Université de Lille, Sciences et Technologies
École Doctorale des Sciences Pour l'Ingénieur (Lille)

THÈSE DE DOCTORAT

Spécialité : Mathématiques

présentée en vue de l'obtention du grade de docteur par :

Thomas Hélart

Sur l'optimalité de l'inégalité de Bernstein-Walsh à poids et ses applications aux méthodes de Krylov.

On the sharpness of the weighted Bernstein-Walsh inequality and its application to Krylov methods.

Soutenue le 27 septembre 2018

Composition du jury :

Rapporteurs :	Baratchart Laurent	INRIA Sophia Antipolis
	Frommer Andreas	Université de Wuppertal
Examinateurs :	Maïda Mylène	Université de Lille
	Matos Ana	Université de Lille
	Vandebril Raf	Université de Louvain
Directeur :	Beckermann Bernhard	Université de Lille

Université de Lille

Laboratoire Paul Painlevé

Remerciements

Je tiens en premier lieu à exprimer ma profonde gratitude à mon directeur de thèse Bernhard Beckermann pour m'avoir encadré durant mon stage de master 2 et pour avoir dirigé cette thèse. Je le remercie pour sa disponibilité, ses conseils, ses suggestions et son exigence qui m'ont permis d'écrire ce document et de découvrir un sujet passionant.

Je voudrais aussi signaler ma reconnaissance à la région nord-pas-de-calais et à l'université Lille 1 pour avoir financé ce projet, ainsi qu'au laboratoire Paul Painlevé pour m'avoir permis de réaliser mon travail dans d'excellentes conditions.

J'adresse mes sincères remerciements à Laurent Baratchart et Andreas Frommer d'avoir accepté d'être les rapporteurs de ma thèse et pour la rapidité et l'attention avec laquelle ils ont relu ce manuscrit. Leurs rapports de thèse me touchent profondément et leurs remarques pertinentes ont permis d'améliorer le document final. Je suis également honoré par la présence de Mylène Maïda, Ana Matos et Raf Vanderbril parmi les membres du jury.

Je souhaite aussi remercier l'ensemble des personnes que j'ai pu cotoyer au sein de l'université de Lille en tant qu'étudiant, en tant qu'enseignant, ou en tant que chercheur.

Une petite pensée pour l'équipe de mathématique de l'université de Mons qui m'a accueuilli durant l'année 2017-2018.

Et enfin un grand merci à mes amis et à ma famille pour avoir été à mes cotés et m'avoir soutenu. Mes camarades de Kung Fu, qui m'ont permis de me défouler quand j'en avais besoin ; mes amis, toujours d'attaque pour partager des moments houblonnés ; mes parents, pour leur soutien de longue date ; Darina Abderrahmane, pour son soutien, ses encouragements et ses idées de voyages et d'aventures ; et ma fille, sans qui mes nuits auraient été bien trops longues et qui d'un sourire sait me faire oublier tous mes tourments mathématiques.

RESUME

Les méthodes de projection sur des espaces de Krylov ont été employées avec grand succès pour diverses tâches en calcul scientifique, par exemple la résolution de grands systèmes d'équations linéaires, le calcul approché de valeurs propres, ou encore le calcul approché des fonctions de matrices fois un vecteur. L'objectif majeur de cette thèse est d'étudier et d'expliquer la convergence superlinéaire des méthodes de Krylov. La plupart des résultats existants sont asymptotiques avec passage à la racine n-ième et considèrent des suites de matrices. Dans un premier temps, nous généralisons une formule de Ipsen et al. concernant la convergence superlinéaire des méthodes MR valable pour des disques, à l'aide des opérateurs de Hankel et de la théorie AAK. Notre analyse permet aussi d'obtenir des bornes supérieures pour des ensembles convexes en utilisant la transformée de Faber. Ensuite nous énonçons notre principal résultat qui est un théorème d'optimalité en théorie du potentiel logarithmique. Nous montrons, à l'aide d'une nouvelle technique de discrétisation d'un potentiel, que l'inégalité de Bernstein-Walsh à poids sur un intervalle réel est optimale, à un facteur universel près, dans le cas où le champs extérieur est un potentiel d'une mesure à support réel à gauche de l'intervalle, ce qui inclut le cas des poids polynômiaux. Via un lien avec un problème sous contrainte, l'inégalité précédente s'applique à l'analyse de la convergence des méthodes de Krylov, et permet de prédire analytiquement un taux de convergence superlinéaire de la méthode du gradient conjugué et des approximations de Rayleigh-Ritz pour des fonctions de Markov, à chaque étape et pour une seule matrice.

ABSTRACT

Projection methods on Krylov spaces were used with great success for various tasks in scientific computing, for example the resolution of large systems of linear equations, the approximate computation of eigenvalues, or the approximate computation of matrix functions times a vector. The main goal in this thesis is to study and explain superlinear convergence of Krylov methods. Most of the existing formulas provide asymptotic results for the n-th root considering an increasing sequence of matrices. Firstly, we generalize a formula of Ipsen et al. concerning superlinear convergence of MR methods valid for disks using Hankel operators and AAK theory, our analysis also allows to obtain upper bounds for convex sets using the Faber transform. Then we state our main theorem which is a sharpness result in logarithmic potential theory using a new technique of discretization of a logarithmic potential. We prove that the weighted Bernstein-Walsh inequality on a real interval is sharp up to some universal constant, when the external field is given by a potential of a real measure supported at the left of the interval. As a special case this result includes the case of weights given by polynomials. Via a link with a constrained extremal problem our inequality applies to the analysis of the convergence of Krylov methods, and allows us to predict analytically the superlinear convergence of the conjugate gradient method and of the error for Rayleigh-Ritz approximations for Markov functions. Our results apply to a simple matrix, without taking the limit and without n-th root.

Contents

Introduction

Un problème important en mathématiques appliquées est le calcul de

$$f(A)b,$$

où A est une matrice carrée, b un vecteur, et f une fonction telle que $f(A)$ est bien définie. En effet, il est courant de modéliser un phénomène physique par une équation aux dérivées partielles, dont la solution fait souvent apparaître des fonctions d'opérateurs. La discrétisation de ces opérateurs donne des matrices A_N de grande taille N qui ne permettent pas le calcul de $f(A_N)$. Une idée courante en mathématiques est de projeter le problème sur un espace de dimension n beaucoup plus petit dans lequel on pourra évaluer de manière efficace les fonctions de matrices fois un vecteur. Les méthodes de projection qui sont étudiées dans cette thèse sont des méthodes dites de Krylov. Le point commun de ces méthodes est d'utiliser les sous-espaces de dimension (au plus) n engendrés par l'ensemble des $r(A)b$ où r est une fonction rationnelle dont le numérateur p et le dénominateur q sont des polynômes de degré au plus $n-1$. Les sous-espaces de Krylov polynômiaux ($q = 1$) ont été introduits en 1931 par le mathématicien russe A.N. Krylov [73] dans un papier concernant le calcul approché de valeurs propres et portent aujourd'hui son nom.

Les méthodes de Krylov ont été utilisées par une multitude d'auteurs dans des domaines très variés des mathématiques aussi bien théoriques qu'appliquées. Par exemple, pour la fonction inverse, le problème revient à résoudre un système linéaire, et les méthodes de Krylov englobent des algorithmes comme le gradient conjugué [66] ou GMRES [102] . La convergence de ces méthodes n'est pas encore pleinement comprise de nos jours. L'étude de la convergence des méthodes de Krylov (polynomiales) pour les systèmes linéaires est reliée à la quantité [59, 13]

$$E_n(S) = \min\left\{\frac{||p||_S}{|p(0)|}, \quad p \in \Pi_n, \quad \forall \lambda \in \Lambda(A) \setminus S : \quad p(\lambda) = 0\right\}, \qquad (1)$$

où Π_n désigne l'ensemble des polynômes de degré au plus n et S est un sous-ensemble de $\mathbb{C}$ ne contenant pas 0. Dans le cas particuler où S contient le spectre de la matrice en question, on obtient

$$E_n(S) = \min_{p \in \Pi_n} \max_{z \in S} \frac{|p(z)|}{|p(0)|}.$$

Dans ce cas, il est connu que l'on a [35]

$$E_n(S) \geq \rho_S^n, \quad \lim_{n\to\infty} E_n(S)^{1/n} = \rho_S \leq 1. \tag{2}$$

où le nombre ρ_S est appelé le facteur d'estimation de convergence asymptotique associé à S. Il est le point de départ de la plupart des études qui essayent de comprendre le comportement de la suite des $E_n(S)$. Ce nombre ρ_S peut être obtenu à l'aide de la théorie du potentiel et notamment avec les fonctions de Green

$$\rho_S = \exp(-g_S(0,\infty)).$$

Si l'ensemble S est connexe, on peut alors calculer ρ_S à l'aide de l'application conforme de Riemann $\phi_s : \overline{\mathbb{C}} \setminus S \to \overline{\mathbb{C}} \setminus \mathbb{D}$ qui vérifie $\phi_s(\infty) = \infty$ et $\phi_s'(\infty) > 0$

$$\rho_s = \frac{1}{|\phi_s(0)|}.$$

Les bornes supérieures que l'on trouve dans la littérature sont des bornes dites linéaires, c'est-à-dire de la forme une constante fois un terme a^n avec a un réel positif plus petit que 1. Ces bornes sont en général satisfaisantes pour les premières itérations, mais en général, lorsque n grandit, ces bornes fournissent des surestimations trop importantes car la convergence s'accélère, ce que l'on appelle convergence superlinéaire [88], et ce phénomène n'est pas encore bien compris. Une formule expliquant ce phénomène manque toujours. A. Kuijlaars est le premier auteur à avoir quantifié [75] la fameuse « rule of thumb » énoncée par D. Bau et L.N. Trefethen dans [9] qui décrit la répartition d'équilibre des valeurs propres. Inspirés par ces travaux B. Beckermann et A. Kuijlaars ont publié en 2000 [12] un article expliquant la convergence superlinéaire du gradient conjugué. Le cas des fonctions de matrices a été traité par B. Beckermann et S. Güttel [17] dans un article publié en 2012. Tous ces travaux sont des résultats asymptotiques après passage à la racine n-ième. De plus, dans ces travaux, les auteurs considèrent une suite de matrices et donc ne donnent pas d'information sur le taux de convergence d'une seule matrice. L'objectif est d'étudier et d'expliquer la convergence superlinéaire des méthodes de Krylov et de fournir une formule sans racine n-ième et sans passage à la limite pour une matrice fixée. Ce qui précède est détaillé dans les deux premiers chapitres où on présente les outils mathématiques de base. Nous présentons aussi un état de l'art sur l'étude de la convergence des méthodes de résolution de systèmes linéaires MR et OR, définies dans le chapitre 2, avec les principales bornes supérieures que l'on peut trouver dans la littérature. Ces deux premiers chapitres ne contiennent pas de résultats originaux. Précisons que dans cette thèse nous travaillerons toujours en précision infinie et nous ne discuterons pas des phénomènes liés à la précision finie de l'ordinateur.

Dans le chapitre 3, nous présentons un travail qui généralise un article de S.L. Campbell, I.C.F. Ipsen, C.T. Kelley et C.D. Meyer [24] en utilisant l'analyse complexe et les polynômes de Faber, dans lequel les auteurs ont donné des bornes d'erreur pour GMRES. L'idée est de séparer le spectre en regroupant les valeurs

propres proches les unes des autres. Si on a une accumulation de valeurs propres dans une région du plan complexe, on considère un ensemble S qui les contient, et dans l'analyse de la convergence, on ne distingue plus ces valeurs propres. Les valeurs propres isolées, appelées outliers, sont traitées à part dans l'étude. Les auteurs utilisent l'inégalité $\frac{\|r_n\|}{\|r_0\|} \leq CE_n(S)$ pour une certaine constante C, où r_n désigne le résidu à l'étape n et $E_n(S)$ est définit en (1) pour un ensemble S ne contenant pas forcément le spectre. Le travail consiste à borner cette quantité $E_n(S)$. La première étape est de ramener le problème sur le disque unité du plan à l'aide de l'application conforme ϕ. Nous fournissons une borne inférieure dans le théorème 3.1.3 notée

$$E_n(S) \geq \frac{1}{|f_{n,\alpha_1,\ldots,\alpha_d}(\alpha_0)|} = \frac{1}{|\alpha_0|^n} \prod_{j=1}^{d} \left| \frac{1-\bar{\alpha}_j\alpha_0}{\alpha_0-\alpha_j} \right|,$$

avec α_0 l'image par ϕ de zéro et les α_j, $j \geq 1$, les images des outliers par ϕ. Noter que l'on retrouve le facteur d'estimation de convergence asymptotique $\rho_S = \frac{1}{|\alpha_0|}$. L'intérêt de considérer plus d'outliers réside dans le fait que l'on peut choisir des ensembles S emboités qui décroient, ainsi ρ_S est aussi décroissant, et donc on obtient un meilleur taux de convergence. Le prix à payer pour obtenir ce taux est la multiplication par le produit de Blaschke $\prod_{j=1}^{d} \left| \frac{1-\bar{\alpha}_j\alpha_0}{\alpha_0-\alpha_j} \right|$ dont chaque terme est supérieur à 1. Le nombre optimal d'outliers à choisir dépend donc de l'indice d'itération n. Dans cette partie nous n'expliquons pas comment choisir les outliers. Dans [24] les ensembles S considérés sont des disques et les auteurs obtiennent

$$E_n(S) \leq \frac{C_{Ipsen}}{|f_{n,\alpha_1,\ldots,\alpha_d}(\alpha_0)|},$$

avec C_{Ipsen} une constante strictement supérieure à 1 qui dépend de S et d. Nous proposons dans le théorème 3.1.4 une borne d'erreur où l'on remplace C_{Ipsen} par une quantité plus générale qui a l'avantage de tendre vers 1 lorsque n grandit et d et S sont fixés. Pour obtenir cette borne supérieure, nous faisons un lien entre les opérateurs de Hankel et un problème de minimisation. Nous généralisons aussi ce travail pour des ensembles S convexes à l'aide de la transformée de Faber avec le défaut d'apporter un facteur 3 dans la borne supérieure par outlier considérée. Et donc cette borne explose avec un nombre important d'outliers.

Le principal résultat de cette thèse, énoncé dans le chapitre 4, concerne un théorème d'optimalité de l'inégalité de Bernstein-Walsh à poids sur un intervalle réel, pour un champs extérieur donné par le potentiel d'une mesure à support réel à gauche de l'intervalle. L'outil mathématique principal est la théorie du potentiel logarithmique avec champs extérieurs Q. En se fixant une union finie d'intervalles compacts $\Sigma \subset \mathbb{R}$, un poids $w = \exp(-Q)$ continu, et en notant le potentiel logarithmique U^μ et son énergie $I(\mu)$, on sait [103, Theorem I.1.3 and Theorem I.4.8] qu'il existe une unique mesure μ_w de support $\mathrm{supp}(\mu_w)$ parmi toutes les mesures de probabilité μ à support dans Σ qui minimise

$$\inf\{I(\mu) + 2\int Q \mathrm{d}\mu : \|\mu\|_\Sigma = 1\}.$$

Cette mesure extrémale est caractrisée par l'existence d'une constante $F \in \mathbb{R}$ telle que

$$\Theta(x) := F - U^{\mu_w}(x) - Q(x) \begin{cases} = 0 & \text{pour } x \in \text{supp}(\mu_w), \\ \leq 0 & \text{pour } x \in \Sigma. \end{cases}$$

L'inégalité de Bernstein-Walsh à poids [103, Theorem III.2.1] dit que pour tout poids w continu, pour tout z complexe et pour tout polynôme P de degré au plus $k \in \mathbb{N}$ on a

$$\frac{|w(z)^k P(z)|}{\|w^k P\|_{\text{supp}(\mu_w)}} \leq e^{k\Theta(z)}. \tag{3}$$

Cette inégalité implique des bornes inférieures pour les méthodes de Krylov. On souhaite donc obtenir une inégalité dans l'autre sens. Il existe un résultat d'optimalité après passage à la racine k-ième et passage à la limite [103, Corollary III.1.10] qui donne l'existence d'un polynôme P_k de degré au plus k tel que pour tout z complexe en dehors de $\text{supp}(\mu_w)$ on a

$$\lim_{k\to\infty} \left(\frac{|w(z)^k P_k(z)|}{\|w^k P_k\|_{\text{supp}(\mu_w)}} \right)^{1/k} = e^{\Theta(z)}, \tag{4}$$

avec $\|f\|_\Sigma = \max_{x\in\Sigma} |f(x)|$. Le théorème 4.1.3 améliore (4) pour $\Sigma = [\alpha, \beta]$ et pour une classe particulière de champs extérieurs issus d'un potentiel positif $Q = U^{\rho/k}$ sur $[\alpha, \beta]$, avec ρ une mesure de Borel à support compact dans $(-\infty, \alpha]$. En effet, nous démontrons que la mesure extrémale μ_w vérifie les conditions

$$\text{supp}(\mu_w) = [a, \beta] \subset [\alpha, \beta],$$

pour un a explicitement calculable, et

$$k\Theta(x) = (k + \|\rho\|) g_{[a,\beta]}(x, \infty) - \int g_{[a,\beta]}(x, y) \mathrm{d}\rho(y),$$

où $g_S(., y)$ désigne la fonction de Green de l'ensemble S avec pôle y. Nous montrons aussi que l'inégalité (3) est optimale à une constante universelle C_{BW} près : pour tout $k \geq 2$, il existe un polynôme P_k de degré au plus k tel que pour tout $x_0 \in \mathbb{R} \setminus \text{supp}(\mu_w)$, on a

$$\frac{|w(x_0)^k P_k(x_0)|}{\|w^k P_k\|_{\text{supp}(\mu_w)}} \geq e^{-C_{BW}}\, e^{k\Theta(x_0)}.$$

La constante donnée ici n'est pas optimale, mais elle a comme principal intérêt d'être universelle.

Notre preuve du théorème 4.1.3 est basée sur un nouveau résultat concernant la discrétisation d'un potentiel logarithmique pour une classe de mesures particulière. Nous montrons dans le théorème 4.2.1 que pour une mesure μ qui vérifie

$$k \frac{d\mu}{dx}(t) = g(t) \frac{k}{\pi\sqrt{(t-a)(\beta-t)}}$$

avec g une fonction vérifiant certaines propriétés (ce qui est le cas pour la mesure extrémale de notre théorème d'optimalité), il existe une constante universelle C_{BW} telle que pour tout $k \geq 2$ on peut construire un polynôme unitaire P_k de degré k qui vérifie les deux propriétés suivantes

1. $\forall z \in \mathbb{C} : \log|P_k(z)| + kU^\mu(z) \leq C_{BW}$,
2. $\forall x \in \mathbb{R} \setminus (a,\beta) : \log|P_k(x)| + kU^\mu(x) \geq 0$.

On suppose dans la suite sans perte de généralité que $[a,\beta] = [-1,1]$ (le cas général s'en déduit par changement de variable affine). Les résultats dans [114, Lemme 9.1] ou [78] et [79], donnent des constantes non explicites qui dépendent des données de départ. Les travaux présentés ici sont reliés à ceux de V. Totik et D. Lubinsky à propos de liens avec des formules de quadrature à poids [80]. En effet, le début de l'étude est similaire, on approche le potentiel logarithmique U^μ à l'aide de la formule de quadrature

$$\sum_{j=0}^{k-1} \log \frac{1}{|x-\xi_j|} = -\log|P_k(x)|, \quad P_k(x) = \prod_{j=0}^{k-1}(x-\xi_j),$$

en découpant $[-1,1]$ en k sous-intervalles $[t_j, t_{j+1}]$, $-1 = t_0 < t_1 < ... < t_k = 1$, de masse $\mu([t_j,t_{j+1}]) = 1/k$, et en prenant $\xi_j \in [t_j,t_{j+1}]$ pour $j = 0,...,k-1$. On obtient alors

$$\log|P_k(x)| + kU^\mu(x) = \sum_{j=0}^{k-1} \int_{t_j}^{t_{j+1}} \log\left|\frac{x-\xi_j}{x-t}\right| k \,\mathrm{d}\mu(t).$$

Il y a deux problèmes majeurs, le premier est que la fonction de densité de μ peut avoir des singularités aux points ± 1, le deuxième est que si $x \in [-1,1]$ on a une singularité dans l'intégrande. L'idée est de couper cette somme en trois et de traiter chaque terme séparément. Nous proposons une nouvelle approche pour majorer ces trois termes basée sur une propriété de la fonction de repartition W_g de notre mesure $k\mu$: en utilisant les hypothèses sur la fonction de densité g, nous montrons dans notre théoréme 4.2.6 la formule de la moyenne

$$c_1 W_g'(\frac{t+x}{2}) \leq \frac{W_g(t) - W_g(x)}{t-x} \leq c_2 W_g'(\frac{t+x}{2}),$$

avec $W_g'(t) = \frac{kg(t)}{\pi\sqrt{1-t^2}}$ et $c_1, c_2 \in \mathbb{R}_+^*$ deux constantes universelles.

Dans le chapitre 5, nous appliquons le théorème d'optimalité de l'inégalité de Bernstein-Walsh à CG et aux fonctions de matrices fois un vecteur pour des fonctions de Markov et des matrices hermitiennes. Ces résultats améliorent ceux obtenus dans [12] pour CG, et dans [17] pour les fonctions de matrices et les espaces de Krylov polynomiaux. Dans ces articles les résultats sont asymptotiques, avec une racine n-ième et considérent une suite de matrices d'ordre N, Sous certaines conditions techniques (section 2.3.3), les auteurs ont obtenu dans [12]

$$\limsup_{\substack{n,N\to\infty \\ n/N\to t}} E_n(\Lambda(A_N))^{1/N} \leq \exp\left(-\int_0^t g_{S(\tau)}(0,\infty)\,\mathrm{d}\tau\right), \tag{5}$$

avec $t \in (0,1)$ et $S(t)$ une famille décroissante d'ensembles dépendant de la distribution des valeurs propres. La fonction de droite est donc concave, ce qui explique la convergence superlinéaire. Les expériences numériques laissaient penser que l'on pouvait obtenir une inégalité pour une seule matrice, sans racine n-ième et sans passer à la limite. Dans cette thèse, on fournit une formule théorique qui le démontre. En particulier, pour CG, nous prouvons la conjecture 2.3.1 dans le théorème 5.1.4 pour une sous-classe de distribution de valeurs propres. En effet, pour un certain $d = d_n$, nous obtenons (à comparer avec (5))

$$E_n(\Lambda(A)) \leq E_n([\lambda_{d+1}, \beta]) \leq e^{C_{BW}} \exp\Big(-N \int_0^{n/N} g_{S(\tau)}(0,\infty)\ \mathrm{d}\tau\Big), \quad (6)$$

avec $S(t) = [a(t), \beta]$ une famille d'intervalles décroissante. Remarquons que pour un ensemble S contenant $S(0)$ (et le spectre de A), on a

$$\exp\Big(-N \int_0^{n/N} g_{S(\tau)}(0,\infty)\ \mathrm{d}\tau\Big) \leq \exp\Big(-n g_S(0,\infty)\Big) = \rho_S^n.$$

Par rapport à (2), nous avons remplacé $g_S(0,\infty)$ par une moyenne des fonctions $g_{S(t)}(0,\infty)$ sur $[0, n/N]$. On peut dans le même esprit appliquer l'optimalité de Bernstein-Walsh à poids aux fonctions de matrices fois un vecteur pour des matrices hermitiennes et pour des fonctions de Markov de la forme

$$f(z) = \int_\Gamma \frac{\mathrm{d}\gamma(x)}{x - z},$$

où γ est une mesure supportée sur un fermé inclu dans $[-\infty, x_0]$ situé à gauche du spectre réel. Le corollaire 5.2.9 donne pour les approximations de Rayleigh-Ritz f_n

$$\|f(A)b - f_n\| \leq K \exp\left(-N \int_0^{n/N} g_{S(\tau)}(x_0,\infty)\ \mathrm{d}\tau\right), \quad (7)$$

avec K une constante dépendant de f et de C_{BW}. La fonction

$$\exp\left(-N \int_0^{n/N} g_{S(\tau)}(x_0,\infty)\ \mathrm{d}\tau\right)$$

est concave en n, ce qui explique la convergence superlinéaire. On a ainsi remplacé dans (6) et (7) des résultats asymptotiques avec une n-ième racine pour une suite de matrices par des inégalités pour une seule matrice.

PLAN

chapitre 1

Nous commençons par introduire les outils de base de cette thèse. Nous définissons d'abord les fonctions de matrices et d'opérateurs, ensuite nous présentons les sous-espaces de Krylov, puis nous discutons des méthodes de Krylov pour un produit fonction de matrice-vecteur.

chapitre 2

Nous nous intéressons aux systèmes linéaires, c'est-à-dire à la fonctions $f(z) = z^{-1}$. Nous présentons différentes méthodes de Krylov pour résoudre ces systèmes, nous regardons en particulier deux méthodes de projection qui sont les méthodes OR et les méthodes MR. Nous présentons ensuite un état de l'art sur les bornes linéaires les plus connues et introduisons la notion de convergence superlinéaire.

chapitre 3

Nous améliorons une formule de [24] concernant la convergence superlinéaire des méthodes MR valable pour des disques S en utilisant l'analyse complexe, les polynômes de Faber et un lien avec les opérateurs de Hankel et un problème de minimisation. Nous présentons aussi une borne inférieure et nous généralisons la borne supérieure pour des ensembles S convexes à l'aide de la transformée de Faber.

chapitre 4

Nous démontrons un théorème d'optimalité en théorie du potentiel logarithmique. À l'aide d'une nouvelle technique de discrétisation d'un potentiel, nous prouvons que l'inégalité de Bernstein-Walsh à poids sur un intervalle réel est optimale, à un facteur universel près, dans le cas où le champs extérieur est le potentiel d'une mesure à support réel à gauche de l'intervalle.

chapitre 5

En utilisant un lien entre un problème extrémal avec champs extérieur et un problème extrémal sous contrainte, nous appliquons notre résultat d'optimalité à la méthode du gradient conjugué (CG) et aux fonctions de matrices pour des matrices hermitiennes et des fonctions de Markov.

OUTLINE

chapter 1

We begin by introducing the basic tools used in this thesis. We first define functions of matrices and operators, then we present Krylov subspaces, and finally we discuss Krylov methods for matrix functions times a vector.

chapter 2

We study linear systems, that is, the function $f(z) = z^{-1}$. We present different Krylov methods to solve these systems and look in particular at two projection methods which are the OR methods and the MR methods. Next, we give a state of the art with some well-known linear bounds and introduce the notion of superlinear convergence.

chapter 3

We improve a formula in [24] concerning the superlinear convergence of MR methods valid for disks S by using complex analysis, Faber polynomials and a link with Hankel operators and a minimization problem. We also present a lower bound and an upper bound for convex sets S using the Faber transform..

chapter 4

We prove an optimality theorem in logarithmic potential theory. Using a new technique of discretization of a potential, we show that the weighted Bernstein-Walsh inequality on a real interval is sharp up to some universal constant, in the case when the external field is given by a potential of a real measure supported at the left of the interval

chapter 5

Exploiting a link between an extremal problem with an external field and an extremal problem under constraint, we apply our sharpness result to the convergence of conjugate gradient (CG) and to the functions of matrices for Hermitian matrices and Markov functions.

Chapter 1

Functions of matrices and Krylov spaces

The aim of this chapter is to give the main basic tools used in this thesis. We start by talking about functions of matrices in Section 1.1: we give three equivalent definitions, present some examples where we can find them, and make some remarks on computational aspects. The difficulty to compute functions of matrices and the examples presented motivate the will to work with functions of matrices times a vector instead of just functions of matrices. Section 1.2 is devoted to the presentation of Krylov spaces. We show some basic properties of rational Krylov spaces, introduced by Ruhe in [98], by using a definition based on polynomial Krylov spaces. This approach reveals the relationships between polynomial and rational Krylov spaces. Then we present the Rayleigh Ritz approximations in Section 1.3 and give a characterization property and an exactness property.

1.1 Functions of matrices

1.1.1 Definitions

Matrix functions are useful tools in applied mathematics and scientific computing. Over the last thirty years, one can observe a very important research activity in the field of matrix functions using tools from numerical linear algebra and approximation theory. Whereas it is easy to define the polynomial (of degree k) of a matrix

$$p_k(A) = \sum_{j=0}^{k} a_j A^j,$$

or an entire function as the exponential

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

things become more tricky for functions defined only on subsets of the complex plane. Those expressions can be defined in terms of the Jordan canonical form of A, the minimal polynomial of A, or by a Cauchy-type integral. The latter definition requires f to be analytic in an open set containing the spectrum of A, with the path of integration in this set. Detailed discussions on these definitions, their requirements on f, and their implementation are provided by Golub and Van Loan [55], Higham [68], or Horn and Johnson [70]. One can find a good introduction in the paper [49]. Let us recall in this section the definitions, the main properties and give a brief overview on computational aspects.

Jordan canonical form

A matrix $A \in \mathbb{C}^{n\times n}$ can be expressed in the Jordan canonical form $A = ZJZ^{-1}$ where Z is nonsingular and $J = diag(J_1, \ldots, J_p)$ with each J_l of the form

$$J_l = \begin{pmatrix} \lambda_l & 1 & & \\ & \lambda_l & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_l \end{pmatrix} \in C^{m_l \times m_l}.$$

The Jordan matrix J is unique up to the ordering of the blocks, but the transforming matrix Z is not. Denote by $\{\lambda_k\}_{k=1}^{s}$ the distinct eigenvalues of A and by n_k the index of λ_k (the order of the largest Jordan block in which λ_k appears). A function is said to be defined on the spectrum of A, denoted by $\Lambda(A)$, if the values $f^{(j)}(\lambda_k)$ exist for $k = 1 : s$ and $j = 0 : n_k - 1$. Those values are called the values of the function f on the spectrum of A.

Definition 1.1.1 (Jordan canonical form) *Let f be defined on $\Lambda(A)$ and let A have the Jordan canonical form $A = ZJZ^{-1}$. Then*

$$f(A) = Zf(J)Z^{-1} = Zdiag(f(J_1), \ldots, f(J_p))Z^{-1}$$

where for $l = 1 : p$

$$f(J_l) = \begin{pmatrix} f(\lambda_l) & f'(\lambda_l) & \ldots & \frac{f^{(m_l-1)}(\lambda_l)}{(m_l-1)!} \\ & f(\lambda_l) & \ddots & \vdots \\ & & \ddots & f'(\lambda_l) \\ & & & f(\lambda_l) \end{pmatrix}.$$

It is important to note that it yields an $f(A)$ that can be shown to be independent of the particular Jordan canonical form we used. It is also interesting to note that it requires only the values of f on the spectrum of A. If

A is diagonalizable, the index of all eigenvalues is equal to one and $f(A) = Zdiag(f(\lambda_1), \ldots, f(\lambda_s))Z^{-1}$ (where each eigenvalue is repeated according to its multiplicity). If A is normal (in particular Hermitian), then it is diagonalizable and one can choose a unitary Z.

Polynomial interpolation

A further representation of the matrix function $f(A)$ is based on polynomial interpolation. It is well-known that there exists a unique monic polynomial ψ_A of minimal degree such that $\psi_A(A) = 0$, called the minimal polynomial, which satisfies the formula

$$\psi_A(z) = \prod_{j=1}^{s} (z - \lambda_j)^{n_j}.$$

Definition 1.1.2 (Polynomial interpolation) *Let f be defined on the spectrum of A. Then*

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

where $p_{f,A}$ is the unique polynomial of degree less than $\deg(\psi_A)$ that interpolates f on the spectrum of A in the Hermite sense

$$p_{f,A}^{(j)}(\lambda_k) = f^{(j)}(\lambda_k), \text{ for } j = 0 : n_k - 1 \text{ and } k = 1 : s.$$

This definition tells us that for computing matrix functions, at least in theory, it is sufficient to know how to evaluate a polynomial of a matrix. It is important to note that the polynomial $p_{f,A}$ given in the definition depends on A, so it is not the case that $f(A) = q(A)$ for some fixed polynomial q independant of A. Sometimes, it is convenient to impose more interpolation conditions than necessary, but it does not affect the ability of the polynomial to produce $f(A)$. Indeed, if q is a polynomial that interpolates f at the spectrum of A in the sense of Hermite and has some additional interpolation conditions, then $q(A) = p_{f,A}(A) = f(A)$.

Cauchy integral

Another way of representing the matrix function $f(A)$ is based on the Cauchy integral formula which is particularly useful for error estimates.

Definition 1.1.3 (Cauchy integral) *Let f be analytic on and inside Γ, a system of Jordan curves encircling each $\lambda_j \in \Lambda(A)$ exactly once (with mathematically positive orientation). Then*

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

This definition is independent of the particular choice of Γ and has the advantage that it can be generalized to operators in Banach spaces thanks to the Riesz-Dunford functional calculus [40].

Remarks

The preceding definitions are equivalent modulo the requirement in the Cauchy integral definition that f is analytic in a region containing the spectrum [68, Theorem 1.12]. No given definition should be used like a black box procedure for computing matrix functions for the following reasons.

- The size of a Jordan block is not (in general) stable under perturbations.
- The interpolation polynomial suffers from the same drawback, and in addition one should know how to efficiently evaluate $p(A)$.
- With the Cauchy integral definition, if one wants to define multi-valued functions like $\log(A)$ or $\sqrt{A}$, one has to choose correctly the set Ω and should know how to select the contour Γ. In general one has to exclude some matrices. We also note that errors are introduced by applying quadrature formulae.

For algebraic operators [88, section 2.8] we can generalize the interpolating polynomial definition. An operator A is said to be algebraic of degree n if there exists a monic minimal polynomial ψ_A of degree n such that $\psi_A(A) = 0$. For such operators we can use the polynomial definition 1.1.2, and every function of an algebraic operator is representable as a polynomial depending on f and ψ_A. With this definition, f need not be analytic in a neighborhood of $\Lambda(A)$, instead it is only required that f possesses derivatives up to a finite order.

1.1.2 Motivation and examples

A nice exposition about various applications of matrix functions for different tasks of scientific computing can be found in Higham's book [68, chapter 2].

Linear systems

The most famous function of matrices is the inverse function $f(z) = z^{-1}$ which appears for example in linear systems. For large matrices, it is in general a very difficult task to compute an inverse. If we want to solve the linear system $Ax = b$ ($f(A) = A^{-1}$), it is well-known that we should not compute A^{-1} (in general).

Differential equation

Many problems in science and engineering are modeled by partial differential equations (using unbounded operators). After a discretization in space, for example, by finite differences, finite elements or pseudospectral methods, such problems can be written as a semilinear system of ordinary differential equations

$$\frac{\mathrm{d}y}{\mathrm{d}t} = Ay + f(t, y),\ y(0) = y_0,\ y \in \mathbb{C}^{n\times n},\ A \in \mathbb{C}^{n\times n}.$$

The idea for this problem is to solve the linear part exactly by the matrix exponential and to integrate the remaining nonlinear part by an explicit scheme.

Here, the so-called φ-functions come into play, which are closely related to the exponential function.

Definition 1.1.4 *For $k \geq 0$, we defined the φ-functions*

$$\varphi_k(z) = \sum_{j=0}^{\infty} \frac{z^j}{(j+k)!}.$$

We note that $\varphi_0 = \exp$. Simple examples of this type of problems are

$$\begin{aligned} \frac{\mathrm{d}y}{\mathrm{d}t} &= Ay,\ y(0) = y_0 \Rightarrow y(t) = \exp(tA)y_0, \\ \frac{\mathrm{d}y}{\mathrm{d}t} &= Ay + b,\ y(0) = 0 \Rightarrow y(t) = t\varphi_1(tA)b, \\ \frac{\mathrm{d}y}{\mathrm{d}t} &= Ay + ct,\ y(0) = 0 \Rightarrow y(t) = t^2\varphi_2(tA)c. \end{aligned}$$

Other examples

We have seen the two most popular functions of matrices, but a lot of functions play a central role in different problems. Among them we can cite the logarithm, the cosine and the sine (second order differential equation), the sign function, the square root and Markov functions.

1.1.3 Computational aspects

It is not necessarily a good idea to stick to one of the definitions of matrix function given previously when it comes to numerically compute a matrix function $f(A)$. We will briefly describe several numerical approaches having their advantages in different situations, basically depending on spectral properties of A, on the dimension and sparsity of A and on the function f. A good reference for the computation of $f(A)$ is [68].

Decomposition of the matrix

A wide class of methods is based on exploiting the fact that if $A = ZBZ^{-1}$ then we have the relation $f(A) = Zf(B)Z^{-1}$. The idea is to factor A with B of a form that allows easy computation of $f(B)$. For example, if $A \in \mathbb{C}^{n\times n}$ is diagonalizable and $\Lambda(A) = \{\lambda_1, \ldots, \lambda_n\}$, we have the representation $A = ZDZ^{-1}$ where $D = diag(\lambda_1, \ldots, \lambda_n)$ is diagonal. If A is not diagonalizable, we can in theory evaluate $f(A)$ with the Jordan canonical form. However, the Jordan form cannot be reliably computed. This approach is recommended only if the matrix Z is well conditioned. An important class of matrices for which this happens is the class of normal matrices which are unitary diagonalizable: $A = UDU^*$ where U is unitary and D is diagonal. In this case the preceding method is feasible if the diagonalization can be computed efficiently. If we want to restrict to unitary

transformation, we recall that for every matrix there is the Schur decomposition $A = QSQ^*$ where Q is unitary and S is upper triangular. Computation of a Schur decomposition is achieved with perfect backward stability by the QR algorithm. In [30], we can find the Schur-Parlett algorithm which computes $f(S)$. Once the Schur decomposition is done, the idea is to reorder S to $\tilde{S}$ in a block form, then to compute the diagonal blocks and use a Parlett recurrence to compute the rest of $f(\tilde{S})$, and finally the unitary similarity transformations from the Schur decomposition and the reordering are applied.

Rational approximation

An other possibility is to obtain a good approximation r of f (in a sense to define) and then to compute $r(A)$. Polynomial approximations for a function often require a quite high degree of the approximating polynomial in order to achieve a reasonable quality of approximation. Rational approximations typically obtain the same quality with substantially fewer degrees of freedom. Moreover we can use the important property that rational functions can be expressed as partial fraction expansions. We can cite two main classes of rational approximations. The first is the best L^∞ (or minimax, or Chebyshev) approximation of f on a compact. If the compact is an interval of $\mathbb{R}$, the best L^∞ rational approximation can be constructed using the Remez algorithm [93]. In the complex case it is far more complicated. The second class is the Padé approximation [7]. In this case we can work with complex functions. Padé approximants of some important functions are known explicitly, sometimes in several representations. The theory of Padé approximation is very well developed and is very attractive in certain cases.

Quadrature rules

Using the Cauchy integral definition

$$f(A) = \frac{1}{2i\pi} \int_\Gamma f(\xi)(\xi - A)^{-1} \mathrm{d}\xi,$$

we can apply a quadrature rule:

$$f(A) \simeq \sum_{j=1}^{p} w_j f(t_j)(t_j - A)^{-1}$$

to get an approximation [31]. For specific functions, other integral representation may be used. In the quadrature framework, the major computational cost is usually due to the inversion of several matrices.

Specific functions

In the case of the exponential we can refer to the famous papers [84] and [85] (cited more than a thousand times). They propose nineteen ways to compute

the exponential in [84] and added a twentieth approach in [85] which use Krylov spaces. We also rmention the reference [68] where we can find specific techniques for the logarithm, the exponential, the sign function, the square root and cosine and sine.

1.1.4 Functions of matrices times a vector

While the appearance of $f(A)$ in a formula may be natural and useful from a theoretical point of view, in practice, it does not always mean that it is necessary or desirable to compute $f(A)$. In general (as seen for the linear system or for differential equations), it is not $f(A)$ that is required, but rather its action on a vector b: $f(A)b$. Moreover, if A is sparse, then $f(A)$ may be dense and for large dimensions, it may be too expensive to compute or store $f(A)$, while computing and storing $f(A)b$ may be feasible. So we are interested in computing the vector $f(A)b$ without explicitly computing $f(A)$. We have seen in the definition using interpolation polynomials, that $f(A) = p_{f,A}(A)$ where $p_{f,A}$ is a Hermite interpolating polynomial determined by the values of f on the spectrum of A. The degree of $p_{f,A}$ is at most equal to the degree of the minimal polynomial ψ_A of A minus one, and it may be greater than necessary in order to produce $f(A)b$. Indeed the good notion here is the minimal polynomial with respect to b, denoted by $\psi_{A,b}$, which is the unique monic polynomial of lowest degree such that $\psi_{A,b}(A)b = 0$. Denoting by $\{\lambda_1, \ldots, \lambda_s\}$ the distinct eigenvalues of A and n_j their corresponding index, we set

$$\psi_{A,b}(z) = \prod_{j=1}^{s} (z - \lambda_j)^{l_j}$$

with $0 \leq l_j \leq n_j$. The degree of the minimal polynomial of b with respect to A is often called the grade of b. We say that f is defined on the spectrum of A with respect to b if the values $f^{(j)}(\lambda_k)$ exist for $k = 1 : s$ and $j = 0 : l_k - 1$. Now we can define $f(A)b$ [68, chap 13].

Definition 1.1.5 (Polynomial interpolation) *Let f be defined on the spectrum of A with respect to b. Then*

$$f(A)b = p_{f,A,b}(A)b$$

where $p_{f,A,b}$ is the unique polynomial of degree less than $\deg(\psi_{A,b})$ *that satisfies the interpolation conditions*

$$p^{(j)}_{f,A,b}(\lambda_k) = f^{(j)}(\lambda_k), \text{ for } j = 0 : l_k - 1 \text{ and } k = 1 : s.$$

We will see that Krylov subspace methods are suitable to approximate $f(A)b$ efficiently and have been extensively used to this purpose, due to their favourable computational and approximation properties. The idea is to project the problem onto some Krylov subspace of smaller dimension where we can use standard algorithms for dense matrices of moderate size. We will see this in the following chapters.

1.2 Krylov spaces

This section is devoted to the presentation of Krylov spaces. Krylov spaces have significant advantages like low memory requirements and good approximation properties, which make them very popular. They are widely used in applications throughout science and engineering. In 1931 Krylov published the paper [73] in which he considered the problem of computing eigenvalues of a square matrix, and introduced those subspaces which now bear his name.

1.2.1 Definitions

Definition 1.2.1 *Given a Hilbert space $\mathcal{H}$, a bounded operator $A \in \mathcal{B}(\mathcal{H})$ and a vector $b \in \mathcal{H}$, we define the Krylov subspaces by*

$$\mathcal{K}_m(A, b) = span\{b, Ab, \ldots, A^{m-1}b\}.$$

A first property is that $\mathcal{K}_m(A, b)$ is the subspace of all vectors that can be written as $x = p(A)b$, where p is a polynomial of degree not exceeding $m - 1$. With increasing order m, polynomial Krylov spaces are nested subspaces of $\mathcal{H}$. If there is no room for ambiguity, we will often write $\mathcal{K}_m$ instead of $\mathcal{K}_m(A, b)$. In this thesis we use the following notations for sets of polynomials.

1. Π_m denotes the set of polynomials of degree at most m,
2. Π_m^α denotes the set of all polynomials p of degree at most m such that $p(\alpha) = 1$,
3. Π_m^∞ denotes the set of monic polynomials of degree at most m.

The key role for the numerical behaviour of a Krylov method is played by the choice of the bases. The Krylov basis $\{b, Ab, \ldots, A^{m-1}b\}$ is not very attractive from a numerical point of view, since the vector $A^j b$ points more and more in the direction of the dominant eigenvector for increasing j (power method), and hence the basis vectors become dependant in finite precision arithmetic. We then need to construct a better basis. Arnoldi's algorithm [3] builds an orthonormal basis which, in exact arithmetic, spans the Krylov subspace. It consists of a modified version of the Gram-Schmidt method to compute an increasing orthonormal system spanning the Krylov spaces. Arnoldi's algorithm gives the following polynomial Krylov decomposition

$$AV_m = V_m H_m + h_{m+1,m} v_{m+1} e_m^* = V_{m+1} \underline{H_m}, \tag{1.1}$$

where $V_m = [v_1 \ldots v_m]$ has orthonormal columns, e_m denotes the m-th column of the identity matrix, $\underline{H_m}$ is a Hessenberg matrix, and $H_m = A_m = V_m^* A V_m$ is obtained from $\underline{H_m}$ by deleting its last row. When A is a Hermitian matrix (or a self-adjoint operator), the Arnoldi process reduces to the more economical Lanczos process [77]. In this case, H_m reduces to a tridiagonal matrix.

