N°ORDRE : 40065

Université Lille 1 Sciences et Technologies
Ecole Doctorale régionale Sciences Pour l'Ingénieur
Lille Nord-de-France

Thèse pour obtenir le grade de Docteur en Sciences de l'Université Lille 1

présentée par **Alexis Gryson**
le 1er Juillet 2009

Discipline : **Mathématiques Appliquées**

Minimisation d'énergie sous contraintes Applications en algèbre linéaire et en contrôle linéaire

---

Directeur de Thèse : Bernhard Beckermann

Jury :

Hervé Queffélec	Professeur, président du jury	Université Lille 1
Laurent Baratchart	Dir. Recherche	Inria Sophia-Antipolis
Andreas Frommer	Professeur	Bergische Universität Wuppertal
Christian Gout	Professeur	Université de Valenciennes
Arno Kuijlaars	Professeur	Katholieke Universiteit Leuven
Miloud Sadkane	Professeur	Université de Brest
Franck Wielonsky	Maître de conférences	Université Lille 1

Résumé

Le problème de Zolotarev pour des ensembles discrets apparaît pour décrire le taux de convergence de la méthode ADI, dans l'approximation de certaines fonctions matricielles ou encore pour quantifier le taux de décroissance des valeurs singulières de certaines matrices structurées.
De plus, la réduction de modèle constitue un enjeu important en théorie du contrôle linéaire, et on peut prédire la qualité de l'approximation d'un système dynamique linéaire continu stationnaire de grande dimension donné grâce à la résolution approchée d'une équation de Sylvester.
Après avoir prouvé l'existence d'un minimiseur pour le troisième problème de Zolotarev pour des ensembles discrets, on détermine dans cette thèse le comportement asymptotique faible de ce problème sous certaines hypothèses de régularité. Pour mener cette étude, on considère un problème de minimisation d'énergie sous contraintes pour des mesures signées en théorie du potentiel logarithmique.
On discute également la précision de nos résultats asymptotiques pour des ensembles discrets généraux du plan complexe, et une formule intégrale explicite est établie dans le cas particulier de deux sous-ensembles discrets de l'axe réel symétriques par rapport à l'origine.
L'impact de nos résultats théoriques pour l'analyse du taux de convergence de la méthode ADI appliquée pour la résolution approchée d'une équation de Lyapounov est estimé à l'aide de plusieurs exemples numériques après avoir exposé l'algorithme nous permettant d'obtenir les paramètres utilisés.
Mots clés : Théorie du potentiel logarithmique, Zolotarev, Minimisation d'énergie sous contraintes, ADI, Théorie du contrôle linéaire

Abstract

The Zolotarev problem with respect to discrete sets arises naturally to describe both the convergence rate of the ADI method, to compute approximation of various functions of matrices and to quantify the decreasing rate of singular values of structured matrices.
Moreover, the theory of model reduction is a key problem in linear control theory, and the quality of the approximation of continuous stationnary linear dynamical system might be predicted with the computation of the solution of a Sylvester equation.
Once proved the existence of a minimizer for the third Zolotarev problem with respect to discrete sets, we give the weak asymptotic behaviour of the Zolotarev quantity under some regularity hypothesis. In this purpose, we introduce a problem of energy minimization with constraints in logarithmic potential theory with respect to signed measures.
We discuss the accuracy of our results for general discrete sets in the complex plane, and we prove an explicit integral formula in the particular case of two discret subsets of the real axis symmetric with respect to the imaginary axis.
Then, the impact of our theoretical results concerning the analysis of the convergence rate of the ADI method applied to solve a Sylvester equation is estimated with various numerical examples after the description of the algorithm which we used to compute the parameters.
Keywords : Logarithmic potential theory, Zolotarev, Minimal energy problem with constraints, ADI, Linear control theory

Remerciements

Mes premiers remerciements vont à Bernhard Beckermann, pour son encadrement durant ces trois années de thèse, sa disponibilité, la confiance qu'il m'a accordée. J'ai beaucoup apprécié les conférences auxquelles j'ai participé ainsi que mon séjour à Moscou.
Je remercie Laurent Bratchart, Andreas Frommer et Arno Kuijlaars d'avoir accepté d'être les rapporteurs de cette thèse, ainsi que Christian Gout, Hervé Quéffelec, Miloud Sadkane et Franck Wielonsky d'avoir accepté de faire partie du jury.
Un grand merci également à Nikolai Nikolski qui m'a mis en contact avec mon directeur de thèse et à Michel Balazard mon directeur de Mémoire de M2. Tous deux m'ont permis d'acquérir un bagage mathématique bien utile pour mener à bien ma thèse.
Merci également à Alexandre Aptekarev et Vladimir Sorokine qui m'ont beaucoup aidé durant mon séjour à Moscou sur un plan pratique autant que mathématique.
Je remercie Jean d'Almeida pour le soutien au séminaire des doctorants dont j'ai été co-responsable durant la dernière année de ma thèse.
Merci à Franck Wielonsky, mon tuteur pédagogique avec qui j'ai eu plaisir à travailler dans le cadre des enseignements que j'ai fait pour le monitorat.
J'ai aussi une pensée particulière pour le laboratoire Paul Painlevé, l'UFR de mathématiques, et l'Ecole doctorale de l'université de Lille 1.
Je remercie chaleureusement Alain Juhel, Patrick Caron, et Franck Bertrand pour l'aide qu'ils m'ont, chacun à leur manière, apportée.
Enfin, je remercie les thésards du laboratoire Paul Painlevé pour leur soutien, et plus particulièrement Amandine, Anne, Bénédicte, Manal, Saja, Shuyan, Alexandre, Eric, Fabien, Houcine, Léon, Martin, Patrick, Qidi, Raphaël, Vincent, Youcef et Benoît bien évidemment, ainsi que tous ceux qui se sont de près ou de loin investis dans le séminaire des doctorants.

Table des matières

Introduction

De nombreux problèmes de mathématiques appliquées nécessitent la résolution approchée d'une équation de Sylvester en grande dimension

$$AX - XB = C, \tag{1}$$

d'inconnue X. Citons en particulier la discrétisation d'une équation aux dérivées partielles elliptique à variables séparables sur un domaine rectangulaire, où A et B représentent respectivement des opérateurs de différences finies dans les directions d'abscisse et d'ordonnée. Lorsque les matrices A et B sont creuses, la méthode ADI [PeRa55] ou ses variantes, [LeRe93], [Pe00b], [Sa06] sont les plus utilisées dans ce contexte, et le taux de convergence de cette méthode pour un choix de paramètres optimaux fait apparaître un problème d'approximation rationnelle, le troisième problème de Zolotarev pour des ensembles discrets. L'étude de la décroissance des valeurs singulières de certaines matrices structurées [Be04], l'approximation du signe sur des ensembles discrets, problème important en chromodynamique quantique [EFLSV02] et l'évaluation de certaines fonctions de matrices [Hi08], [DrKnZa08] sont autant de problèmes dont l'étude fait apparaître un problème de Zolotarev.

Pour prévoir le comportement de systèmes dynamiques aussi divers qu'un lecteur de disques compacts ou une réaction chimique, il est nécessaire d'effectuer une modélisation mathématique du problème. Une telle modélisation est souvent obtenue à l'aide d'une méthode aux éléments finis, ce qui fournit un modèle de très grande complexité, alors que pour des objectifs de contrôle et de simulation, des modèles de plus petites dimensions sont nécessaires. On souhaite de plus que ces modèles reproduisent fidèlement le comportement du système originel de grande dimension, voir [AnIoRo08] pour des exemples numériques de réductions de modèles. Dans le cas d'un système dynamique continu linéaire stationnaire, on peut prédire la qualité de l'approximation d'un modèle de grande dimension donné grâce à la résolution approchée d'une équation de Sylvester, voir [An05], [AnSo02] et [GuAn04].

Le troisième problème de Zolotarev est défini comme suit : pour E et F deux ensembles disjoints du plan complexe et $\mathcal{R}_n$ l'ensemble des fractions rationnelles de numérateur et dénominateur de degré au plus n, on cherche à déterminer la *quantité de Zolotarev*

$$Z_n(E,F) := \min_{r\in\mathcal{R}_n} \|r\|_{L^\infty(E)} \|r^{-1}\|_{L^\infty(F)}.$$

Ce problème a été introduit par Zolotarev, élève de Tchebychev, en 1877 dans l'article *Applications of elliptic functions to questions of functions deviating least or most from zero* pour illustrer l'apport des fonctions elliptiques de Legendre dans l'étude de certains problèmes de minimisation en théorie d'approximation rationnelle.

L'étude de ce problème dans le cas d'ensembles continus est connue et a donnée lieu à de nombreux travaux. Par exemple dans le cas d'ensembles disjoints de capacité positive E et F, Gonchar [Go78] a montré que la quantité $Z_n(E,F)^{1/n}$ admet une limite pour $n \to +\infty$ décrite en considérant un condensateur de plateau positif E de charge unitaire, et de plateau négatif F également de charge unitaire. Cela généralise les résultats originaux de Zolotarev [Zo32] pour le cas de deux intervalles symétriques par rapport à l'origine.

Le taux de convergence de la méthode ADI est determiné par un problème de Zolotarev sur des ensembles discrets égaux aux spectres des matrices A et B coefficients de l'équation (1). Pour déterminer le taux de convergence de cette méthode, on remplace les ensembles discrets par leurs enveloppes convexes, et on sait dans ce cas déterminer des points asymptotiquement optimaux pour la convergence de la méthode ADI appelés points de Leja-Bagby, voir par exemple à ce sujet [BaTh00, CaLeRe97, IsTh95], et [LeSa01].

Cela dit, aucune étude ne porte à notre connaissance sur l'asymptotique du problème de Zolotarev pour des ensembles discrets, et le fait de remplacer des ensembles discrets par leurs enveloppes convexes peut modifier sensiblement le comportement asymptotique du problème considéré. Notre travail est dédié à l'étude de l'asymptotique de la quantité $Z_n(E_N, F_N)$ pour des familles d'ensembles discrets $(E_N)_N$ et $(F_N)_N$ de cardinal N, où n est choisi de façon à ce que le ratio

$$\frac{\text{nombre d'itérations effectuées}}{\text{taille des matrices considérées}} = \frac{n}{N}$$

admette une limite $t > 0$ pour $n, N \to +\infty$.

De plus, l'existence d'une interaction fructueuse entre l'analyse de la convergence de certaines méthodes itératives en algèbre linéaire numérique et la théorie de l'approximation complexe [Tr90] est un fait connu de longue date, en particulier à travers l'utilisation d'outils provenant de la théorie du potentiel logarithmique [DrToTr98, TrBa97]. Ainsi, les travaux de Beckermann et Kuijlaars portant sur l'étude de la convergence des valeurs de Ritz [Ku00] et de la convergence superlinéaire de l'algorithme du gradient conjugué [BeKu01a, BeKu01b, BeKu02] ont montré que l'étude de l'asymptotique du problème de min max polynômial pour une famille d'ensembles discrets $(E_N)_N$

$$\mathcal{E}_n\left(E_N\right) := \min_{p \in \mathcal{P}_n} \max_{z \in E_N} \left|p(z)\right|,$$

où $\mathcal{P}_n$ désigne l'ensemble des polynômes de degré au plus n prenant la valeur 1 en zéro, forme une étape décisive vers l'étude du comportement asymptotique du problème considéré.

Dans beaucoup d'applications comme la discrétisation d'une équation différentielle où l'on fait varier le pas, on connaît une suite d'ensembles discrets $(E_N)_N$ dont la distribution limite est donnée par une mesure σ. On montre alors en utilisant des résultats précédents concernant l'asymptotique faible de polynômes orthogonaux discrets [Ra96, DrSa97, Be00] que la connaissance de σ permet de décrire l'asymptotique du problème de min max polynômial évoqué ci-dessus.

L'expression asymptotique pour le problème min max polynômial fait intervenir la solution d'un problème extrémal en théorie du potentiel logarithmique qui consiste parmi toutes les mesures μ de masse t satisfaisant la contrainte $\mu \leqslant \sigma$ à rechercher la mesure d'énergie logarithmique minimale. En terme d'électrostatique, cela revient à rechercher l'état d'équilibre pour un conducteur de charge t unités soumis à une contrainte de charge

surfacique. En conséquence, cela permet de quantifier le taux de convergence superlinéaire du gradient conjugué en terme de la distribution σ.

On étend dans cette thèse la démarche décrite ci-dessus pour déterminer l'asymptotique au sens de la racine N-ième de la quantité de Zolotarev pour des ensembles discrets. Le cadre en théorie du potentiel logarithmique pour l'étude de notre problème est donné par un problème de minimisation d'énergie sous contraintes pour des mesures signées. Cela constitue une généralisation des travaux actuels qui considèrent des mesures signées [SaTo97], une formulation vectorielle [NiSo91] ou des mesures positives contraintes [DrSa97].

On s'intéresse au problème de minimisation suivant : pour la donnée d'une fonction continue Q appelée champ extérieur, de deux mesures σ_1, σ_2 et de deux réels t_1, t_2, on cherche à minimiser l'énergie logarithmique avec champ extérieur Q

$$\int\int \log \frac{1}{|z-t|}\,\mathrm{d}\mu(t)\,\mathrm{d}\mu(z) + 2\int Q(z)\,\mathrm{d}\mu(z),$$

sur l'ensemble de mesures sous contraintes

$$\{\mu := \mu_1 - \mu_2,\ \mu_j \text{ mesure de Borel},\ \mu_j(\mathbb{C}) = t_j,\ 0 \leqslant \mu_j \leqslant \sigma_j \text{ pour } j \in \{1,2\}\}.$$

On prouve un résultat d'existence et d'unicité pour la mesure extrémale pour ce problème que l'on caractérise par ses conditions d'équilibre. La généralisation d'autres outils issus de la théorie du potentiel logarithmique comme la fonctionnelle de Mhaksar, Saff et Rakhmanov [MhSa85] ou la formule de Buyarov-Rakhmanov [BuRa99] permet ensuite de quantifier le taux de convergence superlinéaire de la méthode ADI, jusqu'alors observé empiriquement.

Pour poursuivre la généralisation du cas d'ensembles continus, on définit finalement l'analogue des points de Leja-Bagby rationnels dans notre cadre discret, ce qui nous permet de définir un nouvel algorithme de détermination des paramètres pour la méthode ADI. On analyse alors le comportement asymptotique de ces points dans le cas discret grâce aux résultats de théorie du potentiel obtenus précédemment.

Notre travail se divise en huit chapitres dont on résume le contenu ci-dessous.

On définit dans le premier chapitre le troisième problème de Zolotarev pour des ensembles discrets, et l'on fait le lien avec diverses questions d'algèbre linéaire et d'approximation. On montre ainsi que la quantité de Zolotarev pour des ensembles discrets intervient naturellement dans les trois problème suivants : décrire le taux de convergence de la méthode ADI pour la résolution approchée d'une équation de Sylvester, quantifier le taux de décroissance des valeurs singulières des matrices de petit rang de déplacement, et enfin, calculer une approximation rationnelle de la fonction signe sur des ensembles discrets.

Le chapitre suivant est consacré à exposer le lien entre le problème de la réalisation partielle d'un système dynamique continu et la résolution approchée d'une équation de Sylvester. On reprend dans ce chapitre diverses notions bien connues en théorie du contrôle linéaire dans le but d'expliquer ce que nos travaux concernant le taux de convergence de la méthode ADI peuvent apporter dans ce contexte.

Dans le troisième chapitre, on s'intéresse au problème de Zolotarev pour des ensembles discrets en tant que problème d'approximation rationnelle, et non plus comme objet lié à certaines questions de théorie du contrôle ou d'algèbre linéaire. On prouve l'existence

d'un minimiseur avant de s'intéresser aux questions, classiques en théorie d'approximation rationnelle, de localisation des zéros et des pôles et de dégénérescence des solutions.

Le quatrième chapitre est dédié à l'étude d'un problème de minimisation d'énergie sous contraintes en théorie du potentiel logarithmique pour des mesures signées. On suit dans ce chapitre la structure classique d'une étude de minimisation d'énergie en théorie du potentiel logarithmique : après quelques rappels élémentaires en analyse complexe, on définit précisément le problème sous contraintes pour des mesures signées qui fait l'objet de notre étude avant de prouver l'existence et l'unicité de la solution. On établit ensuite les conditions d'équilibre avant de prouver que celles-ci caractérisent la mesure extrémale pour notre problème de minimisation.

Le cinquième chapitre présente le passage à la limite entre notre problème discret et l'étude continue du chapitre précédent. On détermine en effet l'asymptotique de la quantité de Zolotarev pour des ensembles discrets grâce aux résultats de théorie du potentiel logarithmique pour des mesures signées obtenus au chapitre 4. Les hypothèses techniques concernant les familles d'ensembles discrets que l'on impose pour obtenir notre résultat asymptotique sont présentées et commentées.

Le résultat principal du chapitre 5 qui donne l'asymptotique de la quantité de Zolotarev pour des ensembles discrets fait intervenir les constantes extrémales du problème de théorie du potentiel étudié chapitre 4. On prouve dans le sixième chapitre une formule intégrale explicite qui donne ces constantes dans un cas particulier du problème de minimisation d'énergie sous contraintes. Cette formule fait intervenir la partie libre du support de la mesure extrémale pour le problème de minimisation considéré, et on prouve enfin dans le cas symétrique une condition suffisante pour que celle-ci soit égale à la réunion de deux intervalles réels.

La formulation explicite du chapitre 6 fait intervenir la partie libre du support de la mesure extrémale pour le problème de minimisation considéré pour une famille de mesures extrémales de différentes masses. L'objet du septième chapitre est alors de caractériser cette famille d'ensembles sous l'hypothèse que ceux-ci sont donnés par l'union de deux intervalles réels. On introduit à cet effet la fonctionnelle de Mhaskar-Saff-Rakhmanov dont on étudie les points critiques. On se ramène par transformation conforme à un problème plus simple, ce qui nous permet d'obtenir un système de quatre équations intégrales vérifiées par les bornes des intervalles recherchés.

Le huitième et dernier chapitre est dévolu à la confrontation de nos résultats théoriques avec plusieurs expériences numériques. On y étudie en particulier la résolution approchée d'une équation de Sylvester par la méthode ADI pour différents exemples dont le cas d'un exemple modèle, des valeurs propres équidistantes et de la discrétisation du Laplacien $2D$ et $4D$. On compare les taux de convergence que nous obtenons avec l'état de l'art en ce domaine avant d'exposer l'algorithme utilisé pour la détermination des paramètres. On interprète enfin cet algorithme grâce aux résultats de théorie du potentiel logarithmique du chapitre 4.

Chapitre 1

Équation de Sylvester et méthode ADI

De nombreux problèmes de modélisation en mathématiques appliquées débouchent sur la résolution approchée d'une équation de Sylvester $AX - XB = C$ d'inconnue X en grande dimension. En particulier, la discrétisation d'une équation aux dérivées partielles elliptique à variables séparables sur un domaine rectangulaire, où A et B représentent respectivement des opérateurs de différences finies dans les directions d'abscisse et d'ordonnée produit naturellement ce type d'équation.

Dans ce chapitre, on donne la définition et quelques propriétés élémentaires des équations de Sylvester puis on expose rapidement quelques méthodes classiques de résolution d'une telle équation. On s'intéresse ensuite à la méthode ADI qui jouera un rôle important dans cette thèse, on présente donc cette méthode ainsi que quelques-unes de ces variantes dédiées à la résolution approchée d'une équation de Sylvester.

La majoration de l'erreur commise par la méthode ADI pour la résolution d'une équation de Sylvester après n itérations donne lieu à un problème d'approximation rationnelle discret dans le plan complexe, le troisième problème de Zolotarev. On définit alors ce problème avant de donner quelques autres connexions de celui-ci avec des problémes d'algèbre linéaire.

1.1 Équation de Sylvester

Définition 1.1.1. *On note* équation de Sylvester *l'équation*

$$AX - XB = C, \tag{1.1}$$

où $A \in \mathbb{C}^{N\times N}$, $B \in \mathbb{C}^{M\times M}$, $C \in \mathbb{C}^{N\times M}$ sont trois matrices à coefficients complexes données, et X l'inconnue, également élément de $\mathbb{C}^{N\times M}$.

Dans le cas où $A = -B$, l'équation (1.1) *devient*

$$AX + XA = C$$

et est appelée équation de Lyapounov *ou parfois équation de Sylvester symétrique.*

Donnons le résultat classique d'existence et d'unicité de la solution d'une équation de Sylvester après la définition du produit de Kronecker qui permet de reformuler une telle équation en système linéaire.

Définition 1.1.2. *Soit M une matrice carrée à coefficients complexes.*
On note dorénavant $\Lambda(M)$ le spectre de M.

Définition 1.1.3. *Soient*

$$P := [p_{u,v}]_{1\leqslant u\leqslant j, 1\leqslant v\leqslant k} \in M_{j,k}(\mathbb{C}),\ Q = [q_{u,v}]_{1\leqslant u\leqslant l, 1\leqslant v\leqslant m} \in M_{l,m}(\mathbb{C}),$$

on définit le produit de Kronecker *de P et Q noté $P \otimes Q$ comme l'élément de $M_{jl,km}(\mathbb{C})$ défini par blocs successifs de taille $l \times m$ dont le bloc d'indice (u, v) est donné par $p_{u,v}Q$.*

Lemme 1.1.4. *Soient A, B, C trois matrices complexes comme ci-dessus.*
L'équation (1.1) *admet une solution unique si et seulement si $\Lambda(A) \cap \Lambda(B) = \emptyset$.*

PREUVE : Pour toute matrice $X \in \mathbb{C}^{N\times M}$, on note $v(X)$ le vecteur de $\mathbb{C}^{NM}$ défini coordonnée par coordonnée par $v(X)_{kN+j} = X_{j,k}$ pour $0 \leqslant k \leqslant M-1$ et $1 \leqslant j \leqslant N$ par la relation $v_{kN+j} = X_{j,(k+1)}$ (autrement dit, on range les éléments de la matrice X colonne par colonne dans un vecteur $v(X)$).

Alors, l'équation (1.1) est équivalente au système linéaire

$$\big((I_M \otimes A) - (B^T \otimes I_N)\big)\, v(X) = v(C),$$

où l'opérateur $\otimes$ désigne le produit de Kronecker, et on a

$$\Lambda\left((I_M \otimes A) - (B^T \otimes I_N)\right) = \{x - y,\ x \in \Lambda(A),\ y \in \Lambda(B)\},$$

ce qui prouve l'existence et l'unicité de la solution si et seulement si les spectres de A et B sont disjoints. □

Voici maintenant un cas particulier dans lequel la solution de l'équation de Sylvester s'écrit sous forme intégrale.

Définition 1.1.5. *On dit qu'une matrice $X \in M_N(C)$ est* stable *si toutes ses valeurs propres sont de partie réelle strictement négative.*

On remarque que si deux matrices de dimension $N \times N$, A et $-B$ sont stables, alors l'équation $AX - XB = C$ admet une unique solution d'après le lemme 1.1.4.

Lemme 1.1.6. *Si A et $-B$ sont deux matrices stables de dimension $N \times N$, alors la solution X de l'équation de Sylvester $AX - XB = -C$ s'écrit sous la forme suivante :*

$$X = \int_0^{+\infty} \mathrm{e}^{tA}\, C\, \mathrm{e}^{-tB}\, \mathrm{d}t.$$

PREUVE : Il suffit d'utiliser l'égalité suivante valable pour $t \geqslant 0$ pour conclure :

$$\frac{d}{dt}\, \mathrm{e}^{tA}\, C\, \mathrm{e}^{-tB} = A\, \mathrm{e}^{tA}\, C\, \mathrm{e}^{-tB} - \mathrm{e}^{tA}\, C\, \mathrm{e}^{-tB}\, B.$$

□

1.2 Méthodes de résolution d'une équation de Sylvester

1.2.1 Méthodes directes de résolution d'une équation de Sylvester

Il existe beaucoup de méthodes conçues pour la résolution approchée d'une équation de Sylvester (1.1)

$$AX - XB = C,$$

où A et B sont de tailles respectives $N \times N$ et $M \times M$, mais même les méthodes les plus efficaces nécessitent $\mathrm{O}(M^3 + N^3)$ opérations, et surtout, ne tiennent pas compte de la structure de la matrice C qui est de rang petit devant N et M dans beaucoup d'applications.

La formulation en produit de Kronecker permet de transformer la résolution d'une équation de Sylvester en résolution de système linéaire, et cette résolution nécessite $\mathrm{O}(N^3M^3)$ opérations par la méthode d'élimination de Gauss.

Citons ici les méthodes directes décrites dans la littérature données par les algorithmes de Bartels-Stewart [BaSt72], Hessenberg-Schur [GoNaVL79] et Hammarling [Ham82]. Ces algorithmes reposent tous sur le même principe, à savoir transformer A en une matrice triangulaire ou une matrice de Hessenberg par une suite de transformations élémentaires.

Si $A = U^\star TU$ et $B = VRV^\star$ où U et V sont unitaires et T et R sous forme de Schur ou de Hessenberg, on obtient en multipliant l'équation de Sylvester (1.1) par $U^\star$ à gauche et par V à droite l'équation de Sylvester

$$T\left(U^\star XV\right) + \left(U^\star XV\right)R = U^\star CV, \tag{1.2}$$

à coefficients des matrices de Schur ou de Hessenberg.

On peut maintenant obtenir la solution de l'équation modifiée (1.2) en résolvant des systèmes triangulaires supérieurs, et retrouver ensuite la matrice X par multiplication par des matrices unitaires.

Ces algorithmes de résolution directe ont tous une complexité qui exclut *de facto* toute tentative de résolution d'une équation de Sylvester pour des matrices de grandes dimensions. Dans le cas de deux matrices A et B creuses, cas par exemple d'un schéma de discrétisation aux différences finies, ces algorithmes présentent enfin l'inconviénient majeur de ne pas conserver cette structure creuse, ce qui impose le stockage de beaucoup plus de coefficients lors de la transformation de A et B sous forme de Schur ou de Hessenberg.

La méthode ADI de Peaceman et Rachford [PeRa55] et ses variantes forment la classe des algorithmes les plus utilisés pour la résolution approchée d'une équation de Sylvester dans le cas de matrices creuses de grande dimension, on présente ainsi celle-ci dans la section suivante.

1.2.2 Itération ADI, estimation de l'erreur

Avant de définir la méthode ADI et d'évaluer l'erreur commise lors de la résolution approchée d'une équation de Sylvester, on rappelle ici quelques faits bien connus concernant l'algorithme du gradient conjugué afin de mettre en perspective les résultats obtenus.

L'algorithme du gradient conjugué est un algorithme très utilisé pour la résolution approchée du système linéaire $AX = b$ où $A \in \mathbb{R}^{N\times N}$ est symétrique définie positive.

On connaît le résultat classique suivant de majoration de la norme de l'erreur relative e_n commise après n itérations de cet algorithme :

$$\|e_n\| \leqslant E_n\left(\Lambda(A)\right), \tag{1.3}$$

avec pour S un sous-ensemble fermé de l'axe réel

$$E_n(S) := \min_{p\in\mathcal{P}_n} \max_{\lambda\in S} |p(\lambda)|, \tag{1.4}$$

où $\mathcal{P}_n$ désigne l'ensemble des polynômes de degré au plus n à coefficients réels prenant la valeur 1 en 0.

Ainsi, la majoration de l'erreur relative commise après n itérations de la méthode du gradient conjugué donne lieu à un problème de min max polynômial sur un ensemble discret.

Dans le même ordre d'idée, on montre dans ce qui suit que l'étude de l'erreur relative commise lors de la résolution approchée d'une équation de Sylvester par l'intermédiaire de la méthode ADI donne lieu à un problème d'approximation rationnelle de type min max, le problème de Zolotarev.

La méthode ADI, ou *alternating direction implicit method*, a été conçue par Peaceman et Rachford en 1955 [PeRa55] pour résoudre des équations aux dérivées partielles elliptiques, et a été adaptée par Wachspress pour l'équation de Sylvester générale présentée ici, voir par exemple [Wa63], [Wa69] et également à ce sujet les travaux de Birkhoff et Varga [BiVa59].

Ayant choisi une valeur intiale X_0 et déterminé des paramètres correspondant aux $k-1$ premières itérations, pour effectuer la k-ième itération de la méthode, on résout les systèmes linéaires présentés lors des deux étapes suivantes :

$$\begin{aligned} (A - p_k I_N)X_{k+1/2} &= X_k(B - p_k I_M) + C, \\ X_{k+1}(B - q_k I_M) &= (A - q_k I_N)X_{k+1/2} - C, \end{aligned} \tag{1.5}$$

itération qui peut également s'écrire en une seule étape

$$\begin{aligned} X_{k+1} \quad &= \left[(A - q_k I_N)(A - p_k I_N)^{-1}\left(X_k(B - p_k I_M) + C\right) - C\right](B - q_k I_M)^{-1}, \\ &= (A - q_k I_N)(A - p_k I_N)^{-1} X_k (B - p_k I_M)(B - q_k I_M)^{-1}, \\ &+ (p_k - q_k)(A - p_k I_N)^{-1} C (B - q_k I_M)^{-1}, \end{aligned} \tag{1.6}$$

où on a utilisé l'identité

$$(A - q_k I_N)(A - p_k I_N)^{-1} = I + (p_k - q_k)(A - p_k I_N)^{-1}.$$

On omettra dorénavant les indices N et M de dimension d'espace dans l'écriture de la matrice identité pour alléger les notations utilisées.

On vérifie simplement qu'un point X satisfaisant l'équation de Sylvester (1.1) est un point fixe de l'itération ADI : si $X_k = X$, alors

$$\begin{aligned} X_{k+1} &= \left[(A-q_kI)(A-p_kI)^{-1}\left(X(B-p_kI)+AX-XB\right)-AX+XB\right](B-q_kI)^{-1}, \\ &= \left[(A-q_kI)(A-p_kI)^{-1}(AX-p_kX)-AX+XB\right](B-q_kI)^{-1}, \\ &= \left((A-q_kI)X-AX+XB\right)(B-q_kI)^{-1}, \\ &= \left(-q_kX+XB\right)(B-q_kI)^{-1}, \\ &= X, \end{aligned}$$

et $X_{k+1} = X$, la solution de l'équation de Sylvester (1.1) est bien un point fixe de l'itération ADI.

Ainsi, on obtient l'égalité pour $k \geqslant 0$

$$X - X_{k+1} = (A-q_kI)(A-p_kI)^{-1}(X-X_k)(B-p_kI)(B-q_kI)^{-1},$$

ce qui donne par récurrence immédiate la majoration

$$\frac{\|X-X_n\|}{\|X-X_0\|} \leqslant \|r_n(A)\|\,\|r_n(B)^{-1}\|,$$

avec l'élément de $\mathcal{R}_n$ unitaire dont les pôles et les zéros sont donnés par les paramètres utilisés

$$r_n(z) := \prod_{k=0}^{n-1} \frac{z-q_k}{z-p_k}.$$

Hypothèse 1.1. *On supposera dorénavant que les matrices A et B mises en jeu dans l'équation de Sylvester $AX - XB = C$ sont normales.*

Proposition 1.2.1. *Avec les notations précédentes, on a la majoration suivante pour l'erreur commise après n iterations de la méthode ADI :*

$$\frac{\|X-X_n\|}{\|X-X_0\|} \leqslant \|r_n\|_{L^\infty(\Lambda(A))}\|r_n^{-1}\|_{L^\infty(\Lambda(B))}. \tag{1.7}$$

PREUVE : Avec les notations précédentes, on a obtenu en étudiant l'erreur commise lors de n itérations de la méthode ADI la borne supérieure pour l'erreur commise

$$\frac{\|X-X_n\|}{\|X-X_0\|} \leqslant \|r_n(A)\|\,\|r_n(B)^{-1}\|.$$

Comme les matrices A et B sont supposées normales par hypothèse, elle sont diagonalisables en base orthonormée, ce qui donne immédiatement le résultat. □

1.2.3 Problème de Zolotarev

On définit ici le problème de Zolotarev dont l'étude fait l'objet de la majeure partie de l'étude théorique menée dans cette thèse.

Définition 1.2.2. *Notons $\mathbb{C}_\infty$ la sphère de Riemann. Le* problème de Zolotarev classique *consiste, étant donnés deux ensembles disjoints E et F de $\mathbb{C}_\infty$ et un entier $n \geqslant 0$, à déterminer*

$$Z_n(E,F) := \inf_{r\in\mathcal{R}_n} \frac{\max_{z\in E}|r(z)|}{\min_{z\in F}|r(z)|} = \inf_{r\in\mathcal{R}_n} \|r\|_{L^\infty(E)}\|r^{-1}\|_{L^\infty(F)}$$

où $\mathcal{R}_n$ désigne l'ensemble des fractions rationnelles à coefficients complexes de numérateur et dénominateur de degré au plus n.

On peut également considérer le problème de Zolotarev généralisé

$$Z_{n,m}(E,F) := \inf_{r\in\mathcal{R}_{n,m}} \frac{\max_{z\in E}|r(z)|}{\min_{z\in F}|r(z)|}$$

où $\mathcal{R}_{n,m}$ désigne l'ensemble des fractions rationnelles à coefficients complexes de numérateur de degré au plus n et de dénominateur de degré au plus m.

Corollaire 1.2.3. *Dans le cas de paramètres $p_1,\dots p_n$ et $q_1,\dots q_n$ optimaux ci-dessus, on obtient*

$$\frac{\|X-X_n\|}{\|X-X_0\|} \leqslant Z_n\left(\Lambda(A)\Lambda(B)\right).$$

1.2.4 Variantes de la méthode ADI

On présente ici quelques variantes de la méthode ADI.

Smith a par exemple proposé en 1968 dans [Sm68] une méthode basée sur la répétition d'un paramètre reprise dans [Sa06] : l'équation (1.1) est équivalente à l'équation

$$X = X_1 + r(A)Xr(-B),\ \text{où } r(z) := \frac{z+p}{z-p},\ p>0, \tag{1.8}$$

où X_1 est le résultat de la première itération donné en (1.6) pour $k=0$ et $X_0 = 0$.

La solution formelle de l'équation (1.8) est donnée par

$$X = \sum_{\ell=0}^{+\infty} r(A)^\ell X_1 r(-B)^\ell,$$

et la méthode de Smith consiste à approcher la solution de (1.8) par troncature à partir de cette série.

Cette méthode converge dès lors que A et B sont deux matrices stables du fait que l'homographie r envoie conformément demi-plan gauche sur le disque unité.

Pour des équations de Sylvester dont le second membre est de rang petit, la méthode de Smith nécessite la résolution de systèmes linéaires dont les coefficients sont des matrices creuses, et le taux de convergence de cette méthode est linéaire. Il existe de plus une formule simple pour les itérées d'ordre une puissance de 2 : en notant

$$Y_0 = X_1,\ Y_{k+1} = Y_k + r(A)^{2^k} Y_k r(B)^{2^k},$$

alors

$$Y_k = \sum_{\ell=0}^{2^k-1} r(A)^\ell X_1 r(B)^\ell,$$

où l'on obtient bien sûr $r(A)^{2^k}$ et $r(B)^{2^k}$ à partir de $r(A)^{2^{k-1}}$ et $r(B)^{2^{k-1}}$ par simple élévation au carré.

On note cependant que cette méthode nécessite le calcul explicite de $r(A)$, lui-même impliquant la résolution de N systèmes linéaires. De plus, les matrices $r(A)$ élevées à une puissance de 2 sont souvent pleines même si A et B sont creuses, ce qui présente le même type d'inconvénient de place mémoire que lors d'une tentative de résolution par une méthode directe.

L'article de Penzl [Pe00b] généralise la méthode de Smith pour une équation de Lyapounov en considérant une suite périodique de paramètres, ce qui permet d'accélérer la convergence pour un choix judicieux de la périodicité, on rencontre ce type de méthode sous le nom de *méthode de Smith cyclique*. On mentionne également l'article récent [BeLiTr08] qui généralise les travaux de [Pe00b] au cas d'une équation de Sylvester et fait le point sur les différentes généralisations de la méthode ADI présentées ici.

Les travaux de Benner et de ses collaborateurs [BeQu02] et [BeQu06] étudient de plus l'utilisation de la méthode ADI et ses généralisations dans un contexte de réduction de système en théorie du contrôle linéaire.

La thèse récente de Sabino [Sa06] fait l'inventaire de l'ensemble des algorithmes évoqués dans cette section, les compare à travers de nombreux exemples numériques et propose une méthode de résolution d'une équation de Lyapounov basée sur une méthode de Smith par blocs.

On cite enfin la méthode ADI *généralisée*, notée dans la littérature GADI, où l'on n'impose plus l'alternance stricte entre la résolution des deux équations de (1.5).

Cette méthode a été étudiée dans [LeRe93] et donne lieu à un problème d'approximation rationnelle dans $\mathcal{R}_{n,m}$, où les degrés du numérateur et du dénominateur des fractions rationnelles mises en jeu ne sont plus supposés égaux.

On pourrait d'ailleurs dans ce qui suit généraliser l'étude effectuée ici pour déterminer l'asymptotique faible du *problème de Zolotarev généralisé* défini en 1.2.2 ce qui permettrait sans doute d'améliorer l'asymptotique obtenue dans [LeRe93] donnée au chapitre suivant à la proposition 3.1.3.

1.3 Problème de Zolotarev et algèbre linéaire

Le problème de Zolotarev pour des ensembles discrets a été défini en 1.2.2, et apparaît dans la majoration de l'erreur après n itérations de la méhode ADI pour la résolution approchée d'une équation de Sylvester. On explicite dans cette section d'autres problèmes d'algèbre linéaire qui font apparaître ce problème d'approximation rationnelle dans le plan complexe.

1.3.1 Valeurs singulières des matrices de petit rang de déplacement

Le lien exposé ci-dessous entre le problème de Zolotarev pour un ensemble discret et la décroissance des valeurs singulières de matrices de petit rang de déplacement est important dans les applications, mais nous sera également utile à plusieurs reprises dans l'étude théorique que nous mènerons par la suite. En particulier, nous utiliserons ce lien pour la preuve d'un cas de dégénérescence de la solution du problème de Zolotarev pour

des ensembles discrets, mais également pour établir une borne inférieure dans l'étude asymptotique de ce problème.

Définition 1.3.1. *Pour A, B, deux matrices comme en* (1.1)*, le* rang de déplacement *par rapport au couple (A, B) de la matrice X est l'entier $\rho_{(A,B)}$ défini par*

$$\rho_{(A,B)}(X) := \mathrm{rg}(AX - XB).$$

Pour certains choix simples de (A, B), par exemple des matrices diagonales, des matrices structurées comme les matrices de Vandermonde, de Krylov, de Cauchy, de Pick, de Loewner, de Hankel, et d'autres encore, le rang de déplacement par rapport à (A, B) est égal à 1 ou 2, voir [Be04].

Avec les notations classiques $\sigma_1(X) \geqslant \sigma_2(X) \geqslant \sigma_3(X) \geqslant ...$ pour les valeurs singulières de la matrice X, on donne dans [Be04] des éléments de preuve du résultat suivant : pour tous les entiers $n \geqslant 1$ tels que $1 + n\rho < \min(M, N)$, on a

$$\sigma_{1+n\rho}(X) \leqslant Z_n(\Lambda(A), \Lambda(B))\, \sigma_1(X), \quad \rho = \mathrm{rg}(AX - XB). \tag{1.9}$$

Pour prouver cette inégalité reprise dans le cas symétrique dans [Pe00a, Théorème 1], on reprend le fait classique suivant dont on trouvera la preuve dans [GoVLo96, Théorème 2.5.3], où l'on note $\|X\|$ la norme spectrale d'une matrice de taille $n \times n$.

Théorème 1.3.2. *Soit $X \in M_N(\mathbb{R})$ une matrice non nulle, on a*

$$\inf_{\tilde{X} \in M_N(\mathbb{R}),\ \mathrm{rg}(\tilde{X}) \leqslant n} \frac{\|X - \tilde{X}\|}{\|X\|} = \frac{\sigma_{n+1}(X)}{\sigma_1(X)}.$$

Voici maintenant la démonstration de l'inégalité 1.9.

PREUVE : Le théorème 1.3.2 cité ci-dessus ainsi que le corollaire 1.2.3 nous suggèrent de nous intéresser au rang de la matrice obtenue après n itérations de la méthode ADI notée X_n, on note ainsi $C := AX - XB$.

On a alors pour tout $n \geqslant 0$ d'après 1.6 l'inégalité

$$\mathrm{rg}\,(X_{n+1}) \leqslant \mathrm{rg}\,(X_n) + \mathrm{rg}(C),$$

d'où en initialisant à $X_0 = 0$,

$$\mathrm{rg}\,(X_n) \leqslant n\,\mathrm{rg}(C).$$

On rappelle alors que X étant solution de l'équation de Sylvester $AX - XB = C$, on a par définition $\mathrm{rg}(C) = \rho$, d'où d'après 1.3.2 et 1.2.3,

$$\begin{aligned} \frac{\sigma_{1+n\rho}(X)}{\sigma_1(X)} &\leqslant \frac{\|X - X_n\|}{\|X\|}, \\ &\leqslant Z_n(\Lambda(A), \Lambda(B)). \end{aligned}$$

□

Ainsi, nos résultats asymptotiques concernant le problème de Zolotarev sur des ensembles discrets permettent de quantifier le taux de décroissance des valeurs singulières de X.

Remarque. Avec les notations précédentes, dans le cas d'une matrice de rang de déplacement par rapport à A et B égal à 1, on obtient le résultat suivant :

$$\frac{1}{\|X\|\|X^{-1}\|} \leqslant Z_n(\Lambda(A), \Lambda(B)), \tag{1.10}$$

car

$$\sigma_1(X) = \|X\| \text{ et } \sigma_N(X) = \frac{1}{\|X^{-1}\|}.$$

Nous exploiterons ce résultat à plusieurs reprises dans ce travail, en particulier dans la preuve de l'inégalité 5.4 au chapitre 5.

1.3.2 Approximation du signe

Définition 1.3.3. *Pour un ensemble discret donné E, les fonctions* signe *et* Heaviside *correspondantes sont données par*

$$signe(z) = 2H(z) - 1, \text{ avec } H(z) = \begin{cases} 1 & si\ z \in E, \\ 0 & sinon. \end{cases}$$

Dans tout ce qui suit, on note indifféremment (a, b) ou $]a, b[$ un intervalle ouvert.

Explicitons le lien entre le problème de Zolotarev et l'approximation des fonctions définies en 1.3.3. On note

$$E = \Lambda(A) \cap (0, +\infty), \quad F = \Lambda(A) \setminus E,$$

les parties positives et négatives du spectre d'une matrice hermitienne inversible A.

L'expression $H(A)$ s'interprète alors comme la projection spectrale sur la somme de sous-espaces propres correspondant aux valeurs propres positives de A, alors que signe(A) donne la fonction signe classique.

L'approximation de la fonction signe(A) par des fractions rationnelles est une tâche importante en chromodynamique quantique, voir par exemple [EFLSV02].

D'autres applications sont dévelopées dans le livre récent de N. Higham, voir en particulier [Hi08, Chapitre 5], et pour une étude du lien entre le problème de Zolotarev et l'approximation de la fonction signe dans le cas classique d'ensembles continus, voir [BaTh00] et [IsTh95].

La preuve de la proposition suivante nécessite des résultats concernant le problème de Zolotarev que l'on prouvera ultérieurement et qu'on se contente d'admettre pour le moment.

Proposition 1.3.4. *Soient $n \geqslant 1$, $E \subset (0, +\infty)$ et $F \subset (-\infty, 0)$ deux ensembles discrets de cardinal strictement supérieur à n.*

Pour une fraction rationnelle R de $\mathcal{R}_n$ à coefficients réels, on note

$$S_R := \max_{x \in E \cup F} |R(x) - signe(x)|,$$

et

$$S_n(E, F) := \inf \{S_R,\ R \in \mathcal{R}_n,\ \text{à coefficients réels}\}$$

qui donne la valeur optimale de l'approximation du signe sur $E \cup F$ en norme L^∞ pour des fractions rationnelles à coefficients réels.

Alors, on a l'égalité

$$S_n(E,F) := \frac{\sqrt{Z_n(E,F)}}{1+Z_n(E,F)}.$$

PREUVE : On prouvera en 3.2.1 qu'il existe une fraction rationnelle extrémale pour le problème de Zolotarev discret dans ce cas, mais également en 3.2.4 que les zéros de celle-ci sont dans l'enveloppe convexe de E, et ses pôles dans l'enveloppe convexe de F, et enfin en 3.2.12 que dans ce cas $Z_n(E,F) < 1$.

On se contentera donc dans la suite de cette preuve de considérer des fractions rationnelles à coefficients réels, on note enfin

$$Z^\star := Z_n(E,F) \text{ et } S^\star := S_n(E,F).$$

Soit $r \in \mathcal{R}_n$. Quitte à multiplier r par une constante, ce qui ne change pas la valeur de $\|r\|_{L^\infty(E)}\|r^{-1}\|_{L^\infty(F)}$, on suppose que $\|r\|_{L^\infty(E)} = \|r^{-1}\|_{L^\infty(F)} =: M_r$, et comme $Z^\star < 1$, on peut supposer $0 < M_r < 1$.

Par décroissance de $x \mapsto \frac{1-x}{1+x}$ sur $(-1,+\infty)$, on a pour $z \in E$

$$\frac{1-M_r}{1+M_r} \leqslant \frac{1-r(z)}{1+r(z)} \leqslant \frac{1+M_r}{1-M_r},$$

on définit alors la fraction rationnelle de $\mathcal{R}_n$

$$R := \frac{1-M_r^2}{1+M_r^2}\frac{1-r}{1+r},$$

et comme

$$\frac{1}{2}\left(\frac{1-M_r}{1+M_r} + \frac{1+M_r}{1-M_r}\right) = \frac{1+M_r^2}{1-M_r^2},$$

on obtient l'inégalité pour $z \in E$

$$\frac{(1-M_r)^2}{1+M_r^2} \leqslant R(z) \leqslant \frac{(1+M_r)^2}{1+M_r^2},$$

ce qui donne, toujours pour $z \in E$,

$$\frac{-2M_r}{1+M_r^2} \leqslant R(z) - 1 \leqslant \frac{2M_r}{1+M_r^2}.$$

En raisonnant de même pour $z \in F$, on prouve l'inégalité

$$-\frac{1+M_r}{1-M_r} \leqslant \frac{1-r(z)}{1+r(z)} \leqslant -\frac{1-M_r}{1+M_r},$$

ce qui donne bien

$$\frac{-2M_r}{1+M_r^2} \leqslant R(z)+1 \leqslant \frac{2M_r}{1+M_r^2},\ z \in F.$$

On en déduit finalement

$$S_R = \frac{2M_r}{1+M_r^2}.$$

Ainsi, pour tout $M > \sqrt{Z^\star}$, il existe une fraction rationnelle R de $\mathcal{R}_n$ à coefficients réels telle que

$$S_R = \frac{2M}{1+M^2},$$

et par croissance sur $(0,1)$ de $x \mapsto \frac{2x}{1+x^2}$, on a

$$S^\star \leqslant \frac{2\sqrt{Z^\star}}{1+Z^\star}.$$

Notons que d'après l'inégalité précédente, $S^\star < 1$.

On fait maintenant le raisonnement inverse : étant donnée une fraction rationnelle R de $\mathcal{R}_n$ à coefficients réels, on note S_R comme précédemment, et on suppose d'après la remarque précédente sans perte de généralité que $S_R \in (0,1)$.

Maintenant, en définissant

$$M := \frac{1-\sqrt{1-S^2}}{S} \text{ et } r := \frac{1-\frac{1-M^2}{1+M^2}R}{1+\frac{1-M^2}{1+M^2}R} \in \mathcal{R}_n,$$

on a $M \in (0,1)$, $S = \frac{2M}{1+M^2}$ et

$$\|r\|_{L^\infty(E)} = \|r^{-1}\|_{L^\infty(F)} = M,$$

ce qui donne l'inégalité

$$S^\star \geqslant \frac{2\sqrt{Z^\star}}{1+Z^\star},$$

ce qui permet de conclure. □

La preuve précédente possède l'avantage d'être constructive, ce qui signifie qu'étant donnée une fraction rationnelle quasi-optimale pour le problème de Zolotarev sur $Z_n(E,F)$, on sait construire une fraction rationnelle quasi-optimale pour le problème de l'approximation du signe, cette constatation nous sera utile dans le chapitre 8 pour le calcul numérique de la quantité $Z_n(E,F)$ pour certains choix de n et des ensembles E, F.

On relève enfin une observation intéressante faite dans [EFLSV02, Fin de section 6] : plutôt que de travailler avec les enveloppes convexes dans le plan complexe des ensembles discrets considérés E et F, il semble utile de prendre en compte auparavant les valeurs propres proches de zéro. Nous formaliserons cette observation par la suite dans le chapitre 8.

Chapitre 2

Équation de Sylvester et réalisation d'un système dynamique

2.1 Introduction

On reprend dans ce chapitre un certain nombre de faits classiques dans le but d'exposer l'intérêt de notre travail concernant la résolution approchée d'équations de Sylvester dans le cadre de la réalisation partielle d'un système dynamique continu en théorie du contrôle linéaire.

Pour présenter ces notions classiques, on s'inspire dans un premier temps de la thèse de Laurent Baratchart [Ba87], dont le premier chapitre donne une introduction à la théorie des systèmes dynamiques linéaires stationnaires dans un cadre continu et dans un cadre discret, ainsi que de [Ni01b, Chapitre 5] qui fournit une introduction au contrôle H^∞ dans un contexte plus théorique.

On se limitera par ailleurs dans ce chapitre à la considération de systèmes dynamiques continus, une théorie parallèle pouvant être développée dans un cadre discret. La théorie discrète donne lieu à une classe d'équations du type $X - AXB = C$ nommées équations de Sylvester en temps discret ou encore équations de Stein. Ce type d'équation, bien que parfois formellement équivalent à une équation de Sylvester, nécessite le développement de méthodes de résolution approchée qui leurs sont propres, voir par exemple [CaLeRe97] qui reprend des techniques de théorie du potentiel logarithmique semblables à celles que nous développerons ultérieurement dans ce but.

Une fois terminée la présentation des concepts mis en jeu dans le cadre continu, on s'inspirera de [Sa06, Chapitre 1], [AnSo02, Sections 1 et 2] et [An98] pour décrire ce que peuvent apporter des techniques de résolution approchée d'une équation de Sylvester dans le cadre de la réduction de modèle pour la théorie du contrôle linéaire à travers la notion de réalisation équilibrée d'un système dynamique continu.

L'objectif de ce chapitre étant de nature descriptive, on ne prétend pas présenter ici une introduction exhaustive à la théorie du contrôle linéaire, et on se contentera de renvoyer le lecteur aux ouvrages classiques de ce domaine pour une présentation de ce type, par exemple [PoWi98] pour l'introduction à la notion de système dans un contexte très général, [An05] pour une présentation centrée sur l'approximation de systèmes dynamiques en grande dimension et [Ni01b, Partie D] pour un point de vue issu de la théorie des opérateurs.

2.2 Système dynamique continu

Définissons pour commencer ce qu'on entend par *système dynamique continu, linéaire et stationnaire.*

Définition 2.2.1. *Un* système dynamique *est la donnée d'une application* $\Sigma : \mathcal{F}_0^m \to \mathcal{F}_0^p$ *où* $\mathcal{F}_0^\ell$ *désigne l'ensemble des fonctions à support contenu dans* $[0, +\infty)$ *et à valeurs dans* $\mathbb{R}^\ell$.

On définit sur chaque $\mathcal{F}^\ell$ *l'opérateur de translation* Δ_r *défini par*

$$\Delta_r(f)(x) := f(x - r), \ \ f \in \mathcal{F}^\ell. \tag{2.1}$$

Un système Σ *est dit* linéaire *si l'application* Σ *est linéaire, et* stationnaire *si* Σ *commute avec chaque opérateur de translation* Δ_r *pour* $r \in \mathbb{R}$.

On note $\mathcal{D}^m$ *l'ensemble des fonctions* $\mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^m)$ *à support compact et pour* $r \in \mathbb{R}$, $\mathcal{D}_r^m$ *le sous-ensemble de* $\mathcal{D}^m$ *des fonctions dont le support est contenu dans* $[r, +\infty)$. *On munit* $\mathcal{D}^m$ *de la topologie de la convergence* L^1 *et* $\mathcal{F}^p$ *de la topologie de la convergence ponctuelle.*

Autrement dit, pour $u := (u_\ell)_{1 \leqslant \ell \leqslant m} \in \mathcal{D}^m$ *une entrée,*

$$\|u\|_{L^1(\mathbb{R}, \mathbb{R}^m)} := \sqrt{\sum_{\ell=1}^m \|u_\ell\|^2_{L^1(\mathbb{R})}}, \tag{2.2}$$

et pour $y = (y_\ell)_{1 \leqslant \ell \leqslant p} \in \mathcal{F}^p$ *une sortie,*

$$\|y\| := \sqrt{\sum_{\ell=1}^p \|y_\ell\|^2_{L^\infty(\mathbb{R})}}. \tag{2.3}$$

Enfin, on appelle système dynamique continu *la donnée d'une application continue au sens des topologies données ci-dessus,*

$$\Sigma \ : \ \mathcal{D}_0^m \to \mathcal{F}_0^p.$$

Le choix de la norme 2 sur les espaces $\mathbb{R}^\ell$ n'a d'autre justification qu'une plus grande commodité dans les calculs qui suivront.

Tout système Σ considéré dorénavant sera supposé continu, linéaire et stationnaire en l'absence de mention contraire explicite.

Proposition 2.2.2. *L'application* Σ *admet un prolongement sur* $\mathcal{D}^m$ *à valeurs dans* $\mathcal{F}^p$ *donné naturellement par l'équation de commutation aux translations* (2.1).

La restriction de Σ *à chaque espace* $\mathcal{D}_r^m$ *est continue, mais ce prolongement n'assure pas nécessairement la continuité de* Σ *sur* $\mathcal{D}^m$ *tout entier.*

La proposition suivante permet de décrire dans le même esprit un moyen naturel de prolonger la définition de Σ aux fonctions L^1 à support compact et donne ensuite une représentation par convolution de l'évaluation en un instant t de la sortie $\Sigma(u)$ pour une entrée u.

Définition 2.2.3. *Soit* $i, j \geqslant 1$. *On dit qu'une fonction* f *définie sur* $\mathbb{R}$ *à valeurs dans* $\mathbb{R}$ *est dans* L^∞_{loc} *si elle est bornée sur tout compact de* $\mathbb{R}$.

