Université de Lille
Sciences et Technologies
Ecole Doctorale
Sciences Pour l'Ingénieur
Lille Hauts-de-France

Laboratoire Paul Painlevé

Thèse pour obtenir le grade de docteur en sciences de l'Université de Lille

Présentée par Joanna Bisch
Soutenue le 22 octobre 2021

Discipline : **Mathématiques appliquées**

Fonctions de Matrices de Toeplitz symétriques

Thèse dirigée par **Bernhard Beckermann**

Jury :

Rapporteurs :		
Mme Paola BOITO	Professeur des Universités	Università di Pisa
M. Andreas FROMMER	Professeur des Universités	Bergische Universität Wuppertal
Directeur de thèse :		
M. Bernhard BECKERMANN	Professeur des Universités	Université de Lille
Examinateurs :		
Mme Ana MATOS	Maîtresse de conférences	Université de Lille
Mme Martine OLIVI	Chargée de recherches	INRIA Sophia Antipolis
Président du Jury :		
M. Emmanuel CREUSE	Professeur des Universités	Université Polytechnique Hauts-de-France

RESUME

Le calcul numérique de fonctions de matrices $f(A)$ avec A matrice carrée de Toeplitz ou Toeplitz-like de taille $n \times n$ trouve son intérêt dans divers domaines mathématiques, que ce soit pour la discrétisation d'une équation integro-différentielle partielle ou la construction de systèmes de filtres. Or, les méthodes classiques de calcul de fonctions de matrices n'utilisant aucune structure particulière de la matrice A, celles-ci sont alors de complexité $\mathcal{O}(n^3)$. Dans cette thèse nous cherchons à réduire cette complexité à $\mathcal{O}(n^2)$ voire $\mathcal{O}(n \log^2(n))$ en exploitant la structure Toeplitz-like de A, notamment à l'aide de l'approximation rationnelle de notre fonction f. Après avoir donné quelques rappels concernant les fonctions de matrices et leur approximation, nous définissons une arithmétique sur les matrices Toeplitz-like, formant un sous-ensemble de matrices permettant des opérations rapides, comme l'addition, la multiplication ou l'inversion, réduisant le calcul d'une fonction rationnelle de matrice à une complexité $\mathcal{O}(n^2)$ voire $\mathcal{O}(n \log^2(n))$. A l'aide de cette nouvelle arithmétique, nous nous attardons ensuite sur l'approximation des fonctions de matrice racine carrée et signe pour lesquelles nous reprenons et accélérons la méthode de Newton sous ses différentes formes. Enfin nous nous intéressons à l'approximation des fonctions de matrices $f(A)$ par $r(A)$ lorsque f est une fonction de Markov avec $f : z \mapsto \int_\alpha^\beta \frac{d\mu(x)}{z-x}$ où μ est une mesure positive à support dans $[\alpha; \beta]$ et A une matrice de Toeplitz symétrique. Nous énonçons alors l'un de nos principaux résultats qui est une borne supérieure pour l'erreur relative d'interpolation rationnelle $1 - r/f$ sur l'intervalle spectral de A. Cette borne est ensuite optimisée par un choix particulier des points d'interpolation. Nous discutons alors de la précision de trois représentations différentes de nos interpolants rationnels et appuyons nos résultats d'applications numériques pour un argument scalaire. Un résultat sur la positivité des paramètres de la fraction continue de Thiele dans le cas de fonctions de Markov est également démontré. Ces résultats sont enfin appliqués au cas matriciel et nous proposons diverses reformulations a priori et a posteriori de notre borne sur $1 - r/f$ pour une fonction de Markov f et un argument matriciel. De nombreuses expériences numériques illustrent notre démarche.

ABSTRACT

The numerical computation of matrix functions $f(A)$ where A is a square Toeplitz or Toeplitz-like matrix of size $n \times n$ is usefull in many mathematical fields, such a the discretization of a partial integro-differential equation, or the construction of digital filters. However, classical methods to compute functions of matrices $f(A)$ don't use any particular structure of the matrix A and have a complexity $O(n^3)$. In this thesis we seek at reducing this complexity to $\mathcal{O}(n^2)$ or $\mathcal{O}(n \log^2(n))$ by exploiting the Toeplitz-like structure of A, in particular with the help of rational approximation of our function f. After giving a short summary of the theory of matrix functions and their approximation, we define an arithmetic on Toeplitz-like matrices, forming a subset of matrices allowing fast operations, such as the addition, the multiplication or the inversion, reducing the calculation of rational functions of a matrix to a complexity $\mathcal{O}(n^2)$ or $\mathcal{O}(n \log^2(n))$. With the help of this new arithmetic, we then focus on the approximation to the matrix square root and sign functions for which we recall and accelerate the Newton method and its variants. Finally we look at the approximation to matrix functions $f(A)$ by $r(A)$ when f is a Markov function $f : z \mapsto \int_\alpha^\beta \frac{d\mu(x)}{z-x}$ where μ is a positive measure with support in $[\alpha; \beta]$ and a symmetric Toeplitz matrix A. We then state one of our main results which is an upper bound on the relative error of rational interpolation $1 - r/f$ over the spectral interval of A. This bound is then optimized by a particular choice of the interpolation points. We then discuss the accuracy of three different representations of our rational interpolants and illustrate our results with numerical applications on a scalar argument. A result on the positivity of parameters of the Thiele continued fraction representation in the case of Markov functions is also demonstrated. These results are then applied on matrices et we provide various a priori and a posteriori reformulations of our bound on $1 - r/f$ for a Markov function f and a matricial argument. Several numerical experiments illustrate our approach.

Table des matières

1 Structure de Toeplitz et fonctions de matrices **11**
1.1 La structure Toeplitz . . . 11
1.2 Définitions et propriétés des fonctions de matrices . . . 15
1.2.1 La forme canonique de Jordan . . . 15
1.2.2 Définition par interpolation polynomiale . . . 16
1.2.3 Définition intégrale . . . 18
1.2.4 Généralités . . . 20
1.3 Quelques méthodes de calcul . . . 22
1.3.1 Evaluation polynômiale et méthode de Horner . . . 22
1.3.2 La méthode de Schur-Parlett . . . 23
1.3.3 Méthodes d'approximation . . . 24
1.3.4 Implémentation des interpolants rationnels . . . 30
1.4 Motivations . . . 31
1.5 Conclusion . . . 34

2 Arithmétique Toeplitz-like **35**
2.1 Opérateur de déplacement . . . 36
2.1.1 Opérateur sous forme Sylvester . . . 36
2.1.2 Opérateur de déplacement sur les opérations de matrices . . . 38
2.2 Rang de déplacement et générateurs : les matrices Toeplitz-like . . . 39
2.2.1 Rang de déplacement . . . 40
2.2.2 Construction des générateurs . . . 41
2.2.3 Reconstruction d'une matrice Toeplitz-like à partir de ses générateurs . . . 42
2.2.4 Résolution de systèmes Toeplitz-like . . . 44
2.3 Opérations sur les générateurs de matrices Toeplitz-like . . . 48
2.3.1 Complément de Schur d'une matrice Toeplitz-like . . . 49
2.3.2 Somme, produit et inverse des matrices Toeplitz-like . . . 50
2.3.3 Rang de déplacement numérique et compression des générateurs . . . 53
2.4 Conclusion . . . 54

3 Application aux fonctions de matrices racine carrée et signe **57**
3.1 Racine carrée principale de matrice . . . 57
3.2 Rappels sur la méthode de Newton pour la racine carrée d'une matrice non-structurée . . . 60
3.2.1 Itération de Newton . . . 60
3.2.2 Stabilité et précision asymptotique de l'itération de Newton . . . 62
3.3 Amélioration de la méthode . . . 65
3.3.1 Choix d'un premier terme . . . 65
3.3.2 Introduction de paramètres . . . 66
3.4 Newton pour les matrices Toeplitz-like et expériences numériques . . . 69

3.4.1 Générateurs associés aux itérations de Newton . . . 69
3.4.2 Expériences numériques . . . 72
3.5 La fonction de matrice signe . . . 82
3.5.1 Rappel sur la méthode de Newton pour la fonction signe . . . 82
3.5.2 Introduction de paramètres . . . 83
3.5.3 Newton signe pour les matrices Toeplitz-like et expériences numériques . . . 84
3.6 Conclusion . . . 86

4 Fonctions de Markov appliquées aux matrices de Toeplitz 89
4.1 Erreur d'approximation des fonctions de Markov par interpolation rationnelle . . . 90
4.1.1 Etat de l'art sur la meilleure approximation rationnelle des fonctions de Markov . . . 91
4.1.2 Borne supérieure pour l'erreur d'approximation des fonctions de Markov par interpolation rationnelle . . . 93
4.1.3 Le cas d'un ou 2 points d'interpolation multiples . . . 97
4.1.4 Le cas général . . . 99
4.1.5 Le cas du disque . . . 101
4.2 Représentation des interpolants rationnels aux points optimaux . . . 103
4.2.1 Représentation en éléments simples des interpolants rationnels . . . 103
4.2.2 Représentation barycentrique des interpolants rationnels . . . 104
4.2.3 Représentation en fraction continue des interpolants rationnels . . . 105
4.3 Applications aux matrices de Toeplitz et expériences numériques . . . 112
4.3.1 Interpolants rationnels en arithmétique Toeplitz-like . . . 112
4.3.2 Expériences numériques . . . 113
4.4 Conclusion . . . 125

5 Conclusion et problèmes ouverts 128

Introduction

Soient $A \in \mathbb{C}^{n\times n}$ une matrice et f une fonction. Un problème important en mathématiques appliquées est le calcul de la fonction de matrice

$$f(A) \in \mathbb{C}^{n\times n}$$

lorsque $f(A)$ existe. En effet, ce problème peut apparaître nécessaire pour la résolution de certains systèmes d'équations différentielles ordinaires. De tels systèmes sont obtenus par exemple après discrétisation en espace par différences finies de certaines équations aux dérivées partielles comme l'équation des ondes ou l'équation de la chaleur où apparaissent naturellement des fonctions de matrices de Toeplitz, ainsi que pour une multitude d'autres problèmes [15]. On peut également citer l'exemple de l'équation integro-différentielle partielle donné à l'exemple 1.4.7 en section 1.4. De plus, la discrétisation de l'intervalle d'étude en $n+1$ sous-intervalles de longueur identique fait apparaître lors de la résolution par différences finies pour discrétiser l'équation une matrice de Toeplitz, c'est-à-dire une matrice avec diagonale et sur/sous-diagonales constantes. Ce phénomène s'observe également lors de prélèvements à intervalle de temps régulier dans le domaine de l'analyse stochastique.

Par conséquent, il est courant de rechercher un algorithme de calcul de la fonction de matrice $f(A)$. Dans le cas où $f = p \in \mathcal{P}_m = \{p \text{ polynôme tel que } \deg(p) \leqslant m\}$ avec $m \geqslant 0$, si $p(z) = \sum_{j=0}^{m} a_j z^j$, on sait que la fonction de matrice $p(A)$ est alors donnée par $p(A) = \sum_{j=0}^{m} a_j A^j$ pour toute matrice $A \in \mathbb{C}^{n\times n}$. Si $f(z) = \exp(z)$, on sait que f possède le développement en série entière $\exp(z) = \sum_{j=0}^{\infty} \frac{z^j}{j!}$, et on définit alors la fonction de matrice $\exp(A)$ pour tout matrice $A \in \mathbb{C}^{n\times n}$ par $\exp(A) = \sum_{j=0}^{\infty} \frac{1}{j!} A^j$. Cependant, toute fonction ne pouvant pas forcément être exprimée sous la forme d'une somme (finie ou infinie), il nous faut considérer d'autres définitions d'une fonction de matrice, données en section 1.2 et lorsque $A \in \mathbb{C}^{n\times n}$ avec $n \gg 1$, il se pose alors la question de la complexité de ce calcul. L'un des premiers à étudier les fonctions de matrices dans un cadre général d'une matrice $A \in \mathbb{C}^{n\times n}$ fut Cayley qui en 1858 dans *A Memoir on the theory of Matrices* [21] s'intéressa à la fonction de matrices racine carrée. Peu de temps après apparaissent les premières définitions générales d'une fonction de matrices donnés notamment par Sylvester et d'autres auteurs. Ce sujet a ensuite suscité de nombreuses recherches et on peut trouver dans la littérature plusieurs références sur les fonctions de matrices, notamment Gantmacher [40], Horn et Johnson [59], Lancaster et Tismenetsky [66], Golub et Van Loan [43]. On peut également citer [57] qui constitue une référence plus récente consacré au calcul effectif des fonctions de matrices.

Dans le cas général, il existe différentes méthodes de calcul d'une fonction de matrice $f(A) \in \mathbb{C}^{n\times n}$, un de ces algorithme étant l'algorithme de Schur-Parlett [29] composé de deux étapes : une première étape où on exécute la décomposition de Schur de la matrice A, c'est-à-dire une décomposition sous la forme $A = U^*TU$ avec U unitaire et T triangulaire supérieure, cette décomposition pouvant par exemple être calculé par l'algorithme QR en complexité $\mathcal{O}(n^3)$, voir $\mathcal{O}(n^2)$ dans le cas de matrices de Hessenberg [43, Sections 7.4]. Puis une deuxième étape où, comme $f(A) = U^* f(T) U$, il nous reste à calculer $f(T)$. Or, les éléments diagonaux de la matrice $f(T)$ vérifiant $f(T)_{j,j} = f(T_{j,j})$, nous obtenons les autres éléments de la matrice successivement à l'aide de la relation de Sylvester $f(T)T - Tf(T) = 0$. Notons qu'il existe des versions blocs pour la résolution de l'équation $f(T)T - Tf(T) = 0$ [57, Section 9.2]. Cependant, il n'est

pas clair comment transformer une matrice de Toeplitz en une matrice de Hessenberg en complexité $\mathcal{O}(n^2)$. Donc l'algorithme de Schur-Parlett pour une matrice de Toeplitz reste de complexité globale de $\mathcal{O}(n^3)$, ce coût étant trop élevé lorsque $n \gg 1$. Dans cette thèse, nous allons plutôt approcher $f(A)$ par une expression $g(A)$ au lieu de calculer $f(A)$ explicitement. Ici, g sera une fonction "plus simple", par exemple un polynôme ou une fonction rationnelle suffisamment proche de f dans un sens à définir pour que l'on puisse contrôler l'erreur absolue $f(A) - g(A)$ ou l'erreur relative $I - g(A)f(A)^{-1}$. On peut alors considérer pour fonction g une somme partielle d'une série de Taylor ou d'une série de Faber, ou encore une fonction rationnelle comme un approximant de Padé [12], mais ces choix de g entraînent toujours une complexité d'ordre $\mathcal{O}(n^3)$ opérations élémentaires.

Or, il est important de remarquer que les méthodes de calcul citées précédemment ne prennent pas en compte la structure particulière de la matrice considérée et donc pas la structure à diagonales et sur/sous-diagonales constantes des matrices de Toeplitz. Cependant, la structure particulière de ces matrices a déjà montré quelques avantages en termes de coût de calcul puisqu'un produit entre une matrice de Toeplitz et un vecteur quelconque peut être effectué en $\mathcal{O}(n \log n)$ opérations élémentaires [63, Section 5.3.3], et on peut donc espérer dégager des algorithmes à coût réduit pour d'autres opérations impliquant des matrices de Toeplitz. S. Massei [69, Chap. 7] s'attarde à définir une arithmétique pour les matrices dite quasi-Toeplitz, c'est-à-dire des matrices de la forme $A = T(a) + E$ où $T(a)$ est une matrice de Toeplitz semi-infinie et E est non-Toeplitz avec $\sum_{i,j\in\mathbb{Z}^+} |E_{i,j}| < \infty$. Par approximation des matrices QT (matrices quasi-Toeplitz semi-infinie) par des matrices CQT (matrices quasi-Toeplitz), S. Massei avec Bini et al. [17] démontrent qu'une arithmétique adaptée à ces matrices permet d'exécuter les opérations telles que la somme, la multiplication ou l'inversion avec un coût réduit.

Plus généralement, le cas de matrices structurées a suscité l'intérêt de différents auteurs pour le développement de méthodes de calcul à faible coût. C'est notamment le cas des auteurs S. Massei, L. Robol et D. Kressner qui se sont intéressé aux matrices hiérarchiques [70], c'est-à-dire aux matrices pour lesquelles il existe un r petit devant la dimension n de sorte que toute sous-matrice de A formée de lignes et colonnes successives et ne comportant pas un élément diagonal de A soit de rang au plus r. En stockant ces matrices sous un format HODLR ou HSS, les auteurs développent des algorithmes de résolution de systèmes linéaires dont les coefficients forment une matrice hiérarchisée avec une complexité de $\mathcal{O}(n \log^2 n)$ opérations élémentaires. Notons ici que, à l'aide d'un changement de base donné explicitement par la transformée de Fourier discrète, on peut se ramener d'une matrice de Toeplitz à une matrice dit Cauchy-like, qui est une matrice hiérarchisée. Par conséquent, en se basant sur [70], on peut résoudre des systèmes de Toeplitz en complexité $\mathcal{O}(n \log^2 n)$ par un algorithme déterministe.

Dans cette thèse, nous ne nous intéressons pas au calcul du produit matrice-vecteur $f(A)b$ avec $b \in \mathbb{C}^n$ pour lequel les techniques des sous-espaces de Krylov comme la méthode d'Arnoldi ou encore la méthode de Lanczos dans le cas d'une matrice symétrique, ayant comme but de se ramener au calcul d'une fonction de matrice pour une matrice de dimension petite devant n (voire [53] ou GMRES [80] ou d'autres méthodes [4]).

Le point de départ de nos recherches est l'article *Fast computation of the matrix exponential for a Toeplitz matrix* de D. Kressner et R. Luce [65] dans lequel les auteurs s'intéressent au calcul de la fonction de matrice $\exp(T)$ avec $T \in \mathbb{C}^{n\times n}$ une matrice de Toeplitz. Ils y fournissent des algorithmes rapides basées sur une arithmétique issue de la structure de déplacement, découverte par [51, 52] puis reprise dans une multitude de travaux de Kailath et autres auteurs [60, 24, 62, 77], pour laquelle les différentes opérations sur les matrices de Toeplitz peuvent être effectuées avec un coût de calcul bien inférieur à l'arithmétique pleine. En particulier toute fonction rationnelle exprimée en une matrice de Toeplitz $T \in \mathbb{C}^{n\times n}$ peut être calculée avec une complexité de $\mathcal{O}(n^2)$ opérations élémentaires. A partir de l'approximation rationnelle de l'exponentielle, ils en déduisent l'existence d'une approximation rationnelle $R(T)$ avec $R \in \mathcal{R}_{s,s} = \{r = \frac{p}{q} \text{ avec } p \in \mathcal{P}_s, q \in \mathcal{P}_s\}$ de l'exponentielle de matrice de Toeplitz, facilement calculable sur machine, de sorte que pour toute matrice de Toeplitz T définie négative.

$$\| \exp(T) - R(T)\|_2 \leqslant CV^{-s}$$

avec C une constante, $V \approx 9,28903...$ et $s \in \mathbb{N}$ tel que $R \in \mathcal{R}_{s,s}$ en appliquant le résultat de Gonchar et Rakhmanov [46]. Les auteurs considèrent alors pour fonction rationnelle R un approximant de Padé sans rechercher l'expression d'une fonction rationnelle R optimisant l'approximation, pour laquelle nous connaissons des résultats d'existence et d'unicité d'après [72].

Le but de cette thèse est pour une matrice de Toeplitz A symétrique définie positive avec intervalle spectral $[c;d]$ et une fonction f analytique sur un ouvert contenant l'intervalle $[c;d]$ de construire des approximants rationnels r de sorte à contrôler l'erreur absolue $\|f(A)-r(A)\|$ ou relative $\|I-f(A)^{-1}r(A)\|$ mais également de fournir des techniques efficaces et fiables pour évaluer $r(A)$. De plus, on sait d'après la théorie des ensemble K-spectraux que pour tout ensemble $\mathbb{E}$ contenant le spectre de A tel que f est analytique sur un voisinage de $\mathbb{E}$,

$$\|f(A) - r(A)\| \leqslant \|f - r\|_{L^\infty(\mathbb{E})} \text{ ou } \|I - f(A)^{-1}r(A)\| \leqslant \|1 - r/f\|_{L^\infty(\mathbb{E})},$$

et on pourra alors rechercher un meilleur approximant rationnel de f sur $\mathbb{E}$ pour borner l'erreur relative ou absolue pour les matrices. Nous allons en particulier considérer des interpolants afin de pouvoir sélectionner nos points d'interpolation en vue de diminuer notre borne supérieure. Notons que par le théorème de Chebyshev rationnel [87], un meilleur approximant est nécessairement un interpolant et on ne perd donc pas la généralité en se limitant aux interpolants. Pour pouvoir faire le lien avec la littérature sur l'approximation rationnelle, nous allons nous limiter à une classe de fonctions f dites de Markov, ainsi que des produits de fonctions de Markov avec une fonction rationnelle, incluant les fonctions $z \mapsto 1/\sqrt{z}$, $z \mapsto \sqrt{z}$, $z \mapsto \log(z)$ et d'autres fonctions élémentaires.

Dans le premier chapitre de nature introductive et sans résultats originaux, nous revisitons en section 1.1 les matrices de Toeplitz, et en section 1.3 les définitions et quelques propriétés des fonctions de matrices. Dans cette section, nous donnons également un aperçu de quelques méthodes de calcul des fonctions de matrices, et en particulier la méthode de Schur-Parlett mentionnée un peu plus tôt dans l'introduction. Finalement, nous énumérons quelques applications où l'occurence des fonctions de matrices de Toeplitz est naturelle et nous terminons, comme pour les autres chapitres, par quelques conclusions.

Au chapitre 2, nous reprenons l'idée de [65] de décrire une nouvelle arithmétique pour laquelle les opérations usuelles telles que la somme, le produit ou l'inversion sur des matrices de Toeplitz peuvent être effectuées avec une complexité réduite. Une itération de ces opérations nous permettra alors d'évaluer d'une manière efficace l'expression $r(A)$ pour A une matrice de Toeplitz et $r \in \mathcal{R}_{m,n}$. Suivant [61, 41], l'opérateur de déplacement d'une matrice carrée A est l'expression

$$S(A) = Z_1 A - A Z_{-1}, \text{ avec } Z_{\pm 1} = \begin{bmatrix} 0 & \dots & \dots & 0 & \pm 1 \\ 1 & \ddots & & & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & 0 \end{bmatrix}$$

introduit en section 2.1. A est dite Toeplitz-like si le rang de $S(A)$ (appelé rang de déplacement de A et noté $\rho(A)$) est faible devant la dimension. Par exemple, une matrice de Toeplitz est de rang de déplacement $\leqslant 2$ et de même pour son inverse même si l'expression de A^{-1} n'a à première vue pas de structure de Toeplitz apparente. On appelle générateurs de A les deux matrices apparaissant dans une décomposition de rang plein de $S(A)$. Il est bien connu est assez facilement vérifiable que l'on puisse reconstruire de manière unique la matrice A à partir de ses générateurs. Passer d'une matrice à ses générateurs est alors une sorte de compression des données en complexité $\mathcal{O}(\rho(A)n)$ (les auteurs de [65] ont travaillé avec un opérateur non pas de type Sylvester mais de type Stein, qui n'est pas injectif), ce passage est expliqué dans la section 2.2. Inspiré d'un raisonnement similaire mais pas identique à [65], nous montrons en section 2.3 que toute opération élémentaire entre deux matrices Toeplitz-like (addition, produit et inverse) donne à nouveau une matrice

Toeplitz-like. Grâce à ces formules, chacune de ces trois opérations élémentaires peut alors être effectuée en prenant comme entrée les générateurs des opérants, et en calculant les générateurs du résultat, en complexité $\mathcal{O}(n \log^2 n)$. Une compression supplémentaire basée sur une SVD de $S(A)$ permet de maîtriser la croissance du rang de déplacement. Par exemple, le rang de déplacement de $r(A)$ pour une fonction rationnelle $r \in \mathcal{R}_{m,\ell}$ peut être borné par $\mathcal{O}(\max\{m,\ell\}\rho(A))$ et donc $r(A)$ peut être évalué en complexité $\mathcal{O}(\max\{m,\ell\} n \log^2 n)$ pour une matrice de Toeplitz. Dans le cadre de cette thèse, l'ensemble des procédures de notre arithmétique Toeplitz-like a été implémentée sous Matlab. Pour faciliter la reproductibilité de nos expériences numériques, nous avons finalement décidé de nous baser dans toutes nos expériences numériques sur le récent paquetage *TLComp*[1] de Robert Luce qui suit les mêmes idées mathématiques.

Dans le chapitre 3, nous entamons la recherche d'une fonction rationnelle de matrice $r(A)$ en adaptant la méthode itérative de Newton pour la racine carrée principale de matrice $\sqrt{A}$ ainsi que le signe de matrice $\text{sign}(A)$ [57, Section 6.3] à notre nouvelle arithmétique. Chaque matrice itérée de la méthode de Newton pour la racine carrée principale étant une fonction rationnelle en la matrice A d'ordre $[2^k|2^k-1]$ avec $k \geqslant 0$, d'après notre proposition 3.2.5, elle peut être calculée, lorsque $k << n$, en $\mathcal{O}(n \log^2 n)$ opérations élémentaires. L'auteur considère dans [57, Section 6.3] une itération avec un premier terme égal à la matrice A, ce qui malgré la convergence quadratique de la méthode pourrait entraîner une convergence lente vers la fonction de matrice concernée. Dans le cas de la racine carrée principale, nous introduisons en sous-section 3.3.1 d'autres choix d'un premier terme sous forme d'un multiple scalaire ou d'approximant de Padé d'ordre concordant avec l'ordre des matrices itérées de la méthode de Newton soit d'ordre $[2^\ell|2^\ell-1]$ permettant une bonne approximation dès le premier terme, ainsi que des paramètres optimaux [9] introduits en sous-section 3.3.2 à chaque itération de cette méthode, accélérant la convergence. Notons ici que notre but n'est pas d'utiliser la méthode de Newton avec ou sans paramètres ou même le choix du premier terme, mais plutôt de comprendre et comparer l'arithmétique Toeplitz-like et l'arithmétique pleine. Pour ce faire, nous effectuons plusieurs expérience numériques en section 3.4 pour lesquelles les calculs en arithmétique Toeplitz-like sont plus rapides pour n assez grand, convergeant en particulier vers un bon approximant rationnel d'ordre $[2^m|2^m-1]$ avec $m = \ell + k$ de la fonction de matrice $\sqrt{T}$ pour $T \in \mathbb{R}^{n\times n}$ matrice de Toeplitz symétrique définie positive. Cet approximant est calculable en notre nouvelle arithmétique avec une complexité d'ordre $\mathcal{O}(\rho(T)^2 n \log^2 n)$ opérations élémentaires, mais est moins précise qu'en arithmétique pleine. La section 3.5 est ensuite dédiée à l'étude de la méthode de Newton pour la fonction de matrice signe, dont une étude plus générale peut être retrouvé dans [37, 4] pour l'approximation de la fonction de matrice $\text{sign}(A)$ ou le produit $\text{sign}(A)x$ pour tout $x \in \mathbb{C}^n$. Comme pour le cas de la racine, le but ici n'est pas d'utiliser la méthode de Newton en elle-même mais encore une fois de comprendre et de comparer les arithmétiques Toeplitz-like et pleine au travers de plusieurs expériences numériques, données en sous-section 3.5.3.

Au chapitre 4, nous nous intéressons au cas particulier d'une fonction de Markov $f^{[\mu]}(z) = \int_\alpha^\beta \frac{d\mu(x)}{z-x}$ avec μ mesure positive à support dans l'intervalle $[\alpha;\beta]$, que l'on souhaite approcher sur un ensemble $\mathbb{E} \subset \mathbb{R}\backslash[\alpha;\beta]$ par un interpolant $r^{[\mu]}$ de $f^{[\mu]}$ de type $[m-1|m]$, en prenant par exemple $\mathbb{E} = [c;d]$, l'intervalle spectral ou encore le spectre de notre matrice de Toeplitz A symétrique. Nous énonçons dans la section 4.1 une partie importante de nos résultats originaux et donnons dans notre théorème 4.1.4 une nouvelle écriture pour une borne supérieure de l'erreur $\|1 - r^{[\mu]}/f^{[\mu]}\|_{L^\infty(\mathbb{E})}$ en terme de produits de Blaschke pour des points d'interpolation assez quelconques. Cette écriture nous permet de conclure que, quelque soit le choix des points d'interpolation, il existe une pire mesure ν (la mesure d'équilibre renormalisée sur l'intervalle $[\alpha;\beta]$) maximisant (à un facteur 3 près) l'erreur relative d'interpolation, résultat que nous n'avons pas vu avant dans la littérature. Nous considérons ensuite différents choix optimaux des points d'interpolation, notre nouvelle borne nous donnant en corollaire 4.1.8 une borne dite a priori en termes de la capacité ϱ du condensateur formé par des plateaux $[\alpha;\beta]$ et $[c;d]$,

$$\Big\|1 - \frac{r_m^{[\mu]}}{f^{[\mu]}}\Big\|_{L^\infty([c;d])} \leqslant 8\varrho^{2m}/(1-2\varrho^{2m})^2, \quad \varrho = \exp\Big(\frac{-1}{cap([\alpha;\beta],[c;d])}\Big),$$

1. paquetage disponible a l'adresse https ://github.com/rluce/tlcomp

la quantité ϱ étant une valeur de référence dans la théorie de meilleure approximation rationnelle des fonctions de Markov. On peut notamment citer [45, Theorem 1] qui décrit le comportement asymptotique de l'erreur absolue par un meilleur approximant sous la forme

$$\lim_{m\to\infty} \min_{r\in\mathcal{R}_{m-1,m}} \|1 - r^{[\mu]}/f^{[\mu]}\|^{1/m}_{L^\infty([c;d])} = \exp\left(-\frac{2}{cap([\alpha;\beta],[c;d])}\right),$$

ou encore T. Ganelius [39], qui exprime également le comportement asymptotique de l'erreur d'approximation absolue par un interpolant rationnel avec points d'interpolation double, ainsi que D. Braess [19] qui borne l'erreur d'approximation sur le disque

$$\|f^{[\mu]} - r^{[\mu]}_m\|_{L^\infty(\overline{\mathbb{D}})} \leqslant Cst.\varrho^{-2m}$$

avec Cst une constante. Enfin, citons H. Stahl et V. Totik [81] qui généralisent le résultat de [45] en exprimant une borne supérieure au comportement asymptotique de l'erreur d'approximation sur tout ensemble compact $\mathbb{E} \subseteq \overline{\mathbb{C}}\setminus[\alpha;\beta]$ symétrique par rapport à l'axe réel par un meilleur approximant de $f^{[\mu]}$ en terme de la capacité logarithmique du condenseur formé par $supp(\mu)$ et l'ensemble $\mathbb{E}$

$$\lim_{m\to\infty} \|f^{[\mu]} - r^{[\mu]}_m\|^{1/2m}_{L^\infty(\mathbb{E})} = e^{-1/cap(\mathbb{E},supp(\mu))}.$$

En exprimant notre théorème 4.1.4 en termes d'une matrice symétrique A, nous obtenons dans le corollaire 4.3.2 une estimation d'erreur résiduelle

$$\|I - r^{[\mu]}_m(A)\Big(f^{[\mu]}(A)\Big)^{-1}\| \leqslant \|I - r^{[\nu]}_m(A)^2\frac{1}{|\alpha|}(A-\alpha I)(A-\beta I)\|,$$

où on observe que la borne dépend uniquement de l'expression calculable $r^{[\nu]}_m$ de la pire mesure ν ce qui en fait donc un second membre que l'on peut calculer, et qui dans nos expériences est du même ordre de grandeur que la borne a priori, au moins en arithmétique exacte. Nous proposons dans notre remarque 4.3.3 une nouvelle condition d'arrêt pour le choix du degré m basée sur l'idée que l'erreur en précision finie l'emporte si la borne a priori s'écarte de la borne résiduelle. Cette idée à priori très simple s'avère concluante dans toutes nos expériences numériques.

Finalement, nous donnons dans la deuxième partie du corollaire 4.3.2 une troisième borne dit a posteriori basée sur l'idée que l'erreur relative entre $r^{[\mu]}_m$ et $f^{[\mu]}$ devrait être du même ordre de grandeur que l'erreur relative entre $r^{[\mu]}_m$ et $r^{[\mu]}_{m+m'}$ lorsque m est grand. Cette idée de borner l'erreur est utilisée avec succès dans les méthodes de Krylov [11], bien qu'elle soit souvent heuristique. Dans notre contexte, nous proposons une preuve de cette estimation a posteriori dans le corollaire 4.3.2. Dans la section 4.2, nous envisageons trois représentations pour notre interpolant $r^{[\mu]}_m$ résultant de trois approches différentes du calcul de l'interpolant, avec l'espoir qu'une des trois s'avère particulièrement stable pour évaluer $r^{[\mu]}_m$, dans un premier temps pour un argument scalaire. Pour la représentation sous forme de fraction continue de Thiele positive, nous donnons dans le théorème 4.2.2 un nouveau résultat sur le lien entre positivité des paramètres et le fait que l'on interpole une fonction de Markov, résultat bien connu dans le cas d'une fraction continue de Stieltjes. Ce résultat de positivité nous permet au théorème 4.2.4 d'affiner un résultat de stabilité backward des interpolants représentés sous forme d'une fraction continue de Thiele [48], en éliminant un facteur qui potentiellement croît exponentiellement. Finalement dans la section 4.3 nous présentons de multiples expériences numériques montrant que la stabilité numérique de notre approche par arithmétique Toeplitz-like dépend fortement du conditionnement de notre matrice de Toeplitz symétrique A. Si la matrice est bien conditionnée alors la façon d'évaluer $r^{[\mu]}_m(A)$ ne joue pas un rôle primordial. Pour un conditionnement moyen, seule une représentation sous forme d'éléments simples de nos interpolants donne des résultats probants, surtout si on combine avec des techniques avec des méthodes de scaling and squaring qui eux-même font appel à la méthode de Newton étudiée au chapitre 3. Par contre, si A est très mal conditionnée, l'ensemble de nos méthodes pour approcher $f^{[\mu]}(A)$ à une précision correcte échouent.

Chapitre 1

Structure de Toeplitz et fonctions de matrices

Calculer une fonction de matrice de Toeplitz $f(A)$ pour $A \in \mathbb{C}^{n\times n}$ et f une fonction apparaît nécessaire dans plusieurs applications mathématiques comme les systèmes de préfiltres [3, Section 7.2.1.2], les résolutions d'EDP obtenues par semi-discrétisation des equations intégrales [67, example 3], [73] ou encore des problèmes aux valeurs propres [57, Section 2.10]. Cependant les méthodes classiques de calcul de fonctions de matrices ne prenant pas compte de la structure de la matrice A, celles-ci vont généralement nécessiter un coût de l'ordre de n^3 opérations élémentaires. Dans le cas de matrices structurées comme les matrices à structure hiérarchique, il est apparu que les propriétés particulières de celles-ci pouvaient être exploitées pour apporter des méthodes de calcul numériques à coût réduit. De cette observation, nous cherchons donc si pour le cas des matrices de Toeplitz, cette structure particulière pourrait présenter certains avantages numériques à développer dans de nouveaux algorithmes. Dans ce chapitre, pour poser les bases de notre étude, nous nous intéressons dans une première section à la structure Toeplitz en rappelant sa définition ainsi qu'un premier exemple d'application de cette structure au calcul d'un produit matrice-vecteur avec un coût numérique réduit par rapport au calcul classique. Dans une deuxième section nous passons en revue les différentes définitions équivalentes de fonctions de matrices ainsi que quelques propriétés de celles-ci déductibles de ces définitions. Puis nous rappelons en section 3 quelques méthodes d'implémentation directe des fonctions de matrices, dans un premier temps dans le cas de fonctions polynomiales, puis la méthode classique de Schur-Parlett, valable pour un spectre plus large de fonctions de matrices. Motivés par le fait qu'une fonction polynomiale ou rationnelle est plus facilement implémentable, nous énonçons ensuite en sous-section 1.3.3 quelques méthodes d'approximation polynomiale et rationnelle des fonctions de matrices et leur différentes formes en sous-section 1.3.4. Enfin, en dernière section, nous développons plus explicitement quelques exemples motivant la recherche pour le calcul d'une fonction de matrice de Toeplitz.

1.1 La structure Toeplitz

En 1911, Otto Toeplitz introduit les matrices de Toeplitz après l'étude de formes quadratiques $\sum \varphi_{i,j} x_i y_j$ pour lesquelles les coefficients de ces formes vérifient la propriété particulière $\varphi_{i,j} = \varphi_{i-j}$ et ainsi obtient des matrices associées $T = (\varphi_{i,j})_{i,j}$ à diagonales et sur/sous-diagonales constantes que l'on appelle à présent matrices de Toeplitz [85],[86]. Plus récemment, la structure particulière à diagonales et sur/sous-diagonales constantes des matrices de Toeplitz apparaît dans divers problèmes suite aux différentes méthodes de réso-

lution employées ou par les méthodes de prélèvement de données à intervalle de temps ou d'espace régulier qui vont faire apparaître des systèmes avec matrices à diagonales et sur/sous-diagonales constantes.

Définition 1.1.1. *On appelle matrice de Toeplitz, toute matrice $T \in \mathbb{C}^{n\times n}$ à diagonales et sur/sous-diagonales constantes, c'est-à-dire toute matrice de la forme*

$$T = \begin{bmatrix} t_0 & t_{-1} & \dots & t_{-n+1} \\ t_1 & t_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & t_{-1} \\ t_{n-1} & \dots & t_1 & t_0 \end{bmatrix}, \quad t_i \in \mathbb{C} \ \ \forall i = -n+1, \dots, n-1.$$

Dans le cas particulier où $t_{n-j} = t_{-j}$ pour $j = 0, \dots, n$, la matrice est dite circulante et si $t_{n-j} = -t_{-j}$ pour $j = 0, \dots, n$, alors on parle de matrice anti-circulante.

Le cas particulier des matrices circulantes et anti-circulantes est intéressant dans le cadre de la diagonalisation. En effet, ces matrices à structure particulière peuvent être diagonalisées à l'aide de la FFT (Fast Fourier Transform) avec une complexité inférieure par rapport à toute matrice sans structure particulière. Pour ce faire, il nous faut employer une matrice à la structure tout aussi particulière : la DFT.

Définition 1.1.2. *Soit $n \in \mathbb{N}$. On appelle DFT (Discrete Fourier Transform) la matrice F_n de taille $n \times n$ où $(F_n)_{j,k} = \frac{1}{\sqrt{n}} e^{-2i\pi(j-1)(k-1)/n}$.*

Pour $n \in \mathbb{N}$, on sait que la matrice F_n est une matrice unitaire, orthogonale et que son produit avec un vecteur quelconque $x \in \mathbb{C}^{n\times 1}$ peut être effectué en $\mathcal{O}(n \log n)$ opérations élémentaires par l'algorithme FFT . Nous renvoyons à [25] pour cet algorithme. Dans le cas d'une matrice circulante ou anti-circulante, c'est-à-dire pour une matrice de la forme

$$C_n = C_n(c_0, \dots, c_{n-1}) = \begin{bmatrix} c_0 & c_1 & \dots & c_{n-1} \\ c_{n-1} & c_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & c_1 \\ c_1 & \dots & c_{n-1} & c_0 \end{bmatrix} \in \mathbb{C}^{n\times n}$$

ou

$$\tilde{C}_n = \tilde{C}_n(c_0, \dots, c_{n-1}) = \begin{bmatrix} c_0 & -c_1 & \dots & -c_{n-1} \\ c_{n-1} & c_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & -c_1 \\ c_1 & \dots & c_{n-1} & c_0 \end{bmatrix} \in \mathbb{C}^{n\times n}$$

respectivement, celles-ci se diagonalisent à l'aide de la DFT de la manière suivante :

Proposition 1.1.3. *Notons $\omega := e^{-i\pi/n}$.*

i. Soit $C_n = C_n(c_0, \dots, c_{n-1}) \in \mathbb{C}^{n\times n}$ une matrice circulante. Alors $C_n = F_n \Delta F_n^*$ où $\Delta = diag(p(\omega^{2(l-1)}))_{l=1,..,n}$ avec $p(x) = c_0 + c_1 x + \dots + c_{n-1}x^{n-1} \in \mathbb{C}_{n-1}[x]$.

ii. Soit $\tilde{C}_n = \tilde{C}_n(c_0, \dots, c_{n-1}) \in \mathbb{C}^{n\times n}$ une matrice anti-circulante. Alors $\tilde{C}_n = \widehat{F}_n \widehat{\Delta} \widehat{F}_n^*$ où $\widehat{\Delta} = diag(p(\omega^{2l-1}))_{l=1,..,n}$ avec $p(x) = c_0 - c_1 x - \dots - c_{n-1}x^{n-1} \in C_{n-1}[x]$ et $\widehat{F}_n = D \cdot F_n$ avec $D = diag(\omega^{(l-1)})_{l=1,..,n}$.

Démonstration.

i. Soit $C_n = C_n(c_0, \dots, c_{n-1}) \in \mathbb{C}^{n\times n}$ une matrice circulante. Posons $p(x) = c_0 + c_1 x + ... + c_{n-1}x^{n-1}$. Alors on a $\forall k = 1, .., n$,

$$C_n \cdot F_n \cdot e_k = \frac{1}{\sqrt{n}} C_n \cdot (e^{\frac{-2i\pi(j-1)(k-1)}{n}})_{j=1,\dots,n} = \frac{1}{\sqrt{n}} \begin{bmatrix} c_0\omega^0 + c_1\omega^{2(k-1)} + ... + c_{n-1}\omega^{2(n-1)(k-1)} \\ c_{n-1}\omega^0 + c_0\omega^{2(k-1)} + ... + c_{n-2}\omega^{2(n-1)(k-1)} \\ \vdots \\ c_1\omega^0 + c_2\omega^{2(k-1)} + ... + c_0\omega^{2(n-1)(k-1)} \end{bmatrix}$$

$$= \frac{1}{\sqrt{n}} \begin{bmatrix} p(\omega^{2(k-1)}) \\ \omega^{2(k-1)} p(\omega^{2(k-1)}) \\ \vdots \\ \omega^{2(n-1)(k-1)} p(\omega^{2(k-1)}) \end{bmatrix} = F_n \cdot e_k \cdot p(\omega^{2(k-1)}).$$

D'où $C_n \cdot F_n = F_n \cdot diag(p(\omega^{2(k-1)}))_{k=1,..,n}$ et par multiplication à droite par F_n^* orthogonale à F_n, on obtient la diagonalisation $C_n = F_n \cdot \mathrm{diag}(p(\omega^{2(k-1)}))_{k=1,..,n} \cdot F_n^*$.

ii. Soit $\tilde{C_n} = \tilde{C_n}(c_0, \dots, c_{n-1}) \in \mathbb{C}^{n\times n}$ une matrice anti-circulante. Posons $p(x) = c_0 - c_1 x - \dots - c_{n-1}x^{n-1}$ et soit $\tilde{F_n} = diag(\omega^0, \omega^1, .., \omega^{n-1}) \cdot F_n$, ce qui nous donne $\tilde{F_n} = \frac{1}{\sqrt{n}} \left(e^{-i\pi(j-1)(2k-1)/n}\right)_{j,k=1,\dots,n}$, d'où $\forall k = 1, .., n,,$

$$\begin{aligned}
\tilde{C_n} \cdot \tilde{F_n} \cdot e_k &= \frac{1}{\sqrt{n}} \tilde{C_n} \cdot \left(e^{-i\pi(j-1)(2k-1)/n}\right)_{j=1,\dots,n} \\
&= \begin{bmatrix} c_0 - c_1\omega^{2k-1} - c_2\omega^{2(2k-1)} - \dots - c_{n-1}\omega^{(n-1)(2k-1)} \\ c_{n-1} + c_0\omega^{2k-1} - c_1\omega^{2(2k-1)} - \dots - c_{n-2}\omega^{(n-1)(2k-1)} \\ \vdots \\ c_1 + c_2\omega^{2k-1} + c_3\omega^{2(2k-1)} + \dots + c_0\omega^{(n-1)(2k-1)} \end{bmatrix} \\
&= \begin{bmatrix} p(\omega^{2k-1}) \\ \omega^{2k-1}p(\omega^{2k-1}) + c_{n-1} + c_{n-1}\omega^{(2k-1)(n-1)}\omega^{2k-1} \\ \vdots \\ \omega^{(2k-1)(n-1)}p(\omega^{2k-1}) + \sum_{l=1}^{n-1} c_l\omega^{(2k-1)(l-1)} + \sum_{l=1}^{n-1} c_l\omega^{(2k-1)(n-1+l)} \end{bmatrix} \\
&= \begin{bmatrix} \omega^0 \\ \omega^{2k-1} \\ \vdots \\ \omega^{(2k-1)(n-1)} \end{bmatrix} p(\omega^{2k-1}) \\
&= \tilde{F_n} \cdot e_k \cdot p(\omega^{2k-1})
\end{aligned}$$

D'où $\tilde{C_n} \cdot \tilde{F_n} = \tilde{F_n} \cdot e_k \cdot p(\omega^{2k-1})$ soit $\tilde{C_n} = \tilde{F_n} \cdot \tilde{\Delta} \cdot \tilde{F_n}^*$.

□

Supposons à présent que l'on souhaite effectuer un calcul matrice-vecteur $T \cdot x$ avec T une matrice de Toeplitz. Sans prendre en compte la structure particulière de la matrice, ce produit matrice-vecteur en dimension n serait effectué en $\mathcal{O}(n^2)$ opérations élémentaires. Or, on peut montrer à l'aide de la proposition précédente que pour toute matrice de Toeplitz $T \in \mathbb{C}^{n\times n}$ et pour tout vecteur x, le produit matrice-vecteur $T \cdot x$ peut être effectué avec un coût réduit par utilisation de la FFT [63, Section 5.3.3].

Lemme 1.1.4 (produit matrice de Toeplitz vecteur)**.** *Pour toute matrice de Toeplitz $T \in \mathbb{C}^{n\times n}$ et pour tout vecteur $x \in \mathbb{C}^n$, le produit matrice-vecteur $T \cdot x$ s'effectue avec une complexité $\mathcal{O}(n \log n)$.*

Démonstration. Notons $T \in \mathbb{C}^{n\times n}$ la matrice de Toeplitz donnée par

$$T = \begin{bmatrix} t_0 & t_{-1} & \dots & t_{-n+1} \\ t_1 & t_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & t_{-1} \\ t_{n-1} & \dots & t_1 & t_0 \end{bmatrix}, \quad \text{avec } t_j \in \mathbb{C},\ \forall j = -n+1, \dots, n-1.$$

On commence par intégrer la matrice T dans une matrice circulante de taille $2n \times 2n$, de sorte à obtenir le système linéaire matriciel :

$$C_{2n} \cdot y = \begin{bmatrix} T & B \\ B & T \end{bmatrix} \cdot \begin{bmatrix} x \\ 0 \end{bmatrix} = \begin{bmatrix} T \cdot x \\ * \end{bmatrix}, \text{ avec } B = \begin{bmatrix} t_0 & t_{n-1} & \dots & t_2 & t_1 \\ t_{-n+1} & t_0 & \ddots & & t_2 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ t_{-2} & & \ddots & \ddots & t_{n-1} \\ t_{-1} & t_{-2} & \dots & t_{-n+1} & t_0 \end{bmatrix}.$$

On effectue alors la FFT sur la matrice C_{2n}, ce qui nous donne d'après la proposition 1.1.3 *i.* la décomposition

$$C_{2n} \cdot y = F_{2n} \cdot \Delta \cdot F_{2n}^* \cdot y.$$

avec Δ, F_{2n} définis comme en proposition 1.1.3. Il ne nous reste alors plus qu'à identifier

$$C_{2n} \cdot y = \begin{bmatrix} T \cdot x \\ * \end{bmatrix} \Leftrightarrow F_{2n} \cdot \Delta \cdot F_{2n}^* \cdot y = \begin{bmatrix} T \cdot x \\ * \end{bmatrix}.$$

Or le produit de la matrice F_{2n} avec un vecteur $y \in \mathbb{C}^{2n}$ revenant à la DFT (Discrete Fourier Transform) du vecteur y, ce produit est effectué en $\mathcal{O}(n \log n)$ opérations, et on peut alors se servir de la décomposition précédente pour réduire le coût de calcul du produit matrice-vecteur $T \cdot x$ à l'aide du schéma de calcul suivant :

i. $F_{2n}^* \cdot y$ en $\mathcal{O}(n \log n)$ opérations ;
ii. $\Delta \cdot (F_{2n}^* \cdot y)$ en $\mathcal{O}(n)$ opérations ;
iii. $F_{2n} \cdot [\Delta \cdot (F_{2n}^* \cdot y)]$ en $\mathcal{O}(n \log n)$ opérations ;

d'où le calcul du produit $C_{2n} \cdot y$ s'effectue avec une complexité $\mathcal{O}(n \log n)$. □

Remarque 1.1.5. *De part la structure à diagonales et sur/sous-diagonales constantes des matrices de Toeplitz, on peut remarquer que la sous-classe des matrices de Toeplitz dans $\mathbb{C}^{n\times n}$ est un espace vectoriel puisque la somme de matrices de Toeplitz est une matrice de Toeplitz, calculable, de part la structure Toeplitz, avec une complexité de $\mathcal{O}(n)$, et la multiplication par un scalaire (calculable avec une complexité de $\mathcal{O}(n)$) est encore une matrice de Toeplitz. Cependant, l'inverse d'une matrice de Toeplitz ou le produit de deux d'entre elles n'est pas de Toeplitz. En effet, dans $\mathbb{C}^{3\times 3}$, on a par exemple*

$$\textit{si } T_1 = \begin{bmatrix} 3 & 5 & 1 \\ -2 & 3 & 5 \\ 1 & -2 & 3 \end{bmatrix} \textit{ et } T_2 = \begin{bmatrix} -4 & 1 & 3 \\ 0 & -4 & 1 \\ -1 & 0 & -4 \end{bmatrix}, \textit{ alors } T_1 \times T_2 = \begin{bmatrix} -13 & -17 & 10 \\ 3 & -14 & -23 \\ -7 & 9 & -11 \end{bmatrix}$$

qui n'est pas une matrice de Toeplitz.

1.2 Définitions et propriétés des fonctions de matrices

Le premier à étudier le domaine des fonctions de matrices fut Cayley dans *A Memoir on the Theory of Matrices* paru en 1858 ([21]), dans lequel il étudie la fonction de matrice racine carrée. Puis les ingénieurs aérospatiaux Frazer, Duncan, et Collar ont montré dans *Applications to Dynamics and Differential Equations* ([35]) l'importance et l'utilité de la fonction exponentielle de matrice. Nous renvoyons le lecteur à [15] pour d'autres exemples concrets de l'application des fonctions de matrices.
Si les fonctions sur des matrices telles que le déterminant ou la trace sont déjà bien connues, ces fonctions sont à valeurs dans le corps de base. Or ici nous nous intéressons à des fonctions à valeurs dans l'espace des matrices, soit des fonctions $f : \mathbb{K}^{n\times n} \to \mathbb{K}^{n\times n}$ pour tout $n \geqslant 1$.

Une première idée pourrait être de définir une fonction de matrice en appliquant la fonction sur chaque coefficient (comme le fait le langage de programmation *Fortran 95*). Par exemple, si $f(x) = \sin(x)$ et $A = (a_{i,j})_{i,j} \in \mathbb{C}^{n\times n}$ avec $n > 1$, on est alors tenté de définir la fonction de matrice $f(A)$ par $f(A) = (\sin(a_{i,j}))_{i,j}$. Cependant cette définition ne s'accorde pas bien avec l'algèbre des matrices puisque si on prend le cas de la fonction carrée, il est facilement démontrable que le carré d'une matrice n'est pas toujours égale à la matrice de ses coefficients élevés au carré, c'est-à-dire que $A^2 \neq (a_{i,j}^2)_{i,j}$, en prenant par exemple la matrice $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$. Par conséquent une telle définition n'est pas envisageable. En revanche, lorsque f est un polynôme $f(x) = \sum_{j=0}^{k} c_j x^j$, on peut définir $f(A)$ en substituant la variable x dans l'expression de f par la matrice A et le coefficient 1 par la matrice identité, ce qui nous donne alors la fonction de matrice $f(A) = \sum_{j=0}^{k} c_j A^j$. Dans le cas des fonctions rationnelles, si on prend par exemple une matrice $A \in \mathbb{C}^{n\times n}$ et la fonction $f(t) = \frac{1+t^2}{1-t}$, alors en supposant que la matrice $I - A$ soit inversible (par exemple si $1 \notin \sigma(A)$), on peut identifier $f(A) = (I - A)^{-1}(I + A^2)$.
Ainsi, une large partie des fonctions de matrices est déjà définie. Mais toute fonction n'est pas un polynôme ou une fonction rationnelle, comme la fonction signe ou la racine carrée. De plus, comme dans notre exemple, la définition donnée ne sera pas valable pour toute matrice (si $1 \in \sigma(A)$, et on ne peut utiliser la définition que nous avons donné pour $f(t) = \frac{1+t^2}{1-t}$). Dans cette section, nous rappelons donc la définition générale d'une fonction de matrice ainsi que ses différentes formes équivalentes.

1.2.1 La forme canonique de Jordan

D'après l'algèbre linéaire, nous savons que toute matrice $A \in \mathbb{C}^{n\times n}$ peut s'exprimer sous sa forme canonique de Jordan

$$Z^{-1}AZ = J := diag(J_1, .., J_p) \tag{1.1}$$

avec

$$J_k = J_k(\lambda_k) = \begin{bmatrix} \lambda_k & 1 & 0 & \dots & 0 \\ 0 & \lambda_k & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \lambda_k & 1 \\ 0 & \dots & \dots & 0 & \lambda_k \end{bmatrix} \in \mathbb{C}^{m_k\times m_k},$$

où Z est une matrice inversible et $m_1 + \ldots + m_p = n$, $\{\lambda_1, \ldots, \lambda_p\}$ sont les valeurs propres distinctes de A pour tout $k = 1, \ldots, p$. La matrice de Jordan J est unique à permutation près des blocs tandis que la matrice Z n'est pas unique, et les sous matrices $J_k(\lambda_k)$ sont appelés blocs de Jordan [57, Section 1.2.1].

Définition 1.2.1. *[57, Definition 1.1] Soit $A \in \mathbb{C}^{n\times n}$ avec valeurs propres distinctes $\lambda_1, \ldots, \lambda_p$. On dit qu'une fonction f est définie sur le spectre de A si les valeurs*

$$f^{(j)}(\lambda_i), \;\; j = 1, \ldots, n_i - 1, \; i = 1, \ldots, p$$

existe, avec n_i la taille du plus gros bloc de Jordan où apparaît la valeur propre λ_i dans la forme canonique de Jordan de A.

De cette décomposition découle une première définition d'une fonction de matrice :

Définition 1.2.2. *Soient $A \in \mathbb{C}^{n\times n}$ avec valeurs propres distinctes $\lambda_1, \ldots, \lambda_p$ et f une fonction définie sur le spectre de A et supposons que A admet la forme canonique de Jordan* (1.1)*. Alors la fonction de matrice $f(A)$ est donnée par*

$$f(A) = Z^{-1} f(J) Z = Z^{-1} diag(f(J_1), .., f(J_p)) Z,$$

où

$$f(J_k) = f(J_k(\lambda_k)) = \begin{bmatrix} \frac{f(\lambda_k)}{0!} & \frac{f^{'}(\lambda_k)}{1!} & \cdots & \frac{f^{(m_k-1)}(\lambda_k)}{(m_k-1)!} \\ 0 & \frac{f(\lambda_k)}{0!} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \frac{f^{'}(\lambda_k)}{1!} \\ 0 & \cdots & 0 & \frac{f(\lambda_k)}{0!} \end{bmatrix} \in \mathbb{C}^{m_k \times m_k} \tag{1.2}$$

et m_k est la taille du bloc de Jordan $J_k = J_k(\lambda_k)$.

Remarque 1.2.3. *Si $A \in \mathbb{C}^{n\times n}$ est diagonalisable sous la forme $A = Z^{-1}\,\mathrm{diag}(\lambda_1, \ldots, \lambda_n)Z$, alors $f(A) = Z^{-1} diag(f(\lambda_1), .., f(\lambda_n))Z$.*

Remarque 1.2.4. *La définition de $f(A)$ ne dépend pas du choix de la position des blocs. En effet, supposons que $A = ZJZ^{-1} = YJ^{'}Y^{-1} = YPJP^{-1}Y^{-1} = WJW^{-1}$ avec P matrice de permutation telle que $J^{'} = PJP^{-1}$. Alors en définissant $f_1(A) = Zf(J)Z^{-1}$ et $f_2(A) = Wf(J)W^{-1}$, on peut montrer que $f_1(A) = f_2(A)$. En effet, on a $W^{-1}Zf(J)Z^{-1}W = f(J)$ c'est-à-dire $X^{-1}f(J)X = f(J)$ où $X = Z^{-1}W$. D'où on a $X^{-1}JX = J \Longrightarrow f(J) = f(X^{-1}JX) = X^{-1}f(J)X \Longrightarrow f_1(A) = f_2(A)$.*

1.2.2 Définition par interpolation polynomiale

Une deuxième définition pour une fonction de matrice fut inspirée par le cas des polynômes de matrices. On sait que pour toute matrice $A \in \mathbb{C}^{n\times n}$, il existe un polynôme ψ appelé polynôme minimal de A qui est l'unique polynôme unitaire de degré minimal tel que $\psi(A) = 0$ et divisant tout autre polynôme p satisfaisant $p(A) = 0$. Ce polynôme minimal est alors donné par $\psi(t) = \sum_{i=1}^{s}(t - \lambda_i)^{n_i}$ où $\lambda_1, \ldots, \lambda_s$ sont les valeurs propres distinctes de la matrice A et n_i la dimension du plus grand block de Jordan associé à la valeur propre λ_i, d'où d'après la définition précédente, $\psi(A) = 0$. Cette formule implique alors que pour deux polynômes p et q, $p(A) = q(A)$ si et seulement si p et q prennent les mêmes valeurs sur le spectre de A [57, Theorem 1.3]. On peut alors en déduire que la matrice $p(A)$ est complètement déterminée par les valeurs du polynôme p en les valeurs propres de A et il apparaît alors naturel de vouloir étendre cette propriété à toute fonction f définie sur le spectre d'une matrice pour laquelle on pourrait caractériser $f(A)$ par ses valeurs sur $\sigma(A)$. Or, on sait que pour une ensemble de valeurs $\{x_1, \ldots, x_s\}$ et une fonction f, on peut trouver un polynôme interpolateur de la fonction f en les valeurs x_i :

Lemme 1.2.5 (polynômes d'Hermite)**.** *Soient $A \in \mathbb{C}^{n\times n}$ avec valeurs propres distinctes $\lambda_1, \dots, \lambda_p$, n_i la taille du plus gros bloc de Jordan associé à la valeur propre λ_i pour $i = 1, \dots, p$ et f une fonction définie sur le spectre de A. Alors il existe un unique polynôme $p_{f,A}$ de degré inférieur au degré du polynôme minimal de A tel que*

$$p_{f,A}^{(k)}(\lambda_i) = f^{(k)}(\lambda_i), \quad i = 1, .., p, \;\; k = 0, 1, .., n_i - 1. \tag{1.3}$$

Ce polynôme est appelé polynôme interpolateur d'Hermite (voir [82, Théorème 2.1.5.2]).

Démonstration.

— unicité : Supposons qu'il existe deux polynômes p_1, p_2 vérifiant la condition (1.3). Alors $Q(x) = p_1(x) - p_2(x)$ vérifie

$$Q^{(k)}(\lambda_i) = 0, \quad i = 1, .., p, \;\; k = 0, 1, .., n_i - 1,$$

d'où λ_i est un racine d'ordre au moins n_i de $Q(x)$ qui possède donc au moins $\sum_{i=1}^{p} n_i = \deg(\psi)$ racines comptées avec leur multiplicité, ce qui entraîne que $Q(x)$ est identiquement nulle puisque de degré $\leq \deg(\psi) - 1$.

— existence : c'est une conséquence de (1.3). En effet, de cette condition, on obtient un système de $d+1 = \deg(\psi)+1$ équations à $d = \deg(\psi)$ inconnues qui sont les coefficients c_j de $p(x) = c_0 + c_1 x + .. + c_{d-1}x^{d-1}$, c'est-à-dire que ces coefficients vérifient

$$\begin{bmatrix} 1 & \lambda_1 & \lambda_1^2 & \dots & \dots & \lambda_1^{d-1} \\ 0 & 1 & 2\lambda_1 & \dots & \dots & (d-1)\lambda_1^{d-2} \\ \vdots & \ddots & \ddots & & & \vdots \\ 0 & \dots & 0 & 1 & \dots & \frac{(d-1)!}{(d-n_1-1)!}\lambda_1^{d-n_1} \\ \vdots & & & & & \vdots \\ 1 & \lambda_p & \lambda_p^2 & \dots & \dots & \lambda_p^{d-1} \\ 0 & 1 & 2\lambda_p & \dots & \dots & (d-1)\lambda_p^{d-2} \\ \vdots & \ddots & \ddots & & & \vdots \\ 0 & \dots & 0 & 1 & \dots & \frac{(d-1)!}{(d-n_p-1)!}\lambda_p^{d-n_p} \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ c_{d-1} \end{bmatrix} = \begin{bmatrix} f_0^{(0)} \\ f_0^{(1)} \\ \vdots \\ f_0^{(n_0-1)} \\ \vdots \\ f_p^{(0)} \\ f_p^{(1)} \\ \vdots \\ f_p^{(n_p-1)} \end{bmatrix}.$$

Or, d'après l'étude de l'unicité, notre matrice est injective et comme elle est carrée ($d = \sum_{i=0}^{p} n_i$) par le théorème du rang, elle est surjective et donc bijective, d'où l'existence et unicité des coefficients c_j.

□

Le lemme 1.2.5 nous fournit alors une deuxième définition pour une fonction de matrice :

Corollaire 1.2.6. *Soient $A \in \mathbb{C}^{n\times n}$, f une fonction définie sur le spectre de A et $p_{f,A}$ le polynôme interpolateur de f au sens d'Hermite donné par*(1.3)*. Alors*

$$f(A) := p_{f,A}(A).$$

Démonstration. Soit $A \in \mathbb{C}^{n\times n}$ avec décomposition de Jordan $A = Z \operatorname{diag}(J_1, \dots, J_p)Z^{-1}$ où $J_k = J_{m_k}(\lambda_k) \in \mathbb{C}^{m_k \times m_k}$, f une fonction définie sur $\sigma(A)$ et p polynôme interpolateur d'Hermite de f satisfaisant (1.3) aux points λ_i, pour $i = 1, .., p$. Alors d'après les propriétés sur les polynômes de matrices, on a

$$p(A) = Zp(\operatorname{diag}(J_1, \dots, J_p))Z^{-1} = Z \operatorname{diag}\left(p(J_1), \dots, p(J_p)\right)Z^{-1}.$$

Or $p(J_k)$ est donné pour tout $k = 1, \dots, p$ par

$$p(J_k) = \begin{bmatrix} p(\lambda_k) & p^{'}(\lambda_k) & \dots & \frac{p^{(m_k-1)}(\lambda_k)}{(m_k-1)!} \\ 0 & p(\lambda_k) & \ddots & \vdots \\ \vdots & \ddots & \ddots & p^{'}(\lambda_k) \\ 0 & \dots & 0 & p(\lambda_k) \end{bmatrix} = \begin{bmatrix} f(\lambda_k) & f^{'}(\lambda_k) & \dots & \frac{f^{(m_k-1)}(\lambda_k)}{(m_k-1)!} \\ 0 & f(\lambda_k) & \ddots & \vdots \\ \vdots & \ddots & \ddots & f^{'}(\lambda_k) \\ 0 & \dots & 0 & f(\lambda_k) \end{bmatrix} = f(J_k),$$

où la deuxième égalité provient de la condition (1.3) et la dernière égalité vient de (1.2). Par égalité sur tous les blocs de Jordan dans la décomposition de la matrice A, on retrouve bien la définition d'une fonction de matrice donnée en définition 1.2.2. □

Notons que le polynôme $p_{f,A}$ dépend des propriétés de la matrice A.

Lemme 1.2.7. *Si $f(z) = \sum_{j=0}^{\infty} a_j z^j$ admet un rayon de convergence $R > \varrho(A) = \max\{|\lambda| : \lambda \in \sigma(A)\}$, alors $f(A) = \sum_{j=0}^{\infty} a_j A^j$ c'est-à-dire que l'on a la convergence $\|f(A) - \sum_{j=0}^{k} a_j A^j\|_2 \underbrace{\to}_{k\to\infty} 0$.*

Donc pour toute fonction f possédant un développement de Taylor, on peut définir $f(A)$, lorsque le rayon de convergence de la fonction est plus grand que le rayon spectral de la matrice, en remplaçant dans cette série la variable t par notre matrice A.

Exemple 1.2.8. *On sait par exemple que l'exponentielle sert souvent pour résoudre les systèmes d'équations différentielles ordinaires à coefficients constants $X^{'} = AX$ avec $A \in \mathbb{C}^{n\times n}$, $X = X(t) \in \mathbb{C}^n$. On sait alors qu'une solution est donnée par $X(t) = \exp(tA)$. Or la fonction exponentielle possède le développement de Taylor*

$$\exp(t) = \sum_{j=0}^{\infty} \frac{t^j}{j!}, \text{ avec rayon de convergence infini}$$

Par le lemme précédent, on a donc que pour toute matrice $A \in \mathbb{C}^{n\times n}$,

$$exp(A) = \sum_{j=0}^{\infty} \frac{A^j}{j!},$$

c'est-à-dire que l'on remplace dans la série de Taylor la variable scalaire par notre matrice A.

Exemple 1.2.9. *Si on souhaite travailler avec la fonction logarithme, on sait que celle-ci possède un développement de Taylor sur le disque de convergence $\mathbb{D} = \{t \in \mathbb{C} : |t| < 1\}$. On peut alors d'après ce qui précède identifier la fonction de matrice logarithme de la manière suivante si $\varrho(A) < 1$:*

$$f(t) = log(1+t) = t - \frac{t^2}{2} + \frac{t^3}{3} - .. \quad lorsque\ |t| < 1$$
$$\Longrightarrow f(A) = A - \frac{A^2}{2} + \frac{A^3}{3} - ...\ lorsque\ \varrho(A) < 1.$$

1.2.3 Définition intégrale

On définit l'intégrale d'une matrice comme la matrice des intégrales des coefficients, c'est-à-dire que si $A = (a_{i,j})_{i,j=1\dots,n}$, alors $\int A = \left(\int a_{i,j}\right)_{i,j=1,\dots,n}$.

Corollaire 1.2.10. *Soit f une fonction analytique sur un ensemble ouvert $\Omega \subseteq \mathbb{C}$ et $\Gamma \subset \Omega$ un système de courbes de Jordan entourant chaque valeur propre $\lambda \in \sigma(A)$ exactement une fois dans le sens positif. Alors*

$$f(A) := \frac{1}{2i\pi} \int_\Gamma f(\zeta)(\zeta I_n - A)^{-1} d\zeta.$$

Démonstration. Soit $A \in \mathbb{C}^{n\times n}$. Si $A = Z \operatorname{diag}(J_1(\lambda_1), \dots, J_p(\lambda_p))Z^{-1}$, alors

$$\begin{aligned}
\frac{1}{2i\pi} \int_\Gamma f(\zeta)(\zeta I - A)^{-1} d\zeta &= Z\Big(\frac{1}{2i\pi} \int_\Gamma f(\zeta)(\zeta I - \operatorname{diag}(J_1(\lambda_1), \dots, J_p(\lambda_p))^{-1} d\zeta\Big) Z^{-1} \\
&= Z\Big(\frac{1}{2i\pi} \int_\Gamma f(\zeta)(\operatorname{diag}(\zeta I - J_1(\lambda_1), \dots, \zeta I - J_p(\lambda_p))^{-1} d\zeta\Big) Z^{-1} \\
&= Z\Big(\frac{1}{2i\pi} \int_\Gamma f(\zeta) \operatorname{diag}\left((\zeta I - J_1(\lambda_1))^{-1}, \dots, (\zeta I - J_p(\lambda_p))^{-1}\right) d\zeta\Big) Z^{-1}.
\end{aligned}$$

Or, pour un bloc de Jordan $(\zeta I - J_k(\lambda_k)) = (\zeta I - J_{m_k}(\lambda_k))$, on a

$$(\zeta I - J_{m_k}(\lambda_k))^{-1} = \begin{bmatrix} \frac{1}{\zeta-\lambda_k} & \frac{1}{(\zeta-\lambda_k)^2} & \cdots & \frac{1}{(\zeta-\lambda_k)^{m_k}} \\ 0 & \frac{1}{\zeta-\lambda_k} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \frac{1}{(\zeta-\lambda_k)^2} \\ 0 & \cdots & \cdots & \frac{1}{\zeta-\lambda_k} \end{bmatrix}.$$

En intégrant terme à terme, on obtient

$$\begin{aligned}
\frac{1}{2i\pi} \int_\Gamma f(\zeta)(\zeta I - A)^{-1} d\zeta &= Z\Big(\frac{1}{2i\pi} \int_\Gamma f(\zeta) \operatorname{diag}\left((\zeta I - J_1(\lambda_1))^{-1}, \dots, (\zeta I - J_p(\lambda_p))^{-1}\right) d\zeta\Big) Z^{-1} \\
&= Z \operatorname{diag}\left(\begin{bmatrix} \frac{1}{2i\pi}\int_\Gamma \frac{f(\zeta)}{z-\lambda_j} & \cdots & \frac{1}{2i\pi}\int_\Gamma \frac{f(\zeta)}{(z-\lambda_j)^{m_j}} \\ & \ddots & \vdots \\ & & \frac{1}{2i\pi}\int_\Gamma \frac{f(\zeta)}{z-\lambda_j} \end{bmatrix} \right)_{j=1,\dots,p} Z^{-1}
\end{aligned}$$

ce qui, d'après la formule de Cauchy, nous donne

$$\begin{aligned}
\frac{1}{2i\pi} \int_\Gamma f(\zeta)(\zeta I - A)^{-1} d\zeta &= Z \operatorname{diag}\left(\begin{bmatrix} f(\lambda_j) & f'(\lambda_j) & \dots & f(\lambda_j)^{(m_j-1)} \\ & \ddots & \ddots & \vdots \\ & & & f'(\lambda_j) \\ & & & f(\lambda_j) \end{bmatrix} \right)_{j=1,\dots,p} Z^{-1} \\
&= Z \operatorname{diag}(f(J_1(\lambda_1)), \dots, f(J_p(\lambda_p)))Z^{-1} = f(A)
\end{aligned}$$

d'après la définition 1.2.2. □

Proposition 1.2.11. *Ces 3 définitions sont équivalentes pour toute fonction f analytique sur un ouvert contenant le spectre de A.*

Démonstration. La définition 1.2.6 nous dit que pour une matrice $A \in \mathbb{C}^{n\times n}$ et une fonction f définie sur $\sigma(A)$, $f(A) = p(A)$. Or par décomposition de Jordan de la matrice A, nous avons vu qu'évaluer p sur la matrice A revenait à évaluer p sur chaque bloc de Jordan et que pour chaque bloc de Jordan J_k, $p(J_k)$ était entièrement déterminée par les valeurs de p et de ses dérivées sur le spectre de A et donc entièrement déterminée par les valeurs de f et de ses dérivées sur le spectre de A, et ainsi on retrouve les mêmes blocs que pour la définition 1.2.2, ce qui montre l'équivalence entre les définitions 1.2.2 et 1.2.6.