Definition 1.2.2 *Given a Hilbert space $\mathcal{H}$, a bounded operator $A \in \mathcal{B}(\mathcal{H})$, a vector $b \in \mathcal{H}$, and a polynomial $q_{m-1} \in \Pi_{m-1}$ which has no zeros in the spectrum $\Lambda(A)$, we define the Krylov subspace of order m associated with (A, b, q_{m-1}) by*

$$\mathcal{Q}_m = \mathcal{Q}_m(A, b) = q_{m-1}(A)^{-1}\mathcal{K}_m(A, b)$$

When $q_{m-1} = 1$, we obtain the standard polynomial Krylov subspaces. It obviously suffices to consider monic polynomials q_{m-1} only. For computations it is convenient to have nested spaces $\mathcal{Q}_1 \subset \mathcal{Q}_2 \subset \dots$ Such nested spaces are obtained if the polynomials q_{m-1} divides q_m for each m. For a given sequence of poles $\{\xi_1, \xi_2, \dots\} \subset \overline{\mathbb{C}} \setminus \Lambda(A)$, we define

$$q_{m-1}(z) = \prod_{\substack{j=1 \\ \xi_j \neq \infty}}^{m-1} (z - \xi_j) \tag{1.2}$$

once and for all. If all $\xi_j = \infty$, then $\mathcal{Q}_m = \mathcal{K}_m$ is a polynomial Krylov subspace ($q_{m-1} = 1$). The space $\mathcal{Q}_m$ is (theorically) independent of the particular ordering of the poles, although in finite precision it can play a crucial role.

Definition 1.2.3 *By M we denote the smallest integer such that $\mathcal{Q}_{M-1} \subsetneq \mathcal{Q}_M = \mathcal{Q}_{M+1}$. M is called the invariance index. If there exists no such integer we set $M = \infty$.*

In the polynomial case, this means that $\mathcal{K}_M$ is A-invariant. We have the following basic properties of rational Krylov spaces [64, Lemma 4.2].

Lemma 1.2.4 *There holds*

1. $\mathcal{Q}_m(A, b) = \mathcal{K}_m(A, q_{m-1}(A)^{-1}b)$.
2. $b \in \mathcal{Q}_m(A, b)$.
3. $\dim(\mathcal{Q}_m(A, b)) = \dim(\mathcal{K}_m(A, b)) = \min(m, M)$.

If q_{m-1} is defined as in (1.2) and if $M < \infty$, then

$$span(b) = \mathcal{Q}_1 \subset \mathcal{Q}_2 \subset \cdots \subset \mathcal{Q}_M = \mathcal{K}_M,$$

otherwise

$$span(b) = \mathcal{Q}_1 \subset \mathcal{Q}_2 \subset \dots$$

Depending on the sequence $\{\xi_j\}$, various special cases of rational Krylov spaces exist. The two most famous methods are the Shift and Invert method where $\xi_j = \xi$ is fixed [45] and the so-called extended Krylov subspace method [37], which is equivalent to the rational Arnoldi method with cyclic pole sequence $\xi_{2j} = \infty$ and $\xi_{2j+1} = 0$. There also exists block versions of rational Krylov spaces which will not be treated here, further details can be found in the references [43, 51, 67, 97, 107] The choice of optimal poles is a difficult problem which has not been solved yet, but a lot of works have been done to find them [38, 39, 53, 65].

1.2.2 Rayleigh-Ritz quotient

Let $\mathcal{H}$ denote a Hilbert space and suppose that the vectors $v_1, \ldots, v_m$ form a basis of an m-dimensional subspace $\mathcal{V}_m \subset \mathcal{H}$. We define the bounded linear operator V_m

$$\begin{array}{rccl} V_m: & \mathbb{C}^m & \to & \mathcal{H} \\ & x & \mapsto & V_m x = x_1 v_1 + \cdots + x_m v_m \end{array}.$$

We define by the same letter V_m the corresponding quasi-matrix [111, lecture 5] $V_m = [v_1, \ldots, v_m]$ with columns v_j. We will consider arbitrary bases (not only orthonormal), and to work with such bases, we need the notion of Moore-Penrose inverse [21, chap 9].

Definition 1.2.5 *Let T be an operator on $\mathcal{H}$. The Moore-Penrose inverse $T^\dagger$ of T is defined by four criteria*

1. $TT^\dagger T = T$
2. $T^\dagger T T^\dagger = T^\dagger$
3. $(TT^\dagger)^* = TT^\dagger$
4. $(T^\dagger T)^* = T^\dagger T$.

Since $\mathcal{V}_m$ is finite dimensional (and therefore closed), there exists a unique Moore-Penrose inverse $V_m^\dagger : \mathcal{H} \to \mathbb{C}^m$ and $V_m V_m^\dagger$ is the orthogonal projection operator onto $\mathcal{V}_m$. Moreover, since V_m has full rank, we have two more useful properties of the Moore-Penrose inverse:

$$(V_m S)^\dagger = S^{-1} V_m^\dagger \quad \text{for every invertible } S \in \mathbb{C}^{m\times m},$$

$$V_m^\dagger V_m = I_m \quad \text{where } I_m \text{ denotes the } m \times m \text{ identity matrix.}$$

Definition 1.2.6 *For a given quasi-matrix $V_m = [v_1, \ldots, v_m]$ of full column rank, the Rayleigh quotient for (A, V_m) is defined as*

$$A_m = V_m^\dagger A V_m \in \mathbb{C}^{m\times m}.$$

The eigenvalues of the Rayleigh quotient $A_m = V_m^\dagger A V_m$ are called the Ritz values.

The properties of the Rayleigh quotient and the Ritz values do not depend on the basis since if V_m and U_m are bases of $\mathcal{V}_m$, then the Rayleigh quotients for (A, V_m) and (A, U_m) are similar matrices [64, Lemma 3.3].

For any quasi-matrix $V_m = [v_1, \ldots, v_m]$ whose columns v_j form a basis of $\mathcal{Q}_m$, we can consider the Rayleigh quotients associated with $\mathcal{Q}_m$. Arnoldi's algorithm gives a special case where V_m is an orthonormal basis of $\mathcal{V}_m = \mathcal{K}_m$, and in this particular case we have $V_m^\dagger = V_m^*$ and $A_m = H_m$. We have the following interesting lemma [64, Lemma 4.5].

Lemma 1.2.7 *Let V_m be a basis of $\mathcal{Q}_m(A,b)$, $A_m = V_m^\dagger A V_m$, and let χ_m denote the characteristic polynomial of A_m. Then the following statements hold.*

1. *A_m is nonderogatory.*
2. *$\chi_m(A)q_{m-1}^{-1}(A)b \perp \mathcal{Q}_m(A,b)$.*
3. *χ_m minimizes $||s_m(A)q_{m-1}^{-1}(A)b||$ among all $s_m \in \Pi_m^\infty$.*

1.2.3 Rational Krylov decomposition

We can generalize the relation (1.1) to the rational case in different (equivalent) ways. Let us propose the following which was used for example in [65].

Definition 1.2.8 *A relation*

$$AV_{m+1}\underline{K_m} = V_{m+1}\underline{H_m} \tag{1.3}$$

where $V_{m+1} = [V_m, v_{m+1}]$ has $m+1$ linearly independent columns such that range$(V_{m+1}) = \mathcal{Q}_{m+1}$, range$(V_m) = \mathcal{Q}_m$, *$\underline{K_m} \in C^{(m+1)\times m}$, $\underline{H_m} \in C^{(m+1)\times m}$, and H_m is of rank m, is called a rational Krylov decomposition.*

If the last row of $\underline{K_m}$ contains only zeros we have

$$AV_m K_m = V_{m+1}\underline{H_m} \tag{1.4}$$

and we say that this decomposition is reduced.

Let us collect some useful facts about rational Krylov decompositions which can be found in [64, Lemma 5.6].

Lemma 1.2.9 *With the preceding notations, we have the following properties.*

1. *The matrix $\underline{K_m}$ of (1.3) is of rank m. In particular, the matrix K_m of the reduced rational Krylov decomposition (1.4) is invertible.*
2. *The validity of (1.4) implies $v_{m+1} \in A\mathcal{Q}_m \setminus \mathcal{Q}_m$.*
3. *If (1.4) is an orthonormal decomposition, i.e., $V_{m+1}^* V_{m+1} = I_{m+1}$, then the Rayleigh quotient A_m can be computed as $A_m = H_m K_m^{-1}$.*

The rational Krylov decomposition and their properties are usefull to compute the examples presented in this thesis.

1.2.4 Complements on Ritz values

Approximations of eigenvalues are often computed with variants of the Arnoldi process, for instance, the Matlab function *eigs* uses the Sorensen's implicitely restarted Arnoldi method [110]. So it is of basic importance to understand which eigenvalues of A are well approximated by the Ritz values (see Definition 1.2.6).

In the Hermitian case, there are many results on the behavior of Ritz values. We assume that the eigenvalues of an Hermitian matrix A are $\lambda_1 < \cdots < \lambda_N$

and denote by $\theta_1^{(i)} < \cdots < \theta_i^{(i)}$ the Ritz values at step i. As is well known in the polynomial case, $\theta_k^{(i)}$ decreases to λ_k as i increases, with equality for $i = N$ [116]. There is an interlacing property which states that between two Ritz values, there is at least one eigenvalue of A ([90, Theorem 10.1.2] for the polynomial case and [18, Theorem 3.1] for the rational case). Early a priori upper bounds for the distance of a polynomial Ritz value to an eigenvalue have been given, most notably, by Kaniel [72], Paige [89], and Saad [99]. Those bounds lead to the well-known Kaniel-Page-Saad estimate for extremal eigenvalues [55], which was generalized in [18, Lemma 2.3]. In [116] the authors studied the behavior of polynomial Ritz values when two eigenvalues are very close and gave a lot of examples.

More involved results using potential theory are given in [75] and [10] for the Lanczos method, which explains the rule of thumb stated in [9]: "the Lanczos iteration tends to converge to eigenvalues in regions of too little charge for an equilibrium distribution". Those results are theoretical and asymptotic. For rational Ritz values, in [18], the authors characterize (again in an asymptotic sense) which eigenvalues are approximated and studied the rate of convergence of Ritz values to eigenvalues.

In contrast, few results about the Ritz values of non Hermitian matrices are known beyond their containment within the numerical range. Using the cartesian decomposition of the matrix ($A = H + iS$, with H hermitian and S skew-hermitian) the authors in [25] gave lower and upper bounds for the real part of the polynomial Ritz values. We can also find some results in [22], [26] or [83], but the general case remains a challenging problem.

1.3 Rayleigh-Ritz methods

Many applications in science and engineering require the evaluation of expressions of the form $f(A)b$ where A is a large (sparse) matrix and f is a nonlinear function. Let A be a bounded linear operator on a complex Hilbert space $\mathcal{H}$ and let f be a complex valued function such that $f(A)$ is defined. Let a vector $b \in \mathcal{H}$ be given. Our aim is to obtain an approximation for $f(A)b$ from a Krylov subspace $\mathcal{Q}_m \subset \mathcal{H}$ of small dimension m, while avoiding the explicit evaluation of $f(A)$, which is usually unfeasible or even impossible.

Definition 1.3.1 *Let V_m be a basis of $\mathcal{Q}_m$ and denote by A_m the Rayleigh-Ritz quotient for (A, V_m). Provided that $f(A_m)$ exists, the Rayleigh-Ritz approximation for $f(A)b$ from $\mathcal{Q}_m$ is defined as*

$$f_m = V_m f(A_m) V_m^\dagger b.$$

If we use the (rational) Arnoldi algorithm [64, Section 5.1] to obtain an orthogonal basis of $\mathcal{Q}_m(A, b)$, then we call f_m the Arnoldi Rayleigh-Ritz approximation.

To justify this definition note that f_m is independent of the choice of the basis V_m [64, Lemma 3.3]. For a given space $\mathcal{Q}_m$ a method for obtaining a Rayleigh-Ritz approximation is referred to as a Rayleigh-Ritz method. Note that for the Arnoldi Rayleigh-Ritz approximation, $V_m^\dagger = V_m^*$ since V_m has orthonormal columns ($V_m^* V_m = I$).

Now let us prove that the Rayleigh-Ritz approximation is exact for rational functions [64, Lemma 4.6]. This exactness property is well known for polynomial approximations [100].

Proposition 1.3.2 (Exactness) *Let V_m be a basis of $\mathcal{Q}_m$ and $A_m = V_m^\dagger A V_m$ the Rayleigh quotient for (A, V_m). Then for every rational function $\tilde{r}_m = p_m/q_{m-1} \in \Pi_m/q_{m-1}$ there holds*

$$V_m V_m^\dagger \tilde{r}_m(A)b = V_m \tilde{r}_m(A_m) V_m^\dagger b.$$

In particular, for every rational function $r_m = p_{m-1}/q_{m-1} \in \Pi_{m-1}/q_{m-1}$ there holds

$$r_m(A)b = V_m r_m(A_m) V_m^\dagger b,$$

i.e., the Rayleigh-Ritz approximation for $r_m(A)b$ is exact (provided that $r_m(A_m)$ is defined).

Proof : Recall that $V_m V_m^\dagger$ is the orthogonal projection onto $\mathcal{Q}_m$. We start by proving the theorem in the polynomial case. It suffices to show that for all $j \leq m$

$$V_m V_m^\dagger A^j b = V_m A_m^j V_m^\dagger b.$$

The proof is by induction on the monomials A^j. This property is clear for $j = 0$. Assume that it is true for some $j \leq m-1$, then

$$\begin{aligned} V_m V_m^\dagger A^{j+1} b &= V_m V_m^\dagger A A^j b = V_m V_m^\dagger A V_m V_m^\dagger A^j b \quad (\text{since } A^j b \in \mathcal{K}_m) \\ &= V_m V_m^\dagger A V_m A_m^j V_m^\dagger b \quad (\text{by induction hypothesis}) \\ &= V_m A_m^{j+1} V_m^\dagger b. \end{aligned}$$

The rational case is a direct consequence of the polynomial case. Setting $c = q_{m-1}(A)^{-1} b$ we know that $\mathcal{Q}_m(A, b) = \mathcal{K}_m(A, c)$, and we obtain

$$\begin{aligned} V_m V_m^\dagger \tilde{r}_m(A)b &= V_m V_m^\dagger p_m(A) q_{m-1}(A)^{-1} b = V_m V_m^\dagger p_m(A) c \\ &= V_m p_m(A_m) V_m^\dagger c. \end{aligned}$$

Or $b = q_{m-1}(A)c = V_m q_{m-1}(A_m) V_m^\dagger c$, which implies

$$V_m^\dagger c = q_{m-1}(A_m)^{-1} V_m^\dagger b, \tag{1.5}$$

and thus

$$V_m V_m^\dagger \tilde{r}_m(A)b = V_m p_m(A_m) q_{m-1}(A_m)^{-1} V_m^\dagger b = V_m \tilde{r}_m(A_m) V_m^\dagger b.$$

□

Remark 1.3.3 *It is possible that $r_m(A)$ is defined but $r_m(A_m)$ is not.*

The next property states an interpolation characterization of the Rayleigh approximation: it is mathematically equivalent to interpolating f over the Ritz values [64, Theorem 4.8].

Proposition 1.3.4 (Characterization) *Let V_m be a basis of $\mathcal{Q}_m$ and $A_m = V_m^\dagger A V_m$ the Rayleigh quotient for (A, V_m). Let f be a function such that $f(A_m)$ is defined. Then*

$$f_m = V_m f(A_m) V_m^\dagger b = r_m(A)b,$$

where $r_m = p_{m-1}/q_{m-1} \in \mathcal{P}_{m-1}/q_{m-1}$ interpolates f at the Ritz values $\Lambda(A_m)$.

Proof : The polynomial case is easy to prove. Indeed, by definition of a function of a matrix, we have $f(A_m) = p_{m-1}(A_m)$, with p_{m-1} interpolating f at the Ritz values (in the Hermite sense). Hence

$$f_m = V_m f(A_m) V_m^\dagger b = V_m p_{m-1}(A_m) V_m^\dagger b = p_{m-1}(A)b,$$

where in the last equality we have used the exactness property (Proposition 1.3.2).

In the rational case, setting $c = q_{m-1}(A)^{-1}b$ and $\tilde{f} = fq_{m-1}$, we have $f(A)b = \tilde{f}(A)c$. Using the fact that $V_m^\dagger c = q_{m-1}(A_m)^{-1} V_m^\dagger b$ (see (1.5)), we obtain

$$\begin{aligned} f_m &= V_m f(A_m) V_m^\dagger b = V_m \tilde{f}(A_m) q_{m-1}^{-1}(A_m) V_m^\dagger b \\ &= V_m \tilde{f}(A_m) V_m^\dagger c \\ &= p_{m-1}(A)c = p_{m-1}(A) q_{m-1}(A)^{-1} b, \end{aligned}$$

where p_{m-1} interpolates $\tilde{f}$ at the Ritz values. This implies that $f_m = r_m(A)b$ where $r_m = p_{m-1}/q_{m-1}$ interpolates f at the Ritz values.

□

Now an interesting question is to determine when the approximation becomes exact. Recall that we denote by M the invariance index (Definition 1.2.3). $M < \infty$ if and only if there exists a unique polynomial $\psi_{A,b} \in \Pi_m^\infty$ such that $\psi_{A,b}(A)b = 0$. In this case the Rayleigh-Ritz approximation $f_M = f(A)b$ is exact [64, Lemma 3.11]. In practice one expects usually to terminate much before the exact termination property takes over.

Example 1.3.5 *In general, the Rayleigh-Ritz approximation differs from $f(A)b$ until the invariance index M is reached, even if $f(A)b \in \mathcal{K}_m$ for $m < M$. For example we take $A = diag(1, 2, 3, 4, 5)$, $b = (1, 1, 1, 1, 1)$ and $f(x) = xcos(2\pi x)$. It is clear that $f(A) = A$, and then $f(A)b = Ab \in \mathcal{K}_2$. We collect the approximations f_m in the following array*

$m=1$	$m=2$	$m=3$	$m=4$	$m=5$
3	−0.8582	0.3216	1.0211	1.0000
3	−1.7164	2.0522	0.0201	2.0000
3	−2.5746	3	−0.9809	3.0000
3	−3.4329	3.1650	0.0402	4.0000
3	−4.2911	2.5472	5.1056	5.0000

which shows that that we do not have the exact approximation before f_5.

Example 1.3.6 *We can obtain $f_m = f(A)b$ before the invariance index. Indeed, taking the same A and b with $f(x) = x^2$, we obtain $f(A)b = f_3 \in \mathcal{K}_3$ before the invariance index. We give the approximations f_m in the following array*

$m = 1$	$m = 2$	$m = 3$
9	-1	1
9	5	4
9	11	9
9	17	16
9	23	25

If $M = \infty$, we set $\mathcal{K}_\infty = \overline{span\{b, Ab, \dots\}}$. There are two possibilities:

1. $\mathcal{K}_\infty = \mathcal{H}$, i.e. A is cyclic for b.
2. $\mathcal{K}_\infty \neq \mathcal{H}$.

Using the holomorphic functional calculus or the continuous functional calculus for normal operators we can obtain that $f(A)b = \lim_{k\to\infty} p_k(A)b$ for $p_k \in \Pi_k$ and thus $f(A)b \in \mathcal{K}_\infty$.

Chapter 2

Krylov methods for linear systems

One of the most powerful tool for solving large and sparse systems of linear algebraic equations is a class of iterative methods called Krylov subspace methods. The use of the Krylov subspaces in iterative methods for linear systems is even counted among the top 10 algorithmic ideas of the 20th century [27]. Convergence analysis of these methods is not only of a great theoretical importance but it can also help to answer practically relevant questions about improving the performance of these methods. As we will see, the question about the convergence behavior leads to complicated nonlinear problems. Despite intense research efforts, these problems are not well understood in the general case. Linear systems are useful for computation of functions of matrices and appear in the algorithms to compute bases of rational Krylov spaces.

After a review on projection methods, we present an application of the Krylov spaces to the resolution of linear systems. In particular, we present and characterize two methods and give some known linear convergence bounds using different kinds of techniques. In the last section, we discuss the notion of superlinear convergence and state a conjecture which will be proved later.

2.1 Krylov methods

2.1.1 General projection methods

Consider the linear system $Ax = b$ where A is a bounded invertible linear operator on a complex Hilbert space $\mathcal{H}$, or simply a $n \times n$ complex matrix. There are several reasons why to choose an abstract, possibly infinite dimensional setting. Indeed, for the algorithms to be considered, it makes no essential difference whether or not the underlying spaces have finite dimension. Moreover the dimension of the search space (where the approximation lies) to be used in practical applications is always much smaller than that of the space $\mathcal{H}$, so, in comparison,

$\dim(\mathcal{H})$ may be considered infinite. Furthermore, a large class of linear systems arises from discretization of operators between infinite dimensional spaces, in which case, a sequence of problems corresponding to a sequence of discretization parameters is the natural object of study, and the later elements of such a sequence inherit many important properties of the infinite-dimensional problem under approximation.

The idea of projection techniques is to extract an approximate solution $\tilde{x}$ to the above problem from a subspace of $\mathcal{H}$. If $\mathcal{S}$ is the finite dimensional search subspace and m is its dimension, then, in general, m constraints must be imposed to be able to extract such an approximation. A typical way is to impose m orthogonality conditions. Specifically, the residual vector $r = b - A\tilde{x}$ is constrained to be orthogonal to m linearly independant vectors. This defines another subspace $\mathcal{C}$ of dimension m. The problem can be written as follows:

$$\text{find } \tilde{x} \in \mathcal{S} \text{ such that } r = b - A\tilde{x} \perp \mathcal{C}.$$

There are two broad classes of projection methods : orthogonal ($\mathcal{S} = \mathcal{C}$) and oblique ($\mathcal{S} \neq \mathcal{C}$). If we wish to exploit the knowledge of an initial guess x_0 to the solution, then the problem should be redefined as follows:

$$\text{find } \tilde{x} \in x_0 + \mathcal{S} \text{ such that } r = b - A\tilde{x} \perp \mathcal{C}$$

Set $\tilde{x} = x_0 + \tilde{u}$, with $\tilde{u} \in \mathcal{S}$ and $r_0 = b - Ax_0$ (initial residual). Existence and uniqueness of $\tilde{x}$ are summarized in the following lemma [42, Proposition 2.2].

Lemma 2.1.1 *Let A, $\tilde{x}$, $\tilde{u}$, x_0, $\mathcal{S}$, $\mathcal{C}$ and r_0 defined as above. Then*

1. *$\exists \tilde{u} \in \mathcal{S}$ such that $r_0 - A\tilde{u} \perp \mathcal{C} \iff r_0 \in A\mathcal{S} + \mathcal{C}^\perp$.*
2. *Such an $\tilde{u}$ is unique if and only if $A\mathcal{S} \cap \mathcal{C}^\perp = \{0\}$.*

We always suppose implicitely that we are in those conditions when we speak about an approximation. If $\mathcal{C} = A\mathcal{S}$ it is clear that there is always a unique approximation. We have the following two well-known optimality results which can be found in [101, Propositions 5.3 and 5.2].

Proposition 2.1.2 *Assume that A is a bounded linear operator and $\mathcal{C} = A\mathcal{S}$. Then $\tilde{x}$ is the result of an oblique projection method onto $\mathcal{S}$ orthogonally to $A\mathcal{S}$ with the starting vector x_0 if and only if*

$$\|b - A\tilde{x}\| = \min_{x \in x_0 + \mathcal{S}} \|b - Ax\|.$$

A need not be invertible in the above proposition, but when A is singular, there may be infinitely many vectors $\tilde{x}$ satisfying the optimality condition. The name MR (Minimum Residual) method comes from this property. For a Hermitian positive definite operator A, we define the A-norm of a vector by $\|x\|_A = \sqrt{x^T Ax}$.

Proposition 2.1.3 *Assume that A is a self-adjoint positive definite operator and $\mathcal{C} = \mathcal{S}$. Then $\tilde{x}$ is the result of an orthogonal method onto $\mathcal{S}$ with the starting vector x_0 if and only if*

$$\|x^* - \tilde{x}\|_A = \min_{x \in x_0 + \mathcal{S}} \|x^* - x\|_A,$$

where x^* denotes the exact solution.

Typically, a new projection step uses a new pair of subspaces and an inital guess equal to the most recent approximation obtained. If no vector of the subspace $\mathcal{S}$ comes close to the exact solution x^*, then it is impossible to find a good approximation. So a question that arises immediately is how good the approximate solution can be? In the particular case when $\mathcal{S}$ is invariant under A, we have the following result [101, Proposition 5.6].

Lemma 2.1.4 *Assume that $\mathcal{S}$ is invariant under A and r_0 belongs to $\mathcal{S}$. Then the approximate solution obtained from any projection method (oblique or orthogonal) onto $\mathcal{S}$ is exact.*

This is a rare occurrence in practice, but the result helps in understanding the breakdown behavior of Krylov methods.

2.1.2 Krylov methods

The Krylov methods are examples of projection methods for solving linear systems $Ax = b$. Krylov spaces are certainly the most widely used spaces in projection methods. A generic Krylov method consist in taking an approximation

$$x_m = x_0 + v_m \in x_0 + \mathcal{S}_m$$

in an affine space of dimension m related to a Krylov subspace. As we have m degree of liberty, we impose m constraints on the residual

$$r_m = r_0 - Av_m \perp \mathcal{C}_m$$

with $\mathcal{C}_m$ a subspace of dimension m also related to a Krylov subspace.

The Krylov subspace methods can be distinguished in four main different classes:

1. The Ritz-Galerkin approach or orthogonal methods (OR methods): construct the $x_m \in x_0 + \mathcal{K}_m$ for which the residual is orthogonal to $\mathcal{K}_m$.
2. The minimum norm residual approach (MR methods): construct the $x_m \in x_0 + \mathcal{K}_m$ for which the residual is minimal.
3. The Petrov-Galerkin approach: construct the $x_m \in x_0 + \mathcal{K}_m$ for which the residual is orthogonal to some other suitable subspace.
4. The minimum norm error approach: determine x_m in $A^T\mathcal{K}_m(A^T, b)$ for which the error norm is minimal.

We are interested in the two first methods. The Ritz-Galerkin approach leads to well-known algorithms such as Conjugate Gradients (CG) or FOM. The minimum norm residual approach leads to methods such as GMRES, MINRES, and ORTHODIR. The choice of a method is a delicate problem. If the matrix A is symmetric positive definite, then the choice is easy: Conjugate Gradients [66]. For other types of matrices the situation is very diffuse. GMRES, proposed in 1986 by Saad and Schultz in [102], is the most robust method, but in terms of work per iteration step it is also relatively expensive. For nice reviews on the different possibilities we refer to [117] and [59].

Using the Drazin inverse, we can also use Krylov spaces to solve linear systems with singular operators [71, 87].

2.1.3 OR and MR methods

We denote by x_m^{OR} and r_m^{OR} the approximation and the residual of the orthogonal method at step m. Respectively we denote by x_m^{MR} and r_m^{MR} the approximation and the residual of the minimum residual method at step m. Let us set some notations to be clear:

$$\begin{cases} x_m^{OR} = x_0 + v_m^{OR} \in x_0 + \mathcal{K}_m(A, r_0) \\ r_m^{OR} = r_0 - Av_m^{OR} = r_0 - h_m^{OR} \perp \mathcal{K}_m(A, r_0) \end{cases}$$

and

$$\begin{cases} x_m^{MR} = x_0 + v_m^{MR} \in x_0 + \mathcal{K}_m(A, r_0) \\ r_m^{MR} = r_0 - Av_m^{MR} = r_0 - h_m^{MR} \perp A\mathcal{K}_m(A, r_0). \end{cases}$$

Remark 2.1.5 *Assume that A is invertible. The polynomial Rayleigh-Ritz approximation (Definition 1.3.1) for $f(z) = z^{-1}$ coincides with the approximation obtained by the OR method applied to the operator equation $Ax = b$ for an initial guess $x_0 = 0$. Indeed, $x_m^{OR} = V_m y_m^{OR} \in \mathcal{K}_m$ and $r_m^{OR} = b - AV_m y_m^{OR} \perp \mathcal{K}_m$ implies that $y_m^{OR} = A_m^{-1} V_m^* b$, and thus we obtain $x_m^{OR} = V_m A_m^{-1} V_m^* b$.*

We have seen in Propositons 2.1.2 and 2.1.3 that

$$h_m^{MR} = P_{A\mathcal{K}_m} r_0$$

is the orthogonal projection of r_0 onto $A\mathcal{K}_m$, and

$$h_m^{OR} = P_{A\mathcal{K}_m}^{\mathcal{K}_m} r_0$$

is the oblique projection of r_0 onto $A\mathcal{K}_m$ orthogonal to $\mathcal{K}_m$.

It is clear that $r_m^{OR} = \varphi_m^{OR}(A) r_0$ and $r_m^{MR} = \varphi_m^{MR}(A) r_0$ for some polynomials in Π_m^0. Those two polynomials solve minimization problems. For the MR method we have [101, Lemma 6.31]

$$\|\varphi_m^{MR}(A) r_0\| = \min_{p \in \Pi_m^0} \|p(A) r_0\|.$$

For the OR method we have a link with an other normalization problem [101, Lemma 6.28]

$$\|\varphi_m^{OR}(A)e_0\|_A = \min_{p\in\Pi_m^0} \|p(A)e_0\|_A.$$

Those two polynomials can be characterized by their zeros which are explicitely known. If the m-th OR approximation is defined, we have [57, Theorem 3.1]

$$\varphi_m^{OR}(z) = \prod_{j=1}^{m}(1-\frac{z}{\theta_j}),$$

where the θ_j designate the Ritz values (Definition 1.2.6). For the MR polynomial, let us define the harmonic Ritz values.

Definition 2.1.6 *Let W_m designate a quasi-matrix whose columns form a basis of $A\mathcal{K}_m$. We define the harmonic Ritz values $\hat{\theta}_j$ as the inverses of the eigenvalues of the matrix $W_m^\dagger A^{-1} W_m$.*

In other words, harmonic Ritz values are the inverses of the eigenvalues of the Rayleigh quotient of (A^{-1}, W_m). If an eigenvalue of $W_m^\dagger A^{-1} W_m$ is zero, we set the corresponding harmonic Ritz-value equal to infinity. In this case the MR method stagnates and the residual polynomial is not changed: $r_m^{MR} = r_{m-1}^{MR}$. Denoting by $\{\hat{\theta}_j\}_{j=1}^m$ the harmonic Ritz values, we have [57, Theorem 7.1]

$$\varphi_m^{MR}(z) = \prod_{j=1}^{m}(1-\frac{z}{\hat{\theta}_j}).$$

The OR and MR methods can be viewed in a more abstract way, without reference to the operator, and we cite the paper [42] which analyses the geometry of Krylov spaces to obtain well-known formulae on those two methods. Their approach has the advantage of simplicity and generality: they express the methods by an approximation scheme in a Hilbert space $\mathcal{H}$ based on orthogonal and oblique projection onto a finite dimensional subspace.

We can connect OR residual and MR residual by a projector. Setting $h_m^{OR} = AV_m y_m^{OR}$ and using the fact that r_m^{OR} is orthogonal to $\mathcal{K}_m$, i.e. $V_m V_m^\dagger r_m^{OR} = 0$ where V_m is a quasi-matrix whose columns form a basis of $\mathcal{K}_m$, we obtain

$$V_m V_m^\dagger r_0 = V_m V_m^\dagger h_m^{OR} = V_m A_m y_m^{OR},$$

which implies ($V_m^\dagger V_m = I$ as V_m has linearly independent columns)

$$V_m^\dagger r_0 = A_m y_m^{OR}.$$

If A_m is invertible, then $y_m^{OR} = A_m^{-1} V_m^\dagger r_0$ and we can write

$$r_m^{OR} = (I - AV_m A_m^{-1} V_m^\dagger) r_0 = P_m r_0,$$

where $P_m^2 = P_m$ and $P_m A V_m x = 0$, for every $x \in \mathbb{C}^m$. Now using that $r_m^{MR} = r_0 - AV_m y_m^{MR}$, it is clear that

$$P_m r_m^{MR} = P_m r_0 = r_m^{OR}.$$

We refer to [42, Theorem 3.4] for others well-known relations between the two methods.

2.2 Linear convergence bounds

In this section we give a brief overview of the most well-known linear convergence bounds for the error of the OR methods and for the residuals of the MR methods. In the analysis of the convergence of Krylov methods we usually encounter the value (defined for example in [13])

$$E_n(S) = \min \left\{ \frac{\|p\|_S}{|p(0)|}, \quad p \in \Pi_n, \quad \forall \lambda \in \Lambda(A) \setminus S : \quad p(\lambda) = 0 \right\}. \tag{2.1}$$

The eigenvalues outside S, denoted by Λ_0, will be called outliers and will play a particular role in our analysis. Obviously we have the inequalities

$$E_n(\Lambda(A)) \leq E_n(S \cup \Lambda_0) \leq E_n(S), \tag{2.2}$$

which have been used for example in [24] or [115] in order to derive a CG convergence bound taking into account few outliers, where typically S is the convex hull of the remaining eigenvalues.

If S contains the spectrum (no outliers), we find the classical quantity

$$E_n(S) = \min_{p \in \Pi_n^0} \max_{\lambda \in S} |p(\lambda)|.$$

If $S \supset \Lambda(A)$ is a compact set in the complex plane with $0 \notin S$, it is known that [35, Theorem 1]

$$E_n(S) \geq \rho_S^n, \tag{2.3}$$

where $\rho_S = \exp(-g_S(0, \infty))$ is usually called the estimated asymptotic convergence factor, and where $g_S(., \infty)$ is the Green function for the complement of S with pole at ∞ (Definition A.1.4). If S is connected we have a link with the conformal map ϕ_S from the exterior of S to the exterior of the closed unit disk that satisfies $\phi_S(\infty) = \infty$ and $\phi_S'(\infty) > 0$, via the formula $g_S(0, \infty) = \log |\phi_S(0)|$ (Eqn. (A.4)). The sharpness of this minimization problem is not easy to handle, except for certain special sets such as when S is an interval or a disk in the complex plane. Using tools from potential theory it is known that [35, Equation 4]

$$\lim_{n \to \infty} E_n(S)^{\frac{1}{n}} = \rho_S \leq 1,$$

which means that asymptotically the bound is sharp in the sense of the n-th root. For special sets, we can find

$$E_n(S) \leq C \rho_S^n.$$

For example, C is 1 in the case of a disk by Zarantonello's lemma [101, Proposition 6.26], in the case of a segment $C \leq 2$ [101] and when S is convex, Eiermann [41] obtained that $C \leq \frac{2}{1-\rho_S^n}$, by using an approximation theorem on Faber polynomials obtained by Kovari and Pommerenke [74, Theorem 2]. In the case of polygons, the rate can be computed using Driscoll's Schwarz-Christoffel Matlab toolbox for numerical conformal mapping (for more details see [33]). We can note that $\rho_S < 1$ except in the case when S completely surrounds the origin in the sense of separating it from the point at infinity.

The concept of K-spectral sets

Let us introduce the concept of K-spectral sets (see for instance [91]).

Definition 2.2.1 *We say that a set X in the complex plane is K-spectral for an operator A between Hilbert spaces, if it is closed, it contains the spectrum of A, and for every complex-valued bounded rational function f on X we have $\|f(A)\| \leq K\|f\|_X$. In the case we can take $K = 1$, we say that X is a spectral set for A.*

In our case, if $S \cup \Lambda_0$ is $C_{S\cup\Lambda_0}$-spectral, then the inequalities

$$\min_{p\in\Pi_n^0} \|p(A)\| \leq C_{S\cup\Lambda_0} E_n(S \cup \Lambda_0) \leq C_{S\cup\Lambda_0} E_n(S)$$

are satisfied. A lot of work has been done on the theory of K-spectral sets, let us give some examples which can all be found in [6].

It is a very well-known fact that if A is a normal operator, then the spectrum of A is a spectral set. This implies in our context that for normal operators we have

$$\min_{p\in\Pi_n^0} \|p(A)\| \leq E_n(\Lambda(A)) \leq E_n(S),$$

and thus in terms of residuals of MR methods, this gives

$$\frac{\|r_n^{MR}\|}{\|r_0\|} \leq E_n(\Lambda(A)) \leq E_n(S).$$

The spectrum of a matrix A is K-spectral if and only if A is diagonalizable, and in this case, we can take K equal to the condition number of the matrix of eigenvectors.

The closed disk $\{z \in \mathbb{C}/|z-a| \leq r\}$ is spectral for A if and only if $\|A-a\| \leq r$.

We will see that the numerical range (see (2.5)) is $(1+\sqrt{2})$-spectral (Crouzeix's theorem), and that the ϵ-pseudospectrum (see (2.6)) is K-spectral for a constant K that depends on ϵ.

2.2.1 OR methods

Let us suppose that A is a Hermitian positive definite matrix so that we can use Proposition 2.1.3. Denoting by e_n the error after n steps, a standard convergence analysis leads to

$$\frac{\|e_n\|_A}{\|e_0\|_A} \leq \min_{p\in\Pi_n^0} \frac{\|p(A)e_0\|_A}{\|e_0\|_A} \leq E_n(\Lambda(A)) \leq E_n(S)$$

for each set S that contains the spectrum. Recall that $\|x\|_A = \sqrt{x^T Ax}$ is the A-norm. It is interesting to remark that the inequality $\frac{\|e_n\|_A}{\|e_0\|_A} \leq E_n(\Lambda(A))$ is sharp in the sense that for given A and n, there exists a starting vector such that equality holds [58]. By setting $S = [\lambda_{min}, \lambda_{max}]$ and using scaled and shifted

Chebyshev polynomials we can obtain the most famous result about convergence of OR methods [101, Equation (6.128)]

$$\frac{\|e_n\|_A}{\|e_0\|_A} \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^n \tag{2.4}$$

with κ the condition number of A. We note that for $S = [\lambda_{min}, \lambda_{max}]$ we have [35, Equation (14)]

$$g_S(0,\infty) = \log\left(\frac{\sqrt{\lambda_{max}}+\sqrt{\lambda_{min}}}{\sqrt{\lambda_{max}}-\sqrt{\lambda_{min}}}\right) = \log\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right),$$

and thus $\rho_S = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$, which means that the lower bound (2.3) is sharp up to a factor 2.

When A is not Hermitian positive definite, we can use the algorithm FOM [96, 101], which is mathematically equivalent to the CG method when A is Hermitian positive definite.

2.2.2 MR methods

Despite their popularity, the convergence of MR methods are not completely understood. While the spectrum can in some cases be a good indicator of convergence, it has been shown that in general, it does not provide sufficient information to fully explain the behavior. Understanding the convergence of MR methods which is facilitated by its optimality property is an important step towards convergence analysis for general methods. It can also inform the construction and evaluation of preconditioners for non symmetric problems. We have

$$\|r_n\| = \min_{p\in\Pi_n^0} \|p(A)r_0\| \leq \|r_0\| \min_{p\in\Pi_n^0} \|p(A)\|$$

All the analysis described here will be done on the value

$$\min_{p_n\in\Pi_n^0} \|p_n(A)\|$$

which is frequently the basis for discussions of the convergence of MR methods. The problem of approximating the polynomial that minimizes this value is refered in the litterature as the ideal MR approximation problem [61]. When A is normal, the convergence is governed by the eigenvalues, and we have

$$\frac{\|r_n\|}{\|r_0\|} \leq E_n(\Lambda(A)).$$

If the spectrum is known, then $E_n(\Lambda(A))$ can be computed (at least numerically). However, usually, we do not know exactly the spectrum, we only know that the spectrum is contained in a certain set S. Depending on the informations we have on the spectrum, we obtain more or less accurate information

on the convergence. When A is not normal, the convergence behavior may not be related to the eigenvalues in any simple way, and understanding the convergence in the general nonnormal case still remains a largely open problem. As an example, in [60], the authors showed that for any prescribed sequence of non-increasing residual norms, there exists a class of right hand sides and matrices whose (non zero) eigenvalues can be chosen arbitrarily, giving residual norms that coincide with the given non-increasing sequence.

Linear bounds in the general case

Let us present three familiar convergence bounds for MR methods based on the eigenvalues, the field of values and the pseudospectra.

In the case A is diagonalizable ($A = ZDZ^{-1}$), we can obtain

$$\frac{\|r_n\|}{\|r_0\|} \leq \kappa(Z) E_n(\Lambda(A)),$$

where $\kappa(Z) = \|Z\|\|Z^{-1}\|$ denotes the condition number of Z. This can be generalized for any matrix in terms of the Jordan form, but the Jordan decomposition is numerically unstable.

In [41], Eiermann developed a bound based on the field of values of A (also called numerical range)

$$\mathbb{W}(A) = \{(Ax, x), \|x\| = 1\}. \tag{2.5}$$

This bound was improved in [11] and can be formulated as

$$\frac{\|r_n\|}{\|r_0\|} \leq \frac{2}{1 - \rho_S^{n+1}} \rho_S^n,$$

with $\mathbb{W}(A) \subset S$, $0 \notin S$ and $\rho_S = \frac{1}{|\phi_S(0)|}$. Another bound which can be obtained easily in terms of the numerical range with the help of Crouzeix's theorem [28, 29] is

$$\frac{\|r_n\|}{\|r_0\|} \leq C_{Crouzeix} E_n(\mathbb{W}(A)),$$

where $C_{Crouzeix} \in [2, 1+\sqrt{2}]$ is the constant of Crouzeix which is conjectured to be 2. It is interesting to note that the residual norms decrease strictly monotonically whenever zero is outside the field of values of A [62]. However, in general, no strict monotonicity is guaranteed.

A third type of bounds can be obtained using the notion of ϵ-pseudospectra [44] defined by

$$\Lambda_\epsilon(A) = \{z \in \mathbb{C} / ||(zI - A)^{-1}|| \geq \frac{1}{\epsilon}\}, \tag{2.6}$$

where I designates the identity and can be omited. Indeed, using the Dunford-Taylor integral, we have for any union of Jordan curves Γ containing the spectrum of A and any polynomial p

$$p(A) = \frac{1}{2i\pi} \int_\Gamma p(z)(z - A)^{-1} \mathrm{d}z,$$

which implies that

$$\frac{\|r_n\|}{\|r_0\|} \leq \frac{L(\Gamma)}{2\pi\epsilon} E_n(\Lambda_\epsilon(A)),$$

where $L(\Gamma)$ is the length of Γ.

We can find many linear bounds for those methods in the literature, including some with angles between Krylov spaces in [42].

Linear bounds for MINRES

In the particular case when A is symmetric and indefinite, MINRES has become the leading algorithm among MR methods. If $\Lambda(A) \subset [a, b] \cup [c, d]$ where $a < b < 0 < c < d$, under the constraint $b - a = d - c$, we have the upper bound

$$\frac{\|r_n\|}{\|r_0\|} \leq 2 \left(\frac{\sqrt{|ad|} - \sqrt{|bc|}}{\sqrt{|ad|} + \sqrt{|bc|}} \right)^{[n/2]},$$

where $[n/2]$ denotes the integer part of $n/2$. This can be proved using an appropriate transformation of the intervals and Tchebyshev polynomials [59, Section 3.1]. See also the paper [108] which goes a little further or the recent paper [104].

2.3 Superlinear convergence

All the bounds we proposed before show a linear convergence. But we can expect to see three phases for the actual convergence [88], the sublinear phase (in the early steps the convergence is very rapid), the linear phase (the convergence settles down to a roughly linear rate), and then the superlinear phase (the convergence of the process accelerates again). In practice all phases need not be identifiable, nor need they appear only once and in this order. In the following we are interested by the superlinear convergence.

2.3.1 Notion of superlinear convergence

In practice, the linear bounds given in the preceding section are too pessimistic, and one observes that the speed of convergence of the process improves as the process proceeds, see for instance Figure 2.3. This phenomenon is known as superlinear convergence of the Krylov methods. Superlinear upper bounds for operators of the form $\alpha I + K$, with K a compact operator, can be found in [120] for OR methods and in [86] for MR methods. If the compact operator K is a p-th Schatten class operator [54], which means that its singular values are p-summable, then the bounds are explicit. In the case of Hermitian matrices for OR methods, an explanation for this behavior, in an asymptotic sense, can be found in [12]. In the general case the phenomenon is not well understood yet.

A remark we can do here is that to obtain linear bounds, we always included the spectrum in a larger fixed set S. The approximation is based on the assumption that S is a good approximation of the spectrum, which in this case

means that the optimal polynomial, which is small on $\Lambda(A)$, is also small on S. This assumption may actually be valid for small n. However for larger values of n it may be more efficient for the optimal polynomial to have some of its zeros very close to some of the eigenvalues, without being small in the full set S. Among other things, this can be viewed as responsible for the phenomenon of superlinear convergence often observed in the later stages of a matrix iteration.

Another reason which is responsible for the superlinear convergence is that some eigenvalues are well approximated during the iterations. Indeed, Ritz values turn out to be accurate approximations of some eigenvalues of A [9], [10], [18], [109]. And once the method has found an eigenvalue, the procedure converges as a process in which this eigenvalue is absent [115], [118].

The observations made above are well known in the numerical linear algebra community, see for instance the monographs [46], [88], [101], [108] or the first articles on this phenomenon [4], [5], [109], [115], [118].

2.3.2 Notion of outliers

As already said, at the beginning of the process, the linear bounds are typically accurate, but as the iterations go on, they can become a great overestimation of the error. An idea to study the superlinear convergence is to choose a set S which will depend on the number of eigenvalues which have been found by considering them as outliers. This set will be smaller as the process goes on, it will exclude the outliers (which are known) and contain the rest of the spectrum (not found precisely yet). To do this, we have to understand which eigenvalues are found by Krylov methods (i.e. Arnoldi or Lanczos algorithms). Trefethen and Bau observed [9, lecture 36] the following rule of thumb for Hermitian matrices: Ritz values tend to converge to eigenvalues in regions of "too little charge" for an equilibrium distribution. This means that we need to compare the distribution of the eigenvalues with the equilibrium distribution, and if the eigenvalues are distributed like the equilibrium distribution, then the Lanczos iteration does not find any eigenvalue until the number of iterations is near the size of the matrix. On the other hand, the Lanczos iteration will find very quickly eigenvalues in the region where their distribution is less dense than the equilibirum distribution. It was Kuijlaars [75] who first quantified this heuristic rule of thumb in the case of Hermitian matrices, by considering a sequence of matrices and working in an asymptotic sense, and refined later in [18].