Proposition 2.2.4. *L'application Σ admet un prolongement à l'ensemble $L^1_c(\mathbb{R})$ des fonctions L^1 à support compact. En notant $y = \Sigma(u)$ pour $u \in L^1_c(\mathbb{R})$, il existe des fonctions h_{ij}, $1 \leqslant i \leqslant p$, $1 \leqslant j \leqslant m$ dans L^∞_{loc} telles que si $h := [h_{ij}]_{1\leqslant i\leqslant m, 1\leqslant j\leqslant p}$, on ait*

$$y(t) = \int_{-\infty}^{t} h(t-\tau)u(\tau)\,\mathrm{d}\tau \ \in \mathbb{R}^p. \tag{2.4}$$

PREUVE : Supposons $p = m = 1$.

Soient $N \geqslant 1$ et $u \in L^1([-N, N])$.

L'ensemble des fonctions $\mathcal{C}^\infty([-N, N])$ étant dense dans $L^1([-N, N])$, il existe pour tout $\varepsilon > 0$ une fonction $u_\varepsilon \in \mathcal{C}^\infty([-N, N])$ telle que $\|u - u_\varepsilon\|_{L^1} \leqslant \varepsilon$. Comme $\Sigma : \mathcal{D}_{-N} \to \mathcal{F}$ est continue où $\mathcal{F}$ est muni de la topologie de la convergence ponctuelle et $\mathcal{D}_{-N}$ de la topologie de la convergence en norme L^1, l'application

$$u \mapsto \Sigma(u)(t)$$

est continue pour tout $t \in [-N, N]$, ce qui permet bien de prolonger Σ sur $L^1([-N, N])$ et donc sur L^1_c.

On définit maintenant sur $L^1([-N, N])$

$$\phi_N(u) := \Sigma(u)(0).$$

La fonction ϕ est une forme linéaire continue d'après ce qui précède, et il existe donc une fonction $g_N \in L^\infty([-N, N])$ qui représente ϕ_N :

$$\phi_N(u) = \int_{-N}^{N} g_N(\tau)u(\tau)\,\mathrm{d}\tau.$$

La quantité N étant choisie arbitrairement grande, on définit donc une fonction g localement L^∞ telle que pour toute fonction u dans L^1_c on ait

$$\Sigma(u)(0) = \int_{\mathbb{R}} g(\tau)u(\tau)\,\mathrm{d}\tau.$$

En définissant $v := \Delta_{-t}(u)$ avec les notation précédentes, on a par hypothèse de stationnarité

$$\Sigma(u) = \Delta_t\left(\Sigma(v)\right), \text{ d'où } \Sigma(u)(t) = \Sigma(v)(0),$$

et on obtient finalement

$$\Sigma(u)(t) = \int_{\mathbb{R}} g(\tau)u(t+\tau)\,\mathrm{d}\tau.$$

Enfin, le support de g est contenu dans $(-\infty, 0]$ par définition même d'un système dynamique, et en posant $h(t) := g(-t)$, le support de h est contenu dans le demi-axe réel positif, et on obtient par changement de variable

$$y(t) = \int_{-\infty}^{t} h(t-\tau)u(\tau)\,\mathrm{d}\tau,$$

ce qui termine la preuve.

Pour $m, p \geqslant 1$, on raisonne de même en considérant des entrées u dont toutes les composantes sont nulles sauf la i-ème. Si on observe la j-ième composante de la sortie y_j,

on peut appliquer la preuve ci-dessus au système scalaire $u_i \mapsto y_j$, ce qui donne l'existence d'une fonction h_{ij} dans L^∞_{loc} telle que

$$y_j(t) = \int_{-\infty}^{t} h_{ij}(t-\tau)u_i(\tau)\,\mathrm{d}\tau.$$

Alors, la matrice $[h_{ij}]_{1\leqslant i\leqslant m, 1\leqslant j\leqslant p}$ est une matrice de fonctions localement L^∞, et on montre (2.4) par linéarité. □

Définition 2.2.5. *La fonction h obtenue à la proposition 2.2.4 est appelée* réponse impulsionnelle *du système Σ.*

L'appellation *réponse impulsionnelle* peut sembler étrange car aucune entrée admissible pour le système Σ ne donnera lieu à la sortie h dans le cadre théorique que nous avons choisi qui n'inclut pas le formalisme de la théorie des distributions. De fait, la réponse impulsionnelle correspond formellement à la réponse que le système produirait s'il était soumis à une entrée impulsionnelle dont toutes les composantes sont données par une distribution de Dirac à l'origine. On peut cependant montrer qu'il existe une suite d'entrées admissibles telle que la suite des sorties associée converge presque partout vers la réponse impulsionnelle, voir [Ba87].

On s'intéresse maintenant à la notion de stabilité d'un système dynamique stationnaire linéaire continu. Il existe plusieurs définitions de cette notion dans la littérature, la notion suivante dite *stabilité BIBO* (pour *bounded input bounded output*) semble être la plus adaptée à notre contexte. Le lien entre la notion de stabilité d'un système dynamique stationnaire linéaire continu et la notion de stabilité d'une matrice sera donné ultérieurement à la proposition 2.3.7.

Définition 2.2.6. *Le système Σ est dit* stable *si sa réponse impulsionnelle est L^1.*

On supposera dorénavant le système Σ stable.

Proposition 2.2.7. *Pour tout $\alpha \geqslant 1$, on peut étendre la définition de Σ en tant qu'application continue de $L^\alpha(\mathbb{R}_+, \mathbb{R}^m) \to L^\alpha(\mathbb{R}_+, \mathbb{R}^p)$ où les espaces L^α sont munis de leur topologie naturelle.*

PREUVE : Supposons $p = m = 1$, soit $u \in L^\alpha_{loc}(\mathbb{R}_+, \mathbb{R}^m)$, notons $f \star g$ le produit de convolution usuel de deux fonctions f et g. Alors, d'après [Ru75, Chapitre 7 Exercice 4], $h \star u$ est défini presque partout et appartient à $L^\alpha(\mathbb{R}_+, \mathbb{R})$ et

$$\|h \star u\|_{L^\alpha(\mathbb{R}_+)} \leqslant \|h\|_{L^1(\mathbb{R}_+)}\|u\|_{L^\alpha(\mathbb{R}_+)}, \tag{2.5}$$

ce qui prouve le résultat d'après 2.2.4.

Si $m, p \geqslant 1$, on définit la norme suivante induite par la norme euclidienne,

$$\|h\|_{L^\alpha(\mathbb{R}_+, \mathbb{R}^{m\times p})} := \left(\int_0^\infty \sigma_{\max}\left(h(t)\right)^\alpha \mathrm{d}t\right)^{\frac{1}{\alpha}}, \tag{2.6}$$

où $\sigma_{\max}(h(t))$ désigne la plus grande valeur singulière de la matrice $h(t)$. Cela donne l'inégalité

$$\|h \star u\|_{L^\alpha(\mathbb{R}^p_+)} \leqslant \|h\|\|u\|_{L^\alpha(\mathbb{R}^m_+)}, \tag{2.7}$$

qui prouve le résultat d'après (2.5). □

On remarque alors que l'équation de convolution entrée-sortie (2.4) permet de généraliser considérablement la classe des entrées admissibles : les fonctions localement L^α pour $\alpha \geqslant 1$ dont le support est minoré forment dorénavant une classe d'entrées admissibles, et l'application Σ est alors continue pour la topologie induite par ces normes.

Pour ce qui suit, nous avons également besoin d'étendre la définition du système Σ à $L^2(\mathbb{R}, \mathbb{R}^m)$ tout entier, en définissant l'image $y = \Sigma(u) \in L^2(\mathbb{R}_+, \mathbb{R}^m)$ par convolution et projection sur $L^2(\mathbb{R}_+, \mathbb{R}^m)$, c'est-à-dire, par la formule $y(t) = 0$ si $t < 0$, et sinon

$$y(t) = \int_{-\infty}^{t} h(t-\tau)u(\tau)\, d\tau,$$

ce qui permet en particulier de définir Σ pour des entrées dites *du passé* dans $L^2(\mathbb{R}_-, \mathbb{R}^p)$.

On définit également S_Σ l'opérateur de convolution qui à un signal dans $L^2(\mathbb{R}, \mathbb{R}^m)$ associe $S_\Sigma(u) := h \star u$. Le lemme suivant donne une inégalité entre les normes d'opérateurs associées à Σ et S_Σ dont la démonstration résulte directement des définitions ci-dessus.

Lemme 2.2.8. *Notons*

$$\|S_\Sigma\|_{\mathcal{L}(L^2(\mathbb{R},\mathbb{R}^m),L^2(\mathbb{R},\mathbb{R}^p))} := \sup_{u \in L^2(\mathbb{R},\mathbb{R}^m), u \neq 0} \frac{\|h \star u\|}{\|u\|}$$

et

$$\|\Sigma\|_{\mathcal{L}(L^2(\mathbb{R},\mathbb{R}^m),L^2(\mathbb{R},\mathbb{R}^p))} := \sup_{u \in L^2(\mathbb{R},\mathbb{R}^m), u \neq 0} \frac{\|\Sigma(u)\|}{\|u\|}.$$

On a alors l'inégalité

$$\|\Sigma\|_{\mathcal{L}(L^2(\mathbb{R},\mathbb{R}^m),L^2(\mathbb{R},\mathbb{R}^p))} \leqslant \|S_\Sigma\|_{\mathcal{L}(L^2(\mathbb{R},\mathbb{R}^m),L^2(\mathbb{R},\mathbb{R}^p))}.$$

Cette extension sera essentielle par la suite pour la définition de la norme d'un système dynamique, et a motivé notre choix de définition de stabilité d'un système.

2.3 Fonction de transfert

La classe des entrées admissibles ayant été généralisée aux fonctions L^2 dont le support est minoré à la section précédente nous sommes à même d'accomplir le premier pas vers la représentation fréquentielle d'un système dynamique continu stationnaire linéaire grâce à la définition de la fonction de transfert d'un tel système. L'introduction à cette représentation nécessite quelques rappels concernant les espaces de Hardy du demi-plan droit, rappels que l'on donne sous la forme la plus concise possible ci-dessous, on renvoie le lecteur vers [Ru75, Chapitre 17] ainsi que [Ni01a, Chapitre 3] pour une présentation détaillée de ces espaces très utilisés en analyse complexe. Pour une présentation adaptée à notre étude des espaces de Hardy d'un demi-plan, étude qui se déduit essentiellement par transformation conforme depuis l'étude réalisée pour le disque unité, voir en particulier [Ni01a, Chapitre 6].

Définition 2.3.1. *Notons $\mathbb{C}_+$ le demi-plan ouvert droit des nombres complexes de partie réelle strictement positive.*

Pour $1 \leqslant \alpha \leqslant \infty$, on définit l'espace de Hardy $H^\alpha(\mathbb{C}_+)$ comme l'ensemble des fonctions holomorphes sur $\mathbb{C}_+$ dont la norme L^α est uniformément majorée sur les droites verticales de $\mathbb{C}_+$.

On recense quelques propriétés classiques des espaces de Hardy du demi-plan dans le théorème suivant. On a transposé dans le demi-plan droit l'énoncé relatif au disque donné dans [Ru75, Théorème 17.7].

Théorème 2.3.2. *Soit f dans $H^\alpha(\mathbb{C}_+)$.*

Alors, f admet presque partout sur l'axe imaginaire une limite que l'on note également f :

$$f(\imath y) := \lim_{x\to 0,\ x>0} f(x+\imath y).$$

La fonction f ainsi étendue est dans $L^\alpha(\imath\mathbb{R})$, on a

$$\lim_{x\to 0,\ x>0} \int_{\imath\mathbb{R}} |f(\imath y) - f(x+\imath y)|^\alpha \, \mathrm{d}y = 0,$$

et on munit $H^\alpha(\mathbb{C}_+)$ d'une norme par

$$\|f\|_{H^\alpha(\mathbb{C}_+)} = \frac{1}{2\pi}\|f\|_{L^\alpha(\imath\mathbb{R})}.$$

Dans le cas d'une fonction à valeurs dans $\mathbb{C}^{p\times m}$, la définition de $H^\alpha\left(\mathbb{C}_+, \mathbb{C}^{p\times m}\right)$ se généralise pour le choix d'une norme quelconque sur $\mathbb{C}^{p\times m}$, on choisira dans ce qui suit pour f dans $H^\alpha\left(C_+, \mathbb{C}^{p\times m}\right)$ la norme

$$\|f\|_{H^\alpha(\mathbb{C}_+, \mathbb{C}^{m\times p})} := \left(\frac{1}{2\pi}\int_0^\infty \sigma_{\max}\left(f(t)\right)^\alpha \, \mathrm{d}t\right)^{\frac{1}{\alpha}}$$

Définition 2.3.3. *Notons ainsi la transformée de Laplace de la réponse impulsionnelle :*

$$s \mapsto \mathcal{H}(s) := \int_{\mathbb{R}} \mathrm{e}^{-st}\, h(t)\mathrm{d}t.$$

On appelle cette matrice la fonction de transfert *du système Σ.*

Dans le cas ou la fonction de transfert est une matrice de fractions rationnelles, on dit que le système Σ est rationnel.

La proposition et le corollaire suivant sont adaptés de [Ni01b, Théorème 5.1.5] et reposent sur un théorème de Payley-Wiener qui décrit l'image par transformation de Laplace des fonctions de $L^2(\mathbb{R}_+)$ en terme d'espaces de Hardy ainsi que sur le théorème de Plancherel [Ru75, Théorème 9.13]. On utilise également le fait les espaces $L^2\left(\mathbb{R}, \mathbb{R}^m\right)$ et $L^2\left(i\mathbb{R}, \mathbb{R}^p\right)$ admettent les décompositions orthogonales

$$L^2\left(\mathbb{R}, \mathbb{R}^m\right) = L^2\left(\mathbb{R}_-, \mathbb{R}^m\right) \oplus_\perp L^2\left(\mathbb{R}_+, \mathbb{R}^m\right)$$

et

$$L^2\left(i\mathbb{R}, \mathbb{C}^p\right) = H^2\left(\mathbb{C}_-, \mathbb{C}^p\right) \oplus_\perp H^2\left(\mathbb{C}_+, \mathbb{C}^p\right)$$

qui sont préservées par la transformation de Laplace.

Proposition 2.3.4. *Soit $y : \mathbb{R}_+ \to \mathbb{C}^p$, notons $\mathcal{Y}$ sa transformée de Laplace*

$$\mathcal{Y}(s) := \int_{\mathbb{R}} \mathrm{e}^{-st}\, y(t)\mathrm{d}t.$$

Alors, $y \in L^2(\mathbb{R}_+, \mathbb{C}^p)$ si et seulement si $\mathcal{Y} \in H^2(\mathbb{C}_+, \mathbb{C}^p)$ et dans ce cas,

$$\|y\|^2_{L^2(\mathbb{R}_+,\mathbb{C}^p)} = \sum_{\ell=1}^{p} \|y_\ell\|^2_{L^2(\mathbb{R}_+)} = \|\mathcal{Y}\|^2_{H^2(\mathbb{C}_+,\mathbb{C}^p)} = \sum_{\ell=1}^{p} \frac{1}{2\pi} \int_{\imath\mathbb{R}} |\mathcal{Y}_\ell|^2 (s) \, |ds|.$$

De même, si $y : \mathbb{R} \to \mathbb{C}^p$, $y \in L^2(\mathbb{R}, \mathbb{C}^p)$ si et seulement si $\mathcal{Y} \in L^2(i\mathbb{R}, \mathbb{C}^p)$ et dans ce cas,

$$\|y\|^2_{L^2(\mathbb{R},\mathbb{C}^p)} = \sum_{\ell=1}^{p} \|y_\ell\|^2_{L^2(\mathbb{R})} = \|\mathcal{Y}\|^2_{L^2(i\mathbb{R},\mathbb{C}^p)} = \sum_{\ell=1}^{p} \frac{1}{2\pi} \int_{\imath\mathbb{R}} |\mathcal{Y}_\ell|^2 (s) \, |ds|.$$

Corollaire 2.3.5. *Soient $u \in L^2(\mathbb{R}_+, \mathbb{C}^m)$, $y = \Sigma(u) \in L^2(\mathbb{R}_+, \mathbb{C}^p)$. En notant $\mathcal{U}$ et $\mathcal{Y}$ les transformées de Laplace respectives de u et y, on a*

$$\mathcal{Y}(s) = \mathcal{H}(s)\mathcal{U}(s). \tag{2.8}$$

De plus, la fonction $\mathcal{H}$ appartient à l'espace de Hardy $H^\infty(\mathbb{C}_+, \mathbb{C}^{p\times m})$ et pour les normes précédemment définies,

$$\|h\|_{L^1(\mathbb{R}_+,\mathbb{R}^{p\times m})} \geqslant \|S_\Sigma\|_{\mathcal{L}(L^2(\mathbb{R},\mathbb{R}^m),L^2(\mathbb{R},\mathbb{R}^p))} = \sup_{\mathcal{U}\neq 0, \mathcal{U}\in L^2(i\mathbb{R},\mathbb{C}^m)} \frac{\|\mathcal{H}\mathcal{U}\|_{L^2(i\mathbb{R},\mathbb{C}^p)}}{\|\mathcal{U}\|_{L^2(i\mathbb{R},\mathbb{C}^m)}} = \|\mathcal{H}\|_{L^\infty(i\mathbb{R},\mathbb{C}^{p\times m})}.$$

Pour la preuve de la dernière égalité du corollaire précédent, voir par exemple [GrSz58, Chapitre 8].

On remarque que par injectivité de la transformation de Laplace, la donnée de la réponse impulsionnelle h est équivalente à celle de la fonction de transfert $\mathcal{H}$.

On donne alors la proposition suivante qui fait le lien avec une représentation courante d'un système dynamique continu rationnel dans la littérature.

Proposition 2.3.6. *Soit u une fonction de $\mathcal{D}_0^m$, si on définit x et y par $x(0) = 0$,*

$$\begin{aligned} \frac{dx}{dt}(t) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t), \end{aligned} \tag{2.9}$$

on a alors $y = \Sigma(u)$.

Réciproquement, pour $(A, B, C) \in M_n(\mathbb{C}) \times M_{n,m}(\mathbb{C}) \times M_{p,n}(\mathbb{C})$, (2.9) définit un système dynamique continu rationnel Σ dont la réponse impulsionnelle est $Ce^{tA}B$ et la fonction de transfert $\mathcal{H}$ est donnée par $z \mapsto C\left[zI - A\right]^{-1} B$.

Ce type de représentation pour un système dynamique continu est essentiel pour les applications et apparaît en physique, chimie, biologie, économie...

La proposition 2.3.6 nous permet alors de considérer indifféremment la donnée d'un système dynamique continu rationnel ou la donnée d'un triplet de matrices de dimensions adéquates, représentation qui se prêtera mieux à la majeure partie de ce qui suit.

Voici maintenant une caractérisation plus maniable de la stabilité d'un système, caractérisation qui nous sera utile par la suite.

Proposition 2.3.7. *Le système Σ est stable si et seulement sa matrice d'état A est stable.*

Définition 2.3.8. *On note* $A \in M_n(\mathbb{C})$ *la* matrice d'état, $B \in M_{n,m}(\mathbb{C})$ *la* matrice de commande *et* $C \in M_{p,n}(\mathbb{C})$ *la* matrice de sortie.

On appelle x *la* variable d'état, y *la* variable de sortie, *et* u *la* variable d'entrée *du système précédent. De même, on appelle* espace d'états *l'espace dans lequel évolue la variable* x.

Dans les applications, la matrice d'état A est souvent hermitienne.

Typiquement, la modélisation de phénomènes physiques produit des systèmes dynamiques Σ avec une matrice d'état creuse et de grande dimension, avec $m, p << n$.

Ici, *grande dimension* signifie que l'on considère des matrices d'ordre $n \geqslant 10^6$, ce qui exclut toute tentative directe de calcul numérique de la solution du système 2.9, même si l'on dispose de la formule explicite suivante pour l'état x et la sortie y du système Σ en fonction du signal d'entrée u et des matrices A, B, C :

$$x(t) = \exp(tA)x(0) + \int_0^t \exp\left((t-\tau)A\right) Bu(\tau)\,\mathrm{d}\tau,\ t \geqslant 0, \tag{2.10}$$

et

$$y(t) = C\exp(tA)x(0) + \int_0^t C\exp\left((t-\tau)A\right) Bu(\tau)\,\mathrm{d}\tau,\ t \geqslant 0.$$

La simulation directe d'un système dynamique d'une telle dimension s'avère souvent impossible compte tenu des dimensions des matrices mises en jeu, mais il existe une méthode d'approximation d'un système dynamique de grande dimension par un système de bien plus petite dimension, méthode que nous allons décrire maintenant.

Cette méthode consiste à projeter l'état x sur un espace d'états de plus petite dimension dans lequel la sortie $\tilde{y}$ du système dynamique projeté approche en un sens que l'on définira ultérieurement la solution y du système dynamique originel de grande dimension.

2.4 Commandabilité et observabilité

On présente ici les notions de commandabilité et d'observabilité pour un système linéaire.

Définition 2.4.1. *Le système* Σ *est commandable si pour tout* $x_1, x_2 \in \mathbb{R}^n$, *il existe* $T \geqslant 0$ *et* u *une fonction* $\mathcal{C}^\infty([0,T],\mathbb{R})$ *telle que la solution de*

$$\frac{d}{dt}x = Ax + Bu,\ x(0) = x_1$$

vérifie $x(T) = x_2$.

La proposition suivante est due à Kalman et donne une condition nécessaire et suffisante algébrique simple de commandabilité.

Proposition 2.4.2. *On appelle la* matrice de commandabilité *du système* Σ

$$C := [B\ AB \dots A^{n-2}B\ A^{n-1}B].$$

Le système Σ *est commandable si et seulement si sa matrice de commandabilité est de rang* n.

Passons maintenant à la définition de l'observabilité, souvent présentée dans les traités de théorie du contrôle linéaire comme une notion duale de la commandabilité.

Il existe de nombreuses définitions équivalentes de cette notion, on choisit ici de présenter une définition classique dans la littérature adoptée dans [PoWi98].

Définition 2.4.3. *On définit pour le système Σ l'ensemble*

$$\mathcal{B} := \left\{(u,x,y) \in \mathcal{C}^\infty(\mathbb{R},\mathbb{R}^m) \times \mathcal{C}^\infty(\mathbb{R},\mathbb{R}^p) \times \mathcal{C}^\infty(\mathbb{R},\mathbb{R}^n),\ \frac{d}{dt}x = Ax + Bu,\ y = Cx\right\}.$$

Alors, Σ est dit observable *si le fait que deux triplets (u,y,x_1) et (u,y,x_2) soient dans $\mathcal{B}$ entraîne la relation $x_1 = x_2$.*

Autrement dit, le fait de connaître (u,y) et la dynamique du système détermine x.

Proposition 2.4.4. *On appelle la* matrice d'observabilité *du système Σ la matrice*

$$O := \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{pmatrix}.$$

Le système Σ est observable *si et seulement si sa matrice d'observabilité est de rang n.*

Ces hypothèses classiques et importantes en théorie du contrôle linéaire étant définies, passons maintenant à la question de la réalisation minimale d'un système dynamique.

On supposera dorénavant Σ observable et commandable.

2.5 Réalisation d'un système dynamique

Soit $\mathcal{H}$ la fonction de transfert d'un système dynamique continu linéaire stationnaire rationnel, on dit alors que le triplet de matrices associé *réalise* le système Σ ou la fonction de transfert $\mathcal{H}$. Dans ce cas, la dimension de la matrice d'état A est appelée *ordre de réalisation* de Σ. On définit ci-dessous la notion de réalisation minimale d'un système avant d'expliciter le processus de réalisation équilibrée dans la section suivante.

Définition 2.5.1. *Le système Σ est appelé* réalisation minimale *de la fonction*

$$\mathcal{H}(z) = C(zI - A)^{-1}B$$

si la dimension de l'espace d'états est minimale parmi toutes les réalisations possibles de Σ de fonction de transfert $\mathcal{H}$.

On dispose alors du résultat suivant, dont la première partie a été publiée dans le premier *journal of control* en 1965.

Théorème 2.5.2. *Le système Σ est une réalisation minimale de la fonction de transfert $\mathcal{H}$ si et seulement si il est commandable et observable.*

Si on connaît une réalisation minimale Σ associée à un triplet de matrices (A,B,C), alors l'ensemble des réalisations minimales de Σ est le suivant :

$$\left\{(SAS^{-1}, SB, CS^{-1}),\ S \in \mathbb{R}^{n\times n},\ \det(S) \neq 0\right\}.$$

On peut alors interpréter ce résultat de la façon suivante : connaissant une réalisation minimale d'un système Σ, le seul degré de liberté dont on dispose parmi l'ensemble des réalisations minimales de ce système consiste à changer de base dans l'espace d'états. Cette possibilité sera exploitée dans le cadre de la réalisation équilibrée de Σ.

Pour écrire la proposition suivante, on utilise [An98, Section 3.1.6] qui fait le lien entre les objets intervenant dans le cadre des systèmes dynamiques discrets et dans celui des systèmes dynamiques continus.

Cette transformation s'effectue à l'aide du changement de variable

$$z = \frac{1-s}{1+s},$$

pour passer du continu au discret, changement de variable qui envoie conformément le demi-plan droit dans le disque unité.

Soit u une entrée de $L^2(\mathbb{R}_-, \mathbb{R}^m)$ complétée par zéro sur le demi-axe réel positif. Sa transformée de Laplace $\mathcal{U}$ appartient à $H^2(\mathbb{C}_-, \mathbb{C}^m)$, et le produit $\mathcal{HU}$ est dans l'espace $L^2(\imath\mathbb{R}, \mathbb{C}^p)$. La transformation de Laplace de la sortie $y = S_\Sigma(u)$ est la projection orthogonale du produit $\mathcal{HU}$ sur l'espace de Hardy du demi-plan droit $H^2(\mathbb{C}_+, \mathbb{C}^p)$.

Pour donner quelques arguments de preuve de ce résultat, on utilise le changement de variable ci-dessus pour se ramener au disque unité, et on trouve après calculs

$$\mathcal{U}(s) = \frac{\sqrt{2}}{1+s}\sum_{k=-\infty}^{-1}\mathcal{U}_k\Big(\frac{1-s}{1+s}\Big)^k = \frac{\sqrt{2}}{1-s}\sum_{k=0}^{\infty}\mathcal{U}_{-1-k}\Big(\frac{1+s}{1-s}\Big)^k,$$

$$\mathcal{H}(s) - \mathcal{H}(1) = (1-s)C(sI-A)^{-1}(I-A)^{-1}B = \frac{1-s}{1+s}\widetilde{C}\Big(I - \frac{1-s}{1+s}\widetilde{A}\Big)^{-1}\widetilde{B},$$

avec

$$\widetilde{C} = \sqrt{2}\,C(I-A)^{-1}, \quad \widetilde{B} = \sqrt{2}\,(I-A)^{-1}B, \quad \text{et} \quad \widetilde{A} = (I-A)^{-1}(I+A).$$

Comme A est stable par hypothèse, $\widetilde{A}$ a toutes ses valeurs propres dans le disque unité ouvert, et en notant Π la projection sur $H^2(\mathbb{C}_+, \mathbb{C}^p)$, on a

$$\Pi\Big(\mathcal{HU}\Big)(s) = \Pi\Big(\big(\mathcal{H}-\mathcal{H}(1)\big)\mathcal{U}\Big)(s) = \frac{\sqrt{2}}{1+s}\sum_{j=0}^{\infty}\mathcal{Y}_j\Big(\frac{1-s}{1+s}\Big)^j, \quad \mathcal{Y}_j := \sum_{k=0}^{\infty}\widetilde{C}\widetilde{A}^{k+j}\widetilde{B}.U_{-1-k}$$

Autrement dit, l'application $\mathcal{U} \mapsto \mathcal{Y} = \Pi(\mathcal{HU})$ admet une représentation matricielle, avec une matrice de Hankel par blocs, que l'on peut écrire comme produit $\widetilde{\mathcal{O}}\widetilde{\mathcal{C}}$, avec

$$\widetilde{\mathcal{O}} = \begin{pmatrix} \widetilde{C} \\ \widetilde{C}\widetilde{A} \\ \widetilde{C}\widetilde{A}^2 \\ \vdots \end{pmatrix} \in \mathbb{R}^{\infty\times n}, \quad \widetilde{\mathcal{C}} = (\widetilde{B}, \widetilde{A}\widetilde{B}, \widetilde{A}^2\widetilde{B}, ...) \in \mathbb{R}^{n\times\infty},$$

Définition 2.5.3. *Avec les notations précédentes, la matrice $\widetilde{\mathcal{O}}\widetilde{\mathcal{C}}$ est une matrice de Hankel par blocs et on note $(\sigma_\ell)_\ell$ ses valeurs sigulières non nulles rangées par ordre décroissant.*

Les éléments $(\sigma_\ell)_\ell$ sont les valeurs singulières de Hankel *du système Σ.*

La proposition suivante résulte directement des hypothèses de stabilité, commandabilité et observabilité vérifiées par Σ.

Proposition 2.5.4. *La matrice* $\mathrm{Hank}\,(\mathcal{H})$ *est de rang* n.

Pour la définition des valeurs singulières d'un opérateur de Hankel, voir [Ni01a, Définition 7.1.3].

On présente dans la prochaine section un premier lien entre la résolution approchée d'équations de Sylvester et les notions de théorie du contrôle linéaire présentées jusqu'à maintenant.

2.6 Grammiens

On définit dans cette section les grammiens de commandabilité et d'observabilité du système Σ dont on donne une interprétation physique avant d'étudier ceux-ci en tant que solution d'une équation de Lyapounov.

Définition 2.6.1. *On définit le* grammien de commandabilité *(reachability gramian)* $P \in M_n(\mathbb{C})$ *et le* grammien d'observabilité *(observability gramian)* $Q \in M_n(\mathbb{C})$ *du système* Σ *:*

$$P := \int_0^{+\infty} \mathrm{e}^{tA}\, BB^\star\, \mathrm{e}^{tA^\star}\, \mathrm{d}t \text{ et } Q := \int_0^{+\infty} \mathrm{e}^{tA^\star}\, C^\star C\, \mathrm{e}^{tA}\, \mathrm{d}t.$$

On remarque que P et Q sont bien définis grâce à l'hypothèse de stabilité de la matrice A. Cette définition s'explique par l'observation suivante : la réponse impulsionnelle du système Σ est donnée par

$$h(t) := C\, \mathrm{e}^{tA}\, B.$$

En notant

$$\chi(t) := \mathrm{e}^{tA}\, B \text{ et } \eta(t) := C\, \mathrm{e}^{tA},$$

la fonction χ correspond à l'état du système avec une entrée impulsionnelle, alors que pour une condition initiale $x(0)$, en l'absence de fonction de forçage u, la sortie du système est donnée par $y(t) = \eta(t)x(0)$, c'est pourquoi on appelle dans la littérature χ l'application *input to state* et η l'application *state-to-output*.

On a alors

$$P = \int_0^{+\infty} \chi(t)\chi(t)^T \mathrm{d}t \text{ et } Q = \int_0^{+\infty} \eta(t)^T\eta(t) \mathrm{d}t.$$

Un état $\hat{x}$ est considéré comme difficile à atteindre si conduire un système à travers un signal d'entrée $u(t)$ depuis $x(0) = 0$ jusqu'à l'état $\hat{x}$ nécessite une énergie importante en comparaison à l'énergie nécessaire aux autres états pour accomplir une telle transformation.

L'énergie minimale parmi tous les signaux d'entrée u nécessaire pour x pour approcher $\hat{x}$ pour $t \mapsto +\infty$ est donnée par

$$\mathcal{E}_r := \left(\hat{x}^\star P^{-1} \hat{x}\right)^{1/2},$$

où $P \in M_n(C)$ est le grammien de commandabilité défini ci-dessus.

De plus, un état $\hat{x}$ est considéré comme difficile à observer si l'énergie produite par le système avec état intitial $x(0) = \hat{x}$ et $u = 0$ est petite comparée aux autres états.

L'énergie totale observée relativement à $\hat{x}$ est alors donnée par la quantité

$$\mathcal{E}_0 := \left(\int_0^{+\infty} \|C\,\mathrm{e}^{tA}\,\hat{x}\|^2 \mathrm{d}t \right)^{1/2} = (\hat{x}^\star Q \hat{x})^{1/2}\,.$$

D'après le lemme 1.1.6 P et Q sont solutions des équations de Lyapounov suivantes

$$\begin{aligned} AP + PA^\star &= -BB^\star \\ A^\star Q + QA &= -C^\star C. \end{aligned} \tag{2.11}$$

On remarque enfin que les grammiens P et Q sont des matrices positives d'après leur définition, et que par unicité de la solution des équations de Lyapounov ci-dessus, P et Q sont hermitiennes.

Définition 2.6.2. *Soit X une matrice complexe carrée d'ordre n. On note alors l'*inertie *de la matrice X*

$$\mathrm{in}(X) := (\mathrm{in}_-(X), \mathrm{in}_0(X), \mathrm{in}_+(X))$$

le nombre de valeurs propres de X situées respectivement dans le demi-plan ouvert gauche, l'axe imaginaire et dans le demi-plan ouvert droit du plan complexe.

La proposition suivante donne une relation entre l'inertie d'une matrice M et l'inertie de la solution de l'équation de Lyapounov de coefficient M. On utilise dans cette proposition les hypothèses de commandabilité, observabilité et stabilité vérifiées par Σ.

Proposition 2.6.3. *Les solutions P et Q des équations de Lyapounov 2.11 vérifient* $\mathrm{in}(P) = \mathrm{in}(Q) = \mathrm{in}(-A)$.

Ainsi, P et Q sont définis positifs.

Pour conclure cette section, la proposition suivante fait le lien entre les valeurs singulières de Hankel de Σ et les grammiens d'observabilité et de commandabilité.

Proposition 2.6.4. *Les valeurs singulières non nulles du système Σ coïncident avec les racines des valeurs propres de QP.*

PREUVE : Commençons par écrire nos deux grammiens en terme des variables discrètes définies en 2.5.3 : comme

$$(I - A)P(I - A^T) - (I + A)P(I + A^T) = -2(AP + PA^T),$$

le grammien de commandabilité est solution de l'équation de Stein

$$P - \widetilde{A} P \widetilde{A}^T = \widetilde{B}\widetilde{B}^T.$$

Ainsi, on a le développement

$$P = \sum_{\ell=0}^{\infty} \widetilde{A}^\ell \widetilde{B}\widetilde{B}^T \left(\widetilde{A}^T\right)^\ell = \widetilde{C}\widetilde{C}^T,$$

valable car toutes les valeurs propres de $\widetilde{A}$ sont dans le disque unité ouvert par hypothèse de stabilité, et on montre de même que

$$Q = \widetilde{O}^T \widetilde{O}.$$

Maintenant, PQ est semblable à une matrice définie positive :

$$\widetilde{C}^T \widetilde{O}^T \widetilde{O}\widetilde{C} = \left(\widetilde{O}\widetilde{C}\right)^\star \widetilde{O}\widetilde{C},$$

et PQ admet donc n valeurs propres strictement positives, d'où le résultat d'après la définition 2.5.3. □

2.7 Réalisation équilibrée

D'après le théorème 2.5.2, le seul degré de liberté dont nous disposons parmi l'ensemble des réalisations minimales d'un système Σ est donné par un changement de base dans l'espace d'états. De plus, d'après les définitions des grammiens, quand le système est commandable et observable et que P et Q sont donc définies positives, les états avec une haute énergie de commande $\mathcal{E}_r$ se trouvent dans les sous-espaces propres correspondant aux petites valeurs propres de P, et les états avec une petite énergie d'observabilité $\mathcal{E}_0$ dans les sous-espaces propres correspondant aux petites valeurs propres de Q.

Ainsi, dans une base donnée, certains états peuvent être difficiles à atteindre mais faciles à observer, et vice-versa.

Il existe cependant une base dans laquelle les états difficiles à atteindre sont également difficiles à observer, voir [An05, Section 7.1], on explique dans cette partie comment construire une telle base pour le système Σ.

Définition 2.7.1. *Une réalisation minimale de Σ dans laquelle les grammiens P et Q sont diagonaux et égaux s'appelle réalisation équilibrée de Σ.*

On peut en effet montrer qu'il existe une base dans laquelle les grammiens s'écrivent

$$\hat{P} = \hat{Q} = \Sigma := diag\left(\sigma_1, \ldots, \sigma_n\right),$$

où les $(\sigma_\ell)_\ell$ sont les valeurs singulières de Hankel définies en 2.5.3, les grammiens sont donc dans cette base égaux et diagonaux, on dit dans ce cas que le système est balancé selon les directions principales.

Une fois la réalisation équilibrée du système obtenue, on cherche à réduire la dimension du système afin de pouvoir le simuler numériquement tout en étant en mesure de quantifier la perte d'information liée à cette réduction de dimension appelée *réduction de modèle*.

Les valeurs singulières de l'opérateur de Hankel associé à Σ obtenues par réalisation équilibrée jouent un rôle essentiel dans le choix du seuil de troncature pour la réduction de modèle, et le fait de quantifier leur taux de décroissance constitue ainsi un enjeu essentiel pour le choix de ce seuil.

Pour construire cette base, on décompose le grammien de commandabilité P en utilisant la décomposition de Cholesky $P = UU^\star$, et on écrit la décomposition en valeurs singulières de $U^\star QU$ sous la forme $U^\star QU = V\Gamma^2 V^\star$. Alors, en définissant $T := \Gamma^{1/2} V^\star U^{-1}$, on a

$$TPT^\star = \left(T^\star\right)^{-1} QT^{-1} = \Gamma.$$

En effectuant le changement de base donné par l'équation $\hat{x}(t) := Tx(t)$, on obtient d'après le théorème 2.5.2 le système suivant équivalent au système de départ Σ

$$\begin{aligned} \frac{d\hat{x}}{dt}(t) &= \hat{A}\hat{x}(t) + \hat{B}u(t) \\ y(t) &= \hat{C}\hat{x}(t) + Du(t), \end{aligned} \tag{2.12}$$

avec $\hat{A} := TAT^{-1}$, $\hat{B} := BT$ et $\hat{C} = CT^{-1}$, et on obtient bien une réalisation équilibrée de Σ.

On observe alors que la signal de sortie y est indépendant de notre choix de base dans l'espace des états, mais ce choix affecte A, B, C, et modifie par conséquent les grammiens du système.

Toutes les données nécessaires sont maintenant réunies pour tronquer le système : on partitionne A, B, C et Γ comme suit :

$$\hat{A} = \begin{pmatrix} \hat{A}_{11} & \hat{A}_{12} \\ \hat{A}_{21} & \hat{A}_{22} \end{pmatrix},$$

$$\hat{B} = \begin{pmatrix} \hat{B}_1 \\ \hat{B}_2 \end{pmatrix}, \ \hat{C} = \begin{bmatrix} \hat{C}_1 \hat{C}_2 \end{bmatrix}, \ \Gamma = \begin{pmatrix} \Gamma_1 & 0 \\ 0 & \Gamma_2 \end{pmatrix},$$

où $\Gamma_1 \in M_r(\mathbb{C})$.

On s'intéresse dorénavant au système réduit suivant noté Σ_r

$$\begin{aligned} \frac{d\tilde{x}}{dt}(t) &= \hat{A}_1 \tilde{x}(t) + \hat{B}_1 u(t) \\ \tilde{y}(t) &= \hat{C}_1 \tilde{x}(t) + Du(t) \end{aligned} \tag{2.13}$$

qui a été obtenu à partir de Σ à partir de la réduction de modèle qui vient d'être réalisée.

L'objectif est maintenant de trouver une valeur de r telle que le système réduit Σ_r soit suffisamment proche du système originel Σ pour une certaine norme, mais avec un entier r assez petit pour envisager d'effectuer des calculs numériques à partir du système réduit.

Cette opération de tronquage possède deux propriétés remarquables que l'on reprend ici depuis [An05, p. 212].

Proposition 2.7.2. *La stabilité de la matrice A entraîne la stabilité de $\hat{A}_1$.*

Cette propriété importante étant donnée, on s'intéresse à la distance entre le système Σ et son modèle réduit Σ_r. Cette section permet de formaliser l'idée selon laquelle deux systèmes dont les réponses fréquentielles sont proches sont en un certain sens à définir eux-mêmes proches.

La proposition suivante permet maintenant de quantifier l'erreur obtenue après réduction du système et sa démonstration utilise la définition de la norme de Σ et la majoration données au lemme 2.2.8 ainsi que le résultat du corollaire 2.3.5.

Proposition 2.7.3. *On a alors la borne suivante :*

$$\|\Sigma - \Sigma_r\|_{\mathcal{L}(L^2(\mathbb{R},\mathbb{R}^m),L^2(\mathbb{R},\mathbb{R}^p))} \leqslant 2\left(\sigma_{r+1} + \ldots + \sigma_n\right).$$

Les valeurs de Hankel du système Σ sont ainsi un paramètre essentiel pour la construction d'un modèle réduit pour ce système.

2.8 Grammien croisé

La proposition 2.6.4 permet de ramener le problème de la détermination des valeurs singulières de Σ en un problème de calcul approché des valeurs propres du produit de deux solutions d'équations de Lyapounov.

Définition 2.8.1. *On dit que le système Σ est un* système carré *lorqu'on a $m = p$.*

On définit dans cette section le *grammien croisé*, objet permettant dans le cas d'un système carré de transformer le problème de détermination des valeurs singulières de Hankel du système Σ en un problème de recherche de valeurs propres pour la solution d'une seule équation de Sylvester.

On note que le passage de deux équations de Lyapounov à une seule équation de Sylvester, s'il réduit le nombre d'équations à traiter sans changer la dimension des objets considérés, présente l'inconvénient de perdre la symétrie des équations de Lyapounov.

Citons la proposition suivante depuis [AnSo02, Lemme 2.2].

Proposition 2.8.2. *Supposons que le système Σ est carré.*
Soit X la solution de l'équation de Sylvester

$$AX + XA = -BC.$$

Alors, l'ensemble des valeurs singulières de Hankel du système Σ et l'ensemble des valeurs propres de X coïncident.

L'utilité concrète de la réduction équilibrée dépend ainsi du taux de décroissance des valeurs singulières de Hankel (σ_ℓ).

Pour évaluer celles-ci dans le cas d'un système Σ carré, la proposition 2.8.2 permet de se ramener à un problème de résolution approchée d'une équation de Sylvester.

Ce chapitre a finalement permis de mettre en avant l'interêt de l'étude approfondie des méthodes de résolution approchée d'une équation de Sylvester dans le contexte de la réalisation équilibrée et de la réduction de modèle d'un système dynamique continu. Les matrices d'état intervenant dans ce type de système dans les applications sont souvent creuses et de grandes dimensions, ce qui se trouve justement être le champ d'application naturel de la méthode ADI pour la résolution approchée d'une équation de Sylvester. Dans le but d'affiner notre étude concernant le taux de convergence de la méthode ADI, le chapitre suivant est dédié à l'étude du problème de Zolotarev pour des ensembles discrets en tant que problème d'approximation rationnelle discrète du plan complexe, étape nécessaire avant de mener à bien par la suite l'étude asymptotique de ce problème.

Chapitre 3

Problème de Zolotarev pour des ensembles discrets

Le problème de Zolotarev pour des ensembles discrets a été défini en 1.2.2 et est apparu naturellement comme majorant pour l'erreur commise dans l'application de la méthode ADI pour la résolution approchée d'une équation de Sylvester ou de Lyapounov. On a également présenté le lien entre ce problème et d'autres questions, dans un contexte d'algèbre linéaire numérique comme la décroissance des valeurs singulières d'une matrice de petit rang de déplacement, ou dans le contexte de l'approximation rationnelle de la fonction signe sur des ensembles discrets. Nos motivations étant établies, on s'y intéresse dans ce chapitre en tant que problème d'approximation rationnelle du plan complexe pour des ensembles discrets, en étudiant les questions d'existence d'une solution et de dégénérescence éventuelle de celle-ci.

3.1 Présentation du problème

In spite of the presently known applications (which are remarkable to the highest degree) of elliptic functions to number theory, geometry and mechanics, I submit that the theory of elliptic functions still leaves much to be desired from the aspect of applications.

Therefore I did not regard as superfluous to consider certain smallest-quantity problems that can be solved with the help of basic formulas in the theory of elliptic functions. These questions belong to the class of smallest quantity problems for which methods of solution were first given by P. L. Chebychev.

Cette citation de Zolotarev, élève de Tchebychev, introduit l'article de 1877 *Applications of elliptic functions to questions of functions deviating least or most from zero* et est reprise par Akhieser [Ak90, Section 9].

3.1.1 Cas classique

On s'intéresse ici au *troisième problème de Zolotarev* qui a été défini en 1.2.2, les autres problèmes de Zolotarev n'entrant pas en considération ici, on notera plus brièvement *problème de Zolotarev* en lieu et place de *troisième problème de Zolotarev*.

Pour un exposé des différents problèmes de Zolotarev, voir [Ac56, section E].

On expose ici quelques résultats bien connus établis pour le problème de Zolotarev sur des ensembles continus.

Dans le cas de deux intervalles réels E et F, Zolotarev a résolu explicitement le problème classique en terme de fonctions elliptiques de Legendre et de Jacobi. À ce sujet, on rappelle que la fonction elliptique de Legendre de première espèce K est définie pour $k \in (0,1)$ par

$$K(k) := K\left(k'\right) = \int_0^1 \frac{\mathrm{d}t}{\sqrt{(1-t^2)(1-t^2)}},$$

on définit en outre $k' := \sqrt{1-k^2}$ ainsi que

$$K'(k) := K\left(k'\right) = \int_0^1 \frac{\mathrm{d}t}{\sqrt{(1-t^2)(1-(1-k^2)t^2)}}.$$

Par ailleurs, un rappel plus complet concernant les fonctions elliptiques de Legendre est donné à la proposition 7.2.1.

Enfin, l'inégalité plus forte énoncée dans la proposition est tirée de [Br80, Théorème V.5.D.5] et [SaTo97, Théorème VIII.3.1].

Proposition 3.1.1. *Si $E = -F = [\alpha, \beta] \subset (0, +\infty)$, on a*

$$\lim_{n\to+\infty} Z_n(E,F)^{\frac{1}{n}} = \rho,\ \ \rho := \exp\left(-2\pi\frac{K(k)}{K'(k)}\right),\ \ k = \frac{\alpha}{\beta}.$$

On a dans ce cas le résultat plus précis suivant :

$$\rho^{n-1} \leqslant Z_n(E,F) \leqslant 16\rho^{n-1}. \tag{3.1}$$

Pour des ensembles E et F plus généraux, Gonchar a donné dans [Go78] le résultat suivant concernant l'asymptotique de la quantité de Zolotarev.

Proposition 3.1.2. *Soient E et F deux compacts disjoints du plan complexe de capacité logarithmique positive et de complémentaire connexe.*

On a alors

$$\lim_{n\to+\infty} Z_n(E,F)^{\frac{1}{n}} = \exp\left(-\frac{1}{\operatorname{cap}(E,F)}\right).$$

où $\operatorname{cap}(E,F)$ désigne la capacité du condensateur à plateau E et F, définie en 6.2.2.

On remarque ici l'apparition d'un terme dont on peut donner une interprétation électrostatique, la preuve de 3.1.2 consiste en effet à décrire l'asymptotique du problème considéré par un problème de théorie du potentiel logarithmique.

Mentionnons également l'étude du problème de Zolotarev généralisé faite dans [LeSa01] dont on cite la proposition suivante.

Proposition 3.1.3. *On a pour des ensembles E et F vérifiant les hypothèses de la proposition précédente*

$$\lim_{n,m\to+\infty,\, n/m\to\lambda} Z_{n,m}(E,F)^{\frac{1}{n+m}} = \exp\left(-G(\tau)\right),\ \ \tau = \frac{\lambda}{\lambda+1}$$

où la fonction G est définie sur $[0,1)$ concave, continue et positive.

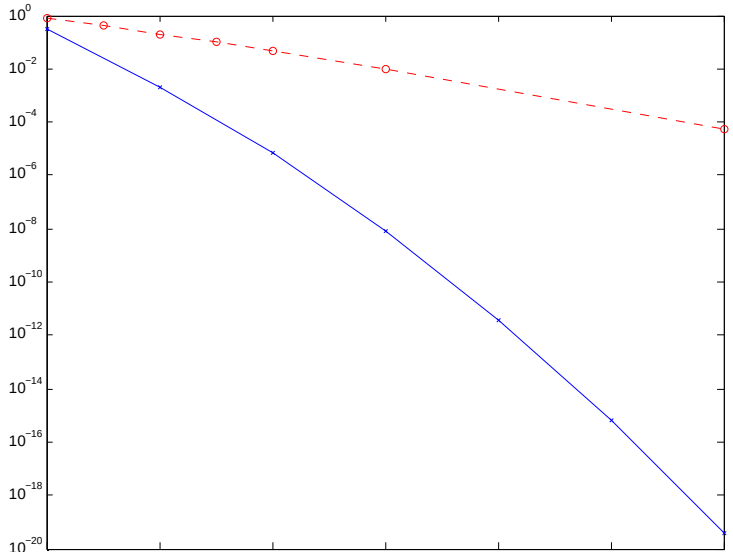

FIG. 3.1 – *$Z_n(E,-E)$ (trait plein) et $Z_n(conv(E),-conv(E))$ (trait pointillé), $E = \{1/N^4 + \cos(k\pi/(2N))\}_{1\leqslant k\leqslant N}$, $N = 20$, en abscisse le degré des fractions rationnelles considérées n, où $1 \leqslant n \leqslant 14$.*

La fonction G utilisée dans la proposition précédente est construite à partir d'un problème extrémal en théorie du potentiel logarithmique.

Les principaux éléments énoncés ci-dessus sont également repris dans [SaTo97, Chapitre VIII annexe].

3.1.2 Cas discret

On s'intéressera dans ce qui suit au problème de Zolotarev pour des ensembles discrets.

La remarque suivante est essentielle et permet de restreindre notre étude aux cas d'ensembles compacts du plan complexe dans tout ce qui suit.

Définition 3.1.4. *Une* homographie *ou* transformation de Moëbius h *est une application de l'ensemble $\mathbb{C}_\infty$ dans lui-même donnée par*

$$h(z) := \frac{\alpha z + \beta}{\gamma z + \delta}, \text{ où } (\alpha,\beta,\gamma,\delta) \in \mathbb{C} \text{ tels que } \alpha\delta - \beta\gamma \neq 0.$$

La condition de non nullité $\alpha\delta - \beta\gamma$ est imposée pour éviter le cas des applications constantes.

Remarque. Soient $n \geqslant 1$, E, F deux ensembles discrets disjoints du plan complexe, h une homographie. On remarque qu'une fraction rationnelle r est extrémale pour le problème de Zolotarev $Z_n(E,F)$ si et seulement si la fraction rationnelle $r\circ h^{-1}$ de $\mathcal{R}_n$ est extrémale pour le problème $Z_n\left(h(E),h(F)\right)$. On supposera donc dorénavant les ensembles E et F compacts, sans perte de généralité pour l'étude du problème $Z_n(E,F)$.

Le lemme suivant est simple, mais essentiel.

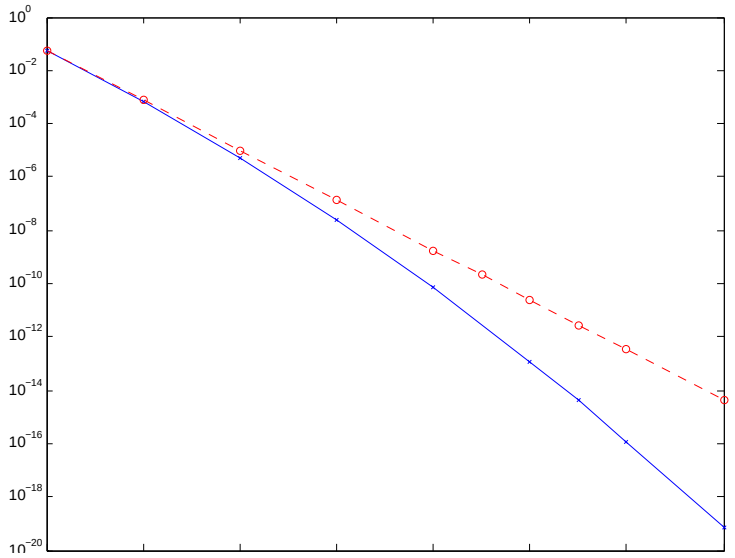

FIG. 3.2 – *$Z_n(E,-E)$ (trait plein) et $Z_n(conv(E),-conv(E))$ (trait pointillé), $E = \{k/N\}_{1\leqslant k\leqslant N}$, $N = 25$, en abscisse le degré des fractions rationnelles considérées n, où $1 \leqslant n \leqslant 16$.*

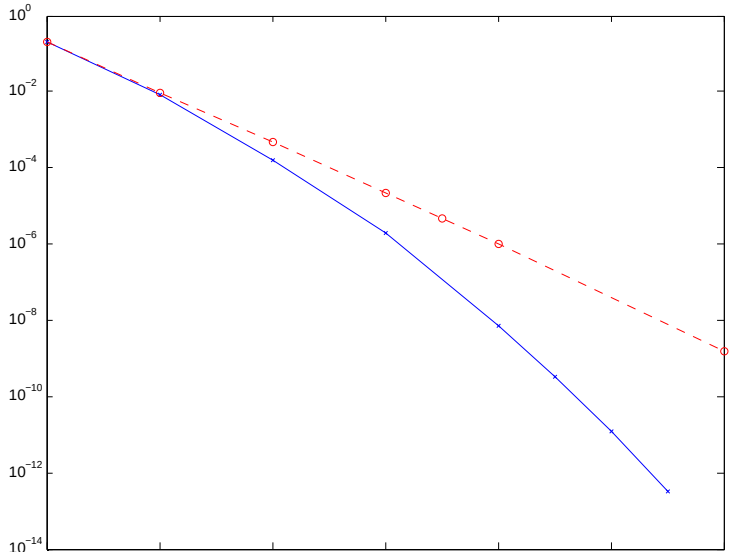

FIG. 3.3 – *$Z_n(E,-E)$ (trait plein) et $Z_n(conv(E),-conv(E))$ (trait pointillé), $E = \{2-2\cos(k\pi/(N+1))\}_{1\leqslant k\leqslant N}$, $N = 20$, en abscisse le degré des fractions rationnelles considérées n, où $1 \leqslant n \leqslant 14$.*

Lemme 3.1.5. *Soit E, F deux ensembles discrets disjoints du plan complexe et conv(E), conv(F) leurs enveloppes convexes respectives. Alors, pour tout $n, m \geqslant 1$ on a*

$$Z_{m,n}(E,F) \leqslant Z_{m,n}(conv(E), conv(F)).$$

PREUVE : Comme par définition de l'enveloppe convexe $E \subset \mathrm{conv}(E)$ et $F \subset \mathrm{conv}(F)$, on a pour toute fraction rationnelle de $\mathcal{R}_{m,n}$

$$\|r\|_{L^\infty(E)}\|r^{-1}\|_{L^\infty(F)} \leqslant \|r\|_{L^\infty(\mathrm{conv}(E))}\|r^{-1}\|_{L^\infty(\mathrm{conv}(F))}$$

d'où le résultat en passant à l'infimum sur $\mathcal{R}_{m,n}$. □

Jusqu'à maintenant, dans une situation donnant lieu à l'étude d'un problème de Zolotarev sur des ensembles discrets, on procède ainsi : on remplace les ensembles discrets E et F naturellement liés au problème considéré -typiquement, tout ou partie du spectre de certaines matrices mises en jeu- par leur enveloppe convexe. On utilise ensuite la relation du lemme 3.1.5 pour appliquer les travaux cités dans le paragraphe précédent, par exemple la proposition 3.1.2 afin d'en déduire une borne supérieure pour l'asymptotique de la quantité de Zolotarev, voir par exemple [EFLSV02].