Montrons à présent l'équivalence entre les définitions 1.2.2 et 1.2.10. Pour toute matrice $M = (m_{i,j})_{i,j=1,\dots,n} \in \mathbb{C}^{n\times n}$, on note $\int_C M dz = (\int_C (m_{i,j}))_{i,j}$. Soit maintenant $A \in \mathbb{C}^{n\times n}$ et Γ un système de courbes de Jordan entourant chaque $\lambda \in \sigma(A)$ une seule fois. Considérons la décomposition de Jordan de A : $A = Z\Lambda Z^{-1} = Z\,\mathrm{diag}(J_1(\lambda_1), \dots, J_p(\lambda_p))Z^{-1}$, avec Z inversible. Alors $\frac{1}{2i\pi}\int_\Gamma f(z)(zI - A)^{-1}dz = Z(\frac{1}{2i\pi}\int_\Gamma f(z)(zI-\Lambda)^{-1}dz)Z^{-1}$. Puisque f est analytique, il nous suffit de calculer cette intégrale sur chaque bloc de Jordan de la décomposition. Soit donc $\lambda = \lambda_i$ pour un $i \in \{1, \dots, n\}$ et $J = J_i(\lambda_i)$ le bloc de Jordan associé. Alors $zI - J = -J_i(\lambda - z)$. Or l'inverse de ce "bloc de Jordan" est donné par

$$(-J_i(\lambda - z))^{-1} = \begin{pmatrix} \frac{1}{z-\lambda} & \frac{1}{(z-\lambda)^2} & \cdots & \cdots & \frac{1}{(z-\lambda)^m} \\ 0 & \frac{1}{z-\lambda} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \frac{1}{(z-\lambda)^2} \\ 0 & \dots & \dots & 0 & \frac{1}{z-\lambda} \end{pmatrix}$$

où m est la dimension du bloc de Jordan, et alors

$$\frac{1}{2i\pi}\int_\Gamma f(z)(zI-J)^{-1}dz$$
$$= \begin{pmatrix} \frac{1}{2i\pi}\int_\Gamma f(z)\frac{1}{z-\lambda}dz & \frac{1}{2i\pi}\int_\Gamma f(z)\frac{1}{(z-\lambda)^2}dz & \dots & \frac{1}{2i\pi}\int_\Gamma f(z)\frac{1}{(z-\lambda)^m}dz \\ 0 & \frac{1}{2i\pi}\int_\Gamma f(z)\frac{1}{z-\lambda}dz & \ddots & \vdots \\ \vdots & \ddots & \ddots & \frac{1}{2i\pi}\int_\Gamma f(z)\frac{1}{(z-\lambda)^2}dz \\ 0 & \dots & 0 & \frac{1}{2i\pi}\int_\Gamma f(z)\frac{1}{z-\lambda}dz \end{pmatrix}$$

(car f est analytique) et d'après la formule de Cauchy, on obtient

$$\frac{1}{2i\pi}\int_\Gamma f(z)(zI-J)^{-1}dz = \begin{pmatrix} f(\lambda) & f'(\lambda) & \dots & \frac{f^{(m-1)}(\lambda)}{(m-1)!} \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & f'(\lambda) \\ 0 & \dots & 0 & f(\lambda) \end{pmatrix}.$$

En rassemblant tous les termes de la décomposition, on retombe sur la définition précédente de fonction de matrice. □

1.2.4 Généralités

Une fois que nous possédons la fonction de matrice $f(A)$, il nous faudra peut-être également considérer des opérations supplémentaires sur la fonction f ou sur la matrice A. En effet, nous pouvons être amenés à considérer la matrice conjuguée de A ou toute matrice équivalente à A. D'autre part, si une fonction f s'écrit comme la somme ou le produit de deux fonctions pour lesquelles les fonctions de matrices sont connues, plutôt que de recalculer une nouvelle fonction de matrice, on souhaiterait pouvoir l'obtenir à partir des fonctions de matrices dont nous disposons déjà. Pour résoudre ce problème, nous pouvons déduire des définitions précédentes quelques propriétés des fonctions de matrices que nous énonçons ici.

Proposition 1.2.12. *i.* $\forall A \in \mathbb{C}^{n\times n}, f(A^*) = (f(A))^*$ *si* $\forall z \in \mathbb{C},\ f(\bar{z}) = \overline{f(z)}$.

ii. $\forall A \in \mathbb{C}^{n\times n}, f(\bar{A}) = \overline{(f(A))}$ *si* $\forall z,\ f(\bar{z}) = \overline{f(z)}$.

iii. $\forall A \in \mathbb{C}^{n\times n}, X \in GL_n(\mathbb{C})$, *on a* $f(XAX^{-1}) = Xf(A)X^{-1}$.

iv. $XA = AX \Longrightarrow Xf(A) = f(A)X$.

v. $f(\mathrm{diag}(A, B)) = \mathrm{diag}(f(A), f(B))$.

Démonstration. i. Puisque $f(\bar{\lambda}) = \overline{f(\lambda)}$, on a, en notant p le polynôme d'interpolation d'Hermite de f sur $\sigma(A^*) = \overline{\sigma(A)}$ que

$$\overline{p(\lambda)} = p(\bar{\lambda}) = f(\bar{\lambda}) = \overline{f(\lambda)}, \ \ \forall \lambda \in \sigma(A).$$

De plus, la propriété $f(\bar{z}) = \overline{f(z)}$ entraîne que $f^{(i)}(\bar{z}) = \overline{f^{(i)}(z)}$. On a alors les mêmes résultats précédents, d'où

$$p \text{ interpole } f \text{ sur } \sigma(A^*) \Rightarrow \bar{p} \text{ interpôle } f \text{ sur } \sigma(A).$$

et donc

$$f(A^*) = p(A^*) = (\bar{p}(A))^* = f(A)^*.$$

ii. En utilisant la propriété précédente et en sachant que $f(A^T) = f(A)^T$ (par l'interpolation polynomiale), on obtient

$$f(\bar{A}) = f((A^*)^T) = f(A^*)^T = [(f(A))^*]^T = \overline{f(A)}.$$

iii. Montrons que le résultat est vrai pour tout polynôme p :
— supposons que $p(x) = x^j$ monôme de degré j. Alors on a :

$$\begin{aligned} p(XAX^{-1}) = (XAX^{-1})^j &= XA\underbrace{X^{-1}X}_{=I}A\underbrace{X^{-1}X}_{=I}A...\underbrace{X^1X}_{=I}AX^{-1} \\ &= XA^jX^{-1} = Xp(A)X^{-1}. \end{aligned}$$

.
— Supposons maintenant que $p(x) = \sum_{j=0}^{deg(p)} a_j x^j$. Alors

$$\begin{aligned} p(XAX^{-1}) = \sum_{j=0}^{deg(p)} a_j(XAX^{-1})^j = \sum_{j=0}^{deg(p)} a_j XA^jX^{-1} &= X(\sum_{j=0}^{deg(p)} a_j A^j)X^{-1} \\ = Xp(A)X^{-1}, & \end{aligned}$$

ainsi donc le résultat est vrai pour tout polynôme, en particulier pour l'unique polynôme interpolateur de Hermite de f, et donc puisque $\forall X \in GL_n(\mathbb{C}), \sigma(A) = \sigma(XAX^{-1})$ (on peut le montrer par double inclusion), on obtient :

$$f(XAX^{-1}) = p(XAX^{-1}) = Xp(A)X^{-1} = Xf(A)X^{-1}$$

.

iv. Le résultat est vrai pour les polynômes : on teste d'abord sur les monômes ($XA^j = AXA^{j-1} = ... = A^{j-1}XA = A^jX$) puis sur un polynôme quelconque en utilisant le résultat sur les monômes.

v. On peut montrer par récurrence que pour tout $j \geqslant 1$, $\begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}^j = \begin{bmatrix} A^j & 0 \\ 0 & B^j \end{bmatrix}$ puis étendre le résultat à tout polynôme et en particulier au polynôme d'interpolation de la fonction étudiée, d'où le résultat.

□

Proposition 1.2.13. *Soient $A \in \mathbb{C}^{n\times n}$, f et g deux fonctions définies sur $\sigma(A)$. Alors*

i. Si $f(z) \equiv 1$, alors $f(A) = I$ avec I la matrice identité pour toute matrice A ;

ii. $(\lambda.f)(A) = \lambda.(f(A))$ *;*

iii. $(f + g)(A) = f(A) + g(A)$ *;*

iv. $(f \times g)(A) = f(A) \times g(A)$;

Démonstration. Le premier point est évident en prenant le polynôme égal à 1 et la définition par polynôme interpolateur d'Hermite.
Soit $p_{f,A}$ polynôme interpolateur de f au sens d'Hermite. Alors $\lambda.p_{f,A}$ interpôle $\lambda.f$ au sens d'Hermite, d'où $(\lambda.f)(A) = (\lambda.p_{f,A})(A) = \lambda.p_{f,A}(A) = \lambda.(f(A))$.
Soient $p_{f,A}$ et $p_{g,A}$ les polynômes interpolateurs de f et g respectivement au sens d'Hermite. Alors $p_{f,A}+p_{g,A}$ interpôle $f+g$ au sens d'Hermite par les propriétés de la dérivée et alors $(f+g)(A) = (p_{f,A}+p_{g,A})(A) = p_{f,A}(A) + p_{g,A}(A) = f(A) + g(A)$.
Pour le produit, d'après la règle de différentiation, on a en notant $\sigma(A) = \{\lambda_1, \ldots, \lambda_m\}$ que

$$(f \times g)^{(k)}(\lambda_i) = \sum_{j=0}^{k} \binom{k}{j} f^{(j)}(\lambda_i) \times g^{(k-j)}(\lambda_i) = \sum_{j=0}^{k} \binom{k}{j} p_{f,A}^{(j)}(\lambda_i) \times p_{g,A}^{(k-j)}(\lambda_i) = (p_{f,A} \times p_{g,A})^{(k)}(\lambda_i)$$

pour tout $i = 1, \ldots, m$ et $k = 0, \ldots, n(\lambda_i) - 1$, d'où $p_{f,A} \times p_{g,A}$ interpôle $f \times g$ au sens d'Hermite et donc $(f \times g)(A) = (p_{f,A} \times p_{g,A})(A) = p_{f,A}(A) \times p_{g,A}(A) = f(A) \times g(A)$. □

1.3 Quelques méthodes de calcul

L'emploi des définitions précédentes de fonctions de matrices n'est généralement pas une bonne idée pour une implémentation numérique. Cependant, des alternatives pour l'implémentation numérique d'une fonction de matrice existent, certaines ne prenant en compte ni la structure de la matrice ni la fonction considérée et d'autres plus spécifiques à certaines fonctions. Nous décrivons brièvement ici quelques méthodes numériques pour le calcul de la fonction de matrice $f(A)$, avec l'étude du cas particulier des fonctions polynomiales à l'aide la méthode de Horner, puis dans un cadre plus général, nous rappelons la méthode de Schur-Parlett applicable à toute fonction f. Nous renvoyons également au livre de Higham [57] pour d'autres références sur le calcul de $f(A)$.

1.3.1 Evaluation polynômiale et méthode de Horner

Lorsque l'on considère un polynôme $p \in \mathcal{P}_m$ avec $p(x) = \sum_{j=0}^{m} b_j x^j$ et $A \in \mathbb{C}^{n\times n}$, l'évaluation de $p(A)$ est alors donnée par

$$p(A) = \sum_{j=0}^{m} b_j A^j.$$

Pour implémenter ce polynôme de matrices, plutôt que de calculer chaque puissance A^j et de les stocker, on peut faire appel à une autre méthode.

En effet, la méthode de Horner permet d'éviter le stockage de plusieurs puissances de A en construisant $p(A)$ par récurrence de la manière suivante [57, Section 4.2] :

Algorithm 1 Algorithme de Horner

Require: $b_1, \ldots ; b_m$;
Ensure: $p(A) = \sum_{j=0}^{m} b_j A^j$;
$S_m = b_m A + b_{m-1} I$;
for $k = m-2 : -1 : 0$ **do**
$S_k = A S_{k+1} + b_k I$;
end for
$p(A) = S_0$;

Cet algorithme permet de construire $p(A)$ en $\mathcal{O}((m-1)n^3)$ opérations élémentaires lorsque $A \in \mathbb{C}^{n\times n}$.

Soit à présent $p \in \mathcal{P}_m$ un polynôme de degré m. Quitte à ajouter des coefficients nuls au polynôme p, on peut supposer que $m = rs - 1$ avec $r, s \in \mathbb{N}^*$. On note alors que le polynôme $p(A)$ peut être calculé à l'aide de la méthode de Paterson et Stockmeyer, en décomposant $p(A)$ sous la forme

$$p(A) = \sum_{j=0}^{r-1} B_j (A^s)^j, \quad B_j = \sum_{k=0}^{s-1} b_{sj+k} A^k, \tag{1.4}$$

où les puissances $A^2, \ldots, A^s$ sont calculées puis les termes de (1.4) sont calculés à l'aide de l'algorithme de Horner. On arrive alors à une complexité $(r+s)n^3$.

1.3.2 La méthode de Schur-Parlett

Pour une fonction générale f définie sur le spectre d'une matrice $A \in \mathbb{C}^{n\times n}$ où $A = ZBZ^{-1}$, une méthode pour calculer la fonction de matrice $f(A)$ est l'utilisation de la propriété

$$f(ZBZ^{-1}) = Zf(B)Z^{-1}, \ \ \forall Z, B \in \mathbb{C}^{n\times n} \text{ avec } Z \text{ inversible}$$

lorsque $f(B)$ est facilement implémentable [29, Section 1]. Par exemple lorsque A est diagonalisable avec valeurs propres $\lambda_1, \ldots, \lambda_n$, alors $f(A) = Z \operatorname{diag}(f(\lambda_1), \ldots, f(\lambda_n) Z^{-1}$.
L'erreur d'évaluation dépendant du conditionnement $\kappa(Z) = \|Z\| \, \|Z^{-1}\| \geqslant 1$, il est par conséquent préférable de trouver une décomposition $A = ZBZ^{-1}$ avec un conditionnement faible. Pour ce faire, on considère la décomposition de Schur :

Définition 1.3.1. *Soit $A \in \mathbb{C}^{n\times n}$ une matrice. Alors existe une matrice unitaire $Q \in \mathbb{C}^{n\times n}$ et une matrice triangulaire supérieure [29, Section 1.1] $X \in \mathbb{C}^{n\times n}$ telles que*

$$A = QXQ^*.$$

Cette décomposition est appelée décomposition de Schur. Cette méthode nous impose alors une complexité de $\mathcal{O}(n^3)$.

Cette complexité peut être réduite à $\mathcal{O}(n^2)$ en sélectionnant la première matrice Q_0 dans la décomposition $H_0 = Q_0 A Q_0^*$ de sorte à ce que H_0 soit Hessenberg, ce qui réduit le coût des décompositions QR dans les étapes suivantes de la décomposition de Schur.

A partir de la décomposition de Schur, nous utilisons ensuite une décomposition en blocs de la matrice

X,

$$X = \begin{pmatrix} X_{1,1} & X_{1,2} & \dots & X_{1,l} \\ 0 & X_{2,2} & \dots & X_{2,l} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & X_{l,l} \end{pmatrix}$$

où les $X_{j,j}$ sont des sous-matrices carrées et les $X_{j,k}$ pour $k > j$ sont des matrices rectangulaires. $F = f(X)$ possède alors la même structure en blocs que la matrice X et on calcule $f(X)$ à l'aide de la récurrence de Parlett par blocs :

— Premièrement, on calcule les blocs diagonaux $F_{j,j} = T(X_{j,j})$ qui sont obtenus par une méthode directe [57, Section 9.1] comme un développement de Taylor, une interpolation polynômiale...

— On calcule les blocs restants $F_{j,k}$ par indice $j - k \geqslant 1$ croissant en résolvant l'équation de Sylvester

$$X_{j,j}F_{j,k} - F_{j,k}X_{k,k} = \sum_{i=j}^{k-1} F_{j,i}X_{i,k} - \sum_{i=j+1}^{k} X_{j,k}F_{i,k}$$

d'après l'identité $Xf(X) = f(X)X$ soit $XF = FX$. Pour avoir l'existence d'une solution à l'équation de Sylvester il nous faut nous assurer que les blocs diagonaux $X_{j,j}$ n'ont pas de valeur propre commune. On prendra donc soin au préalable de réorganiser la matrice X pour éviter que les blocs diagonaux aient des valeurs propres communes. Cette étape peut être réalisée à l'aide de techniques standards [57, Section 9.3].

Remarque 1.3.2. *Le problème de cette méthode bien qu'efficace est qu'elle nécessite encore une complexité d'ordre n^3 lorsque n est la dimension. Par conséquent, il nous faut envisager d'autres méthodes. Nous allons donc considérer dans la suite l'approximation des fonctions de matrices.*

1.3.3 Méthodes d'approximation

Plutôt que de calculer nos fonctions de matrices à l'aide d'algorithmes directes, puisque les fonctions polynomiales ou rationnelles sont généralement plus facilement calculables que les fonctions elles-mêmes, on peut alors chercher si un approximant polynômial ou rationnel pour une fonction de matrice existe. Comme l'approximation polynomiale et rationnelle en dimension 1 dispose d'une large littérature, nous pouvons envisager de transposer ces approximants au cas matriciel et ainsi considérer la fonction de matrice $g(A)$ approchant $f(A)$ lorsque g est un approximant polynomial ou rationnel de la fonction f en dimension 1. Dans cette section, nous motivons l'étude des approximants de fonctions à l'aide des ensembles K-spectraux, nous permettant de mesurer l'erreur d'approximation de notre fonction de matrice $f(A)$ par $g(A)$ à l'aide de l'erreur d'approximation en dimension 1. Ce résultat nous permet alors de considérer plusieurs approximants potentiels de la fonction tels que les polynômes de Faber ou les approximants de Padé.

Sauf indication dans le reste de la thèse, nous notons pour une matrice $A \in \mathbb{C}^{n\times n}$, $\|A\| := \|A\|_2 = \sup\{\|A \cdot x\|_2, \|x\| = 1\}$ la norme 2 sur les matrices.

Définition 1.3.3. *Un ensemble fermé $\mathbb{E} \subset \mathbb{C}$ est un ensemble K-spectral pour une matrice $X \in \mathbb{C}^{n\times n}$ si il existe une constante K telle que $\sigma(X) \subseteq \mathbb{E}$ et pour toute fonction g analytique sur un voisinage de $\mathbb{E}$ on a*

$$\|g(X)\| \leqslant K\|g\|_{L^\infty(\mathbb{E})}$$

Exemple 1.3.4.

— *Tout disque fermé* $\{z \in \mathbb{C} : |z-\alpha| \leqslant r\}$ *avec* α, r *tels que* $\|X - \alpha I\| \leqslant r$ *est un ensemble* 1*-spectral [5, Section 107.2].*
— *Lorsque* X *est normale, le spectre* $\sigma(X)$ *est un ensemble* 1*-spectral pour* X *[8, Lemma 4.3].*
— *Le rang numérique* $W(X) = \left\{ \frac{y^*Xy}{y^*y} : y \in \mathbb{C}^n \backslash \{0\} \right\}$ *est* K*-spectral avec* $K \leqslant 11,08$*. En particulier, le disque centré en* 0 *et de rayon* $\max\{|z| : z \in W(X)\}$ *est* 2*-spectral pour* X *[26, Section 3]. De plus, pour tout ensemble borné convexe* Ω *avec bords réguliers tel que* $\overline{W(X)} \subseteq \Omega$*,* $K \leqslant 1 + \sqrt{2}$ *[27].*

Nous renvoyons à [5] pour d'autres exemples d'ensembles K*-spectraux.*

La définition d'un ensemble K-spectral $\mathbb{E}$ ne dépendant que des propriétés spectrales de la matrice X, pour toute fonction f analytique sur $\mathbb{E}$, nous pouvons considérer n'importe quel approximant g de f et majorer l'erreur sans pôles dans $\mathbb{E}$

$$\|f(X) - g(X)\| \leqslant K \|f - g\|_{L^\infty(\mathbb{E})}.$$

Le challenge est à présent de déterminer, pour une matrice $X \in \mathbb{C}^{n\times n}$ et une fonction f analytique sur un ouvert $\Omega \supseteq \sigma(X)$, un ensemble $\mathbb{E} \subseteq \Omega$ avec $\sigma(X) \subseteq \mathbb{E}$, un approximant g minimisant l'erreur $\|f - g\|_{L^\infty(\mathbb{E})}$. Nous rappelons à présent quelques exemples classiques d'approximants polynomiaux et rationnels.

Polynômes de Faber

Soit $\mathbb{E} \subseteq \overline{\mathbb{C}}$ un ensemble convexe compact et f une fonction analytique dans un voisinage de $\mathbb{E}$. Ici nous cherchons à minimiser $\|f - p\|_{L^\infty(\mathbb{E})}$ avec $p \in \mathcal{P}_n$. Soient $\phi : \overline{\mathbb{C}}\backslash\mathbb{E} \to \overline{\mathbb{C}}\backslash\overline{\mathbb{D}}$ l'application de Riemann vérifiant $\phi(\infty) = \infty$, $\phi'(\infty) > 0$ et $\phi'(z) \neq 0$ $\forall z$. Enfin soit $\varphi = \phi^{-1} : \overline{\mathbb{C}}\backslash\overline{\mathbb{D}} \to \overline{\mathbb{C}}\backslash\mathbb{E}$. On définit les polynômes $F_j(z)$ pour tout $j \geqslant 0$ et tout $z \in Int(\mathbb{E})$ et $|w| \geqslant 1$ par la fonction génératrice [28, equation 2.3]

$$\frac{w\varphi'(w)}{\varphi(w) - z} = \sum_{j=0}^{\infty} \frac{F_j(z)}{w^j}.$$

Par exemple, lorsque $\mathbb{E} = \overline{\mathbb{D}}$, $\varphi(w) = w$ et $F_j(z) = z^j$.

Pour tout $j \geqslant 0$, les F_j sont des polynômes de degré j et pour tout $j \geqslant 1$, $F_j(\varphi(w)) - w^j$ est analytique sur $\overline{\mathbb{C}}\backslash\overline{\mathbb{D}}$ et s'annule en ∞. De plus, $\|F_j\|_{\mathbb{E}} \leqslant 2$.

Pour $\mathbb{E}_R = \{z \notin \mathbb{E}, |\phi(z)| \leqslant R\}$ et f analytique dans un voisinage de $\mathbb{E}_R$ pour un $R > 1$, on définit les coefficients de Faber

$$f_j = \frac{1}{2i\pi} \int_{|w|=1} \frac{f(\varphi(w))}{w^j} \frac{dw}{w}, \quad j \geqslant 0.$$

Alors pour tout $j \geqslant 0$, $f_j = \mathcal{O}(R^{-j})$ et les sommes partielles de la série de Faber $\sum_{j=0}^{\infty} f_j F_j(z)$ convergent uniformément vers f dans $\mathbb{E}$ [28, p. 590]. De plus, ces coefficients vérifient

$$|f_{m+1}| \leqslant \sqrt{\sum_{j=m+1}^{\infty} |f_j|^2} \leqslant \min_{P \in \Pi_m} \|f - P\|_{L^\infty(\mathbb{E})} \leqslant \|f - \sum_{j=0}^{m} f_j F_j\|_{L^\infty(\mathbb{E})} \leqslant 2 \sum_{j=m+1}^{\infty} |f_j|.$$

Donc lorsque les coefficients f_j décroissent rapidement vers 0, nous avons un bon approximant.

Transformée de Faber et pôles prescrits

Soient $w_1, \ldots, w_m \in \overline{\mathbb{C}}\backslash\overline{\mathbb{D}}$ distincts, $Q(w) = \prod_{j=1}^{m}(1 - w/w_j)$ et $q(z) = \prod_{j=1}^{m}(z - \varphi(w_j))$ où $\psi = \overline{\mathbb{C}}\backslash\overline{\mathbb{E}} \to$

$\overline{\mathbb{C}}\backslash\overline{\mathbb{D}}$ application conforme.

Proposition 1.3.5. *Soient f analytique dans un voisinage de $\mathbb{E}$, $R_m = P_m/Q$ l'interpolant polynômial de $F(w) = f_0/2 + \sum_{j=1}^{\infty} f_j w^j$ aux points $0,\ 1/w_1, \ldots, 1/w_m$ où $f_j = \frac{1}{2i\pi}\int_{|w|=1} \frac{f(\psi(w))}{w^j}\frac{dw}{w}$. Définissons*

$$B(w) = w\prod_{j=1}^{m}\frac{w-1/\bar{w}_j}{1-w/w_j},\ \ \frac{p_m}{q} := \mathcal{F}\Big(\frac{P_m}{Q}\Big),\ \ b_j := \frac{1}{2i\pi}\int_{|u|=1}\frac{f(\varphi(w))du}{B(u)u^j},$$

avec $\mathcal{F}$ la transformée de Faber. Alors

$$|b_1| \leqslant \sqrt{\sum_{j=1}^{\infty}|b_j|^2} \leqslant \min_{p\in P_m}\Big\|f-\frac{p}{q}\Big\|_{\mathbb{E}} \leqslant \Big\|f-\frac{p_m}{q}\Big\|_{\mathbb{E}} \leqslant 2\sum_{j=1}^{\infty}|b_j|$$

On obtient ainsi lorsque $W(A)\subseteq\mathbb{E}$ et A hermitienne

$$\|f(A)-p_m(A)q(A)^{-1}\| \leqslant \|F-P_m/Q\|_{\overline{\mathbb{D}}} \leqslant \sum_{j=1}^{\infty}|b_j|.$$

Démonstration. Montrons d'abord que $p_m \in \mathcal{P}_m$. Considérons la décomposition en éléments simples de $\frac{P_m}{Q}$. Alors

$$\begin{aligned}\mathcal{F}\Big(\frac{P_m(w)}{Q(w)}\Big)(z) &= \mathcal{F}\Big(c_0+\sum_{j=1}^{m}\frac{c_j}{w-w_j}\Big)(z) = \mathcal{F}\Big(c_0-\sum_{j=1}^{m}\frac{c_j}{w_j}\sum_{k=0}^{\infty}\frac{w^k}{w_j^k}\Big)(z)\\ &= c_0\mathcal{F}(1)(z)-\sum_{j=1}^{m}\frac{c_j}{w_j}\sum_{k=0}^{\infty}\frac{\mathcal{F}(w^k)(z)}{w_j^k} = c_0+\frac{P_m}{Q}(0)-\sum_{j=1}^{m}\frac{c_j}{w_j}\frac{w_j\psi'(w_j)}{\psi(w_j)-z}\\ &= c_0+\frac{P_m}{Q}(0)+\sum_{j=1}^{m}\frac{c_j\psi'(w_j)}{z-\psi(w_j)}\end{aligned}$$

qui est un élément de $\mathcal{P}_m/q$. Pour la dernière inégalité, $F-P_m/Q$ est analytique dans un voisinage de $|w|\leqslant 1+\varepsilon$ pour un $\varepsilon>0$. D'après la formule d'Hermite, pour f analytique sur un $\Omega_0\supseteq\Omega$ avec $1/w_j\in Int(\Omega_0)$,

$$f(z)-R_{m,Q}(z) = B(z)\frac{1}{2i\pi}\int_{\partial\Omega_0}\frac{f(\zeta)}{B(\zeta)}\frac{d\zeta}{\zeta-z}$$

et on obtient pour $|w|=1$ sachant que $|B(w)|=1$,

$$\begin{aligned}|F(w)-\frac{P_m}{Q}(w)| &= |B(w)|\Big|\frac{1}{2i\pi}\int_{|u|=1+\varepsilon}\frac{F(u)}{B(u)}\frac{du}{u-w}\Big| = \Big|\frac{1}{2i\pi}\int_{|u|=1+\varepsilon}\frac{f(\psi(u))}{B(u)}\frac{du}{u-w}\Big|\\ &= \Big|\sum_{j=1}^{\infty}b_jw^{j-1}\Big| \leqslant \sum_{j=1}^{\infty}|b_j|,\end{aligned}$$

où la deuxième égalité provient du fait que

$$u\mapsto\frac{F(u)-f(\psi(u))}{B(u)}\frac{1}{u-w}$$

est analytique dans $|u|\geqslant 1=\varepsilon$ inclus ∞ avec un double zéro en ∞. Or, pour tout polynôme P, $\|\mathcal{F}(P)\|_{\mathbb{E}}\leqslant 2\|P\|_{\overline{\mathbb{D}}}$, d'où

$$\|f-\frac{p_m}{q}\|_{\mathbb{E}} = \|\mathcal{F}\Big(F-\frac{P_m}{Q}\Big)\|_{\mathbb{E}} \leqslant 2\|F-\frac{P_m}{Q}\|_{\overline{\mathbb{D}}} \leqslant 2\sum_{j=1}^{\infty}|b_j|. \tag{1.5}$$

Démontrons à présent la deuxième inégalité. Pour tout polynôme $p \in \mathcal{P}_m$, nous pouvons écrire

$$(f - \frac{p}{q})(\psi(w)) = w(\tilde{F}(w) - \frac{P}{Q}(w)) + H(w), \quad \tilde{F}(w) = \sum_{j=1}^{\infty} F_j w^j = \frac{F(w) - F(0)}{w},$$

avec $P \in \mathcal{P}_{m-1}$ et H analytique dans $|u| > 1$, obtenu par développement en série de Faber. Comme le terme à gauche du second membre est analytique dans un voisinage du disque et s'annule en 0, nous obtenons alors

$$\|f - \frac{p}{q}\|_{\mathbb{E}}^2 = \|(f - \frac{p}{q}) \circ \psi\|_{\partial\mathbb{D}}^2 \geqslant \|(f - \frac{p}{q}) \circ \psi\|_2^2 = \|\tilde{F} - \frac{P}{Q}\|_2^2 + \|H\|_2^2.$$

Notons qu'il existe un polynôme $\tilde{P} \in \mathcal{P}_{m-1}$ tel que

$$\frac{P_m}{Q}(w) - F(0) = \frac{P_m}{Q}(w) - \frac{P_m}{Q}(0) = w\frac{\tilde{P}}{Q}(w) \text{ et alors } \|\tilde{F} - \frac{\tilde{P}}{Q}\|_2^2 = \sum_{j=1}^{\infty} |b_j|^2,$$

la dernière égalité provenant de la représentation intégrale de $|F - P_m/Q|$ donné ci-dessus, $\tilde{P}/Q$ étant l'interpolant de $\mathcal{P}_m/Q$ de $\tilde{F}$ aux points $1/\overline{w}_1, \ldots, 1/\overline{w}_m$. En combinant ces 2 chaînes d'inégalités, il est suffisant de démontrer que

$$\|\tilde{F} - \frac{\tilde{P}}{Q}\|_2 = \min_{P \in \mathcal{P}_{m-1}} \|\tilde{F} - \frac{P}{Q}\|_2,$$

c'est-à-dire qu'on connaît le meilleur approximant par rapport à la norme induite par un produit scalaire. Au sens des moindres carrés,

$$\frac{\tilde{P}}{Q}(w) = \sum_{j=1}^{m} \frac{e_j}{w - w_j}$$

est meilleur approximant par rapport à $\|.\|_2$ de $\tilde{F}$ si et seulement si l'erreur $\tilde{F} - \tilde{P}/Q$ est orthogonal à toute fonction dans $\mathcal{P}_{m-1}/Q$ avec base $1/(w - w_j)$, $j = 1, \ldots, m$. Il suffit alors que

$$\Big[\langle \tilde{F}, \frac{1}{w - w_j} \rangle\Big]_{j=1,\ldots,m} = \Big[\langle \frac{1}{w - w_\ell}, \frac{1}{w - w_j} \rangle\Big]_{\ell,j=1,\ldots,m} [e_\ell]_{\ell=1,\ldots,m}.$$

Un rapide calcul des résidus montre que

$$\langle \tilde{F}, \frac{1}{w - w_j} \rangle = -F(1/\overline{w}_j)\overline{w}_j, \;\; \langle \frac{1}{w - w_\ell}, \frac{1}{w - w_j} \rangle = -\frac{1}{\overline{w}_j - w_\ell}/\overline{w}_j$$

et donc notre système est équivalent au fait que $\tilde{P}/Q$ interpôle $\tilde{F}$ aux points $1/\overline{w}_1, \ldots, 1/\overline{w}_m$. Puisque A est normale, d'après [8, Section 4.7], $\|f(A) - p_m(A)/q(A)\| \leqslant \|f - p_m/q\|_{\mathbb{W}(A)} \leqslant \|f - p_m/q\|_{\mathbb{E}} \leqslant 2\|F - P_m/Q\|_{\overline{\mathbb{D}}}$ et la dernière inégalité provient de (1.5). □

Approximation de Padé

Un exemple d'approximants rationnels que l'on retrouve régulièrement dans la littérature est le cas des approximants de Padé.

Définition 1.3.6 (Définition de Frobenius et Padé-Frobenius)**.** *Soit f développable en série entière au voisinage de 0, L, M deux entiers $\geqslant 0$ et soient $p_{L,M}$ et $q_{L,M}$ deux polynômes de degré respectifs L et M, tels que*

$$q_{L,M}(z)f(z) - p_{L,M}(z) = O(z^{L+M+1}).$$

Alors le couple $(p_{L,M}, q_{L,M})$ est appelé approximant de Padé de f d'ordre $[L, M]$.

De tels polynômes $p_{L,M}$ et $q_{L,M}$ peuvent toujours être trouvés : si on développe f sous la forme $f(z) = c_0 + c_1 z + \ldots + c_n z^n + \ldots$, $p_{L,M}(z) = p_0 + p_1 z + \ldots + p_L z^L$ et $q_{L,M}(z) = q_0 + q_1 z + \ldots + q_M z^M$, alors on a le système suivant :

$$\begin{aligned}
c_0 q_0 &= p_0 \\
c_0 q_1 + c_1 q_0 &= p_1 \\
&\vdots \\
c_0 q_M + c_1 q_{M-1} + \ldots c_M q_0 &= p_M \\
c_1 q_M + c_2 q_{M-1} + \ldots + c_{M+1} q_0 &= p_{M+1} \\
&\vdots \\
c_{L-M} q_M + c_{L-M+1} q_{M-1} + \ldots + c_L q_0 &= p_L \\
c_{L-M+1} q_M + c_{L-M} q_{M-1} + \ldots + c_{L+1} q_0 &= 0 \\
&\vdots \\
c_L q_M + c_{L+1} q_{M-1} + \ldots + c_{L+M} q_0 &= 0
\end{aligned}$$

soit en posant $q_0 = 1$,

$$\begin{pmatrix}
c_0 & 0 & \ldots & 0 \\
c_1 & c_0 & \ddots & \vdots \\
\vdots & \vdots & \ddots & 0 \\
\vdots & \vdots & & \ddots \\
c_L & c_{L-1} & \ldots & c_{L-M} \\
c_{L+1} & c_L & \ldots & c_{L-M+1} \\
\vdots & \vdots & \ddots & \vdots \\
c_{L+M} & c_{L+M-1} & \ldots & c_L
\end{pmatrix}
\begin{pmatrix} 1 \\ q_1 \\ \vdots \\ q_M \end{pmatrix}
=
\begin{pmatrix} p_0 \\ p_1 \\ \vdots \\ \vdots \\ p_L \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

On décompose alors ce système linéaire en 2 sous-problèmes : premièrement, en prenant les M dernières lignes de la matrice $(c_{i-j})_{i=0,\ldots,L+M, j=0,\ldots,M}$ avec $c_k = 0$ si $k < 0$, on obtient le système

$$\begin{pmatrix}
c_{L+1} & c_L & \ldots & c_{L-M+1} \\
c_{L+2} & c_{L+1} & \ldots & c_{L-M} \\
\vdots & \vdots & & \vdots \\
c_{L+M} & c_{L+M-1} & \ldots & c_L
\end{pmatrix}
\begin{pmatrix} 1 \\ q_1 \\ \vdots \\ q_M \end{pmatrix}
=
\begin{pmatrix} 0 \\ \vdots \\ \vdots \\ 0 \end{pmatrix}$$

ce qu'on peut ré-écrire sous forme d'un système linéaire avec matrice de Toeplitz :

$$\begin{pmatrix}
c_L & c_{L-1} & \ldots & c_{L-M+1} \\
c_{L+1} & c_L & \ldots & c_{L-M} \\
\vdots & \vdots & & \vdots \\
c_{L+M-1} & c_{L+M-2} & \ldots & c_L
\end{pmatrix}
\begin{pmatrix} q_1 \\ q_2 \\ \vdots \\ q_M \end{pmatrix}
= -
\begin{pmatrix} c_{L+1} \\ c_{L+2} \\ \vdots \\ c_{L+M} \end{pmatrix}$$

et on obtient ainsi les coefficients $q_1, \ldots, q_M$ en résolvant ce système linéaire. Puis on obtient les coefficients $p_0, \ldots, p_L$ par l'égalité

$$\begin{pmatrix}
c_0 & 0 & \ldots & 0 \\
c_1 & c_0 & \ddots & \vdots \\
\vdots & \vdots & \ddots & 0 \\
\vdots & \vdots & & \ddots \\
c_L & c_{L-1} & \ldots & c_{L-M}
\end{pmatrix}
\begin{pmatrix} 1 \\ q_1 \\ \vdots \\ q_M \end{pmatrix}
=
\begin{pmatrix} p_0 \\ p_1 \\ \vdots \\ \vdots \\ p_L \end{pmatrix}.$$

Exemple 1.3.7. *Soit* $f(z) = \log(1-z) = -z - \frac{z^2}{2} - \frac{z^3}{3} - \frac{z^4}{4} - ...$, *alors les approximants de Padé diagonaux de* f *(c'est-à-dire lorsque* $L = M$*) sont donnés par*

$$r_{11} = \frac{-2z}{2-z}, \quad r_{22} = \frac{-6z+3z^2}{6-6z+z^2}, \quad r_{33} = \frac{-60z+60z^2-11z^2}{60-90z+36z^2-3z^3} \ , \dots$$

En revanche, si l'on veut utiliser des fonctions rationnelles pour approcher la fonction f au voisinage de 0, il nous faudrait alors considérer le quotient $\frac{p_{L,M}}{q_{L,M}}$. Or, des polynômes $p_{L,M}$ et $q_{L,M}$ tels que

$$f(z) - \frac{p_{L,M}(z)}{q_{L,M}(z)} = O(z^{M+L+1})$$

n'existent pas toujours. En voici un contre-exemple :

Exemple 1.3.8. *Soit* $f(z) = 1 + z^2$. *Nous cherchons* p *et* q *de degré 1 tous les deux de sorte que*

$$\frac{p(z)}{q(z)} = \frac{p_0 + p_1 z}{q_0 + q_1 z} = 1 + z^2 + O(z^3).$$

Alors $p_0 + p_1 z = q_0 + q_1 z + q_0 z^2 + O(z^3)$, *soit*

$$p_0 = q_0, \ \ p_1 = q_1, \ \ q_0 = 0 \ \ d'où \ \frac{p(z)}{q(z)} = \frac{0 + q_1 z}{0 + q_1 z} = 1.$$

Mais $1 \neq 1 + z^2 + O(z^3)$ *!*

Remarque 1.3.9. *Ainsi, par identification des coefficients dans les deux séries, si* $f(z) = \sum_{j=0}^{\infty} c_j z^j$ *au voisinage de 0, on obtient les deux équations suivantes([57]) :*

$$\forall 0 \leq s \leq M, \quad \sum_{t=0}^{\min\{s,N\}} q_t c_{s-t} = p_s \ \ et \ \ \forall M < s \leq M+N, \quad \sum_{t=0}^{\min\{s,N\}} q_t c_{s-t} = 0$$

soit le système

$$\begin{pmatrix} c_0 & 0 & \dots & 0 \\ c_1 & c_0 & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ c_M & c_{M-1} & \dots & c_0 \\ c_{M+1} & c_M & \dots & c_1 \\ \vdots & \vdots & \ddots & \vdots \\ c_{M+N} & c_{M+N-1} & \dots & c_M \end{pmatrix} \begin{pmatrix} 1 \\ q_1 \\ \vdots \\ q_N \end{pmatrix} = \begin{pmatrix} p_0 \\ p_1 \\ \vdots \\ p_M \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Définition 1.3.10. *[6, Section 1.4] Si des polynômes* $A_{L,M}$ *et* $B_{L,M}$ *de degrés respectifs* L *et* M *peuvent être trouvés de sorte que*

$$\frac{A_{L,M}(z)}{B_{L,M}(z)} = f(z) + O(z^{L+M+1})$$

avec $B_{L,M}(0) = 1$, *alors* $\frac{A_{L,M}(z)}{B_{L,M}(z)}$ *est un approximant de Padé de* f.

Cette définition est alors équivalente à ce que $B_{L,M}(z)f(z) - A_{L,M}(z) = O(z^{L+M+1})$ si $B_{L,M}(0) = 1$. Si $q_{L,M}(0) \neq 0$, alors les deux définitions précédentes sont équivalentes. Mais si $q_{L,M}(0) = 0$ il se peut qu'un approximant de Padé d'ordre $[L, M]$ n'existe pas.

Bien sûr en pratique, on ne peut obtenir $q_{L,M}$ puis calculer $q_{L,M}(0)$ pour finalement dire que l'approximant de Padé n'existe pas. Mais à partir du développement en série formelle d'une fonction, on peut déterminer pour n'importe quel ordre si un approximant de Padé existe ou pas pour cet ordre. En effet, pour f développable en série entière autour de 0, notons $f(z) = \sum_{j=0}^{\infty} c_j z^j$ et définissons

$$C(m,n) = q_{L,M}(0) = \begin{vmatrix} c_{L-M+1} & c_{L-M+2} & \cdots & c_L \\ c_{L-M+2} & c_{L-M+3} & \cdots & c_{L+1} \\ \vdots & \vdots & & \vdots \\ c_L & c_{L+1} & \cdots & c_{L+M+1} \end{vmatrix}.$$

Proposition 1.3.11. *([6, p. 22]) Si $C(m,n) \neq 0$, alors il existe un approximant de Padé d'ordre $[L,M]$. En revanche, si $C(m,n) = 0$, on ne peut pas assurer l'existence d'un approximant de Padé d'ordre $[L,M]$.*

Proposition 1.3.12. *([6, Theorem 1.4.3]) L'approximant de Padé s'il existe est unique.*

Démonstration. Soit $[r,s]$ fixé. Supposons qu'il existe deux approximants de Padé $\frac{p}{q}$ et $\frac{a}{b}$ de degré $[L,M]$. Alors on a

$$\frac{p(z)}{q(z)} - \frac{a(z)}{b(z)} = o(z^{L+M+1})_{z\to 0}$$

En multipliant par $q(z)b(z)$, on obtient $p(z)b(z) - a(z)q(z) = o(z^{L+M+1})$. Or le polynôme de gauche est un polynôme de degré au plus $L+M$ et donc est identiquement nul d'où $\frac{p}{q} = \frac{a}{b}$. Par hypothèse sur les approximants de Padé, p, q et a, b n'ont pas de facteurs en communs respectivement, donc on a l'unicité. □

Maintenant, dans le cas de racine carrée, existe-t-il un approximant de Padé pour tout ordre ? On rappelle que si $f(z) = z^\gamma$ avec $-1 < \gamma < 0$, alors f est une fonction de Markov [8, Section 3.1]. Nous allons voir que pour toute fonction de Markov, tout approximant de Padé de n'importe quel ordre existe (voire Section 4).

Proposition 1.3.13. *Si r est un approximant de Padé de f d'ordre $[m,n]$, alors $\frac{1}{r}$ est un approximant de Padé d'ordre $[L,M]$ de $\frac{1}{f}$.*

Démonstration. Si $f(z) - r(z) = O(z^{L+M+1})$, $f(0), r(0) \neq 0$, alors

$$\frac{1}{f(z)} - \frac{1}{r(z)} = \frac{r(z) - f(z)}{f(z)r(z)} = O(z^{L+M+1})$$

ce qui nous donne la propriété. □

Remarque 1.3.14. *Au-delà de l'existence ou non d'un approximant de Padé, le cas échéant il faudra s'assurer que celui-ci possède une bonne stabilité, et on pourra alors considérer une sous-classe des approximants de Padé "bien conditionnés" sans pôles parasites [12].*

1.3.4 Implémentation des interpolants rationnels

Pour une fonction rationnelle $r \in \mathcal{R}_{l,m}$, il est peu conseillé d'évaluer $r(A)$ en identifiant r comme un quotient de deux polynômes $p \in \mathcal{P}_l$ et $q \in \mathcal{P}_m$, puisque le calcul cumulé des puissances de la matrice A et l'inversion de $q(A)$ peuvent entraîner une perte de précision. En revanche, il existe dans la littérature de nombreuses formes possibles pour une fonction rationnelle. Nous nous intéressons ici principalement à deux

formes de ces fonctions rationnelles : une forme en éléments simples ou sous forme de fraction continue. Si $r = \frac{p}{q} \in \mathcal{R}_{l,m}$ avec $\deg(p) \leqslant l$ et $\deg(q) \leqslant m$, alors en supposant que tous les pôles de r sont simples, ce qui est toujours vrai pour certaines classes de fonctions (voir 4.1.2 ou [44, equations (8)-(9)]), la décomposition en éléments simples de r est donnée par $r(x) = \frac{p}{q}(x) = P(x) + \sum_{j=1}^{m} \frac{a_j}{x-x_j}$ où les x_j sont les pôles de r. On peut alors étendre cette décomposition par application sur un matrice $A \in \mathbb{C}^{n\times n}$

$$r(A) = P(A) + \sum_{j=1}^{m} a_j (A - x_j I)^{-1}.$$

Cette décomposition nécessite le calcul de chaque résolvante $(A - x_j I)^{-1}$. De ce fait, nous gagnons en stabilité par rapport à une implémentation sous forme d'un quotient de polynômes.

Définition 1.3.15. *Soit $r = \frac{p}{q} \in \mathcal{R}_{m,m}$. Alors la décomposition en fraction continue de r est la forme*

$$r(x) = c_0 + \cfrac{x - x_1}{c_1} + \cfrac{x - x_2}{c_2} + \ldots + \cfrac{x - x_t}{c_t}$$

avec $t \leqslant m$.

Remarque 1.3.16. *Les coefficients de cette décomposition sont évalués par récurrence inverse : commençons par $C_m^{(m)} = c_m$ puis pour $k = m-1, m-2, \ldots, 1$,*

$$C_k^{(m)} = c_k + \frac{x - x_{k+1}}{C_{k+1}^{(m)}}.$$

En supposant connus les paramètres x_k, c_k, la fraction rationnelle r sous forme de fraction continue peut alors être évalué avec une complexité de $M(n) + mI(n)$ avec $M(n)$ la complexité pour la multiplication de deux matrices de taille $n \times n$ et $I(n)$ la complexité pour l'inversion d'une matrice de taille $n \times n$.

$$C_k^{(m)}(A) = c_k I + (A - x_{k+1} I).(C_{k+1}^{(m)}(A))^{-1}.$$

1.4 Motivations

Les fonctions de matrices concernant un large spectre de domaines, il s'avère alors essentiel de savoir bien les implémenter ou les approcher. Plusieurs applications justifiant cette étude se trouvent dans [57]. Nous exposons ici quelques exemples d'applications des fonctions de matrices de Toeplitz.

Exemple 1.4.1 (Discrétisation en espace de l'équation des ondes)**.** *Considérons l'équation des ondes :*

$$\begin{cases} \frac{d^2y}{dt^2}(t,x) + \frac{d^2y}{dx^2}(t,x) = 0, \quad t \in \mathbb{R}, x \in [0, L] \\ y(0,x) = y_0(x), \quad \frac{dy}{dt}(0,x) = y_0'(x) \\ y(t,0) = 0 \end{cases}$$

On discrétise alors l'espace $[0, L]$ en N sous-intervalles de longueur $\delta x = L/N$ $[x_i, x_{i+1}]$ pour $i = 0, .., N-1$ avec $0 = x_0 < x_1 < ... < x_N = L$. A t fixé, pour modéliser les dérivées $(\frac{d^2}{dx^2})(y)$ aux points de l'espace (t, x_i), on utilise généralement des formule de quotient de différence fini de la manière suivante :

$$\frac{d}{dx} y(t,x) \approx \frac{y(x + \delta x) - y(x)}{\delta x}, \quad \textit{lorsque } N \to \infty$$

d'où en chaque point du maillage, on a $\frac{d^2y}{dx^2}(t, x_i) \approx \frac{y(t,x_{i-1}) - 2y(t,x_i) + y(t,x_{i+1})}{\delta x^2}$.
En notant $y_i = u(t, x_i), f_i = f(t, x_i)$ pour $i = 0, .., N$, on obtient la formulation suivante :

$$\frac{d^2y_i}{dt^2} + \frac{1}{\delta x^2}(y_{i-1} - 2y_i + y_{i+1}) = f_i$$

avec $y_0 = y_N = 0$ *d'après les conditions initiales.*

D'où en notant $Y = (y_i)_{i=0,..,N} \in \mathbb{C}^{N+1}$*, on peut réécrire le problème sous la forme :*

$$\frac{d^2}{dt^2}(Y(t)) + AY(t) = F(t) \text{ où } A = \frac{1}{\delta x^2}\begin{bmatrix} -2 & 1 & 0 & \dots & 0 \\ 1 & -2 & 1 & \ddots & \vdots \\ 0 & 1 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 1 \\ 0 & \dots & 0 & 1 & -2 \end{bmatrix}$$

On a alors sur $\mathbb{R}^n$ *l'équation suivante :*

$$\begin{cases} \frac{d^2y}{dt^2} + Ay = 0 \quad \text{sur } \mathbb{R} \\ y(0) = y_0 \\ y^{'}(0) = y_0^{'} \end{cases} \tag{1.6}$$

avec $y \in \mathbb{C}^n, A \in \mathbb{C}^{n\times n}, n \in \mathbb{N}^*$*. Alors la solution est donnée par [2]*

$$y(t) = \cos(\sqrt{A}t)y_0 + (\sqrt{A})^{-1}\sin(\sqrt{A}t)y_0^{'}$$

Exemple 1.4.2 (Matrice de covariance)**.** *Dans le domaine de l'analyse stochastique, lorsque les données sont basées sur des observations biométriques ou des séries temporelles apparaissant dans la nature, les statisticiens font souvent face à des matrices de covariance avec une structure Toeplitz symétrique. En effet, considérons un échantillon de p prélèvements* $X_0, X_1, \dots, X_p$ *effectués à intervalles de temps réguliers. Notons* $\varrho(k) = \varrho(X_t, X_{t+k})$ *la corrélation entre* X_t *et une valeur ultérieure* X_{t+k}*. Lorsque le processus stochastique est stationnaire,* $\varrho(X_t, X_{t+k}) = \varrho(X_{t+k}, X_t)$ *et* $\varrho(k) = \varrho(-k)$ *pour tout* $k = 0, \dots, p-1$*. Par conséquent la matrice d'auto-corrélation des p variables de ce processus est une matrice de Toeplitz symétrique.*
Ce type de processus peut être observé dans les études de croissance ou les observations des processus psychologique comme l'apprentissage ou le développement de certaines habilités [74, Section 1.1]. B.N. Mukhexjee en 1988 [74] donne de nombreuses propriétés et applications de l'étude des matrices de Toeplitz liées aux études de problèmes stochastiques.

Exemple 1.4.3 (Détermination des coefficients des approximants de Padé)**.** *Soit f une fonction analytique au voisinage de* 0*. Supposons qu'au voisinage de* 0*, f développable en série entière* $\sum_{j=0}^{\infty} c_j z^j$*. Alors pour tout* $m, n \in \mathbb{N}$ *et en identifiant f à sa série au voisinage de 0, on peut définir les approximants de Padé d'ordre* $[m, l]$ *de f le couple de polynômes* (p, q) *avec* $p(z) = p_0 + p_1 z + \dots + p_m z^m$ *et* $q(x) = q_0 + q_1 z + \dots + q_l z^l$ *satisfaisant*

$$q(z)f(z) - p(z) = \mathcal{O}(z^{m+n+1}) \quad \text{au voisinage de 0.}$$

En identifiant les coefficients dans cette équation, on obtient le système linéaire suivant :

$$\begin{bmatrix} c_0 & 0 & \dots & \dots & 0 \\ c_1 & c_0 & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & 0 \\ c_l & c_{l-1} & \dots & \dots & c_0 \\ \vdots & \vdots & & & \vdots \\ c_{m+l-1} & c_{m+l-2} & \dots & \dots & c_{l-1} \\ c_{m+l} & c_{m+l-1} & \dots & \dots & c_l \end{bmatrix} \cdot \begin{bmatrix} q_0 \\ q_1 \\ \vdots \\ q_l \end{bmatrix} = \begin{bmatrix} p_0 \\ p_1 \\ \vdots \\ \vdots \\ p_m \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

système linéaire avec matrice de Toeplitz rectangulaire.

Exemple 1.4.4 (Moyenne Géométrique)**.** *Soient $A, B \in \mathbb{C}^{n\times n}$ hermitiennes définies positives. La moyenne géométrique $A\#B$ de A et B, définie comme l'unique solution définie positive hermitienne de l'équation $XA^{-1}X = B$, [57, Section 2.10] est donnée par*

$$X = B^{1/2}(B^{-1/2}AB^{-1/2})^{1/2}B^{1/2} = B(B^{-1}A)^{1/2} = (AB^{-1})^{1/2}B.$$

On peut également considérer la définition suivante de la moyenne géométrique

$$X = \exp\Big(\frac{1}{2}\log(A) + \log(B)\Big)$$

où le log est le logarithme principal.

Exemple 1.4.5 (Problème aux valeurs propres)**.** *Une première application des fonctions de matrices peut se trouver dans la résolution des problèmes généralisés aux valeurs propres [57, Section 2.10], c'est-à-dire la résolution du problème $Ax - \lambda Bx$ avec A hermitienne, B hermitienne définie positive. En supposant connue la racine carrée d'une matrice, celui-ci peut se ré-écrire sous la forme*

$$B^{-1/2}AB^{-1/2}(B^{1/2}x) = \lambda(B^{1/2}x) \Leftrightarrow Cy = \lambda y,\ y = B^{1/2}x,\ C = B^{-1/2}AB^{-1/2},$$

le terme de droite dans l'équation correspondant à un problème standard hermitien aux valeurs propres.

Exemple 1.4.6 (Préfiltre)**.** *On peut citer un exemple appliqué au problème d'egaliseur dans les standards modernes de télécommunication [3, Section 7.2.1.2]. A cause des interférences entre les différents utilisateurs, la transmission en sens descendant WCDMA (Wideband Code Division Multiple Access) doit être modifiée et il faut ainsi créer un egaliseur dans la liaison descendante. Une solution est d'employer un egaliseur d'erreur quadratique minimum linéaire, qui fournit une estimation du k^e symbole du m^e utilisateur*

$$y_k(k) = c_m^T(k)\Theta^T(\sigma_y^2\Theta\Theta^T + \sigma_\nu^2 I)^{-1}x(k)$$

où $x(k)$ est le vecteur échantillon reçu, Θ est la matrice de canal, $c_m^T(k)$ est la séquence d'étalement du k^e utilisateur du me symbol, σ_y^2 et σ_ν^2 sont les variables de la séquence transmise et le bruit receveur respectivement. Si l'on se penche sur le préfiltre $(\sigma_y^2\Theta\Theta^T + \sigma_\nu^2 I)^{-1}$, celui-ci correspond à une matrice d'autocorrélation inverse T^{-1} de la séquence $x(k)$. En considérant une approximation de Toeplitz de T et si l'on note s_d la ligne centrale de la matrice S^{-1} racine carrée de T^{-1}, alors les coefficients du préfiltre sont les éléments du vecteur

$$v = S^{-1}s_d.$$

Exemple 1.4.7 (Option pricing avec le modèle de Merton). *Considérons dans un modèle de Merton [73] le problème d'évaluation du prix d'option pour seul actif sous-jacent. La valeur de l'option $\omega(\zeta, t)$ sur un domaine $(-\infty; \infty) \times [0; T]$ avec $T > 0$ satisfait l'équation integro-differentielle partielle*

$$\omega_t = \frac{\nu^2}{2}\omega_{\zeta\zeta} + (r - \lambda K - \frac{\nu^2}{2})\omega_\zeta - (r + \lambda)\omega + \int_{-\infty}^{\infty} \omega(\zeta + \eta, t)\phi(\eta)d\eta \tag{1.7}$$

où T dénote le temps à échéance, $\nu \geqslant 0$ la versatilité, r le taux d'intérêt sans risque, λ l'intensité d'arrivée d'un processus de Poisson, ϕ la distribution normale avec moyenne μ et dérivation standard σ et $K = e^{\mu+\sigma^2/2-1}$. Après troncature du domaine infini $(-\infty; \infty) \times [0; T]$ en $(\zeta_{\min}; \zeta_{\max}) \times [0; T]$, on discrétise l'intervalle $(\zeta_{\min}; \zeta_{\max})$ en $n+1$ sous-intervalles de longueur Δ_ζ. La discrétisation de la partie différentielle de (1.7) par différences centrées puis la discrétisation de la partie intégrale de (1.7) à l'aide de la formule triangulaire nous donne respectivement une matrice tridiagonale de Toeplitz D_n et une matrice de Toeplitz T_n, de sorte que la matrice de Toeplitz réelle non-symétrique $A_n = D_n + \lambda T_n$ est la matrice de la semi-discrétisation du système par rapport à t. Le prix de l'option en $t = T$ nécessite alors le calcul de $\exp(TA_n)\omega_0$ [65] [67, example 3], où ω_0 est la valeur discrétisée de la valeur initiale $\omega_0 = \omega(\zeta, 0) = \max(Ke^\zeta - K, 0)$ avec K le prix d'exercice.

Exemple 1.4.8 (Racine carrée pour le log). *Pour calculer $\log(A)$ l'unique matrice réelle avec valeurs propres sans partie imaginaire dans $[-\pi; \pi]$, appelé logarithme principal avec $A \in \mathbb{C}^{n\times n}$ sans valeurs propres sur l'axe réel négatif, on peut transformer la matrice A à l'aide des propriétés du logarithme et calculer $2^\ell \log(A^{1/2^\ell})$ de sorte à ce que la matrice $A^{1/2^\ell}$ soit proche de la matrice identité [54]. Cette méthode de scaling and squaring va donc nécessiter le calcul cumulées de racines carrées de la matrice A.*

1.5 Conclusion

Nous avons vu que plusieurs méthodes de calcul ou d'approximation polynômiale ou rationnelle des fonctions de matrices existent déjà. Cependant, lorsque la matrice étudiée est de taille $n \times n$, une implémentation sur machine requiert une complexité de $\mathcal{O}(n^3)$ opérations arithmétiques élémentaires. Or lorsque la dimension n est très grande ($n \approx 5000$ par exemple) ce nombre d'opérations devient lourd sur machine et nécessite un certain temps de calcul. De plus, pour un grand nombre d'opérations nécessaires au calcul, le risque de perte de précision augmente et le résultat pourrait être trop éloigné de la fonction de matrice que nous cherchons à implémenter. Au cours de cette thèse, nous avons donc cherché une autre méthode pour calculer des fonctions de matrices de Toeplitz et les implémenter avec un nombre d'opérations réduit. Lorsque nous travaillons avec des matrices de dimension $n \times n$, nous pouvons espérer réduire le nombre d'opérations arithmétiques au minimum à un ordre n^2. Pour ce faire, comme de nombreux auteurs avant nous, nous allons tenter d'exploiter la structure à diagonales et sur/sous diagonales constantes des matrices de Toeplitz.

Chapitre 2

Arithmétique Toeplitz-like

Nous avons vu au chapitre 1 en remarque 1.1.5 que l'ensemble des matrice de Toeplitz muni de la somme et de la multiplication par un scalaire formait un sous-espace vectoriel de $\mathbb{C}^{n\times n}$ mais que cet espace n'était pas stable par multiplication entre elles ou par inversion alors que les applications numériques vont généralement nécessiter des multiplications entre matrices de Toeplitz ou leur inversion. Les algorithmes généraux ne prenant pas en compte la structure particulière des matrices nécessitant un nombre trop important d'opérations, il nous faut donc développer des méthodes d'implémentation pour effectuer ces opérations à moindre coût. Ici, nous ne nous intéressons pas au calcul du produit d'une fonction de matrice $f(T) \in \mathbb{C}^{n\times n}$ avec un vecteur $b \in \mathbb{C}^{n\times 1}$, opération pour laquelle de nombreuses méthodes ont déjà été développées dans la littérature notamment avec l'emploi des sous-espace de Krylov, mais à l'implémentation de la fonction de matrice elle-même ou la résolution des systèmes linéaires de Toeplitz, cette dernière étant sujet de nombreuses études dans la littérature récente.
Récemment, D. Kressner et R. Luce [65] se sont intéressés au cas particulier de la fonction de matrice exponentielle sur les matrices de Toeplitz en développant une arithmétique adaptée à la structure des matrices de Toeplitz et pour laquelle les opérations usuelles telles que la multiplication ou l'inversion sur des matrices de Toeplitz sont effectuées avec une complexité bien inférieure à celle rencontrée avec les algorithmes déjà existants. En nous en inspirant, nous développons dans ce chapitre une arithmétique similaire pour laquelle les opérations usuelles sont effectuées avec une complexité réduite. Plus particulièrement, nous allons voir comment une fonction rationnelle de faible ordre sur une matrice de Toeplitz peut être implémentée avec une complexité $\mathcal{O}(n\log^2 n)$ dans cette arithmétique et $\mathcal{O}(n^2)$ si l'on souhaite reconstruire cette fonction rationnelle de matrice en arithmétique pleine.
Pour ce faire, nous introduisons dans une première partie la notion d'opérateur de déplacement avec matrices shift, en considérant le cas général d'un opérateur de déplacement. Nous montrons que l'image d'une matrice de Toeplitz par cet opérateur possède une structure particulière en sous-section 2.1.1 ainsi que quelques propriétés remarquables sur le comportement de l'opérateur de déplacement sur les opérations usuelles sur les matrices en sous-section 2.1.2. Puis dans une deuxième section, nous introduisons en sous-section 2.2.1 la notion de rang de déplacement qui nous permettra de caractériser un ensemble de matrices contenant les matrices de Toeplitz, que nous appellerons Toeplitz-like vérifiant des propriétés communes. Ce rang de déplacement va nous permettre de construire en sous-section 2.2.2 une nouvelle classe de matrices associées aux matrices Toeplitz-like de taille considérablement réduite, pour lesquelles une méthode de reconstruction de la matrice Toeplitz-like associée est énoncée en sous-section 2.2.3. Ce nouveau type de matrices de taille inférieure permet alors en sous-section 2.2.4 un nouvel algorithme de résolution des systèmes de Toeplitz et Toeplitz-like, fournissant une solution avec une complexité de $\mathcal{O}(n\log^2 n)$. Cet algorithme nous permet ensuite de décrire en troisième section une arithmétique Toeplitz-like, reprenant en sous-section 2.3.2 les opérations usuelles sur les matrices mais cette fois-ci avec une complexité bien inférieure de l'ordre de

$\mathcal{O}(n \log^2 n)$, lorsque n désigne la dimension dans le cas de matrices Toeplitz-like. Un algorithme de compression est énoncé en sous-section 2.3.3 nous permettant de conserver cette faible complexité au cours des opérations.

2.1 Opérateur de déplacement

Nous avons vu qu'un algorithme pour le calcul d'un produit entre une matrice de Toeplitz $T \in \mathbb{C}^{n\times n}$ et un vecteur $x \in \mathbb{C}^n$ avec une complexité de $\mathcal{O}(n \log n)$ au lieu de $\mathcal{O}(n^2)$ habituellement pouvait être employé en exploitant la structure à diagonales et sur/sous-diagonales constantes des matrices de Toeplitz. Cette structure est un cas particulier de la structure plus générale dite de déplacement, que l'on peut identifier dans plusieurs applications numériques [62, Section 1]. Si pour des matrices de Toeplitz $T_1, T_2 \in \mathbb{C}^{n\times n}$, $T_1, T_1 \times T_2$ ne sont pas Toeplitz, celles-ci ne manquent pas pour autant d'une structure particulière. T. Kailath, S.-Y. Kung et M. Morf [61] montrent que pour une matrice de Toeplitz T, son inverse T^{-1} peut s'écrire sous la forme $T^{-1} = \sum_{j=1}^{m} L_j U_j$ avec $m \ll n$, avec L_j matrice de Toeplitz triangulaire inférieure et U_j matrice de Toeplitz triangulaire supérieure. m peut être défini comme le rang de l'image par un opérateur dit de déplacement [61, Lemma 1] donné par

$$T \mapsto T - ZTZ^T$$

où Z est une matrice shift avec $Z_{i,j} = \delta_{i-j,1}$, l'opération ZTZ^T consistant à déplacer les diagonales de la matrice T d'un rang vers le bas, donnant ainsi le nom d'opérateur de déplacement. On dit alors que la matrice est structurée lorsque le rang de l'image par application de l'opérateur est négligeable devant la dimension n. Plus généralement, on appelle opérateur de déplacement tout opérateur de la forme

$$\nabla_{Z_\theta, Z_\alpha}(T) := T - Z_\theta T Z_\alpha^* \tag{2.1}$$

avec $Z_\theta, Z_\alpha \in \mathbb{C}^{n\times n}$ matrice shift avec élément 1 sur la première sous-diagonale, $\theta, \alpha \in \mathbb{C}$ en position $(1, n)$ et zéro ailleurs.

2.1.1 Opérateur sous forme Sylvester

Plutôt que de travailler avec l'opérateur de déplacement sous la forme (2.1), nous allons considérer un opérateur sous la forme d'un opérateur de Sylvester. Avant cela, nous rappelons quelques propriétés sur les équations de Sylvester :

Définition 2.1.1. *Soient $A \in \mathbb{C}^{p\times p}$, $B \in \mathbb{C}^{q\times q}$ et $C \in \mathbb{C}^{p\times q}$ avec $p, q \geqslant 1$. On appelle équation de Sylvester une equation de la forme*

$$AX - XB = C$$

avec $X \in \mathbb{C}^{p\times q}$ comme inconnue.

En 1884, J.J. Sylvester donna une condition d'existence et d'unicité d'une solution à cette équation [83] :

Lemme 2.1.2. *Soient $A \in \mathbb{C}^{p\times p}$, $B \in \mathbb{C}^{q\times q}$ et $C \in \mathbb{C}^{p\times q}$. L'équation de Sylvester $AX - XB = C$ admet une unique solution $X \in \mathbb{C}^{p\times q}$ pour tout membre de droite C si et seulement si $\sigma(A) \cap \sigma(B) = \emptyset$.*

Démonstration. Supposons que $\sigma(A) \cap \sigma(B) \neq \emptyset$ et soient $\lambda \in \sigma(A) \cap \sigma(B)$, $x \in \mathbb{C}^{n\times 1}$ vecteur propre à droite de A pour la valeur propre λ et $y^* \in \mathbb{C}^{1\times n}$ vecteur propre à gauche de B pour la valeur propre λ. Si

on note $X = xy^*$, alors $AX - XB = \lambda xy^* - \lambda xy^* = 0$ et donc pour $C = 0$ nous obtenons 2 solutions $X = 0$ et $X = xy^*$. Inversement, supposons que $\sigma(A) \cap \sigma(B) = \emptyset$. Quitte à employer une décomposition de Schur, on peut supposer que B est triangulaire supérieure et alors la j^{e} colonne de l'équation de $AX - XB = C$ s'écrit

$$(A - b_{j,j}I)X(:,j) = C(:,j) + \sum_{k=1}^{j-1} X(:,k)B(k,j)$$

et comme $B(j,j) \in \sigma(B)$ et que $A - B(j,j)I$ est inversible d'après nos hypothèses, on peut déterminer de façon unique chaque colonne de la matrice X solution de l'équation $AX - XB = C$. □

Considérons à présent l'opérateur de Sylvester :

$$S_{\theta,\gamma} : \begin{cases} \mathbb{C}^{n\times n} & \to \mathbb{C}^{n\times n} \\ X & \mapsto Z_\theta X - XZ_\gamma \end{cases}$$

où pour tout $\theta \in \mathbb{C}$, Z_θ est la matrice à diagonales constantes, appelée matrice shift, donnée par

$$Z_\theta = \begin{pmatrix} 0 & \dots & \dots & 0 & \theta \\ 1 & \ddots & & & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & 0 \end{pmatrix}.$$

D'après le lemme 2.1.2, nous pouvons dégager une condition d'existence et d'unicité de l'équation de Sylvester $Z_\theta X - XZ_\gamma = C$ pour toute matrice $C \in \mathbb{C}^{n\times n}$.