As an example, assume that the eigenvalues of A are located in $[-1, 1]$. The equilibrium measure has density $1/(\pi\sqrt{1-t^2})$ which is infinite at the endpoints. Thus if the eigenvalues are equidistant, the Lanczos method tends to find the eigenvalues near the endpoints [75, Section 4.2], and we observe the superlinear convergence. And if the eigenvalues are the zeros of the Chebyshev polynomials of the first kind, there is no superlinear convergence for $E_n(S)$.

Let us consider the matrix

$$A_N = \begin{pmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix}.$$

The eigenvalues of A_N are given by $\lambda_{j,N} = 2 - 2\cos(\pi j/(N+1))$ leading to the asymptotic eigenvalue distribution given by the equilibrium measure of the interval $[0, 4]$. In this case we do not have superlinear convergence for the quantity $E_n(S)$. For example, in Figure 2.1 on the left there is no superlinear convergence and one has to reach approximately the dimension of the matrix to achieve full precision. On the other hand, for a specific right-hand side b, we may still observe superlinear convergence [14] as confirmed by the plot in Figure 2.1 on the right.

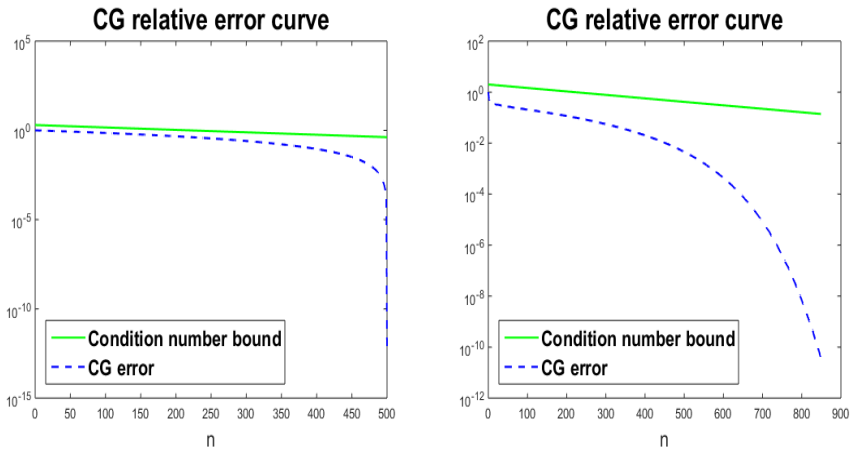

Figure 2.1: *CG relative error curve for the one dimensional Poisson problem discretized on a uniform grid of size* $N = 1000$ *with initial vector* $x_0 = 0$. *On the left* $b = (1, \ldots, 1)^T$ *and on the right* $b_k = (N+1)^2 \sum_{j=1}^{N} 2^{-j} \sin(j\pi k/(N+1))$ *for* $k = 1 : N$.

An isolated outlier has no effect on the asymptotic convergence factor. The reason is that any isolated eigenvalue can be annihilated by a single zero of a polynomial; the rest of the zeros can be devoted to achieving minimal norm on the rest of S. The price to pay to annihilate an outlier grows as it approaches the origin, indeed, we can observe a delay of a number of steps of the convergence on Figure 2.2 on the left. When the outlier is far from the origin it delays the convergence by only few steps. If we consider several outliers the effects are the same, see Figure 2.2 on the right.

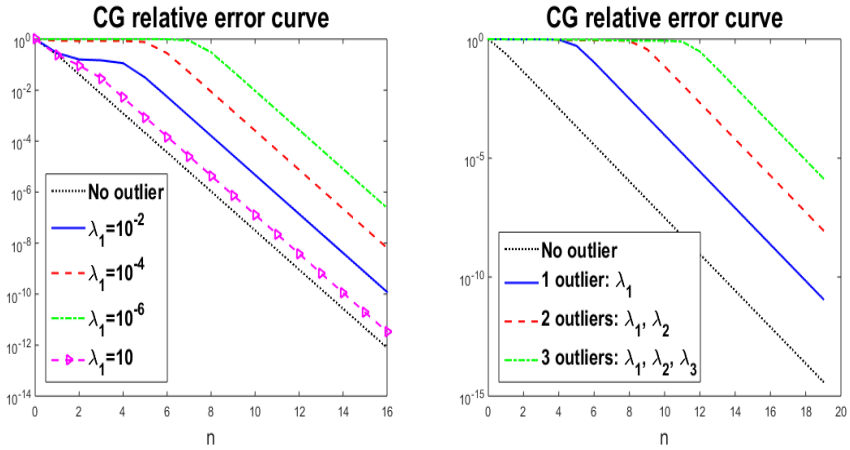

Figure 2.2: *CG relative error curve for diagonal matrices with spectrum related to the equilibrium distribution in* $[2, 4]$*, with* $x_0 = 0$ *and* $b = (1, \ldots, 1)^T$*. On the left we plot the relative error curve without outlier (black dotted line) and compare with the addition of one outlier with different values. On the right we compare the relative error (black dotted line) with the cases of several outliers,* $\lambda_1 = 10^{-3}$*,* $\lambda_2 = 1/50$*,* $\lambda_3 = 10^{-1}$*.*

2.3.3 Superlinear convergence for CG (conjugate gradients)

People have been aware of superlinear convergence for CG for more than forty years. A first attempt to quantify such a convergence behavior was suggested by Kuijlaars and Beckermann [12], see also the review [76] for a more comprehensive summary, or the review [15] from the perspective of discrete orthogonal polynomials.

The key ingredient of this theory is to dispose of a measure σ with continuous potential U^σ (see Definition A.1.3) and compact support describing the eigenvalue distribution. In [12], this is quantified by the following condition.

Condition 1 There is a sequence of systems $A_N x_N = c_N$, where N denotes the size of the matrix, such that the eigenvalues of the matrices A_N are uniformly bounded (i.e. they are all in a fixed interval $[0, R]$), and they have asymptotic distribution σ, which means that σ is the weak-star limit of normalized counting measures of the spectra of the A_N

$$\lim_{N\to\infty} \frac{1}{N} \sum_{\lambda\in\Lambda(A_N)} \delta_\lambda = \sigma. \tag{2.7}$$

It follows that σ must have a compact support in $[0, R]$ and has total mass at most one.

Let us consider two more technical conditions.

Condition 2 The logarithmic potential U^σ is continuous and real-valued. It is a regularity condition on σ, it prevents σ to have point masses. This condition is satisfied for example if the density is continuous.

Condition 3 The condition 1 is satisfied also for the function log, i.e.

$$\lim_{N\to\infty} \frac{1}{N} \sum_{\lambda\in\Lambda(A_N)} \log(\lambda) = \int \log(\lambda) \mathrm{d}\sigma(\lambda).$$

This condition follows from the condition 1 if we know that the eigenvalues are in $[a, R]$ for all N with $a > 0$. In fact this condition prevents eigenvalues from approaching 0 too fast as $N \to \infty$.

In many applications, the matrices appear as discretizations of a continuous operator, and these three conditions are natural. Under those weak assumptions for small eigenvalues, the authors establish in [12, Theorem 2.1] for the nth iterate $x_{n,N}^{CG}$ of conjugate gradients applied to the system $A_N x_N = c_N$ the asymptotic upper bound

$$\begin{aligned} \limsup_{\substack{n,N\to\infty \\ n/N\to t}} \left(\frac{\|x_{n,N}^{CG} - A_N^{-1}c_N\|_{A_N}}{\|x_{0,N}^{CG} - A_N^{-1}c_N\|_{A_N}} \right)^{1/n} &\leq \limsup_{\substack{n,N\to\infty \\ n/N\to t}} \left(E_n(\Lambda(A_N))\right)^{1/n} \\ &\leq \exp\left(-\frac{1}{t} \int_0^t g_{S(\tau)}(0,\infty) d\tau \right), \quad (2.8) \end{aligned}$$

where $(S(t))_{0<t<\|\sigma\|}$ is a decreasing family of compact subsets of the convex hull of the spectra, obtained from some constrained extremal problem in logarithmic potential theory, which we explain below (see also Appendix A.3).

For measures σ with compact support and continuous potential, and $0 < t < \|\sigma\|$, according to [32, 94] there exists a unique minimizer $\nu_{t,\sigma}$ of $I(\nu)$ under all candidates $\nu \in \mathcal{M}_1(\mathrm{supp}(\sigma))$ with $\nu \leq \sigma/t$. This minimizer is uniquely characterized by the existence of a constant $C_{t,\sigma} \in \mathbb{R}$ such that

$$U^{\nu_{t,\sigma}}(x) = C_{t,\sigma} \quad \text{for } x \in \mathrm{supp}(\sigma/\mathrm{t} - \nu_{\mathrm{t},\sigma}), \quad U^{\nu_{t,\sigma}}(x) \leq C_{t,\sigma} \quad \text{for } x \in \mathrm{supp}(\sigma).$$

Many Buyarov-Rakhmanov [23] type properties are known about the measures $\nu_{t,\sigma}$ for fixed σ and varying t, we just recall here from [12, Proof of Theorem 2.1] the fact that the measures $t\nu_{t,\sigma}$ are increasing in t, and hence

$$S(t) := \mathrm{supp}(\sigma/\mathrm{t} - \nu_{\mathrm{t},\sigma}) \quad \text{decreases in } t. \tag{2.9}$$

As a consequence, the map $n \mapsto -N \int_0^{n/N} g_{S(\tau)}(0,\infty) d\tau$ is concave and describes superlinear convergence behavior. The compact sets $S(t)$ may have a quite complicated shape, and the main finding of [75] roughly says that the nth Ritz values of A_N approach well all eigenvalues in $\Lambda(A_N) \setminus S(n/N)$. There is a similar (rough) interpretation of (2.8): so-called "converged" eigenvalues which are already well approached by nth Ritz values should no longer contribute (in exact arithmetic) to the convergence of CG at later stages.

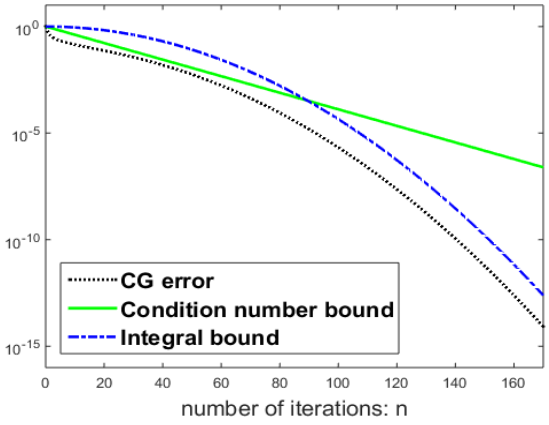

Figure 2.3: Lower and upper bounds for $n \mapsto E_n(\Lambda(A))$. Here $\lambda_j = (j/N)\,(2 - j/N)$ for $j = 1, ..., N = 1000$. As lower bound we draw the relative CG error in energy norm, with $A = \mathrm{diag}(\lambda_1, ..., \lambda_N)$, $c = (1, ..., 1)^T$, and starting vector $x_0^{CG} = 0$ (black dotted line). The upper bounds come from (2.4) with $b/a = \lambda_N/\lambda_1$ for the condition number bound (green solid line)

Let us consider a fourth condition.

Condition 4

$$\lim_{N\to\infty} \frac{1}{(\#\Lambda(A_N))^2} \sum_{\lambda\in\Lambda(A_N)} \sum_{\substack{\lambda'\in\Lambda(A_N)\\ \lambda'\neq\lambda}} \log\frac{1}{|\lambda-\lambda'|} = \frac{1}{\|\sigma\|^2} I(\sigma).$$

Under this additional condition, the inequality (2.8) can be improved to give equality [12, Theorem 2.2].

The drawback of this theory is that all results in [12] study only the so-called asymptotic convergence factor. In addition, this theory requires to consider sequences of systems of equations with a joint eigenvalue distribution, and thus gives not so much information about the actual rate of convergence for a single matrix. Numerical evidence in [12, 13, 14] did let to conjecture that the above upper bound (2.8) even holds (up to some modest constant) for a single matrix A, without limits and without taking the n-th root, see for instance [12, Eqn.(1.9) and Figures 1 and 4], [14, Eqn. (1.3)], or Figure 2.3. Of course, for a single matrix we cannot define σ through (2.7). This gives the following conjecture.

Conjecture 2.3.1 *There is a (modest) constant $C \in \mathbb{R}$ and a technique of associating a measure σ with compact support and continuous potential to the spectrum of a positive definite matrix A such that, for all n sufficiently small,*

$$E_n(\Lambda(A)) \leq \exp\left(C - N\int_0^{n/N} g_{S(t)}(0,\infty)dt\right).$$

It may be that this conjecture is wrong for measures where $S(t)$ has a complicated shape. In our proof of the conjecture (chapter 5), following [12, Lemma 3.1(a)], we will impose sufficient conditions on σ such that $S(t) = [a(t), b]$ for all t.

Chapter 3

Convergence of Minimal Residual methods in the presence of few outliers

In this chapter we present a study of the convergence behavior of Minimal Residual (MR) methods, like GMRES, for solving nonsingular systems of linear equations, which is an improvement and a generalization of the work of Ipsen et al. in [24]. In fact we will study the quantity $E_n(S)$ introduced in Equation (2.1). In [24], the authors obtained upper bounds of this quantity for disks S. Using tools from complex analysis, Faber polynomials and AAK theory, we construct polynomials leading to better upper bounds, and generalize the work to convex sets. Our upper bound provides a better understanding of the superlinear convergence of MR methods in presence of few outliers. We also give a lower bound in Theorem 3.1.3 which can be seen as a generalization of the classical asymptotic convergence factor.

3.1 Introduction

In this chapter we concentrate our study on the value $E_n(S)$ introduced in equation (2.1)

$$E_n(S) = \min \left\{ \frac{\|p\|_S}{|p(0)|}, \quad p \in \Pi_n, \quad \forall \lambda \in \Lambda(A) \setminus S : \quad p(\lambda) = 0 \right\}.$$

Let $\{\lambda_j\}_{j=1}^J$ denote the distinct eigenvalues of a matrix A, and define $\Lambda_0 = \{\lambda_j\}_{j=1}^d$, a subset of the spectrum which will play a particular role in our analysis. The set S will enclose the set $\{\lambda_j\}_{j=d+1}^J$.

Definition 3.1.1 *The eigenvalues in Λ_0 will be called outliers. We say that $S = S_d$ is an inclusion set if it is a closed neighborhood of the remaining spectrum*

$\{\lambda_j\}_{j=d+1}^{J}$, *non-empty and different of* $\overline{\mathbb{C}}$, *such that* $\overline{\mathbb{C}} \setminus S$ *is simply connected and* $0, \lambda_1, \ldots, \lambda_d \notin S$.

In [24] S is called a cluster. The main idea is to treat separately outliers and eigenvalues in the inclusion set. In this chapter we will not discuss how to choose outliers, but once fixed, an outlier should not control the rate of convergence. In the chapter 5 we will come back to this problem for a particular distribution of the spectrum in Theorem 5.1.4. So their choice will depend on which step of the process we are. There is a strong link between the choice and the number of the outliers and the inclusion set S. The more informations we have on the spectrum of A, the more we will be able to choose a good set S.

For connected inclusion sets S that contain the spectrum, it is known [35, Theorem 1 and 2] that

$$E_n(S) \geq \frac{1}{|\alpha_0|^n},$$

where the inverse of $|\alpha_0|$ (see Definition 3.1.2) is the asymptotic convergence factor ρ_S given in Equation (2.3). We give a lower bound in Theorem 3.1.3 which allows some eigenvalues to be outside S, and so generalize this lower bound. So we can choose smaller sets included in the preceding one which will give larger $|\alpha_0|$ (see Remark 3.2.1). The price to pay is a factor of Blaschke products (independent of n) of modulus greater than one.

Recall that we have the inequalities (2.2)

$$E_n(\Lambda(A)) \leq E_n(S \cup \Lambda_0) \leq E_n(S).$$

Our goal is to find a polynomial p_n $(d \leq n)$ in

$$\Pi_n^0(\Lambda_0) = \{p \in \Pi_n^0 / p(\lambda) = 0 \text{ for } \lambda \in \Lambda_0\}, \tag{3.1}$$

and to give an overestimation of $E_n(S)$ with the inequality

$$E_n(S) \leq \frac{\|p_n\|_S}{|p_n(0)|},$$

where $\|p_n\|_S = \sup_{z \in S} |p_n(z)|$ is the ∞-norm. Our overestimation will be related to the asymptotic convergence factor and will lead to sharp upper bounds (in the sense that the ratio of the upper and the lower bounds tends to one).

3.1.1 Results in the paper of Ipsen et al.

The first step in our analysis is to bring the problem onto the open unit disk in the complex plane, denoted by $\mathbb{D}$, by a change of variable. Consider the conformal map ϕ_S (see Equation (A.2)) its inverse $\psi_S : \overline{\mathbb{C}} \setminus \overline{\mathbb{D}} \to \overline{\mathbb{C}} \setminus S$. With this change of variable and the maximum principle, we can consider the problem on the unit disk

$$E_n(S) = \min_{p \in \Pi_n^0(\Lambda_0)} \max_{z \in S} |p(z)| = \min_{p \in \Pi_n^0(\Lambda_0)} \max_{w \in \partial\mathbb{D}} |p(\psi_S(w))|,$$

which will allow us to use results from complex analysis and approximation theory in the complex plane.

Definition 3.1.2 *Let us set the notations $\alpha_0 = \phi_S(0)$ ($0 \notin S$) and $\alpha_j = \phi_S(\lambda_j)$ ($j = 1 : d$), for the images of zero and of the eigenvalues outside S under ϕ_S. We define the following function on $\overline{\mathbb{C}} \setminus \mathbb{D}$*

$$f_{n,\alpha_1,\ldots,\alpha_d}(w) = w^n \prod_{j=1}^{d} \left(\frac{w - \alpha_j}{1 - \bar{\alpha}_j w} \right) \tag{3.2}$$

It is important to note that ϕ_S depends strongly on the choice of S, and thus if we change S, all the α_j change. The function $f_{n,\alpha_1,\ldots,\alpha_d}$ will play a central role in the sequel.

In [24] the authors considered the case when the inclusion set S is a disk centered at γ of radius $r = |\gamma|\rho$ ($\rho < 1$)

$$\{\lambda_j, d+1 \leq j \leq J\} \subset S = D(\gamma, r), \quad r = |\gamma|\rho.$$

Considering the product $p_n = q_d s_{n-d}$ where

$$q_d(z) = \prod_{j=1}^{d} \left(1 - \frac{z}{\lambda_j}\right) = \prod_{j=1}^{d} \frac{\alpha_j - \phi_S(z)}{\alpha_j - \alpha_0} \in \Pi_d^0,$$

and $s_{n-d} \in \Pi_{n-d}^0$, they overestimated the maximum of the product by the product of the maxima. Then they choose the best s_{n-d} on a disk which is explicitly known by Zarantonello's lemma [101, Proposition 6.26]:

$$s_{n-d}(z) = \left(1 - \frac{z}{\gamma}\right)^{n-d} = \left(\frac{\phi_S(z)}{\phi_S(0)}\right)^{n-d}.$$

As a consequence they obtained in [24, Proposition 4.1] the upper bound for disks (reformulated in our terms)

$$E_n(S) \leq \frac{C_{Ipsen}}{|f_{n,\alpha_1,\ldots,\alpha_d}(\alpha_0)|}, \tag{3.3}$$

where $1/|f_{n,\alpha_1,\ldots,\alpha_d}(\alpha_0)|$ is the lower bound given in Theorem 3.1.3, and

$$C_{Ipsen} = \prod_{j=1}^{d} \frac{1 + \frac{1}{|\alpha_j|}}{|1 - \frac{1}{\overline{\alpha}_j \alpha_0}|}.$$

C_{Ipsen}, which depends on S and d (not on n), represents the ratio between the upper bound and the lower bound, and is strictly greater than one as

$$1 + \frac{1}{|\alpha_j|} \geq 1 + \frac{1}{|\overline{\alpha}_j \alpha_0|} \geq |1 - \frac{1}{\overline{\alpha}_j \alpha_0}|.$$

In fact in [24], they gave the bound

$$E_n(S) \leq \frac{1}{|f_{n,\alpha_1,\ldots,\alpha_d}(\alpha_0)|} \left(\max_{j=1:d} \frac{1 + \frac{1}{|\alpha_j|}}{|1 - \frac{1}{(\alpha_0 \overline{\alpha_j})}|} \right)^d,$$

but there is no reason to take the maximum instead of the product.

3.1.2 Improvements

The aim of this chapter is to improve the results in [24]. We remark that Zarantonello's lemma does not apply if we take into account q_d. So we need another technique to find a better approximation which takes into consideration the outliers. We prove in Section 3.2 that we have the following lower bound which can be seen as a generalization of the classical asymptotic convergence factor.

Theorem 3.1.3 (Lower bound) *For $d \leq n$ and for every inclusion set S (Definition 3.1.1), we have an explicit lower bound for our min-max problem*

$$E_n(S) \geq \frac{1}{|f_{n,\alpha_1,\ldots,\alpha_d}(\alpha_0)|} = \frac{1}{|\alpha_0|^n} \prod_{j=1}^{d} \left| \frac{1-\bar{\alpha}_j \alpha_0}{\alpha_0 - \alpha_j} \right|,$$

where $f_{n,\alpha_1,\ldots,\alpha_d}$ is defined in (3.2).

Then we suggest a choice of polynomial in w taking into account that we impose certain zeros (s_k will depend on S and q_d). In the case of a disk, the maps ϕ and ψ are linear, and thus our choice leads to a polynomial in z. This allows us to obtain in Section 3.3 an upper bound in the case of a disk stated in the following theorem.

Theorem 3.1.4 (Upper bound for a disk and d outliers) *Let the inclusion set S be a closed disk and consider d outliers ($d \leq n$). For each α_j we associate an integer $n_j \geq 1$ such that $\sum_{j=1}^{d} n_j \leq n$. Then we can obtain the upper bound*

$$E_n(S) \leq \Upsilon_n(n_1, \ldots, n_d) := \frac{C_{n_1,\ldots,n_d}}{|f_{n,\alpha_1,\ldots,\alpha_d}(\alpha_0)|}, \tag{3.4}$$

where $C_{n_1,\ldots,n_d} = \prod_{j=1}^{d} \frac{1+\frac{1}{|\alpha_j|^{n_j}}}{|1-\frac{1}{(\alpha_0\overline{\alpha}_j)^{n_j}}|}$.

The ratio of our upper and lower bound $C_{n_1,\ldots,n_d}$ depends on S, d and the choice of the n_j. Provided that all $n_j \to \infty$, $C_{n_1,\ldots,n_d}$ tends to 1, showing that our upper bound is asymptotically sharp. If all the n_j are equal to one we find the upper bound given in Equation (3.3), and thus this theorem contains the Proposition 4.1 in [24] as a special case. With a good choice of n_j, our theorem may allow us to obtain better upper bounds.

For more general inclusion sets S, a polynomial in w do not give a polynomial in z, so in this case we use the Faber transform which transforms polynomials in w into polynomials in z. In Section 3.4 we consider convex sets with one outlier which can be extended for several outliers, but the constant will explode as the number of outliers grows, this is why this bound can only be satisfactory for few outliers.

Theorem 3.1.5 (Upper bound for a convex set and one outlier) *If the inclusion set S is convex, then for n such that $\frac{1+\frac{1}{|\alpha_1|^n}}{|1-\frac{1}{(\alpha_0\overline{\alpha_1})^n}|} < \frac{|\alpha_0|^n}{2}$, we have the*

upper bound

$$E_n(S) \leq \frac{3C_{conv}}{|f_{n,\alpha_1}(\alpha_0)|},$$

where $C_{conv} = \frac{1+\frac{1}{|\alpha_1|^n}}{|1-\frac{1}{(\alpha_1\alpha_0)^n}|}\left[\left(1-\frac{2}{|\alpha_0|^n}\frac{1+\frac{1}{|\alpha_1|^n}}{|1-\frac{1}{(\alpha_0\alpha_1)^n}|}\right)\right]^{-1}$.

C_{conv} tends to 1 as $n \to \infty$, showing that the preceding upper bound is asymtotically sharp up to a factor 3.

In the last section we present some openings and some generalizations which can be done.

3.2 Lower bound

In this section we consider d (simple) outliers $\{\lambda_j, 1 \leq j \leq d\}$, an inclusion set S (see Definition 3.1.1), and we prove the lower bound given in Theorem 3.1.3. The main idea of the proof is to solve another problem of minimization on the unit disk by using the maximum principle.

Proof of Theorem 3.1.3 : Consider a function f which is analytic and bounded in $\overline{\mathbb{C}} \setminus \mathbb{D}$ such that $f(\alpha_j) = 0$, for $j = 1 : d$. Then the function $g(w) = f(w)\prod_{j=1}^{d}\left(\frac{1-\bar{\alpha_j}w}{w-\alpha_j}\right)$ is also analytic and bounded in $\overline{\mathbb{C}} \setminus \mathbb{D}$. Using that the Blaschke factors are of modulus one on $\partial\mathbb{D}$ and the maximum principle for analytic functions, we have $\sup_{\overline{\mathbb{C}}\setminus\mathbb{D}}|g(w)| = \|f\|_{\partial\mathbb{D}}$, which implies

$$\forall w \in \overline{\mathbb{C}} \setminus \mathbb{D}, \quad |f(w)| \leq \|f\|_{\partial\mathbb{D}} \prod_{j=1}^{d}\left|\frac{w-\alpha_j}{1-\bar{\alpha_j}w}\right|.$$

Moreover, if we ask that $f(\alpha_0) = 1$, the preceding formula permits to obtain

$$\|f\|_{\partial\mathbb{D}} \geq \prod_{j=1}^{d}\left|\frac{1-\bar{\alpha_j}\alpha_0}{\alpha_0-\alpha_j}\right|.$$

Now we observe that for every $p \in \Pi_n^0(\Lambda_0)$ (defined in (3.1)), the function $f(w) = \frac{p(\psi(w))\alpha_0^n}{w^n}$ is analytic and bounded outside the unit disk, and satisfies the conditions $f(\alpha_j) = 0$ and $f(\alpha_0) = 1$. This allows to deduce that

$$\min_{p\in\Pi_n^0(\Lambda_0)}\max_{w\in\partial\mathbb{D}}\left|\frac{p(\psi(w))\alpha_0^n}{w^n}\right| \geq \prod_{j=1}^{d}\left|\frac{1-\bar{\alpha_j}\alpha_0}{\alpha_0-\alpha_j}\right|$$

which leads to the lower bound given in Theorem 3.1.3.

□

Remark 3.2.1 *If $S_2 \subset S_1$, then $|\phi_{S_1}(0)| \leq |\phi_{S_2}(0)|$. Indeed, since by definition (see Definition A.1.4) the function $g_{S_2}(.,\infty) - g_{S_1}(.,\infty)$ has nonnegative*

boundary values on the boundary of S_1, it is nonnegative everywhere on $\overline{\mathbb{C}} \setminus S_1$ (minimum principle). This property leads us to consider a sequence of decreasing inclusion sets S_j, i.e. $S_{j+1} \subset S_j$ for all j, to obtain a decreasing sequence $\frac{1}{|\phi_{S_j}(0)|}$.

We see from the lower bound that the choice of S and the number d of outliers will be very important to obtain satisfying bounds for MR methods.

Remark 3.2.2 *We note that the function $w \mapsto \prod_{j=1}^{d} \frac{w-\alpha_j}{1-\bar{\alpha_j}w} \frac{1-\bar{\alpha_j}\alpha_0}{\alpha_0-\alpha_j}$ attains the lower bound given in the proof of the theorem. Unfortunately this is not a polynomial. So an idea to find an upper bound is to choose a polynomial $p_n \in \Pi_n^0(\Lambda_0)$ such that $p_n \circ \psi(w) \simeq \frac{w^n}{\alpha_0^n} \prod_{j=1}^{d} \frac{w-\alpha_j}{1-\bar{\alpha_j}w} \frac{1-\bar{\alpha_j}\alpha_0}{\alpha_0-\alpha_j}$. We can do it by searching a polynomial p_n such that $p_n \circ \psi$ approximates the function $f_{n,d}$ and then normalize at α_0.*

Interest of considering more outliers.

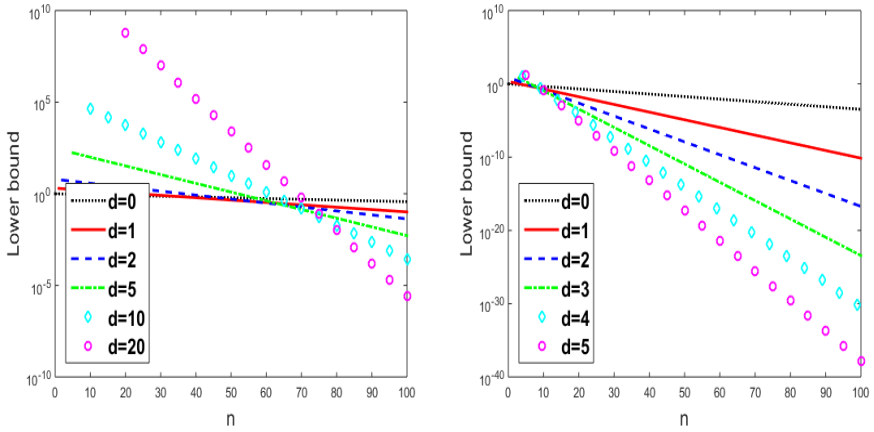

Figure 3.1: *Lower Bounds with different d. On the left we consider a matrix of order 100 with equidistant eigenvalues $\lambda_1 = 1, \lambda_2 = 2, \ldots, \lambda_{100} = 100$. For each d, we take $S_d = \overline{D}((\lambda_{d+1} + \lambda_{100})/2, (\lambda_{100} - \lambda_{d+1} + 1)/2)$. On the right we consider a matrix of order 100 with five eigenvalues $\lambda_1 = 1, \lambda_2 = 2, \lambda_3 = 3, \lambda_4 = 4, \lambda_5 = 5$, and all the others contained in a closed disk $\overline{D}(9, 3)$ centered at 9 and of radius 3 with no more information. We take as inclusion sets for $d = 0 : 4$, $S_d = \overline{D}((\lambda_{d+1} + 9)/2, (9 - \lambda_{d+1} + 1)/2)$, and $S_5 = \overline{D}(9, 3)$.*

Let us illustrate some properties of the lower bound, depending on the number of outliers and the choice of S, with two academic examples in Figure 3.1. The product of the inverses of Blaschke factors is constant for S and d fixed, and

it can grow fastly with d as we can see from the first example in Figure 3.1 on the left. But this product can be compensated for large n by the fact that if we have more outliers, the inclusion set S can be chosen smaller, which can lead to a smaller $\frac{1}{|\alpha_0|}$ (Remark 3.2.1). For example, if we have a decreasing sequence of inclusion sets S_j (which is the case in all our examples), then $\frac{1}{|\phi_{S_j(0)}|} \geq \frac{1}{|\phi_{S_{j+1}(0)}|}$.

We see on the first example on the left that it can take a lot of iterations before it becomes interesting to consider more outliers. In the second example (Figure 3.1 on the right), five eigenvalues are well separated from the others, and we see that it is attractive to consider them as outliers. Indeed, the case $d = 5$ is rapidly better than the others.

In both examples, for small n, it is not interesting to consider a lot of outliers, but as n increases, we can obtain a better rate of convergence if we take into account more and more outliers. So this notion of outliers will allow us to capture the superlinear convergence explained in the preceding chapter.

As for a fixed n the lower bound is true for every $d \leq n$ (and is sharp up to some constant as will be seen later), it is interesting to search the best d which gives the best lower bound at step n, but it seems to be a difficult task. On the graph this means to take the lower envelope of all the lines.

An idea to make the lower bound smaller could be to choose ϕ and the α_j such that the Blaschke product $\prod_{j=1}^{d} \left|\frac{1-\bar{\alpha}_j\alpha_0}{\alpha_0-\alpha_j}\right|$ is as small as possible. This kind of minimization problem has received considerable attention in complex approximation theory, as for example in [8, 47] or [103, Theorem VIII.3.1].

3.3 Upper bound for a disk

In this section we talk about the particular case when the inclusion sets are disks $S = \overline{D}(\gamma, r)$. The conformal maps

$$\phi(z) = \frac{z-\gamma}{r} \quad \text{and} \quad \psi(w) = wr + \gamma$$

are linear, and then a polynomial in w will lead to a polynomial in z. We begin in §3.3.1 with an introduction to AAK theory and Nehari's theorem, and propose a candidate for the approximation of f_{n,α_1}. This polynomial leads to an upper bound for one outlier given in §3.3.2. Inspired by this case $d = 1$, we prove and discuss Theorem 3.1.4 concerning the case of a disk with several outliers in §3.3.3.

3.3.1 AAK theory

In the case $d = 1$, AAK theory allows us to prove that the best approximation of f_{n,α_1} by an H^∞ function is achieved by a polynomial, which will be explicitely given. Let us recall briefly some basic properties of Hankel operators and state Nehari's theorem [92, chap 1]. We denote by $H^2(\mathbb{D})$ the classical Hardy space which consists of the holomorphic functions in $\mathbb{D}$ whose power series coefficients at the origin are square summable. A function $f \in H^2(\mathbb{D})$ has

an associated boundary function on the unit circle $\mathbb{T}$, also denoted by f, defined almost everywhere by means of nontangential limits. By Fatou's theorem we have $\|f\|_{H^2(\mathbb{D})} = \|f\|_{L^2(\mathbb{T})}$, where $L^2(\mathbb{T})$ is the classical Hilbert space endowed with the normalized Lebesgue measure. The boundary functions contain the subspace of $L^2(\mathbb{T})$, also called H^2, of functions whose Fourier coefficients with negative indices vanish. The orthogonal complement of H^2 in L^2, denoted by $(H^2)^\perp = H^2_-$, consists of the functions whose Fourier coefficients with nonnegative indices vanish.

Definition 3.3.1 *For $\varphi \in L^\infty(\mathbb{T})$, the Hankel operator H_φ is defined to be the operator from H^2 to H^2_- given by*

$$H_\varphi f = \mathbb{P}_-(\varphi f),$$

where $\mathbb{P}_-$ is the orthogonal projection onto H^2_-.

The function φ is called a symbol of the Hankel operator H_φ. There are infinitely many different symbols that produce the same Hankel operator, indeed it is known that $H_\varphi = 0$ is tantamount to $\varphi \in H^\infty$. We have the following important theorem [92, Theorem 1.3].

Theorem 3.3.2 (Nehari) *Let $\varphi \in L^\infty(\mathbb{T})$. Then*

$$\|H_\varphi\| = \inf\{\|\varphi - h\|_\infty, h \in H^\infty\} = dist(\varphi, H^\infty).$$

In the following we need the norm of an Hankel operator with a symbol equal to a Blaschke factor.

Lemma 3.3.3 *For $|\alpha| > 1$, we have*

$$\|H_{\frac{w-\alpha}{1-\overline{\alpha}w}}\| = 1$$

Proof of the lemma : Using the equality $\frac{w-\alpha}{1-\overline{\alpha}w} = \frac{-1}{\overline{\alpha}} + \frac{|\alpha|^2-1}{\overline{\alpha}^2}\frac{1}{w-\frac{1}{\overline{\alpha}}}$, we obtain

$$\|H_{\frac{w-\alpha}{1-\overline{\alpha}w}}\| = \frac{|\alpha|^2-1}{|\alpha|^2}\|H_{\frac{1}{w-\frac{1}{\overline{\alpha}}}}\|,$$

and it is known that $\|H_{\frac{1}{w-\frac{1}{\overline{\alpha}}}}\| = \frac{|\alpha|^2}{|\alpha|^2-1}$ [121, Example 15.17]. Let us give another proof of this equality here. The Hankel operator $H_{\frac{1}{w-\frac{1}{\overline{\alpha}}}}$ can be computed explicitely, indeed for every $f \in H^2$ we can write

$$H_{\frac{1}{w-\frac{1}{\overline{\alpha}}}} f(w) = \mathbb{P}_-\left(\frac{f(w) - f(\frac{1}{\overline{\alpha}})}{w - \frac{1}{\overline{\alpha}}} + \frac{f(\frac{1}{\overline{\alpha}})}{w - \frac{1}{\overline{\alpha}}}\right) = f(\frac{1}{\overline{\alpha}})\frac{1}{w - \frac{1}{\overline{\alpha}}}.$$

We can see this rank one operator as the tensor product $\varphi_1 \otimes \varphi_2$ defined by (see [1, Section 0.6])

$$(\varphi_1 \otimes \varphi_2)f = \langle \varphi_2, f\rangle \varphi_1,$$

with $\varphi_1(w) = \frac{1}{w-\frac{1}{\overline{\alpha}}} = \sum_{p=0}^{\infty} \frac{1}{\overline{\alpha}^p} \frac{1}{w^{p+1}} \in H^2_-$ and φ_2 the linear functional defined by $\varphi_2(f) = f(\frac{1}{\overline{\alpha}})$. We can compute easily the norms of φ_1 and φ_2. For φ_1, we have

$$\|\varphi_1\|^2 = \sum_{p=0}^{\infty} \left(\frac{1}{|\alpha|^p}\right)^2 = \sum_{p=0}^{\infty} \left(\frac{1}{|\alpha|^2}\right)^p = \frac{|\alpha|^2}{|\alpha|^2-1}.$$

For φ_2, it is clear by Cauchy-Schwarz that

$$\|\varphi_2\| \leq \frac{|\alpha|}{\sqrt{|\alpha|^2-1}},$$

and this norm is reached for the function which has the Fourier coefficents $(\frac{1}{\alpha^j})$. Now using the equality $\|\varphi_1 \otimes \varphi_2\| = \|\varphi_1\| \, \|\varphi_2\|$ [1, Section 0.6], we obtain

$$\|H_{\frac{1}{w-\frac{1}{\overline{\alpha}}}}\| = \|\varphi_1\| \, \|\varphi_2\| = \frac{|\alpha|^2}{|\alpha|^2-1}$$

which ends the proof.

□

Remark 3.3.4 *If we set $\beta = 1/\overline{\alpha}$, we have $\frac{w-\alpha}{1-\overline{\alpha}w} = \frac{\beta}{\overline{\beta}} \frac{1-\overline{\beta}w}{w-\beta}$ and thus*

$$\|H_{\frac{w-\alpha}{1-\overline{\alpha}w}}\| = \|H_{\frac{1-\overline{\beta}w}{w-\beta}}\|.$$

As $|\beta| < 1$, $\frac{w-\beta}{1-\overline{\beta}w}$ is a classical Blaschke factor. It is known in the theory of SISO systems (of finite rank), that for lossless transfer functions of type C/B, where C is a product of at most $m-1$ (classical) Blaschke factors and B a product of m (classical) Blaschke factors, that $\|H_{C/B}\| = 1$. This implies the preceding lemma for $m = 0$. This result was proved in personal notes[1] with Martine Olivi[2].

This short introduction to Hankel operators, will help us to propose a good polynomial to approximate the function

$$f_{n,\alpha_1}(w) = w^n \frac{w-\alpha_1}{1-\overline{\alpha}_1 w} = -\frac{w-\alpha_1}{\overline{\alpha}_1^n} \sum_{j=0}^{n-1} (\overline{\alpha}_1 w)^j + \frac{1}{\overline{\alpha}_1^n} \frac{w-\alpha_1}{1-\overline{\alpha}_1 w}. \tag{3.5}$$

Considering the polynomial

$$p_{n,\alpha_1}(w) = -\frac{w-\alpha_1}{\overline{\alpha}_1^n} \sum_{j=0}^{n-1} (\overline{\alpha}_1 w)^j = f_{n,\alpha_1}(w) \left(1 - \frac{1}{(\overline{\alpha}_1 w)^n}\right), \tag{3.6}$$

we obtain

$$\|f_{n,\alpha_1} - p_{n,\alpha_1}\|_{\partial\mathbb{D}} = \frac{1}{|\alpha_1|^n}.$$

[1] The inequality $\|H_{C/B}\| \leq 1$ is clear, and by Nehari's theorem, we know that $\|H_{C/B}\| = \inf\{\|C - hB\|, h \in H^\infty\}$. Then the idea is to make a link with a Nevanlinna-Pick problem: for $f \in H^\infty$ find $\min\{\|f - hB\|, h \in H^\infty\}$. This minimum is attained for $h = 0$ if and only if f is a Blaschke product of degree $< m$ [52, Corollary I.2.3].

[2] EPI-APICS, Inria, BP 93, Sophia-Antipolis cedex (Martine.Olivi@inria.fr, https://www-sop.inria.fr/members/Martine.Olivi/).

Proposition 3.3.5 *With the preceding notations,*

$$\inf\{\|f_{n,\alpha_1} - h\|_\infty, h \in H^\infty\} = \frac{1}{|\alpha_1|^n},$$

and the infimum is achieved by the polynomial p_{n,α_1}.

Proof of Proposition 3.3.5 : This is a direct consequence of (3.5) and of Lemma 3.3.3

$$H_{f_{n,\alpha_1}} = H_{f_{n,\alpha_1} - p_{n,\alpha_1}} = \frac{1}{\overline{\alpha}_1^n} H_{\frac{w-\alpha_1}{1-\overline{\alpha}_1 w}}.$$

□

So we found that the polynomial p_{n,α_1} solves a problem of minimization over a class of analytic functions. After normalization at α_0, we obtain

$$\frac{p_{n,\alpha_1}(w)}{p_{n,\alpha_1}(\alpha_0)} = \frac{w-\alpha_1}{\alpha_0-\alpha_1} \frac{1+\overline{\alpha}_1 w + \cdots + (\overline{\alpha}_1 w)^{n-1}}{1+\overline{\alpha}_1\alpha_0 + \cdots + (\overline{\alpha}_1\alpha_0)^{n-1}},$$

and thus in term of z

$$\frac{p_{n,\alpha_1}(\phi(z))}{p_{n,\alpha_1}(\phi(0))} = \frac{\phi(z)-\phi(\lambda_1)}{\phi(0)-\phi(\lambda_1)} \frac{1+\overline{\phi(\lambda_1)}\phi(z) + \cdots + (\overline{\phi(\lambda_1)}\phi(z))^{n-1}}{1+\overline{\phi(\lambda_1)}\phi(0) + \cdots + (\overline{\phi(\lambda_1)}\phi(0))^{n-1}},$$

which is a polynomial in z since ϕ is linear.

3.3.2 Upper bound for one outlier

In this subsection we use the polynomial p_{n,α_1} given by AAK theory to prove Theorem 3.1.4 for one outlier stated in Proposition 3.3.6. We also discuss and compare our result with the bound which comes from [24] and defined in Equation (3.3).

Proposition 3.3.6 (Upper bound for a disk and one outlier) *In the case of one (simple) outlier* α_1 *and* $1 \leq n_1 \leq n$*, if the inclusion set* S *is a disk, we can obtain the upper bound*

$$E_n(S) \leq \Upsilon_n(n_1) := \frac{1}{|f_{n,\alpha_1}(\alpha_0)|} \frac{1+\frac{1}{|\alpha_1|^{n_1}}}{|1-\frac{1}{(\alpha_0\overline{\alpha_1})^{n_1}}|}.$$

This upper bound is asymptotically sharp since the ratio between the upper and the lower bounds $\frac{1+\frac{1}{|\alpha_1|^{n_1}}}{|1-\frac{1}{(\alpha_0\overline{\alpha_1})^{n_1}}|}$ tends to 1 as $n_1 \to \infty$.

Proof of Proposition 3.3.6 : Inspired by the polynomial given in (3.6), we set for $1 \leq n_1 \leq n$ the polynomial

$$p_{n_1,\alpha_1}(w) = -\frac{w-\alpha_1}{\overline{\alpha}_1^{n_1}} \sum_{j=0}^{n_1-1} (\overline{\alpha}_1 w)^j = f_{n_1,\alpha_1}(w)\left(1 - \frac{1}{(\overline{\alpha}_1 w)^{n_1}}\right).$$

Multiplying this polynomial with w^{n-n_1} gives a polynomial of degree n which allows to conclude. Indeed, for

$$p_n(w) = w^{n-n_1} p_{n_1,\alpha_1}(w) = f_{n,\alpha_1}(w)\left(1 - \frac{1}{(\overline{\alpha}_1 w)^{n_1}}\right),$$

we have the inequalities

$$\|p_n\|_{\partial\mathbb{D}} \leq 1 + \frac{1}{|\alpha_1|^{n_1}} \quad \text{and} \quad |p_n(\alpha_0)| = |f_{n,\alpha_1}(\alpha_0)|\left|1 - \frac{1}{(\alpha_0\overline{\alpha_1})^{n_1}}\right|, \tag{3.7}$$

and using the fact that $p_n \circ \phi_S \in \Pi_n^0(\Lambda_0)$ we conclude

$$E_n(S) \leq \frac{\|p_n\|_{\partial\mathbb{D}}}{|p_n(\alpha_0)|} \leq \frac{1}{|f_{n,\alpha_1}(\alpha_0)|}\frac{1 + \frac{1}{|\alpha_1|^{n_1}}}{|1 - \frac{1}{(\alpha_0\overline{\alpha_1})^{n_1}}|}.$$

□

Remark 3.3.7 *It is clear that we have the following overestimation*

$$\Upsilon_n(n_1) \leq \frac{1}{|f_{n,\alpha_1}(\alpha_0)|}\frac{1 + \frac{1}{|\alpha_1|^{n_1}}}{1 - \frac{1}{|\alpha_0\alpha_1|^{n_1}}}.$$

This overestimation is the worst case possible, which is attained when $\alpha_0\overline{\alpha_1} \in \mathbb{R}_+$, a case we will study in the sequel. This upper bound is decreasing with n_1, and thus to minimize it, we have to take the $n_1 = n$. This implies that this upper bound is easier to handle, specially when we will work with several outliers.