En procédant ainsi, on peut cependant obtenir une très grande surestimation de l'asymptotique de la quantité de Zolotarev pour des ensembles discrets, ceci étant dû au fait qu'une fraction rationnelle petite sur un ensemble discret ne l'est pas forcément sur toute l'enveloppe convexe de cet ensemble.

Pour illustrer le phénomène de surestimation décrit au paragraphe précédent, on présente dans les figures 3.1, 3.2 et 3.3 la quantité $Z_n(E,-E)$ (trait plein marqué par des croix) contre la quantité $Z_n(\mathrm{conv}(E), -\mathrm{conv}(E))$ (trait pointillé marqué par des cercles) sur une échelle semi-logarithmique pour des degrés n croissants et les ensembles discrets -dont le choix sera explicité dans la section consacrée aux exemples numériques-

$$E = \{1/N^4 + \cos(k\pi/(2N))\}_{1\leqslant k\leqslant N}\ (N=20), \quad E = \{k/N\}_{1\leqslant k\leqslant N}\ (N=25)$$

et enfin

$$E = \{2 - 2\cos(k\pi/(2N+1))\}_{1\leqslant k\leqslant N}\ (N=20).$$

La distance entre les deux courbes semble plus importante si E est plus proche de $-E$, et on obtient en accord avec (3.1) des quasi-droites dans le cas des enveloppes convexes alors que la courbe pour le cas discret semble concave.

L'explication des procédures numériques qui ont conduit à l'obtention de ces figures sera également détaillée au chapitre 8.

Le travail qui suit est voué à l'étude asymptotique de la quantité de Zolotarev sur des ensembles discrets moyennant certaines hypothèses concernant la répartition asymptotique des familles d'ensembles discrets considérées. On établit ainsi un taux de convergence plus précis pour la méthode ADI, ce qui met en évidence un phénomène de convergence superlinéaire de cette méthode jusqu'à maintenant constaté empiriquement. Avant de mener à bien cette étude asymptotique, on s'intéresse aux questions naturelles d'existence et d'éventuelle dégénérescence d'une solution pour le problème de Zolotarev sur des ensembles discrets à degré fixé.

3.2 Existence et dégénérescence de la solution

3.2.1 Existence du minimiseur pour le problème de Zolotarev

L'adaptation au contexte de l'approximation rationnelle sur des ensembles discrets d'arguments permettant d'assurer l'existence d'un minimiseur dans le cadre continu n'est pas toujours aisée, et il est bien connu que certains problèmes de minimisation en approximation rationnelle discrète n'admettent pas de solution.

Dans le cas d'ensembles continus, il existe cependant des résultats classiques d'existence et d'équioscillation, voir par exemple [Ac56, Section 32] et [Br80, section 5.1] pour une présentation de ce type de résultat. On y présente par exemple la démonstration pour l'existence dans le cas classique d'un problème du type

$$\min_{r\in\mathcal{R}_n}\max_{z\in[a,b]}|f(z)-r(z)|$$

où f est par exemple continue sur $[a,b]$. Ce type de preuve utilise l'existence et la convergence uniforme d'une suite de minimiseurs sur $[a,b]$ privé d'un ensemble fini de points, et ce type de raisonnement ne passe bien entendu pas aisément aux ensembles discrets.

Dans le cas continu, on peut ensuite montrer qu'en cas de non-dégénérescence de la solution du problème de minimisation, ce qui signifie que celle-ci est bien de degré (n,n) s'il s'agit d'un problème de minimisation sur $\mathcal{R}_n$, on peut former une suite minimisante convergeant uniformément sur tout l'intervalle $[a,b]$ considéré, et étudier ensuite les questions d'équioscillation et d'unicité des solutions.

Pour mettre en évidence les difficultés spécifiques à la théorie de l'approximation sur des ensembles discrets, on cite l'exemple suivant depuis [Br80, p. 109] : soient $m,n\geqslant 1$ deux entiers et $E:=\left\{0,\frac{1}{N},\dots,1\right\}$ où $N>m+n$. On définit la fonction f sur E par

$$f(x)=1,\ \text{si } x=0,\ \ f(x)=0 \text{ sinon.}$$

Alors, f n'est à l'évidence pas un élément de $\mathcal{R}_{m,n}$, et la suite

$$u_k(x):=\frac{1}{1+kx},\ \ k\geqslant 1$$

est une suite minimisante au sens où la quantité $dist(u_k,\mathcal{R}_{m,n})$ est de limite nulle pour $k\to+\infty$.

On peut cependant montrer qu'il existe un minimiseur pour le problème de Zolotarev discret généralisé, et cette section est dédiée à la preuve de ce résultat. On prouve d'abord ce résultat en toute généralité grâce à la reformulation de celui-ci assortie de considérations de compacité. On utilise ensuite une hypothèse de séparation sur les ensembles discrets considérés pour donner une autre preuve d'existence en se rattachant à des arguments d'analyse classiques grâce à quelques considérations de géométrie élémentaire.

Cette démonstration géométrique présente en outre l'avantage d'offrir un résultat de localisation des pôles et des zéros du minimiseur, et c'est sous cette forme que nous l'énoncerons.

On caractérise enfin l'éventuelle dégénérescence de la solution dans le cas du problème $Z_1(E,F)$ pour le cas de deux ensembles E et F à deux éléments, avant de donner une condition pour que la solution d'un problème de Zolotarev sur $\mathcal{R}_n$ soit bien de degré maximal.

On note cependant que l'existence d'un minimiseur pour le problème de Zolotarev ne joue pas de rôle pour établir le comportement asymptotique de celui-ci, mais cette question s'impose naturellement dès lors que l'on a affaire à un problème d'approximation rationnelle.

On utilise dans la première preuve de l'existence d'un minimiseur la notion de pseudo-inverse de Moore-Penrose, voir par exemple [GoVLo96, p. 257] pour plus de précisions à ce sujet.

Proposition 3.2.1. *Soient $n, m \geqslant 1$, E et F des ensembles discrets disjoints du plan complexe tels que* $\mathrm{card}(E) \geqslant m+1$ *et* $\mathrm{card}(F) \geqslant n+1$. *Alors, il existe un minimiseur pour $Z_{m,n}(E,F)$, autrement dit, il existe $r \in \mathcal{R}_{m,n}$ telle que*

$$Z_{n,m}(E,F) = \|r\|_{L^\infty(E)} \|r^{-1}\|_{L^\infty(F)}.$$

PREUVE : Considérons le problème de minimisation suivant : on note $N := card(E) > m$, $M := card(F) > n$, $E = \{a_1, \ldots, a_M\}$, $F = \{b_1, \ldots, b_N\}$, et on veut résoudre

$$\begin{cases} \min Z^2, \\ \vec{p} \in \mathbb{C}^{m+1},\ \vec{q} \in \mathbb{C}^{n+1},\ Z \in \mathbb{R}, \\ \forall j = 1, ..., M : \quad |(1, a_j, ..., a_j^m)\vec{p}|^2 \leqslant Z\,|(1, a_j, ..., a_j^n)\vec{q}|^2, \\ \forall j = 1, ..., N : \quad |(1, b_j, ..., b_j^n)\vec{q}|^2 \leqslant Z\,|(1, b_j, ..., b_j^m)\vec{p}|^2, \\ \|\vec{q}\|^2 = 1. \end{cases}$$

On adopte ici les conventions de notations suivantes :

$$p(x) = \sum_{j=0}^{m} p_j x^j,$$

et

$$\vec{p} = \begin{pmatrix} p_0 \\ \vdots \\ p_m \end{pmatrix},$$

donc avec ces conventions de notations, $(1, a_j, ..., a_j^m)\vec{p} = p(a_j)$.

Sans changer la valeur optimale, on peut ajouter la contrainte $0 \leqslant Z \leqslant 1$ d'après les inégalités énoncées dans le cas $p = (1, 0, \ldots, 0)$ et $q = (1, 0, \ldots, 0)$. De plus, en notant V la matrice de Vandermonde

$$\begin{pmatrix} 1 & a_0 & \ldots & a_0^m \\ 1 & a_1 & \ldots & a_1^m \\ \vdots & \vdots & \vdots & \vdots \\ 1 & a_M & \ldots & a_M^m \end{pmatrix},$$

on a

$$\begin{pmatrix} p(a_0) \\ \vdots \\ p(a_M) \end{pmatrix} = V\vec{p}.$$

La matrice V admet un pseudo-inverse au sens de Moore-Penrose V^+, et comme $M > m$, les colonnes de V sont indépendantes et la matrice V^+V n'est autre que la matrice identité

de dimension $(m+1)\times(m+1)$, d'où

$$\|\vec{p}\| \leqslant \left\| V^+ \begin{pmatrix} q(a_0) \\ \vdots \\ q(a_M) \end{pmatrix} \right\|,$$

et d'après l'hypothèse $\|\vec{q}\| = 1$, on en déduit que $\|\vec{p}\|$ est borné supérieurement.

Le problème de minimisation sur le triplet $(Z, \vec{p}, \vec{q})$ est un problème de minimisation sur un compact de l'espace de dimension finie sur $\mathbb{R}$, $\mathbb{R}\times\mathbb{C}^{m+1}\times\mathbb{C}^{n+1}$, ce problème admet donc un minimiseur que l'on note $(Z^\star, \vec{p}^\star, \vec{q}^\star)$.

Quitte à renormaliser $\vec{q}^\star$, on peut supposer sans perte de généralité la fraction rationnelle $r^\star := \frac{p^\star}{q^\star}$ irréductible, et par conséquent, les polynômes $p^\star$ et $q^\star$ n'ont pas de zéro commun. Comme $Z^\star \leqslant 1$ est une quantité finie, on en déduit que $p^\star(b_k) \neq 0$ pour $k \in \{1, N\}$ et $q^\star(a_j) \neq 0$ pour $j \in \{1, M\}$.

Compte-tenu de ce qui précède, en considérant la fraction rationnelle $r^\star$, on obtient $Z^\star \geqslant Z_{m,n}(E,F)$.

Soit $\varepsilon > 0$: il existe par définition une fraction rationelle r de $\mathcal{R}_{m,n}$ vérifiant

$$\|r\|_{L^\infty(E)}\|r^{-1}\|_{L^\infty(F)} \leqslant Z_{m,n}(E,F) + \varepsilon.$$

On écrit $r = \frac{p}{q}$ sous forme irréductible et en multipliant p par un scalaire, on normalise les coefficients de q sous la contrainte $\|\vec{q}\| = 1$.

Maintenant, quitte à multiplier r par une constante non nulle, ce qui ne modifie pas la quantité

$$\|r\|_{L^\infty(E)}\|r^{-1}\|_{L^\infty(F)},$$

on peut supposer

$$\|r\|_{L^\infty(E)} = \|r^{-1}\|_{L^\infty(F)} = \sqrt{Z_{m,n}(E,F)+\varepsilon}.$$

Comme p et q n'ont pas de zéros communs et que la quantité $Z_{m,n}(E,F)$ est finie, p ne s'annule pas sur F et q ne s'annule pas sur E, on obtient donc les inégalités

$$\forall a \in E,\ |p(a)| \leqslant \sqrt{(Z_{m,n}(E,F)+\varepsilon)}\,|q(a)|\,,$$

ainsi que

$$\forall b \in F,\ |q(b)| \leqslant \sqrt{(Z_{m,n}(E,F)+\varepsilon)}\,|p(b)|\,,$$

et

$$\begin{pmatrix} Z_{m,n}(E,F)+\varepsilon \\ p \\ q \end{pmatrix}$$

est donc un candidat au problème de minimisation ci-dessus. On a prouvé l'inégalité valable pour tout $\varepsilon > 0$

$$Z^\star \leqslant Z_{m,n}(E,F) + \varepsilon,$$

ce qui permet de conclure à l'existence d'un minimiseur pour le problème de Zolotarev généralisé. □

Voici une propriété de localisation des zéros et des pôles d'un minimiseur pour le problème de Zolotarev sous une hypothèse supplémentaire de séparation des ensembles considérés.

Définition 3.2.2. *On appelle* disque généralisé *sur* $\mathbb{C}_\infty$ *un disque de centre un point de* $\mathbb{C}_\infty$ *et de rayon strictement positif, ou un demi-plan.*

Définition 3.2.3. *Soient* D_1 *et* D_2 *deux disques généralisés sur la sphère de Riemann. Pour* $\ell = 1, 2$*, on note* $\mathcal{C}_\ell$ *la frontière de* D_ℓ*.*

On dit que D_1 *et* D_2 *sont* strictement séparés *s'ils sont d'intersection vide et que* $\mathcal{C}_1$ *et* $\mathcal{C}_2$ *sont soit deux cercles concentriques disincts soit deux droites parallèles distinctes.*

On remarque que dans le cas classique où E est contenu dans le demi-plan ouvert droit et F dans le demi-plan ouvert gauche, il existe deux demi-plans strictement séparés contenant E et F, ce qui entre bien dans le cadre de la définition précédente.

On adopte dorénavant la convention de notation classique qui consiste pour S un sous-ensemble donné du plan complexe à noter $conv(S)$ son enveloppe convexe.

Proposition 3.2.4. *Soient* E*,* F *deux ensembles discrets finis de cardinal strictement supérieur à un entier* $n \geqslant 1$ *donné. On suppose qu'il existe deux disques généralisés* D_E *et* D_F *strictement séparés, de frontières respectives* $\mathcal{C}_E$ *et* $\mathcal{C}_F$*, tels que les ensembles discrets* E *et* F *vérifient* $E \subset D_E$ *et* $F \subset D_F$*.*

Il existe un minimiseur $r^\star$ *pour* $Z_n(E, F)$ *dont les zéros sont dans* $\mathrm{conv}(E)$ *et les pôles dans* $\mathrm{conv}(F)$*.*

PREUVE : Le problème de Zolotarev considéré dans le cadre de cet énoncé admet un minimiseur d'après 3.2.1, mais on peut grâce à des arguments élémentaires de géométrie se ramener à un cas standard d'existence d'un minimiseur en théorie de l'approximation pour en déduire l'existence d'une fraction rationnelle réalisant l'infimum pour le problème de Zolotarev, et cet argument nous permet de surcroît d'obtenir le résultat de localisation énoncé.

On prouve cet énoncé dans le cas de d'une séparation stricte par des demi-plans.

Soit une fraction rationnelle $r \in \mathcal{R}_n$ candidate admettant un zéro $z_0 \notin D_E$. On construit à partir de r la fraction rationnelle $\tilde{r}$ en remplaçant z_0 par son image par la réflexion par rapport à la droite $\mathcal{C}_F$ notée $\tilde{z}_0$, et comme

$$\forall (x, y) \in E \times F, \ \left|\frac{x - \tilde{z}_0}{x - z_0}\right| \left|\frac{y - z_0}{y - \tilde{z}_0}\right| \leqslant 1,$$

on montre facilement que celle-ci vérifie

$$\|\tilde{r}\|_{L^\infty(E)} \|\tilde{r}^{-1}\|_{L^\infty(F)} \leqslant \|r\|_{L^\infty(E)} \|r^{-1}\|_{L^\infty(F)},$$

ce qui permet ainsi de se restreindre uniquement aux fractions rationnelles admettant tous leurs zéros dans D_E.

On procède de même pour les pôles en considérant la fraction rationnelle construite avec l'image des pôles de la fraction rationnelle originelle par la réflexion par rapport à $\mathcal{C}_E$, d'où

$$Z_n(E, F) = \inf_{r \in \mathcal{R},\ \text{zéros}(r) \subset D_E,\ \text{pôles}(r) \subset D_F} \|r\|_{L^\infty(E)} \|r^{-1}\|_{L^\infty(F)}.$$

Dans le cas de disques dont les frontières sont des cercles concentriques, le même raisonnement utilisant les inversions par rapport à ceux-ci permet de conclure.

La preuve de l'existence d'un minimiseur pour un tel problème est alors classique grâce à la séparation des zéros et des pôles, voir par exemple [Ac56, Section 33] ou encore [Br80, Lemme V.1.1]. □

Une fois établie l'existence d'une solution, la théorie de l'approximation rationnelle pose le problème de la dégénérescence éventuelle d'une solution : pour un problème de minimisation sur l'ensemble $\mathcal{R}_{m,n}$, peut-on assurer que le minimiseur soit bien de degré (m,n) ? On donne dans la section suivante l'étude du cas du problème de Zolotarev sur l'ensemble $\mathcal{R}_1$ qui nous permettra d'en déduire une condition de ce type par la suite.

3.2.2 Étude exhaustive d'un cas simple

On définit pour commencer la notion de deux ensembles du plan complexe dont les points sont entrelacés.

Définition 3.2.5. *Soient $E = \{a,b\}$ et $F = \{c,d\}$ quatre points distincts du plan complexe cocycliques ou alignés, autrement dit, sur un même cercle généralisé noté $\mathcal{C}$.*

On note $\widehat{ab}$ l'arc de cercle -le segment dans le cas de 4 points alignés- joignant a à b de mesure minimale pour la mesure de Lebesgue sur le cercle $\mathcal{C}$, de même pour $\widehat{cd}$.

On dit que les points des ensembles E et F sont entrelacés *si et seulement si les points (a,b,c,d) sont sur un même cercle généralisé $\mathcal{C}$ et si les arcs $\widehat{ab}$ et $\widehat{cd}$ sont d'intersection non vide et non inclus l'un dans l'autre.*

On considère maintenant le problème de Zolotarev pour des fractions rationnelles de l'ensemble $\mathcal{R}_1$ dans le cas de deux ensembles réels de deux points alignés entrelacés.

Lemme 3.2.6. *Pour $\kappa \in (0,1)$, on définit les ensembles discrets $E = \{-1,\kappa\}$ et $F = \{-\kappa,1\}$. On a alors*

$$Z_1(E,F) = 1,$$

ce qui signifie que ce problème admet une solution dégénérée égale à la fraction rationnelle constante 1.

PREUVE : On a tout d'abord l'inégalité $Z_1(E,F) \leqslant 1$ en rappelant que la fraction rationnelle constante 1 est un candidat pour le problème $Z_1(E,F)$, il suffit donc de prouver l'inégalité $Z_1(E,F) \geqslant 1$ pour conclure.

On définit les matrices

$$X := \begin{pmatrix} \frac{1-\kappa}{1+\kappa} & \frac{2\sqrt{\kappa}}{1+\kappa} \\ \frac{2\sqrt{\kappa}}{1+\kappa} & -\frac{1-\kappa}{1+\kappa} \end{pmatrix},$$

$$A := diag(-1,\kappa) \text{ et } B := diag(1,-\kappa).$$

La matrice X est unitaire, et par conséquent le conditionnement de X vaut 1, et on montre par simple calcul matriciel que la matrice $AX - XB$ est de rang 1.

On a alors d'après l'équation (1.10) comme $\rho_{A,B}(X) = 1$ l'inégalité

$$Z_1(E,F) \geqslant \frac{1}{\|X\|\|X^{-1}\|} = 1,$$

ce qui permet de conclure. □

Les points de E et F dans le cas du lemme 3.2.6 sont entrelacés, et dans ce cas, le problème de Zolotarev admet un minimiseur dégénéré.

On cherche maintenant à généraliser ce résultat en explorant les cas de dégénérescence dans le cas de deux ensembles de deux points distincts quelconques du plan complexe, en tirant partie de l'invariance de la quantité de Zolotarev par transformation homographique. Pour ce faire, on précise quelques définitions et autres résultats classiques de géométrie dans le plan complexe dans ce qui suit.

On définit pour commencer le birapport de quatre points distincts du plan complexe, quantité qui permettra de simplifier considérablement les calculs explicites que l'on mènera pour un condensateur réel dans le chapitre 7.

Proposition 3.2.7. *On rappelle la définition du birapport de quatre nombres complexes distincts.*

$$[a,b,c,d] := \frac{c-a}{c-b}\frac{d-b}{d-a}.$$

Si l'un des points est infini, on étend la définition de la quantité ci-dessus par passage à la limite.

Si h est une homographie, alors

$$[h(a),h(b),h(c),h(d)] = [a,b,c,d].$$

Le lemme suivant est cité de [Han04, p. 188].

Lemme 3.2.8. *Soit a,b,c,d quatre réels distincts tels que $[a,b,c,d] > 0$.*

Il existe un unique élément k de $(0,1)$ tel qu'on ait l'égalité des birapports

$$[a,b,c,d] = [-1,-k,k,1],$$

et celui-ci est donné par

$$k := k(a,b,c,d) = \frac{1-\sqrt{[a,d,b,c]}}{1+\sqrt{[a,d,b,c]}}.$$

Lemme 3.2.9. *Soient (a,b,c,d) quatre points disctincts du plan complexe.*

Alors,

$$[a,b,c,d] \in \mathbb{R} \Leftrightarrow \ (a,b,c,d) \text{ sont sur un même cercle généralisé,}$$

et dans ce cas,

$$[a,b,c,d] < 0 \Leftrightarrow \text{ les points de } E \text{ et } F \text{ sont entrelacés.}$$

PREUVE : La première partie de l'énoncé est classique et ne sera pas redémontrée ici.

Si (a,b,c,d) sont sur un cercle $\mathcal{C}$ de centre z, en se placant dans un repère de centre z et d'axe des abscisses la droite (za), un calcul simple donne en notant θ_w la mesure dans $[0,2\pi)$ de l'angle $\widehat{azw}$ dans ce repère où $\mathcal{C}$ est orienté dans le sens trigonométrique,

$$[a,b,c,d] = \frac{\sin\left(\frac{\theta_c-\theta_a}{2}\right)}{\sin\left(\frac{\theta_c-\theta_b}{2}\right)}\frac{\sin\left(\frac{\theta_d-\theta_b}{2}\right)}{\sin\left(\frac{\theta_d-\theta_a}{2}\right)},$$

ce qui prouve le résultat dans le cas de points cocycliques. Un raisonnement similaire permet de conclure dans le cas de points alignés. □

Citons enfin le résultat suivant depuis [Si67, Chapitre 5 Théorème 5.5].

Proposition 3.2.10. *Soient $\mathcal{C}_1$ et $\mathcal{C}_2$ deux cercles généralisés, z_1, z_2, z_3 et w_1, w_2, w_3 deux triplets de points distincts appartenant respectivement à $\mathcal{C}_1$ et $\mathcal{C}_2$.*

Alors, il existe une homographie h envoyant $\mathcal{C}_1$ sur $\mathcal{C}_2$ de façon à ce que

$$h(z_\ell) = w_\ell,\ \ell \in \{1,2,3\}.$$

Les rappels de géométrie nécessaires à ce qui suit étant réalisés, voici maintenant la proposition principale de cette section.

Proposition 3.2.11. *Avec les mêmes notations que précédemment,*

$$Z_1(E,F) = 1 \ \Leftrightarrow\ \textit{les points de } E \textit{ et } F \textit{ sont entrelacés.}$$

PREUVE : Supposons les points de E et F entrelacés, soit $\mathcal{C}$ le cercle généralisé auquel appartiennent ces points.

Alors, $[a,b,c,d]$ est un réel strictement négatif d'après 3.2.9, donc $[a,c,b,d] > 0$ et on note $k := k(a,c,b,d)$ défini en 3.2.8.

Alors, en notant

$$K := \frac{\sqrt{\frac{c-a}{b-a}\frac{d-c}{d-b}} - 1}{\sqrt{\frac{c-a}{b-a}\frac{d-c}{d-b}} + 1}$$

et

$$h_1(z) = k\frac{2z - (c+b) + K(b-c)}{2zK + (b-c) - K(c+b)},$$

cette homographie vérifie

$$h_1(a) = -1,\ h_1(b) = k,\ h_1(c) = -k \text{ et } h_1(d) = 1,$$

et les points des ensembles $\{-1,k\}$ et $\{-k,1\}$ sont entrelacés d'après le signe de leur birapport.

Pour ce choix de k, on a alors d'après 3.1.2

$$Z_1\left(\{a,b\},\{c,d\}\right) = Z_1\left(\{-1,k\},\{-k,1\}\right),$$

ce qui permet de conclure d'après le lemme 3.2.6.

Supposons maintenant les points de E et F non entrelacés. Daprès 3.2.10, il existe une homographie h telle que $h(a) = -1$, $h(b) = 1$ et $h(c) = 2$. Quitte à échanger la valeur de $h(c)$ pour une autre valeur réelle strictement supérieure à 1, on suppose que $z := h(d)$ n'est pas infini.

Si $[a,b,c,d] \in \mathbb{R}$, les points (a,b,c,d) sont cocycliques ou alignés et non entrelacés, et dans ce cas z est réel et n'appartient pas à $[-1,1]$. Pour $\varepsilon > 0$ assez petit, les disques généralisés

$$D_E := \left\{|w| \leqslant 1 + \frac{\varepsilon}{2}\right\} \text{ et } D_F := \{|w| \geqslant 1+\varepsilon\}$$

contiennent respectivement $h(E) = \{-1,1\}$ et $h(F) = \{2,z\}$.

Alors, la fraction rationnelle $r(z) = z$ vérifie

$$\max_{z\in h(E)} |r(z)| < 1 + \frac{\varepsilon}{2} \text{ et } \max_{z\in h(E)} |r^{-1}(z)| < \frac{1}{1+\varepsilon},$$

et par conséquent, $Z_1\left(h(E), h(F)\right) < 1$. Enfin, comme $Z_1\left(h(E), h(F)\right) = Z_1(E, F)$ par invariance de la quantité de Zolotarev par transformation homographique d'après 3.1.2, cela permet de conclure dans ce cas.

On suppose maintenant $[a, b, c, d] \notin \mathbb{R}$ et on adapte la preuve précédente.

On choisit alors un cercle $\mathcal{C}$ dont le centre est sur l'axe imaginaire de rayon suffisamment grand pour que z ne soit pas pas dans le disque $D := conv(\mathcal{C})$, notons ω le centre de $\mathcal{C}$ et R son rayon. Pour $\varepsilon > 0$, on note D_ε l'image de D par l'homothétie de rapport $1 + \varepsilon$ et de centre ω.

Pour $\varepsilon > 0$ suffisamment petit, les points -1 et 1 sont dans l'intérieur de D_ε et les points 2 et z dans son complémentaire D_ε^c, et on a même

$$\{-1, 1\} \in D_{\varepsilon/2} \text{ et } \{2, z\} \in D_\varepsilon^c.$$

On définit alors la fraction rationnelle $r(z) := z - \omega$, et celle-ci vérifie par construction

$$\max_{z \in h(E)} |r(z)| < R\left(1 + \frac{\varepsilon}{2}\right) \text{ et } \max_{z \in h(E)} |r^{-1}(z)| < \frac{1}{R(1+\varepsilon)},$$

d'où le résultat. □

Nous disposons donc maintenant d'un critère géométrique simple de dégénérescence dans le cas $n = 1$ et E et F de cardinal 2. On généralise ce critère dans la section suivante.

3.2.3 Dégénérescence d'un minimiseur

On s'intéresse maintenant à la caractérisation de l'existence d'un minimiseur dégénéré dans le cas général $Z_n(E, F)$ en utilisant le critère établi en 3.2.11.

Proposition 3.2.12. *Soient E, F deux ensembles de $N \geqslant 1$ points distincts du plan complexe.*

la suite $(Z_n(E, F))_{0 \leqslant n \leqslant N}$ est strictement décroissante si et seulement si $Z_1(E, F) < 1$.

PREUVE : Si la suite $(Z_n(E, F))_{0 \leqslant n \leqslant N}$ est strictement décroissante, comme $Z_0(E, F) = 1$, cela fournit directement une implication.

Réciproquement, supposons qu'il existe n dans $\{1, \ldots, N-1\}$ tel que

$$Z_n(E, F) = Z_{n-1}(E, F).$$

Soient $\rho \in \mathcal{R}_1$ et $r \in \mathcal{R}_{n-1}$ telle que

$$Z_n(E, F) = \|r\|_{L^\infty(E)} \|r^{-1}\|_{L^\infty(F)} :$$

une telle fraction rationnelle existe d'après ce qui précède, et comme $r\rho \in \mathcal{R}_n$, on a

$$\|r\rho\|_{L^\infty(E)} \|(r\rho)^{-1}\|_{L^\infty(F)} \geqslant \|r\|_{L^\infty(E)} \|r^{-1}\|_{L^\infty(F)}.$$

On a de plus les inégalités

$$\|r\rho\|_{L^\infty(E)} \leqslant \|r\|_{L^\infty(E)} \|\rho\|_{L^\infty(E)} \text{ et } \|(r\rho)^{-1}\|_{L^\infty(F)} \leqslant \|r^{-1}\|_{L^\infty(F)} \|\rho^{-1}\|_{L^\infty(F)},$$

et

$$\|r\rho\|_{L^\infty(E)}\|(r\rho)^{-1}\|_{L^\infty(F)} > 0,$$

car $r \in \mathcal{R}_{n-1}$, ce qui permet d'en déduire l'inégalité

$$\|\rho\|_{L^\infty(E)}\|\rho^{-1}\|_{L^\infty(F)} \geqslant 1.$$

Ce résultat étant valable pour tout $\rho \in \mathcal{R}_1$, cela prouve que $Z_1(E,F) = 1$ et permet de conclure par contraposition. □

On en déduit immédiatement le corollaire suivant.

Corollaire 3.2.13. *Avec les notations précédentes, le problème de minimisation $Z_n(E,F)$ admet une solution dégénérée si et seulement si $Z_1(E,F) = 1$.*

PREUVE : Il suffit de remarquer que la stricte décroissance de la suite considérée à la proposition 3.2.12 est équivalente à la non-existence d'une solution dégénérée pour conclure. □

Remarque. Pour terminer, nous ne disposons à présent d'aucun résultat ou contre-exemple concernant l'unicité éventuelle d'une solution au problème de Zolotarev dans les conditions du lemme 3.2.1.

Nous en sommes à ce sujet réduits à constater que toute fraction rationnelle qui s'annule en zéro est un minimiseur pour $Z_1(\{0\}\{-2,-1\}) = 0$, ce qui ne permet pour le moment que de fournir un exemple trivial de non-unicité de la solution au problème de Zolotarev pour des ensembles discrets.

Les questions de l'existence et de la dégénérescence d'un minimiseur pour le problème de Zolotarev sur des ensembles discrets ayant été étudiées, on cherche maintenant à mettre en évidence de façon théorique la différence de comportement asymptotique observée figures 3.1, 3.2 et 3.3 entre les quantités de Zolotarev pour des ensembles discrets et pour des ensembles continus. On développe dans le chapitre suivant les outils de théorie du potentiel logarithmique adaptés à cet effet.

Chapitre 4

Minimisation d'énergie sous contrainte : cas des mesures signées

En appliquant l'algorithme du gradient conjugué pour la résolution approchée du système linéaire $AX = b$ où A est une matrice symétrique définie positive de taille $N \times N$, on aboutit comme exposé au chapitre 1 à l'étude du problème de min max polynômial $E_n(\Lambda(A))$ défini en (1.4).

L'ensemble $\Lambda(A)$ est habituellement remplacé par son enveloppe convexe dans la littérature, ce qui permet d'en déduire des majorations classiques en terme de conditionnement pour l'erreur relative commise par cet algorithme.

Dans l'article [BeKu01a], on cherche à étudier l'asymptotique de la quantité $E_n(A_N)$ pour une suite (A_N) de matrices définies positives en conservant la nature discrète des ensembles $\Lambda(A_N)$ grâce à un problème de théorie du potentiel logarithmique sous contrainte pour des mesures positives. Cela permet de quantifier le phénomène de convergence superlinéaire de l'algorithme du gradient conjugué.

Comme présenté dans le chapitre 1, l'étude de l'erreur commise lors de l'application de la méthode ADI aboutit à un problème de Zolotarev sur des ensembles discrets, et les majorations habituellement utilisées consistent alors à remplacer les ensembles discrets obtenus par leurs enveloppes convexes. En s'inspirant de l'article [BeKu01a], on cherche maintenant à développer des techniques de théorie du potentiel logarithmique adaptées à l'étude de l'asymptotique de notre problème pour des ensembles discrets.

Dans [BeKu01a], on s'intéresse au problème de minimisation suivant : pour la donnée d'une fonction continue Q appelée champ extérieur, d'une mesure σ appelée contrainte et d'un réel $t \leqslant \sigma(\mathbb{C})$, on cherche à minimiser une quantité dite énergie logarithmique, donnée pour une mesure μ par

$$\int\int \log \frac{1}{|z-u|}\, \mathrm{d}\mu(u)\, \mathrm{d}\mu(z) + 2\int Q(z)\, \mathrm{d}\mu(z)$$

sur un ensemble de mesures μ donné par $\{\mu(\mathbb{C}) = t,\ 0 \leqslant \mu \leqslant \sigma\}$.

De la même façon que l'analogue du problème de min max polynômial E_n est donné dans notre cadre par un problème min max d'approximation rationnelle, le cadre théorique naturel pour l'étude de notre problème est donné par l'étude d'un problème de théorie du potentiel logarithmique sous contrainte pour des mesures signées. Cela nous amène à considérer des généralisations du problème de minimisation énoncé ci-dessus.

En particulier, pour deux mesures σ_1, σ_2 et un couple de réels (t_1, t_2) donné tel que pour $j = 1, 2$, $0 \leqslant t_j \leqslant \sigma_j(\mathbb{C})$, on s'intéresse à la minimisation de l'énergie logarithmique donnée ci-dessus où μ appartient à une classe élargie de mesures candidates décrite par

$$\{\mu := \mu_1 - \mu_2,\ \mu_j \text{ mesure},\ \mu_j(\mathbb{C}) = t_j,\ 0 \leqslant \mu_j \leqslant \sigma_j \text{ pour } j \in \{1, 2\}\}.$$

La théorie du potentiel logarithmique a été étudiée en connexion avec de nombreux domaines, en particulier en théorie de l'approximation ou pour l'étude de l'asymptotique des suites de polynômes orthogonaux -pour ne citer que les domaines les plus proches de notre étude- comme dans le livre de Saff et Totik [SaTo97], mais également avec un point de vue plus centré sur l'analyse complexe à une variable, comme par exemple dans le livre de Ransford [Ra95].

Il apparaît ici nécessaire de mentionner la théorie du potentiel vectorielle où l'on ne considère plus seulement des espaces de mesures ou de mesures signées, mais plutôt un problème de minimisation d'énergie sur un espace produit de mesures avec une matrice d'interaction généralement supposée inversible, point de vue généralement adopté par l'École Russe comme dans le livre de Nikishin et Sorokin [NiSo91].

Notre étude aurait ainsi pu se formuler dans ce cadre vectoriel, même si la matrice d'interaction correspondant à notre contexte égale à

$$\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

n'est pas inversible, d'où un ensemble de difficultés techniques qui expliquent notre parti pris d'une étude construite sur un ensemble de mesures signées sous contrainte qui permet de se soustraire de ce contexte vectoriel, même si celui-ci permettrait très probablement de généraliser notre étude, au moins sur le plan théorique.

Concernant la structure, on suit dans ce chapitre le schéma classique d'un traité comportant une étude de problème de minimisation d'énergie en théorie du potentiel logarithmique : après avoir donné quelques rappels d'analyse complexe, la première partie 4.1 est dédiée à la définition et aux premières propriétés des objets mis en jeu comme le potentiel ou l'énergie logarithmique d'une mesure ainsi qu'à la formulation et l'étude des hypothèses de régularité que l'on souhaite imposer.

On établit ensuite dans la partie 4.2 un résultat d'existence et d'unicité de la solution de notre problème de minimisation. On obtient ce résultat grâce à l'étude de la fonctionnelle énergie dans notre contexte qui présente une difficulté technique par rapport à un contexte classique due au fait que l'on permet ici au support de la mesure contrainte positive et de la mesure contrainte négative d'être d'intersection non vide. On étudie ici cette situation d'une part parce que l'impact numérique de notre étude sera plus particulièrement significatif dans ce cas, et d'autre part parce que celle-ci se produit naturellement dans bien des problèmes d'algèbre linéaire en grande dimension, par exemple dès que l'on cherche à la solution approchée d'une équation de Lyapounov $AX + XA = B$ pour une matrice A définie positive ayant une ou plusieurs valeurs propres très proches de l'origine, ce qui sera détaillé dans le chapitre 8 dédié aux expériences numériques.

Une fois ce résultat d'existence et d'unicité établi, on suit les grandes lignes de la preuve de [DrSa97, Théorème 2.6], pour établir les conditions d'équilibre vérifiées par le minimiseur, conditions d'équilibre qui seront simplifiées grâce aux hypothèses de régularité imposées précédemment et dont on prouvera ensuite comme à l'accoutumée qu'elles caractérisent la mesure minimisante.

On signale enfin que les principaux résultats de ce chapitre seront récapitulés dans le cadre du théorème 4.3.1, ce qui permet au lecteur plus intéressé par les résultats concernant l'asymptotique faible du problème de Zolotarev pour des ensembles discrets de s'appuyer sur ce théorème durant la lecture du chapitre suivant qui lui sera consacré.

Un autre chapitre sera ensuite consacré à une étude en théorie du potentiel logarithmique où l'on tentera d'obtenir le maximum de résultats dans un cas particulier de l'étude menée dans ce chapitre.

4.1 Contexte, première définitions

On consacre la première partie de ce chapitre au rappel et à l'exposé des notions d'analyse complexe nécessaires à la définition du problème de minimisation dont ce chapitre fait l'objet.

4.1.1 Superharmonicité, principe du maximum

On donne ici quelques rappels élémentaires qui seront constamment utilisés dans toute la suite du chapitre de théorie du potentiel logarithmique.

L'objectif de cette section n'est pas d'entrer dans la théorie liée aux notions mises en jeu ici, mais d'énoncer quelques outils importants pour les démonstrations qui suivront, le lecteur familier avec ces notions pourra donc sans dommage passer à la section suivante s'il le souhaite !

Voici pour commencer la définition d'une *fonction semi-continue inférieurement* pour une étude détaillée des propriétés de ces fonctions dans le contexte de la théorie du potentiel logarithmique, on pourra par exemple consulter les livres de Ransford [Ra95, Chapitres 1 et 2] ou de Saff et Totik [SaTo97, Chapitre 0].

Définition 4.1.1. *Soit u une fonction définie sur X un sous-ensemble de $\mathbb{C}$.*

La fonction u est semi-continue inférieurement *si pour tout $z \in X$,*

$$\liminf_{x \to z} u(x) \geqslant u(z).$$

Une fonction v est semi-continue supérieurement *si $-v$ est semi-continue inférieurement.*

On rappelle maintenant la définition d'un domaine et d'une fonction superharmonique sur un domaine de $\mathbb{C}$.

Définition 4.1.2. *Un* domaine *du plan complexe est un sous-ensemble ouvert et connexe de celui-ci.*

Définition 4.1.3. *Une fonction à valeurs dans $(-\infty, +\infty]$ définie sur un domaine D du plan complexe est dite* superharmonique *si elle est semi-continue inférieurement et vérifie la sur-inégalité locale de la moyenne :*

$$f(z) \geqslant \frac{1}{2\pi} \int_{-\pi}^{\pi} f\left(z + r\,\mathrm{e}^{i\theta}\right) \mathrm{d}\theta,$$

pour $z \in D$ et $r > 0$ tel que le disque fermé centré en z de rayon r soit contenu dans D.

De plus, $v : \mathbb{C} \to [-\infty, +\infty)$ est sous-harmonique *si $-v$ est superharmonique.*

On donne enfin une version du principe du minimum adaptée à notre étude (voir [SaTo97][Théorème 0.5.2]).

Théorème 4.1.4 (Principe du minimum)**.** *Soit D un domaine borné du plan complexe et g une fonction superharmonique sur D telle que*

$$\liminf_{z\to z', z\in D} g(z) \geqslant m$$

pour tout $z' \in \partial D$.

Alors, soit g est une constante soit $g(z) > m$ pour $z \in D$.

Remarque. En changeant g en $-g$ dans le théorème précédent, on en déduit immédiatemment le *principe du maximum*, valable pour des fonctions sous-harmoniques.

4.1.2 Potentiel et énergie logarithmique

On donne maintenant les définitions des objets qui seront utilisés dans ce chapitre ainsi que les hypothèses de régularité que l'on souhaite imposer dans notre contexte.

Dans tout ce qui suit, on appellera *mesure* une mesure de Borel positive finie à support compact dans le plan complexe et *mesure signée* une différence de deux telles mesures.

On utilise par ailleurs le mot *positif* au sens de *positif ou nul*.

Si S est un sous-ensemble compact du plan complexe $\mathbb{C}$, on note $\mathcal{M}(S)$ l'ensemble des mesures signées sur S et $\mathcal{M}^+(S)$ l'ensemble des mesures sur S.

On reprend ci-dessous les notations classiques de [Ru75, Chapitre 6] à propos des mesures signées.

Définition 4.1.5. *Pour μ dans $\mathcal{M}(S)$, on note $|\mu|$ la mesure de variation totale de μ, et $\mu = \mu_1 - \mu_2$ sa* décomposition de Jordan*, définie par*

$$\mu_1 := \frac{|\mu| + \mu}{2} \text{ et } \mu_2 := \frac{|\mu| - \mu}{2}$$

Les deux objets principaux utilisés en théorie du potentiel sont respectivement l'énergie et le potentiel logarithmique que l'on définit ci-dessous. L'énergie et le potentiel seront reliés par la suite à l'asymptotique du problème de Zolotarev sur des ensembles discrets, et permettront en outre de se forger une intuition physique de la situation.

Définition 4.1.6. *On définit le* potentiel logarithmique *d'une mesure ρ de $\mathcal{M}^+(S)$:*

$$U^\rho(z) := \int \log \frac{1}{|z-t|} \, \mathrm{d}\rho(t),$$

Si μ appartient à $\mathcal{M}(S)$, on définit le potentiel logarithmique *de la mesure signée μ par*

$$U^\mu(z) := U^{\mu_1}(z) - U^{\mu_2}(z)$$

défini pour z tel que $U^{\mu_1}(z) \neq +\infty$ ou $U^{\mu_2}(z) \neq +\infty$.

Proposition 4.1.7. *Soit $\rho \in \mathcal{M}^+(S)$, la fonction U^ρ est superharmonique sur $\mathbb{C}$, sous-harmonique sur $\mathbb{C} \setminus \mathrm{supp}\,(\rho)$ et à valeurs dans $(-\infty, +\infty]$,.*

PREUVE : Voir [NiSo91, Théorème 5.2.1], [Ra95, Théorème 3.1.2] ou encore [SaTo97, Théorème 0.5.6]. □

Définition 4.1.8. *Pour une fonction Q continue sur $\mathbb{C}$, on définit comme l'*énergie logarithmique *avec champ extérieur Q d'une mesure $\rho \in \mathcal{M}^+(S)$:*

$$I_Q(\rho) := \int\int \log \frac{1}{|z-t|} \,\mathrm{d}\rho(t)\,\mathrm{d}\rho(z) + 2\int Q(z)\,\mathrm{d}\rho(z).$$

Lorsque la fonction Q est nulle, on note I en lieu et place de I_0 pour la fonctionnelle énergie.

On définit alors l'énergie mutuelle d'un couple de mesures (ρ_1, ρ_2) de $\mathcal{M}^+(S)$:

$$I(\rho_1, \rho_2) := \int\int \log \frac{1}{|z-t|} d\rho_1(t)\,\mathrm{d}\rho_2(z).$$

La mesure signée μ de $\mathcal{M}(S)$ est dite d'énergie finie *dès lors que la quantité $I(|\mu|)$ est finie.*

Pour $\nu = \nu_1 - \nu_2$ et $\mu = \mu_1 - \mu_2$ dans $\mathcal{M}(S)$ deux mesures signées d'énergie finie, on définit l'énergie logarithmique de μ avec champ extérieur Q

$$I_Q(\mu) := I(\mu_1, \mu_1) - 2I(\mu_1, \mu_2) + I(\mu_2, \mu_2) + 2\int Q(z)\,\mathrm{d}\mu(z)$$

et l'énergie mutuelle des mesures signées μ et ν

$$I(\mu, \nu) := I(\mu_1, \nu_1) - I(\mu_1, \nu_2) - I(\mu_2, \nu_1) + I(\mu_2, \nu_2).$$

Après avoir défini le potentiel logarithmique et l'énergie logarithmique, on peut maintenant définir la capacité d'un ensemble du plan complexe.

La capacité logarithmique d'un sous-ensemble du plan complexe est un autre objet de base de la théorie du potentiel logarithmique, et on peut estimer à bien des égards que les ensembles de capacité nulle jouent en théorie du potentiel logarithmique un rôle comparable aux ensembles de mesure nulle dans la théorie de l'intégration.

Définition 4.1.9. *Pour E un sous-ensemble du plan complexe, on note* $\mathrm{cap}\,(E)$ *sa* capacité logarithmique*, définie par*

$$\mathrm{cap}\,(E) := \exp\left(-\inf\{I(\mu),\ \mu \text{ mesure de probabilité telle que } \mathrm{supp}\,(\mu) \subset E\}\right).$$

On dit qu'une propriété donnée a lieu quasi-partout *-que l'on abrégera par* qp- *si celle-ci a lieu partout sauf sur un ensemble de capacité logarithmique nulle.*

On adopte dans la définition précédente la convention $\exp(-\infty) = 0$, et on peut par exemple citer la proposition suivante depuis [Ra95, Théorème 3.2.4].

Proposition 4.1.10. *Tout sous-ensemble du plan complexe de mesure de Lebesgue sur $\mathbb{C}$ strictement positive est également de capacité logarithmique strictement positive, ou de façon équivalente, un évènement ayant lieu quasi-partout a lieu presque-partout.*

Le même résultat est valable sur la droite réelle munie de la mesure de Lebesgue sur $\mathbb{R}$.

Il existe beaucoup de résultats concernant les propriétés topologiques des ensembles de capacité logarithmique nulle, par exemple que ceux-ci sont de dimension de Hausdorff nulle, voir [Ra95, p 56,57], mais cela dépasse le cadre de notre travail.

On rappelle ici la convention classique concernant l'arithmétique dans $(0, +\infty)$ dans le but de pouvoir étendre nos résultats à des mesures éventuellement infinies sur une partie de leur support par la suite.

Cette extension théorique permettra de modéliser via de la théorie du potentiel logarithmique un ensemble plus large de situations, typiquement d'accepter un problème de minimisation d'énergie partiellement contraint, situation naturelle dans certains cas par la suite.

Remarque. Dans tout ce qui suit, on adopte les conventions de [Ru75, Chapitre 1] concernant l'arithmétique dans $[0, +\infty]$, à savoir

$$a + \infty = \infty + a = \infty, \text{ si } 0 \leqslant a \leqslant \infty$$

et

$$a.\infty = \infty.a = \begin{cases} \infty & \text{si } 0 < a \leqslant \infty, \\ 0 & \text{si } a = 0 \end{cases}$$

ce qui permet de conserver la commutativité, l'associativité et la distributivité des lois sur $[0, \infty]$ et ainsi de considérer des mesures prenant la valeur ∞ sur une partie de leur support. En l'absence d'ambiguïté, en notera indifféremment ∞ et $+\infty$.

Définition 4.1.11. *On définit maintenant σ_1 et σ_2 deux mesures du plan complexe, on note pour $j \in \{1, 2\}$, $\Sigma_j := \text{supp}\,(\sigma_j)$ et E_j le sous-ensemble maximal de Σ_j sur lequel cette mesure est infinie.*

On notera enfin pour $j = 1, 2$, $E_j^c := \Sigma_j \setminus E_j$, on a donc

$$\begin{cases} \sigma_j|_{E_j} & = \infty, \\ \sigma_j|_{E_j^c} & \textit{est une mesure finie.} \end{cases}$$

On attire ici l'attention du lecteur sur le fait suivant : l'ensemble E_j^c n'a *a priori* pas de propriété topologique de type ouvert ou fermé, contrairement à Σ_j qui est un sous-ensemble compact du plan complexe, mais le support de la mesure $\sigma_j|_{E_j^c}$ est par définition fermé et égal à l'adhérence de E_j^c.

On s'attachera dorénavant dans un but de clarification des notations à utiliser l'indice 1 (resp. 2) pour désigner la partie positive (resp. négative) des mesures étudiées.

Hypothèse 4.1. *On suppose les ensembles Σ_j compacts et de capacité strictement positive pour $j = 1, 2$.*

On suppose en outre que $\text{dist}\,(E_1, \Sigma_2) > 0$, $\text{dist}\,(E_2, \Sigma_1) > 0$, *et que l'ensemble $E_1^c \cap E_2^c$ est de capacité logarithmique nulle.*

On suppose enfin que pour $j = 1, 2$, toutes les composantes connexes de $\mathbb{C} \setminus E_j$ sont régulières par rapport au problème de Dirichlet.

Cette hypothèse technique permettra de définir le problème de minimisation d'énergie logarithmique dans sa forme la plus générale, avec des mesures partiellement contraintes, la contrainte σ_j n'étant en effet active que sur l'ensemble E_j^c.

On permet ici aux ensemble E_1^c et E_2^c sur lesquels vit respectivement la partie positive et la partie négative de la contrainte considérée d'être d'intersection non vide, ce qui ajoutera

quelques points de difficulté technique dans les démonstrations de théorie du potentiel. Cependant, cela permettra par la suite, en particulier dans le chapitre 8, de considérer des applications pour lesquelles l'impact de notre étude sera plus significatif, typiquement le cas de deux intervalles symétriques de l'axe réel $supp\,(\sigma_1) = -supp\,(\sigma_2) = [0, A]$.

Les détails techniques concernant la définition de la régularité d'un ensemble du plan complexe par rapport au problème de Dirichlet seront rappelés à la proposition 4.2.8.

On rappelle simplement ici qu'il s'agit d'un critère de régularité local vérifié par un domaine G du plan complexe si et seulement s'il est vérifié par tous les points de sa frontière.

Le résultat suivant tiré de [Ra95, Théorème 4.2.1] donne une large catégorie d'ensembles réguliers par rapport au problème de Dirichlet.

Proposition 4.1.12. *Un domaine simplement connexe tel que $\mathbb{C}_\infty \setminus G$ contienne au moins deux points est régulier par rapport au problème de Dirichlet.*

Définition 4.1.13. *Dans le cas où σ_1 et σ_2 sont des mesures finies, on note $\sigma := \sigma_1 - \sigma_2$ que l'on appelle* mesure contrainte signée.

Pour la majeure partie des applications dans ce travail, les mesures σ_1 et σ_2 considérées seront finies, le lecteur pourra donc considérer des mesures σ_1 et σ_2 finies s'il le souhaite.

On fait maintenant une hypothèse de régularité sur les potentiels logarithmiques des mesures restreintes $\sigma_1|_{E_1^c}$ et $\sigma_2|_{E_2^c}$. Cette hypothèse permettra d'en déduire des propriétés de régularité à propos de la mesure d'équilibre du problème de minimisation d'énergie que l'on considérera, c'est pourquoi l'hypothèse suivante, même si elle porte sur des objets de nature différente, joue un rôle proche de l'hypothèse de régularité par rapport au problème de Dirichlet faite plus haut.

Hypothèse 4.2. *Pour $j \in \{1, 2\}$, on suppose que $\sigma_j|_{E_j^c}$ est une mesure telle que le potentiel logarithmique $U^{\sigma_j|_{E_j^c}}$ soit continu, et on suppose également le champ extérieur Q continu.*

Du point de vue de l'interprétation électrostatique de notre problème, l'indice 1 désignera ainsi dorénavant le plateau chargé positivement du condensateur étudié et l'indice 2 la plateau chargé négativement.

Remarque. L'hypothèse 4.2 n'est pas très restrictive. En effet, si par exemple les mesures $\sigma_1|_{E_1^c}$ et $\sigma_2|_{E_2^c}$ sont absolument continues par rapport à la mesure de Lebesgue du plan complexe et que leurs densités ne comportent qu'un nombre fini de singularités de type logarithmique ou puissance, alors l'hypothèse 4.2 est satisfaite. Cependant, l'hypothèse 4.2 n'est plus vérifiée dès lors que les mesures $\sigma_1|_{E_1^c}$ ou $\sigma_2|_{E_2^c}$ ont des atomes.

Définition 4.1.14. *On définit pour t_1, t_2, t réels positifs les ensembles*

$$\mathcal{M}_{\sigma_1,\sigma_2}^{t_1,t_2} := \{\mu := \mu_1 - \mu_2,\ \mu_j \text{ mesure},\ \mu_j(\mathbb{C}) = t_j, \\ 0 \leqslant \mu_j \leqslant \sigma_j \text{ pour } j \in \{1, 2\}\},$$

et dans le cas $t_1 = t_2 = t$ et σ_1 et σ_2 deux mesures finies, on note

$$\mathcal{M}_\sigma^t := \{\mu := \mu_1 - \mu_2,\ \mu_j \text{ mesure},\ \mu_j(\mathbb{C}) = t,\ 0 \leqslant \mu_j \leqslant \sigma_j \text{ pour } j \in \{1, 2\}\}.$$

qui constitueront par la suite des ensembles de mesures candidates pour certains problèmes de minimisation d'énergie sous contrainte.

L'ensemble $\mathcal{M}^{t_1,t_2}_{\sigma_1,\sigma_2}$ est formé de mesures signées sous les contraintes $\sigma_j,\ j=1,2$, contraintes effectives seulement sur les ensembles E_j^c. On note que toute mesure μ de $\mathcal{M}^{t_1,t_2}_{\sigma_1,\sigma_2}$ admet une unique décomposition de Jordan sous la forme $\mu_1-\mu_2$, notation que l'on adopte dès à présent, et on a de plus pour $j\in\{1,2\}$,

$$\mu_j=\mu_j|_{E_j}+\mu_j|_{E_j^c}.$$

On s'intéresse dans le lemme suivant aux propriétés topologiques de ces ensembles de candidats au sens de la topologie faible-$\star$.

