Algebraic properties of robust Padé approximants

Bernhard Beckermann, Ana Matos

To cite this version:

Bernhard Beckermann, Ana Matos. Algebraic properties of robust Padé approximants. 2013. hal-00871140v1

HAL Id: hal-00871140
https://hal.archives-ouvertes.fr/hal-00871140v1

Submitted on 9 Oct 2013 (v1), last revised 29 Nov 2013 (v2)
ALGEBRAIC PROPERTIES OF ROBUST PADÉ APPROXIMANTS
BERNHARD BECKERMANN & ANA C. MATOS∗

Abstract. For a recent new numerical method for computing so-called robust Padé approximants through SVD techniques, the authors gave numerical evidence that such approximants are insensitive to perturbations in the data, and do not have so-called spurious poles, that is, poles with a close-by zero or poles with small residuals. A black box procedure for eliminating spurious poles would have a major impact on the convergence theory of Padé approximants since it is known that convergence in capacity plus absence of poles in some domain D implies locally uniform convergence in D.

In the present paper we provide a proof for forward stability (or robustness), and show absence of spurious poles for the subclass of so-called well-conditioned Padé approximants. We also give a numerical example of some robust Padé approximant which has spurious poles, and discuss related questions. It turns out that it is not sufficient to discuss only linear algebra properties of the underlying rectangular Toeplitz matrix, since in our results other matrices like Sylvester matrices also occur. These types of matrices have been used before in numerical greatest common divisor computations.

Key words. Padé approximation, SVD, regularization, Froissart doublet, spurious poles.

AMS subject classifications. 41A21, 65F22

1. Introduction and statement of the main results. A popular method for approximation, for analytic continuation or for detection of singularities of a function f knowing a (formal) power series expansion at zero \( f(z) \approx \sum_{j=0}^{\infty} c_j z^j \) is to compute its \([m/n]\) Padé approximant at zero \( p/q \), namely a rational function satisfying

\[
p(z) = \sum_{j=0}^{m} p_j z^j, \quad q(z) = \sum_{j=0}^{n} q_j z^j \neq 0, \quad \text{with expansion} \quad f(z)q(z) - p(z) = \sum_{j=m+n+1}^{\infty} c_j z^j. \tag{1.1}
\]

It is well known [2] Section 1 that there always exists an \([m/n]\) Padé approximant: we just have to find a non-trivial solution of the homogeneous system of \( n \) equations and \( n + 1 \) unknowns with Toeplitz structure

\[
C \operatorname{vec}(q) = 0, \quad C = \begin{bmatrix}
c_{m+1} & \cdots & c_{m-n+2} & c_{m-n+1} \\
c_{m+2} & \cdots & c_{m-n+3} & c_{m-n+2} \\
\vdots & \ddots & \vdots & \vdots \\
c_{m+n} & \cdots & c_{m+1} & c_{m}
\end{bmatrix}, \quad \operatorname{vec}(q) = \begin{bmatrix}
q_0 \\
q_1 \\
\vdots \\
q_n
\end{bmatrix} \tag{1.2}
\]

with the convention \( c_j = 0 \) for \( j < 0 \), and then find the coefficients of \( p \) from (1.1). Whereas (1.2) has infinitely many solutions, it is also known [2] that the rational function \( p/q \) is unique.

Though many theoretical results [2] Section 6 show the usefulness of sequences of Padé approximants in approximating \( f \) or its singularities, there are drawbacks making it somehow difficult to interpret correctly the approximation power of such approximants: it might happen that the rational function has poles at places where the function \( f \) has no singularities, so-called spurious poles. This somehow vague notion needs some more explanation, for a precise (asymptotic) definition see the work [19] Definition 8 or [21] of Stahl: the Padé convergence theory like the Nuttall-Pommerenke Theorem for meromorphic functions \( f \) [2] Theorem 6.5.4 or the celebrated Stahl Theorem for algebraic functions \( f \) [20] Theorem 1.2, [2] Theorem 6.6.9 (or more general multivalued functions) tells us that there are domains \( D \) of analyticity of \( f \) such that the \([n/n]\) Padé approximants tend for \( n \to \infty \) to \( f \) in capacity on any compact subset \( K \) of \( D \). That is, given any threshold \( \epsilon > 0 \), the set of exceptional points in \( K \) where the error is larger than \( \epsilon \) becomes quickly "small", see, e.g., [2] Section 6.6 and the references therein. By the Gonchar Lemma [14] Lemma 1, convergence in capacity and absence of poles implies uniform convergence, but there are examples showing that there might be poles of an infinite subsequence of \([n/n]\) Padé approximants in \( K \), which of course makes it impossible to have uniform convergence in \( K \). Stahl shows in [18] Theorem 3.7

∗ Laboratoire Painlevé UMR 8524, UFR Mathématiques – M3, Université de Lille, F-59655 Villeneuve d’Ascq CEDEX, France. E-mail: {bbecke,matos}@math.univ-lille1.fr. Supported in part by the Labex CEMPI (ANR-11-LABX-0007-01).

†Version September 21, 2013
that one can establish uniform convergence for the special case of hyperelliptic functions by simply dropping all terms in a partial fraction decomposition with poles in \( D \). More generally, for algebraic functions, Stahl mentions in [20] Remark (8) for Theorem 1.2 an elimination procedure for spurious poles, but without giving details.

The notion [21] of asymptotically spurious poles of course is intractable on a computer since we are able to compute only finitely many approximants. In addition, the computed approximants will be affected by finite precision arithmetic, or by noise on the given Taylor coefficients. It was suggested by Froissart [10] and further analyzed for particular functions in [6, 12, 13] that instead we should identify poles of Padé approximants which come along with a ”close-by” zero, so-called Froissart doublets. The occurrence of such doublets is observed experimentally to increase in case of noise on the Taylor coefficients [6]. Stahl shows in [21] that in fact asymptotically spurious poles give raise to ”asymptotical Froissart doublets”.

Another popular method to detect ”doubtful” (generically simple) poles, adapted for instance in [15], is to identify poles \( z_k \) which have ”small” residuals \( a_k \) corresponding to terms \( \frac{a_k}{z-z_k} \) in the partial fraction decomposition of a Padé approximant \( p/q \).

Before going further, some notation. We denote by \( R_{m,n} \) the set of rational functions with numerator (and denominator) degree not exceeding \( m \) (and \( n \), respectively). In what follows, \( \| \cdot \| \) will always denote Euclidian norm together with the induced spectral matrix norm, and \( \kappa(A) = \| A \| \| A^\dagger \| \) the spectral condition number of a possibly rectangular matrix \( A \), with the pseudoinverse \( A^\dagger \). We notice that a change of norms might improve some of our estimates below, in particular we do not claim that any of the powers of \( m+n+1 \) occurring below are optimal. Hence we will sometimes use the writing \( a_1 \lesssim a_2 \) meaning that there exist modest constants \( b, r > 0 \) (not depending on \( f \) or \( m, n \)) such that \( a_1 \leq b(m+n+1)^r a_2 \), and \( a_1 \sim a_2 \) meaning that \( a_1 \lesssim a_2 \) and \( a_2 \lesssim a_1 \). As suggested in [15], before computing Padé approximants of \( f \) it is suitable to replace \( f \) by a suitably scaled counterpart \( af(bz) \) with nonzero scalars \( a, b \) chosen such that

\[
\sum_{j=0}^{m+n} |c_j|^2 = 1. \tag{1.3}
\]

Here the rescaling factor \( b \) should be chosen in order to obtain quantities \( |c_j| \leq 1 \) of comparable size, which asymptotically means that we rescale the complex plane in a way such that a meromorphic function \( f \) becomes analytic in \( |z| < 1 \). Finally, in order to simplify notation, in what follows we always fix \( m \) and \( n \) and drop these indices.

1.1. Robust Padé approximants, degeneracy and related matrices. Recently [15], Gonnet, Gütte and Trefethen suggested some new and quite interesting concept of a robust \([m’n’] \) Padé approximant \( p/q \) based on SVD computations. This object essentially is a \([m’n] \) Padé approximant (at least for exact arithmetic) for suitably chosen \( m \leq m’ \) and \( n \leq n’ \). Though the suggested numerical method to find \( m, n \) from \( m’, n’ \) is much more elaborate, one may get an idea of the method by thinking of \((m, n)\) as being the upper left corner of a ”numerical block” of the Padé table containing the coordinate \((m’, n’)\), or being on the upper or left border of such a ”block” and on the same diagonal \( m’ - n’ = m - n \). In the numerical experiences reported in [15], the shape of such a ”numerical block” is either a (finite or infinite) square or an infinite diagonal. Their robust \([m’n] \) Padé approximant \( p/q \) has the following properties

- (P1) it is nondegenerate in the sense that the polynomials \( p \) and \( q \) are co-prime, and that the defect \( \min \{m - \deg p, n - \deg q\} \) is equal to zero;
- (P2) the \( n \)th largest singular value \( \sigma_n(C) \) is larger than a certain threshold;
- (P3) the denominator is given by choosing as vec\((q)\) a right singular vector of norm 1 corresponding to the singular value \( \sigma_{n+1}(C) = 0 \).

We can read from (P2), (P3) that indeed \( C \) has maximal (unstructured) numerical rank \( n \), and thus vec\((q)\) spans the numerical kernel of \( C \), see also [1]. Moreover, according to (1.3) and (P2), the (unstructured) condition number \( \kappa(C) = \sigma_1(C)/\sigma_n(C) = \|C\|\|C^\dagger\| \) with the pseudoinverse \( C^\dagger = C^* (CC^*)^{-1} \) will be of moderate size.

The authors in [15] use analogies from well-known regularization techniques for linear algebra problems in order to justify theoretically their approach. Their paper contains many numerical examples which let believe that these new “regularized” approximants are indeed robust, that is, small perturbations in the input (like noisy Taylor coefficients) produce similar approximants (see also [1, 2] below for this
notion of robustness or forward stability). Also, in all numerical experiences reported in [15], these robust approximants do no longer have Froissart doublets nor small residuals. The aim of the present paper is to give some theoretical results complementing these numerically observed phenomena. For instance, we present a numerical example of robust approximants where spurious poles have not been eliminated. In addition, we describe a subclass of robust approximants where we can insure that we have eliminated spurious poles. All our statements only apply to nondegenerate $[m/n]$ Padé approximants $p/q$, and we will see that (P2) will enable us to show that the underlying nonlinear map is forward well-conditioned, whereas for the backward condition number, for Froissart doublets or for small residuals, other matrices do occur.

We first observe that (1.1), (1.2) is equivalent to solving

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

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

where $T$ is block upper triangular, with the lower right block given by $-C$. We will also require the two matrices

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

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

with $Q \in \mathbb{C}^{(m+n+1) \times (m+n+1)}$, and $S = S(q, -p) \in \mathbb{C}^{(m+n+1) \times (m+n+2)}$ having one more row and two more columns than the usual Sylvester matrix of two polynomials. Notice that these matrices are related through

$$S = QT.$$

1.2. Continuity and conditioning of the Padé map. For defining a (nonlinear) Padé map $F$

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

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 $p/q$ we have to be a bit careful due to degeneracies in the Padé table, also we have to fix the normalization (norm and phase) of the coefficients. Uniqueness is obtained by taking any $p, q$ of degree at most $m$, and $n$, respectively, satisfying (1.4), by canceling out a possible non-trival greatest common divisor such that $q(0) \neq 0$ (since $p(0) = c_0q(0)$), and then normalize in a suitable manner by a complex scalar, here

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

Notice that a non-trivial greatest common divisor only occurs for degenerate Padé approximants, and only here it might happen that $TF(c) \neq 0$. Also, $F$ is neither injective nor surjective. By adapting the techniques of [25], one may show the following result which is stated here without proof and which shows the importance of degeneracy.