Now let us compare the upper bounds (3.3) ($n_1 = 1$) and (3.4) with $n_1 = n$. Although for large n we should choose $n_1 = n$ since our upper bound is asymptotically sharp, for small n the choice of n_1 is more tricky. We illustrate this in Figure 3.2. On the left we are in a case where $n_1 = n$ is always better. On the right, the case $n_1 = n$ is worse than the case $n_1 = 1$ up to $n = 3894$, which makes the new bound for $n_1 = n$ less attractive for small n. Such a phenomenon is due to the facts that α_0 and α_1 are of opposite signs, and α_1 is very near one (which means that λ_1 should perhaps not have been considered as an outlier).

Case $\overline{\alpha}_1\alpha_0 \in \mathbb{R}$

Now let us look at a special case when $\overline{\alpha}_1\alpha_0 \in \mathbb{R}$, which is a case we can encounter in a lot of situations like for example for Hermitian matrices. Setting $\alpha_0 = r_0 e^{i\theta_0}$ and $\alpha_1 = r_1 e^{i\theta_1}$, we easily obtain

$$\begin{aligned}\bar{\alpha}_1\alpha_0 \in \mathbb{R} &\iff \theta_1 \equiv \theta_0 \pmod{\pi} \\ &\iff \alpha_1, \alpha_0 \text{ and } 0 \text{ are aligned on the same straight line} \\ &\iff \lambda_1, 0 \text{ and } \gamma \text{ are aligned on the same straight line.}\end{aligned}$$

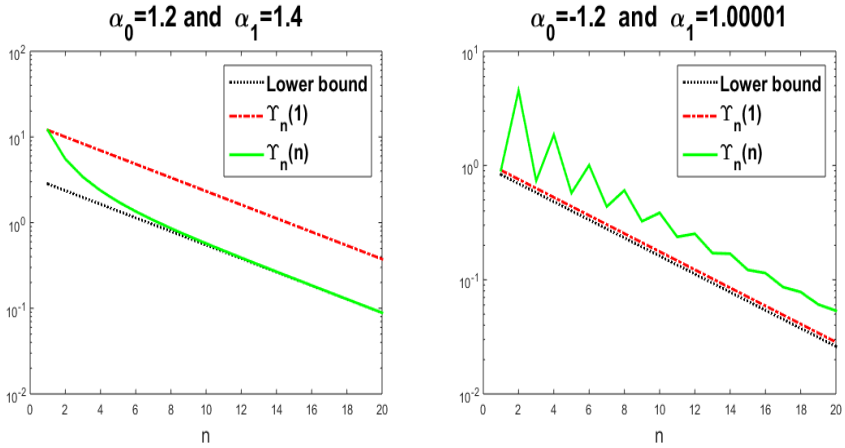

Figure 3.2: *Comparison between the case $n_1 = 1$ and $n_1 = n$, in the case of a disk with one outlier.*

If $\theta_1 \equiv \theta_0 \pmod{2\pi}$, i.e. α_1 and α_0 are on the same straight line passing through zero and on the same side with respect to zero, then $e^{i(\theta_1-\theta_0)} = 1$ and we have

$$\Upsilon_n(n_1) = \frac{1}{|f_{n,\alpha_1}(\alpha_0)|} \frac{1 + \frac{1}{r_1^{n_1}}}{1 - \frac{1}{(r_0 r_1)^{n_1}}}.$$

So in this case, it is clear that $\Upsilon_n(n) \geq \Upsilon_n(1)$ for every n. In Figure 3.3 we consider two academic examples to illustrate this situation. The size of $|\alpha_0|$ gives the slope of the lower bound. The upper bound $\Upsilon_n(n)$ depends on the size of $|\alpha_1|$, larger it is, faster $\Upsilon_n(n)$ tends to the lower bound.

If $\theta_1 \equiv \theta_0 + \pi \pmod{2\pi}$, i.e. α_1 and α_0 are on the same straight line containing zero but on different side with repsect to zero, then $e^{i(\theta_1-\theta_0)} = -1$, and we have

$$\Upsilon_n(n_1) = \frac{1}{|f_{n,\alpha_1}(\alpha_0)|} \frac{1 + \frac{1}{r_1^{n_1}}}{1 + \frac{(-1)^{n_1+1}}{(r_0 r_1)^{n_1}}}.$$

We clearly see the influence of the parity of n_1 in this case in Figure 3.4. The choice $n_1 = 1$ is better than the choice $n_1 = n$ at the beginning and the convergence of $\Upsilon_n(n_1)$ depends on the size of $|\alpha_1|$.

Those two cases are extreme cases. If α_0 and α_1 are not aligned, the bounds have the same behaviors, more or less pronounced.

3.3.3 Upper bound for a disk and several outliers

When $d \geq 2$, AAK theory does not give a polynomial as we can see in the following example. We consider the function $h(w) = w^n + \sigma \frac{1-cw}{w-c}$ with $c > 1$ and

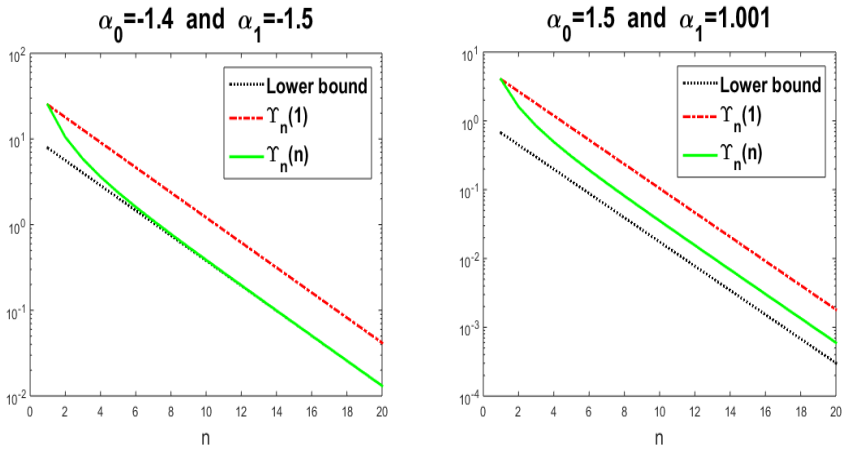

Figure 3.3: *Comparison between the cases $n_1 = 1$ and $n_1 = n$, in the case of a disk with one outlier, and with $\theta_1 \equiv \theta_0 \pmod{2\pi}$.*

$\sigma \in (0,1)$. By verifying that for n even we have $h(-1) = 1-\sigma$, $h(0) = -\sigma/c$ and $h(1) = 1+\sigma$, we see that h has two zeros in $(-1,1)$. We call $\frac{1}{\alpha_1}$ and $\frac{1}{\alpha_2}$ those (real) zeros. This implies that $f_{n,\alpha_1,\alpha_2}(w)\left(1+\frac{\sigma}{w^n}\frac{1-cw}{w-c}\right)$ is an H^∞ function, but this is not a polynomial. As seen in Remark 3.3.4, the Hankel operator with symbol $\frac{1-cw}{w-c}\frac{w-\alpha_1}{1-\overline{\alpha_1}w}\frac{w-\alpha_2}{1-\overline{\alpha_2}w}$ has norm one, and thus we conclude that for even n, the norm in Nehari's theorem is not reached by a polynomial. So we cannot use this theory for $d \geq 2$.

Inspired by our analysis for one outlier, we give a proof of Theorem 3.1.4 which is similar to the proof of Proposition 3.3.6.

Proof of Theorem 3.1.4 : Let us consider the polynomials

$$p_{n_j,\alpha_j}(w) = -\frac{w-\alpha_j}{\overline{\alpha}_j^{n_j}} \sum_{p=0}^{n_j-1} (\overline{\alpha}_1 w)^p = f_{n_j,\alpha_j}(w)\left(1 - \frac{1}{(\overline{\alpha}_j w)^{n_j}}\right), \quad j = 1:d,$$

defined in (3.6), and

$$p_n(w) = w^{n-n_1-\cdots-n_d} \prod_{j=1}^{d} p_{n_j,\alpha_j}(w) = f_{n,\alpha_1,\ldots,\alpha_d}(w) \prod_{j=1}^{d}\left(1 - \frac{1}{(\overline{\alpha}_j w)^{n_j}}\right),$$

which is a polynomial of degree n. The preceding relations permit to deduce that

$$\|p_n\|_{\partial\mathbb{D}} \leq \prod_{j=1}^{d}\left(1 + \frac{1}{|\alpha_j|^{n_j}}\right)$$

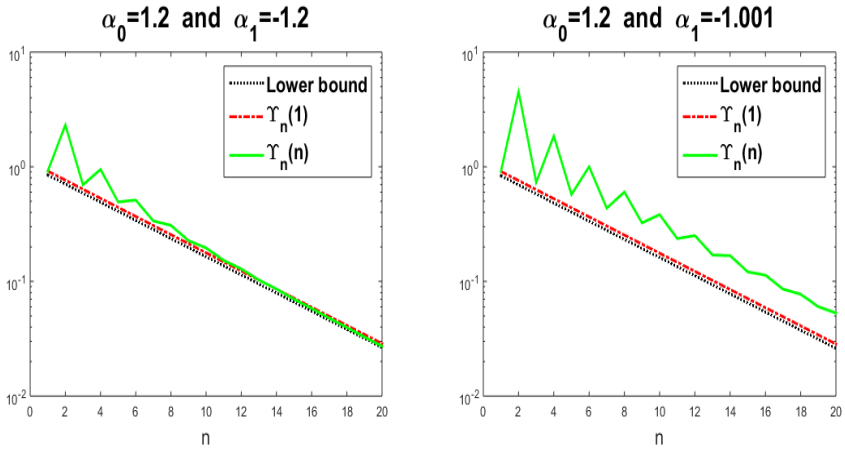

Figure 3.4: *Comparison between the cases $n_1 = 1$ and $n_1 = n$, in the case of a disk with one outlier, and with $\theta_1 \equiv \theta_0 + \Pi \pmod{2\pi}$.*

and

$$|p_n(\alpha_0)| = |f_{n,\alpha_1,\dots,\alpha_d}(\alpha_0)| \prod_{j=1}^{d} \left| 1 - \frac{1}{(\overline{\alpha}_j \alpha_0)^{n_j}} \right|,$$

and the upper bound follows.

□

The choice of the n_j clearly plays a role in the convergence of our upper bound. We will present two choices. The first one consists in taking $n_j = n/d$ when d divides n, and in adding one to the n_j in a certain order when d does not divide n. For example we can add one to the n_j that correspond to the larger $|\alpha_j|$. This choice will be referred to as the distributed choice because we give the same importance to each outlier when d divides n.
For the second choice, using the overestimation discussed in Remark 3.3.7, we can also obtain

$$\Upsilon_n(n_1, \dots, n_d) \leq \frac{1}{|f_{n,\alpha_1,\dots,\alpha_d}(\alpha_0)|} \prod_{j=1}^{d} \frac{1 + \frac{1}{|\alpha_j|^{n_j}}}{1 - \frac{1}{|\alpha_j|^{n_j}}}.$$

The lower bound does not depend on the choice of the n_j, and thus we would like to compute the n_j that minimize the product, but it seems hard to obtain a formula for those n_j. Instead, motivated by the fact that $(1+\epsilon)/(1-\epsilon) \sim 1+2\epsilon$ for ϵ near zero, we will choose the n_j that nearly minimize the quantity $\sum_{j=1}^{d} \frac{1}{|\alpha_j|^{n_j}}$.

Lemma 3.3.8 *For fixed α_j and under the constraint $\sum_{j=1}^{d} x_j = n$, the quantity*

$\sum_{j=1}^{d} \frac{1}{|\alpha_j|^{x_j}}$ *is minimized by the following* x_j*'s:*

$$x_j = \frac{1}{\log|\alpha_j| \sum_{p=1}^{d} \frac{1}{\log|\alpha_p|}} \left(n + \sum_{p=1}^{d} \frac{\log\log|\alpha_j| - \log\log|\alpha_p|}{\log|\alpha_p|} \right). \tag{3.8}$$

Proof of Lemma 3.3.8 : For $x_j \in \mathbb{R}_+$, consider the function $g(x) = g(x_1, \ldots, x_d) = \sum_{j=1}^{d} \frac{1}{|\alpha_j|^{x_j}}$, and the constraint $h(x) = h(x_1, \ldots, x_d) = n - \sum_{j=1}^{d} x_j$, we can find a local minimum of g subject to the constraint $h = 0$ by choosing the x_j's defined in (3.8). Indeed, if x is a local extremum of g subject to the constraint $h = 0$, there exists $c \in \mathbb{R}$ such that $\nabla g(x) = c\nabla h(x)$. By computing, we obtain $c = \frac{\log|\alpha_j|}{|\alpha_j|^{x_j}}$ and thus $x_j = \frac{\log\log|\alpha_j| - \log c}{\log|\alpha_j|}$. The condition $\sum_{j=1}^{d} x_j = n$ implies

$$\log(c) = \frac{\sum_{p=1}^{d} \frac{\log\log|\alpha_p|}{\log|\alpha_p|} - n}{\sum_{p=1}^{d} 1/\log|\alpha_p|}.$$

This allows to conclude.

□

As n_j must be an integer, we can take the nearest integer of this quantity. If needed we have to adjust the n_j such that their sum is equal to n. This case is refered to as the Lagrange case. If $|\alpha_0| \simeq 1$, this choice of n_j will lead to a better choice than the distributed choice.

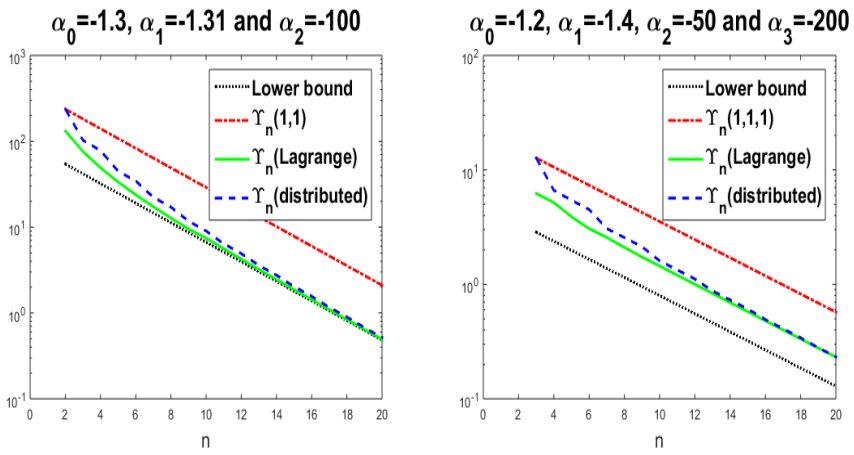

Figure 3.5: *Comparison between different choices of* n_j*.*

In Figure 3.5 we compare the two choices presented with the upper bound given in [24, Proposition 4.1] by (3.3) and the lower bound. If the absolute values of the α_j differs a lot, the Lagrange choice is better at the beginning, but up to a certain step, it seems that the two choices proposed are very near.

As in the case of one outlier, when the α_j $(j = 0 : d)$ are on the same straight line passing through zero and on the same side with respect to zero, we have

$$\Upsilon_n(n_1, \ldots, n_d) = \frac{1}{|f_{n,\alpha_1,\ldots,\alpha_d}(\alpha_0)|} \prod_{j=1}^{d} \frac{1 + \frac{1}{|\alpha_j|^{n_j}}}{1 - \frac{1}{|\alpha_0\alpha_j|^{n_j}}}.$$

In general, we can obtain similar behaviors as in the case with one outlier. Note that each term in the product $\prod_{j=1}^{d} \frac{1+\frac{1}{|\alpha_j|^{n_j}}}{1-\frac{1}{|\alpha_0\alpha_j|^{n_j}}}$ tends to one, but more we have outliers, more the ratio between the upper and lower bounds will take time to tend to one.

An interesting question is to ask if it is advantageous to take more outliers. The product $\prod_{j=1}^{d} \frac{1+\frac{1}{|\alpha_j|^{n_j}}}{|1-\frac{1}{(\overline{\alpha_j}\alpha_0)^{n_j}}|}$ tends to one and thus will not interfere with the rate of convergence. On the other side, the term $\frac{1}{|f_{n,\alpha_1,\ldots,\alpha_d}(\alpha_0)|}$ depends strongly on the choice of S as we have already seen in Section 3.2 concerning the lower bound, in particular in Figure 3.1. Thus we have the same behavior as for the lower bound. For a fixed d, we have obtained an upper bound in terms of a straight line which clearly decreases with d. By taking the minimum over all d, we hope to capture superlinear convergence through the (concave) lower envelope of these straight lines. The problem is that we do not know how to choose the best d. This question of choosing an optimal d for given n is closely related to how many eigenvalues could be considered as outliers since they are well approached by Ritz values.

We can see the improvement taking more outliers in Figure 3.6 on the left. We do not need any information on the localization of the eigenvalues in the disk to plot this graph. To obtain better rates of convergence, we need to know more outliers, and thus we need more information on the spectrum.

An important fact in all the bounds presented is that they all have a linear convergence (asymptotically they have the behavior of a straight line). But it is well-known that Krylov methods often has a so-called superlinear convergence behaviour (see Section 2.3). This is illustrated in Figure 3.6 on the right, where the graph is the same as on the left, but we add the plot of GMRES $\|r_n\|/\|r_0\|$, with starting vector zero. We do not consider the constant C_S in our graphs. The gap between the plots of GMRES and the upper bounds in Figure 3.6 on the right is due to the fact that considering a disk is not a great idea to enclose 95 points. This is why it can be interesting to consider more general inclusion sets S. As an example, we have plotted the upper bound given in Theorem 3.1.5 by considering an ellipse and one outlier.

So an idea to obtain better error bounds is to choose sets that better follow the shape of the spectrum. We can also be in the case where a disk prevent to consider eigenvalues as outliers but other shapes of sets could.

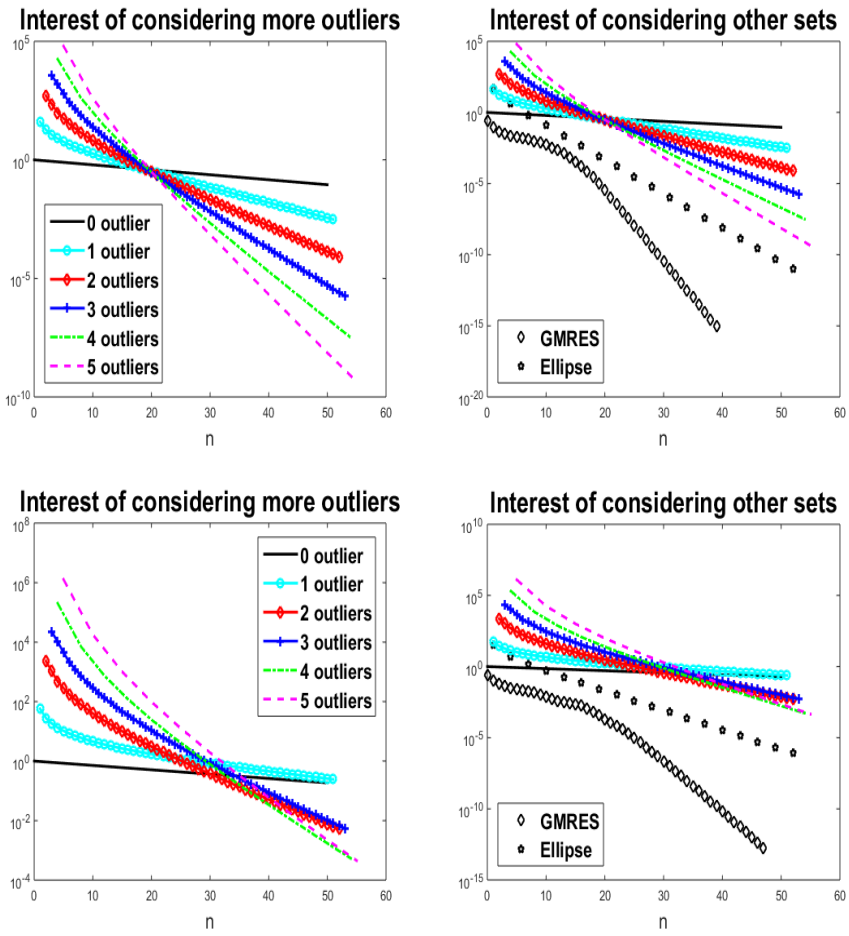

Figure 3.6: *Here we present the interest of considering several outliers for a matrix of size* 100, $b = (1, \ldots, 1)^T$ *and* $x_0 = 0$. *In the top row on the left, we consider that the matrix has eigenvalues* $\lambda_1 = 1$, $\lambda_2 = 3$, $\lambda_3 = 5$, $\lambda_4 = 7$, $\lambda_5 = 9$, *and all the others in the disk* $D(26, 15)$. *Below on the left, we consider that the matrix has eigenvalues* $\lambda_1 = 1$, $\lambda_2 = 1.3$, $\lambda_3 = 3.8$, $\lambda_4 = 4.1$, $\lambda_5 = 5.6$, *and all the others in the disk* $D(23, 17)$. *On the right we suppose that we know all the eigenvalues such that we can plot GMRES and compare with an ellipse and one outlier. In those examples all the eigenvalues are on the real axis.*

3.4 Convex inclusion set and one outlier

The choice of S is a delicate problem which depends on our knowledge of the spectrum. For example, in [11, Theorem 2], the author supposed that S contains the field of values and $0 \notin S$, and obtained the bound

$$\frac{\|r_n^{MR}\|}{\|r_0\|} \leq \frac{1}{|\phi_S(0)|^n} \frac{2}{1 - \frac{1}{|\phi_S(0)|^{n+1}}}.$$

In the case we consider zero outlier and S is a disk, we have the similar upper bound

$$\frac{\|r_n^{MR}\|}{\|r_0\|} \leq C_S \frac{1}{|\phi_S(0)|^n}.$$

To generalize our idea to convex sets, we make use of Faber polynomials, which give a polynomial in $w = \phi_S(z)$. Unfortunately, when S is not a disk, after the change of variable, we do not find a polynomial in z (ϕ_S is not linear). To work around this difficulty, we use the Faber transform.

3.4.1 Faber polynomials

In this section we suppose that S is convex, and we write $\phi_S = \phi$ and $\psi_S = \psi$ (we forget the dependence on S) for the functions given by the Riemann mapping theorem. The p-th Faber polynomial of S denoted by F_p (or F_p^S if S is not explicit) is defined as the polynomial part of the Laurent expansion at infinity of ϕ^p

$$F_p(z) = \frac{1}{c^p} z^p + a_{p-1}^{(p)} z^{p-1} + \cdots + a_1^{(p)} z + a_0^{(p)}.$$

Good surveys of their properties are provided by [50] or [112]. Let us give three fondamental examples of Faber polynomials.

If S is a disk $D(\gamma, r)$ we are in the case discussed before and we have $\phi(z) = \frac{z-\gamma}{r}$ which is a polynomial, and then

$$F_p(z) = \phi(z)^p = \left(\frac{z-\gamma}{r}\right)^p.$$

If $S = [-1, 1]$, it is known that ψ is the Joukowski function

$$\psi(w) = \frac{1}{2}\left(w + \frac{1}{w}\right)$$

and

$$\phi(z) = z + \sqrt{z^2 - 1}$$

where the branch of the root is taken so that the condition $\lim_{z\to\infty} \frac{1}{z}\sqrt{z^2-1} = 1$ holds. We can prove that

$$F_0(z) = T_0(z) \quad \text{and} \quad F_p(z) = 2T_p(z),\ p \geq 1,$$

where T_p denotes the p-th Chebyshev polynomial of the first kind, i.e. $T_p(x) = \cos(p \arccos x)$.

If S is an ellipse with foci at the points ± 1 and with semi-axes $a = \frac{1}{2}(R + \frac{1}{R})$ and $b = \frac{1}{2}(R - \frac{1}{R})$ where $R > 1$, then

$$\psi(w) = \frac{1}{2}\left(Rw + \frac{1}{Rw}\right)$$

and

$$F_p(z) = \frac{2}{R^{2p}} T_p(z) \text{ for } p \geq 1.$$

If we know that

$$\psi(z) = cz + c_0 + \frac{c_1}{z} + \ldots,$$

we can compute the Faber polynomials consecutively by the following recursion relation [112, Eqn (II.2.6)]

$$cF_{p+1} = zF_p(z) - pc_p - \sum_{s=0}^{p} c_s F_{p-s}(z), \text{ for } p \geq 2.$$

We define the level curve

$$\Gamma_R = \{z : |\phi(z)| = R\} = \{\psi(\omega) : |\omega| = R\}$$

and the inclusion set

$$S_R = \mathbb{C} \setminus \{z \in \mathbb{C} \setminus S : |\phi(z)| > R\}.$$

The conformal map ϕ_R associated to S_R is easy to obtain, indeed we have the relation $\phi_R(z) = \phi(z)/R$. For ψ_R we have the relation $\psi_R(w) = \psi(Rw)$.

Now let us define the Faber transform [50, I.6.C] which is a linear bijection from $\Pi_k(\mathbb{D})$ to $\Pi_k(S)$:

$$\begin{array}{rccc} \mathcal{F} & : & \Pi_k(\mathbb{D}) & \to & \Pi_k(S) \\ & & \sum_{j=0}^{k} a_j w^j & \mapsto & \sum_{j=0}^{k} a_j F_j(z). \end{array}$$

This map transforms polynomials in w into polynomials in z. If $\mathcal{F}$ is continuous, then $\mathcal{F}$ admits a unique extension that is continuous from $\mathbb{A}(\mathbb{D})$ to $\mathbb{A}(S)$, where $\mathbb{A}(S)$ denotes the Banach algebra of functions analytic in the interior of S and continuous on S, equipped with the uniform norm. This is for example the case for convex sets [50, Theorem 2 page 48-49].

3.4.2 Upper bound for a convex inclusion set and one outlier

Our polynomial in the case S is convex is inspired by the polynomial p_{n,α_1} given in (3.6) which solves the minimization problem related to AAK theory. We obtain Theorem 3.1.5 using the modified Faber transform

$$\begin{array}{rccc} \mathcal{F}_+ & : & \mathbb{A}(\mathbb{D}) & \to & \mathbb{A}(S) \\ & & f & \mapsto & \mathcal{F}(f) + f(0). \end{array}$$

To prove this theorem we need the following lemma which can be found in [74].

Lemma 3.4.1 *For convex sets S, we have*

$$Re\left(\frac{u\psi'(u)}{\psi(u)-\psi(w)}-\frac{1}{2}\frac{u+w}{u-w}\right)\geq 0$$

for $|u|>1$ and $|w|=1$.

Proof : As the level sets S_R are convex, by a geometric argument, we can prove that for $|u|=R\geq 1$ and all $z\in S_R$, $z\neq\psi(u)$

$$Re\left(\frac{u\psi'(u)}{\psi(u)-z}\right)\geq 0.$$

In particular, for $|u|\geq|w|=1$, $u\neq w$ we have

$$Re\left(\frac{u\psi'(u)}{\psi(u)-\psi(w)}\right)\geq 0.$$

It is not hard to see that $\frac{u+w}{u-w}\in i\mathbb{R}$ for $|u|=|w|=1$, $u\neq w$. Setting $h(u,w)=\frac{u\psi'(u)}{\psi(u)-\psi(w)}-\frac{1}{2}\frac{u+w}{u-w}$, we have that h is analytic in u for $|u|>|w|=1$. Thus by the minimum principle

$$\min_{|u|>|w|=1} Re\left(h(u,w)\right)=\min_{|u|=|w|=1} Re\left(h(u,w)\right)\geq 0.$$

□

Now we are able to give a proof for Theorem 3.1.5.

Proof of Theorem 3.1.5 : The first step is to prove that for every polynomial $p(w)=\sum_{j=0}^{l}a_jw^j$ we have

$$\|\mathcal{F}_+(p)\circ\psi-p\|_{\partial\mathbb{D}}\leq\|p\|_{\partial\mathbb{D}},\tag{3.9}$$

which was already proved in [16, Theorem 3.4]. Let us give another proof here by working on the level sets S_R. As the Faber polynomials $F_{n,R}$ associated to S_R are related to the Faber polynomials associated to S via the formula $F_{n,R}(z)=F_n(z)/R^n$, we obtain for the Faber transfrorm $\mathcal{F}_R$ associated to S_R for $z\in int(S_R)$

$$\begin{aligned}\mathcal{F}_R(p)(z)&=\sum_{j=0}^{l}a_jF_{j,R}(z)=\sum_{j=0}^{l}a_jF_j(z)/R^j=\frac{1}{2\pi}\int_{|u|=1}p(u)\frac{u\psi_R'(u)}{\psi_R(u)-z}\frac{\mathrm{d}u}{iu}\\&=\frac{1}{2\pi}\int_{|y|=R}p(y/R)\frac{y\psi'(y)}{\psi(y)-z}\frac{\mathrm{d}y}{iy},\end{aligned}\tag{3.10}$$

where in the third inequality we have used [50, Eqn (I.6.16)]. The function

$$g(u)=\overline{p(1/\overline{u})}$$

is analytic in $\overline{\mathbb{C}} \setminus \mathbb{D}$ and continuous on $|u| \geq 1$, with $g(u) = \overline{p(u)}$ for $u \in \partial\mathbb{D}$. Noting that

$$\frac{1}{2i\pi}\int_{|u|=1} \overline{p(u)}\frac{\psi_R'(u)}{\psi_R(u)-z}\mathrm{d}u = \frac{1}{2i\pi}\int_{|u|=1} g(u)\frac{\psi_R'(u)}{\psi_R(u)-z}\mathrm{d}u = g(\infty) = \overline{p(0)},$$

where the second inequality follows from the residue theorem, we obtain

$$p(0) = \frac{1}{2\pi}\int_{|u|=1} p(u)\overline{\left(\frac{u\psi_R'(u)}{\psi_R(u)-z}\right)}\frac{\mathrm{d}u}{iu} = \frac{1}{2\pi}\int_{|y|=R} p(y/R)\overline{\left(\frac{y\psi'(y)}{\psi(y)-z}\right)}\frac{\mathrm{d}y}{iy} \tag{3.11}$$

Adding (3.10) and (3.11) yields for $z \in int(S_R)$

$$\mathcal{F}_{R,+}(p)(z) = \mathcal{F}_R(p)(z) + p(0) = \frac{1}{2\pi}\int_{|y|=R} p(y/R)\ 2Re\left(\frac{y\psi'(y)}{\psi(y)-z}\right)\frac{\mathrm{d}y}{iy}. \tag{3.12}$$

For $|w| = 1$, we have the Poisson integral formula [103, Section 0.4]

$$\begin{aligned} p(w/R) &= \frac{1}{2\pi}\int_{|u|=1} p(u)\ Re\left(\frac{u+w/R}{u-w/R}\right)\frac{\mathrm{d}u}{iu} \\ &= \frac{1}{2\pi}\int_{|y|=R} p(y/R)\ Re\left(\frac{y+w}{y-w}\right)\frac{\mathrm{d}y}{iy}. \end{aligned} \tag{3.13}$$

Combining the two integrals (3.12) and (3.13), we conclude that for $|w| = 1$

$$\mathcal{F}_{R,+}(p)(\psi(w)) - p(w/R) = \frac{1}{2\pi}\int\limits_{|y|=R} p(y/R)\ Re\left(2\frac{y\psi'(y)}{\psi(y)-\psi(w)} - \frac{y+w}{y-w}\right)\frac{\mathrm{d}y}{iy}. \tag{3.14}$$

This implies that

$$\max_{|w|=1}|\mathcal{F}_{R,+}(p)(\psi(w)) - p(w/R)| \leq \|p\|_{\partial\mathbb{D}} \frac{1}{2\pi}\int_{|y|=R} Re\left(2\frac{y\psi'(y)}{\psi(y)-\psi(w)} - \frac{y+w}{y-w}\right)\frac{\mathrm{d}y}{iy},$$

where we have dropped the modulus inside the integral thanks to Lemma 3.4.1. The relation (3.14) with $p = 1$ gives

$$\frac{1}{2\pi}\int_{|y|=R} Re\left(\frac{y\psi'(y)}{\psi(y)-\psi(w)} - \frac{1}{2}\frac{y+w}{y-w}\right)\frac{\mathrm{d}y}{iy} = 1.$$

Now by taking the limit as $R \to 1$, we obtain (3.9).

Let us consider the polynomial

$$t_n(z) = \mathcal{F}_+(p_{n,\alpha_1})(z) - \mathcal{F}_+(p_{n,\alpha_1})(\lambda_1),$$

with p_{n,α_1} defined in (3.6). The second step is to prove the inequality

$$\|t_n\|_S \leq 3\|p_{n,\alpha_1}\|_{\mathbb{D}} = 3(1+\frac{1}{|\alpha_1|^n}). \tag{3.15}$$

The function $\mathcal{F}_+(p_{n,\alpha_1})\circ\psi - p_{n,\alpha_1}$ is analytic in $\overline{\mathbb{C}}\backslash\mathbb{D}$. The fact that $p_{n,\alpha_1}(\alpha_1)=0$ and the maximum principle give

$$|\mathcal{F}_+(p_{n,\alpha_1})(\lambda_1)| = |\mathcal{F}_+(p_{n,\alpha_1})(\psi(\alpha_1)) - p_{n,\alpha_1}(\alpha_1)| \leq \|p_{n,\alpha_1}\|_{\mathbb{D}}. \tag{3.16}$$

Equations (3.9) and (3.16) permits to conclude that

$$\begin{aligned}
\|t_n\|_S &= \max_{|w|=1} |\mathcal{F}_+(p_{n,\alpha_1})(\psi(w)) - \mathcal{F}_+(p_{n,\alpha_1})(\lambda_1)| \\
&= \max_{|w|=1} |\mathcal{F}_+(p_{n,\alpha_1})(\psi(w)) - p_{n,\alpha_1}(w) + p_{n,\alpha_1}(w) - \mathcal{F}_+(p_{n,\alpha_1})(\lambda_1)| \\
&\leq 3\|p_{n,\alpha_1}\|_{\mathbb{D}} \leq 3(1+\frac{1}{|\alpha_1|^n})
\end{aligned}$$

as claimed in (3.15). The last inequality being a consequence of (3.7).

The third step is to prove

$$|t_n(0)| \geq |p_{n,\alpha_1}(\alpha_0)| \left(1 - \frac{2}{|\alpha_0|^n} \frac{1+\frac{1}{|\alpha_1|^n}}{|1-\frac{1}{(\alpha_0\overline{\alpha_1})^n}|}\right). \tag{3.17}$$

Noticing that

$$\frac{w^n(t_n(\psi(w)) - p_{n,\alpha_1}(w))}{f_{n,\alpha_1}(w)}$$

is analytic for $|w|>1$ including $w=\infty$ and $w=\alpha_1$, it follows by the maximum principle that

$$\begin{aligned}
\left|\frac{\alpha_0^n(t_n(\psi(\alpha_0)) - p_{n,\alpha_1}(\alpha_0))}{f_{n,\alpha_1}(\alpha_0)}\right| &\leq \max_{w\in\partial\mathbb{D}} \left|\frac{w^n(t_n(\psi(w)) - p_{n,\alpha_1}(w))}{f_{n,\alpha_1}(w)}\right| \\
&= \|t_n\circ\psi - p_{n,\alpha_1}\|_{\partial\mathbb{D}} \leq 2\|p_{n,\alpha_1}\|_{\partial\mathbb{D}},
\end{aligned}$$

where for the last inequality we used (3.9) and (3.16). By rearranging the terms and using (3.7) we have

$$|t_n(0) - p_{n,\alpha_1}(\alpha_0)| \leq \frac{|p_{n,\alpha_1}(\alpha_0)|}{|\alpha_0|^n|1-\frac{1}{(\alpha_0\overline{\alpha_1})^n}|} 2(1+\frac{1}{|\alpha_1|^n}),$$

which by the triangle inequality leads to (3.17).

In conclusion, if the right hand side of (3.17) is positive (for n sufficiently

large) we obtain the upper bound

$$\begin{aligned} E_n(S) &\leq \frac{\|t_n\|_S}{|t_n(0)|} \\ &\leq \frac{3(1+\frac{1}{|\alpha_1|^n})}{|p_{n,\alpha_1}(\alpha_0)|}\left[1-\frac{2}{|\alpha_0|^n}\frac{1+\frac{1}{|\alpha_1|^n}}{|1-\frac{1}{(\alpha_0\overline{\alpha_1})^n}|}\right]^{-1} \\ &\leq \frac{3}{|f_{n,\alpha_1}(\alpha_0)|}\frac{1+\frac{1}{|\alpha_1|^n}}{|1-\frac{1}{(\alpha_0\overline{\alpha_1})^n}|}\left[1-\frac{2}{|\alpha_0|^n}\frac{1+\frac{1}{|\alpha_1|^n}}{|1-\frac{1}{(\alpha_0\overline{\alpha_1})^n}|}\right]^{-1} \end{aligned}$$

which behaves when n tends to infinity asymptotically like $\frac{3}{|f_{n,\alpha_1}(\alpha_0)|}$.

□

Let us illustrate Theorem 3.1.5 in Figure 3.7. We consider the block tridiagonal matrix of order $n = p^2$ resulting from discretizing the 2D Poisson's equation on a square with the 5-point operator on an p-by-p mesh (given in Matlab by *gallery('poisson',p)*). This matrix has a spectrum on the real axis, thus a circle

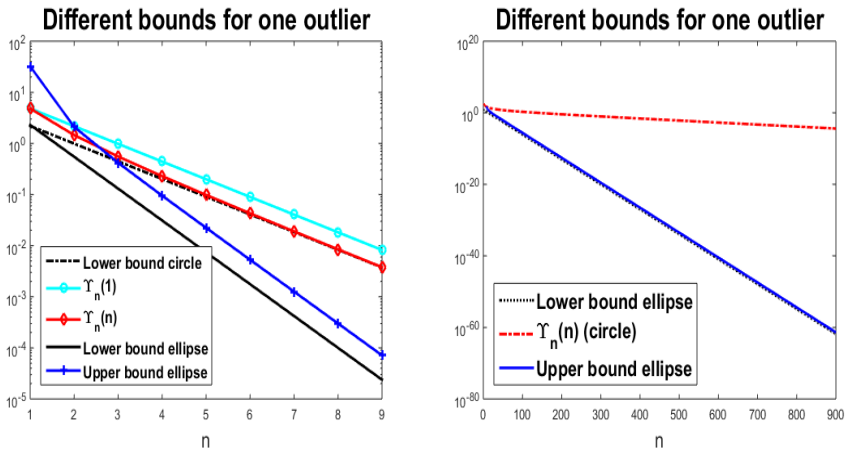

Figure 3.7: *We compare the bounds found for disks and for ellipses. The disks considered have center* $(\lambda_2+\lambda_n)/2$ *and radius* $(\lambda_n-\lambda_2)/2+10^{-5}$. *The ellipses have the same center as the disks, their semi major axis is equal to the radius of the disk, and their minor semi axis is set to* 10^{-5}.

is clearly a bad approximation of the spectrum. On the left we present for $p = 3$ the lower bounds for a disk and an ellipse and the upper bounds in [24] and in Theorems 3.4 and 3.1.5. On the right we compare for $p = 30$ the upper bounds for a disk and of an ellipse with the lower bound for an ellipse. We clearly have an improvement if we choose a good ellipse (an interval in limit case) instead of a circle.

Remark 3.4.2 *In Corollary 5.1.1 and Remark 5.1.2, we will see that in the*

case of an interval we can obtain an explicit universal constant independent of d.

3.5 Open problems

3.5.1 Convex inclusion set and several outliers

Applying the proof of Theorem 3.1.5 to the polynomial

$$t_n(z) = \prod_{j=1}^{d} (\mathcal{F}_+(p_{n_j,\alpha_j})(z) - \mathcal{F}_+(p_{n_j,\alpha_j})(\lambda_j)),$$

we obtain a similar conclusion as in Theorem 3.1.5 but with a constant 3^d, this is why this bound cannot be interesting for a lot of outliers. The factor 3 found in (3.15) is clearly not optimal, and we believe that we can obtain a factor near one (at least 2). If we forget this factor, we obtain the same behavior as in the case of a disk if we consider several outliers.

If $S = [\alpha, \beta]$ is an interval, and the outliers are in $(0, \alpha)$, we will find later (see Corollary 5.1.1 and Remark 5.1.2) a better result with a universal constant that does not depend on d.

3.5.2 More general inclusion sets

The convexity of S and the condition $0 \notin S$ are difficult to reconcile. We could consider polygons which can better fit the shape of the eigenvalues without the preceding conditions. Indeed, we can compute Faber polynomials for polygons by using the Schwarz-Christoffel toolbox developped by Driscoll in [33, 34].

3.5.3 Inclusion set with several connected components

For some problems, it can be interesting to consider the inclusion set S as a disjoint union of N connected subsets S_j of $\mathbb{C}$. In [24, Proposition 5.1], the authors generalized their result to several inclusion sets (called clusters) but the bounds given seriously deteriorates for multiple clusters.

It is an open problem to solve a minimization problem on a union of disconnected sets. An idea could be to divide the polynomial q_d into a product $q_{d,1} \dots q_{d,N}$ with the $q_{d,j}$ having their zeros (outliers) "near" the set S_j. And then to choose a polynomial $p_k = p_{k,1} \dots p_{k,N}$ in such a way that we try to minimize $\|q_{d,j} p_{k,j}\|_{S_j}$ with the preceding method. We make an error for each j and the error at the end can be very large. Moreover we have to clearly define what is "close" to the set S_j.

If we have a good polynomial for disks, an idea to generalize to more general sets could be to use the Faber-Walsh polynomials [112, chap 13]. Indeed, they allow to consider sets S consisting of N disjoint compact sets $S_1, \dots, S_N$, with the complement of S a N-connected open set containing the point ∞. Those

polynomials are a generalization of the Faber polynomials but it is difficult to work with them. We cite the papers [105] and [106] where we can find explicit examples of such polynomials and optimality properties.

3.5.4 Infinite dimensional analysis

We can extend the analysis of the previous section to infinite dimensional case for certain operators. Suppose $\mathcal{H}$ is a Hilbert space and A a bounded operator on $\mathcal{H}$. If $A = I + K$ is a sum of the identity and of a compact operator K, it has a countable sequence of eigenvalues with one as the only accumulation point. So we can consider the inclusion set as a disk centered at one, outliers outside this disk, and all our theory holds in this case. A major reference for the infinite dimensional analysis of Krylov methods is [88].

Chapter 4

On the sharpness of the weighted Bernstein-Walsh inequality

In this chapter we show that the weighted Bernstein-Walsh inequality in logarithmic potential theory is sharp up to some new universal constant, provided that the external field is given by a logarithmic potential. Our main tool for such results is a new technique of discretization of logarithmic potentials, where we take the same starting point as in earlier work of Totik and of Levin & Lubinsky, but add an important new ingredient, namely some new mean value property for the cumulative distribution function of the underlying measure. This work is a conjoint work done with Beckermann [19].

4.1 Introduction

We discuss in §4.1.1 the sharpness of the so-called weighted Bernstein-Walsh inequality for the particular case where the external field is the logarithmic potential of some measure. Here our main result in Theorem 4.1.3 indicates the existence of some new universal constant. Our main technical result stated and proved in §4.2 is Theorem 4.2.1 on a new fine discretization of logarithmic potentials for a suitable class of measures, where in contrast to preceding work of Totik, Lubinsky and others we get (large but) explicit constants. Here an essential tool is a new mean value property stated in Theorem 4.2.6.

4.1.1 The weighted Bernstein-Walsh inequality

Given a finite union of compact intervals $\Sigma \subset \mathbb{R}$, we denote by $\mathcal{M}_1(\Sigma)$ the set of Borel measures μ with support $\mathrm{supp}(\mu)$ in Σ and of total mass 1, and consider the logarithmic potential U^μ and energy $I(\mu)$ (see Definition A.1.3).