Proposition 2.1.3. *Soient $\theta, \gamma \in \mathbb{C}$. Alors l'opérateur de déplacement $S_{\theta,\gamma}$ est bijectif si et seulement si $\theta \neq \gamma$.*

Démonstration. Pour touts $\theta, \gamma \in \mathbb{C}$ et $n \geqslant 1$, les matrices $Z_\theta \in \mathbb{R}^{n\times n}$ et $Z_\theta \in \mathbb{R}^{n\times n}$ vérifient $Z_\theta^n = \theta I$ et $Z_\gamma^n = \gamma I$. Donc en notant $\theta = |\theta|e^{i\varphi_\theta}$, on obtient que $\sigma(Z_\theta) = \{|\theta|^{1/n} e^{i\nu_\theta} e^{2ij\pi/n}, j = 0, \dots, n-1\}$ avec $\nu_\theta = \varphi_\theta/n$. Par conséquent, $\sigma(Z_\theta) \cap \sigma(Z_\gamma) \neq \emptyset$ si et seulement si il existe $k \in \{0, 1, \dots, n-1\}$ tel que

$$|\theta|^{1/n} e^{i\nu_\theta} e^{2ik\pi/n} = |\gamma|^{1/n} e^{i\nu_\gamma} \Leftrightarrow \begin{cases} |\theta|^{1/n} = |\gamma|^{1/n} \\ \nu_\theta + 2k\pi/n = \nu_\gamma \end{cases} \Leftrightarrow \begin{cases} |\theta| = |\gamma| \\ \varphi_\theta + 2k\pi = \varphi_\gamma \end{cases} \Leftrightarrow \theta = \gamma,$$

et on obtient la proposition énoncée. □

A présent, considérons une matrice $X = (x_{i,j})_{i,j} \in \mathbb{C}^{n\times n}$ avec $n \geqslant 1$ et déterminons l'action de l'opérateur de déplacement sur cette matrice . On a pour tout $\theta, \gamma \in \mathbb{C}$,

$$Z_\theta X = \begin{bmatrix} \theta x_{n,1} & \theta x_{n,2} & \dots & \theta x_{n,n} \\ x_{1,1} & x_{1,2} & \dots & x_{1,n} \\ \vdots & \vdots & & \vdots \\ x_{n-1,1} & x_{n-1,2} & \dots & x_{n-1,n} \end{bmatrix} \text{ et } XZ_\gamma = \begin{bmatrix} x_{1,2} & x_{1,3} & \dots & x_{1,n} & \gamma x_{1,1} \\ x_{2,2} & x_{2,3} & \dots & x_{1,n} & \gamma x_{2,1} \\ \vdots & \vdots & & \vdots & \vdots \\ x_{n,2} & x_{n,3} & \dots & x_{n,n} & \gamma x_{n,1} \end{bmatrix}$$

c'est-à-dire que la multiplication à gauche par la matrice Z_θ décale d'une ligne vers le bas chaque ligne et place en première ligne la dernière ligne multipliée par θ tandis que la multiplication à droite par la matrice

Z_γ décale d'une colonne vers la gauche chaque colonne et place en dernière colonne la première colonne multipliée par γ, ce qui nous donne au final

$$Z_\theta X - X Z_\gamma = \begin{bmatrix} \theta x_{n,1} - x_{1,2} & \theta x_{n,2} - x_{1,3} & \dots & \theta x_{n,n} - \gamma x_{1,1} \\ x_{1,1} - x_{2,2} & x_{1,2} - x_{2,3} & \dots & x_{1,n} - \gamma x_{2,1} \\ \vdots & \vdots & & \vdots \\ x_{n-1,1} - x_{n,2} & x_{n-1,2} - x_{n,3} & \dots & x_{n-1,n} - \gamma x_{n,1} \end{bmatrix}$$

. Lorsque $X = T$ est une matrice de Toeplitz c'est-à-dire

$$T = \begin{bmatrix} t_0 & t_{-1} & \dots & t_{-n+1} \\ t_1 & t_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & t_{-1} \\ t_{n-1} & \dots & t_1 & t_0 \end{bmatrix},$$

la structure à diagonales et sur/sous-diagonales constantes nous permet d'obtenir une structure toute particulière de l'image de T par l'opérateur de déplacement, donnée par

$$S_{\theta,\gamma}(T) = Z_\theta T - T Z_\gamma = \begin{bmatrix} \theta t_{n-1} - t_{-1} & \theta t_{n-2} - t_{-2} & \dots & \theta t_1 - t_{-n+1} & (\theta - \gamma) t_0 \\ 0 & \dots & \dots & 0 & t_{-n+1} - \gamma t_1 \\ \vdots & & & \vdots & \vdots \\ \vdots & & & \vdots & t_{-2} - \gamma t_{n-2} \\ 0 & \dots & \dots & 0 & t_{-1} - \gamma t_{n-1} \end{bmatrix}. \tag{2.2}$$

2.1.2 Opérateur de déplacement sur les opérations de matrices

Nous avons parlé au chapitre 1 de matrices pouvant apparaître sous la forme de somme, produit ou inverse de matrices de Toeplitz. Si celle-ci ne sont pas nécessairement de Toeplitz, on peut alors se demander si elles possèdent des propriétés similaires, en particulier si la matrice image par l'opérateur de déplacement possède un faible rang. Par la définition de cet opérateur et des matrices shift Z_θ, Z_γ, on peut démontrer les propriétés suivantes :

Proposition 2.1.4. *Soit $\theta \neq \gamma$. De la décomposition $Z_\theta = Z + \theta e_1 e_n^*$, nous obtenons*

i. $S_{\theta,\gamma}(sX) = sS_{\theta,\gamma}(X)$ *pour tout* $s \in \mathbb{C}$ *et* $X \in \mathbb{C}^{n\times n}$;
ii. $S_{\theta,\gamma}(X_1 + X_2) = S_{\theta,\gamma}(X_1) + S_{\theta,\gamma}(X_2)$;
iii. $S_{\theta,\gamma}(X^{-1}) = -X^{-1}S_{\theta,\gamma}(X)X^{-1} + (\theta-\gamma)X^{-1}e_1e_n^* + (\theta-\gamma)e_1e_n^*X^{-1}$;
iv. $S_{\theta,\gamma}(X_1X_2) = S_{\theta,\gamma}(X_1)X_2 + X_1S_{\theta,\gamma}(X_2) - (\theta-\gamma)X_1e_1e_n^*X_2$;
v. $\forall k \in \mathbb{N}$, $S_{\theta,\gamma}(X^k) = \sum_{j=0}^{k-1} X^j S_{\theta,\gamma}(X) X^{k-j-1} - (\theta-\gamma)\sum_{j=1}^{k-1} X^j e_1 e_n^* X^{k-j}$;

Démonstration. i. est trivial.
Pour la somme, on remarque que pour toutes matrices $X_1, X_2 \in \mathbb{C}^{n\times n}$, $Z_\theta(X_1+X_2) - (X_1+X_2)Z_\gamma = Z_\theta X_1 - X_1 Z_\gamma + Z_\theta X_2 - X_2 Z_\gamma = S_{\theta,\gamma}(X_1) + S_{\theta,\gamma}(X_2)$.
Pour *iii.*, il suffit de remarquer que $XS(X^{-1})X = XZ_\theta - Z_\gamma X = XZ_\gamma + X(\theta-\gamma)e_1e_n^* - Z_\theta X + (\theta-\gamma)e_1e_n^*X$ $= -S(X) - X(\theta-\gamma)e_1e_n^* + (\theta-\gamma)e_1e_n^*X$ et on obtient *iii.*
Pour *iv.*, on fait apparaître les termes $S(X_1)$ et $S(X_2)$ dans $S(X_1X_2)$, ce qui nous donne

$$\begin{aligned} S(X_1X_2) &= Z_\theta X_1X_2 - X_1X_2Z_\gamma = S(X_1)X_2 + X_1Z_\gamma X_2 - X_1X_2Z_\gamma \\ &= S(X_1)X_2 + X_1S(X_2) - (\theta-\gamma)X_1e_1e_n^*X_2. \end{aligned}$$

Enfin, pour une puissance k de X, lorsque $k = 1$ v. est trivial et pour $k = 2$, on reprend le cas du produit. Pour tout $k \geqslant 3$, par induction

$$\begin{aligned}
S(X^k) &= Z_\theta X^k - X^k Z_\gamma = (Z_\theta X^{k-1} - X^{k-1} Z_\gamma)X + X^{k-1} Z_\gamma X - X^{k-1} X Z_\gamma \\
&= S(X^{k-1})X + X^{k-1}(Z_\gamma X - X Z_\gamma) \\
&= S(X^{k-1})X + X^{k-1}(Z_\theta X - X Z_\gamma) - (\theta - \gamma) X^{k-1} e_1 e_n^* X \\
&= (\sum_{j=0}^{k-2} X^j S(X) X^{k-j-2} - (\theta - \gamma) \sum_{j=1}^{k-2} X^j e_1 e_n^* X^{k-j-1}) X + X^{k-1} S(X) \\
&\qquad - (\theta - \gamma) X^{k-1} e_1 e_n^* X \\
&= \sum_{j=0}^{k-2} X^j S(X) X^{k-j-1} - (\theta - \gamma) \sum_{j=1}^{k-2} X^j e_1 e_n^* X^{k-j} + X^{k-1} S(X) \\
&\qquad - (\theta - \gamma) X^{k-1} e_1 e_n^* X \\
&= \sum_{j=0}^{k-1} X^j S(X) X^{k-j-1} - (\theta - \gamma) \sum_{j=1}^{k-1} X^j e_1 e_n^* X^{k-j}
\end{aligned}$$

ce qui nous donne v. □

Nous considérons à partir de maintenant l'opérateur de déplacement avec $\theta = 1$ et $\gamma = -1$, soit l'opérateur

$$S(X) := S_{1,-1}(X) = Z_1 X - X Z_{-1}.$$

Proposition 2.1.5. *Pour tout $n \geqslant 1$, $Z_1 Z_1^* = Z_{-1} Z_{-1}^* = I$, $Z_1^n = I$ et $Z_{-1}^n = -I$.*

Démonstration. On vérifie ces propriétés par calcul matriciel. □

2.2 Rang de déplacement et générateurs : les matrices Toeplitz-like

Nous venons de voir que l'image par l'opérateur de déplacement d'une matrice de Toeplitz possédait une structure particulière avec une colonne et une ligne non nulles. Il nous reste alors à voir comment exploiter cette structure. D. Kressner et R. Luce, en considérant l'opérateur de déplacement sous la forme Stein $T \mapsto T - ZTZ$ avec $Z = Z_0$ matrice shift [65], définissent un ensemble de matrices à partir du rang de l'image par cet opérateur de déplacement. Lorsque celui-ci est petit, l'image est alors décomposée comme un produit de matrices de dimension inférieure, permettant par la suite d'obtenir une arithmétique particulière sur laquelle les différentes opérations telles que la somme, le produit ou l'inversion peuvent être effectuées avec une complexité réduite par rapport à l'arithmétique pleine. Nous allons dans cette section nous inspirer de cette démarche, mais en considérant notre opérateur de déplacement $T \mapsto Z_1 T - T Z_{-1}$ qui de part la structure particulière des matrices Z_1 et Z_{-1} nous permettra d'obtenir un algorithme de reconstruction de la matrice en arithmétique pleine. Nous commençons dans une première sous-section par définir le rang de déplacement d'une matrice et l'étudions dans le cas des matrices de Toeplitz. Ce rang de déplacement est alors employé en sous-section 2.2.2 pour définir une nouvelle famille de matrices pour laquelle nous démontrons ensuite en sous-section 2.2.3 qu'il existe un algorithme de reconstruction nous permettant de ré-obtenir la matrice pleine associée. En sous-section 2.2.4, nous montrons comment cette nouvelle famille de

matrices est employée dans des algorithmes de résolution de systèmes Toeplitz à complexité d'ordre $\mathcal{O}(n^2)$ ou $\mathcal{O}(n \log^2 n)$.

2.2.1 Rang de déplacement

Nous venons de voir que l'opérateur de déplacement appliqué à une matrice de Toeplitz donne une matrice image avec structure particulière. On peut alors tenter d'exploiter cette structure en observant en particulier le rang des matrices images par cet opérateur.

Définition 2.2.1. *[61, Lemma 1] Soit $X \in \mathbb{C}^{n\times n}$. Alors le rang de la matrice image $S(X)$ est appelé le rang de déplacement de X, noté $\rho = \rho(X) := \text{rang}(S(X))$.*

Exemple 2.2.2.

- *La matrice identité, cas particulier des matrices de Toeplitz, est de rang de déplacement 1.*
- *Pour toute matrice $X \in \mathbb{C}^{n\times n}$ de rang de déplacement ρ et pour tout scalaire $s \neq 0$, $s \in \mathbb{C}$, $\rho(sX) = \rho(X)$.*

Proposition 2.2.3. *Soit $T \in \mathbb{C}^{n\times n}$ matrice de Toeplitz. Alors T est de rang de déplacement inférieur ou égal à 2.*

Démonstration. D'après (2.2), pour toute matrice de Toeplitz $T = (t_{i-j})_{i,j=1,\dots,n}$, $S(T)$ comporte au plus une ligne et une colonne non nulles. Par la méthode du pivot de Gauss, on réduit cette matrice à au plus 2 lignes ou 2 colonnes non nulles. □

Remarque 2.2.4. *Dans le cas où $t_0 \neq 0$ et $t_{i,j} = 0$ pour tout $i \neq j$, $T = t_0 I$ et $S(T)$ est constitué d'un seul élément non-nul et donc T est de rang de déplacement 1. Si T est une matrice circulante non nulle, $S(T)$ est composé d'une seule colonne non nulle et peut se réduire à un seul élément non nul par pivot de Gauss et est donc $S(T)$ de rang de déplacement 1. Si T est anti-circulante, alors $S(T)$ est composée d'une seule ligne non-nulle et est donc de rang de déplacement 1.*

Lorsque $T \in \mathbb{C}^{n\times n}$ est de Toeplitz, son inverse ou son produit avec une autre matrice de Toeplitz n'est pas nécessairement une matrice de Toeplitz. Cependant ces matrices possèdent des caractéristiques communes en termes du rang de déplacement :

Définition 2.2.5. *Soient $n \in \mathbb{N}^*$ et $T \in \mathbb{C}^{n\times n}$. Alors T est dite Toeplitz-like si $\rho(T)$ est négligeable devant la dimension n. En particulier, toute matrice de Toeplitz est une matrice Toeplitz-like [61],[41, Section 3].*

Exemple 2.2.6. *L'introduction de cette famille nous permet d'intégrer les matrices de Toeplitz dans une famille plus grande de matrices pouvant être obtenues par opérations élémentaires sur celle-ci : en effet, supposons que $T \in \mathbb{C}^{n\times n}$ est une matrice Toeplitz-like avec $n \gg 1$. Alors d'après le point iii. de la proposition 2.1.4, $\rho(T^{-1}) \leqslant \rho(T) + \text{rang}(T^{-1}e_1e_n^*) + \text{rang}(e_1e_n^*T^{-1}) = \rho(T) + 2 \ll n$ et donc T^{-1} est une matrice Toeplitz-like. De même lorsque T_1 et T_2 sont 2 matrices Toeplitz-like, alors d'après le point iv. de la proposition 2.1.4, $\rho(T_1T_2) \leqslant \rho(T_1) + \rho(T_2) + \text{rang}(T_1e_1e_n^*T_2)$, or $\text{rang}(T_1e_1e_n^*T_2) \leqslant 1$, d'où $\rho(T_1T_2) \leqslant \rho(T_1) + \rho(T_2) + 1 \ll n$ puisque $\rho(T_1), \rho(T_2) \ll n$ et donc le produit de matrice T_1T_2 est une matrice Toeplitz-like. On vérifie également rapidement que $T_1 + T_2$ est Toeplitz-like dès que T_1 et T_2 sont Toeplitz-like.*

2.2.2 Construction des générateurs

Nous venons de voir que les matrices de Toeplitz et Toeplitz-like ont la particularité d'avoir un rang de déplacement faible par rapport à la dimension. Or pour toute matrice $A \in \mathbb{C}^{n\times n}$, on peut décomposer la matrice A en un produit de 2 matrices $G \cdot B^*$ où $G, B \in \mathbb{C}^{n\times \text{rang}(A)}$ et donc si A est Toeplitz-like, c'est-à-dire si $\rho(A) \ll n$, nous pouvons donc décomposer son image par l'opérateur de déplacement sous la forme d'un produit $G.B^*$ où G et B sont de dimension $n \times \rho(A)$. En particulier lorsque A est une matrice de Toeplitz, l'image par l'opérateur de déplacement peut se décomposer en produit de 2 matrices de taille $n \times 2$. Par conséquent, travailler avec ce couple de matrices va nous permettre de construire les opérations usuelles sur les matrices avec un coût réduit.

Définition 2.2.7. *Soit $X \in \mathbb{C}^{n\times n}$. Alors il existe deux matrices $G, B \in \mathbb{C}^{n\times\rho}$ où $\rho = \rho(X)$ telles que $Z_1X - XZ_{-1} = GB^*$, appelés générateurs de X (associés à l'opérateur S).*

Exemple 2.2.8. *Soit $T = (t_{i-j})_{i,j=1\ldots,n} \in \mathbb{C}^{n\times n}$ une matrice de Toeplitz. Alors d'après* (2.2)*, il existe deux matrices $G, B \in \mathbb{C}^{n\times 2}$ telles que $Z_1T - TZ_{-1} = GB^*$ données par*

$$G = \begin{bmatrix} 1 & 2t_0 \\ 0 & t_{-n+1} + t_1 \\ \vdots & \vdots \\ 0 & t_{-1} + t_{n-1} \end{bmatrix}, \quad B = \begin{bmatrix} t^*_{n-1} - t^*_{-1} & 0 \\ \vdots & \vdots \\ t^*_1 - t^*_{-n+1} & 0 \\ 0 & 1 \end{bmatrix}.$$

Notons que, contrairement à la matrice image $S(X)$, le couple de générateurs n'est pas unique puisque pour toute matrice $J \in \mathbb{C}^{\rho\times\rho}$ inversible, le couple de matrices (GJ, BJ^{-*}) est également un couple de générateurs dès que $Z_1X - XZ_{-1} = GB^*$. Notons de plus que pour une matrice Toeplitz-like T avec générateurs (G, B), nous pouvons directement en déduire les générateurs de la matrice T^{-*} puisque $S(T^{-*}) = -(Z_1T^{-*}B)(Z^*_{-1}T^{-1}G)^*$. Ces générateurs peuvent alors être obtenus par résolution de 2 systèmes linéaires avec T^* et T qui, nous allons le voir plus loin dans ce chapitre, peut être effectuée à faible coût.

Après cette observation, du fait de la taille réduite des générateurs par rapport à la matrice de départ, nous pouvons envisager de travailler avec les générateurs d'une matrice plutôt qu'avec la matrice elle-même. De plus, à partir de tout couple de générateurs $(G, B) \in \mathbb{C}^{n\times\rho}$, on peut retrouver la matrice $X \in \mathbb{C}^{n\times n}$ telle que $Z_1X - XZ_{-1} = GB^*$, dont l'existence est assurée par le lemme 2.1.2 et le choix des matrices Z_1, Z_{-1}.

Proposition 2.2.9. *Soit $X \in \mathbb{C}^{n\times n}$ et $G, B \in \mathbb{C}^{n\times\rho}$ les générateurs associés à X avec $\rho = \rho(X)$. Alors*

$$X = \frac{1}{2}\sum_{j=0}^{n-1} Z_1^j(Z_1X - XZ_{-1})Z_{-1}^{n-j-1},$$

soit

$$X = \frac{1}{2}\sum_{j=0}^{n-1} Z_1^j GB^* Z_{-1}^{n-j-1} \tag{2.3}$$

Démonstration. Pour tout $n \geqslant 1$, on sait que $Z_1^n = I$ et $Z_{-1}^n = -I$, d'où $2X = Z_1^nX - XZ_{-1}^n$. On peut alors

vérifier que

$$\begin{aligned}\sum_{j=0}^{n-1} Z_1^j(Z_1X - XZ_{-1})Z_{-1}^{n-j-1} &= \sum_{j=0}^{n-1} Z_1^{j+1}XZ_{-1}^{n-j-1} - \sum_{j=0}^{n-1} Z_1^jXZ_{-1}^{n-j} \\ &= \sum_{j=1}^{n} Z_1^jXZ_{-1}^{n-j} - \sum_{j=0}^{n-1} Z_1^jXZ_{-1}^{n-j} \\ &= X + \sum_{j=1}^{n-1} Z_1^jXZ_{-1}^{n-j} - \sum_{j=1}^{n-1} Z_1^jXZ_{-1}^{n-j} - (-1)X \\ &= 2X.\end{aligned}$$

□

Corollaire 2.2.10. *Pour toute matrice $X \in \mathbb{C}^{n\times n}$ et toute norme unitairement invariante $\|.\|_*$,*

$$\frac{1}{2}\|S_{1,-1}(X)\|_* \leqslant \|X\|_* \leqslant \frac{n}{2}\|S_{1,-1}(X)\|_*. \tag{2.4}$$

Démonstration. Puisque Z_1 et Z_{-1} sont unitaires, Z_1^k et Z_{-1}^{n-k-1} le sont aussi pour tout $k = 0, \ldots, n-1$. D'où $\forall X \in \mathbb{C}^{n\times n}$, on a par les propriétés d'une norme unitairement invariante,

$$\|Z_1^kX\|_* = \|XZ_{-1}^{n-k-1}\|_* = \|Z_1^kXZ_{-1}^{n-k-1}\|_* = \|X\|_*, \forall k = 0, \ldots, n-1.$$

Pour la première inégalité, il suffit de calculer : $\|S(X)\|_* = \|Z_1X - XZ_{-1}\|_* \leq \|Z_1X\|_* + \|XZ_{-1}\|_* = 2\|X\|_*$. Puis d'après la formule de reconstruction précédente, on a

$$\begin{aligned}\|X\|_* = \frac{1}{2}\|\sum_{k=0}^{n-1} Z_1^k(Z_1X - XZ_{-1})Z_{-1}^{n-k-1}\|_* &\leq \frac{1}{2}\sum_{j=0}^{n-1}\|Z_1^k(Z_1X - XZ_{-1})Z_{-1}^{n-k-1}\|_* \\ &= \frac{n}{2}\|S(X)\|_*\end{aligned}$$

□

2.2.3 Reconstruction d'une matrice Toeplitz-like à partir de ses générateurs

La reconstruction de X à partir de ses générateurs à l'aide de la formule (2.3) nous donnerait à première vue une complexité de l'ordre n^4, que l'on pourrait réduire à $\mathcal{O}(\rho(X)n^3)$ opérations élémentaires en considérant la formule (2.3) comme $\rho \times n$ produits matrices-vecteurs. Cependant, à l'aide de la structure particulière des matrices Z_1 et Z_{-1}, la formule de reconstruction (2.3) nous permet d'écrire la matrice X comme somme de $\rho(X)$ produits entre une matrice circulante et une matrice anti-circulante. En effet, notons $X \in \mathbb{C}^{n\times n}$, $G = (g_{i,j})_{\substack{i=1,\ldots,n\\ j=1,\ldots,\rho}}$ et $B = (b_{i,j})_{\substack{i=1,\ldots,n\\ j=1,\ldots,\rho}}$ les générateurs associés à X et g_l, b_l les colonnes respectives de G et B avec $l = 1, \ldots, \rho$. Alors

$$2X = \sum_{j=0}^{n-1} Z_1^j G B^* Z_{-1}^{n-j-1} = -\sum_{j=0}^{n-1} Z_1^j (\sum_{l=1}^{\rho} g_l b_l^*)(Z_{-1}^{j+1})^* = -\sum_{l=1}^{\rho}\sum_{j=0}^{n-1} (Z_1^j g_l)(Z_{-1}^{j+1} b_l)^*$$

$$= -\sum_{l=1}^{\rho}\sum_{j=0}^{n-1} \begin{bmatrix} g_{n-j+1,l} \\ \vdots \\ g_{n,l} \\ g_{1,l} \\ \vdots \\ g_{n-j,l} \end{bmatrix} \begin{bmatrix} -b_{n-j,l}^* & \dots & -b_{n,l}^* & b_{1,l}^* & \dots & b_{n-j-1,l}^* \end{bmatrix} \tag{2.5}$$

$$= -\sum_{l=1}^{\rho} \begin{bmatrix} g_{1,l} & g_{n,l} & \dots & g_{2,l} \\ g_{2,l} & g_{1,l} & \dots & g_{3,l} \\ \vdots & \vdots & \ddots & \vdots \\ g_{n,l} & g_{n-1,l} & \dots & g_{1,l} \end{bmatrix} \begin{bmatrix} -b_{n,l}^* & b_{1,l}^* & \dots & b_{n-1,l}^* \\ -b_{n-1,l}^* & -b_{n,l}^* & \dots & b_{n-2,l}^* \\ \vdots & \vdots & \ddots & \vdots \\ -b_{1,l}^* & -b_{2,l}^* & \dots & -b_{n,l}^* \end{bmatrix} = -\sum_{l=1}^{\rho} A_l \ , \tag{2.6}$$

où $A_l = \begin{bmatrix} g_{1,l} & g_{n,l} & \dots & g_{2,l} \\ g_{2,l} & g_{1,l} & \dots & g_{3,l} \\ \vdots & \vdots & \ddots & \vdots \\ g_{n,l} & g_{n-1,l} & \dots & g_{1,l} \end{bmatrix} \begin{bmatrix} -b_{n,l}^* & b_{1,l}^* & \dots & b_{n-1,l}^* \\ -b_{n-1,l}^* & -b_{n,l}^* & \dots & b_{n-2,l}^* \\ \vdots & \vdots & \ddots & \vdots \\ -b_{1,l}^* & -b_{2,l}^* & \dots & -b_{n,l}^* \end{bmatrix}$.

Lemme 2.2.11. *Pour $l = 1, \dots, \rho(X)$, A_l est construit avec une complexité de $\mathcal{O}(n^2)$ opérations arithmétiques.*

Démonstration. Définissons $\tilde{g}_{j,l} = g_{j,l}$ si $1 \leqslant j \leqslant n$, $\tilde{g}_{j,l} = g_{j-n,l}$ si $n \leqslant j \leqslant 2n$, $\tilde{g}_{j,l} = g_{j+n,l}$ si $-n \leqslant j \leqslant 0$. De même pour les coefficients $b_{j,l}$. On voit alors que $\forall i = 2, \dots, n, \ \forall k = 1, \dots, n-1$,

$$\begin{aligned}(A_l)_{i,k} &= \sum_{j=0}^{n-1} \tilde{g}_{i-j,l}\tilde{b}_{k-1-j,l} \\ &= (A_l)_{i+1,k+1}\tilde{g}_{i+1-n,l}\tilde{b}_{(k+1)-1-n,l} - \tilde{g}_{i+1,l}\tilde{b}_{(k+1)-2,l}\end{aligned}$$

D'où pour tout $k = 1, \dots, n-1$ et $\forall i = 2, \dots, n$,

$$(A_l)_{i+1,k+1} = (A_l)_{i,k} + \tilde{g}_{i+1,l}\tilde{b}_{(k+1)-2,l} + \tilde{g}_{i+1-n,l}\tilde{b}_{(k+1)-1-n,l}$$

Ainsi pour construire A_l pour $l = 1, \dots, \rho$ à partir des générateurs, il nous suffit de construire les coefficients de la première colonne et première ligne en effectuant n produits de coefficients $\tilde{g}_{i-j,l}\tilde{b}_{k-1-j,l}$ pour les $2n-1$ coefficients de la première ligne et première colonne de la matrice A_l, soit un premier coût de $\mathcal{O}(n^2)$ opérations élémentaires, puis le calcul des coefficients $(A_l)_{i,j}$ pour $i, j = 2, \dots, n$ va nécessiter un coût de l'ordre de $\mathcal{O}(n^2)$ par addition de 2 produits de coefficients. Chaque matrice A_l pour $l = 1, \dots, \rho(X)$ peut être donc construite en un $\mathcal{O}(n^2)$ opérations élémentaires. □

Corollaire 2.2.12. *Soient $X \in \mathbb{C}^{n\times n}$ une matrice avec rang de déplacement $\rho(X)$ avec $\rho(X) \ll n$ et $G, B \in \mathbb{C}^{n\times\rho(X)}$ ses générateurs associés. Alors X peut être construite avec une complexité de $\mathcal{O}(\rho(X)n^2)$ à partir de G et B.*

Démonstration. D'après le lemme 2.2.11, le calcul de A_l pour $l = 1, \ldots, \rho(X)$ possède une complexité de $\mathcal{O}(n^2)$. Par conséquent, il nous faut effectuer $\rho(X)$ fois ce calcul, ce qui nous donne une complexité globale pour la reconstruction de X de $\mathcal{O}(\rho n^2)$ et $\mathcal{O}((\rho-1)n^2)$ additions pour la somme. □

Ainsi, si nous possédons les générateurs associés à une fonction de matrice, nous obtenons cette fonction de matrice en arithmétique pleine avec un nombre d'opérations de l'ordre de n^2 lorsque celle-ci admet un faible rang de déplacement. Nous allons donc dans la suite étudier ce rang de déplacement pour des fonctions de matrices. Plus précisément, nous allons voir comment se construisent les générateurs d'une fonction de matrice à partir de ceux de la matrice de départ.

2.2.4 Résolution de systèmes Toeplitz-like

Les systèmes de Toeplitz $Tx = b$ avec $T \in \mathbb{C}^{n\times n}$ et $x, b \in \mathbb{C}^n$ survenant dans plusieurs applications numériques, il apparaît essentiel de pouvoir les résoudre le plus efficacement possible. La première méthode de résolution de ces systèmes linéaires de Toeplitz fut l'algorithme de Levinson [68], sous sa forme la plus connue appelée algorithme de Levinson-Durbin [63, Section 1.2], qui résout le système linéaire en fournissant les facteurs d'une factorisation triangulaire de l'inverse T^{-1}, tandis que l'algorithme de Schur [63, Sections 1.6.3, 1.7.6] calcule celle de la matrice T elle-même pour résoudre $Tx = b$. Ces deux premiers algorithmes ont permis de réduire la complexité à $\mathcal{O}(n^2)$ opérations arithmétiques en exploitant la structure Toeplitz de la matrice, contre $\mathcal{O}(n^3)$ habituellement. On parle alors d'algorithmes rapides. Cependant, la stabilité pour l'algorithme de Levinson-Durbin n'est pas assurée pour toute matrice de Toeplitz et l'algorithme de Schur est également instable. Heinig [50, Section 7] propose de transformer les systèmes de Toeplitz en systèmes Cauchy-like, c'est à dire en systèmes de la forme $D\hat{T} - \hat{T}\tilde{D} = CH^*$ avec $D, \tilde{D}$ deux matrices diagonales et $C, H \in \mathbb{C}^{n\times r}$ où $r \ll n$ (la matrice $\hat{T}$ est alors dite Cauchy-like), pour lesquels on peut introduire un pivot partiel dans la résolution [63, Section 1.13.1], mais ne suffit pas à garantir la stabilité numérique pour toute matrice de Toeplitz. En revanche, on trouve dans la littérature plus récente un algorithme, appelé algorithme GKO après I. Gohberg, T. Kailath et V. Olshevsky [41], pour la résolution d'un système linéaire dont les coefficients formeraient une matrice de Toeplitz T satisfaisant l'équation de Sylvester

$$Z_1 T - T Z_{-1} = GB^*. \tag{2.7}$$

Cet algorithme se décompose en 2 étapes ; la première consiste à transformer le système Toeplitz-like (2.7) en un système Cauchy-like comme pour l'algorithme de Schur, à l'aide d'une Fast Fourier Transform (FFT) : si $Z_1 T - T_{-1} Z_{-1} = GB^*$, alors

$$\begin{aligned} Z_1 T - T Z_{-1} = GB^* &\Leftrightarrow F_n \Delta F_n^* T - T\widehat{F}_n \widehat{\Delta} \widehat{F}_n^* = GB^* \\ &\Leftrightarrow \Delta F_n^* T \widehat{F}_n - F_n^* T \widehat{F}_n \widehat{\Delta} = F_n^* GB^* \widehat{F}_n \\ &\Leftrightarrow \Delta \widehat{T} - \widehat{T}\widehat{\Delta} = F_n^* GB^* \widehat{F}_n = \tilde{G}\tilde{B}^* \end{aligned}$$

où Δ et $\widehat{\Delta}$ sont donnés par la proposition 1.1.3. En posant $\widehat{T} = F_n^* T \widehat{F}_n, \tilde{G} = F_n^* G$, $\tilde{B} = \widehat{F}_n^* B$, et

$$\widehat{T} = \Big(\frac{\tilde{G}_j \tilde{B}_k^*}{\lambda_j - \tilde{\lambda}_k}\Big)_{j,k=1,..,n} \tag{2.8}$$

est Cauchy-like. Dans une deuxième étape, on note

$$\hat{T} = \begin{bmatrix} d & u \\ l & T_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \frac{1}{d} l & I \end{bmatrix} \begin{bmatrix} d & u \\ 0 & T^{(2)} \end{bmatrix} \tag{2.9}$$

où $\hat{T}^{(2)}$ est le complément de Schur de T_2 dans $\hat{T}$ c'est-à-dire que $T^{(2)} = T_2 - \frac{1}{d} lu$.

Alors $T^{(2)}$ vérifie

$$\Delta_{2:n,2:n} T^{(2)} - T^{(2)} \widehat{\Delta}_{2:n,2:n} = \tilde{G}\tilde{B}^*$$

où

$$\begin{bmatrix} 0 \\ \tilde{G} \end{bmatrix} = G - \begin{bmatrix} 1 \\ \frac{1}{d} l \end{bmatrix} G_{1,:} \text{ et } \begin{bmatrix} 0 \\ \tilde{B} \end{bmatrix} = B - \begin{bmatrix} 1 \\ \frac{1}{d^*} u^* \end{bmatrix} B_{1,:}.$$

Ensuite on répète l'opération sur $T^{(2)}$ et ainsi de suite. On obtient ainsi la décomposition suivante :

$$\begin{aligned}
\hat{T} = \begin{pmatrix} d & u \\ l & T^{(2)} \end{pmatrix} &= \begin{pmatrix} 1 & 0 \\ \frac{1}{d} l & I \end{pmatrix} \begin{pmatrix} d & u \\ 0 & T^{(2)} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \frac{1}{d_1} l_1 & I \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{1}{d_2} l_2 & I \end{pmatrix} \begin{pmatrix} d & & u \\ 0 & d_2 & u_2 \\ 0 & 0 & T^{(3)} \end{pmatrix} \\
&= \dots \begin{pmatrix} 1 & 0 \\ \frac{1}{d_1} & I \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{1}{d_2} l_2 & I \end{pmatrix} \dots \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & 1 & 0 \\ 0 & \dots & 0 & \frac{1}{d_n} l_n \end{pmatrix} \begin{pmatrix} d_1 & & & u \\ 0 & \ddots & & \\ \vdots & & d_{n-1} & u_{n-1} \\ 0 & \dots & \dots & 0 \quad u_n \end{pmatrix} \\
&= LU.
\end{aligned}$$

Afin d'éviter une croissance trop importante dans les coefficients de la factorisation et apporter une stabilité, on peut déterminer le pivot, c'est-à-dire l'élément de plus grande amplitude dans la première colonne en position $(k, 1)$ puis effectuer un échange de lignes dans $\hat{T}$ en multipliant $\hat{T}$ à gauche par la matrice de permutation $P(1, k)$. On l'appelle décomposition LU avec pivot partiel.

Au cours des étapes on peut alors ré-écrire

$$\begin{aligned}
\Delta\hat{T} - \hat{T}\widehat{\Delta} = \tilde{G}\tilde{B}^* \\
&\Leftrightarrow (P(1,k)\Delta P(1,k_1))(P(1,k_1)\hat{T}) - (P(1,k_1)\hat{T})\widehat{\Delta} = (P(1,k_1)\tilde{G})\tilde{B}^* \\
&\Leftrightarrow (P(2,k_2)P(1,k_1))\Delta P(1,k_1)P(2,k_2)(P(2,k_2)P(1,k_1)\hat{T}) \\
&\qquad\qquad - (P(2,k_2)P(1,k_1)\hat{T})\widehat{\Delta} = (P(2,k_2)P(1,k_1)\tilde{G})\tilde{B}^* \\
&\qquad\qquad \vdots \\
&\Leftrightarrow (P(n-1,k_{n-1})..P(1,k_1)\Delta P(1,k_1)..P(n-1,k_{n-1}))(P(n-1,k_{n-1})..P(1,k_1)\hat{T}) \\
&\qquad\qquad - (P(n-1,k_{n-1})..P(1,k_1)\hat{T})\widehat{\Delta} = (P(n-1,k_{n-1})..P(1,k_1)\tilde{G})\tilde{B}^* \\
&\Leftrightarrow (P\Delta P^{-1})(P\hat{T}) - (P\hat{T})\widehat{\Delta} = (P\tilde{G})\tilde{B}^*
\end{aligned}$$

où $P = P(n-1, k_{n-1})P(n-2, k_{n-2}) \dots P(2, k_2)P(1, k_1)$ est une matrice de permutation.

Lemme 2.2.13. *Tout système de n équations linéaires à n inconnues avec comme matrice de coefficients une matrice Toeplitz-like de rang de déplacement ρ peut être résolu en une complexité de $\mathcal{O}(\rho n^2)$ opérations élémentaires à l'aide de l'algorithme GKO.*

Algorithm 2 Algorithme GKO

Require: $\hat{T} \in \mathbb{C}^{n\times n}$ Cauchy-like , $\tilde{G}, \tilde{B}$ générateurs associés
Ensure: Factorisation LU de $\hat{T}$
1: $L, U \leftarrow 0^{n\times n}, P \leftarrow I$
2: **for** $k = 1, .., n-1$ **do**
3: $L_{k:n,k} \leftarrow (\text{diag}(\lambda_k, .., \lambda_n) - \tilde{\lambda}_k I)^{-1} \tilde{G}_{k:n,:} \tilde{B}^*_{k,:}$
4: on sélectionne l'élément de plus grande amplitude : $k_{max} = \max_{j=k,..,n} |\hat{T}_{j,k}|$
5: **if** $k_{max} > k$ **then**
6: $P = PP(k, k_{max})$
7: $\hat{T} \leftarrow P(k, k_{max})\hat{T}, \Delta \leftarrow P(k, k_{max})\Delta P(k, k_{max}), L \leftarrow P(k, k_{max})L$
8: **end if**
9: $U_{k,k} \leftarrow L_{k,k}, U_{k,k+1:n} \leftarrow \tilde{G}_{k,:} \tilde{B}^*_{:,k+1:n} (\lambda_k I - \text{diag}(\tilde{\lambda}_{k+1}, .., \tilde{\lambda}_n))^{-1}$
10: $L_{k,k} \leftarrow 1$, $L_{k+1:n,k} \leftarrow L_{k+1:n,k}/U_{k,k}$.
11: $\tilde{G}_{k+1:n,:} \leftarrow \tilde{G}_{k+1:n,:} - L_{k+1:n,k} \tilde{G}_{k,:}$
12: $\tilde{B}_{k+1:n,:} \leftarrow \tilde{B}_{k+1:n,:} - U^*_{k,k+1:n} \tilde{B}_{k+1,:}/U_{k,k}$
13: **end for**

Remarque 2.2.14. *L'algorithme GKO constitue un algorithme efficace pour la résolution des système Toeplitz vérifiant la même stabilité que l'élimination de Gauss avec pivot [41]. Cet algorithme est implémenté dans une fonction* ***tsolve****.*

L'algorithme GKO est donc un algorithme rapide pour la résolution des systèmes linéaires Toeplitz et Toeplitz-like. Plus récemment, cet algorithme a été amélioré afin d'obtenir un algorithme avec complexité $\mathcal{O}(n \log^2 n)$: on parle alors d'algorithme super rapide. A partir du système Cauchy-like

$$\Delta\widehat{T} - \widehat{T}\widehat{\Delta} = F_n^* G B^* \widehat{F}_n = \tilde{G}\tilde{B}^*,$$

S. Massei, L. Robol et D. Kressner [70, Section 5.1] compressent la matrice Cauchy-like

$$\widehat{T} = \Big(\frac{\tilde{G}_j \tilde{B}^*_k}{\lambda_j - \tilde{\lambda}_k}\Big)_{j,k=1,..,n} \tag{2.10}$$

au format HSS (Hierarchically Semi-Separable) et à l'aide de la FFT, implémentent les produits matrice-vecteur $\widehat{T}x$ et $\widehat{T}^*x$ adaptés au format HSS, le tout permettant de résoudre le système linéaire $Tx = b$ par passage au format HSS et réduisant la complexité de la résolution de $Tx = b$ à $\mathcal{O}(\rho^2 n \log^2 n)$ [49, Section 3.6], [70, Section 3.4 et 5.1]. Cette procédure est implémentée dans une fonction MATLAB **toeplitz_solve** disponible dans le package ***hm-toolbox*** que l'on retrouve sur *https ://github.com/numpi/hm-toolbox.*

Lemme 2.2.15. *Tout système d'équations linéaires avec comme matrice de coefficients une matrice Toeplitz-like de rang de déplacement ρ peut être résolu en une complexité de $\mathcal{O}(\rho^2 n \log^2 n)$ opérations élémentaires.*

Remarque 2.2.16. *Lorsque la dimension $n \in [10^3; 10^5]$, de vastes expériences numériques ont montré que l'algorithme GKO classique avec un coût de $\mathcal{O}(\rho n^2)$ opérations est plus rapide que la combinaison avec* ***hm-toolbox*** *(voire [70, sections 3.4 et 5.1]).*

Exemple 2.2.17. *Pour avoir un aperçu du gain engendré par l'algorithme GKO, nous testons cet algorithme de résolution sur plusieurs matrices : sur MATLAB en prenant 10 matrices de Toeplitz de dimensions $m = 1000, \ldots, 10000$, nous mesurons le temps de calcul des solutions des systèmes linéaires $Tx = b$ avec $b \in \mathbb{C}^m$ un vecteur aléatoire. Pour chaque matrice, on résout dans un premier temps $Tx = b$ par un backslash $T\backslash b$, commande pré-définie de MATLAB qui nous donnant une première solution x_1. Puis on procède à une*

résolution du système linéaire Toeplitz par l'algorithme GKO, nous donnant une solution x_2 à l'aide de la commande ***tsolve*** *du package* ***TLComp***[1]*. Enfin, nous employons la résolution par l'algorithme combinant transformation en système Cauchy-like et passage à la structure HSS, ce qui nous donne une 3ème solution x_3 à l'aide de la commande* ***toeplitz_solve****. D'après les propriétés des méthodes de résolution énoncées précédemment, on s'attend à ce que les temps de calcul nécessaires à la résolution d'un système Toeplitz $Tx = b$ suivent une loi $C \times m^\beta$ avec $\beta = 1$ pour* ***toeplitz_solve*** *(du package* ***TLComp****), $\beta = 2$ pour* ***tsolve*** *également du package* ***TLComp*** *et $\beta = 3$ pour le backslash. Nous obtenons ainsi le graphique suivant :*

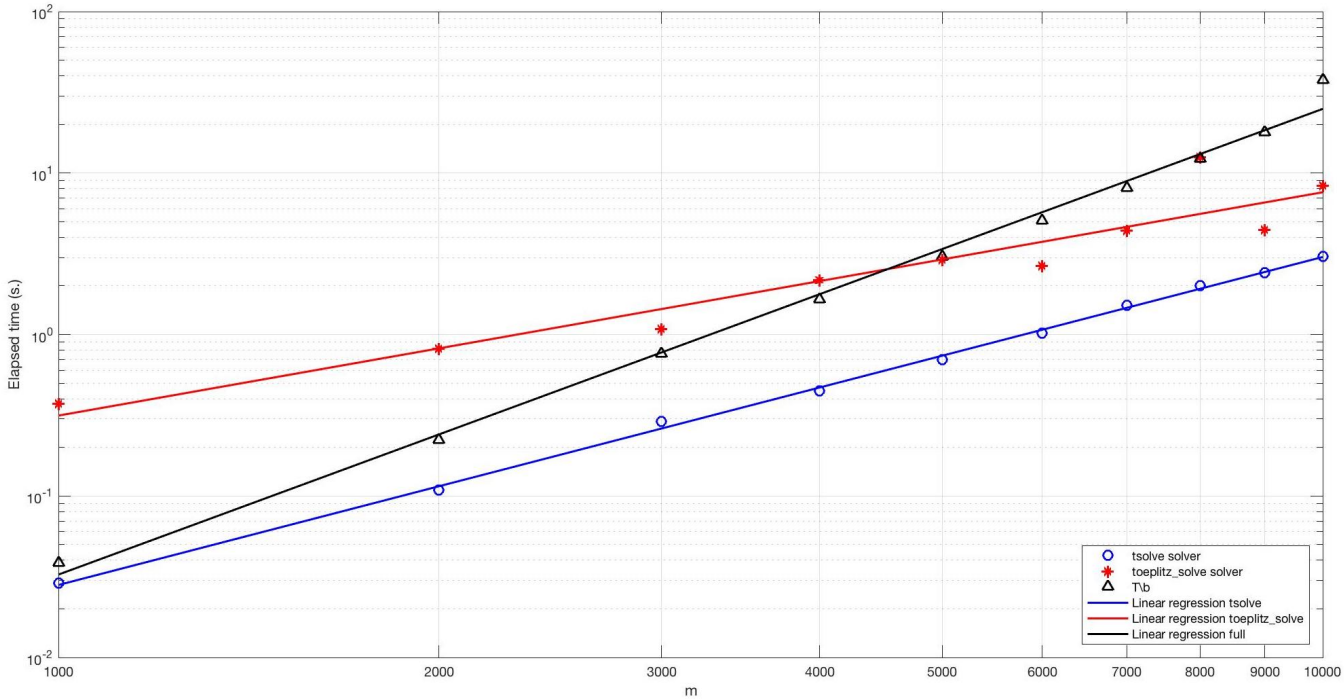

FIGURE 2.1 – Temps nécessaire sur échelle doublement logarithmique pour la résolution du système de Toeplitz $Tx = b$ avec $T \in \mathbb{C}^{m\times m}$ pour $m = 1000, 2000, \ldots, 10000$ à l'aide de la commande backslash (marqueurs noirs), ***tsolve*** (marqueurs bleus) et ***toeplitz_solve*** (marqueurs rouges) et droites de régression associées.

Dans cet exemple, on peut relever que, à l'échelle doublement logarithmique, la loi $C \times m^\beta$ devient une droite de régression sous la forme $\log(C) + \beta \log(m)$, et qu'alors le coefficient directeur β est de $2,8842$ pour la résolution par backslash, $2,0313$ pour la résolution par la commande ***tsolve*** *soit par l'algorithme GKO et $1,3831$ par la commande* ***toeplitz_solve*** *soit par l'algorithme Cauchy+HSS. Ces résultats semblent alors en accord avec les complexités énoncées précédemment pour les différentes méthodes de résolution des systèmes de Toeplitz.*

Exemple 2.2.18. *Pour voir plus loin, nous testons à présent les deux méthodes basées sur l'arithmétique Toeplitz-like pour des matrices de plus grande dimension : prenons 4 matrices de Toeplitz aléatoires de tailles $m = 10000, 20000, 30000, 40000$, $b \in \mathbb{C}^{m\times 1}$ et mesurons le temps nécessaire à la résolution des systèmes de Toeplitz $Tx = b$ à l'aide de l'algorithme GKO et Cauchy+HSS. On obtient alors le graphique en figure 2.2 : toujours sur une échelle doublement logarithmique, les coefficients directeurs β des droite de régression $\log(C)+\beta\log(m)$ valent $2,6697$ pour l'algorithmes GKO et $1,7548$ pour Cauchy+HSS. On remarque alors que les taux de croissance varient de ceux en figure 2.1. Ceci pourrait s'expliquer par le fait que lorsque n croît,*

1. package disponible à l'adresse https ://github.com/rluce/tlcomp

il y a de plus en plus d'appel récursifs dans l'approche HSS et le stockage des quantités intermédiaires. Ici se trouvent des tests effectués pour les deux méthodes sur MATLAB qui ne semble pas être le langage le plus adapté pour pouvoir comparer correctement les deux méthodes. Par conséquent, nous notons ici simplement des observations dans le cas d'une implémentation sur MATLAB qui nous intéresse et employons donc plutôt l'algorithme GKO d'après des résultats expérimentaux.

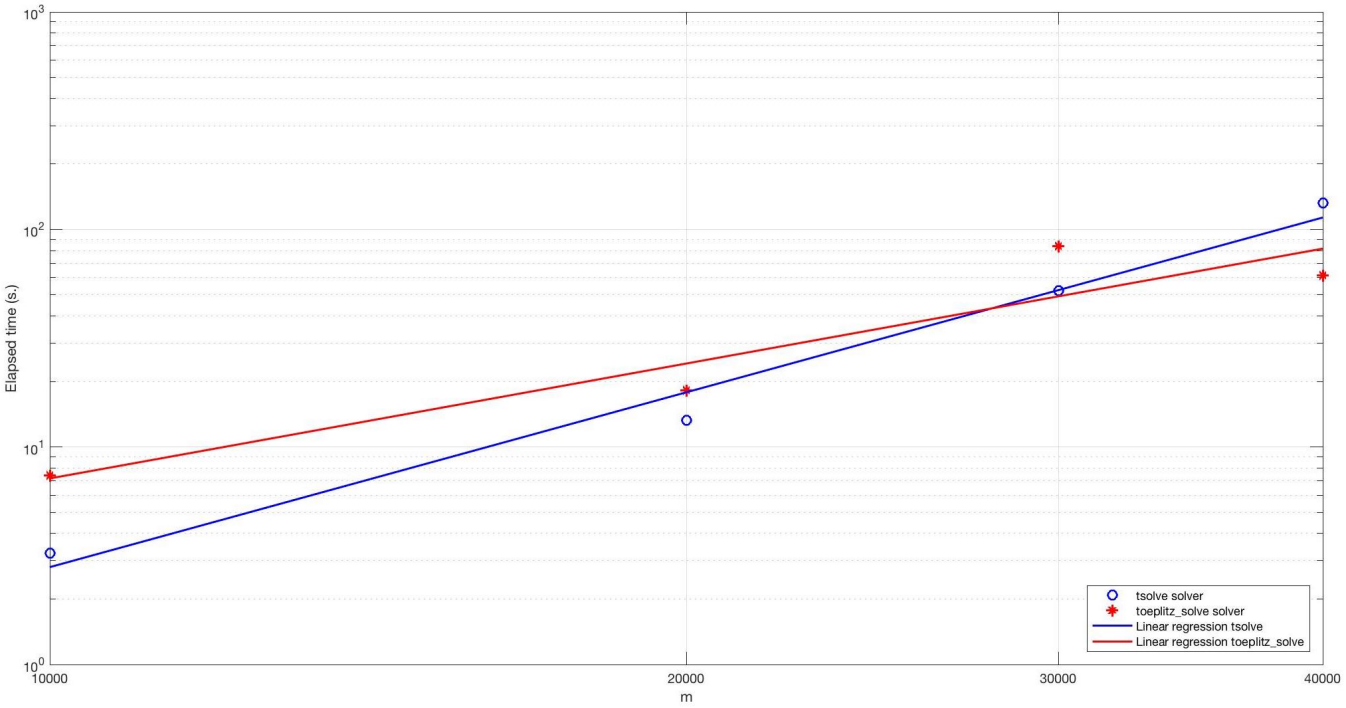

FIGURE 2.2 – Temps nécessaire sur une échelle doublement logarithmique pour la résolution du système de Toeplitz $Tx = b$ avec $T \in \mathbb{C}^{m\times m}$ pour $m = 10000, 20000, 30000$ et 40000 à l'aide de la commande ***tsolve*** (marqueurs bleus) et ***toeplitz_solve*** (marqueurs rouges) et droites de régression associées.

Dans les prochaines expériences, étant donnée la taille de nos matrices, nous emploierons donc l'algorithme GKO classique.

2.3 Opérations sur les générateurs de matrices Toeplitz-like

Considérons une matrice Toeplitz-like $T \in \mathbb{C}^{n\times n}$ et $(G, B) \in (\mathbb{C}^{n\times\rho(T)})^2$ un couple de générateurs associés à T et supposons que l'on souhaite calculer une somme ou un produit avec une autre matrice Toeplitz-like ou encore son inverse. Il nous faut alors expliciter les générateurs associés à ces opérations et connaître le rang de déplacement associé afin d'assurer des algorithmes rapides et super-rapides en cas de reconstruction ou résolution de système. Dans cette section, nous étudions donc l'impact des opérations usuelles sur les générateurs et leur rang de déplacement.

2.3.1 Complément de Schur d'une matrice Toeplitz-like

Pour répondre à notre problème, nous allons nous servir du résultat suivant sur le complément de Schur [42] :

Lemme 2.3.1. *Soit* $X \in \mathbb{C}^{n\times n}$ *sous la forme* $X = \begin{bmatrix} D & U \\ L & X_1 \end{bmatrix}$ *avec* $D \in \mathbb{C}^{k\times k}$ *inversible où* $k < n$ *vérifiant l'équation de Sylvester*

$$FX - XA = GB^*, \quad , F, A \in \mathbb{C}^{n\times n}, \ \ G, B \in \mathbb{C}^{n\times r}, \ \ r \leqslant n.$$

où $F = \begin{bmatrix} F_{1,1} & F_{1,2} \\ F_{2,1} & F_{2,2} \end{bmatrix}, A = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix}$ *avec* $F_{1,1}, A_{1,1} \in \mathbb{C}^{k\times k}, F_{1,2}, A_{1,2} \in \mathbb{C}^{k\times n-k}$ *et*

$$G = \begin{bmatrix} \widehat{G} \\ G_2 \end{bmatrix} \text{ et } B = \begin{bmatrix} \widehat{B} \\ B_2 \end{bmatrix} \tag{2.11}$$

avec $\widehat{G}, \widehat{B} \in \mathbb{C}^{k\times n}$ *et* $G_2, B_2 \in \mathbb{C}^{(n-k)\times n}$. *Alors le complément de Schur* $X_2 := X_1 - LD^{-1}U$ *de* D *dans* X *vérifie l'équation de Sylvester*

$$(-LD^{-1}F_{1,2} + F_{2,2})X_2 - X_2(-A_{2,1}D^{-1}U + A_{2,2}) = (G_2 - LD^{-1}\widehat{G})(B_2^* - \widehat{B}^*D^{-1}U). \tag{2.12}$$

Démonstration. Décomposons X sous la forme

$$X = \begin{bmatrix} I & 0 \\ LD^{-1} & I \end{bmatrix} \begin{bmatrix} D & 0 \\ 0 & X_2 \end{bmatrix} \begin{bmatrix} I & D^{-1}U \\ 0 & I \end{bmatrix}.$$

En multipliant l'équation par la gauche par la matrice $\begin{bmatrix} I & 0 \\ -LD^{-1} & I \end{bmatrix}$ et par la droite par la matrice $\begin{bmatrix} I & -D^{-1}U \\ 0 & I \end{bmatrix}$, on obtient

$$\begin{aligned} &\begin{bmatrix} I & 0 \\ -LD^{-1} & I \end{bmatrix} \begin{bmatrix} F_{1,1} & F_{1,2} \\ F_{2,1} & F_{2,2} \end{bmatrix} \begin{bmatrix} I & 0 \\ LD^{-1} & I \end{bmatrix} \begin{bmatrix} D & 0 \\ 0 & X_2 \end{bmatrix} \\ &\qquad - \begin{bmatrix} D & 0 \\ 0 & X_2 \end{bmatrix} \begin{bmatrix} I & D^{-1}U \\ 0 & I \end{bmatrix} \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix} \begin{bmatrix} I & -D^{-1}U \\ 0 & I \end{bmatrix} \\ &\qquad = \begin{bmatrix} I & 0 \\ -LD^{-1} & I \end{bmatrix} \begin{bmatrix} \widehat{G} \\ G_2 \end{bmatrix} \begin{bmatrix} \widehat{B} & B_2 \end{bmatrix} \begin{bmatrix} I & -D^{-1}U \\ 0 & I \end{bmatrix}, \end{aligned}$$

alors

$$\begin{aligned} &\begin{bmatrix} * & * \\ * & -LD^{-1}F_{1,2} + F_{2,2} \end{bmatrix} \begin{bmatrix} D & 0 \\ 0 & X_2 \end{bmatrix} - \begin{bmatrix} D & 0 \\ 0 & X_2 \end{bmatrix} \begin{bmatrix} * & * \\ * & -A_{2,1}D^{-1}U + A_{2,2} \end{bmatrix} \\ &= \begin{bmatrix} * \\ -LD^{-1}\widehat{G} + G_2 \end{bmatrix} \begin{bmatrix} * & -\widehat{B}^*D^{-1}U + B_2^* \end{bmatrix}. \end{aligned}$$

Ainsi par identification sur le block inférieur droit de l'équation, on obtient la propriété

$$(-LD^{-1}F_{1,2} + F_{2,2})X_2 - X_2(-A_{2,1}D^{-1}U + A_{2,2}) = (-LD^{-1}\widehat{G} + G_2)(-\widehat{B}^*D^{-1}U + B_2^*).$$

□

En adaptant la méthode du complément de Schur au cas de l'opérateur de déplacement, on obtient le corollaire suivant :

Corollaire 2.3.2. *Lorsque* $F = \mathrm{diag}(Z_1, Z_1)$ *et* $A = \mathrm{diag}(Z_{-1}, Z_{-1})$*, le complément de Schur* $X_2 = X_1 - LD^{-1}U$ *vérifie*

$$Z_1X_2 - X_2Z_{-1} = (-LD^{-1}\widehat{G} + G_2)(-\widehat{B}^*D^{-1}U + B_2^*). \tag{2.13}$$

Ainsi, nous pouvons déterminer des générateurs pour le complément de Schur X_2 à partir des générateurs de $X = \begin{bmatrix} D & U \\ L & X_1 \end{bmatrix}$. Soit à présent X_1 une matrice quelconque. A l'aide d'un choix particulier des éléments L, U, D, nous pouvons tenter de construire une matrice $X = \begin{bmatrix} D & U \\ L & X_1 \end{bmatrix}$ dont les générateurs soient facilement identifiables et obtenir un complément de Schur $X_2 = X_1 - LD^{-1}U$ fonction de matrice de X_1.

2.3.2 Somme, produit et inverse des matrices Toeplitz-like

Proposition 2.3.3. *Soient* $X_1, X_2 \in \mathbb{C}^{n\times n}$ *ayant pour générateurs respectifs* $G_1, B_1 \in \mathbb{C}^{n\times\rho_1}$ *et* $G_2, B_2 \in \mathbb{C}^{n\times\rho_2}$ *où* $\rho_1 = \rho(X_1)$ *et* $\rho_2 = \rho(X_2)$*. Alors* X_1+X_2 *admet pour générateurs* $G = \begin{bmatrix} G_1 & G_2 \end{bmatrix}$ *et* $B = \begin{bmatrix} B_1 & B_2 \end{bmatrix}$*. De plus* $\rho(X_1 + X_2) \leqslant \rho_1 + \rho_2$ *et les générateurs de* $X_1 + X_2$ *sont calculés en* $\mathcal{O}(\max(\rho_1, \rho_2)n)$ *opérations arithmétiques élémentaires.*

Proposition 2.3.4. *Soit* $X \in \mathbb{C}^{n\times n}$ *une matrice inversible avec générateurs* $G, B \in \mathbb{C}^{n\times\rho}$ *où* $\rho = \rho(X)$*. Alors des générateurs de* X^{-1} *sont donnés* $\widetilde{G} = \begin{bmatrix} -X^{-1}G & 2X^{-1}e_1 & 2e_1 \end{bmatrix}$ *et* $\widetilde{B} = \begin{bmatrix} X^{-*}B & e_n & X^{-*}e_n \end{bmatrix}$*. De plus,* $\rho(X^{-1}) \leqslant \rho + 2$ *et si* $e_1 \in \Im(G)$ *ou* $e_n \in \Im(B)$*, alors* $\rho(X^{-1}) \leqslant \rho(X) + 1$*.*

Démonstration. Considérons les matrices $M = \begin{bmatrix} -X & I \\ I & 0 \end{bmatrix}$, $F = \begin{bmatrix} Z_1 & 0 \\ 0 & Z_1 \end{bmatrix}$ et $A = \begin{bmatrix} Z_{-1} & 0 \\ 0 & Z_{-1} \end{bmatrix}$. Alors

$$\begin{aligned} FM - MA &= \begin{bmatrix} -Z_1X + XZ_{-1} & (\theta-\gamma)e_1e_n^* \\ 2e_1e_n^* & 0 \end{bmatrix} = \begin{bmatrix} -GB^* & 2e_1e_n^* \\ 2e_1e_n^* & 0 \end{bmatrix} \\ &= \begin{bmatrix} -G & 2e_1 & 0 \\ 0 & 0 & 2e_1 \end{bmatrix} \begin{bmatrix} B & 0 & e_n \\ 0 & e_n & 0 \end{bmatrix}^*. \end{aligned}$$

En posant

$$\widehat{G} = \begin{bmatrix} -G & 2e_1 & 0 \end{bmatrix}, G_2 = \begin{bmatrix} 0 & 0 & 2e_1 \end{bmatrix}$$
$$\widehat{B} = \begin{bmatrix} B & 0 & e_n \end{bmatrix}, B_2 = \begin{bmatrix} 0 & e_n & 0 \end{bmatrix},$$

on obtient par la formule du complément de Schur

$$Z_1(0 - I(-X)^{-1}I) - (0 - I(-X)^{-1}I)Z - 1 = (-I(-X)^{-1}\widehat{G} + G_2)(-\widehat{B}^*(-X)^{-1}I + B_2^*)$$

et le résultat énoncé est prouvé en posant $\widetilde{G} = (-I(-X)^{-1}\widehat{G} + G_2)$ et $\widetilde{B} = (-\widehat{B}^*(-X)^{-1}I + B_2^*)$. □

D'après la proposition 2.3.4, la construction des générateurs de l'inverse nécessitent la résolution de systèmes Toeplitz-like pour obtenir les éléments $-X^{-1}G$, $X^{-1}e_1$, $-X^{-*}B$ et $X^{-*}e_n$. Or, à l'aide de la

proposition 2.2.15, on peut en conclure qu'avec l'emploi de l'algorithme combinant une transformation Toeplitz-like en Cauchy-like et le solveur de ***hm-toolbox***, le calcul des générateurs de l'inverse bénéficie d'un coût réduit. Plus précisément, on obtient le corollaire suivant :

Corollaire 2.3.5. *Les générateurs de X^{-1} peuvent être calculés avec une complexité de $\mathcal{O}(\rho^2 n \log^2 n)$ lorsque $\rho = \rho(X)$.*

Proposition 2.3.6. *Soient $X_1, X_2 \in \mathbb{C}^{n\times n}$ deux matrices avec générateurs respectifs $G_1, B_1 \in \mathbb{C}^{n\times\rho_1}$ et $G_2B_2 \in \mathbb{C}^{n\times\rho_2}$ où $\rho_1 = \rho(X_1)$ et $\rho_2 = \rho(X_2)$. Alors X_1X_2 admet pour générateurs $G = \begin{bmatrix}-2X_2e_1 & X_2G_1 & G_2\end{bmatrix}$ et $B = \begin{bmatrix}X_1^*e_n & B_1 & X_1^*B_2\end{bmatrix}$, $\rho(X_1X_2) \leqslant \rho_1 + \rho_2 + 1$. Si de plus X_1 et X_2 sont Toeplitz, alors les générateurs de X_1X_2 sont calculés en $\mathcal{O}(\max(\rho_1, \rho_2)n \log n)$ opérations élémentaires. Si de plus $e_1 \in \Im(G_1)$ ou $e_n \in \Im(B_2)$, alors $\rho(X_1X_2) \leqslant \rho_1 + \rho_2$.*

Démonstration. Considérons les matrices $M = \begin{bmatrix}-I & X_1\\ X_2 & 0\end{bmatrix}$, $F = \begin{bmatrix}Z_1 & 0\\ 0 & Z_1\end{bmatrix}$ et $A = \begin{bmatrix}Z_{-1} & 0\\ 0 & Z_{-1}\end{bmatrix}$. Alors

$$\begin{aligned} FM - MA &= \begin{bmatrix}Z_1 & 0\\ 0 & Z_1\end{bmatrix}\begin{bmatrix}-I & X_1\\ X_2 & 0\end{bmatrix} - \begin{bmatrix}-I & X_1\\ X_2 & 0\end{bmatrix}\begin{bmatrix}Z_{-1} & 0\\ 0 & Z_{-1}\end{bmatrix} \\ &= \begin{bmatrix}-Z_1 + Z_1 & Z_1X_1 - X_1Z_{-1}\\ Z_1X_2 - X_2Z_{-1} & 0\end{bmatrix} = \begin{bmatrix}-2e_1e_n^* & G_1B_1^*\\ G_2B_2^* & 0\end{bmatrix} \\ &= \begin{bmatrix}-2e_1 & G_1 & 0\\ 0 & 0 & G_2\end{bmatrix}\begin{bmatrix}e_n & 0 & B_2\\ 0 & B_1 & 0\end{bmatrix}^*. \end{aligned}$$

Ainsi d'après la formule du complément de Schur, on a

$$\begin{aligned} &Z_1X_1X_2 - X_1X_2Z_{-1} = \\ &\quad (X - 2\begin{bmatrix}-2e_1 & G_1 & 0\end{bmatrix} + \begin{bmatrix}0 & 0 & G_2\end{bmatrix})(\begin{bmatrix}e_n & 0 & B_2\end{bmatrix}^* X_1 + \begin{bmatrix}0 & B_1 & 0\end{bmatrix}^*) \\ &\quad \begin{bmatrix}-2X_2e_1 & X_2G_1 & G_2\end{bmatrix}\begin{bmatrix}X_1^*e_n & B_1 & X_1^*B_2\end{bmatrix}^* \end{aligned}$$

et on obtient les générateurs annoncés. Pour le calcul des termes X_2G_1 et $X_1^*B_2$, X_1, X_2 étant Toeplitz-like, on on sait d'après la sous-section 2.2.3 que celles-ci se décomposent en sommes de produits de matrices de Toeplitz, et alors tout produit de X_1 ou X_2 peut se réduire à une complexité de $\mathcal{O}(\max\{\rho(X_1), \rho(X_2)\}n \log n)$ à l'aide de la FFT 1.1.4 et on obtient le coût numérique énoncé pour la construction. □

Corollaire 2.3.7. *Soient $X \in \mathbb{C}^{n\times n}$ avec générateurs $G, B \in \mathbb{C}^{n\times\rho}$ où $\rho = \rho(X)$ et $s \in \mathbb{N}^*$. Alors X^s admet pour générateurs*

$$G_s = \begin{bmatrix}-2X^{s-1}e_1 & \dots & -2Xe_1 & G & XG & \dots & X^{s-1}G\end{bmatrix}$$

et

$$B_s = \begin{bmatrix}(X^*)e_n & \dots & (X^*)^{s-1}e_n & (X^*)^{s-1}B & (X^*)^{s-2}B & \dots & B\end{bmatrix}$$

qui sont calculés en $\mathcal{O}(s\rho n \log n)$. De plus, si $e_1 \in \Im(G)$ ou $e_n \in \Im(B)$, alors $\rho(X^s) \leqslant s \times \rho$.

Proposition 2.3.8. *Soient $X \in \mathbb{C}^{n\times n}$ avec générateurs $G, B \in \mathbb{C}^{n\times\rho}$ avec $\rho = \rho(X)$ et $p \in \mathcal{P}_m$ avec $p(x) = a_0 + a_1x + \ldots + a_mx^m$. Notons $G_i, B_i \in \mathbb{C}^{n\times\rho_i}$ les générateurs de X^i avec $\rho_i = \rho(X^i)$ donnés dans le corollaire précédent pour tout $i = 1, \ldots, m$. Alors $p(X)$ admet pour générateurs*

$$\widehat{G} = \begin{bmatrix}a_0e_1 & a_1G_1 & \dots & a_sG_s\end{bmatrix} \text{ et } \widehat{B} = \begin{bmatrix}e_n & B_1 & \dots & B_s\end{bmatrix}.$$

De plus, si $e_1 \in \Im(G)$ ou $e_n\Im(B)$, alors $\rho(p(X)) \leqslant m \times \rho + 1$.

Démonstration. Par somme, les générateurs de $a_0 I + a_1 X + \ldots + a_m X^m$ sont donnés par

$$\begin{bmatrix} a_0 e_1 & a_1 G_1 & \ldots & a_m G_m \end{bmatrix} \text{ et } \begin{bmatrix} e_n & B_1 & \ldots & B_m \end{bmatrix}.$$

□

Proposition 2.3.9. *Soient $X \in \mathbb{C}^{n\times n}$ avec générateurs $G, B \in \mathbb{C}^{n\times\rho}$ où $\rho = \rho(X)$ et $r = \frac{p}{q} \in \mathcal{R}_{k,m}$. Alors $r(X)$ admet pour générateurs*

$$G = \begin{bmatrix} q(X)^{-1} G_q & -q(X)^{-1} G_p & 2e_1 \end{bmatrix},$$
$$B = \begin{bmatrix} -p(X)^* q(X)^{-*} B_q & B_p & -p(X)^* q(X)^{-*} e_n \end{bmatrix}$$

qui sont calculés en $\mathcal{O}(\rho^2 n \log^2 n)$ opérations élémentaires lorsque $k, m \ll n$ De plus, si $e_1 \in \Im(G)$ ou $e_n \in \Im(B)$, alors $\rho(r(X)) \leqslant \rho \max\{\deg(p), \deg(q)\} + 1$.

Démonstration. Définissons $M = \begin{bmatrix} -q(X) & p(X) \\ I & 0 \end{bmatrix}$ et posons $F = \begin{bmatrix} Z_1 & 0 \\ 0 & Z_1 \end{bmatrix}$ et $A = \begin{bmatrix} Z_{-1} & 0 \\ 0 & Z_{-1} \end{bmatrix}$. Par la formule du complément de Schur, on obtient

$$\begin{aligned} Y_\theta X - X Y_\alpha &= \begin{bmatrix} -G_q B_q^* & G_p B_p^* \\ 2e_1 e_n^* & 0 \end{bmatrix} \\ &= \begin{bmatrix} -G_q & G_p & 0 \\ 0 & 0 & 2e_1 \end{bmatrix} \begin{bmatrix} B_q^* & 0 \\ 0 & B_p^* \\ e_n^* & 0 \end{bmatrix} \end{aligned}$$

et par la décomposition

$$\widehat{G} = \begin{bmatrix} G_q & G_p & 0 \end{bmatrix}, G_2 = \begin{bmatrix} 0 & 0 & 2e_1 \end{bmatrix}$$
$$\widehat{B} = \begin{bmatrix} B_q & 0 & e_n \end{bmatrix}, B_2 = \begin{bmatrix} 0 & B_p & 0 \end{bmatrix}$$

on obtient

$$\begin{aligned} Z_\theta X - X Z_\alpha &= (-q(X)^{-1}\widehat{G} + G_2)(-p(X)^* q(X)^{-*}\widehat{B} + B_2)^* \\ &= \begin{bmatrix} q(X)^{-1} G_q & -q(X)^{-1} G_p & (\theta - \gamma) e_1 \end{bmatrix} \\ &\qquad (\begin{bmatrix} -p(X)^* q(X)^{-*} B_q & B_p & -p(X)^* q(X)^{-*} e_1 \end{bmatrix})^* \end{aligned}$$

□

Nous venons donc de voir que les générateurs associés aux opérations matricielles telles que la somme, le produit et même une fonction rationnelle peuvent être donnés explicitement. De plus dans le cas de polynômes ou de fonction rationnelle de matrices, lorsque la ou les matrices considérées sont de faible rang de déplacement, la matrice obtenue par ces opérations reste de faible rang de déplacement lorsque les degrés de polynômes considérés reste faible. Nous l'avons vu précédemment, nous disposons d'une formule de reconstruction nécessitant $\mathcal{O}(\rho n^2)$ opérations arithmétiques élémentaires dès que les générateurs sont de tailles $n \times \rho$.

Corollaire 2.3.10. *Soit $T \in \mathbb{C}^{n\times n}$ matrice Toeplitz-like de rang de déplacement $\rho = \rho(T)$ et $r \in \mathcal{R}_{l,m}$ avec $l < m$. Alors $r(T)$ sous forme d'éléments simples possède un rang de déplacement inférieur ou égal à $(\rho + 1)m$ et peut être implémenté en $\mathcal{O}(\rho^2 mn \log^2 n)$.*

Démonstration. Soit $T \in \mathbb{C}^{n\times n}$ matrice Toeplitz-like de rang de déplacement $\rho(T)$ et r fonction rationnelle d'ordre quelconque. En réécrivant cette dernière, lorsque cela est possible, sous la forme d'une décomposition en éléments simples, on obtient par application à la matrice T l'égalité

$$p(T)q(T)^{-1} = P(T) + \sum_{j=1}^{k} \zeta_j(\beta_j I - T)^{-1}$$

avec P polynôme, $k = \deg(q)$, β_j les pôles de notre fraction rationnelle et ζ_j les résidus associés. On a

- $\rho(\beta_i I - T) \leq \rho(T)$.
- $\rho(\beta_i I - T)^{-1} \leq \rho(T) + 1$ d'après les résultats sur l'inverse.
- $\rho(\sum_{j=1}^{k} \zeta_j(\beta_j I - T)^{-1}) \leq k(\rho(T) + 1)$.
- $\rho(P(T) + \sum_{j=1}^{k} \zeta_j(\beta_j I - T)^{-1}) \leq \rho(T)deg(P) + k(\rho(T) + 1) = \rho(T)(deg(P) + k) + k$.

Dans le cas où $\deg(p) < \deg(q)$, la décomposition en éléments simples se réduit à une somme de k résolvantes de la forme $(\beta_j I - T)^{-1}$ et donc à l'aide des algorithmes de solutions de systèmes Toeplitz-like, on obtient le résultat énoncé. □

2.3.3 Rang de déplacement numérique et compression des générateurs

Comme nous venons de le voir, effectuer des opérations sur les générateurs augmente le rang de déplacement. Par conséquent, on peut se retrouver avec un nombre de colonnes pour nos générateurs et un rang de déplacement plus grands que nécessaire en considérant la précision machine. Pour remédier à ce problème, on peut procéder à chaque opération à une compression en tronquant les valeurs singulières de nos générateurs. Cette démarche est notamment motivée par le résultat suivant :

Proposition 2.3.11. *Soit $X \in \mathbb{C}^{n\times n}$ une matrice Toeplitz-like de rang de déplacement ρ et considérons la décomposition en valeurs singulières*

$$S(X) = U\Sigma V^* = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \begin{bmatrix} V_1^* \\ V_2^* \end{bmatrix}$$

avec $\Sigma_1 = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$, $\Sigma_2 = \mathrm{diag}(\sigma_{r+1}, \ldots, \sigma_\rho)$ où $\sigma_r + 1, \ldots, \sigma_\rho < \varepsilon$. Alors en notant $\tilde{X} = \Gamma(U_1\Sigma_1, V_1)$, on obtient

$$\|X - \tilde{X}\|_2 \leq \frac{n}{2}\sigma_{r+1} \text{ et } \|X - \tilde{X}\|_F \leq \frac{n}{2}\sqrt{\sigma_{r+1}^2, \ldots, \sigma_\rho^2}.$$

Démonstration. Par linéarité de l'opérateur de déplacement S,

$$S(X - \tilde{X}) = S(X) - S(\tilde{X}) = U\Sigma V^* - U_1\Sigma_1 V_1^* = U_2\Sigma_2 V_2^*$$

et on obtient notre résultat en appliquant l'inégalité (2.4). □

Ainsi, le rang de déplacement, correspondant au nombre de valeurs singulières non nulles, peut être réduit en remplaçant les valeurs propres plus petites ou égales à une tolérance à 0, tout en préservant une bonne estimation de la matrice associée aux générateurs après reconstruction. Nous faisons donc appel dans nos essais numériques à l'algorithme de compression suivant :

Algorithm 3 Generators compression algorithm

Require: matrice A et générateurs associés $G, B \in \mathbb{C}^{n\times\rho}$ avec $\rho = \rho(A)$ et $r \leqslant \rho$ tel que $\sigma_{r+1}, \ldots, \sigma_\rho < tol$
Ensure: Générateurs compressés $\tilde{G}, \tilde{B} \in \mathbb{C}^{n\times r}$ tels que $\Gamma(G, B) \approx \Gamma(\tilde{G}, \tilde{B})$ avec $\Gamma = S^{-1}$;

1: Calculer la décomposition QR de G et B : $G = Q_1 R_1$, $B = Q_2 R_2$;
2: Calculer $S = R_1 R_2^* \in \mathbb{C}^{\rho\times\rho}$;
3: Calculer la SVD $U_1 \Sigma_1 V_1 \approx S$, $\Sigma_1 \in \mathbb{C}^{r\times r}$;
4: Calculer $\tilde{G} = Q_1 U_1 \Sigma_1^{1/2} \in \mathbb{C}^{n\times r}$ et $\tilde{B} = Q_2 V_1 \Sigma_1^{1/2} \in \mathbb{C}^{n\times r}$;

Remarque 2.3.12. *De par la dimension des générateurs, l'algorithme de compression est alors de complexité $\mathcal{O}(\rho^2 n)$ lorsque ρ est le rang de déplacement. Ainsi l'introduction de la compression dans nos futures algorithmes n'entraînera pas une augmentation considérable de la complexité globale et est donc un outil indispensable pour des calculs fiables en réduisant les opérations inutiles.*

Revenons à présent à notre problème d'implémentation de fonctions de matrices de Toeplitz. Par exemple, si l'on cherche l'implémentation de la racine carrée d'une matrice de Toeplitz, celle-ci n'est généralement pas de rang de déplacement faible devant la dimension, mais une approximation rationnelle pourrait cependant constituer une alternative. On recherche alors plutôt à vérifier la propriété de faible ε-rang de déplacement [65] :

Définition 2.3.13. *Soit $A \in \mathbb{C}^{n\times n}$ et $\varepsilon > 0$. On dit que la matrice A possède un ε-rang de déplacement ρ si*

$$\min_{B\in\mathbb{C}^{n\times n}} \{\|S(A-B)\|_2 : \rho(B) \leqslant \rho\} \leqslant \varepsilon.$$

Ainsi, à l'aide d'un bon approximant rationnel de la fonction de matrices, si celui-ci possède un faible rang de déplacement, typiquement si l'ordre de l'approximant n'explose pas, alors on pourra considérer celui-ci à la place de la fonction de matrice recherchée pour effectuer les calculs souhaités avec un coût réduit.