**Theorem 1.1.** $F$ is continuous in a neighborhood of $c$ if and only if its $[m/n]$ Padé approximant $F(c)$ is nondegenerate.
For studying the (first order) forward and backward conditioning we will restrict ourselves to the real Padé map, namely the restriction of $F$ onto $\mathbb{R}^{m+n+1}$ (also denoted by $F$) and hence $F(c) \in \mathbb{R}^{m+n+2}$. Here one is interested in studying for $y = F(c)$ the perturbed equation

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

For the forward conditioning $\kappa_{for}(F)(c)$, one asks (for $\eta = 0$) how large the change of values $F(\tilde{c}) - F(c)$ can be for a worst case infinitesimal perturbation $\tilde{c} - c$ of the argument, whereas for backward conditioning $\kappa_{back}(F)(c)$ one knows an infinitesimal perturbation $\tilde{y} - y$ of the value, and one asks whether the perturbed value $\tilde{y}$ is the right answer for a nearby argument. In the latter case, due to the lack of surjectivity, the perturbed value $\tilde{y}$ first has to be projected on the image of $F$, and in general, due to the lack of injectivity, there might be several arguments $c$ and we have to find the best one. All errors occurring in the next statement are relative errors.

**Theorem 1.2.** Suppose that $F$ is continuous in a neighborhood of $c \in \mathbb{R}^{m+n+1}$, that $F(\mathbb{R}^{m+n+1})$ holds, and that the matrix $T$ is defined by $c$ and $Q$ by $F(c)$. Then we have the following forward amplification of real errors

$$\kappa_{for}(F)(c) := \limsup_{\tilde{c} \to c} \frac{\|F(\tilde{c}) - F(c)\|}{\|F(c)\|} \frac{\|\tilde{c} - c\|}{\|c\|} = \|TQ\|.$$

Moreover, the backward conditioning for real perturbations\(^2\) is given by\(^3\)

$$\kappa_{back}(F)(c) := \limsup_{\tilde{y} \to F(c)} \inf \left\{ \frac{\|\tilde{c} - c\|}{\|\tilde{y} - F(c)\|} : \|F(c) - \tilde{y}\| = \text{dist}(\tilde{y}, F(\mathbb{R}^{m+n+1})) \right\} = \|Q^{-1}T\|.$$

We know from \(^1\text{3}\) and \(^1\text{8}\) (see also Lemma \(^3\) below) that both matrices $T$ and $Q$ have a norm not larger than $\sqrt{m+n+2}$. Thus we learn from Theorem 1.2 that the real Padé map is forward (backward) well-conditioned at $c$ provided that the smallest singular value of $T$ (and of $Q$, respectively) is not too small. It is shown in Lemma \(^3\) below that the smallest singular values of $T$ and $C$ are of the same magnitude. Thus condition \(^P2\) insures that the real Padé map is forward well-conditioned.

In our proof of Theorem 1.2 we exploit a well-known formula for $\kappa_{for}(F)(c)$ in terms of the Jacobian of $F$. To our knowledge, similar formulas for $\kappa_{back}(F)(c)$ in terms of the pseudoinverse of the Jacobian have not been established before in the literature, we will formulate at the end of \(^3\) a conjecture on this matter.

### 1.3. Well-conditioned rational functions and spurious poles

Let us now turn to the subject of spurious poles, which in the present paper we study for general rational functions and not only for Padé approximants. It will be shown in Lemma \(^3\) below that $p/q$ is nondegenerate if and only if the corresponding matrix $S$ has full row rank. In a numerical setting, rank deficiency is typically excluded in requiring a condition number of modest size. In what follows, we will refer to rational functions as well-conditioned if the corresponding matrix $S$ has a modest condition number. As we show in the next theorem, for well-conditioned rational functions we are able to control the occurrence both of Froissart doublets and of small residuals. In the statement, we will make use of the uniform chordal metric in the set $\mathcal{M}_K$ of functions meromorphic in some compact $K \subset \mathbb{C}$ being defined by

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

Such a metric is of course useful to study questions of uniform convergence for rational or meromorphic functions since such functions are continuous in $K$ with respect to the chordal metric. A (different)

---

\(^2\)The occurrence of (a sub-matrix of) $Q$ in the backward conditioning of the Padé map has been noticed before by S. Güttes (personal communication).

\(^3\) Our definition for backward conditioning requires that, for $\tilde{y}$ being sufficiently close to $F(c)$, the infimum in the definition of the distance $\text{dist}(\tilde{y}, F(\mathbb{R}^{m+n+1}))$ is attained. This will be true for the Padé map under the assumptions of Theorem 1.2. In the general case it is sufficient to suppose that there is a sufficiently small compact neighborhood $N$ of $c$ such that $F(F(\mathbb{R}^{m+n+1}) \setminus F(N))$ does not have the accumulation point $F(c)$ since then we may replace $\text{dist}(\tilde{y}, F(\mathbb{R}^{m+n+1}))$ in the definition for backward conditioning by $\text{dist}(\tilde{y}, F(N))$ with the compact set $F(N)$. If such a neighborhood does not exist, then one naturally defines the backward conditioning to be $+\infty$. 

---
uniform metric has been also employed in [25] for measuring the distance of two rational functions for the continuity of the Padé map. We will discuss the link with the distance of two coefficient vectors in more detail in [4]. Notice that the next statement does not only cover Froissart doublets and small residuals of \( r = p/q \) but also of rational functions \( \tilde{r} = \tilde{p}/\tilde{q} \) close to \( r \), as those constructed in [15] where small leading coefficients in \( p \) or \( q \) are replaced by 0, compare with Remark 5.2 below.

**Theorem 1.3.** Let the two polynomials \( p \) of degree \( \leq m \) and \( q \) of degree \( \leq n \) be such that \( r = p/q \) is nondegenerate. Then the following statements hold for the matrix coefficients in \( p/q \).

(a) For any meromorphic function \( \tilde{r} \in \mathcal{M}_2 \) with \( \chi_D(r, \tilde{r}) \leq 1/3 \), the Euclidian distance of any pair of zeros and poles of \( \tilde{r} \) in the unit disk is bounded below by \( 1/(3\sqrt{2}(m+n+1)^{3/2}\kappa(S)) \).

(b) For any rational function \( \tilde{r} \in \mathcal{R}_{m,n} \) with \( 2(m+n+1)^2\kappa(S)^2\chi_D(r, \tilde{r}) \leq 1/3 \), the modulus of any residual of a simple pole in the unit disk of \( \tilde{r} \) is bounded below by \( 1/((2(m+n+1)^{3/2}\kappa(S)) \).

Numerical results presented in Example 2.3 below indicate that both lower bounds of Theorem 1.3 can be approximately attained. It seems for us that, due to the use of the basis of monomials, the occurrence of the unit disk \( \mathbb{D} \) in Theorem 1.3 is natural. For the case \( m = n \) of diagonal rational functions \( r, \tilde{r} \in \mathcal{R}_{n,n} \), we could also obtain results outside of the unit disk, by considering the reversed numerator and denominator polynomials (for which \( \kappa(S) \) remains unchanged).

Let us finally turn to convergence questions for robust Padé approximants. In [15, §8], Gonnet, Güttel, and Trefethen asked whether there are analogues of classical convergence theorems by Stahl and Pomerencne for robust Padé approximants where the absence of spurious poles would ensure to obtain not only convergence in capacity but uniform convergence. To be more precise, the authors suggest to compute robust Padé approximants of type \( [m_k,n_k] \) for increasing sequences of numbers \( m_k, n_k \), where each approximant is computed using a threshold \( tol_k \) possibly tending to zero for \( k \to \infty \) (compare with condition (P2)). Of course, a variable threshold does no longer allow a simple control of spurious poles through our Theorem 1.3. But quite often there are only a finite number of distinct robust Padé approximants following for instance a diagonal path \( (m_k = n_k = k) \) if one uses a fixed threshold for all approximants. For instance, the numerical experiments for the exponential function with \( tol_k = 10^{-14} \) as reported in [15, Fig. 5.1] tell us that there are only 8 distinct robust Padé approximants on the diagonal, since all approximants of type \( [n/n] \) for \( n \geq 8 \) reduce to the one for \( n = 7 \).

This vague observation can be made more explicit for Stieltjes functions \( f \), since here the matrix \( C \) has a condition number which grows quickly with \( n \), see [3] for results on the condition number of positive definite Hankel matrices. For general functions \( f \), we have the following result (for a related result on two succeeding Padé approximants on a diagonal see also Corollary 6.1 below).

**Theorem 1.4.** Let \( r = p/q \in \mathcal{R}_{m,n} \) be nondegenerate and \( \tilde{r} = \tilde{p}/\tilde{q} \in \mathcal{R}_{\tilde{m},\tilde{n}} \) with \( \tilde{m} < m \) and \( \tilde{n} < n \). Then \( 2\chi_D(r, \tilde{r})\kappa(S)^2 \geq (m+n+1)^{-2} \) for the matrix \( S = S(q, -p) \).

Roughly speaking, we learn from Theorem 1.4 that for any function \( f \) which can be well approximated by some element of \( \mathcal{R}_{m-1,n-1} \) with respect to the uniform chordal metric in the unit disk, its \( [n/n] \) Padé approximant \( r \) either does not have a small approximation error \( \chi_D(f, r) \), or otherwise the number \( \kappa(S) \) is necessarily "large". We are not aware of a systematic study of the error of best rational approximants with respect to the uniform chordal metric, beside the negative result [3] Theorem 3.1.

The remainder of the paper is organized as follows. §2 contains some numerical experiments which confirm our theoretical findings. In §3 we give some auxiliary statements and provide a proof of Theorem 1.2 on the conditioning of the real Padé map. We also formulate some general conjecture how to compute the backward conditioning of a nonlinear locally differentiable map. §4 is devoted to the study of distances of rational functions, we will show in Theorem 1.7 that in some cases the uniform chordal metric is close to forming differences of scaled coefficient vectors. A proof of Theorem 1.3 and Theorem 1.4 is provided in §5. In Section 6 we report about some previous work on related fields like numerical GCDs, condition number estimators, look-ahead procedures for computing Padé approximants, and finally the sensitivity of eigenvalues of a Hankel pencil. A summary of our work and concluding remarks can be found in §7.

2. Some numerical experiments. In this section we present examples of subdiagonal Padé approximants \( (m = n - 1) \) for three functions, namely

\[
 f_1(z) = \int_1^1 \frac{dx}{\sqrt{1-x^2} - 1-xz}, \quad f_2(z) = \exp(z), \quad f_3(z) = \sum_{j=0}^{2N} c_3(j)z^j, \quad c_3 = \text{randn}(2N), \tag{2.1}
 \]
the first one a Stieltjes function analytic in \(|z| < 1\), and the second (and third) one an entire function with quickly decaying Taylor coefficients (and random coefficients, respectively). For each \(m + n = n = 1, \ldots, N\), we first normalize the vector of the first \(m + n + 1\) Taylor coefficients following (1.3) by dividing by the norm. Subsequently, we compute the denominator coefficients through svd computations, the corresponding coefficients of the numerator by multiplying with a submatrix of \(T\), and then normalize following (1.8) by dividing by the norm. It turns out that all approximants are nondegenerate (though there are \(2 \times 2\) blocks in the Padé table of the even function \(f_1\)).