Given a weight w defined on $\mathbb{R}$ and continuous on Σ (and thus admissible, see Definition A.2.1) together with an external field $Q(x) = -\log(w(x))$, it is known [103, Theorem I.1.3 and Theorem I.4.8] that there is a unique minimizer $\mu_w \in \mathcal{M}_1(\Sigma)$ of the extremal problem

$$\inf\{I(\mu) + 2\int Q\mathrm{d}\mu : \mu \in \mathcal{M}_1(\Sigma)\} \tag{4.1}$$

which is uniquely characterized by the existence of a constant $F \in \mathbb{R}$ such that

$$\Theta(x) := F - U^{\mu_w}(x) - Q(x) \begin{cases} = 0 & \text{for } x \in \mathrm{supp}(\mu_w), \\ \leq 0 & \text{for } x \in \Sigma. \end{cases} \tag{4.2}$$

Logarithmic potential theory with external fields has been applied with success for getting asymptotics for various polynomial extremal problems [103], maybe one of the most prominent results being the weighted Bernstein-Walsh inequality [103, Theorem III.2.1]

$$\forall x_0 \in \mathbb{R} \quad \forall P \in \Pi_k : \quad \frac{|w(x_0)^k P(x_0)|}{\|w^k P\|_{\mathrm{supp}(\mu_w)}} \leq e^{k\Theta(x_0)}, \tag{4.3}$$

and its sharpness, see, e.g., [103, Corollary III.1.10],

$$\exists P_k \in \Pi_k \quad \forall x_0 \in \mathbb{R} \setminus \mathrm{supp}(\mu) : \quad \lim_{k\to\infty} \left(\frac{|w(x_0)^k P_k(x_0)|}{\|w^k P_k\|_{\mathrm{supp}(\mu_w)}} \right)^{1/k} = e^{\Theta(x_0)}, \tag{4.4}$$

where Π_k denotes the set of polynomials of degree at most k, and $\|f\|_\Sigma = \max_{x\in\Sigma} |f(x)|$. One aim of this chapter is to improve (4.4) for a particular class of external fields, see Theorem 4.1.3 below, namely to show that (4.3) is sharp up to some constant. Before giving some more details, let us first have a look at other classes of external fields where such constants are explicitly known. In what follows we will write $g_S(\cdot, \zeta)$ to denote the Green function in $\overline{\mathbb{C}} \setminus S$ for a compact set $S \subset \mathbb{R}$ with pole at $\zeta \in \overline{\mathbb{C}} \setminus S$ (see Definition A.1.4 and Equation (A.1)). We will be mainly interested in the special case of an interval S where the Green function vanishes on S and is strictly positive outside S, and where explicit formulas are available.

Example 4.1.1 *Consider $\Sigma = [\alpha, \beta]$ and $Q = 0$, then an explicit formula is known for the minimizer in (4.1) denoted by $\omega_{[\alpha,\beta]}$ and called Robin equilibrium measure of the interval $[\alpha, \beta]$ (see Eqn. (A.5))*

$$\mathrm{supp}(\omega_{[\alpha,\beta]}) = \Sigma = [\alpha, \beta], \quad \frac{\mathrm{d}\omega_{[\alpha,\beta]}}{\mathrm{dx}}(\mathrm{x}) = \frac{1}{\pi\sqrt{(\mathrm{x}-\alpha)(\beta-\mathrm{x})}}.$$

It is also known from, e.g., [103, Eqn. (I.4.8)] that $\Theta(z) = g_{[\alpha,\beta]}(z, \infty)$, and thus (4.3) becomes the classical Bernstein-Walsh inequality. Taking $P_k(x) = T_k(\frac{2x-\alpha-\beta}{\beta-\alpha})$ with T_k the Chebyshev polynomial of the first kind, one may also show that (4.3) is sharp up to a factor $1/2$.

Example 4.1.2 *Consider* $\Sigma = [\alpha, \beta]$, *and* $w(x)^k = 1/\sqrt{q(x)}$ *with* q *being a polynomial of degree* $\ell \leq 2k$, *strictly positive on* $[\alpha, \beta]$, *compare with [82, chap 4.4]. Thus* $Q = -U^\rho$ *with* ρ *an atomic measure of mass* $\ell/(2k) \leq 1$. *Here the extremal measure* μ_w *in* (4.1), (4.2) *is given in [103, Example II.4.8] in terms of balayage (Section A.4) onto* $\mathrm{supp}(\mu_w) = \Sigma$, *and it follows from [103, Eqn. (II.4.32)] that*

$$\Theta(x) = (1 - \frac{\ell}{2k}) g_{[\alpha,\beta]}(x, \infty) + \int g_{[\alpha,\beta]}(x, y) \mathrm{d}\rho(y).$$

Moreover, with help of the factorization

$$\widetilde{q}(y)\widetilde{q}(\frac{1}{y}) = q(x), \quad \frac{2x - \alpha - \beta}{\beta - \alpha} = \frac{1}{2}(y + \frac{1}{y}) \in \mathbb{R},$$

$|y| \geq 1$, *the polynomial* $\widetilde{q}$ *of degree* ℓ *having all its roots outside the unit circle, it is known that* P_k *defined by*

$$w(x)^k P_k(x) = \frac{1}{2}(e^{k\Theta(x)} + e^{-k\Theta(x)}), \quad e^{2k\Theta(x)} = \frac{y^{2k}\widetilde{q}(\frac{1}{y})}{\widetilde{q}(y)},$$

is a polynomial of degree k, *showing that again* (4.3) *is sharp up to a factor* $1/2$.

We are interested in the case where the external field is a positive potential $U^{\rho/k}$ (not necessarily of an atomic measure), for instance if w^k is a (power of a) polynomial. This includes the particular case $w(x) = |x|^\theta$ on $\Sigma = [0, 1]$ for $\theta > 0$, starting point of an important research area about incomplete polynomials [103, §VI.1.1]. For external fields being a positive potential, we recall below how to solve the extremal problem, including the well-known pushing effect that the support of the equilibrium measure may be a proper subset of Σ. We then state our main result on the sharpness of the weighted Bernstein-Walsh inequality.

Theorem 4.1.3 *Let* $k \geq 1$ *be some integer, and* $Q = U^{\rho/k}$ *on* $\Sigma = [\alpha, \beta]$, *with the Borel measure* ρ *being compactly supported on* $(-\infty, \alpha]$. *Consider on* Σ *the strictly decreasing function*

$$\eta(z) := \int \sqrt{\frac{\beta - y}{z - y}} \mathrm{d}\rho(y), \tag{4.5}$$

and set $a = \alpha$ *if* $k + ||\rho|| \geq \eta(\alpha)$, *and else denote by* a *the unique solution of* $k + ||\rho|| = \eta(a)$ *in* Σ. *Then the extremal measure in* (4.1), (4.2) *is given by*

$$\mathrm{supp}(\mu_w) = [a, \beta], \quad k\Theta(x) = (k + \|\rho\|) g_{[a,\beta]}(x, \infty) - \int g_{[a,\beta]}(x, y) \mathrm{d}\rho(y). \tag{4.6}$$

Moreover, the weighted Bernstein-Walsh inequality (4.3) *is sharp up to some constant, that is, there exists a universal real constant* $C_{BW} > 0$ *such that, for all* $k \geq 2$, *we may construct a polynomial* P_k *of degree* k *such that, for all* $x_0 \in \mathbb{R} \setminus [a, \beta]$,

$$\frac{|w(x_0)^k P_k(x_0)|}{\|w^k P_k\|_{\mathrm{supp}(\mu_w)}} \geq e^{-C_{BW}} e^{k\Theta(x_0)}. \tag{4.7}$$

Our proof of Theorem 4.1.3 presented in §4.3 is based on a fine discretization of the logarithmic potential $U^{k\mu_w}$. We will show in this chapter that $C_{BW} \leq 15383$, but this is by no means optimal. The most remarkable fact for us seems to be that such a constant does not depend on the data ρ, a, β nor on k. In particular, we do not need any further assumptions on smoothness of ρ, which is probably required by other techniques like a Riemann-Hilbert approach (which in any case would only allow to discuss asymptotics).

4.1.2 Structure of this chapter

The reminding of this chapter is organized as follows. Section 2 contains our results on discretizing the logarithmic potential of a class of measures including the extremal measure of Theorem 4.1.3. We first state our main Theorem 4.2.1, and then report in §4.2.1 about related results of Totik and of Lubinsky, and about the link with weighted quadrature formulas. Subsequently, we give in §4.2.2 the structure of the proof of Theorem 4.2.1, where following Totik we write the discretization error as a sum of three sums. We then state our original approach for dealing with these three sums, namely the mean value property of Theorem 4.2.6, and describe in §4.2.3 how to bound each of the three sums, with explicit constants.

In Section 4.3 we explain how to deduce Theorem 4.1.3 from Theorem 4.2.1. Our (quite technical) proof of Theorem 4.2.6 is postphoned to Section 4.4, and in Section 4.5 we gather some further technical results for dealing with our three sums. We end this chapter by giving some concluding remarks.

4.2 Discretization of a potential

Our proof of Theorem 4.1.3 is based on the approximation of kU^{μ_w} with μ_w the equilibrium measure as in Theorem 4.1.3 by $-\log|P_k(z)|$ with P_k a suitable monic polynomial of degree k. We will show the following.

Theorem 4.2.1 *Consider a measure $\mu \in \mathcal{M}_1([a,\beta])$ which has the density*

$$k\frac{d\mu}{dx}(t) = g(t)\frac{k}{\pi\sqrt{(t-a)(\beta-t)}}$$

for a function g which is non negative, concave and increasing[1] *on (a,β), such that $t \mapsto \frac{g(t)}{t-a}$ is convex on (a,β). Then there exists a universal explicit constant C_{BW} such that for each $k \geq 2$ we may construct a monic polynomial P_k of degree k such that*

(a) $\forall z \in \mathbb{C}$: $\log|P_k(z)| + kU^{\mu}(z) \leq C_{BW}$,

(b) $\forall x \in \mathbb{R} \setminus (a,\beta) : \log|P_k(x)| + kU^{\mu}(x) \geq 0$.

[1] In particular, g is continuous and bounded on (a,β), thus we may extend g to become a continuous, non-negative, concave and increasing function in $[a,\beta]$.

We will show in the proof of Theorem 4.1.3 that the extremal measure μ_w of Theorem 4.1.3 satisfies the assumptions of Theorem 4.2.1.

Example 4.2.2 *Another class of functions g satisfying the assumptions of Theorem 4.2.1 for* $[a,\beta]=[-1,1]$ *is given by*

$$\begin{aligned} g(x) &= (x+1)^{\theta}\pi/\int_{-1}^{1}(t+1)^{\theta-1/2}(1-t)^{-1/2}\mathrm{d}t \\ &= \frac{\pi}{2^{\theta}}\frac{\Gamma(\theta+1)}{\Gamma(1/2)\Gamma(\theta+1/2)}(x+1)^{\theta} \end{aligned}$$

for $\theta\in[0,1]$.

We will describe in §2.1 related work for discretizing potentials under various assumptions, but here the constants in general depend on μ, see for instance [103, §VI.4] for a summary. In §2.2 we give a proof of Theorem 4.2.1, where we initially follow the approach of Totik in [114, §2 and §9], see also the very accessible reference [80, Method 1] for the particular case $g(t)=2t$ on $[a,\beta]=[0,1]$ (up to a quadratic change of variables). Subsequently, we give in §2.3 a proof of three upper bounds we used in §2.2. Since the general case follows from a linear change of variables, we will suppose in what follows that $[a,\beta]=[-1,1]$ in Theorem 4.2.1.

4.2.1 How to discretize a potential?

It is natural to approach the logarithmic potential $U^{\mu}(x)=\int\log(1/|x-t|)\mathrm{d}\mu(t)$ by a quadrature rule of the form

$$\sum_{j=0}^{k-1}\log\frac{1}{|x-\xi_j|}=-\log|P_k(x)|,\quad P_k(x)=\prod_{j=0}^{k-1}(x-\xi_j), \tag{4.8}$$

for instance a weighted rectangular or midpoint rule, where we first cut $[-1,1]$ into k subintervals $[t_j,t_{j+1}]$, $-1=t_0<t_1<\ldots<t_k=1$, of equal mass $\mu([t_j,t_{j+1}])=1/k$, and chose $\xi_j\in[t_j,t_{j+1}]$ for $j=0,\ldots,k-1$. As long as $x\notin[-1,1]$ and the density of μ does not vary too much, we may bound the error $kU^{\mu}(x)+\log|P_k(x)|$ above and below, and may even show convergence to 0 for $k\to\infty$ for suitable choices of ξ_j. In our case we have the additional difficulties that the density of μ may have singularities at ±1, showing that the interval lengths $t_{j+1}-t_j$ may strongly vary in size for $j=0,1,\ldots,k-1$, and in addition in case $x\in[-1,1]$ we have to deal with a logarithmic singularity of the integrand.

Totik in [80, Method 1] used the weighted midpoint rule

$$\xi_j=\int_{t_j}^{t_{j+1}}t\mathrm{d}\mu(t)/\int_{t_j}^{t_{j+1}}\mathrm{d}\mu(t)=k\int_{t_j}^{t_{j+1}}t\mathrm{d}\mu(t) \tag{4.9}$$

for $j=0,1,\ldots,k-1$. In the particular case $[a,b]=[0,1]$ and $g(t)=2t$, a proof of Theorem 4.2.1 can be found in [80, §2], which strongly relies on the explicit

knowledge of asymptotics for the points ξ_j and t_j as a function of j and k for $k \to \infty$, and thus on the explicit knowledge of μ. In [103, Theorem VI.4.2] (see also the related result [114, Lemma 9.1] where the roots of P_k are slightly shifted into the complex plane), Totik considered probability measures μ with densities which are continuous up to a finite number of singularities of the form $|t-a_j|^{\delta_j}$ for $\delta_j > -1$. These assumptions are true in the setting of Theorem 4.2.1. He then shows the existence of (non explicit) constants $C_{T,1}, C_{T,2}$ depending on μ but not on k such that, for all $x \in \mathbb{R}$,

$$\log|P_k(x)| + kU^\mu(x) \leq C_{T,1},$$
$$\log|P_k(x)| + kU^\mu(x) \geq C_{T,2} + \max\Big\{0, \log(\mathrm{dist}(x, \{\xi_0, ..., \xi_{k-1}\}))\Big\}.$$

We see that the first inequality is as in Theorem 4.2.1(a), whereas the second one is clearly weaker than Theorem 4.2.1(b) for $x \in \mathbb{R} \setminus (-1,1)$ close to $[-1,1]$, since we get an additional term $\log(1/k^\beta)$ for some $\beta > 0$. Again, a proof of these statements uses heavily asymptotics for the points ξ_j and t_j as a function of j and k for $k \to \infty$, and thus quite a bit of information on μ.

Another technique of discretization has been considered by Lubinsky & Levin in [78] and [79], see also the very accessible reference [80, Method 2] for the particular case $g(t) = 2t$ on $[a,b] = [0,1]$ (up to a quadratic change of variables). With $t_0, ..., t_k$ as before, consider intermediate abscissa $t_{j+1/2} \in (t_j, t_{j+1})$ such that all intervals $[t_{j/2}, t_{(j+1)/2}]$ have the same mass $1/(2k)$. Given $x_0 \in \mathbb{R}$, the authors then apply trapezian rule on most of the subintervals $[t_{j-1/2}, t_{j+1/2}]$ corrected with suitable rectangle rules on the remaining 2 or 3 subintervals such that $\{\xi_0, ..., \xi_{k-1}\} \subset \{\pm 1, t_{1/2}, t_{3/2}, ..., t_{k-1/2}\}$. Up to a (quadratic) change of variables, the authors of [78, Theorem 9.1] suppose that

$$\frac{d\mu}{dx}(t) = \frac{(t+1)h(t)}{\pi\sqrt{1-t^2}}$$

with h continuous and > 0 on $[-1,1]$, and the modulus of continuity satisfies that $\log(1/\delta)\omega(h,\delta)$ is bounded above by some $\Gamma > 0$ for $\delta \in (0,1)$. In this case, for all $x \in \mathbb{R}$,

$$\log|P_k(x)| + kU^\mu(x) \leq C_{LL,1},$$
$$\log|P_k(x_0)| + kU^\mu(x_0) \geq C_{LL,2},$$

where $C_{LL,1}, C_{LL,2}$ are (non explicit) constants depending only on Γ and the minimum and maximum of h on $[-1,1]$. Note that the assumptions of [78, Theorem 9.1] and those of Theorem 4.2.1 are different and do not imply each other, see for instance Example 4.2.2 for $\theta < 1$. However, the above inequalities are quite close to those of Theorem 4.2.1, though our constants do not depend on μ, and our P_k does not depend on x_0, and we only allow $x_0 \in \mathbb{R} \setminus (a,b)$.

Example 4.2.3 *In the particular case $[\alpha,\beta] = [-1,1]$ and $g = 1$ in Theorem 4.2.1, we have explicit formulas*

$$t_j = -\cos(\pi\frac{j}{k}), \quad \xi_j = -c_k\cos(\pi\frac{2j+1}{2k}), \quad c_k = \frac{2k}{\pi}\sin(\frac{\pi}{2k}).$$

Here the midpoint approach of Totik gives the monic polynomial

$$P_k(x) = 2(\frac{c_k}{2})^k T_k(x/c_k)$$

which is not optimal for the one-sided approximation of $kU^{\omega_{[-1,1]}}(x)$ *in Theorem 4.2.1 or the sharpness of the classical Bernstein-Walsh inequality as discussed in Example 4.1.1, but good enough for concluding in Theorem 4.2.1.*

The previous example is misleading in the sense that in general there is no such sufficiently explicit formula for the t_j nor the ξ_j which will allow us to conclude in Theorem 4.2.1.

4.2.2 Structure of the proof of Theorem 4.2.1

We start by observing that, with the choices (4.8), (4.9),

$$\log|P_k(x)| + kU^{\mu}(x) = k\sum_{j=0}^{k-1}\int_{t_j}^{t_{j+1}} \log\left|\frac{x-\xi_j}{x-t}\right| \mathrm{d}\mu(t).$$

The following classical lemma shows Theorem 4.2.1(b).

Lemma 4.2.4

$$k\int_{t_j}^{t_{j+1}} \log\left|\frac{x-\xi_j}{x-t}\right| \mathrm{d}\mu(t) \begin{cases} \geq 0 & \text{for } x \in \mathbb{R}\setminus(t_j,t_{j+1}), \\ \leq \frac{1}{4}\dfrac{(t_{j+1}-t_j)^2}{(x-t_j)(x-t_{j+1})} & \text{for } x \in \mathbb{R}\setminus[t_j,t_{j+1}]. \end{cases}$$

Proof : Using the fact that $m(t) = \log|\frac{x-\xi_j}{x-t}|$ is convex on $[t_j, t_{j+1}]$ by assumption on x, we know that $m(t) \geq m(\xi_j) + m'(\xi_j)(t-\xi_j) = m'(\xi_j)(t-\xi_j)$, and thus

$$k\int_{t_j}^{t_{j+1}} \log\left|\frac{x-\xi_j}{x-t}\right| \mathrm{d}\mu(t) \geq m'(\xi_j)\int_{t_j}^{t_{j+1}} (t-\xi_j)\,\mathrm{d}\mu(t) = 0,$$

where in the last equality we have used (4.9). Also, using the convexity of m and the inequality $\log(x) \leq x-1$ we obtain

$$\begin{aligned} m(t) &\leq m(t_j)\frac{t_{j+1}-t}{t_{j+1}-t_j} + m(t_{j+1})\frac{t-t_j}{t_{j+1}-t_j} \\ &\leq \frac{t_j-\xi_j}{x-t_j}\frac{t_{j+1}-t}{t_{j+1}-t_j} + \frac{t_{j+1}-\xi_j}{x-t_{j+1}}\frac{t-t_j}{t_{j+1}-t_j}. \end{aligned}$$

Integrating and using again (4.9) we conclude that

$$\begin{aligned} k\int_{t_j}^{t_{j+1}} \log|\frac{x-\xi_j}{x-t}|\,\mathrm{d}\mu(t) &\leq \frac{t_j-\xi_j}{x-t_j}\frac{t_{j+1}-\xi_j}{t_{j+1}-t_j} + \frac{t_{j+1}-\xi_j}{x-t_{j+1}}\frac{\xi_j-t_j}{t_{j+1}-t_j} \\ &= \frac{(t_{j+1}-\xi_j)(\xi_j-t_j)}{(x-t_j)(x-t_{j+1})} \leq \frac{1}{4}\frac{(t_{j+1}-t_j)^2}{(x-t_j)(x-t_{j+1})}. \end{aligned}$$

□

Remark 4.2.5 *(a) By the same argument, the inequality of Theorem 4.2.1(b), namely* $\log|P_k(x)| + kU^\mu(x) \geq 0$, *also holds for* $x \in \{t_0, t_1, ..., t_k\}$.
(b) For $x > 1$ *(and similarly for* $x < -1$*), the right-hand side of Theorem 4.2.1(b) cannot be improved since, by Lemma 4.2.4 and Lemma 4.5.8(c),*

$$\begin{aligned}\log|P_k(x)| + kU^\mu(x) &\leq \max_{\ell=0,\dots,k-1} \frac{t_{\ell+1}-t_\ell}{4} \sum_{j=0}^{k-1} \frac{t_{j+1}-t_j}{(x-t_j)(x-t_{j+1})} \\ &= \max_{\ell=0,\dots,k-1} \frac{t_{\ell+1}-t_\ell}{2(x^2-1)} \leq \frac{1}{(x^2-1)} \left(\frac{3\pi}{2k}\right)^{1/3}.\end{aligned}$$

(c) For $x \in \mathbb{C} \setminus \mathbb{R}$, *it is not too difficult to show that* $m(t) = \log|\frac{x-\xi_j}{x-t}|$ *satisfies*

$$|m(t) - m(\xi_j) - (t-\xi_j)m'(\xi_j)| \leq \frac{(t_{j+1}-t_j)^2}{2\, dist(x,[-1,1])^2},$$

and hence by Lemma 4.5.8(c)

$$|\log|P_k(x)| + kU^\mu(x)| \leq \sum_{j=0}^{k-1} \frac{(t_{j+1}-t_j)^2}{2\, dist(x,[-1,1])^2} \leq \frac{1}{dist(x,[-1,1])^2} \left(\frac{12\pi}{k}\right)^{1/3}.$$

Thus, for sufficiently large k, *the inequality of Theorem 4.2.1(b) also holds for non-real* x *up to some arbitrarily small constant.*

Let us now turn to a proof of Theorem 4.2.1(a). We claim that it is sufficient to show Theorem 4.2.1(a) for $x \in [-1,1] = \mathrm{supp}(\mu)$, since then for μ-almost all x

$$kU^\mu(x) \leq C_{BW} - \log|P_k(z)| = C_{BW} + \sum_{j=0}^{k-1} U^{\delta_{\xi_j}}(x),$$

and thus this inequality holds for all $x \in \mathbb{C}$ by the principle of domination [103, Theorem II.3.2] and the finiteness of $I(\mu)$. Therefore, let $x \in [-1,1]$ and, more precisely,

$$j_0 \in \{0,1,...,k-1\} \quad \text{with} \quad x \in [t_{j_0}, t_{j_0+1}]. \tag{4.10}$$

According to Lemma 4.2.4, we get the following upper bound

$$\log|P_k(x)| + kU^\mu(x) \leq \Sigma_1 + \Sigma_2 + \Sigma_3 \tag{4.11}$$

with

$$\begin{aligned}\sum\nolimits_1 &= \sum_{j=0}^{j_0-2} \int_{t_j}^{t_{j+1}} \log\left|\frac{x-\xi_j}{x-t}\right| k\, \mathrm{d}\mu(t) \leq \frac{1}{4} \sum_{j=0}^{j_0-2} \frac{(t_{j+1}-t_j)^2}{(t_{j_0}-t_{j+1})^2}, \\ \sum\nolimits_2 &= \sum_{j=\max\{0,j_0-1\}}^{\min\{j_0+1,k-1\}} \int_{t_j}^{t_{j+1}} \log\left|\frac{x-\xi_j}{x-t}\right| k\, \mathrm{d}\mu(t), \\ \sum\nolimits_3 &= \sum_{j=j_0+2}^{k-1} \int_{t_j}^{t_{j+1}} \log\left|\frac{x-\xi_j}{x-t}\right| k\, \mathrm{d}\mu(t) \leq \frac{1}{4} \sum_{j=j_0+2}^{k-1} \frac{(t_{j+1}-t_j)^2}{(t_j-t_{j_0+1})^2}.\end{aligned}$$

Already in the particular Chebyshev case of Example 4.2.3 one may check that such a simple telescop sum trick as in Remark 4.2.5 does not allow to conclude, since in general $|t_j - t_\ell|$ does not behave uniformly for $j, \ell \in \{0, 1, ..., k-1\}$ like $|j-\ell|/k$, as it would be the case for equidistant points. We will discuss our upper bounds for the above three sums in the Propositions 4.2.7–4.2.9 of §4.2.3, which allows us to conclude the proof of Theorem 4.2.1, with the explicit constant

$$C_{BW} = 872 + 32 + 14479 = 15383.$$

So far we followed quite closely the reasoning in the literature, with more explicit constants. In all considerations to follow we will require precise lower and upper bounds for the ratio

$$\frac{j-\ell}{t_j - t_\ell}$$

which will follow from a new mean value property for the cumulative distribution function

$$W_g(x) = k\int_{-1}^{x} \mathrm{d}\mu(t), \quad W_g'(t) = g(t)W_1'(t) = \frac{kg(t)}{\pi\sqrt{1-t^2}}, \tag{4.12}$$

since $W_g(t_j) = j$ for $j = 0, 1, ..., k$.

Theorem 4.2.6 *Under the assumptions of Theorem 4.2.1 with $[a, b] = [-1, 1]$, we have for all distinct $x, t \in [-1, 1]$*

$$c_1 W_g'\left(\frac{t+x}{2}\right) \leq \frac{W_g(t) - W_g(x)}{t-x} \leq c_2 W_g'\left(\frac{t+x}{2}\right).$$

where $c_1 = \frac{1}{4}$ and $c_2 = \pi\sqrt{2}$.

Notice that, even for the particular case $g = 1$ and $W_1(-\cos(\alpha)) = k\alpha/\pi$, this statement is not totally obvious, but can be verified by means of elementary computations with improved constants c_1 and c_2, see Lemma 4.4.1 below. The proof for general g is strongly based on Jensen's inequality, we refer the reader to Subsection 4.4 for details.

4.2.3 Bounding three sums

For concluding our proof of Theorem 4.2.1, it remains to obtain upper bounds for the three terms on the right-hand side of (4.11), where we will proceed in order of difficulty, and apply beside Theorem 4.2.6 a certain number of technical results established in Appendix 4.5, and recalled below. In the reminder of this section we will always suppose that the assumptions of Theorem 4.2.1 hold with $[a, b] = [-1, 1]$ and j_0 is chosen as in (4.10).

We start with the sum

$$\sum\nolimits_3 \leq \frac{1}{4}\sum_{j=j_0+2}^{k-1} \frac{(t_{j+1} - t_j)^2}{(t_j - t_{j_0+1})^2},$$

where beside Theorem 4.2.6 we rely on an upper bound for the quantity

$$(1+\frac{t_j+t_{j+1}}{2})\Big/(1+t_j),$$

see Lemma 4.5.4.

Proposition 4.2.7 *There holds*

$$\sum\nolimits_3 \leq \frac{c_2^2 c_5 \pi^2}{12 c_1^2} \leq 872.$$

Proof : By Theorem 4.2.6

$$\sum\nolimits_3 \leq \frac{c_2^2}{4c_1^2} \sum_{j=j_0+2}^{k-1} \frac{1}{(j-j_0-1)^2} \frac{W_g'(\frac{t_j+t_{j_0+1}}{2})^2}{W_g'(\frac{t_j+t_{j+1}}{2})^2}.$$

As g is increasing and $j > j_0+1$, we have that $g(\frac{t_j+t_{j_0+1}}{2}) \leq g(\frac{t_j+t_{j+1}}{2})$, and thus

$$\begin{aligned}
\frac{W_g'(\frac{t_j+t_{j_0+1}}{2})^2}{W_g'(\frac{t_j+t_{j+1}}{2})^2} &= \frac{g(\frac{t_j+t_{j_0+1}}{2})^2}{g(\frac{t_j+t_{j+1}}{2})^2} \frac{W_1'(\frac{t_j+t_{j_0+1}}{2})^2}{W_1'(\frac{t_j+t_{j+1}}{2})^2} \\
&\leq \frac{W_1'(\frac{t_j+t_{j_0+1}}{2})^2}{W_1'(\frac{t_j+t_{j+1}}{2})^2} \leq \frac{1+\frac{t_j+t_{j+1}}{2}}{1+\frac{t_j+t_{j_0+1}}{2}} \\
&\leq 2\frac{1+\frac{t_j+t_{j+1}}{2}}{1+t_j} \leq 2c_5,
\end{aligned}$$

where in the last inequality we have applied Lemma 4.5.4. Combining these two results yields the claimed upper bound.

□

Let us now turn to the sum

$$\sum\nolimits_1 \leq \frac{1}{4} \sum_{j=0}^{j_0-2} \frac{(t_{j+1}-t_j)^2}{(t_{j_0}-t_{j+1})^2}.$$

Here we require beside Theorem 4.2.6 also upper bounds for the two ratios

$$\frac{1+\frac{t_j+t_{j+1}}{2}}{1+t_{j+1}}, \quad \text{and} \quad \frac{(j+1)^2}{j_0^2} \frac{1+\frac{t_{j+1}+t_{j_0}}{2}}{1+t_{j+1}}$$

for $j \leq j_0-1 \leq k-2$, see Lemma 4.5.5 and Lemma 4.5.6.

Proposition 4.2.8 *There holds*

$$\sum\nolimits_1 \leq \frac{c_2^2 \pi^2}{6c_1^2}(18+\pi^2) \leq 14479.$$

Proof : Using the fact that $j < j_0 - 1$, and that $g(t) = (1+t)h(t)$ with a decreasing function h, we find that

$$\begin{aligned}\frac{g(\frac{t_{j+1}+t_{j_0}}{2})}{g(\frac{t_j+t_{j+1}}{2})} &= \frac{1+\frac{t_{j+1}+t_{j_0}}{2}}{1+\frac{t_j+t_{j+1}}{2}} \frac{h(\frac{t_{j+1}+t_{j_0}}{2})}{h(\frac{t_j+t_{j+1}}{2})} \\ &\leq \frac{1+\frac{t_{j+1}+t_{j_0}}{2}}{1+\frac{t_j+t_{j+1}}{2}}.\end{aligned}$$

This allows us to write

$$\begin{aligned}\frac{W_g'(\frac{t_{j+1}+t_{j_0}}{2})^2}{W_g'(\frac{t_j+t_{j+1}}{2})^2} &\leq \frac{(1+\frac{t_{j+1}+t_{j_0}}{2})^2}{(1+\frac{t_j+t_{j+1}}{2})^2} \frac{W_1'(\frac{t_{j+1}+t_{j_0}}{2})^2}{W_1'(\frac{t_j+t_{j+1}}{2})^2} \\ &\leq \frac{1+\frac{t_{j+1}+t_{j_0}}{2}}{1+\frac{t_j+t_{j+1}}{2}} \frac{1-\frac{t_j+t_{j+1}}{2}}{1-\frac{t_{j+1}+t_{j_0}}{2}} \\ &\leq 2\frac{1+\frac{t_{j+1}+t_{j_0}}{2}}{1+t_{j+1}} \frac{1-\frac{t_j+t_{j+1}}{2}}{1-\frac{t_{j+1}+t_{j_0}}{2}},\end{aligned}$$

and thus, again by Theorem 4.2.6,

$$\sum\nolimits_1 \leq \frac{c_2^2}{2c_1^2} \sum_{j=0}^{j_0-2} \frac{1}{(j_0-j-1)^2} \frac{1+\frac{t_{j+1}+t_{j_0}}{2}}{1+t_{j+1}} \frac{1-\frac{t_j+t_{j+1}}{2}}{1-\frac{t_{j+1}+t_{j_0}}{2}}.$$

The following arguments depend on the sign of t_{j+1}. We therefore set $j_1 = j_0 - 1$ if $t_{j_0-1} < 0$, and else chose $j_1 \in \{0, 1, ..., j_0-2\}$ with $t_{j_1} < 0 \leq t_{j_1+1}$, and cut our sum into two parts $\Sigma_1 = \Sigma_{1,1} + \Sigma_{1,2}$, where in the first sum $j \in \{j_1, ..., j_0 - 2\}$, and in the second one $j \in \{0, ..., j_1 - 1\}$.

If $j \geq j_1$ and thus $t_{j+1} \geq 0$,

$$\frac{1+\frac{t_{j+1}+t_{j_0}}{2}}{1+t_{j+1}} \frac{1-\frac{t_j+t_{j+1}}{2}}{1-\frac{t_{j+1}+t_{j_0}}{2}} \leq 2\frac{1-\frac{t_j+t_{j+1}}{2}}{1-\frac{t_{j+1}+t_{j_0}}{2}} \leq 4\frac{1-\frac{t_j+t_{j+1}}{2}}{1-t_{j+1}} \leq 36,$$

where in the last inequality we have applied Lemma 4.5.5. Hence,

$$\sum\nolimits_{1,1} \leq \frac{18c_2^2}{c_1^2} \sum_{j=j_1}^{j_0-2} \frac{1}{(j_0-j-1)^2}. \tag{4.13}$$

If $j < j_1$ and thus $t_{j+1} < 0$,

$$\frac{1+\frac{t_{j+1}+t_{j_0}}{2}}{1+t_{j+1}} \frac{1-\frac{t_j+t_{j+1}}{2}}{1-\frac{t_{j+1}+t_{j_0}}{2}} \leq 4\frac{1+\frac{t_{j+1}+t_{j_0}}{2}}{1+t_{j+1}} \leq \frac{\pi^2}{2}\frac{j_0^2}{(j+1)^2},$$

where the last inequality follows from Lemma 4.5.6. Thus,

$$\begin{aligned}\sum\nolimits_{1,2} &\leq \frac{c_2^2\pi^2}{4c_1^2}\sum_{j=0}^{j_1-1}\frac{j_0^2}{(j_0-j-1)^2(j+1)^2}\\ &\leq \frac{c_2^2\pi^2}{4c_1^2}\left(\sum_{j=0,j+1<j_0/2}^{j_1-1}\frac{4}{(j+1)^2}+\sum_{j=0,j+1\geq j_0/2}^{j_1-1}\frac{4}{(j_0-j-1)^2}\right).\end{aligned}$$

Since $\pi^2 \leq 18$, a combination with (4.13) gives the upper bound for Σ_1 as claimed in Proposition 4.2.8.

□

We finally discuss in our third proposition the expression

$$\sum\nolimits_2 = \sum_{j=\max\{0,j_0-1\}}^{\min\{j_0+1,k-1\}}\int_{t_j}^{t_{j+1}}\log\left|\frac{x-\xi_j}{x-t}\right|W_g'(t)\,\mathrm{d}t,$$

where we integrate in a neighborhood of x and thus have to deal with the logarithmic singularity of the integrand. Here again Theorem 4.2.6 will be essential. As maybe expected from [114], our proof for $j_0 \in \{1,2,...,k-2\}$ is quite different from that for x close to the endpoints and thus $j_0 \in \{0,k-1\}$: in the first case, we require lower and upper bounds for the ratio of the lengths of two consecutive intervals $[t_j,t_{j+1}]$ established in Lemma 4.5.9, whereas in the second case we require upper bounds for

$$k\sqrt{t_1-t_0}, \quad \text{and} \quad k\sqrt{t_k-t_{k-1}},$$

see Lemma 4.5.8.

Proposition 4.2.9 *There holds*

$$\sum\nolimits_2 \leq 6c_2+\log\left(\frac{6c_2\sqrt{c_5}}{c_1}\right)\leq 32.$$

Proof : By integration by part,

$$\begin{aligned}\int_{t_j}^{t_{j+1}}\log\left|\frac{x-\xi_j}{x-t}\right|W_g'(t)\mathrm{d}t &= \left[\log\left|\frac{x-\xi_j}{x-t}\right|(W_g(t)-W_g(x))\right]_{t_j}^{t_{j+1}}\\ &\quad+\int_{t_j}^{t_{j+1}}\frac{W_g(t)-W_g(x)}{t-x}\mathrm{d}t.\end{aligned}$$

In order to make our formulas a bit easier to read, we write $j_1=\max\{0,j_0-1\}$,

$j_2 = \min\{k-1, j_0+1\}$, and get $\Sigma_2 = \Sigma_{2,1} + \Sigma_{2,2}$, with

$$\sum_{2,1} = \int_{t_{j_1}}^{t_{j_2+1}} \frac{W_g(t) - W_g(x)}{t-x} \mathrm{d}t,$$

$$\sum_{2,2} = \sum_{j=j_1}^{j_2} \left(\log \left| \frac{x-\xi_j}{x-t_{j+1}} \right| (W_g(t_{j+1}) - W_g(x)) + \log \left| \frac{x-\xi_j}{x-t_j} \right| (W_g(x) - W_g(t_j)) \right).$$

The first term is easily bounded. Indeed, using Theorem 4.2.6, we get

$$\begin{aligned} \sum_{2,1} &\leq c_2 \int_{t_{j_1}}^{t_{t_{j_2+1}}} W_g'(\frac{t+x}{2}) \mathrm{d}t \\ &\leq 2c_2 \left(W_g(\frac{x+t_{j_2+1}}{2}) - W_g(\frac{x+t_{j_1}}{2}) \right) \\ &\leq 2c_2 \left(W_g(t_{j_2+1}) - W_g(t_{j_1}) \right) = 2c_2(j_2+1-j_1), \end{aligned}$$

and thus $\Sigma_{2,1} \leq 6c_2$ for $j_0 \in \{1, ..., k-2\}$, and $\Sigma_{2,1} \leq 4c_2$ for $j_0 \in \{0, k-1\}$.

It remains to give an upper bound for $\Sigma_{2,2}$. We first study the case $j_0 \in \{1, ..., k-2\}$ and thus $j_1 = j_0 - 1$, $j_2 = j_0 + 1$. As $W_g(t_{j+1}) - W_g(t_j) = 1$ for every j, we notice that

$$W_g(t_{j_0+2}) - W_g(x) = 2(W_g(t_{j_0+1}) - W_g(x)) + (W_g(x) - W_g(t_{j_0})) \tag{4.14}$$

and

$$W_g(x) - W_g(t_{j_0-1}) = (W_g(t_{j_0+1}) - W_g(x)) + 2(W_g(x) - W_g(t_{j_0})).$$

Inserting this information into $\Sigma_{2,2}$, we obtain, after some elementary computations,

$$\begin{aligned} \sum_{2,2} =& \underbrace{[W_g(t_{j_0+1}) - W_g(x)]}_{\geq 0} \log \frac{|x-\xi_{j_0}|}{|x-t_{j_0+2}|} \underbrace{\frac{|x-\xi_{j_0-1}|}{|x-t_{j_0-1}|}}_{\leq 1} \underbrace{\frac{|x-\xi_{j_0+1}|}{|x-t_{j_0+2}|}}_{\leq 1} \\ &+ \underbrace{[W_g(x) - W_g(t_{j_0})]}_{\geq 0} \log \frac{|x-\xi_{j_0}|}{|x-t_{j_0-1}|} \underbrace{\frac{|x-\xi_{j_0-1}|}{|x-t_{j_0-1}|}}_{\leq 1} \underbrace{\frac{|x-\xi_{j_0+1}|}{|x-t_{j_0+2}|}}_{\leq 1} \\ \leq& [W_g(t_{j_0+1}) - W_g(x)] \log \left| \frac{x-\xi_{j_0}}{x-t_{j_0+2}} \right| + [W_g(x) - W_g(t_{j_0})] \log \left| \frac{x-\xi_{j_0}}{x-t_{j_0-1}} \right| \\ \leq& \log \left(\frac{6c_2\sqrt{c_5}}{c_1} \right). \end{aligned}$$

In order to justify the last inequality, we have to distinguish two cases. In case $x \in [t_{j_0}, \xi_{j_0}]$, we find that $\log |\frac{x-\xi_{j_0}}{x-t_{j_0+2}}| \leq 0$, implying that

$$\sum_{2,2} \leq [W_g(x) - W_g(t_{j_0})] \log \left| \frac{x-\xi_{j_0}}{x-t_{j_0-1}} \right| \leq \log \frac{t_{j_0+1} - t_{j_0}}{t_{j_0} - t_{j_0-1}},$$

and we conclude with help of Lemma 4.5.9. The case $x \in [\xi_{j_0}, t_{j_0+1}]$ is similar, here $\sum_{2,2} \leq \log \frac{t_{j_0+1}-t_{j_0}}{t_{j_0+2}-t_{j_0+1}}$, and we conclude again using Lemma 4.5.9.

Let us now consider the sum $\Sigma_{2,2}$ for the particular case $j_0 = 0$ and thus $j_1 = 0$, $j_2 = 1$. Using (4.14), this sum can be bounded above as before by

$$\begin{aligned}\sum_{2,2} &= \underbrace{[W_g(t_1) - W_g(x)]}_{\geq 0} \log \frac{|x-\xi_0|}{|x-t_2|} \underbrace{\frac{|x-\xi_1|}{|x-t_2|}}_{\leq 1} \\ &\quad + \underbrace{[W_g(x) - W_g(t_0)]}_{\geq 0} \log \frac{|x-\xi_0|}{|x-t_0|} \underbrace{\frac{|x-\xi_1|}{|x-t_2|}}_{\leq 1} \\ &\leq [W_g(t_1) - W_g(x)] \log \left| \frac{x-\xi_0}{x-t_2} \right| + [W_g(x) - W_g(t_0)] \log \left| \frac{x-\xi_0}{x-t_0} \right|.\end{aligned}$$

We have to consider three different cases: if $x \in [\frac{t_0+\xi_0}{2}, \frac{\xi_0+t_1}{2}]$ then $\Sigma_{2,2} \leq 0$. If $x \in [\frac{\xi_0+t_1}{2}, t_1]$, then

$$\sum_{2,2} \leq [W_g(t_1) - W_g(x)] \log \left| \frac{x-\xi_0}{x-t_2} \right| \leq \log \frac{t_1-t_0}{t_2-t_1} \leq \log \left(\frac{6c_2\sqrt{c_5}}{c_1} \right)$$

as before. Finally, in the case $x \in [t_0, \frac{t_0+\xi_0}{2}]$, we use the fact that $|x-\xi_0| \leq t_1 - t_0$, and apply Theorem 4.2.6 in order to get

$$\begin{aligned}\sum_{2,2} &\leq [W_g(x) - W_g(t_0)] \log \left| \frac{t_1-t_0}{x-t_0} \right| \\ &\leq c_2(x-t_0) g(\frac{t_0+x}{2}) W_1'(\frac{t_0+x}{2}) \log \left| \frac{t_1-t_0}{x-t_0} \right|.\end{aligned}$$

Since $\frac{t_0+x}{2} \leq 0$ and $\frac{t_0+x}{2} \leq t_1$, we have that

$$(x-t_0) g(\frac{t_0+x}{2}) W_1'(\frac{t_0+x}{2}) \leq \frac{k}{\pi} g(t_1) \frac{x-t_0}{\sqrt{1+\frac{t_0+x}{2}}} = \frac{k\sqrt{2}}{\pi} g(t_1) \sqrt{x-t_0}.$$

Using the fact that $\max_{y \geq 0} \sqrt{y} \log \frac{1}{y} = 2/e$, we conclude with help of Lemma 4.5.8(a) that

$$\sum_{2,2} \leq \frac{2c_2\sqrt{2}}{\pi e} k g(t_1) \sqrt{t_1 - t_0} \leq \frac{6}{e} c_2.$$

The reasoning for $j_0 = k-1$ is similar and allows for the same conclusion, we just have to replace Lemma 4.5.8(a) by Lemma 4.5.8(b) providing an upper bound for $k\sqrt{t_k - t_{k-1}}$. Thus

$$\sum_2 = \sum_{2,1} + \sum_{2,2} \leq \max \left\{ 6c_2 + \log \left(\frac{6c_2\sqrt{c_5}}{c_1} \right), 4c_2 + \frac{6}{e} c_2 \right\},$$

and the statement follows.

□

4.3 Proof of the main theorem

Let us first show our claim (4.6) for the support of the equilibrium measure μ_w. We observe that the external field $Q = U^{\rho/k}$ is convex on $\Sigma = [\alpha, \beta]$ and hence $\mathrm{supp}(\mu_w) = [a, b']$ for some $\alpha \le a < b' \le \beta$ by [103, Theorem IV.1.10(b)]. Since $U^{\mu_w} + Q$ is strictly decreasing on (b', ∞), the equilibrium condition (4.2) tells us that necessarily $\beta = b'$. We show below the two implications

$$\text{for some } a > \alpha: \qquad \mathrm{supp}(\mu_w) = [a, \beta] \text{ implies that } \eta(a) = k + \|\rho\|, \tag{4.15}$$
$$\mathrm{supp}(\mu_w) = [\alpha, \beta] \text{ implies that } \eta(\alpha) \le k + \|\rho\|, \tag{4.16}$$

with the strictly decreasing η as in (4.5). Since there is exactly one solution $> \alpha$ of the equation $\eta(a) = k + \|\rho\|$ iff $\eta(\alpha) > k + \|\rho\|$, our statement on $\mathrm{supp}(\mu_w)$ follows.