Lemme 4.1.15. *Sous les hypothèses 4.2, l'ensemble $\mathcal{M}^{t_1,t_2}_{\sigma_1,\sigma_2}$ est fermé pour la topologie faible-$\star$ dans l'ensemble $\{\mu:=\mu_1-\mu_2,\ \mu_j$ mesure, $\mu_j(\mathbb{C})=t_j$ pour $j\in\{1,2\}\}$.*

PREUVE : Soient $(\mu^k)_k$ une suite d'éléments de $\mathcal{M}^{t_1,t_2}_{\sigma_1,\sigma_2}$, de limite faible μ et f une fonction continue positive à support compact sur $\mathbb{C}$: alors, on a pour tout $k\geqslant 1$ et pour $j\in\{1,2\}$

$$\int_{E_j^c} f(z)\,\mathrm{d}\mu_j^k(z)\leqslant\int_{E_j^c} f(z)\,\mathrm{d}\sigma_j(z),$$

d'où

$$\int_{E_j^c} f(z)\,\mathrm{d}\mu_j(z)\leqslant\int_{E_j^c} f(z)\,\mathrm{d}\sigma_j(z),$$

et la mesure $\sigma_j-\mu_j$ est donc positive sur E_j^c, ce qui termine notre preuve. □

Remarque. On démontre de même que l'ensemble $\mathcal{M}^t_\sigma$ est faiblement fermé.

Comme remarqué précedemment, il est nécessaire de définir avec soin l'expression du potentiel logarithmique et de l'énergie logarithmique d'une mesure signée, par exemple pour une mesure de l'ensemble $\mathcal{M}^{t_1,t_2}_{\sigma_1,\sigma_2}$.

Dans le cas d'une mesure $\mu\in\mathcal{M}^t_\sigma$, on a d'après le théorème de Fubini l'expression

$$I(\mu):=\int U^\mu(z)\,\mathrm{d}\mu(z),$$

où U^μ est continue et $supp\,(\mu)$ compact, ce qui prouve que l'énergie logarithmique de tout élément de $\mathcal{M}^t_\sigma$ est finie.

Les principaux objets ayant été définis nous pouvons passer aux premiers résultats de théorie du potentiel dans la section suivante.

4.2 Minimisation d'énergie sous contrainte

4.2.1 Semi-continuité inférieure de l'énergie

On s'attache dans cette section à établir la semi-continuité inférieure de la fonctionnelle énergie dans notre contexte, propriété qui garantira l'existence d'un minimiseur pour notre problème de minimisation d'énergie.

Cette propriété est bien connue et est rappelée ci-dessous dans le cadre d'un problème de minimisation d'énergie pour des mesures de probabilité sur un ensemble S, voir [Ra95, Lemme 3.3.3], la démonstration étant identique pour $\{\mu\in\mathcal{M}(S),\ \mu\leqslant\sigma\}$.

Proposition 4.2.1. *Soient S un compact et σ une mesure sur S : la fonctionnelle énergie est semi-continue inférieurement au sens de la topologie associée à la convergence faible-$\star$ sur l'ensemble des mesures de probabilité à support contenu dans S, $\{\mu \in \mathcal{M}(S),\ \mu(\mathbb{C}) = 1\}$ ainsi que sur $\{\mu \in \mathcal{M}(S),\ \mu \leqslant \sigma\}$.*

Cependant, une attention toute particulière doit ici être accordée au fait que l'ensemble $E_1^c \cap E_2^c$ n'est pas supposé vide dans l'hypothèse 4.1, mais seulement de capacité logarithmique nulle, ce qui induit une difficulté technique dans la preuve de la semi-continuité inférieure de la fonctionnelle énergie I faite ci-dessous.

On cherche donc ici à suivre le schéma de la preuve de [Ra95, Lemme 3.3.3] tout en prenant bien soin de lever la difficulté technique induite par notre généralisation, d'où le présence du lemme technique suivant.

Lemme 4.2.2. *Soit pour $0 < \delta < \frac{1}{2}$*

$$U_\delta := \{(z,t) \in \Sigma_1 \times \Sigma_2,\ |z-t| < \delta\}.$$

On a alors

$$\lim_{\delta \to 0} \int \int_{U_\delta} \log \frac{1}{|z-t|}\, \mathrm{d}\sigma_1(t)\, \mathrm{d}\sigma_2(z) = 0. \tag{4.1}$$

PREUVE : D'après l'hypothèse 4.1, $dist\,(\Sigma_1, E_2) > 0$ et $dist\,(\Sigma_2, E_1) > 0$ d'où pour δ assez petit, on a

$$U_\delta := \{(z,t) \in E_1^c \times E_2^c,\ |z-t| < \delta\},$$

ce qui prouve que l'intégrale considérée est bien définie pour δ suffisamment petit.

Pour prouver (4.1), remarquons que la fonction $(t,z) \to \log \frac{1}{|z-t|}$ est positive sur $U_{1/2}$, et d'après le théorème de Fubini applicable ici car pour $\delta > 0$ assez petit les quantités sont de signe constant, l'intégrale

$$\int \int \log \frac{1}{|z-t|}\, \mathrm{d}\sigma_1(t)\, \mathrm{d}\sigma_2(z) = \int_{\mathbb{C}} U^{\sigma_1}(z)\, \mathrm{d}\sigma_2(z)$$

est finie.

De plus, pour tout $\delta \in \left(0, \frac{1}{2}\right)$,

$$0 \leqslant \mathbb{1}_{U_\delta}(t,z) \log \frac{1}{|z-t|} \leqslant \mathbb{1}_{U_{1/2}}(t,z) \log \frac{1}{|z-t|},$$

et la famille d'ensembles U_δ est décroissante pour l'inclusion avec

$$U := \bigcap_{\delta > 0} U_\delta = \{(x,x), x \in E_1^c \cap E_2^c\}.$$

D'après l'hypothèse 4.2, l'ensemble U est de capacité nulle, et comme U^{σ_j} est continu pour $j = 1, 2$, par théorème de convergence monotone,

$$\begin{aligned} \lim_{\delta \to 0} \int_{U_\delta} \log \frac{1}{|z-t|}\, \mathrm{d}\sigma_1(t)\, \mathrm{d}\sigma_2(z) &= \lim_{\delta \to 0} \int_{U_{1/2}} \mathbb{1}_{U_\delta}(t,z) \log \frac{1}{|z-t|}\, \mathrm{d}\sigma_1(t)\, \mathrm{d}\sigma_2(z) \\ &= \int_U \mathbb{1}_U(t,z) \log \frac{1}{|z-t|}\, \mathrm{d}\sigma_1(t)\, \mathrm{d}\sigma_2(z) = 0, \end{aligned}$$

ce qui prouve bien l'assertion (4.1) □

Nous pouvons maintenant passer à la preuve de la proposition motivant cette section.

Proposition 4.2.3. *La fonctionnelle d'énergie I_Q définie sur $\mathcal{M}^{t_1,t_2}_{\sigma_1,\sigma_2}$ à valeurs dans $(-\infty,+\infty]$ est semi-continue inférieurement au sens de la topologie associée à la convergence faible-$\star$.*

PREUVE : Soit $(\mu^k)_{k\geqslant 0}$ une suite d'éléments de $\mathcal{M}^{t_1,t_2}_{\sigma_1,\sigma_2}$ de limite faible μ dont la décomposition de Jordan de chaque élément est donnée par $\mu^k = \mu_1^k - \mu_2^k$.

Quitte à extraire une sous-suite convergente de $(\mu_1^k)_{k\geqslant 0}$ (resp. $(\mu_2^k)_{k\geqslant 0}$), on peut supposer que celle-ci admet une limite faible μ_1 (resp. μ_2), et de plus, $\mu = \mu_1 - \mu_2$ appartient à $\mathcal{M}^{t_1,t_2}_{\sigma_1,\sigma_2}$ qui est fermé pour la topologie faible-$\star$ d'après le lemme 4.1.14.

Le champ extérieur Q étant supposé continu sur $\Sigma_1 \cup \Sigma_2$, on a

$$\liminf_{k\to+\infty} \int Q(z)\,\mathrm{d}\mu^k(z) = \lim_{k\to+\infty} \int Q(z)\,\mathrm{d}\mu^k(z) = \int Q(z)\,\mathrm{d}\mu(z),$$

et il suffit donc de prouver la semi-continuité inférieure de la fonctionnelle énergie sans champ extérieur, notée I.

D'après la semi-continuité inférieure de la fonctionnelle énergie dans le cas des mesures, on a

$$\liminf_{k\to\infty} I(\mu^k) - I(\mu) \geqslant 2\Big(I(\mu_1,\mu_2) - \limsup_{k\to\infty} I(\mu_1^k,\mu_2^k)\Big).$$

On écrit alors avec les notations précédentes pour $\delta > 0$,

$$I(\mu_1^k,\mu_2^k) = \int_{U_\delta} \log\frac{1}{|z-t|}\,\mathrm{d}\mu_1^k(t)d\mu_2^k(z) + \int_{\mathbb{C}\setminus U_\delta} \log\frac{1}{|z-t|}\,\mathrm{d}\mu_1^k(t)\,\mathrm{d}\mu_2^k(z)$$

et par continuité de l'application $(z,t)\mapsto \log\frac{1}{|z-t|}$ restreinte à $\mathbb{C}\setminus U_\delta$, on en déduit que

$$\lim_{k\to+\infty} \int_{\mathbb{C}\setminus U_\delta} \log\frac{1}{|z-t|}\,\mathrm{d}\mu_1^k(t)\,\mathrm{d}\mu_2^k(z) = \int_{\mathbb{C}\setminus U_\delta} \log\frac{1}{|z-t|}\,\mathrm{d}\mu_1(t)\,\mathrm{d}\mu_2(z),$$

et le lemme 4.2.2 nous permet alors de conclure que

$$\lim_{k\to+\infty} I(\mu_1^k,\mu_2^k) = I(\mu_1,\mu_2),$$

d'où le résultat. □

La semi-continuité inférieure de la fonctionnelle énergie est un outil essentiel pour la démonstration de l'existence d'un minimiseur pour le problème de minimisation d'énergie sous contrainte considéré.

4.2.2 Problème de minimisation, existence et unicité du minimiseur

On applique dans cette section des techniques classiques de minimisation d'une forme quadratique sur un ensemble convexe : voir par exemple [SaTo97, Théorème VIII.1.4] pour le cas des mesures signées sans contrainte avec champ extérieur avec lequel on fera le lien dans le chapitre 6, ou encore [NiSo91, Chapitre 5 Paragraphe 4] pour un langage plus proche des techniques hilbertiennes.

On considère le problème de minimisation d'énergie sous contrainte suivant :

$$(P):\ \text{Trouver } \mu^{t_1,t_2} \text{ telle que} \qquad I_Q(\mu^{t_1,t_2}) = \inf\left\{I_Q(\mu),\ \mu\in\mathcal{M}^{t_1,t_2}_{\sigma_1,\sigma_2}\right\},$$

où l'ensemble des mesures candidates $\mathcal{M}^{t_1,t_2}_{\sigma_1,\sigma_2}$ a été défini en 4.1.14.

Ce problème correspond à la formalisation précise du problème exposé dans l'introduction de ce chapitre. On remarque que pour $t_2 = 0$, on retrouve bien un problème de minimisation pour des mesures positives, problème plus classique en théorie du potentiel logarithmique.

Lemme 4.2.4. *Soit* $V_{\sigma,Q} := \inf\left(I_Q(\mu),\ \mu \in \mathcal{M}^{t_1,t_2}_{\sigma_1,\sigma_2}\right)$.

La quantité $V_{\sigma,Q}$ *est finie.*

PREUVE : Le champ extérieur Q étant continu, on a pour toute mesure μ de $\mathcal{M}^{t_1,t_2}_{\sigma_1,\sigma_2}$

$$\left|\int_{\mathbb{C}} Q(z)\,\mathrm{d}\mu(z)\right| \leqslant (t_1+t_2)\max_{z\in\Sigma_1\cup\Sigma_2}|Q(z)|\,,$$

il suffit donc de prouver le résultat pour la fonctionnelle énergie associée à un champ extérieur nul.

Exhibons maintenant une mesure à énergie finie de $\mathcal{M}^{t_1,t_2}_{\sigma_1,\sigma_2}$: comme pour $j \in \{1,2\}$, Σ_j est supposé de capacité logarithmique strictement positive, il existe une mesure à énergie logarithmique finie μ_j dont le support est contenu dans Σ_j telle que $\mu = \mu_1 - \mu_2$ soit dans $\mathcal{M}^{t_1,t_2}_{\sigma_1,\sigma_2}$.

Il suffit alors par définition de l'énergie logarithmique d'une mesure signée de montrer que le terme $I(\mu_1,\mu_2)$ est fini pour conclure, et en écrivant la décomposition

$$I(\mu_1,\mu_2) = \int\int_{U_\delta} \log\frac{1}{|z-t|}\,\mathrm{d}\mu_1(t)\,\mathrm{d}\mu_2(z) + \int\int_{\mathbb{C}\setminus U_\delta} \log\frac{1}{|z-t|}\,\mathrm{d}\mu_1(t)\,\mathrm{d}\mu_2(z),$$

on obtient que la quantité

$$\int\int_{U_\delta} \log\frac{1}{|z-t|}\,\mathrm{d}\mu_1(t)\,\mathrm{d}\mu_2(z)$$

est finie pour δ suffisamment petit d'après le lemme 4.2.2, et $(z,t) \mapsto \log\frac{1}{|z-t|}$ est continue sur $\mathbb{C}\setminus U_\delta$, d'où

$$\int\int_{\mathbb{C}\setminus U_\delta} \log\frac{1}{|z-t|}\,\mathrm{d}\mu_1(t)\,\mathrm{d}\mu_2(z) < \infty,$$

ce qui prouve bien l'existence d'une telle mesure.

Enfin, le noyau logarithmique $(z,t) \mapsto \log\frac{1}{|z-t|}$ est borné inférieurement sur $\Sigma_1\times\Sigma_2$, ce qui conclut la preuve de la finitude de $V_{\sigma,Q}$. □

On cite le lemme [SaTo97, Lemme I.1.8] valable dans le cas particulier des mesures de masse totale nulle que nous utiliserons à plusieurs reprises, notamment pour établir l'unicité du minimiseur.

Lemme 4.2.5 (Lemme 1.1.8, [SaTo97])**.** *Soit* $\mu = \mu_1 - \mu_2$ *une mesure signée à support compact de masse totale nulle telle que les mesures* μ_1 *et* μ_2 *soient d'énergie logarithmique finie.*

Alors, l'énergie logarithmique de μ *est positive :*

$$I(\mu) = \int\int \log\frac{1}{|z-t|}\,\mathrm{d}\mu(t)\,\mathrm{d}\mu(z) \geqslant 0$$

et celle-ci est nulle si et seulement si $\mu = 0$.

Voici maintenant le théorème de Helly (voir [SaTo97, Théorème 0.1.3]), théorème utile dans la démonstration de l'existence d'une mesure minimisante.

Théorème 4.2.6 (Théorème de Helly)**.** *Soient S un compact de $\mathbb{C}$ et $(\mu_n)_n$ une suite de mesures à valeur complexes à support contenu dans S de masse totale $\sup_{n\geqslant 1}|\mu_n|(\mathbb{C})$ bornée, alors on peut extraire de la suite $(\mu_n)_n$ une sous-suite convergente au sens de la topologie faible-$\star$.*

La proposition suivante donne le résultat attendu d'existence et d'unicité de la mesure minimisante en reprenant la preuve de [SaTo97, Théorème I.1.3], le point technique ayant été réglé au paragraphe précédent.

Proposition 4.2.7. *Il existe une unique mesure μ^{t_1,t_2} dans $\mathcal{M}^{t_1,t_2}_{\sigma_1,\sigma_2}$ telle que*

$$I_Q\left(\mu^{t_1,t_2}\right) = V_{\sigma,Q}.$$

PREUVE : Soit $(\mu^k)_{k\geqslant 0}$ une suite de mesures de $\mathcal{M}^{t_1,t_2}_{\sigma_1,\sigma_2}$ telle que $\lim_{k\to+\infty} I_Q(\mu^k) = V_{\sigma,Q}$.

Quitte à extraire une sous-suite, on peut d'après le théorème de Helly 4.2.6 supposer que la suite $(\mu_1^k)_{k\geqslant 0}$ (resp. $(\mu_2^k)_{k\geqslant 0}$) converge faiblement vers une mesure μ_1 (resp. μ_2).

La mesure $\mu := \mu_1 - \mu_2$ appartient à l'ensemble $\mathcal{M}^{t_1,t_2}_{\sigma_1,\sigma_2}$ d'après le lemme 4.1.15 et par semi-continuité inférieure de la fonctionnelle énergie (proposition 4.2.3),

$$I_Q(\mu) \leqslant \liminf_{k\to+\infty}\left(I_Q(\mu_1^k - \mu_2^k)\right),$$

et nécessairement

$$I_Q(\mu) = V_{\sigma,Q}.$$

Supposons maintenant qu'il existe deux mesures μ et ν de $\mathcal{M}^{t_1,t_2}_{\sigma_1,\sigma_2}$ qui réalisent l'infimum :

$$I_Q(\mu) = I_Q(\nu) = V_{\sigma,Q}.$$

Alors,

$$V_{\sigma,Q} = \frac{1}{2}(I_Q(\mu) + I_Q(\nu)), \tag{4.2}$$

$$= I_Q\left(\frac{\mu+\nu}{2}\right) + I\left(\frac{\mu-\nu}{2}\right) \tag{4.3}$$

et la mesure $(\mu+\nu)/2$ appartient à $\mathcal{M}^{t_1,t_2}_{\sigma_1,\sigma_2}$ qui est convexe.

De plus, $\mu-\nu$ est une mesure de Borel de masse totale nulle et d'énergie finie, donc d'après le lemme 4.2.5, la nullité de $I(\mu-\nu)$ nous assure que $\mu=\nu$. □

On appellera indifféremment *mesure d'équilibre* ou *minimiseur* pour le problème (P) la mesure μ^{t_1,t_2} dont l'existence et l'unicité viennent d'être établies.

4.2.3 Régularité du potentiel logarithmique

On s'intéresse dans cette section à la régularité du potentiel logarithmique

Pour donner quelques précisions techniques plus détaillées à propos du problème de Dirichlet, on s'inspire ici de [SaTo97, Annexe A2].

Théorème 4.2.8. *Soient G un domaine borné du plan complexe et f une fonction continue définie sur ∂G.*

On définit alors

$$\mathcal{H}_f^{u,G} := \{g,\ g \text{ superharmonique et bornée inférieurement sur } G,\ \liminf_{z\to x, z\in G} g(z) \geqslant f(x),\ x \in \partial G\},$$

$$\mathcal{H}_f^{l,G} := \{g,\ g \text{ sousharmonique et bornée supérieurement sur } G,\ \limsup_{z\to x, z\in G} g(z) \leqslant f(x),\ x \in \partial G\},$$

$$\overline{H}_f^G(z) := \inf\left\{g(z),\ g \in \mathcal{H}_f^{u,G}\right\}$$

et enfin

$$\underline{H}_f^G(z) := \sup\left\{g(z),\ g \in \mathcal{H}_f^{l,G}\right\}.$$

Alors,

$$\overline{H}_f^G = \underline{H}_f^G := H_f^G,$$

et cette fonction est appelée solution de Perron-Wiener-Brelot pour le problème de Dirichlet sur G pour la fonction f.

De plus, H_f^G est harmonique sur G et pour quasi-tous les points x de ∂G,

$$\lim_{z\to x} H_f^G(z) = f(x).$$

Ce formalisme montre comme énoncé précédemment que le critère de régularité pour le problème de Dirichlet est bien un critère local et que seul un ensemble exceptionnel de points du bord d'un domaine peut éventuellement ne pas être régulier au sens du problème de Dirichlet.

Définition 4.2.9. *Avec les notations du théorème précédent, on dit que $x \in \partial G$ est* régulier par rapport au problème de Dirichlet *si pour toute fonction continue sur ∂G,*

$$\lim_{z\to x} H_f^G(z) = f(x),$$

et on dit qu'un ensemble est régulier par rapport au problème de Dirichlet *si tous les points de sa frontière le sont.*

L'hypothèse de régularité 4.1 est utile ici dans le but de simplifier un résultat ultérieur important, à savoir les conditions d'équilibre qui seront vérifiées par la mesure minimisante, grâce au fait suivant, extrait de [SaTo97, Théorème I.5.1] qui assure la régularité du potentiel sous certaines conditions.

Théorème 4.2.10. *Soit Q un champ extérieur continu et $\Sigma \subset \mathbb{C}$ tel que toutes les composantes de $\Sigma \setminus \mathbb{C}$ soient régulières pour le problème de Dirichlet.*

Alors, si pour $s > 0$ on note μ la mesure d'équilibre pour le problème de minimisation d'énergie logarithmique I_Q sur l'ensemble $\{\nu \in \mathcal{M}^+(\Sigma),\ \nu(\mathbb{C}) = s\}$, le potentiel U^μ est continu sur $\mathbb{C}$.

On connaît d'après la proposition 4.1.7 la continuité du potentiel logarithmique d'une mesure de Borel ρ hors de son support, mais il est souvent agréable de travailler dans un contexte permettant d'assurer la continuité des potentiels considérés sur tout le plan complexe par souci de simplification des propositions obtenues.

Le lemme suivant, dit *lemme de Rakhmanov*, donne un cas de conservation de la continuité du potentiel logarithmique particulièrement utile dans le cadre de notre étude sous contrainte.

De la même façon que les différentes hypothèses de régularité faites précédemment 4.2 et 4.1, ce lemme est à mettre en relation avec le théorème 4.2.10 en tant que moyen d'affirmer une propriété de régularité à propos de la mesure minimisante pour le problème de minimisation (P).

Lemme 4.2.11. *Soit $\mu = \mu_1 - \mu_2$ dans $\mathcal{M}^{t_1,t_2}_{\sigma_1,\sigma_2}$. Alors, les potentiels $U^{\mu_1|_{E_1^c}}$ et $U^{\mu_2|_{E_1^c}}$ sont continus.*

PREUVE : En effet, $U^{\mu_1|_{E_1^c}}$ est semi-continue inférieurement, et comme

$$U^{\mu_1|_{E_1^c}} = U^{\sigma_1|_{E_1^c}} - U^{\sigma_1|_{E_1^c}-\mu_1|_{E_1^c}}$$

où $U^{\sigma_1|_{E_1^c}}$ est par hypothèse continue et $U^{\sigma_1|_{E_1^c}-\mu_1|_{E_1^c}}$ semi-continue inférieurement, $U^{\mu_1|_{E_1^c}}$ est également semi-continue supérieurement, donc continue. Ce résultat se prouve de la même façon pour le potentiel $U^{\mu_2|_{E_2^c}}$. □

4.2.4 Conditions d'équilibre pour le minimiseur

Nous allons maintenant en généralisant le raisonnement de [DrSa97, Théorème 2.6] établir les conditions d'équilibre vérifiées par la mesure minimisante pour le problème (P), avant de montrer dans la prochaine section que celles-ci caractérisent la mesure d'équilibre pour (P).

À la différence de la situation considérée dans [DrSa97], on considère ici des mesures signées, mais l'essentiel du raisonnement reste identique.

Les conditions d'équilibre énoncées ci-dessous en (4.4) seront par la suite simplifiées en (4.9) grâce aux hypothèses de régularité 4.2 et 4.1 que l'on impose sur les ensembles considérés ainsi que sur la mesure contrainte.

Proposition 4.2.12. *Sous l'hypothèse 4.2, le minimiseur $\mu^{t_1,t_2} \in \mathcal{M}^{t_1,t_2}_{\sigma_1,\sigma_2}$ vérifie les conditions d'équilibre*

$$\begin{cases} U^{\mu^{t_1,t_2}}(z) + Q(z) & \geqslant F_1^{t_1,t_2}, \quad \left(\sigma_1 - \mu_1^{t_1,t_2}\right) \ pp, \\ U^{\mu^{t_1,t_2}}(z) + Q(z) & \leqslant F_1^{t_1,t_2}, \quad qp \ sur \ \text{supp}\left(\mu_1^{t_1,t_2}\right), \\ -U^{\mu^{t_1,t_2}}(z) - Q(z) & \geqslant F_2^{t_1,t_2}, \quad \left(\sigma_2 - \mu_2^{t_1,t_2}\right) \ pp, \\ -U^{\mu^{t_1,t_2}}(z) - Q(z) & \leqslant F_2^{t_1,t_2}, \quad qp \ sur \ \text{supp}\left(\mu_2^{t_1,t_2}\right), \end{cases} \tag{4.4}$$

avec des constantes réelles $F_1^{t_1,t_2}, F_2^{t_1,t_2}$.

PREUVE : On note comme précédemment μ^{t_1,t_2} la solution du problème extrémal (P) dont l'existence a été prouvée au lemme 4.2.7.

On suppose tout d'abord que les conditions

$$\begin{cases} U^{\mu^{t_1,t_2}}(z)+Q(z) & \geqslant F_1^{t_1,t_2}, \quad \left(\sigma_1-\mu_1^{t_1,t_2}\right) \text{ pp}, \\ U^{\mu^{t_1,t_2}}(z)+Q(z) & \leqslant F_1^{t_1,t_2}, \quad \text{qp sur } supp\left(\mu_1^{t_1,t_2}\right), \end{cases} \tag{4.5}$$

ne sont pas vérifiées, autrement dit qu'il existe des ensembles

$$K \subset supp\left(\sigma_1-\mu_1^{t_1,t_2}\right),\ \tilde{K} \subset supp\left(\mu_1^{t_1,t_2}\right),$$

vérifiant

$$\left(\sigma_1-\mu_1^{t_1,t_2}\right)(K)>0,\ cap\left(\tilde{K}\right)>0$$

et que l'inégalité

$$U^{\mu^{t_1,t_2}}(y)+Q(y)>l_2>l_1>U^{\mu^{t_1,t_2}}(z)+Q(z) \tag{4.6}$$

soit vérifiée pour certaines constantes l_1 et l_2 pour tout $(y,z)\in\tilde{K}\times K$.

Comme $\tilde{K}$ est de capacité strictement positive et que

$$supp\left(\mu_1^{t_1,t_2}\right)\cap supp\left(\mu_2^{t_1,t_2}\right)\subset\Sigma_1\cap\Sigma_2$$

est de capacité nulle, on peut supposer $\tilde{K}\subset supp\left(\mu_1^{t_1,t_2}\right)$.

Soit $x_0\in\tilde{K}$: pour tout $\varepsilon>0$, on peut alors trouver un ouvert noté $\Omega(x_0,\varepsilon)$ de mesure μ^{t_1,t_2} strictement positive contenu dans la boule $\mathcal{B}(x_0,\varepsilon)$ de centre x_0 de rayon ε tel que

$$\Omega(x_0,\varepsilon)\cap supp\left(\mu_2^{t_1,t_2}\right)=\emptyset.$$

Comme la fonction $U^{\mu^{t_1,t_2}}+Q$ est semi-continue inférieurement sur $\mathbb{C}\setminus supp\left(\mu_2^{t_1,t_2}\right)$, on a

$$\left(U^{\mu^{t_1,t_2}}+Q\right)(y)>l_2 \text{ pour tout } y\in\ K':=\Omega(x_0,\varepsilon)\cap supp\left(\mu_1^{t_1,t_2}\right).$$

On remarque que d'après les inégalités (4.6), $K\cap K'=\emptyset$.

Par construction des ensembles K et K', on a de plus $\left(\sigma_1-\mu^{t_1,t_2}\right)(K)>0$ et $\mu^{t_1,t_2}(K')>0$. Définissons alors

$$\nu:=\left(\sigma_1-\mu^{t_1,t_2}\right)|_K-\alpha\mu^{t_1,t_2}|_{K'},$$

où $\alpha>0$ est un paramètre choisi de façon à ce que ν soit une mesure de masse totale nulle.

Maintenant, pour tout réel $\delta>0$ assez petit, $\mu^{t_1,t_2}+\delta\nu$ appartient à $\mathcal{M}_{\sigma_1,\sigma_2}^{t_1,t_2}$ par construction, et on a

$$\begin{aligned} I_Q\left(\mu^{t_1,t_2}+\delta\nu\right)-I_Q\left(\mu^{t_1,t_2}\right) &= 2\delta\int\left(U^{\mu^{t_1,t_2}}+Q\right)(z)\,\mathrm{d}\nu(z)+\delta^2 I(\nu), \\ &< 2\delta\left[l_1\left(\sigma_1-\mu^{t_1,t_2}\right)(K)-l_2\alpha\mu^{t_1,t_2}(K')\right]+\delta^2 I(\nu), \\ &= 2\delta\left(\sigma_1-\mu^{t_1,t_2}\right)(K)(l_1-l_2)+\delta^2 I(\nu), \\ &< 0, \end{aligned}$$

pour δ suffisamment petit.

Le même raisonnement s'applique dans le cas où les conditions

$$\begin{cases} -U^{\mu^{t_1,t_2}}(z) - Q(z) & \geqslant F_2^{t_1,t_2}, \quad \left(\sigma_2 - \mu_2^{t_1,t_2}\right) \text{ pp}, \\ -U^{\mu^{t_1,t_2}}(z) - Q(z) & \leqslant F_2^{t_1,t_2}, \quad \text{qp sur } supp\left(\mu_2^{t_1,t_2}\right), \end{cases} \tag{4.7}$$

ne sont pas vérifiées.

On a ainsi construit un élément de $\mathcal{M}_{\sigma_1,\sigma_2}^{t_1,t_2}$ d'énergie logarithmique strictement inférieure à celle de μ^{t_1,t_2}, ce qui contredit le fait que cette mesure minimise l'énergie logarithmique parmi tous les éléments de $\mathcal{M}_{\sigma_1,\sigma_2}^{t_1,t_2}$. □

La définition suivante, dont la deuxième partie est tirée de [SaTo97, Section I.3], nous propose une notation différente pour les conditions d'équilibre.

Définition 4.2.13. *Pour une fonction h à valeurs réelles et un sous-ensemble H de $\mathbb{C}$, on note*

$$\text{" inf "}_{z\in H} h(z)$$

le plus grand élément L de $\mathbb{R} \cup \{-\infty, +\infty\}$ tel que h prenne des valeurs strictement inférieures à L seulement sur un ensemble de capacité logarithmique nulle.

On définit de même la notation " sup ".

On remarque que la définition précédente revient en d'autres termes à quotienter par les ensembles de capacité logarithmique nulle, ce qui constitue une approche classique en théorie de la mesure pour les ensembles de mesure nulle.

On remarque également que les équations suivantes extraites des conditions d'équilibre (4.4)

$$\begin{cases} U^{\mu^{t_1,t_2}}(z) + Q(z) & \leqslant F_1^{t_1,t_2}, \quad \text{qp sur } supp\left(\mu_1^{t_1,t_2}\right), \\ -U^{\mu^{t_1,t_2}}(z) - Q(z) & \leqslant F_2^{t_1,t_2}, \quad \text{qp sur } supp\left(\mu_2^{t_1,t_2}\right), \end{cases}$$

peuvent se réécrire avec le formalisme de la définition 4.2.13 sous la forme suivante :

$$\begin{cases} \text{" sup}_{z\in supp\left(\mu_1^{t_1,t_2}\right)}\text{ "} \left(U^{\mu^{t_1,t_2}}(z) + Q(z)\right) & \leqslant F_1^{t_1,t_2}, \\ \text{" sup}_{z\in supp\left(\mu_2^{t_1,t_2}\right)}\text{ "} \left(-U^{\mu^{t_1,t_2}}(z) - Q(z)\right) & \leqslant F_2^{t_1,t_2}. \end{cases}$$

Voici maintenant un lemme de régularité à propos du potentiel logarithmique de la mesure minimisante, lemme qui nous permettra en exploitant pleinement les conditions de régularité imposées aux ensembles et aux mesures considérées de simplifier les conditions d'équilibre vérifiées par la mesure minimisante μ^{t_1,t_2}.

Il est à noter que la preuve de ce résultat utilise la version (4.4) des conditions d'équilibre pour en déduire la version plus simple (4.9) de celles-ci.

La théorie du potentiel propose en effet un vaste éventail de situations (avec ou sans contrainte, avec ou sans champ extérieur, mesures ou mesures signées) et il est bien souvent utile d'identifier une passerelle entre deux situations distinctes pour pouvoir transporter d'un contexte à l'autre des résultats théoriques déjà établis dans un cas connu. Cela sera fait dans la preuve de la proposition suivante à partir des conditions (4.4) à l'aide d'un théorème de continuité du potentiel logarithmique pour des mesures sans contrainte avec champ extérieur énoncé en 4.2.10.

Le même type de raisonnement pour se ramener à une situation connue sera également effectué dans un chapitre suivant sous le nom de *dualité champ-contrainte*.

Lemme 4.2.14. *La fonction $U^{\mu^{t_1,t_2}}$ est continue sur $\mathbb{C}$.*

PREUVE : La fonction $U^{\mu^{t_1,t_2}}$ est continue sur $\mathbb{C} \setminus (\Sigma_1 \cup \Sigma_2)$ puisqu'elle est harmonique sur cet ensemble, et celle-ci est continue sur $E_1^c \cup E_2^c$ d'après le lemme de Rakhmanov 4.2.11, il nous reste donc seulement à prouver le résultat de continuité sur les ensembles E_1 et E_2.

Notons pour $j \in \{1,2\}$, $\tilde{t_j} := t_j - \mu_j^{t_1,t_2}\left(E_j^c\right)$.

En définissant le champ extérieur continu sur E_1

$$\tilde{Q} := Q + U^{\mu_1^{t_1,t_2}|_{E_1^c}} - U^{\mu_2^{t_1,t_2}},$$

on constate d'après la proposition 4.2.12 que la mesure $\mu_1^{t_1,t_2}|_{E_1}$ vérifie les conditions d'équilibre

$$\begin{cases} U^{\mu_1^{t_1,t_2}|_{E_1}}(z) + \tilde{Q}(z) & \geqslant F_1^{t_1,t_2}, \quad z \in E_1, \\ U^{\mu_1^{t_1,t_2}|_{E_1}}(z) + \tilde{Q}(z) & \leqslant F_1^{t_1,t_2}, \quad \text{qp sur } \mathit{supp}\left(\mu_1^{t_1,t_2}|_{E_1}\right), \end{cases} \tag{4.8}$$

et est par conséquent, d'après [SaTo97, Théorème I.3.3], solution du problème de minimisation d'énergie sans contrainte sur l'ensemble E_1 pour des mesures de masse $\tilde{t_1}$ avec champ extérieur $\tilde{Q}$.

Ainsi, toutes les composantes de $\mathbb{C} \setminus E_1$ étant supposées régulières par rapport au problème de Dirichlet , on en déduit la continuité de $U^{\mu^{t_1,t_2}}|_{E_1}$ sur E_1 d'après le théorème 4.2.10, ce qui termine la preuve. □

Voici pour conclure sur cette section les conditions d'équilibre simplifiées vérifiées par le minimiseur.

Corollaire 4.2.15. *Sous l'hypothèse 4.2, le minimiseur $\mu^{t_1,t_2} \in \mathcal{M}_{\sigma_1,\sigma_2}^{t_1,t_2}$ vérifie les conditions d'équilibre*

$$\begin{cases} U^{\mu^{t_1,t_2}}(z) + Q(z) & \geqslant F_1^{t_1,t_2}, \quad z \in \text{supp}\left(\sigma_1 - \mu_1^{t_1,t_2}\right) \\ U^{\mu^{t_1,t_2}}(z) + Q(z) & \leqslant F_1^{t_1,t_2}, \quad z \in \text{supp}\left(\mu_1^{t_1,t_2}\right), \\ -U^{\mu^{t_1,t_2}}(z) - Q(z) & \geqslant F_2^{t_1,t_2}, \quad z \in \text{supp}\left(\sigma_2 - \mu_2^{t_1,t_2}\right) \\ -U^{\mu^{t_1,t_2}}(z) - Q(z) & \leqslant F_2^{t_1,t_2}, \quad z \in \text{supp}\left(\mu_2^{t_1,t_2}\right). \end{cases} \tag{4.9}$$

PREUVE : Toutes les inégalités du système (4.4) sont vraies partout d'après le lemme 4.2.14 par continuité du potentiel $U^{\mu^{t_1,t_2}}$ sur $\mathbb{C}$, ce qui prouve le résultat demandé. □

Remarque. On remarque maintenant que d'après les conditions d'équilibre établies ci-dessus, on a pour $j \in \{1,2\}$

$$z \mapsto U^{\mu^{t_1,t_2}}(z) + Q(z) = (-1)^{j+1} F_j^{t_1,t_2},$$

sur l'ensemble $\mathit{supp}\left(\mu_j^{t_1,t_2}\right)$ ce qui prouve l'unicité du couple de constantes extrémales $\left(F_1^{t_1,t_2}, F_2^{t_1,t_2}\right)$ dans le cas $t_1 > 0$ et $t_2 > 0$.

4.2.5 Un cas particulier important

On présente ici le cas du problème de minimisation d'énergie sous contrainte sans champ extérieur avec des mesures contraintes finies σ_1 et σ_2. Dans ce cas, on peut facilement se ramener à un problème de minimisation d'énergie sous contrainte pour des mesures (et non des mesures signées). En effet, la mesure en laquelle la fonctionnelle énergie atteint son minimum global doit également minimiser les applications partielles associées, ce qui permet d'utiliser des théorèmes déjà disponibles dans la littérature dans ce cadre, par exemple [DrSa97, Théorème 2.6]. Cela fournit une preuve plus rapide des conditions d'équilibre vérifiées par la mesure minimisante dans ce cas particulier.

On rappelle ici le *principe de domination* (voir [SaTo97, Théorème II.3.2]) qui permet d'étendre le domaine de validité des conditions d'équilibre dans le cas particulier considéré ici.

Théorème 4.2.16 (Principe de domination)**.** *Soient μ et ν deux mesures à support compact sur $\mathbb{C}$, telles que la masse totale du ν ne dépasse pas celle de μ.*

Supposons également que μ a une énergie logarithmique finie.

Alors, si pour une constante c l'inégalité

$$U^{\mu}(z) \leqslant U^{\nu}(z) + c$$

a lieu μ-presque partout, alors celle-ci est vérifiée pour tout z de $\mathbb{C}$.

Proposition 4.2.17. *Sous l'hypothèse 4.2, le minimiseur $\mu^t \in \mathcal{M}_\sigma^t$ vérifie les conditions d'équilibre*

$$\begin{cases} U^{\mu^t}(z) & = F_1^t, \quad z \in \operatorname{supp}(\sigma_1 - \mu_1^t), \\ U^{\mu^t}(z) & \leqslant F_1^t, \quad z \in \mathbb{C}, \\ -U^{\mu^t}(z) & = F_2^t, \quad z \in \operatorname{supp}(\sigma_2 - \mu_2^t), \\ -U^{\mu^t}(z) & \leqslant F_2^t, \quad z \in \mathbb{C}. \end{cases} \tag{4.10}$$

avec des constantes réelles positives F_1^t, F_2^t telles que $F_1^t + F_2^t > 0$.

PREUVE : La mesure μ^t vérifie les conditions d'équilibre (4.9), et comme μ_1^t, μ_2^t sont deux mesures à support compact de masse t et d'énergie logarithmique finie, on déduit de la seconde condition d'équilibre d'après le principe de domination rappelé à la proposition 4.2.16 que l'inégalité

$$U^{\mu_1^t}(z) - U^{\mu_2^t}(z) \leqslant F_1^t$$

est valable sur tout le plan complexe, et on procède de même pour l'inégalité

$$U^{\mu_2^t}(z) - U^{\mu_1^t}(z) \leqslant F_2^t.$$

Enfin, comme $\lim_{|z|\to\infty} U^{\mu_t}(z) = 0$, on déduit directement de (4.10) l'inégalité

$$-F_2^t \leqslant 0 \leqslant F_1^t,$$

et comme la mesure μ^t n'est pas nulle, la stricte positivité de $F_1^t + F_2^t > 0$ découle du lemme 4.2.5. □

4.2.6 Caractérisations de la mesure minimisante

Donnons maintenant quelques caractérisations de la mesure minimisante pour le problème (P), le but de ce paragraphe étant de prouver que les conditions d'équilibre (4.9) précédemment énoncées permettent de caractériser la mesure d'équilibre μ^t.

On notera dans toute cette section $\mu^\star = \mu^{t_1,t_2}$ la solution du problème de minimisation d'énergie (P) et $\mathcal{M}^\star = \mathcal{M}^{t_1,t_2}_{\sigma_1,\sigma_2}$ l'ensemble des mesures candidates.

Commençons maintenant par une propriété classique du minimiseur global pour un problème d'optimisation convexe.

Lemme 4.2.18. *La solution $\mu^\star$ au problème (P) est caractérisée par la propriété suivante :*

$$I(\mu^\star) + \int Q(z)\,\mathrm{d}\mu^\star(z) \leqslant I(\nu,\mu^\star) + \int Q(z)\,\mathrm{d}\nu(z), \quad \forall \nu \in \mathcal{M}^\star.$$

PREUVE : Soit $\nu \in \mathcal{M}^\star$ et $s \in (0,1]$: $s\nu + (1-s)\mu^\star$ appartient à $\mathcal{M}^\star$ qui est convexe, et par minimalité de $\mu^\star$,

$$\begin{aligned} 0 &\leqslant I_Q\left(s\nu + (1-s)\mu^\star\right) - I_Q\left(\mu^\star\right), &(4.11)\\ &= s^2 I(\nu) + (s^2-2s)I(\mu^\star) + 2s(1-s)I(\nu,\mu^\star) + 2s\int Q(z)\,\mathrm{d}\left(\nu-\mu^\star\right)(z), &(4.12)\end{aligned}$$

d'où le résultat en divisant les deux membres de l'inégalité obtenue par $2s$ puis en faisant tendre s vers 0^+.

Réciproquement, si $\mu \in \mathcal{M}^\star$ vérifie

$$I(\mu) + \int Q(z)\,\mathrm{d}\mu(z) \leqslant I(\nu,\mu) + \int Q(z)\,\mathrm{d}\nu(z) \quad \forall \nu \in \mathcal{M}^\star,$$

alors comme

$$I(\nu,\mu) = \frac{1}{2}\left(I(\nu) + I(\mu) - I(\mu-\nu)\right),$$

on choisit $\nu = \mu^\star$ solution du problème (P) ce qui donne

$$I_Q(\mu) \leqslant I_Q(\mu^\star) - I(\mu-\mu^\star),$$

d'où $I(\mu-\mu^\star) = 0$ et $\mu = \mu^\star$ d'après le lemme 4.2.5 car $\mu - \mu^\star$ est de masse totale nulle. □

Nous sommes maintenant prêts à prouver le lemme suivant, objectif principal de cette section.

On suit à nouveau le raisonnement de [DrSa97, Théorème 2.6], notre preuve étant légèrement plus longue suite à la prise en compte des ensembles E_1 et E_2 sur lesquels on permet à la contrainte d'être infinie. Cela n'ajoute pas de difficulté technique majeure dans la preuve suivante.

Lemme 4.2.19. *Soit μ une mesure de $\mathcal{M}^\star$ vérifiant les conditions d'équilibre (4.4) pour un couple de constantes (F_1, F_2).*

Alors, μ est égale à la solution du problème de minimisation (P) que l'on note $\mu^\star$.

PREUVE : Soit μ une mesure de $\mathcal{M}^\star$ vérifiant les conditions d'équilibre (4.4) pour un couple de constantes réelles (F_1, F_2), soit

$$\begin{cases} U^\mu(z) + Q(z) & \geqslant F_1 \quad (\sigma_1 - \mu_1) \text{ pp}, \\ U^\mu(z) + Q(z) & \leqslant F_1, \quad \text{qp sur } supp\,(\mu_1)\,, \\ -U^\mu(z) - Q(z) & \geqslant F_2, \quad (\sigma_2 - \mu_2) \text{ pp}, \\ -U^\mu(z) - Q(z) & \leqslant F_2, \quad \text{qp sur } supp\,(\mu_2)\,. \end{cases} \tag{4.13}$$

Montrons que pour toute mesure ν de $\mathcal{M}^\star$ on a l'inégalité

$$I(\mu) + \int Q(z)\,\mathrm{d}\mu(z) \leqslant I(\nu, \mu) + \int Q(z)\,\mathrm{d}\nu(z), \tag{4.14}$$

ce qui permettra d'en déduire que $\mu = \mu^\star$ d'après le lemme 4.2.18.

On commence par remarquer que l'inégalité (4.14) est équivalente à la formulation plus compacte suivante

$$\int \left(U^\mu + Q\right)(z)\,\mathrm{d}\left(\nu - \mu\right)(z) \geqslant 0$$

et comme pour $j = 1, 2$, $\nu_j - \mu_j$ est de masse totale nulle, on peut de façon équivalente démontrer les inégalités suivantes :

$$I_j := \int f_j(z)\,\mathrm{d}\left(\nu_j - \mu_j\right)(z) \geqslant 0, \text{ où } f_j(z) := (-1)^{j+1}\left(U^\mu + Q\right)(z) - F_j, \ j = 1, 2.$$

On s'attache maintenant à prouver la positivité de l'intégrale I_1, la preuve étant la même pour I_2.

On décompose alors I_1 sous la forme

$$I_1 = \int_{E^-} f_1(z)\,\mathrm{d}\left(\nu_1 - \mu_1\right)(z) + \int_{E^+} f_1(z)\,\mathrm{d}\left(\nu_1 - \mu_1\right)(z)$$

où $E^- := f_1^{-1}\left((-\infty, 0)\right)$ et $E^+ := f_1^{-1}\left([0, +\infty)\right)$.

Maintenant, d'après les conditions d'équilibre (4.4) vérifiées par μ, $\mu_1\left(E^+\right) = 0$, d'après [Ra95, Théorème 3.2.3] car μ_1 est une mesure d'énergie logarithmique finie. On a donc

$$\int_{E^+} f_1(z)\,\mathrm{d}\left(\nu_1 - \mu_1\right)(z) = \int_{E^+} f_1(z)\,\mathrm{d}\nu_1(z) \geqslant 0,$$

car ν_1 est une mesure positive.

On décompose alors la deuxième intégrale comme suit :

$$\begin{aligned} \int_{E^-} f_1(z) d\left(\nu_1 - \mu_1\right)(z) = &\int_{E^- \cap E_1^c} f_1(z)\,\mathrm{d}\left(\nu_1 - \sigma_1\right)(z) \\ &+ \int_{E^- \cap E_1^c} f_1(z)\,\mathrm{d}\left(\sigma_1 - \mu_1\right)(z) + \int_{E^- \cap E_1} f_1(z)\,\mathrm{d}\left(\nu_1 - \mu_1\right)(z) \end{aligned}$$

et d'après les conditions d'équilibre on a les relations

$$\left(\sigma_1 - \mu_1\right)\left(E^- \cap E_1^c\right) = 0 \text{ et } E^- \cap E_1 = \emptyset,$$

donc

$$\int_{E^- \cap E_1^c} f_1(z)\,\mathrm{d}\left(\sigma_1 - \mu_1\right)(z) = \int_{E^- \cap E_1} f_1(z)\,\mathrm{d}\left(\nu_1 - \mu_1\right)(z) = 0,$$

et comme ν appartient à $\mathcal{M}^\star$, $\nu_1 \leqslant \sigma_1$ sur E_1^c, d'après les conditions d'équilibre, on a

$$\int_{E^- \cap E_1^c} f_1(z)\,\mathrm{d}\,(\nu_1 - \sigma_1)\,(z) \geqslant 0,$$

ce qui termine la preuve.

□

4.3 Synthèse

Synthétisons dans le théorème suivant les principales propriétés obtenues dans le cadre de notre étude.

Théorème 4.3.1. *On suppose les hypothèses 4.1 et 4.2 vérifiées.*

Alors, pour un couple de réels (t_1, t_2) donné tel que pour $j = 1, 2$, $0 \leqslant t_j \leqslant \sigma_j(\mathbb{C})$, le problème de minimisation d'énergie

$$(P): \ \textit{Trouver } \mu^{t_1,t_2} \textit{ telle que} \qquad I_Q(\mu^{t_1,t_2}) = \inf\left\{I_Q(\mu),\ \mu \in \mathcal{M}^{t_1,t_2}_{\sigma_1,\sigma_2}\right\},$$

où l'ensemble des mesures candidates $\mathcal{M}^{t_1,t_2}_{\sigma_1,\sigma_2}$ a été défini en 4.1.14 admet une unique solution μ^{t_1,t_2} vérifiant les conditions d'équilibre suivantes : il existe deux constantes réelles $F_1^{t_1,t_2}$ et $F_2^{t_1,t_2}$ telles que

$$\begin{cases} U^{\mu^{t_1,t_2}}(z) + Q(z) & \geqslant F_1^{t_1,t_2}, \quad z \in \operatorname{supp}\left(\sigma_1 - \mu_1^{t_1,t_2}\right), \\ U^{\mu^{t_1,t_2}}(z) + Q(z) & \leqslant F_1^{t_1,t_2}, \quad z \in \operatorname{supp}\left(\mu_1^{t_1,t_2}\right), \\ -U^{\mu^{t_1,t_2}}(z) - Q(z) & \geqslant F_2^{t_1,t_2}, \quad z \in \operatorname{supp}\left(\sigma_2 - \mu_2^{t_1,t_2}\right), \\ -U^{\mu^{t_1,t_2}}(z) - Q(z) & \leqslant F_2^{t_1,t_2}, \quad z \in \operatorname{supp}\left(\mu_2^{t_1,t_2}\right). \end{cases} \tag{4.15}$$

Réciproquement, une mesure signée $\mu \in \mathcal{M}^{t_1,t_2}_{\sigma_1,\sigma_2}$ vérifiant le système d'équations

$$\begin{cases} U^{\mu}(z) + Q(z) & \geqslant F_1, \quad (\sigma_1 - \mu_1) \ \ pp, \\ U^{\mu}(z) + Q(z) & \leqslant F_1, \quad qp \ sur \ \operatorname{supp}\left(\mu_1\right), \\ -U^{\mu}(z) - Q(z) & \geqslant F_2, \quad (\sigma_2 - \mu_2) \ \ pp, \\ -U^{\mu}(z) - Q(z) & \leqslant F_2, \quad qp \ sur \ \operatorname{supp}\left(\mu_2\right), \end{cases} \tag{4.16}$$

pour un couple de constantes réelles (F_1, F_2) est égale à la mesure d'équilibre signée μ^{t_1,t_2} et on a de plus les égalités $F_1 = F_1^{t_1,t_2}$ et $F_2 = F_2^{t_1,t_2}$.

Si on suppose en outre le champ extérieur Q identiquement nul, les mesures σ_1, σ_2 finies et $t_1 = t_2 = t \in (0, \min(\sigma_1(\mathbb{C}), \sigma_2(\mathbb{C})))$, on note $\mathcal{M}^t_\sigma$ l'ensemble des mesures candidates, μ^t la mesure minimisante, et les conditions d'équilibre s'écrivent alors

$$\begin{cases} U^{\mu^t}(z) & = F_1^t, \quad z \in \operatorname{supp}(\sigma_1 - \mu_1^t), \\ U^{\mu^t}(z) & \leqslant F_1^t, \quad z \in \mathbb{C}, \\ -U^{\mu^t}(z) & = F_2^t, \quad z \in \operatorname{supp}(\sigma_2 - \mu_2^t), \\ -U^{\mu^t}(z) & \leqslant F_2^t, \quad z \in \mathbb{C}. \end{cases} \tag{4.17}$$

avec des constantes réelles positives F_1^t, F_2^t telles que $F_1^t + F_2^t > 0$.

Chapitre 5

Étude asymptotique du problème de Zolotarev

Le chapitre 4 a permis de préciser le contexte adéquat en théorie du potentiel logarithmique, il reste maintenant à faire le lien entre le problème de Zolotarev pour des ensembles discrets exposé chapitre 3 et les résultats du théorème 4.3.1 qui synthétisent notre étude de théorie du potentiel.

Après avoir précisé le contexte technique de notre étude grâce à quelques hypothèses préalables, le travail consiste à reformuler le problème de Zolotarev sous forme de théorie du potentiel pour des ensembles discrets. Cela permet de faire le lien avec la théorie continue qui décrit l'asymptotique de notre problème discret.

On détaille tout d'abord dans ce chapitre les hypothèses techniques que l'on choisit d'imposer pour obtenir les résultats concernant l'asymptotique du problème de Zolotarev pour des ensembles discrets. En particulier, on fait l'hypothèse que les familles d'ensembles discrets liées au problème de Zolotarev que l'on considère admettent une répartition asymptotique dont la densité donne la mesure contrainte pour le problème de théorie du potentiel considéré au chapitre précédent.

Des hypothèses de séparation sur les familles d'ensembles discrets considérées sont également formulées dans ce chapitre. Celles-ci sont nécessaires dans la seconde partie du chapitre pour décrire l'asymptotique faible du problème de Zolotarev pour des ensembles discrets.

On se trouve alors dans le cadre d'un problème de théorie du potentiel sous contrainte pour des mesures signées où la mesure contrainte σ est finie, problème à propos duquel le théorème 4.3.1 synthétise les résultats de théorie du potentiel du chapitre précédent.

On s'inspire une fois de plus de l'étude fine de l'asymptotique de l'erreur commise lors de l'application de l'algorithme du gradient conjugué dans [BeKu01a], et on commence par construire une fraction rationnelle candidate pour donner une borne supérieure pour l'asymptotique faible de la quantité de Zolotarev. Cela utilise nos résultats de théorie du potentiel sous contrainte qui permettent un choix judicieux des pôles et les zéros de cette fraction rationnelle.

Nous obtenons ainsi une borne supérieure pour le comportement asymptotique de la quantité de Zolotarev pour des ensembles discrets, et sous une hypothèse supplémentaire de séparation sur les ensembles considérés nous prouvons que cette borne supérieure décrit effectivement l'asymptotique de notre problème.

L'établissement de la borne inférieure pour la quantité de Zolotarev constitue le point le plus technique du chapitre. Pour obtenir cette minoration, on définit les points de Fekete rationnels pour des ensembles discrets. Ces points sont fréquemment considérés dans la littérature dans le contexte d'ensembles continus. La détermination de l'asymptotique du problème de Zolotarev pour les ensembles de points de Fekete ne pose alors pas de difficulté technique majeure, en tant qu'elle permet de passer d'un problème de minimisation avec un ensemble infini de candidats à un problème de minimisation fini avec peu de degrés de liberté. Le point essentiel de la démonstration de la borne inférieure réside dans la preuve du lien entre la quantité de Zolotarev $Z_n(E_N, F_N)$ et la quantité de Zolotarev pour les sous-ensembles de points de Fekete rationnels discrets de E_N et F_N qui fait intervenir le conditionnement d'une matrice de Cauchy construite à partir de ces points.

