)$, in presence of noise $\Rightarrow f(z) + \epsilon(z) = \sum_{j=0}^{\infty}(f_j + \epsilon_j)z^j$
  From theoretical results on the convergence of Padé approximants

  $$[m-1/m] \to_{m\to\infty} f(z) + \epsilon(z)$$

  $\Rightarrow$ noise can be modelled by the $(m-n)$ spurious poles which come along with $(m-n)$ close zeros $\Rightarrow$ we can filter the noise by identifying and eliminating the Froissart doublets (unstable poles) (A.Cuyt & al.)
- numerical instabilities in the computation of the value of the function: small change in arguments give rise to large variation of the function values - large Lipschitz constant

Some definitions

- recall the uniform chordal metric: $K \subset \mathbb{C}$ compact set, r,$\tilde{r}$ rational functions

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

well adapted to study (uniform) convergence questions (since meromorphic functions are continuous on the Riemann sphere)

- Lipschitz constants

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

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

Consequences: numerical instabilities

- $\chi(r(z_p), r(z_q)) = 1$ if $p(z_p) = 0, q(z_q) = 0$, and so if $z_p, z_q \in K$,

$$L_K(r) \geq \frac{1}{\chi(z_p, z_q)}$$

$\Rightarrow$ very large Lipschitz constant if there is a Froissart doublet in K.

- if $z_q \in K$ is a simple pole then

$$\rho(r)(z_q) = \frac{1}{\mathrm{res}(z_q)} \leq \rho_K(r)$$

$\Rightarrow \rho_K(r)$ is large if there is a small residue

small variations in the argument can be amplified in the computation of function values.

Consequences: numerical instabilities

- $\chi(r(z_p), r(z_q)) = 1$ if $p(z_p) = 0, q(z_q) = 0$, and so if $z_p, z_q \in K$,

$$L_K(r) \geq \frac{1}{\chi(z_p, z_q)}$$

$\Rightarrow$ very large Lipschitz constant if there is a Froissart doublet in K.

- if $z_q \in K$ is a simple pole then

$$\rho(r)(z_q) = \frac{1}{\mathrm{res}(z_q)} \leq \rho_K(r)$$

$\Rightarrow$ $\rho_K(r)$ is large if there is a small residue

small variations in the argument can be amplified in the computation of function values.

How to control the existence of Froissart doublets?

AIM: find lower bounds for the distance zero-pole

$$\left\{ \begin{array}{c} |z_p - z_q| \\ \\ \chi(z_p, z_q) \end{array} \right\} \gtrsim \left\{ \begin{array}{l} 1/\mathrm{cond}(S(p,q)) \\ {\epsilon_i}^K(p,q) \text{ numerical coprimeness of } p, q \end{array} \right.$$

- we set $a_1 \lesssim a_2$ meaning that there exist modest constants $b, r > 0$ not depending on m, n such that $a_1 \leq b(m+n+1)^r a_2$.

How to control the existence of Froissart doublets?

1. Conditioning of the Sylvester type matrix $S(p, q)$
2. Numerical coprimeness $\epsilon_i^K(p, q)$

Lower bounds for the distance pole-zero based on $\mathrm{cond}(S(p,q))$

Theorem [Beckermann, AM]

Let $r \in \mathbb{C}_{m,n}[z]$ be such that $r = p/q$ is nondegenerate. Then the distance of any couple of zeros and poles (z_p, z_q) of r in the unit disk is bounded below by

$$|z_p - z_q| \gtrsim 1/\mathrm{cond}(S).$$

here $\mathrm{cond}(S) = \| S \|_2 \| S^\dagger \|_2$

$\Rightarrow$ for a modest condition number of S there are no Froissart doublets

Robustness for cond$(S(p,q))$

the indicators are not sensitive with respect to a small perturbation of the numerator and denominator

Theorem [Beckermann, AM]