For a proof of (4.15), suppose that $\mathrm{supp}(\mu_w) = [a, \beta]$ for some $a > \alpha$. Then, by [103, Theorem IV.1.11(ii)], the derivative of the F-functional of Mhaskar and Saff

$$y \mapsto \log\frac{\beta - y}{4} - \int Q d\omega_{[y,\beta]} = (1 + \frac{\|\rho\|}{k}) \log\frac{\beta - y}{4} + \frac{1}{k}\int g_{[y,\beta]}(x, \infty) d\rho(x)$$

must vanish at $y = a$, and a small calculation gives the necessary condition

$$0 = \frac{1}{k(\beta - a)}\Big(k + \|\rho\| - \eta(a)\Big)$$

and thus $\eta(a) = k + \|\rho\|$, implying (4.15).

In order to show (4.16) together with the representation (4.6) of Θ, let $\mathrm{supp}(\mu_w) = [a, \beta]$ for some $a \in [\alpha, \beta)$. We denote by $Bal(\rho, [a, \beta])$ the measure obtained by balayage onto the interval $[a, \beta]$, see [103, §II.4] or Section A.4. Then, by construction,

$$k\mu_w + Bal(\rho, [a, \beta])$$

is a positive measure of mass $k + \|\rho\|$ having a constant potential on $[a, \beta]$, and thus $k\mu_w + Bal(\rho, [a, \beta]) = (k + \|\rho\|)\omega_{[a,\beta]}$. We apply the explicit formula for balayage onto an interval given in [103, Eqn. (II.4.47)] or (A.7), and get for $t \in [a, \beta]$

$$g(t) := \frac{d\mu_w}{d\omega_{[a,\beta]}}(t) = \frac{k + \|\rho\|}{k} - \frac{1}{k}\int_{-\infty}^{a} \frac{\sqrt{(\beta - y)(a - y)}}{t - y} \mathrm{d}\rho(y).$$

As a consequence

$$0 \le \lim_{t \to a+0} g(t) = \frac{k + \|\rho\| - \eta(a)}{k},$$

showing that $\eta(a) \le k + \|\rho\|$ is finite. In particular, in case $a = \alpha$ we get (4.16). Moreover, by [103, Eqn. (II.5.4)], with a suitable $F \in \mathbb{R}$,

$$\begin{aligned} k(F - U^{\mu_w}(x) - Q(x)) &= kF - U^{(k+\|\rho\|)\omega_{[a,\beta]} + \rho - Bal(\rho,[a,\beta])}(x) \\ &= (k + \|\rho\|)g(x, \infty) - \int g(x, y)\, \mathrm{d}\rho(y), \end{aligned}$$

the right-hand side vanishing on $[a, \beta]$, and thus the constant F coincides with the one in (4.2). Hence, the above expression equals $k\Theta(x)$, showing (4.6).

It remains to show that Theorem 4.2.1 implies (4.7), where we start to verify the hypotheses on

$$g(t) = \frac{d\mu_w}{d\omega_{[a,\beta]}}(t) = \frac{k + \|\rho\| - \eta(a)}{k} + \frac{t-a}{k}\int_{-\infty}^{a} \sqrt{\frac{\beta - y}{a - y}}\frac{\mathrm{d}\rho(y)}{t-y}.$$

We first observe that g is differentiable on $(a, \beta]$, with derivative

$$g'(t) = \frac{1}{k}\int_{-\infty}^{a} \sqrt{\frac{\beta - y}{a - y}}\frac{a-y}{(t-y)^2}\mathrm{d}\rho(y),$$

which is both ≥ 0 and decreasing in $t \in (a, \beta]$. Hence g is increasing and concave in (a, β), and, by a similar argument, $h(t) := g(t)/(t-a)$ is convex on (a, β). Thus the assumptions of Theorem 4.2.1 hold. With $P_k \in \Pi_k$ as in Theorem 4.2.1 we have that

$$\begin{aligned}\log \|w^k P_k\|_{[a,\beta]} &= \max_{x\in[a,\beta]} -kQ(x) + \log|P_k(x)| \\ &\leq \max_{x\in[a,\beta]} -kQ(x) - kU^{\mu_w}(x) + C_{BW} = -kF + C_{BW},\end{aligned}$$

where for obtaining the inequality we have applied Theorem 4.2.1(a), and in the last equality we have used (4.2) and in particular the fact that Θ vanishes on $[a, \beta]$. Also, for $x \in \mathbb{R} \setminus (a, \beta)$, we deduce from Theorem 4.2.1(b) and (4.2) that

$$\log w(x)^k|P_k(x)| \geq -kQ(x) - kU^{\mu_w}(x) = k\Theta(x) - kF.$$

Combining these two inequalities gives (4.7).

4.4 Proof of the mean value property of Theorem 4.2.6

As said before, a central role in our analysis is played by the mean value property of the cumulative distribution function W_g stated in Theorem 4.2.6: there exist constants $c_1 = \frac{1}{4}$ and $c_2 = \pi\sqrt{2}$ such that, for all $x, t \in [-1, 1]$,

$$c_1 W_g'(\frac{t+x}{2}) \leq \frac{W_g(t) - W_g(x)}{t-x} \leq c_2 W_g'(\frac{t+x}{2}).$$

The aim of this section is to provide a proof of this mean value property. We will first consider the two particular cases $g = 1$ in Lemma 4.4.1 and $g(t) = 1 + t$ in Lemma 4.4.2. The general case then will follow by concavity of g and by convexity of $h(t) = g(t)/(t+1)$. In what follows it will be convenient to consider the substitution $t = -\cos(\theta_t)$ and $x = -\cos(\theta_x)$, $\theta_t, \theta_x \in [0, \pi]$, where we can suppose without loss of generality that $t > x$, and thus $0 \leq \theta_x < \theta_t \leq \pi$.

Lemma 4.4.1 *For every $x, t \in [-1, 1]$, we have for $c_3 = \pi/\sqrt{2}$*

$$W_1'(\frac{t+x}{2}) \leq \frac{W_1(t) - W_1(x)}{t - x} \leq c_3 W_1'(\frac{t+x}{2}). \tag{4.17}$$

Proof : Elementary trigonometric formulas give

$$\frac{W_1(t) - W_1(x)}{t - x} = \frac{k}{\pi} \frac{\theta_t - \theta_x}{\cos(\theta_x) - \cos(\theta_t)} = \frac{k}{\pi} \frac{\frac{\theta_t - \theta_x}{2}}{\sin(\frac{\theta_t - \theta_x}{2})} \frac{1}{\sin(\frac{\theta_t + \theta_x}{2})}.$$

Observing that $\frac{\theta_t - \theta_x}{2} \in [0, \frac{\pi}{2}]$ and thus

$$\sin(\frac{\theta_t - \theta_x}{2}) \leq \frac{\theta_t - \theta_x}{2} \leq \frac{\pi}{2} \sin(\frac{\theta_t - \theta_x}{2}),$$

we deduce that

$$\frac{k}{\pi} \frac{1}{\sin(\frac{\theta_t + \theta_x}{2})} \leq \frac{W_1(t) - W_1(x)}{t - x} \leq \frac{k}{2} \frac{1}{\sin(\frac{\theta_t + \theta_x}{2})}.$$

Since

$$W_1'(\frac{t+x}{2}) = \frac{k}{\pi} \frac{1}{\sqrt{1 - \cos^2(\frac{\theta_t + \theta_x}{2}) \cos^2(\frac{\theta_t - \theta_x}{2})}},$$

the left-hand inequality in (4.17) immediately follows.

If $\frac{\theta_t + \theta_x}{2} \leq \frac{\pi}{2}$, then $0 \leq \frac{\theta_t - \theta_x}{2} \leq \frac{\theta_t + \theta_x}{2} \leq \frac{\pi}{2}$. If $\frac{\theta_t + \theta_x}{2} \geq \frac{\pi}{2}$, then $0 \leq \frac{\theta_t - \theta_x}{2} \leq \pi - \frac{\theta_t + \theta_x}{2} \leq \frac{\pi}{2}$. In both cases we find that

$$1 - \cos^2(\frac{\theta_t + \theta_x}{2}) \cos^2(\frac{\theta_t - \theta_x}{2}) \leq 1 - \cos^4(\frac{\theta_t + \theta_x}{2}) \leq 2 \left(1 - \cos^2(\frac{\theta_t + \theta_x}{2})\right),$$

which implies the right-hand side of (4.17).

□

We now turn to the special case $g(y) = 1 + y$ where we only require one inequality for $W_g = W_{1+y}$.

Lemma 4.4.2 *For every $x, t \in [-1, 1]$, we have for $c_4 = 1/2$*

$$c_4 W_{1+y}'(\frac{x+t}{2}) \leq \frac{W_{1+y}(t) - W_{1+y}(x)}{t - x}.$$

Proof : By Lemma 4.4.1,

$$\frac{W_{1+y}(t) - W_{1+y}(x)}{t - x} \geq \frac{W_{1+y}(t) - W_{1+y}(x)}{W_1(t) - W_1(x)} W_1'(\frac{x+t}{2}).$$

Thus it is sufficient to show that

$$\frac{W_{1+y}(t) - W_{1+y}(x)}{W_1(t) - W_1(x)} \geq \frac{1}{2}(1 + \frac{x+t}{2}).$$

By definition of W_{1+y},

$$\begin{aligned}\frac{W_{1+y}(t)-W_{1+y}(x)}{W_1(t)-W_1(x)} &= \frac{1}{W_1(t)-W_1(x)}\int_x^t W'_{1+y}(s)\mathrm{d}s \\ &= \frac{\theta_t-\theta_x+\sin(\theta_x)-\sin(\theta_t)}{\theta_t-\theta_x} \\ &= 1-\frac{2}{\theta_t-\theta_x}\sin(\frac{\theta_t-\theta_x}{2})\cos(\frac{\theta_t+\theta_x}{2}).\end{aligned}$$

Hence it remains to show that

$$\cos(\frac{\theta_t+\theta_x}{2})\left(2\frac{\sin(\frac{\theta_t-\theta_x}{2})}{\frac{\theta_t-\theta_x}{2}}-\cos(\frac{\theta_t-\theta_x}{2})\right)\leq 1.$$

Since $\gamma\mapsto 2\sin(\gamma)-\gamma\cos(\gamma)$ is increasing in $[0,\pi/2]$, the factor in large brackets is ≥ 0, and $\cos((\theta_t+\theta_x)/2)\leq\cos((\theta_t-\theta_x)/2$. Thus we only have to consider the worst case $\gamma=(\theta_t+\theta_x)/2=(\theta_t-\theta_x)/2\in[0,\pi/2]$, with

$$\cos(\gamma)\left(2\frac{\sin(\gamma)}{\gamma}-\cos(\gamma)\right)\leq 2\cos(\gamma)-\cos^2(\gamma)\leq 1.$$

□

We are now prepared to give a proof of Theorem 4.2.6. For the upper bound, we use Lemma 4.4.1 in order to conclude that

$$\begin{aligned}\frac{W_g(t)-W_g(x)}{t-x} &= \frac{W_g(t)-W_g(x)}{W_1(t)-W_1(x)}\frac{W_1(t)-W_1(x)}{t-x} \\ &\leq \frac{W_g(t)-W_g(x)}{W_1(t)-W_1(x)}\,c_3\,W'_1(\frac{t+x}{2}).\end{aligned}$$

Recalling that g is concave, we get from the Jensen inequality

$$\begin{aligned}\frac{W_g(t)-W_g(x)}{W_1(t)-W_1(x)} &= \int_x^t \frac{W'_g(s)}{W_1(t)-W_1(x)}\mathrm{d}s \\ &= \int_x^t g(s)\frac{W'_1(s)}{W_1(t)-W_1(x)}\mathrm{d}s \\ &\leq g\left(\int_x^t s\frac{W'_1(s)}{W_1(t)-W_1(x)}\mathrm{d}s\right) \\ &\leq 2g(\frac{t+x}{2}),\end{aligned}$$

the last inequality being established in Lemma 4.4.3 below. Thus we obtain the upper bound with $c_2=2c_3=\pi\sqrt{2}$. For the lower bound, our argument is similar, but now we use Lemma 4.4.2 in order to get

$$\begin{aligned}\frac{W_g(t)-W_g(x)}{t-x} &= \frac{W_g(t)-W_g(x)}{W_{1+y}(t)-W_{1+y}(x)}\frac{W_{1+y}(t)-W_{1+y}(x)}{t-x} \\ &\geq \frac{W_g(t)-W_g(x)}{W_{1+y}(t)-W_{1+y}(x)}c_4W'_{1+y}(\frac{x+t}{2}).\end{aligned}$$

Recalling that $h(y) = g(y)/(1+y)$ is convex, we get from the Jensen inequality

$$\begin{aligned}\frac{W_g(t) - W_g(x)}{W_{1+y}(t) - W_{1+y}(x)} &= \int_x^t h(s) \frac{W'_{1+y}(s)}{W_{1+y}(t) - W_{1+y}(x)} \mathrm{d}s \\ &\geq h\left(\int_x^t s \frac{W'_{1+y}(s)}{W_{1+y}(t) - W_{1+y}(x)} \mathrm{d}s \right) \\ &\geq \frac{1}{2} \frac{g(\frac{t+x}{2})}{1 + \frac{t+x}{2}},\end{aligned}$$

where for the last inequality we apply Lemma 4.4.4 below. This gives us the lower bound with $c_1 = c_4/2 = 1/4$.

For concluding, it remains to establish two technical results.

Lemma 4.4.3 *For $-1 \leq x < t \leq 1$ we have*

$$g\left(\int_x^t \frac{sW'_1(s)}{W_1(t) - W_1(x)} \mathrm{d}s \right) \leq 2g(\frac{x+t}{2}).$$

Proof : Elementary trigonometric computations give

$$\overline{x} := \int_x^t \frac{sW'_1(s)}{W_1(t) - W_1(x)} \mathrm{d}s = -\cos\left(\frac{\theta_t + \theta_x}{2}\right) \frac{\sin\left(\frac{\theta_t - \theta_x}{2}\right)}{\frac{\theta_t - \theta_x}{2}} \in [-1, 1],$$

and

$$\frac{x+t}{2} = -\cos\left(\frac{\theta_t + \theta_x}{2}\right) \cos\left(\frac{\theta_t - \theta_x}{2}\right).$$

We now have to distinguish two cases. If $\frac{x+t}{2} \leq 0$ or, equivalently, $\cos(\frac{\theta_t + \theta_x}{2}) \geq 0$ then, using the fact that $\sin(y) \geq y \cos(y) \geq 0$ for $y \in [0, \frac{\pi}{2}]$, we get $\overline{x} \leq \frac{x+t}{2}$, and the statement $g(\overline{x}) \leq g(\frac{x+t}{2})$ follows (without a factor 2) by monotonicity of g. If however $\frac{x+t}{2} \geq 0$, then $1 \geq \overline{x} \geq \frac{x+t}{2} \geq 0$ by the same argument as in the first case. By concavity of g and (4.20) we get

$$\begin{aligned}g(\overline{x}) &\leq g(\frac{x+t}{2}) + g'(\frac{x+t}{2} - 0)\left(\overline{x} - \frac{x+t}{2}\right) \\ &\leq g(\frac{x+t}{2}) + g'(\frac{x+t}{2} - 0) \leq 2g(\frac{x+t}{2}),\end{aligned}$$

the last inequality being shown in Lemma 4.5.1(b) below. Thus Lemma 4.4.3 holds.

□

Lemma 4.4.4 *For $-1 \leq x < t \leq 1$ we have*

$$h\left(\int_x^t s \frac{W'_{1+y}(s)}{W_{1+y}(t) - W_{1+y}(x)} \mathrm{d}s \right) \geq \frac{1}{2} h(\frac{x+t}{2}).$$

Proof : Let us first show that

$$\overline{x} := \int_x^t s \frac{W'_{1+y}(s)}{W_{1+y}(t) - W_{1+y}(x)} \mathrm{d}s \geq \frac{x+t}{2}. \tag{4.18}$$

We write shorter $w(s) = \frac{W'_{1+y}(s)}{W_{1+y}(t)-W_{1+y}(x)}$ being increasing in s. Hence

$$\begin{aligned}\overline{x} - \frac{x+t}{2} &= \int_x^t (s - \frac{x+t}{2}) w(s) \mathrm{d}s \\ &= \int_x^t (s - \frac{x+t}{2})(w(s) - w(\frac{x+t}{2})) \mathrm{d}s \geq 0,\end{aligned}$$

as claimed in (4.18). Also, by definition, $\overline{x} \leq t$, and thus

$$h(\overline{x}) = \frac{g(\overline{x})}{\overline{x}+1} \geq \frac{g(\frac{x+t}{2})}{\overline{x}+1} \geq \frac{g(\frac{x+t}{2})}{t+1} \geq \frac{1}{2} \frac{g(\frac{x+t}{2})}{1+\frac{x+t}{2}} = \frac{h(\frac{x+t}{2})}{2}.$$

□

4.5 Some further technical lemmata

After having established the mean value property of W_g in §4.4, we gather in this section all the other technical properties of the abscissa $t_j = -\cos(\theta_j)$ needed in §3.

In the sequel of this section we always suppose the conditions on k, g, h of Theorem 4.2.1 to be true, that is, k is some integer ≥ 2, g is non-negative, increasing and concave, and $h(t) = g(t)/(t+1)$ is convex.

The first result summarizes some properties of the function g.

Lemma 4.5.1 *The following properties hold:*

(a) $h(t) = \frac{g(t)}{1+t}$ *is decreasing on* $(-1, 1]$*;*

(b) $g'(t-0)(1+t) \leq g(t)$ *for* $t \in (-1, 1)$*, and* $g'(t-0) \leq g(t)$ *for* $t \in [0, 1)$*;*

(c) $g(1) \leq 2$*;*

(d) $g(1) \geq g(0) = h(0) \geq 1$.

Proof : We first recall that, by concavity of g on $[-1, 1]$, we have for all $-1 \leq x_1 < x_2 < x_3 \leq 1$ that

$$\frac{g(x_3) - g(x_2)}{x_3 - x_2} \leq \frac{g(x_3) - g(x_1)}{x_3 - x_1} \leq \frac{g(x_2) - g(x_1)}{x_2 - x_1}. \tag{4.19}$$

Since $g(-1) \geq 0$, we may therefore write

$$h(t) = \frac{g(t) - g(-1)}{t - (-1)} + \frac{g(-1)}{t+1}$$

as a sum of two decreasing functions, implying (a). Passing to the limit in (4.19), we also have that the directional derivatives $g'(x_2 - 0)$ and $g'(x_2 + 0)$ exist for all $x_2 \in (-1, 1)$, with

$$\frac{g(x_3) - g(x_2)}{x_3 - x_2} \leq g'(x_2 + 0) \leq g'(x_2 - 0) \leq \frac{g(x_2) - g(x_1)}{x_2 - x_1},$$

and in particular

$$g(x) \leq g(x_2) + g'(x_2 - 0)(x - x_2) \quad \text{for all } x \in [-1, 1]. \tag{4.20}$$

Setting $x = -1$, $x_2 = t$ in (4.20) leads to (b) since $g(-1) \geq 0$. Furthermore, using the concavity of g and setting $x_2 = 0$ in (4.20), we get for all $t \in [-1, 1]$ that

$$g(-1)\frac{1-t}{2} + g(1)\frac{1+t}{2} \leq g(t) \leq g(0) + tg'(0-).$$

Taking into account (4.12), multiplying by $W_1'(t)$ and integrating from -1 to 1 gives

$$k\frac{g(1)}{2} \leq k\frac{g(1) + g(-1)}{2} \leq k \leq kg(0),$$

implying parts (c) and (d).

□

The following elementary lemma will be helpful in what follows.

Lemma 4.5.2 *For $\gamma \geq 0$ and $0 \leq \delta \leq \theta \leq \pi/2$ there holds*

$$\frac{\sin(\gamma)}{\sin(\delta)} \leq \frac{\theta}{\sin(\theta)}\frac{\gamma}{\delta} \leq \frac{\pi}{2}\frac{\gamma}{\delta}.$$

Proof : Since $x \mapsto x/\sin(x)$ is increasing in $[0, \pi/2]$, we have that

$$\frac{\sin(\gamma)}{\sin(\delta)} \leq \frac{\gamma}{\sin(\delta)} = \frac{\delta}{\sin(\delta)}\frac{\gamma}{\delta} \leq \frac{\theta}{\sin(\theta)}\frac{\gamma}{\delta} \leq \frac{\pi}{2}\frac{\gamma}{\delta}.$$

□

The following result tells us that the angles θ_j defined by $t_j = -\cos(\theta_j)$ for $j = 0, 1, ..., k$ have a quite regular behavior.

Lemma 4.5.3 **(a)** *The sequence $(\theta_{j+1} - \theta_j)_{0 \leq j < k-1}$ is decreasing.*
(b) *For $j \in \{1, ..., k-1\}$ there holds*

$$\frac{\theta_{j+1} - \theta_j}{\theta_j} \leq \frac{1}{j},$$

(c) *For $j \in \{0, ..., k-1\}$ we have*

$$\frac{\theta_{j+1} - \theta_j}{\theta_{j+1}} \leq \frac{1}{j+1}.$$

(d) *For $j \in \{0, 1, ..., k-2\}$*

$$\frac{\theta_k - \theta_j}{\theta_k - \theta_{j+1}} \leq 4.$$

Proof : Using that g is increasing, we get for $j \in \{0, ..., k-1\}$

$$1 = \int_{t_j}^{t_{j+1}} W_g'(t)\mathrm{d}t = \frac{k}{\pi}\int_{t_j}^{t_{j+1}} \frac{g(t)}{\sqrt{1-t^2}}\mathrm{d}t$$
$$\begin{cases} \geq \frac{k}{\pi} g(t_j) \int_{t_j}^{t_{j+1}} \frac{\mathrm{d}t}{\sqrt{1-t^2}} = \frac{k}{\pi} g(t_j)(\theta_{j+1} - \theta_j), \\ \leq \frac{k}{\pi} g(t_{j+1}) \int_{t_j}^{t_{j+1}} \frac{\mathrm{d}t}{\sqrt{1-t^2}} = \frac{k}{\pi} g(t_{j+1})(\theta_{j+1} - \theta_j), \end{cases}$$

implying that

$$\frac{\pi}{kg(t_{j+1})} \leq \theta_{j+1} - \theta_j \leq \frac{\pi}{kg(t_j)}. \tag{4.21}$$

Thus (a) holds. For a proof of (b), we apply (a) to conclude that, for $j \in \{1, 2, ..., k-1\}$,

$$\frac{\theta_{j+1} - \theta_j}{\theta_j} = \frac{\theta_{j+1} - \theta_j}{\theta_j - \theta_0} = \frac{\theta_{j+1} - \theta_j}{\sum_{p=0}^{j-1} \theta_{p+1} - \theta_p}$$
$$\leq \frac{\theta_{j+1} - \theta_j}{j(\theta_j - \theta_{j-1})} \leq \frac{1}{j}.$$

A proof of part (c) follows the same lines, we omit details. Let us finally show (d). In case $k = 2$, we know from (4.21) and Lemma 4.5.1(c) that $\theta_2 - \theta_1 \geq \frac{\pi}{2g(1)} \geq \pi/4$, implying (d). In case $k \geq 3$ we can write

$$\frac{\theta_k - \theta_j}{\theta_k - \theta_{j+1}} = 1 + \frac{\theta_{j+1} - \theta_j}{\theta_k - \theta_{j+1}}$$
$$= 1 + \frac{\theta_{j+1} - \theta_j}{\sum_{\ell=j+1}^{k-1} (\theta_{\ell+1} - \theta_\ell)}$$
$$\leq 1 + \frac{1}{k-1-j} \frac{\theta_{j+1} - \theta_j}{\theta_k - \theta_{k-1}},$$

where in the last inequality we have applied (a). By part (c), $\theta_{j+1} - \theta_j \leq \theta_{j+1}/(j+1) \leq \pi/(j+1)$, and $\theta_k - \theta_{k-1} \geq \pi/(kg(t_k)) \geq \pi/(2k)$ by (4.21) and Lemma 4.5.1(c). Hence using that $k \geq 3$, we obtain

$$\frac{\theta_k - \theta_j}{\theta_k - \theta_{j+1}} \leq 1 + 2\frac{k}{(j+1)(k-1-j)} \leq 1 + 2\frac{k}{k-1} \leq 4.$$

□

The following result is used in our proof of Proposition 4.2.7.

Lemma 4.5.4 *For $j \in \{1, 2, ..., k-1\}$ there holds*

$$1 + \frac{t_j + t_{j+1}}{2} \leq c_5(1 + t_j)$$

where $c_5 = \frac{3\pi}{4} + 1$.

Proof : By Lemma 4.5.3(b), $\theta_{j+1} - \theta_j \leq \theta_j/j \leq \pi$, implying that

$$\sin(\frac{\theta_{j+1}-\theta_j}{2}) \leq \sin(\frac{\theta_j}{2j}) \leq \sin(\frac{\theta_j}{2}).$$

Moreover, since $\theta_{j+1} + \theta_j \leq (2 + 1/j)\theta_j$, we get by Lemma 4.5.2

$$\begin{aligned}\frac{t_{j+1}-t_j}{t_j - t_0} &= \frac{\sin(\frac{\theta_{j+1}-\theta_j}{2})}{\sin(\frac{\theta_j}{2})}\frac{\sin(\frac{\theta_{j+1}+\theta_j}{2})}{\sin(\frac{\theta_j}{2})} \\ &\leq \frac{\pi}{2}(2+\frac{1}{j}) \leq \frac{3\pi}{2},\end{aligned}$$

and Lemma 4.5.4 follows.

□

Let us now show the two main properties required for our proof of Proposition 4.2.8.

Lemma 4.5.5 *For $j \in \{0, 1, ..., k-2\}$ we have*

$$1 - \frac{t_j + t_{j+1}}{2} \leq 9(1 - t_{j+1}).$$

Proof : We will show the equivalent statement

$$t_{j+1} - t_j \leq 16(t_k - t_{j+1}).$$

If $t_{j+1} \leq 1/\sqrt{2}$, we obtain

$$\frac{t_{j+1}-t_j}{1-t_{j+1}} \leq \frac{t_{j+1}+1}{1-t_{j+1}} \leq (1+\sqrt{2})^2 \leq 6 \leq 16.$$

It remains to consider the case $t_{j+1} \geq 1/\sqrt{2}$, and thus $\theta_{j+1} \geq 3\pi/4$, or $(\pi - \theta_{j+1})/2 \leq \pi/8$. Using first Lemma 4.5.2 and then Lemma 4.5.3(d), we obtain

$$\begin{aligned}\frac{t_{j+1}-t_j}{1-t_{j+1}} &= -1 + \frac{1-t_j}{1-t_{j+1}} = -1 + \frac{1+\cos(\theta_j)}{1+\cos(\theta_{j+1})} \\ &= -1 + \frac{\cos^2(\frac{\theta_j}{2})}{\cos^2(\frac{\theta_{j+1}}{2})} = -1 + \frac{\sin^2(\frac{\pi}{2}-\frac{\theta_j}{2})}{\sin^2(\frac{\pi}{2}-\frac{\theta_{j+1}}{2})} \\ &\leq -1 + \left(\frac{\pi/8}{\sin(\pi/8)}\frac{\pi-\theta_j}{\pi-\theta_{j+1}}\right)^2 \leq -1 + 16\left(\frac{\pi/8}{\sin(\pi/8)}\right)^2 \leq 16.\end{aligned}$$

□

Lemma 4.5.6 *For $j \leq j_0 - 1 \leq k-2$ we have*

$$\frac{1+\frac{t_{j+1}+t_{j_0}}{2}}{1+t_{j+1}} \leq \frac{\pi^2}{8}\frac{j_0^2}{(j+1)^2}.$$

Proof : Notice that, by Lemma 4.5.2,

$$\frac{1+\frac{t_{j+1}+t_{j_0}}{2}}{1+t_{j+1}} = 1+\frac{t_{j_0}-t_{j+1}}{2(1+t_{j+1})} = 1+\frac{\sin(\frac{\theta_{j+1}+\theta_{j_0}}{2})\sin(\frac{\theta_{j_0}-\theta_{j+1}}{2})}{2\sin^2(\frac{\theta_{j+1}}{2})}$$
$$\leq 1+\frac{\pi^2}{8}\frac{(\theta_{j+1}+\theta_{j_0})(\theta_{j_0}-\theta_{j+1})}{\theta_{j+1}^2}.$$

Applying Lemma 4.5.3(a), and recalling that $j_0 \geq j+1$, we obtain

$$\frac{\theta_{j_0}-\theta_{j+1}}{\theta_{j+1}} = \frac{\sum_{\ell=j+1}^{j_0-1}(\theta_{\ell+1}-\theta_\ell)}{\sum_{\ell=0}^{j}(\theta_{\ell+1}-\theta_\ell)}$$
$$\leq \frac{(j_0-j-1)(\theta_{j+2}-\theta_{j+1})}{(j+1)(\theta_{j+1}-\theta_j)} \leq \frac{j_0-j-1}{j+1},$$

and

$$\frac{\theta_{j_0}+\theta_{j+1}}{\theta_{j+1}} = \frac{\theta_{j_0}-\theta_{j+1}}{\theta_{j+1}}+2 \leq \frac{j_0+j+1}{j+1}.$$

Combining the three inequalities, we deduce that

$$\frac{1+\frac{t_{j+1}+t_{j_0}}{2}}{1+t_{j+1}} \leq 1+\frac{\pi^2}{8}\frac{j_0^2-(j+1)^2}{(j+1)^2} \leq \frac{\pi^2}{8}\frac{j_0^2}{(j+1)^2}.$$

□

The three following results are required in our proof of Proposition 4.2.9.

Lemma 4.5.7 *For $k \geq 2$, we have*

$$t_{k-1} \geq 0.$$

Proof : Suppose that $t_{k-1} < 0$. Then using Lemma 4.5.1(d), and the fact that g is increasing allows us to find a contradiction

$$1 = \int_{t_{k-1}}^1 W_g'(t)\mathrm{d}t > \int_0^1 W_g'(t)\mathrm{d}t = \int_0^1 \frac{g(t)k}{\pi}\frac{1}{\sqrt{1-t^2}}\mathrm{d}t$$
$$\geq \frac{g(0)k}{\pi}\int_0^1 \frac{1}{\sqrt{1-t^2}}\mathrm{d}t \geq \frac{k}{2}.$$

□

Lemma 4.5.8 *There holds*

(a) $\sqrt{t_1-t_0} \leq \frac{3\pi}{\sqrt{2}g(t_1)k}$;

(b) $\sqrt{t_k-t_{k-1}} \leq \frac{\pi}{kg(1)}$;

(c) For all $j \in \{0,1,...,k-1\}$ we have $t_{j+1}-t_j \leq (\frac{12\pi}{k})^{1/3}$.

Proof : By Lemma 4.5.1(a), we find for $t \in (t_0, t_1]$ that

$$W_g'(t) = h(t)\frac{k}{\pi}\frac{\sqrt{1+t}}{\sqrt{1-t}} \geq \frac{h(t_1)k}{\pi\sqrt{2}}\sqrt{1+t}.$$

Integrating over the interval $[t_0, t_1] = [-1, t_1]$ gives

$$1 \geq \frac{h(t_1)k\sqrt{2}}{3\pi}(1+t_1)^{3/2} = \frac{g(t_1)k\sqrt{2}}{3\pi}(1+t_1)^{1/2},$$

which implies part (a). By Lemma 4.5.1(a) and Lemma 4.5.7, there holds for $t \in [t_{k-1}, t_k] \subset [0,1]$,

$$W_g'(t) \geq \frac{k}{\pi}\frac{h(1)}{\sqrt{1-t}},$$

and by integrating over the interval $[t_{k-1}, t_k] = [t_{k-1}, 1]$ we get

$$1 \geq \frac{2kh(1)}{\pi}(1-t_{k-1})^{1/2} = \frac{kg(1)}{\pi}(1-t_{k-1})^{1/2},$$

as required for part (b). For a proof of (c), we observe that, by Lemma 4.5.3(a),

$$\begin{aligned} t_{j+1} - t_j &= 2\sin(\frac{\theta_{j+1}-\theta_j}{2})\sin(\frac{\theta_{j+1}+\theta_j}{2}) \\ &\leq 2\sin(\frac{\theta_1-\theta_0}{2}) = \sqrt{2(t_1-t_0)}. \end{aligned}$$

By concavity and positivity of g and Lemma 4.5.1(d),

$$g(t_1) \geq g(t_0)\frac{1-t_1}{2} + g(1)\frac{t_1-t_0}{2} \geq \frac{t_1-t_0}{2}.$$

Multiplying with $\sqrt{t_1-t_0}$ and applying part (a) we arrive at

$$(t_1-t_0)^{3/2} \leq 2g(t_1)\sqrt{t_1-t_0} \leq \frac{3\pi\sqrt{2}}{k},$$

which yields part (c).

□

Lemma 4.5.9 *For every* $j \in \{0, \ldots, k-2\}$

$$\frac{2}{3\pi} \leq \frac{t_{j+1}-t_j}{t_{j+2}-t_{j+1}} \leq \frac{6c_2\sqrt{c_5}}{c_1}.$$

Proof : In order to show the left-hand inequality, we write

$$\frac{t_{j+2}-t_{j+1}}{t_{j+1}-t_j} = \frac{\sin(\frac{\theta_{j+1}+\theta_{j+2}}{2})}{\sin(\frac{\theta_j+\theta_{j+1}}{2})}\frac{\sin(\frac{\theta_{j+2}-\theta_{j+1}}{2})}{\sin(\frac{\theta_{j+1}-\theta_j}{2})} \leq \frac{\sin(\frac{\theta_{j+1}+\theta_{j+2}}{2})}{\sin(\frac{\theta_j+\theta_{j+1}}{2})},$$

where we have applied Lemma 4.5.3(a). We claim that the right-hand term is $\leq 3\pi/2$. Indeed, if $(\theta_j + \theta_{j+1})/2 \geq \pi/2$, then this quotient is less than one. Else, by using Lemma 4.5.2, we obtain

$$\frac{\sin(\frac{\theta_{j+1}+\theta_{j+2}}{2})}{\sin(\frac{\theta_j+\theta_{j+1}}{2})} \leq \frac{\pi}{2}\frac{\theta_{j+1}+\theta_{j+2}}{\theta_j+\theta_{j+1}},$$

and from Lemma 4.5.3(b) we get that $\theta_{j+1} + \theta_{j+2} = \theta_{j+2} - \theta_{j+1} + 2\theta_{j+1} \leq 3\theta_{j+1} \leq 3(\theta_j + \theta_{j+1})$.

To prove the right-hand inequality in Lemma 4.5.9 we use Theorem 4.2.6 and Lemma 4.5.1(a) in order to obtain

$$\begin{aligned}\frac{t_{j+1}-t_j}{t_{j+2}-t_{j+1}} &\leq \frac{c_2}{c_1}\frac{W_g'(\frac{t_{j+1}+t_{j+2}}{2})}{W_g'(\frac{t_j+t_{j+1}}{2})} \\ &\leq \frac{c_2}{c_1}\frac{h(\frac{t_{j+1}+t_{j+2}}{2})(1+\frac{t_{j+1}+t_{j+2}}{2})}{h(\frac{t_j+t_{j+1}}{2})(1+\frac{t_j+t_{j+1}}{2})}\frac{W'(\frac{t_{j+1}+t_{j+2}}{2})}{W'(\frac{t_j+t_{j+1}}{2})} \\ &\leq \frac{c_2}{c_1}\sqrt{\frac{1+\frac{t_{j+1}+t_{j+2}}{2}}{1+\frac{t_j+t_{j+1}}{2}}\frac{1-\frac{t_j+t_{j+1}}{2}}{1-\frac{t_{j+1}+t_{j+2}}{2}}}.\end{aligned}$$

With help of Lemma 4.5.4 and Lemma 4.5.5 we obtain

$$\frac{t_{j+1}-t_j}{t_{j+2}-t_{j+1}} \leq \frac{c_2}{c_1}3\sqrt{c_5}\sqrt{\underbrace{\frac{1+t_{j+1}}{1+\frac{t_j+t_{j+1}}{2}}}_{\leq 2}\underbrace{\frac{1-t_{j+1}}{1-\frac{t_{j+1}+t_{j+2}}{2}}}_{\leq 2}} \leq \frac{6c_2\sqrt{c_5}}{c_1}$$

□

4.6 Open problems

We believe that, with an optimal choice of C_{BW}, the quantity $e^{C_{BW}}$ is of modest size. This is clearly not true for our present explicit upper bound of C_{BW}, and remains a direction of future research, maybe asymptotic analysis could be helpful.

We also believe that our result on the discretization of a potential can be generalized to more general measures, for example without the assumption that $t \mapsto \frac{g(t)}{t-a}$ is convex on (a, b), which is used only once. This possibly would allow us to consider both small and large eigenvalues as outliers.

Finally, the above-mentioned conjecture on the CG convergence remains open for general sets $S(t)$.

Chapter 5

Applications of the sharpness of the Bernstein-Walsh inequality

In this chapter we apply Theorem 4.1.3 to find superlinear convergence bounds for Krylov methods. In the first section, we prove Conjecture 2.3.1 stated before, at least for a class of measures which is of particular interest for applications. In the second section we derive a superlinear convergence bound for polynomial approximations of Markov functions of Hermitian matrices.

5.1 Superlinear convergence for conjugate gradients

Our aim in this section is to explain superlinear convergence of Conjugate Gradients (CG). In terms of the chapter 3, we consider as inclusion set an interval. Up to now, our work is valid only for few outliers. As we want to explain this phenomenon for a lot of outliers, we need to consider other techniques. One of the appealing aspects of CG convergence is that there is a close link with polynomial extremal problems and extremal problems in logarithmic potential theory. We give and discuss in Section 5.1 some new upper bound for the rate of convergence of conjugate gradients, and show in our Theorem 5.1.4 the Conjecture 2.3.1 for a particular class of eigenvalue distributions, which is illustrated by some (academic) numerical examples. To our knowledge, the present work is the first which deals with Conjecture 2.3.1, at least for a suitable subclass of eigenvalue distributions. In the first subsection we explain and state our results, and give some examples. In Subsection 5.1.2 we explain how to deduce Theorem 5.1.4 from Theorem 4.1.3.

5.1.1 Superlinear convergence for conjugate gradients

Conjugate gradients is a popular method for solving large sparse linear systems $Ax = b$ with symmetric positive definite A, with spectrum $\Lambda(A) = \{\lambda_j\}$, $0 < \lambda_1 < \lambda_2 < ... \leq \beta$. It is known for a long time that there are eigenvalue distributions which lead to convergence which is faster than the one described in (2.4), namely so-called superlinear convergence, see for instance Figure 2.3.

In what follows we consider $S = [\lambda_{d+1}, \beta]$, and thus we prescribe as roots of the polynomial (of degree $n \geq d$) the smallest d eigenvalues $\lambda_1, ..., \lambda_d$. Understanding the modulus of the product of the corresponding linear factors as a weight, and setting $\rho = \delta_{\lambda_1} + ... + \delta_{\lambda_d}$, $\alpha = \lambda_{d+1}$ and $n = k + d = k + \|\rho\|$, Theorem 4.1.3 and the equality $\|w^k p\|_\Sigma = \|w^k p\|_{[a,\beta]}$ (consequence of the weighted Bernstein-Walsh inequality) give the following upper bounds in terms of Green functions. The sharpness follows from the weighted Bernstein-Walsh inequality (4.3).

Corollary 5.1.1 *For any integer $n > d + 1 \geq 1$, let $a = a_{d,n}$ be equal to λ_{d+1} if $n \geq \sum_{j=1}^{d} \sqrt{\frac{b-\lambda_j}{\lambda_{d+1}-\lambda_j}} = \eta(\lambda_{d+1})$, and else let a be the unique solution $> \lambda_{d+1}$ of the equation $n = \sum_{j=1}^{d} \sqrt{\frac{\beta-\lambda_j}{a-\lambda_j}}$. Then*

$$E_n([\lambda_{d+1}, \beta]) \leq e^{C_{BW}} \exp\Big(-n g_{[a,\beta]}(0, \infty) + \sum_{j=1}^{d} g_{[a,\beta]}(0, \lambda_j)\Big),$$

being sharp up to the factor $e^{C_{BW}}$.

Remark 5.1.2 *Using the formulas (A.3) and (A.4) for the Green function of a closed interval $S = [a, \beta]$ (with complement being open and simply connected), we note that*

$$\exp\Big(-n g_{[a,\beta]}(0, \infty) + \sum_{j=1}^{d} g_{[a,\beta]}(0, \lambda_j)\Big) = \frac{1}{|f_{n,\alpha_1,...,\alpha_d}(\alpha_0)|},$$

where $f_{n,\alpha_1,...,\alpha_d}$ is defined in Definition 3.1.2. So we have obtained the inequalities

$$\frac{1}{|f_{n,\alpha_1,...,\alpha_d}(\alpha_0)|} \leq E_n([\lambda_{d+1}, \beta]) \leq \frac{e^{C_{BW}}}{|f_{n,\alpha_1,...,\alpha_d}(\alpha_0)|},$$

with $e^{C_{BW}}$ a universal constant. Let us emphasize the fact that for small n, a can be larger than λ_{d+1}, and the function $f_{n,\alpha_1,...,\alpha_d}$ is related to $S = [a, \beta]$. This was not the case in the chapter 3 where our set S included the rest of the spectrum.

Corollary 5.1.1 gives us for each $d < n - 1$ an upper bound for the function $n \mapsto \log E_n(\Lambda(A))$, each of them having the shape of a straight line for sufficiently large n, with the slope $-g_{[\lambda_{d+1},\beta]}(0, \infty)$ of these straight lines decreasing with d, but the abscissa in general increases. We thus hope that $\log E_n(\Lambda(A))$

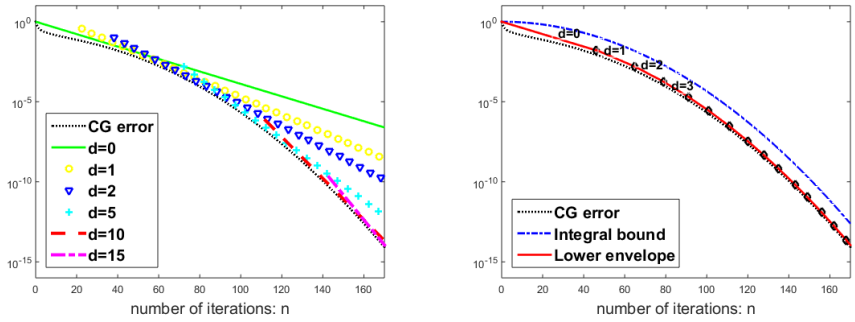

Figure 5.1: *Illustration of Corollary 5.1.1. As lower bound we draw on both plots the relative CG error in energy norm, with* $\lambda_j, N, A, b, x_0^{CG}$ *as in Figure 2.3 (black dotted line). The straight lines on the left correspond to the bounds for* $d \in \{0, 1, 2, 5, 10, 15\}$ *given in Corollary 5.1.1, each time for* n *sufficiently large such that* $a_{n,d} = \lambda_{d+1}$ *and thus* $S = S(t) = [\lambda_{d+1}, \beta]$*. Since it is difficult to see details, we have drawn on the right only the CG error and the concave lower envelope of all straight lines in Corollary 5.1.1, where we indicate in the plot the correspondence between a segment and the choice of* d*. To compare, we also have reproduced on the right from Figure 2.3 the integral bound from Conjecture 2.3.1 with* $C = 0$ *(blue dash-dotted line), verifying numerically that Corollary 5.1.1 is the right tool to prove the conjecture.*

is close to the value of the concave lower envelope of these straight lines, which is true for the particular example of Figure 5.1. In fact, finding an optimal $d = d_n < n$ with minimal $E_n([\lambda_{d+1}, \beta])$ for given n seems to be a difficult task, we will suggest an approximate solution in order to solve the above conjecture.

To prove Conjecture 2.3.1, following [12, Lemma 3.1(a)], we will impose sufficient conditions on σ such that $S(t) = [a(t), \beta]$ (defined in Eqn. (2.9)) for all t.

Lemma 5.1.3 *Suppose that* σ *is supported on the interval* $[a(0), \beta]$ *with density with respect to Lebesgue measure denoted by* σ'*, and suppose*[1] *that* $x \mapsto \sqrt{(x - a(0))(\beta - x)}\sigma'(x)$ *vanishes at* $x = a$*, and is strictly increasing in* $[a(0), \beta]$*. Then for all* $t \in (0, \|\sigma\|)$ *we have* $S(t) = [a(t), \beta]$*, with* $a(t)$ *being the unique solution of the equation*

$$t = \int_{a(0)}^{a(t)} \sqrt{\frac{\beta - x}{a(t) - x}} d\sigma(x),$$

in particular $t \mapsto a(t)$ *is strictly increasing.*

Roughly speaking, having $S(t) = [a(t), \beta]$ for sufficiently small t means that there are so few eigenvalues around 0 that they are the first eigenvalues which are well approached by Ritz values of low order. One of the reasons to consider such sets $S(t)$ is that, in any case, the superlinear convergence rate is only pronounced

[1] It follows that σ has compact support and continuous potential.

if small eigenvalues are well approached by Ritz values, and the rate depends not as much on other "converging" eigenvalues, which in first order could be neglected. Another reason is that, if the system $Ax = b$ comes from discretizing an elliptic PDE, we might have only asymptotic knowledge on small eigenvalues of A through a so-called Weyl formula. The final reason is that in the particular case $S(t) = [a(t), \beta]$ the analysis becomes simpler, and also the upper bound is more explicit, since, by (2.4),

$$\begin{aligned}\exp\Big(-N\int_0^{n/N} g_{S(\tau)}(0,\infty)\ \mathrm{d}\tau\Big) &= \exp\left(N\int_0^{n/N}\log\Big(\frac{\sqrt{\beta/a(\tau)}-1}{\sqrt{\beta/a(\tau)}+1}\Big)\ \mathrm{d}\tau\right)\\ &\le \prod_{j=0}^{n-1}\frac{\sqrt{\beta/a(j/N)}-1}{\sqrt{\beta/a(j/N)}+1}\end{aligned}$$

in terms of some "effective condition number" $\beta/a(j/N)$, compare with [14, Eqn. (2.27)].