We draw in Fig 2.1, Fig 2.2 and Fig 2.3 the (spectral) condition number of the four matrices \(C\), \(T\), \(S\) and \(Q\), as well as the norm of the two matrices \(T^\dagger Q\) and \(Q^{-1} T\) occurring in Theorem 1.2. One observes that always \(\kappa(C)\) and \(\kappa(T)\) are of the same magnitude, and that \(\max\{\kappa(Q), \kappa(T)\} \ll \kappa(S)\) (these properties are shown analytically in Lemma 3.3 below). More surprisingly, we also observe (without proof) that \(\|Q^{-1} T\|\) and \(\kappa(Q)\) are always of the same magnitude, and the same is true for \(\|T^\dagger Q\|\) and \(\kappa(T)\) (up to some artifacts for the exponential function and \(n \geq 11\) in Fig 2.2 which we believe are due to rounding errors).

In order to discuss the sharpness of Theorem 1.3 we also draw the reciprocal values of

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

\[
\text{Residual} = \min\{|p(\tau) q'(|\tau|) : q(\tau) = 0, |\tau| \leq 1\},
\]

in case where the \([n - 1]\) Padé approximant has at least one pole in the unit disk. Below we give some specific comments for each of the three functions.

\[\text{Fig. 2.1. Condition numbers related to the Stieltjes function } f_1.\]

**Example 2.1.** The \([n - 1]\) Padé approximant for \(n = 1, \ldots, N = 15\) of the Stieltjes function \(f_1\) in (2.1) does not have poles in the unit disk, even in presence of rounding errors. We observe from Fig. 2.1 that \(\kappa(S)\) and \(\kappa(T)\) have the same magnitude, and are growing exponentially in \(n\). Also, \(\kappa(Q)\) is growing exponentially in \(n\), but less quickly. This example clearly shows that \(\kappa(S)\) large does not imply the existence of a Froissart doublet or a small residual in the disk.

**Example 2.2.** The \([n - 1]\) Padé approximant for \(n = 1, \ldots, 10\) of the exponential function \(f_2\) in (2.1) does not have poles in the (open) unit disk, even in presence of rounding errors. We observe from

\[\text{Fig. 2.2.} \text{ Condition numbers related to the exponential function } f_2.\]

\[\text{Fig. 2.3.} \text{ Condition numbers related to the random function } f_3.\]

In exact arithmetic, it is known that all poles tend to infinity for \(n \to \infty\). Thus we believe that, due to rounding errors,
Fig. 2.2. Condition numbers related to the exponential function $f_2$.

Fig. 2.2 for $n \leq 10$ that $\kappa(Q)$ is close to 1, and thus $\kappa(S)$ and $\kappa(T)$ have the same magnitude, which is growing quickly with $n$.

Example 2.3. The numerical results reported in Fig. 2.3 for the random function $f_3$ in (2.1) for our $[n-1][n]$ Padé approximants for $n \geq 11$ having poles in the disk are badly computed. Also, Matlab gives warnings that the condition numbers and norms for $n \geq 11$ are badly computed.
n = 1, 2, ..., N = 30 depend of course on the realization of the random Taylor coefficients, but for about 10 realizations we found each time a similar behavior: all approximants are robust since \( \kappa(T) \) is always not too far from 1. As a consequence, \( \kappa(S) \) and \( \kappa(Q) \) have the same magnitude, the dependence on \( n \) being quite erratic, in this example between 1 and 10\(^{20} \). This shows that there are cases where a Padé approximant is robust but not well-conditioned.

Even more striking, in this example the curves for 1/Froissart and 1/Residual follow quite closely that of \( \kappa(S) \), showing that, for this example, Theorem 1.3 is essentially sharp.

In the context of Example [2,3] we should also mention the very recent paper [10] where, given arbitrary nonzero complex numbers \( z_k \) of modulus \( \leq 1/3 \), the author explicitly gives a function \( f \) analytic in \( |z| < 1 \) where the subsequence of \([n_k|n_k]\) diagonal Padé approximants, \( n_k = 2^k - 2 \), are robust (with the condition number of \( C \) being bounded by 5) but have a (spurious) pole at \( z_k \). His function \( f \) is resulting from a smart modification of Gammel’s counterexample [2, §6.7], where the \([n_k|n_k]\) approximant coincides with the \([n_k|1]\) Padé approximant, leading to a rich block structure in the Padé table. It can be shown that in this case both \( Q \) and \( S \) have a condition number being of the same magnitude as \( |z_k|^{-2n_k} \), hence these approximants are not well-conditioned.

3. Conditioning of the Padé map and proof of Theorem 1.2

The aim of this section is to analyze the conditioning of the real Padé map and in particular to provide a proof of Theorem 1.2. In this section we consider a kind of left inverse real Padé map \( G \) being defined by

\[
G : \mathbb{R}^{m+n+2} \ni y = \begin{bmatrix} \text{vec}(p) \\ \text{vec}(q) \end{bmatrix} \mapsto c = \begin{bmatrix} c_0 \\ \vdots \\ c_{m+n} \end{bmatrix} \in \mathbb{R}^{m+n+1}, \quad \frac{p(z)}{q(z)} = \sum_{j=0}^{m+n} c_j z^j + O(z^{m+n+1}) z \to 0,
\]

which is defined at least for pairs of polynomials with \( q(0) \neq 0 \) (as it is true for a neighborhood of any value \( F(c) \)).

We start with three technical statements, the first one relating nondegeneracy (or by Theorem 1.1 the continuity of \( F \) at \( c \)) to the matrix \( S \), the second dealing with the differentiability of the maps \( F \) and \( G \) and their Jacobian, and finally the third relating the smallest and largest singular values of \( C \) to those of \( T \).

**Lemma 3.1.** Let \( p, q \) be two polynomials, \( p \) of degree \( \leq m \) and \( q \) of degree \( \leq n \). Then \( p/q \) is nondegenerate if and only if the matrix \( S \) defined in (1.5) has full (row) rank \( m+n+1 \).

*Proof.* Suppose that \( p/q \) is degenerate. Then either \( p_n = q_m = 0 \) (implying that the last row of \( S \) is zero), or there exists \( \gamma \in \mathbb{C} \) with \( p(\gamma) = q(\gamma) = 0 \), implying that \((1, \gamma, ..., \gamma^{m+n}) S = 0 \). Thus, in both cases \( S \) does not have full row rank.

Conversely, suppose that \( p/q \) is nondegenerate, then at least one of the leading coefficients \( p_m \) or \( q_n \) is not vanishing, without loss of generality \( p_m \neq 0 \). Notice that, up to permutation of columns, \( S \) equals

\[
\begin{bmatrix}
S & * & * \\
0 & q_n & -p_m \\
\end{bmatrix}
\]

with the classical square Sylvester matrix \( S \in \mathbb{C}^{(m+n) \times (m+n)} \), obtained from \( S \) by dropping the last row, and last column in each column block. With \( x_1, ..., x_m \) the roots of \( p \), observe that by assumption \( q(x_j) \neq 0 \). We use the formula [11, Theorem 9.3(ii)]

\[
\det S = \pm (p_m)^n \prod_{j=1}^{m} q(x_j) \neq 0
\]

in order to conclude that \( S \) and thus \( S \) has full row rank. \( \Box \)

Recall from (1.6) that \( S \) having rank \( m+n+1 \) implies that the matrix \( Q \) defined in (1.5) is invertible, and the matrix \( T \) defined in (1.4) also has full rank \( m+n+1 \).

**Lemma 3.2.** Under the assumptions of Theorem 1.3, \( G(F(c)) = \tilde{c} \) for all \( \tilde{c} \) in some neighborhood of \( c \). Moreover, \( G \) is differentiable in some neighborhood of \( F(c) \), and \( F \) is differentiable in some neighborhood of \( c \), with Jacobians

\[
J_G(F(c)) = Q^{-1}T, \quad J_F(c) = T^\dagger Q.
\]  

(3.1)
Proof. We write more explicitly
\[ \tilde{y} = \begin{bmatrix} \text{vec}(\tilde{p}) \\ \text{vec}(\tilde{q}) \end{bmatrix}, \quad \tilde{c} = G(\tilde{y}) = Q(\tilde{q})^{-1} \begin{bmatrix} \text{vec}(\tilde{p}) \\ 0 \end{bmatrix}, \]
where we have written \( Q = Q(q) \) for the triangular Toeplitz matrix in (1.5). Notice that the map \( \text{vec}(\tilde{q}) \mapsto Q(\tilde{q}) \) clearly is differentiable. Since by assumption \( Q(q) \) is invertible, we conclude that \( \text{vec}(\tilde{q}) \mapsto Q(\tilde{q})^{-1} \) is differentiable in a neighborhood of \( \text{vec}(q) \), and thus \( \tilde{y} \mapsto G(\tilde{y}) \) is differentiable in a neighborhood of \( y \).

In order to compute its Jacobian \( J_G(y) \), we write \( T_0(c) \) for the submatrix of \( T = T(c) \) in (1.4) formed by the last \( n + 1 \) columns, and we notice that \( c = G(y) \) does satisfy
\[ \begin{bmatrix} \text{vec}(\tilde{p}) \\ 0 \end{bmatrix} = Q(\tilde{q})G(\tilde{y}) = -T_0(\tilde{c})\text{vec}(\tilde{q}). \]

Taking the product rule for partial derivatives, we obtain
\[ \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} = Q(\tilde{q})J_G(\tilde{y}) - T_0(\tilde{c}) \begin{bmatrix} 0 & I \end{bmatrix} \]

implying that \( Q(\tilde{q})J_G(\tilde{y}) = T(\tilde{c}) \), as claimed in (3.1).

For studying \( F \), consider now the function
\[ H : \mathbb{R}^{m+n+2} \times \mathbb{R}^{m+n+1} \ni (\tilde{y}, \tilde{c}) \mapsto H(\tilde{y}, \tilde{c}) = \begin{bmatrix} G(\tilde{y}) - \tilde{c} \\ \frac{\tilde{y}^T}{\tilde{y}} \tilde{y} - 1 \end{bmatrix}, \]
defined and being \( C^1 \) in a neighborhood of \( (y, c) = (F(c), c) \), and \( H(F(c), c) = 0 \). Since
\[ \frac{\partial H}{\partial \tilde{y}}(\tilde{y}, \tilde{c}) = \begin{bmatrix} J_G(\tilde{y}) \\ 2\tilde{y} \end{bmatrix} = \begin{bmatrix} Q(\tilde{q})^{-1}T(\tilde{c}) \\ 2\tilde{y} \end{bmatrix}, \]
and
\[ \begin{bmatrix} Q(q)^{-1}T(c) \\ 2y^T \end{bmatrix} \left[ \begin{bmatrix} Q(q)^{-1}T(c) \\ 2y^T \end{bmatrix} \right]^* = \begin{bmatrix} Q(q)^{-1}T(c)T(c)^*Q(q)^{-*} & 0 \\ 0 & 4y^T y \end{bmatrix} \]
is invertible by assumption on \( S \) and thus on \( T \) and \( Q \), the theorem of implicit functions tells us that there is a function \( \tilde{F} : \tilde{c} \mapsto \tilde{y} \) defined in a neighborhood \( N \) of \( c \) such that \( \tilde{F}(c) = F(c) \), and \( H(\tilde{F}(c), \tilde{c}) = 0 \) for all \( \tilde{c} \in N \). Thus \( G(\tilde{F}(\tilde{c})) = \tilde{c} = G(F(\tilde{c})) \), implying that the vectors \( \tilde{F}(\tilde{c}) \) and \( F(\tilde{c}) \) represent the same rational function. By continuity of \( F \) at \( c \), Theorem 1.1 and Lemma 3.1, we may conclude that the Sylvester matrix associated to \( F(\tilde{c}) \) for \( \tilde{c} \in N \) sufficiently close to \( c \) is of full row rank, and thus \( \tilde{F}(\tilde{c}) \) equals \( F(\tilde{c}) \) up to a multiplication with a constant (possibly depending on \( \tilde{c} \)). The additional relation \( \|F(\tilde{c})\|^2 = 1 \) allows to conclude that both vectors have the same norm, and thus might only differ by sign. In order to check the Padé sign convention (1.8), let \( e_{m+2} \) be the \((m+2)\)th canonical vector in \( \mathbb{R}^{m+n+2} \). Then \( e_{m+2}^*F(c) = e_{m+2}^*F(c) = q(0) > 0 \), and hence by continuity of \( \tilde{F} \) also \( e_{m+2}^*\tilde{F}(\tilde{c}) > 0 \) for \( \tilde{c} \in N \) sufficiently close to \( c \), implying that \( \tilde{F}(\tilde{c}) = F(\tilde{c}) \).