Il nous restera alors à obtenir dans le chapitre suivant 6 une formulation intégrale pour les constantes extrémales qui décrivent le comportement asymptotique de la quantité de Zolotarev. Cette formulation sera utilisée pour comparer nos résultats aux résultats classiques dans le chapitre 8.

5.1 Hypothèses techniques

On présente les hypothèses techniques que l'on choisit d'imposer à notre étude pour pouvoir obtenir des résultats concernant l'asymptotique faible du problème de Zolotarev pour des ensembles discrets.

Dans ce qui suit, on considère des familles d'ensembles discrets $(E_N)_N$ et $(F_N)_N$ - typiquement données par tout ou partie du spectre d'une famille de matrices de taille $N \times N$- et on donne maintenant les hypothèses que l'on juge utile d'imposer afin de pouvoir décrire l'asymptotique de la quantité $Z_n(E_N, F_N)$ en fonction de la distribution asympotique des familles $(E_N)_N$ et $(F_N)_N$ que l'on suppose connue et de la limite du quotient n/N donnée par un paramètre $t \in (0, 1)$.

Hypothèse 5.1. *Pour tout N, les ensembles E_N et F_N sont disjoints, de cardinal N, on suppose également $\bigcup_N (E_N \cup F_N)$ borné.*

L'hypothèse de bornitude peut parfois être réalisée en utilisant une transformation de Moëbius.

Définition 5.1.1. *La* mesure de comptage normalisée *d'un ensemble discret E est donnée par*

$$\nu_n(E) := \frac{1}{n} \sum_{\lambda \in E} \delta_\lambda$$

où $n \in \mathbb{N}$.

Hypothèse 5.2. *On suppose alors que les mesures de comptage normalisées des familles d'ensembles discrets $(E_N)_N$ et $(F_N)_N$ admettent une limite faible pour $N \to +\infty$:*

$$\nu_N(E_N) \xrightarrow{\star} \sigma_1 \text{ et } \nu_N(F_N) \xrightarrow{\star} \sigma_2.$$

Donnons maintenant le cadre dans lequel s'insére l'hypothèse 5.2.

Définition 5.1.2. *Soit $(A_n)_{n\geqslant 0}$ une suite de matrices hermitiennes, on note comme précédemment pour $N \geqslant 0$, $\Lambda(A_N)$ le spectre de la matrice A_N.*

La suite $(A_N)_N$ admet une distribution spectrale asymptotique *donnée par la mesure σ si pour toute fonction continue à support compact f on a*

$$\lim_{N\to+\infty} \sum_{\lambda\in\Lambda(A_N)} f(\lambda) = \int f(z)\mathrm{d}\sigma(z)$$

où chaque valeur propre est comptée avec son ordre de multiplicité.

On nomme σ la distribution asymptotique *ou encore la* mesure limite *de la famille $(\Lambda(A_N))_N$.*

Dans le cas présent, le choix d'une matrice A_N diagonale formée des éléments de l'ensemble E_N permet de se retrouver dans le cadre de l'hypothèse 5.2.

Cependant, dans la plupart des cas, on connait la suite de matrices (A_N) sans pour autant connaître le spectre de chacune des matrices formant cette suite.

Voici un exemple classique de suite de matrices admettant une distribution spectrale asymptotique : pour une fonction ω à valeurs réelles définie et intégrable au sens de Lebesgue sur $(-\pi, \pi)$, on note

$$\omega_j := \int_{-\pi}^{\pi} f(t)\,\mathrm{e}^{\imath jt}\,\mathrm{d}t,\ \imath^2 = -1,\ -\infty < j < +\infty$$

son j-ième coefficient de Fourier, et on construit à partir des coefficients de Fourier de ω la suite de matrices $(T_N(\omega))_{N\geqslant 0}$ définie par

$$T_N(\omega) := [\omega_{j-i}]_{1\leqslant j,k\leqslant N}\,.$$

On a alors le résultat suivant dû à Szegö, voir par exemple [BöSi97, Théorème 1.9].

Théorème 5.1.3. *Avec les notations précédentes, on a pour toute fonction f continue à support compact*

$$\lim_{N\to+\infty} \frac{1}{N} \sum_{\lambda\in\Lambda(T_N(\omega))} f(\lambda) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f\left(\omega(t)\right)\mathrm{d}t.$$

Ce résultat classique d'existence d'une distribution spectrale asymptotique a connu de multiples généralisations, par exemple pour le cas d'une fonction ω à plusieurs variables ou à valeur dans un espace de matrices hermitiennes.

Dans le cas d'une fonction ω à valeurs complexes, il mentionne également l'existence d'un théorème de Tyrtyshnikov semblable au théorème 5.1.3 où le rôle des valeurs propres est tenu par les valeurs singulières de la suite de matrices de Toeplitz qui n'est plus *a priori* une suite de matrices hermitiennes.

Dans le cas d'un schéma de discrétisation aux différences finies pour un opérateur différentiel, on peut parfois calculer explicitement la distribution asymptotique spectrale de la suite de matrices associée à la discrétisation, on consultera le chapitre 8 pour un exemple simple d'un tel calcul.

Des travaux récents ont été menés pour le cas de la méthode des éléments finis et permettent d'assurer l'existence d'une telle distribution spectrale asymptotique dans certains cas pour les matrices associées à cette méthode, voir par exemple [BeSC07, Théorème 1.1].

Même si les ensembles discrets considérés dans le cadre du problème de Zolotarev ne sont pas nécessairement donnés en tant que spectre d'une matrice hermitienne, on pourra garder en tête cet exemple dans la suite du chapitre. Le cas de l'application à la résolution approchée d'une équation de Sylvester par la méthode ADI se situe par exemple naturellement dans ce cadre.

Voici maintenant l'hypothèse imposée sur la mesure limite des familles $(E_N)_N$ et $(F_N)_N$:

Hypothèse 5.3. *Pour $j \in \{1,2\}$, on suppose que σ_j est une mesure dont le potentiel logarithmique U^{σ_j} est continu, et dont le support Σ_j est de capacité logarithmique strictement positive.*

On suppose également $\Sigma_1 \cap \Sigma_2$ de capacité logarithmique nulle et on note $\sigma = \sigma_1 - \sigma_2$.

Cette hypothèse technique reprend les hypothèses 4.1 et 4.2 dans un cadre légèrement simplifié, voir par ailleurs la remarque 4.1.2 pour les commentaires à propos de l'hypothèse de continuité sur les potentiels logarithmiques U^{σ_1} et U^{σ_2}.

Définition 5.1.4. *On définit l'*énergie mutuelle régularisée *des mesures de comptage normalisées de deux ensembles discrets E et F :*

$$I^\star\left(\nu_N(E), \nu_N(F)\right) := \frac{1}{N^2} \sum_{x\in E, y\in F, x\neq y} \log \frac{1}{|x-y|}.$$

*ainsi que l'*énergie régularisée *de la mesure de comptage de E :*

$$I^\star\left(\nu_N(E)\right) := I^\star\left(\nu_N(E), \nu_N(E)\right).$$

On peut donner l'interprétation physique suivante de la définition précédente : l'énergie régularisée $I^\star\left(\nu_N(E)\right)$ est l'énergie d'un système constitué de N particules chargées de masses égales et de même signe dans lequel chaque particule interagit avec tous les autres éléments du système à l'exception d'elle-même, alors que dans le cas de l'énergie mutuelle régularisée, une particule de E interagit avec toutes les particules de F sans exception.

On remarque finalement que pour des ensembles discrets disjoints E et F, l'énergie mutuelle régularisée et l'énergie mutuelle logarithmique usuelle des mesures $\nu_N(E)$ et $\nu_N(F)$ coïncident.

Dans le cas d'une mesure de comptage normalisée signée de deux ensembles discrets disjoints E et F, l'interprétation physique précédente nous incite à la définition qui suit.

Définition 5.1.5. *On définit l'*énergie régularisée pour une mesure de comptage normalisée signée *:*

$$I^\star\left(\nu_N(E) - \nu_N(F)\right) := I^\star\left(\nu_N(E)\right) + I^\star\left(\nu_N(F)\right) - 2I^\star\left(\nu_N(E), \nu_N(F)\right),$$

*et on donne de même la définition de l'*énergie mutuelle régularisée pour des mesures de comptage normalisées signées *pour (E,F), (E',F') deux paires d'ensembles discrets disjoints*

$$\begin{aligned} 2I^\star\left(\nu_N\left(E\right) - \nu_N\left(F\right), \nu_N\left(E'\right) - \nu_N\left(F'\right)\right) \quad &= I^\star\left(\nu_N\left(E\cup E'\right) - \nu_N\left(F\cup F'\right)\right) \\ &- I^\star\left(\nu_N\left(E\right) - \nu_N\left(F\right)\right) \\ &- I^\star\left(\nu_N\left(E'\right) - \nu_N\left(F'\right)\right). \end{aligned} \tag{5.1}$$

On donne maintenant une condition de séparation asymptotique sur les familles d'ensembles discrets $(E_N)_N$ et $(F_N)_N$, condition exprimée en terme d'énergie régularisée.

Hypothèse 5.4. *On a*

$$\lim_{N\to+\infty} I^\star\left(\nu_N(E_N)+\nu_N(F_N)\right) = I(\sigma_1+\sigma_2).$$

Hypothèse 5.5. *On suppose que* $\operatorname{dist}(E_N,F_N)^{1/N} \to 1$ *pour* $N\to\infty$.

L'hypothèse 5.4 est une condition de séparation pour les éléments de l'ensemble $E_N \cup F_N$.

D'autres conditions suffisantes de séparation ont été considérées auparavant, voir [Be00] pour une comparaison des différentes hypothèses existantes.

Par exemple, l'hypothèse 5.4 a lieu dès lors que l'on suppose vérifiée la condition suivante issue de [Ra96] vérifiée :

$$\exists C>0,\ \forall N\geqslant 1, \qquad \inf_{x\neq y\in E_N\cup F_N}|x-y|\geqslant \frac{C}{N}.$$

On a également la condition suffisante suivante que l'on cite depuis [Be00, p. 4] après adaptation à notre contexte. Il s'agit d'un condition suffisante pour l'hypothèse d'énergie 5.4 dont la vérification s'avérera la plus difficile parmi toutes les hypothèses effectuées pour les exemples considérés dans le chapitre 8.

Proposition 5.1.6. *Si on a*

$$\lim_{N\to+\infty}\max_{y\in -E_N\cup E_N}\left|\prod_{x\in -E_N\cup E_N,\,x\neq y}|y-x|^{1/N}-\mathrm{e}^{-U^{|\sigma|}(y)}\right|=0, \tag{5.2}$$

où la mesure $|\sigma|$ *décrit d'après 5.2 la distribution asymptotique de la famille* $(-E_N\cup E_N)_N$*, alors l'hypothèse 5.4 est vérifiée.*

Sous certaines conditions, on peut d'ailleurs montrer que ce type d'hypothèse de séparation renforcée entraîne la continuité du potentiel logarithmique de σ, voir par exemple [BeGuVa09, Théorème 3.2].

Remarque. Les hypothèses 5.4 et 5.5 sont indépendantes, considérons par exemple

$$E_N=-F_N=\{\exp(-N^\alpha)\}\cup\{1+j/N : 0\leqslant j<N\}.$$

Dans ce cas, 5.5 est vérifiée si et seulement si $\alpha<1$, alors que 5.4 a lieu pour $\alpha<2$.

Au contraire, pour

$$E_N=-F_N=\{1+\exp(-N^\alpha)\}\cup\{j/N : 1<j\leqslant N\},$$

la condition 5.5 est vérifiée pour tout $\alpha>0$ alors que 5.4 a lieu seulement pour $\alpha<2$.

5.2 Résultat principal

Voici le résultat principal concernant la quantité de Zolotarev pour des familles $(E_N)_N$ et $(F_N)_N$ d'ensembles discrets.

On considère comme précédemment la limite pour $n, N \to \infty$, où on choisit $n = n(N)$ de façon que $n/N \to t > 0$ où t est un élément de $(0,1)$ fixé à l'avance.

Pour l'exemple de l'application à la méthode ADI, t représente le ratio entre le nombre d'itérations que l'on souhaite effectuer et la taille de la matrice considérée, ce qui explique pourquoi dans la plupart des applications t est petit devant 1.

Théorème 5.2.1. *Sous les hypothèses 5.1, 5.2 et 5.3, il existe $T \in (0,1)$ tel que pour tout $t \in (0,T)$, on ait*

$$\limsup_{n,N\to+\infty,n/N\to t} Z_n(E_N,F_N)^{1/N} \leqslant e^{-(F_1^t+F_2^t)}, \tag{5.3}$$

où F_1^t, F_2^t sont deux constantes réelles positives définies par les conditions d'équilibre du problème de théorie du potentiel considéré au theorème 4.3.1.

Dans le cas où on suppose également vérifiées les conditions de séparation 5.4 et 5.5, on a le résultat supplémentaire

$$\lim_{n,N\to+\infty,n/N\to t} Z_n(E_N,F_N)^{1/N} = e^{-(F_1^t+F_2^t)}. \tag{5.4}$$

5.2.1 Borne supérieure

Dans cette section, on utilise les résultats de théorie du potentiel pour établir une borne supérieure concernant la quantité asymptotique de Zolotarev.

Voici un lemme préliminaire vers la démonstration de la majoration (5.3) pour la quantité de Zolotarev sous les hypothèses 5.1, 5.2 et 5.3.

On suppose dorénavant que le paramètre t appartient à $(0,T)$, avec $T := \min(\sigma_1(\mathbb{C}), \sigma_2(\mathbb{C}))$.

Le lemme suivant montre que toute mesure candidate de $\mathcal{M}_\sigma^t$ peut s'écrire comme différence de deux limites faibles de suites de mesures de comptage normalisées de sous-ensembles de E_N et F_N.

Lemme 5.2.2. *Soient σ la mesure contrainte et $\mu = \mu_1 - \mu_2$ dans $\mathcal{M}_\sigma^t$.*

Alors, pour tout $n = n(N) \leqslant N$ on peut choisir deux suites d'ensembles discrets $E_N^\star \subset E_N$ (resp. $F_N^\star \subset F_N$) telles que pour tout N, $\mathrm{card}(E_N^\star) = \mathrm{card}(F_N^\star) = n$, $n/N \to t$,

$$\nu_N(E_N^\star) \overset{\star}{\rightharpoonup} \mu_1 \quad et \quad \nu_N(F_N^\star) \overset{\star}{\rightharpoonup} \mu_2.$$

De plus, si K_1, K_2 sont deux ensembles fermés tels que $\sigma_i(\partial K_i) = 0$ et $\sigma_i(K_i) = \mu_i(K_i)$, on peut choisir $E_N^\star, F_N^\star$ tels que pour N suffisamment grand on ait

$$E_N \cap K_1 \subset E_N^\star \subset E_N \quad et \quad F_N \cap K_2 \subset F_N^\star \subset F_N.$$

PREUVE : Voir [Be00, Lemme 2.1] □

Citons également [SaTo97, Théorème I.6.8] le *principe de descente* qui jouera un rôle central dans la preuve de (5.3).

Théorème 5.2.3 (Principe de descente)**.** *Soient $(\mu_n)_n$ des mesures de probabilité ayant toutes leur support dans un sous-ensemble fixé compact de $\mathbb{C}$ et convergeant vers une mesure μ au sens de la topologie faible-$\star$. Supposons de plus que pour tout n un point z_n est donné tel que $z_n \to z^\star$ où $z^\star$ appartient à $\mathbb{C}$. Alors,*

$$U^\mu(z^\star) \leqslant \liminf_{n\to+\infty} U^{\mu_n}(z_n) \quad \textit{et} \quad I(\mu) \leqslant \liminf_{n\to+\infty} I(\mu_n).$$

Remarque. Sous les hypothèses du principe de descente, si on suppose de plus que pour tout n, z_n appartient à un ensemble Σ_1 et que $supp(\mu_n) \subset \Sigma_2$ où Σ_1 et Σ_2 sont deux ensembles disjoints du plan complexe, alors

$$U^\mu(z^\star) = \lim_{n\to+\infty} U^{\mu_n}(z_n), \tag{5.5}$$

d'après la convergence uniforme des intégrandes.

Maintenant que toutes les conditions sont réunies, démontrons (5.3).

PREUVE : Définissons tout d'abord les ensembles

$$K_1^\varepsilon := \{\lambda \in \mathbb{C},\ U^{\mu^t}(\lambda) \leqslant F_1^t - \varepsilon\}, \quad K_2^\varepsilon := \{\lambda \in \mathbb{C},\ -U^{\mu^t}(\lambda) \leqslant F_2^t - \varepsilon\}.$$

D'après le lemme de Rakhmanov 4.2.11, U^{μ^t} est une fonction continue, et les ensembles K_i^ε, $i = 1, 2$ sont donc fermés.

Les conditions d'équilibre (4.10) vérifiées par μ^t assurent que les deux ensembles K_ε^i et $supp\,(\sigma_i - \mu_i^t)$ sont disjoints, ainsi $\sigma_i(K_i^\varepsilon) = \mu_i^t(K_i^\varepsilon)$.

Comme les ensembles K_i^ε, $i = 1, 2$ admettent des frontières disjointes pour ε assez petit, il existe une quantité au plus dénombrable de $(K_i^\varepsilon)_\varepsilon$ vérifiant $\sigma_i(\partial K_i^\varepsilon) > 0$, et quitte à diminuer le paramètre ε on peut supposer

$$\sigma_i(\partial K_i^\varepsilon) = 0 \text{ pour } i \in \{1, 2\}.$$

On définit maintenant

$$r_n(\lambda) := \prod_{\lambda' \in E_N^\star} \left(1 - \frac{\lambda}{\lambda'}\right) \prod_{\lambda'' \in F_N^\star} \left(1 - \frac{\lambda}{\lambda''}\right)^{-1},$$

où les ensembles $E_N^\star$ et $F_N^\star$ sont choisis comme dans le lemme 5.2.2 appliquée à la mesure d'équilibre signée μ^t.

Le candidat r_n pour établir une borne supérieure pour la quantité de Zolotarev étant construit, établissons maintenant une borne supérieure pour

$$\left\|r_n^{-1}\right\|_{L^\infty(F_N)} \|r_n\|_{L^\infty(E_N)}.$$

On définit $(x_1^N, x_2^N) \in E_N \times F_N$ tels que

$$\|r_n\|_{L^\infty(E_N)} = |r_n(x_1^N)| \qquad \text{et} \qquad \|r_n^{-1}\|_{L^\infty(F_N)} = |r_n^{-1}(x_2^N)|.$$

Comme r_n (resp. r_n^{-1}) s'annule sur $E_N \cap K_1^\varepsilon$ (resp. $F_N \cap K_2^\varepsilon$) d'après le choix des ensembles $E_N^\star$ et $F_N^\star$ décrit dans le lemme 5.2.2, on a les inégalités

$$x_1^N \in \mathbb{C} \setminus K_1^\varepsilon \ = \{\lambda \in \mathbb{C}, \quad U^{\mu^t}(\lambda) > F_1^t - \varepsilon\}, \tag{5.6}$$
$$x_2^N \in \mathbb{C} \setminus K_2^\varepsilon \ = \{\lambda \in \mathbb{C},\ -U^{\mu^t}(\lambda) > F_2^t - \varepsilon\}. \tag{5.7}$$

Les suites $(x_j^N)_N$, $j \in \{1,2\}$, sont bornées d'après l'hypothèse 5.1, et quitte à extraire une sous-suite convergente, on peut supposer que $(x_j^N)_N$ converge pour $N \to +\infty$ avec la limite

$$x_j^\star := \lim_{N\to+\infty} x_j^N.$$

De plus, d'après le lemme 5.2.2 appliqué à la mesure signée minimisante μ^t,

$$\nu_N(E_N^\star) \overset{\star}{\rightharpoonup} \mu_1^t \qquad \text{et} \qquad \nu_N(F_N^\star) \overset{\star}{\rightharpoonup} \mu_2^t.$$

En vertu du principe de descente énoncé théorème 5.2.3, on a

$$U^{\mu_1^t}(x_1^\star) \leqslant \liminf_N U^{\nu_N(E_N^\star)}(x_1^N) \quad \text{et} \quad U^{\mu_2^t}(x_2^\star) \leqslant \liminf_N U^{\nu_N(F_N^\star)}(x_2^N). \tag{5.8}$$

et on a de plus, pour ε suffisamment petit,

$$\overline{\mathbb{C}\setminus K_1^\varepsilon} \cap \overline{\mathbb{C}\setminus K_2^\varepsilon} = \emptyset$$

car d'après le théorème 4.3.1 les constantes $F_1^t \geqslant 0$ et $F_2^t \geqslant 0$ sont de somme strictement positive.

On obtient donc d'après (5.5)

$$U^{\mu_1^t}(x_2^\star) = \lim_N U^{\nu_N(E_N^\star)}(x_2^N) \qquad \text{et} \qquad U^{\mu_2^t}(x_1^\star) = \lim_N U^{\nu_N(F_N^\star)}(x_1^N). \tag{5.9}$$

De plus,

$$\begin{aligned}\frac{1}{N}\log\left|r_n(x_1^N)\right| - \frac{1}{N}\log\left|r_n(x_2^N)\right| \; &= \tfrac{1}{N}\textstyle\sum_{\lambda\in E_N^\star}\log|\lambda - x_1^N| - \tfrac{1}{N}\sum_{\lambda\in F_N^\star}\log|\lambda - x_1^N| \\ &\quad + \tfrac{1}{N}\textstyle\sum_{\lambda\in F_N^\star}\log|\lambda - x_2^N| - \tfrac{1}{N}\sum_{\lambda\in E_N^\star}\log|\lambda - x_2^N|,\end{aligned}$$

ce qui donne avec (5.8) et (5.9) pour $N \to \infty$,

$$-\frac{1}{N}\sum_{\lambda\in E_N^\star}\log|\lambda - x_2^N| \to U^{\mu_1^t}(x_2^\star), \qquad -\frac{1}{N}\sum_{\lambda\in F_N^\star}\log|\lambda - x_1^N| \to U^{\mu_2^t}(x_1^\star)$$

et

$$\limsup_N \frac{1}{N}\sum_{\lambda\in E_N^\star}\log|\lambda - x_1^N| \leqslant -U^{\mu_1^t}(x_1^\star), \; \limsup_N \frac{1}{N}\sum_{\lambda\in F_N^\star}\log|\lambda - x_2^N| \leqslant -U^{\mu_2^t}(x_2^\star).$$

En combinant ces résultats avec ceux établis précédemment dans l'équation (5.6), on conclut que

$$\begin{aligned}&\limsup_N \left[\frac{1}{N}\log\left|r_n(x_1^N)\right| - \frac{1}{N}\log\left|r_n(x_2^N)\right|\right], \\ \leqslant\; & -U^{\mu_1^t}(x_1^\star) + U^{\mu_2^t}(x_1^\star) - U^{\mu_2^t}(x_2^\star) + U^{\mu_1^t}(x_2^\star), \\ =\; & -U^{\mu^t}(x_1^\star) + U^{\mu^t}(x_2^\star) \leqslant -F_1^t - F_2^t + 2\varepsilon.\end{aligned}$$

Cette inégalité étant vérifiée pour tout $\varepsilon > 0$, notre résultat (5.3) en découle immédiatement. □

5.2.2 Points de Fekete rationnels discrets

Les *points de Fekete rationnels* sont définis en [SaTo97, Corollaire 8.3.2], mais sur des ensembles compacts quelconques du plan complexe, et sont très utilisés pour construire des fractions rationnelles asymptotiquement extrémales pour le problème de Zolotarev classique.

On suit dans cette section le même raisonnement pour les points de Fekete discrets dont la définition s'inspire du cas continu en analysant leur comportement asymptotique en terme de mesure minimisante pour le problème extrémal en théorie du potentiel (P).

Définissons tout d'abord les points de Fekete rationnels qui s'avéreront d'une utilité essentielle dans la démonstration de (5.4).

Définition 5.2.4. *Les points de Fekete rationnels d'ordre n maximisent la quantité*

$$\mathcal{F}\left(\left(\lambda_1^0,\dots,\lambda_1^n\right),\left(\lambda_2^0,\dots,\lambda_2^n\right)\right) := \prod_{0\leqslant i\neq j\leqslant n} \frac{|\lambda_1^i-\lambda_1^j|}{|\lambda_1^i-\lambda_2^j|}\frac{|\lambda_2^i-\lambda_2^j|}{|\lambda_2^i-\lambda_1^j|}$$

parmi tous les ensembles $(\lambda_1^i)_{0\leqslant i\leqslant n}\subset E_N$, $(\lambda_2^i)_{0\leqslant i\leqslant n}\subset F_N$.

On notera dans tout ce qui suit $E_{n,N} := (\lambda_1^i)_{i=0}^n$ (resp. $F_{n,N} := (\lambda_2^i)_{i=0}^n$) l'ensemble des points rationels de Fekete d'ordre n de l'ensemble E_N (resp. F_N).

Remarquons qu'on ne dispose pas de résultat d'unicité concernant les points de Fekete rationnels discrets, on supposera dans ce qui suit qu'on choisit un couple de $(n+1)$-uplets de points sataifaisant la définition ci-dessus.

On commet ici l'abus de notation qui consiste à supprimer la dépendance en n, N dans la notation pour les points de Fekete rationels de E_N et F_N d'ordre n, ceci afin de simplifier les notations utilisées.

Le lemme suivant montre en suivant l'idée de [Be00, Lemme 2.2.b)] qu'une fois l'hypothèse 5.4 vérifiée, celle-ci est valable pour toutes familles de sous-ensembles de E_N et F_N dont les mesures de comptage normalisées admettent une limite faible.

Lemme 5.2.5. *Sous l'hypothèse de séparation 5.4, pour toute suite d'ensembles*

$$(G_{1,N},G_{2,N})\subset E_N\cup F_N \text{ telle que pour } j\in\{1,2\},$$

$$\nu_N(G_{j,N})\overset{\star}{\rightharpoonup}\nu_j,\quad \text{on a}\quad \lim_{N\to+\infty} I^\star\left(\nu_N\left(G_{1,N}\right),\nu_N\left(G_{2,N}\right)\right)=I(\nu_1,\nu_2).$$

PREUVE : En notant $k_M(z)=\min\{M,\log\frac{1}{|z|}\}$ le noyau logarithmique tronqué, on a pour tout $M>0$

$$I^\star\left(\nu_N\left(G_{1,N}\right),\nu_N\left(G_{2,N}\right)\right)=\int\int k_M(x-y)d\nu_N\left(G_{1,N}\right)(x)d\nu_N\left(G_{2,N}\right)(y)$$
$$+\frac{1}{N^2}\sum_{(x,y)\in U_M\left(G_{1,N},G_{2,N}\right)}\left(\log\frac{1}{|x-y|}-M\right)-M\frac{card\left(G_{1,N}\cap G_{2,N}\right)}{N^2},$$

où $U_M\left(G_{1,N},G_{2,N}\right) := \{(x,y)\in G_{1,N}\times G_{2,N},\ x\neq y \text{ et } |x-y|<e^{-M}\}$.

Ainsi, puisque

$$\sum_{(x,y)\in U_M\left(G_{1,N},G_{2,N}\right)}\left(\log\frac{1}{|x-y|}-M\right)\geqslant 0$$

et par continuité de k_M sur $\mathbb{C} \times \mathbb{C}$, on a

$$\lim_{N \to +\infty} \int \int k_M(x-y) d\nu_N\left(G_{1,N}\right)(x) d\nu_N\left(G_{2,N}\right)(y) = \int \int k_M(x-y) d\nu_1(x) d\nu_2(y),$$

et d'après le théorème de convergence monotone,

$$\lim_{M \to +\infty} \int \int k_M(x-y) d\nu_1(x) d\nu_2(y) = I(\nu_1, \nu_2).$$

Ceci prouve l'inégalité de semi-continuité

$$\liminf_{N \to +\infty} I^\star\left(\nu_N\left(G_{1,N}\right), \nu_N\left(G_{2,N}\right)\right) \geqslant I(\nu_1, \nu_2) \tag{5.10}$$

pour toute suite d'ensembles $(G_{1,N}, G_{2,N}) \subset E_N \cup F_N$ dont la mesure de comptage normalisée admet une limite-$\star$ notée (ν_1, ν_2).

On écrit maintenant

$$\begin{aligned}
&I^\star\left(\nu_N\left(G_{1,N}\right), \nu_N\left(G_{2,N}\right)\right) \\
&= I^\star\left(\nu_N\left(E_N \cup F_N\right)\right) - I^\star\left(\nu_N\left(G_{1,N}\right), \nu_N\left(E_N \cup F_N \setminus G_{2,N}\right)\right), \\
&- I^\star\left(\nu_N\left(E_N \cup F_N \setminus G_{1,N}\right), \nu_N\left(G_{2,N}\right)\right), \\
&- I^\star\left(\nu_N\left(E_N \cup F_N \setminus G_{1,N}\right), \nu_N\left(E_N \cup F_N \setminus G_{2,N}\right)\right)
\end{aligned}$$

ce qui permet de conclure d'après l'hypothèse 5.4 et (5.10) que

$$\lim_{N \to +\infty} I^\star\left(\nu_N\left(G_{1,N}\right), \nu_N\left(G_{2,N}\right)\right) = I(\nu_1, \nu_2).$$

□

Remarque. La démonstration du lemme 5.2.5 prouve que les inégalités

$$\liminf_N I^\star\left(\nu_N(E_N)\right) \geqslant I(\sigma_1), \ \liminf_N I^\star\left(\nu_N(F_N)\right) \geqslant I(\sigma_2)$$

et

$$\liminf_N I^\star(\nu_N(E_N) + \nu_N(F_N)) \geqslant I(\sigma_1 + \sigma_2)$$

découlent de la seule hypothèse 5.2 pour toute famille d'ensembles discrets comme en 5.2, l'hypothèse 5.4 est donc équivalente aux égalités

$$\lim_N I^\star\left(\nu_N(E_N)\right) = I(\sigma_1), \ \lim_N I^\star\left(\nu_N(F_N)\right) = I(\sigma_2)$$

et

$$\lim_N I^\star\left(\nu_N(E_N), \nu_N(F_N)\right) = I(\sigma_1, \sigma_2).$$

Remarque. Le lemme 5.2.5 peut-être généralisé de la même façon au cadre de l'énergie mutuelle régularisée pour des mesures discrètes admettant une limite au sens de la topologie faible-$\star$ en écrivant le terme d'énergie mutuelle régularisée discrète de deux mesures de comptage signées comme dans (5.1) et en appliquant notre lemme à chaque terme de cette expression.

Le lemme suivant montre que l'asymptotique des points de Fekete rationnels discrets est décrit par la solution de notre problème de minimisation d'énergie (P), ce qui met en lumière leur importance, au moins dans un contexte théorique.

Lemme 5.2.6. *On a le résultat suivant pour les mesures de comptage normalisées des points de Fekete quand $n/N \to t$:*

$$\nu_N(E_{n,N}) \overset{\star}{\rightharpoonup} \mu_1^t \qquad et \qquad \nu_N(F_{n,N}) \overset{\star}{\rightharpoonup} \mu_2^t. \tag{5.11}$$

PREUVE : Rapellons que d'après 5.1 les ensembles discrets E_N et F_N sont contenus dans deux ensembles compact fixés, ce qui permet ainsi d'appliquer le théorème de Helly 4.2.6 et d'extraire deux sous-suites $(n_k)_k$ et $(N_k)_k$ telles que $\nu_{N_k}(E_{n_k,N_k}) \overset{\star}{\rightharpoonup} \mu_1$ et $\nu_{N_k}(F_{n_k,N_k}) \overset{\star}{\rightharpoonup} \mu_2$.

Notons maintenant $\mu := \mu_1 - \mu_2$ la mesure signée limite obtenue ci-dessus.

Alors, μ appartient à $\mathcal{M}_\sigma^t$ et d'après le lemme 5.2.5 et la remarque précédente, on a

$$\lim_{k\to+\infty} I^\star\left(\nu_N\left(E_{n_k,N_k}\right) - \nu_N\left(F_{n_k,N_k}\right)\right) = I(\mu).$$

Soient $E_n^\star \subset E_N$, $F_n^\star \subset F_N$, tous deux de cardinal $n+1$.

D'après les hypothèses 5.1 et 5.5, on a

$$\lim_{N\to\infty} \log\left(\mathcal{F}\left(E_n^\star, F_n^\star\right)^{1/N^2}\right) + I^*(\nu_N(E_n^\star) - \nu_N(F_n^\star)) = 0.$$

De plus, d'après la définition 5.2.4, $\mathcal{F}\left(E_{n,N}, F_{n,N}\right) \geqslant \mathcal{F}\left(E_n^\star, F_n^\star\right)$.

Ainsi, dans le cas des ensembles $E_N^\star, F_N^\star$ choisis comme dans le lemme 5.2.2 avec limite faible-$\star$, μ^t, on obtient en appliquant de nouveau le lemme 5.2.5

$$\begin{aligned} I(\mu) &= \lim_{k\to+\infty} I^\star\left(\nu_N\left(E_{n_k,N_k}\right) - \nu_N\left(F_{n_k,N_k}\right)\right) \\ &\leqslant \lim_{k\to+\infty} I^\star\left(\nu_N\left(E_{N_k}^*\right) - \nu_N\left(F_{N_k}^*\right)\right) \\ &= I(\mu^t), \end{aligned}$$

et par conséquent $\mu = \mu^t$ par unicité du minimiseur dans le théorème 4.3.1. □

Remarque. On voit dès la définition des points de Fekete rationnels 5.2.4 que la détermination numérique de ceux-ci semble vouée à l'échec : il s'agit de maximiser une fraction rationnelle parmi tous les $(n+1)\times(n+1)$-uplets de points de $E_N \times F_N$ et pire encore, la détermination des points de Fekete d'ordre n n'aide en rien à la détermination des points de Fekete d'ordre $n+1$.

On pourra heureusement par la suite construire des points asymptotiquement optimaux pour la quantité de Zolotarev, adaptés au cas discret, bien plus aisés à déterminer numériquement et utiles dans la partie numérique de ce travail.

5.2.3 Quantité de Zolotarev contrainte par localisation des pôles et zéros

On définit maintenant pour E, F ensembles discrets disjoints et $n \geqslant 1$

$$Z_n^\star(E,F) := \inf_{r\in\mathcal{R}_n(E,F)} \|r\|_{L^\infty(E)} \left\|r^{-1}\right\|_{L^\infty(F)},$$

où l'ensemble des candidats $\mathcal{R}_n(E,F) := \{r \in \mathcal{R}_n,\ r \text{ a ses zéros dans } E \text{ et ses pôles dans } F\}$ est fini.

On s'attend de plus à être en mesure de relier les quantités $Z_n^*(E,F)$ et $Z_n(E,F)$, car il semble intuitivement simple pour obtenir une fraction rationnelle r petite sur un ensemble E de contraindre les zéros de r à appartenir à E.

La proposition suivante formalise le lien existant entre les quantités $Z_n(E_N,F_N)$ et $Z_n^\star(E_{n,N},F_{n,N})$.

Commençons par un lemme adapté à la preuve de la borne inférieure asymptotique pour la quantité de Zolotarev et consacré au calcul de l'inverse d'une matrice de type Cauchy.

Lemme 5.2.7. *Soient $(x_i)_{0\leqslant i\leqslant n}$, $(y_j)_{0\leqslant j\leqslant n}$ deux $n+1$-uplets de points tous distincts du plan complexe et pour un choix donné de la racine carrée sur le plan complexe privé d'une demi-droite issue de l'origine d'intersection vide avec $(x_i)_{0\leqslant i\leqslant n}\cup(y_j)_{0\leqslant j\leqslant n}$, on définit*

$$\omega_1(z):=\prod_{\ell=0}^{n}(z-x_\ell),\ \omega_2(z):=\prod_{\ell=0}^{n}(z-y_\ell),$$

et

$$X:=\left[\frac{1}{x_i-y_j}\sqrt{\frac{\omega_1(y_j)\,\omega_2(x_i)}{\omega_1'(x_i)\,\omega_2'(y_j)}}\right]_{0\leqslant i,j\leqslant n}.$$

Alors, X est inversible et $X^{-1}=-X^T$.

PREUVE : On définit tout d'abord

$$X_0:=\left[\frac{1}{x_j-y_k}\right]_{0\leqslant j,k\leqslant n}.$$

Montrons que X_0 est inversible d'inverse

$$X_0^{-1}=\left[Res_{\,z=y_j}\left(\frac{\omega_1(z)}{z-x_k}\frac{1}{\omega_1'(x_k)}\frac{\omega_2(x_k)}{\omega_2(z)}\right)\right]_{0\leqslant j,k\leqslant n}.$$

En effet, en définissant pour $(\ell,j)\in\{0,n\}$ la fraction rationnelle

$$f_{\ell,j}(z):=-\frac{1}{(z-x_\ell)(z-x_j)}\frac{\omega_1(z)}{\omega_1'(x_j)}\frac{\omega_2(x_j)}{\omega_2(z)},$$

on a en calculant $\omega_2'(z)$

$$Res_{\,z=y_k}(f_{\ell,j})=\frac{1}{x_j-y_k}\frac{1}{y_k-x_\ell}\frac{\omega_1(y_k)\omega_2(x_j)}{\omega_1'(x_j)\omega_2'(y_k)},$$

et

$$Res_{\,z=x_j}(f_{\ell,j})=\begin{cases}0 & \text{si } \ell\neq j,\\ -1 & \text{sinon.}\end{cases}$$

Notons Γ_R le cercle $\{|z|=R\}$ orienté positivement. On a pour R suffisamment grand d'après le théorème des résidus appliqué à la fonction $f_{\ell,j}$,

$$\sum_{0\leqslant k\leqslant n}Res_{\,z=y_k}(f_{\ell,j})+\sum_{0\leqslant j\leqslant n}Res_{\,z=x_j}(f_{\ell,j})=\frac{1}{2i\pi}\oint_{\Gamma_R}f_{\ell,j}(z)dz,$$

et après passage à la limite $R \to +\infty$,

$$\sum_{k=0}^{n} \frac{1}{x_\ell - y_k} Res_{z=y_k} \left(\frac{\omega_1(z)}{z - x_j} \frac{1}{\omega_1'(x_j)} \frac{\omega_2(x_j)}{\omega_2(z)} \right) = \begin{cases} 0 & \text{if } \ell \neq j, \\ 1 & \text{sinon,} \end{cases}$$

d'où le résultat pour X_0.

Maintenant, en posant

$$D := diag \left(\sqrt{\frac{\omega_2(x_k)}{\omega_1'(x_k)}} \right)_{0 \leqslant k \leqslant n} \text{ et } D' := diag \left(\sqrt{\frac{\omega_1(y_j)}{\omega_2'(y_j)}} \right)_{0 \leqslant j \leqslant n},$$

on a

$$X = DX_0D' \text{ d'où } X^{-1} = D'^{-1}X_0^{-1}D^{-1},$$

et

$$\begin{aligned} X^{-1} &= \left[\sqrt{\frac{\omega_2'(y_j)}{\omega_1(y_j)}} \sqrt{\frac{\omega_1'(x_k)}{\omega_2(x_k)}} \frac{\omega_1(y_j)\omega_2(x_k)}{\omega_1'(x_k)\omega_2'(y_j)} \frac{1}{y_j - x_k} \right]_{0 \leqslant j,k \leqslant n}, \\ &= \left[\sqrt{\frac{\omega_1(y_j)\omega_2(x_k)}{\omega_1'(x_k)\omega_2'(y_j)}} \frac{1}{y_j - x_k} \right]_{0 \leqslant j,k \leqslant n}, \\ &= -X^T. \end{aligned}$$

D'où le résultat. □

Proposition 5.2.8. *On a*

$$\frac{1}{(n+1)^2} Z_n^\star(E_{n,N}, F_{n,N}) \leqslant Z_n(E_{n,N}, F_{n,N}) \leqslant Z_n(E_N, F_N).$$

PREUVE : L'inégalité $Z_n(E_{n,N}, F_{n,N}) \leqslant Z_n(E_N, F_N)$ est une conséquence triviale des inclusions $E_{n,N} \subset E_N$, $F_{n,N} \subset F_N$. Il reste ainsi à prouver que

$$Z_n^\star(E_{n,N}, F_{n,N}) \leqslant (n+1)^2 Z_n(E_{n,N}, F_{n,N}). \tag{5.12}$$

On définit la matrice étudiée précédemment dans le cas des points de Fekete rationnels discrets

$$X := \left[\frac{1}{\lambda_1^i - \lambda_2^j} \sqrt{\frac{\omega_1\left(\lambda_2^j\right)\omega_2\left(\lambda_1^i\right)}{\omega_1'\left(\lambda_1^i\right)\omega_2'\left(\lambda_2^j\right)}} \right]_{0 \leqslant i,j \leqslant n}, \quad \omega_j(z) := \prod_{\ell=0}^{n} (z - \lambda_i^\ell).$$

On vérifie que si on définit pour $i = 1, 2$,

$$D_i := diag\left(\lambda_0^i, \lambda_1^i, \ldots, \lambda_n^i\right),$$

la matrice $D_1X_0 - X_0D_2$ est de rang 1, et d'après la démonstration du lemme 5.2.7, comme $X = DX_0D'$ où D et D' sont diagonales, on obtient ainsi que $D_1X - XD_2$ est également de rang 1.

On obtient alors par application de (1.9) que

$$\frac{1}{\|X\|\|X^{-1}\|} \leqslant Z_n(E_{n,N}, F_{n,N}), \tag{5.13}$$

où $\|\cdot\|$ désigne la norme spectrale sur l'ensemble des matrices carrées de taille $(n+1)\times(n+1)$ à coefficients complexes.

L'inégalité

$$\frac{1}{\|X\|\|X^{-1}\|} \leqslant \|r\|_{L^\infty(E_{n,N})}\|r^{-1}\|_{L^\infty(F_{n,N})} \tag{5.14}$$

pour toute fraction rationnelle de $r \in \mathcal{R}_n$ a déjà été obtenue par [Pe00a, Theorème 1] pour des ensembles symétriques $E_{n,N} = -F_{n,N}$, et on vérifie maintenant qu'aucun des arguments de la démonstration ne nécessite une telle hypothèse de symétrie.

Pour $i,j \in \{0,\dots,n\}$, on définit

$$r_{i,j}(z) := \frac{\prod_{\ell\neq i}(z-\lambda_1^\ell)}{\prod_{\ell\neq j}(z-\lambda_2^\ell)}.$$

On remarque ainsi que

$$\begin{aligned} Z_n^\star(E_{n,N},F_{n,N}) &= \min_{0\leqslant i,j\leqslant n}\left|\frac{r_{i,j}\left(\lambda_1^i\right)}{r_{i,j}\left(\lambda_2^j\right)}\right|, \\ &= \min_{0\leqslant i,j\leqslant n}\left|\left(\lambda_1^i-\lambda_2^j\right)^2\frac{\omega_1'\left(\lambda_1^i\right)\omega_2'\left(\lambda_2^j\right)}{\omega_2\left(\lambda_1^i\right)\omega_1\left(\lambda_2^j\right)}\right| = \min_{0\leqslant i,j\leqslant n}\frac{1}{|X_{ij}|^2}. \end{aligned}$$

On a $X^{-1} = -X^t$ d'après le lemme 5.2.7, et en rappelant que

$$\max_{0\leqslant i,j\leqslant n}|X_{ij}|^2 \geqslant \frac{1}{(n+1)^2}\|X\|_2^2,$$

on obtient

$$\frac{1}{\|X\|\|X^{-1}\|} \geqslant \frac{1}{(n+1)^2}Z_n^\star(E_{n,N},F_{n,N}), \tag{5.15}$$

ce qui compte tenu de (5.13) prouve l'inégalité souhaitée (5.12). □

5.2.4 Points de Fekete et fractions rationnelles

Tout le travail précédemment effectué dans ce chapitre nous a permis de nous ramener à l'étude de la quantité $Z_n^\star(E_{n,N},F_{n,N})$ au lieu de $Z_n(E_N,F_N)$. La détermination de l'asymptotique faible de $Z_n^\star(E_{n,N},F_{n,N})$ pour $n/N\to t$ devrait être plus aisée : le fait de considérer une quantité de Zolotarev sur les points de Fekete rationnels discrets nous permet de passer d'un problème admettant une infinité de candidats $r\in\mathcal{R}_{n,n}$ à un problème où les fractions rationnelles candidates admettent n zéros et n pôles, ces zéros et ces pôles étant tenus d'appartenir à un ensemble de cardinal $n+1$, ce qui ne laisse que très peu de degrés de liberté pour les minimiseurs potentiels pour $Z_n^\star(E_{n,N},F_{n,N})$.

Montrons maintenant que la famille de fractions rationnelles contruite avec des zéros et des pôles choisis parmi les points de Fekete rationnels discrets de E_N et F_N est asyptotiquement optimale au sens du problème de Zolotarev considéré ici, ce qui nous permet en outre d'obtenir une relation impliquant la constante extrémale obtenue dans le cadre du problème de minimisation d'énergie sous contrainte (P).

Ce passage met à nouveau en valeur l'interêt théorique des points de Fekete, et souligne ainsi la nécessité de construire des points qui permettent d'obtenir un comportement quasi-optimal pour l'asymptotique de la quantité de Zolotarev mais dont l'évaluation numérique soit envisageable.

Lemme 5.2.9. *Avec les notations qui précèdent, on a*

$$\liminf_{n,N,n/N\to t} \min_{0\leqslant i,j\leqslant n} \|r_{i,j}\|_{L^\infty(E_{n,N})}^{1/N} \geqslant e^{-F_1^t},\ \liminf_{n,N,n/N\to t} \min_{0\leqslant i,j\leqslant n} \|r_{i,j}^{-1}\|_{L^\infty(F_{n,N})}^{1/N} \geqslant e^{-F_2^t}.$$

PREUVE : On remarque tout d'abord que pour tous les entiers i, j de $\{0, n\}$ et pour tout $x \in E_N$,

$$|r_{i,j}(x)| = |r_{i,i}(x)| \left|\frac{x-\lambda_2^j}{x-\lambda_2^i}\right|,$$

d'où

$$\|r_{i,j}\|_{L^\infty(E_{n,N})} = \left|r_{i,j}\left(\lambda_1^i\right)\right| \geqslant \frac{\delta_N}{\Delta}\|r_{i,i}\|_{L^\infty(E_{n,N})} = \frac{\delta_N}{\Delta}\left|r_{i,i}\left(\lambda_1^i\right)\right|,$$

où l'on a noté $\delta_N = dist(E_N, F_N)$, et $\Delta := \sup_{N\geqslant 1} \max_{x\in E_N, y\in F_N} |x-y|$ quantité finie d'après l'hypothèse 5.1,

$$\begin{aligned} &\sqrt{\frac{\mathcal{F}\left(\left(\lambda_1^0,\dots,\lambda_1^{i-1},x,\lambda_1^{i+1},\dots,\lambda_1^n\right),\left(\lambda_2^0,\dots,\lambda_2^n\right)\right)}{\mathcal{F}\left(\left(\lambda_1^0,\dots,\lambda_1^n\right),\left(\lambda_2^0,\dots,\lambda_2^n\right)\right)}} \\ = &\prod_{0\leqslant\ell\leqslant n,\ell\neq i}\frac{|x-\lambda_1^\ell|}{|\lambda_1^i-\lambda_1^\ell|}\prod_{0\leqslant\ell\leqslant n,\ell\neq i}\frac{|\lambda_1^i-\lambda_2^\ell|}{|x-\lambda_2^\ell|} = \left|\frac{r_{i,i}(x)}{r_{i,i}(\lambda_1^i)}\right|. \end{aligned}$$

Compte tenu de la définition des points de Fekete rationnels discrets, on a pour $i \in \{0, n\}$,

$$\left|\frac{r_{i,i}(x)}{r_{i,i}(\lambda_1^i)}\right| \leqslant 1,$$

ce qui donne pour tout $x \in E_N$,

$$\left|r_{i,j}\left(\lambda_1^i\right)\right| \geqslant \frac{\delta_N}{\Delta}\left|r_{i,i}(x)\right|.$$

En prenant le produit de ces quantités pour $x \in E_N \setminus E_{n,N}$, on en déduit que

$$\begin{aligned} &\frac{N-n-1}{N^2}\log\left|r_{i,j}\left(\lambda_1^i\right)\right| \geqslant \frac{N-n-1}{N^2}\log\frac{\delta_N}{\Delta} + \frac{1}{N^2}\sum_{x\in E_N\setminus E_{n,N}}\log\left|r_{i,i}(x)\right|, \\ &= \frac{N-n-1}{N^2}\log\frac{\delta_N}{\Delta} - \frac{1}{N^2}\sum_{x\in E_N\setminus E_{n,N}}\Big(\sum_{y\in E_{n,N}\setminus\lambda_1^i}\log\frac{1}{|x-y|} - \sum_{y\in F_{n,N}\setminus\lambda_2^i}\log\frac{1}{|x-y|}\Big), \\ &\geqslant \frac{1}{N}\log\frac{\delta_N}{\Delta} - I^\star\left(\nu_N\left(E_N\setminus E_{n,N}\right),\nu_N\left(E_{n,N}\setminus\{\lambda_1^i\}\right) - \nu_N\left(F_{n,N}\setminus\{\lambda_2^i\}\right)\right). \end{aligned}$$

Maintenant, comme la limite faible-$\star$ d'une suite de mesures de comptage normalisées n'est pas affectée par la suppression d'un point, on obtient d'après le lemme 5.2.6 que

$$\nu_N\left(E_N\setminus E_{n,N}\right) \overset{\star}{\rightharpoonup} \sigma_1 - \mu_1^t,\quad \nu_N\left(E_{n,N}\setminus\{\lambda_1^i\}\right) \overset{\star}{\rightharpoonup} \mu_1^t,\quad \nu_N\left(F_{n,N}\setminus\{\lambda_2^i\}\right) \overset{\star}{\rightharpoonup} \mu_2^t,$$

et le lemme 5.2.5 donne

$$\lim_{n,N\to+\infty,n/N\to t} I^\star\left(\nu_N\left(E_N\setminus E_{n,N}\right),\nu_N\left(E_{n,N}\setminus\{\lambda_1^i\}\right)\right) = I\ \left(\sigma_1-\mu_1^t,\mu_1^t\right),$$
$$\lim_{n,N\to+\infty,n/N\to t} I^\star\left(\nu_N\left(E_N\setminus E_{n,N}\right),\nu_N\left(F_{n,N}\setminus\{\lambda_2^i\}\right)\right) = I\ \left(\sigma_1-\mu_1^t,\mu_2^t\right).$$

On a de plus d'après l'hypothèse 5.5 le fait que $\lim_{N\to+\infty}\delta_N^{1/N}=1$ et on obtient donc d'après les conditions d'équilibre du théorème 4.3.1,

$$\begin{aligned}&(1-t)\liminf_{n,N,n/N\to t}\frac{1}{N}\log\left|r_{i,j}\left(\lambda_1^i\right)\right| \geqslant -I\ \left(\sigma_1-\mu_1^t,\mu^t\right)\\ &=-\int U^{\mu^t}(z)d\left(\sigma_1-\mu_1^t\right)(z)=-(\sigma_1-\mu_1^t)(\mathbb{C})F_1^t=-(1-t)F_1^t,\end{aligned}$$

comme énoncé dans la première partie du lemme 5.2.9.

La seconde partie en découle immédiatement en inversant les rôles joués par les ensembles E_N et F_N. □

5.2.5 Borne inférieure

Tous les éléments sont maintenant réunis pour prouver l'inégalité (5.4) sous les hypothèses 5.1, 5.2, 5.3, 5.4, et 5.5, en voici la preuve.

PREUVE : On a vu précédemment que

$$Z_n^\star(E_{n,N},F_{n,N})=\min_{0\leqslant i,j\leqslant n}\left|\frac{r_{ij}\left(\lambda_1^i\right)}{r_{ij}\left(\lambda_2^j\right)}\right|.$$

En combinant le lemme 5.2.9 et la proposition 5.2.8, on obtient

$$\begin{aligned}\liminf_N Z_n(E_N,F_N)^{1/N} &\geqslant \liminf_N Z_n^*(E_{n,N},F_{n,N})^{1/N},\\ &= \liminf_N \min_{0\leqslant i,j\leqslant n}\|r_{i,j}\|_{L^\infty(E_{n,N})}^{1/N}\left\|r_{i,j}^{-1}\right\|_{L^\infty(F_{n,N})}^{1/N},\\ &\geqslant e^{-\left(F_1^t+F_2^t\right)},\end{aligned}$$

comme annoncé dans (5.4). □

Le prochain chapitre est dédié à l'étude plus approfondie de notre problème de théorie du potentiel, qui permettra de donner des conditions suffisantes pour obtenir une formulation intégrale de la constante extrémale $F_1^t+F_2^t$, et par conséquent une formulation plus explicite et exploitable numériquement.

Chapitre 6

Formulation intégrale pour la constante extrémale

Comme expliqué en fin du chapitre 4 consacré à l'étude de notre problème de minimisation d'énergie en théorie du potentiel logarithmique pour des mesures signées, on cherche dorénavant à étudier plus en profondeur ce problème dans un cas particulier.

Le résultat principal de ce chapitre est donné au théorème 6.3.4 où l'on obtient une formulation intégrale explicite de la constante extrémale $F_1^t + F_2^t$ qui gouverne l'asymptotique faible du problème de Zolotarev pour des ensembles discrets.

On suit dans ce chapitre la démarche de [BuRa99], article lui-même construit dans le but d'obtenir la formule dite *formule de Buyarov-Rakhmanov* issue de [BuRa99, Théorème 2] qui donne une formulation intégrale pour la mesure d'équilibre d'un problème sans contrainte avec champ extérieur.