Let $K \subset \mathbb{C}$ and $\frac{p}{q}, \frac{\widetilde{p}}{\widetilde{q}} \in \mathbb{C}_{m,n}[z]$. If $\frac{p}{q}$ is nondegenerate and

$$\|(p-\widetilde{p}, q-\widetilde{q})\|_2 \leq \frac{1}{3\sqrt{m+n+1}\,\|S(p,q)^{\dagger}\|_2}$$

then $\frac{1}{2} \leq \mathrm{cond}(S(\widetilde{p},\widetilde{q}))/\mathrm{cond}(S(p,q)) \leq 2.$

Furthermore, let $z_p, z_q \in \mathbb{C}$ with $\widetilde{p}(z_p) = \widetilde{q}(z_q) = 0$. Then,

$$|z_p - z_q| \geq \frac{1}{6\sqrt{2}(m+n+1)^{3/2}\mathrm{cond}(S(p,q))},$$

Lower bounds on residuals

Theorem [Beckermann, AM]

Let $r \in \mathbb{C}_{m,n}[z]$ be such that $r = p/q$ is nondegenerate. Then the modulus of any residual of a simple pole z_q of r in the unit disk is bounded below by

$$\mathrm{res}(z_q) \gtrsim 1/\ \mathrm{cond}(S).$$

Moreover this result is still true for any rational function $\widetilde{r} \in \mathbb{C}_{m,n}[z]$ in a neighbourhood of r, with $\chi_{\mathbb{D}}(r, \widetilde{r}) \lesssim 1/\mathrm{cond}(S)^2$,

$\Rightarrow$ if $\chi_{\mathbb{D}}(r, \widetilde{r})$ is sufficiently small then

r has a Froissart doublet iff $\widetilde{r}$ has one.

Summary of results so far: what can cond($S(p, q)$) control?

Well conditioned rational functions $\Leftrightarrow$
cond($S(p, q)$) of moderate size

$$\Downarrow$$

- forward and backward stability of Padé map
- equivalence of the distances $d(r, \tilde{r})$ and $\chi_K(r, \tilde{r})$
- stop criterium in the computation of sequence of Padé approximants
- no presence of Froissart doublets and small residuals

How to control the existence of Froissart doublets?

1. Conditioning of the Sylvester type matrix $S(p,q)$
2. Numerical coprimeness $\epsilon_i^K(p,q)$

How to determine coprimeness of two numeric polynomials?

$$c(z) = c_0 + c_1 z + \cdots + c_n z^n, \quad \rightsquigarrow \vec{c} = (c_0, c_1, \cdots, c_n)^T$$

Definition

let $p \in \mathbb{C}_m[z], q \in \mathbb{C}_n[z]$

$$\boxed{\begin{aligned}\epsilon_i(p,q) = \quad & \inf\{\| (p - p^*, q - q^*) \|_i : (p^*, q^*) \in \mathbb{C}_m[z] \times \mathbb{C}_n[z] \text{ have} \\ & \text{a common root, }\}, i = 1, 2\end{aligned}}$$

with

$$\begin{cases} \| (p,q) \|_1 = \max(\| \vec{p} \|_1, \| \vec{q} \|_1) \\ \| (p,q) \|_2 = \sqrt{\sum_{j=0}^{m} |p_j|^2 + \sum_{j=0}^{n} |q_j|^2} \end{cases}$$

How to determine coprimeness of two numeric polynomials

Lemma (relationship with Sylvester matrix)

$$\begin{aligned}
\epsilon_1(p,q) &= \inf\{\| S_*(p,q) - S_*(\tilde{p},\tilde{q}) \|_1 : \quad S_*(\tilde{p},\tilde{q}) \; \textit{singular}\} \geq \\
&\geq \min\{\| S_*(p,q) - B \|_1 : \quad B \; \textit{singular}\} = \| S_*(p,q)^{-1} \|_1^{-1} \\
\epsilon_2(p,q) &\geq 1/(\sqrt{m+m+1}) \| S(p,q)^{\dagger} \|_2^{-1}
\end{aligned}$$