Thus \( F \) is \( C^1 \), with Jacobian
\[ J_F(c) = -\frac{\partial H}{\partial \tilde{y}}(y, c)^{-1}\frac{\partial H}{\partial \tilde{c}}(y, c) = \begin{bmatrix} Q^{-1}T \\ 2y^T \end{bmatrix} \left[ \begin{bmatrix} Q^{-1}TT^*Q^{-*} & 0 \\ 0 & 4y^T y \end{bmatrix} \right]^{-1} \begin{bmatrix} I \\ 0 \end{bmatrix} = T^*(TT^*)^{-1}Q = T^1Q. \]

\[ \square \]

Lemma 3.3. Suppose that \( S \) has rank \( m + n + 1 \). Then for the matrices \( C \) of (1.2) and \( T \) of (1.4) we have that
\[ \max\{1, \|C\|\} \leq \|T\| \leq \sqrt{m + n + 2}, \] (3.2)
\[ \|C^\dagger\| \leq \|T^\dagger\| \leq \sqrt{2(m + n + 2)}\|C^\dagger\|. \] (3.3)
Furthermore, for the matrices \( S, Q \) of (1.5) with the normalization (1.8) there holds

\[
\|Q\| \leq \sqrt{m+n+1}, \quad \frac{1}{\sqrt{2}} \leq \|S\| \leq \sqrt{m+n+1}, \quad \|Q^{-1}\| \leq \|T\| \|S^\dagger\|, \quad \|T\| \leq \|Q\| \|S\|.
\]

**Proof.** Since 1 is an entry and \(-C\) a submatrix of \( T \), we obtain the first inequality of (3.2), and the second follows from the scaling (1.3) and the general fact that any matrix \( \in \mathbb{C}^{(m+n+1) \times (m+n+2)} \) with columns of norm \( \leq 1 \) has a Frobenius norm \( \leq \sqrt{m+n+2} \).

For a proof of (3.3), we first recall that by assumption both \( C \) and \( T \) have full row rank, and hence

\[
\frac{1}{\|C\|} = \min_{y \in \mathbb{C}^{m+n+1}} \frac{\|y^T T\|}{\|y\|} \leq \min_{x \in \mathbb{C}^n} \frac{\|(0,x^\star)T\|}{\|x\|} = \frac{\|x^\ast C\|}{\|x\|} = \frac{1}{\|C\|},
\]

implying the first inequality. For the second, recall that \( C^\dagger = C^*(CC^*)^{-1} \) and hence the two matrices

\[
T = \begin{bmatrix} I & -L \\ 0 & -C \end{bmatrix}, \quad T^R = \begin{bmatrix} I & -LC^\dagger \\ 0 & -C^\dagger \end{bmatrix} = \begin{bmatrix} C & -L \\ 0 & -I \end{bmatrix} \begin{bmatrix} C^\dagger & 0 \\ 0 & C^\dagger \end{bmatrix}
\]

satisfy \( TT^R = I \). Since the orthogonal projector \( T^T T \) is of norm 1, we conclude that \( \|T^\dagger\| = \|T^\dagger TT^R\| \leq \|T^R\| \). It remains to observe that the right-hand factor in the above factorization of \( T^R \) has norm \( \|C^\dagger\| \), and the left-hand factor has rows of norm \( \leq 2 \) (in fact \( \leq 1 \) provided that \( n \leq m \)) due to (1.3).

We finally turn to a proof of (3.4), the upper bound for \( \|Q\| \) following as before from the scaling (1.8). Using (1.8) we also observe that the sum of the squares of the norms of all columns of the matrix \( S \) equals \( \|S\|^2_F = (m+1)\|\text{vec}(q)\|^2 + (n+1)\|\text{vec}(p)\|^2 \leq m+n+1 \) and the sum of the squares of the norms of the first and \( (n+2) \)nd column of \( S \) equals \( \|\text{vec}(q)\|^2 + \|\text{vec}(p)\|^2 = 1 \leq 2 \|S\|^2 \), implying the claimed inequalities for \( \|S\| \). For the upper bound for \( \|Q^{-1}\| \) (which we suspect to be not very sharp), we use (1.6) in order to conclude that \( I = SS^\dagger = QT\)_{S^\dagger} and thus \( Q^{-1} = TS^\dagger \). Finally, since (1.6) is a full rank decomposition, we also have that \( S^\dagger = T^\dagger Q^{-1} \) and thus \( T^\dagger = S^\dagger Q \), implying the claimed bound for \( \|T^\dagger\| \).

We are now prepared to give the proof of the main statement.

**Proof.** of Theorem 1.2 According to Lemma 3.2 our nonlinear map \( F \) is differentiable in \( c \). Thus from [24] we have the following well-known relation for the forward conditioning

\[
\limsup_{\hat{\gamma} \rightarrow c} \frac{\|F(\hat{\gamma}) - F(c)\|}{\|F(c)\|} = \frac{\|c\|}{\|J_F(c)\|} = \|T^\dagger Q\|,
\]

where for the last equality we used the facts that \( \|c\| = 1 \) according to (1.3), \( \|F(c)\| = 1 \) by definition (1.8), and that we have the explicit formula (3.1) for the Jacobian. This shows (1.9).

For a proof of (1.10), let \( \hat{y} \) be sufficiently close to \( y = F(c) \). We first claim that

\[
\text{dist}(\hat{y}, F(\mathbb{R}^{m+n+1})) = \|\hat{y}\| - 1, \quad \text{attained for} \quad F(\hat{c}) = \hat{y}/\|\hat{y}\|.
\]

Indeed, according to (1.8), \( F(\mathbb{R}^{m+n+1}) \) is a subset of the sphere \( S = \{ z \in \mathbb{R}^{m+n+2} : \|z\| = 1 \} \), and thus

\[
\text{dist}(\hat{y}, F(\mathbb{R}^{m+n+1})) \geq \text{dist}(\hat{y}, S) = \|\hat{y}\| - 1,
\]

the second equality being attained for \( \hat{y}/\|\hat{y}\| \in S \). Write more explicitly

\[
y = F(c) = \begin{bmatrix} \text{vec}(p) \\ \text{vec}(q) \end{bmatrix}, \quad \hat{y}/\|\hat{y}\| = \begin{bmatrix} \text{vec}(\hat{p}) \\ \text{vec}(\hat{q}) \end{bmatrix},
\]

with \( q(0) > 0 \) by (1.8). Since \( \hat{q}(0) \in \mathbb{R} \) is close to \( q(0) \), we conclude that \( \hat{q}(0) > 0 \) and hence \( \hat{y}/\|\hat{y}\| = F(\hat{c}) \) for a unique \( \hat{c} = G(\hat{y}) \), implying (3.5). Since for \( y = F(c) \) we also have that \( c = G(y) \), we find that

\[
\limsup_{\hat{y} \rightarrow F(c)} \frac{\|\hat{c} - c\|}{\|c\|} = \limsup_{\hat{y} \rightarrow y} \frac{\|G(\hat{y}) - G(y)\|}{\|G(\hat{y})\|} = \limsup_{\hat{y} \rightarrow y} \frac{\|y\|}{\|G(y)\|} \|J_G(y)\|,
\]

where

\[
\limsup_{\hat{y} \rightarrow y} \frac{\|G(\hat{y}) - G(y)\|}{\|G(\hat{y})\|} = \limsup_{\hat{y} \rightarrow y} \frac{\|y\|}{\|G(y)\|} \|J_G(y)\|,
\]

and

\[
\limsup_{\hat{y} \rightarrow F(c)} \frac{\|\hat{c} - c\|}{\|c\|}.
\]
i.e., the backward conditioning of $F$ at $c$ is the same as the forward conditioning of $G$ at $y = F(c)$. Thus again (1.10) follows from (1.3), (1.8), and (3.1).

To our knowledge, the backward conditioning of a nonlinear locally differentiable map did not receive as much attention in the literature as it should. In accordance with part (1.10) of Theorem 1.2, it is possible to show the following result which is stated here without proof.

**Theorem 3.4.** Let $F : \mathbb{R}^M \rightarrow \mathbb{R}^N$ be differentiable and different from a constant in a closed neighborhood $N$ of $c$, and $c \neq 0$, $F(c) \neq 0$, and suppose that $F(\mathbb{R}^M) \setminus F(N)$ does not accumulate at $F(c)$. Then we have the following backward conditioning at $c$

$$
\kappa_{\text{back}}(F)(c) = \lim_{\bar{y} \rightarrow F(c)} \inf \left\{ \frac{\|\bar{y} - \bar{c}\|}{\|\bar{y} - F(\bar{c})\|} : \|F(\bar{c}) - \bar{y}\| = \text{dist}(\bar{y}, F(\mathbb{R}^M)) \right\} = \frac{\|F(c)\|}{\|c\|} \|J_F(c)\|^{1/2}
$$

provided that $\bar{c} \mapsto \text{rank } J_F(\bar{c})$ is continuous in $\bar{c}$.

It is easy to check by employing least squares techniques that this result holds for affine but not necessarily injective or surjective functions $F$, here of course $J_F(c)$ does not depend on $c$. Thus our statement essentially says that, provided that $\bar{c} \mapsto \text{rank } J_F(\bar{c})$ is continuous in $c$, we are allowed to linearize $F$ around $c$ for studying the backward condition number $\kappa_{\text{back}}(F)(c)$.

The example

$$
\bar{c} = \begin{bmatrix} \bar{c_1} \\ \bar{c_2} \end{bmatrix} \mapsto F(\bar{c}) = \begin{bmatrix} \bar{c_1} \\ \bar{c_2} \end{bmatrix}, \quad c = \begin{bmatrix} 1 \\ 0 \end{bmatrix}
$$

shows (after some elementary computations) that our assumption on the rank cannot be dropped in Theorem 3.4 since here $\kappa_{\text{back}}(F)(c) = \infty$. We have the following.

**Conjecture 3.5.** Under the assumptions of Theorem 3.4, if $\bar{c} \mapsto \text{rank } J_F(\bar{c})$ is not continuous in $c$, then $\kappa_{\text{back}}(F)(c) = \infty$.