Theorem 5.1.4 *Let σ and $S(t) = [a(t), \beta]$ for $0 < t < \|\sigma\|$ be as in Lemma 5.1.3, and A be a symmetric positive definite matrix of size N with spectrum $\lambda_1 < \lambda_2 < .... \le \beta$. If the integers $n \ge 2$ and $d = d_n \in \{0, 1, ..., n-2\}$ are such that*

$$\text{for } j = 1, 2, ..., d: \quad \sigma((-\infty, \lambda_j]) \ge j/N, \tag{5.1}$$

$$\lambda_d < a(n/N) \le \lambda_{d+1} \quad (\text{or } a(n/N) \le \lambda_1 \text{ in case } d = 0), \tag{5.2}$$

then

$$E_n(\Lambda(A)) \le E_n([\lambda_{d+1}, \beta]) \le \exp\Big(C_{BW} - N\int_0^{n/N} g_{S(\tau)}(0,\infty)\ \mathrm{d}\tau\Big),$$

and thus Conjecture 2.3.1 holds.

Note that the above choice (5.2) of d is nearly optimal in the following sense: consider diagonal A_N with eigenvalues satisfying $\sigma((-\infty, \lambda_{j,N}]) = j/N$ for $j = 1, ..., N$. Furthermore, let $d = d_{n,N}$ with $\lambda_{d,N} < a(n/N) \le \lambda_{d+1,N}$, then[2]

$$\lim_{\substack{n,N\to\infty\\ n/N\to t}} E_{n,N}([\lambda_{d_{n,N}+1,N}, \beta])^{1/n} = \liminf_{\substack{n,N\to\infty\\ n/N\to t\\ 0\le d<n-1}} E_{n,N}([\lambda_{d+1,N}, \beta])^{1/n}.$$

We first observe that our assumptions of Lemma 5.1.3 on σ and the choice of the eigenvalues $\lambda_{1,N} < \lambda_{2,N} < ...$ of A_N, allows to show that not only (2.7) but also the quite technical four conditions given in Subsection 2.3.3 or in [12, Conditions (i)–(iv)] hold [19, Theorem 1.7(b)]. As a consequence of [12, Theorem 2.2],

$$\lim_{\substack{n,N\to\infty\\ n/N\to t}} E_n(\Lambda(A_N))^{1/n} = \exp\Big(-\frac{1}{t}\int_0^t g_{S(\tau)}(0,\infty)\ \mathrm{d}\tau\Big),$$

[2] We write $E_{n,N}$ instead of E_n in order to indicate that here we consider the spectrum of A_N depending on N.

that is, we have equality in (2.8). Then, using Theorem 5.1.4 and the simple inequality $E_n(\Lambda(A_N)) \leq E_n([\lambda_{d+1}, \beta])$,

$$\liminf_{\substack{n,N\to\infty \\ n/N\to t}} E_{n,N}([\lambda_{d_{n,N}+1,N}, \beta])^{1/n} \leq \limsup_{\substack{n,N\to\infty \\ n/N\to t}} E_{n,N}([\lambda_{d_{n,N}+1,N}, \beta])^{1/n}$$

$$\leq \exp\Big(-\frac{1}{t}\int_0^t g_{S(\tau)}(0,\infty)\ \mathrm{d}\tau\Big) = \lim_{\substack{n,N\to\infty \\ n/N\to t}} E_n(\Lambda(A_N))^{1/n}$$

$$\leq \liminf_{\substack{n,N\to\infty \\ n/N\to t \\ 0\leq d<n-1}} E_{n,N}([\lambda_{d+1,N}, \beta])^{1/n}. \leq \liminf_{\substack{n,N\to\infty \\ n/N\to t}} E_{n,N}([\lambda_{d_{n,N}+1,N}, \beta])^{1/n},$$

and the statement follows.

It is also interesting to compare Theorem 5.1.4 with [13, Theorem 3.1] which showed under the sole assumption (2.7) (and for quite general measures σ) that, for any fixed compact set S, the quantity $E_{n,N}(S)^{1/n}$ is asymptotically greater than or equal to the right-hand side of (2.8). One of the consequences of our Theorem 5.1.4 is that, roughly, we can achieve equality for the interval $S = S(n/N)$.

Our proof of Theorem 5.1.4 will be presented in §5.1.2, let study here some examples.

Example 5.1.5 *Consider the probability density*

$$\frac{d\sigma}{dx}(x) = \frac{1}{2\sqrt{1-x}} \quad \text{on} \quad [a(0), \beta] = [0, 1].$$

For this measure we may apply Lemma 5.1.3, and a simple computation shows for $0 < t < \|\sigma\| = 1$ that $a(t) = t^2$. We may also compute eigenvalues λ_j satisfying equality in (5.1):

$$\sigma([0, \lambda_j]) = \frac{j}{N} \quad \textit{iff} \quad \lambda_j = \frac{j}{N}\Big(2 - \frac{j}{N}\Big),$$

which behave like equidistant points for $j \ll N$. These are the eigenvalues used in Figure 2.3 and Figure 5.1. In this special example we even have an explicit formula for the quantity $d = d_n$ of Theorem 5.1.4, namely

$$d_n + 1 = \lceil N(1 - \sqrt{1 - (n/N)^2}) \rceil,$$

in particular $d_n = 0$ for $n \leq 45$, $d_n = 1$ for $46 \leq n \leq 64$, and $d_n = 2$ for $65 \leq n \leq 78$, in accordance with the right-hand plot of Figure 5.1. Note that for small n we have $d_n + 1 \approx \lceil \frac{n^2}{2N} \rceil$.

In the previous example the small eigenvalues were approximately equidistant, with stepsize $2/N$, and the convex hull of the spectrum given approximately by $[2/N, 1]$. Up to correct scaling, a similar behavior is true for the eigenvalues of the finite difference discretization of the 2D Laplacian on the unit square with Dirichlet boundary conditions, and thus the convergence curves should be similar. However, this is no longer true for higher dimensions $D \geq 3$, where we expect that $\sigma'(x)$ grows like a constant times $x^{(D-2)/2}$ for small x, which motivates the following example.

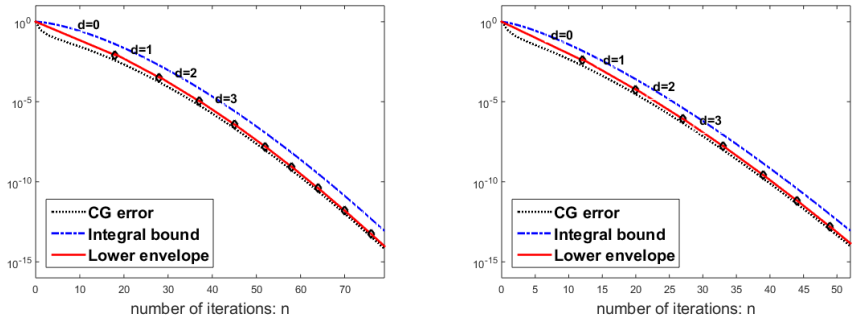

Figure 5.2: *Illustration of Theorem 5.1.4, with λ_j for $j = 1, ..., N = 1000$ as in Example 5.1.6, where on the left $s = 0.5$ and on the right $s = 1$. As lower bound we draw on both plots the relative CG error in energy norm, with A, b, x_0^{CG} as in Figure 2.3 (blue dotted line). The concave lower envelope is obtained from Theorem 5.1.4 (red solid line), where we indicate in the plot the correspondence between a segment and the choice of d. We also draw the integral bound (blue dash-dotted line), showing that Conjecture 2.3.1 holds with $C = 0$.*

Example 5.1.6 *For a parameter $\gamma > 0$, consider the density*

$$\frac{d\sigma}{dx}(x) = \frac{\gamma x^s}{\sqrt{\beta - x}} \quad on \quad [0, \beta].$$

In this example we only consider probability measures σ and thus $\gamma\beta^{s+1/2}B(s+1, 1/2) = 1$, with $B(\cdot, \cdot)$ the beta function. Notice that, for $s = 0$, we recover Example 5.1.5. A small computation using Lemma 5.1.3 gives

$$a(t)/\beta = t^{\frac{1}{s+1/2}}.$$

We again choose λ_j for $j = 1, 2, ..., N = 1000$ attaining equality in (5.1), however, there are no longer explicit formulas, and thus the λ_j have to be computed numerically. In Figure 5.2 we have plotted two examples for $\beta = 1$, on the left for $s = 0.5$ and on the right for $s = 1$, where in both cases we have chosen the approximately optimal $d = d_n$ of Theorem 5.1.4. Notice also the well-known phenomena that the convergence of CG improves with s getting larger.

5.1.2 Proof of Theorem 5.1.4

For our proof of Theorem 5.1.4, we choose $n > d + 1$ as in the statement such that (5.1) and (5.2) hold. By our assumption (5.1) on λ_j, we may choose $\widetilde{\lambda}_j \leq \lambda_j$ such that

$$\sigma((-\infty, \widetilde{\lambda}_j]) = \frac{j}{N} \quad \text{for } j = 1, 2, ..., d + 1.$$

Consider $k = n - d \geq 2$, and the two measures of mass d

$$\rho = \delta_{\lambda_1} + ... + \delta_{\lambda_d}, \quad \widetilde{\rho} = N\sigma|_{(-\infty, \widetilde{\lambda}_d]}.$$

We also consider an artificial point $\lambda_0 = \widetilde{\lambda}_0$ such that $\lambda_0 < \widetilde{\lambda}_1$ and $\sigma((-\infty, \lambda_0]) = 0$. Before proving Theorem 5.1.4, let us begin by giving two lemmas.

Lemma 5.1.7 *Under the preceding notations we have for $x \in \mathbb{C}$*

$$U^{\widetilde{\rho}}(x) - U^{\rho}(x) \leq 0, \quad \text{if } Re(x) \geq \lambda_{d+1},$$

and

$$U^{\widetilde{\rho}}(x) - U^{\rho}(x) \geq 0, \quad \text{if } Re(x) \leq \lambda_0.$$

Proof : Let $Re(x) \geq \lambda_{d+1}$. For $t \in [\widetilde{\lambda}_j, \widetilde{\lambda}_{j+1}]$, $j = 0 : d-1$, we have $|x-t| \geq |x - \lambda_{j+1}|$, and thus

$$N \int_{\widetilde{\lambda}_j}^{\widetilde{\lambda}_{j+1}} \log \frac{1}{|x-t|} \mathrm{d}\sigma(t) \leq \log \frac{1}{|x - \lambda_{j+1}|}.$$

Taking the sum for $j = 0 : d-1$, we obtain $U^{\widetilde{\rho}}(x) \leq U^{\rho}(x)$.

Let $Re(x) \leq \lambda_0$. For $t \in [\widetilde{\lambda}_j, \widetilde{\lambda}_{j+1}]$, $j = 0 : d-1$, we have $|x-t| \leq |x - \lambda_{j+1}|$, and thus with the same reasoning as before, we obtain $U^{\widetilde{\rho}}(x) \geq U^{\rho}(x)$.

□

Lemma 5.1.8 *Consider the external field $Q(x) = U^{\widetilde{\rho}/k}(x)$, the extremal measure μ and the constant F are as in (4.1), (4.2). Let $t = n/N$, $\nu_{t,\sigma}$ be the solution of the constrained equilibrium problem mentioned in the paragraph after (2.8), and $C_{t,\sigma}$ the corresponding constant. Then for all $x \in \mathbb{C}$ we have*

$$kU^{\mu}(x) + U^{\widetilde{\rho}}(x) - kF \leq n(U^{\nu_{t,\sigma}}(x) - C_{t,\sigma}). \tag{5.3}$$

Proof : Let us first show that (5.3) holds for $x \in \mathrm{supp}(\mu)$. Indeed, since $\mathrm{supp}(\mu) \subset [\lambda_{d+1}, b] \subset [a(t), \beta] = \mathrm{supp}(\sigma/t - \nu_{t,\sigma})$ by assumption (5.2), we find from the respective equilibrium conditions for both extremal problems that both expressions on the left-hand side and on the right-hand side of (5.3) vanish for $x \in \mathrm{supp}(\mu)$. We also know that all measures involved in (5.3) have finite energy, with masses $\|k\mu + \widetilde{\rho}\| = k + d = n = n\|\nu_{t,\sigma}\|$. Let us show that $\widetilde{\rho} \leq n\nu_{t,\sigma}$. Indeed, $\widetilde{\lambda}_d \leq \lambda_d < a(t)$ by construction and (5.2), and thus, by definition of $S(t) = [a(t), \beta] = \mathrm{supp}(\sigma/t - \nu_{t,\sigma})$,

$$n\nu_{t,\sigma}|_{(-\infty, \widetilde{\lambda}_d]} = \frac{n}{t}\sigma|_{(-\infty, \widetilde{\lambda}_d]} = N\sigma|_{(-\infty, \widetilde{\lambda}_d]} = \widetilde{\rho}.$$

Hence, by subtracting $U^{\widetilde{\rho}}(x)$ from both sides of (5.3), we get from the principle of domination for logarithmic potentials [103, Theorem II.3.2] that (5.3) holds for all $x \in \mathbb{C}$.

□

Now we are able to prove Theorem 5.1.4.

Proof of Theorem 5.1.4: Using the definition of $E_n([\lambda_{d+1}, \beta])$ and Lemma 5.1.7 (with $\lambda_0 = 0$), we get the chain of inequalities

$$\begin{aligned} E_n([\lambda_{d+1}, \beta]) &= \min_{p \in \Pi_k} \frac{\|e^{-U^\rho} p\|_{[\lambda_{d+1}, b]}}{e^{-U^\rho(0)}|p(0)|} \leq \min_{p \in \Pi_k} \frac{\|e^{-U^{\widetilde{\rho}}} p\|_{[\lambda_{d+1}, b]}}{e^{-U^{\widetilde{\rho}}(0)}|p(0)|} \\ &\leq \exp\Big(C_{BW} + kU^\mu(0) + U^{\widetilde{\rho}}(0) - kF\Big), \end{aligned}$$

where in the last inequality we have applied Theorem 4.1.3 with $\alpha = \lambda_{d+1}$, the external field $Q(x) = U^{\widetilde{\rho}/k}(x)$, and where the extremal measure μ and the constant F are as in (4.1), (4.2). On the other hand, with $t = n/N$, we know from [12, Theorem 2.1] (see (A.6)) that

$$\exp\Big(C_{BW} - N \int_0^{n/N} g_{S(\tau)}(0, \infty)\, \mathrm{d}\tau\Big) = \exp\Big(C_{BW} + n(U^{\nu_{t,\sigma}}(0) - C_{t,\sigma})\Big), \quad (5.4)$$

with $\nu_{t,\sigma}$ the solution of the constrained equilibrium problem, and $C_{t,\sigma}$ the corresponding constant. For establishing Theorem 5.1.4, it only remains to use the inequality given in Lemma 5.1.8 for $x = 0$.

□

5.2 Superlinear convergence for the approximation of matrix functions

In this section a superlinear convergence bound for polynomial Arnoldi approximations to functions of matrices is derived. This bound generalizes the preceding superlinear convergence bound for the CG method to more general functions with finite singularities. We consider the quantity

$$\|f(A)b - f_m\|,$$

where f_m is a Rayleigh approximation for $f(A)b$ given in Definition 1.3.1. This quantity was analyzed by several authors [20], [36], [63], [69], [100]. In the first part we recall some upper bounds which can be found in the literature. In Subsection 5.2.2 we explain and state our Theorem 5.2.8, and give some numerical examples. In Subsection 5.2.3 we show how to deduce Theorem 5.2.8 from Theorem 4.1.3.

5.2.1 Approximation of matrix functions

The accuracy of an approximation obtained by some rational Krylov method is determined by the "quality" of the rational Krylov space $\mathcal{Q}_m$ and the extraction. Of course, an approximation f_m can only be as good as the search space it is extracted from, i.e.,

$$\|f(A)b - f_m\| \geq \min_{u \in \mathcal{Q}_m} \|f(A)b - u\| = \min_{r \in \Pi_{m-1}/q_{m-1}} \|f(A)b - r(A)b\|.$$

The minimum is achieved by the orthogonal projection of $f(A)b$ which is given by a rational function evaluated at a matrix. In a rational Krylov method it is therefore necessary to make the right-hand side of this inequality as small as possible by choosing the poles of q_{m-1} suitably. If the extraction is the Rayleigh-Ritz approximation, with some conditions on f, we can obtain a near-best approximation $f_m \in \mathcal{Q}_m$ [20, Proposition 3.1].

Proposition 5.2.1 *Let V_m be a basis of $\mathcal{Q}_m$ and $A_m = V_m^\dagger A V_m$. Let f be analytic in a neighborhood of convex compact set $\Sigma \supseteq \mathbb{W}(A)$ and consider $f_m = V_m f(A_m) V_m^\dagger b$. There holds*

$$\|f(A)b - f_m\| \leq 2\|b\|(1+\sqrt{2}) \min_{r \in \Pi_{m-1}/q_{m-1}} \|f - r\|_\Sigma.$$

If A is Hermitian, the result holds with 1 instead of $1+\sqrt{2}$.

Proof : Using the exactness property and the fact that $\mathbb{W}(A_m) \subseteq \mathbb{W}(A) \subseteq \Sigma$, we obtain for every $r \in \Pi_{m-1}/q_{m-1}$

$$\begin{aligned}\|f(A)b - f_m\| &= \|(f-r)(A)b - V_m(f-r)(A_m)V_m^\dagger b\| \\ &\leq (\|(f-r)(A)\| + \|(f-r)(A_m)\|)\,\|b\| \\ &\leq 2\|b\|(1+\sqrt{2})\|f-r\|_\Sigma.\end{aligned}$$

The last inequality is a consequence of Crouzeix's theorem.

□

This proposition suggests that we choose the poles q_{m-1} such that

$$\min_{r \in \Pi_{m-1}/q_{m-1}} \|f - r\|_\Sigma$$

becomes as small as possible, which is a rational best uniform approximation problem.

Remark 5.2.2 *We cannot expect the bound in the preceding proposition to be sharp if A is highly nonnormal since it is based on the numerical range. Unfortunately, this bound can be crude even for a self-adjoint operator. Indeed, for Hermitian operators the inequality in the proof can be improved to*

$$\|f(A)b - f_m\| \leq 2 \min_{r \in \Pi_{m-1}/q_{m-1}} \|f - r\|_{\Lambda(A)\cup\Lambda(A_m)}\|b\|$$

which is a min-max problem on a discrete set. So if the spectrum of A does not "fill" the numerical range sufficiently well, we cannot expect the bound to be sharp.

Now let χ_m denote the characteristic polynomial of A_m and Γ be an integration contour (finite union of nonintersecting regular Jordan curves) such that $\Lambda(A_m) \subset int(\Gamma)$. If f is analytic in $int(\Gamma)$ and extends continuously to Γ, then

so does $\tilde{f} = fq_{m-1}$. Owing to Hermite's formula [119, page 50], the polynomial $p_{m-1} \in \Pi_{m-1}$ interpolating $\tilde{f}$ at $\Lambda(A_m)$ can be expressed as

$$p_{m-1}(z) = \frac{1}{2i\pi} \int_\Gamma \frac{\chi_m(\xi) - \chi_m(z)}{\xi - z} \frac{\tilde{f}(\xi)}{\chi_m(\xi)} \mathrm{d}\xi,$$

where χ_m is the characteristic polynomial of A_m. For the interpolation error we have by Cauchy's formula

$$\tilde{f}(z) - p_{m-1}(z) = \frac{1}{2i\pi} \int_\Gamma \frac{\chi_m(z)}{\xi - z} \frac{\tilde{f}(\xi)}{\chi_m(\xi)} \mathrm{d}\xi.$$

Dividing this equation by q_{m-1} and setting $s_m = \chi_m / q_{m-1}$ we obtain

$$f(z) - r_m(z) = \frac{1}{2i\pi} \int_\Gamma \frac{s_m(z)}{s_m(\xi)} \frac{f(\xi)}{\xi - z} \mathrm{d}\xi, \tag{5.5}$$

where $r_m = p_{m-1}/q_{m-1}$. Thus, by choosing r_m such that it interpolates f at $\Lambda(A_m)$, the error of the Rayleigh-Ritz approximation can be expressed by (recall that $f_m = r_m(A)b$)

$$f(A)b - f_m = \frac{1}{2i\pi} \int_\Gamma \frac{s_m(A)}{s_m(\xi)} f(\xi)(\xi - A)^{-1} b \mathrm{d}\xi.$$

This way of writing the error can be very useful to obtain upper error bounds. For instance, we can obtain

$$\|f(A)b - f_m\| \leq C \frac{\|s_m(A)b\|}{\min_{\xi \in \Gamma} |s_m(\xi)|},$$

where $C = \frac{L(\Gamma)}{2\pi} \max_{\xi \in \Gamma} \left\| f(\xi)(\xi - A)^{-1} \right\|$ is a constant, with $L(\Gamma)$ the length of Γ. It is remarkable that s_m is the minimizer of the factor $\|s_m(A)b\|$ among all $s_m \in \Pi_m^\infty / q_{m-1}$.

If f is analytic in a neighborhood of $\mathbb{W}(A)$ containing an integration contour Γ such that $\mathbb{W}(A) \subseteq \Sigma \subset int(\Gamma)$, then we can write (Crouzeix's theorem)

$$\|f(A)b - f_m\| \leq C(1 + \sqrt{2}) \|b\| \times \|s_m\|_\Sigma \times \|s_m^{-1}\|_\Gamma,$$

noting that the zeros of s_m are rational Ritz values, hence contained in $\mathbb{W}(A)$, and therefore $\|s_m^{-1}\|_\Sigma < \infty$. This bound suggests to consider rational functions s_m that are "smallest possible" on the set Σ and "largest possible" on the integration contour Γ winding around Σ. The zeros of the denominator q_{m-1} of such an optimal rational function should constitute "good" poles for the rational Krylov space $\mathcal{Q}_m$. This problem is known as a Zolotarev problem [56], [113], [122].

5.2.2 Superlinear convergence for the approximation of matrix functions

In [17] the authors derived a superlinear convergence bound for rational Arnoldi approximations to Markov functions of Hermitian matrices. This bound generalizes the superlinear convergence bound for the CG method [12] to more general functions and to rational Krylov spaces.

Consider a Hermitian matrix A with spectrum $\Lambda(A) = \{\lambda_j\}$, $\lambda_1 < \lambda_2 < ... < \lambda_N \leq b$, and a Markov function

$$f(z) = \int_\Gamma \frac{\mathrm{d}\gamma(x)}{x - z}, \tag{5.6}$$

where γ is a complex measure supported on a closed set $\Gamma \subset \overline{\mathbb{C}} \setminus [\lambda_1, \lambda_N]$. The associated Newton potential given by

$$\hat{f}(z) = \int_\Gamma \frac{\mathrm{d}|\gamma|(x)}{|x - z|}, \tag{5.7}$$

is finite on $[\lambda_1, \lambda_N]$. To see the result given in this section as a generalization of the result concerning CG, we can consider z^{-1} as a Markov function with $\gamma = \delta_0$ the Dirac measure supported at 0 and thus $\Gamma = \{0\}$. Other examples of Markov functions [37, Example 1 and 2] are

$$\frac{\log(1+z)}{z} = \int_{-\infty}^{-1} \frac{1/x}{x - z}\mathrm{d}x,$$

$$\frac{\exp(\theta\sqrt{z}) - 1}{z} = \int_{-\infty}^{0} \frac{\sin(\theta\sqrt{-x}}{\pi x}\frac{\mathrm{d}x}{x - z},$$

or [48]

$$z^{-\alpha} = \frac{\sin(\alpha\pi)}{\pi}\int_{-\infty}^{0} \frac{x^{-\alpha}}{z - x}\mathrm{d}x \quad , \alpha \in (0, 1).$$

In [17] the authors considered a sequence of matrices A_N, where N indicates the size of the matrix, whose eigenvalues $\{\lambda_{j,N}\}$ have an asymptotic distribution given by a probability measure σ having compact support. They associated to those matrices a sequence of vectors b_N of unit length. Accordingly, they considered the sequence of rational Krylov spaces $\mathcal{Q}_{n,N}(A_N, b_N)$ of order $n = n(N)$ such that $n/N \to t \in (0, 1)$ as $N \to \infty$, and such that the poles are asymptotically distributed according to a measure ν_t. The family of measures $t \mapsto \nu_t$ is supposed increasing and differentiable with derivative $\tilde{\nu}_t$. They denoted by $f_{n,N}$ the n-th rational Arnoldi approximation. Under some technical assumptions on the eigenvalues, the poles, and the vectors b_N, and provided that the intersection of Γ with all the eigenvalues is empty, the authors established in [17, Theorem 3.1] the asymptotic upper bound

$$\limsup_{\substack{n,N\to\infty \\ n/N\to t}} \left(\|f(A_N)b_N - f_{n,N}\|\right)^{1/N} \leq \max_{x\in\Gamma} \exp\left(-\int_0^t \int g_{S(\tau)}(x, y)\mathrm{d}\tilde{\nu}_\tau(y)\mathrm{d}\tau\right).$$

In terms of polynomial approximations, as the poles are ∞, the preceding bound can be written

$$\limsup_{\substack{n,N\to\infty \\ n/N\to t}} \left(\|f(A_N)b_N - f_{n,N}\|\right)^{1/N} \leq \max_{x\in\Gamma} \exp\left(-\int_0^t g_{S(\tau)}(x,\infty)\mathrm{d}\tau\right).$$

This formula is a generalization of the formula 2.8 for CG ($f(z) = z^{-1}$ and $\Gamma = \{0\}$) to Markov functions. As in the case of CG, although those results are of an asymptotic nature, there is numerical evidence that the superlinear convergence phenomenon also occur for finite N, without limits and without taking the n-th root, see for instance [17, Figure 1.1]. In the case of a single matrix, we have the same problem as before since we cannot define σ as an asymptotic distribution of the eigenvalues.

In order to prove a similar result as Theorem 5.1.4 for Markov functions, we first extend our definition of $E_n(S)$. For a fixed matrix A, a compact set S, sufficiently large n, $\Gamma \subset \overline{\mathbb{C}} \setminus [\lambda_1, \ldots, \lambda_N]$ and $\Gamma \cap S = \emptyset$, let

$$E_n(S,\Gamma) = \min_{p\in\Pi_n(\Lambda_0)} \left\{ \frac{\|p\|_S}{\min_{x\in\Gamma}|p(x)|} \right\},$$

where $\Pi_n(\Lambda_0)$ is defined in (3.1). Clearly we have $E_n(S,\{0\}) = E_n(S)$.

The first step is to make a link between the quantities $\|f(A)b - f_n\|$ and $E_n(\Lambda(A),\Gamma)$ which can be done by using Hermite's formula given in (5.5), and we obtain the following theorem [17, Theorem 2.2].

Theorem 5.2.3 *Consider a Markov function f given by (5.6) analytic in $\overline{\mathbb{C}}\setminus\Gamma$ containing the interval $[\lambda_1,\lambda_N]$, and let $\hat{f}$ be its associated Newton potential (5.7). For any $\tilde{s}_n \in \Pi_n/q_{n-1}$ we have*

$$\|f(A)b - f_n\| \leq 2 \max_{z\in[\lambda_1,\lambda_N]} \hat{f}(z) \frac{\|\tilde{s}_n(A)b\|}{\min_{x\in\Gamma}|\tilde{s}_n(x)|}. \tag{5.8}$$

Proof : Let $\{\tilde{\theta}_j\}$ be the zeros of $\tilde{s}_n$. If for some j, $\tilde{\theta}_j \in \Gamma$, then the theorem is clear because the right-hand side equals $+\infty$. So it is sufficient to consider $\tilde{\theta}_j \notin \Gamma$.

The Rayleigh-Ritz approximation is independent of the particular basis, hence let Vn be orthonormal.

Denote by $\tilde{r}_n$ the rational interpolant of f with (fixed) denominator q_{n-1} and numerator in Π_{n-1} interpolating f at each $\tilde{\theta}_j$. Then the exactness property (Proposition 1.3.2) gives $\tilde{r}_n(A)b = V_n\tilde{r}_n(A_n)V_n^*b$. As the interpolation error can be represented as (see (5.5))

$$f(z) - \tilde{r}_n(z) = \tilde{s}_n(z)\int_\Gamma \frac{\mathrm{d}\gamma(x)}{\tilde{s}_n(x)(x-z)},$$

we obtain

$$\begin{aligned}\|f(A)b - f_n\| &\leq \|f(A)b - \tilde{r}_n(A)b\| + \|\tilde{r}_n(A)b - r_n(A)b\| \\ &= \|(f-\tilde{r}_n)(A)b\| + \|V_n(\tilde{r}_n - f)(A_n)V_n^*b\| \\ &\leq \|\hat{f}(A)\|\frac{\|\tilde{s}_n(A)b\|}{\min_{x\in\Gamma}|\tilde{s}_n(x)|} + \|\hat{f}(A_n)\|\frac{\|V_n\tilde{s}_n(A_n)V_n^*b\|}{\min_{x\in\Gamma}|\tilde{s}_n(x)|}.\end{aligned}$$

By exactness property (Proposition 1.3.2) we have the equality $V_n\tilde{s}_n(A_n)V_n^*b = V_nV_n^*\tilde{s}_n(A)b$, and as A is supposed Hermitian, it follows that

$$\|f(A)b - f_n\| \leq \max_{z\in\Lambda(A)}|\hat{f}(z)|\frac{\|\tilde{s}_n(A)b\|}{\min_{x\in\Gamma}|\tilde{s}_n(x)|} + \max_{z\in\Lambda(A_n)}|\hat{f}(z)|\frac{\|\tilde{s}_n(A)b\|}{\min_{x\in\Gamma}|\tilde{s}_n(x)|}.$$

The observation $\Lambda(A)\cup\Lambda(A_n)\subset[\lambda_1,\lambda_N]$ allows to conclude.

□

Remark 5.2.4 *For the special case $\Gamma\subset[-\infty,\lambda_1)$ (or $\Gamma\subset(\lambda_N,+\infty]$) and γ positive (or negative), we have $\hat{f}(z) = |f(z)|$ for $z\in[\lambda_1,\lambda_N]$. Hence we can replace $\hat{f}$ by f in the upper bound of the theorem.*

In the following of this section, we restrict our study to polynomial approximations (no poles) and to $\Gamma\subset[-\infty,\lambda_1)$. We can reformulate (5.8) ($q_{n-1}=1$)

$$\|f(A)b - f_n\| \leq 2\|b\|\max_{z\in[\lambda_1,\lambda_N]}\hat{f}(z)\frac{\|\tilde{\chi}_n(A)\|}{\min_{x\in\Gamma}|\tilde{\chi}_n(x)|},$$

for every $\tilde{\chi}_n\in\Pi_n$. As A is supposed Hermitian, using Definition 5.2.2 we obtain

$$\|f(A)b - f_n\| \leq 2\|b\|\max_{z\in[\lambda_1,\lambda_N]}\hat{f}(z)E_n(\Lambda(A),\Gamma).$$

Now, like in Corollary 5.1.1, we consider $S=[\lambda_{d+1},\lambda_N]$, $\rho=\delta_{\lambda_1}+\ldots+\delta_{\lambda_d}$, $\alpha=\lambda_{d+1}$ and $n=k+d=k+\|\rho\|$.

Lemma 5.2.5 *Suppose we are under the assumptions of Theorem 4.1.3. If $\Gamma\subset[-\infty,\lambda_1)$ and x_0 designates the right extremity of Γ, then*

$$\begin{aligned}\max_{x\in\Gamma}\exp\Big(-k\Theta(x)\Big) &= \max_{x\in\Gamma}\exp\Big(-ng_{[a,\beta]}(x,\infty) + \sum_{j=1}^{d} g_{[a,\beta]}(x,\lambda_j)\Big) \\ &= \exp\Big(-ng_{[a,\beta]}(x_0,\infty) + \sum_{j=1}^{d} g_{[a,\beta]}(x_0,\lambda_j)\Big)\end{aligned}$$

Proof : Under our hypotheses, we have $\Theta(x) = F - U^{\mu+\rho/k}(x)$. For $x\leq\lambda_1$ it is not hard to see that $U^{\mu+\rho/k}$ is increasing, and thus Θ is decreasing. This implies that

$$\max_{x\in\Gamma} -k\Theta(x) = -k\Theta(x_0),$$

and we conclude with the relation (4.6).

□

Remark 5.2.6 *If $\Gamma \subset (\lambda_N, +\infty]$, we have the same result with x_0 the left extremity of Γ. In the case $\Gamma \subset \mathbb{R} \setminus [\lambda_1, \lambda_N]$ being at both sides of the spectral interval, the maximum is attained at one of the two nearest point of $[\lambda_1, \lambda_N]$.*

Theorem 4.1.3 and Lemma 5.2.5 give the following upper bounds in terms of Green functions. The sharpness follows from the weighted Bernstein-Walsh inequality (4.3).

Corollary 5.2.7 *For any integer $n > d+1 \geq 1$, let $a = a_{d,n}$ be equal to λ_{d+1} if $n \geq \sum_{j=1}^{d} \sqrt{\frac{\beta-\lambda_j}{\lambda_{d+1}-\lambda_j}} = \eta(\lambda_{d+1})$, otherwise let a be the unique solution $> \lambda_{d+1}$ of the equation $n = \sum_{j=1}^{d} \sqrt{\frac{\beta-\lambda_j}{a-\lambda_j}}$. If $\Gamma \subset [-\infty, \lambda_1)$ and x_0 designates the right extremity of Γ, then*

$$E_n([\lambda_{d+1}, \lambda_N], \Gamma) \leq e^{C_{BW}} \exp\Big(-n g_{[a,\beta]}(x_0, \infty) + \sum_{j=1}^{d} g_{[a,\beta]}(x_0, \lambda_j)\Big),$$

being sharp up to the factor $e^{C_{BW}}$. Thus

$$\|f(A)b - f_n\| \leq K \exp\Big(-n g_{[a,\beta]}(x_0, \infty) + \sum_{j=1}^{d} g_{[a,\beta]}(x_0, \lambda_j)\Big),$$

with $K = e^{C_{BW}} 2\|b\| \max_{z\in[\lambda_1,\lambda_N]} |\hat{f}(z)|$.

Similarly to Corollary 5.1.1, Corollary 5.2.7 gives us for each $d < n-1$ an upper bound for the function $n \mapsto \log E_n(\Lambda(A), \Gamma)$, each of them having the shape of a straight line for sufficiently large n, with the slope $-g_{[\lambda_{d+1},\beta]}(x_0, \infty)$ of these straight lines decreasing with d, but the abscissa in general increases. We thus hope that $\log E_n(\Lambda(A), \Gamma)$ is close to the value of the concave lower envelope of these straight lines (see the particular example of Figure 5.3). Let us suggest an approximate solution for a near optimal d.

Theorem 5.2.8 *Let σ and $S(t) = [a(t), \beta]$ for $0 < t < \|\sigma\|$ be as in Lemma 5.1.3, A be a Hermitian matrix with spectrum $\lambda_1 < \lambda_2 < ... < \lambda_N \leq \beta$, $\Gamma \subset [-\infty, \lambda_1)$ and x_0 designates the right extremity of Γ. If the integers $n \geq 2$ and $d = d_n \in \{0, 1, ..., n-2\}$ are such that*

$$\text{for } j = 1, 2, ..., d: \quad \sigma((-\infty, \lambda_j]) \geq j/N, \tag{5.9}$$

$$\lambda_d < a(n/N) \leq \lambda_{d+1} \quad (\text{or } a(n/N) \leq \lambda_1 \text{ in case } d = 0), \tag{5.10}$$

then

$$E_n(\Lambda(A), \Gamma) \leq e^{C_{BW}} \exp\left(-N \int_0^{n/N} g_{S(\tau)}(x_0, \infty) \mathrm{d}\tau\right).$$

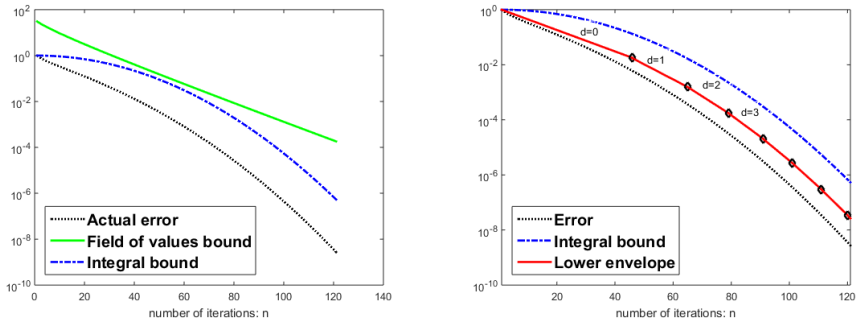

Figure 5.3: *Illustration of Corollary 5.2.7. As lower bound we draw on both plots the error* $\|f(A)b - f_n\|/\|f_1\|$*, with* A *the diagonal matrix with eigenvalues* $\lambda_j = \frac{j}{N}\left(2 - \frac{j}{N}\right)$ *given by Example 5.1.5,* $b = (1, \dots, 1)^T$ *and* $f(x) = 1/\sqrt{x}$ *(black dotted line). We also have reproduced on both plots the integral bound from Theorem 5.2.8(a) with* $C_{BW} = 1$ *(blue dash-dotted line). On the left we compare the two preceding plots with the upper bound given in Proposition 5.2.1 divided by* $\|f_1\|$ *(green solid line). On the right we have drawn the polygon obtained from the lower envelope of all straight lines in Corollary 5.2.7, where we indicate in the plot the correspondence between a segment and the choice of* d *(red solid line).*

Let us note that the above choice (5.10) of d is nearly optimal in the sense of the N-th root with a proof similar to the one given in the case of CG after Theorem 5.1.4

$$\lim_{\substack{n,N\to\infty \\ n/N\to t}} E_{n,N}([\lambda_{d_{n,N}+1,N}, \beta], \Gamma)^{1/N} = \liminf_{\substack{n,N\to\infty \\ n/N\to t \\ 0\le d<n-1}} E_{n,N}([\lambda_{d+1,N}, \beta], \Gamma)^{1/N}.$$

In our Figures 5.3 and 5.4, for the function $z^{-1/2}$, the maximum is attained at $x_0 = 0$. This explains that we find the same behaviors as for CG. A direct consequence of Theorem 5.2.8 and Theorem 5.2.3 is the following result.

Corollary 5.2.9 *Under the assumptions of Theorems 5.2.3 and 5.2.8, we have*

$$\|f(A)b - f_n\| \le K \exp\left(-N \int_0^{n/N} g_{S(\tau)}(x_0, \infty) \mathrm{d}\tau\right)$$

where $K = 2\|b\| \max_{z\in[\lambda_1,\lambda_N]} |\hat{f}(z)| e^{C_{BW}}$.

5.2.3 Proof of Theorem 5.2.8

By definition, it is clear that $E_n(\Lambda(A), \Gamma) \le E_n([\lambda_{d+1}, \beta], \Gamma)$. Using the same notations as in Subsection 5.1.2 and similar arguments as in the proof of Theorem 5.1.4, we prove Theorem 5.2.8. Indeed, let us apply Theorem 4.1.3 with

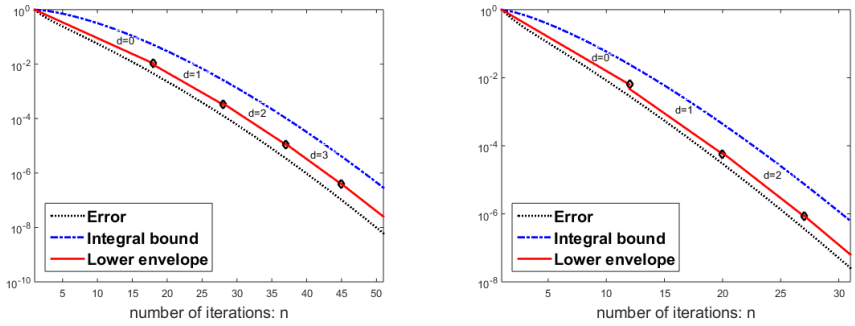

Figure 5.4: *Illustration of Theorem 5.2.8(a) and Corollary 5.2.9. We consider* λ_j *for* $j = 1, ..., N = 1000$ *as in Example 5.1.6, where on the left* $s = 0.5$ *and on the right* $s = 1$ *(s is based on the density). As lower bound we draw on both plots the actual error* $\|f(A)b - f_n\|$*, with* A *the diagonal matrix with eigenvalues the* λ_j *(black dotted line). The polygons are obtained from Theorem 5.2.8(a) (red solid line), where we indicate in the plot the correspondence between a segment and the choice of* d*. We also draw the integral bound (blue dash-dotted line) with* $C_{BW} = 0$*.*

$\alpha = \lambda_{d+1}$, the external field $Q(x) = U^{\widetilde{\rho}/k}(x)$, and where the extremal measure μ and the constant F are as in (4.1), (4.2), to get the chain of inequalities

$$\begin{aligned} E_n([\lambda_{d+1}, \beta], \Gamma) &= \min_{p \in \Pi_k} \frac{\|e^{-U^\rho} p\|_{[\lambda_{d+1},\beta]}}{\min_{x\in\Gamma} e^{-U^\rho(x)}|p(x)|} \\ &\leq \min_{p \in \Pi_k} \frac{\|e^{-U^{\widetilde{\rho}}} p\|_{[\lambda_{d+1},\beta]}}{\min_{x\in\Gamma} e^{-U^{\widetilde{\rho}}(x)}|p(x)|} \text{ (Lemma 5.1.7)} \\ &\leq \frac{\|e^{-U^{\widetilde{\rho}}} P_k\|_{[\lambda_{d+1},\beta]}}{\min_{x\in\Gamma} e^{-U^{\widetilde{\rho}}(x)}|P_k(x)|} \\ &\leq e^{C_{BW}} \max_{x\in\Gamma} \exp\Big(kU^\mu(x) + U^{\widetilde{\rho}}(x) - kF\Big), \end{aligned}$$

where in the last inequality we used that

$$\begin{aligned} \|e^{-U^{\widetilde{\rho}}} P_k\|_{[\lambda_{d+1},\beta]} &\leq e^{C_{BW}} \min_{x\in\Gamma} \Big\{ e^{-U^{\widetilde{\rho}}(x)}|P_k(x)|e^{-k\Theta(x)} \Big\} \\ &\leq e^{C_{BW}} \min_{x\in\Gamma} \Big\{ e^{-U^{\widetilde{\rho}}(x)}|P_k(x)| \Big\} \max_{x\in\Gamma} e^{-k\Theta(x)}. \end{aligned}$$

Now Lemma 5.1.8 implies

$$E_n([\lambda_{d+1}, \beta], \Gamma) \leq e^{C_{BW}} \max_{x\in\Gamma} \exp\Big(n(U^{\nu_{t,\sigma}}(x) - C_{t,\sigma})\Big)$$

On the other hand, with $t = n/N$, we deduce as in (5.4) that

$$\exp\Big(-N \int_0^{n/N} g_{S(\tau)}(x, \infty) d\tau\Big) = \exp\Big(n(U^{\nu_{t,\sigma}}(x) - C_{t,\sigma})\Big),$$

with $\nu_{t,\sigma}$ the solution of the constrained equilibrium problem, and $C_{t,\sigma}$ the corresponding constant. The fact that the maximum is attained at x_0 is a consequence of the monotony of the Green functions. So Theorem 5.2.8 is proved.

5.3 Single repeated pole

In the particular case of one repeated pole $\xi \in \mathbb{C} \setminus \Lambda(A)$, Theorem 5.2.3 says that for any $\tilde{s}_n \in \Pi_n/q_{n-1}$, with $q_{n-1}(z) = (z-\xi)^{n-1}$, we have

$$\|f(A)b - f_n\| \leq 2 \max_{z\in[\lambda_1,\lambda_N]} |\hat{f}(z)| \frac{\|\tilde{s}_n(A)b\|}{\min_{x\in\Gamma} |\tilde{s}_n(x)|}.$$

In particular, if the numerator of $\tilde{s}_n$ has degree $n-1$, we have $\tilde{s}_n(z) = p_{n-1}(\frac{1}{z-\xi})$ for a certain $p_{n-1} \in \Pi_{n-1}$. In the following we consider the pole $\xi \in \mathbb{R}\setminus\Lambda(A)$ and we introduce the transformation $T(z) = (\xi - z)^{-1}$, which implies the inequality

$$\|f(A)b - f_n\| \leq 2\|b\| \max_{z\in[\lambda_1,\lambda_N]} |\hat{f}(z)| E_{n-1}(\Lambda(T(A)), T(\Gamma)), \tag{5.11}$$

where $T(A) = (\xi - A)^{-1}$ is Hermitian.