- $\| (p,q) \|_1 / \epsilon_1(p,q)$ is a structured condition number of $S_*(p,q)$ in the class of Sylvester matrices
- if we perturb the coefficients of the polynomials by $\delta < 1/ \| S_*(p,q)^{-1} \|$ we still have coprime polynomials
- as $\| S(p,q) \|_2$ is not far from $\| (p,q) \|_2$, then $\epsilon_2(p,q)$ is a kind of smallest structured singular value

How to determine coprimeness of two numeric polynomials

Lemma (relationship with Sylvester matrix)

$$\begin{aligned}\epsilon_1(p,q) &= \inf\{\| S_*(p,q) - S_*(\tilde p,\tilde q)\|_1 : \quad S_*(\tilde p,\tilde q) \textit{ singular}\} \geq \\ &\geq \min\{\| S_*(p,q) - B\|_1 : \quad B \textit{ singular}\} = \| S_*(p,q)^{-1}\|_1^{-1} \\ \epsilon_2(p,q) &\geq 1/(\sqrt{m+m+1})\ \| S(p,q)^{\dagger}\|_2^{-1}\end{aligned}$$

- $\| (p,q)\|_1 / \epsilon_1(p,q)$ is a structured condition number of $S_*(p,q)$ in the class of Sylvester matrices
- if we perturb the coefficients of the polynomials by $\delta < 1/\| S_*(p,q)^{-1}\|$ we still have coprime polynomials
- as $\| S(p,q)\|_2$ is not far from $\| (p,q)\|_2$, then $\epsilon_2(p,q)$ is a kind of smallest structured singular value

Another expression for $\epsilon_i(p,q)$

Definition

For $K \subset \mathbb{C}$ we set

$$\epsilon_1^K(p,q) := \inf_{z\in K} \max\left\{ \frac{|p(z)|}{\max(1,|z|^m)}, \frac{|q(z)|}{\max(1,|z|^n)} \right\} =$$

$$\epsilon_2^K(p,q) := \inf_{z\in K} \left(\frac{|p(z)|^2}{\sum_{i=0}^m |z|^{2i}} + \frac{|q(z)|^2}{\sum_{i=0}^n |z|^{2i}} \right)^{1/2}$$

$\Rightarrow$ minimisation with only one parameter

Theorem [Beckermann & Labahn '98]

$$\epsilon_i(p,q) = \epsilon_i^{\overline{\mathbb{C}}}(p,q)$$

.

Link between Froissart doublets and numerical coprimeness

Theorem [Beckermann, Labanh, AM]

Let $K \subset \mathbb{C}$ and consider two polynomials $(p, q) \in \mathbb{C}_m[z] \times \mathbb{C}_n[z]$ defining a non degenerate rational function $r = p/q$. Let $z_p, z_q \in K$ such that $p(z_p) = q(z_q) = 0$. Then

$$\chi(z_p, z_q) \geq \frac{1}{2} \frac{\epsilon_i^K(p, q)}{\max\left(m \parallel \vec{p} \parallel_i, n \parallel \vec{q} \parallel_i\right)} \quad i = 1, 2$$

- (p, q) numerically relatively prime $\Rightarrow$ $r = p/q$ doesn't have Froissart doublets.
- this inequality is sharper than the one involving $\mathrm{cond}(S(p, q))$
- $\epsilon_i^K(p, q)$ and $\mathrm{cond}(S(p, q))$ can be of different order.

Link between residuals and numerical coprimeness

Theorem [Beckermann, Labahn, AM]

Let z_q be a simple pole of $r = p/q$ in $\mathbb{D}$,
Then the residual of z_q, $\text{res}(z_q)$, is bounded by

$$\text{res}(z_q) \geq \frac{\epsilon_1^{\mathbb{D}}(p,q)}{(m+n) \parallel (p,q) \parallel_1}$$

(p,q) numerically coprime $\Rightarrow r = p/q$ doesn't have small residuals

Robustness for $\epsilon_i(p, q)$

this indicator is not sensitive with respect to a small perturbation of the numerator and denominator

Theorem [Beckermann, Labahn, AM]

Let $K \subset \mathbb{C}$ and $\frac{p}{q}, \frac{\widetilde{p}}{\widetilde{q}} \in \mathbb{C}_{m,n}[z]$. Let $i \in \{1, 2\}$. If

$$\|(p - \widetilde{p}, q - \widetilde{q})\|_i \leq \frac{1}{2}\epsilon_i^K(p, q)$$

then $\frac{1}{2} \leq \epsilon_i^K(\widetilde{p}, \widetilde{q})/\epsilon_i^K(p, q) \leq 3/2$.