2.4 Conclusion

Dans ce chapitre, nous avons vu comment exploiter la structure particulière des matrices de Toeplitz. A partir de la définition de l'opérateur de déplacement $T \mapsto Z_1 T - T Z_{-1}$ que nous avons étudié en première section, nous avons vu que le cas des matrices de Toeplitz donnait en image par cet opérateur une matrice avec une structure particulière.
Cette structure est ensuite exploitée à l'aide de la définition du rang de déplacement en section 2.2 où nous avons vu que pour des matrices à faible rang de déplacement que l'on appelle alors Toeplitz-like, leur image pouvait être découpée sous la forme d'un produit de matrices, que l'on appelle générateurs et de taille considérablement réduite. En particulier, ces générateurs nous permettent alors d'implémenter des algorithmes de résolution de tout système d'équations linéaires avec comme matrice de coefficients une matrice Toeplitz-like de rang de déplacement ρ en complexité $\mathcal{O}(\rho n^2)$ à l'aide de l'algorithme GKO ou $\mathcal{O}(\rho^2 n \log^2 n)$ opérations élémentaires par transformation du système Toeplitz-like en système Cauchy-like puis par compression au format HSS.
En section 2.3, nous nous sommes intéressés au développement d'une arithmétique Toeplitz-like, en définissant pour une matrice T la construction des générateurs associés à différentes opérations sur la matrice T. L'emploi des algorithmes super-rapides de résolution de système d'équations linéaires avec comme matrice de coefficients une matrice Toeplitz-like nous permet alors d'effectuer les calculs de ces générateurs avec une complexité de l'ordre $\mathcal{O}(\rho(T)^2 n \log^2 n)$ opérations élémentaires. En particulier, toute fonction polynômiale

ou rationnelle en la matrice Toeplitz-like T est calculée en arithmétique Toeplitz-like avec une complexité de $\mathcal{O}(\rho(T)^2 n \log^2 n)$, réduisant considérablement le coût par rapport à une arithmétique pleine.
De plus, pour assurer que le rang de déplacement de ces opérations sur une matrice T n'explose pas, nous avons montré que nous disposons en dernière partie d'un algorithme de compression des générateurs.
Ces résultats nous amènent donc à considérer une chose : étant donné que le calcul d'une fonction rationnelle de matrice de Toeplitz $T \in \mathbb{C}^{n \times n}$ en arithmétique Toeplitz-like est de complexité $\mathcal{O}(\rho(T)^2 n \log^2 n)$ et que nous disposons également de l'approximation rationnelle de fonctions, plutôt que de calculer la fonction de matrice $f(T)$ pour une certaine fonction f et une matrice de Toeplitz T, il nous serait plus facile de déterminer un approximant rationnelle de $f(T)$, construit en arithmétique Toeplitz-like.
Dans la suite de cette thèse, nous allons tester cette nouvelle arithmétique dans différentes expériences numériques qui sont réalisées à l'aide du package ***TLComp***[2], reprenant l'arithmétique Toeplitz-like exposée dans ce chapitre.

2. disponible à l'adresse https ://github.com/rluce/tlcomp

Chapitre 3

Application aux fonctions de matrices racine carrée et signe

La fonction de matrice racine carrée est l'une des fonctions de matrices les plus utilisées, généralement dans le cas de matrices symétriques, que ce soit dans une nécessité d'obtenir la fonction de matrice en elle-même ou pour améliorer les algorithmes de calcul pour d'autres fonctions de matrices comme le log (voire section 1.4). Si la définition intégrale de la racine carrée d'une matrice $A \in \mathbb{C}^{n\times n}$ sans valeur propre dans $]-\infty;0]$ est donnée par $\sqrt{A} = \frac{2}{\pi} A \int_0^\infty (t^2 I + A)^{-1} dt$, il nous faut concevoir des méthodes de calcul ou d'approximation pour un calcul numérique. Dans ce chapitre, nous commençons en section 3.1 par rappeler la définition et quelques propriétés de la fonction de matrice racine carrée principale $\sqrt{A}$ d'une matrice $A \in \mathbb{C}^{n\times n}$ et montrons comment obtenir la forme intégrale de cette fonction de matrice. Puis en sous-section 3.2.1, nous rappelons la méthode itérative de Newton pour la racine carrée principale sans prendre en compte la structure particulière de la matrice concernée. Ses variantes Newton-BD et Newton-DB produit pour la racine carrée principale sont également étudiées ainsi que leur convergence et leur stabilité asymptotique en sous section 3.2.2. Afin de réduire le nombre d'itérations nécessaires à la convergence des matrices itérées de la méthode de Newton pour la racine carrée principale, nous tentons d'accélérer celle-ci en proposant différents choix pour un premier terme en sous-section 3.3.1 ainsi qu'en introduisant des paramètres dans ces méthodes en sous-section 3.3.2 permettant l'accélération de la convergence. En section 3.4, nous prenons en compte le cas des matrices Toeplitz-like en étudiant la méthode de Newton pour la racine carrée principale et ses variantes calculées en arithmétique Toeplitz-like. Plus précisément, les générateurs associés aux différentes formes de cette méthode sont explicités et le rang de déplacement de chaque matrice itérée est alors majoré en fonction de celui de la matrice étudiée et plusieurs expériences numériques sont observées par la suite en sous-section 3.4.2. En section 3.5, après quelques rappels sur la fonction de matrice signe, nous passons en revu en sous-section 3.5.1 la méthode de Newton pour la fonction signe, son adaptation en arithmétique Toeplitz-like et fournissons quelques expériences numériques en sous-section 3.5.3.

3.1 Racine carrée principale de matrice

Considérons la fonction de matrice racine carrée. Pour une matrice $A \in \mathbb{C}^{n\times n}$, certains auteurs appellent racine carrée de A toute matrice $B \in \mathbb{C}^{n\times n}$ telle que $B^2 = A$ et donc A peut posséder jusqu'à 2^n racines différentes [57, Theorem 1.26]. Ici dans notre travail, nous allons exclusivement considérer la fonction racine carrée principale $f(z) = z^{1/2} = \sqrt{z}$ définie pour un argument scalaire $z = \varrho e^{i\theta} \in \mathbb{C} \setminus (-\infty;0]$ par $f(\varrho e^{i\theta}) =$

$\sqrt{\varrho}e^{i\theta/2}$ où $\varrho > 0$, $\sqrt{\varrho} > 0$ et $\theta \in (-\pi;\pi)$. Pour une matrice A avec $\sigma(A) \subseteq \mathbb{C} \setminus (-\infty;0]$, la racine carrée principale X de A est l'unique matrice X solution de l'équation $X^2 = A$ avec toutes ses valeurs propres de partie réelle strictement positive, voire [57, Theorem 1.29]. Nous la notons $X = \sqrt{A}$.

Lemme 3.1.1. *Soit $f : z \mapsto \frac{1}{\sqrt{(z-\alpha)(z-\beta)}}$ avec $-\infty < \alpha < \beta < +\infty$ où $x \mapsto \sqrt{x}$ est la racine carrée principale. Alors*

$$f(z) = \frac{1}{\pi}\int_\alpha^\beta \frac{1}{\sqrt{(t-\alpha)(\beta-t)}}\frac{dt}{z-t}.$$

Démonstration. Considérons la fonction $f : z \mapsto \frac{1}{\sqrt{(z-\alpha)(z-\beta)}} = \frac{1}{z-\alpha}\frac{\sqrt{z-\alpha}}{\sqrt{z-\beta}} = \frac{1}{z-\beta}\frac{\sqrt{z-\beta}}{\sqrt{z-\alpha}}$ avec $x \mapsto \sqrt{x}$ la racine carrée principale. La fonction f est analytique sur $(\mathbb{C}\setminus[\alpha;\beta])\cup\{\infty\}$ par analyticité de la racine carrée principale. De plus, $\frac{1}{\sqrt{(z-\alpha)(z-\beta)}} = \frac{1}{\pm z}\frac{1}{\sqrt{1-\frac{\alpha+\beta}{z}+\frac{\alpha\beta}{z^2}}}$ au voisinage de $\pm\infty$ et ainsi $\frac{1}{|\sqrt{(z-\alpha)(z-\beta)|}}$ se comporte comme $\frac{1}{|z|} + \mathcal{O}(\frac{1}{z^2})$ lorsque $|z| \to +\infty$. De plus, $f(z) \in \mathbb{R}$ dès que $z \in \mathbb{R} \setminus [\alpha;\beta]$. Plus précisément, $f(z) > 0$ pour $z > \beta$ et $f(z) < 0$ dès que $z < \alpha$.
Ensuite, d'après la formule de Cauchy, pour tout $z \in \mathbb{C} \setminus [\alpha;\beta]$ et pour tout contour γ n'entourant pas $[\alpha;\beta]$ et entourant z au sens mathématique positif, on peut écrire

$$\frac{1}{\sqrt{(z-\alpha)(z-\beta)}} = \frac{1}{2i\pi}\int_\gamma \frac{1}{\sqrt{(x-\alpha)(x-\beta)}}\frac{dx}{x-z}$$

Soit $0 < r < \operatorname{dist}(z,[\alpha;\beta])/2$ et considérons le contour $\gamma = \gamma_1 + \gamma_2 + \gamma_3 + \gamma_4$ donnée en figure 3.1 où

— $\gamma_1 : x = x(t) = \alpha + re^{it}$, avec $t : 3\pi/2 \to \pi/2$ dans le sens négatif ;
— $\gamma_2 = x = x(t) = t + ir$, $t : \alpha \to \beta$;
— $\gamma_3 : x = x(t) = \beta + re^{it}$, $t : \pi/2 \to -\pi/2$ dans le sens négatif ;
— $\gamma_4 : x = x(t) = t - ir$, $t \in \beta \to \alpha$;

Alors $\int_\gamma f(x)\frac{dx}{x-z} = \int_{\gamma_1} f(x)\frac{dx}{x-z} + \int_{\gamma_2} f(x)\frac{dx}{x-z} + \int_{\gamma_3} f(x)\frac{dx}{x-z} + \int_{\gamma_4} f(x)\frac{dx}{x-z}$. Or,

i. $\forall x \in \gamma_1$, $f(x) = f(x(t)) = \frac{1}{\sqrt{(\alpha+re^{it}-\alpha)(\alpha+re^{it}-\beta)}} = \frac{1}{\sqrt{(re^{it})(\alpha+re^{it}-\beta)}}$ d'où $|f(x)| \leqslant \frac{1}{\sqrt{r|\beta-\alpha|}} = \frac{1}{\sqrt{\beta-\alpha}}\frac{1}{\sqrt{r}}$ et $\frac{1}{|z-x|} \leqslant \frac{1}{||z-\alpha|-r|}$. Or $|z-\alpha| \geqslant \operatorname{dist}(z,[\alpha;\beta])$ d'où $\frac{1}{|z-x|} \leqslant \frac{2}{\operatorname{dist}(z,[\alpha;\beta])}$ et donc

$$\begin{aligned}\Big|\int_{\gamma_1} f(x)\frac{dx}{z-x}\Big| &\leqslant \int_{\gamma_1} \frac{1}{\sqrt{\beta-\alpha}}\frac{1}{\sqrt{r}}\frac{2}{\operatorname{dist}(z,[\alpha;\beta])}dx \\ &\leqslant \frac{1}{\sqrt{\beta-\alpha}}\frac{1}{\sqrt{r}}\frac{2}{\operatorname{dist}(z,[\alpha;\beta])}r\int_{3\pi/2}^{\pi/2} ie^{it}dt \to 0 \text{ quand } r \to 0;\end{aligned}$$

ii. De même, $\int_{\gamma_3} f(x)\frac{dx}{x-z} \to 0$ quand $r \to 0$;

iii. De plus

$$\begin{aligned}\frac{1}{2i\pi}\int_\gamma f(x)\frac{dx}{x-z} &= \frac{1}{2i\pi}\Big(\int_{\gamma_2} f(x)\frac{dx}{x-z} + \int_{\gamma_4} f(x)\frac{dx}{x-z}\Big) \\ &= \frac{1}{2i\pi}\Big(\int_\alpha^\beta \frac{f(t+ir)}{t+ir-z}dt + \int_\beta^\alpha \frac{f(t-ir)}{t-ir-z}dt\Big) \\ &= \frac{1}{2i\pi}\Big(\int_\alpha^\beta \frac{f(t+ir)}{t+ir-z}dt - \int_\alpha^\beta \frac{f(t-ir)}{t-ir-z}dt\Big) \\ &= \frac{1}{2i\pi}\int_\alpha^\beta \frac{f(t+ir)}{t+ir-z} - \frac{\overline{f(t+ir)}}{\overline{t+ir}-z}dt \\ &= -\frac{1}{\pi}\int_\alpha^\beta \frac{\Re(f(t+ir))r}{(z-t)^2+r^2}dt + \frac{1}{\pi}\int_\alpha^\beta \Im(f(t+ir))\frac{t-z}{(t-z)^2+r^2}dt.\end{aligned}$$

Or, $\Im(f(t+ir)) = -\Im(f(t-ir)) = -|f(t+ir)| \to -|f(t)|$ quand $r \to 0$ et $\frac{t-z}{(t-z)^2+r^2} \to \frac{1}{t-z}$ quand $r \to 0$. De plus,

$$\Big|2i\int_\alpha^\beta \Re(f(t+ir))\frac{r}{(t-z)^2+r^2}dt\Big| \leqslant 2\int_\alpha^\beta |\Re(f(t+ir))|\frac{r}{\mathrm{dist}(z,[\alpha;\beta])^2-r^2}dt \to 0$$

lorsque $r \to 0$ d'où on peut conclure que

$$\lim_{r\to 0}\frac{1}{2i\pi}\Big(\int_{\gamma_2} f(x)\frac{1}{x-z}+\int_{\gamma_4} f(x)\frac{dx}{x-z}\Big) = \frac{1}{\pi}\int_\alpha^\beta -|f(t)|\frac{dt}{t-z} = \frac{1}{\pi}\int_\alpha^\beta \frac{1}{\sqrt{(t-\alpha)(\beta-t)}}\frac{dt}{z-t}.$$

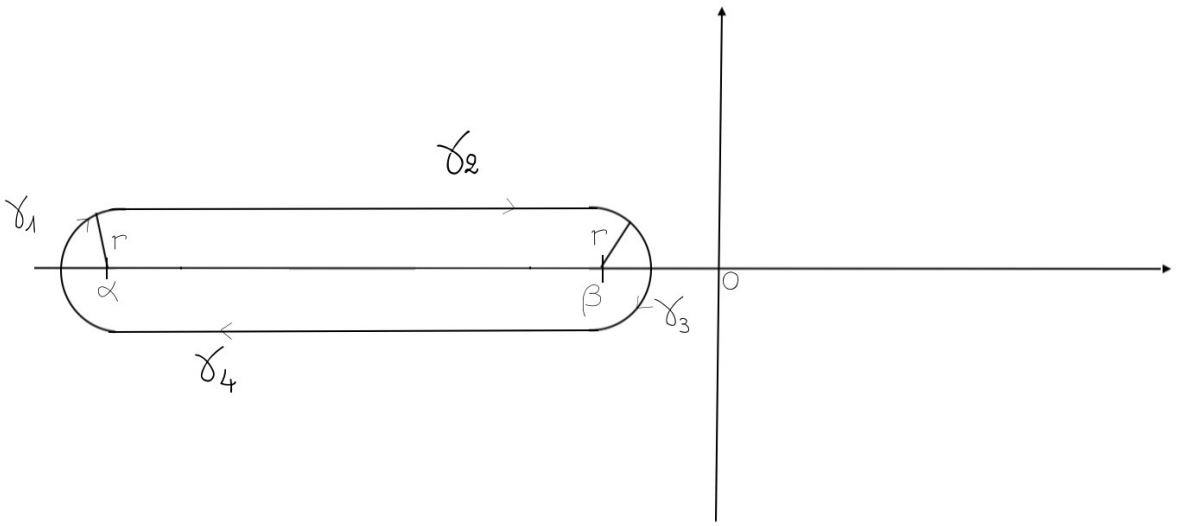

FIGURE 3.1 – Contour γ.

□

Lemme 3.1.2. *La fonction de matrice $f(A) = \sqrt{A}$ est donnée par la représentation intégrale $f(A) = \frac{2}{\pi}A\int_0^\infty (t^2I+A)^{-1}dt$.*

Démonstration. D'après le lemme 3.1.1, pour tout $-\infty < \alpha < \beta < +\infty$, $\frac{1}{\sqrt{(\alpha-z)(\beta-z)}} = \frac{1}{\pi}\int_\alpha^\beta \frac{1}{\sqrt{(t-\alpha)(\beta-t)}}\frac{dt}{z-t}$. Or, en multipliant la fonction $f(z) = \frac{1}{\sqrt{(\alpha-z)(\beta-z)}}$ par $\sqrt{|\alpha|}$, on obtient $\sqrt{|\alpha|}f(z) = \frac{1}{\sqrt{(1-z/\alpha)(\beta-z)}}$, et en faisant tendre $\alpha \to -\infty$ et en considérant $\beta = 0$, on obtient une expression intégrale de la racine carrée inverse $z^{-1/2} = \lim_{\alpha\to\infty}\sqrt{|\alpha|}\frac{1}{\sqrt{(\alpha-z)(0-z)}} = \lim_{\alpha\to\infty}\frac{1}{\pi}\int_\alpha^0 \frac{\sqrt{|\alpha|}}{\sqrt{(t-\alpha)(-t)}}\frac{dt}{z-t} = \frac{1}{\pi}\int_{-\infty}^0 \frac{1}{\sqrt{|t|}}\frac{dt}{z-t}$.
Prenons à présent la définition intégrale de la fonction de matrice. On obtient pour toute matrice A et pour tout contour Γ entourant le spectre de A une fois dans le sens positif,

$$\begin{aligned}\sqrt{A} = AA^{-1/2} &= A\Big(\frac{1}{2i\pi}\int_\Gamma f(\zeta)(\zeta I-A)^{-1}d\zeta\Big) = A\Big(\frac{1}{2i\pi}\int_\Gamma \frac{1}{\pi}\int_{-\infty}^0 \frac{1}{\sqrt{|t|}}\frac{dt}{\zeta-t}(\zeta I-A)^{-1}d\zeta\Big)\\ &= A\Big(\frac{1}{\pi}\int_{-\infty}^0 \frac{1}{2i\pi}\int_\Gamma (\zeta I-A)^{-1}\frac{1}{\zeta-t}d\zeta\frac{1}{\sqrt{|t|}}dt\Big) = A\Big(\frac{1}{\pi}\int_{-\infty}^0 (A-t)^{-1}\frac{1}{\sqrt{|t|}}dt\Big)\end{aligned}$$

Par changements de variables $u = -t$ puis $x = \sqrt{u}$, on obtient

$$\sqrt{A} = A\Big(-\frac{1}{\pi}\int_{+\infty}^{0} (A+u)^{-1}\frac{1}{\sqrt{u}}du\Big) = A\Big(\frac{1}{\pi}\int_{0}^{+\infty}(A+x^2)^{-1}\frac{2x}{x}dx\Big) = \frac{2}{\pi}A\int_{0}^{+\infty}(x^2I+A)^{-1}dx$$

et nous obtenons ainsi l'expression intégrale de la fonction de matrice racine carrée principale. □

3.2 Rappels sur la méthode de Newton pour la racine carrée d'une matrice non-structurée

Pour construire la racine carrée principale d'une matrice, l'algorithme de Schur peut être employé comme expliquée par N. I. Higham [57, Section 6.2]. Cependant, pour une matrice $A \in \mathbb{C}^{n\times n}$, utiliser un tel algorithme nécessite un coût numérique de $\frac{28}{3}n^3$ opérations [57, Algorithm 6.3]. Si l'on souhaite utiliser notre arithmétique Toeplitz-like, étant donné que nous pouvons implémenter des fonctions rationnelles avec cet arithmétique à faible complexité, nous allons ici plutôt rechercher des approximants rationnels à la fonction de matrice racine carrée principale. L'une des méthodes les plus connues pour l'approximation de la fonction de matrice racine carrée principale est la méthode de Newton. Celle-ci ne nécessitant pas de structure particulière de la matrice étudiée, elle peut donc être employée pour toute matrice A satisfaisant $\sigma(A)\cap\mathbb{R}_- = \emptyset$. Cependant, pour des matrices non structurées, passer par cette méthode n'aurait pas d'intérêt puisqu'il nous en coûterait alors une complexité de l'ordre de celle de la méthode de Schur-Parlett. Or dans notre contexte, la méthode itérative de Newton produit des fonctions rationnelles en la matrice étudiée, et l'on sait alors d'après la proposition 2.3.9 que ces matrices itérées pourront être construites avec une complexité bien inférieure. Dans cette section donc, nous rappelons dans un premier temps cette méthode ainsi que ses variantes, puis dans une deuxième sous-section nous rappelons les propriétés de stabilité de chacune de ces variantes.

3.2.1 Itération de Newton

En dérivant l'équation $Y^2 = A$ avec $A \in \mathbb{C}^{n\times n}$ sans valeur propre dans $\mathbb{R}_-$ et en supposant que X avec $Y = X + E$ est une solution approchée avec $\|E\| < \varepsilon$, on a $A = (X+E)^2 = X^2 + XE + EX + E^2$. En ignorant le terme E^2, on obtient alors l'équation $XE + EX = A - X^2$. Définissons donc X_0 et E_0 avec X_0 premier terme fixe et E_0 solution de l'équation de Sylvester associée $X_0E_0 + E_0X_0 = A - X_0^2$, et posons $X_1 = X_0 + E_0$. On obtient une récurrence $X_kE_k + E_kX_k = A - X_k^2$ avec inconnue E_k, où les matrices itérées E_k sont bien définies si et seulement si X_k et $-X_k$ n'ont pas de valeur propre commune, soit lorsque X_k est inversible, et $X_{k+1} = X_k + E_k$. En supposant que X_0 commute avec E_0, on peut définir les matrices itérées de Newton de la manière suivante :

Définition 3.2.1. *Soient $A \in \mathbb{C}^{n\times n}$ et $X_0 \in \mathbb{C}^{n\times n}$ tel que X_0 commute avec A. Alors la méthode itérative de Newton pour la matrice A est donnée par*

$$\begin{cases} X_0 \text{ donné} \\ X_{k+1} = \frac{1}{2}(X_k + AX_k^{-1}) \end{cases}. \tag{3.1}$$

Lemme 3.2.2. *[57, Lemma 6.8] Soit $(X_k)_k$ la suite des matrices itérées de la méthode de Newton (3.1) où X_0 commute avec A et supposons que toutes les matrices itérées soient bien définies. Alors pour tout k, X_k commute avec A. De plus, X_k est hermitienne pour tout $k \geqslant 0$ dès que A l'est [57, Lemme 6.8].*

Dans la suite, nous sélectionnerons uniquement des premiers termes X_0 commutant avec A.

Lemme 3.2.3. *Pour toute matrice $A \in \mathbb{C}^{n\times n}$ telle que $\sigma(A) \cap \mathbb{R}_- = \emptyset$, la méthode de Newton converge (localement) quadratiquement vers la racine carrée principale de A avec [57, equation (6.13)]*

$$\|X_{k+1} - \sqrt{A}\| \leqslant \frac{1}{2}\|X_k^{-1}\|\|X_k - \sqrt{A}\|^2, \quad \textit{pour tout } k \geqslant 0. \tag{3.2}$$

Démonstration. Pour tout $k \geqslant 0$,

$$\begin{aligned} X_{k+1} - \sqrt{A} &= \frac{1}{2}(X_k + X_k^{-1}A) - \sqrt{A} = \frac{1}{2}(X_k - 2\sqrt{A} + X_k^{-1}A) \\ &= \frac{1}{2}X_k^{-1}(X_k^2 - 2X_k\sqrt{A} + A) = \frac{1}{2}X_k^{-1}(X_k - \sqrt{A})^2 \end{aligned}$$

d'où $\|X_{k+1} - \sqrt{A}\| \leqslant \frac{1}{2}\|X_k^{-1}\|\|X_k - \sqrt{A}\|^2$. □

Cette méthode de Newton est très peu stable [57, Chap. 6.4] comme nous le rappelons en section 3.2.2 et ne converge pas nécessairement pour chaque matrice $A \in \mathbb{C}^{n\times n}$ avec spectre dans $\overline{\mathbb{C}}\backslash\mathbb{R}_-$. Pour remédier à ce problème, des variantes de la méthode de Newton permettent d'assurer la convergence vers la racine carrée principale de la matrice A tout en assurant la stabilité. Nous nous intéressons ici à deux de ces formes :

Méthode de Newton-DB : Soit X_0 commutant avec A. En définissant $Y_k = A^{-1}X_k$ pour tout $k \geqslant 0$, alors $X_{k+1} = \frac{1}{2}(X_k + Y_k^{-1})$ et $Y_{k+1} = A^{-1}X_{k+1} = \frac{1}{2}(Y_k + X_k^{-1})$. Cette méthode itérative dérivée de Newton est appelée itération de Denman et Beavers [30] :

$$\begin{cases} X_0, \ \ Y_0 = A^{-1}X_0, \\ X_{k+1} = \frac{1}{2}(X_k + Y_k^{-1}), \\ Y_{k+1} = \frac{1}{2}(Y_k + X_k^{-1}), \end{cases} \tag{3.3}$$

et vérifie $X_k \to \sqrt{A}$ la racine carrée principale de A et $Y_k \to A^{-1/2}$. De plus, les X_k définis par Newton-DB sont identiques aux X_k de Newton.

Méthode de Newton-DB produit : Soit X_0 commutant avec A. En notant $M_k = X_kY_k$ pour tout $k \geqslant 0$ avec X_k, Y_k définis par la méthode de Newton-DB, on obtient la méthode de Newton-DB produit [1], identifiée par Cheng, Higham, Kenney et Laub [23] :

$$\begin{cases} X_0, \ \ M_0 = X_0, \\ M_{k+1} = \frac{1}{2}\Big(I + \frac{1}{2}(M_k + M_k^{-1})\Big), \\ X_{k+1} = \frac{1}{2}(I + M_k^{-1})X_k, \end{cases} \tag{3.4}$$

Remarque 3.2.4. *Il est à noter que pour chacune des formes de la méthode de Newton, les matrices itérées X_k sont les mêmes. En effet, à partir de X_0, pour la méthode Newton-DB, $X_{k+1} = \frac{1}{2}(X_k + Y_k^{-1}) = \frac{1}{2}(X_k + X_k^{-1}A) = \frac{1}{2}(X_k + AX_k^{-1})$ puisque X_k commute avec A d'où X_k^{-1} également. Ensuite, pour la méthode de Newton-DB produit, par définition de M_k, $X_{k+1} = \frac{1}{2}(I + M_k^{-1})X_k = \frac{1}{2}(I + Y_k^{-1}X_k^{-1})X_k = \frac{1}{2}(X_k + Y_k^{-1})$ avec Y_k défini d'après la méthode de Newton-DB et on retrouve donc bien les matrices itérées X_k de la méthode classique.*

1. Dans son ouvrage [57], Higham considère l'itération $X_{k+1} = \frac{1}{2}X_k(I + M_k^{-1})$ mais notre récurrence semble plus adaptée au cas où la matrice ne commute plus avec A à cause de l'arithmétique en précision finie. Nous verrons que cette modification n'a pas d'impact sur la stabilité et la précision asymptotique.

Si X_0 est une fonction rationnelle d'ordre $[2^\ell|2^\ell-1]$ exprimée en la matrice A, soit $X_0 = r_{2^\ell,2^\ell-1}(A)$ pour un $\ell \geqslant 0$, alors cela reste vraie pour chaque matrice itérée X_k.

Proposition 3.2.5. *Soit $A \in \mathbb{C}^{n\times n}$ avec $\sigma(A) \subseteq \mathbb{C}\backslash\mathbb{R}_-$ et soit $X_0 = r_{2^\ell,2^\ell-1}(A) \in \mathbb{C}^{n\times n}$ avec $r_\ell \in \mathcal{R}_{2^\ell,2^\ell-1}$ pour $\ell \geqslant 0$, un terme initial pour la méthode de Newton (et ses variantes). Alors*

$$X_k = r_{2^{k+\ell},2^{k+\ell}-1}(A)$$

avec $r_{2^{k+\ell},2^{k+\ell}-1} \in \mathcal{R}_{2^{k+\ell},2^{k+\ell}-1}$ pour tout $k \geqslant 1$.

Démonstration. On démontre par récurrence. Soit $X_0 = r_{2^\ell,2^\ell-1}(A) \in \mathbb{C}^{n\times n}$ le premier terme de la méthode itérative de Newton et ses variantes. Alors $X_1 = \frac{1}{2}(X_0 + AX_0^{-1}) = \frac{1}{2}(X_0^2 + A)X_0^{-1} = r_{2,1}(X_0) = r_{2^{\ell+1},2^{\ell+1}-1}(A)$. Supposons à présent que pour un $k \geqslant 1$, $X_k = r_{2^{k+\ell},2^{k+\ell}-1}(A)$. Alors $X_{k+1} = \frac{1}{2}(X_k + AX_k^{-1}) = r_{k+1}(X_0) = r_{k+1}(r_{2^{\ell+1},2^{\ell+1}-1}(A))$ est d'ordre $[2\max(\deg(p_k),\deg(q_k)); \deg(p_k) + \deg(q_k)] = [2\times 2^{k+\ell}, 2^{k+\ell} + 2^{k+\ell} - 1] = [2^{k+\ell+1}, 2^{k+\ell+1} - 1]$ et donc

$$X_{k+1} = r_{2^{k+\ell+1},2^{k+\ell+1}-1}(A)$$

et le résultat est démontré. □

3.2.2 Stabilité et précision asymptotique de l'itération de Newton

Malgré la convergence quadratique de la méthode de Newton, en fonction des matrices considérées, il nous faut nous assurer de la stabilité numérique de ces algorithmes récursifs. Pour ce faire, Higham [57, Chap. 6.4] considère la dérivée de Fréchet de la fonction associée aux itérations de Newton.

Définition 3.2.6. *[57, Section 3.1] La dérivée de Fréchet d'une fonction de matrice $f : \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$ en un point $X \in \mathbb{C}^{n\times n}$ est l'application linéaire*

$$L_f : \begin{cases} \mathbb{C}^{n\times n} & \to \mathbb{C}^{n\times n} \\ E & \mapsto L_f(X,E) \end{cases}$$

telle que pour tout $E \in \mathbb{C}^{n\times n}$,

$$f(X+E) - f(X) - L_f(X,E) = o(\|E\|).$$

Considérons une itération $X_{k+1} = G(X_k)$ avec point fixe X et supposons que G admet une dérivée de Fréchet $L_G(X)$ en X. Alors l'itération est stable au voisinage du point fixe X s'il existe une constante c telle que $\|L_G^i(X)\| \leqslant c$ pour tout $i > 0$ avec $L_G^i(X) = (L_G(X))^i$. Pour obtenir la dérivée de Fréchet de l'itération $X_{k+1} = G(X_k)$, on définie dans un premier temps $X_0 = X + E_0$ pour E_0 tel que $\|E_0\| < \varepsilon$ avec $\varepsilon > 0$ suffisamment petit. Par récurrence, on définit $E_k = X_k - X$, ce qui nous donne pour tout $k \geqslant 0$,

$$X_{k+1} = G(X_k) = G(X + E_k) = G(X) + L_G(X,E_k) + o(\|E_k\|)$$

et comme $G(X) = X$, on observe que

$$E_{k+1} = L_G(X,E_k) + o(\|E_k\|), \tag{3.5}$$

et alors pour tout $i \geqslant 1$,

$$E_k = L_G^{k+1}(X,E_0) + \sum_{j=0}^{k} L_G^j(X, o(\|E_{k-j}\|)) \tag{3.6}$$

ce qui nous donne pour une itération stable

$$\|E_{k+1}\| \leqslant c\|E_0\| + c\sum_{j=0}^{k} o(\|E_{k-j}\|) \leqslant c\|E_0\| + c(k+1)\, o(\|E_0\|).$$

Définition 3.2.7 (Précision asymptotique)**.** *Soit $X_{k+1} = G(X_k)$ une itération avec point fixe X. Alors on appelle précision asymptotique (relative) la quantité $\varepsilon\|L_G(X)\|$ avec ε la précision machine.*

Cette précision asymptotique nous permet de mesurer l'erreur relative pour les itérations consécutives : en effet, si $X_0 = X + E_0$ avec $\|E_0\| \leqslant \varepsilon\|X\|$, alors pour $X_1 = X + E_1$, on a d'après (3.5),

$$\|X_1 - X\| = \|E_1\| \lesssim \|L_G(X, E_0) \leqslant \|L_G(X)\|\, \|E_0\| \leqslant \varepsilon\|L_G(X)\|\, \|X\|,$$

et ainsi $\varepsilon\|L_G\|$ borne l'erreur relative $\frac{\|X_1 - X\|}{\|X\|}$.
A partir de ces critères, N. J. Higham mesure la stabilité et la précision asymptotique des trois itérations de Newton. Nous notons pour toute matrice $A \in \mathbb{C}^{n\times n}$,

$$\kappa(A) = \|A\|_2\|A^{-1}\|_2$$

le conditionnement de la matrice A. Lorsque A est hermitienne, le conditionnement vérifie $\kappa(A) = \frac{\lambda_{\max}}{\lambda_{\min}}$ avec $\lambda_{\max} = \max\sigma(A)$ et $\lambda_{\min} = \min\sigma(A)$.

Lemme 3.2.8. *[57, Section 6.4.1] Soit $A \in \mathbb{C}^{n\times n}$ matrice diagonalisable avec valeurs propres $\lambda_1, \ldots, \lambda_n$. Alors la méthode de Newton est stable pour A si et seulement si*

$$\max_{i,j} |\frac{1}{2}(1 - \lambda_i^{1/2}\lambda_j^{-1/2})| < 1. \tag{3.7}$$

De plus, la précision asymptotique pour la méthode de Newton pour toute matrice A avec $\sigma(A) \cap \mathbb{R}_- = \emptyset$ est de $\frac{1}{2}(1 + \kappa(\sqrt{A}))\varepsilon$ où ε est la précision.

Démonstration. Soit $A \in \mathbb{C}^{n\times n}$. Supposons que $A = Z\Lambda Z^{-1}$ avec $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ et soit $X_0 = ZD_0Z^{-1}$ où $D_0 = \mathrm{diag}(d_1, \ldots, d_n)$. Pour une perturbation de rang 1 $E_0 = \varepsilon u_i v_j := \varepsilon(Ze_i)(e_j^T Z^{-1})$ de X_0 avec $i \neq j$. Alors $(X_0 + E_0)^{-1} = X_0^{-1} - X_0^{-1}E_0X_0^{-1}$ d'après la formule de Sherman-Morrison, et alors la perturbation induite pour X_1 vérifie $E_1 = \frac{1}{2}(E_0 - X_0^{-1}E_0X_0^{-1}A) = \frac{1}{2}(E_0 - \varepsilon\frac{1}{d_i}Ze_ie_j^T\frac{\lambda_j}{d_j}Z^{-1}) = \frac{1}{2}(1 - \frac{\lambda_j}{d_id_j})E_0$. En notant $X_0 = \sqrt{A}$ de sorte à ce que $d_i = \lambda_i^{1/2}$, alors $E_1 = \frac{1}{2}(1 - \lambda_i^{1/2}\lambda_j^{-1/2})E_0$, et après k itérations on a $X_k + E_k = \sqrt{A} + \left[\frac{1}{2}(1 - -\lambda_i^{1/2}\lambda_j^{-1/2})\right]^k E_0$, ce qui montre que l'itération de Newton peut diverger lorsque lorsque la condition (3.7) n'est pas vérifiée.
Pour mesurer la précision asymptotique, notons $G(X) = \frac{1}{2}(X + X^{-1}A)$ et E une perturbation de X. Alors $G(X + E) = \frac{1}{2}(X + X^{-1}A) + \frac{1}{2}(E - X^{-1}EX^{-1}A) + \mathcal{O}(\|E\|^2)$ où on identifie la dérivée de Fréchet au point fixe $X = \sqrt{A}$,

$$L_G(\sqrt{A}, E) = \frac{1}{2}(E - \sqrt{A}^{-1}E\sqrt{A})$$

et on obtient alors $\|L_G(\sqrt{A}, E)\| \leqslant \frac{1}{2}(1 + \kappa(\sqrt{A}))\|E\|$ ce qui nous donne l'estimation énoncée. □

Dans le cas d'une matrice A hermitienne définie positive, il nous faudra alors la condition $\kappa(A) = \frac{\max(\sigma(A))}{\min(\sigma(A))} < 9$, ce qui restreint considérablement le champs des matrices pour lesquelles la méthode de Newton est exploitable. En revanche, les méthodes dérivées Newton DB et Newton DB produit permettent d'élargir considérablement ce champs de matrices de par leur stabilité et leur précision asymptotique respectives.

Lemme 3.2.9. *Les méthode de Newton DB et Newton DB produit sont stables pour toute matrice et satisfont une précision asymptotique d'ordre $\kappa(A^{1/2})\varepsilon$ et $\frac{3}{2}\varepsilon$ respectivement.*

Démonstration. Notons $G(X,Y) = \frac{1}{2}\begin{bmatrix} X+Y^{-1} \\ Y+X^{-1} \end{bmatrix}$ la fonction associée à l'itération de Newton DB. Sa dérivée de Fréchet en (X,Y) dans une direction (E,F) est donnée par

$$L_g(X,Y)(E,F) = L_g(X,Y;E,F) = \frac{1}{2}\begin{bmatrix} E - Y^{-1}FY^{-1} \\ F - X^{-1}EX^{-1} \end{bmatrix}$$

et on peut rapidement vérifier qu'en un point fixe (B,B^{-1}) de G, sa dérivée de Fréchet $L_g(B,B^{-1})$ est idempotent et ceci prouve que l'itération de Newton DB est stable au point fixe $(\sqrt{A},\sqrt{A}^{-1})$. De plus, pour toutes matrices E et F avec $\|E\| \leqslant \varepsilon\|\sqrt{A}\|$ et $\|F\| \leqslant \varepsilon\|\sqrt{A}^{-1}\|$, $\|L_g(\sqrt{A},\sqrt{A}^{-1})(E,F)\| \leqslant \frac{1}{2}\|\sqrt{A}\|(1+\kappa(\sqrt{A}))\varepsilon$ et ainsi la précision asymptotique est bornée par $\frac{1}{2}(1+\kappa(\sqrt{A}))\varepsilon$.

Considérons à présent la méthode de Newton-DB produit donnée par 3.4

$$\begin{cases} X_0 = A, \;\; M_0 = A, \\ M_{k+1} = \frac{1}{2}\Big(I + \frac{1}{2}(M_k + M_k^{-1})\Big), \\ X_{k+1} = \frac{1}{2}(I + M_k^{-1})X_k, \end{cases}$$

Notons pour toutes matrices $M, X \in \mathbb{C}^{n\times n}$,

$$G(M,X) = \frac{1}{2}\begin{bmatrix} I + \frac{1}{2}(M + M^{-1}) \\ (I+M^{-1})X \end{bmatrix}.$$

Sa dérivée de Fréchet vérifie pour toutes matrices $M, X \in \mathbb{C}^{n\times n}$ fixées et $E, F \in \mathbb{C}^{n\times n}$ vérifie l'égalité

$$L_g(M,X)(E,F) = L_g(M,X,E,F) = \frac{1}{2}\begin{bmatrix} \frac{1}{2}(E - M^{-1}EM^{-1}) \\ (I+M^{-1})F - (M^{-1}EM^{-1})X \end{bmatrix}. \tag{3.8}$$

Pour $M = I$ et $X = \sqrt{A}$, on a alors $L_g(I,\sqrt{A})(E,F) = \frac{1}{2}\begin{bmatrix} 0 \\ 2F - E\sqrt{A} \end{bmatrix}$, et en notant $\tilde{E} = \frac{1}{4}(E - M^{-1}EM^{-1}) = 0$ et $\tilde{F} = \frac{1}{2}(I+M^{-1})F - (M^{-1}EM^{-1})X = \frac{1}{2}(2F - E\sqrt{A})$, on observe alors que

$$\begin{aligned}(L_g(I,\sqrt{A}))^2(E,F) &= \frac{1}{2}\begin{bmatrix} \frac{1}{2}(\tilde{E} - M^{-1}\tilde{E}M^{-1}) \\ (I+M^{-1})\tilde{F} - (M^{-1}\tilde{E}M^{-1})\sqrt{A} \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 0 \\ 2F - E\sqrt{A} \end{bmatrix} \\ &= L_g(I,\sqrt{A})(E,F).\end{aligned}$$

Par récurrence, on obtient plus généralement que

$$L_g(I,\sqrt{A})^n = L_g(I,\sqrt{A}), \quad \text{pour tout } n \geqslant 1$$

d'où $L_g(I,\sqrt{A})$ est idempotente et ainsi la méthode de Newton-DB produit est stable. Enfin d'après (3.8),

$$\|L_g(I,\sqrt{A})(E,F)\| \leqslant \frac{1}{2}(2\|F\| + \|E\|\|\sqrt{A}\|) \leqslant \frac{3}{2}\|\sqrt{A}\|\varepsilon$$

lorsque $\|E\| \leqslant \varepsilon$ et $\|F\| \leqslant \|\sqrt{A}\|\varepsilon$ avec ε la précision machine et on obtient une précision asymptotique de $\frac{3}{2}\varepsilon$. □

3.3 Amélioration de la méthode

Nous avons vu que, pour une matrice $A \in \mathbb{C}^{n\times n}$ avec $\sigma(A) \cap \mathbb{R}_- = \emptyset$, la méthode de Newton et ses variantes construisaient des matrices itérées X_k (avec X_0 premier terme commutant avec A) qui convergent (localement) quadratiquement vers la racine carrée principale de A d'après (3.2). Or si le premier terme X_0 est très éloigné de $\sqrt{A}$ soit lorsque $\|X_0 - A\| \gg 1$, le nombre d'itérations nécessaires à la convergence des X_k vers $\sqrt{A}$ peut alors exploser, entraînant une complexité importante. Pour remédier à ce problème, nous exposons dans cette section quelques choix de premier terme X_0 de sorte à garantir que $\|X_0 - \sqrt{A}\| \approx 1$ dans un premier temps. Puis, pour apporter une convergence plus rapide, nous rappelons quelques choix de paramètres que l'on peut introduire dans les différentes formes de la méthode de Newton afin d'accélérer la convergence.

3.3.1 Choix d'un premier terme

Soit $A \in \mathbb{C}^{n\times n}$ avec $\sigma(A) \subseteq [\alpha;\beta]$ et X_0 le premier terme de la méthode de Newton pour l'approximation de la fonction de matrice $\sqrt{A}$. Habituellement le premier terme choisi est la matrice A elle-même. Or, si la matrice A est très éloignée de sa racine carrée principale $\sqrt{A}$ alors la convergence risque de nécessiter un nombre important d'itérations et nous ferait perdre l'intérêt d'une application en arithmétique Toeplitz like. Pour réduire l'erreur initiale $\|X_0 - \sqrt{A}\|$, nous sélectionnons deux types de premier terme autre que A :

- un premier terme $X_0 = cA$ multiple scalaire de la matrice $A \in \mathbb{C}^{n\times n}$ avec $c \in \mathbb{R}_+^*$. Nous sélectionnons alors un scalaire $c > 0$ de sorte à normaliser le premier terme X_0.
- nous avons vu en sous-section 1.3.3 que nous disposons d'approximants de Padé. Chaque matrice itérée X_k de la méthode de Newton étant fonction rationnelle d'ordre $[2^k; 2^k - 1]$, nous prenons X_0 sous la forme d'un approximant de Padé d'ordre $[2^\ell; 2^\ell - 1]$ en la matrice A pour $\ell \geqslant 1$. Nous considérons donc dans nos expériences numériques des approximants de Padé p/q d'ordre $[2,1]$, $[8,7]$ et $[16,15]$ appliqués à la matrice A. La fonction $z \mapsto \sqrt{z}$ n'étant pas développable en série entière au voisinage de 0, on considère plutôt la fonction $z \mapsto \sqrt{1+z}$ et on note également que pour toute matrice $A \in \mathbb{C}^{n\times n}$ et pour tout $c > 0$, $\sqrt{A} = \sqrt{c}(I + (\frac{1}{c}A - I))^{1/2}$. En employant un approximant de Padé p/q d'ordre $[2^\ell; 2^\ell - 1]$ en 0 de la fonction $\sqrt{1+z}$, on considère le premier terme

$$X_0 = \sqrt{c}\,\frac{p}{q}(\frac{1}{c}A - I).$$

Remarque 3.3.1. *Lorsque X_0 est un approximant de Padé d'ordre $[2^k; 2^k - 1]$ en la matrice $\frac{1}{c}A - I$, on n'évalue pas $\frac{p}{q}$ en cette matrice sous la forme d'un quotient de polynôme, mais on détermine les coefficients de sa forme en éléments simples, lorsque celle-ci existe.*

$$X_0 = P(B) + \sum_{j=0}^{s} a_j (B - b_j)^{-1}, \quad \text{où } B = \frac{1}{c}A - I, s \in \mathbb{N}, a_j, b_j \in \mathbb{C}, P \in \mathcal{P}_1[x].$$

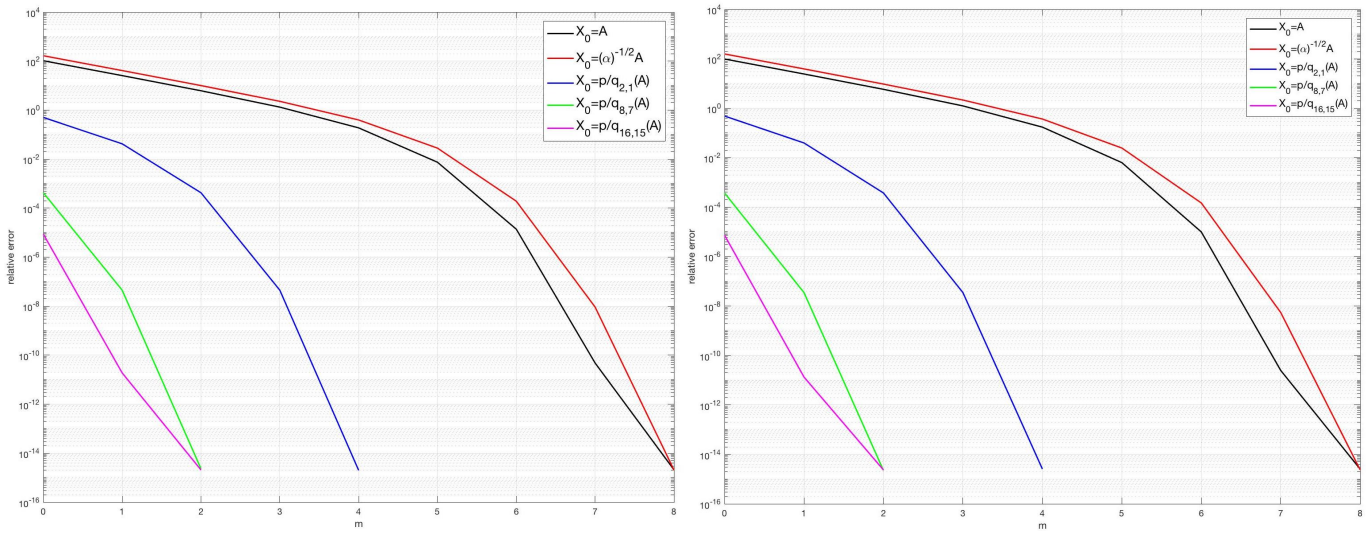

FIGURE 3.2 – Erreurs relatives $\|I - X_m A^{-1/2}\|_2$ avec X_m matrices itérées des méthodes de Newton et Newton-DB produit pour l'approximation de $\sqrt{A}$ où $A \in \mathbb{R}^{3000\times 3000}$ matrice de Toeplitz symétrique définie positive à spectre dans $[0,86; 6,33]$ et conditionnement $7,321$ (à gauche), spectre dans $[0,62; 104,27]$ et conditionnement $166,242$ (à droite) pour différents premiers termes X_0.

Exemple 3.3.2. *Pour la figure 3.2 nous avons implémenté les méthodes de Newton et Newton-DB avec différents premiers termes X_0 en arithmétique pleine pour 2 matrices de Toeplitz symétriques définies positives de taille 3000×3000 avec spectre dans $[0,75; 5,69]$ (à gauche avec Newton) et $[0,34; 59,37]$ (à droite avec Newton-DB) et où nous notons m l'indice des matrices itérées calculées X_m par les méthodes de Newton et Newton-DB. Nous affichons sur ces graphiques les erreurs relatives $\|I - X_m A^{-1/2}\|_2$ en fonction de l'indice m où $A^{-1/2}$ est calculée à l'aide de la commande $inv(sqrtm(A))$ de MATLAB. Nous pouvons observer ici la convergence quadratique des matrices itérées des méthodes de Newton et Newton-DB. Nous voyons également sur ces 2 graphiques que le choix d'un premier terme sous la forme d'un approximant de Padé d'ordre $[2^\ell, 2^\ell - 1]$ en la matrice $A \in \mathbb{C}^{n\times n}$ au lieu de $X_0 = cA$ permet de réduire considérablement le nombre d'itérations nécessaires à la convergence, passant de 6 à 4 ou 2 itérations nécessaires pour la méthode de Newton avec $\sigma(A) \subseteq [0,75; 5,69]$ pour le graphique de gauche et 9 à 6 ou 3 itérations pour la méthode de Newton-DB avec $\sigma(A) \subseteq [0,34; 59,37]$ pour le graphique de droite. En revanche, on s'aperçoit qu'employer un ordre important des approximants de Padé n'est pas nécessaire puisque prendre $X_0 = (p/q)_{8,7}(A)$ ou $X_0 = (p/q)_{16,15}(A)$ n'accélère pas la convergence dans le premier cas à gauche, et ne réduit que d'un le nombre d'itérations nécessaires à la convergence à droite.*

3.3.2 Introduction de paramètres

Nous avons vu que les différentes formes de la méthode de Newton offrent une convergence quadratique vers la racine carrée principale d'une matrice A. Cependant, pour accélérer encore la convergence, on peut remplacer les termes X_k, Y_k, M_k dans chaque itération par des matrices $\mu_k X_k$ ($\mu_k Y_k$ et $\mu_k M_k$), comme discuté par [79] dans le cas scalaire (discussion que l'on peut trouver dans [18, Chap. V.5.D]). La méthode de Newton et ses variantes s'écrivent alors

Méthode de Newton :

$$X_0 \text{ donné}, \quad X_{k+1} = \frac{1}{2}(\mu_k X_k + \mu_k^{-1} X_k^{-1} A) \tag{3.9}$$

Méthode de Newton-DB :

$$\begin{cases} X_0, \ Y_0 = A^{-1} X_0 \\ X_{k+1} = \frac{1}{2}(\mu_k X_k + \mu_k^{-1} Y_k^{-1}), \\ Y_{k+1} = \frac{1}{2}(\mu_k Y_k + \mu_k^{-1} X_k^{-1}), \end{cases} \tag{3.10}$$

Méthode de Newton-DB produit :

$$\begin{cases} X_0, \ M_0 = X_0 \\ M_{k+1} = \frac{1}{2}\Big(I + \frac{\mu_k^2 M_k + \mu_k^{-2} M_k^{-1}}{2}\Big), \\ X_{k+1} = \frac{1}{2}\mu_k (I + \mu_k^{-2} M_k^{-1}) X_k, \end{cases} \tag{3.11}$$

Higham suggère des paramètres déterminantaux

$$\mu_k = |\det\Big(X_k/\sqrt{A}\Big)|^{-1/n} = |\det(X_k)/\det(A)^{1/2}|^{-1/n}$$

dans le cas Newton et Newton-DB et $\mu_k = |\det(M_k)|^{-1/(2n)}$ dans le cas Newton produit [57, Section 6.5], de tels paramètres permettant de rapprocher les valeurs propres des X_k sur le disque de rayon $\varrho(\sqrt{A})$. Cependant, le calcul du déterminant à chaque étape des méthodes de Newton nous fait perdre l'intérêt de la structure avec générateurs qui sera appliquée plus tard. Nous prenons donc des paramètres suggérés par Byers et Xu [20], mais également Higham, Hale et Trefethen [57], dépendant uniquement du spectre de la matrice initiale A [9] : pour $\sigma(A) \subseteq [\alpha;\beta] \subset (0;+\infty)$,

$$\mu_0 = \frac{1}{\sqrt[4]{\alpha\beta}}, \quad \mu_1 = \sqrt{\frac{2\sqrt{\kappa}}{1+\kappa}}, \quad \kappa = \sqrt{\frac{\alpha}{\beta}}, \quad \forall k \geqslant 1, \ \mu_{k+1} = \sqrt{\frac{2\mu_k}{1+\mu_k^2}}. \tag{3.12}$$

Dans le cas d'une matrice symétrique définie positive, ce scaling permet d'obtenir [9, Corollary 2] la borne supérieure sur l'erreur relative

$$\Big\|I - X_k \sqrt{A}^{-1}\Big\| \leqslant \frac{2e_k}{1-e_k}$$

où $e_k = \min\Big\{\Big\|\frac{1}{\sqrt{z}}\big(\sqrt{z} - \frac{p(z)}{q(z)}\big)\Big\|_{L^\infty(\alpha;\beta)} : \deg p \leqslant 2^{k-1}, \deg q \leqslant 2^{k-1}-1\Big\}$. De plus, les matrices $\frac{2\mu_k^2}{1+\mu_k^2} X_k$ vérifient

$$\Big\|I - \frac{2\mu_k^2}{1+\mu_k^2} X_k A^{-1/2}\Big\| \leqslant e_k$$

Dans le cas de la méthode de Newton-DB produit les matrices itérées X_k et M_k vérifient alors

$$\sigma(X_k A^{-1/2}) \subset [1; \frac{1}{\mu_k^2}], \quad \sigma(M_k) \subset [1; \frac{1}{\mu_k^4}], \quad \|X_k A^{-1} X_k - I\| = \|M_k - I\| \leqslant \frac{1-\mu_k^4}{\mu_k^4}. \tag{3.13}$$

Pour conserver la stabilité et la précision asymptotique obtenues avec paramètres $\mu_k = 1$, nous procédons en 2 phases comme suggéré par Higham [57, Section 6.6] : en première phase, on utilise les paramètres

décrits par (3.12) jusqu'à ce que $\frac{1-\mu_k^4}{\mu_k^4} \leqslant 10^{-3}$, puis on choisit $\mu_k = 1$ pour la suite des itérations. Ainsi les matrices $M_k = X_k A^{-1} X_k$ vérifient pour tout $k \geqslant 0$,

$$\|M_{k+1} - I\| \leqslant \frac{1}{4}\|M_k - I\|^2.$$

En effet, on a pour tout $k \geqslant 0$,

$$\begin{aligned}\|M_{k+1} - I\| &= \|\frac{1}{2}I + \frac{1}{4}(M_k + M_k^{-1}) - I\| = \frac{1}{4}\|M_k + M_k^{-1} - 2I\| \\ &= \frac{1}{4}\|M_k^{-1}(M_k^2 + I - 2M_k)\| = \frac{1}{4}\|M_k^{-1}(M_k - I)^2\|,\end{aligned}$$

soit

$$\|M_{k+1} - I\| \leqslant \frac{1}{4}\|M_k^{-1}\|\,\|M_k - I\|^2. \tag{3.14}$$

Or, on peut démontrer par récurrence que si pour un $K \geqslant 0$, $\|I - M_K\| \leqslant \frac{1}{4}$, alors pour tout $k \geqslant K$, $\|I - M_k\| \leqslant \frac{1}{4}$. En effet, si $\|I - M_K\| \leqslant \frac{1}{4}$, alors par les séries de Neumann, $\|M_K^{-1}\| = \|(I - I + M_k)^{-1}\| \leqslant \frac{1}{1-1/4} = \frac{4}{3}$, d'où d'après (3.14),

$$\|I - M_{K+1}\| \leqslant \frac{1}{4}\frac{4}{3}\|M_k - I\|^2 = \frac{1}{3}\|M_k - I\|^2 \leqslant \frac{1}{12}\|I - M_K\| \leqslant \frac{1}{4}$$

d'où $\|I - M_k\| \leqslant \frac{1}{4}$ pour tout $k \geqslant K$ dès que $\|I - M_K\| \leqslant \frac{1}{4}$. Or, d'après (3.14), $\sigma(M_k) \subset [1; \frac{1}{\mu_k^4}]$, et avec M_0 et A symétrique, $\|M_k^{-1}\| \leqslant 1$ et alors

$$\|M_{k+1} - I\| \leqslant \frac{1}{4}\|M_k - I\|^2.$$

A l'aide de la convergence quadratique, on espère n'avoir besoin que de trois itérations supplémentaires pour la phase 2.

Considérons à présent le cas $X_0 = \frac{p}{q}(A)$ où $\frac{p}{q}$ est un approximant de Padé d'ordre $[2^m; 2^m - 1]$ avec $m \in \mathbb{N}$ et A une matrice symétrique définie positive avec $\sigma(A) \subseteq [\alpha; \beta]$. On souhaite en théorie que $A^{-1/2}X_k$ soit proche de l'identité, soit que $\|I - A^{-1/2}X_k\| = \varrho(I - A^{-1/2}X_k) \approx 0$. Or, on peut noter que $X_0 = \frac{p_{\ell,\ell-1}}{q_{\ell,\ell-1}}(\frac{1}{c}A - I) = A^{1/2}g_\ell(A)$ avec $g_\ell(z) = \frac{1+\left(\frac{\sqrt{c}-\sqrt{z}}{\sqrt{c}+\sqrt{z}}\right)^\ell}{1-\left(\frac{\sqrt{c}-\sqrt{z}}{\sqrt{c}+\sqrt{z}}\right)^\ell}$ et $\left|\frac{\sqrt{c}-\sqrt{z}}{\sqrt{c}+\sqrt{z}}\right| \leqslant \gamma < 1$ pour tout $z \in [\alpha; \beta]$ si et seulement si $c = \sqrt{\alpha\beta}$ et $\gamma = \frac{1-\sqrt[4]{\alpha/\beta}}{1+\sqrt[4]{\alpha/\beta}}$. Avec de tels paramètres, on obtient pour tout $m \in \mathbb{N}^*$,

$$g_{2^m}(z) \in \left[1; \frac{1+\gamma^{2^m}}{1-\gamma^{2^m}}\right] = [1; g_{2^m}(\alpha)].$$

On pose alors

$$\mu_0 := \frac{1}{\sqrt{g_{2^m}(\alpha)}} \text{ et } \mu_{j+1} = \sqrt{\frac{2\mu_j}{1+\mu_j^2}}, \tag{3.15}$$

nous assurant que pour tout $k \geqslant 0$, $g_{2^m}(z) \in [1; \frac{1}{\mu_k^2}]$ pour tout $z \in [\alpha; \beta]$.

3.4 Newton pour les matrices Toeplitz-like et expériences numériques

Nous venons de voir comment approcher la racine carrée principale d'une matrice A avec spectre dans $\overline{\mathbb{C}}\backslash\mathbb{R}_-$ à l'aide de la méthode de Newton et ses variantes en arithmétique pleine. Bien que la méthode converge quadratiquement et peut être accélérée par un choix de premier terme X_0 et l'introduction de paramètres μ_k, il n'en reste pas moins que le calcul des matrices itérées va nécessiter l'inversion et la multiplication de matrices, ce qui en arithmétique pleine va nécessiter une complexité d'ordre n^3 si l'on ne prend pas en compte de la structure de la matrice. Considérons $T \in \mathbb{C}^{n\times n}$ une matrice de Toeplitz ou Toeplitz-like pour laquelle nous cherchons à approcher $\sqrt{T}$ par la méthode de Newton. En démarrant avec un premier terme $X_0 = cT$ ou $X_0 = p/q(T)$ avec p/q approximant de Padé, chaque matrice itérée X_k est une fonction rationnelle de la matrice T de la forme $r_{2^k,2^k-1}(X_0)$. La méthode de Newton peut donc être implémentée à l'aide l'arithmétique Toeplitz-like du chapitre 2. De plus, en choisissant un premier terme X_0 de sorte que $\|X_0 - \sqrt{A}\| \leqslant 1$ et en introduisant des paramètres μ_k pour accélérer la convergence, on espère obtenir un bon approximant de la racine carrée principale en peu d'itérations de sorte à ce que le rang de déplacement de la matrice itérée final soit négligeable devant la dimension n. Dans cette section, nous étudions donc la méthode de Newton et ses variantes en arithmétique Toeplitz-like, les générateurs associés aux matrices itérées et le rang de déplacement de chaque matrice itérée. Puis dans une deuxième sous-section, nous réalisons quelques expériences numériques de la méthode de Newton (ainsi que Newton DB et Newton-DB produit) en arithmétique Toeplitz-like. En particulier, le rang de déplacement numérique des matrices itérées de Newton est mesuré pour chaque X_k afin de déterminer si la racine carrée principale d'une matrice peut être approchée par une matrice Toeplitz-like.

3.4.1 Générateurs associés aux itérations de Newton

Soit $T \in \mathbb{C}^{n\times n}$ matrice Toeplitz-like et $G, B \in \mathbb{C}^{n\times\rho(T)}$ des générateurs associés. Soit $X_0 \in \mathbb{C}^{n\times n}$ le premier terme de la méthode de Newton avec $X_0 = cT$ avec $c \in \mathbb{R}$ ou $X_0 = \frac{p}{q}(\frac{1}{c}T - I)$ avec $\frac{p}{q}$ approximant de Padé d'ordre $[2^\ell, 2^\ell - 1]$, que l'on suppose Toeplitz-like. Pour tout $k \geqslant 0$ chaque matrice itérée X_k de Newton est donnée en fonction des matrices itérées précédentes. Par conséquent, on peut démontrer que les générateurs G_k et B_k associés à X_k sont donnés à partir des générateurs G_{k-1} et B_{k-1} de X_{k-1} d'après les propositions 2.3.3, 2.3.4 et 2.3.6 de la manière suivante :

Proposition 3.4.1. *Soient $A \in \mathbb{C}^{n\times n}$ avec générateurs $G, B \in \mathbb{C}^{n\times\rho(A)}$ et $(X_k)_k$ la suite des matrices itérées de la méthode de Newton* (3.9) *avec X_0 premier terme donné. Notons pour tout $k \geqslant 0$, G_k, B_k des générateurs associés à X_k. Alors G_k, B_k vérifient pour tout $k \geqslant 1$,*

$$G_k = \frac{1}{2}\begin{bmatrix}\mu_k G_{k-1} & \mu_k^{-1}X_{k-1}^{-1}G & -\mu_k^{-1}X_{k-1}^{-1}G_{k-1} & 2\mu_k^{-1}e_1\end{bmatrix}$$

et

$$B_k = \begin{bmatrix}B_{k-1} & B & A^*X_{k-1}^{-*}B_{k-1} & A^*X_{k-1}^{-*}e_n\end{bmatrix}.$$

Démonstration. Soient G_{k-1}, B_{k-1} générateurs de X_{k-1}. D'après la proposition 2.3.4, les générateurs de X_{k-1}^{-1} sont donnés par $G_1 = \begin{bmatrix}-X_{k-1}^{-1}G_{k-1} & 2X_{k-1}^{-1}e_1 & 2e_1\end{bmatrix}$ et $B_1 = \begin{bmatrix}X_{k-1}^{-*}B_{k-1} & e_n & X_{k-1}^{-*}e_n\end{bmatrix}$ d'où d'après la proposition 2.3.6, les générateurs de AX_{k-1}^{-1} sont donnés par

$$\tilde{G}_{k-1} = \begin{bmatrix}-2X_{k-1}^{-1}e_1 & X_{k-1}^{-1}G & G_1\end{bmatrix} = \begin{bmatrix}-2X_{k-1}^{-1}e_1 & X_{k-1}^{-1}G & -X_{k-1}^{-1}G_{k-1} & 2X_{k-1}^{-1}e_1 & 2e_1\end{bmatrix}$$

et

$$\tilde{B}_{k-1} = \begin{bmatrix} A^*e_n & B & A^*B_1 \end{bmatrix} = \begin{bmatrix} A^*e_n & B & A^*X_{k-1}^{-*}B_{k-1} & A^*e_n & A^*X_{k-1}^{-*}e_n \end{bmatrix}.$$

Après calcul du produit $\tilde{G}_{k-1}\tilde{B}_{k-1}^*$ et élimination des termes en trop, on peut prendre pour générateurs de AX_{k-1}^{-1}

$$\tilde{G}_{k-1} = \begin{bmatrix} X_{k-1}^{-1}G & -X_{k-1}^{-1}G_{k-1} & 2e_1 \end{bmatrix} \quad \text{et} \quad \tilde{B}_{k-1} = \begin{bmatrix} B & A^*X_{k-1}^{-*}B_{k-1} & A^*X_{k-1}^{-*}e_n \end{bmatrix}.$$

Comme $X_k = \frac{1}{2}(\mu_{k-1}X_{k-1} + \mu_{k-1}^{-1}X_{k-1}^{-1}A)$, on trouve finalement d'après la proposition 2.3.3,

$$G_k = \frac{1}{2}\begin{bmatrix} \mu_{k-1}G_{k-1} & \mu_{k-1}^{-1}X_{k-1}^{-1}G & -\mu_{k-1}^{-1}X_{k-1}^{-1}G_{k-1} & 2\mu_{k-1}^{-1}e_1 \end{bmatrix}$$

et

$$B_k = \begin{bmatrix} B_{k-1} & B & A^*X_{k-1}^{-*}B_{k-1} & A^*X_{k-1}^{-*}e_n \end{bmatrix}.$$

□

Proposition 3.4.2. *Soit $(X_k)_k$ et $(Y_k)_k$ les matrices itérées de la méthode de Newton-DB* (3.10) *avec X_0, Y_0 premiers termes donnés. Notons pour tout $k \geqslant 0$, G_k, B_k les générateurs associés à X_k et $\tilde{G}_k, \tilde{B}_k$ les générateurs associés à Y_k. Alors G_k, B_k vérifient pour tout $k \geqslant 1$,*

$$G_k = \frac{1}{2}\begin{bmatrix} \mu_k G_{k-1} & -\mu_k^{-1}Y_{k-1}^{-1}\tilde{G}_{k-1} & 2\mu_k^{-1}Y_{k-1}^{-1}e_1 & 2\mu_k^{-1}e_1 \end{bmatrix}$$
$$B_k = \begin{bmatrix} B_{k-1} & Y_{k-1}^{-*}\tilde{B}_{k-1} & e_n & Y_{k-1}^{-*}e_n \end{bmatrix}$$

et $\tilde{G}_k, \tilde{B}_k$ vérifient

$$\tilde{G}_k = \frac{1}{2}\begin{bmatrix} \mu_k\tilde{G}_{k-1} & -\mu_k^{-1}X_{k-1}^{-1}\tilde{G}_{k-1} & 2\mu_k^{-1}X_{k-1}^{-1}e_1 & 2\mu_k^{-1}e_1 \end{bmatrix}$$
$$\tilde{B}_k = \begin{bmatrix} \tilde{B}_{k-1} & X_{k-1}^{-*}B_{k-1} & e_n & X_{k-1}^{-*}e_n \end{bmatrix}.$$

Démonstration. Comme pour la proposition précédente, on reprend les résultats des propositions 2.3.4 et 2.3.3. □

Proposition 3.4.3. *Soit $(X_k)_k$ et $(M_k)_k$ les matrices itérées de la méthode de Newton-DB produit* (3.11) *avec X_0, M_0 premiers termes donnés. Notons pour tout $k \geqslant 0$, G_k, B_k les générateurs associés à X_k et G_{M_k}, B_{M_k} les générateurs associés à M_k pour tout $k \geqslant 0$. Alors G_k, B_k vérifient*

$$G_k = \frac{1}{2}\begin{bmatrix} -\mu_{k-1}^{-2}X_{k-1}M_{k-1}^{-1}G_{M_{k-1}} & 2\mu_{k-1}^{-2}X_{k-1}M_{k-1}^{-1}e_1 & G_{k-1} \end{bmatrix}$$
$$B_k = \begin{bmatrix} M_{k-1}^{-*}B_{M_{k-1}} & e_n & (I + \mu_{k-1}^{-2}M_{k-1}^{-1})^*B_{k-1} \end{bmatrix}$$

et G_{M_k}, B_{M_k} vérifient

$$G_{M_k} = \frac{1}{2}\begin{bmatrix} (2I + \mu_{k-1}^{-2}M_{k-1}^{-1})e_1 & \mu_{k-1}^2G_{M_{k-1}} & -\mu_{k-1}^{-2}M_{k-1}^{-1}G_{M_{k-1}} & 2\mu_{k-1}^{-2}e_1 \end{bmatrix}$$
$$B_{M_{k-1}} = \begin{bmatrix} e_n & B_{M_{k-1}} & M_{k-1}^{-*}B_{M_{k-1}} & M_{k-1}^{-*}e_n \end{bmatrix}$$

Démonstration. — Soient $G_{M_{k-1}}$ et $B_{M_{k-1}}$ des générateurs de la matrice M_{k-1}.A l'aide des propositions 2.3.4 et 2.3.3, les générateurs de la matrice $\frac{1}{2}(\mu_{k-1}^2M_{k-1} + \mu_{k-1}^{-2}M_{k-1}^{-1})$ sont donnés par

$$\frac{1}{2}\begin{bmatrix} \mu_{k-1}^2G_{M_{k-1}} & -\mu_{k-1}^{-2}M_{k_1}^{-1}G_{M_{k-1}} & 2\mu_{k-1}^{-2}M_{k-1}^{-1}e_1 & 2\mu_{k-1}^{-2}e_1 \end{bmatrix}$$

et

$$\begin{bmatrix} B_{M_{k-1}} & M_{k-1}^{-*}B_{M_{k-1}} & e_n & M_{k-1}^{-*}e_n \end{bmatrix}$$

d'où les générateurs de $M_k = \frac{1}{2}\Big(I + \frac{1}{2}(\mu_{k-1}^2 M_{k-1} + \mu_{k-1}^{-2}M_{k-1}^{-1})\Big)$ sont donnés par

$$G_{M_k} = \frac{1}{4}\begin{bmatrix} 4e_1 & \mu_{k-1}^2 G_{M_{k-1}} & -\mu_{k-1}^{-2}M_{k_1}^{-1}G_{M_{k-1}} & 2\mu_{k-1}^{-2}M_{k-1}^{-1}e_1 & 2\mu_{k-1}^{-2}e_1 \end{bmatrix}$$

et

$$B_{M_k} = \begin{bmatrix} e_n & B_{M_{k-1}} & M_{k-1}^{-*}B_{M_{k-1}} & e_n & M_{k-1}^{-*}e_n \end{bmatrix}.$$

Or, en effectuant le produit $G_{M_k}B_{M_k}^*$, on obtient

$$\begin{aligned} G_{M_k}B_{M_k}^* &= \frac{1}{4}\Big(4e_1e_n^* + \mu_{k-1}^2 G_{M_{k-1}}B_{M_{k-1}}^* - \mu_{k-1}^{-2}M_{k-1}^{-1}G_{M_{k-1}}B_{M_{k-1}}^*M_{k-1}^{-1} \\ &\qquad + 2\mu_{k-1}^{-2}M_{k-1}^{-1}e_1e_n^* + 2\mu_{k-1}^{-2}e_1e_n^*M_{k-1}^*\Big) \\ &= \frac{1}{4}\Big((4I + 2\mu_{k-1}^{-2}M_{k-1}^{-1})e_1e_n^* + \mu_{k-1}^2 G_{M_{k-1}}B_{M_{k-1}}^* - \mu_{k-1}^{-2}M_{k-1}^{-1}G_{M_{k-1}}B_{M_{k-1}}^*M_{k-1}^{-1} \\ &\qquad + 2\mu_{k-1}^{-2}e_1e_n^*M_{k-1}^*\Big) \end{aligned}$$

et donc les générateurs de M_k sont donnés par

$$\begin{aligned} G_{M_k} &= \frac{1}{2}\begin{bmatrix} (2I + \mu_{k-1}^{-2}M_{k-1}^{-1})e_1 & \mu_{k-1}^2 G_{M_{k-1}} & -\mu_{k-1}^{-2}M_{k-1}^{-1}G_{M_{k-1}} & 2\mu_{k-1}^{-2}e_1 \end{bmatrix} \\ B_{M_{k-1}} &= \begin{bmatrix} e_n & B_{M_{k-1}} & M_{k-1}^{-*}B_{M_{k-1}} & M_{k-1}^{-*}e_n \end{bmatrix} \end{aligned}$$

— Soient à présent G_{k-1}, B_{k-1} des générateurs de X_k pour un $k \geqslant 0$. A l'aide des proposition 2.3.4 et 2.3.3, les générateurs de la matrice $I + \mu_{k-1}^{-2}M_{k-1}^{-1}$ sont donnés par

$$\begin{bmatrix} 2(I + \mu_{k-1}^{-2}M_{k-1}^{-1})e_1 & -\mu_{k-1}^{-2}M_{k-1}^{-1}G_{M_{k-1}} & 2\mu_{k-1}^{-2}e_1 \end{bmatrix}$$

et

$$\begin{bmatrix} e_n & M_{k-1}^{-*}B_{M_{k-1}} & M_{k-1}^{-*}e_n \end{bmatrix}$$

d'où les générateurs du produit $\frac{1}{2}(I + \mu_{k-1}^{-2}M_{k-1}^{-1})X_{k-1} = X_k$ sont donnés d'après la proposition 2.3.6 par

$$G_k = \frac{1}{2}\begin{bmatrix} -2X_{k-1}e_1 & 2X_{k-1}(I + \mu_{k-1}^{-2}M_{k-1}^{-1})e_1 & -\mu_{k-1}^{-2}X_{k-1}M_{k-1}^{-1}G_{M_{k-1}} & 2\mu_{k-1}^{-2}X_{k-1}e_1 & G_{k-1} \end{bmatrix}$$

et

$$B_k = \begin{bmatrix} (I + \mu_{k-1}^{-2}M_{k-1}^{-1})^*e_n & e_n & M_{k-1}^{-*}B_{M_{k-1}} & M_{k-1}^{-*}e_n & (I + \mu_{k-1}^{-2}M_{k-1}^{-1})^*B_{k-1} \end{bmatrix}.$$

Or, en effectuant le produit $G_kB_k^*$, on obtient

$$\begin{aligned} G_kB_k^* &= \frac{1}{2}\Big(-2X_{k-1}e_1e_n^*(I + \mu_{k-1}^{-2}M_{k-1}^{-1}) + 2X_{k-1}(I + \mu_{k-1}^{-2}M_{k-1}^{-1})e_1e_n^* \\ &\qquad - \mu_{k-1}^{-2}X_{k-1}M_{k-1}^{-1}G_{M_{k-1}}B_{M_{k-1}}^*M_{k-1}^{-1} + 2\mu_{k-1}^{-2}X_{k-1}e_1e_n^*M_{k-1}^{-1} \\ &\qquad + G_{k-1}B_{k-1}^*(I + \mu_{k-1}^{-2}M_{k-1}^{-1})\Big) \\ &= \frac{1}{2}\Big(2\mu_{k-1}^{-2}X_{k-1}M_{k-1}^{-1}e_1e_n^* - \mu_{k-1}X_{k-1}M_{k-1}^{-1}G_{M_{k-1}}B_{M_{k-1}}^*M_{k-1}^{-1} \\ &\qquad + G_{k-1}B_{k-1}^*(I + \mu_{k-1}^{-2}M_{k-1}^{-1})\Big) \end{aligned}$$

et on retrouve le résultat énoncé.