4. Distances between two rational functions and their coefficient vectors. A central question in this paper is how to measure the distance between two rational functions

$$
r = p/q \in \mathcal{R}_{m,n}, \quad \bar{r} = \bar{p}/\bar{q} \in \mathcal{R}_{m,n},
$$

with coefficient vectors

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

A natural metric in the set $\mathcal{M}_K$ of functions meromorphic in some compact $K \subset \mathbb{C}$ would be the uniform chordal metric $\chi_K(r, \bar{r})$ introduced in (1.11). This metric is well adapted to study (uniform) convergence questions (since meromorphic functions are continuous on the Riemann sphere), and we will also see that it enables us to study Froissart doublets and small residuals. However, it is not so clear how to relate such a metric to the coefficient vectors (in the basis of monomials) of numerators and denominators of rational functions, which are used to parametrize rational functions in the Padé map. This is essentially due to the fact that there are several coefficient vectors $x(r)$ representing the same rational function $r$: even if we suppose that $r$ is nondegenerate, we still may multiply $x(r)$ by an arbitrary complex scalar. As before, we will always suppose that coefficient vectors are of norm 1, but this fixes only the absolute value but not the phase of the scalar normalization constant. For defining a metric between rational functions it will therefore be suitable to measure the distance of coefficient vectors with optimal phase

$$
d(r, \bar{r}) := \min\{\|x(r) - ax(\bar{r})\| : a \in \mathbb{C}, |a| = 1\}. \quad (4.1)
$$

The reader easily checks that $\|x(r) - ax(\bar{r})\|$ does not depend on $a$ if $x(r)$ and $x(\bar{r})$ are mutually orthogonal, and else

$$
\arg\min\{\|x(r) - ax(\bar{r})\| : a \in \mathbb{C}, |a| = 1\} = \frac{x(\bar{r})^*x(r)}{|x(\bar{r})^*x(r)|}. \quad (4.2)
$$

In particular, if both $x(r)$ and $x(\bar{r})$ are real then

$$
d(r, \bar{r}) = \min\{\|x(r) - x(\bar{r})\|, \|x(r) + x(\bar{r})\|\},
$$

and else
and more precisely \(d(r, \bar{r}) = ||x(r) - x(\bar{r})||\) provided that \(x(\bar{r})^* x(r) \geq 0 \) or \(||x(r) - x(\bar{r})|| \leq \sqrt{2}\), as it was the case in our study of the continuity and the conditioning of the real Padé map.

Recall from the introduction that we called a rational function \(r = p/q\) well-conditioned if the condition number \(\kappa(S)\) (which does not depend on the normalization of the coefficient vector occurring in \([1.5]\)) is not too large. The following result shows that the two distances \(d(r, \cdot)\) and \(\chi_D(r, \cdot)\) for the closed unit disk \(D\) introduced above are of comparable size provided that \(r\) is well-conditioned.

**Theorem 4.1.** Let \(r = p/q\) be nondegenerate, then for all \(\bar{r} \in \mathcal{R}_{m,n}\)

\[
\frac{(m+n+1)^{-3/2}}{\sqrt{2\kappa(S)}} d(r, \bar{r}) \leq \chi_D(r, \bar{r}) \leq \sqrt{2(m+n+1)} \kappa(S) d(r, \bar{r}). 
\]

**Proof.** According to \([4.1],[4.2]\), and our convention on the norm we may choose the phase of \(x(\bar{r})\) such that

\[x(r) = \begin{bmatrix} \text{vec}(p) \\ \text{vec}(q) \end{bmatrix}, \quad x(\bar{r}) = \begin{bmatrix} \text{vec}(-\bar{p}) \\ \text{vec}(-\bar{q}) \end{bmatrix}\]

are of norm 1, \(||x(r) - x(\bar{r})|| = d(r, \bar{r})\) and hence \(x(r)^* x(\bar{r}) \geq 0\). Hence we may repeat the arguments in the proof of \([3.4]\) and get the inequalities

\[1/\sqrt{2} \leq \|S\| \leq \sqrt{m+n+1}.\]

In order to establish the right-hand inequality of \([4.3]\), it is sufficient to show the relation

\[z \in \mathbb{D}: \quad \chi(r(z), \bar{r}(z)) \leq \sqrt{2(m+n+1)} \kappa(S) ||x(r) - x(\bar{r})||.\]

By definition of the chordal metric and the Cauchy-Schwarz inequality,

\[
\chi(r(z), \bar{r}(z)) = \left\| \frac{(p(z) - \bar{p}(z)) \bar{q}(z) - \bar{p}(z)(q(z) - \bar{q}(z))}{\sqrt{p(z)^2 + |q(z)|^2}} \right\| \leq \left\| \frac{p(z) - \bar{p}(z)}{q(z) - \bar{q}(z)} \right\| \frac{\sqrt{p(z)^2 + |q(z)|^2}}{\sqrt{p(z)^2 + |q(z)|^2}}
\]

\[= \left\| \begin{bmatrix} 1, \ldots, z^m \\ 0 \end{bmatrix}, \ldots, \begin{bmatrix} 1, \ldots, z^n \\ 0 \end{bmatrix} \right\| \left\| \begin{bmatrix} 1, \ldots, z^m \\ 0 \end{bmatrix} - \begin{bmatrix} 1, \ldots, z^n \\ 0 \end{bmatrix} \right\|\]

\[
\left\| \left(1, z, \cdots, z^{n+m} \right) \right\| \leq \sqrt{p(z)^2 + |q(z)|^2}.\]

Let us study separately the term in the denominator. We remark that

\[
\left(1, z, \cdots, z^{n+m} \right) S = \left(-q(z), -zq(z), \cdots, -z^m q(z), p(z), \cdots, z^n p(z) \right).
\]

By Lemma 3.1 we know that the Sylvester-like matrix \(S\) has full row rank and hence \(SS^\dagger = I\). Multiplying the above relation on the right by \(S^\dagger\) and taking norms we arrive at

\[
\left\| \left(1, z, \cdots, z^{n+m} \right) \right\|^2 \leq ||S||^2 \left\| \left(-q(z), -zq(z), \cdots, -z^m q(z), p(z), \cdots, z^n p(z) \right) \right\|^2
\]

\[
\leq ||S||^2 \left\| \begin{bmatrix} 1, z, \ldots, z^m \\ 0 \end{bmatrix} \right\| \left\| \begin{bmatrix} 1, \ldots, z^m \\ 0 \end{bmatrix} \right\| \left\| \begin{bmatrix} p(z)^2 + |q(z)|^2 \right\|^2,\]

which implies that

\[\forall z \in \mathbb{C}: \quad 1 \leq ||S|| \sqrt{p(z)^2 + |q(z)|^2}.\]

Inserting \([4.9]\) into \([4.7]\) and using \([4.5]\) and the fact that \(z \in \mathbb{D}\) implies \([4.6]\).

It remains the left-hand inequality of \([4.3]\), for which it is sufficient to show

\[
d(r, \bar{r}) \leq \sqrt{2} (m+n+1)^{3/2} \kappa(S) \chi_K(r, \bar{r}),\]

with \(K\) the set of \((m+n+1)\)th roots of unity. Denote by \(\Omega = (\frac{1}{\sqrt{m+n+1}} \exp(\frac{2\pi i j k}{m+n+1}))_{j,k=0, \ldots, m+n}\) the unitary DFT matrix of order \(m+n+1\). By Lemma 3.1 the unit norm vector \(x(r)\) generates the (right)
kernel of $S$, and thus $S^\dagger S = I - x(r)x(r)^\ast$. Since $x(r)^*x(\bar{r}) \geq 0$, we find an angle $\alpha \in (0, \pi/2]$ such that $\cos(\alpha) = x(r)^*x(\bar{r})/(\|x(r)\|\|x(\bar{r})\|) = x(r)^*x(\bar{r})$. Thus
\[ d(r, \bar{r}) = \|x(r) - x(\bar{r})\| = \sqrt{2 - 2\cos(\alpha)} = 2\sin(\alpha/2), \]
whereas
\[ \|S^\dagger S(x(r) - x(\bar{r}))\| = \|x(\bar{r}) - x(r)\cos(\alpha)\| = \sqrt{1 - \cos^2(\alpha)} = \sin(\alpha) = 2\sin(\alpha/2)\cos(\alpha/2). \]

Thus $\|S^\dagger S(x(r) - x(\bar{r}))\| = \cos(\alpha/2) d(r, \bar{r}) \geq d(r, \bar{r})/\sqrt{2}$, implying that
\[ d(r, \bar{r})/\sqrt{2} \leq \|S^\dagger S(x(r) - x(\bar{r}))\| \leq \|S^\dagger\| \|S(x(r) - x(\bar{r}))\| \]
\[ = \|S^\dagger\| \|\Omega S(x(r) - x(\bar{r}))\| \leq \|S^\dagger\| \max_{z \in \mathcal{R}} |p(z)\bar{q}(z) - q(z)\bar{p}(z)|. \]
\hfill (4.11)

Returning to (4.8), we also find that
\[ \forall |z| \leq 1 : \quad (m + n + 1)\|S\|^2 \geq \|1, z, ..., z^{m+n}\|^2 \]
\[ = |p(z)|^2 \sum_{j=0}^{n} |z|^2j + |q(z)|^2 \sum_{j=0}^{m} |z|^2j \geq |p(z)|^2 + |q(z)|^2. \]

A similar bound is obtained for $\bar{p}(z), \bar{q}(z)$, which combined with (4.5) becomes
\[ \forall |z| \leq 1 : \quad (m + n + 1) \geq \sqrt{|p(z)|^2 + |q(z)|^2}. \]

Inserting these two relations into the right-hand side of (4.11) implies (4.10). 

5. Proofs of Theorem 1.3 and of Theorem 1.4. We start by establishing a technical result on the condition number of Sylvester-like matrices close to $S$.

**Lemma 5.1.** Let $r = p/q$ be nondegenerate. If $\bar{r} = \bar{p}/\bar{q} \in \mathcal{R}_{m,n}$ satisfies
\[ \sqrt{2(m + n + 1)} d(r, \bar{r}) \kappa(S) \leq 1/3, \]
then it is nondegenerate, and $\kappa(\tilde{S}) \leq 2\kappa(S)$ for the Sylvester-like matrix $\tilde{S} = S(-\bar{q}, \bar{p})$ constructed as in (1.5).

More generally, if $\bar{r}$ is degenerate then $\sqrt{2(m + n + 1)} d(r, \bar{r}) \kappa(S) \geq 1$.

**Proof.** For a proof of the first statement, write $E := S^\dagger(S - \tilde{S})$, and denote by $x(r), x(\bar{r})$ the corresponding coefficient vectors with unit norm and particular phase such that $\|x(r) - x(\bar{r})\| = d(r, \bar{r})$. Then $S(I - E) = S - SS^\dagger(S - \tilde{S}) = \tilde{S}$. Using the same arguments as in the proof of (4.5), we obtain
\[ \|E\| \leq \|S^\dagger\| \|S - \tilde{S}\| \leq \sqrt{m + n + 1}\|S^\dagger\| \|x(r) - x(\bar{r})\| \]
\[ \leq \sqrt{2(m + n + 1)} \kappa(S) d(r, \bar{r}) \leq 1/3 \]
by assumption (5.1) on $\bar{r}$. Hence $\|\tilde{S}\| \leq (1 + \|E\|) \|S\| \leq \frac{3}{2} \|S\|$. Also, $(I - E)^{-1}S^\dagger$ is a right inverse of $\tilde{S}$, showing that $\tilde{S}$ has full row rank, and that
\[ \|\tilde{S}^\dagger\| = \|\tilde{S}^\dagger(I - E)^{-1}S^\dagger\| \leq \|(I - E)^{-1}\| \|S^\dagger\| \leq \frac{3}{2}\|S^\dagger\|, \]
from which the first assertion follows.

For the second part, we know from Lemma 3.1 that rank $\tilde{S} < m + n + 1$, and hence for the smallest singular value of $S$ by the Eckhard-Young Theorem
\[ \frac{1}{\|S^\dagger\|} = \sigma_{\min}(S) \leq \|\tilde{S} - S\| \leq \sqrt{2(m + n + 1)} \|S\| d(r, \bar{r}), \]
as claimed above. 