Furthermore, let $z_p, z_q \in \mathbb{C}$ with $\widetilde{p}(z_p) = \widetilde{q}(z_q) = 0$. Then

$$\chi(z_p, z_q) \geq \frac{\epsilon_i^K(p, q)}{6(m + n)\,\|(p, q)\|_i}.$$

An alternative way of defining "good" properties for rational functions

- $$\epsilon_i^{\mathbb{D}}(p,q) \gtrsim \frac{1}{\mathrm{cond}(S(p,q))}$$
- quantities $\mathrm{cond}(S(p,q)$ and $\epsilon_i^{\mathbb{D}}(p,q)$ can be of different order
- we can define a larger class of rational functions with numerator and denominator being numerically co-prime in the sense of $\epsilon_i^{\mathbb{D}}$ not too small that do not have neither Froissart doublets nor small residues.

How to control the existence of Froissart doublets?

3 - Estimates with the spherical derivative $\rho_K(r)$

Estimate with spherical derivative

Recall

$$\rho(r)(z) := \frac{|r'(z)|}{1+|r(z)|^2}(\text{spherical derivative})$$

and define

$$\rho_K(r) := \sup_{z\in K} \rho(r)(z)$$

Theorem

Let $K \subset \mathbb{C}$ and $r = \frac{p}{q} \in \mathbb{C}_{m,n}[z]$ with p and q coprime and $z_p, z_q \in \mathbb{C}$ with $p(z_p) = q(z_q) = 0$.
If K is convex then

$$\rho_K(r) = \sup_{z_1,z_2\in K} \frac{\chi(r(z_1), r(z_2))}{|z_1 - z_2|}.$$

In particular, $|z_p - z_q| \geq \frac{1}{\rho_K(r)}$, $\quad \operatorname{res}(z_q) \geq \frac{1}{\rho_K(r)}$

Comparing with numerical coprimeness

Theorem

Let $K \subset \mathbb{C}$ and $r = \frac{p}{q} \in \mathbb{C}_{m,n}[z]$. If $K \subset \mathbb{D}$ or $m = n$ then

$$\frac{1}{2} \frac{\epsilon_1^K(p,q)}{\max(\ m \parallel \vec{p} \parallel_1,\ n \parallel \vec{q} \parallel_1\ )} \leq \frac{1}{\rho_K(r)}.$$

This estimate can be sharper as

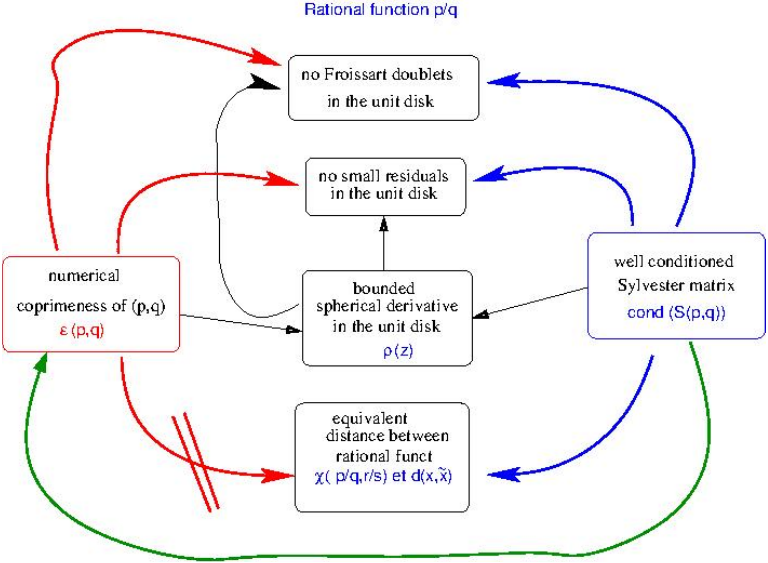
$$\epsilon_K(p^m, q^m) = \epsilon_K(p,q)^m, \quad \nu_K(r^m) \leq m\,\nu_K(r).$$