□

Proposition 3.4.4. *Soit $X_{k+1} = G(X_k)$ une suite itérative donnée par la méthode de Newton, Newton-DB ou Newton-DB produit pour une matrice Toeplitz-like A avec premier terme X_0. Alors pour tout $k \geqslant 0$, $\rho(X_k) \leqslant 2^k \varrho(X_0) + (2^k - 1)(\rho(A) + 1)$.*

Démonstration. Rappelons tout d'abord que pour les trois méthodes (Newton, Newton-DB et Newton-DB produit), les matrices X_k vérifient $X_{k+1} = \frac{1}{2}(X_k + AX_k^{-1})$ d'après les définitions de chacune des méthodes. Or, d'après les propriétés du rang de déplacement, $\rho(X_{k+1}) \leqslant 2\rho(X_k) + \rho(A) + 1$. Par récurrence, on peut alors démontrer que $\rho(X_k) \leqslant 2^k \rho(X_0) + (2^k - 1)(\rho(A) + 1)$ pour tout $k \geqslant 0$. □

Proposition 3.4.5. *Comparons à présent les coûts numériques de la méthode de Newton selon le premier terme considéré pour une matrice Toeplitz-like $A \in \mathbb{C}^{n\times n}$. Soit X_0 le premier terme avec $X_0 = cA$ ou $X_0 = p/q(A)$ avec p/q approximant de Padé d'ordre $[2^\ell, 2^\ell - 1]$, $\ell \geqslant 1$. Notons $K_1, K_2 > 0$ les indices pour lesquels X_{K_1} et X_{K_2} atteignent la convergence pour la racine carrée de A avec premier terme $X_0 = cA$ et $X_0 = p/q(A)$ respectivement. Alors cette précision est atteinte en $\mathcal{O}\big(2^{2K_1}\rho(A)^2 n \log^2 n\big)$ opérations lorsque $X_0 = cA$ et $\mathcal{O}\big(2^{2(\ell+K_2)}\rho(A)^2 n \log^2 n\big)$ lorsque $X_0 = p/q(A)$.*

Démonstration. Notons $X_{k+1} = \frac{1}{2}(X_k + AX_k^{-1})$, X_0 donné, une itération de Newton satisfaisant la condition 3.7 et notons pour tout $k \geqslant 0$, $\rho(X_k)$ le rang de déplacement de X_k. D'après les coûts numériques estimées pour les différentes opérations pour les matrices Toeplitz-like, le calcul de X_{k+1} à partir de X_k en arithmétique Toeplitz-like nécessite $\mathcal{O}(\rho(X_k)^2 n \log^2 n)$, coût issu principalement de l'inversion de la matrice X_k, ce qui nous donne par récurrence que le coût de calcul de X_k à partir de X_0 est effectué en $\mathcal{O}(2^{2k}\rho(X_0)^2 n \log^2 n)$ et le coût total dépend donc du rang de déplacement du terme initial X_0. Or, si $X_0 = c.A$, alors $\rho(X_0) = \rho(A)$ pour toute constante $c \neq 0$. Si $X_0 = p/q(A)$, approximant de Padé d'ordre $[2^\ell, 2^\ell - 1]$, alors $\rho(X_0) \leqslant 2^\ell(\rho(A) + 1)$, et nous pouvons conclure avec le résultat énoncé. □

3.4.2 Expériences numériques

Dans cette sous-section, nous présentons plusieurs résultats numériques sur l'implémentation de la méthode itérative de Newton et de ses variantes sur des matrices $A \in \mathbb{C}^{n\times n}$ Toeplitz symétriques définies positives. Afin de mesurer l'efficacité de notre arithmétique Toeplitz-like, nous calculons dans un premier temps les matrices itérées de Newton et ses variantes à l'aide de l'arithmétique Toeplitz-like et nous affichons en ligne continue à chaque itération l'erreur relative $\|I - X_k A^{-1/2}\|_2$ par reconstruction des X_k en arithmétique pleine et où $A^{-1/2}$ est calculée à l'aide des commandes MATLAB, puis nous effectuons la même méthode sur la même matrice pleine A et calculons les matrices itérées de Newton en arithmétique pleine et traçons en ligne pointillée la même erreur relative $\|I - X_k A^{-1/2}\|_2$. Dans les deux cas, nous employons un critère d'arrêt heuristique pour la méthode de Newton et ses variantes à l'aide des résidus. En effet, si $X_k \to \sqrt{A}$ avec la méthode de Newton ou ses variantes, alors $\|X_{k+1}^2 - A\| = \frac{1}{4}\|X_k^{-2}(X_k^4 - 2AX_k^2 + A^2)\|$ d'où $\|I - X_{k+1}A^{-1}X_{k+1}\| = \|(A - X_{k+1}^2)A^{-1}\| = \frac{1}{4}\|X_k^{-2}A(I - X_k A^{-1}X_k)^2\| \leqslant \frac{1}{4}\|X_k^{-2}A\|\,\|I - X_k A^{-1}X_k\|^2$. Or, comme dans le cas de l'étude d'erreur de la méthode de Newton-DB produit, s'il existe $K \geqslant 0$ tel que $\|I - X_K A^{-1}X_K\| \leqslant \frac{1}{4}$, alors $\|I - X_k A^{-1}X_k\| \leqslant \frac{1}{4}$ pour tout $k \geqslant K$. et on s'arrête donc quand $\|I - M_k\| \leqslant \frac{1}{4}$ et

$$\|I - X_{k+1}A^{-1}X_{k+1}\| \geqslant \|I - X_k A^{-1}X_k\|, \tag{3.16}$$

c'est-à-dire que l'on arrête l'itération dès que le résidu croît. De plus, pour chacune des méthodes étudiées en arithmétique Toeplitz-like, nous incorporons l'algorithme de compression des générateurs.

Remarque 3.4.6. *Dans les graphiques suivants, les matrices itérées X_m des méthodes de Newton, Newton-DB et Newton-DB produit sont calculées en arithmétique Toeplitz-like, mais sont reconstruites en arithmétique pleine afin de calculer les erreurs relatives $\|I-X_mA^{-1/2}\|$ et d'afficher le comportement de cette erreur. En pratique, nous pourrons éviter cette reconstruction et construire les matrices itérées X_m, Y_m et M_m en arithmétique Toeplitz-like et nous arrêter lorsque le critère d'arrêt $\|I-X_{m+1}A^{-1}X_{m+1}\| > \|I-X_mA^{-1}X_m\|$ est atteint, soit lorsque le résidu augmente, résidus qui pourront également être calculés avec une complexité $\mathcal{O}(\rho(X_m)n^2)$ pour X_m à l'aide de la commande toeplknorm du package **TLComp**.*

Algorithm 4 Méthode de Newton

Require: $A \in \mathbb{C}^{n\times n}$ symétrique définie positive, X_0 premier terme et μ_k donnés par (3.12) ou (3.15).
Ensure: $X_{\max}$ approximation de $\sqrt{A}$
while $\|I - X_{k+1}A^{-1}X_{k+1}\| \leqslant \frac{1}{2}\|I - X_kA^{-1}X_k\|$ **do**
$X_{k+1} = \frac{1}{2}(\mu_k X_k + \mu_k^{-1}X_k^{-1}A)$;
$(G_{k+1}, B_{k+1}) \leftarrow$ générateurs de X_{k+1} ;
$(\tilde{G}_{k+1}, \tilde{B}_{k+1}) \leftarrow compress(G_{k+1}, B_{k+1})$;
$X_{k+1} \leftarrow \Gamma(\tilde{G}_{k+1}, \tilde{B}_{k+1})$;
end while

Algorithm 5 Méthode de Newton-DB

Require: $A \in \mathbb{C}^{n\times n}$ symétrique définie positive, X_0 premier terme et μ_k donnés par (3.12) ou (3.15).
Ensure: $X_{\max}$ approximation de $\sqrt{A}$
while $\|I - X_{k+1}A^{-1}X_{k+1}\| \leqslant \frac{1}{2}\|I - X_kA^{-1}X_k\|$ **do**
$X_{k+1} = \frac{1}{2}(\mu_k X_k + \mu_k^{-1}Y_k^{-1})$;
$Y_{k+1} = \frac{1}{2}(\mu_k Y_k + \mu_k^{-1}X_k^{-1})$;
$(G_{k+1}, B_{k+1}) \leftarrow$ générateurs de X_{k+1}, $(F_{k+1}, A_{k+1}) \leftarrow$ générateurs de Y_{k+1} ;
$(\tilde{G}_{k+1}, \tilde{B}_{k+1}) \leftarrow compress(G_{k+1}, B_{k+1})$, $(\tilde{F}_{k+1}, \tilde{A}_{k+1}) \leftarrow compress(F_{k+1}, A_{k+1})$;
$X_{k+1} \leftarrow \Gamma(\tilde{G}_{k+1}, \tilde{B}_{k+1})$, $Y_{k+1} \leftarrow \Gamma(\tilde{F}_{k+1}, \tilde{A}_{k+1})$;
end while

Algorithm 6 Méthode de Newton-DB produit

Require: $A \in \mathbb{C}^{n\times n}$ symétrique définie positive, X_0 premier terme, $M_0 = X_0$ et μ_k donnés par (3.12).
Ensure: $X_{\max}$ approximation de $\sqrt{A}$
while $\|I - X_{k+1}A^{-1}X_{k+1}\| \leqslant \frac{1}{2}\|I - X_kA^{-1}X_k\|$ **do**
$M_{k+1} = \frac{1}{2}\Big(I + \frac{\mu_k^2 M_k + \mu_k^{-2}M_k^{-1}}{2}\Big)$;
$X_{k+1} = \frac{1}{2}\mu_k(I + \mu_k^{-2}M_k^{-1})X_k$;
$(G_{k+1}, B_{k+1}) \leftarrow$ générateurs de X_{k+1}, $(F_{k+1}, A_{k+1}) \leftarrow$ générateurs de M_{k+1} ;
$(\tilde{G}_{k+1}, \tilde{B}_{k+1}) \leftarrow compress(G_{k+1}, B_{k+1})$, $(\tilde{F}_{k+1}, \tilde{A}_{k+1}) \leftarrow compress(F_{k+1}, A_{k+1})$;
$X_{k+1} \leftarrow \Gamma(\tilde{G}_{k+1}, \tilde{B}_{k+1})$, $M_{k+1} \leftarrow \Gamma(\tilde{F}_{k+1}, \tilde{A}_{k+1})$;
end while

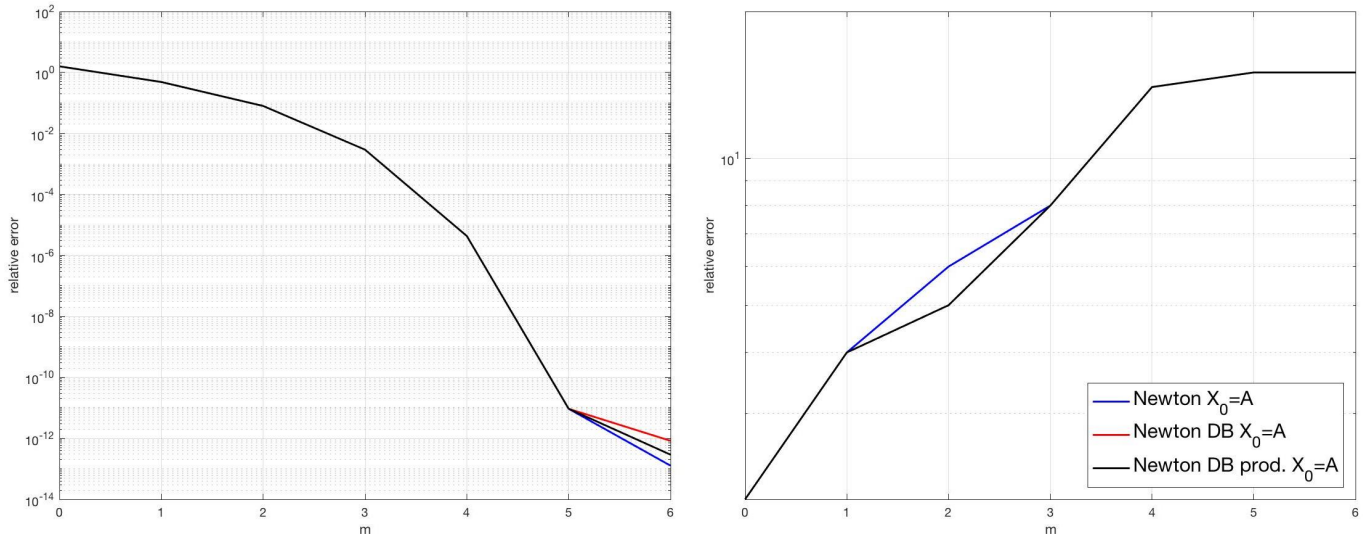

FIGURE 3.3 – Erreurs relatives $\|I - X_m A^{-1/2}\|_2$ à l'aide des méthodes de Newton, Newton DB et Newton DB produit en arithmétique Toeplitz-like pour l'approximation de $\sqrt{A}$ avec $A \in \mathbb{R}^{3000\times 3000}$ matrice de Toeplitz symétrique définie positive à spectre dans $[0,878; 6,696]$ et conditionnement $8,79$ avec $X_0 = A$ et paramètres $\mu_m = 1$.

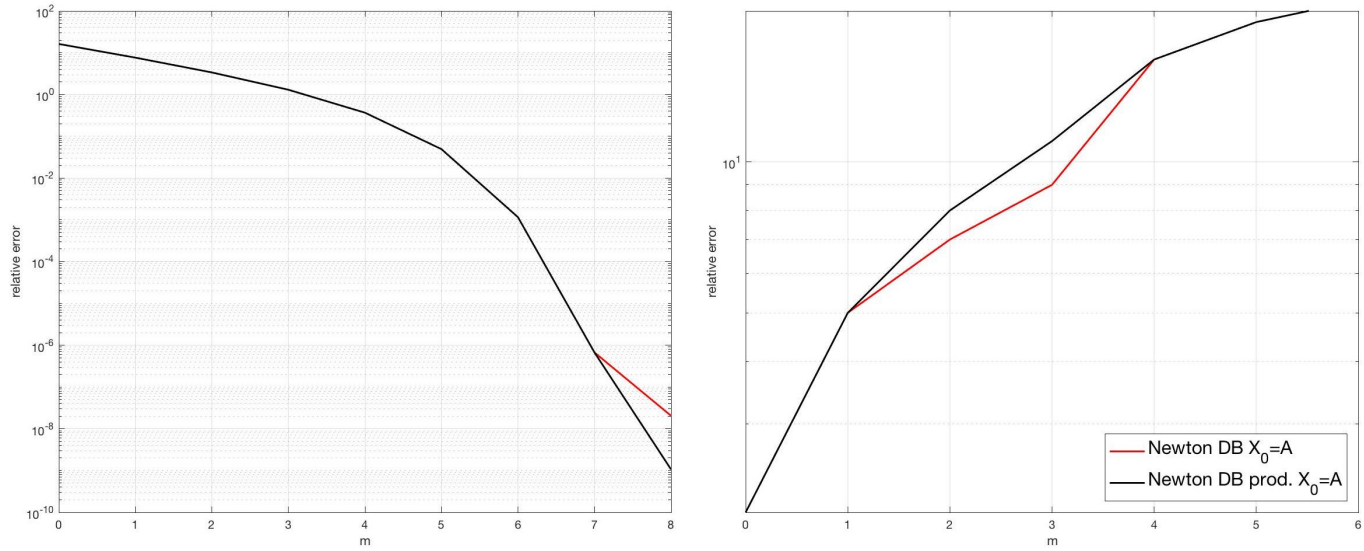

FIGURE 3.4 – Erreurs relatives $\|I - X_m A^{-1/2}\|_2$ à l'aide des méthodes de Newton DB et Newton DB produit en arithmétique Toeplitz-like pour l'approximation de $\sqrt{A}$ avec $A \in \mathbb{R}^{3000\times 3000}$ matrice de Toeplitz symétrique définie positive à spectre dans $[0,169; 295,531]$ et conditionnement $1,7433 \times 10^3$ avec $X_0 = A$ et paramètres $\mu_m = 1$.

Exemple 3.4.7. *En figure 3.3 où nous avons tracé les erreurs relatives* $\|I - X_m A^{-1/2}\|_2$ *avec* $A \in \mathbb{C}^{3000\times 3000}$ *Toeplitz symétrique définie positive avec* $\sigma(A) \subseteq [0,878; 6,696]$ *pour les 3 variantes de la méthode de Newton à gauche et les rang de déplacement des matrices* X_m *de chaque méthode à droite avec* $\mu_m = 1$ *pour tout* m *et* $X_0 = A$*, on observe que les 3 méthodes sont équivalentes dans la convergence pour une matrice avec faible conditionnement . D'après la proposition 3.4.4, on sait que pour tout* $m \geqslant 0$ *le rang de déplacement des matrices* X_m *des différentes formes de la méthode de Newton vérifient* $\rho(X_m) \leqslant 2^m\rho(X_0)+(2^m-1)(\rho(A)+1)$*. Or on observe en figure 3.3 sur le graphique de droite que le rang de déplacement calculé sur machine semble se stabiliser après 5 itérations. Ce phénomène s'explique par le fait que l'on observe ici non pas le*

rang de déplacement théorique mais le rang de déplacement numérique de nos matrices itérées qui lui est inférieur. Plus précisément, ce rang de déplacement semble se stabiliser à partir d'un certain rang < 20*. Par conséquent, chacune de nos matrices itérées pour les différentes variantes de la méthode itérative de Newton est construite en un* $\mathcal{O}(n \log^2 n)$ *opérations arithmétiques.*

En figure 3.4, en raison du conditionnement supérieur à 9, on ne peut plus utiliser la méthode de Newton, et nous employons donc les méthodes de Newton-DB et Newton-DB produit avec $\mu_m = 1$ *et* $X_0 = A$ *en observant les mêmes données que pour la figure précédente à savoir les erreurs relatives* $\|I - X_m A^{-1/2}\|_2$ *pour* $A \in \mathbb{C}^{3000\times 3000}$ *Toeplitz symétrique définie positive avec* $\sigma(A) \subseteq [0,169; 295,531]$ *et le rang de déplacement des matrices* X_m*. Si globalement les observations restent les mêmes que pour le graphique précédent, on voit cependant ici qu'à partir de m=7, on perd un petit peu en précision par la méthode de Newton-DB par rapport à la méthode de Newton-DB produit.*

Dans les graphiques suivants, nous présentons comme précédemment les erreurs relatives $\|I - X_m A^{-1/2}\|$ avec X_m matrice itérée de la méthode de Newton avec le critère d'arrêt (3.16) mais cette fois-ci en comparant 2 arithmétiques : nous représentons les erreurs relatives pour les X_k calculées en arithmétique Toeplitz-like en ligne pleine et les erreurs relatives pour les X_k calculées en arithmétique pleine en ligne pointillée afin de les comparer.

Exemple 3.4.8. *En figure 3.5 nous présentons des essais numériques des trois différentes formes de la méthode de Newton sur des matrices A de taille* 3000×3000 *Toeplitz symétriques définies positives de différents conditionnements avec différents premiers termes* X_0 *:* $X_0 = A$*,* $X_0 = \frac{1}{\sqrt{\lambda_{\min}}} A$*,* $X_0 = (p/q)_{1,2}(A)$*,* $X_0 = (p/q)_{7,8}(A)$ *et* $X_0 = (p/q)_{15,16}(A)$ *avec* $(p/q)_{m-1,m}$ *approximant de Padé d'ordre* $[m|m-1]$*. Nous affichons l'erreur relative* $\|I - X_k(\sqrt{A})^{-1}\|_2$ *pour les 2 arithmétiques à gauche et le rang de déplacement numérique de chaque matrice* X_k *en arithmétique Toeplitz-like à droite. Pour la première matrice* $A \in \mathbb{C}^{n\times n}$ *avec* $\sigma(A) \subseteq [0,75; 5,69]$ *et conditionnement* $7,537$ *(graphe supérieur), on observe d'abord que la convergence des itérées de Newton en arithmétique Toeplitz-like est quasi identique à l'arithmétique pleine et la méthode converge de manière générale vers la fonction de matrice* $\sqrt{A}$ *jusqu'à une tolérance* $< 10^{-12}$*. On observe également que le choix du premier terme* X_0 *va grandement impacter sur le nombre d'itérations nécessaires à la convergence : en effet, pour un premier terme* $X_0 = cA$ *avec* $c = 1, (\lambda_{\min})^{-1/2}$*, il est nécessaire d'effectuer 5 ou 6 itérations avant d'atteindre une tolérance* 10^{-12}*, alors que pour un premier terme* $X_0 = (p/q)_{8,7}(A)$ *ou* $X_0 = (p/q)_{16,15}(A)$*, une seule itération suffit pour atteindre cette tolérance. On peut noter ici qu'un ordre* $(8,7)$ *pour nos approximants de Padé est suffisant et que prendre un approximant de Padé d'ordre* $(16,15)$ *ne peut accélérer la convergence et augmente le nombre d'opérations à effectuer pour obtenir le premier terme* $X_0 = (p/q)_{16,15}(A)$ *par rapport à un choix* $X_0 = (p/q)_{8,7}(A)$*. Par conséquent, dans le cas de cette matrice, le choix d'un premier terme* $X_0 = (p/q)_{8,7}(A)$ *semble être préférable. D'autre part, le rang de déplacement numérique bien que croissant, stagne à partir d'un certain rang à une valeur de 16 et ce pour tout type de premier terme* X_0*. Cet observation nous permet de conclure qu'une matrice itérée atteignant la tolérance* 10^{-12} *en arithmétique Toeplitz-like peut alors être reconstruite à l'aide de nos algorithmes avec une complexité de l'ordre de* n^2*, assurant ainsi dans le cas de cette matrice bien conditionnée, une méthode d'approximation de la racine carrée avec une complexité globale de l'ordre de* n^2*.*
Des résultats similaires peuvent être observées pour la matrice suivante $A \in \mathbb{C}^{3000\times 3000}$ *où* $\sigma(A) \subseteq [0,72; 228,54]$ *et de conditionnement de* $317,416$ *avec une tolérance de* 10^{-11} *pouvant être atteinte à l'aide de la méthode Newton DB.*

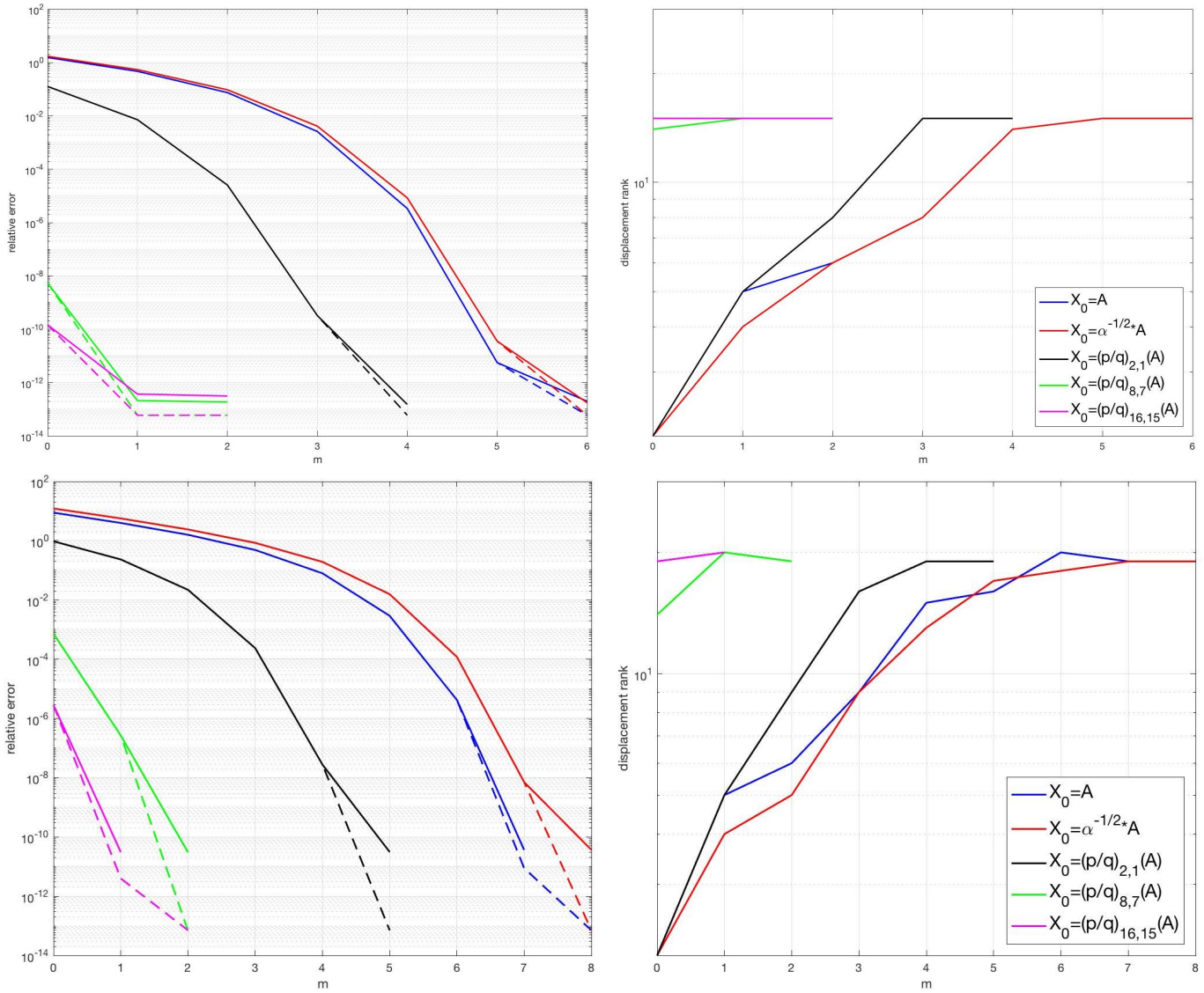

FIGURE 3.5 – Erreurs relatives $\|I - X_m A^{-1/2}\|_2$ et rangs de déplacement des matrices X_m par les méthodes de Newton et Newton DB en arithmétique Toeplitz-like (ligne pleine) et arithmétique pleine (ligne pointillée) pour l'approximation de $\sqrt{A}$ avec $A \in \mathbb{R}^{3000\times 3000}$ matrices de Toeplitz symétrique définie positive respectivement à spectre dans $[0,55; 96,52]$ et conditionnement $8,0445$ (graphique supérieur), $[0,62; 115,4]$ et conditionnement $173,7936$ (graphique inférieur) pour différents premiers termes X_0 et paramètres $\mu_m = 1$.

Remarque 3.4.9. *En pratique, plutôt que de reconstruire chaque matrice X_m en arithmétique pleine pour déterminer l'erreur relative pour la racine carrée principale à l'aide de la norme 2 nécessitant $\mathcal{O}(n^3)$ opérations, nous proposons le schéma suivant : en partant d'un premier terme X_0 pour une matrice $A \in \mathbb{C}^{n\times n}$ dont on souhaite approcher la racine carrée à l'aide des différentes formes de la méthode de Newton énoncées dans ce chapitre, on calcule les matrices X_m d'après les schémas itératifs en arithmétique Toeplitz-like. A chaque itération, nous mesurons le résidu $\|I - X_m A^{-1} X_m\|$ et si $\|I - X_m A^{-1} X_m\| \leqslant \|I - X_{m-1} A^{-1} X_{m-1}\|$, alors on continu en arithmétique Toeplitz-like, sinon on arrête l'itération, la tolérance minimale possible ayant été atteinte. Pour mesure ce résidu, on emploie la commande toeplknorm((I − $X_m A^{-1} X_m$).G,(I − $X_m A^{-1} X_m$).B,inf) du package* ***TLComp*** *qui mesure le résidu $\|I - X_m A^{-1} X_m\|_{\text{inf}}$ en arithmétique Toeplitz-like avec une complexité $\mathcal{O}(\rho(I - X_m A^{-1} X_m)n^2)$, qui étant donné que $I - X_m A^{-1} X_m$ est Toeplitz-like à chaque itération pour un faible nombre m, réduit donc la complexité. On peut alors reconstruire ou non par la suite la matrice X_m ayant atteint la plus faible tolérance en fonction des besoins.*

Pour rendre compte du gain engendré par ce schéma, nous le testons sur plusieurs matrices de Toeplitz symétriques définies positives de taille 3000×3000*, en mesurant le temps nécessaire (en secondes) pour atteindre le critère d'arrêt* (3.16) *en arithmétique Toeplitz-like et pleine en fonction de trois premiers termes* X_0*, dont les résultats pour trois matrices sélectionnées sont entrées dans le tableau suivant :*

Matrice $A \in \mathbb{C}^{n\times n}$	*premier terme*	*arithmétique TL*	*arithmétique pleine*
$\mathrm{cond}(A) = 5,82$	$X_0 = A$	$12,86$ *s.*	$841,01$ *s.*
	$X_0 = \frac{1}{\sqrt{\lambda_{\min}}}A$	$10,52$ *s.*	$809,38$ *s.*
	$X_0 = (p/q)_{2,1}(A)$	$10,23$	$477,45$
$\mathrm{cond}(A) = 70,28$	$X_0 = A$	$121,26$ *s.*	$1,04 \times 10^3$ *s.*
	$X_0 = \frac{1}{\sqrt{\lambda_{\min}}}A$	$92,44$ *s.*	$880,91$ *s.*
	$X_0 = (p/q)_{2,1}(A)$	$20,03$ *s.*	$566,58$
$\mathrm{cond}(A) = 3,24 \times 10^3$	$X_0 = A$	$37,68$ *s.*	$1082,2$ *s.*
	$X_0 = \frac{1}{\sqrt{\lambda_{\min}}}A$	$69,11$ *s.*	$1176,4$ *s.*
	$X_0 = (p/q)_{2,1}(A)$	$138,7$	$815,90$

On peut alors observer dans ce tableau que le temps nécessaire pour une implémentation en arithmétique Toeplitz-like est significativement plus petit que le temps nécessaire pour une implémentation en arithmétique pleine, rendant compte ainsi du gain sur le nombre d'opérations nécessaires à effectuer pour atteindre une convergence.

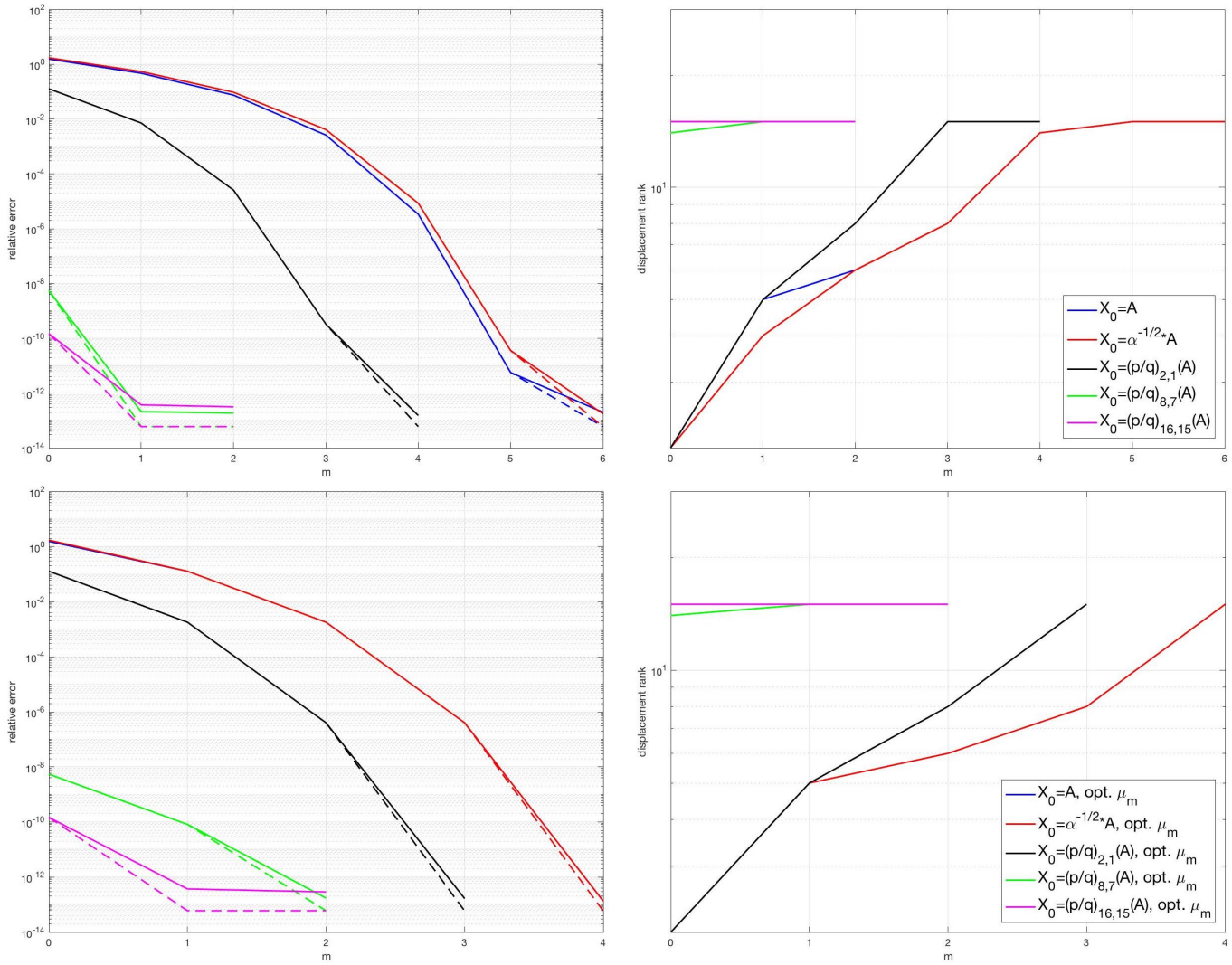

FIGURE 3.6 – Erreurs relatives $\|I - X_m A^{-1/2}\|_2$ et rangs de déplacement des itérées X_m calculées à l'aide de la méthode de Newton-DB en arithmétique Toeplitz-like (ligne pleine) et arithmétique pleine (ligne pointillée) avec paramètres $\mu_m = 1$ (graphique supérieur) et μ_k donnés par (3.12) ou (3.15) (graphique inférieur) pour l'approximation de $\sqrt{A}$ avec $A \in \mathbb{R}^{3000\times 3000}$ matrice de Toeplitz symétrique définie positive à spectre dans $[0,75; 5,69]$ et conditionnement $7,537$ pour différents premiers termes X_0.

Exemple 3.4.10. *Considérons à présent l'introduction de paramètres μ_k dans la méthode de Newton. Nous testons en figure 3.6 la méthode de Newton classique avec et sans paramètres* (3.12) *ou* (3.15)*, toujours avec les différents choix possibles de premier terme X_0, pour la matrice $A \in \mathbb{C}^{3000\times 3000}$ de Toeplitz symétrique définie positive avec $\sigma(A) \subseteq [0,75; 5,69]$ et conditionnement $7,537$. L'introduction de paramètres μ_k donnés par* (3.12) *permet dans le cas où $X_0 = A$ ou $X_0 = \lambda_{\min}^{-1/2} A$ de réduire de deux itérations le nombre d'itérations nécessaires pour atteindre une tolérance 10^{-12}. De même dans le cas où $X_0 = \lambda_{\min}^{-1/2} A$, les paramètres μ_k donnés par* (3.15) *permet dans le cas où $X_0 = (p/q)_{2,1}(A)$ de réduire de deux itérations le nombre d'itérations nécessaires pour atteindre cette même tolérance. Nous observons donc bien une accélération de la convergence à l'aide des paramètres que ce soit en arithmétique Toeplitz-like ou pleine, mais ce phénomène ne peut s'observer dans le cas où $X_0 = (p/q)_{8,7}(A)$ ou $X_0 = (p/q)_{16,15}(A)$ puisque pour ces choix de premier terme, la convergence est déjà très rapide. On remarque également que le rang de déplacement n'est pas influencé par l'introduction de paramètres et stagne à la même valeur que pour la méthode de Newton sans paramètres. Des observations similaires peuvent être faites en figure 3.7 pour une matrice $A \in \mathbb{C}^{3000\times 3000}$ de Toeplitz*

symétrique définie positive avec $\sigma(A) \subseteq [0, 14, 47, 01]$ *et conditionnement* $335, 78$ *pour laquelle on emploie la méthode de Newton DB avec ou sans paramètres* (3.12) *et* (3.15).

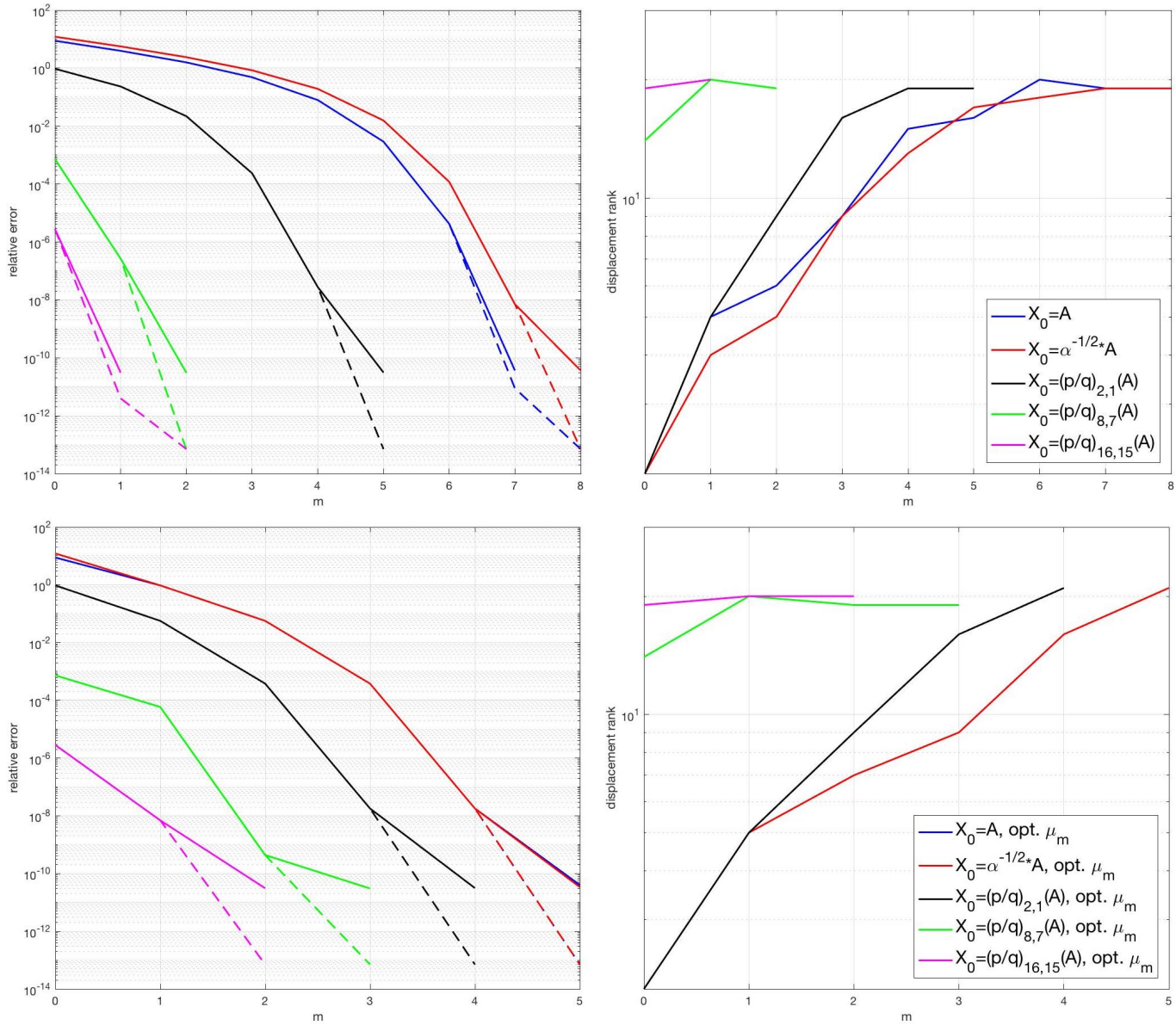

FIGURE 3.7 – Erreurs relatives $\|I - X_m A^{-1/2}\|_2$ et rangs de déplacement des itérées X_m calculées à l'aide de la méthode de Newton-DB avec ou sans paramètres (3.12) ou (3.15) en arithmétique Toeplitz-like (ligne pleine) et arithmétique pleine (ligne pointillée) pour l'approximation de $\sqrt{A}$ avec $A \in \mathbb{R}^{3000\times 3000}$ matrice de Toeplitz symétrique définie positive à spectre dans $[0, 55; 96, 52]$ et conditionnement $173, 7936$ pour différents premiers termes X_0 avec paramètres $\mu_m = 1$ (graphique supérieur) et μ_m donnés par (3.12) ou (3.15) (graphique inférieur).

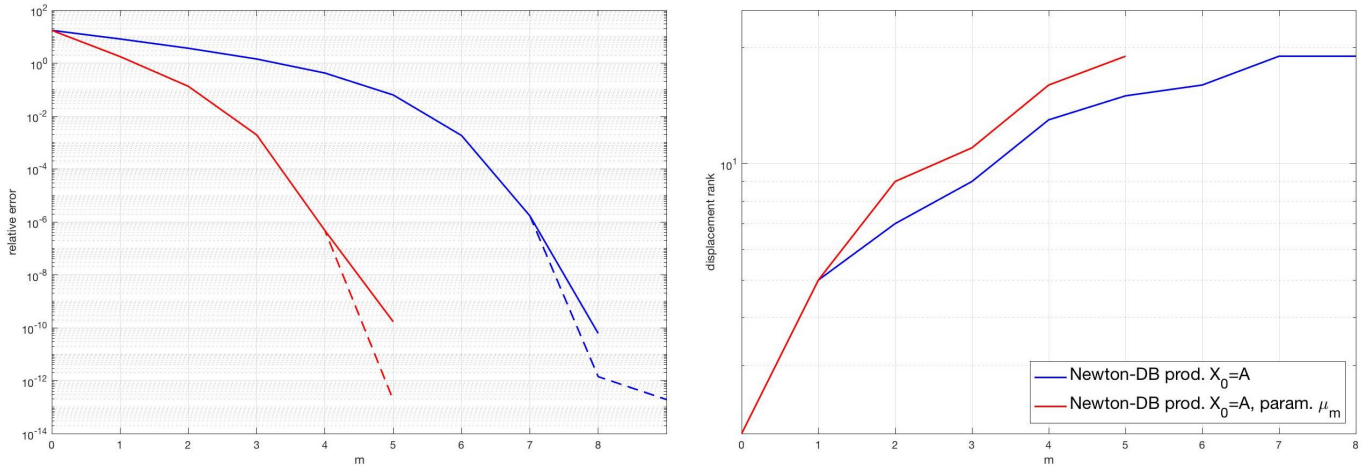

FIGURE 3.8 – Erreurs relatives $\|I - X_m A^{-1/2}\|_2$ à l'aide de la méthode de Newton DB produit en arithmétique Toeplitz-like (ligne pleine) et arithmétique pleine (ligne pointillée) pour l'approximation de $\sqrt{A}$ avec $A \in \mathbb{R}^{3000\times 3000}$ matrice de Toeplitz symétrique définie positive à spectre dans $[0,4; 335,771]$ et conditionnement $821,0866$ avec $\mu_m = 1$ pour tout m.

Exemple 3.4.11. *En figure 3.8 nous implémentons la méthode de Newton-DB produit sur une matrice $A \in \mathbb{C}^{3000\times 3000}$ de Toeplitz symétrique définie positive avec $\sigma(A) \subseteq [0,4; 335,771]$ et conditionnement $821,0866$ en arithmétique Toeplitz-like et pleine, avec et sans paramètres. On observe alors deux phénomènes pour l'erreur d'approximation : premièrement, l'introduction des paramètres nous permet de réduire considérablement le nombre d'itérations nécessaires pour la méthode Newton-DB produit, passant de 8 à 5 itérations nécessaires pour atteindre la convergence en arithmétique Toeplitz-like. Deuxièmement, on voit ici que l'arithmétique Toeplitz-like ne permet pas d'atteindre la même tolérance qu'avec l'arithmétique pleine, puisqu'en arithmétique Toeplitz-like nous atteignons une tolérance de 10^{-10} alors que l'arithmétique pleine permet d'atteindre une tolérance de 10^{-12}. Cette différence peut être s'explique par la perte de précision en passant à l'arithmétique Toeplitz-like.*

Exemple 3.4.12. *En figure 3.9 nous calculons les matrices itérées X_m les méthode Newton DB et Newton DB produit pour une matrice A avec conditionnement d'ordre 10^3 avec différents premiers termes et affichons les erreurs relative $\|I - X_m A^{-1/2}\|$ et les rangs de déplacement de chaque matrice X_m. On remarque alors encore une fois que dans le cas de l'arithmétique Toeplitz-like, la convergence des matrices itérées vers la racine carrée de matrice ne peut être assurée pour toute précision souhaitée contrairement à l'arithmétique pleine.*

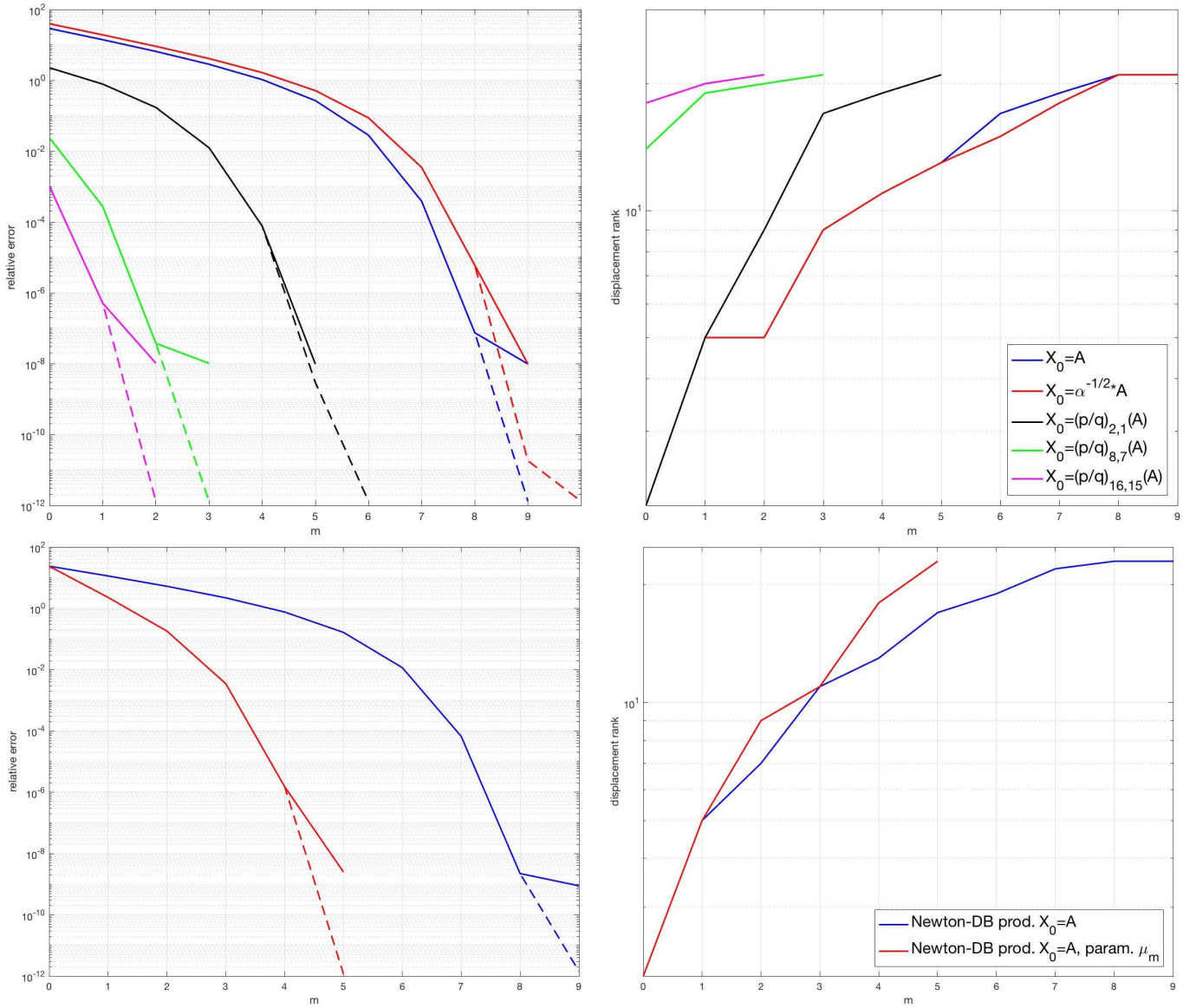

FIGURE 3.9 – Erreurs relatives $\|I - X_m A^{-1/2}\|_2$ et rang de déplacement des matrices itérées des méthodes Newton-DB et Newton-DB produit en arithmétique Toeplitz-like (lignes pleines) et arithmétique pleine (lignes pointillées) pour l'approximation de $\sqrt{A}$ avec $A \in \mathbb{R}^{3000\times 3000}$ matrices de Toeplitz symétrique définie positive respectivement à spectre dans $[0,552; 896,714]$ et conditionnement d'ordre $1,6226 \times 10^3$ pour différents premiers termes X_0 et paramètres $\mu_m = 1$.

En conclusion, nous avons donc vu que la méthode de Newton et ses variantes implémentées en arithmétique Toeplitz-like dans le cas de matrices $A \in \mathbb{C}^{n\times n}$ de Toeplitz symétriques définies positives bien conditionnées permettaient d'obtenir un approximant de la fonction de matrice $\sqrt{A}$. En particulier dans le cas des méthode de Newton et Newton-DB, en prenant $X = (p/q)_{8,7}(A)$, nous obtenons une bonne approximation en deux ou trois itérations en moyenne, et pouvons accélérer la convergence à l'aide des paramètres (3.15) sans modifier le rang de déplacement de la matrice itérée finale. Dans le cas de la méthode de Newton-DB produit, nous démarrons avec $X_0 = A$ et employons les paramètres (3.15) afin d'obtenir une bonne approximation en 5 ou 6 itérations. La méthode de Newton-DB produit n'améliore pas la convergence par rapport à la méthode de Newton-DB. Par conséquent nous pouvons conclure que pour approcher la fonction de matrice $A \in \mathbb{C}^{n\times n}$ Toeplitz symétrique définie positive à l'aide de l'arithmétique Toeplitz-like,

un des meilleurs moyens est d'employer la méthode de Newton-DB avec premier terme $X_0 = (p/q)_{8,7}(A)$. En supposant que l'on utilise uniquement l'arithmétique Toeplitz-like avec le critère d'arrêt (3.16) sans reconstruction, on obtient la matrice X_K de faible rang de déplacement étant donné le faible nombre d'itérations nécessaires avant d'atteindre le critère d'arrêt, construite avec une complexité $\mathcal{O}(n \log^2 n)$ voir $\mathcal{O}(n^2)$ si l'on souhaite la reconstruire en arithmétique pleine. Pour des matrices moins bien conditionnées, la méthode de Newton et ses variantes ne fonctionnent pas aussi bien qu'en arithmétique pleine et on perd de la précision en atteignant la convergence. Cette observation provient de la stabilité de l'arithmétique Toeplitz-like étant donné que l'on peut observer en figure 3.9 que la méthode de Newton-DB et Newton-DB produit calculées en arithmétique pleine permettent d'atteindre une tolérance de 10^{-12} contrairement à leur calcul en arithmétique Toeplitz-like.

3.5 La fonction de matrice signe

Considérons à présent la fonction signe, définie pour tout $z \in \mathbb{C}\backslash i\mathbb{R}$ par $\text{sign}(z) = 1$ si $\Re(z) > 0$ et $\text{sign}(z) = -1$ si $\Re(z) < 0$. Pour toute matrice $A \in \mathbb{C}^{n\times n}$ avec $\sigma(A) \subset \Omega = \mathbb{C}\backslash i\mathbb{R}$, la fonction de matrice $\text{sign}(A)$ introduite par Roberts [78] comme outil pour la résolution des équations de Lyapounov et équations algébriques de Riccati peut être définie à partir de sa décomposition de Jordan par :

$$\text{sign}(A) = Z \begin{bmatrix} -I_p & 0 \\ 0 & I_q \end{bmatrix} Z^{-1}, \text{ lorsque } A = Z\,\text{diag}(J_1, J_2)Z^{-1}$$

avec $\sigma(J_1) = \{\lambda \in \sigma(A) : \Re(\lambda) < 0\}$ et $\sigma(J_2) = \{\lambda \in \sigma(A) : \Re(\lambda) > 0\}$.

Comme pour le cas de la racine carrée principale, pour calculer $\text{sign}(A)$, il est possible d'utiliser l'algorithme de Schur-Parlett, mais qui possède une complexité de $28\frac{2}{3}n^3$. L'approximation de la fonction de matrice $\text{sign}(A)$ trouve plusieurs applications en mathématiques appliquées, notamment en chromodynamique quantique (voir [88]). D'autres applications peuvent être trouvées dans [57, Chap. 5]. Or, la fonction signe est toujours reliée à la fonction racine carrée principale puisque pour tout $a \in \mathbb{C}\backslash i\mathbb{R}$, $a = \text{sign}(a)(a^2)^{1/2}$ avec $z \mapsto z^{1/2}$ la racine carrée principale, et dans le cas matriciel, on peut noter [57, Section 5.1],

$$A = SN, \text{ avec } S = \text{sign}(A) \text{ et } N = (A^2)^{1/2}$$

où les matrices S, N et A commutent entre elles et cette définition est unique dès que A est hermitienne, sinon on considère la matrice $H = (A^*A)^{1/2}$.
De cette observation les méthodes précédentes pour la racine carrée principale peuvent être adaptées pour la fonction signe. Dans cette section, nous voyons comment une méthode de Newton similaire à celle pour la racine carrée nous permet l'approximation de la fonction de matrices signe. Puis nous étudions l'adaptation de cette méthode à l'arithmétique Toeplitz-like, les générateurs associés et le comportement du rang de déplacement, le tout étant ensuite illustré par plusieurs expériences numériques.

3.5.1 Rappel sur la méthode de Newton pour la fonction signe

Du lien entre le signe et la racine carrée principale d'une matrice, les méthodes de Newton pour ces deux fonctions peuvent être reliées de la manière suivante :

Théorème 3.5.1. *[57, Theorem 6.11] Soit $A \in \mathbb{C}^{n\times n}$ telle que $\sigma(A) \cap \mathbb{R}_- = \emptyset$. Considérons toute itération $X_{k+1} = g(X_k) = X_k h(X_k^2)$ convergeant vers $\text{sign}(X_0)$ pour $X_0 = \begin{bmatrix} 0 & A \\ I & 0 \end{bmatrix}$ à l'ordre m. Alors les matrices*

itérées définies par

$$\begin{cases} Y_{k+1} = Y_k h(Z_k Y_k), & Y_0 = A, \\ Z_{k+1} = h(Z_k Y_k) Z_k & Z_0 = I \end{cases}$$

vérifient $Y_k \to A^{1/2}$ *et* $Z_k \to A^{-1/2}$ *lorsque* $k \to \infty$, Y_k *commute avec* Z_k *et* $Y_k = AZ_k$ *pour tout* k.

Ainsi, comme pour la racine carrée, on peut faire appel à une version de la méthode de Newton pour la fonction signe :

$$X_{k+1} = \frac{1}{2}(X_k + X_k^{-1}), \;\; X_0 = A, \tag{3.17}$$

introduite par Roberts et obtenue par dérivation de l'équation $X^2 = I$.

Théorème 3.5.2. *Soit* $A \in \mathbb{C}^{n\times n}$ *telle que* $\sigma(A) \cap i\mathbb{R} = \emptyset$. *Alors les matrices itérées de Newton données par* (3.17) *convergent quadratiquement vers* $S = \operatorname{sign}(A)$ *avec*

$$\|X_{k+1} - S\| \leqslant \frac{1}{2}\|X_k^{-1}\|\|X_k - S\|^2 \tag{3.18}$$

pour toute norme consistante.

On peut de plus mesurer la vitesse de convergence des matrices itérées de Newton plus précisément [57, Theorem 5.6] : en effet, considérons la suite de matrices $G_0 = (A-S)(A+S)^{-1}, G_k = (X_k - S)(X_k + S)^{-1}$. Or, pour tout $k \geqslant 0$,

$$X_{k+1} \pm S = \frac{1}{2}X_k^{-1}(X_k^2 \pm 2X_k S + I) = \frac{1}{2}X_k^{-1}(X_k \pm S)$$

d'où $(X_{k+1}-S)(X_{k+1}+S)^{-1} = ((X_k-S)(X_k+S)^{-1})^2$, ce qui nous permet de noter $G_{k+1} = G_k^2 = \ldots = G_0^{2^{k+1}}$ et $X_k = (I - G_k)^{-1}(I + G_k)S$ pour tout $k \geqslant 0$. De plus, G_0 a pour valeurs propres les $\gamma = \frac{\lambda - \operatorname{sign}(\lambda)}{\lambda + \operatorname{sign}(\lambda)}$ où $\lambda \in \sigma(A)$ d'où $\gamma \in Int(\mathbb{D})$, soit $\rho(G_0) < 1$ puisque A ne possède aucune valeur propre imaginaire pure et donc $G_k \underset{k\to+\infty}{\longrightarrow} 0$ et pour tout $k \geqslant 0$,

$$\|G_0^{2^k}\| \geqslant \varrho(G_0^{2^k}) = \Big(\max_{\lambda\in\sigma(A)} \Big|\frac{\lambda - \operatorname{sign}(\lambda)}{\lambda + \operatorname{sign}(\lambda)}\Big|\Big)^{2^k}.$$

Il apparaît alors que si $\varrho(A) \gg 1$ ou si A possède des valeurs propres proches de l'axe imaginaire alors la convergence sera lente.

3.5.2 Introduction de paramètres

Pour accélérer la convergence des matrices itérées de la méthode de Newton, on peut également introduire des paramètres μ_k et remplacer X_k dans (3.17) par $\mu_k X_k$, ce qui nous donne l'itération

$$X_{k+1} = \frac{1}{2}(\mu_k X_k + \mu_k^{-1} X_k^{-1}), \quad X_0 = A, \tag{3.19}$$

où les μ_k doivent être réels positifs, de sorte à préserver le signe des matrices X_k. N. J. Higham [57, Section 5.5] suggère 3 choix de paramètres :

i. paramètres déterminantaux : $\mu_k = |\det(X_k)|^{-1/n}$;

ii. paramètres spectraux : $\mu_k = \sqrt{\varrho(X_k^{-1})/\varrho(X_k)}$,

iii. paramètres normalisés : $\mu_k = \sqrt{\|X_k^{-1}\|/\|X_k\|}$.

3.5.3 Newton signe pour les matrices Toeplitz-like et expériences numériques

De la même manière que pour la méthode de Newton pour la fonction racine carrée, on peut déterminer les générateurs associés à chaque itération : soit $A \in \mathbb{C}^{n\times n}$ Toeplitz-like avec $\sigma(A) \in \mathbb{C}\backslash i\mathbb{R}$ et supposons que l'on souhaite approcher $\text{sign}(A)$ à l'aide de la méthode de Newton pour le signe et notons $(X_k)_k$ la suite des matrices itérées de Newton définis par (3.19) et G_k, B_k les générateurs associés.

Proposition 3.5.3. *Les suites de générateurs $(G_k)_k$ et $(B_k)_k$ sont donnés par récurrence*

$$G_k = \frac{1}{2}\begin{bmatrix}\mu_k G_{k-1} & -\mu_k^{-1}X_{k-1}^{-1}G_{k-1} & 2\mu_k^{-1}X_{k-1}^{-1}e_1 & 2\mu_k^{-1}e_1\end{bmatrix}$$
$$B_k = \begin{bmatrix}B_{k-1} & X_{k-1}^{-*}B_{k-1} & e_n & X_{k-1}^{-*}e_n\end{bmatrix}$$

En particulier, si $e_1 \in \Im(G)$ ou $e_n \in \Im(B)$, alors $e_1 \in \Im(G_k)$ ou $e_n \in \Im(B_k)$ pour tout $k \geqslant 0$.

Démonstration. Pour la démonstration, on utilise les propositions 2.3.4 et 2.3.3 et on trouve le résultat énoncé. □

On peut noter pour tout $k \geqslant 1$ que

$$X_{k+1} = r(X_k) = r_k(A) \text{ avec } r(x) = \frac{1}{2}(\frac{x^2+1}{x}) \in \mathcal{R}_{2,1}.$$

En notant $(a_k, b_k) = (\deg(p_k), \deg(q_k)$ lorsque $r_k = \frac{p_k}{q_k}$, on peut montrer par récurrence que $(a_k, b_k) = (2\times 3^{k-1}, 3^{k-1})$. Par conséquent, en considérant que le nombre d'itérations K nécessaires à la convergence soit négligeable devant la dimension, le calcul de l'approximant rationnel $r_K(A)$ de $\text{sign}(A)$ peut être effectué en $\mathcal{O}(2\times 3^{K-2}\varrho n\log^2 n)$ avec ϱ le rang de déplacement de la matrice A. Il nous reste alors à voir comment les générateurs associés aux itérations se comportent, en particulier dans le cas de Toeplitz.

Proposition 3.5.4. *Soit $A \in \mathbb{C}^{n\times n}$ matrice de Toeplitz. Alors en notant $X_k = r_k(A)$ les matrices itérées de Newton pour le signe avec $X_0 = A$ et $K \geqslant 1$ le rang de la matrice itérée approximant du signe de A satisfait $\rho(r_K(A)) \leqslant 2^{K+1}$.*

Démonstration. Soit $(X_k)_k$ la suite des matrices itérées de Newton pour la fonction signe. Alors d'après les propriétés du rang de déplacement, on a pour tout $k \geqslant 0$,

$$\rho(X_{k+1}) \leqslant \rho(X_k) + \rho(X_k) + 1 = 2\rho(X_k) + 1.$$

Par récurrence, on montre que $\rho(X_K) \leqslant 2^{K-1}\rho(X_0) + \sum_{j=0}^{K-1} 2^j = \sum_{j=0}^{K-1} 2^j = 2^{K+1}$. □

Pour le choix des paramètres μ_k, les paramètres déterminantaux de ne sont pas compatibles avec notre arithmétique Toeplitz-like puisque le calcul du déterminant par transformation des matrices en arithmétique Toeplitz-like en matrices pleines nous ferait perdre gain de complexité de l'arithmétique Toeplitz-like. Les paramètres spectraux sont également écartés, puisque si en théorie la symétrie des matrices itérées est assurée, en pratique sur machine la propriété de symétrie n'est pas nécessairement conservée avec la précision machine. En revanche, nous pouvons considérer les paramètres normalisés à l'aide de la commande $toeplknorm(T.G, T.B, p)$ où $p = 1, \infty, 'fro'$ ($'fro'$ pour la norme de Fröbenius) indique le type de norme mesurée, commande issue du package *TLComp* pour l'arithmétique Toeplitz-like, avec un coût de l'ordre $\mathcal{O}(\varrho n^2)$ avec ϱ le rang de déplacement de la matrice considérée.

Algorithm 7 Méthode de Newton signe

Require: $A \in \mathbb{C}^{n\times n}$ symétrique définie positive, X_0 premier terme et $\mu_k = 1$ ou $\mu_k = \sqrt{\|X_k^{-1}\|/\|X_k\|}$ pour tout $k \geqslant 0$.

Ensure: $X_{\max}$ approximation de $\mathrm{sign}(A)$

while $\|I - X_{k+1}^2\| \leqslant \frac{1}{2}\|I - X_k^2\|$ **do**

$X_{k+1} = \frac{1}{2}(\mu_k X_k + \mu_k^{-1} X_k^{-1})$;

$\mu_{k+1} = \sqrt{\|X_{k+1}^{-1}\|/\|X_{k+1}\|}$

end while

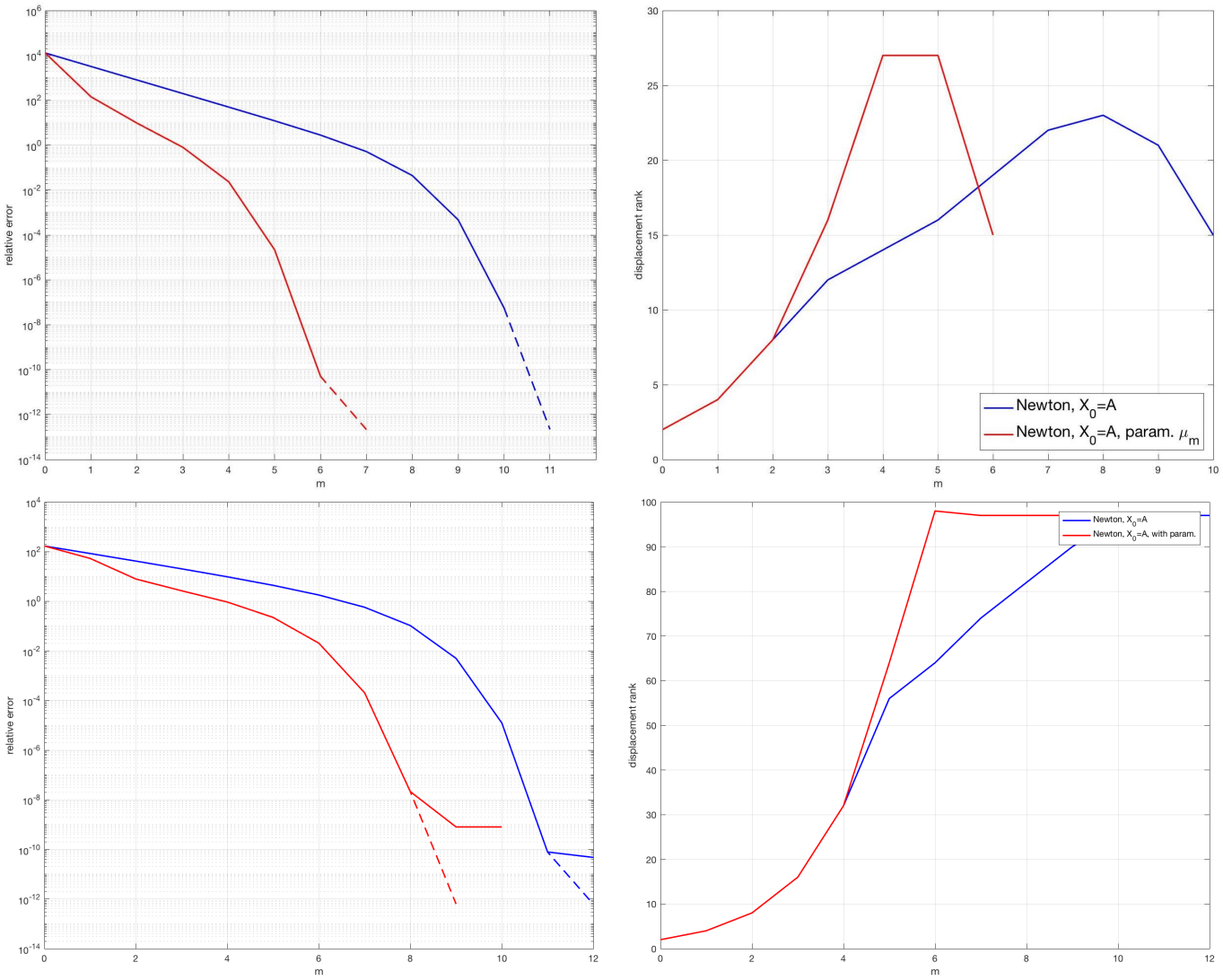

FIGURE 3.10 – Erreur relative L^2 et rang de déplacement des matrices itérées X_m de la méthode de Newton avec et sans paramètres $\mu_m = \sqrt{\|X_m^{-1}\|/\|X_m\|}$ pour l'approximation de $\mathrm{sign}(A)$ avec $A \in \mathbb{R}^{3000\times 3000}$ matrice de Toeplitz symétrique à spectre dans $[-0,941; 113,297]$ et conditionnement $121,02$ (graphique supérieur), et $A \in \mathbb{R}^{3000\times 3000}$ matrice de Toeplitz symétrique à spectre dans $[-173,999; 162,7]$ et conditionnement $3,37 \times 10^3$ (graphique inférieur).