The preceding equation and the sharpness of the Bernstein-Walsh inequality (Theorem 4.1.3) give the following result.

Proposition 5.3.1 *For any integer $n > d+1 \geq 1$, let $a = a_{d,n}$ be equal to λ_{d+1} if $n \geq \sum_{j=1}^{d} \sqrt{\frac{(\beta-\lambda_j)(\xi-\lambda_{d+1})}{(\lambda_{d+1}-\lambda_j)(\xi-\beta)}}$, and else let a be the unique solution $> \lambda_{d+1}$ of the equation $n = \sum_{j=1}^{d} \sqrt{\frac{(\beta-\lambda_j)(\xi-a)}{(a-\lambda_j)(\xi-\beta)}}$. Moreover, if $\Gamma \subset [-\infty, \lambda_1)$, we obtain*

$$E_n(\Lambda(T(A), T(\Gamma)) \leq e^{C_{BW}} \max_{x\in\Gamma} \exp\Big(-n g_{[a,\beta]}(x,\xi) + \sum_{j=1}^{d} g_{[a,\beta]}(x,\lambda_j)\Big),$$

being sharp up to the factor $e^{C_{BW}}$. And thus

$$\|f(A)b - f_n\| \leq K \max_{x\in\Gamma} \exp\Big(-(n-1) g_{[a,\beta]}(x,\xi) + \sum_{j=1}^{d} g_{[a,\beta]}(x,\lambda_j)\Big), \tag{5.12}$$

with $K = e^{C_{BW}} 2\|b\| \max_{z\in[\lambda_1,\lambda_N]} \hat{f}(z)$.

Proof : We easily obtain $\eta(T(z)) = \sum_{j=1}^{d} \sqrt{\frac{T(\beta)-T(\lambda_j)}{T(z)-T(\lambda_j)}} = \sum_{j=1}^{d} \sqrt{\frac{(\beta-\lambda_j)(\xi-z)}{(z-\lambda_j)(\xi-\beta)}}$. Noting that $T(\xi) = \infty$, Theorem 4.1.3 implies

$$\begin{aligned} E_n(\Lambda(T(A), T(\Gamma)) &\leq e^{C_{BW}} \max_{y\in T(\Gamma)} \exp\Big(-n g_{T([a,\beta])}(y,\infty) \\ &\qquad + \sum_{j=1}^{d} g_{T([a,\beta])}(y, T(\lambda_j))\Big) \\ &\leq e^{C_{BW}} \max_{x\in\Gamma} \exp\Big(-n g_{[a,\beta]}(x,\xi) + \sum_{j=1}^{d} g_{[a,\beta]}(x,\lambda_j)\Big), \end{aligned}$$

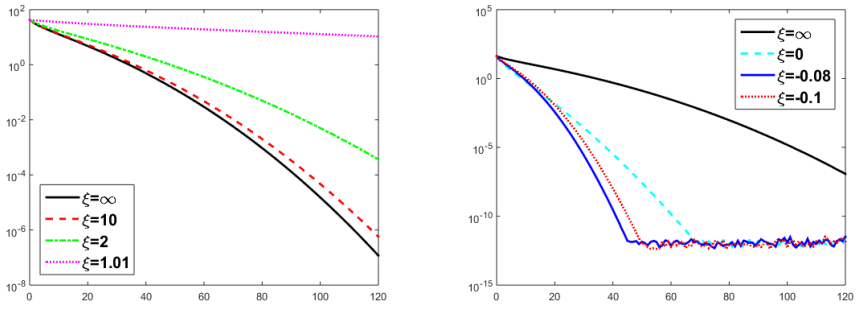

Figure 5.5: *Illustration of Proposition 5.3.1. We draw on both plots the error* $\|f(A)b - f_n\|$*, with* A *the diagonal matrix with eigenvalues* $\lambda_j = \frac{j}{N}\left(2 - \frac{j}{N}\right)$ *given by Example 5.1.5 for different values of* ξ*, and for* $f(z) = z^{-1/2}$*. On the left we consider* $\xi \geq \lambda_N$ *and on the right* $\xi \leq \lambda_1$*.*

where the last inequality is a consequence of [103, Eqn. (II.4.4)]. Now Equation (5.12) is a consequence of (5.11). □

If ξ is on the right of the spectral interval, the function T is positive and increasing on $(-\infty, \xi]$, thus the maximum is attained at the right extremity of Γ by the same arguments as before. By derivating the Green function $g_{[a,\beta]}(x, \xi)$ with respect to ξ we can prove that the best choice is $\xi = \infty$, which means that we do not improve the convergence compared to the polynomial case (see Figure 5.5 on the left). On the other hand, if we choose ξ on the left of the spectral interval, we see on Figure 5.5 on the right that we can obtain better rates of convergence. It seems to be a difficult problem to find the best ξ as a, β and d depend on ξ. On the example proposed the best ξ lies in Γ.

Let us make an heuristic remark. We have

$$\sum_{j=1}^{d} \sqrt{\frac{(\beta - \lambda_j)(\xi - a)}{(a - \lambda_j)(\xi - \beta)}} = \frac{\xi - a}{\xi - \beta} \sum_{j=1}^{d} \sqrt{\frac{(\beta - \lambda_j)}{(a - \lambda_j)}},$$

so up to the factor $\frac{\xi - a}{\xi - \beta}$ we find the same quantity as in Corollary 5.2.7. We observe that if ξ is on the right of the spectral interval, the quotient $\frac{\xi - a}{\xi - \beta}$ is greater than one, and thus we consider less outliers. This explains why we should take $\xi = \infty$ in this case. On the other hand, if ξ is on the left of the spectral interval, the quotient $\frac{\xi - a}{\xi - \beta}$ is less than one, and we can consider more outliers. This explains why we should consider $\xi \leq \lambda_1$ to improve the bound given by the polynomial case.

5.4 Extension to complex contour

For general functions f being analytic in some neighborhood of $[\lambda_1, \lambda_N]$, by choosing a suitable contour Γ encircling $[\lambda_1, \lambda_N]$, we still obtain from the Cauchy

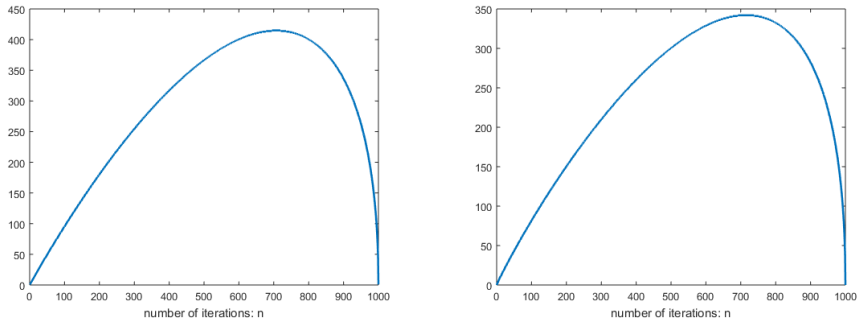

Figure 5.6: *Illustration of Lemma 5.4.1. We draw on both plots the value of $k = n - d$ for each n, with $d = d_n$ given in Theorem 5.1.4. On the left we consider Example 5.1.5 and on the right Example 5.1.6 with $s = 1$.*

integral formula a representation as in (5.6), for z in the interior of Γ.

Recall that from Remark 4.2.5(c), in the case of an interval $[a, \beta]$, we have for $x \in \mathbb{C} \setminus \mathbb{R}$,

$$|\log|P_k(x)| + kU^\mu(x)| \leq \frac{((\beta - a)/2)^2}{\mathrm{dist}(x, [a, \beta])^2} \left(\frac{12\pi}{k}\right)^{1/3}.$$

Thus the inequality of Theorem 4.2.1(b) also holds for non-real x up to some constant C which depends on the distance between x and $[a, \beta]$. This implies that in Theorem 4.1.3, equation (4.7), for $x \in \mathbb{C} \setminus \mathbb{R}$ and for a certain constant C, we have

$$\frac{\|w^k P_k\|_{\mathrm{supp}(\mu)}}{|w(x)^k P_k(x)|} \leq e^{C_{BW}+C}\, e^{-k\Theta(x)}. \tag{5.13}$$

Hence, up to e^C, inequalities in Corollary 5.2.7 hold. If Γ encircles $[\lambda_1, \lambda_N]$ Lemma 5.1.7 is false and our proof of Theorem 5.2.8 does not hold.

Note that for sufficiently large k the quantity $\frac{((\beta-a)/2)^2}{\mathrm{dist}(\Gamma,[a,\beta])^2} \left(\frac{12\pi}{k}\right)^{1/3}$ becomes small. Sufficiently large k makes sense. Indeed, in all our examples $k = n - d_n$ is increasing with n for a ratio n/N not too large, say $n \leq N/2$, and we have the following result.

Lemma 5.4.1 *If we suppose $\lambda_{d_n+1} \leq (\lambda_1 + \beta)/2$, then the number $k = n - d_n$ associated to our choice of near optimal d_n in Theorem 5.2.8 becomes large with n.*

Proof : We have by Lemma 5.1.3

$$n = \int_{a(0)}^{a(n/N)} \sqrt{\frac{\beta - y}{a(n/N) - y}} N \mathrm{d}\sigma(y),$$

which implies

$$n \geq \int_{a(0)}^{\lambda_d} \sqrt{\frac{\beta - y}{a(n/N) - y}} \mathrm{d}\rho(y).$$

And thus using the decay of the function η (defined in Equation 4.5), we obtain

$$n \geq \sum_{j=1}^{d} \sqrt{\frac{\beta - \lambda_j}{\lambda_{d+1} - \lambda_j}} \geq d\sqrt{\frac{\beta - \lambda_1}{\lambda_{d+1} - \lambda_1}}.$$

So $\lambda_{d+1} \leq (\lambda_1 + \beta)/2$ implies

$$n \geq d\sqrt{2},$$

and thus $k \geq (\sqrt{2} - 1)d_n$ becomes large with n as d_n is increasing with n.

□

For general analytic functions f such as the exponential function, the equality $|f(z)| = |\hat{f}(z)|$ for $z \in \Lambda(A)$ is no longer true (compare to Remark 5.2.4). Although the integral does not depend on the choice of Γ, the functions $\hat{f}$ of (5.7) might be of much larger modulus than f, depending on the choice of Γ. For example, for large spectral intervals, the function f may highly vary and our overestimation becomes crude. It can be useful to choose the integration contour dependent on n.

Conclusion

In this dissertation, we have presented new formulas to explain the superlinear convergence of Krylov methods.

In Chapters 1 and 2 we have introduced the basic tools used in this thesis, we have given a state of the art of linear bounds for MR and OR methods, and we have introduced the notions of outliers and of superlinear convergence.

In Chapter 3 we have improved and generalized a paper of Ipsen et al. concerning the convergence of MR methods. In particular, when the inclusion set is a disk, we have given a more general upper bound such that the ratio with the lower bound tends to one. We also gave an upper bound in the case of convex sets. We believe that the factor 3^d obtained in the convex case can be reduced to apply our analysis to several outliers with a more modest constant still depending on d. Another perspective of research could be to generalize our work to more general sets, like non-convex sets, or sets having several connected components by using Faber-Walsh polynomials.

In Chapter 4 we have shown that the weighted Bernstein-Walsh inequality is sharp up to some new universal constant C_{BW}, in the particular case of an external field being given by the logarithmic potential of some positive measure supported on the left of Σ. We believe that, with an optimal choice of C_{BW}, the quantity $e^{C_{BW}}$ is of modest size. This is clearly not true for our present explicit upper bound of C_{BW}, and remains a direction of future research, maybe asymptotic analysis could be helpful. Our main tool is a variation of the technique of Totik of discretizing a logarithmic potential, provided that the underlying measure has a weight satisfying some monotonicity and/or convexity assumptions. We also believe that our result on the discretization of a potential can be generalized to more general measures, for example without the assumption that $t \mapsto \frac{g(t)}{t-a}$ is convex on (a, b), which is used only once in the proof of Lemma 4.4.2. This possibly would allow us to consider both small and large eigenvalues as outliers.

Finally, in chapter 5, we have seen that our new sharpness result for the weighted Bernstein-Walsh inequality leads to a variety of new explicit bounds for the convergence of Krylov methods. By approximately optimizing the number of outliers, we are able to partly show Conjecture 2.3.1 formulated by Beckermann and Kuijlaars in terms of means of Green functions. We establish a new upper bound for CG and for the approximation of matrix functions in form of an

inequality for every iteration index n. In addition, our bounds are valid for a single matrix and do no longer require to consider sequences of systems of equations with a joint eigenvalue distribution. Note that the above-mentioned conjecture on the CG convergence remains open for general sets $S(t)$.

Bibliography

[1] J. Agler, J.E. McCarthy, Pick Interpolation and Hilbert Function Spaces, Grad. Stud. Math. **44** (2002).

[2] A.C. Antoulas, Approximation of large-scale dynamical systems, Advances in Design and Control, SIAM (2005).

[3] W.E. Arnoldi, The principle of minimized iteration in the solution of the matrix eigenvalue problem, *Quart. Appl. Math.* **9** (1951) 17-29.

[4] O. Axelsson, G. Lindskog, On the eigenvalue distribution of a class of preconditioning methods, *Numer. Math.* **48** (1986) 479-498.

[5] O. Axelsson, G. Lindskog, On the rate of convergence of the preconditioned conjugate gradient method, *Numer. Math.* **48** (1986) 499-523.

[6] C. Badea, B. Beckermann, Spectral sets, Chapter 37 in Handbook of Linear Algebra, 2nd edition (2014).

[7] G.A. Baker, P. Graves-Morris, Padé approximants, Encyclopedia of Mathematics and its Applications **59** Cambridge University Press, 2nd edition (2009).

[8] L. Baratchart, V.A. Prokhorov, E.B. Saff, On Blaschke products associated with n-widths, *J. Approx. Theory* **126** (2004) 40-51.

[9] D. Bau and L. N. Trefethen, Numerical linear algebra, SIAM (1997).

[10] B. Beckermann, A note on the convergence of Ritz values for sequences of matrices, *Technical Report ANO 408* (2000).

[11] B. Beckermann, Image numérique, GMRES et polynômes de Faber, *C. R. Math. Acad. Sci. Paris* **340** (2005) 855-860.

[12] B. Beckermann, A.B.J. Kuijlaars, Superlinear convergence of conjugate gradients, *SIAM J. Numer. Anal.* **39** (2001) 300-329.

[13] B.Beckermann, A.B.J. Kuijlaars, On the sharpness of an asymptotic error estimate for conjugate gradients, *BIT* **41** (2001) 856-867.

[14] B. Beckermann, A.B.J. Kuijlaars, Superlinear CG convergence for special right-hand sides, *Electron. Trans. Numer. Anal.* **14** (2002) 1-19.

[15] B. Beckermann, Discrete orthogonal polynomials and superlinear convergence of Krylov subspace methods in numerical linear algebra, *Lecture Notes in Mathematics* **1883**, Springer-Verlag (2006) 119-185.

[16] B. Beckermann, An error analysis for rational Galerkin projection applied to the Sylvester equation, *SIAM J. Numer. Anal.* **49** (2011) 2430-2450.

[17] B. Beckermann, S. Güttel, Superlinear convergence of the rational Arnoldi method for the approximation of matrix functions, *Numer. Math.* **121** (2012) 205-236.

[18] B. Beckermann, S. Güttel, R. Vandebril, On the convergence of rational Ritz values, *SIAM J. Matrix Anal. Appl.* **31** (2010) 1740-1774.

[19] B. Beckermann, T. Hélart, On the sharpness of the weighted Bernstein-Walsh inequality, with applications to the superlinear convergence of conjugate gradients, *arXiv:1707.07871* (2017).

[20] B. Beckermann, L. Reichel, Error estimates and evaluation of matrix functions via the Faber transform, *SIAM J. Num. Anal.* **47** (2009) 3849-3883.

[21] A. Ben-Israel, T.N.E. Greville, Generalized inverses, Springer-Verlag (2003).

[22] Z. Bujanović, Krylov type methods for large scale eigenvalue computations, PhD thesis, University of Zagreb (2011).

[23] V.S. Buyarov, E.A. Rakhmanov, Families of equilibrium measures in an external field on the real axis, *Sb. Mat.* **190** (1999) 791-802.

[24] S.L. Campbell, I.C.F. Ipsen, C.T. Kelley , C.D. Meyer, GMRES and the minimal polynomial, *BIT* **36** (1996) 664-675.

[25] R. Carden, M. Embree, Ritz value localization for non-Hermitian matrices, *SIAM J. Matrix Anal. Appl.* **33** (2012) 1320-1338.

[26] R. Carden, D.J. Hansen, Ritz values of normal matrices and Ceva's theorem, *Linear Algebra Appl.* **438** (2013) 4114-4129.

[27] B.B.A. Cipra, The best of the 20th century: Editors name top 10 algorithms, *SIAM News* **33** (2000) 20-21.

[28] M. Crouzeix, Numerical range, holomorphic calculus and applications, Manuscript (2005).

[29] M. Crouzeix, C. Palencia, The numerical range is a (1+sqrt(2))-spectral set, *SIAM J. Matrix Anal.* **38** (2017) 649-655.

[30] P.I. Davies and N.J. Higham, A Schur-Parlett algorithm for computing matrix functions, *SIAM J. Matrix Anal. Appl.* **25** (2003) 464-485.

[31] P.I. Davies and N.J. Higham, Computing $f(A)b$ for Matrix Functions f. *Lect. Notes Comput. Sci. Eng.* **47** (2005) 15-24.

[32] P.D. Dragnev, E.B. Saff, Constrained energy problems with applications to orthogonal polynomials of a discrete variable, *J. Anal. Math.* **72** (1997) 223-259.

[33] T.A. Driscoll, Algorithm 756: A MATLAB toolbox for Schwarz-Christoffel mapping, *ACM Trans. Math. Software* **22** (1996) 168-186.

[34] T.A. Driscoll, Algorithm 843: improvements to the Schwarz-Christoffel toolbox for MATLAB, *ACM Trans. Math. Software* **31** (2005) 239-251.

[35] T.A. Driscoll, K-C. Toh, L.N. Trefethen, From potential theory to matrix iterations in six steps, *SIAM Rev.* **40** (1998) 547-578.

[36] V.L. Druskin, L.A. Knizhnerman, Two polynomial methods of calculating functions of symmetric matrices, *USSR Comput. Maths. Math. Phys.* **29** (1989) 112-121.

[37] V. Druskin, L. Knizhnerman, Extended Krylov subspaces: approximation of the matrix square root and related functions, *SIAM J. Matrix Anal. Appl.* **19** (1998) 755-771.

[38] V. Druskin, C. Lieberman, M. Zaslavsky, On adaptive choice of shifts in rational Krylov subspace reduction of evolutionary problems, *SIAM J. Sci. Comput.* **32** (2010) 2485-2496.

[39] V. Druskin, V. Simoncini, Adaptive rational Krylov subspaces for large-scale dynamical systems, *Systems Control Lett.* **60** (2011) 546-560.

[40] N. Dunford, J.T. Schwartz, Linear Operators I, II, John Wiley & Sons (1988).

[41] M. Eiermann, On semi-iterative methods generated by Faber polynomials, *Numer. Math.* **56** (1989) 139-156.

[42] M. Eiermann, O. Ernst, Geometric aspects of the theory of Krylov subspace methods, *Acta Numer.* **10** (2001) 251-312.

[43] A. El Guennouni, K. Jbilou, A.J. Riquet, Block Krylov Subspace Methods for Solving Large Sylvester Equations, *Numerical Algorithms* **29** (2002) 75-96.

[44] M. Embree, L.N. Trefethen, Spectra and pseudospectra: the behavior of nonnormal matrices and operators, Princeton University Press (2005).

[45] J.van den Eshof, M. Hochbruck, Preconditioning Lanczos approximations to the matrix exponential, *SIAM J. Sci. Comput.* **27** (2006) 1438-1457.

[46] B. Fischer, Polynomial based iteration methods for symmetric linear systems, Wiley & Teubner (1996).

[47] S.D. Fisher, E.B. Saff, The asymptotic distribution of zeros of minimal Blaschke products, *J. Approx. Theory* **98** (1999) 104-116.

[48] A. Frommer, S. Güttel, M. Schweitzer, Convergence of Restarted Krylov Subspace Methods for Stieltjes Functions of Matrices, *SIAM J. Matrix Anal. and Appl.* **35** (2014) 1602-1624.

[49] A. Frommer, V. Simoncini, Matrix functions, Spinger (2008) 275-303.

[50] D. Gaier, Lectures on complex approximation, Birkhäuser (1987).

[51] E. Gallopoulos, V. Simoncini, Convergence properties of block GMRES and matrix polynomials, *Linear Algebra Appl.* **247** (1996) 97-119.

[52] J.B. Garnett, Bounded analytic functions , Graduate Texts in Mathematics **236**, Springer (2007).

[53] T. Göckler, V. Grimm, Uniform approximation of φ-functions in exponential integrators by a rational Krylov subspace method with simple poles, *SIAM J. Matrix Anal. Appl.* **35** (2014) 1467-1489.

[54] I. Gohberg, S. Goldberg, M.A. Kaashoek, Classes of linear operators. Vol. I, Birkhäuser (1990).

[55] G.H. Golub, C.F. Van Loan, Matrix computations, Johns Hopkins University Press (2013).

[56] A.A. Gonchar, Zolotarev problems connected with rational functions, *Math. USSR Sb* **7** (1969) 623-635.

[57] S Goossens, D. Roose, Ritz and harmonic Ritz values and the convergence of FOM and GMRES, *Num. Linear Algebra Appl.* **6** (1999) 281-293.

[58] A. Greenbaum, Comparison of splittings used with the conjugate gradient algorithm, *Numer. Math.* **33** (1979) 181-193.

[59] A. Greenbaum, Iterative methods for solving linear systems, SIAM (1997).

[60] A. Greenbaum, V. Ptàk, Z. Strakos, Any nonincreasing convergence curve is possible for GMRES, *SIAM J. Matrix Anal. Appl.* **17** (1996) 465-469.

[61] A. Greenbaum and L.N. Trefethen, GMRES/CR and Arnoldi/Lanczos as matrix approximation problems, *SIAM J. Sci. Comput.* **15** (1994) 359-368.

[62] A. Greenbaum, Z. Strakoš, Matrices that generate the same Krylov residual spaces, *IMA Vol. Math. Appl.* **60** (1994) 95-95.

[63] V. Grimm, M. Gugat, Approximation of semigroups and related operator functions by resolvent series, *SIAM J. Numer. Anal.* **48** (2010) 1826-1845.

[64] S. Güttel, Rational Krylov methods for operator functions, PhD Thesis, Technische Universität Bergakademie Freiberg (2010).

[65] Guttel, Rational Krylov approximation of matrix functions: Numerical methods and optimal pole selection. *GAMM* **36** (2013) 8-31.

[66] M.R. Hestenes, E. Stiefel, Methods of conjugate gradients for solving linear systems, *J. research Nat. Bur. Standards* **49** (1952) 409-436.

[67] M. Heyouni, K. Jbilou, An extended block Arnoldi algorithm for large-scale solutions of the continuous-time algebraic Riccati equation, *Electron. Trans. Numer. Anal.* **33** (2008) 53-62.

[68] N.J. Higham, Functions of matrices, SIAM (2008).

[69] M. Hochbruck, C. Lubich, On Krylov subspace approximation to the matrix exponential operator, *SIAM J. Numer. Anal.* **34** (1997) 1911-1925.

[70] R.A. Horn, C.R. Johnson, Matrix analysis, Cambridge University Press (2013).

[71] I.C.F. Ipsen, C.D. Meyer, The idea behind Krylov methods, *American Mathematical Monthly* **105** (1998) 889-899.

[72] S. Kaniel, Estimates for some computational techniques in linear algebra, *Math. Comp.* **20** (1966) 369-378.

[73] A.N. Krylov, On the numerical solution of the equation by which, in technical matters, frequencies of small oscillations of material systems are determined, *Izv. Akad. Nauk SSSR* **1** (1931) 555-570.

[74] T. Kövari, Ch. Pommerenke, On Faber polynomials and Faber expansions, *Math. Z.* **99** (1967) 193-206.

[75] A.B.J. Kuijlaars, Which eigenvalues are found by the Lanczos method?, *SIAM J. Matrix Anal. Appl.* **22** (2000) 306-321.

[76] A.B.J. Kuijlaars, Convergence analysis of Krylov subspace iterations with methods from potential theory, *SIAM Rev.* **48** (2006) 3-40.

[77] C. Lanczos, An iteration method for the solution of the eigenvalue problem of linear differential and integral operators, *J. Research Nat. Bur. Standards* **45** (1950) 255-282.

[78] A.L. Levin, D.S. Lubinsky, Christoffel functions, orthogonal polynomials, and Nevai's conjecture for Freud weights, *Constr. Approx.* **8** (1992) 463-535.

[79] A.L. Levin, D.S. Lubinsky, Christoffel functions and orthogonal polynomials for exponential weights on $[-1, 1]$, *Mem. Amer. Math. Soc.* **111**, AMS (1994).

[80] D.S. Lubinsky, V. Totik, How to discretize a logarithmic potential?, *Acta Sci. Math. (Szeged)* **57** (1993) 419-428.

[81] A. Martínez Finkelshtein, Equilibrium problems of potential theory in the complex plane, *Lecture Notes in Mathematics* **1883**, Springer (1994) 79-117.

[82] G. Meinardus, Approximation of functions: theory and numerical methods, Springer-Verlag (1967).

[83] G. Meurant, On the location of the Ritz values in the Arnoldi process, *Electron. Trans. Numer. Anal.* **43** (2014) 188-212.

[84] C. Moler, C. Van Loan, Nineteen dubious ways to compute the exponential of a matrix, *SIAM Rev.* **20** (1978) 801-836.

[85] C. Moler, C. Van Loan, Nineteen dubious ways to compute the exponential of a matrix, *SIAM Rev.* **45** (2003) 3-49.

[86] I. Moret, A note on the superlinear convergence of GMRES, *SIAM J. Numer. Anal.* **34** (1997) 513-516.

[87] Y.M. Nechepurenko, M. Sadkane, A generalization of matrix inversion with application to linear differential-algebraic systems, *Electron. J. Linear Algebra* **23** (2012) 831-844.

[88] O. Nevanlinna, Convergence of iterations for linear equations, Birkhäuser (1993).

[89] C. C. Paige, The computation of eigenvalues and eigenvectors of very large sparse matrices, PhD Thesis, Univ. of London (1971).

[90] N.B. Parlett, The symmetric eigenvalue problem, Prentice-Hall, Inc., Englewood Cliffs, N.J. (1998).

[91] V. Paulsen, Completely bounded maps and operators algebras, *Cambridge studies in advanced mathematics* **78** (2002).

[92] V. Peller, Hankel operators and their applications, Springer-Verlag (2003).

[93] M.J.D. Powell, Approximation theory and methods, Cambridge University Press (1981).

[94] E.A. Rakhmanov, Equilibrium measure and the distribution of zeros of the extremal polynomials of a discrete variable, *Sb. Mat.* **187** (1996) 1213-1228.

[95] T. Ransford, Potential theory in the complex plane, Cambridge University Press (1995).

[96] M. Robbé, M. Sadkane, A Convergence Analysis of GMRES and FOM Methods for Sylvester Equations, *Numerical Algorithms* **30** (2002) 71-89.

[97] M. Robbé, M. Sadkane, Exact and inexact breakdowns in the block GMRES method, *Linear Algebra Appl.* **419** (2006) 265-285.

[98] A. Ruhe, Rational Krylov sequence methods for eigenvalue computation, *Linear Algebra Appl.* **58** (1984) 391-405.

[99] Y. Saad, On the rates of convergence of the Lanczos and the block-Lanczos methods, *SIAM J. Numer. Anal.* **17** (1980) 687-706.

[100] Y. Saad, Analysis of some Krylov subspace approximations to the matrix exponential operator, *SIAM J. Numer. Anal.* **29** (1992) 209-228.

[101] Y. Saad, Iterative methods for sparse linear systems, SIAM (2003).

[102] Y. Saad, M.H. Schultz, GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems, *SIAM J. Sci. Statist. Comput.* **7** (1986) 856-869.

[103] E.B. Saff, V. Totik, Logarithmic potentials with external fields, Springer-Verlag (1997).

[104] K. Schiefermayr, An upper bound for the norm of the Chebyshev polynomial on two intervals, *J. Math. Anal. Appl.* **445** (2017) 871-883.

[105] O. Sète, Some properties of Faber-Walsh polynomials. arXiv:1306.1347v1, 2013.

[106] O. Sète and J. Liesen, Faber-Walsh polynomials on two disjoint intervals. arXiv:1502.07633v1, 2015.

[107] V. Simoncini, On the numerical solution of $AX - XB = C$, *BIT* **36** (1996) 814-830.

[108] V. Simoncini, D.B. Szyld, On the superlinear convergence of MINRES, *Numerical Mathematics and Advanced Applications 2011*, Springer, Berlin, Heidelberg (2013) 733-740.

[109] G.L.G. Sleijpen, A. Van der Sluis, Further results on the convergence behavior of conjugate-gradients and Ritz values *Linear algebra Appl.* **246** (1996) 233-278.

[110] D.C. Sorensen, Implicit application of polynomial filters in a k-step Arnoldi method, *SIAM J. Matrix Anal. Appl.* **13** (1992) 357-385.

[111] G.W. Stewart, Afternotes goes to graduate school, SIAM (1998).

[112] P.K. Suetin, Series of Faber polynomials, Gordon and Breach Science Publishers (1998).

[113] J. Todd, Applications of transformation theory: A legacy from Zolotarev (1847-1878), *Springer, Dordrecht* Approximation Theory and Spline Functions (1984) 207-245.

[114] V. Totik, Weighted approximation with varying weight, *Lecture Notes in Mathematics* **1569**, Springer-Verlag (1994).

[115] A. van der Sluis, H.A. van der Vorst, The rate of convergence of conjugate gradients, *Numer. Math.* **48** (1986) 543-560.

[116] A. van der Sluis, H.A. van der Vorst, The convergence behavior of Ritz values in the presence of close eigenvalues, *Linear Algebra Appl.* **88/89** (1987) 651-694.

[117] H.A. van der Vorst, Iterative Krylov methods for large linear systems, Cambridge University Press (003).

[118] H.A. van der Vorst and C. Vuik, The superlinear convergence of GMRES, *J. Comput. Appl. Math.* **48** (1993) 327-341

[119] J.L. Walsh, Interpolation and approximation by rational functions in the complex domain, *AMS Colloquium Publications* **XX** AMS (1960).

[120] R. Winther, Some superlinear convergence results for the conjugate gradient method, *SIAM J. Numer. Anal.* **17** (1980) 14-17.

[121] N. Young, An introduction to Hilbert space, Cambridge University Press (1988).

[122] D.I. Zolotarev, Application of elliptic functions to questions of functions deviating least and most from zero, *Zap. Imp. Akad. Nauk. St. Petersburg* **30** (1877) 1-59.

Appendix A

Potential theory in the complex plane

In this appendix, we recall definitions and basic facts about logarithmic potential theory. Our review on potential theory is based on the three references [81], [95], and [103]. In [95] you can find a very good introduction to classical potential theory, [81] is a course given during a summer school which summarize the most relevant properties of potentials, and the book [103] is a full account of the potential theory with an external field.

A.1 Logarithmic potentials

Let D be a domain of $\mathbb{C}$ and E a compact of $\mathbb{C}$.

Definition A.1.1 *A function $u : D \to \mathbb{R} \cup \{+\infty\}$ is lower semi continuous (l.s.c.) at $z \in \mathbb{C}$ if*

$$\liminf_{t \to z, t \in D} u(t) \geq u(z).$$

u is l.s.c. in D if it is l.s.c. at every $z \in D$.

Equivalently, u is l.s.c. on D if $\{z \in D / u(z) > \alpha\}$ is open for every $\alpha \in \mathbb{R}$.

Definition A.1.2 *A function $u : D \to \mathbb{R} \cup \{+\infty\}$ is superharmonic on D if*

1. *$u \not\equiv +\infty$,*
2. *u is l.s.c. on D,*
3. *for every open disk $D(a, r)$ contained in D,*

$$u(a) \geq \frac{1}{2\pi} \int_0^{2\pi} u(a + re^{i\theta}) \mathrm{d}\theta.$$

It is important to note that if f is holomorphic in a domain D, then $\log(\frac{1}{|f|})$ is superharmonic on D and harmonic where $f \neq 0$. We can also remark that if $u \in \mathcal{C}^2(D)$, then u is superharmonic if and only if $\Delta u(z) \leq 0$ for $z \in D$.

Let $\mathcal{M}(E)$ denote the collection of all positive measures supported on E, and $\mathcal{M}_1(E)$ the collection of all positive unit measures supported on E.

Definition A.1.3 *The logarithmic potential associated with a positive measure $\mu \in \mathcal{M}(E)$ is defined by*

$$U^\mu(z) = \int \log \frac{1}{|z-t|} \mathrm{d}\mu(t),$$

and its energy is defined by

$$I(\mu) = \int U^\mu \mathrm{d}\mu.$$

The logarithmic potential is harmonic outside the support $\mathrm{supp}(\mu)$ of μ and superharmonic in $\mathbb{C}$ [81, Theorem 2.2]. The minimization energy problem consists in the determination of

$$V_E = \inf\{I(\mu), \mu \in \mathcal{M}_1(E)\}.$$

The constant $V_E \in (-\infty, +\infty]$ is called the Robin constant of E, and the logarithmic capacity of E is defined by

$$cap(E) = e^{-V_E}.$$

If $cap(E) > 0$ (E is non-polar), then there exists a unique measure $\omega_E \in \mathcal{M}_1(E)$ such that $I(\omega_E) = V_E$, which is called the equilibrium measure [103, Theorem I.1.3(b)]. The potential U^{ω_E} associated with ω_E is called the equilibrium potential for E. A fundamental theorem of Frostman [81, Eqn. (3.11)-(3.13)] asserts that if $cap(E) > 0$ (E is non-polar), then

1. $U^{\omega_E}(z) \leq V_E$ everywhere in $\mathbb{C}$,
2. $U^{\omega_E}(z) = V_E$ quasi-everywhere in E,

where quasi-everywhere means everywhere except on a subset of capacity zero. Those properties characterized the equilibrium measure U^{ω_E} [81, Proposition 3.7]. We define the outer domain Ω_∞ of E as the unbounded component of $\mathbb{C} \setminus E$. It is known that $\mathrm{supp}(\omega_\mathrm{E}) \subseteq \partial\Omega_\infty$ and if we have a strict inequality, the set $\partial\Omega_\infty \setminus \mathrm{supp}(\omega_\mathrm{E})$ has capacity zero, and thus $cap(E) = cap(\partial\Omega_\infty)$ [103, Corollary I.4.5].

We have an important link between the equilibrium potential and the Green function with pole at infinity $g_E(z, \infty)$ [81, Section 3.3].

Definition A.1.4 *The Green function with pole at infinity $g_E(z, \infty)$ is defined in the outer domain Ω_∞ of E with the following properties*

1. $g_E(.,\infty)$ *is nonnegative and harmonic in* Ω_∞,
2. $g_E(z,\infty) - \log|z|$ *is harmonic in a neighborhood of* ∞,
3. $\lim_{z\to z_0, z\in\Omega_\infty} g_E(z,\infty) = 0$ *for quasi-every* $z_0 \in \partial\Omega_\infty$.

One can extend g_E to the whole $\mathbb{C}$ by

$$g_E(z,\infty) = \begin{cases} \limsup_{z\to z_0, z\in\Omega_\infty} g_E(z,\infty), \text{ if } z_0 \in \partial\Omega_\infty \\ 0, \text{ if } z \in \text{interior of } P_c(E) \end{cases},$$

where $P_c(E) = \mathbb{C} \setminus \Omega_\infty$ is called the polynomial convex hull of E. We can also define the Green function with pole at $a \neq \infty$. If $cap(\partial E) > 0$ and $a \in \Omega_\infty$, then we have [103, Eqn. (4.4)].

$$g_E(z,a) = g_{E'}(\frac{1}{z-a}, \infty), \tag{A.1}$$

where E' is the domain obtained from E under the mapping $\frac{1}{z-a}$. So questions concerning $g_E(z,a)$ can be transformed into related ones concerning $g_{E'}(z,\infty)$. We have the following important relation [81, Eqn. (3.15)]

$$U^{\omega_E}(z) + g_E(z,\infty) = \log\frac{1}{cap(E)} = V_E.$$

If we suppose that the outer domain is simply connected in $\bar{\mathbb{C}}$, then we have the existence and unicity of a conformal map $\phi : \Omega_\infty \to \mathbb{C}\setminus\mathbb{D}$ such that $\phi(\infty) = \infty$ and $\phi'(\infty) > 0$ (by the Riemann mapping theorem). In the neighborhood of infinity we have

$$\phi(z) = \frac{1}{cap(E)} z + d_0 + \frac{d_{-1}}{z} + \dots \tag{A.2}$$

and we have an important relation between conformal mappings and Green functions given by [103, Section II.4]

$$g_E(z,a) = \log\left|\frac{1-\overline{\phi(a)}\phi(z)}{\phi(a)-\phi(z)}\right|, \tag{A.3}$$

if $a \neq \infty$, and

$$g_E(z,\infty) = \log|\phi(z)|, \tag{A.4}$$

if $a = \infty$. In the particular case of an interval $[a,b]$, we have [103, Example I.3.5]

$$g_{[a,b]}(z,\infty) = \log\left|\frac{2z-a-b}{b-a} + \sqrt{\left(\frac{2z-a-b}{b-a}\right)^2 - 1}\right|,$$

and

$$\text{supp}(\omega_{[a,b]}) = [a,b], \quad \frac{d\omega_{[a,b]}}{dt}(t) = \frac{1}{\pi\sqrt{(t-a)(b-t)}}. \tag{A.5}$$

A.2 Logarithmic potentials with external field

We will briefly describe the mathematical model corresponding to the existence of an external field for a closed set Σ.

Definition A.2.1 *A weight function $w : \Sigma \to [0, +\infty)$ is admissible if*

1. *w is upper semi-continuous on Σ,*
2. *$E_0 = \{z \in \Sigma / w(z) > 0\}$ has positive capacity,*
3. *if Σ is unbounded, then $|z|w(z) \to 0$ as $|z| \to \infty$, $z \in \Sigma$.*

The weights considered in this dissertation are continuous, and the two first conditions are given.

Definition A.2.2 *We define the external field by*

$$Q(z) = \log \frac{1}{w(z)}.$$

For admissible weights we have the following definition.

Definition A.2.3 *For a Borel measure $\mu \in \mathcal{M}(\Sigma)$ we define the weighted energy by*

$$\begin{aligned} I_w(\mu) &= \int \int \log \frac{1}{|z-t|w(z)w(t)} \mathrm{d}\mu(t)\mathrm{d}\mu(z) \\ &= \int U^{\mu}(z)\mathrm{d}\mu(z) + 2\int Q\mathrm{d}\mu. \end{aligned}$$

Note that the last inequality holds when both integrals are finite. The classical case corresponds to choose $\Sigma = E$ compact and $w = 1$ on E. The minimization weighted energy problem is the determination of

$$V_w = \inf\{I_w(\mu), \mu \in \mathcal{M}_1(\Sigma)\}.$$

The constant $V_w \in (-\infty, +\infty)$ is called the weighted Robin constant of Σ.

There exists a unique measure $\mu_w \in \mathcal{M}_1(\Sigma)$ such that $I(\mu_w) = V_w$ [103, Theorem I.1.3(b)], which is called the weighted equilibrium measure (associated with w). The support $S_w = \mathrm{supp}(\mu_w)$ is compact, contained in Σ_0 and has positive capacity [103, Theorem I.1.3(c)]. The potential U^{μ_w} associated with μ_w is called the weighted equilibrium potential for Σ. Setting $F_w = V_w - \int Q\mathrm{d}\mu_w$, we have the following most important properties [103, Theorem I.1.3(d)-(f)],

1. $U^{\mu_w}(z) + Q(z) \geq F_w$ quasi-everywhere in Σ.
2. $U^{\mu_w}(z) + Q(z) \leq F_w$ on S_w.

Hence we have equality quasi-everywhere on S_w. We remark that for a continuous weight w, the first inequality holds in Σ, and thus we have equality on S_w.

A.3 Logarithmic potentials with constraint

Let Q be a continuous external field on a compact set E and σ a finite measure such that $\text{supp}(\sigma) = \text{E}$, $\sigma(E_0) > 1$ and U^σ is continuous. We define

$$\mathcal{M}_1^\sigma(E) = \{\mu \in \mathcal{M}_1(E) \mid 0 \leq \mu \leq \sigma\}.$$

We have the constraint Robin's constant

$$V_w^\sigma = \inf_{\mu \in \mathcal{M}_1^\sigma(E)} I_w(\mu).$$

As before, it can be proved [32, Theorem 2.1] that there exists a unique measure μ_w^σ such that $V_w^\sigma = I(\mu_w^\sigma)$.

Let us look at a constrained energy problem without external field, which is the case in paragraph 2.3.3. We consider σ a positive Borel measure with compact support E on $\mathbb{R}$ which has total mass at most one. We will introduce a parameter $t \in (0, ||\sigma||)$ and define the class

$$\mathcal{M}_{1,t}^\sigma(E) = \{\mu \in \mathcal{M}_1(E)/0 \leq t\mu \leq \sigma\}.$$

We have a unique measure $\nu_{t,\sigma}$ [32, 94] minimizing the logarithmic energy in $\mathcal{M}_{1,t}^\sigma(E)$

$$I(\nu_{t,\sigma}) = \inf\{I(\mu),\ \mu \in \mathcal{M}_{1,t}^\sigma(E)\}.$$

It is characterized by the fact that there exists a constant $C_{t,\sigma}$ such that

$$\begin{aligned} U^{\nu_{t,\sigma}}(z) &= C_{t,\sigma} \text{ for } z \in \text{supp}(\sigma/\text{t} - \nu_{\text{t},\sigma}) = \text{S(t)} \\ U^{\nu_{t,\sigma}}(z) &\leq C_{t,\sigma} \text{ for } z \in \text{supp}(\sigma). \end{aligned}$$

It was observed and proved in [12, Proof of Theorem 2.1]. that if $0 < t_1 < t_2 < \|\sigma\|$, then $S(t_2) \subset S(t_1)$.

There is a connection between the constrained minimization problem and the energy problem in the presence of an external field [12, Proof of Theorem 2.1]

$$-\frac{1}{t}\int_0^t g_{S(\tau)}(0,\infty)\text{d}\tau = -C_{t,\sigma} + U^{\nu_{t,\sigma}}(0). \tag{A.6}$$

This link is essential in our proof of Theorem 5.1.4.

A.4 Notion of balayage

Given a domain $G \subset \bar{\mathbb{C}}$ such that its boundary is a compact of $\mathbb{C}$ of positive capacity, and given μ a Borel measure on G, the balayage problem consists of finding another measure, denoted by $Bal(\mu, \partial G)$, supported on ∂G which has the same mass and such that the potentials conincide on ∂G quasi everywhere (up to some constant). $Bal(\mu, \partial G)$ is called the balayage measure. Such a measure always exists [103, Theorem 4.1 an 4.4].

The notion of balayage can be usefull in the important case of an external field given by a potential of some measure ν with compact support outside of Σ [103, Example II.4.8]. Consider an external field of the form $Q = -cU^{\nu}$, with ν a probability measure with compact support disjoint form Σ and $c \in [0, 1]$. Then we have

$$\mu_w = cBal(\nu, E) + (1 - c)\omega_E.$$

This relation is used in Example 4.1.2 for $E = [a, b]$, $c = l/2k$, and $c\nu = \rho$, where $\rho = (1/2k)\sum_{j=1}^{l} \delta_{q_j}$ with the q_j designating the zeros of the polynomial q.

We have a strong link between the balayage of a Dirac measure δ_a with $a \in \Omega_\infty$ and Green functions via the formula [103, Eqn. (II.4.32)]

$$g_E(z, a) = \log \frac{1}{|z - a|} - \int_{\partial\Omega_\infty} \log \frac{1}{|z - t|} \mathrm{d}Bal(\delta_a, \partial\Omega_\infty)(t) + g_E(a, \infty).$$

We also use this equality in Example 4.1.2.

For a measure ρ supported in $\mathbb{R}$ and having no mass on the interval $[a, b]$, we know that the balayage of ρ on $[a, b]$ is given for $t \in [a, b]$ by [103, Eqn. (IV.4.47)]

$$\frac{\mathrm{d}Bal(\rho, [a, b])}{\mathrm{d}t}(t) = \frac{1}{\pi} \int \frac{|\sqrt{(y - a)(y - b)}|}{|t - y|\sqrt{(t - a)(b - t)}} \mathrm{d}\rho(y). \tag{A.7}$$

This equality is used in the proof of Theorem 4.2.1 to explicitely compute the solution of an extremal problem with a continuous external field.