Example

Consider $r = (\frac{p}{q})^m$ for $p(z) = z$, $q(z) = \frac{z-1}{2}$ with $m \geq 0$ an integer.

$$\epsilon_1(p^m, q^m) = \epsilon_1(p,q)^m = 3^{-m} \quad \text{and } \rho_K(r) \leq 2m\,\rho_K(\frac{p}{q}) = \frac{9m}{2}.$$

Summary of results



How can we use these results?

Constructing new rational approximants

- propose an easily computable estimate $E(p,q)$ of one of the previous quantities $\mathrm{cond}(S(p,q))$, $1/\epsilon_i^K(p,q)$, $\nu_K(r)$
- prevent from computing "bad" approximants: don't use the function $\frac{p}{q}$ if $E(p,q)$ is large.
- construct approximants with a penalizing term

Example: Padé approximants satisfy $T\begin{pmatrix}\vec{p}\\ \vec{q}\end{pmatrix} = 0$ for a matrix T constructed from the series coefficients.
Construct a regularized approximant satisfying the optimization problem

$$\min_{(\vec{p},\vec{q})}\left(\parallel T\begin{pmatrix}\vec{p}\\ \vec{q}\end{pmatrix}\parallel + \rho E(p,q)\right)$$

with ρ a penalization factor.

How can we use these results?

Constructing new rational approximants

- propose an easily computable estimate $E(p, q)$ of one of the previous quantities $\mathrm{cond}(S(p, q))$, $1/\epsilon_i^K(p, q)$, $\nu_K(r)$
- prevent from computing "bad" approximants: don't use the function $\frac{p}{q}$ if $E(p, q)$ is large.
- construct approximants with a penalizing term

  Example: Padé approximants satisfy $T \begin{pmatrix} \vec{p} \\ \vec{q} \end{pmatrix} = 0$ for a matrix T constructed from the series coefficients.

  Construct a regularized approximant satisfying the optimization problem

$$\min_{(\vec{p},\vec{q})} \left(\| \, T \begin{pmatrix} \vec{p} \\ \vec{q} \end{pmatrix} \| + \rho E(p, q) \right)$$

with ρ a penalization factor.

How can we use these results? Future work

Future work:

- consider other polynomial basis (Tchebyshev, Legendre ...)
- representation of a rational function in a barycentric form
- generalize to multivarite approximation
- ....

THANK YOU !

- B. Beckermann, G. Labahn , A. Matos, On rational functions without Froissart doublets *(submitted)* http://arxiv.org/abs/1605.00506
- B. Beckermann, A. Matos, Algebraic properties of robust Padé approximants *Journal of Approx. Theory* **190**, 91-115 (2015)
- B. Beckermann, G. Labahn, When are two numerical polynomials relatively prime? *J. Symbolic Computations* **26**, 677-689 (1998).
- S. Cabay and R. Meleshko, A weakly stable Algorithm for Padé Approximants and the Inversion of Hankel matrices, *SIAM J. Matrix Analysis and Applications* **14** (1993) 735-765.
- R.M. Corless, P.M. Gianni, B.M. Trager, S.M. Watt , The singular value decomposition for polynomial systems, *Proceedings ISSAC'95*,ACM Press, 195–207 (1995)
- J. Gilewicz and M. Pindor, Padé approximants and noise: a case of geometric series, J. Comput. Appl. Math., 87 (1997), pp. 199-214.
- J. Gilewicz and M. Pindor, Padé approximants and noise: rational functions, J. Comput. Appl. Math., 105 (1999), pp. 285-297.