We are now prepared to proceed with a proof of Theorem 1.3.
Proof. of Theorem 1.3 (a). Let \( \sigma, \tau \in \mathbb{D} \) with \( \tilde{r}(\sigma) = 0, \tilde{r}(\tau) = \infty \), then \( \chi(r(\tau), r(\sigma)) \geq 1/3 \) because of

\[
1 = \chi(\tilde{r}(\tau), \tilde{r}(\sigma)) \leq \chi(r(\tau), r(\sigma)) + \chi(\tilde{r}(\tau), r(\tau)) + \chi(r(\sigma), \tilde{r}(\sigma)) \\
\leq \chi(r(\tau), r(\sigma)) + 2\chi_B(r, \tilde{r}) \leq \chi(r(\tau), r(\sigma)) + \frac{2}{3}.
\]

Consider the spherical derivative

\[
r^{#}(z) := \frac{|r'(z)|}{1 + |r(z)|^2}.
\] (5.2)

We claim that

\[
\chi(r(\tau), r(\sigma)) \leq \max_{z \in \mathbb{D}} r^{#}(z) \leq \sqrt{2}(m + n + 1)^{3/2} \kappa(S)
\] (5.3)

which implies that \( |\tau - \sigma| \geq 1/(3\sqrt{2}(m + n + 1)^{3/2} \kappa(S)) \), as claimed in Theorem 1.3.

In order to show the left-hand inequality of (5.3), recall from [17] that the chordal metric is dominated by

\[
\forall w_1, w_2 \in \mathbb{C} : \chi(w_1, w_2) \leq \int_{\gamma} \frac{|dw|}{1 + |w|^2},
\]

where \( \gamma \) is any differentiable curve in the extended complex plane joining \( w_1 \) with \( w_2 \). Taking \( \gamma : \mathbb{D} \supset |\sigma, \tau| \ni z \mapsto r(z) \in \mathbb{C} \), we conclude that

\[
\chi(r(\sigma), r(\tau)) \leq \int_{\gamma} \frac{|dw|}{1 + |w|^2} = \int_{z \in [\sigma, \tau]} r^{#}(z) \, dz \leq |\sigma - \tau| \max_{z \in \mathbb{D}} r^{#}(z),
\]

as claimed above. It remains to give an upper bound for \( r^{#}(z) \) for \( z \in \mathbb{D} \), here we closely follow arguments of the proof of Theorem 4.1. We have

\[
r^{#}(z) = \frac{|p'(z)q(z) - q'(z)p(z)|}{|p(z)|^2 + |q(z)|^2} \leq \frac{\| \begin{bmatrix} p'(z) \\ q'(z) \end{bmatrix} \|}{\sqrt{|p(z)|^2 + |q(z)|^2}} \leq \| S^{\#} \| \left\| \begin{bmatrix} p'(z) \\ q'(z) \end{bmatrix} \right\|,
\]

where in the last step we have applied (4.9). Since

\[
\left\| \begin{bmatrix} p'(z) \\ q'(z) \end{bmatrix} \right\| = \left\| \begin{bmatrix} 0, 1, 2z, ..., m z^{m-1} \\ 0, 1, 2z, ..., n z^{n-1} \end{bmatrix} \right\| \leq (m + n + 1)^{3/2}
\]

and \( 1 \leq \sqrt{2} \| S^{\#} \| \) by (4.5), we obtain the second inequality claimed in (5.3), and hence the part of Theorem 1.3 on Froissart doublets is shown. \( \square \)

Proof. of Theorem 1.3 (b). We start by observing that for the residual \( \alpha_0 \) of a simple pole \( z_0 \in \mathbb{D} \) of \( r = p/\tilde{q} \in \mathcal{R}_{m,n} \) there holds

\[
\frac{1}{|\alpha_0|} = \frac{|q'(z_0)|}{|p(z_0)|} = r^{#}(z_0) \leq \sqrt{2}(m + n + 1)^{3/2} \kappa(S)
\]

where for the last inequality we have applied (5.3). The assumption \( 2(m + n + 1)^2 \kappa(S) \chi_B(r, \tilde{r}) \leq 1/3 \) together with Theorem 4.1 tells us that (5.1) holds, and thus also \( \tilde{r} \) is nondegenerate. By applying the same reasoning as for \( r \), we obtain for the residual \( \tilde{\alpha}_0 \) of a simple pole \( \tilde{z}_0 \in \mathbb{D} \) of \( \tilde{r} = \tilde{p}/\tilde{q} \in \mathcal{R}_{m,n} \) the claimed inequality

\[
\frac{1}{|\tilde{\alpha}_0|} \leq \sqrt{2}(m + n + 1)^{3/2} \kappa(\tilde{S}) \leq 2\sqrt{2} (m + n + 1)^{3/2} \kappa(S)
\]

where for the last inequality we have applied the first part of Lemma 5.1. \( \square \)

Remark 5.2. Recall from the above proof of Theorem 1.3 (b) that we have shown the lower bound \( 1/((2(m + n + 1))^{3/2} \kappa(S)) \) for the modulus of any residual of a simple pole in the unit disk of any \( \tilde{r} \in \mathcal{R}_{m,n} \).
solely under the hypothesis $\sqrt{2(m+n+1)} d(r, \tilde{r}) \kappa(S) \leq 1/3$, which according to Theorem 4.1 is weaker than the hypothesis $2(m+n+1)^2 \kappa(S)^2 \chi_\Delta(r, \tilde{r}) \leq 1/3$ stated in Theorem 1.3(b), and stronger than the hypothesis $\chi_\Delta(r, \tilde{r}) \leq 1/3$ of Theorem 1.3(a).

In the numerical procedure described in [15], the authors do not necessarily return the $[m/n]$ Padé approximant $r = p/q$ but $\tilde{r} = \tilde{p}/\tilde{q}$ obtained by replacing the $\ell$ leading coefficients of $p$ (or of $q$, but not of both since otherwise $\kappa(S)$ would be large) of modulus $\leq \epsilon$ by 0. Thus $d(r, \tilde{r}) \leq \|x(r) - x(\tilde{r})\| \leq \sqrt{\epsilon} \kappa, \epsilon$ and Theorem 1.3(a),(b) do apply provided that $\sqrt{2(m+n+1)} \kappa(S) \leq 1/3$.}

**Remark 5.3.** By examining the above proofs and using elementary techniques of complex analysis we see that it is possible to generalize Theorem 1.3 to the case $r, \tilde{r} \in \mathcal{M}(\mathbb{D})$ of general meromorphic functions (at least if $r$ has no zeros/poles on the unit circle), but the price to pay is that the constants become less explicit, in particular there is no longer the condition number of a matrix.

For instance, by examining the proof of Theorem 1.3(a) we see that we can give a lower bound for the Euclidian distance between a pole and a zero of $\tilde{r}$ in terms of the reciprocal of the maximum spherical derivative of $r$ on the unit disk $\mathbb{D}$ provided that $\chi_\Delta(r, \tilde{r}) \leq 1/3$. Moreover, from the Rouché Theorem we see that for any sufficiently small $\epsilon > 0$ there exists a (computable) $\delta > 0$ depending on $r$ and $\epsilon$ such that, for any $\tilde{r} \in \mathcal{M}(\mathbb{D})$ with $\chi_\Delta(r, \tilde{r}) \leq \delta$ we have that the $\epsilon$-neighborhood of any pole or zero of $r$ contains the same number of poles or zeros of $\tilde{r}$ counting multiplicities as $r$, and $\tilde{r}$ has no other poles and zeros in $\mathbb{D}$. This constitutes an alternative approach to control Froissart doublets of $\tilde{r}$.

In addition, by possibly choosing a smaller $\delta > 0$ we may insure that, for a simple pole of $r$, the residual of the corresponding simple pole of $\tilde{r}$ differs from that of $r$ at most by $\epsilon$, giving a possibility to exclude small residuals for $\tilde{r}$. Thus we may roughly summarize by saying that if $\chi_\Delta(r, \tilde{r})$ is sufficiently small then $r$ has a spurious pole if and only if $\tilde{r}$ has.

**Proof.** of Theorem 1.4. By assumption and the second part of Lemma 5.1

$$\sqrt{2(m+n+1)} d(r, \tilde{r}) \kappa(S) \geq 1,$$

and a combination with Theorem 4.1 implies that $2(m+n+1)^2 \chi_\Delta(r, \tilde{r}) \kappa(S)^2 \geq 1$, as claimed in Theorem 1.4. ☐

6. Numerical GCD and other related results.

6.1. Froissart doublet and numerical GCD. One could wonder whether the existence of Froissart doubles of a rational function $r = p/q \in \mathcal{R}_{m,n}$, namely the existence of a zero $\sigma$ and a pole $\tau$ of $r$ with small Euclidian distance $|\sigma - \tau|$, is related to the fact that the pair $(p, q)$ is close to a similar pair $(\tilde{p}, \tilde{q})$ with non-trivial greatest common divisor (GCD), or more generally being degenerate, that is, the quantity

$$\epsilon(p, q) := \min \left\{ \|x(r) - \left[ \begin{array}{c} \text{vec}(\tilde{p}) \\ \text{vec}(\tilde{q}) \end{array} \right] \| : \tilde{p}/\tilde{q} \in \mathcal{R}_{m,n} \text{ is degenerate} \right\}$$

is small. This quantity has been discussed in [5]. According to [5] Theorem 4.1 and Remark 4.3] we have

$$\epsilon(p, q) = \inf_{z \in \mathbb{C}} \sqrt{\frac{|p(z)|^2}{1 + |z|^2 + ... + |z|^{2m}} + \frac{|q(z)|^2}{1 + |z|^2 + ... + |z|^{2n}}} \quad (6.1)$$

the argument $z^*$ where the infimum is attained being called the closest common root (which is indeed a common root of the closest degenerate pair). Since $p'(\sigma) = 0$, we get in first order that

$$|p(\sigma)|^2 \approx |p'(\sigma)(\sigma - \tau)|^2 \leq m^2 \|\text{vec}(p)\|^2 (1 + |\tau|^2 + ... + |\tau|^{2m}) |\sigma - \tau|^2$$

with $\|\text{vec}(p)\| \leq 1$ because of the normalization of $x(r)$. Using a similar argument for $q(\sigma)$, we conclude that, in first order, $\epsilon(p, q) \lesssim \min(m, n) |\sigma - \tau|$ for small $|\sigma - \tau|$, in other words, a Froissart doublet implies a non-trivial numerical GCD.

6.2. Numerical GCD and (structured) smallest singular values. Recall from Lemma 3.1 that $\tilde{r} = \tilde{p}/\tilde{q}$ is degenerate if and only if the corresponding Sylvester-like matrix $\tilde{S}$ is not of full rank. According to the arguments in the proof of, e.g., [15] or Lemma 5.1 the expression $\|x(r) - x(\tilde{r})\|$ in the definition of $\epsilon(p, q)$ can be replaced (up to some modest power of $(m+n+1)$) by $\|S - \tilde{S}\|$ or by
In other words, $\epsilon(p,q)$ is essentially the (absolute or relative) distance of $S$ to the set of singular Sylvester-like matrices, a kind of smallest structured singular value of $S$ (or reciprocal structured condition number). Thus, if we forget about structures, we get $\frac{1}{\kappa(S)} \lesssim \epsilon(p,q)$, what is essentially the finding of the second part of Lemma 5.1. Also, combining with the heuristic arguments of 46.1 we again find $1 \lesssim \kappa(S)|\sigma-\tau|$, a result which is established rigorously in Theorem 1.3(a).