Exemple 3.5.5. *En figure 3.10, nous commençons par étudier la méthode de Newton avec ou sans paramètres* $\mu_m = \sqrt{\|X_m^{-1}\|/\|X_m\|}$ *pour le signe sur une matrice* $A \in \mathbb{C}^{3000\times 3000}$ *de Toeplitz symétrique où*

$\sigma(A) \subseteq [-1,44; 260,21]$ *et conditionnement* $181,73$ *sans valeurs propres dans* $i\mathbb{R}$. *Comme pour le cas de la fonction de matrice racine carrée, nous imposons un critère d'arrêt en fonction du résidu de chaque matrice itérée* X_m *de la méthode de Newton et stoppons l'itération dès que*

$$\|I - X_{m+1}^2\| > \frac{1}{2}\|I - X_m^2\|.$$

On remarque alors sur le graphique de gauche que les matrices itérées de Newton pour le signe permettent d'atteindre une tolérance de 10^{-12} *que ce soit en arithmétique Toeplitz-like ou en arithmétique pleine. Or avec paramètres* $\mu_m = \sqrt{\|X_m^{-1}\|/\|X_m\|}$, *on observe en particulier que le nombre d'itérations nécessaires à la convergence est réduit de presque moitié, passant ainsi de 12 à 7 itérations nécéssaires, justifiant ainsi l'importance de l'introduction de ces paramètres dans la méthode de Newton pour le signe. Concernant le rang de déplacement, celui-ci non seulement n'explose pas mais surtout décroît pour atteindre un rang de déplacement de 5, plus proche du rang de déplacement de la fonction de matrice* $\mathrm{sign}(A)$. *Cette observation permet donc de justifier l'approche par la méthode de Newton pour l'approximation de la fonction de matrice* $\mathrm{sign}(A)$ *à l'aide de l'arithmétique Toeplitz-like.*
Des résultats similaires peuvent également être observés sur le graphique inférieur pour une deuxième matrice $A \in \mathbb{C}^{3000\times 3000}$ *Toeplitz symétrique avec* $\sigma(A) \subseteq [-173,999; 162,7]$ *et conditionnement* $3,37 \times 10^3$, *malgré une perte de tolérance, ici de* 10^{-10} *sans paramètres et* 10^{-9} *avec paramètres.*

Remarque 3.5.6. *Comme pour le cas de la fonction de matrice racine carrée principale, nous proposons le schéma suivant : à partir de* $X_0 = A \in \mathbb{C}^{n\times n}$, *on calcule les matrices* X_k *de la méthode de Newton pour le signe en arithmétique Toeplitz-like et on mesure le résidu à l'aide de la commande toeplknorm((*$I - X_k^2$*).G,(*$I - X_k^2$*).B,inf) qui mesure le résidu* $\|I - X_k^2\|_\infty$ *en arithmétique Toeplitz-like avec une complexité* $\mathcal{O}(\rho(I - X_k^2)n^2)$, *qui étant donné que* $I - X_k^2$ *est Toeplitz-like à chaque itération pour un faible nombre* m, *réduit donc la complexité. On peut alors reconstruire ou non la matrice ayant atteint la tolérance minimale en arithmétique en fonction des besoin, nous assurant ainsi un schéma global avec une complexité d'ordre* n^2.

3.6 Conclusion

Dans ce chapitre, nous avons vu comment approcher les fonctions de matrices racine carrée principale et signe avec un coût réduit. Nous avons rappelé en première section quelques propriétés de la fonction de matrice en nous concentrant essentiellement sur la racine carrée principale.
Puis nous avons repris dans une deuxième section la méthode Newton ainsi que ses variantes Newton-DB et Newton-DB produit explicitées dans le cas de matrices non structurées pour lesquelles nous connaissons les propriétés de convergence quadratique, de stabilité et de précision asymptotique.
Dans le but d'accélérer la convergence, nous proposons en section 3 quelques choix de premiers termes pour la méthode itérative de Newton afin d'exploiter rapidement la convergence quadratique des matrices X_k, ainsi qu'un choix de paramètres optimisés permettant d'accélérer la convergence.
Ces résultats ont ensuite été étudiés dans le cas de l'arithmétique Toeplitz-like en section 4 où à partir des générateurs d'une matrice Toeplitz-like $A \in \mathbb{C}^{n\times n}$ nous avons explicité les générateurs associés à chaque matrice itérée X_k de la méthode de Newton et de ses variantes, fonction rationnelle d'ordre $[2^k|2^k - 1]$ en la matrice A, à l'aide de nos données sur l'arithmétique Toeplitz-like du chapitre 2. Cette étude nous a alors permis de déterminer le rang de déplacement à chaque itération, ce qui est ensuite illustré par plusieurs expériences numériques sur des matrices de Toeplitz symétrique définies positives, montrant la convergence quadratique et le rang de déplacement numérique pour chaque terme de nos itérations de Newton en arithmétique Toeplitz-like. Ces observations nous permettent de conclure que dans le cas de matrices assez bien conditionnées, nous pouvons approcher la fonction de matrice racine carrée $\sqrt{A}$, où $A \in \mathbb{R}^{n\times n}$ Toeplitz

symétrique définie positive, avec une complexité de l'ordre de $\mathcal{O}(n^2)$ opérations élémentaires à l'aide d'un choix particulier du premier terme et des paramètres.
Enfin en dernière section, nous avons fait de même avec la fonction de matrice signe, en adaptant la méthode de Newton pour le signe à l'arithmétique Toeplitz-like nous permettant d'approcher la fonction de matrice $\mathrm{sign}(A)$ avec une faible complexité d'ordre $\mathcal{O}(n \log^2 n)$.
Dans ce chapitre, nous n'avons considéré pour le cas de la fonction de matrice racine carrée principale que des matrices symétriques définies positives. Le cas de matrices non symétriques pourra être considéré dans de futures recherches.

Chapitre 4

Fonctions de Markov appliquées aux matrices de Toeplitz

Nous avons vu au chapitre précédent que la racine carrée principale d'une matrice de Toeplitz pouvait être approchée avec une complexité d'ordre $\mathcal{O}(n \log^2 n)$ ou $\mathcal{O}(n^2)$ en passant par la méthode itérative de Newton et ses variantes calculées en arithmétique Toeplitz-like. Or la fonction racine carrée principale est un cas particulier des fonctions de Markov, c'est-à-dire des fonction de la forme

$$f^{[\mu]}(z) = \int_\alpha^\beta \frac{d\mu(x)}{z-x}$$

où $-\infty \leqslant \alpha < \beta$ et μ est une mesure positive à support dans $[\alpha; \beta]$. Cette famille de fonctions inclut par exemple des fonctions avec support infini tels que

$$f^{[\mu]}(z) = \frac{\log(z)}{z-1}, \quad f(z) = z^\gamma, \gamma \in (-1; 0), \text{ avec } \quad d\mu(x) = \frac{\sin(|\gamma|\pi)}{\pi}|x|^\gamma$$

ainsi que des fonctions à d'autres supports tels que les fonctions de la forme

$$f^{[\nu]}(z) = \frac{\sqrt{|\alpha|}}{\sqrt{(z-\alpha)(z-\beta)}}, \quad d\nu(x) = \frac{\sqrt{|\alpha|}dx}{\pi\sqrt{(x-\alpha)(\beta-x)}}. \tag{4.1}$$

avec $\alpha < \beta$ (voir lemme 3.1.1).

Plutôt que d'utiliser des méthodes itératives comme pour la racine carrée principale, nous cherchons ici à calculer directement un approximant rationnel pour les fonctions de Markov que nous appliquons ensuite sur nos matrices. Or, dans la littérature actuelle, les résultats sur un meilleur approximant rationnel sont généralement étudiés dans le cas de fonctions rationnelles d'ordre $[m-1|m]$ pour $m \geqslant 1$. Afin de pouvoir comparer nos futures résultats avec la théorie déjà existante sur la meilleure approximation rationnelle, nous allons donc ici considérer des fonctions rationnelles d'ordre $[m-1|m]$.

Dans ce chapitre, nous allons nous servir de la propriété des ensembles K-spectraux dont nous rappelons la définition : un ensemble fermé $\mathbb{E}$ est un ensemble K-spectral pour une matrice carrée A si pour toute fonction g analytique sur un voisinage de $\mathbb{E}$ on a

$$\|g(A)\| \leqslant K\|g\|_{L^\infty(\mathbb{E})}$$

et nécessairement $\sigma(A) \subseteq \mathbb{E}$. De plus, pour une matrice $A \in \mathbb{C}^{n\times n}$ symétrique, tout ensemble $\mathbb{E}$ contenant $\sigma(A)$ est un ensemble K-spectral avec $K = 1$. Par conséquent, nous pouvons sélectionner un ensemble $\mathbb{E}$

contenant le spectre de A et alors pour toute fonction f analytique sur un voisinage de $\mathbb{E}$ et toute fonction rationnelle r_m d'ordre $[m-1|m]$ sans pôles dans $\mathbb{E}$, nous avons

$$\|f(A) - r_m(A)\| \leqslant K\|f - r_m\|_{L^\infty(\mathbb{E})} \quad \text{et} \quad \|I - r_m(A)f(A)^{-1}\| \leqslant K\|\frac{f - r_m}{f}\|_{L^\infty(\mathbb{E})}.$$

Ici, nous allons considérer pour une fonction de Markov $f^{[\mu]}$ et une matrice $A \in \mathbb{R}^{n\times n}$, un ensemble $\mathbb{E} \supseteq \sigma(A)$ et des Padé multipoints d'ordre $[m-1|m]$ sur l'ensemble $\mathbb{E}$ en $2m$ points $z_1, \ldots, z_{2m}$ c'est-à-dire des interpolants rationnelles d'odre $[m-1|m]$ de la forme $r_m(f^{[\mu]})(z) = P_{m-1}(z)/Q_m(z)$ avec $P_{m-1} \in \mathcal{P}_{m-1}$ et $Q_m \in \mathcal{P}_m$ vérifiant

$$f^{[\mu]}(z_j) = r_m(f^{[\mu]})(z_j) \quad \text{soit} \quad Q_m(z_j)f(z_j) - P_{m-1}(z_j) = 0 \tag{4.2}$$

pour $j = 1, \ldots, 2m$ et étudier les erreurs $\|f^{[\mu]} - r_m(f^{[\mu]})\|_{L^\infty(\mathbb{E})}$ et $\|\frac{f^{[\mu]} - r_m(f^{[\mu]})}{f^{[\mu]}}\|_{L^\infty(\mathbb{E})}$.

Dans une première sous-section, nous rappelons quelques résultats déjà existants sur la meilleure approximation rationnelle des fonctions de Markov. Puis la sous-section 4.1.2 est consacrée à l'étude de l'erreur absolue et relative d'approximation d'une fonction de Markov par un Padé multipoint d'ordre $[m-1|m]$. En particulier, une première borne supérieure pour l'erreur relative d'approximation par interpolation d'une fonction de Markov $f^{[\mu]}$ sur un ensemble $\mathbb{E}$ est donné au théorème 4.1.4, borne dépendante de l'approximation par interpolation de la fonction de Markov $f^{[\nu]}$ où $\nu(x)$ dénote la mesure d'équilibre renormalisée définie par (4.1), nous permettant ensuite de donner une borne supérieure a priori dépendante uniquement des points d'interpolation. Nous cherchons ensuite à optimiser cette borne en fonction de nos choix de points d'interpolation. Notons que différents choix de points d'interpolation ont été évoqués précédemment dans la littérature, et on peut notamment citer [7]. Ici nous considérons d'abord en sous-section 4.1.3 le cas d'un ou deux points d'interpolation multiples. Puis en considérant tout choix possible en sous-section 4.1.4 nous obtenons une borne supérieure a priori optimisée par un choix particulier de nos points d'interpolation dans le cas où $\mathbb{E} = [c; d] \subset \mathbb{R}$. Ces points seront appelés points d'interpolation optimaux pour l'erreur relative d'approximation sur l'intervalle $[c; d]$ d'une fonction de Markov avec mesure à support dans $[\alpha; \beta]$ par l'interpolant rationnel associé, borne supérieure a priori dépendante de la quantité $\varrho = \exp\left(\frac{-1}{cap([\alpha;\beta],[c;d])}\right)$. Nous nous intéressons également en sous-section 4.1.5 au cas particulier où $\mathbb{E} = \overline{\mathbb{D}}$. Puis dans le but de pouvoir implémenter de manière efficace et robuste nos Padé multipoints avec points d'interpolation optimaux, nous passons en revu en section 4.2 différentes formes de ces interpolants en rappelant quelques résultats de stabilité pour l'évaluation de telles fonctions rationnelles. En sous-section 4.2.3, nous nous intéressons en particulier à la représentation en fraction continue de Thiele pour laquelle un nouveau résultat de positivité des paramètres de cette représentation est démontré ainsi qu'un résultat original sur la backward stabilité des fractions continues de Thiele positives. Tous ces résultats sont ensuite appliqués en section 4.3 à l'implémentation des Padé multipoints en une matrice de Toeplitz symétrique $A \in \mathbb{R}^{n\times n}$ en arithmétique Toeplitz-like. Le rang de déplacement des différentes représentations est alors donné en sous-section 4.3.1 puis nous énonçons en sous-section 4.3.2 deux critères d'arrêt : un critère a posteriori prenant en compte le fait que l'erreur relative n'est pas beaucoup modifié pour un indice un peu plus élevé et un critère résiduel en observant que la fonction de matrice $f^{[\mu]}(A)^{-1}$ est généralement plus facile à calculer afin de pouvoir mesure l'erreur $\|I - r_m(f^{[\mu]})(A)f^{[\mu]}(A)^{-1}\|$, suivi de plusieurs expériences numériques où nous comparons l'erreur relative d'approximation $\|I - r_m(f^{[\mu]})(A)f^{[\mu]}(A)^{-1}\|$ par nos Padé multipoints sur les différentes arithmétiques. Une version abrégée de ce chapitre a été soumise à publication [10].

4.1 Erreur d'approximation des fonctions de Markov par interpolation rationnelle

Dans cette section, nous cherchons à borner l'erreur d'approximation dans $\mathcal{R}_{m-1,m}$ d'une fonction de Markov $f^{[\mu]}$ sur tout ensemble $\mathbb{E}$. A cet effet, nous commençons par rappeler quelques résultats déjà

connus sur le sujet et expliquons pourquoi nous avons cherché à apporter un autre résultat. Puis dans une deuxième sous-section, nous établissons une nouvelle borne supérieure pour l'erreur relative d'approximation par interpolation en donnant l'un de nos principaux résultat au théorème 4.1.4 qui consiste en une borne supérieure exprimée en fonction de l'approximation relative par interpolation de la fonction de Markov $f^{[\nu]}$ où ν désigne la mesure d'équilibre renormalisée. Nous cherchons ensuite en sous-sections 4.1.3 et 4.1.4 à optimiser cette borne dans le cas où $\mathbb{E}$ est un intervalle fini de l'axe réel disjoint du support de la mesure associé à la fonction $f^{[\mu]}$ à l'aide d'un choix particulier des points d'interpolations dans différents cas. Enfin, nous nous intéressons au cas du disque fermé, cas permettant de généraliser une borne supérieure à tout ensemble $\mathbb{E}$ convexe compact symétrique par rapport à l'axe réel.

4.1.1 Etat de l'art sur la meilleure approximation rationnelle des fonctions de Markov

L'approximation rationnelle des fonctions de Markov a suscité l'intérêt de nombreux auteurs. En 1978, A. A. Gonchar s'est intéressé au comportement asymptotique de l'erreur de meilleure approximation rationnelle sur un intervalle $[c;d]$ pour une fonction de Markov avec mesure à support dans $[\alpha;\beta]$ où $[\alpha;\beta]\cap[c;d]=\emptyset$ et a obtenu [45, Theorem 1]

$$\lim_{m\to\infty}\min_{r\in\mathcal{R}_{m-1,m}}\|f-r\|^{1/m}_{L^\infty([c;d])}=\exp\left(-\frac{2}{cap([\alpha;\beta],[c;d])}\right)$$

où $cap([\alpha;\beta],[c;d])$ désigne la capacité logarithmique des intervalles $[\alpha;\beta]$ et $[c;d]$:

Définition 4.1.1. *Soient K et S deux ensembles compacts disjoints dans $\overline{\mathbb{C}}$. Pour deux mesures de probabilité μ_1 et μ_2 telles que $supp(\mu_1)\subseteq K$ et $supp(\mu_2)\subseteq S$, on considère la mesure $\sigma:=\mu_1-\mu_2$ et l'énergie $I(\sigma)=\int\int\log\left(\frac{1}{|z-t|}\right)d\sigma(z)d\sigma(t)$ associée, puis on pose*

$$V_{KS}:=\inf\{I(\sigma):\sigma=\mu_1-\mu_2, supp(\mu_1)\subseteq K, supp(\mu_2)\subseteq S,\mu_1,\mu_2\geqslant 0,\|\mu_1\|=\|\mu_2\|=1\}.$$

La capacité associée à K et S est alors donnée par

$$cap(K,S)=\frac{1}{V_{KS}}.$$

Cependant, le résultat énoncé par A. A. Gonchar est donné en terme du comportement asymptotique de $\min_{r\in\mathcal{R}_{m-1,m}}\|f-r\|^{1/m}_{L^\infty([c;d])}$ sans expliciter la fonction rationnelle minimisante pour chaque m et n'est donc pas exploitable pour une implémentation numérique. Dans un cadre plus général, il traite le cas de l'approximation sur un ensemble $\mathbb{E}\subseteq\overline{\mathbb{D}}$ compact et symétrique par rapport à l'axe réel en fournissant une borne supérieure asymptotique à la limite sup. de l'erreur d'approximation dans $\mathcal{R}_{m-1,m}$ [44, Section 4].
En 1982, T. Ganelius a également étudié le cas des interpolants rationnels $\mathcal{R}_{m-1,m}$ et a donné une borne supérieure à l'erreur absolue minimisée $\min_{r\in\mathcal{R}m-1,m}\|f-r\|_{L^\infty([c;d])}$ en utilisant des interpolants avec points d'interpolation doubles [39, Chap. 4]. Cependant, ces points d'interpolation ne sont pas explicités ce qui nous empêche également de les utiliser sur machine.
En 1987, D. Braess a considéré le cas particulier d'une approximation sur le disque unité et des points d'interpolation doubles [19, Theorem 2.1], et a ainsi obtenu la borne supérieure pour l'approximation

$$\min_{r\in\mathcal{R}_{m-1,m}}\|f-r\|_{L^\infty(\overline{\mathbb{D}})}\leqslant\frac{32\mu([\phi(\alpha),\phi(\beta)])}{dist(\overline{\mathbb{D}},[\phi(\alpha),\phi(\beta)])^2}\varrho^{2m+4}$$

lorsque ϕ est l'application conforme de $\overline{\mathbb{C}}\backslash[-1;1]$ dans $\overline{\mathbb{C}}\backslash\overline{\mathbb{D}}$ et f est une fonction de Markov avec mesure μ à support dans $[\phi(\alpha);\phi(\beta)]$ et $\varrho=\exp\left(\pi\frac{K(k)}{K(k')}\right)$ avec, k,k' donné par [19, equation (1.8)] et K l'intégrale

elliptique complète. Cependant, on souhaiterait obtenir une meilleure constante en facteur de ϱ^{2m+4}.
En 1992, H. Stahl et V. Totik donnent une caractérisation d'une meilleure approximation rationnelle [81, Theorem 6.2.2] d'une fonction de Markov $f : z \mapsto \int_\alpha^\beta \frac{d\mu(x)}{z-x}$ avec μ une mesure positive à support dans $[\alpha;\beta]$ sur un ensemble $\mathbb{E} \subseteq \overline{\mathbb{C}}\backslash[\alpha;\beta]$ symétrique par rapport à l'axe réel et $r_m^* \in \mathcal{R}_{m,m}$ une fonction rationnelle telle que

$$\|f - r_m^*\|_{L^\infty(\mathbb{E})} = \inf_{r_m \in \mathcal{R}_m} \|f - r_m\|_{L^\infty(\mathbb{E})}.$$

et alors

$$\limsup_{m\to\infty} \|f - r_m^*\|_{L^\infty(\mathbb{E})}^{1/2m} \leqslant e^{-1/cap(\mathbb{E},supp(\mu))}.$$

De plus, si $cap(\mathbb{E}, supp(\mu)) > 0$, alors

$$\lim_{m\to\infty} \|f - r_m^*\|_{L^\infty(\mathbb{E})}^{1/2m} = e^{-1/cap(\mathbb{E},supp(\mu))}.$$

On peut également citer B. Beckermann et L. Reichel [13] qui pour des pôles fixes $z_1, \ldots, z_m \notin \mathbb{E}$ avec $\mathbb{E}$ ensemble convexe compact de $\mathbb{C}$, étudient le problème de minimisation $\eta_m^k(f, \mathbb{E}) = \min\left\{\|f - \frac{p}{q}\|_{L^\infty(\mathbb{E})} : p \in \mathcal{P}_k\right\}$ où $q(z) = \prod_{j=1}^m (z - z_j)$, f fonction de Markov. Sur $\mathbb{E} = \overline{\mathbb{D}}$, ils obtiennent la borne supérieure [13, Theorem 6.1]

$$\min_{p_m \in \mathcal{P}_m} \|f - \frac{p_m}{q_m}\|_{L^\infty(\overline{\mathbb{D}})} \leqslant \frac{\|f\|_{L^\infty(\overline{\mathbb{D}})}}{\beta} \max_{z\in[\alpha;\beta]} \frac{1}{|B(z)|}$$

avec $B(z) = \prod_{j=1}^m \frac{1-zz_j}{z-z_j}$, et ce résultat se généralise alors à tout ensemble $\mathbb{E}$ convexe, compact et symétrique par rapport à l'axe réel à l'aide de la transformée de Faber $\mathcal{F}$ pour $\mathbb{E}$ grâce à laquelle $\min_{p_m\in\mathcal{P}_m} \|f - \frac{p_m}{q_m}\|_{L^\infty(\mathbb{E})} \leqslant 2\min_{p_m\in\mathcal{P}_m} \|\mathcal{F}^{-1}(f) - \frac{p_m}{q_m}\|_{L^\infty(\overline{\mathbb{D}})}$ où $\mathcal{F}^{-1}(f)$ est encore une fonction de Markov [13, Theorem 6.1] permettant alors d'obtenir pour une définition particulière des points d'interpolation [13, equation (6.11)] la borne supérieure [13, Theorem 6.6]

$$\min_{p_m \in \mathcal{P}_m} \|f - \frac{p_m}{q_m}\|_{L^\infty(\overline{\mathbb{D}})} \leqslant 4\frac{\|f\|_{L^\infty(\mathbb{E})}}{\phi(\beta)} R([\alpha;\beta], \mathbb{E})^{-m}$$

où $\phi : \overline{\mathbb{C}}\backslash\mathbb{E} \to \overline{\mathbb{C}}\backslash\overline{\mathbb{D}}$ est une application conforme avec normalisation $\phi(\infty) = \infty$ et $\phi'(\infty) > 0$ et $R([\alpha;\beta], \mathbb{E})^{-m} = \exp\left(\frac{-m}{cap([\alpha;\beta],\mathbb{E})}\right)$. Cependant les pôles sont fixes, ce qui ne nous permet pas une optimisation des interpolants rationnels pour optimiser l'approximation rationnelle d'une fonction de Markov.
En 2009, L.A. Knizhnerman [64] s'est intéressé aux approximants de Padé-Faber de type $[m-1, m]$ $(m \geqslant 1)$, notés $\{r_m\}_m$, des fonctions de Markov de la forme $f(z) = \int_{-\infty}^0 \frac{d\mu(x)}{z-x}$ sur un ensemble compact $\mathbb{E} \subset \overline{\mathbb{C}}$ à complémentaire simplement connexe dans $\overline{\mathbb{C}}$ et symétrique par rapport à l'axe réel tel que $\mathbb{E} \cap (-\infty; 0] = \emptyset$. Notons $\psi : \overline{\mathbb{C}}\backslash\overline{\mathbb{D}} \to \overline{\mathbb{C}}\backslash\mathbb{E}$ l'application conforme telle que $\psi(\infty) = \infty, \psi'(\infty) > 0$ et $\phi = \psi^{-1}$ son application inverse. Alors en utilisant l'égalité sur les coefficients de Faber définis $\forall n$ par

$$a_n(g) = \frac{1}{2i\pi} \int_{|w|=r} \frac{g(\psi(w))}{w^{n+1}} dw$$

lorsque $g \in \{h \in Hol(\mathbb{D}) : \exists R > 1, h$ prolongeable en une fonction $h \in Hol(\{w : |w| < R\})$ pour tout $r \leqslant R\}$. On sait que $a_k(g) = a_k(r_m)$ pour tout $k \leqslant 2m-1$ (plus généralement, pour toute fonction rationelle $r_{m,n} \in \mathcal{R}_{m,n}$, $a_k(r_{m,n})$ existe et est égal à $a_k(g)$ dès que $m \geqslant n-1$ et $k \leqslant m+n$[32, Theorem 1.1]), et Knizhnermann démontre alors que ces approximants sont eux aussi des fonctions de Markov. Puis par la transformée de Faber $\mathcal{F}$ pour $\mathbb{E}$ [31] définie par

$$\mathcal{F}(g)(z) = \frac{1}{2i\pi} \int_{|w|=1} g(\phi(w)) \frac{dw}{w - w'}, \quad \mathcal{F}(\sum_{j=0}^k a_j w^j)(z) = \sum_{j=0}^k a_j F_j(z)$$

(où les F_j sont les polynômes de Faber), Knizhnerman donne une estimation de l'erreur des approximants de Padé $R_m = \mathcal{F}^{-1}(r_m)$ en l'infini (où R_m est bien définie puisque pour tout $R_{m,n} = \frac{P_m}{Q_n} \in \mathcal{R}_{m,n}$ avec $deg(P_m) \leqslant m$ et $deg(Q_n) \leqslant n$, $\mathcal{F}^{-1}(R_{m,n}) \in \mathcal{R}_{\tilde{m},n}$ avec $\tilde{m} = \max\{m, n-1\}$ [32, Thm 1.1]) de la fonction de Markov $F = \mathcal{F}^{-1}(f)$ [13, Théorème 6.1] associée à une mesure ν à support dans $[\varrho, 0]$ où $\varrho = \phi(0)^{-1}$. Par changements de variables successifs $u = \phi(x)$ puis $w = \frac{1}{u}$ et par la relation entre polynômes orthogonaux par rapport à une mesure et l'approximation de Padé [6, Part.2, Section 3.1], il obtient finalement l'estimation

$$\|f - r_m\|_{L^\infty(\mathbb{E})} = \mathcal{O}(|\tilde{\phi}(-1)|^{-2m})$$

où $\tilde{\phi}(u) = 1 - 2\phi(0)u + \sqrt{[1-2\phi(0)u]^2 - 1}$ (application conforme associée à l'intervalle $[\rho, 0]$) et où $\mathcal{O}$ cache la quantité $\int_\varrho^0 d\nu(u) + \int_\varrho^0 |u|^m d\nu(u)$.
Nous avons alors remarqué une chose : la notion de $\mathcal{O}(.)$ est un peu floue, et pourrait contenir des quantités trop grandes pour assurer une convergence suffisante. Autrement dit, nous ne pouvons pas utiliser ce résultat pour des simulations numériques lorsque l'on souhaite passer par des approximations rationnels pour approcher la fonction de Markov initiale.

Notre travail ici va donc être de proposer une autre approximation dans $\mathcal{R}_{m-1,m}$ par interpolation, facilement implémentable pour pouvoir par la suite être appliquée aux matrices.

4.1.2 Borne supérieure pour l'erreur d'approximation des fonctions de Markov par interpolation rationnelle

Soit $f^{[\mu]} : z \mapsto \int_\alpha^\beta \frac{d\mu(x)}{z-x}$ fonction de Markov avec μ une mesure positive à support dans $[\alpha;\beta]$ où $-\infty \leqslant \alpha < \beta < +\infty$ et $m \in \mathbb{N}^*$ fixé. Notons $\mathcal{A}_m := \{z_1, \ldots, z_{2m}\} \subseteq \overline{\mathbb{C}}\backslash[\alpha;\beta]$ où les points z_j pour $j = 1, \ldots, 2m$ sont réels ou non-réels apparaissant par paires de conjugués dans $\mathcal{A}_m$ et définissons

$$\omega(z) = \pm \prod_{j=1}^{2m} (z - z_j). \tag{4.3}$$

Soit $r_m(f^{[\mu]})$ le Padé multipoint d'ordre $[m-1|m]$ de $f^{[\mu]}$ aux points $z_1, \ldots, z_{2m}$, c'est-à-dire l'interpolant dans $\mathcal{R}_{m-1,m}$ de $f^{[\mu]}$ aux points $z_1, \ldots, z_{2m}$. Par hypothèse sur les points $z_1, \ldots, z_{2m}$, la fonction ω est réelle sur $\mathbb{R}$, ne s'annulant pas sur l'intervalle $[\alpha;\beta]$. Les points $z_1, \ldots, z_{2m}$ n'étant pas forcément doubles, nous avons défini ω en $\pm$ de sorte à ce que l'on puisse assumer que $\omega(x) > 0$ pour tout $x \in [\alpha;\beta]$ et on obtient d'après [44, equations (8)-(9)] lorsque $r_m(f^{[\mu]}) = \frac{P_{m-1}}{Q_m} \in \mathcal{R}_{m-1,m}$ la propriété

$$\forall z \in \mathbb{C}\backslash[\alpha;\beta], \quad f^{[\mu]}(z) - r_m(f^{[\mu]})(z) = \frac{\omega(z)}{Q_m^2(z)} \int_\alpha^\beta \frac{Q_m^2(x)}{\omega(x)} \frac{d\mu(x)}{z-x} \tag{4.4}$$

avec Q_m un polynôme admettant m zéros simples, tous localisés dans l'intervalle ouvert $(\alpha;\beta)$ et $\forall k = 0, 1, \ldots, m-1$,

$$\int_\alpha^\beta Q_m(x) x^k \frac{d\mu(x)}{\omega(x)} = 0 \tag{4.5}$$

c'est-à-dire que pour tout $m \in \mathbb{N}^*$, Q_m est le m^{e} polynôme orthogonal par rapport à la mesure $\frac{d\mu(x)}{\omega(x)}$. De plus, l'interpolant rationnel $r_m(f^{[\mu]})$ est unique [44] avec résidus positifs [81, Lemma 6.1.2]. Notre étude d'erreur est basée sur une formule bien connues énoncées dans le lemme suivant :

Lemme 4.1.2. *Soient $m \in \mathbb{N}^*$, $-\infty \leqslant \alpha < \beta < c \leqslant d < +\infty$, $f^{[\mu]} : z \mapsto \int_\alpha^\beta \frac{d\mu(x)}{z-x}$ fonction de Markov avec μ une mesure positive à support dans $[\alpha;\beta]$ et $z_1, \ldots, z_{2m} \in \overline{\mathbb{C}}\backslash[\alpha;\beta]$, $\omega(z) = \pm\prod_{j=1}^{2m}(z - z_j)$ que l'on*

suppose positif sur $[\alpha;\beta]$. *Soit* $r_m(f^{[\mu]})$ *le Padé multipoint de* $f^{[\mu]}$ *d'ordre* $[m-1|m]$ *associé aux points* z_j, $j=1,\dots,2m$. *Alors* $\forall z\in\mathbb{R}\backslash[\alpha;\beta]$,

$$|f^{[\mu]}(z)-r_m(f^{[\mu]})(z)| = \min_{\deg(P)\leq m}\frac{\omega(z)}{|P(z)|^2}\int_\alpha^\beta\frac{|P(x)|^2}{\omega(x)}\frac{d\mu(x)}{|z-x|} \tag{4.6}$$

$$\leqslant |f^{[\mu]}(z)|\min_{\deg P\leqslant m}\frac{|\omega(z)|}{|P(z)|^2}\Big\|\frac{P^2}{\omega}\Big\|_{L^\infty([\alpha;\beta])} \tag{4.7}$$

et le minimum de (4.6) *est atteint en* $P=Q_m$.

Démonstration. Nous démontrons ici uniquement l'équation (4.6) car (4.7) est une conséquence directe. Soient $f^{[\mu]}, r_m(f^{[\mu]})$ définis d'après nos hypothèses. Alors d'après (4.4),

$$|f^{[\mu]}(z)-r_m(f^{[\mu]})(z)|\leqslant\frac{\omega(z)}{Q_m^2(z)}\int_\alpha^\beta\frac{Q_m^2(x)}{\omega(x)}\frac{d\mu(x)}{|z-x|}.$$

Soit ν une mesure positive à support dans l'intervalle $[\alpha;\beta]$ et considérons $\{\phi_0,\phi_1,\dots\}$ un ensemble de polynômes orthonormés par rapport à la mesure ν. Alors la fonction $K_m^\nu(x,y):=\sum_{j=0}^m\phi_j(x)\overline{\phi_j(y)}$ vérifie $\forall x,z$

$$\int_\alpha^\beta K_m^\nu(x,y)K_m^\nu(y,z)d\nu(y)=\sum_{j=0}^m\phi_j(x)\overline{\phi_j(z)}\int_\alpha^\beta\phi_j(y)\overline{\phi_j(y)}d\nu(y)=K_m^\nu(x,z).$$

Par l'inégalité de Cauchy-Schwarz et par orthogonalité, on a $\forall P\in\mathcal{P}_m$ donné par $P(x)=\sum_{j=0}^m a_j\phi_j(x)$,

$$\begin{aligned}\frac{\int_\alpha^\beta|P(x)|^2d\nu(x)}{|P(z)|^2}&=\frac{\int_\alpha^\beta|\sum_{j=0}^m a_j\phi_j(x)|^2d\nu(x)}{|\sum_{j=0}^m a_j\phi_j(z)|^2}\geqslant\frac{\sum_{j=0}^m\sum_{k=0}^m a_j\overline{a_k}\int_\alpha^\beta\phi_j(x)\overline{\phi_k(x)}d\nu(x)}{\sum_{j=0}^m|a_j|^2\sum_{j=0}^m|\phi_j(z)|^2}\\&=\frac{\sum_{j=0}^m|a_j|^2}{\sum_{j=0}^m|a_j|^2\sum_{j=0}^m\phi_j(z)\overline{\phi_j(z)}}=\frac{1}{K_m^\nu(z,z)}.\end{aligned}\tag{4.8}$$

Or à z fixé, l'application $x\mapsto K_m^\nu(z,x)$ est un multiple scalaire non trivial du k^{e} polynôme orthogonal par rapport à la mesure $x\mapsto|z-x|d\nu(x)$. En effet, par la formule de Darboux-Christoffel [84, p.42-44], on peut montrer que $(x-z)K_m^\nu(z,x)=C(\phi_{m+1}(x)\phi_m(z)-\phi_m(x)\phi_{m+1}(z))$ d'où

$$\int_\alpha^\beta K_m^\nu(z,x)x^j(x-z)d\nu(x)=0,\ \ \forall j=0,1,\dots,m-1.$$

Ainsi, si on pose $d\nu(x):=\frac{d\mu(x)}{|z-x|}\frac{1}{\omega(x)}$, ν est une mesure et $K_m^\nu(z,x)$ est un multiple scalaire non trivial du m^{e} polynôme orthogonal par rapport à la mesure donnée par $|z-x|d\nu(x)=\frac{d\mu(x)}{\omega(x)}$. Par conséquent, d'après l'équation (4.8), on a $\forall z\in\mathbb{R}\backslash[\alpha;\beta]$:

$$\begin{aligned}\Big|\frac{\omega(z)}{Q_m^2(z)}\Big|\int_\alpha^\beta\Big|\frac{Q_m^2(x)}{\omega(x)}\Big|\frac{d\mu(x)}{|z-x|}&=\Big|\frac{\omega(z)}{cQ_m^2(z)}\Big|\int_\alpha^\beta\Big|\frac{cQ_m^2(x)}{\omega(x)}\Big|\frac{d\mu(x)}{|z-x|}\\&\leqslant|\omega(z)|\Big(\frac{\int_\alpha^\beta|P(x)^2|d\nu(x)}{|P(z)|^2}\Big)=\Big|\frac{\omega(z)}{P^2(z)}\Big|\int_\alpha^\beta\Big|\frac{P^2(x)}{\omega(x)}\Big|\frac{d\mu(x)}{|z-x|}\end{aligned}$$

pour tout polynôme $P\in\mathcal{P}_m$, incluant le polynôme minimisant. □

A présent, plutôt que de conserver une borne supérieure en terme du dénominateur, on cherche à exprimer cette borne uniquement en fonction des points d'interpolation afin de pouvoir sélectionner à l'avenir des points pour optimiser notre interpolation. Lorsque $[\alpha;\beta]=[-1;1]$, ce problème peut se rapporter au problème de minimisation de G. Meinardus [72, Section 4.1 et 4.4], en utilisant la définition suivante (voire [72, Section 3.2, Theorem 23] pour la preuve) :

Définition 4.1.3. *Soit e une fonction continue sur $[\alpha;\beta]$. On dit que e admet une alternante de longueur m si il existe $x_1 < \ldots < x_m \in [\alpha;\beta]$ tels que*

$$e(x_j) = \varepsilon(-1)^j \|e\|_{L^\infty(\|[\alpha;\beta])}$$

pour $j = 1,\ldots,m$ pour un $\varepsilon \in \{-1;1\}$. Etant donné une fonction poids σ positive et continue sur $[\alpha;\beta]$, un polynôme unitaire T de degré m est appelé polynôme de Chebyshev pondéré de degré m sur $[\alpha;\beta]$ à poids σ si

$$\|\sigma T\|_{L^\infty([\alpha;\beta])} \leqslant \|\sigma P\|_{L^\infty([\alpha;\beta])}$$

pour tout polynôme P unitaire de degré m.

A l'aide de cette définition, nous énonçons ci-dessous l'un des résultats principaux de cette thèse :

Théorème 4.1.4. *Soit $f^{[\mu]}(z) = \int_\alpha^\beta \frac{d\mu(x)}{z-x}$ une fonction de Markov avec $supp(\mu) \subset [\alpha;\beta]$, $m \in \mathbb{N}^*$, $z_1,\ldots,z_{2m}$ $2m$ points de $\overline{\mathbb{C}}\backslash[\alpha;\beta]$ réels ou non-réels apparaissant par paires de conjugués et $\mathbb{E} \subset \mathbb{R}\backslash[\alpha;\beta]$. Lorsque les points ont une multiplicité paire, nous parlons de cas positif. Alors le Padé multipoint $r_m(f^{[\mu]})$ de type $[m-1|m]$ de $f^{[\mu]}$ aux points $z_1,\ldots,z_{2m}$ vérifie*

$$\left\|\frac{f^{[\mu]} - r_m(f^{[\mu]})}{f^{[\mu]}}\right\|_{L^\infty(\mathbb{E})} \leqslant \begin{cases} 2\left\|\frac{f^{[\nu]} - r_m(f^{[\nu]})}{f^{[\nu]}}\right\|_{L^\infty(\mathbb{E})} \leqslant 4\eta_{2m}, & \text{dans le cas positif} \\ \left\|1 - \left(\frac{r_m(f^{[\nu]})}{f^{[\nu]}}\right)^2\right\|_{L^\infty(\mathbb{E})} \leqslant 4\frac{\eta_{2m}}{(1-\eta_{2m})^2} & \text{dans le cas général} \end{cases}$$

où $\frac{d\nu}{dx}(x) = \frac{\sqrt{|\alpha|}}{\pi\sqrt{(z-\alpha)(\beta-z)}}$ dénote la mesure d'équilibre renormalisée sur l'intervalle $[\alpha;\beta]$ pour la fonction de Markov $f^{[\nu]}(z) = \frac{\sqrt{\alpha}}{\sqrt{(z-\alpha)(z-\beta)}}$ et

$$\eta_{2m} = \max_{z\in\mathbb{E}} |G_{2m}(z)|, \;\; G_{2m}(z) = \prod_{j=1}^{2m} \frac{\varphi(z) - \varphi(z_j)}{1 - \varphi(z)\varphi(z_j)}$$

avec φ l'application conforme de $\overline{\mathbb{C}}\backslash[\alpha;\beta]$ dans $\overline{\mathbb{C}}\backslash\overline{\mathbb{D}}$ avec normalisation $\varphi(\infty) = \infty$ et $\varphi'(\infty) > 0$.

Démonstration. D'après (4.7),il nous faut rechercher un polynôme de Chebyshev pondéré par $1/\sqrt{\omega}$ de degré m sur l'intervalle $[\alpha;\beta]$. Or, d'après le théorème d'equi-oscillation de Chebyshev [72, Section 3.2, Theorem 23], on sait qu'un polynôme Q unitaire de degré m est un polynôme de Chebyshev pondéré par un poids σ de degré m sur l'intervalle $[\alpha;\beta]$ si et seulement si σQ admet une alternante de degré m. On sait également que ce polynôme de Chebyshev pondéré Q de degré m vérifie

$$\frac{\|\sigma Q\|_{L^\infty([\alpha;\beta])}}{|Q(z)|} \leqslant \frac{\|\sigma P\|_{L^\infty([\alpha;\beta])}}{|P(z)|}$$

pour tout polynôme P de degré $\leqslant m$ et pour tout $z \in \mathbb{R} \setminus [\alpha;\beta]$ [72, Section 4.4]. Par conséquent, pour trouver un polynôme de degré $\leqslant m$ extrémal pour (4.7), il nous suffit de trouver un polynôme P unitaire de degré $\leqslant m$ avec $\frac{P}{\sqrt{\omega}}$ admettant une alternante de longueur m sur $[\alpha;\beta]$. Montrons qu'il suffit de considérer l'intervalle $[-1;1]$. En effet, soit $T \in \mathcal{R}_{1,1}$ transformation de Moebius telle que

$$T([-1;1]) = [\alpha;\beta] \;\; , T(\mathbb{R}) = \mathbb{R}. \tag{4.9}$$

Alors pour tout polynôme $p \in \mathcal{P}_m$, il existe $P \in \mathcal{P}_m$ tel que

$$\frac{P(T(y))}{\sqrt{\omega(T(y))}} = \frac{p(y)}{\sqrt{\rho(y)}}, \;\; \text{où } \rho(y) = \pm\prod_{j=1}^{2m}(y - T^{-1}(z_j)). \tag{4.10}$$

Il reste alors à construire p avec $p/\sqrt{\rho}$ ayant une alternante sur $[-1;1]$. Une telle construction se trouve dans [72, Section 4.4] si ρ est le carré d'un polynôme de degré $\leqslant m$ (c'est le cas de points d'interpolation avec multiplicité paire) et peut être généralisé à notre cas d'une multiplicité quelconque. On peut factoriser

$$\rho\Big(\frac{1}{2}(w+\frac{1}{w})\Big) = H(w)H(\frac{1}{w}),\ \ H(w) = \sum_{j=0}^{2m} H_j w^j = H_{2m}\prod_{j=1}^{2m}(w-w_j),\ \ |w_j|>1$$

où $\frac{1}{2}(w_j+\frac{1}{w_j}) = T^{-1}(z_j)$, soit $w_j = \varphi(z_j)$ pour tout $j=1,\dots,2m$. Or $w\mapsto w^j+w^{-j}$ est un polynôme de degré j en $w+w^{-1}$, d'où l'application p définie par

$$p\Big(\frac{1}{2}(w+\frac{1}{w})\Big) = \frac{1}{2}(w^{-m}H(w)+w^m H(\frac{1}{w}))$$

est un polynôme de degré m. Il nous reste maintenant à vérifier la propriété d'alternance. Considérons le produit de Blaschke

$$B(w) = \frac{w^{2m}H(\frac{1}{w})}{H(w)} = \prod_{j=1}^{2m}\frac{1-w_j w}{w-w_j} = \frac{1}{G_{2m}(\varphi^{-1}(w))} \tag{4.11}$$

avec zéros dans le disque fermé $\overline{\mathbb{D}}$, réels ou par paires de conjugués. Alors en notant $x=\cos(t)$ et $w=e^{it}$, $t\in[0;2\pi[$, on observe que

$$\frac{w^{-m}H(w)}{|w^{-m}H(w)|} = e^{-is},\ \ \frac{w^m H(1/w)}{|w^m H(1/w)|} = e^{is},\ \ B(w) = e^{2is}$$

et donc $p(\cos(t))/\sqrt{\rho(\cos(t))} = \cos(s)$. Or pour un produit de Blaschke et pour $t\in[0;\pi]$, son argument croît de 0 à $2m\pi$ lorsque t croît de 0 à π, ce qui nous amène à l'oscillation désirée. Nous avons donc démontré que

$$\min_{\deg(Q)\leqslant m}\Big\|\frac{\omega}{Q^2}\Big\|_{L^\infty(\mathbb{E})}\Big\|\frac{Q^2}{\omega}\Big\|_{L^\infty([\alpha;\beta])} = \max_{x\in\mathbb{E}}\min_{\deg(Q)\leqslant m}\Big|\frac{\omega(x)}{Q^2(x)}\Big|\Big\|\frac{Q^2}{\omega}\Big\|_{L^\infty([\alpha;\beta])} = \max_{x\in\mathbb{E}}\frac{4|G_{2m}(x)|}{(1+G_{2m}(x))^2} \tag{4.12}$$

avec $G_{2m}(x)\in(-1;1)$ pour $x\in\mathbb{E}$ dans le cas général. Dans le cas positif, $G_{2m}(\beta)=1>0$, G_{2m} possède un nombre paire de changements de signe dans tout sous-intervalle de $\mathbb{R}\backslash Int(\mathbb{E})$ et zéros avec multiplicités paires dans $Int(\mathbb{E})$. Par conséquent, $G_{2m}(x)\in[0;1)$ pour $x\in\mathbb{E}$ dans le cas positif.

Posons à présent $w=\varphi(z)$, $z=T(y)$, $y=\frac{1}{2}(w+1/w)$. Alors d'après (4.11), pour tout $z\notin[\alpha;\beta]$,

$$\frac{1-G_{2m}(z)}{1+G_{2m}(z)} = \frac{B(w)-1}{B(w)+1} = \frac{w^m H(\frac{1}{w})-w^{-m}H(w)}{w^m H(\frac{1}{w})+w^{-m}H(w)} = \frac{1}{2}\frac{w^m H(\frac{1}{w})-w^{-m}H(w)}{q(y)}$$

avec $q\in\mathcal{P}_m$. Or, $w^m H(\frac{1}{w})-w^{-m}H(w) = \sum_{j=0}^{2m}H_j(w)(w^{m-j}-w^{j-m})$, et pour tout $j=0,\dots,2m$,

$$\begin{aligned} w^{m-j}-w^{j-m} &= w^{m-j}(1-w^{2j-2m}) = w^{m-j}(1-w^{-2})\sum_{k=0}^{m-j-1}w^{-2k} = (w-w^{-1})\sum_{k=0}^{m-j-1}w^{m-j-1-2k} \\ &= (w-w^{-1})\sum_{k\in\Omega_{m-j-1}}\tilde{p}(w+w^{-1}) = (w-w^{-1})p_{m-j-1}(\frac{1}{2}(w+w^{-1})) \end{aligned}$$

avec Ω_{m-j-1} est l'ensemble des nombres impairs entre 1 et $m-j-1$ si $m-j-1$ est pair ou Ω_{m-j-1} l'ensemble des nombres pairs entre 0 et $m-j-1$ si $m-j-1$ est impair et p_k un polynôme de degré k. D'après notre choix de la racine, $\frac{1}{2}(w-w^{-1}) = \sqrt{y^2-1}$, d'où

$$\frac{1-G_{2m}(z)}{1+G_{2m}(z)} = \sqrt{y^2-1}\frac{\sum_{j=0}^{m-1}H_j p_{m-j-1}(y)-\sum_{j=1}^{m}H_{j+m}p_{j-1}(y)}{q(y)} = \sqrt{y^2-1}r(y)$$

avec $r \in \mathcal{R}_{m-1,m}$ de dénominateur q défini précédemment. De plus, à l'aide de la transformée de Moebius T définie en (4.9), on peut noter

$$\begin{aligned}\sqrt{y^2-1}r(y) &= \sqrt{T^{-2}(z)-1}\ \ r(T^{-1}(z)) = \sqrt{(T^{-1}(z)-1)(T^{-1}(z)+1)}\ \ r(T^{-1}(z)) \\ &= \sqrt{(T^{-1}(z)-T^{-1}(\alpha))(T^{-1}(z)-T^{-1}(\beta))}\ \ r(T^{-1}(z)).\end{aligned}$$

et en fonction du support fini ou infini $[\alpha;\beta]$ de la mesure de départ, on montre qu'il existe $R \in \mathcal{R}_{m-1,m}$ tel que

$$\frac{1-G_{2m}(z)}{1+G_{2m}(z)} = \sqrt{y^2-1}r(y) = \frac{R(z)}{f^{[\nu]}(z)}.$$

Or, $G_{2m}(z) = 1$ lorsque $z = z_j$ pour tout $j = 1,\dots,2m$, d'où $r(y)$ interpôle $1/\sqrt{y^2-1}$ aux points $y_j = T^{-1}(z_j)$ et $R(z) = r_m^{[\nu]}(z)$ est un interpolant rationnel de type $[m-1|m]$ de $f^{[\nu]}(z)$ aux points z_j. En particulier, pour $z \in \mathbb{E}$,

$$\frac{4|G_{2m}(z)|}{(1+G_{2m}(z))^2} \leqslant \frac{4|G_{2m}(z)|}{1+G_{2m}(z)} \leqslant 2|1-\frac{r_m^{[\nu]}(z)}{f^{[\nu]}(z)}| \leqslant 4\eta_{2m} \quad \text{dans le cas positif,}$$

$$\frac{4|G_{2m}(z)|}{(1+G_{2m}(z))^2} = |1-\Big(\frac{r^{[\nu]}(z)}{f^{[\nu]}(z)}\Big)^2| \leqslant \frac{4\eta_{2m}}{(1-\eta_{2m})^2}, \quad \text{dans le cas général.}$$

En combinant (4.6) et (4.12), on obtient le résultat énoncé. □

4.1.3 Le cas d'un ou 2 points d'interpolation multiples

La quantité η_{2m} énoncée au théorème 4.1.4 ne dépendant que du choix des points d'interpolation et de l'ensemble $\mathbb{E}$ sur lequel on approche la fonction $f^{[\mu]}$, nous optimisons le choix des points d'interpolation $z_1,\dots,z_{2m}$. Dans le cas de pôles fixes, B. Beckermann et L. Reichel [13, Corollary 6.4] déterminent des points d'interpolation optimisés. En reprenant cette idée et en considérant le cas où $\mathbb{E}$ est un intervalle $[c;d]$, nous déterminons des choix optimisés des points d'interpolation dans les cas d'un ou deux points d'interpolation distincts.

Corollaire 4.1.5. *Soit $f^{[\mu]}(z) = \int_\alpha^\beta \frac{d\mu(x)}{z-x}$ fonction de Markov, μ mesure positive avec $supp(\mu) \subset [\alpha;\beta]$ et $r_m^{[\mu]}$ le Padé multipoint de $f^{[\mu]}$ en un point $z_1 \in \overline{\mathbb{R}}\backslash[\alpha;\beta]$ d'ordre $2m$. Alors pour tout $m \geqslant 1$,*

$$\|1-f^{[\mu]}/r_m^{[\mu]}\|_{L^\infty([c;d])} \leqslant 4\frac{|y_{opt}|^{-2m}}{(1-|y_{opt}|^{-2m})^2}$$

où

$$z_1 = \varphi^{-1}\Big(\frac{1+\varphi(c)y_{opt}}{\varphi(c)+y_{opt}}\Big), \quad y_{opt} = -\frac{1}{k}-\sqrt{\frac{1}{k^2}-1}, \quad k = \frac{\varphi(d)-\varphi(c)}{\varphi(d)\varphi(c)-1}$$

avec $\varphi : \overline{\mathbb{C}}\backslash[\alpha;\beta] \to \overline{\mathbb{C}}\backslash\overline{\mathbb{D}}$ l'application conforme associée à $[\alpha;\beta]$ avec normalisation $\varphi(\infty) = \infty$ et $\varphi'(\infty) > 0$. Cette borne supérieure est minimale parmi les bornes supérieures pour un point d'interpolation multiple.

Démonstration. Soit $f_w(z) = \frac{\varphi(z)-\varphi(w)}{1-\varphi(z)\varphi(w)}$ pour tout $z \in [c;d]$ et $w \in \overline{\mathbb{R}}\backslash[\alpha;\beta]$. Alors la fonction f_w vérifie
- f est décroissante (de dérivée négative) et $f_z(w) = -f_w(z)$ pour tout z, w :
- $|f_w(z)| = f_w(z)$ décroissante sur $[c;d]$ si $z \leqslant w$;
- $|f_w(z)| = -f_w(z)$ croissante sur $[c;d]$ si $z \geqslant w$.

Définissons une fonction g de la manière suivante :

$$g(x) := \max_{x\in[c;d]} |f_w(x)| = \max\{|f_w(c)|, |f_w(d)|\} \tag{4.13}$$

où la deuxième égalité provient de la décroissance de la fonction f_w. D'après la proposition 4.1.4, avec un point d'interpolation unique de multiplicité $2m$, il nous faut donc rechercher un minimiseur w^* de g.
On remarque dans un premier temps que l'équation $|f_w(c)| = |f_w(d)|$ admet une solution en w. En effet, l'équation $|f_w(c)| = |f_w(d)|$ n'est possible que si $f_w(c) = -f_w(d)$ par décroissance de f_w, et on obtient l'équation $(\varphi(c)-\varphi(d))-2\varphi(w)(1+\varphi(c)\varphi(d))+\varphi(w)^2(\varphi(c)\varphi(d)) = 0$ avec déterminant $4(\varphi(c)^2-1)(\varphi(d)^2-1) > 0$ et on peut vérifier que l'une des deux solutions se trouve dans l'intervalle $[\varphi(c);\varphi(d)]$ et notons w^* son image réciproque par φ. w^* est alors un minimiseur de la fonction g. En effet, si $|f_w(c)| = |f_w(d)|$, par décroissance de f_w,

— si $w \leqslant w^*$, $f_c(w) \leqslant f_c(w^*)$ d'où $f_w(c) \geqslant f_{w^*}(c)$ et $f_{w^*}(c) > 0$ d'où $f_w(c) > 0$ et donc

$$|f_w(c)| \geqslant |f_{w^*}(c)| = |f_{w^*}(d)| \Rightarrow g(w) \geqslant g(w^*);$$

— si $w \geqslant w^*$, $f_d(w) \geqslant f_d(w^*)$ d'où $f_w(d) \leqslant f_{w^*}(d)$ et $f_{w^*}(d) < 0$ d'où $f_w(d) < 0$ et donc

$$|f_w(d)| \geqslant |f_{w^*}(d)| = |f_{w^*}(c)| \Rightarrow g(w) \geqslant g(w^*);$$

Montrons alors que w^* donné par

$$w^* = \varphi^{-1}\Big(\frac{1+\varphi(c)y_{opt}}{\varphi(c)+y_{opt}}\Big), \quad y_{opt} = -\frac{1}{k} - \sqrt{\frac{1}{k^2}-1}, \quad k = \frac{\varphi(d)-\varphi(c)}{\varphi(d)\varphi(c)-1}$$

est notre minimiseur.

— si $\varphi(c) = \infty$, alors $|f_c(w)| = |f_w(c)| = \frac{1}{|w|}$ pour tout w, d'où $g(w^*) = \frac{1}{|w^*|}$ et on vérifie alors que w^* satisfait $|f_{w^*}(c)| = |f_{w^*}(d)| = \frac{1}{|y_{opt}|}$;

— si $\varphi(c) < \infty$, alors $|f_{w^*}(c)| = |f_{w^*}(d)| = \frac{1}{|y_{opt}|}$ également.

Et donc w^* est notre minimiseur.

□

Corollaire 4.1.6. *Soit $f^{[\mu]}(z) = \int_\alpha^\beta \frac{d\mu(x)}{z-x}$ fonction de Markov avec μ mesure positive avec $supp(\mu) \subset [\alpha;\beta]$ et $r_m^{[\mu]}$ le Padé multipoint de $f^{[\mu]}$ aux points $z_1 = c$ et $z_2 = d$ de multiplicité m. Alors*

$$\left\|1 - r_m^{[\mu]}/f^{[\mu]}\right\|_{L^\infty([c;d])} \leqslant 4\frac{|y_{opt}|^{-2m}}{(1-|y_{opt}|^{-2m})^2}$$

avec $y_{opt} = -\frac{1}{k} - \sqrt{\frac{1}{k^2}-1}$, $k = \frac{\varphi(d)-\varphi(c)}{\varphi(d)\varphi(c)-1}$ et φ l'application conforme de $\overline{\mathbb{C}}\backslash[\alpha;\beta]$ dans $\overline{\mathbb{C}}\backslash\overline{\mathbb{D}}$ avec normalisation $\varphi(\infty) = \infty$ et $\varphi'(\infty) > 0$.

Démonstration. Définissons $f_{c,d}(z) = \frac{\varphi(c)-z}{1-\varphi(c)z}\frac{\varphi(d)-z}{1-\varphi(d)z}$. Alors $f_{c,d}$ est dérivable sur $[c;d]$ et les racines de $f'_{c,d}$ sont les

$$z_{1,2} = \frac{\varphi(c)\varphi(d)+1 \pm \sqrt{(\varphi(d)^2-1)(\varphi(c)^2-1)}}{\varphi(c)+\varphi(d)}.$$

Or $z_1 = \frac{\varphi(c)\varphi(d)+1-\sqrt{(\varphi(d)^2-1)(\varphi(c)^2-1)}}{\varphi(c)+\varphi(d)} < \varphi(c)$ d'où $z_1 \notin [\varphi(c);\varphi(d)]$. De plus, puisque $f_{c,d}(\varphi(c)) = f_{c,d}(\varphi(d)) = 0$ et $f_{c,d} \neq 0$, $z_2 \in [\varphi(c);\varphi(d)]$. Notons $f_c(z) = \frac{\varphi(c)-z}{1-\varphi(c)z}$ et $f_d(z) = \frac{\varphi(d)-z}{1-\varphi(d)z}$, on peut vérifier que

$$f_c(z_2) = -y_{opt}^{-1} \quad \text{et} \quad f_d(z_2) = y_{opt}^{-1}$$

d'où $f_{c,d}(z_2) = f_c(z_2)f_d(z_2) = -y_{opt}^2$ et donc

$$\max_{z\in[c;d]} |G_{2m}(z)| = |f_{c,d}(z_2)|^m = |y_{opt}|^{-2m}.$$

□

4.1.4 Le cas général

B. Beckermann et L. Reichel considèrent dans [34] le cas de points de pôles fixes et mesurent l'erreur d'approximation associée. Ici, nous disposons de pôles libres, ce qui augmente le nombre de possibilités et va nous permettre de sélectionner des points d'interpolation minimisant la quantité η_{2m}.

Remarque 4.1.7. *Puisque la composition de deux facteurs de Blaschke est un facteur de Blaschke, il advient que le taux G_{2m} dans la proposition 4.1.4 ne dépend que du choix de l'application de Moebius T, i.e. que φ n'est pas nécessairement normalisée à l'infini, et il nous faut simplement vérifier que $T(\mathbb{R}) = \mathbb{R}$, $T(-1) = \alpha$, $T(1) = \beta$ et que T soit croissante sur $[-1;1]$. Il existe cependant une unique application T satisfaisant les conditions précédentes et les conditions supplémentaires $T(1/\kappa) = c$ et $T(-1/\kappa) = d$ où la valeur $\kappa \in (0;1)$ est obtenu à partir du birapport de 4 réels qui est invariant sous transformation linéaire, donnée*

$$T(-1) = \alpha, T(1) = \beta, T(1/\kappa) = c, T(-1/\kappa) = d, \frac{(c-\alpha)(d-\beta)}{(c-\beta)(d-\alpha)} = \Big(\frac{1+\kappa}{1-\kappa}\Big)^2 =: \frac{1}{k^2}. \tag{4.14}$$

De plus, $T([-1;1]) = [\alpha;\beta]$, $T([1/\kappa;-1/\kappa]) = [c;d]$ où $[1/\kappa;-1/\kappa] = \overline{\mathbb{R}}\backslash(-1/\kappa;1/\kappa)$. En utilisant l'application de Moebius, on obtient l'expression simplifiée

$$\eta_{2m}([c;d]) = \max_{w\in[\frac{1}{\lambda};-\frac{1}{\lambda}]} \Big|\prod_{j=1}^{2m} \frac{w-w_j}{1-w_j w}\Big|, \;\; w_j = \varphi(z_j), \;\; \lambda = \frac{1}{\varphi(c)} = -\frac{1}{\varphi(d)} = \frac{1-\sqrt{k}}{1+\sqrt{k}} \tag{4.15}$$

Minimiser (4.15) sur le choix des points d'interpolation w_j de module > 1 nous amène au problème de minimisation des produits de Blaschke (après substitution $u = 1/w$) sur l'intervalle $[-\lambda;\lambda]$, ce qui a été repris récemment par [75]. Ces résultats nous permettent alors d'énoncer le corollaire du théorème 4.1.4 suivant :

Corollaire 4.1.8. *Soient $\alpha, \beta, c, d, k, \lambda, \varphi$ avec $\mathbb{E}$=$[c;d]$ comme dans la remarque 4.1.7. Alors les points optimaux minimisant $\eta_{2m}([c;d])$ sont donnés en termes des fonctions elliptiques de Jacobi $sn(.)$ et l'intégrale elliptique complète $K(.)$, avec $K(x) = \int_0^1 \frac{dt}{(1-t^2)(1-x^2t^2)}$, par*

$$\frac{1}{\varphi(z_j)} = \frac{1}{w_j} = \lambda sn\Big(K(\lambda^2)(\frac{2j-1}{2m} - 1), \lambda^2\Big) \in (-\lambda;\lambda) \tag{4.16}$$

pour $j = 1, 2, \ldots, 2m$, et on obtient, pour ces points d'interpolation, la borne supérieure

$$\|\frac{f^{[\mu]} - r^{[\mu]}}{f^{[\mu]}}\|_{L^\infty([c;d])} \leqslant 8\varrho^{2m}/(1-2\varrho^{2m})^2, \;\; \varrho = \exp\Big(\frac{-1}{cap([\alpha;\beta],[c;d])}\Big), \tag{4.17}$$

en supposant que $2\varrho^{2m} < 1$.

Démonstration. D'après le théorème 4.1.4, il nous faut minimiser un produit de Blaschke à l'aide d'un choix particulier des points d'interpolation. Or, la composition d'un produit de Blaschke avec un facteur de Blaschke restant un produit de Blaschke, le changement de variable $z = T(x)$ entraîne en notant

$[1/\lambda;-1/\lambda]=(-\infty;-1/\lambda]\cup[1/\lambda;+\infty)$,

$$\min_{z_1,\dots,z_{2m}\in\overline{\mathbb{C}}\setminus[\alpha;\beta]}\max_{z\in[c;d]}\prod_{j=1}^{2m}\Big|\frac{\varphi(z)-\varphi(z_j)}{1-\varphi(z)\overline{\varphi(z_j)}}\Big|=\min_{x_1,\dots,x_{2m}\in\overline{\mathbb{C}}\setminus\overline{\mathbb{D}}}\max_{x\in[1/\lambda;-1/\lambda]}\prod_{j=1}^{2m}\Big|\frac{x-x_j}{1-x\overline{x}_j}\Big|$$
$$=\min_{w_1,\dots,w_{2m}\in\overline{\mathbb{D}}}\max_{u\in[-\lambda;\lambda]}\prod_{j=1}^{2m}\Big|\frac{u-w_j}{1-u\overline{w}_j}\Big|.$$

Ce problème de minimisation d'un produit de Blaschke d'ordre plus général N sur un intervalle symétrique $[-\lambda;\lambda]$ a notamment été étudié par T.W. Ng et C.Y. Tsang [75] qui donnent une formule explicite des racines d'un tel produit de Blaschke minimal [75, Section 3.2]. Or, ces racines sont définies sous la formes $w_j=cd(\frac{(2j-1)\omega_1(\tau)}{2N},\tau)$ pour $j=1,\dots,N$ avec $\tau\in i(0,+\infty)$, minimisant le produit de Blaschke sur un intervalle symétrique $[-\sqrt{k(\tau)};\sqrt{k(\tau)}]$. En comparant les définitions de Ng et Tsang avec la terminologie des fonctions elliptiques de Jacobi [76], on s'aperçoit alors que pour $\tau=iK^{'}(k)/K(k)$ avec K l'intégrale elliptique complète de module k [76, (19.2.8)] et $K^{'}(k)=K(\sqrt{1-k^2})$, les quantités $\omega_1(\tau)$ et $k(\tau)$ peuvent alors être remplacées par $2K(k)$ et $k=\lambda^2\in(0;1)$ respectivement. Ainsi, les racines du produit de Blaschke minimisant notre problème sont données par

$$w_j=\lambda cd\Big(K(\lambda^2)\frac{2j-1}{2m},\lambda^2\Big)=-\lambda sn\Big(K(\lambda^2)(-1+\frac{2j-1}{2m}),\lambda^2\Big),\ \ j=1,\dots,2m$$

où l'égalité provient de [76, Table 22.4.3]. La fonction sn étant impaire, on obtient les points optimisés (4.16). De plus d'après [75, Proposition 4.1.(a)], il existe un lien entre notre problème de minimisation et le troisième problème de Zolotarev, donné par

$$\eta_{2m}([c;d])=\min_{B\text{ produit de Blaschke d'ordre 2m}}\|B\|_{L^\infty([-\lambda;\lambda])}=\sqrt{\min_{R\in\mathcal{R}_{2m,2m}}\|R\|_{L^\infty([-\lambda;\lambda])}\|1/R\|_{L^\infty([1/\lambda;-1/\lambda])}}.$$

Or d'après [14, equations (3.7) et (3.8)],

$$\eta_{2m}([c;d])\leqslant 2\exp\Big(-\frac{m}{cap([-\lambda;\lambda],[1/\lambda;-1/\lambda])}\Big).$$

Par invariance du birapport sous application conforme et symétrie, on obtient

$$cap([-\lambda;\lambda],[1/\lambda;-1/\lambda])=cap([-k;-1],[k,1])=cap([\alpha;\beta],[c;d])$$

d'après les propriétés du module de Groetsch, d'où

$$\eta_{2m}([c;d])\leqslant 2\exp\Big(-\frac{m}{cap([-\lambda;\lambda],[1/\lambda;-1/\lambda])}\Big)=2\varrho^{2m}\tag{4.18}$$

□

Remarque 4.1.9. *Au corollaire 4.1.8 nous avons énoncé un choix de points dits optimaux. Ici, il est à noter que nous désignons par ce terme des points d'interpolation $z_1,\dots,z_{2m}$ minimisant la borne supérieure a priori en minimisant le produit de Blaschke associé.*

Nous venons de voir qu'un choix optimisé des points d'interpolation existe dans le cas de l'approximation sur un intervalle réel $[c;d]$, minimisant notre borne supérieure, mais pas nécessairement l'erreur relative étudiée. En effet, on peut espérer une borne plus précise par exemple si le support de la mesure associée à la fonction de Markov est un sous-ensemble propre de $[\alpha;\beta]$, formé par exemple de deux intervalles disjoints ; cependant dans le cas où $supp(\mu)=[\alpha;\beta]$ et μ est régulière au sens de [81], alors Gonchar [45, Theorem 1] (voir également [81, Theorem 6.2.2]) a montré que la $2m^{\text{e}}$ racine de l'erreur de la meilleure approximation rationnelle dans $\mathcal{R}_{m-1,m}$ de f sur un intervalle $[c;d]$ tend vers ϱ lorsque m tendait vers l'infini, soit

$$\lim_{m\to\infty}\min_{r\in\mathcal{R}_{m-1,m}}\|f-r\|^{1/2m}=\varrho.$$

Plus précisément, dans le cas de la fonction de Markov $f(z) = 1/\sqrt{z}$ avec $[\alpha;\beta] = [-\infty;0]$, Zolotarev [89], qui a exprimé de nombreux problèmes extrémaux en termes des nombres de Zolotarev, fut l'un des premiers à donner une inéquation pour l'erreur relative de meilleure approximation dans $\mathcal{R}_{m-1,m}$. On peut également citer Achieser [1, p. 147] et l'Appendix de [14] pour lesquels on obtient une borne supérieure $4\varrho^{2m}$ pour l'erreur relative de meilleure approximation dans $\mathbb{R}_{m-1,m}$ qui se comporte comme $4\varrho^{2m}(1+o(\varrho^{2m})_{m\to+\infty})$ (voir également [18, Theorem V.5.5]). Par conséquent, si l'on désire des points d'interpolation optimisant la borne supérieure pour toute fonction de Markov, les points (4.16) sont optimaux à un facteur au plus 2.

4.1.5 Le cas du disque

Dans le cas complexe, l'égalité (4.4) reste valide mais il n'en est pas de même pour l'inégalité (4.6). Il nous faut donc adapter nos résultats à l'aide du Lemme de Freud [36, Section III.7]. Ce résultat a notamment déjà été exploité par Ganelius [39] et Braess [19, Theorem 2.1] dont le travail sur l'interpolation rationnelle avec points d'interpolation de multiplicité paire nous a inspiré pour la preuve de nos résultats après correction d'une erreur dans l'utilisation du Lemme de Freud (un facteur $\beta-\alpha$ manquait), et l'avons amélioré [19, eqn (2.7)] d'un facteur 2. Il apparaît que pour des points de multiplicité pas nécessairement paire, on améliore l'estimation.

L'étude de l'erreur d'approximation rationnelle sur le disque fermé $\overline{\mathbb{D}}$ nous permet de généraliser notre étude à n'importe quel ensemble convexe compact $\mathbb{E}$ symétrique par rapport à l'axe réel. En effet, en définissant la transformée de Faber $\mathcal{F}$ (voir [38] ou [13]) et sa forme modifiée $\mathcal{F}_+$ par $\mathcal{F}_+(h) = \mathcal{F}(h) + h(0)$ pour toute fonction h analytique sur $\mathbb{D}$, on peut obtenir un bon approximant sur $\mathbb{E}$ à partir de $\overline{\mathbb{D}}$. Plus précisément, Ellacott [32, Theorem 1.1] a montré que $r_m \in \mathcal{R}_{m-1,m}$ si et seulement si $\mathcal{F}(r_m) \in \mathcal{R}_{m-1,m}$ et Kniznherman [64] ainsi que Beckermann et Reichel [13] ont montré simultanément que l'image réciproque par $\mathcal{F}$ d'une fonction de Markov est une fonction de Markov dont la mesure associée peut facilement être explicitée. De ces observations, on peut alors se servir de l'inégalité

$$\|\mathcal{F}_+(f) - \mathcal{F}_+(r_m)\|_{L^\infty(\mathbb{E})} \leqslant 2\|f - r_m\|_{L^\infty(\overline{\mathbb{D}})}$$

démontré dans [38, Theorem 2] et du théorème 4.1.10 pour déterminer un bon approximant rationnel sur $\mathbb{E}$ d'une fonction de Markov. En passant aux fonctions de matrices, nous pouvons citer le résultat de Beckermann et Reichel [13, Theorem 2.1] qui nous donne

$$\|\mathcal{F}_+(f)(A) - \mathcal{F}_+(r_m)(A)\|_{L^\infty(\mathbb{E})} \leqslant 2\|f - r_m\|_{L^\infty(\overline{\mathbb{D}})} \tag{4.19}$$

en supposant que le spectre de la matrice carrée A est un sous-ensemble de $\mathbb{E}$. Par conséquent, l'étude de l'erreur d'approximation par interpolation des fonctions de Markov sur le disque unité fermé est indispensable.