Pour le cas des mesures contraintes par γ une mesure de masse $T > t$, on considère dans [BeKu01a] le problème de minimisation d'énergie $\left(\widehat{P}_+\right)$ cité dans la preuve du théorème 4.2.17 sur l'ensemble

$$\mathcal{M}_\gamma^t := \{\nu,\ \nu \text{ mesure},\ \nu(\mathbb{C}) = t,\ 0 \leqslant \nu \leqslant \gamma\} \text{ où } 0 < t < T.$$

On note ν^t la mesure minimisante pour ce problème de minimisation d'énergie et on note $S(t) := supp\ (\gamma - \nu^t)$ la partie de $supp\,(\gamma)$ libre de contrainte. On définit enfin $\omega_{S(\tau)}$ la mesure de Robin de l'ensemble $S(\tau)$, autrement dit la mesure minimisante pour l'énergie logarithmique sur l'ensemble

$$\{\lambda \text{ mesure},\ supp\,(\lambda) \subset S(\tau),\ \lambda(\mathbb{C}) = 1\}.$$

On a alors la formulation intégrale suivante donnée dans [BeKu01a, p. 9] pour la mesure d'équilibre, formulation appelée *formule de Buyarov-Rakhmanov modifiée* dans cet article :

$$\nu^t = \int_0^t \omega_{S(\tau)}\, \mathrm{d}\tau. \tag{6.1}$$

Cette formule pour le cas des mesures sous contrainte sans champ extérieur est obtenue grâce à un résultat dit *de dualité champ-contrainte* qui permet de reformuler un problème de minimisation d'énergie sous contrainte sans champ extérieur sous la forme d'un problème de minimisation d'énergie sans contrainte avec champ extérieur. On commence

donc par prouver un résultat de ce type afin d'adapter des travaux de [BuRa99] et [La06] conçus pour des problèmes sans contrainte avec champ extérieur, respectivement pour des mesures et pour un problème de théorie du potentiel vectoriel.

Pour suivre la démarche de la preuve de [BuRa99, Théorème 2] on définit ensuite une fonctionnelle adaptée à la détermination des ensembles $supp\left(sigma_j - \mu_j^t\right)$, $j = 1, 2$, dont on détaille la construction au cours de ce chapitre avant de passer au résultat principal de cette section.

La formulation intégrale étant bien plus explicite dans le cas où les ensembles

$$supp\left(\sigma_j - \mu_j^\tau\right)_{0 \leqslant \tau \leqslant t}, \ j = 1, 2$$

sont des intervalles réels, on donne en fin de chapitre une condition suffisante pour se trouver dans ce cas, condition valable uniquement dans le cas symétrique et adaptée du cas des mesures sous contraintes donné dans [BeKu01a, Lemme 3.1].

On impose dans tout ce chapitre les deux hypothèses supplémentaires suivantes que l'on supposera dorénavant vérifiées. On considère maintenant des contraintes finies σ_1 et σ_2 de même masse sur le support desquelles on impose une restriction de nature topologique.

Hypothèse 6.1. *On suppose dorénavant qu'on se trouve dans le cadre plus restrictif suivant :*

$$E_1 = E_2 = \emptyset, \ \sigma_1(\mathbb{C}) = \sigma_2(\mathbb{C}), \ \text{et } t_1 = t_2 = t \in (0, \sigma_1(\mathbb{C}))\,.$$

Hypothèse 6.2. *On suppose en outre que les ensembles* supp (σ_j) *pour* $j \in \{1, 2\}$ *sont de complémentaire connexe et d'intérieur vide.*

L'hypothèse 6.2 est vérifiée dans le cas d'une contrainte à support contenu dans une union finie d'arcs de Jordan du plan complexe, en particulier dans le cas d'une contrainte à support réel.

6.1 Dualité champ/contrainte

L'objectif de cette section est de reformuler notre problème de minimisation d'énergie sous contrainte sans champ extérieur sous la forme d'un problème de minimisation d'énergie sans contrainte avec champ extérieur. Ceci permet ensuite d'utiliser et d'adapter des outils conçus pour la théorie du potentiel sans contrainte à notre problème de référence. On appelle habituellement ce type de résultat un résultat de *dualité champ/contrainte*, la nouveauté principale de cette section consiste à généraliser ce résultat de dualité au cas des mesures signées, on s'inspire dans ce paragraphe des résultats établis pour le cas des mesures dans [Be06, Lemme 2.6].

Soit Q une fonction continue sur $\mathbb{C}$.

Définition 6.1.1. *Pour $s > 0$, on définit l'ensemble*

$$\mathcal{Q}^s := \{\tilde{\mu} = \tilde{\mu}_1 - \tilde{\mu}_2, \ \tilde{\mu}_i \text{ est une mesure sur } \Sigma_i : \tilde{\mu}_i(\mathbb{C}) = s\},$$

et le problème de minimisation d'énergie suivant

$$(\widetilde{P}) \qquad \text{Trouver } \tilde{\mu}^s \in \mathcal{Q}^s \text{ tel que } \quad I_Q(\tilde{\mu}^s) = \inf\left\{I_Q(\tilde{\mu}), \tilde{\mu} \in \mathcal{Q}^s\right\},$$

avec l'énergie logarithmique tenant compte de l'influence du champ extérieur définie comme précédemment par

$$I_Q(\tilde{\mu}) = I(\tilde{\mu}) + 2\int Q(z)\,\mathrm{d}\tilde{\mu}(z).$$

On s'intéresse dorénavant au champ extérieur donné par $Q = -U^\sigma$, continu d'après l'hypothèse 4.2.

Pour une étude détaillée générale des problèmes de minimisation énergie sans contrainte pour des mesures signées, voir par exemple [SaTo97, Theorème VIII.1.4].

Adaptons maintenant le [SaTo97, Theorème VIII.2.2] au cas particulier de notre problème $(\widetilde{P})$.

Théorème 6.1.2. *Le problème $(\widetilde{P})$ admet une unique solution $\tilde{\mu}^s$ caractérisée par les conditions d'équilibre suivantes :*

$$\begin{cases} U^{\tilde{\mu}^s}(z) + Q(z) \geqslant \widetilde{F}^s_{1,Q}, & z \in \Sigma_1, \\ U^{\tilde{\mu}^s}(z) + Q(z) \leqslant \widetilde{F}^s_{1,Q}, & z \in \mathrm{supp}\,(\tilde{\mu}_1), \\ -U^{\tilde{\mu}^s}(z) - Q(z) \geqslant \widetilde{F}^s_{2,Q}, & z \in \Sigma_2, \\ -U^{\tilde{\mu}^s}(z) - Q(z) \leqslant \widetilde{F}^s_{2,Q}, & z \in \mathrm{supp}\,(\tilde{\mu}_2), \end{cases} \tag{6.2}$$

pour certaines constantes $\widetilde{F}^s_{1,Q}, \widetilde{F}^s_{2,Q}$.

Proposition 6.1.3. *Soient $s, t > 0$ tels que $s + t = \sigma_1(\mathbb{C}) = \sigma_2(\mathbb{C})$.*

Alors, la solution $\mu^t = \mu^t_1 - \mu^t_2$ au problème extrémal sous contrainte sans champ extérieur (P) et la solution $\tilde{\mu}^s = \tilde{\mu}^s_{Q,1} - \tilde{\mu}^s_{Q,2}$ au problème avec champ extérieur $Q = -U^\sigma$ sans contrainte $(\widetilde{P})$ sont liées par la formule suivante :

$$\tilde{\mu}^s = \sigma - \mu^t.$$

PREUVE : Soit $\tilde{\mu} = \sigma - \mu^t = \tilde{\mu}_1 - \tilde{\mu}_2$ avec $\tilde{\mu}_j = \sigma_j - \mu^t_j$ pour $j = 1, 2$.

Remarquons que l'appartenance de μ^t à l'ensemble $\mathcal{M}^t_\sigma$ implique que $\tilde{\mu}$ est un élément de $\mathcal{Q}^s$, et donc en particulier que celle-ci a une énergie logarithmique finie.

En notant $Q(z) = -U^\sigma(z)$ comme précédemment, on observe alors que

$$U^{\tilde{\mu}}(z) + Q(z) = -U^{\mu^t}(z).$$

Ainsi, (6.2) se déduit immédiatement de (4.10) en définissant $\widetilde{F}^s_{j,Q} = -F^t_j$ pour $j = 1, 2$, ce qui prouve bien

$$\tilde{\mu}^s = \sigma - \mu^t.$$

□

Remarque. A propos de la première partie de la preuve précédente, il nous paraît pertinent de mentionner le fait que, d'après [SaTo97, Théorème VIII.2.2], l'assertion réciproque est également vérifiée : la solution $\tilde{\mu} = \tilde{\mu}^s$ au problème extrémal $(\widetilde{P})$ avec champ extérieur $Q = -U^\sigma$ vérifie toujours les conditions d'équilibre (6.2), où les constantes extrémales $\widetilde{F}^s_{1,-U^\sigma}, \widetilde{F}^s_{2,-U^\sigma}$ sont uniques, car d'après la proposition 6.1.3, on a $\widetilde{F}^s_{j,-U^\sigma} = -F^t_j$ pour $j \in \{1, 2\}$), et l'unicité des F^t_j a été établie précédemment dans le cadre de la démonstration du théorème 4.3.1.

Notre proposition 6.1.3 nous permet alors d'affirmer des propriétés *a priori* inattendues pour ce problème avec le champ extérieur $Q = -U^\sigma$: on sait maintenant que la première et la troisième inégalité du système (6.2) sont valables pour tout $z \in \mathbb{C}$, ou encore que $\tilde{\mu}^s_j \leqslant \sigma_j$ pour $j = 1, 2$, ce qui découle également de [SaTo97, Théorème IV.4.5].

6.2 $\mathfrak{F}_t$-fonctionnelle de Mhaskar-Saff-Rakhmanov

On remarque dans le théorème 4.3.1 que les conditions d'équilibre (4.17) font intervenir les ensembles *supp* $(\sigma_j - \mu_j)$, $j = 1, 2$, qui *ne sont pas connus à l'avance.*

Ces ensembles joueront un rôle crucial par la suite pour le choix des paramètres pour l'asymptotique faible du problème de Zolotarev pour des ensembles discrets. Il s'agit en effet des ensembles en lesquels la contrainte σ n'est pas atteinte, ce qui justifie au moins intuitivement l'attention particulière à porter à ces ensembles. Ce type de situation de recherche de support de mesure extrémale se produit en théorie du potentiel logarithmique dans un contexte bien plus général que notre étude : dans le cas d'un problème sans contrainte avec champ extérieur pour des mesures positives, on est amené à chercher à déterminer le support de la mesure d'équilibre.

En l'absence de champ extérieur, il existe des propriétés de localisation du support d'une mesure extrémale, celui-ci se situe par exemple sur la frontière extérieure du compact considéré. Cependant, ces propriétés ne s'adaptent pas au cas d'un champ extérieur non nul, ce qui correspond tout à fait à l'intution électrostatique que l'on peut avoir de la situation, d'où la nécessité de créer un outil voué à la détermination du support de la mesure d'équilibre.

L'outil le plus fréquemment utilisé à cet effet, la $\mathfrak{F}$-*fonctionnelle* a été considéré pour la première fois dans [MhSa85] et permet de caractériser aux ensembles de capacité logarithmique nulle près le support de la mesure extrémale pour un problème de théorie du potentiel logarithmique sous la forme d'un problème de minimisation de ladite fonctionnelle.

Cet outil a ensuite été développé et généralisé à un contexte bien plus large, on citera par exemple l'article de Buyarov et Rakhmanov [BuRa99, Lemme 1] dont on s'inspirera largement de la preuve dans cette section, ou encore [SaTo97, Théorème IV.1.5]. De plus, [LeLu01] étudie celle-ci du point de vue de la théorie du potentiel de Green, point de vue qui s'adapte à notre étude dans le cas d'une mesure contrainte dont la densité est impaire.

Pour commencer ce travail d'adaptation de la preuve de [BuRa99, Lemme 1], citons tout d'abord une partie du théorème [SaTo97, Théorème VIII.2.6] adapté à notre contexte pour définir la mesure d'équilibre d'un condensateur à plateaux K_1 et K_2.

Proposition 6.2.1. *Soient K_1 et K_2 deux compacts disjoints réguliers du plan complexe de capacité logarithmique strictement positive. Le problème de minimisation d'énergie suivant admet une unique solution notée ω_{K_1,K_2} et appelée mesure d'équilibre de K_1, K_2 :*

$$\textit{Trouver } \omega_{K_1,K_2} \in \mathcal{M}^{1,1}(K_1,K_2) \textit{ telle que } \quad I(\omega_{K_1,K_2}) = \inf\left(I(\omega),\ \omega \in \mathcal{M}^{1,1}(K_1,K_2)\right),$$

où

$$\mathcal{M}^{1,1}(K_1,K_2) := \{\omega := \omega_1 - \omega_2,\ \omega_j \textit{ mesure de probabilité},\\ \operatorname{supp}(\omega_j) \subset K_j \textit{ pour } j \in \{1,2\}\}$$

d'après la définition 4.1.14.

La mesure ω_{K_1,K_2} est caractérisée par les conditions d'équilibre

$$\begin{cases} U^{\omega_{K_1,K_2}}(z) = \phi_1, & qp \text{ sur } K_1, \\ U^{\omega_{K_1,K_2}}(z) \leqslant \phi_1, & z \in \mathbb{C}, \\ -U^{\omega_{K_1,K_2}}(z) = \phi_2, & qp \text{ sur } K_2, \\ -U^{\omega_{K_1,K_2}}(z) \leqslant \phi_2, & z \in \mathbb{C}, \end{cases} \tag{6.3}$$

vérifiées par un couple de constantes réelles (ϕ_1, ϕ_2).

On peut également définir la capacité au sens d'un condensateur en théorie du potentiel logarithmique, définition à relier à 4.1.9.

Définition 6.2.2. *En utilisant les notations précédentes, on définit alors la* capacité du condensateur *à plateau positif K_1 et négatif K_2 par*

$$\operatorname{cap}(K_1, K_2) := \frac{1}{\phi_1 + \phi_2}.$$

La définition suivante reprend un concept défini précedemment dans le cas d'un domaine de $\mathbb{C}$, et nous sera nécessaire pour la définition de la $\mathfrak{F}$-fonctionnelle.

Définition 6.2.3. *On dit qu'un compact du plan complexe est régulier si son complémentaire est régulier au sens du problème de Dirichlet, notion définie en 4.2.9.*

On a alors la propriété suivante tirée de [Ra95, Définition 3.8.1 et Théorème 4.2.4] pour un compact régulier.

Proposition 6.2.4. *Un compact régulier du plan complexe n'admet pas de point isolé.*

Proposition 6.2.5. *Si (K_1, K_2) est un couple de compacts réguliers disjoints du plan complexe, tous deux de complémentaire connexe et d'intérieur vide, alors* $\operatorname{supp}(\omega_{K_1,K_2}) = K_1 \cup K_2$.

PREUVE : Pour commencer, rappelons que le potentiel $U^{\omega_{K_1,K_2}}$ est une fonction continue sur $\mathbb{C}$ d'après 4.2.14. Ainsi, les conditions d'équilibre vérifiées par ω_{K_1,K_2} se réécrivent

$$\begin{cases} U^{\omega_{K_1,K_2}}(z) = \phi_1, & z \in K_1, \\ U^{\omega_{K_1,K_2}}(z) \leqslant \phi_1, & z \in \mathbb{C}, \\ -U^{\omega_{K_1,K_2}}(z) = \phi_2, & z \in K_2, \\ -U^{\omega_{K_1,K_2}}(z) \leqslant \phi_2, & z \in \mathbb{C}, \end{cases} \tag{6.4}$$

Par définition de la mesure ω_{K_1,K_2}, on a l' inclusion $supp\,(\omega_{K_1,K_2}) \subset K_1 \cup K_2$. Supposons cette inclusion stricte, par exemple supposons qu'il existe $z_0 \in K_1 \cup K_2 \setminus supp\,(\omega_{K_1,K_2})$, par exemple $z_0 \in K_1$.

D'après le principe du maximum 4.1.4, comme K_1 est de complémentaire connexe et d'intérieur vide, $U^{\omega_{K_1,K_2}}$ est sous-harmonique sur $\mathbb{C} \setminus K_1$, on a $U^{\omega_{K_1,K_2}}(z_0) < \phi_1$, ce qui contredit les conditions d'équilibre (6.4). On raisonne en supposant que $z_0 \in K_2$, ce qui termine la preuve. □

Nous sommes maintenant à même de définir la $\mathfrak{F}$-*fonctionnelle* de Mhaskar-Saff-Rakhmanov dans une formulation adaptée à notre étude.

Définition 6.2.6. *Pour $j \in \{1,2\}$, on note $S_1^t := \operatorname{supp}(\sigma_1 - \mu_1^t)$, $S_2^t := \operatorname{supp}(\sigma_2 - \mu_2^t)$ et $S^t := S_1^t \cup S_2^t$.*

Pour deux compacts disjoints du plan complexe K_1 et K_2, on définit la $\mathfrak{F}$-fonctionnelle

$$\mathfrak{F}_t(K_1,K_2) := (\sigma_1(\mathbb{C}) - t)\frac{1}{\operatorname{cap}(K_1,K_2)} - \int U^\sigma \,\mathrm{d}\omega_{K_1,K_2}(z).$$

On prouve maintenant que les ensembles en lesquels la $\mathfrak{F}$-fonctionnelle atteint son minimum sont donnés à ensemble de capacité nulle près par *supp* $\left(\sigma_j - \mu_j^t\right)$, pour $j \in \{1,2\}$, ce qui justifie la construction et le rôle de la $\mathfrak{F}$-fonctionnelle dans la suite de notre travail. Le but final de cette étude de fonctionnelle est d'utiliser des techniques classique de minimisation afin de déterminer ces ensembles.

Définissons les ensembles $\widehat{S_1^t} := \{z \in \Sigma_1 : U^{\mu^t} = F_1^t\}$ et $\widehat{S_2^t} := \{z \in \Sigma_2 : U^{\mu^t} = -F_2^t\}$.

La prochaine définition n'a pour objet qu'un allégement des notations.

Définition 6.2.7. *On notera dorénavant*

$$\omega^t := \omega_{S_1^t,S_2^t}$$

la mesure d'équilibre du condensateur (S_1^t, S_2^t).

Lemme 6.2.8. *On a les inclusions*

$$S_1^t \subset \widehat{S_1^t} \quad et \quad S_2^t \subset \widehat{S_2^t}.$$

PREUVE : Les inclusions énoncées résultent directement des conditions d'équilibre (4.10) vérifiées par la mesure minimisante μ^t. □

Proposition 6.2.9. *Notons $F^t := F_1^t + F_2^t$.*

On a $\mathfrak{F}_t(S_1^t, S_2^t) = -F^t$ et pour tout compacts réguliers disjoints du plan complexe K_1 et K_2 tels que pour $j \in \{1,2\}$, $K_j \subset \Sigma_j$,

$$\mathfrak{F}_t(K_1,K_2) \geqslant \mathfrak{F}_t(S_1^t,S_2^t).$$

De plus,

$$\mathfrak{F}_t(K_1,K_2) = \mathfrak{F}_t(S_1^t,S_2^t)$$

si et seulement si les inclusions

$$S_1^t \subset K_1 \subset \widehat{S_1^t} \quad et \quad S_2^t \subset K_2 \subset \widehat{S_2^t}$$

sont vérifiées.

PREUVE : Soient K_1 et K_2 deux compacts réguliers disjoints du plan complexe tels que pour $j \in \{1,2\}$, $K_j \subset \Sigma_j$.

Notons $\tilde{\mu} = \tilde{\mu}_Q^s$ où $s = \sigma_1(\mathbb{C}) - t$ la mesure solution du problème de minimisation $(\widetilde{P})$ défini dans le cadre de la dualité champ/contrainte lors de la proposition 6.1.3.

Montrons pour commencer que

$$\mathfrak{F}_t(K_1,K_2) \geqslant -F^t.$$

D'après la proposition 6.1.3 en conservant les mêmes notations que précédemment,

$\tilde{\mu} = \sigma - \mu^t$. En intégrant la première inégalité des conditions d'équilibre (6.2) par rapport à la mesure $d\omega_1$ où $\omega_{K_1,K_2} = \omega_1 - \omega_2$ est définie en 6.2.1, on obtient

$$\int U^{\tilde{\mu}} d\omega_1 - \int U^{\sigma} \, \mathrm{d}\omega_1 \geqslant -F_1^t.$$

On applique deux fois le théorème de Fubini, ce qui est licite car les fonctions $U^{\tilde{\mu_j}}$, $j \in \{1,2\}$ sont continues sur le compact K_1, et on a alors

$$\begin{aligned} \int U^{\tilde{\mu}} \, \mathrm{d}\omega_1 &= \int U^{\tilde{\mu}_1} \, \mathrm{d}\omega_1 - \int U^{\tilde{\mu}_2} \, \mathrm{d}\omega_1, \\ &= \int U^{\omega_1} \, \mathrm{d}\tilde{\mu}_1 - \int U^{\omega_1} \, \mathrm{d}\tilde{\mu}_2, \\ &= \int U^{\omega_1} \, \mathrm{d}\tilde{\mu}, \end{aligned}$$

d'où

$$\int U^{\omega_1} \, \mathrm{d}\tilde{\mu} - \int U^{\sigma} \, \mathrm{d}\omega_1 \geqslant -F_1^t.$$

On raisonne de même à partir de la troisième inégalité des conditions d'équilibre (6.2), d'où

$$-\int U^{\omega_2} \, \mathrm{d}\tilde{\mu} + \int U^{\sigma} \, \mathrm{d}\omega_2 \geqslant -F_2^t,$$

et comme $\omega_{K_1,K_2} = \omega_1 - \omega_2$,

$$\int U^{\omega_{K_1,K_2}} \, \mathrm{d}\tilde{\mu} - \int U^{\sigma} \, \mathrm{d}\omega_{K_1,K_2} \geqslant -F^t.$$

D'après les conditions d'équilibre (6.3) énoncées proposition 6.2.1,

$$\begin{aligned} \int U^{\omega_{K_1,K_2}} \, \mathrm{d}\tilde{\mu} &\leqslant \phi_1 \tilde{\mu}_1(\mathbb{C}) + \phi_2 \tilde{\mu}_2(\mathbb{C}), \\ &= \frac{1}{cap\ (K_1,K_2)} (\sigma_1(\mathbb{C}) - t), \end{aligned}$$

d'où

$$\mathfrak{F}_t(K_1,K_2) \geqslant -F^t.$$

Montrons maintenant que $\mathfrak{F}_t(S_1^t, S_2^t) = -F^t$.

Les deux premières inégalités des conditions d'équilibre (6.2) donnent

$$U^{\tilde{\mu}_Q^t}(z) - U^{\sigma}(z) = -F_1^t \quad \forall z \in S_1^t,$$

d'où

$$\int U^{\omega_1^t} \, \mathrm{d}\tilde{\mu} - \int U^{\sigma} \, \mathrm{d}\omega_1^t = -F_1^t.$$

On intègre par rapport à $\omega_{S^t,1}$ partie positive de la mesure d'équilibre du condensateur S_1^t, S_2^t, et d'après les deux dernières inégalités des conditions d'équilibre, on obtient

$$\int U^{\tilde{\mu}} \, \mathrm{d}\omega^t - \int U^{\sigma} \, \mathrm{d}\omega^t = -F^t.$$

Le potentiel U^{ω^t} est constant quasi-partout sur S_1^t et S_2^t, d'après les conditions d'équilibre (6.3) vérifiées par ω^t, et d'après le théorème de Fubini,

$$\int U^{\tilde{\mu}} \, \mathrm{d}\omega^t = \int U^{\omega^t} \, \mathrm{d}\tilde{\mu} = \frac{1}{cap\ (S_1^t, S_2^t)} (\sigma_1(\mathbb{C}) - t),$$

on en déduit que

$$\mathfrak{F}_t(S_1^t, S_2^t) = -F^t.$$

Si on suppose pour $i \in \{1, 2\}$ l'inclusion $S_i^t \subset K_i \subset \widehat{S_i^t}$ vérifiée, en raisonnant comme précédemment à partir des conditions d'équilibre, on en déduit que

$$\mathfrak{F}_t\,(K_1, K_2) = -F^t.$$

Supposons maintenant que l'on ait l'égalité pour la $\mathfrak{F}$-fonctionnelle

$$\mathfrak{F}_t\,(K_1, K_2) = -F^t$$

et que l'une des inclusions $S_i^t \subset K_i \subset \widehat{S_i^t}$ ne soit pas vérifiée, par exemple pour $i = 1$.

Si S_1^t n'est pas contenu dans K_1, on choisit un élément $z_0 \in S_1^t \backslash K_1$, et tous les ensembles considérés étant compacts, il existe un réel ε positif tel que la boule

$$\mathcal{B}(z_0, \varepsilon) := \{z \in \mathbb{C}, |z - z_0| < \varepsilon\}$$

soit d'intersection avec K_1 vide.

On rappelle que par hypothèse 6.2, Σ_1 est de complémentaire connexe et d'intérieur vide, ce qui est également le cas de K_1 car $K_1 \subset \Sigma_1$. D'après le principe du maximum 4.1.4, comme U^{ω_K} est sous-harmonique sur $\mathbb{C} \setminus K_1$,

$$U^{\omega_K}(z) < \phi_1 \text{ pour } z \in \mathcal{B}(z_0, \varepsilon),$$

et par choix de z_0,

$$(\sigma_1 - \mu_1^t)\,(\mathcal{B}(z_0, \varepsilon)) > 0,$$

ce qui donne l'inégalité stricte

$$\int U^{\omega_{K_1,K_2}} \, \mathrm{d}(\sigma_1 - \mu_1^t) < \phi_1(\sigma_1(\mathbb{C}) - t),$$

et contredit l'inégalité $\mathfrak{F}_t\,(K_1, K_2) = -F^t$.

On suppose maintenant que K_1 n'est pas contenu dans $\widehat{S_1^t}$. On choisit alors $z_0 \in K_1 \backslash \widehat{S_1^t}$. Par continuité de U^{μ^t}, il existe $\varepsilon > 0$ tel que

$$U^{\mu^t}(z) < F_1^t \quad \forall z \in \mathcal{B}(z_0, \varepsilon).$$

D'après la proposition 6.2.5, $z_0 \in supp\ (\omega_{K_1,K_2})$ et par conséquent,

$$\omega_{K_1,K_2}\,(\mathcal{B}(z_0, \varepsilon) \cap K_1) > 0.$$

Ainsi,

$$-\int U^{\mu^t} \, \mathrm{d}\omega_1 > -F_1^t,$$

et

$$\int U^{\sigma-\mu^t}\,\mathrm{d}\omega_1 - \int U^{\sigma}\,\mathrm{d}\omega_1 > -F_1^t,$$

ce qui permet d'en déduire que

$$\mathfrak{F}_t(K_1,K_2) > -F^t,$$

d'où une contradiction.

On montre de même dans le cas où l'on suppose que l'une des inclusions $S_2^t \subset K_2 \subset \widehat{S_2^t}$ n'est pas vérifiée que nécessairement

$$\mathfrak{F}_t(K_1,K_2) > -F^t,$$

ce qui termine la preuve de la proposition. □

6.3 Formulation intégrale pour la constante extrémale

On donne dans cette section le résultat majeur de ce chapitre permettant d'obtenir une formulation intégrale pour les constantes extrémales mises en jeu dans notre problème de théorie du potentiel décrit dans le chapitre précédent afin d'expliciter la constante $F_1^t + F_2^t$ qui donne l'asymptotique faible du problème de Zolotarev pour des ensembles discrets. Nous nous inspirerons principalement ici de l'article de Lapik [La06], notre travail consiste à adapter les preuves issues de cette article à notre langage de problème sous contrainte pour des mesures signées.

Voici dans cette direction un résultat de monotonie concernant les solutions du problème de minimisation $(\widetilde{P})$, résultat adapté de [La06, Proposition 1]. Ce résultat technique nous permettra d'obtenir un autre résultat de monotonie au sens de l'inclusion concernant les supports d'une famille de mesures d'équilibre $\left(supp\,\left(\sigma_j - \mu_j^\tau\right)\right)_{0\leqslant\tau\leqslant t}$, $j=1,2$, résultat capital dans la preuve de la formulation intégrale pour la constante extrémale.

Lemme 6.3.1. *Avec les notations précédentes, on a*

$$0 < s_1 < s_2 \;\Rightarrow\; \widetilde{\mu}^{s_1} < \widetilde{\mu}^{s_2}.$$

On donne ici des éléments de preuve de ce lemme dont il sera fait usage ultérieurement dans la preuve de la proposition 6.3.2.

PREUVE : On note $\varepsilon := s_2 - s_1 > 0$ et α^ε la mesure d'équilibre pour le problème avec champ extérieur $\widetilde{Q} := Q + U^{\widetilde{\mu}^{s_1}}$ sur $\mathcal{Q}^\varepsilon$. Il nous suffit de prouver la relation suivante pour conclure :

$$\nu := \widetilde{\mu}^{s_1} + \alpha^\varepsilon = \widetilde{\mu}^{s_2}.$$

Pour prouver cette égalité, remarquons tout d'abord que $\nu \in \mathcal{Q}^{s_2}$, et considérons les conditions d'équilibre vérifiées par U^{α^ε} :

$$\left\{\begin{array}{ll} U^{\alpha^\varepsilon}(z) + U^{\tilde{\mu}^{s_1}}(z) + Q(z) \geqslant F_1^\varepsilon, & z\in\Sigma_1,\\ U^{\alpha^\varepsilon}(z) + U^{\tilde{\mu}^{s_1}}(z) + Q(z) \leqslant F_1^\varepsilon, & z\in supp\,(\alpha_1^\varepsilon),\\ -U^{\alpha^\varepsilon}(z) - U^{\tilde{\mu}^{s_1}}(z) - Q(z) \geqslant F_2^\varepsilon, & z\in\Sigma_2,\\ -U^{\alpha^\varepsilon}(z) - U^{\tilde{\mu}^{s_1}}(z) - Q(z) \leqslant F_2^\varepsilon, & z\in supp\,(\alpha_2^\varepsilon), \end{array}\right. \tag{6.5}$$

pour certaines constantes $F_1^\varepsilon, F_2^\varepsilon$.

Il nous suffit alors de montrer l'inclusion $supp\ (\widetilde{\mu}^{s_1}) \subset supp\ (\alpha^\varepsilon)$ pour en conclure que ν vérifie les conditions d'équilibre caractérisant $\widetilde{\mu}^{s_2}$.

On prouvera simplement $supp\ (\widetilde{\mu}_1^{s_1}) \subset supp\ (\alpha_1^\varepsilon)$, la preuve de l'autre inclusion étant similaire.

On remarque alors que la fonction U^{α^ε} est sous-harmonique sur $\mathbb{C} \setminus supp\ (\alpha_1^\varepsilon)$, et par conséquent, pour $z \in \mathbb{C} \setminus supp\ (\alpha_1^\varepsilon)$ on a

$$\begin{aligned} U^{\alpha^\varepsilon}(z) &< \sup_{x \in \mathbb{C} \setminus supp\left(\alpha_1^\varepsilon\right)} \left(U^{\alpha^\varepsilon}(x)\right), \\ &\leqslant \sup_{x \in supp\left(\alpha_1^\varepsilon\right)} \left(U^{\alpha^\varepsilon}(x)\right), \\ &= F_1^\varepsilon - \inf_{x \in supp\left(\alpha_1^\varepsilon\right)} \left(U^{\widetilde{\mu}^{s_1}}(x) + Q(x)\right), \\ &\leqslant F_1^\varepsilon - \widetilde{F}_{1,Q}^{s_1}, \end{aligned}$$

où la première et la deuxième inégalité proviennent du principe du minimum pour les fonctions sous-harmoniques 4.1.4, et les autres relations découlent des conditions d'équilibre vérifiés par U^{α^ε} et $U^{\widetilde{\mu}^{s_1}}$.

Finalement, sur $supp\ (\widetilde{\mu}_1^{s_1}) \setminus supp\ (\alpha_1^\varepsilon)$ on a

$$U^{\alpha^\varepsilon}(z) + U^{\tilde{\mu}^{s_1}}(z) + Q(z) < F_1^\varepsilon - \widetilde{F}_{1,Q}^{s_1} + \widetilde{F}_{1,Q}^{s_1},$$

ce qui achève notre preuve. □

Proposition 6.3.2. *Sous les hypothèses précédentes, on a pour tout $t, \delta > 0$ tels que $t + \delta < \sigma_1(\mathbb{C})$ et pour $j \in \{1, 2\}$ la chaîne d'inclusions*

$$S_j^{t+\delta} \subset \{z \in \mathbb{C} : U^{\mu^{t+\delta}}(z) = (-1)^{j-1} F_j^t\} \subset S_j^t \subset \{z \in \mathbb{C} : U^{\mu^t}(z) = (-1)^{j-1} F_j^t\}. \quad (6.6)$$

On donne des éléments de preuve de cette proposition adaptés de [La06].

PREUVE : Il suffit de prouver les inclusions

$$\{z \in \mathbb{C} : U^{\mu^{t+\delta}}(z) = (-1)^{j-1} F_j^t\} \subset S_j^t \ j = 1, 2$$

pour conclure, on démontrera uniquement l'inclusion écrite ci-dessus pour $j = 1$, la preuve pour $j = 2$ est similaire.

On utilise les notations de la preuve du lemme 6.3.1, avec $Q = -U^\sigma$, $s_2 := \sigma_1(\mathbb{C}) - t$ et $s_1 := \sigma_1(\mathbb{C}) - t - \delta$, on a alors d'après 6.1.3

$$\tilde{\mu}^{s_1} = \sigma - \mu^{t+\delta},\ \tilde{\mu}^{s_2} = \sigma - \mu^t,$$

et il suffit ainsi de prouver l'inclusion

$$\{z \in \mathbb{C} : U^{\tilde{\mu}^{s_1}}(z) - U^\sigma(z) = F_1^t\} \subset supp\ (\tilde{\mu}_1^{s_2}).$$

On a établi lors de la preuve du lemme 6.3.1 l'inégalité

$$U^{\tilde{\mu}^{s_2}}(z) - U^\sigma(z) < F_1^\varepsilon,\ z \in \{z \in \mathbb{C} : U^{\tilde{\mu}^{s_1}}(z) - U^\sigma(z) = F_1^t\} \cap (\mathbb{C} \setminus supp\ (\tilde{\mu}_1^{s_2})),$$

ce qui prouve d'après les conditions d'équilibre vérifiées par $U^{\tilde{\mu}^{s_2}}$ que

$$\{z \in \mathbb{C} : U^{\tilde{\mu}^{s_1}}(z) - U^{\sigma}(z) = F_1^t\} \cap (\mathbb{C} \setminus supp\ (\tilde{\mu}_1^{s_2})) = \emptyset,$$

d'où le résultat. □

Le lemme de régularité suivant est utile dans la démonstration de la formulation intégrale pour la constante extrémale de notre problème de minimisation d'énergie sous contrainte.

Lemme 6.3.3. *La fonction*

$$t \mapsto \text{cap}\ \left(S_1^t, S_2^t\right)$$

définie sur $(0, \sigma_1(\mathbb{C}))$ *est continue sauf sur un ensemble de points au plus dénombrable.*

PREUVE : La chaîne d'inclusions (6.6) énoncée à la proposition 6.3.2 nous donne pour $\varepsilon > 0$ assez petit et $j = 1, 2$

$$supp\ \left(\omega_j^{t+\varepsilon}\right) \subset S_j^t,$$

et par conséquent, par définition de la mesure d'équilibre, $I\left(\omega^{t+\varepsilon}\right) \geqslant I\left(\omega^t\right)$. Or,

$$I\left(\omega^t\right) = \int U^{\omega^t}(z)\, \mathrm{d}\omega^t(z) = \frac{1}{cap\ (S_1^t, S_2^t)}$$

et la fonction $t \mapsto cap\ (S_1^t, S_2^t)$ est donc décroissante en t. Ainsi, l'ensemble de points de discontinuité sur $(0, \sigma_1(\mathbb{C}))$ de cette fonction que l'on note D est au plus dénombrable. □

Voici maintenant le théorème principal de cette section, théorème par la suite essentiel pour donner des valeurs explicites pour le comportement asymptotique faible de la quantité de Zolotarev. On suit dans la preuve de ce théorème la preuve de [BuRa99, Théorème 2], preuve qui repose sur d'élégants arguments de convexité pour la mesure et la constante extrémale.

Théorème 6.3.4. *Pour* $t \in (0, \sigma_1(\mathbb{C}))$ *on a la formulation intégrale suivante :*

$$F^t = \int_0^t \frac{\mathrm{d}\tau}{\text{cap}\ \left(supp\left(\sigma_1 - \mu_1^\tau\right), supp\left(\sigma_2 - \mu_2^\tau\right)\right)}. \tag{6.7}$$

Remarque. Cette représentation intégrale devient explicite en terme de fonctions elliptiques de Legendre dès lors que l'on peut affirmer que les ensembles $supp\left(\sigma_j - \mu_j^\tau\right)$ sont des intervalles réels, ce qui donnera lieu à des calculs explicites chapitre 7.

PREUVE : Soient t, ε, η tels que $0 < t - \eta < t < t + \varepsilon < \sigma_1(\mathbb{C})$.

On a alors

$$\frac{F^t - F^{t-\eta}}{\eta} \leqslant \frac{1}{cap\ (S_1^t, S_2^t)} \leqslant \frac{F^{t+\varepsilon} - F^t}{\varepsilon}. \tag{6.8}$$

En effet, d'après la proposition 6.3.2, on a

$$\begin{aligned} -F^{t+\varepsilon} &= \mathfrak{F}_{t+\varepsilon}(S_1^{t+\varepsilon}, S_2^{t+\varepsilon}) \leqslant \mathfrak{F}_{t+\varepsilon}(S_1^t, S_2^t), \\ &= \mathfrak{F}_t(S_1^t, S_2^t) - \frac{\varepsilon}{cap\,(S_1^t, S_1^t)} = -F^t - \frac{\varepsilon}{cap\,(S_1^t, S_2^t)}, \end{aligned}$$

et de même,

$$-F^{t-\eta} \leqslant \mathfrak{F}_{t-\eta}(S_1^t, S_2^t) = -F^t + \frac{\eta}{cap\,(S_1^t, S_2^t)},$$

ce qui prouve bien 6.8, et prouve ainsi la convexité de la fonction

$$\phi : (0, \sigma_1(\mathbb{C})) \ni t \mapsto F^t.$$

De plus, avec les notations de la démonstration du lemme 6.3.3, on sait d'après (6.8) que pour $t \notin D$, ϕ est différentiable en t, avec

$$\phi'(t) = \frac{\partial F^t}{\partial t} = \frac{1}{cap\ (S_1^t, S_2^t)}.$$

Comme la fonction ϕ est convexe, elle est absolument continue et il ne nous reste maintenant qu'à montrer que

$$\lim_{t \to 0^+} \phi(t) = 0$$

pour prouver l'identité (6.7) énoncée théorème 6.3.4.

Pour démontrer la limite $\phi(0+) = 0$, on remarque que d'après (4.10),

$$\begin{aligned} I\left(\mu^t, \sigma - \mu^t\right) &= \int U^{\mu^t} d(\sigma_1 - \mu_1^t) - \int U^{\mu^t} d(\sigma_2 - \mu_2^t), \\ &= (\sigma_1(\mathbb{C}) - t)\, F^t. \end{aligned}$$

On a de plus

$$\begin{aligned} 0 &\leqslant (\sigma_1(\mathbb{C}) - t)\, F^t \leqslant \left| I\left(\mu^t, \sigma - \mu^t\right)\right|, \\ &\leqslant \left|I\left(\mu^t, \mu^t\right)\right| + \left|I\left(\mu^t, \sigma\right)\right| \leqslant \left|I\left(\mu^t, t\frac{\sigma}{\sigma_1(\mathbb{C})}\right)\right| + \left|I\left(\mu^t, \sigma\right)\right|, \\ &\leqslant \left(1 + \frac{t}{\sigma_1(\mathbb{C})}\right)\left|I\left(\mu^t, \sigma\right)\right| \leqslant 4t \sup_{z \in \mathbb{C}} |U^\sigma(z)|, \end{aligned}$$

ce qui fournit la relation $F^t \to 0$ pour $t \to 0+$ et termine par là-même la preuve. □

Remarque. Le théorème 6.3.4 met en valeur l'importance de la détermination des ensembles $supp\,(\sigma_j,\, -\mu_j^t)$.

En réalité, ceux-ci sont très difficiles à déterminer en pratique et peuvent même avoir une structure de type Cantor.

Pour comprendre comment construire de tels exemples, considérons $S_j(t)$ pour $0 < t < T$ et $j = 1, 2$ des sous-ensembles compacts de l'axe réel (que l'on suppose tout de même réguliers par rapport au problème de Dirichlet pour assurer une certaine régularité), décroissants en t avec $S_1(t) \cap S_2(t)$ vide et continus au sens où la fermeture de $\cup_{\tau > t} S_j(\tau)$ est égale à $S_j(t)$.

Soit une mesure contrainte donnée par

$$\sigma = \int_0^T \omega_{S_1(t), S_2(t)} \mathrm{d}t,$$

ce qui signifie que pour tout borélien A de $\mathbb{R}^2$, on a

$$\sigma(A) = \int_0^T \omega_{S_1(t), S_2(t)}(A) \mathrm{d}t,$$

où l'intégrande est monotone. On a alors $supp\,(\sigma_j - \mu_j^t) = S_j(t)$ pour tout $0 < t < T$ et $j = 1, 2$.

Pour prouver cette affirmation, soit $\mu = \int_0^t \omega_{S_1(\tau),S_2(\tau)}\, \mathrm{d}\tau$.

Alors, $\mu \in \mathcal{M}_\sigma^t$, avec $supp\,(\sigma_j - \mu_j) = S_j(t)$. Pour calculer son potentiel logarithmique, on trouve d'après le théorème de Fubini

$$\begin{aligned} U^\mu(z) &= \int \log \frac{1}{|z-x|}\, \mathrm{d}\mu(x) \\ &= \int_0^t \mathrm{d}\tau \int \log \frac{1}{|z-x|}\, \mathrm{d}\omega_{S_1(\tau),S_2(\tau)}(x) \\ &= \int_0^t U^{\omega_{S_1(\tau),S_2(\tau)}}(z)\, \mathrm{d}\tau \end{aligned}$$

et les relations (6.3) ainsi que la structure des ensembles $S_j(\tau)$ nous permettent de vérifier les conditions d'équilibre (4.10), ce qui prouve d'après le théorème 4.3.1 que $\mu = \mu^t$.

6.4 Conditions suffisantes pour le cas de deux intervalles réels

Toujours dans le souci d'obtenir une formulation intégrale de la constante extrémale la plus explicite possible, on cherche à établir des conditions suffisantes qui garantissent qu'on se trouve dans le cas où les ensembles $supp\left(\sigma_j - \mu_j^\tau\right)$ sont des intervalles réels. On adapte pour ceci des conditions établies dans [BeKu01a, Lemme 3.1] pour le cas de mesures sous contraintes.

L'inconvénient majeur de ce procédé (l'adaptation du cas des mesures) est qu'il nous limite pour le moment aux mesures contraintes à densité impaire, autrement dit du point de vue de la méthode ADI au cadre de la résolution approchée d'une équation de Lyapounov plutôt que d'une équation de Sylvester. Malheureusement, nous ne parvenons pas pour le moment à établir de condition dans le cas d'une densité non symétrique.

Définition 6.4.1. *On dit que l'on se trouve dans le* cas de deux intervalles réels *lorsque la contrainte σ est supportée par l'axe réel et que les ensembles $S_j^\tau := \mathrm{supp}\,(\sigma_j - \tau\mu_j^\tau)$, sont des intervalles réels compacts pour $j \in \{1,2\}$:*

$$S_1^\tau = [a_\tau, b_\tau] \text{ et } S_2^\tau = [c_\tau, d_\tau], \text{ pour } 0 \leqslant \tau \leqslant t.$$

On dit que l'on se trouve dans le cas de deux intervalles réels symétriques *lorsque l'on se trouve dans le cas de deux intervalles réels et que les ensembles S_j^τ sont symétriques par rapport à l'origine pour $j = 1, 2$ et $0 \leqslant \tau \leqslant t$.*

Remarque. On rappelle que d'après la proposition 6.3.2, la famille d'ensembles $(S(\tau))_{0<\tau<t}$ est décroissante au sens de l'inclusion, ce qui prouve que les fonctions $\tau \mapsto a_\tau$, $\tau \mapsto c_\tau$ sont croissantes sur $(0,t)$ et les fonctions $\tau \mapsto b_\tau$, $\tau \mapsto d_\tau$ décroissantes sur $(0,t)$.

Lemme 6.4.2. *On suppose que l'on se trouve dans le cas de deux intervalles réels et que la densité σ admet une densité impaire.*

Alors, on se trouve dans le cas de deux intervalles réels symétriques.

PREUVE : Si la contrainte σ supportée par l'axe réel a une densité impaire, par unicité de la mesure d'équilibre μ^t du théorème 4.3.1, on en déduit que μ^t admet également une

densité impaire, d'où avec les notation précédentes,

$$S_1^\tau = -S_2^\tau \text{ pour } 0 \leqslant \tau \leqslant t.$$

□

La condition suivante est inspirée par le [BeKu01a, Lemme 3.1] que l'on cite en partie ci-dessous avec les notations précédant l'énoncé de (6.1) dans l'introduction du chapitre.

Lemme 6.4.3. *Si la fonction*

$$\lambda \mapsto \gamma'(\lambda)\sqrt{(\lambda-a)(b-\lambda)} \text{ est strictement croissante sur } [a,b],$$

alors $S(t)$ est un intervalle contenant b pour tout $t \in (0,T)$.

Si la fonction

$$\lambda \mapsto \gamma'(\lambda)\sqrt{(\lambda-a)(b-\lambda)} \text{ est strictement décroissante sur } [a,b],$$

alors $S(t)$ est un intervalle contenant a pour tout $t \in (0,T)$.

Proposition 6.4.4. *Soit une contrainte symétrique σ de support* $\operatorname{supp}(\sigma_1) = [A,B] \subset [0,+\infty)$ *et de densité σ_1'.*

Si la fonction

$$x \mapsto \sqrt{(x^2-A^2)(B^2-x^2)}\sigma_1'(x)$$

est croissante sur $[A,B]$, alors le cas de deux intervalles réels symétriques a lieu avec $B = b_\tau$ pour $0 < \tau < t$.

De même, si la fonction

$$x \mapsto \sqrt{(x^2-A^2)(B^2-x^2)}\sigma_1'(x)$$

est décroissante sur $[A,B]$, alors le cas de deux intervalles réels symétriques a lieu avec $A = a_\tau$ pour $0 < \tau < t$.

PREUVE : On prouve seulement la première affirmation, la preuve de la deuxième est similaire.

Définissons la mesure γ supportée par $[A^2,B^2]$ avec la densité $\gamma'(x) := \sigma_1'(\sqrt{x})$ et la masse

$$T := \int \gamma'(x)dx = \int_{A^2}^{B^2} \sigma_1'\left(\sqrt{x}\right)dx = \int_A^B 2x\sigma_1'(x)\,\mathrm{d}x.$$

Alors, $x \mapsto \sqrt{(x-A^2)(B^2-x)}\gamma'(x)$ est croissante sur $[A^2,B^2]$ par hypothèse.

En considérant le problème de minimisation d'énergie sous la contrainte γ, on sait d'après le lemme 6.4.3 et la formule de Buyarov-Rakhmanov modifiée (6.1) qu'il existe une fonction croissante $t \mapsto \alpha(t) \in [A^2,B^2]$ telle que

$$\gamma'(x) = \int_0^T \frac{1}{\pi}\frac{\mathbb{1}_{[\alpha(t),B^2]}(x)}{\sqrt{(B^2-x)(x-\alpha(t))}}\,\mathrm{d}t,$$

et par conséquent,

$$\begin{aligned}\sigma_1'(x) &= \gamma'(x^2)\\ &= \int_0^T \frac{1}{\pi}\frac{\mathbb{1}_{[\alpha(t),B^2]}(x^2)}{\sqrt{(B^2-x^2)(x^2-\alpha(t))}}\,\mathrm{d}t\\ &= \int_0^T \frac{1}{\pi}\frac{\mathbb{1}_{[\sqrt{\alpha(t)},B]}(x)}{\sqrt{(B^2-x^2)(x^2-\alpha(t))}}\,\mathrm{d}t.\end{aligned}$$

On note $\omega_{a,b}$ la mesure d'équilibre signée du condensateur symétrique à plateau positif $[a,b]$ et plateau négatif $[-b,-a]$. On connaît la formule explicite suivante pour la densité d'après [SaTo97, p 413] :

$$\omega_{a,b}'(x) = \frac{\mathbb{1}_{[a,b]}(x)}{\sqrt{(x^2-a^2)(b^2-x^2)}}\frac{b}{K'(a/b)}. \tag{6.9}$$

Les arguments de la preuve de cette formule sont repris dans la preuve de la proposition 7.4.1.

La fonction

$$\phi(u) = \frac{1}{\pi B}\int_0^u K'\left(\frac{\sqrt{\alpha(\tau)}}{B}\right)\,\mathrm{d}\tau$$

est continue et strictement croissante, et on montre par changement de variable que σ_1' se réécrit

$$\sigma_1'(x) = \sigma_2'(-x) = \int_0^{\sigma_1(\mathbb{C})} \omega_{a_t,B}'(x)\,\mathrm{d}t \text{ avec } a_t = \alpha(\phi^{-1}(t)),$$

et, comme dans la remarque 6.3, on en conclut que $supp\,(\sigma_1 - \mu_1^t) = [a_t, B]$, pour tous les réel t en lesquels la fonction $t \mapsto a_t$ est continue.

Il suffit de considérer les limites à droites en t de cette fonction en ses points de discontinuité pour conclure. □

Toujours en supposant les hypothèses 5.1, 5.2, 5.3, 5.4, et 5.5 vérifiées, on obtient le corollaire suivant qui quantifie de façon plus explicite l'asymptotique faible du problème de Zolotarev dès lors qu'on se trouve dans le cas du condensateur symétrique défini en 6.4.1 pour lequel on a donné une condition suffisante en 6.4.4 ci-dessus. Le calcul de la capacité d'un condensateur est connu, voir par exemple [LeLu01, Section "An example"], et sera par ailleurs redémontré à la proposition 7.4.1.

Corollaire 6.4.5. *On suppose en plus que l'on se trouve dans le cas de deux intervalles symétriques défini en 6.4.1 pour des ensembles $E_N = -F_N \subset (0,+\infty)$. On a alors la formule suivante :*

$$\lim_{n,N\to+\infty,n/N\to t} \frac{1}{N}\log Z_n(E_N,-E_N) = -2\pi\int_0^t \frac{K(a_\tau/b_\tau)}{K'(a_\tau/b_\tau)}d\tau,$$

où $\tau \mapsto a_\tau$ est croissante et $\tau \mapsto b_\tau$ décroissante sur $(0,T)$.

PREUVE : Il suffit d'appliquer les théorèmes 5.2.1 et 6.3.4 pour en déduire le corollaire, avec $supp\ (\sigma_1 - \mu_1^\tau) = [a_\tau, b_\tau]$, d'où la monotonie des fonctions mises en jeu. □

Proposition 6.4.6. *On a au voisinage de zéro l'inégalité*

$$4\log\left(\frac{(1+k')^2}{k}\right) < 2\pi\frac{K'(k)}{K(k)} < 4\log\left(\frac{4}{k}\right) \tag{6.10}$$

ce qui donne l'équivalent au voisinage de zéro

$$\frac{\pi}{2}\frac{K'(k)}{K(k)} \sim_{k\to 0} 4\log\left(\frac{4}{k}\right). \tag{6.11}$$

On en déduit ainsi l'équivalent suivant pour la constante extrémale dans le cas d'une contrainte à support contenu dans $(0,+\infty)$.

Remarque. Dans le cas $a \neq 0$, la proposition 6.4.6 donne l'équivalent

$$\begin{aligned} F^t &= -2\pi\int_0^t \frac{K(a_\tau/b)}{K'(a_\tau/b)}d\tau \\ &\sim_{t\to 0} -8\pi t\log\left(\frac{4b}{a}\right), \end{aligned}$$

valable pour dans le cas symétrique pour $S_1^\tau = [a_\tau, b]$ pour τ proche de zéro.

Les résultats de ce chapitre nous permettront d'effectuer des estimations numériques de l'asymptotique de la quantité de Zolotarev et de les confronter aux figures 3.1, 3.2 et 3.3 dans le chapitre 8.

Chapitre 7

Cas du condensateur réel : équations intégrales

7.1 Introduction

La $\mathfrak{F}_t$-fonctionnelle a été définie au chapitre précédent, et la proposition 6.2.9 nous incite à rechercher ses minima dans le but d'obtenir des informations concernant la partie libre de la contrainte pour des mesures de différentes masses donnée par les ensembles $\left(supp\left(\sigma_j - \mu_j^\tau\right)\right)_{0\leqslant\tau\leqslant t}$, $j = 1, 2$. La détermination de ces ensembles est nécessaire pour appliquer la formulation explicite du corollaire 6.4.5.

Les informations à propos de la localisation de la famille $\left(supp\left(\sigma_j - \mu_j^\tau\right)\right)_{0\leqslant\tau\leqslant t}$, $j = 1, 2$ nous seront également utiles par la suite pour l'exploitation numérique de nos résultats dans le chapitre 8.

Ce chapitre est destiné à caractériser les éventuels points critiques de la $\mathfrak{F}_t$-fonctionelle par un système d'équations intégrales dans le cas du condensateur réel. Pour caractériser ce cas, on a donné une condition suffisante à la proposition 6.4.4.

Même si les conditions suffisantes énoncées en 6.4.4 ne sont valables que pour le cas d'un condensateur réel symétrique, les calculs de cette partie seront faits dans le cadre du condensateur réel général.

Notons maintenant

$$K := [a, b] \cup [c, d]$$

où $a < b < c < d$,

$$\mathcal{T}_1 := (b, c),\ \mathcal{T}_2 := (-\infty, a) \cup (d, +\infty),$$

ainsi que

$$E := \mathcal{T}_1 \cup \mathcal{T}_2.$$

On emploiera l'abus de notation qui consiste à noter

$$\mathfrak{F}_t(a, b, c, d)$$

en lieu et place de

$$\mathfrak{F}_t\left([a, b], [c, d]\right).$$

On suppose dans toute cette section que le support de la contrainte σ est contenu dans l'axe réel, et plus précisément que si l'on écrit comme précédemment la décomposition de Jordan de la contrainte sous la forme $\sigma = \sigma_1 - \sigma_2$, on a

$$supp\,(\sigma_1) \subset (0, +\infty) \text{ et } supp\,(\sigma_2) \subset (-\infty, 0).$$

Dans tout ce chapitre, on suppose la mesure signée σ suffisamment régulière pour que les potentiels U^{σ_j}, $j = 1, 2$ soient continus, cette hypothèse de régularité a été discutée après son énoncé à l'hypothèse 5.3.

On fait enfin la convention de notation suivante dans tout le chapitre : pour $u, v \in \mathbb{R}^2$, on note

$$\sigma(u, v) := \sigma\left((u, v)\right).$$

On remarque que d'après l'hypothèse de continuité des potentiels logarithmiques U^{σ_j}, $j = 1, 2$, σ n'a pas d'atomes et par conséquent, pour $u, v \in \mathbb{R}^2$,

$$\sigma(u, v) = \sigma\left([u, v]\right).$$

Le principe des calculs qui vont être menés maintenant est le suivant : l'expression de la $\mathfrak{F}_t$-fonctionnelle dépend de quatre paramètres a, b, c et d, mais on peut exprimer celle-ci en fonction du birapport des réels $[a, b, c, d]$ afin de faciliter le calcul du gradient de cette fonctionnelle en dérivant par rapport à ce paramètre.