We should mention the relation with [3] who both do not argue in terms of our matrix $S$ defined in (1.5), but in terms of the classical square Sylvester matrix $S$ of order $m + n$ obtained from $S$ by dropping the last column in each column block and the last row (but we believe that this difference is not important). The authors in [3] suggested to look at a gap in the singular values of $S$ in order to find the degree of a numerical GCD, in particular, (normalized) pairs $(p,q)$ of polynomials with sufficiently “large” $\sigma_{\text{min}}(S) \sim 1/\kappa(S)$ should be considered as numerically coprime. This has to be compared with our notion of well-conditioned rational functions where $\kappa(S)$ is modest. In [3] (while working with different vector norms) one gives a lower bound $1/\kappa_{BL}$ for $\epsilon(p,q)$ with the estimator $\kappa_{BL} := \max(||S^{-1}e_1||, ||S^{-1}e_{m+n}||)$, $e_j$ denoting the $j$th canonical vector, being shown to be roughly between $\kappa(S)$ and its square root.

We have the following sharper complement of Theorem 1.4.

Corollary 6.1. For the nondegenerate $[m\mid n]$ Padé approximant $r = p/q$ and the (possibly degenerate) $[m-1\mid n-1]$ Padé approximant $\tilde{r} = u/v$ we have that for all $|z| \leq 1$

$$2(m+n+1)^{3/2} \kappa(S) \chi(r(z),\tilde{r}(z)) \geq |z|^{m+n-1}, \quad \kappa_{BL} \chi(r(z),\tilde{r}(z)) \geq |z|^{m+n-1}.$$  

Proof. Notice that $S^{-1}e_{m+n}$ is a (not normalized) coefficient vector of the degenerate rational function $u/v \in \mathcal{R}_{m,n}$ (more precisely, $u/v \in \mathcal{R}_{m-1,n-1}$) satisfying $q(z)u(z) - p(z)v(z) = z^{m+n-1}$ and hence $q(z)(f(z)v(z) - u(z)) = v(z)(f(z)q(z) - p(z)) + O(z^{m+n-1}) = O(z^{m+n-1})$. Then the relation $q(0) \neq 0$ implies that $\tilde{r} = u/v$ is the $[m-1\mid n-1]$ Padé approximant of $f$. Writing in this proof $\hat{S} \in \mathbb{C}^{(m+n-1)\times (m+n)}$ for the Sylvester matrix of $(u,v)$, we find as in the proof of (3.4) that $||\hat{S}|| \leq \sqrt{m+n+1 ||S^{-1}e_{m+n}||} \leq \sqrt{m+n+1} \kappa_{BL} \leq \sqrt{m+n+1} ||S^{-1}||$. Also, since one is a submatrix of the other, we also have that $||\hat{S}|| \leq ||S|| \leq \min\{\sqrt{m+n+1}, 2||\hat{S}||\}$. Consequently, for all $|z| \leq 1$

$$\min\left\{ (m+n+1)^{1/2} \kappa_{BL}, 2(m+n+1)^{3/2} \kappa(S) \right\} \chi(r(z),\tilde{r}(z)) \geq (m+n+1)||S|| ||\hat{S}|| \chi(r(z),\tilde{r}(z)) \geq |z|^{m+n+1} - \frac{\sqrt{m+n+1} ||\hat{S}||}{\sqrt{\|p(z)^2 + q(z)^2\|}} \geq |z|^{m+n+1},$$  

where in the last inequality we have applied twice (4.12). \hfill \Box

Taking the maximum for $z \in \mathbb{D}$, Corollary 6.1 implies that $2(m+n+1)^{3/2} \kappa(S) \chi(\mathbb{D}(r,\tilde{r})) \geq 1$, which we expect to be sharper than Theorem 1.4 since there the factor $\kappa(S)^2$ did occur. Moreover, we may even take the minimum over $|z| = 1$ in Corollary 6.1 showing that, at each point $z$ of the unit circle, $\chi(r(z),\tilde{r}(z)) \geq 1/(2(m+n+1)^{3/2} \kappa(S))$. Thus, roughly speaking, if $\chi(\mathbb{D}(f,\tilde{f}))$ is small and $\kappa(S)$ not too large then even the minimal chordal distance between $f$ and $\tilde{f}$ on the unit circle is not small.

6.3. The work of Cabay and Meleshko. In order to jump over “numerical blocks” in the Padé table by some look-ahead procedure, Cabay and Meleshko [7] (see also [2], Section 3.6) needed to decide whether $[m\mid n]$ Padé approximant $r = p/q$ of $f$ is significantly different from the the $[m-1\mid n-1]$ Padé approximant $\tilde{r} = \tilde{p}/\tilde{q}$. Denoting by $e$ the first column of the rectangular matrix $C$ introduced in (1.2) and by $\tilde{C}$ the square Toeplitz matrix of order $n$ formed by the other columns, we know from (1.2) that, with a suitable scalar $\tilde{e}$,

$$\text{vec}(q) = q(0) \begin{bmatrix} 1 \\ -C^{-1}e \end{bmatrix}, \quad \text{vec}(\tilde{q}) = \begin{bmatrix} 0 \\ \tilde{e} \end{bmatrix}, \quad \text{and thus vec}(\tilde{q}) = \tilde{r}C^{-1}e_n$$

where $\tilde{q}(z)f(z) - \tilde{p}(z) = \tilde{e}z^{m+n+1} + O(z^{m+n})$. The authors in [7] suggested to use the normalization $\|\text{vec}(q)\| = \|\text{vec}(\tilde{q})\| = 1$ and used the indicator

$$\kappa_{CM} = \frac{1}{|q(0)e|}$$

\footnote{Compare also (6.1) with (4.3).}

\footnote{Up to permutations and changes of sign of some columns which does not affect norms.}
as an estimator for $\|C^{-1}\|$, motivated by parts (i),(ii) of the following statement.

**Lemma 6.2.** We have that (i) $\kappa_{CM} \geq \|C^{-1}\|/n$, (ii) $\kappa_{CM} \leq \sqrt{n} \|C^{-1}\|/r$, (iii) $\sigma_{\min}(C) \geq 1/(n\kappa_{CM})$, and (iv) $\kappa_{CM} \sim \|S^{-1}e_{m+n}\| \leq \kappa_{BL}$.

**Proof.** The Gohberg-Semencul formula [2, Theorem 3.6.2] tells us that $q(0)\overline{e}C^{-1} = A_1A_2 - A_3A_4$, with the four matrices $A_1, A_2, A_3, A_4$ given by the triangular Toeplitz matrices

$$
\begin{bmatrix}
q_0 & 0 & \cdots & 0 \\
q_1 & q_0 & \cdots & 0 \\
& \ddots & \ddots & \vdots \\
& & q_{n-1} & q_0 \\
& & & \tilde{q}_0
\end{bmatrix},
\begin{bmatrix}
\tilde{q}_{n-1} & \cdots & \tilde{q}_1 & \tilde{q}_0 \\
0 & \tilde{q}_{n-1} & q_1 & 0 \\
& \ddots & \ddots & \vdots \\
& & 0 & \tilde{q}_{n-2} & \cdots & \tilde{q}_0 \\
& & & 0 & \cdots & \cdots & \cdots \\
& & & & 0 & \cdots & 0 & q_n
\end{bmatrix}.
$$

Hence

$$
\kappa_{CM} \|C^{-1}\| = \|A_1\|_F\|A_2\|_F + \|A_4\|_F \|A_3\|_F \leq \sqrt{\|A_1\|_F^2 + \|A_4\|_F^2} \sqrt{\|A_1\|_F^2 + \|A_4\|_F^2} = n,
$$

as claimed in part (i). By the normalization of the denominators we also find that

$$
\kappa_{CM} = \frac{1}{\|q(0)e\|} = \|C^{-1}\|/r \leq \sqrt{\|C^{-1}\|_F^2 + \|C^{-1}\|_F^2} \leq \|C^{-1}\|_F^2 \leq \|C^{-1}\|_F, 
$$

where $\|C\|_F \leq \sqrt{n}$ by (1.3), implying (ii). In view of (i), for establishing (iii) it is sufficient to notice that

$$
\sigma_{\min}(C) = \min_{x \not= 0} \frac{\|x^*C\|}{\|x\|} \leq \min_{x \not= 0} \frac{\|x^*C\|}{\|x\|} \leq \sigma_{\min}(C).
$$

A proof of part (iv) is slightly more involved. Notice first that the normalization $\|\overline{\text{vec}}(\tilde{q})\| = 1$ of [7] does not lead to coefficient vectors $x(\tilde{r})$ of norm 1, but $\|\overline{\text{vec}}(\tilde{q})\| \leq \|x(\tilde{r})\| \leq (1 + \|T\|)\|\overline{\text{vec}}(\tilde{q})\| \leq (1 + \sqrt{m + n + 2})\|\overline{\text{vec}}(\tilde{q})\|$ by Lemma 3.3 and thus $\|x(\tilde{r})\| \sim 1$. We have

$$
p(z)\tilde{q}(z) - \overline{p}(z)q(z) = q(z)(\tilde{q}(z)f(z) - \overline{p}(z)) - \overline{q}(z)(q(z)f(z) - p(z)) = q(0)\overline{e}z^{m+n-1}
$$

since it is a polynomial of degree at most $m + n - 1$, and the powers $z^j$ vanish for $j < m + n - 1$. This latter identity can be rewritten as $\overline{S}x(\tilde{r}) = -q(0)\overline{e}e_{m+n}$, and thus

$$
\kappa_{CM} = \frac{\|S^{-1}e_{m+n}\|}{\|x(\tilde{r})\|} \sim \|S^{-1}e_{m+n}\| \leq \kappa_{BL}
$$

the last inequality following directly from the definition of $\kappa_{BL}$. This shows part (iv). \qed

The algorithm presented in [7] tries out all Padé approximants of type $[m - j]n - j$ for integer $j$ (i.e., on the same diagonal), and accepts to compute the $[m/n]$ Padé approximant if $\kappa_{CM}$ is sufficiently small. By Lemma 6.2(iii), this means that in the Cabay-Meleshko algorithm we only compute robust Padé approximants in the sense of (P2), that is, in the sense of Gonnet, Gütel and Trefethen [15].

Finally, as in the proof of Corollary 6.1 we get from Lemma 6.2(iv) that $|z|^{m+n-1} \leq \kappa_{CM}x(r(z), \tilde{r}(z))$ for all $|z| \leq 1$. In particular, a sufficiently small $\kappa_{CM}$ implies that $r$ and $\tilde{r}$ are indeed significantly different.

**6.4. Forward conditioning of Padé values and simple Padé poles.** The fact that we know the Jacobian of the real Padé map $J_F(c) = T^tQ = S^tQ^2$ in the nondegenerate case (see Lemma 3.2) allows us to study the forward conditioning of other related maps. Write for a moment more explicitly the dependency on the vector $c$ of $m + n + 1$ Taylor coefficients as

$$
p_c(z) = (1, z, \ldots, z^m, 0, \ldots, 0)F(c), \quad q_c(z) = (0, \ldots, 0, 1, z, \ldots, z^n)F(c).
$$

**Corollary 6.3.** The forward conditioning of the value map $\mathbb{R}^{m+n+1} \ni c \mapsto v(c) = p_c(z)/q_c(z)$ at a fixed argument $z$ (different from a pole or a zero) is given by

$$
\kappa_{for}(v)(c) = \frac{\|(1, z, z^2, \ldots, z^{m+n})Q^2\|}{|p_c(z)q_c(z)|}.
$$
Provided that \( m \geq n - 1 \), the forward conditioning of the root map \( \mathbb{R}^{m+n+1} \ni c \mapsto \rho(c) \in \mathbb{C} \) with \( \rho(c) \) being a simple root of \( q_c \) is given by

\[
\kappa_{for}(\rho)(c) = |\rho(c)|^{m+n} \frac{\|\text{vec}(\tilde{\ell})\|}{|\alpha_c|}, \quad \tilde{\ell}(z) = (\rho(c) \ell(z))^2, \quad \ell(z) = \frac{q_c(z)}{(z - \rho(c)) q_c'(\rho(c))},
\]

with \( \alpha_c \) the corresponding residual of \( p_c/q_c \) at \( \rho(c) \).