En nous inspirant des travaux de [39] et [19, Theorem 2.1] qui à partir du lemme de Freud [36, Section III.7] dans le cas de points avec multiplicité paire, nous énonçons à présent un nouveau théorème sur l'erreur d'approximation par interpolation sur le disque unité fermé des fonctions de Markov pour tout choix des points d'interpolation :

Théorème 4.1.10. *Soit $f(z) = \int_\alpha^\beta \frac{d\mu(x)}{z-x}$ une fonction de Markov avec $supp(\mu) \subseteq [\alpha;\beta]$ et soient $z_1, \ldots, z_{2m} \in \overline{\mathbb{C}}\backslash[\alpha;\beta]$ des points d'interpolation que l'on suppose réels ou non-réels apparaissant par paires de conjugués et $-\infty \leqslant \alpha < \beta < -1$. Si tous les points d'interpolation sont de multiplicité paire, alors*

$$\|\frac{f^{[\mu]} - r_m(f^{[\mu]})}{f^{[\mu]}}\|_{L^\infty(\overline{\mathbb{D}})} \leqslant C \max_{z\in[\alpha;\beta]} \Big|\prod_{j=1}^{2m} \frac{1-zz_j}{z-z_j}\Big|,\ \ C = \Big(\frac{1-\beta}{-1-\beta}\Big)^2.$$

Dans le cas général,

$$\|\frac{f^{[\mu]}-r_m(f^{[\mu]})}{f^{[\mu]}}\|_{L^\infty(\overline{\mathbb{D}})} \leqslant C\frac{4\eta_{2m}^{'}}{(1-\eta_{2m}^{'})^2},\ \ \eta_{2m}^{'}=\max_{z\in[\alpha;\beta]}G_{2m}(z)$$

avec C défini précédemment et G_{2m} comme au théorème 4.1.4.

Démonstration. Soient ω défini par (4.3) et $Q\in\mathcal{P}_m$ à coefficients réels. Alors $x\mapsto\frac{Q(x)/Q(z)-1}{z-x}\in\mathcal{P}_{m-1}$ est orthogonal à Q_m par rapport à la mesure μ/ω. On obtient alors

$$\begin{aligned}f^{[\mu]}(z)-r_m(f^{[\mu]})(z)=\frac{\omega(z)}{Q_m^2(z)}\int_\alpha^\beta\frac{Q_m^2(x)}{\omega(x)}\frac{d\mu(x)}{z-x}&=\frac{\omega(z)}{Q_m(z)}\int_\alpha^\beta\frac{Q_m(x)}{\omega(x)}\frac{d\mu(x)}{z-x}\\&=\frac{\omega(z)}{Q_m(z)Q(z)}\int_\alpha^\beta\frac{Q_m(x)Q(x)}{\omega(x)}\frac{d\mu(x)}{z-x}.\end{aligned}$$

Pour tout $z\in\mathbb{C}$ avec $|z|=1$ et pour tout $x\leqslant\beta<-1$:

$$\frac{1}{1-x}\leqslant Re(\frac{1}{z-x})\leqslant\frac{1}{|z-x|}\leqslant\frac{1}{-1-x}\leqslant\frac{1}{1-x}\frac{1-\beta}{-1-\beta},\tag{4.20}$$

et en appliquant l'inégalité de Cauchy-Schwarz dans la dernière intégrale, on obtient

$$|f^{[\mu]}(z)-r_m(f^{[\mu]})(z)|^2\leqslant\frac{\omega(z)}{Q_m^2(z)}\int_\alpha^\beta\frac{Q_m^2(x)}{\omega(x)}\frac{d\mu(x)}{|z-x|}\frac{\omega(z)}{|Q^2(z)|}\int_\alpha^\beta\frac{|Q^2(x)|}{\omega(x)}\frac{d\mu(x)}{|z-x|}.$$

Par conséquent, le premier facteur est $\leqslant\frac{1-\beta}{-1-\beta}|f(z)$ d'après (4.20) et on obtient les erreurs absolues et relative :

$$|z|=1:\ \ |f^{[\mu]}(z)-r_m(f^{[\mu]})(z)|\leqslant C\min_{\deg Q\leqslant m}\Big\|\frac{\omega}{Q^2}\Big\|_{L^\infty(\partial\mathbb{D})}\Big\|\frac{Q^2}{\omega}\Big\|_{L^\infty([\alpha;\beta])}\tag{4.21}$$

avec $C=\frac{1-\beta}{-1-\beta}f^{[\mu]}(-1)\leqslant(\frac{1-\beta}{-1-\beta})^2Re(f^{[\mu]}(z))\leqslant(\frac{1-\beta}{-1-\beta})^2|f^{[\mu]}(z)|$.

Lorsque les points d'interpolation sont de multiplicité paire, disons $z_{m+j}=z_j$ pour $j=1,2,\ldots,m$ alors on obtient la borne supérieure du théorème 4.1.10 en prenant $Q(x)=(1-z_1z)(1-z_2z)\ldots(1-z_{2m}z)$ qui, par utilisation du principe de maximum pour les fonctions analytiques, est un polynôme extremal pour (4.21) dans ce cas. Enfin dans le cas général, on utilise le même polynôme que pour l'intervalle $[\alpha;\beta]$ dans le théorème 4.1.4 qui est optimal pour (4.20) à un facteur $4/(1-\eta_{2m}^{'})^2$ près. □

Remarque 4.1.11. *Comme dans la partie précédente, nous pouvons chercher à optimiser la borne supérieure énoncée avec un choix particulier des points d'interpolation, notamment dans le cas d'un unique point d'interpolation multiple z_1 d'ordre $2m$ ou de points libres. Les résultats de la section précédente restent valables dans ce cas en choisissant $[c;d]=[1/\beta;1/\alpha]$.*
Dans le cas d'approximant de Padé-Faber, Knizhnermann [64] considère le cas de points d'interpolation $z_1=z_2=\ldots=z_{2m}=0$ et obtient alors $\eta_{2m}^{'}=(1/\beta)^{2m}$.

Dans le cas plus général où $z_j\in[\frac{1}{\alpha};0]$, l'erreur se simplifie :

Lemme 4.1.12. *Soit $m\geqslant1$ et supposons que $[\alpha;\beta]\subseteq[-\infty;-1[$ et que $\forall j=1,\ldots,2m$, $z_j\in[\frac{1}{\alpha};0]$. Alors*

$$\max_{|z|=1}\Big|\frac{\omega(z)}{Q_m(z)^2}\Big|=\Big|\frac{\omega(-1)}{Q_m(-1)^2}\Big|$$

et donc

$$\|f^{[\mu]}-r_m(f^{[\mu]})\|_{L^\infty(\overline{\mathbb{D}})}=|f^{[\mu]}(-1)-r_m(f^{[\mu]})(-1)|\leqslant|f^{[\mu]}(-1)|\max_{z\in\overline{\mathbb{D}}}\Big|\prod_{j=1}^{2m}\frac{1-zz_j}{z-z_j}\Big|.$$

Démonstration. Pour $x \in [\alpha, \beta]$, $\min_{z\in\partial\mathbb{D}} z - x$ est atteint pour $z = -1$ d'où étant donné les propriétés de ω et Q_m, $\max_{|z|\leqslant 1} \int_\alpha^\beta \left|\frac{Q_m^2(x)}{\omega(x)}\right| \frac{d\mu(x)}{|z-x|} = \int_\alpha^\beta \left|\frac{Q_m^2(x)}{\omega(x)}\right| \frac{d\mu(x)}{|-1-x|}$. Il nous reste alors à démontrer que $\max_{|z|=1} \left|\frac{\omega(z)}{Q_m(z)^2}\right| = \left|\frac{\omega(-1)}{Q_m(-1)^2}\right|$. Soit z_0 un pôle et z_1 un point d'interpolation (les points d'interpolation doivent être réels ou par paire de conjugués d'après nos hypothèses de départ). Alors $\max_{|z|=1} |\frac{\omega(z)}{Q_m^2(z)}| = |\frac{\omega(-1)}{Q_m^2(-1)}|$. En effet, puisque $z_1 \in [1/\alpha; 0]$ et $z_0 \in [\alpha; \beta]$, en notons $z = \cos(t) + i\sin(t)$ on obtient

$$\left|\frac{z - z_1}{z - z_0}\right|^2 = \frac{(\cos(t) - z_1)^2 + \sin(t)^2}{(\cos(t) - z_0)^2 + \sin(t)^2} = \frac{1 + z_1^2 - 2z_1\cos(t)}{1 + z_0^2 - 2z_0\cos(t)} =: h(t)$$

h ainsi définie est une fonction continue en t, dérivable et vérifie

$$h^{'}(t) = \frac{2\sin(t)(z_1 - z_0)(1 - z_1 z_0)}{(1 + z_0^2 - 2z_0\cos(t))^2}$$

Or par hypothèse $z_1 \geqslant \frac{1}{\beta}$, $z_0 \in [\alpha; \beta] \subseteq]-\infty; -1]$, d'où $(z_1 - z_0) \geqslant 0$ car $z_0 \leqslant \beta \leqslant \frac{1}{\beta} \leqslant z_1$ et $1 - z_1 z_0 \leqslant 0$ si $z_1 \leqslant \frac{1}{z_0}$, ou $1 - z_1 z_0 \geqslant 0$ si $z_1 \geqslant \frac{1}{z_0}$, d'où le signe de $h^{'}(t)$ dépend du signe de $\sin(t)$ et h atteint son maximum en valeur absolue en $t = \pi$, soit lorsque $z = -1$.

En conclusion, lorsque $z_j \in [\frac{1}{\alpha}; 0]$, $\max_{|z|=1} \left|\frac{\omega(z)}{Q_m(z)}\right| \leqslant \left|\frac{\omega(-1)}{Q_m^2(-1)}\right|$ et notre lemme est démontré. □

4.2 Représentation des interpolants rationnels aux points optimaux

Nous supposons ici que les points d'interpolation sont réels distincts, $\mathbb{E} = [c; d] \subset \mathbb{R}\backslash[\alpha; \beta]$ en corrélation avec le choix optimisé des points d'interpolation, et ordonnés tels que

$$\beta < c \leqslant z_1 < z_2 < \ldots < z_{2m} \leqslant d \tag{4.22}$$

avec $c, d \in \mathbb{R}$ tels que $-\infty \leqslant \alpha < \beta < c \leqslant d < +\infty$, $m \in \mathbb{N}^*$ fixé. Le calcul du Padé multipoint $r_m(f^{[\mu]}) = P_m/Q_m$ de type $[m-1|m]$ ou $[m|m]$ est fortement lié à la représentation de celui-ci. Une première idée serait de calculer $r_m(f^{[\mu]})$ sous la forme d'un quotient de polynômes en résolvant un système linéaire homogène de $2m$ équations et $2m+1$ inconnues qui sont les coefficients de P_{m-1} et Q_m issu de la condition d'interpolation (4.2), $f(z_j)Q_m(z_j) - P_m(z_j) = 0$ pour $j = 1, \ldots, 2m$. Or en pratique, la résolution de ce système nous amène à considérer une matrice très mal conditionnée et l'on perd alors la précision sur les coefficients et donc également sur l'approximation. Il existe d'ailleurs une règle générale pour les approximants de Padé exprimés dans la base des monômes [6, Section 2.1] selon laquelle on perd jusqu'à m décimales de précision lors de la résolution du sytème homogène. D'autres bases de polynômes peuvent être employées comme les polynômes de Chebyshev sur l'intervalle $[c; d]$ ou une base de Newton correspondant à ordonner les points d'interpolation de manière optimisée conduisant à un meilleur conditionnement. Cependant, ces bases vont dépendre uniquement de l'intervalle $[c; d]$ sur lequel on approche notre fonction. Or un 'bon' interpolant rationnel devrait dépendre non seulement de l'intervalle d'étude, mais également de la fonction f. Pour ces raisons, nous n'implémentons pas l'interpolant rationnel sous forme d'un quotient de polynômes et recherchons d'autres formes du Padé multipoint aux points d'interpolation optimisés.

4.2.1 Représentation en éléments simples des interpolants rationnels

En considérant des Padé multipoints d'ordre $[m-1|m]$, on peut rechercher une forme de ces fonctions exploitant cet ordre particulier. Une toute première idée est de prendre une représentation en éléments

simples. Etant donné que les pôles de nos interpolants sont simples, il apparaît que pour $r_m(f^{[\mu]})$ Padé multipoint avec m pôles simples $x_1, \ldots, x_m$ dans $(\alpha;\beta)$, $r_m(f^{[\mu]})$ admet une représentation en éléments simples sous la forme

$$r_m(f^{[\mu]})(z) = \frac{a_1}{z-x_1} + \frac{a_2}{z-x_2} + \ldots + \frac{a_m}{z-x_m} \tag{4.23}$$

avec résidus $a_j > 0$ pour $j = 1, \ldots, m$. D'après les travaux de Mayo et Antoulas [71], résumé récemment dans [33], l'interpolant r_m sous forme d'éléments simples peut être représentée comme une fonction de transfert d'un système dynamique SISO à l'aide d'un faisceau de matrices : en effet, on peut noter que $r_m(f^{[\mu]})(z) = W(\mathbb{L}_s - z\mathbb{L})^{-1}V^T$ où on définit les vecteurs lignes $W = (f(z_{2j-1}))_{j=1,\ldots,m}$, $V = (f(z_{2j}))_{j=1,\ldots,m}$ et les matrices de Loewner

$$\mathbb{L} = \left(\frac{f(z_{2j}) - f(z_{2k-1})}{z_{2j} - z_{2k-1}}\right)_{\substack{j=1,\ldots,m \\ k=1,\ldots,m}} \text{ et } \mathbb{L}_s = \left(\frac{z_{2j}f(z_{2j}) - z_{2k-1}f(z_{2k-1})}{z_{2j} - z_{2k-1}}\right)_{\substack{j=1,\ldots,m \\ k=1,\ldots,m}}$$

d'après les conditions d'interpolation. Ainsi les pôles de $r_m(f^{[\mu]})$ sont les valeurs propres du faisceau de matrices $\mathbb{L}_s - z\mathbb{L}$, que l'on peut déterminer à l'aide de commandes standards, ici **eig** sur Matlab. Une fois que ces pôles ont été déterminés, il ne nous reste plus qu'à trouver les résidus $a_1, \ldots, a_m$ en résolvant un problème de moindres carrés :

$$MX = F, \text{ avec } M = \left(\frac{1}{z_j - x_k}\right)_{\substack{j=1,\ldots,2m \\ k=1,\ldots,m}}, \quad X = (a_j)_{j=1,\ldots,m}, \quad F = (f(z_j))_{j=1,\ldots,2m}.$$

Algorithm 8 f fonction donnée et $2m$ points d'interpolation $z_1, ..., z_{2m}$, calcule les pôles x_j et les résidus a_j de la décomposition en éléments simples (4.23) du Padé multipoint aux points optimisés de f d'ordre $[m-1|m]$.

Ensure: Pôles $x_1, ..., x_m$ et les résidus $a_1, ..., a_m$ dans (4.23).

Déterminer les valeurs propres $x_1, ..., x_m$ du faisceau de matrices $\mathbb{L}_s - z\mathbb{L}$ de Mayo et Antoulas ;

Calculer la solution $y = (a_1, ..., a_m)^T$ du problème de moindres carrés

$$\min_{y \in \mathbb{R}^m \setminus \{0\}} \| \Big(\frac{1}{z_j - x_k}\Big)_{j=1,\ldots,2m, k=1,\ldots,m} y - \Big(f(z_j)\Big)_{j=1,\ldots,2m} \|;$$

4.2.2 Représentation barycentrique des interpolants rationnels

Pour une fonction rationnelle $r_m = p_m/q_m \in \mathcal{R}_{m,m}$, en représentant le numérateur p_m et le dénominateur q_m sous forme barycentrique, on obtient une nouvelle représentation de cette fonction rationnelle

$$r_m(z) = \sum_{j=0}^{m} \frac{\alpha_j}{z-t_j} \Bigg/ \sum_{j=0}^{m} \frac{\beta_j}{z-t_j}, \tag{4.24}$$

appelée représentation barycentrique [16] où les points t_j sont appelés points de support et $\alpha_j, \beta_j \in \mathbb{C}$.
Il existe des résultats connus sur l'erreur commise si on évalue sur l'axe réel notre fonction rationnelle sous forme barycentrique $r_m(z) = \sum_{j=0}^{m} \frac{\alpha_j}{z-t_j} \Big/ \sum_{j=0}^{m} \frac{\beta_j}{z-t_j}$ en précision finie : la backward stability a été démontrée par [34, Section 2.3], [56, p. 551] et la forward stability par [34, Lemma 2.1] obtenue en adaptant l'analyse de O. Celis [22, Proposition 2.4.3].

Lorsque r_m satisfait la condition d'interpolation $r_m(t_j) = f(t_j)$, il nous faut alors imposer $\alpha_j = \beta_j f(t_j)$ pour $j = 0, \ldots, m$. Dans [34], ces interpolants rationnels avec points de support optimisés sont exploités dans une nouvelle implémentation appelée *minimax* de l'algorithme de Remez rationnel [34] qui donne en sortie un meilleur approximant rationnel de type $[m'|m]$ pour des fonctions de Markov avec $m', m \leqslant 40$ à précision machine près.
Pour un interpolant rationnel de type $[m|m]$, on impose un choix des points de support parmi les points d'interpolation $z_1, \ldots, z_{2m+1}$ en sélectionnant $t_j = z_{2j+1}$ pour $j = 0, \ldots, m$, puis on résout le système linéaire en les coefficients $\beta_0, \ldots, \beta_m$ pour obtenir la propriété d'interpolation aux points z_{2j} pour $j = 1, \ldots, m$, voir algorithme 9. On peut noter que cette résolution implique l'emploi de la transposé de la matrice de Loewner $\mathbb{L}$ que nous avons déjà vu en sous-section 4.2.1 à laquelle on ajoute une colonne puisque $\alpha_j = f(t_j)\beta_j$ pour tout $j = 0, \ldots, m$. Cette répartition des $m + 1$ points de support parmi les $2m$ points d'interpolation est motivée par [34, Corollary 4.5] montrant qu'une certaine matrice de Cauchy après multiplication à gauche et à droite par une matrice diagonale soit à lignes orthonormées
Dans notre cas de Padé multipoints de type $[m-1|m]$, nous considérons les points de support $t_0 = z_1$ et $t_j = z_{2j}$ pour $j = 1, \ldots, m$ pour assurer l'entrelacement avec les points d'interpolation restants. Puis comme pour le cas d'interpolant d'ordre $[m|m]$, nous résolvons un système linéaire en les coefficients $\beta_0, \ldots, \beta_m$ à l'aide de la transposé de la matrice de Loewner $\mathbb{L}$ pour assurer l'interpolation aux points restants z_{2j+1} pour $j = 1, \ldots, m-1$, où on impose également que $f(t_0)\beta_0 + f(t_1)\beta_1 + \ldots + f(t_m)\beta_m = 0$ pour s'assurer que les degrés sont corrects.

r ainsi définie vérifie donc

$$\sum_{k=0}^{m} \frac{f(t_k)\beta_j}{z_{2j+1} - t_k} = f(z_{2j+1}) \sum_{k=0}^{m} \frac{\beta_k}{z_{2j+1} - t_k} \quad \Rightarrow \quad \sum_{k=0}^{m} \beta_k \frac{f(z_{2j+1}) - f(t_k)}{z_{2j+1} - t_k} = 0$$

$$\Rightarrow \quad \begin{pmatrix} \frac{f(z_{2j+1})-f(t_k)}{z_{2j+1}-t_k} \\ 1 \ldots 1 \end{pmatrix}_{\substack{j=1,\ldots,m-1 \\ k=1,\ldots,m+1}} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_m \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

et l'on obtient les paramètres β_j par SVD de la matrice du système homogène.

Algorithm 9 f une fonction donnée et $2m + 1$ points d'interpolation $z_1, ..., z_{2m+1}$, calcule les points de support t_j et les poids β_j, $\alpha_j = f(t_j)\beta_j$ de l'interpolant rationnel (4.24) de f d'ordre $[m|m]$.

Ensure: Pour $j = 0, 1, ..., m$: points de support t_j et les poids β_j, $\alpha_j = f(t_j)\beta_j$ de (4.24).
Sélectionner les points de support $t_0 = z_1$ et $t_j = z_{2j}$ pour $j = 1, \ldots, m$;
Calculer une solution $y = (\beta_0, ..., \beta_m)^T$ du système d'équations linéaires homogène

$$\begin{pmatrix} \frac{f(z_{2j+1})-f(t_k)}{z_{2j+1}-t_k} \\ 1 \ldots 1 \end{pmatrix}_{\substack{j=1,\ldots,m-1 \\ k=1,\ldots,m+1}} y = 0.$$

4.2.3 Représentation en fraction continue des interpolants rationnels

Nous nous intéressons à présent à une représentation sous forme de fraction continue de Thiele de nos Padé multipoints. Selon [6, Section 7.1], pour des points d'interpolation $z_1, z_2, \ldots$ et paramètres

$f_1^{(1)}, f_2^{(2)}, \ldots \in \mathbb{C}$, la M^{e} fraction continue de Thiele est la fonction rationnelle

$$R_M^{(1)} = f_1^{(1)} + \frac{z - z_1}{\Big|\ f_2^{(2)}} + \ldots + \frac{z - z_{M-1}}{\Big|\ f_M^{(M)}}. \tag{4.25}$$

On parle de fraction continue de Thiele positive si tous les paramètres $f_j^{(j)}$ sont strictement positifs et (4.22) est vérifié.

Etant donnée une fonction $f^{(1)}$, on définit ses différences réciproques par

$$\forall 1 \leqslant k \leqslant M : \quad f_k^{(1)} = f^{(1)}(z_k) \text{ et } \forall 1 \leqslant j < k \leqslant M : \quad f_k^{(j+1)} = \frac{z_k - z_j}{f_k^{(j)} - f_j^{(j)}}. \tag{4.26}$$

en supposant qu'il n'y ait pas de division par zéro. Alors $R_M^{(1)}(z_k) = f^{(1)}(z_k)$ pour tout $k = 1, \ldots, M$. Plus précisément, $R_M^{(1)} = R_{2m+1}^{(1)}$ est un interpolant de type $[m|m]$ de $f^{(1)}$ aux points $z_1, \ldots, z_{2m+1}$ et $R_M^{(1)} = R_{2m}^{(1)}$ est un interpolant de type $[m|m-1]$ de $f^{(1)}$ aux points $z_1, \ldots, z_{2m}$. Enfin, en posant $f^{(1)} := 1/f(z)$, alors $1/R_{2m}(z)$ est un interpolant rationnel de type $[m-1, m]$ de f. Cette propriété d'interpolation devient immédiate en introduisant la famille de fonctions

$$\forall 1 \leqslant j < k \leqslant M : \quad f^{(j+1)}(z) = \frac{z - z_j}{f^{(j)}(z) - f^{(j)}(z_j)}, \quad R_M^{(j+1)}(z) = \frac{z - z_j}{R_M^{(j)}(z) - R_M^{(j)}(z_j)}, \tag{4.27}$$

et alors $f_k^{(j)} = f^{(j)}(z_k) = R_M^{(j)}(z_k)$ pour $1 \leqslant j \leqslant k \leqslant M$, et

$$R_M^{(j)}(z) = f_j^{(j)} + \frac{z - z_j}{\Big|\ f_{j+1}^{(j+1)}} + \cdots + \frac{z - z_{M-1}}{\Big|\ f_M^{(M)}}.$$

En particulier, $R_M^{(j)}$ interpôle $f^{(j)}$ aux points $z_j, z_{j+1}, \ldots, z_M$. Le schéma d'évaluation inverse en un argument fixe z d'une fraction continue de Thiele avec coefficients $f_j^{(j)}$ est donnée par

$$R_M^{(M)}(z) = f_M^{(M)}, \quad \text{et pour } j = M-1, M-2, \ldots, 1 : \quad R_M^{(j)} = f_j^{(j)} + \frac{z - z_j}{R_M^{(j+1)}(z)}. \tag{4.28}$$

Pour cet algorithme de construction, explicité dans [48], Graves-Morris suggère de réordonner sous forme d'un pivot de Gauss les couples de valeurs $(z_k, f_k^{(j)})$ pour $k = j, j+1, \ldots, M$ de sorte que pour tout $j = 1, \ldots, M$,

$$|f_j^{(j)}| = \min\{|f_k^{(j)}| : k = j, j+1, \ldots, M\}. \tag{4.29}$$

En combinant (4.26), (4.29) et (4.28), on obtient ainsi l'algorithme modifié 10 de Thacher-Tukey [48] et [6] en omettant le cas d'une division par zéro.

Algorithm 10 Pour une fonction $f^{(1)}$ et M points d'interpolation $z_1, \ldots, z_M$, calcule et évalue en une valeur z la représentation sous forme de fraction continue de Thiele (4.25) via l'algorithme modifié de Thacher-Tukey [48] de l'interpolant rationnel de $f^{(1)}$ d'ordre $[m|m-1]$ (si $M = 2m$) ou d'ordre $[m|m]$ (si $M = 2m+1$).

Ensure: Coefficients $f_1^{(1)}, \ldots, f_M^{(M)}$ de la représentation (4.25) et valeur $R_M^{(1)}(z)$ de l'interpolant.
  for $k = 1, \ldots, M$, **do**
    Initialiser $f_k^{(1)} = f^{(1)}(z_k)$;
  end for
  for $j = 1, \ldots, M-1$ **do**
    Permuter les couples $(f_k^{(j)}, z_k)$ pour $k = j, j+1, \ldots, M$ de sorte à vérifier la condition (4.29) ;
    for $k = j, j+1, \ldots, M$ **do**
      $f_k^{(j+1)} = (z_k - z_j)/(f_k^{(j)} - f_j^{(j)})$;
    end for
  end for
  Initialiser $R_M^{(M)} = f_M^{(M)}$;
  for $j = M-1, M-2, \ldots, 1$ **do**
    $R_M^{(j)}(z) = f_j^{(j)} + (z - z_j)/R_M^{(j+1)}(z)$;
  end for

Remarque 4.2.1. *On peut noter que si $R_M^{(1)}$ est un fraction continue positive, alors par récurrence sur $k-j$, à l'aide de* (4.22) *et* (4.26) *on peut montrer que $0 < f_j^{(j)} < f_k^{(j)}$ pour $1 \leqslant j < k \leqslant M$ de sorte qu'on ne divise jamais par zéro dans* (4.26) *et alors* (4.29) *est vérifié sans pivotage nécessaire. Cependant, nous n'avons pas connaissance dans la littérature de l'existence d'une classe de fonctions dont les fractions continues de Thiele soient positives. Nous donnons ici une telle classe de fonctions, résultat que nous démontrons à la fin de cette section.*

Théorème 4.2.2. *Si $1/f^{(1)}$ est une fonction de Markov avec mesure à support infini dans $[\alpha; \beta]$, alors toutes les fonction $1/f^{(j)}$ données par* (4.26) *sont également des fonctions de Markov avec mesure respectives μ_j avec support infini dans $[\alpha; \beta]$. Par conséquent, puisque ces fonctions de Markov sont positives décroissantes sur $(\beta; +\infty)$, $f_k^{(j)} > f^{(j)}(z_j) > 0$ pour tout $k > j$ et alors la représentation en fraction continue de Thiele de l'interpolant de $f^{(1)}$ est positive.*

Exemple 4.2.3. *Soit $f^{(1)}(z) = \sqrt{z}$ de sorte que $1/f^{(1)}(z)$ est une fonction de Markov avec support $[\alpha; \beta] = (-\infty; 0]$. Alors $f_k^{(1)} = \sqrt{z_k}$ et pour tout $j \geqslant 2$,*

$$f^{(j)}(z) = \sqrt{z} + \sqrt{z_{j-1}}, \quad f_k^{(j)} = \sqrt{z_k} + \sqrt{z_{j-1}} > 0.$$

En particulier, $1/f^{(j)}(z)$ est également une fonction de Markov avec support $[\alpha; \beta] = (-\infty; 0]$ et la fraction continue de Thiele est positive. Une telle formule explicite ne se trouve pas, à notre connaissance, dans la littérature actuelle, excepté dans le cas limite des approximants de Padé ([57, Theorem 5.9]).

Nous donnons à présent un résultat sur la stabilité backward de l'algorithme de Thacher-Tukey modifié : en précision finie, (4.26) donne les valeurs $f_k^{(j)}$ de la fraction continue avec valeurs exactes aux points z_k proches des valeurs recherchées $f^{(j)}(z_k)$. Pour démontrer ce résultat, nous nous sommes inspirés d'un résultat similaire [47, Theorem 4.1] de Graves-Morris qui considère les fractions continues de Thiele non nécessairement positives avec pivotage (4.29). Notre apport principal est de supprimer un coefficient 2^{k-j} qui croît exponentiellement avec M.

D'après [55, equation (2.4)], on peut noter lorsque $\varepsilon > 0$ désigne la précision machine que les opérations

élémentaires en précision arithmétique finie vérifient

$$\widetilde{(x \pm y)} = (x \pm y)(1+\delta), \ \ |\delta| < \varepsilon$$
$$\widetilde{(xy)} = (xy)(1+\delta), \ \ |\delta| < \varepsilon$$
$$\widetilde{(x/y)} = (x/y)(1+\delta), \ \ |\delta| < \varepsilon.$$

Théorème 4.2.4. *Soit $1/f^{(1)}$ fonction de Markov et supposons que les valeurs $\tilde{f}_k^{(j)}$ pour $1 \leqslant j \leqslant k \leqslant M$ sont calculées via (4.26) en utilisant la précision arithmétique finie avec précision machine ε. Notons $\tilde{R}_M^{(1)}$ la fonction continue exacte construite avec les coefficients (inexactes) $\tilde{f}_1^{(1)}, \tilde{f}_2^{(2)}, \ldots, \tilde{f}_M^{(M)}$ supposées strictement positives. Alors*

$$k = 1, \ldots, m: \quad |\tilde{R}_M^{(1)}(z_k) - f^{(1)}(z_k)| \leqslant \frac{3k\varepsilon}{1-3k^2\varepsilon} |\tilde{R}_M^{(1)}(z_k)|.$$

Démonstration. Considérons les fonctions rationnelles issues par (4.28) et données par

$$\tilde{R}_M^{(M)}(z) = \tilde{f}_M^{(M)}, \text{ et pour tout } j = M-1, M-2, \ldots, 1: \ \ \tilde{R}_M^{(j)}(z) = \tilde{f}_j^{(j)} + \frac{z - z_j}{\tilde{R}_M^{(j+1)}(z)}. \tag{4.30}$$

On montre alors par récurrence sur $k-j$ que

$$\tilde{f}_k^{(j)} = (1+\delta_{j,k})\tilde{R}_M^{(j)}(z_k), \ \ |\delta_{j,k}| \leqslant \gamma_{k-j}, \text{ avec } \gamma_l = \frac{3l\varepsilon}{1-3l^2\varepsilon}. \tag{4.31}$$

Le cas $k = j$ est évident puisque par définition, $\tilde{f}_j^{(j)} = \tilde{R}_M^{(j)}(z_j)$. Supposons à présent $k > j$. On peut noter pour tout $k > j$

$$\widetilde{z_k - z_j} = (z_k - z_j)(1+\delta_{j,k}), \ \widetilde{f_k^{(j)} - f_j^{(j)}} = (\tilde{f}_k^{(j)} - \tilde{f}_j^{(j)})(1+\Delta_{j,k}), \tag{4.32}$$

$$\Big(\widetilde{\frac{z_k - z_j}{f_k^{(j)} - f_j^{(j)}}}\Big) = \Big(\frac{\widetilde{z_k - z_j}}{\widetilde{f_k^{(j)} - f_j^{(j)}}}\Big)(1+\theta_{j,k}) \tag{4.33}$$

où $|\delta_{j,k}|, |\Delta_{j,k}|, |\theta_{j,k}| < \varepsilon$ et alors

$$f^{(1)}(z_k) = \tilde{f}_k^{(1)}(1+\varepsilon_{1,k}), \ \ \tilde{f}_k^{(j+1)} = \frac{z_k - z_j}{\tilde{f}_k^{(j)} - \tilde{f}_k^{(j)}}(1+\varepsilon_{j+1,k})$$

où $\varepsilon_{1,k}$ provient de l'arrondi de $f^{(1)}(z_k)$ et $\varepsilon_{j+1,k}$ provient des erreurs dans (4.32) et (4.33) et on montre alors que $|\varepsilon_{j+1,k}| < \frac{3\varepsilon}{1-3\varepsilon}$ [55, Lemma 3.5].
On peut noter à l'aide de (4.30) et (4.31) valable pour $j+1$ par hypothèse de récurrence que

$$\begin{aligned}
\tilde{f}_k^{(j)} - \tilde{R}_M^{(j)}(z_k) &= \tilde{f}_k^{(j)} - \tilde{f}_j^{(j)} - (\tilde{R}_M^{(j)}(z_k) - \tilde{f}_j^{(j)}) \\
&= \frac{z_k - z_j}{\tilde{f}_k^{(j+1)}}(1+\varepsilon_{j+1,k}) - (\tilde{R}_M^{(j)}(z_k) - \tilde{f}_j^{(j)}) \\
&= \frac{z_k - z_j}{\tilde{R}_M^{(j+1)}(z_k)}\Big(\frac{1+\varepsilon_{j+1,k}}{1+\delta_{j+1,k}}\Big) - (\tilde{R}_M^{(j)}(z_k) - \tilde{f}_j^{(j)}) \\
&= \Big(\frac{1+\varepsilon_{j+1,k}}{1+\delta_{j+1,k}} - 1\Big)(\tilde{R}_M^{(j)}(z_k) - \tilde{f}_j^{(j)}),
\end{aligned}$$

On obtient finalement (4.31) en observant que $|\tilde{R}_M^{(j)}(z_k) - \tilde{f}_j^{(j)}| = \tilde{R}_M^{(j)}(z_k) - \tilde{R}_M^{(j)}(z_j) \leqslant \tilde{R}_M^{(j)}(z_k)$, par hypothèse $\tilde{f}_j^{(j)} > 0$ pour $j = 1, \ldots, M$ et par l'inégalité

$$\begin{aligned}
\Big|\Big(\frac{1+\varepsilon_{j+1,k}}{1+\delta_{j+1,k}} - 1\Big)\Big| &\leqslant \Big|\Big(\frac{\varepsilon_{j+1,k} - \delta_{j+1,k}}{1+\delta_{j+1,k}}\Big)\Big| \leqslant \frac{|\varepsilon_{j+1,k}| + |\delta_{j+1,k}|}{|1+\delta_{j+1,k}|} \leqslant \frac{|\varepsilon_{j+1,k}| + \gamma_{k-j-1}}{1-\gamma_{k-j-1}} \\
&\leqslant \frac{|\varepsilon_{j+1,k}| + \frac{3(k-j-1)\varepsilon}{1-3(k-j-1)^2\varepsilon}}{1 - \frac{3(k-j-1)\varepsilon}{1-3(k-j-1)^2\varepsilon}} = \frac{1 - 3(k-j-1)^2\varepsilon + 3(k-j-1)\varepsilon}{1 - 3(k-j-1)^2\varepsilon - 3(k-j-1)\varepsilon} \\
&\leqslant \frac{3(k-j)\varepsilon}{1-3(k-j)^2\varepsilon} = \gamma_{k-j}.
\end{aligned}$$

On en déduit alors le résultat du théorème pour le cas $j = 1$. □

Remarque 4.2.5. *Des expériences numériques répétées ont fait apparaître le fait que les coefficients $\tilde{f}_k^{(j+1)}$ du théorème 4.2.4 ne vérifient plus la positivité lorsque l'erreur $R_j^{(1)}(z) - f^{(1)}(z)$ pour $z \in [c;d]$ est proche de la précision machine.*

Afin de pouvoir prouver le théorème 4.2.2, nous énonçons ici un lemme, partiellement connu du problème des moments classique de Stieltjes à changement de variable près, voir [6, Section 5.2 et 5.3] ou [18, Theorem V.4.4]. Dans la suite de cette section, on suppose que $\alpha < \beta < z_1$.

Pour toute fonction g analytique dans un voisinage d'un point z_1, on définit la matrice de Hankel à l'aide des coefficients de Taylor en z_1 par

$$\mathcal{H}_n^{(l)}(g) = \begin{bmatrix} g_l & g_{l+1} & \cdots & g_{n+l} \\ g_{l+1} & g_{l+2} & \cdots & g_{n+l+1} \\ \vdots & \vdots & & \vdots \\ g_{n+l} & g_{n+l+1} & \cdots & g_{2n+l} \end{bmatrix}, \quad g(z) = \sum_{j=0}^{\infty} g_j (z - z_1)^j. \tag{4.34}$$

Dans [6, Theorem 5.3.1], les auteurs donnent une condition nécessaire et suffisante sur les matrices de Hankel associées pour qu'une fonction analytique soit une fonction de Stieltjes. En nous en inspirant, nous donnons un résultat similaire pour le cas des fonction de Markov.

Lemme 4.2.6. *Si f est une fonction de Markov avec mesure positive à support infini dans $[\alpha;\beta]$, alors pour tout $n \geqslant 0$, les matrices de Hankel $\mathcal{H}_n^{(0)}(f)$ sont définies positives, et les matrices de Hankel $\mathcal{H}_n^{(1)}(f)$ sont définies négatives. Inversement, si f est analytique dans $\overline{\mathbb{C}}\backslash[\alpha;\beta]$, avec matrices de Hankel $\mathcal{H}_n^{(0)}(f)$ définies positives et $\mathcal{H}_n^{(1)}(f)$ définies négatives pour tout $n \geqslant 0$, alors f est une fonction de Markov avec mesure positive μ à support infini dans $[\alpha;\beta]$.*

Démonstration. Supposons dans un premier que f est une fonction de Markov avec mesure à support infini inclus dans $[\alpha;\beta]$ et notons pour tout $j \geqslant 0$, $f_j := \int_\alpha^\beta \frac{d\mu(x)}{(z_1-x)(x-z_1)^j}$ les moments de μ qui sont les coefficients du développement de Taylor de f au voisinage de z_1. Alors pour tout $p = (p_0, p_1, \ldots, p_n)^T \in \mathbb{R}^{n+1}$,

$$\begin{aligned}
p^T \mathcal{H}_n^{(0)}(f) p = \sum_{j=0}^{n}\sum_{k=0}^{n} p_j f_{j+k} p_k &= \int_\alpha^\beta \frac{1}{z_1 - x}\Big(\sum_{j=0}^{n}\sum_{k=0}^{n} \frac{p_j}{(x-z_1)^j}\frac{p_k}{(x-z_1)^k}\Big) d\mu(x) \\
&= \int_\alpha^\beta \frac{1}{z_1 - x} P(\frac{1}{x-z_1})^2 d\mu(x) > 0
\end{aligned}$$

avec $P(x) = \sum_{j=0}^{n} p_j x^j \in \mathcal{P}_n$. De la même manière on montre que pour tout $p = (p_0, p_1, \ldots, p_n)^T \in \mathbb{R}^{n+1}$, $p^T \mathcal{H}_n^{(1)}(f) p < 0$. Plus généralement, pour tout $p = (p_0, \ldots, p_n)^T \in \mathbb{R}^{n+1}$ et touts $l, n \geqslant 0$,

$$p^T \mathcal{H}_n^{(l)}(f) p = \int_\alpha^\beta \frac{1}{(z_1 - x)(x - z_1)^l} P(\frac{1}{x - z_1})^2 d\mu(x) \text{ avec } P \in \mathcal{P}_n. \tag{4.35}$$

Ainsi, si l est paire alors $p^T \mathcal{H}_n^{(l)}(f) p > 0$ et si l est impaire, alors $p^T \mathcal{H}_n^{(l)}(f) p < 0$ d'après nos hypothèses de départ. On en conclut donc la première partie de notre théorème lorsque $l = 0$ et $l = 1$.

Supposons à présent que nous disposions d'une fonction f analytique dans $\overline{\mathbb{C}} \backslash [\alpha; \beta]$ avec matrices de Hankel $\mathcal{H}_n^{(0)}(f)$ définie positive et $\mathcal{H}_n^{(1)}(f)$ définie négative pour tout $n \geqslant 0$. Considérons les approximants de Padé p_m/q_m de type $[m-1|m]$ de f en z_1. D'après [6, Section 1.1, equation (1.8)], le dénominateur q_m est donné par

$$\begin{aligned} q_m(z) &= \begin{vmatrix} f_0 & \cdots & f_{m-1} & f_m \\ \vdots & & \vdots & \vdots \\ f_{m-1} & \cdots & f_{2m-2} & f_{2m-1} \\ (z-z_1)^{m+1} & \ldots & (z-z_1) & 1 \end{vmatrix} \\ &= \begin{vmatrix} f_0 - (z-z_1)f_1 & f_1 - (z-z_1)f_2 & \cdots & f_{m-1} - (z-z_1)f_m \\ f_1 - (z-z_1)f_2 & f_2 - (z-z_1)f_3 & \cdots & f_m - (z-z_1)f_{m+1} \\ \vdots & \vdots & & \vdots \\ f_{m-1} - (z_1)f_m & f_m - (z-z_1)f_{m+1} & \cdots & f_{2m-2} - (z-z_1)f_{2m-1} \end{vmatrix} \\ &= \det\Big(\mathcal{H}_{m-1}^{(0)}(f) - (z-z_1)\mathcal{H}_{m-1}^{(1)}(f)\Big). \end{aligned}$$

De plus, d'après l'identité du déterminant de Sylvester [6, Theorem 1.4.1], il existe une récurrence à 3 termes pour tout triplets de dénominateurs consécutifs une fois le signe des coefficients connu. On voit alors que la suite des dénominateurs est une suite de Sturm. Plus précisément pour tout $z \in \mathbb{R}$, la suite $\mathcal{M}(z) = \{q_m(z), m \geqslant 1\}$ est une suite de Sturm c'est-à-dire que $q_{m-1}(z) q_{m+1}(z) < 0$ dès que $q_m(z) = 0$. En effet, on peut vérifier d'après l'identité du déterminant que

$$q_{m-1}(z) q_{m+1}(z) < 0, \text{ dès que } q_m(z) = 0 \tag{4.36}$$

pour tout $m \geqslant 1$ et donc pour tout $z \in \mathbb{R}$, $\mathcal{M}(z)$ est une suite de Sturm. p_m/q_m possède alors m pôles distincts $x_{1,m}, \ldots, x_{m,m} \in (-\infty; z_1)$ et ses résidus sont positifs. En effet, on peut noter que

— $q_0(z) = 1$;
— $q_1(z) = f_0 - (z - z_1) f_1$;
— $q_2(z) = q_1(z)(f_2 - (z-z_1)f_3) - (f_1 - (z-z_1)f_2)^2$

et ainsi de suite. De plus, $q_m(z_1) = \det\Big(\mathcal{H}_m^{(0)}(f\Big) > 0$ et pour tout z, $q_m(z) = \det\Big(- z\mathcal{H}_m^{(1)}(f) - (-\mathcal{H}_m^{(0)} - z_1 \mathcal{H}_m^{(1)}(f))\Big)$ d'où $q_m(-\infty) = (-1)^m \infty$. Or, $q_1(z_{1,1}) = 0$ lorsque $z_{1,1} := z_1 + \frac{f_0}{f_1} < z_1$ par hypothèse, $q_2(z_{1,1}) < 0$ d'après (4.36) et $q_2(z_1) = \det\Big(\mathcal{H}_2^{(0)}(f)\Big) > 0$ et $q_2(-\infty) > 0$, d'où il existe $z_{2,1} \in (z_{11}; z_1)$ et $z_{2,2} \in (-\infty; z_{11})$ tels que $q_2(z_{2,1}) = 0$ et $q_2(z_{2,2}) = 0$. Par récurrence immédiate, on trouve que q_m possède m racines distinctes $z_{m,j}$ pour $j = 1, \ldots, m$ dans l'intervalle $(-\infty; z_1)$ avec $\text{sign}(q_{m-1}(z_{m,k})) = (-1)^{k+1}$. Or d'après [6, Section 3.5, equation (5.18)],

$$p_{m+1}/q_{m+1}(z) - p_m/q_m(z) = \frac{\det\Big(\mathcal{H}_m^{(0)}(f)\Big)^2 (z - z_1)^{2m}}{q_{m+1}(z) q_m(z)} \tag{4.37}$$

d'où $p_{m+1}(z) q_m(z) - p_m(z) q_{m+1}(z) = (z - z_1)^{2m} \det\Big(\mathcal{H}_m^{(0)}(f)\Big) > 0$ d'où $\text{sign}(p_{m+1}(z)) = \text{sign}(q_m(z))$ si

$q_{m+1}(z) = 0$. Or, $\text{sign}(q_{m+1}(z_{m+1,k})) = (-1)^{k+1}$, et comme $q_{m+1}(z_1) > 0$,

$$\text{sign}\Big(\frac{d}{dx} q_{m+1}(z)\Big|_{z=z_{m+1,k}} \Big) = (-1)^{k+1}, \quad \text{et donc} \quad \frac{p_{m+1}(z)}{q'_{m+1}(z)}\Big|_{q_{m+1}(z)=0} > 0$$

d'où les résidus sont positifs. On peut donc écrire pour tout m,

$$\frac{p_m(z)}{q_m(z)} = \int_\alpha^\beta \frac{d\nu_m(x)}{z-x}, \;\; \nu_m = \sum_{j=1}^m a_{j,m}\delta_{x_{j,m}}.$$

En prenant ν une limite faible de la suite $(\nu_m)_m$ avec $supp(\nu) \subset (-\infty; z_1)$, on peut conclure que p_m/q_m tend vers la fonction de Markov $g(z) = \int \frac{d\nu(x)}{z-x}$ sur tout sous-ensemble compact de $\overline{\mathbb{C}}\backslash(-\infty, z_1]$. On observe à présent que, pour tout $\gamma > z_1$ et $k \geqslant 0$,

$$0 < (-1)^k \frac{g^{(k)}(\gamma)}{k!} = \int \frac{d\nu(x)}{(\gamma - x)^{k+1}} \leqslant (-1)^k \frac{g^{(k)}(z_1)}{k!} = (-1)^k \frac{f^{(k)}(z_1)}{k!}$$

où la dernière égalité provient de la condition d'interpolation des approximants de Padé. Par hypothèse sur f, la k^{e} racine de l'expression de droite a une lim sup > 0 ne dépendant pas du choix de γ. En choisissant γ suffisamment proche de z_1, on conclut que g est analytique sur un voisinage de z_1, et alors $g = f$ par unicité. Enfin, le support de ν est inclus dans l'intervalle $[\alpha; \beta]$ puisqu'une fonction de Markov ne peut pas avoir de continuation analytique sur le voisinage d'un point de $supp(\nu)$. □

Démonstration du Théorème 4.2.2. Nous démontrons le résultat uniquement pour le cas $j = 1$, les autres cas s'en suivent. Soit donc $f^{(1)} = 1/f$ avec f fonction de Markov avec mesure μ à support infini inclus dans $[\alpha; \beta]$. Puisque $f(z) \neq 0$ pour $z \notin [\alpha; \beta]$, $f^{(1)}$ est analytique sur $\overline{\mathbb{C}}\backslash[\alpha; \beta]$. Il en est de même pour la fonction

$$g(z) := \frac{f^{(1)}(z) - f^{(1)}(z_1)}{z - z_1}.$$

De plus, puisque f est une fonction non-réelle sur $\overline{\mathbb{C}}\backslash\mathbb{R}$ et strictement décroissante sur $\overline{\mathbb{R}}\backslash[\alpha; \beta]$, on observe également, d'après (4.27), que $f^{(2)} = 1/g$ est analytique sur $\overline{\mathbb{C}}\backslash[\alpha; \beta]$. Par conséquent, d'après la première partie du Lemme 4.2.6 appliquée à f, les matrices de Hankel associées à $1/f^{(1)}$

$$\mathcal{H}_n^{(0)}(\frac{1}{f^{(1)}}) = \mathcal{H}_n^{(0)}(f), \;\; \text{et} \;\; -\mathcal{H}_n^{(1)}(\frac{1}{f^{(1)}}) = -\mathcal{H}_n^{(1)}(f)$$

sont définies positives pour tout $n \geqslant 0$. D'après la seconde partie du Lemme 4.2.6 appliquée à la fonction $g = 1/f^{(2)}$, il ne nous reste plus qu'à montrer que les matrices de Hankel

$$\mathcal{H}_n^{(0)}(\frac{1}{f^{(2)}}) = \mathcal{H}_n^{(1)}(f^{(1)}), \;\; \text{et} \;\; -\mathcal{H}_n^{(1)}(\frac{1}{f^{(2)}}) = -\mathcal{H}_n^{(2)}(f^{(1)})$$

sont définies positives pour tout $n \geqslant 0$.

A l'aide de la formule d'identité du bigradient d'Hadamard [6, Theorem 2.4.1], on obtient pour touts $m, n \geqslant 0$ les égalités de déterminant

$$\det\Big(\mathcal{H}_n^{(m)}(f^{(1)})\Big) = (-1)^{(n+1)+(m-1)(m-2)/2} f^{(1)}(z_1)^{m+2n+1} \det\Big(\mathcal{H}_{m+n-1}^{(2-m)}(\frac{1}{f^{(1)}})\Big),$$

et alors

$$\det\Big(\mathcal{H}_n^{(0)}(\frac{1}{f^{(2)}})\Big) = \det\Big(\mathcal{H}_n^{(1)}(f^{(1)})\Big) = (-1)^{n+1} f^{(1)}(z_1)^{2n+2} \det\Big(\mathcal{H}_n^{(1)}(\frac{1}{f^{(1)}})\Big) > 0 \tag{4.38}$$

$$(-1)^{n+1} \det\Big(\mathcal{H}_n^{(2)}(\frac{1}{f^{(1)}})\Big) = (-1)^{n+1} \det\Big(\mathcal{H}_n^{(2)}(f^{(1)})\Big) = f^{(1)}(z_1)^{2n+3} \det\Big(\mathcal{H}_{n+1}^{(0)}(\frac{1}{f^{(1)}})\Big) > 0 \tag{4.39}$$

ce qui nous permet de conclure avec le résultat du théorème. □

4.3 Applications aux matrices de Toeplitz et expériences numériques

Considérons à présent une matrice $A \in \mathbb{R}^{n\times n}$ Toeplitz symétrique définie positive et $f^{[\mu]} : z \mapsto \int_\alpha^\beta \frac{d\mu(x)}{z-x}$ fonction de Markov avec $-\infty \leqslant \alpha < \beta < +\infty$ et μ une mesure positive à support dans $[\alpha;\beta]$. Nous venons de voir comment approcher la fonction $f^{[\mu]}$ à l'aide d'interpolants rationnels d'ordre $[m-1|m]$. En combinant cette information avec la théorie des ensembles K-spectraux, l'erreur absolue et relative d'approximation de la fonction de matrices $f^{[\mu]}(A)$ par cet interpolant appliqué à la matrice A sera donc bornée par un multiple l'erreur absolue ou relative respectivement obtenue sur l'intervalle contenant le spectre de la matrice A, soit

$$\|f^{[\mu]}(A) - r_m(f^{[\mu]})(A)\| \leqslant K\|f^{[\mu]} - r_m(f^{[\mu]})\|_{\infty,\mathbb{E}} \tag{4.40}$$

ou

$$\|I - r_m(f^{[\mu]})(A)f^{[\mu]}(A)^{-1}\| \leqslant K\Big\|\frac{f^{[\mu]} - r_m(f^{[\mu]})}{f^{[\mu]}}\|_{\infty,\mathbb{E}} \tag{4.41}$$

pour un ensemble $\mathbb{E}$ contenant le spectre de la matrice A. Comme la matrice considérée A est symétrique, elle est donc normale et alors d'après l'exemple 1.3.4, $\|f^{[\mu]}(A) - r_m(f^{[\mu]})(A)f^{[\mu]}(A)^{-1}\| \leqslant \|f^{[\mu]} - r_m(f^{[\mu]})\|_{\infty,\sigma(A)} \leqslant \|f^{[\mu]} - r_m(f^{[\mu]})\|_{\infty,\mathbb{E}}$ et $\|I - r_m(f^{[\mu]})(A)\| \leqslant \|f^{[\mu]} - r_m(f^{[\mu]})\|_{\infty,\sigma(A)} \leqslant \|\frac{f^{[\mu]} - r_m(f^{[\mu]})}{f^{[\mu]}}\|_{\infty,\mathbb{E}}$ d'où $K = 1$ dans (4.40). A l'aide des différentes formes de nos interpolants énoncées en sous-section 4.2, l'implémentation de la fonction de matrice $f^{[\mu]}(A)$ va s'avérer moins coûteuse que le calcul direct de la fonction de matrice $f(A)$ en arithmétique Toeplitz-like lorsque A est une matrice Toeplitz-like et m petit. Dans cette section, nous donnons dans un premier temps des résultats sur l'implémentation des $r_m(A)$ pour $m \geqslant 1$, en particulier concernant le rang de déplacement de ceux-ci. Puis nous illustrons nos résultats par plusieurs expériences numériques en implémentant nos interpolants rationnels sous leurs différentes formes en différentes matrices de Toeplitz réelles symétrique définies positives et en comparant les erreurs relatives obtenues pour différents indices m.

4.3.1 Interpolants rationnels en arithmétique Toeplitz-like

Comme nous l'avons cité plus tôt, si $f(A)$ n'est pas nécessairement de faible rang de déplacement, en revanche $r_m(f^{[\mu]})(A)$ un approximant rationnel d'ordre $[m-1|m]$ appliqué à la matrice A est de rang de déplacement au plus $\mathcal{O}(m(\varrho+1))$ d'après le corollaire 2.3.10 lorsque ϱ est le rang de déplacement de A. Nous prenons à chaque fois un interpolant rationnel aux points d'interpolation donnés par (4.16), ce qui nous permet alors d'atteindre une précision $\delta > 0$ pour $m = \mathcal{O}(\log(1/\delta))$, avec une constante cachée dépendant uniquement du bi-rapport de α, β, c et d. En effet, dans nos simulations numériques ci-dessous, le rang de déplacement de $r_m(A)$ (après compression) grandit au plus linéairement avec m, et parfois même moins si la précision augmente.

Théorème 4.3.1. *Soit $f^{[\mu]} : z \mapsto \int_\alpha^\beta \frac{d\mu(x)}{z-x}$ une fonction de Markov avec $supp(\mu) \subset [\alpha;\beta]$, $\delta > 0$, $m \geqslant 1$, $A \in \mathbb{R}^{n\times n}$ matrice Toeplitz-like symétrique avec rang de déplacement ρ et spectre inclus dans l'intervalle réel $[c;d]$ avec $c > \beta$. Notons également $r_m(f^{[\mu]})$ le Padé multipoint de $f^{[\mu]}$ de type $[m-1|m]$ (en arithmétique exacte) aux points d'interpolation* (4.16) *dépendant uniquement de m, α, β, c, d. Alors pour $m = \mathcal{O}(\log(1/\delta))$, $r_m(f^{[\mu]})(A)$ est de rang de déplacement $\mathcal{O}(m\rho)$ et est un approximant de $f^{[\mu]}(A)$ de précision (relative) $\mathcal{O}(\delta)$. De plus, le calcul numérique des générateurs de $r_m(f^{[\mu]})(A)$ avec $r_m(f^{[\mu]})$ donné par les techniques des sections 4.2.1, 4.2.2 et 4.2.3 à l'aide de l'algèbre des matrices Toeplitz-like possède une complexité $\mathcal{O}(m\rho^3 n\log^2(n))$ pour les 2 premières approches, et $\mathcal{O}(m^2\rho^3 n\log^2(n))$ pour la fraction continue de Thiele.*

Démonstration. Nous démontrons la deuxième partie du théorème, où nous omettons le coût de calcul des pôles/résidus ou autres paramètres qui sont de complexité $\mathcal{O}(m^3)$, ce qui est négligeable devant la dimension

n. La décomposition en éléments simples (4.23) dans la partie 4.2.1 semble être la manière la plus facile d'implémenter les générateurs de $r_m(f^{[\mu]})(A)$: en effet, il nous suffit de calculer les générateurs de chaque résolvante $(A - x_j I)^{-1}$ (de rang de déplacement borné par $\rho + 1$ d'après la proposition 2.3.4), combiner et compresser. Ici le travail principal consiste à calculer m fois les générateurs d'une résolvante, puis de résoudre au plus $2m(\varrho + 1)$ systèmes avec matrices de Toeplitz, ce qui nous amène à la complexité énoncée.

La représentation barycentrique de la section 4.2.2 nécessite de calculer séparément les générateurs de

$$P(A) = \sum_{j=0}^{m} f(t_j)\beta_j (A - t_j I)^{-1}, \;\; Q(A) = \sum_{j=0}^{m} \beta_j (A - t_j I)^{-1}$$

de rang de déplacement au plus $(m + 1)(\varrho + 1)$, puis ceux de $Q(A)^{-1}$ et finalement ceux de $P(A)Q(A)^{-1}$ avec un coût de l'ordre de 4 fois celui du cas discuté avant.

Enfin, pour assurer la stabilité de la representation (4.25) en section 4.2.3, on utilise une évaluation inverse de $R_{2m}^{(1)}(A)$ via (4.28), ce qui nous amène à $R_{2m}^{(2m)}(A) = f_{2m}^{(2m)} I$, et $R_{2m}^{(j)}(A) = f_j^{(j)} I + (A - z_j I) R_{2m}^{(j+1)}(A)^{-1}$ pour $j = 2m - 1, 2m - 2, \ldots, 1$. Par conséquent, le coût principal provient du calcul des générateurs de l'inverse de $R_{2m}^{(j+1)}(A)$, de rang de déplacement au plus $(\rho + 1)(2m + 2 - j)/2$, rang obtenue à l'aide de la proposition 2.3.9 en démontrant par récurrence que le numérateur et le dénominateur de $R_{2m}^{(j)}$ sont de degré $\leqslant (2m - j + 1)/2$. □

4.3.2 Expériences numériques

Nous notons à partir de maintenant les Padé multipoints d'une fonction de Markov $f^{[\mu]}$ sous la forme $r_m^{[\mu]}$ au lieu de $r_m(f^{[\mu]})$ afin de faciliter nos futures notations.
Lors de notre étude, pour éviter le calcul explicite de la fonction de matrice $f^{[\mu]}(A)$, nous proposons ici deux types de bornes supérieures dans le cas matriciel. En observant que $(f^{[\nu]}(A))^{-2}$ est plus facilement implémentable que $(f^{[\mu]}(A))^{-1}$ avec $f^{[\nu]}$ défini au théorème 4.1.4, nous présentons une première borne issue de la proposition 4.1.4 sous forme de résidu pour la fonction racine carrée inverse. Dans un deuxième temps, par la proposition 4.1.4, l'erreur relative ne change pas beaucoup si l'on remplace $f^{[\mu]}$ par $r^{[\mu]}_{m+m'}$ pour une petite valeur de m'.

Corollaire 4.3.2. *Sous les mêmes conditions que la proposition 4.1.4, soit $A \in \mathbb{C}^{n\times n}$ une matrice symétrique telle que $\mathbb{E} = \sigma(A) \subset \mathbb{C}\backslash[\alpha;\beta]$. Alors pour tout $m \geqslant 1$, nous avons la borne résiduelle*

$$\|I - r_m^{[\mu]}(A)\Big(f^{[\mu]}(A)\Big)^{-1}\| \leqslant \|I - r_m^{[\nu]}(A)^2 \frac{1}{|\alpha|}(A - \alpha I)(A - \beta I)\|. \tag{4.42}$$

où $\frac{d\nu}{dx}(x) = \frac{\sqrt{|\alpha|}}{\pi\sqrt{(z-\alpha)(\beta-z)}}$ et $r_m^{[\nu]}$ interpolant d'ordre $[m-1|m]$ de la fonction de Markov $f^{[\nu]}$ au théorème 4.1.4. Si de plus $\eta_{2m} \leqslant (\sqrt{2}-1)^2$ et

$$\delta := \frac{4\tilde{\eta}}{(1-\tilde{\eta})^2} \in (0,1), \quad \tilde{\eta} := \max_{z\in\mathbb{E}} \Big| \prod_{j=2m+1}^{2m+2m'} \frac{\varphi(z) - \varphi(z_j)}{1 - \varphi(z)\varphi(z_j)} \Big|,$$

on obtient la borne a posteriori

$$\|I - r_m^{[\mu]}(A)\Big(f^{[\mu]}(A)\Big)^{-1}\| \leqslant \frac{1+\delta}{1-\delta} \|I - r_m^{[\mu]}(A)\Big(r^{[\mu]}_{m+m'}(A)\Big)^{-1}\|$$

Démonstration. D'après la proposition 4.1.4,

$$\Big\|\frac{f^{[\mu]}-r_m^{[\mu]}}{f^{[\mu]}}\Big\|_{L^\infty(\mathbb{E})}\leqslant\Big\|1-\Big(\frac{r_m^{[\nu]}}{f^{[\nu]}}\Big)^2\Big\|_{L^\infty(\mathbb{E})}=\Big\|1-(r_m^{[\nu]})^2(f^{[\nu]})^{-2}\Big\|_{L^\infty(\mathbb{E})}$$

où $\frac{d\nu}{dx}(x)=\frac{\sqrt{|\alpha|}}{\pi\sqrt{(z-\alpha)(\beta-z)}}$. Pour cette mesure, $f^{[\nu]}(A)=\sqrt{|\alpha|}(A-\alpha I)^{-1/2}(A-\beta I)^{-1/2}$, d'où $(f^{[\nu]}(A))^{-2}=\frac{1}{|\alpha|}(A-\alpha I)(A-\beta I)$ et on obtient la première inégalité puisque pour $A\in\mathbb{C}^{n\times n}$ symétrique et $\mathbb{E}=\sigma(A)$, $\|I-r_m^{[\mu]}(A)\Big(f^{[\mu]}(A)\Big)^{-1}\|_2=\rho(I-r_m^{[\mu]}(A)\Big(f^{[\mu]}(A)\Big)^{-1})=\|1-\frac{r_m^{[\mu]}}{f^{[\mu]}}\|_{L^\infty(\mathbb{E})}$, de même pour $\|I-r_m^{[\nu]}(A)^2\frac{1}{|\alpha|}(A-\alpha I)(A-\beta I)\|$.
Pour la deuxième inéquation, on reprend l'argument (4.7) en remplaçant m par $m+m^{'}$: considérons l'ensemble des polynômes de la forme $Q=PQ_m$ avec Q_m polynôme issu de (4.7) et $P\in\mathcal{P}_{m'}$, et on obtient

$$\Big|\frac{f^{[\mu]}(z)-r^{[\mu]}_{m+m'}(z)}{f^{[\mu]}(z)-r_m^{[\mu]}(z)}\Big|\leqslant\min_{\deg P\leqslant m}\frac{|\tilde\omega(z)|}{P^2(z)}\Big\|\frac{P^2}{\tilde\omega}\Big\|_{L^\infty([\alpha;\beta])},\quad\tilde\omega(z)=\prod_{j=2m+1}^{2m+2m^{'}}(z-z_j).$$

En reprenant la preuve de la proposition 4.1.4, on peut conclure que pour $z\in\mathbb{E}$,

$$\Big|\frac{1-r^{[\mu]}_{m+m'}(z)/f^{[\mu]}(z)}{1-r_m^{[\mu]}(z)/f^{[\mu]}(z)}\Big|\leqslant\frac{4\tilde\eta}{(1-\tilde\eta)^2},\quad|1-r_m^{[\mu]}(z)/f^{[\mu]}(z)|\leqslant\frac{4\eta_{2m}}{(1-\eta_{2m})^2}\leqslant1\tag{4.43}$$

avec $\tilde\eta$ défini précédemment. On pose alors $\delta=\frac{4\tilde\eta}{(1-\tilde\eta)^2}\in(0,1)$. Par conséquent

$$\begin{aligned}\Big|\frac{1-r_m^{[\mu]}(z)/f^{[\mu]}(z)}{1-r_m^{[\mu]}(z)/r^{[\mu]}_{m+m'}(z)}\Big|&=\Big|\frac{r^{[\mu]}_{m+m'}(z)}{f^{[\mu]}(z)}\Big|\Big|\frac{1-r_m^{[\mu]}(z)/f^{[\mu]}(z)}{r^{[\mu]}_{m+m'}(z)/f^{[\mu]}(z)-r_m^{[\mu]}(z)/f^{[\mu]}(z)}\Big|\\&\leqslant\Big|\frac{r^{[\mu]}_{m+m'}(z)}{f^{[\mu]}(z)}\Big|\frac{|1-r_m^{[\mu]}(z)/f^{[\mu]}(z)|}{|1-r_m^{[\mu]}(z)/f^{[\mu]}(z)|-|1-r^{[\mu]}_{m+m'}(z)/f^{[\mu]}(z)|}\\&\leqslant\Big|\frac{r^{[\mu]}_{m+m'}(z)}{f^{[\mu]}(z)}\Big|\frac{1}{1-\Big|\frac{1-r^{[\mu]}_{m+m'}(z)/f^{[\mu]}(z)|}{1-r_m^{[\mu]}(z)/f^{[\mu]}(z)}\Big|}\\&\leqslant\frac{1}{1-\delta}\Big|\frac{r^{[\mu]}_{m+m'}(z)}{f^{[\mu]}(z)}\Big|\leqslant\frac{1}{1-\delta}\Big(1+|1-\frac{r^{[\mu]}_{m+m'}(z)}{f^{[\mu]}(z)}|\Big)\\\frac{1}{1-\delta}&\leqslant\Big(1+\delta|1-\frac{r_m^{[\mu]}(z)}{f^{[\mu]}(z)}|\Big)\leqslant\frac{1+\delta}{1-\delta},\end{aligned}$$

et le résultat énoncé sur les matrices est démontré en observant comme pour la première inéquation que pour la matrice symétrique A, $\|I-r_m^{[\mu]}(A)\Big(f^{[\mu]}(A)\Big)^{-1}\|=\|1-r_m^{[\mu]}\Big(f^{[\mu]}\Big)^{-1}\|_{L^\infty(\mathbb{E})}$ et $\|I-r_m^{[\mu]}(A)\Big(r^{[\mu]}_{m+m'}(A)\Big)^{-1}\|=\|1-r_m^{[\mu]}\Big(r^{[\mu]}_{m+m'}\Big)^{-1}\|_{L^\infty(\mathbb{E})}$. □

Remarque 4.3.3. *A l'aide de ce résultat, nous pouvons à présent améliorer la borne supérieure énoncée au corollaire 4.3.2 : en effet, pour toute matrice $A\in\mathbb{C}^{n\times n}$ symétrique telle que $\mathbb{E}=\sigma(A)\subset[c;d]\subseteq\mathbb{C}\backslash[\alpha;\beta]$, en combinant la proposition 4.1.4 avec la borne supérieure* (4.18)*, l'erreur relative pour l'approximant rationnel $r_m^{[\mu]}$, elle-même bornée par* (4.17)

$$\|I-r_m^{[\mu]}(A)\Big(f^{[\mu]}(A)\Big)^{-1}\|\leqslant\|I-r_m^{[\nu]}(A)^2\frac{1}{|\alpha|}(A-\alpha I)(A-\beta I)\|\leqslant8\varrho^{2m}/(1-2\varrho^{2m})^2,\tag{4.44}$$

avec $\varrho = \exp\left(\frac{-1}{cap([\alpha;\beta],[c;d])}\right)$ *pour tout* $m \geqslant 1$. *Nous allons voir plus tard dans ce chapitre que pour une matrice* A *symétrique avec spectre dans un intervalle* $[c;d] \subset \mathbb{R}$, *les expériences numériques font état de perte de cette inéquation à cause de la précision finie lors de l'évaluation sur ordinateur de nos Padé multipoints appliqués à cette matrice. Plus précisément, l'erreur relative pour* $r_m^{[\mu]}$ *se comporte de manière erratique dès que celle-ci vient dépasser notre borne supérieure a priori. Dans le but de ne considérer que des ordres pour lesquels l'erreur relative reste en dessous de cette borne, nous définissons une méthode de recherche d'un ordre* m *pour lequel l'erreur est minimisée tout en restant inférieure à la borne énoncée. Pour ce faire, nous suggérons de calculer* $r_m^{[\mu]}(A)$ *et* $r_m^{[\nu]}(A)$ *pour* $m = 1, 2, \ldots$ *et stoppons au dernier indice après lequel l'erreur résiduelle est plus grande que 5 fois la borne supérieure estimée pour pour* $r_m^{[\nu]}$, *soit quand*

$$\|I - r_m^{[\nu]}(A)\frac{1}{|\alpha|}(A - \alpha I)(A - \beta I)r_m^{[\nu]}(A)\| \geqslant 40\varrho^{2m}/(1 - 2\varrho^{2m})^2. \tag{4.45}$$

Dans les résultats numériques qui suivent, nous indiquons tous les indices m *pour lesquels l'équation précédente est vérifiée à l'aide de marqueurs. Il est à noter que le critère d'arrêt énoncé ne nécessite pas le calcul explicite de la fonction de matrice* $f^{[\mu]}(A)$ *et semble très bien fonctionner en pratique.*

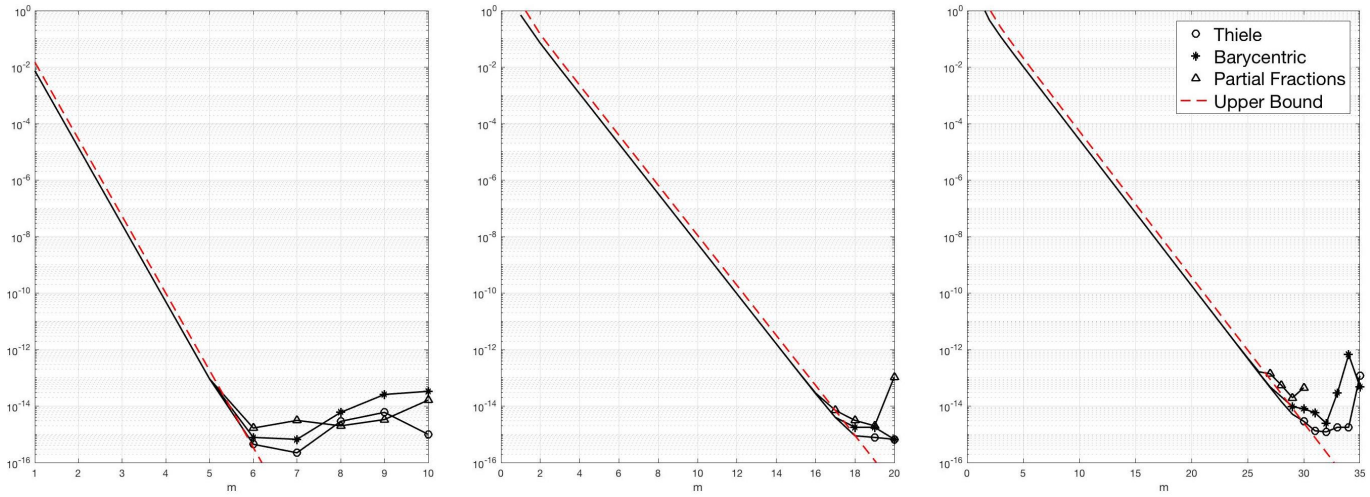

FIGURE 4.1 – Erreur relative L^∞ sur l'intervalle $[c;d]$ des interpolants rationnels de type $[m-1|m]$ de la fonction de Markov $f^{[\mu]}(z) = \frac{1}{\sqrt{z}}$ avec $\alpha = -\infty$, $\beta = 0$, $d = 1$ et $c \in \{1/2, 10^{-3}, 10^{-6}\}$ (de la gauche vers la droite). Pour chaque c et m, nous prenons les points d'interpolation quasi-optimaux (4.16) dépendant de m et $\alpha; \beta, c, d$, et affichons l'erreur relative de ces mêmes interpolants (en ligne pleine noire) composés avec les 3 différentes méthodes d'implémentation : à l'aide d'une représentation en éléments simples de la sous-section 4.2.1 (marqueurs triangulaires), d'une représentation barycentrique (marqueurs étoilés) et d'une représentation en fraction continue (marqueurs circulaires). La 4ème donnée en une ligne pointillée rouge nous donne la borne supérieure a priori (4.17).