Pour faciliter les calculs, une transformation homographique judicieuse permet en effet de se ramener au cas du condensateur $[-k^{-1}, -1] \cup [1, k^{-1}]$, cadre plus simple dans lequel tous les calculs de dérivation de la $\mathfrak{F}_t$-fonctionnelle seront faits.

La résolution de

$$\nabla \mathfrak{F}_t(a, b, c, d) = 0$$

donne lieu à un système de 4 équations intégrales, dont on cite ici les trois plus simples.

Théorème 7.1.1. *Pour $a < b < c < d$, notons $R(z) := (z - a)(z - b)(z - c)(z - d)$.*

Si (a, b, c, d) est un point critique de la $\mathfrak{F}_t$-fonctionnelle, on a

$$\int_{\mathcal{T}_1} \frac{\mathrm{d}\sigma(z)}{\sqrt{R(z)}} = \int_{\mathcal{T}_2} \frac{\mathrm{d}\sigma(z)}{\sqrt{R(z)}},$$

$$\int_{\mathcal{T}_1} \frac{z\,\mathrm{d}\sigma(z)}{\sqrt{R(z)}} = \int_{\mathcal{T}_2} \frac{z\,\mathrm{d}\sigma(z)}{\sqrt{R(z)}},$$

et

$$\int_{\mathcal{T}_1} \frac{z^2\,\mathrm{d}\sigma(z)}{\sqrt{R(z)}} = \int_{\mathcal{T}_2} \frac{z^2\,\mathrm{d}\sigma(z)}{\sqrt{R(z)}}.$$

On note que parmi les trois équations énoncées ci-dessus, deux sont triviales dès que la mesure contrainte σ est de densité impaire, cas important dans les applications du chapitre 8.

Les trois équations intégrales du résultat précédent sont obtenues par combinaisons linéaires en éliminant le terme obtenu par dérivation de $\mathfrak{F}_t$ par rapport au birapport k d'écriture plus complexe. On obtiendra enfin au prix de quelques calculs une quatrième équation plus complexe faisant intervenir ce terme qui est citée dans le théorème récapitulatif en fin de chapitre.

La quatrième équation intégrale s'exprime en terme de fonctions spéciales, les fonctions elliptiques de Legendre, on donne ici quelques rappels à propos de ces fonctions.

7.2 Fonctions elliptiques de Legendre

L'utilisation des fonctions elliptiques de Legendre est un outil indispensable dans ce paragraphe, nous rappelons ici le minimum nécessaire à l'élaboration des calculs qui suivent.

On rappelle ici la définition des fonctions elliptiques de Legendre de première, deuxième et troisième espèce, les définitions sont tirées de [PrBrMa88, p 771-774], on pourra également consulter à ce sujet [AbSt64, Section 17] et [GrRy00, Section 8.11].

Définition 7.2.1. *On définit pour* $k \in (0,1)$ *et* $z \in [-1,1]$ *la* fonction elliptique de Legendre de première espèce incomplète F *ainsi que la* fonction elliptique de Legendre de première espèce complète K *de paramètre* k

$$F(z,k) := \int_0^z \frac{\mathrm{d}t}{\sqrt{(1-t^2)(1-k^2t^2)}} \text{ et } K(k) := F(1,k).$$

On définit également $k' := \sqrt{1-k^2}$ *et*

$$K'(k) := K(k') = \int_0^1 \frac{\mathrm{d}t}{\sqrt{(1-t^2)(1-(1-k^2)t^2)}}.$$

On définit enfin la fonction elliptique de Legendre de troisième espèce incomplète Π *pour* $(z,\nu,k) \in [0,1]\times(0,1)\times(0,1)$,

$$\Pi(z,\nu,k) := \int_0^z \frac{\mathrm{d}t}{(1-\nu t^2)\sqrt{(1-t^2)(1-k^2t^2)}}.$$

ainsi que la fonction auxillaire f *qui sera utilisée lors de la réécriture de la* $\mathfrak{F}$*-fonctionnelle proposition 7.6.2 :*

$$f(u,k) := \frac{F(1,k)-F(u,k)}{K'(k)} \text{ pour } (u,k) \in [0,1]\times(0,1).$$

On note enfin

$$G(u,k) := F\left(\frac{1}{ku},k\right).$$

Proposition 7.2.2. *On a*

$$\frac{\partial F}{\partial u}(u,k) = \frac{1}{\sqrt{(1-u^2)(1-k^2u^2)}}, \text{ pour } (u,k) \in [0,1]\times(0,1) \tag{7.1}$$

et par théorème de dérivation de fonctions composées,

$$\frac{\partial G}{\partial u}(u,k) = -\frac{1}{\sqrt{(1-u^2)(1-k^2u^2)}}, \text{ pour } (u,k) \in (0,1]\times(0,1). \tag{7.2}$$

On a de plus

$$\frac{d}{dk}\frac{K}{K'}(k) = \frac{\pi}{2k(1-k^2)\left(K'(k)\right)^2} \text{ pour } k \in (0,1). \tag{7.3}$$

On a également pour $u \in [0,1)$ *et* $0<k<1$

$$\frac{\partial f}{\partial k}(u,k) = \Pi(1,1-k^2u^2,k')\frac{k^2u\sqrt{1-u^2}}{\sqrt{1-k^2u^2}}\frac{1}{k(1-k^2)K'(k)}. \tag{7.4}$$

PREUVE : L'équation (7.1) résulte directement de la définition de F, (7.2) résulte alors d'un simple calcul, on connait de plus l'expression suivante d'après [PrBrMa88, p 771-774] :

$$\frac{d}{dk}\frac{K}{K'}(k) = \frac{\pi}{2k(1-k^2)\left(K'(k)\right)^2},$$

ainsi que pour $(u,k) \in (0,1] \times (0,1)$,

$$\begin{aligned}\frac{\partial}{\partial k}\left(\frac{F(u,k)}{K'(k)}\right) &= \frac{\pi}{2k(1-k^2)\left(K'(k)\right)^2},\\ &\quad - \frac{1}{k(1-k^2)\left(K'(k)\right)^2}\Pi\left(1, 1-k^2u^2, k'\right)k^2u\sqrt{\frac{1-u^2}{1-k^2u^2}},\end{aligned}$$

ce qui permet d'en déduire le résultat concernant f. □

7.3 Homographies

Les homographies nous ont déjà été utiles pour étudier les cas de dégénérescence de la solution au problème de Zolotarev.

On utilise maintenant celles-ci de façon plus explicite, en faisant intervenir dans les calculs un paramètre construit à partir du birapport $[a,b,c,d]$ des bornes des plateaux du condensateur réel que l'on considère dans les calculs. Le birapport des quatre complexes disctincts a été défini en 3.2.7, on remarque que dans notre situation de quatre réels distincts ordonnés, le birapport est un réel positif.

Les résultats présentés ici sont bien connus, on s'est inspiré de [Han04, p. 188] pour ce paragraphe.

On note dans tout ce qui suit $k = k(a,b,c,d)$ l'unique élément k de $(0,1)$ vérifiant

$$[a,b,c,d] = [-1,-k,k,1] = [-k^{-1},-1,1,k^{-1}]$$

d'après le lemme 3.2.8.

Définition 7.3.1. *Notons pour un quadruplet (a,b,c,d) de réels distincts tels que $[a,b,c,d] > 0$*

$$\mu := (a,b,c,d) \mapsto \frac{\sqrt{\frac{c-a}{b-a}\frac{d-c}{d-b}} - 1}{\sqrt{\frac{c-a}{b-a}\frac{d-c}{d-b}} + 1}.$$

Voici maintenant les homographies pertinentes pour notre problème puisqu'elles permettront de limiter nos calcul au cas du condensateur de plateau positif $[1,k^{-1}]$ et négatif $[-k^{-1},-1]$.

Proposition 7.3.2. *On note k l'unique élément de $(0,1)$ tel que*

$$[a,b,c,d] = [-k^{-1},-1,1,k^{-1}].$$

Il existe une unique transformation homographique T vérifiant

$$\begin{aligned} T : -\frac{1}{k} &\mapsto a, \\ -1 &\mapsto b, \\ 1 &\mapsto c, \\ \frac{1}{k} &\mapsto d, \end{aligned}$$

et on a

$$T(u) := \frac{b+c}{2} + \frac{c-b}{2}\frac{u-\mu(a,b,c,d)}{1-u\mu(a,b,c,d)} \quad \forall\, u \in \mathbb{C} \setminus \left\{ \frac{1}{\mu(a,b,c,d)} \right\}.$$

De plus, il existe une unique transformation homographique Q vérifiant

$$\begin{aligned} Q : -\frac{1}{k} &\mapsto b, \\ -1 &\mapsto a, \\ 1 &\mapsto d, \\ \frac{1}{k} &\mapsto c, \end{aligned}$$

et on a

$$Q(u) := \frac{a+d}{2} + \frac{d-a}{2}\frac{1-u\mu(b,a,d,c)}{u-\mu(b,a,d,c)} \quad \forall\, u \in \mathbb{C} \setminus \{\mu(b,a,d,c)\},$$

et

$$Q(u) = T\left(\frac{1}{ku}\right).$$

Notons maintenant $h := T^{-1}$ et $l := Q^{-1}$.

On a alors les égalités suivantes valables sur les ensembles de définition respectifs des homographies considérées

$$h(z) = \frac{2z-(c+b)+\mu(a,b,c,d)(c-b)}{2z\mu(a,b,c,d)+(c-b)-\mu(a,b,c,d)(c+b)}$$

et

$$l(z) = \frac{2z\mu(b,a,d,c)+(d-a)-\mu(b,a,d,c)(a+d)}{2z-(a+d)+\mu(b,a,d,c)(d-a)}.$$

PREUVE : L'existence et l'unicité de la transformation homographique vérifiant les égalités demandées est une propriété classique de géométrie projective, voir par exemple [Si67, Chapitre 5 Théorème 5.5].

L'expression de la transformation homographique inverse de T est donnée dans [Han04, p. 188].

Il suffit pour l'homographie Q d'observer que l'application $u \longmapsto T(\frac{1}{ku})$ vérifie les égalités énoncées pour Q et ensuite de calculer l'expression de $T\left(\frac{1}{ku}\right)$ qui donne bien le résultat souhaité.

Enfin, l'expression de h est donnée dans [Han04, p. 188] et l'expression de l s'obtient par calcul direct à partir de la relation de la proposition 7.3.2 $Q(u) = T\left(\frac{1}{ku}\right)$. □

On remarque que dans le cas de points symétriques par rapport à l'origine $d = -a = \beta$ et $c = -b = \alpha$ où $0 < \alpha < \beta$, on obtient

$$k = \frac{\alpha}{\beta},\ \mu(-\beta, -\alpha, \alpha, \beta) = 0$$

et dans ce cas T et Q sont des applications linéaires données par $T(u) = \alpha u$ et $Q(u) = \frac{\beta}{u}$, ce qui simplifie les expressions données ci-dessus.

7.4 Potentiel logarithmique d'un condensateur réel

Les calculs de la section précédente ont pour but de simplifier le problème pour se ramener au condensateur $[-k^{-1}, -1] \cup [1, k^{-1}]$ où k est un paramètre adapté à notre problème construit à partir du birapport $[a, b, c, d]$. Il nous faut alors connaître le potentiel logarithmique de ce condensateur pour continuer qui est déterminé en suivant la démarche de [SaTo97, Exemple II.5.14] pour le cas symétrique et en utilisant une homographie adéquate pour le calcul général. L'exemple [SaTo97, Exemple II.5.14] est rédigé dans un vocablaire de potentiel de Green, ce qui revient à considérer des mesures signées à densité impaire, ce qui constitue bien le cadre de notre calcul dans le cas symétrique.

Proposition 7.4.1. *Le potentiel logarithmique de la mesure d'équilibre notée ω_k du condensateur de plateau négatif $[-k^{-1}, -1]$ et positif $[1, k^{-1}]$ pour un paramètre $k \in (0, 1)$ est la fonction impaire vérifiant*

$$U^{\omega_k}(z) = \pi \frac{F(z, k)}{K'(k)} \ si\ 0 \leqslant z < 1,$$
$$U^{\omega_k}(z) = \pi \frac{K(k)}{K'(k)} \ si\ 1 \leqslant z \leqslant k^{-1},$$
$$U^{\omega_k}(z) = \pi \frac{F\left(\frac{1}{kz}, k\right)}{K'(k)} \ si\ k^{-1} < z.$$

La capacité de ce condensateur vaut alors

$$\mathrm{cap}\left([-k^{-1}, -1], [1, k^{-1}]\right) = \frac{K'(k)}{2\pi K(k)}.$$

Dans le cas d'un condensateur non symétrique de plateau négatif $[a, b]$ et positif $[c, d]$ où $a < b < c < d$, on a en notant $k = k(a, b, c, d)$ et avec les notations précédentes

$$U^{\omega_{[a,b],[c,d]}}(z) = \pi \frac{F\left(\frac{1}{kh(z)}, k\right)}{K'(k)} \ si\ z < a\ ou\ z > d,$$
$$U^{\omega_{[a,b],[c,d]}}(z) = -\pi \frac{K(k)}{K'(k)} \ si\ a \leqslant z \leqslant b,$$
$$U^{\omega_{[a,b],[c,d]}}(z) = \pi \frac{F\left(h(z), k\right)}{K'(k)} \ si\ b < z < c,$$
$$U^{\omega_{[a,b],[c,d]}}(z) = \pi \frac{K(k)}{K'(k)} \ si\ c \leqslant z \leqslant d$$

et

$$\text{cap}\,([a,b],[c,d]) = \frac{K'(k)}{2\pi K(k)}. \tag{7.5}$$

PREUVE : Par symétrie de l'ensemble $[-k^{-1},-1]\cup[1,k^{-1}]$ par rapport à l'origine et par unicité la mesure d'équilibre ω_k, celle-ci admet une densité impaire, ce qui prouve l'imparité de la fonction U^{ω_k}.

Maintenant, on sait d'après [SaTo97, p. 413] que la densité $\mathrm{d}\omega_k$ vérifie

$$\frac{\mathrm{d}\omega_k}{\mathrm{d}z} = \frac{C}{-i\pi\sqrt{(1-z^2)(k^{-2}-z^2)}},$$

pour une certaine constante C, la condition $\omega_{[-k^{-1},-1],[1,k^{-1}]}([1,k^{-1}]) = 1$ donne alors par détermination de la racine choisie dans [SaTo97, p. 413]

$$C = \frac{-i\pi}{\int_1^{k^{-1}} \frac{\mathrm{d}z}{\sqrt{(1-z^2)(k^{-2}-z^2)}}}.$$

On a enfin

$$\int_1^{k^{-1}} \frac{\mathrm{d}z}{\sqrt{(z^2-1)(k^{-2}-z^2)}} = k\int_0^1 \frac{\mathrm{d}s}{\sqrt{(1-s^2)\,(1-(1-k^2)s^2)}} = kK'(k)$$

où l'on a fait le changement de variable

$$s^2 = \frac{1-k^2z^2}{1-k^2},$$

d'où le résultat pour le potentiel U^{ω_k} par définition de la fonction elliptique F.

La valeur de la capacité s'ensuit alors par définition de celle-ci (voir 6.2.2), et on en déduit également (7.5) car la capacité est invariante par transformation homographique d'après [Ra95, Théorème 5.2.3].

Connaissant le potentiel U^{ω_k}, on sait donc qu'il vérifie les conditions d'équilibre (6.3), et il est alors aisé de vérifier par définition de h que la fonction $z \mapsto U^{\omega_k\circ h}(z)$ vérifie les conditions d'équilibre requises pour le potentiel $U^{\omega_{[a,b],[c,d]}}$ données en (6.3), ce qui termine la preuve. □

7.5 Calcul de dérivées partielles

Dans l'optique d'obtenir des équations liées aux points critiques de la $\mathfrak{F}$-fonctionnelle, on s'intéresse maintenant aux dérivées partielles des applications T et Q par rapport aux variables a, b, c, d. Cette partie est d'intérêt purement calculatoire et donne les combinaisons linéaires qui permettent d'obtenir les trois premières équations intégrales énoncées dans l'introduction.

Proposition 7.5.1. *On a en tout point où ces expressions ont un sens*

$$\frac{\partial T}{\partial a}(z) = \frac{z^2-1}{4}\frac{c-b}{(b-a)(c-a)}T'(z),$$
$$\frac{\partial T}{\partial b}(z) = \left[-\frac{z^2-1}{4}\frac{d+a-2b}{(d-b)(b-a)} - \frac{(z-1)(1-z\mu(a,b,c,d))}{(c-b)(1-\mu(a,b,c,d))}\right]T'(z),$$
$$\frac{\partial T}{\partial c}(z) = \left[\frac{z^2-1}{4}\frac{d+a-2c}{(d-c)(c-a)} + \frac{(z+1)(1-z\mu(a,b,c,d))}{(c-b)(1+\mu(a,b,c,d))}\right]T'(z),$$
$$\frac{\partial T}{\partial d}(z) = \frac{z^2-1}{4}\frac{c-b}{(d-b)(d-c)}T'(z)$$

et

$$\frac{\partial Q}{\partial a}(z) = \left[\frac{c+b-2a}{(c-a)(b-a)} - \frac{(z-1)(z-\mu(b,a,d,c)}{(d-a)(1-\mu(b,a,d,c))}\right]Q'(z),$$
$$\frac{\partial Q}{\partial b}(z) = -\frac{z^2-1}{4}\frac{d-a}{(b-a)(d-b)}Q'(z),$$
$$\frac{\partial Q}{\partial c}(z) = -\frac{z^2-1}{4}\frac{d-a}{(c-a)(d-c)}Q'(z),$$
$$\frac{\partial Q}{\partial d}(z) = \left[-\frac{z^2-1}{4}\frac{c+b-2d}{(d-c)(d-b)} - \frac{(z+1)(z-\mu(b,a,d,c))}{(d-a)(1+\mu(b,a,d,c))}\right]Q'(z).$$

On donne maintenant des relations utilisant les dérivées partielles du paramètre $k(a,b,c,d)$ afin d'obtenir trois équations intégrales simples par combinaison linéaire comme expliqué précédemment.

Proposition 7.5.2. *Avec les notations du lemme 3.2.8, on a pour $a<b<c<d$ les équations suivantes :*

$$\frac{\partial k}{\partial a} = \frac{(d-c)(c-b)}{(1+[a,d,b,c])^2\,[a,d,b,c](a-c)^2(d-b)}, \tag{7.6}$$

$$(c-a)(d-a)\frac{\partial k}{\partial a} + (c-b)(d-b)\frac{\partial k}{\partial b} = 0, \tag{7.7}$$

$$(b-a)(d-a)\frac{\partial k}{\partial a} - (c-b)(d-c)\frac{\partial k}{\partial c} = 0 \tag{7.8}$$

et

$$(c-a)(b-a)\frac{\partial k}{\partial a} + (d-c)(d-b)\frac{\partial k}{\partial d} = 0. \tag{7.9}$$

Pour effectuer ces combinaisons linéaires, il nous faut maintenant reformuler l'écriture de la $\mathfrak{F}$-fonctionnelle en fonction du paramètre k, ce qui fait intervenir des fonctions elliptiques de Legendre.

7.6 Réécriture de la fonctionnelle

Nous disposons maintenant des outils pour réécrire la $\mathfrak{F}$-fonctionnelle en fonction du paramètre k de façon à faciliter les calculs menant au système d'équations intégrales déterminant les points critiques de celle-ci.

Dans la définition suivante, on relâche la dépendance en (a,b,c,d) du paramètre k pour le considérer comme une variable indépendante.

Définition 7.6.1. *On définit avec les notations de la définition 7.2.1*

$$I(a,b,c,d,k) = \int_0^1 f(u,k)\,\mathrm{d}\sigma\circ T(u) - \int_{-1}^0 f(-u,k)\,\mathrm{d}\sigma\circ T(u)$$

et

$$J(a,b,c,d,k) = \int_{-1}^0 f(-u,k)\,\mathrm{d}\sigma\circ Q(u) - \int_0^1 f(u,k)\,\mathrm{d}\sigma\circ Q(u),$$

où l'on considère T et Q comme fonctions des paramètres (a,b,c,d).

On note dorénavant

$$\frac{1}{\pi}\mathfrak{G}_t(a,b,c,d,k) := -2t\frac{K(k)}{K'(k)} + I(a,b,c,d,k) + J(a,b,c,d,k).$$

Proposition 7.6.2. *Soient $a<b<0<c<d$. On a*

$$\mathfrak{F}_t(a,b,c,d) = \mathfrak{G}_t(a,b,c,d,k)_{|k=k(a,b,c,d)}.$$

PREUVE : On commence par réaménager les termes en utilisant l'expression du potentiel logarithmique du condensateur à plateau positif $[c,d]$ et négatif $[a,b]$ donné à la proposition 7.4.1 :

$$\begin{aligned}
\frac{1}{\pi}\mathfrak{F}_t(a,b,c,d) &= 2\frac{K(k)}{K'(k)}\left(\sigma\left((0,+\infty)\right) - t\right) - \int U^{\omega_{[a,b],[c,d]}}\,\mathrm{d}\sigma(z),\\
&= 2\frac{K(k)}{K'(k)}\left(\sigma\left((0,+\infty)\right) - t\right) - \int_{\mathcal{T}_2}\frac{F\left(\frac{1}{kh(z)},k\right)}{K'(k)}\,\mathrm{d}\sigma(z)\\
&- \int_{\mathcal{T}_1}\frac{F\left(h(z),k\right)}{K'(k)}\,\mathrm{d}\sigma(z) - \frac{K(k)}{K'(k)}\left(\sigma(c,d)-\sigma(a,b)\right),\\
&= \frac{K(k)}{K'(k)}\left(|\sigma|(E) - 2t\right)\\
&- \int_{\mathcal{T}_1}\frac{F\left(h(z),k\right)}{K'(k)}\,\mathrm{d}\sigma(z) - \int_{\mathcal{T}_2}\frac{F\left(\frac{1}{kh(z)},k\right)}{K'(k)}\,\mathrm{d}\sigma(z).
\end{aligned}$$

On effectue le changement de variable $u=h(z)$, d'où

$$\int_{\mathcal{T}_1} F\left(h(z),k\right)\,\mathrm{d}\sigma(z) = \int_{-1}^1 F(u,k)\,\mathrm{d}\sigma\left(T(u)\right)$$

et de même,

$$\int_{\mathcal{T}_2} F\Big(\frac{1}{kh(z)},k\Big)\,\mathrm{d}\sigma(z) = \int_{(-\infty,-k^{-1})\cup(k^{-1},+\infty)} F\Big(\frac{1}{ku},k\Big)\,\mathrm{d}\sigma\left(Q(u)\right).$$

On obtient alors,

$$\begin{aligned}
&\frac{K(k)}{K'(k)}|\sigma|(b,c) - \int_{\mathcal{T}_1} \frac{F\left(h(z),k\right)}{K'(k)}\,\mathrm{d}\sigma(z)\\
&= \int_0^c \frac{F(1,k) - F\left(h(z),k\right)}{K'(k)}\,\mathrm{d}\sigma(z) + \int_b^0 \frac{-F(1,k) - F\left(h(z),k\right)}{K'(k)}\,\mathrm{d}\sigma(z)\\
&= \int_0^1 \frac{F(1,k) - F(u,k)}{K'(k)}\,\mathrm{d}\sigma\circ T(u) + \int_{-1}^0 \frac{-F(1,k) - F(u,k)}{K'(k)}\,\mathrm{d}\sigma\circ T(u)\\
&= \int_0^1 f(u,k)\,\mathrm{d}\sigma\circ T(u) - \int_{-1}^0 f(-u,k)\,\mathrm{d}\sigma\circ T(u)
\end{aligned}$$

et on procède de même pour J, ce qui donne le résultat. □

La réécriture de la $\mathfrak{F}_t$-fonctionnelle a produit des termes donnés par les intégrales I et J que nous allons nous maintenant exprimer par rappport aux variables a, b, c, d. La définition de ces intégrales donne une forme aisée à dériver par rapport à k, mais pas par rapport aux autres paramètres, la proposition suivante donne une écriture plus maniable de I et J pour la dérivation par rapport à (a, b, c, d).

Proposition 7.6.3. *On a*

$$I(a,b,c,d,k) = \frac{1}{K'(k)}\int_{-1}^1 \frac{|\sigma\left(0,T(u)\right)|}{\sqrt{(1-u^2)(1-k^2u^2)}}\,\mathrm{d}u$$

et

$$J(a,b,c,d,k) = \frac{1}{K'(k)}\int_{-1}^1 \frac{\mathbb{1}_{\{Q(u)<0\}}\,|\sigma\left(-\infty,Q(u)\right)| + \mathbb{1}_{\{Q(u)>0\}}\,|\sigma\left(Q(u),+\infty\right)|}{\sqrt{(1-u^2)(1-k^2u^2)}}\,\mathrm{d}u.$$

PREUVE : Par intégration par parties, on obtient

$$\begin{aligned}
\int_{-1}^{h(0)} F(u,k)\,\mathrm{d}\sigma\left(T(u)\right) &= \Big[-F(u,k)\sigma\circ T\left(u,h(0)\right)\Big]_{-1}^{h(0)}\\
&\quad + \int_{-1}^{h(0)} \frac{\partial F}{\partial u}\sigma\circ T\left(u,h(0)\right)\,\mathrm{d}u
\end{aligned}$$

et

$$\begin{aligned}
\int_{h(0)}^{1} F(u,k)\,\mathrm{d}\sigma\left(T(u)\right) &= \Big[F(u,k)\sigma\circ T\left(u,h(0)\right)\Big]_{h(0)}^{1}\\
&\quad - \int_{h(0)}^{1} \frac{\partial F}{\partial u}\sigma\circ T\left(h(0),u\right)\,\mathrm{d}u,
\end{aligned}$$

d'où d'après (7.1)

$$\begin{aligned}
\int_{-1}^1 F(u,k)\,\mathrm{d}\sigma\circ T(u) &= \sigma(0,c)F(1,k) - \sigma(b,0)F(1,k)\\
&\quad - \int_{-1}^1 |\sigma\left(0,T(u)\right)|\,\frac{\mathrm{d}u}{\sqrt{(1-u^2)(1-k^2u^2)}}.
\end{aligned}$$

On effectue le même processus de calcul sur $\mathcal{T}_2$ ce qui donne d'après (7.2)

$$\begin{aligned} I_1 &:= \int_{k^{-1}}^{h(\infty)} F\Big(\frac{1}{ku}, k\Big)\, \mathrm{d}\sigma \circ T(u), \\ &= \Big[-\sigma \circ T\,(u, h(\infty))\, F\Big(\frac{1}{ku}, k\Big)\Big]_{\frac{1}{k}}^{h(\infty)} \\ &+ \int_{k^{-1}}^{h(\infty)} \frac{\partial F}{\partial u}\Big(\frac{1}{ku}, k\Big)\sigma \circ T\,(u, h(\infty))\ \mathrm{d}u, \\ &= \sigma(d, +\infty)F(1,k) - \int_{k^{-1}}^{h(\infty)} \frac{\sigma \circ T\,(u, h(\infty))}{\sqrt{(1-u^2)(1-k^2u^2)}}\,\mathrm{d}u. \end{aligned}$$

De plus,

$$\begin{aligned} I_2 &:= \int_{(-\infty,-k^{-1})\cup(h(\infty),+\infty)} F\Big(\frac{1}{ku}, k\Big)\, \mathrm{d}\sigma \circ T(u) \\ &= \Big[\sigma \circ T\,(h(\infty), u)\, F\Big(\frac{1}{ku}, k\Big)\Big]_{h(\infty)}^{-k^{-1}}, \\ &- \int_{(-\infty,-k^{-1})\cup(h(\infty),+\infty)} \frac{\partial F}{\partial u}\Big(\frac{1}{ku}, k\Big)\sigma \circ T\,(h(\infty), u)\ \mathrm{d}u, \\ &= -\sigma(-\infty, a) - \int_{(-\infty,-k^{-1})\cup(h(\infty),+\infty)} \frac{|\sigma\,(-\infty, T(u))|}{\sqrt{(1-u^2)(1-k^2u^2)}}\,\mathrm{d}u. \end{aligned}$$

Cela donne finalement l'écriture souhaitée grâce au changement de variable $u = \frac{1}{kv}$ dans les intégrales I_1 et I_2. □

Les changements de variables donnés par les homographies h et l nous ont permis de passer de la configuration *non symétrique* -condensateur à plateaux $[a, b]$ et $[c, d]$- en la configuration *symétrique* -plateaux $[-k^{-1}, -1]$ et $[1, k^{-1}]$- où k est donné par la proposition 7.5.2.

On fait ensuite ces changements de variable dans les termes obtenus par dérivation par rapport aux variables a, b, c, d des applications T et Q qui interviennent dans la réécriture de la $\mathfrak{F}_t$-fonctionnelle.

On définit dans ce but les quantités suivantes.

Définition 7.6.4. *On note*

$$M := (a, b, c, d) \mapsto \sqrt{(c-a)(d-b)} + \sqrt{(d-c)(b-a)},$$

$$R(z) := (z-a)(z-b)(z-c)(z-d)$$

et

$$P(z) := \frac{(z-b)(z-c)}{(z-a)(z-d)}.$$

Le lemme suivant prépare les changements de variable évoqués ci-dessus.

Lemme 7.6.5. *On a avec les notations des définitions 7.3.1 et 7.6.4*

$$\begin{aligned}
\frac{(1-h(z)\mu)\,(1-h(z))}{(c-b)(1-\mu)}\frac{1}{\sqrt{(1-h(z)^2)\,(1-k^2h(z)^2)}} &= -\frac{M}{2(c-b)}\frac{z-c}{\sqrt{R(z)}},\\
\frac{(1-h(z)\mu)\,(1+h(z))}{(c-b)(1+\mu)}\frac{1}{\sqrt{(1-h(z)^2)\,(1-k^2h(z)^2)}} &= \frac{M}{2(c-b)}\frac{z-b}{\sqrt{R(z)}},\\
\frac{(l(z)-1)\,(l(z)-\tilde{\mu})}{(d-a)(1-\tilde{\mu})}\frac{1}{\sqrt{(1-l(z)^2)\,(1-k^2l(z)^2)}} &= -\frac{M}{2(d-a)}\frac{z-d}{\sqrt{R(z)}},\\
\frac{(l(z)+1)\,(l(z)-\tilde{\mu})}{(d-a)(1+\tilde{\mu})}\frac{1}{\sqrt{(1-l(z)^2)\,(1-k^2l(z)^2)}} &= \frac{M}{2(d-a)}\frac{z-a}{\sqrt{R(z)}}
\end{aligned}$$

et également

$$\begin{aligned}
\sqrt{\frac{1-h(z)^2}{1-k^2h(z)^2}} &= \frac{M}{c-b}\sqrt{P(z)},\\
\sqrt{\frac{1-l(z)^2}{1-k^2l(z)^2}} &= \frac{M}{d-a}\frac{1}{\sqrt{P(z)}},
\end{aligned}$$

où chaque égalité est vérifiée sur l'ensemble de définition des fonctions considérées.

Utilisons maintenant les résultats des propositions 7.5.1 et 7.6.5 pour évaluer les dérivées partielles des intégrales I et J par rapport aux variables a, b, c et d. On peut dans chaque cas appliquer un théorème de dérivation sous le signe intégrale étant donné que la mesure σ est finie et que les termes mis en jeu sont toujours des fractions rationnelles n'ayant pas de pôle dans les intervalles sur lesquels on travaille. Le résultat s'ensuit alors par changement de variable en utilisant les relations établies précédemment.

Proposition 7.6.6. *Le calcul des dérivées partielles des fonctions I et J par rapport aux paramètres a, b, c, d vérifiant $a < b < c < d$ donne*

$$\begin{aligned}
\frac{\partial I}{\partial a} &= -\frac{1}{K'(k)}\frac{M}{4}\frac{1}{(b-a)(c-a)}\int_{\mathcal{T}_1}\sqrt{P(z)}\,\mathrm{d}\sigma(z),\\
\frac{\partial I}{\partial b} &= \frac{1}{K'(k)}\frac{M}{4}\frac{d+a-2b}{(d-b)(b-a)(c-b)}\int_{\mathcal{T}_1}\sqrt{P(z)}\,\mathrm{d}\sigma(z)\\
&\quad -\frac{1}{K'(k)}\frac{M}{2(c-b)}\int_{\mathcal{T}_1}\frac{z-c}{\sqrt{R(z)}}\,\mathrm{d}\sigma(z),\\
\frac{\partial I}{\partial c} &= -\frac{1}{K'(k)}\frac{M}{4}\frac{d+a-2c}{(d-c)(c-a)(c-b)}\int_{\mathcal{T}_1}\sqrt{P(z)}\,\mathrm{d}\sigma(z)\\
&\quad +\frac{1}{K'(k)}\frac{M}{2(c-b)}\int_{\mathcal{T}_1}\frac{z-c}{\sqrt{R(z)}}\,\mathrm{d}\sigma(z),\\
\frac{\partial I}{\partial d} &= -\frac{1}{K'(k)}\frac{M}{4}\frac{1}{(d-b)(d-c)}\int_{\mathcal{T}_1}\sqrt{P(z)}\,\mathrm{d}\sigma(z),
\end{aligned}$$

et

$$\begin{aligned}
\frac{\partial J}{\partial a} &= -\frac{1}{K'(k)}\frac{M}{4}\frac{c+b-2a}{(b-a)(c-a)(d-a)}\int_{\mathcal{T}_2}\frac{1}{\sqrt{P(z)}}\,\mathrm{d}\sigma(z)\\
&\quad+\frac{1}{K'(k)}\frac{M}{2(d-a)}\int_{\mathcal{T}_2}\frac{z-d}{\sqrt{R(z)}}\,\mathrm{d}\sigma(z),\\
\frac{\partial J}{\partial b} &= \frac{1}{K'(k)}\frac{M}{4}\frac{1}{(b-a)(d-b)}\int_{\mathcal{T}_2}\frac{1}{\sqrt{P(z)}}\,\mathrm{d}\sigma(z),\\
\frac{\partial J}{\partial c} &= \frac{1}{K'(k)}\frac{M}{4}\frac{1}{(c-a)(d-c)}\int_{\mathcal{T}_2}\frac{1}{\sqrt{P(z)}}\,\mathrm{d}\sigma(z),\\
\frac{\partial J}{\partial d} &= \frac{1}{K'(k)}\frac{M}{4}\frac{c+b-2d}{(d-a)(d-b)(d-c)}\int_{\mathcal{T}_2}\frac{1}{\sqrt{P(z)}}\,\mathrm{d}\sigma(z)\\
&\quad-\frac{1}{K'(k)}\frac{M}{2(d-a)}\int_{\mathcal{T}_2}\frac{z-a}{\sqrt{R(z)}}\,\mathrm{d}\sigma(z).
\end{aligned}$$

Le cas symétrique où $d\sigma$ est impaire et $d=-a=B>0$ et fixé et le paramètre $c=-b=\alpha>0$ est à déterminer nous sera utile ultérieurement. Dans les applications du chapitre suivant, on connaît en effet le support de σ_1, égal à $[A,B]$, et les conditions suffisantes du chapitre précédent nous assurent que *supp* $(\sigma_1-\mu_1^t)=[\alpha,B]$.

Définition 7.6.7. *Supposons* $\mathrm{d}\sigma$ *impaire et* $\mathrm{supp}\,(\sigma_1)=[A,B]\subset(0,\infty)$.
On définit dans ce cas

$$\mathfrak{L}_t:\alpha\mapsto\mathfrak{F}_t(-B,-\alpha,\alpha,B).$$

Dans le contexte expliqué ci-dessus, on cherche les points critiques de la fonction $\mathfrak{L}_t$. On a dans ce cas l'énoncé suivant.

Proposition 7.6.8. *On a*

$$I(-B,-\alpha,\alpha,B,k)=\frac{2}{K'(k)}\int_0^1\frac{\sigma\,(0,\alpha u)\;\mathrm{d}u}{\sqrt{(1-u^2)(1-k^2u^2)}}$$

et

$$J(-B,-\alpha,\alpha,B,k)=0.$$

On a en outre

$$\frac{d}{d\alpha}I(-B,-\alpha,\alpha,B,k)=\frac{2}{K'(k)}\int_0^1\frac{u\,\mathrm{d}\sigma(\alpha u)}{\sqrt{(1-u^2)(1-k^2u^2)}}.\tag{7.10}$$

PREUVE : La reformulation des intégrales provient directement de leur définition et de l'hypothèse de symétrie faite à propos de la mesure σ.

Pour obtenir la dérivée énoncée en (7.10), on remarque pour commencer que J ne dépend pas de α, et la dérivation de I est immédiate. □

7.7 Premières équations intégrales

La proposition suivante donne les trois premières équations intégrales obtenues lorsque le quadruplet (a, b, c, d) est un point critique de la $\mathfrak{F}_t$-fonctionnelle.

Ces équations sont obtenues en éliminant les termes en $\frac{\partial F}{\partial k}$ où F est la fonction elliptique utilisée grâce aux équations aux dérivées partielles vérifiées par le paramètre $k(a, b, c, d)$ énoncées à la proposition 7.5.2.

Proposition 7.7.1. *Si le quadruplet (a, b, c, d) est un point critique de la $\mathfrak{F}_t$-fonctionnelle, il vérifie les trois équations intégrales suivantes :*

$$\int_{\mathcal{T}_1} \frac{\mathrm{d}\sigma(z)}{\sqrt{R(z)}} = \int_{\mathcal{T}_2} \frac{\mathrm{d}\sigma(z)}{\sqrt{R(z)}}, \tag{7.11}$$

$$\int_{\mathcal{T}_1} \frac{z\,\mathrm{d}\sigma(z)}{\sqrt{R(z)}} = \int_{\mathcal{T}_2} \frac{z\,\mathrm{d}\sigma(z)}{\sqrt{R(z)}}, \tag{7.12}$$

et

$$\int_{\mathcal{T}_1} \frac{z^2\,\mathrm{d}\sigma(z)}{\sqrt{R(z)}} = \int_{\mathcal{T}_2} \frac{z^2\,\mathrm{d}\sigma(z)}{\sqrt{R(z)}}. \tag{7.13}$$

PREUVE : On cherche maintenant à obtenir les équations intégrales issues de l'équation

$$\nabla \mathfrak{F}_t(a, b, c, d) = 0$$

en éliminant les termes issus de la dérivation par rapport à k. On écrit alors

$$\frac{\partial \mathfrak{G}_t}{\partial a} = \frac{\partial \mathfrak{G}_t}{\partial k}\frac{\partial k}{\partial a} + \frac{\partial I}{\partial a} + \frac{\partial J}{\partial a}, \tag{7.14}$$

$$\frac{\partial \mathfrak{G}_t}{\partial b} = \frac{\partial \mathfrak{G}_t}{\partial k}\frac{\partial k}{\partial b} + \frac{\partial I}{\partial b} + \frac{\partial J}{\partial b}, \tag{7.15}$$

$$\frac{\partial \mathfrak{G}_t}{\partial c} = \frac{\partial \mathfrak{G}_t}{\partial k}\frac{\partial k}{\partial c} + \frac{\partial I}{\partial c} + \frac{\partial J}{\partial c}, \tag{7.16}$$

$$\frac{\partial \mathfrak{G}_t}{\partial d} = \frac{\partial \mathfrak{G}_t}{\partial k}\frac{\partial k}{\partial d} + \frac{\partial I}{\partial d} + \frac{\partial J}{\partial d} \tag{7.17}$$

et d'après les équations de la proposition 7.5.2, on effectue les combinaisons linéaires suivantes :

$$(7.14)(c-a)(d-a) + (7.15)(c-b)(d-b), \tag{7.18}$$

$$(7.14)(b-a)(d-a) - (7.16)(d-c)(c-b), \tag{7.19}$$

$$(7.14)(c-a)(b-a) + (7.17)(d-c)(d-b). \tag{7.20}$$

Cela nous donne les équation suivantes : (7.18) s'écrit

$$\begin{aligned}\frac{d-a}{2(b-a)}\int_{\mathcal{T}_1}\sqrt{P}\,\mathrm{d}\sigma = \frac{c-b}{2(b-a)}\int_{\mathcal{T}_2}\frac{\mathrm{d}\sigma}{\sqrt{P}} - \frac{c+b-2a}{2(b-a)}\int_{\mathcal{T}_2}\frac{\mathrm{d}\sigma}{\sqrt{P}}\\ + (c-a)\int_{\mathcal{T}_2}\frac{z-d}{\sqrt{R(z)}}\,\mathrm{d}\sigma(z) + \frac{d+a-2b}{2(b-a)}\int_{\mathcal{T}_1}\sqrt{P}\,\mathrm{d}\sigma\\ - (d-b)\int_{\mathcal{T}_1}\frac{z-c}{\sqrt{R(z)}}\,\mathrm{d}\sigma(z),\end{aligned}$$

soit après simplification

$$-\int_{\mathcal{T}_1}\sqrt{P}\,\mathrm{d}\sigma - \int_{\mathcal{T}_2}\frac{\mathrm{d}\sigma}{\sqrt{P}} + (c-a)\int_{\mathcal{T}_2}\frac{z-d}{\sqrt{R(z)}}\,\mathrm{d}\sigma(z) - (d-b)\int_{\mathcal{T}_1}\frac{z-c}{\sqrt{R(z)}}\,\mathrm{d}\sigma(z) = 0. \quad (7.21)$$

L'équation (7.19) s'écrit alors

$$\begin{aligned}\frac{d-a}{2(c-a)}\int_{\mathcal{T}_1}\sqrt{P}\,\mathrm{d}\sigma = -\frac{c-b}{2(c-a)}\int_{\mathcal{T}_2}\frac{\mathrm{d}\sigma}{\sqrt{P}} - \frac{c+b-2a}{2(c-a)}\int_{\mathcal{T}_2}\frac{\mathrm{d}\sigma}{\sqrt{P}}\\ + (b-a)\int_{\mathcal{T}_2}\frac{z-d}{\sqrt{R(z)}}\,\mathrm{d}\sigma(z) + \frac{d+a-2c}{2(c-a)}\int_{\mathcal{T}_1}\sqrt{P}\,\mathrm{d}\sigma\\ - (d-c)\int_{\mathcal{T}_1}\frac{z-b}{\sqrt{R(z)}}\,\mathrm{d}\sigma(z),\end{aligned}$$

soit après simplification

$$-\int_{\mathcal{T}_1}\sqrt{P}\,\mathrm{d}\sigma - \int_{\mathcal{T}_2}\frac{\mathrm{d}\sigma}{\sqrt{P}} + (b-a)\int_{\mathcal{T}_2}\frac{z-d}{\sqrt{R(z)}}\,\mathrm{d}\sigma(z) - (d-c)\int_{\mathcal{T}_1}\frac{z-c}{\sqrt{R(z)}}\,\mathrm{d}\sigma(z) = 0 \quad (7.22)$$

et enfin, (7.20) s'écrit

$$\begin{aligned}-\int_{\mathcal{T}_1}\sqrt{P}\,\mathrm{d}\sigma = \frac{c+b-2a}{2(d-a)}\int_{\mathcal{T}_2}\frac{\mathrm{d}\sigma}{\sqrt{P}} - \frac{(b-a)(c-a)}{d-a}\int_{\mathcal{T}_2}\frac{z-d}{\sqrt{R(z)}}\,\mathrm{d}\sigma(z)\\ - \frac{c+b-2d}{2(d-a)}\int_{\mathcal{T}_2}\frac{\mathrm{d}\sigma}{\sqrt{P}} + \frac{(d-c)(d-b)}{d-a}\int_{\mathcal{T}_2}\frac{z-a}{\sqrt{R(z)}}\,\mathrm{d}\sigma(z),\end{aligned}$$

soit à nouveau après simplifications

$$\begin{aligned}&-\int_{\mathcal{T}_1}\sqrt{P}\,\mathrm{d}\sigma - \int_{\mathcal{T}_2}\frac{\mathrm{d}\sigma}{\sqrt{P}} + \frac{(b-a)(c-a)}{d-a}\int_{\mathcal{T}_2}\frac{z-d}{\sqrt{R(z)}}\,\mathrm{d}\sigma(z), \quad (7.23)\\ &-\frac{(d-c)(d-b)}{d-a}\int_{\mathcal{T}_2}\frac{z-a}{\sqrt{R(z)}}\,\mathrm{d}\sigma(z) = 0.\end{aligned}$$

On obtient alors à partir de (7.21), (7.22) et (7.23)

$$\begin{aligned}&(c-a)\int_{\mathcal{T}_2}\frac{z-d}{\sqrt{R(z)}}\,\mathrm{d}\sigma(z) - (d-b)\int_{\mathcal{T}_1}\frac{z-c}{\sqrt{R(z)}}\,\mathrm{d}\sigma(z) =\\ &(b-a)\int_{\mathcal{T}_2}\frac{z-d}{\sqrt{R(z)}}\,\mathrm{d}\sigma(z) - (d-c)\int_{\mathcal{T}_1}\frac{z-b}{\sqrt{R(z)}}\,\mathrm{d}\sigma(z)\end{aligned}$$

et

$$(c-a)\int_{\mathcal{T}_2}\frac{z-d}{\sqrt{R(z)}}\,\mathrm{d}\sigma(z)-(d-b)\int_{\mathcal{T}_1}\frac{z-c}{\sqrt{R(z)}}\,\mathrm{d}\sigma(z)=$$
$$\frac{(b-a)(c-a)}{d-a}\int_{\mathcal{T}_2}\frac{z-d}{\sqrt{R(z)}}\,\mathrm{d}\sigma(z)-\frac{(d-c)(d-b)}{d-a}\int_{\mathcal{T}_2}\frac{z-a}{\sqrt{R(z)}}\,\mathrm{d}\sigma(z),$$

d'où

$$\int_{\mathcal{T}_2}\frac{z}{\sqrt{R(z)}}\,\mathrm{d}\sigma(z)-d\int_{\mathcal{T}_2}\frac{1}{\sqrt{R(z)}}\,\mathrm{d}\sigma(z)=$$
$$\int_{\mathcal{T}_1}\frac{z}{\sqrt{R(z)}}\,\mathrm{d}\sigma(z)-d\int_{\mathcal{T}_1}\frac{1}{\sqrt{R(z)}}\,\mathrm{d}\sigma(z)$$

et

$$\int_{\mathcal{T}_2}\frac{z}{\sqrt{R(z)}}\,\mathrm{d}\sigma(z)-c\int_{\mathcal{T}_2}\frac{1}{\sqrt{R(z)}}\,\mathrm{d}\sigma(z)=$$
$$\int_{\mathcal{T}_1}\frac{z}{\sqrt{R(z)}}\,\mathrm{d}\sigma(z)-c\int_{\mathcal{T}_1}\frac{1}{\sqrt{R(z)}}\,\mathrm{d}\sigma(z),$$

ce qui donne finalement par combinaison linéaire les deux équations intégrales suivantes énoncées ci-dessus : (7.11)

$$\int_{\mathcal{T}_1}\frac{\mathrm{d}\sigma(z)}{\sqrt{R(z)}}=\int_{\mathcal{T}_2}\frac{\mathrm{d}\sigma(z)}{\sqrt{R(z)}}$$

et (7.12)

$$\int_{\mathcal{T}_1}\frac{z\,\mathrm{d}\sigma(z)}{\sqrt{R(z)}}=\int_{\mathcal{T}_2}\frac{z\,\mathrm{d}\sigma(z)}{\sqrt{R(z)}}.$$

De plus, comme

$$\int_{\mathcal{T}_1}\sqrt{P}\,\mathrm{d}\sigma=-\int_{\mathcal{T}_1}\frac{(z-b)(z-c)}{\sqrt{R(z)}}\,\mathrm{d}\sigma(z)$$
$$=-\int_{\mathcal{T}_1}\frac{z^2}{\sqrt{R(z)}}\,\mathrm{d}\sigma(z)+(b+c)\int_{\mathcal{T}_1}\frac{z}{\sqrt{R(z)}}\,\mathrm{d}\sigma(z)-bc\int_{\mathcal{T}_1}\frac{1}{\sqrt{R(z)}}\,\mathrm{d}\sigma(z)$$

et

$$\int_{\mathcal{T}_2}\frac{\mathrm{d}\sigma}{\sqrt{P}}=\int_{\mathcal{T}_2}\frac{z^2\,\mathrm{d}\sigma(z)}{\sqrt{R(z)}}-(a+d)\int_{\mathcal{T}_2}\frac{z\,\mathrm{d}\sigma(z)}{\sqrt{R(z)}}+ad\int_{\mathcal{T}_2}\frac{\mathrm{d}\sigma(z)}{\sqrt{R(z)}},$$

l'équation (7.23) donne alors compte tenu de (7.11) et (7.12) l'égalité (7.13)

$$\int_{\mathcal{T}_1}\frac{z^2\,\mathrm{d}\sigma(z)}{\sqrt{R(z)}}=\int_{\mathcal{T}_2}\frac{z^2\,\mathrm{d}\sigma(z)}{\sqrt{R(z)}},$$

ce qui termine la démonstration de la proposition. □

7.8 Une autre équation intégrale

Il nous reste maintenant à établir l'équation intégrale obtenue par dérivation de la $\mathfrak{F}_t$-fonctionnelle en conservant les termes o0btenus par dérivation par rapport à k.

Proposition 7.8.1. *Si (a,b,c,d) est un point critique de la $\mathfrak{F}_t$-fonctionnelle, alors on a l'équation*

$$\begin{aligned}
t &= \frac{1}{\pi}\frac{M}{c-b}\int_b^c \Pi\big(1,1-k^2h(z)^2,k'\big)k^2h(z)\sqrt{P(z)}\,\mathrm{d}\sigma(z)\\
&+\frac{1}{\pi}\frac{M}{d-a}\int_{(-\infty,a)\cup(d,+\infty)}\Pi\big(1,1-k^2l(z)^2,k'\big)k^2l(z)\frac{\mathrm{d}\sigma(z)}{\sqrt{P(z)}}\\
&-\frac{K'(k)}{\pi}\frac{kM}{(d-a)(c-b)}\bigg((b+c-a-d)\int_b^c\frac{z^2\,\mathrm{d}\sigma(z)}{\sqrt{R(z)}}+2(ad-bc)\int_b^c\frac{z d\sigma(z)}{\sqrt{R(z)}}\\
&+(abc-abd-acd+bcd)\int_b^c\frac{\mathrm{d}\sigma(z)}{\sqrt{R(z)}}\bigg).
\end{aligned}$$

Remarque. Cette équation intégrale fait intervenir la troisième fonction elliptique de Legendre Π, et se prête *a priori* mal à une résolution numérique, même dans le cas de deux intervalles réels symétriques par rapport à l'origine malgré la formulation simplifiée de cette équation pour ce cas particulier énoncée théorème 7.8.2.

On remarque cependant que celle-ci dépend linéairement de t, ce qui permet pour un quadruplet (a,b,c,d) fixé de déterminer le paramètre t associé en évaluant les intégrales mises en jeu ci-dessus.