**Proof.** For establishing the first formula, notice that

\[
\kappa_{for}(v)(c) = \frac{\|J_v(c)\|}{\|v(c)\|} = \frac{\|q_c(z)(1, z, \ldots, z^m, 0, \ldots, 0) J_F(c) - p_c(z)(0, \ldots, 0, 1, z, \ldots, z^n) J_F(c)\|}{|p_c(z) q_c(z)|} = \frac{\|J_v(z, z^2, \ldots, z^{m+n}) SS^t Q^2\|}{|p_c(z) q_c(z)|}.
\]

A proof for the second one is slightly more involved. We have

\[
(0, \ldots, 0, 1, \rho(c), \ldots, \rho(c)^n) J_F(c) + q_c'(\rho(c)) J_F(c) = 0
\]

or, writing \( \alpha_c = \frac{p_c(\rho(c))}{q_c'(\rho(c))} \) for the corresponding Padé residual

\[
\kappa_{for}(\rho)(c) = \frac{\|J_F(c)\|}{\|\rho(c)\|} = \frac{\|q_c(\rho(c)) (1, \rho(c), \ldots, \rho(c)^m, 0, \ldots, 0) J_F(c) - p_c(\rho(c)) (0, \ldots, 0, 1, \rho(c), \ldots, \rho(c)^n) J_F(c)\|}{|p_c(\rho(c)) q_c'(\rho(c))|^2} = \frac{\|1, \rho(c), \rho(c)^2, \ldots, \rho(c)^{m+n}) Q^2\|}{|p_c(\rho(c)) q_c'(\rho(c))|^2}.
\]

Notice that, with \( Q \), also \( Q^2 \) is a triangular Toeplitz matrix, generated by the coefficients of \( q_c(z)^2 \). Hence, since \( m \geq n - 1 \), we obtain for \( \omega \) being an \((n + m + 1)\) root of unity,

\[
(1, \rho(c), \rho(c)^2, \ldots, \rho(c)^{m+n}) Q^2(1, \omega, \ldots, \omega^{m+n})^t = \frac{\rho(c)^{m+n+1} \omega^{m+n+1} q_c(1/\omega)^2 - q(\rho(c))^2}{\rho(c) - 1/\omega} = \frac{\rho(c)^{m+n+1} \omega^{m+n} q_c(1/\omega)^2}{\rho(c) - 1/\omega},
\]

and hence with the DFT matrix \( \Omega = (\exp(2\pi i j k/(m + n + 1)))_{j,k=0,\ldots,m+n} \).

\[
\kappa_{for}(\rho)(c) = \frac{\|1, \rho(c), \rho(c)^2, \ldots, \rho(c)^{m+n}) Q^2\|}{|p_c(\rho(c)) q_c'(\rho(c))|^2} = |\rho(c)|^{m+n} \frac{\|\text{vec}(\tilde{\ell})\|}{|\alpha_c|} = |\rho(c)|^{m+n} \frac{\|\text{vec}(\tilde{\ell})\|}{|\alpha_c|}.
\]

\( \square \)

7. **Conclusions.** In this paper we have presented several results on the sensitivity of \(|m|n|\) Padé approximants, as well as on the occurrence of spurious poles. Our findings are expressed in terms of four matrices, namely a rectangular Toeplitz matrix \( C \) as in \([15]\), a rectangular striped Toeplitz matrix \( T \), a square triangular Toeplitz matrix \( Q \), and a rectangular Sylvester-like matrix \( S = Q T \), see \([12],[14],[16]\). These four matrices satisfy

\[
\|C\| \lesssim 1, \|T\| \sim 1, \|S\| \sim 1, \|Q\| \lesssim 1 \text{ due to scaling (see (1.3) and Lemma 3.3),}
\]

\[
\|C^t\| \sim \|T^t\| \sim \kappa(T) \text{ and } |T^t| \lesssim |S^t| \sim \kappa(S), \text{ and } \|Q^{-1}\| \lesssim \|S^t\| \text{ (see Lemma 3.3).}
\]

We introduced a kind of hierarchical classification of \(|m|n|\) Padé approximants: there are first the so-called nondegenerate Padé approximants \( r = p/q \) considered before in \([25]\) which can characterize equivalently by one of the following properties
• the polynomials $p$ and $q$ are co-prime, and that the defect $\min\{m - \deg p, n - \deg q\}$ is equal to zero (see (P1)), in other words, they correspond to entries located on the left or upper border of a block in the Padé table (in exact arithmetic);
• the Padé map is continuous (see [25] and Theorem 1.1);
• the matrix $S$ (and hence $T$ and $C$) has full row rank (see Lemma 3.1).

Secondly there is the subclass of so-called robust Padé approximants in the sense of [15] characterized by a modest $\sigma_{\text{min}}(C)$, or, equivalently, a modest $\kappa(T)$. We show that here
• the real Padé map is forward well-conditioned (but not necessarily backward, see Theorem 1.2 and Example 2.3);
• the Cabay-Meleshko algorithm [7] of [6.3] computes also robust Padé approximants along a diagonal, it is (most of the times) cheaper than the approach of [15] since it is recursive, but it might miss some robust approximants since the estimator $\kappa_{CM}$ might be larger than $\kappa(T)$;

Finally we have introduced in this paper the class of so-called well-conditioned Padé approximants characterized by a modest $\kappa(S)$ (which is hence a subclass of that of robust approximants). For these approximants we have established the following properties
• we can control both Froissart doublets (Euclidian distance between poles and zeros of $r$ in the unit disk) as well as small residuals in the disk, see Theorem 1.3;
• the real Padé map is backward well-conditioned (since $\|Q^{-1}T\| \leq \kappa(S)$, see Theorem 1.2);
• it is equivalent to measure the distance to $\tilde{r} \in \mathcal{R}_{m,n}$ through the uniform chordal metric in the unit disk or through the difference of normalized coefficient vectors (see Theorem 4.1);
• its numerator and denominator are numerically coprime in the sense of [5, 8], see [6.2].

In the introduction we mentioned the question from [15] whether robust approximants do not have Froissart doublets nor small residuals. Our Example 2.3 shows that such a statement is wrong in general, but we were able to give a positive answer at least for well-conditioned Padé approximants.

We can also draw from Theorem 1.3, Theorem 1.4 and Remark 5.3 the conclusion that it is impossible to find well-conditioned Padé approximants close to $f$ in $\mathbb{D}$ with small error for "wired" functions $f$ having themselves small residuals or Froissart doublets in the disk. However, the scaling assumption (1.3) at least asymptotically scales the complex plane in a way that $f$ will have no singularities in the (open) disk.

More important, Theorem 1.4 and even more Corollary 6.1 seem to indicate that there are only finitely many well-conditioned Padé approximants along a fixed diagonal which are close to $f$ in the whole unit disk, or even only close at one point of the unit circle.

For future work, it seems for us desirable to get a deeper understanding of the link between $\kappa(S)$ and $\kappa(T)$ (beyond the relation $\kappa(T) \leq \kappa(S)$), that is, the link between Padé approximants which are robust and those which are well-conditioned.

Also, it would be nice to know whether the lower bounds of, e.g., Theorem 1.3 are sharp. We feel that the lower bounds should not involve unstructured condition numbers but so-called structured condition numbers, the latter taking into account the particular structure of our matrices $S, C$ or $T$. This will be further analyzed in a future work.

Acknowledgements. The authors gratefully acknowledge valuable discussions with Alexander Aptekarev and Stefan Güttel.

REFERENCES

[1] V. M. Adukov and O. L. Ibryaeva, A new algorithm for computing Padé approximants, J. Comput. Appl. Math., 237 (2013), pp. 529-541.
[2] G. A. Baker, Jr. and P. R. Graves-Morris, Padé Approximants, 2nd ed., Cambridge Univ. Press, 1996.
[3] B. Beckermann, The Condition Number of real Vandermonde, Krylov and positive definite Hankel matrices, Numer. Mathematik 85 (2000) 553-577.
[4] B. Beckermann, G. Golub, G. Labahn, On the numerical condition of a generalized Hankel eigenvalue problem, Numer. Mathematik 106 (2007) 41-68.
[5] B. Beckermann, G. Labahn, When are two numerical polynomials relatively prime? J. Symbolic Computations 26 (1998) 677-689.
[6] D. Bessis, Padé approximations in noise filtering, J. Comput. Appl. Math., 66 (1996), pp. 85-88.
[7] 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.
[8] R.M. Corless, P.M. Gianni, B.M. Trager & S.M. Watt, The Singular Value Decomposition for Polynomial Systems, Proceedings ISSAC ’95, ACM Press (1995) 195-207.
[9] N. Daras, V. Nestoridis, and C. Papadimitropoulos, Universal Padé approximants of Seleznev type, Arch. Math. 100 (2013), 571–585.

[10] M. Froissart, Approximation de Padé: application à la physique des particules élémentaires, in RCP, Programme No. 25, v. 9, CNRS, Strasbourg (1969), pp. 1-13.

[11] K.O. Geddes, S.R. Czapor and G. Labahn, Algorithms for Computer Algebra, Kluwer Academic Publishers, (1992).

[12] J. Gilewicz and M. Pindor, Padé approximants and noise: a case of geometric series, J. Comput. Appl. Math., 87 (1997), pp. 199-214.

[13] J. Gilewicz and M. Pindor, Padé approximants and noise: rational functions, J. Comput. Appl. Math., 105 (1999), pp. 285-297.

[14] A.A. Gonchar, On the convergence of generalized Padé approximants of meromorphic functions, Math. Sbornik 27 (1975) 503-514.

[15] P. Gonnet, S. Güttel and L. N. Trefethen, Robust Padé approximation via SVD, SIAM Review, 55 (2013), pp. 101-117.

[16] W.F. Mascarenhas, Robust Padé Approximants Can Diverge, arXiv:1309.5753 (2013).

[17] J.L. Schiff, Normal Families, Springer Verlag 1993.

[18] H. Stahl, Diagonal Padé approximants to hyperelliptic functions. Annales de la Faculté des Sciences de Toulouse, no spécial Stieltjes (1996) 121-193.

[19] H. Stahl, The convergence of diagonal Padé approximants and the Padé conjecture. J. Comput. Appl. Math. 86 (1997), 287296.

[20] H. Stahl, The convergence of Padé approximants to functions with branch points. J. Approximation Theory 91 (1997), 139-204.

[21] H. Stahl, Spurious poles in Padé approximation, J. Comp. Appl. Math., 99 (1998), 511-527.

[22] L. N. Trefethen, Square blocks and equisoscillation in the Padé, Walsh, and CF tables, in P. R. Graves-Morris, E. B. Saff, and R. S. Varga, eds., Rational Approximation and Interpolation, Springer Lect. Notes in Math. 1105, 1984.

[23] L. N. Trefethen, Approximation Theory and Approximation Practice, SIAM, 2013.

[24] L. N. Trefethen and D. Bau, III, Numerical Linear Algebra, SIAM, 1997.

[25] H. Werner and L. Wuytack, On the continuity of the Padé operator, SIAM J. Numer. Anal., Vol. 20, (1983), 1273-1280

Bernhard Beckermann, Ana C. Matos, 
{bbecker, matos}@math.univ-lille1.fr
Laboratoire de Mathématiques P. Painlevé UMR CNRS 8524 - Bat.M2
Université Lille - Sciences et Technologies
F-59655 Villeneuve d’Ascq Cedex, FRANCE