Exemple 4.3.4. *Dans la figure 4.1, nous avons représenté l'erreur relative* $L^\infty([c;d])$ *des mêmes interpolants rationnels* $r_m^{[\mu]}$ *sous différentes formes (éléments simples, barycentrique, fraction continue de Thiele) pour la même fonction de Markov* $f^{[\mu]}(z) = 1/\sqrt{z}$ *aux points d'interpolation* (4.16) *en fonction de* m. *Nous avons ici discrétisé l'intervalle* $[c;d]$ *en 500 points en cosinus, entrées d'une matrice diagonale* A. *En théorie, toutes les courbes devraient avoir un comportement identique et rester en dessous de la borne supérieure a priori* (4.17). *Cependant, en précision arithmétique machine, on observe qu'une fois que la méthode d'implémentation et* c *sont fixés, la courbe d'erreur traverse une première fois la borne supérieure à priori puis décroît rarement, augmentant même parfois. Nous utilisons des marqueurs sur les différentes courbes pour les indices* m *satisfaisant la condition de la remarque 4.3.3. Si on note* m' *l'indice tel que* $m' + 1$ *est le premier indice satisfaisant cette condition, l'erreur pour* m' *est en-dessous de la borne à priori et la passage au-dessus de*

cette borne s'effectue entre les indices m' et $m'+1$. De plus, on peut observer lors de différentes expériences numériques que l'erreur pour $m > m'$ n'est jamais plus petite que $1/10$ fois que l'erreur pour m'. Ceci confirme alors que notre critère d'arrêt fonctionne bien en pratique. On peut remarquer sur nos graphiques que pour les 3 méthodes, l'erreur relative finale est du même ordre de grandeur, non loin de la précision machine, et croît modestement avec d/c. Nous allons voir plus loin que cette observation n'est plus valable si on évalue nos interpolants sur des matrices générales au lieu de matrices diagonales.

Nous considérons dans les expériences suivantes des matrices de Toeplitz réelles symétriques définies positives avec spectre dans un intervalle donné $[c;d]$, où les différentes opérations se simplifient et on a les estimations d'erreur

$$\|f^{[\mu]}(A) - r_m^{[\mu]}(A)\| \leqslant \|f^{[\mu]} - r_m^{[\mu]}\|_{L^\infty([c;d])}, \quad \|I - r_m^{[\mu]}(A) f^{[\mu]}(A)^{-1}\| \leqslant \Big\| \frac{f^{[\mu]} - r_m^{[\mu]}}{f^{[\mu]}} \Big\|_{L^\infty([c;d])}.$$

Nous présentons à présent des simulations numériques sur matrices de Toeplitz symétriques de taille $n \times n$ pour $n = 500, 2000$ avec spectre dans un intervalle réel $[c;d]$ tel que $c > \beta$. Afin de mettre en lumière toute notre théorie, nous observons dans chaque figure correspondante des valeurs propres $\lambda_{\min}$ et $\lambda_{\max}$ fixées et une fonction de Markov fixée. Dans les graphiques ci-dessous, on observe de gauche à droite les cas

i. arithmétique Toeplitz-like avec $[c;d] = [\lambda_{\min}; \lambda_{\max}]$;
ii. arithmétique Toeplitz-like avec $[c;d] = [\frac{1}{2}\lambda_{\min}; 2\lambda_{\max}]$;
iii. sans arithmétique Toeplitz-like, $[c;d] = [\lambda_{\min}; \lambda_{\max}]$;
iv. $[c;d] = [\lambda_{\min}; \lambda_{\max}]$, matrice diagonal d'ordre $n \times n$ avec éléments diagonaux en cosinus dans $[c;d]$ sans arithmétique Toeplitz-like ;

A l'aide de cette configuration, nous allons pouvoir mesurer l'impact du choix de l'intervalle sur lequel on approche par rapport à l'intervalle spectral de notre matrice en comparant i. et ii.. Egalement nous pourrons comparer l'arithmétique Toeplitz-like par rapport à une arithmétique en matrice pleine à l'aide des affichages i. et iii. . Enfin nous verrons l'impact de la non-commutativité de nos matrices, dû à la précision en arithmétique finie, à l'aide du cas iv. que nous pourrons comparer aux 3 autres. Dans chaque cas, nous étudions les 3 formes possibles (4.23), (4.24) et (4.25) de nos Padé multipoints $r_m^{[\mu]}$ et affichons l'erreur relative $\|I - r_m^{[\mu]}(A) f^{[\mu]}(A)^{-1}\|$ dans chaque cas à l'aide d'une courbe noire. Nous signalons les différentes représentations à l'aide de différents marqueurs : des marqueurs triangulaires pour la représentation en éléments simples, étoilés pour la représentation barycentrique et circulaires pour la forme fraction continue de Thiele. Comme pour la figure 4.1, nous affichons les marqueurs de chaque représentation pour tout indice $m > m'$ tel que m' est le premier indice pour lequel l'équation (4.45) de la remarque 4.3.3 est vérifiée. Nous affichons également en rouge la borne supérieure a priori sur chaque graphique. Notons que le calcul de l'erreur relative nous oblige à calculer $f^{[\mu]}(A)^{-1}$ à l'aide des commandes intégrées de Matlab comme **logm** ou **sqrtm**, ce qui nous a amené à considérer des matrices de taille modérée.

Exemple 4.3.5. *En figure 4.2, nous présentons l'étude de l'erreur d'approximation rationnelle de la fonction de matrice* $\log(A)(A-I)^{-1}$ *avec* $A \in \mathbb{R}^{500\times 500}$ *matrice de Toeplitz symétrique définie positive avec valeurs propres extrémales* $\lambda_{\min} = 25,1918$ *et* $\lambda_{\max} = 129,5678$ *et conditionnement de* $5,1433$*, et où la fonction* $f : z \mapsto \log(z)/(z-1)$ *est une fonction de Markov avec* $\alpha = -\infty$ *et* $\beta = 0$*. En notant* $r_m^{[\mu]}$ *le Padé multipoint d'ordre* $[m-1|m]$ *de* $f^{[\mu]}(z) = \frac{\log(z)}{z-1}$*, on approche* $\log(A)$ *par* $(A-I)r_m^{[\mu]}(A)$*, ce qui nous amène à la même étude que l'approximation de* $f^{[\mu]}(A)$ *par* $r_m^{[\mu]}(A)$*. On observe alors*

a. *Dans les 4 cas i., ii., iii. et iv., le critère d'arrêt de la remarque 4.3.3 fonctionne parfaitement : pour tous les indices m inférieurs à m' le premier indice vérifiant (4.45), l'erreur relative est inférieure à la borne supérieure à priori, et pour les indices supérieurs à m', l'erreur relative reste au dessus de la borne supérieure a priori ;*

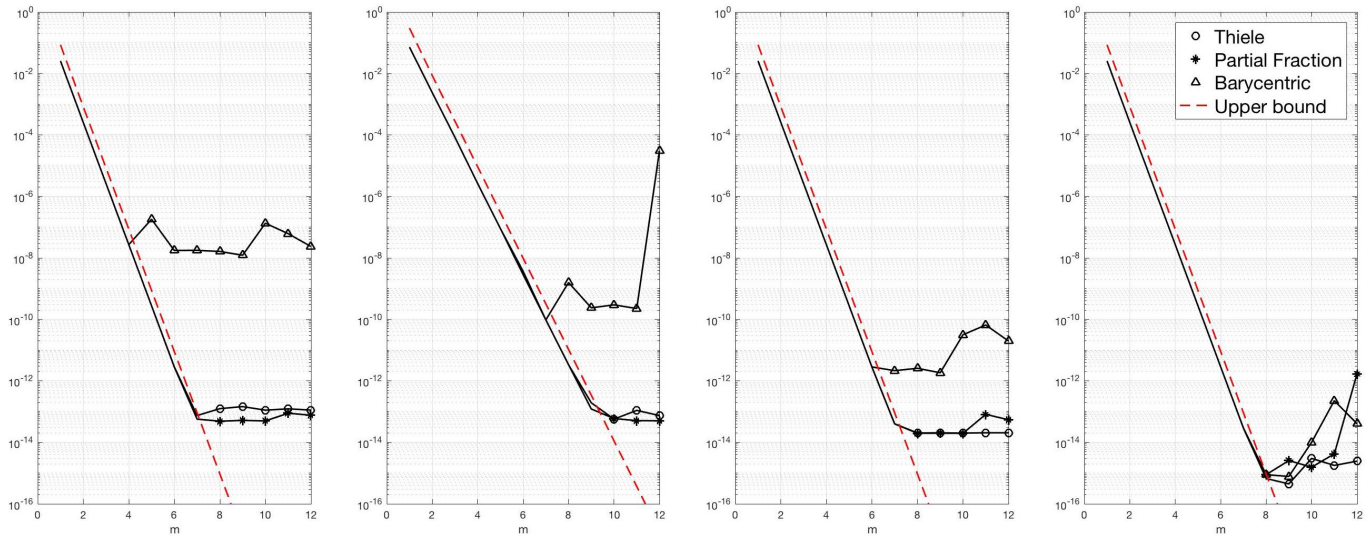

FIGURE 4.2 – Erreurs relatives d'approximation $\|I - r_m^{[\mu]}(A) f^{[\mu]}(A)^{-1}\|_2$ avec $f^{[\mu]}$ la fonction de Markov $f^{[\mu]}(z) = \log(z)/(z-1)$, $A \in \mathbb{R}^{500\times 500}$ matrice de Toeplitz symétrique définie positive avec valeurs propres extrémales $\lambda_{\min} = 25,1918$ et $\lambda_{\max} = 129,5678$ et un conditionnement de $5,1433$ et $r_m^{[\mu]}$ Padé multipoint aux points optimisés (4.16).

b. *Dans le cas iv. avec matrice diagonales à éléments diagonaux en cosinus, les 3 méthodes d'évaluation de $r_m(A)$ sont équivalentes comme pour la figure 4.1 ;*

c. *Dans les cas i. à iii., la représentation barycentrique semble moins bien fonctionner que les autres méthodes et traverse la borne supérieure a priori bien avant les autres méthodes d'implémentation de l'interpolant rationnel ;*

d. *Les représentations barycentrique et fraction continue de Thiele ont un comportement similaire et atteignent une erreur minimale de l'ordre de 10^{-10} ;*

e. *les modèles i. et ii. permettent de conclure que l'élargissement de l'intervalle d'approximation $[c;d]$ ne permet pas de réduire l'erreur, et nécessite un nombre plus grand d'itérations pour obtenir la même erreur que dans le cas $[c;d] = [\lambda_{\min};\lambda_{\max}]$;*

f. *Pour mesurer la complexité, on peut noter que le rang de déplacement de $r_m(A)$ dans les cas i. et ii. croît avec m, d'abord linéairement puis se stabilise autour de 22 (pour la forme Thiele et éléments simples, le double pour la forme barycentrique), une fois une bonne précision atteinte.*

Les observations a. à f. pour la matrice en figure 4.2 peuvent être également faites pour d'autres matrices de Toeplitz symétriques définies positives, du moment que leur conditionnement reste modeste, à savoir inférieur à 10. Notons que pour $\alpha = -\infty$ et $\beta = 0$, le conditionnement d'une matrice A est donné par le birapport donné par (4.14) (pour les cas i., ii. et iv.) et détermine le taux de convergence asymptotique (4.17), essentiellement la pente de notre borne supérieure a priori.
Cependant, l'observation n'est plus valable lorsque le conditionnement de A est plus grand que 10. Pour éviter le problème du conditionnement de nos matrices, nous reprenons la méthode de scaling and squaring énoncée par Higham dans [57, Chapter 11.5] et [58], méthode que l'on peut appliquer à la fonction logarithme et aux puissances fractionnées $x \mapsto x^\gamma$ pour tout $\gamma \in \mathbb{R}$ de la manière suivante :

$$\log(A) = 2^\ell \log(A^{1/2^\ell}), \quad A^\gamma = \left((A^{1/2^\ell})^\gamma\right)^{2^\ell}. \tag{4.46}$$

Il nous faut alors d'abord calculer ℓ racines carrées $A_0 = A$, et $A_j = \sqrt{A_{j-1}} = (A_{j-1})^{1/2}$ pour $j = 1, \ldots, \ell$

avec ℓ sélectionné de telle sorte que

$$\operatorname{cond}\left(A^{1/2^\ell}\right) \leqslant (d/c)^{1/2^\ell} < 10. \tag{4.47}$$

Pour calculer ces racines consécutives, nous choisissons d'utiliser la méthode itérative de Newton sous sa forme Newton DB produit vérifiant la plus grande stabilité donnée au chapitre 3, section 3.2.1. Notons pour B matrice de départ dont on souhaite calculer la racine que les matrices itérées M_k vérifient $M_k - I = X_k B^{-1} X_k - I$, ce qui nous donne le résidu de la racine carrée $\sqrt{B} = B^{1/2}$ en corollaire 4.3.2. Afin d'assurer la stabilité et la précision asymptotique démontrées pour $\mu_k = 1$ au lemme 3.2.9 [57], nous suggérons de procéder en deux phases : dans une première phase nous appliquons la méthode de Newton DB produit avec paramètres (3.12) jusqu'à obtenir $\frac{1-\mu_k^4}{\mu_k^4} \leqslant 10^{-3}$, puis pour $k \geqslant K$, nous prenons les paramètres $\mu_k = 1$, la convergence quadratique des matrices itérées de Newton-DB produit assurant dans une deuxième phase d'atteindre une haute précision après 3 itérations supplémentaires.
Une fois que l'on a calculé $A^{1/2^\ell}$, nous évaluons notre interpolant rationnel de la fonction de Markov f en la matrice $A^{1/2^\ell}$, puis nous appliquons le squaring ou renormalisation afin d'approcher $f(A)$. Dans les cas i. ii., nous effectuons le scaling en arithmétique Toeplitz-like. Notons que le coût de calcul des interpolants rationnels en une matrice est plus élevé que le coût de calcul nécessaire pour la méthode de scaling et squaring, du moins pour des conditionnements inférieurs à 10^6 où il nous faut calculer $\ell \leqslant 3$ racines carrées et au plus 8 itérations de Newton sont nécessaires pour atteindre la convergence pour la racine carrée.

Exemple 4.3.6. *En figure 4.3, nous présentons sur les graphiques de gauche les erreurs relatives* $\|I - r_m^{[\mu]}(A)/\log(A)\|_2$ *avec* $r_m^{[\mu]}$ *l'interpolant rationnel aux points optimaux* (4.16) *sous forme de fraction continue scaled avec la méthode de Newton-DB (ligne bleu) et Newton-DB produit (ligne rouge) avec* X_m *construite à l'aide de l'arithmétique Toeplitz-like (graphique supérieur) et à l'aide de l'arithmétique pleine (graphique inférieur). Nous y représentons également la borne supérieure à priori* (4.17)*. Sur la droite sont affichées les erreurs relatives* $\|I - X_m A^{-1/2}\|_2$ *avec* X_m *obtenue par la méthode de Newton DB et Newton-DB produit en arithmétique pleine. Pour une des 2 variantes de Newton, on démarre avec* X_0 *et on calcule les matrices itérées* X_m *de cette méthode pour approcher* $A^{1/2^1}$ *jusqu'à atteindre la convergence avec une matrice itérée* X_K*, puis on on redémarre une itération de Newton avec un premier terme* $Y_0 = X_K$ *pour approcher* $X_k^{1/2}$ *que l'on suppose proche de* $A^{1/2^2}$*. Comme* $\operatorname{cond}(A) = 1,314 \times 10^3$*, 2 implémentations de la méthode de Newton suffisent à obtenir une racine* 2^2*ème telle que* $\operatorname{cond}(A^{1/2^2}) < 10$ *pour ensuite appliquer nos interpolants rationnels en les points* (4.16) *sous la forme* $2^2 r_m(A^{1/2^2})$*.*

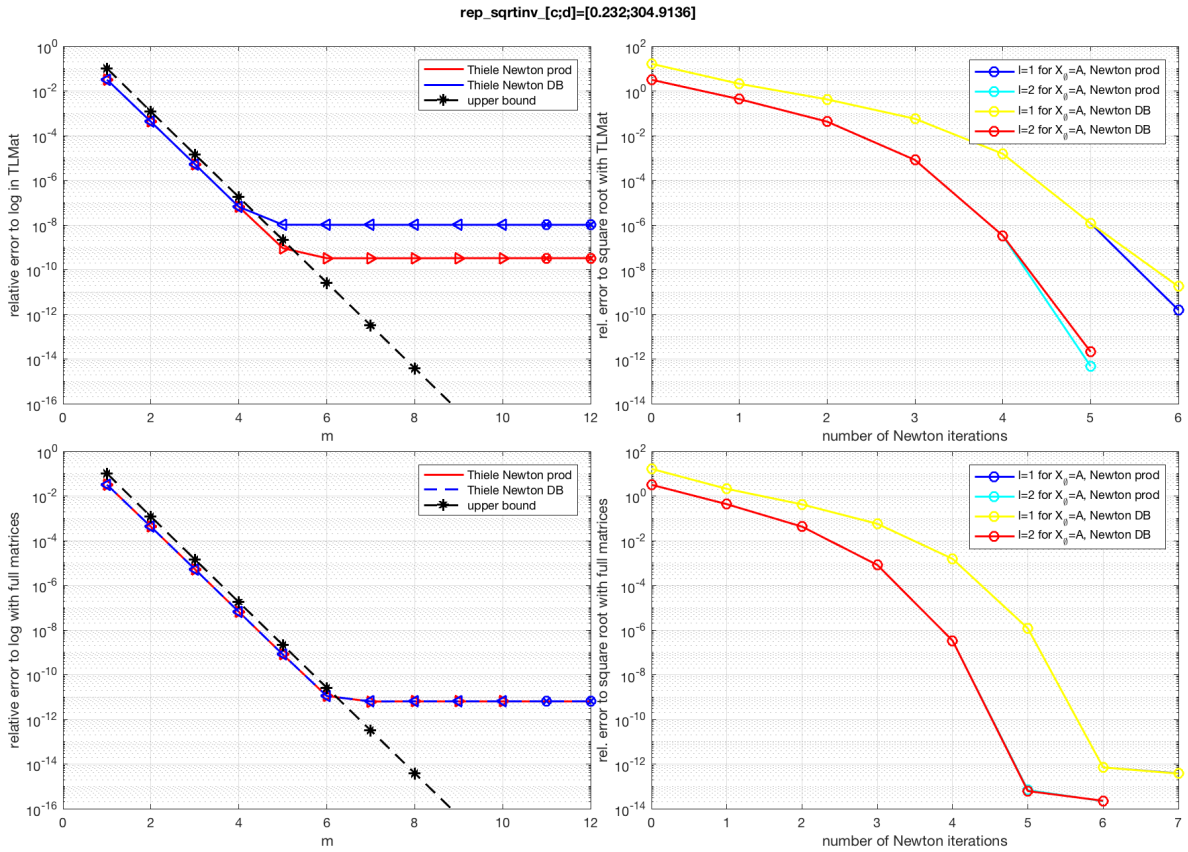

FIGURE 4.3 – Erreurs d'approximation relative $\|I - r_m^{[\mu]}(A) f^{[\mu]}(A)^{-1}\|_2$ pour la fonction de Markov $f^{[\mu]}(z) = \log(z)/(z-1)$ avec $r_m^{[\mu]}$ implémentée sous forme de fraction continue de Thiele scaled (à gauche) pour $A \in \mathbb{R}^{2000\times 2000}$ matrice de Toeplitz symétrique définie positive avec $\sigma(A) \subseteq [0, 232; 304, 92]$ et conditionnement $1,314 \times 10^3$, en l'arithmétique Toeplitz-like (en haut) et arithmétique pleine (en bas), et erreur relative pour l'approximation $A^{1/2^\ell}$ (à droite) par les méthodes de Newton-DB et Newton-DB produit avec $X_0 = A$ et μ_m optimisés jusqu'à atteindre une tolérance de 10^{-3}, puis $\mu_m = 1$ pour tout m.

On peut observer dans nos graphiques que l'erreur relative finale pour les méthodes scaled Thiele et scaled éléments simples est dominée par l'erreur relative finale pour l'approximation de la racine carrée $A_1 = A^{1/2}$. Ce comportement semble logique étant donné que cette matrice possède le pire conditionnement parmi les matrices A_j, ce qui va influencer l'erreur relative après application de nos interpolants rationnels. De plus, nous avons pu observer après de multiples simulations numériques comme présentes ci-dessus que la méthode de Newton DB produit offre des résultats similaires aux autres formulation de la méthode de

Newton.

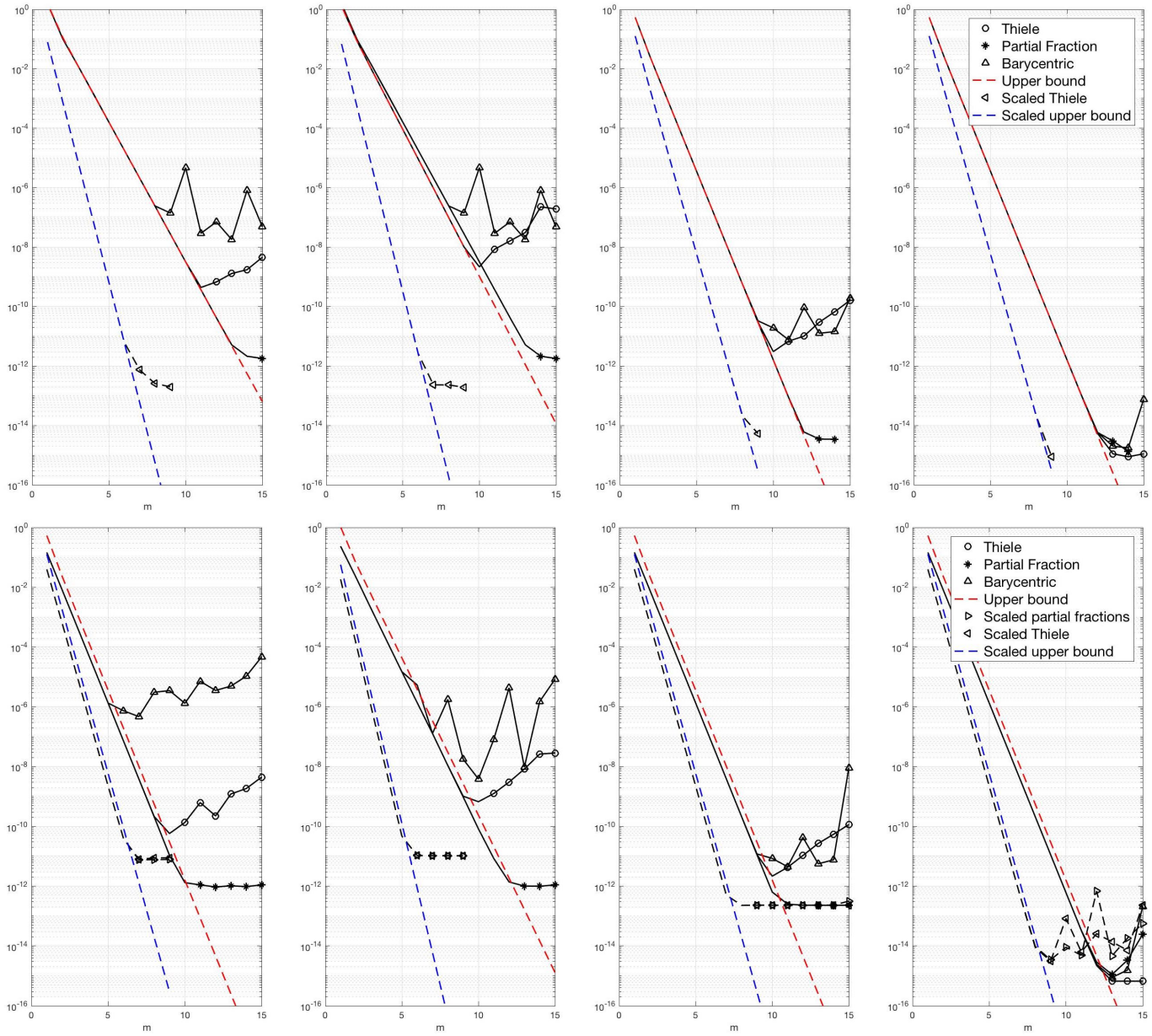

FIGURE 4.4 – Résidus $\|I - r_m^{[\nu]}(A)A^{-1}r_m^{[\nu]}(A)\|_2$ (graphiques supérieurs) et erreurs relatives $\|I - r_m^{[\mu]}(A)f^{[\mu]}(A)^{-1}\|_2$ (graphiques inférieurs) pour $A \in \mathbb{R}^{2000\times 2000}$ Toeplitz symétrique définie positive avec $\lambda_{\min} = 1,3485$, $\lambda_{\max} = 72,6864$ et conditionnement $53,9025$, $f^{[\mu]}(x) = \frac{\log(x)}{x-1}$ fonction de Markov, $r_m^{[\nu]}$ donné en proposition 4.1.4 et $r_m^{[\mu]}$ Padé multipoint de $f^{[\mu]}$ d'ordre $[m-1|m]$ aux points (4.16) sous forme de fraction continue (marqueurs circulaires), éléments simples (marqueurs étoilés) et barycentrique (marqueurs triangulaires verticaux) ainsi que Thiele scaled (marqueurs triangulaires vers la gauche) et éléments simples scaled (marqueurs triangulaires vers la droite).

Exemple 4.3.7. *En figure 4.4 nous affichons pour une matrice $A \in \mathbb{R}^{2000\times 2000}$ Toeplitz symétrique définie positive avec $\lambda_{\min} = 1,3485$ et $\lambda_{\max} = 72,6864$ et conditionnement $53,9025$ et la fonction $f^{[\mu]}(z) = \frac{\log(z)}{z-1}$, les résidus $\|I - r_m^{[\nu]}(A)Ar_m^{[\nu]}(A)\|_2$ (graphes supérieurs) pour les différents cas i. à iv. (de gauche à droite) et les erreurs relatives $\|I - r_m^{[\mu]}(A)ff^{[\mu]}(A)^{-1}\|_2$ (graphes inférieurs) pour ces mêmes cas à l'aide des dif-*

férentes formes d'implémentation de nos interpolants rationnels aux points optimisés (4.16). *En plus de la borne supérieure a priori, nous avons ajouté une borne supérieure associée à la matrice* $A^{1/2^\ell}$ *(en bleu) ainsi que les erreur d'approximation rationnelle (en noire) obtenue en évaluant nos interpolants rationnels de* $f^{[\mu]}(x) = \log(x)/(x-1)$ *en* $A^{1/2^\ell}$ *via une décomposition en éléments simples (marqueurs triangulaires orientés vers la droite) ou via une fraction continue de Thiele (marqueurs triangulaires orientés vers la gauche). Ces 2 approches ont un comportement similaire d'après* (4.47). *Pour les résidus, on note par des marqueurs (triangulaires, circulaires, étoilés) les indices* m *vérifiant* (4.45) *et on peut observer que ce critère fonctionne très bien et nous permet de bien signaler quand le critère* (4.45) *est vérifié. Une fois le premier marqueur (triangulaire, circulaire ou étoilé) trouvé, on peut voir sur les graphes inférieurs que celui-ci désigne bien le premier indice* m *à partir duquel nos Padé multipoints* $r_m^{[\mu]}(A)$ *dépassent la borne supérieure a priori* (4.17). *Pour le choix de la représentation de nos interpolants rationnels, on observe que les représentations barycentrique fonctionnent très peu et qu'on ne peut atteindre qu'une tolérance de l'ordre de* 10^{-7} *pour l'erreur relative dans les 2 premiers cas i. et ii.. La représentation en fraction continue de Thiele permet elle d'atteindre une tolérance de l'ordre de* 10^{-10} *et* 10^{-9} *dans ces 2 mêmes cas. Enfin, la représentation en éléments simples ainsi que la fraction continue de Thiele scaled permettent d'atteindre une meilleure tolérance d'ordre* 10^{-12} *et* 10^{-11} *respectivement. Dans les cas iii. et iv. respectivement sur les mêmes matrices en arithmétique pleine et sur des matrices diagonales avec éléments diagonaux en cosinus avec même intervalle spectral que le cas i., toutes ces représentations se comportent bien, permettant d'atteindre une tolérance de* 10^{-11} *pour les représentations barycentrique et fraction continue de Thiele pour l'implémentation en arithmétique pleine,* 10^{-13} *pour les autres représentations, et une tolérance de* 10^{-14} *pour une matrice diagonale.*
De cette première observation, on voit déjà qu'un choix de représentation sous forme d'éléments simples semble être privilégié, ainsi qu'une représentation en fraction continue de Thiele dans le cas d'une matrice bien conditionnée en arithmétique Toeplitz-like.

Exemple 4.3.8. *En figure 4.5, nous affichons pour une matrice* $A \in \mathbb{R}^{2000\times 2000}$ *Toeplitz symétrique définie positive avec* $\lambda_{\min} = 2,6246$ *et* $\lambda_{\max} = 1224,5$ *et conditionnement* $466,5583$ *et la fonction* $f^{[\mu]}(z) = \frac{\log(z)}{z-1}$ *, les résidus* $\|I - r_m^{[\mu]}(A)Ar_m^{[\mu]}(A)\|_2$ *(graphes supérieurs) pour les différents cas i. à iv. (de gauche à droite) et les erreurs relatives* $\|I - r_m^{[\mu]}(A)f^{[\mu]}(A)^{-1}\|_2$ *(graphes inférieurs) pour ces mêmes cas à l'aide des différentes formes d'implémentation de nos interpolants rationnels aux points optimisés* (4.16). *Comme précédemment nous indiquons par des marqueurs triangulaires, circulaires et étoilés tous les indices supérieurs au premier indice vérifiant* (4.45) *pour toutes nos représentations. Ici pour cette matrice* A*, les représentations barycentrique et fraction continue de Thiele de nos interpolants aux points optimisés* (4.16) *ne sont pas fiables puisque dans les cas i. et ii. celles-ci ne permettent pas d'aller au-delà d'une tolérance de* 10^{-6}*, alors que la représentation en éléments simples ainsi que la représentation en fraction continue de Thiele scaled permettent d'atteindre une tolérance de* 10^{-10}*. Par conséquent, on peut penser à prioriser une représentation en fraction continue de Thiele scaled lors de l'implémentation en arithmétique Toeplitz-like.*

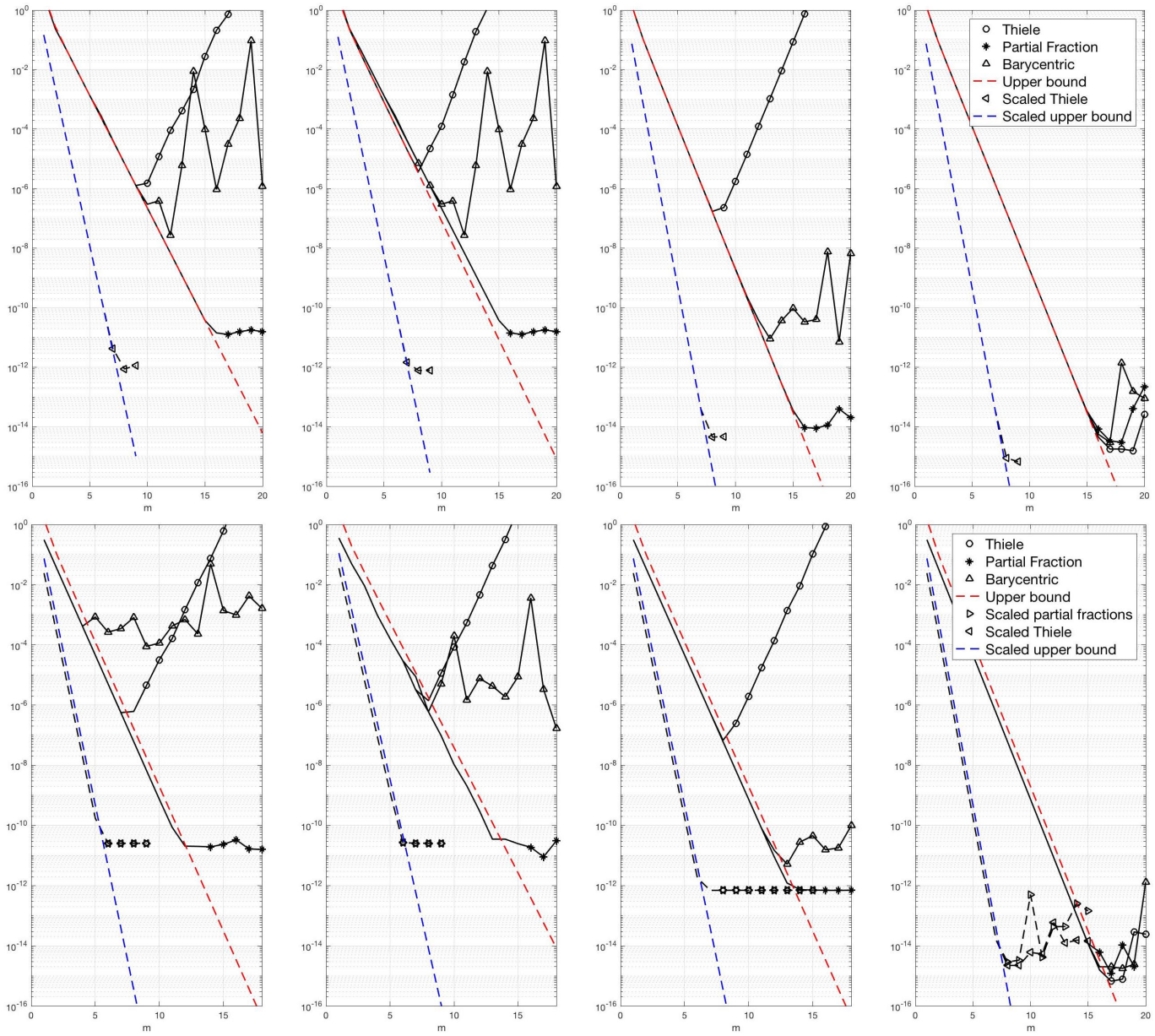

FIGURE 4.5 – Résidus $\|I - r_m^{[\nu]}(A)A^{-1}r_m^{[\nu]}(A)\|_2$ (graphiques supérieurs) et erreurs relatives $\|I - r_m^{[\mu]}(A)f^{[\mu]}(A)^{-1}\|_2$ (graphiques inférieurs) pour $A \in \mathbb{R}^{2000\times 2000}$ Toeplitz symétrique définie positive avec $\lambda_{\min} = 2,6246$, $\lambda_{\max} = 1224,5$ et conditionnement $466,5583$, $f^{[\mu]}(x) = \frac{\log(x)}{x-1}$ fonction de Markov, $r_m^{[\nu]}$ donné par la proposition 4.1.4 et $r_m^{[\mu]}$ Padé multipoints de $f^{[\mu]}$ d'ordre $[m-1|m]$ aux points (4.16) sous forme de fraction continue (marqueurs circulaires), éléments simples (marqueurs étoilés) et barycentrique (marqueurs triangulaires verticaux) ainsi que Thiele scaled (marqueurs triangulaires vers la gauche) et éléments simples scaled (marqueurs triangulaires vers la droite).

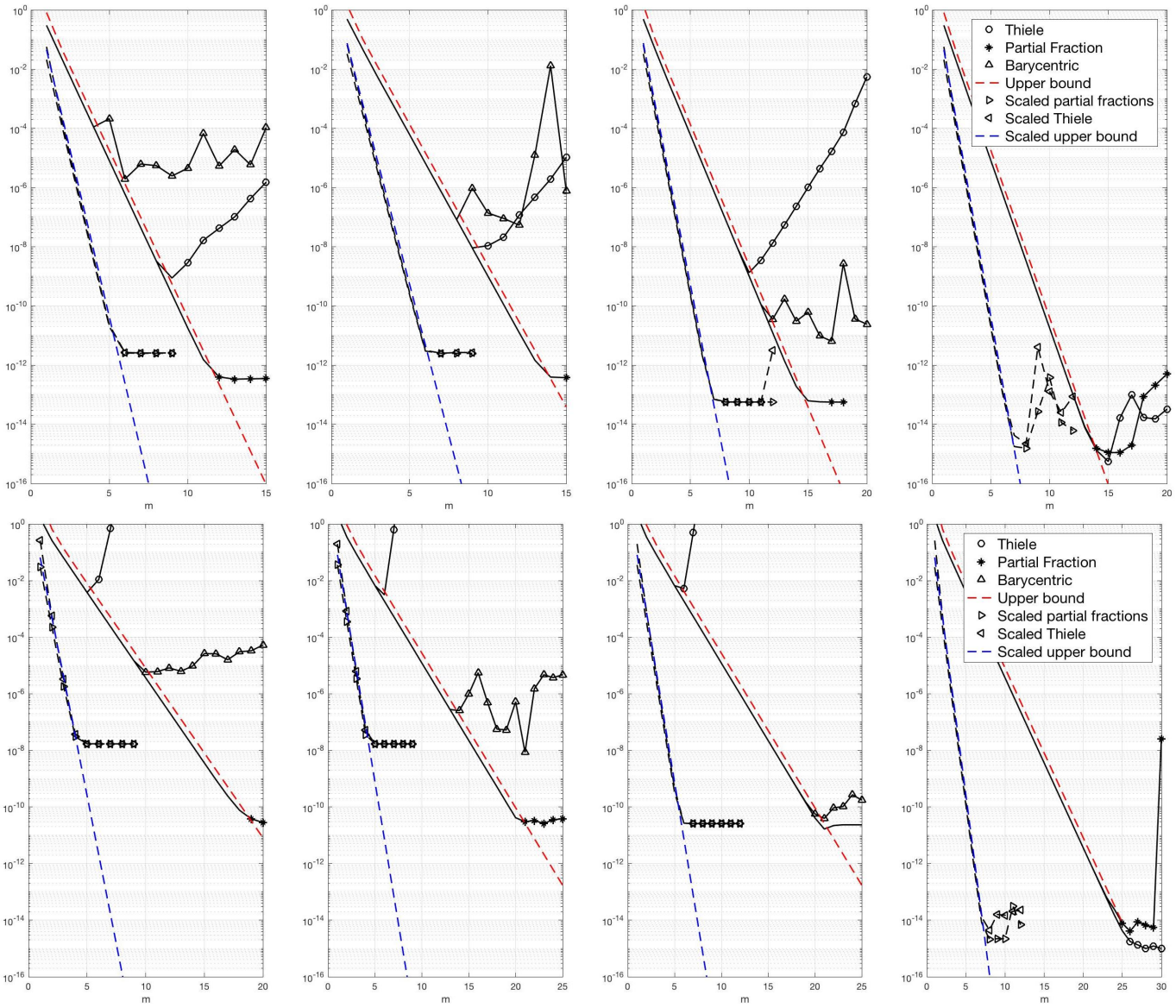

FIGURE 4.6 – Erreurs relatives $\|I - r_m^{[\mu]}(A) f^{[\mu]}(A)^{-1}\|_2$ avec $f^{[\mu]}(A) = A^{-1/3}$ et $r_m^{[\mu]}$ Padé multipoints aux points d'interpolation optimisés (4.16) pour 2 matrices de Toeplitz définies positives : en premier une matrice d'ordre 2000 avec valeurs propres extrémales $\lambda_{\min} = 2,1$ et $\lambda_{\max} = 261,419$ et conditionnement $124,4853$ (en haut) et la matrice du Laplacien en 1D d'ordre 499, i.e. la matrice avec valeur 2 sur la diagonale et valeur -1 sur la sur- et sous-diagonale, avec valeurs propres extrémales $\lambda_{\min} = 3,95 \times 10^{-5}$, $\lambda_{\max} = 4$ et conditionnement $1,01 \times 10^5$.

Exemple 4.3.9. *Pour la figure 4.6, nous étudions l'erreur d'approximation de la fonction de matrice $A \mapsto A^{-1/3}$ pour 2 matrices de Toeplitz symétriques définies positives : A_1 d'intervalle spectral $[2,1; 261,419]$ et conditionnement $124,4853$ et A_2 la matrice du Laplacien de dimension 499, d'intervalle spectral $[3,95 \times 10^{-5}; 4]$ et conditionnement $1,01 \times 10^5$. Etant donné que la fonction $x \mapsto x^\gamma$ est une fonction de Markov uniquement pour $\gamma \in [-1; 0)$, nous modifions légèrement l'approche définie dans (4.46), ce qui s'est par ailleurs vérifié après plusieurs essais numériques qui ont montré que l'approche à l'aide de (4.46) entraîne une perte de précision : pour $\gamma \in \mathbb{R}$, nous notons $2^\ell \gamma = k + \gamma'$ avec $k \in \mathbb{Z}$ et $\gamma' \in [-1; 0)$, de telle sorte que $A^\gamma = g(A^{1/2^\ell})(A^{1/2^\ell})^k$ avec $g(x) = x^{\gamma'}$ fonction de Markov soit approchée par $r_m(A^{1/2^\ell})(A^{1/2^\ell})^k$ avec r_m interpolant rationnel de g. Pour le premier cas avec une matrice de Toeplitz symétrique définie positive avec conditionnement $121,7$, $\ell = 2$ et $k = 2, \gamma' = -2/3$ alors que pour la seconde matrice, Laplacien en 1D*

de conditionnement $1,01 \times 10^5$*, on trouve* $\ell = 3$ *et* $k = 3, \gamma' = -1/3$.
On observe alors que pour la matrice A_1 *comme pour la matrice en figure 4.5 qu'en arithmétique Toeplitz-like, les représentations barycentriques et fraction continue de Thiele ne permettent pas d'atteindre une tolérance suffisante alors qu'une représentation en éléments simples ou en fraction continue de Thiele scaled permettent d'atteindre une tolérance d'ordre* 10^{-12}*. Cette observation change avec la matrice* A_2 *de conditionnement d'ordre* 10^5 *puisque seule une représentation en éléments simples permet d'atteindre une tolérance d'ordre* 10^{-10} *alors que la représentation en fraction continue de Thiele scaled ne descend pas en dessous d'une tolérance de* 10^{-8}*, ce qui est encore pire avec les représentations barycentrique et en fraction continue de Thiele en arithmétique Toeplitz-like.*

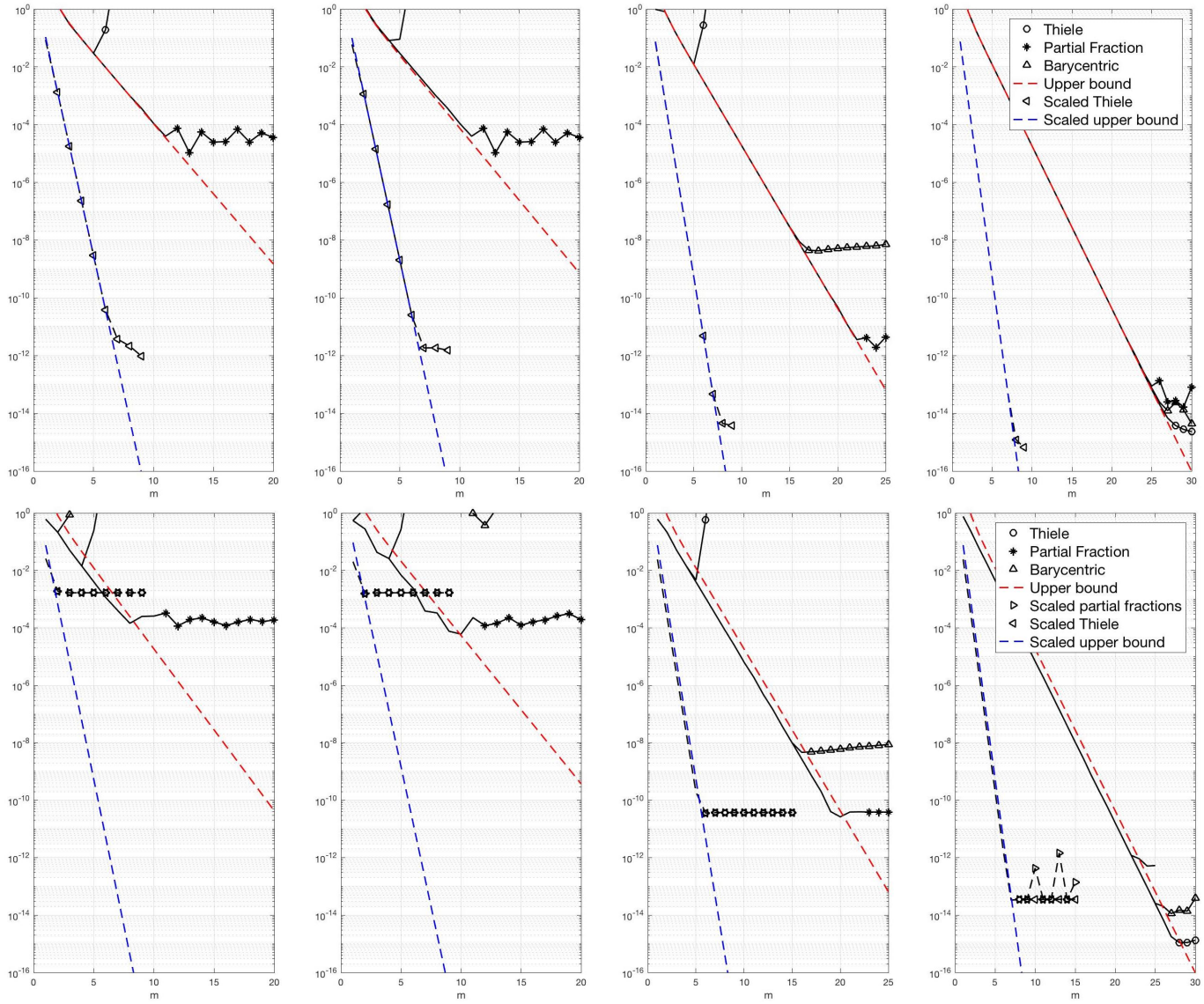

FIGURE 4.7 – Résidus $\|I - r_m^{[\nu]}(A)A^{-1}r_m^{[\nu]}(A)\|_2$ (graphiques supérieurs) et erreurs relatives $\|I - r_m(A)f^{[\mu]}(A)^{-1}\|_2$ (graphiques inférieurs) pour $A \in \mathbb{R}^{2000\times 2000}$ Toeplitz symétrique définie positive avec $\lambda_{\min} = 0,1606$, $\lambda_{\max} = 4,0552 \cdot 10^4$ et conditionnement $2,5247 \cdot 10^5$, $f^{[\mu]}(x) = \frac{\log(x)}{x-1}$ fonction de Markov, $r_m^{[\nu]}$ donné par la proposition 4.1.4 et $r_m^{[\mu]}$ Padé multipoints de f d'ordre $[m-1|m]$ aux points (4.16) sous forme de fraction continue (marqueurs circulaires), éléments simples (marqueurs étoilés) et barycentrique (marqueurs triangulaires verticaux) ainsi que Thiele scaled (marqueurs triangulaires vers la gauche) et éléments simples scaled (marqueurs triangulaires vers la droite).

Exemple 4.3.10. *En figure 4.7, nous affichons pour une matrice* $A \in \mathbb{R}^{2000\times 2000}$ *Toeplitz symétrique définie positive avec* $\lambda_{\min} = 0,1606$ *et* $\lambda_{\max} = 4,0552 \cdot 10^4$ *et conditionnement* $2,5247 \times 10^5$ *et la fonction* $f^{[\mu]}(z) = \frac{\log(z)}{z-1}$*, les résidus* $\|I - r_m^{[\mu]}(A)Ar_m^{[\mu]}(A)\|_2$ *(graphes supérieurs) pour les différents cas i. à iv. (de gauche à droite) et les erreurs relatives* $\|I - r_m^{[\mu]}(A)f^{[\mu]}(A)^{-1}\|_2$ *(graphes inférieurs) pour ces mêmes cas à l'aide des différentes formes d'implémentation de nos interpolants rationnels aux points optimisés* (4.16)*. On observe qu'en arithmétique Toeplitz-like correspondant aux cas i. et ii., aucune des représentations ne permet d'atteindre une tolérance inférieure* 10^{-4}*, ce qui nous montre alors la limite de l'arithmétique Toeplitz-like : pour des matrices avec un conditionnement un peu élevé, il nous est impossible d'approcher correctement la fonction de matrice* $f^{[\mu]}(A)$ *en arithmétique Toeplitz-like, contrairement à l'arithmétique pleine pour laquelle les représentations en éléments simples et en fraction continue de Thiele scaled permettent d'atteindre une tolérance de* 10^{-10}*.*

En résumé, pour une matrice $A \in \mathbb{R}^{n\times n}$ pas trop mal conditionnées, si l'on souhaite passer par l'arithmétique Toeplitz-like pour l'implémentation de $r_m^{[\mu]}(A)$ où $r_m^{[\mu]} = r_m(f^{[\mu]})$ interpolant rationnel de la fonction de Markov $f^{[\mu]}(z) = \int_\alpha^\beta \frac{d\mu(x)}{z-x}$ aux points d'interpolation optimisés (4.16), il est fortement recommandé d'employer une représentation en éléments simples de $r^{[\mu]}$. Pour des matrices bien conditionnées, on peut éventuellement employer une représentation en fraction continue de Thiele scaled combinée avec une méthode de Newton-DB produit. On obtient ainsi dans tous les cas un bon approximant de la fonction de Markov en la matrice avec une complexité d'ordre $\mathcal{O}(n^2)$ opérations élémentaires.

4.4 Conclusion

Dans ce chapitre nous avons démontré que pour toute fonction de Markov $f^{[\mu]}(z) = \int_\alpha^\beta \frac{d\mu(x)}{z-x}$ avec $-\infty \leqslant \alpha < \beta < \infty$ et μ une mesure positive à support dans $[\alpha;\beta]$ et pour tout ensemble $\mathbb{E} \subseteq \mathbb{R} \setminus [\alpha;\beta]$, il existe un meilleur approximant rationnel $r_m^{[\mu]} = r_m(f^{[\mu]})$ d'ordre $[m-1|m]$ sous forme d'un interpolant en $2m$ points distincts donnés par (4.16), pour lequel une borne supérieure a priori vérifiant les propriétés de meilleure approximation observée dans la littérature a pu être dégagé. L'efficacité de ces représentations est illustrée dans un premier temps dans la figure 4.1 où ces différentes formes permettent d'approcher la précision machine pour l'erreur relative.
En section 4.2, nous passons en revue 3 formes d'implémentation de nos interpolants rationnels pour éviter une implémentation sous forme de quotient de polynômes : une représentation en éléments simples pour laquelle les pôles sont les valeurs propres d'un faisceau de Loewner et les résidus sont déterminés par résolution d'un problème de moindre carrés, puis une représentation barycentrique pour laquelle la stabilité backward et forward sont déjà démontré. Enfin, une représentation en fraction continue, généralisant le concept de représentation en fraction continue des approximants de Padé des fonctions de Stieltjes aux fonctions de Markov. Dans cette section, nous apportons notre contribution au travers du théorème 4.2.2, démontrant que les coefficients de la décomposition en fraction continue de notre interpolant rationnel sont positifs, ainsi que le théorème 4.2.4 où nous démontrons la stabilité backward de la représentation en fraction continue de nos interpolants rationnels.
De par l'ordre de ces interpolants $r_m(f^{[\mu]})$ et des différentes formes d'implémentation évoquées pour leur implémentation, nous montrons en section 4.3.1 comment implémenter de manière efficace $r_m(A)$ pour $A \in \mathbb{R}^{n\times n}$ matrice de Toeplitz symétrique définie positive avec une complexité de $\mathcal{O}(n\log^2 n)$ opérations élémentaires en utilisant l'arithmétique Toeplitz-like. Le corollaire 4.3.2 apporte deux nouvelles bornes, résiduelle et a posteriori permettant de déterminer l'indice m pour lequel on maîtrise encore l'erreur. Ces résultats sont alors suivis d'expériences numériques pour différentes fonctions de Markov et matrices de Toeplitz réelles symétriques définies positives, affichant les erreurs relatives d'approximation de la fonction de matrice $f^{[\mu]}(A)$ par $r_m^{[\mu]}(A)$ sous ses différentes formes pour différents $m \geqslant 1$ ainsi que les résidus pour la

fonction de matrice donnée par (4.17).
De ces résultats, nous concluons que pour l'approximation de la fonction de matrice $f^{[\mu]}(A)$, lorsque A n'est pas trop mal conditionné, mieux vaut implémenter $r^{[\mu]}(A)$ avec $r^{[\mu]}$ interpolant rationnel aux points d'interpolation optimisés (4.16) sous forme d'éléments simples en arithmétique Toeplitz-like avec une complexité de $\mathcal{O}(n \log^2 n)$ ou $\mathcal{O}(n^2)$ opérations élémentaires en fonction de si l'on souhaite ou non reconstruire la fonction de matrice en arithmétique pleine, et si la matrice est bien conditionnée, nous pouvons envisager d'employer une représentation sous forme de fraction continue de Thiele scaled.

Chapitre 5

Conclusion et problèmes ouverts

Dans cette thèse, nous avons présenté une nouvelle méthode d'approximation des fonctions de matrices. Après avoir effectué quelques rappels sur les matrices de Toeplitz et les fonctions de matrices au chapitre 1, nous nous sommes attelés au chapitre 2 au développement d'une nouvelle arithmétique appelée Toeplitz-like permettant d'effectuer toute opération élémentaire telle que la somme, le produit ou l'inverse sur ces matrices avec une complexité de $\mathcal{O}(n \log^2 n)$. En particulier, nous avons vu qu'une fonction rationnelle de matrice $r(A)$ avec $r \in \mathcal{R}_{m,\ell}$ peut être évalué en complexité $\mathcal{O}(\max\{m, \ell\} n \log^2 n)$.

Au chapitre 3, nous avons testé notre arithmétique Toeplitz-like pour le cas de l'approximation rationnelle des fonctions de matrices $\sqrt{A}$ et $\mathrm{sign}(A)$ en comparant la méthode de Newton pour ces deux fonctions de matrices exécutée en arithmétique Toeplitz-like et en arithmétique pleine que nous avons appuyé par plusieurs expériences numériques, nous permettant dans le cas de matrices bien conditionnées d'approcher les fonctions de matrices $\sqrt{A}$ et $\mathrm{sign}(A)$ par des fonctions rationnelles de matrices $r_m(A)$ avec $r_m \in \mathcal{R}_{2^k,2^k-1}$ où $k \geqslant 0$, calculée en complexité $\mathcal{O}(\rho(A)^2 n \log^2 n)$ lorsque A est une matrice Toeplitz-like.

Enfin au chapitre 4, nous avons présenté quelques résultats originaux sur l'approximation rationnelle des fonctions de Markov $f^{[\mu]}$ par des interpolants d'ordre $[m-1|m]$ sur un ensemble $\mathbb{E} \subset \mathbb{R} \setminus [\alpha; \beta]$. Nous avons énoncé dans ce chapitre quelques résultats originaux comme une borne supérieure a priori de l'erreur relative d'approximation de $f^{[\mu]}$ par $r_m^{[\mu]}$ sur $\mathbb{E}$, ainsi que des bornes supérieures a posteriori et résiduelle, nous permettant d'introduire un critère d'arrêt pour déterminer un indice m à partir duquel l'erreur n'est plus maîtrisable d'après notre borne supérieure a priori.

Dans le cas où $\mathbb{E} = \overline{\mathbb{D}}$, nous pensons que les bornes énoncées en section 4.1.5 et qu'un choix particulier de points d'interpolation dans le cas du disque pourrait nous amener à minimiser notre borne supérieure. De plus, pour notre borne résiduelle (4.42), on considère dans le terme de droite un interpolant en les points optimisés pour la fonction de Markov $f^{[\mu]}$. Or, cette borne pourrait être optimisée en considérant plutôt des points d'interpolation optimisés pour le cas de la fonction racine carrée comme apparaissant dans cette borne puisque d'après [14, Appendix A], nous pourrions réduire le terme à droite de (4.44) à une borne

$$\|I - r_m^{[\nu]}(A)^2 \frac{1}{|\alpha|}(A - \alpha I)(A - \beta I)\| \leqslant 4\varrho^{2m}$$

en considérant l'approximation sur un intervalle de la forme $[c; d]$ avec $[c; d] \cap [\alpha; \beta] = \emptyset$.

Bibliographie

[1] N. I. Achieser, *Elements of the theory of elliptic functions*, American Mathematical Society, Providence, (1990).

[2] G. Allaire, *Analyse numérique et optimisation : une introduction à la modélisation mathématique et à la simulation numérique*, Mathématiques appliquées, Éditions de l'École Polytechnique, Palaiseau, 2005.

[3] J. J. A. Apolinario, *QRD-RLS adaptive filtering*, Springer-Verlag US, 2009.

[4] G. Arnold, N. Cundy, J. Van den Eshof, A. Frommer, S. Krieg, T. Lippert, and K. Schäfer, *Numerical methods for the QCD overlap operator : II. Optimal Krylov subspace methods*, in QCD and Numerical Analysis III, Berlin, Heidelberg, 2005, Springer, pp. 153–167.

[5] C. Badea and B. Beckermann, *Spectral sets*, in L. Hogben, Handb. Linear Algebr., 2 ed., 2013, ch. 37, pp. 26–37.

[6] G. A. Baker, Jr. and P. Graves-Morris, *Padé approximants*, vol. 59 of Encyclopedia of Mathematics and its Applications, Cambridge Univ. Press, Cambridge, second ed., 1996.

[7] L. Baratchart, M. Olivi, and F. Seyfert, *Boundary Nevanlinna-Pick interpolation with prescribed peak points. Application to impedance matching*, SIAM J. Math. Anal., 49 (2017), pp. 1131–1165.

[8] B. Beckermann, *Numerical linear algebra and functions of matrices, polycopié de cours au M2 de Mathématiques appliquées, Université de Lille*, 2013.

[9] ——, *Optimally scaled Newton iterations for the matrix square root*, in talk at the Advances in Matrix Functions and Matrix Equations workshop, Manchester, UK, 2013.

[10] B. Beckermann, J. Bisch, and R. Luce, *Computing Markov functions of Toeplitz matrices.* submitted for publication, 2021.

[11] B. Beckermann, D. Kressner, and M. Schweitzer, *Low-rank updates of matrix functions*, SIAM J. Matrix Anal. Appl., 39 (2017), pp. 539–565.

[12] B. Beckermann and A. Matos, *Algebraic properties of robust Padé approximants*, J. Approx. Theory, 190 (2015), pp. 91–115.

[13] B. Beckermann and L. Reichel, *Error estimates and evaluation of matrix functions via the Faber transform*, SIAM J. Numer. Anal., 47 (2009), pp. 3849–3883.

[14] B. Beckermann and A. Townsend, *Bounds on the singular values of matrices with displacement structure*, SIAM Rev., 61 (2019), pp. 319–344.

[15] M. Benzi and P. Boito, *Matrix functions in network analysis*, GAMM-Mitteilungen, 43 (2020).

[16] J.-P. Berrut, R. Baltensperger, and H. D. Mittelmann, *Recent developments in barycentric rational interpolation*, in Trends and applications in constructive approximation, Int. Ser. Numer. Math., Birkhäuser Basel, Switzerland, 2005, pp. 27–51.

[17] D. Bini, S. Massei, and B. Meini, *Semi-infinite Quasi-Toeplitz matrices with applications to QBD stochastic processes*, Math. Comp., 87 (2016), p. 2811–2830.

[18] D. Braess, *Nonlinear approximation theory*, vol. 7 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, 1986.

[19] ——, *Rational approximation of Stieltjes functions by the Carathéodory-Fejèr method*, Constr. Approx., 3 (1987), pp. 43–50.

[20] R. Byers and H. Xu, *A new scaling for Newton's iteration for the polar decomposition and its backward stability*, SIAM J. Matrix Anal. Appl., 30 (2008), pp. 822–843.

[21] A. Cayley, *A memoir on the theory of matrices*, Phil. Trans. R. Soc., 148 (1858), pp. 17–37.

[22] O. S. Celis, *Practical rational interpolation of exact and inexact data : theory and algorithms*, PhD thesis, Universiteit Antwerpen, 2008.

[23] S. Cheng, N. J. Higham, C. Kenney, and A. Laub, *Approximating the logarithm of a matrix to specified accuracy*, SIAM J. Matrix Anal. Appl., 22 (2001), p. 1112–1125.

[24] J. Chun and T. Kailath, *Displacement structure for Hankel, Vandermonde, and related (derived) matrices*, Linear Algebra Appl., 151 (1991), pp. 199–227.

[25] J. W. Cooley and J. W. Tukey, *An algorithm for the machine calculation of complex Fourier series*, Math. Comput., 19 (1965), pp. 249–259.

[26] M. Crouzeix, *Bounds for analytical functions of matrices*, Integr. Equ. Oper. Theory, 48 (2004), pp. 461–477.

[27] M. Crouzeix and C. Palencia, *The numerical range is a* $(1+\sqrt{2})$ *spectral set*, SIAM J. Matrix Anal. Appl., 38 (2017), pp. 649–655.

[28] J. H. Curtiss, *Faber polynomials and the Faber series*, Amer. Math. Monthly, 78 (1971), pp. 577–596.

[29] P. I. Davies and N. J. Higham, *A Schur-Parlett algorithm for computing matrix functions*, SIAM J. Matrix Anal. Appl., 25 (2004), pp. 464–485.

[30] E. D. Denman and A. N. Beavers, *The matrix sign function and computations in systems*, Appl. Math. and Comput., 2 (1976), pp. 63–94.

[31] J. Elgin, *The Faber transform and analytic continuation*, Proc. Amer. Math. Soc., 103 (1988), pp. 237–243.

[32] S. W. Ellacott, *On the Faber transform and efficient numerical rational approximation*, SIAM J. Numer. Anal., 20 (1983), pp. 989–1000.

[33] M. Embree and A. C. Ionita, *Pseudospectra of Loewner matrix pencils*, ArXiv, 1910.12153 (2019).

[34] S.-I. Filip, Y. Nakatsukasa, L. N. Trefethen, and B. Beckermann, *Rational minimax approximation via adaptive barycentric representations*, SIAM J. Sci. Comput., 40 (2018), pp. A2427–A2455.

[35] R. A. Frazer, W. J. Duncan, and A. R. Collar, *Elementary matrices and some applications to dynamics and differential equations*, Amer. J. Physics, 29 (1961), pp. 555–556.

[36] G. Freud, *Orthogonal polynomials*, Pergamon Press, Oxford, New York, Toronto, Sydney, 1971.

[37] A. Frommer and V. Simoncini, *Matrix functions*, Model Order Reduction : Theory, Research Aspects and Applications, 13 (2008), pp. 275–303.

[38] D. Gaier, *Lectures on complex approximation*, Birkhäuser Boston, Boston, MA, 1st ed. ed., 1987.

[39] T. Ganelius, *Degree of rational approximation*, in Lectures on approximation and value distribution, vol. 79 of Sém. Math. Sup., Presses Univ. Montréal, Montréal, Que., 1982, pp. 9–78.

[40] F. R. Gantmacher, *The Theory of matrices. Vol. 1*, Chelsea, New York, 1959.

[41] I. Gohberg, T. Kailath, and V. Olshevsky, *Fast Gaussian elimination with partial pivoting for matrices with displacement structure*, Math. Comp., 64 (1995), pp. 1557–1576.

[42] I. Gohberg and V. Olshevsky, *Fast state space algorithms for matrix Nehari and Nehari-Takagi interpolation problems*, Integr. Equ. Oper. Theory, 20 (1994), pp. 44–83.

[43] G. H. Golub and C. F. Van Loan, *Matrix computations*, The Johns Hopkins Univ. Press, Baltimore, London, 3rd edition ed., 1996.

[44] A. A. Gonchar, *On Markov's theorem for multipoint Padé approximants*, Math. USSR-Sb., 34 (1978), pp. 449–459.

[45] ——, *On the speed of rational approximation of some analytic functions*, Math. USSR-Sb., 34 (1978), pp. 131–145.

[46] A. A. Gonchar and E. A. Rakhmanov, *Equilibrium distributions and degree of rational approximation of analytic functions*, Math. USSR-Sb., 62 (1989), pp. 305–348.

[47] P. R. Graves-Morris, *Practical, reliable, rational interpolation*, J. Inst. Math. Appl., 25 (1980), pp. 267–286.

[48] ——, *Efficient reliable rational interpolation*, in Padé approximation and its applications, Amsterdam 1980 (Amsterdam, 1980), vol. 888 of Lecture Notes in Math., Springer, Berlin-New York, 1981, pp. 28–63.

[49] M. Gu, *Stable and efficient algorithms for structured systems of linear equations*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 279–306.

[50] G. Heinig, *Inversion of generalized Cauchy matrices and other classes of structured matrices*, in Linear Algebra for Signal Processing, Springer, New York, 1995, pp. 63–81.

[51] G. Heinig and K. Rost, *Algebraic Methods for Toeplitz-like Matrices and Operators*, Operator Theory : Advances and Applications, 13, Birkhäuser Basel, 1984.

[52] ——, *Matrices with displacement structure, generalized Bezoutians, and Moebius transformations*, in The Gohberg Anniversary Collection, Operator Theory : Advances and Applications, Birkhäuser Basel, 1989, pp. 203–230.

[53] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Res. NIST, 49 (1952), pp. 409–436.

[54] N. J. Higham, *Evaluating Padé approximants of the matrix logarithm*, SIAM J. Matrix Anal. Appl., 22 (2001), pp. 1126–1135.

[55] ——, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, PA, second ed., 2002.

[56] ——, *The numerical stability of barycentric Lagrange interpolation*, IMA J. Numer. Anal., 24 (2004), pp. 547–556.

[57] ——, *Functions of matrices, Theory and computation*, (SIAM), Philadelphia, PA, 2008.

[58] N. J. Higham and L. Lin, *An improved Schur–Padé algorithm for fractional powers of a matrix and their Fréchet derivatives*, SIAM J. Matrix Anal. Appl., 34 (2013).

[59] R. A. Horn and C. R. Johnson, *Topics in matrix analysis*, Cambridge Univ. Press, Cambridge, New York, Melbourne, 1991.

[60] T. Kailath, S.-Y. Kung, and M. Morf, *Displacement ranks of a matrix*, Bull. Amer. Math. Soc. (N.S.), 1 (1979), pp. 769–773.

[61] ——, *Displacement ranks of matrices and linear equations*, J. Math. Anal. Appl., 68 (1979), pp. 395–407.

[62] T. Kailath and A. H. Sayed, *Displacement structure : Theory and applications*, SIAM Rev., 37 (1995), pp. 297–386.

[63] T. Kailath and A. H. Sayed, eds., *Fast reliable algorithms for matrices with structure*, SIAM, Philadelphia, PA, 1999.

[64] L. A. Knizhnerman, *Padé-Faber approximation of Markov functions on real-symmetric compact sets*, Math. Notes, 86 (2009), pp. 81–92.

[65] D. Kressner and R. Luce, *Fast computation of the matrix exponential for a Toeplitz matrix*, SIAM J. Matrix Anal. Appl., 39 (2018), pp. 23–47.

[66] P. Lancaster and M. Tismenetsky, *The theory of matrices : with applications*, Computer science and applied mathematics, Academic Press, Orlando, 2nd ed., 1985.

[67] S. T. Lee, H.-K. Pang, and H.-W. Sun, *Shift-invert Arnoldi approximation to the Toeplitz matrix exponential*, SIAM J. Sci. Comput., 32 (2010), pp. 774–792.

[68] N. Levinson, *The Wiener (root mean square) error criterion in filter design and prediction*, J. Math. Phys., 25 (1946), pp. 261–278.

[69] S. Massei, *Exploiting rank structures in the numerical solution of Markov chains and matrix functions*, PhD thesis, Scuola Normale Superiore di Pisa, 2017.

[70] S. Massei, L. Robol, and D. Kressner, *hm-toolbox : Matlab software for HODLR and HSS matrices*, SIAM J. Sci. Comput., 42 (2020), pp. 43–68.

[71] A. Mayo and A. Antoulas, *A framework for the solution of the generalized realization problem*, Linear Algebra Appl., 425 (2007), pp. 634 – 662.

[72] G. Meinardus, *Approximation of functions : Theory and numerical methods*, Springer Tracts in Natural Philosophy, Vol. 13, Springer, 1 edition, 1967.

[73] R. C. Merton, *Option pricing when underlying stock returns are discontinuous*, Journal of financial economics, 3 (1976), pp. 125–144.

[74] B. N. Mukherjee and S. S. Maiti, *On some properties of positive definite Toeplitz matrices and their possible applications*, Linear Algebra Appl., 102 (1988), pp. 211–240.

[75] T. W. Ng and C. Y. Tsang, *Chebyshev-Blaschke products : solutions to certain approximation problems and differential equations*, J. Comput. Appl. Math., 277 (2015), pp. 106–114.

[76] F. W. J. Olver, D. W. Lozier, A. Olde Daalhuis, B. I. Schneider, R. Boisvert, C. Clark, B. Miller, B. Saunders, H. Cohl, and e. M.A. McClain, *NIST digital library of mathematical functions*. http ://dlmf.nist.gov/, Release 1.1.2 of 2021-06-15.

[77] V. Pan, *Decreasing the displacement rank of a matrix*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 118–121.

[78] J. D. Roberts, *Linear model reduction and solution of the algebraic Riccati equation by use of the sign function*, Internat. J. Control, 32 (1980), pp. 677–687.

[79] H. Rutishauser, *Betrachtungen zur Quadratwurzeliteration*, Monatshefte für Mathematik, 67 (1963), pp. 452–464.

[80] Y. Saad and M. H. Schultz, *GMRES : A generalized minimal residual algorithm for solving non-symmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.

[81] H. Stahl and V. Totik, *General orthogonal polynomials*, vol. 43 of Encyclopedia Math. Appl., Cambridge Univ. Press, Cambridge, 1992.

[82] J. Stoer and R. Bulirsch, *Introduction to numerical analysis*, vol. 12 of Texts in Applied Mathematics, Springer, third ed., 2002.

[83] J. J. Sylvester, *Sur la solution du cas plus général des équations linéaires en quantités binaires, c'est-à-dire en quaternions ou en matrices du second ordres. sur la résolution générale de l'équation linéaire en matrices d'une ordre quelconque. sur l'équation linéaire trinôme en matrices d'une ordre quelconque*, Comptes rendus de l'Académie des Sciences, 99 (1884), pp. 117–118,409–412,432–436,527–529.

[84] G. Szegö, *Orthogonal polynomials*, vol. 23 of A. M. S, fourth ed., 1975.

[85] O. Toeplitz, *Zur Transformation der scharen bilinearer Formen von unendlich vielen Veränderlichen, Nachr. der kgl. Gesellschaft der Wissenschaften zu Göttingen*, Mathematisch-physikalische Klasse, (1907), pp. 110–115.

[86] ——, *Zur theorie der quadratischen und bilinearen Formen von unendlich vielen Veränderlichen*, Math. Ann., 70 (1911), pp. 351–376.

[87] L. N. Trefethen, *Approximation theory and approximation practice*, SIAM, Philadelphia, PA, 2013.

[88] J. Van den Eshof, A. Frommer, T. Lippert, K. Schilling, and H. Van der Vorst, *Numerical methods for the QCD overlap operator. I. Sign-function and error bounds*, Comput. Phys. Commun., 146 (2002), p. 203–224.

[89] E. I. Zolotarev, *Application of elliptic functions to questions of functions deviating least and most from zero*, Zap. Imp. Akad. Nauk, 30 (1877), pp. 1–59.