PREUVE : Rappelons l'expression de la $\mathfrak{F}_t$-fonctionnelle établie à la proposition 7.6.2 :

$$\begin{aligned}
\frac{1}{\pi}\mathfrak{G}_t(a,b,c,d,k) &= -2t\frac{K(k)}{K'(k)}+\int_0^1 f(u,k)\,\mathrm{d}\sigma\circ T(u)-\int_0^1 f(u,k)\,\mathrm{d}\sigma\circ Q(u)\\
&\quad-\int_{-1}^0 f(-u,k)\,\mathrm{d}\sigma\circ T(u)+\int_{-1}^0 f(-u,k)\,\mathrm{d}\sigma\circ Q(u).
\end{aligned}$$

On applique la proposition 7.2.2, d'où

$$\begin{aligned}
\frac{1}{\pi}\frac{\partial\mathfrak{G}_t}{\partial k}(a,b,c,d,k) &= -2t\frac{d}{dk}\frac{K}{K'}(k)+\int_0^1\frac{\partial f}{\partial k}(u,k)\,\mathrm{d}\sigma\circ T(u)-\int_0^1\frac{\partial f}{\partial k}(u,k)\,\mathrm{d}\sigma\circ Q(u)\\
&\quad-\int_{-1}^0\frac{\partial f}{\partial k}(-u,k)\,\mathrm{d}\sigma\circ T(u)+\int_{-1}^0\frac{\partial f}{\partial k}(-u,k)\,\mathrm{d}\sigma\circ Q(u),\\
&=\frac{1}{k(1-k^2)\left(K'(k)\right)^2}\bigg[-\pi t+\int_{-1}^1\Pi(1,1-k^2u^2,k')\frac{k^2u\sqrt{1-u^2}}{\sqrt{1-k^2u^2}}\,\mathrm{d}\sigma\circ T(u)\\
&\quad-\int_{-1}^1\Pi(1,1-k^2u^2,k')\frac{k^2u\sqrt{1-u^2}}{\sqrt{1-k^2u^2}}\,\mathrm{d}\sigma\circ Q(u)\bigg],\\
&\quad\frac{1}{k(1-k^2)\big(K'(k)\big)^2}\bigg[-\pi t+\int_b^c\Pi\left(1,1-k^2h(z)^2,k'\right)\frac{k^2h(z)\sqrt{1-h(z)^2}}{\sqrt{1-k^2h(z)^2}}\,\mathrm{d}\sigma(z)\\
&\quad+\int_{(-\infty,a)\cup(d,+\infty)}\Pi\left(1,1-k^2l(z)^2,k'\right)\frac{k^2l(z)\sqrt{1-l(z)^2}}{\sqrt{1-k^2l(z)^2}}\,\mathrm{d}\sigma(z)\bigg].
\end{aligned}$$

On applique maintenant la proposition 7.6.5, ce qui donne

$$\frac{1}{\pi}\frac{\partial \mathfrak{G}_t}{\partial k}(a,b,c,d,k) = \frac{1}{k(1-k^2)\left(K'(k)\right)^2}\Bigg[-\pi t + \int_b^c \Pi\left(1, 1-k^2h(z)^2, k'\right) k^2 h(z) \frac{M}{c-b}\sqrt{P(z)}\,\mathrm{d}\sigma(z)$$
$$+ \int_{(-\infty,a)\cup(d,+\infty)} \Pi\left(1, 1-k^2l(z)^2, k'\right) k^2 l(z) \frac{M}{(d-a)\sqrt{P(z)}}\,\mathrm{d}\sigma(z)\Bigg].$$

Avec les notations précédentes, il nous reste à écrire que

$$\frac{\partial \mathfrak{F}_t}{\partial a}(a,b,c,d) = 0$$

si et seulement si

$$k(1-k^2)\left(K'(k)\right)^2 \frac{\partial \mathfrak{G}_t}{\partial k} = -\frac{k(1-k^2)\left(K'(k)\right)^2}{\frac{\partial k}{\partial a}}\left(\frac{\partial I}{\partial a} + \frac{\partial J}{\partial a}\right) \tag{7.24}$$

et d'après les équations intégrales précédemment obtenues à la proposition 7.6.6 et le calcul de $\frac{\partial k}{\partial a}$ donné en (7.6), le second membre de (7.24) vaut après calculs

$$\frac{kM}{(d-a)(c-b)}\Bigg((b+c-a-d)\int_b^c \frac{z^2\,\mathrm{d}\sigma(z)}{\sqrt{R(z)}} + 2(ad-bc)\int_b^c \frac{z\mathrm{d}\sigma(z)}{\sqrt{R(z)}}$$
$$+ (abc-abd-acd+bcd)\int_b^c \frac{\mathrm{d}\sigma(z)}{\sqrt{R(z)}}\Bigg),$$

d'où le résultat. □

Théorème 7.8.2. *Soient quatre réels $a<b<c<d$ tels que (a,b,c,d) soit un point critique de la $\mathfrak{F}_t$-fonctionnelle, on a alors les équations intégrales suivantes :*

$$\int_{\mathcal{T}_1} \frac{\mathrm{d}\sigma(z)}{\sqrt{R(z)}} = \int_{\mathcal{T}_2} \frac{\mathrm{d}\sigma(z)}{\sqrt{R(z)}},$$
$$\int_{\mathcal{T}_1} \frac{z\,\mathrm{d}\sigma(z)}{\sqrt{R(z)}} = \int_{\mathcal{T}_2} \frac{z\,\mathrm{d}\sigma(z)}{\sqrt{R(z)}},$$
$$\int_{\mathcal{T}_1} \frac{z^2\,\mathrm{d}\sigma(z)}{\sqrt{R(z)}} = \int_{\mathcal{T}_2} \frac{z^2\,\mathrm{d}\sigma(z)}{\sqrt{R(z)}}.$$

et

$$t = \frac{1}{\pi}\frac{M}{c-b}\int_b^c \Pi\big(1, 1-k^2h(z)^2, k'\big)k^2h(z)\sqrt{P(z)}\,\mathrm{d}\sigma(z)$$
$$+ \frac{1}{\pi}\frac{M}{d-a}\int_{(-\infty,a)\cup(d,+\infty)} \Pi\big(1, 1-k^2l(z)^2, k'\big)k^2l(z)\frac{\mathrm{d}\sigma(z)}{\sqrt{P(z)}}$$
$$- \frac{K'(k)}{\pi}\frac{kM}{(d-a)(c-b)}\Bigg((b+c-a-d)\int_b^c \frac{z^2\,\mathrm{d}\sigma(z)}{\sqrt{R(z)}} + 2(ad-bc)\int_b^c \frac{z\,\mathrm{d}\sigma(z)}{\sqrt{R(z)}}$$
$$+ (abc-abd-acd+bcd)\int_b^c \frac{\mathrm{d}\sigma(z)}{\sqrt{R(z)}}\Bigg).$$

Dans le cas symétrique $d = -a := \beta$, $c = -b := \alpha$, *et* $\mathrm{d}\sigma$ *impaire, ce système se réécrit de façon plus compacte*

$$\int_0^\alpha \frac{z\,\mathrm{d}\sigma(z)}{\sqrt{(z^2-\alpha^2)(z^2-\beta^2)}} = \int_\beta^{+\infty} \frac{z\,\mathrm{d}\sigma(z)}{\sqrt{(z^2-\alpha^2)(\beta^2-z^2)}} \tag{7.25}$$

et

$$t = \frac{2}{\beta\pi}\int_{[0,\alpha]\cup[\beta,+\infty]} \Pi\left(1, 1-\frac{z^2}{\beta^2}, \sqrt{1-\frac{\alpha^2}{\beta^2}}\right) z\sqrt{\left|\frac{z^2-\alpha^2}{z^2-\beta^2}\right|}\,\mathrm{d}\sigma(z), \tag{7.26}$$

$$= \frac{2}{\beta\pi}\int_{[0,\alpha]\cup[\beta,+\infty]} \Pi\left(1, 1-\frac{\alpha^2}{z^2}, \sqrt{1-\frac{\alpha^2}{\beta^2}}\right) z\sqrt{\left|\frac{z^2-\beta^2}{z^2-\alpha^2}\right|}\,\mathrm{d}\sigma(z). \tag{7.27}$$

Avec les notations et sous les hypothèses de 7.6.7, si $\mathfrak{L}_t'(\alpha) = 0$, *le réel* α *vérifie l'équation intégrale suivante :*

$$t = \frac{2}{B\pi}\int_{[0,\alpha]\cup[B,+\infty]} \Pi\left(1, 1-\frac{\alpha^2}{z^2}, \sqrt{1-\frac{\alpha^2}{B^2}}\right) z\sqrt{\left|\frac{z^2-B^2}{z^2-\alpha^2}\right|}\,\mathrm{d}\sigma(z). \tag{7.28}$$

PREUVE : Toutes les assertions relatives au cas non-symétrique ont été prouvées auparavant.

Il suffit alors de constater que dans le cas symétrique les équations (7.11) et (7.13) deviennent triviales par imparité de la contrainte σ, et ensuite d'utiliser le changement de variable issu de [AbSt64, 17.7.9]

$$\Pi\left(1, 1-\frac{\alpha^2}{z^2}, \sqrt{1-\frac{\alpha^2}{\beta^2}}\right) = \frac{z^2}{\alpha^2}\frac{z^2-\alpha^2}{z^2-\beta^2}\Pi\left(1, 1-\frac{z^2}{\beta^2}, \sqrt{1-\frac{\alpha^2}{\beta^2}}\right) + \frac{z^2}{\alpha^2}\frac{\beta^2-\alpha^2}{\beta^2-z^2}K'\left(\frac{\alpha}{\beta}\right),$$

pour obtenir l'expression simplifiée (7.26).

Dans le cas de la fonction $\mathfrak{L}_t$ en tant que fonction de α, on obtient d'après les formules données à la proposition 7.6.8 et le changement de variable ci-dessus l'équation intégrale énoncée. □

Nous avons bien obtenu des expressions intégrales données ci-dessus pour caractériser la famille d'ensembles $\left(supp\left(\sigma_j - \mu_j^\tau\right)\right)_{0\leqslant t\leqslant\tau}$, $j \in \{1,2\}$, étape nécessaire pour les applications numériques dans le chapitre qui suit.

Pour utiliser ces résultats de façon concrète, on dispose donc dans le cas symétrique de deux équations intégrales non triviales, celle qui nous intéresse donne le lien entre le paramètre t et la partie libre de la contrainte et a nécessité le plus de calculs dans ce qui précède, nous expliquerons dans le chapitre qui suit la façon dont nous avons malgré tout pu tirer parti de cette équation.

Chapitre 8

Exemples numériques

On présente dans ce qui suit des exemples numériques dans le but d'illustrer nos résultats théoriques concernant la quantité de Zolotarev pour des ensembles discrets $Z_n(E_N, F_N)$ ainsi que l'impact de notre travail sur le choix des paramètres pour la méthode ADI, même si les exemples étudiés ici restent de nature académique.

Les exemples qui seront étudiés ici concernent des matrices dont on peut déterminer tout le spectre facilement, mais on verra lors des procédures numériques qu'il ne nous est en réalité seulement nécessaire de connaître les valeurs propres proches de zéro des matrices considérées.

8.1 Exemples étudiés

8.1.1 Valeurs propres équidistantes

On choisit pour commencer l'exemple le plus classique des valeurs propres équidistantes : ici,

$$A_N := diag\left(\frac{1}{2N}, \frac{3}{2N}, \ldots, \frac{2N-1}{2N}\right),$$

ce qui revient à considérer un problème de Zolotarev symétrique pour l'ensemble discret

$$E_N^{equi} := \left\{\frac{2k-1}{2N} : 1 \leqslant k \leqslant N\right\}.$$

ou encore à estimer grâce à la méthode ADI la solution de l'équation de Sylvester $A_N X + X A_N = B_N$ où B_N est un second membre choisi aléatoirement. Le calcul de la distribution limite des valeurs propres ne présente ici bien sûr aucune difficulté, et on a ici $\sigma_1'(x) = 1$ sur $[0, 1]$.

8.1.2 Distribution en cosinus perturbé

Les résultats théoriques concernant l'asymptotique faible du problème de Zolotarev pour des ensembles discrets nous incitent à penser qu'un exemple numérique concernant une matrice comportant peu de valeurs propres proches de zéro avec une valeur propre isolée très proche de zéro permettrait de mettre une évidence un comportement asymptotique très différent pour les quantités de Zolotarev discrètes et continues.

On choisit dans ce but l'exemple suivant donné par la résolution approchée d'une équation de Sylvester

$$A_N X + X A_N = B_N \text{ avec } A_N := diag\left(\cos\left(\frac{\pi k}{2N}\right) + \frac{1}{N^4}\right)_{1\leqslant k\leqslant N},$$

ce qui revient à considérer un problème de Zolotarev pour l'ensemble discret

$$E_N^{\cos} := \left\{\cos\left(\frac{\pi k}{2N}\right) + \frac{1}{N^4} : 1 \leqslant k \leqslant N\right\}.$$

Il semble intuitivement que la perturbation en $1/N^4$ imposée à la distribution en cosinus considérée ne devrait pas modifier la répartition asymptotique de cette distribution, ce que nous allons rapidement vérifier.

On note $\mathcal{C}^0(\mathbb{R})$ l'ensemble des fonctions continues sur $\mathbb{R}$ à support compact.

Lemme 8.1.1. *On a pour toute fonction f de $\mathcal{C}_0(\mathbb{R})$*

$$\lim_{N\to+\infty} \int_{\mathbb{R}} f(t) d\,\nu_N\left(E_N^{\cos}\right)(t) = \frac{2}{\pi}\int_0^1 f(t)\frac{dt}{\sqrt{1-t^2}}.$$

PREUVE : Pour la suite d'ensembles discrets $(E_N^1)_N$, on a pour toute fonction de $\mathcal{C}_0(\mathbb{R})$

$$\begin{aligned}\lim_{N\to+\infty} \frac{1}{N}\sum_{k=1}^{N} f\left(\cos\left(\frac{k\pi}{2N}\right)\right) &= \int_0^1 f\left(\cos\left(\frac{\pi u}{2}\right)\right) d\,u \\ &= \frac{2}{\pi}\int_0^1 f(t)\frac{dt}{\sqrt{1-t^2}},\end{aligned}$$

et on conclut en majorant

$$f\left(\cos\left(\frac{k\pi}{2N}\right) + \frac{1}{N^4}\right) - f\left(\cos\left(\frac{k\pi}{2N}\right)\right)$$

par uniforme continuité de f. □

Remarque. Le lemme 8.1.1 donne la limite des suites de mesures de comptage normalisées de $(E_N^{\cos})_N$ ou d'un point de vue orienté vers l'algèbre linéaire, détermine le symbole associé à la suite de matrices de Toeplitz $(A_N)_N$, exemple originel d'une suite de matrices admettant une distribution spectrale asymptotique comme mentionné au théorème 5.1.3.

8.1.3 Discrétisation du Laplacien 2D

On considère maintenant une équation de Poisson bidimensionnelle

$$-\frac{\partial^2 u}{\partial x^2}(x,y) - \frac{\partial^2 u}{\partial y^2}(x,y) = f(x,y)$$

dans le carré unité ouvert $0 < x, y < 1$ avec conditions de Dirichlet au bord. La discrétisation aux différences finies usuelle sur la grille uniforme

$$\left(\frac{j}{N+1}, \frac{k}{N+1}\right),\ 0 \leqslant j,k \leqslant N$$

donne lieu à une équation de Lyapounov pour des matrices de taille $N \times N$

$$A_N^{2D} X + X A_N^{2D} = B_N$$

d'inconnue X, où

$$A_N^{2D} := \begin{pmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \dots & 0 & -1 & 2 \end{pmatrix}.$$

Les valeurs propres de A_N sont bien connues

$$\Lambda\left(A_N^{2D}\right) := E_N^{2D} = \left\{2 - 2\cos\left(\frac{k\pi}{N+1}\right)\right\}_{1 \leqslant k \leqslant N},$$

et on a le résultat suivant, également bien connu, qui donne la distribution asymptotique de la suite de matrices de Toeplitz $\left(A_N^{2D}\right)_N$ et se démontre comme le lemme 8.1.1.

Lemme 8.1.2. *On a pour toute fonction f de $\mathcal{C}_0(\mathbb{R})$*

$$\lim_{N \to +\infty} \int_{\mathbb{R}} f(t) d\,\nu_N\left(E_N^{2D}\right)(t) = \frac{1}{\pi} \int_0^4 f(t) \frac{dt}{\sqrt{t(4-t)}},$$

et le spectre $\Lambda\left(A_N^{2D}\right)$ admet donc une densité asymptotique donnée par

$$\omega^{2D}(z) = \frac{\mathbb{1}_{[0,4]}(z)}{\pi} \frac{1}{\sqrt{z(4-z)}}.$$

On pourrait bien entendu utiliser le produit de Kronecker pour reécrire cette équation de Sylvester sous la forme d'un système linéaire pour appliquer un algorithme adapté à cette configuration, y compris en tenant compte du caractère discret des ensembles considérés comme dans [BeKu01a], mais cela conduit à un sysème linéaire de taille N^2 avec des matrices plus complexe que les matrices tridiagonales mises en jeu ici, et la représentation sous la forme d'une équation de Sylvester permet enfin une conception intuitive de la solution X dont la coordonnée (i, j) approche la solution au voisinage du point d'abscisse $\frac{i}{N+1}$ et d'ordonnée $\frac{j}{N+1}$.

8.1.4 Discrétisation du Laplacien 4D

On considère ici une équation de Poisson en dimension 4

$$-\Delta u(w, x, y, z) = f(w, x, y, z)$$

dans l'hypercube unité ouvert $0 < w, x, y, z < 1$ avec conditions de Dirichlet au bord.

On suppose ici que l'entier $N = m^2$ est un carré, et toujours en discrétisant sur une grille uniforme, on obtient l'équation de Lyapounov pour des matrices de taille $N \times N$

$$A_N^{4D} X + X A_N^{4D} = B_N$$

d'inconnue X, où

$$A_N^{4D} := A_m^{2D} \otimes I_m + I_m \otimes A_m^{2D}, \tag{8.1}$$

avec les notations usuelles pour le produit de Kronecker défini en 1.1.3.

Les valeurs propres de A_N^{4D} sont bien connues

$$\Lambda\left(A_N^{4D}\right) := E_N^{4D} = \left\{4 - 2\left(\cos\left(\frac{k\pi}{N+1}\right) + \cos\left(\frac{j\pi}{N+1}\right)\right)\right\}_{1\leqslant j,k\leqslant m}, \tag{8.2}$$

et on a le résultat suivant cité de [BeKu01a, p 23] qui donne le symbole de la suite de matrices de Toeplitz $(A_N^2)_N$ et se démontre comme le lemme 8.1.1.

Lemme 8.1.3. *Le spectre de la suite de matrices $\left(A_N^{4D}\right)_{N\geqslant 1}$ admet une densité asymptotique ω^{4D} donnée par la convolution de ω^{2D} avec lui même : pour $0 < z < 8$, on a*

$$\begin{aligned}\omega^{4D}(z) &= \frac{1}{2}\int \omega^{2D}(z-t)\omega^{2D}(t)\mathrm{d}t \\ &= \frac{1}{2\pi^2}\int_0^z \frac{\mathrm{d}t}{\sqrt{t(z-t)(4-t)(4-z-t)}}.\end{aligned}$$

La fonction ω^{4D} peut également s'exprimer à l'aide de fonctions hypergéométriques, mais cela ne nous est pas utile pour déterminer les paramètres dans ce qui suit.

8.2 Vérification des hypothèses

Ce paragraphe a pour objet la vérification des hypothèses techniques faites dans le chapitre 5 pour chacun des exemples numériques évoqués dans ce chapitre.

On cite à cet effet le théorème [SaTo97, Théorème VI.4.1] qui donne une estimation de l'erreur commise pour un certain choix de discrétisation d'une mesure donnée, ce qui nous sera utile pour vérifier l'hypothèse 5.4.

On remarque enfin que l'hypothèse 5.4 est une hypothèse concernant la mesure *positive* $\sigma_1+\sigma_2$, et que nous sommes dans chaque exemple de ce chapitre dans un cadre symétrique, cela explique que les critères de vérification que l'on énonce ici soient adaptés à des mesures de ce type.

Théorème 8.2.1. *Soit η une mesure de probabilité sur $[-1,1]$ telle qu'il existe K et ε des réels strictement positifs tels que*

$$\int_{|x-t|\leqslant N^{-K}} |\log|x-t||\ \mathrm{d}\eta(t) \leqslant \varepsilon\frac{\log N}{N}. \tag{8.3}$$

Alors, si on choisit des points $(y_{\ell,N})_{0\leqslant\ell\leqslant N}$ dans $[-1,1]$ tels que

$$\int_{y_{\ell,N}}^{y_{\ell+1,N}} \mathrm{d}\eta(t) = \frac{1}{N},\ 0 \leqslant \ell \leqslant N-1, \tag{8.4}$$

et si

$$P_N(x) := \prod_{\ell=1}^{N-1}(x - y_{\ell,N}),$$

alors

$$\frac{1}{4}\,\mathrm{e}^{-NU^{\eta}(x)}\,|y_{N_x,x}-x| \leqslant |P_N(x)| \leqslant N^{K+\varepsilon}\,\mathrm{e}^{-NU^{\eta}(x)},$$

où $y_{N_x,x}$ désigne le point le plus proche de x.

On utilise dans la preuve qui suit les conditions suffisantes énoncées après l'hypothèse 5.4 pour vérifier celle-ci dans le cadre de nos exemples numériques.

Proposition 8.2.2. *Les cinq hypothèses techniques 5.1, 5.2, 5.3, 5.4 et 5.5 sont vérifiées pour les familles d'ensembles discrets $\left(E_N^{equi}\right)_N$, $\left(E_N^{\cos}\right)_N$ et $\left(E_N^{2D}\right)_N$.*

PREUVE : D'après la description de chacun des exemples qui précède, les hypothèses 5.1, 5.2 et 5.5 sont vérifiées. L'hypothèse 5.3 est vérifiée pour chaque exemple d'après la remarque 4.1.2, il ne nous reste plus qu'à prouver qu'on se trouve bien dans le cadre de l'hypothèse 5.4.

D'après les conditions suffisantes énoncées chapitre 5, l'hypothèse 5.4 est à l'évidence vérifiée pour la famille de valeurs propres équidistantes car dans ce cas,

$$\liminf_{N\to+\infty} [N|x_N - y_N|] = 1 > 0, \text{ pour } x_N \neq y_N \in E_N^{equi}.$$

Il n'est pas difficile de se convaincre que toutes les mesures considérées sont suffisamment régulières pour vérifier (8.3) pour un certain K.

L'exemple de la discrétisation du Laplacien bidimensionnel nécessite l'utilisation des propositions 8.2.1 et 5.1.6 citées ci-dessus, et on entre bien dans ce cas dans le cadre d'une discrétisation vérifiant (8.4).

Pour le cas des valeurs propres en cosinus perturbé, la distribution classique en $\left(\cos\left(\frac{\pi k}{2N}\right)\right)_{1\leqslant k\leqslant N}$ vérifie bien une égalité du type (8.4), et on procède comme pour la preuve de 8.1.1 pour montrer que la perturbation en N^{-4} n'a pas d'incidence sur le résultat. □

Enfin, les arguments de monotonie de la proposition 6.4.4 nous permettent d'affirmer que dans tous les cas considérés ici -à l'exception de la discrétisation du Laplacien $4D$ qui fait l'objet de la remarque qui suit- on a $supp\,(\sigma_1 - \mu_1^t) = [a_t, B]$ pour $0 < t < \sigma_1(\mathbb{C}) = 1$.

Remarque. Pour le cas de la discrétisation du Laplacien $4D$, on vérifie facilement les hypothèses 5.1, 5.2, 5.3 et 5.5, mais la complexité du symbole ω^{4D} ne nous permet pas de conclure au sujet de la vérification de l'hypothèse de séparation 5.4.

La matrice donnée en (8.1) possède des valeurs propres doubles d'après l'écriture du spectre de celle-ci en (8.2), et on conjecture que A_N^{4D} possède $\frac{N}{2} + \circ(N)$ valeurs propres doubles.

Comme chaque valeur propre est comptée avec multiplicité 1 dans le calcul des mesures contraintes σ_1 et σ_2, cela ne contribue pas à simplifier notre calcul.

Cet exemple numérique a encore une spécificité parmi ceux présentés dans ce chapitre : dans ce cas, la proposition 6.4.4 ne s'applique pas directement, et on conjecture que $supp\,(\sigma_1 - \mu_1^t)$ est un intervalle centré en 4 (point milieu du support de la contrainte positive), ce qui constituerait une généralisation de [BeKu01a, Lemme 3.1.c] à notre cas.

8.3 Problème de Zolotarev

On explique dans les deux premiers paragraphes de cette section la détermination des paramètres pour les illustrations liées au problème de Zolotarev et à son asymptotique.

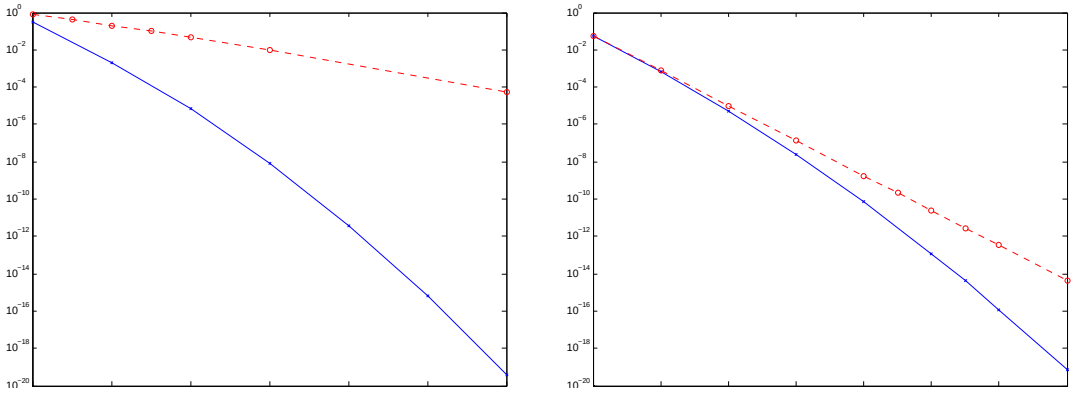

FIG. 8.1 – Rappel des figures 3.1 (à gauche) et 3.2 (à droite) : comparaison de l'asymptotique pour des problèmes de Zolotarev discrets (traits pleins) et continus (traits pointillés)

On compare ensuite la valeur du problème de Zolotarev et de l'asymptotique que l'on a établie dans le cas de l'exemple de la discrétisation du Laplacien bidimensionnel.

8.3.1 Algorithme de Remès rationnel

On reprend ci-dessus les figures 3.1, 3.2 afin d'expliquer plus avant la détermination numérique des paramètres liés à ces figures ainsi qu'à la figure 3.3 qui sera réutilisée ultérieurement dans cette section.

Pour obtenir les figures 3.1, 3.2 et 3.3, il nous a fallu calculer explicitement les quantités de Zolotarev $Z_n(E,F)$ pour $1 \leqslant n \leqslant 20$ pour certains exemples d'ensembles discrets.

Pour cela, nous avons utilisé le lien avec l'approximation rationelle de la fonction signe exposé en 1.3.4. La preuve de cette proposition permet en effet de passer d'un candidat pour le problème de Zolotarev à un candidat pour l'approximation du signe de façon explicite, et notre problème se transforme donc en le calcul de la valeur suivante :

$$S_n(E,F) := \inf_{r\in\mathcal{R}_n} \max_{z\in E\cup F} |r(z) - \text{signe}(z)| \tag{8.5}$$

pour les différentes familles d'ensembles discrets étudiés.

Nous avons pour cela utilisé l'algorithme de Remès rationnel, dit également algorithme d'échange, détaillé dans [Po81, Chapitre 8] pour le cas polynômial et adapté au cas rationnel dans [Po81, Chapitre 10]. Le choix de cet algorithme est justifié par le fait qu'il s'adapte simplement à l'approximation sur des ensembles discrets, on esquisse ici le fonctionnement de cet algorithme dans le contexte de l'approximation d'une fonction continue donnée pour en convaincre le lecteur.

Soient f une fonction continue sur un intervalle I, $(\gamma_\ell)_{1\leqslant\ell\leqslant 2n+1}$ un ensemble de points distincts de I rangés dans l'ordre croissant que l'on appelle *référence* et $r \in \mathcal{R}_n$.

Si les équations

$$r(x_\ell) + (-1)^\ell h = f(x_\ell),\ 0 \leqslant \ell \leqslant 2n+1, \tag{8.6}$$

sont vérifiées pour une certaine constante h, alors d'après [Po81, Théorème 10.1], r est un minimiseur pour la quantité

$$\max_{0\leqslant\ell\leqslant 2n+1} |r(\gamma_\ell) - f(\gamma_\ell)| \tag{8.7}$$

parmi tous les éléments de $\mathcal{R}_n$.

L'algorithme consiste en les étapes suivantes : on montre que h est solution d'un problème aux valeurs propres généralisées

$$\det(A - hB) = 0,$$

pour certaines matrices A positive et B définie positive construites à partir des points de référence $(\gamma_\ell)_{1\leqslant\ell\leqslant 2n+1}$. On trouve alors pour chaque solution h un vecteur de coefficients du dénominateur de r, et on peut montrer qu'il existe une unique solution que l'on note $\vec{b}$ pour laquelle le dénominateur ne s'annule pas. Cela permet ensuite une fois connus h et $\vec{b}$ de retrouver les coefficients du numérateur de r par résolution d'un système linéaire.

La stratégie consiste ensuite à choisir une nouvelle référence en cherchant des points $(\delta_\ell)_{1\leqslant\ell\leqslant 2n+1}$ qui vérifient l'inégalité

$$\max_{0\leqslant\ell\leqslant 2n+1} |r(\gamma_\ell) - f(\gamma_\ell)| \leqslant \max_{0\leqslant\ell\leqslant 2n+1} |r(\delta_\ell) - f(\delta_\ell)|.$$

Dans le cas polynômial où la théorie assure l'existence d'un minimiseur, on sait d'après [Po81, Théorème 8.2] que cette méthode converge vers le meilleur approximant en un nombre fini d'itérations. En l'absence de résultat théorique à ce sujet, nous ne pouvons que constater dans le cas discret que cet algorithme peut fournir une valeur optimale pour 8.7, l'idée étant alors de lancer l'algorithme pour plusieurs choix de références initiales.

En réalité, on utilise plutôt l'algorithme d'échange simple qui consiste pour une itération à changer un seul des points d'une référence donnée.

8.3.2 Asymptotique du problème de Zolotarev

On explique ici comment ont été réalisés les calculs liés à la figure 8.2.

On pose $t = n/N$, et l'on représente figure 8.2 la fonction

$$n \mapsto \exp\Big(-\frac{n}{cap\,(-conv\,(E_{20})\,,\,conv\,(E_{20}))}\Big), \tag{8.8}$$

où $E_N = \{2 - 2\cos(k\pi/(N+1))\}_{1\leqslant k\leqslant N}$ et $N = 20$. Cette fonction décrit l'asymptotique du problème de Zolotarev dans le cas continu d'après (3.1) et correspond à la courbe du dessus de la figure 8.2. On représente également figure 8.2 la fonction

$$n \mapsto \exp\left(-N\int_0^{n/N} \frac{\mathrm{d}\tau}{cap\,([-B, a_\tau], [a_\tau, B])}\right), \tag{8.9}$$

construite pour le cas discret d'après le corollaire 6.4.5 qui correspond à la courbe du dessous de la figure 8.2.

À l'évidence, les deux expressions diffèrent et n'ont pas le même comportement asymptotique lorsque n devient grand, ce qui semble dû au fait que l'expression (8.8) tient compte de façon plus fine que l'expression (8.9) de la structure des ensembles discrets étudiés. On remarque cependant que les deux courbes de la figure 8.2 ont la même pente à l'origine, le régime de convergence superlinéaire ne s'établit pas dès les premières itérations.

On remarque également que la courbe correspondant à l'asymptotique de la quantité de Zolotarev sur des ensembles discrets est concave (sur une échelle semi-logarithmique),

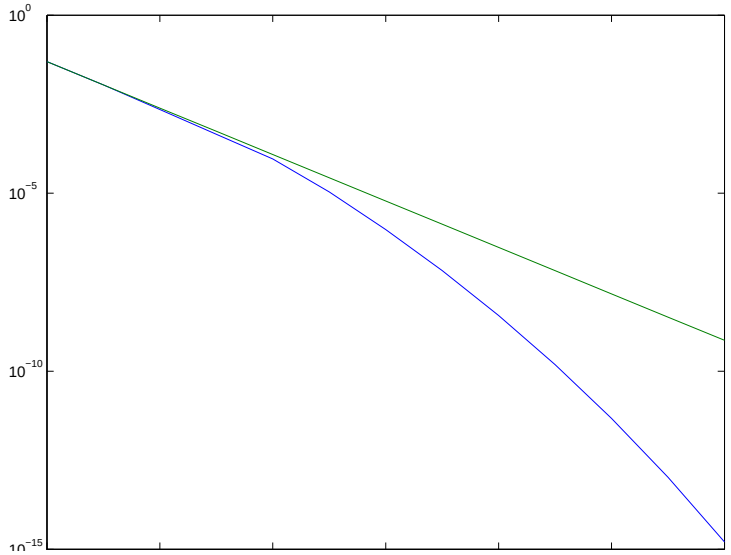

FIG. 8.2 – *Asymptotique du problème de Zolotarev, cas discret et continu,* $N = 20$*, cas du Laplacien* $2D$*, en abscisse le degré des fractions rationnelles considérées* n*, où* $1 \leqslant n \leqslant 14$.

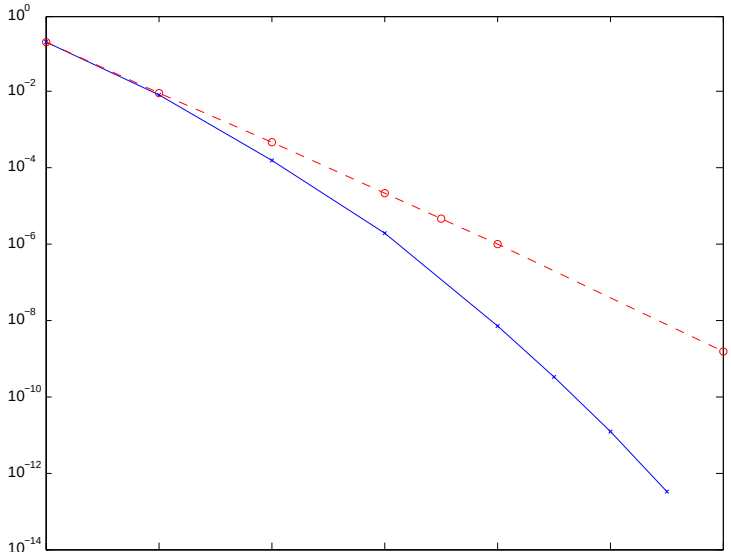

FIG. 8.3 – *Rappel de la figure 3.2 :* $Z_n(E, -E)$ et $Z_n(\mathrm{conv}(E), -\mathrm{conv}(E))$, $E_N = \{2 - 2\cos(k\pi/(N+1))\}_{1 \leqslant k \leqslant N}$, $N = 20$, où $1 \leqslant n \leqslant 14$.

comme prévu par notre théorie : en effet, les fonctions $\tau \mapsto a_\tau$ et $\tau \mapsto b_\tau$ mises en jeu dans le corollaire 6.4.5 qui détermine l'asymptotique de la quantité de Zolotarev pour des ensembles discrets sont respectivement croissante et décroissante, ce qui résulte de la monotonie de la famille $(S^\tau)_\tau$, voir (6.6).

On rappelle que notre théorie nous permet seulement d'établir des propriétés asymptotiques pour l'asymptotique de la quantité de Zolotarev au sens de la racine N-ième, alors que la figure 8.2 tend à conjecturer un résultat d'asymptotique forte.

Il nous faut maintenant expliquer la détermination des valeurs numériques pour les constantes a_t.

D'après la proposition 6.3.2, on peut déterminer a_t à t fixé en recherchant les points critiques de la fonctionnelle

$$a \mapsto \mathfrak{F}_t([-B,-a],[a,B]),$$

et les équations intégrales qui découlent de ce travail ont été énoncées théorème 7.8.2.

Cependant comme l'équation intégrale énoncée avec des paramètres $\alpha = a_t$ et $\beta = B$ en (7.26) est linéaire en le paramètre t mais dépend de façon complexe du paramètre a, il est plus simple de procéder de façon contraire, en fixant $a \in (A,B)$ et ainsi en déterminant le paramètre $t > 0$ tel que $a = a_t$.

À ce sujet, nous ne disposons pas pour le moment d'un algorithme nous permettant de résoudre les équations intégrales en a à t fixé de façon raisonnable.

Dans le cas particulier de deux intervalles symétriques par rapport à l'origine, nous pourrions également déterminer le paramètre a_t en suivant la méthode décrite lors de la preuve de la proposition 6.4.4 : dans les deux cas que nous étudions, la fonction $x \mapsto \sigma_1'(x)\sqrt{(x^2-A^2)(B^2-x^2)}$ s'annule pour $x = A$, et ainsi, d'après [BeKu01a, Lemme 3.1], la quantité $\alpha(t)$ est déterminée par les équations intégrales

$$t = \int_{A^2}^{\alpha(t)} \sqrt{\frac{B^2-x}{\alpha(t)-x}}\sigma_1'(\sqrt{x})\,dx.$$

Ceci donne par exemple pour le cas des valeurs du cosinus perturbé la valeur $\alpha(t) = (4t/\pi)^2$, alors que les expressions liées au cas de la discrétisation du Laplacien sont plus complexes.

8.3.3 Comparaison dans la cas du Laplacien bidimensionnel

Les courbes produites pour l'estimation de l'asymptotique de la quantité de Zolotarev prédisent de façon fiable le comportement de celle-ci pour des ensembles discrets d'après les figures 8.2 et 8.3.

Ces expériences numériques nous incitent à conjecturer notre résultat concernant l'asymptotique de la quantité de Zolotarev pour des ensembles discrets dans un cadre d'asymptotique forte, ce type de résultat d'asymptotique fort étant parfois connu dans le cadre d'ensembles continus d'après 3.1. Cependant, l'obtention d'un tel résultat d'asymptotique faible dépasserait très probablement la cadre de la théorie du potentiel logarithmique qui a été exposé ici et nécessiterait probablement l'utilisation d'un arsenal théorique considérablement plus développé.

8.4 Méthode ADI

Passons maintenant à l'application de nos résultats à la méthode ADI.

Comme mentionné dans la partie décrivant à la méthode ADI, notre point de départ est l'équation de Lyapounov $A_N X + X A_N = B_N$ d'inconnue X avec les suites de matrices considérées dans les exemples évoqués plus haut.

On a choisi pour chaque exemple de présenter la situation correspondant à des matrices de taille $N = 100$ puis de taille $N = 1000$. Dans chaque cas, on fixe une solution choisie aléatoirement X, ce qui nous permet de calculer le membre de droite B_N. Enfin, on a choisi la donnée initiale $X_0 = 0$ pour initialiser l'algorithme ADI.

À propos de la taille des matrices considérées qui peut sembler limitée, on rappelle qu'une solution possible pour la résolution approchée d'une équation de Lyapounov consiste à transformer celle-ci en système linéaire par application du produit de Kronecker, ce qui a pour effet d'élever au carré la dimension des objets mis en jeu. Ainsi, les situations respectives comparables à nos calculs pour des matrices 100×100 et 1000×1000 sont données par la résolution approchée de systèmes linéaires pour des seconds membres de tailles 10^4 et 10^6.

Les quatre courbes de chacune des figures 8.4, 8.5, 8.6, 8.7, 8.8, 8.9, 8.10 et 8.11 ont la signification suivante : deux courbes correspondent aux bornes supérieures pour les erreurs commises calculées à partir des fractions rationnelles et énoncées en (1.7), à savoir la courbe marquée par des croix dans le cas de paramètres classiques calculés pour les enveloppes convexes des ensembles discrets et la courbe marquée par des cercles dans le cas de paramètres adaptés à la nature discrète de nos ensembles, et deux courbes correspondent au calcul de l'erreur relative commise au bout de n itérations $\|X_n - X\|/\|X\|$, la courbe pleine dans le cas de paramètres continus et la courbe pointillée dans le cas de paramètres discrets.

Nos résultats théoriques portent sur l'asymptotique de la racine N-ième de la quantité de Zolotarev, ce sont donc des résultats d'asymptotique faible, alors que l'on représente l'erreur et la borne supérieure de celle-ci pour la méthode ADI, ce qui n'entre donc pas dans le cadre de nos résultats théoriques.

On observe dans presque toutes les figures un phénomène numérique classique, les courbes d'erreur deviennent plates en deçà d'une certaine erreur, ce qui est bien entendu lié aux diverses erreurs numériques qui affectent le calcul.

On remarque deux régimes distincts selon la taille des matrices considérées : pour des matrices 100×100, le phénomène de convergence superlinéaire apparaît très clairement dans les exemples les plus favorables, alors que celui-ci n'est pas évident pour des matrices de taille 1000×1000. En réalité, on peut observer un tel phénomène pour les matrices les plus grandes, mais celui-ci apparaît une fois que l'on a effectué plus d'itérations, ce qui correspond dans notre situation à des approximations de la solution exacte dont la précision dépasse de très loin la précision machine.

Le résultats sont comme prévu les plus spectaculaires pour la distribution en cosinus perturbé ce qui est dû à la présence d'une valeur propre isolée $1/N^4$ proche de 0.

Pour comparer avec des travaux déjà réalisés concernant par exemple la discrétisation du Laplacien bidimensionnel, on cite par exemple [Sa06, Figure 2.4 p. 49] qui est présenté dans un langage de théorie des systèmes mais concerne un exemple numérique quasi-identique.

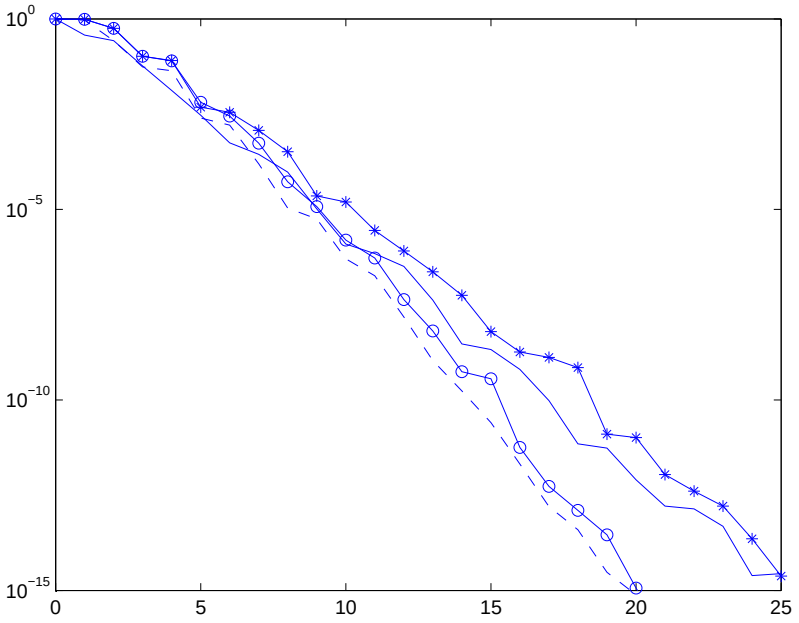

FIG. 8.4 – *Valeurs propres équidistantes, matrice* 100×100.

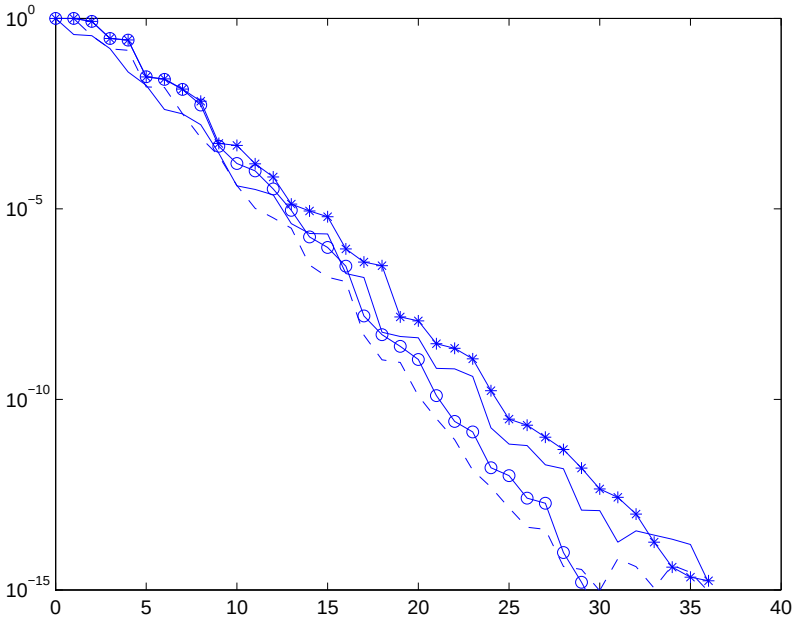

FIG. 8.5 – *Valeurs propres équidistantes, matrice* 1000×1000.

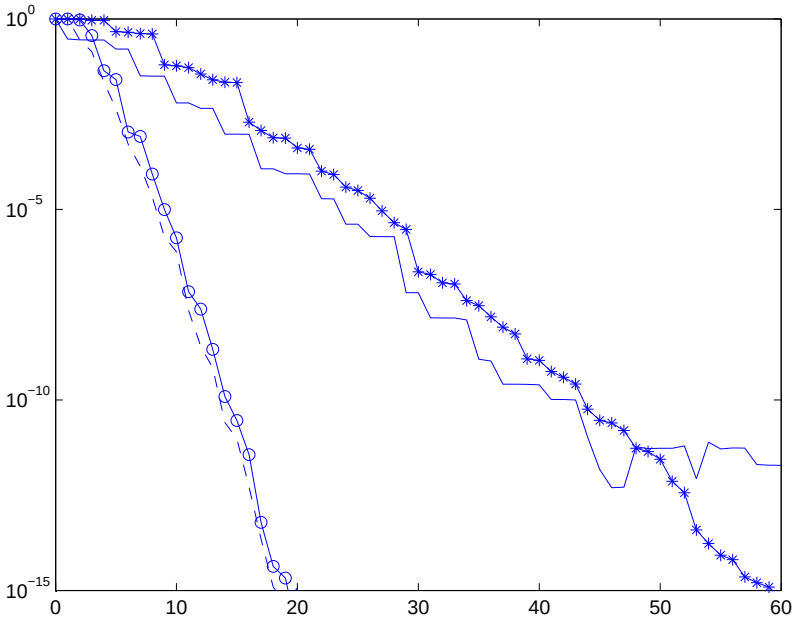

FIG. 8.6 – *Distribution en cosinus perturbée, matrice* 100×100.

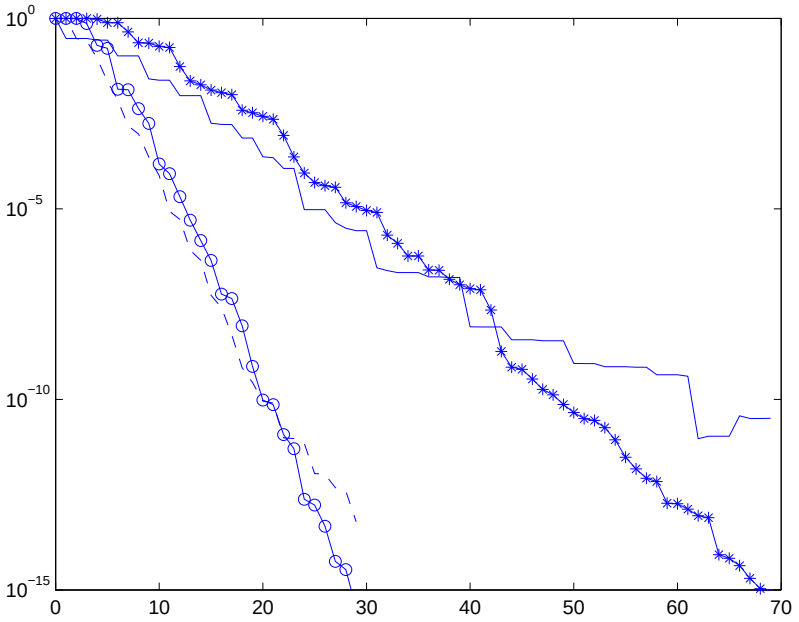

FIG. 8.7 – *Distribution en cosinus perturbée, matrice* 1000×1000.

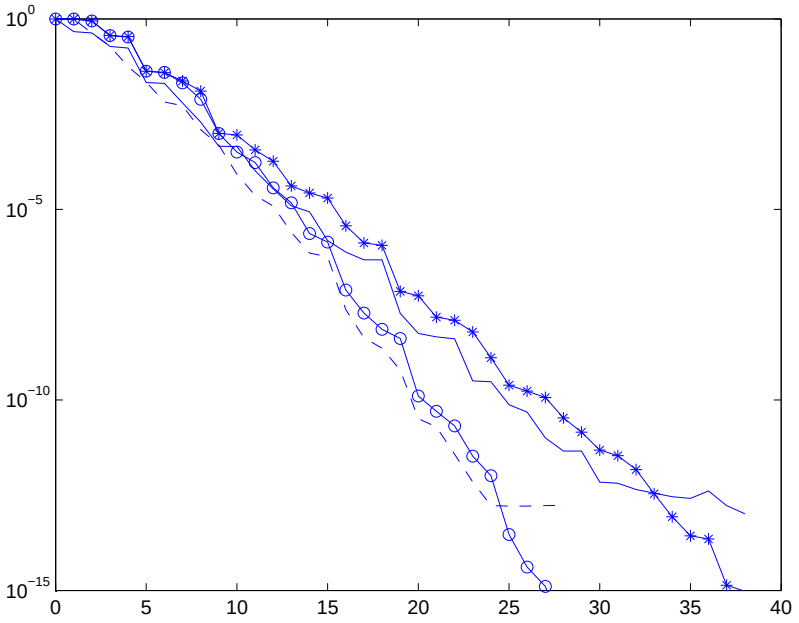

FIG. 8.8 – *Discrétisation du Laplacien* $2D$*, matrice* 100×100.

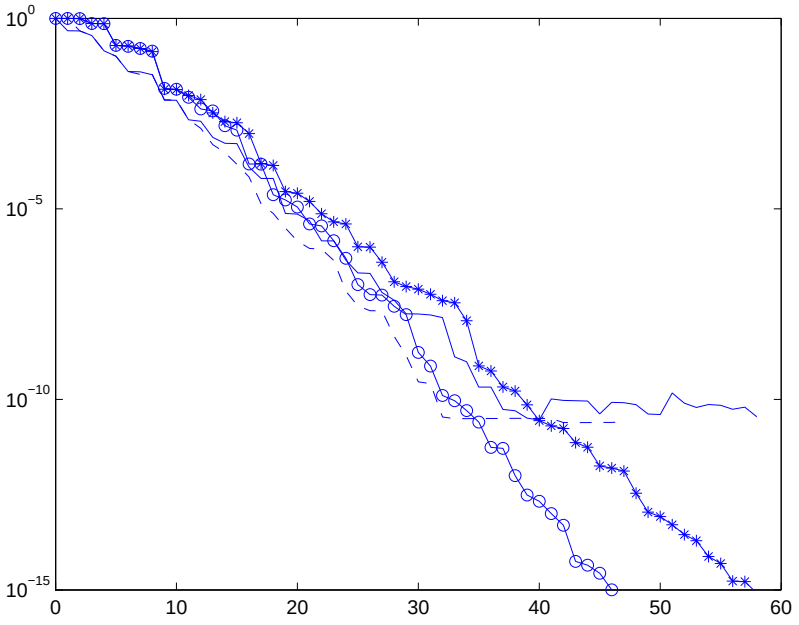

FIG. 8.9 – *Discrétisation du Laplacien* $2D$*, matrice* 1000×1000.

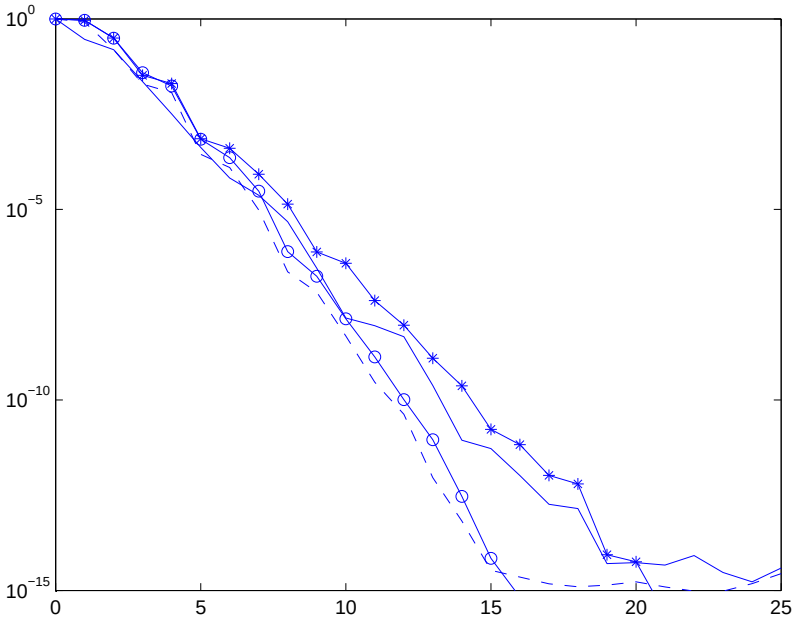

FIG. 8.10 – *Discrétisation du Laplacien* $4D$*, matrice* 100×100.

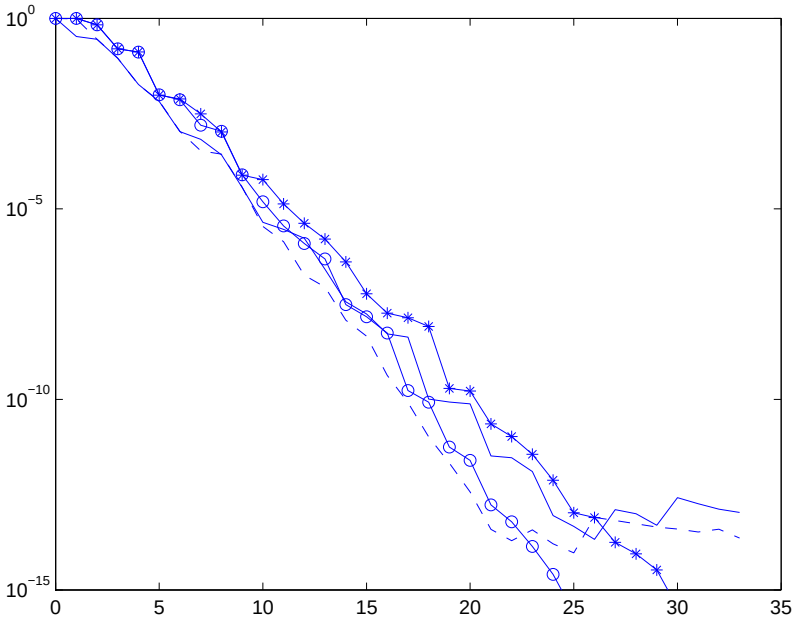

FIG. 8.11 – *Discrétisation du Laplacien* $4D$*, matrice* 1000×1000.

Pour une matrice 100×100, selon l'illustration présentée par Sabino, pour obtenir une erreur relative de 10^{-15}, les bornes supérieures données respectivement par les travaux de Penzl [Pe00b] et l'amélioration apportée par le travail de Sabino nécessitent 70 et 40 itérations, alors que l'erreur commise atteint ce seuil au bout de 25 itérations environ.

Si on compare avec la figure 8.8, on constate la borne supérieure pour l'erreur commise donnée par notre travail prédit que cette erreur relative sera atteinte au bout de 27 itérations environ.

Pour le cas du Laplacien $4D$, nous avons calculé les paramètres de la même façon que précédemment, à savoir en isolant une partie discrète proche de l'origine et en remplaçant la partie droite du spectre par son enveloppe convexe. Notre choix de paramètres n'et alors probablement pas optimal pour des raisons que nous exposerons dans ce qui suit, mais le simple fait de tenir compte de la nature discrète d'une partie du spectre proche de l'origine suffit à améliorer de façon nette le taux de convergence de la méthode ADI d'après les figures 8.10 et 8.11.

Ceci prouve que la prise en compte du caractère discret des spectres des matrices considérés améliore significativement dans certains cas l'estimation de la borne supérieure pour l'erreur commise.

8.5 Points de Léja-Bagby

Il nous reste alors à expliquer le choix des paramètres, qui reste pour le moment en partie heuristique.

Il suit d'après notre travail préalable que les points de Fekete rationnels pour des ensembles discrets sont d'un grand interêt théorique, mais ceux-ci sont difficiles voir impossible à déterminer en pratique.

Voici la définition des points rationnels de Léja-Bagby pour des ensembles discrets.

On compare cette définition avec celle extraite de [SaTo97, Définition précédant le Théorème VIII.3.5].

Définition 8.5.1. *Pour un ensemble E du demi-plan droit, les* points de Léja-Bagby rationnels pour des ensembles discrets $p_j = -q_j$ *pour le couple d'ensembles $(E, -E)$ de poids w donné sont déterminés par la procédure récursive*

$$q_{m+1} = \arg\min_{z \in E} |w(z) r_m(z)|, \quad r_m(z) = \prod_{j=1}^{m} \frac{z - q_j}{z + q_j}, \quad r_0(z) = 1. \tag{8.10}$$

On compare avec la procédure suivante donnée dans [SaTo97, Chapitre VIII] pour le cas d'ensembles continus que l'on écrit dans le cas d'ensembles symétriques :

$$q_{m+1} = \arg\min_{z \in E} |w^m(z) r_m(z)|, \quad r_m(z) = \prod_{j=1}^{m} \frac{z - q_j}{z + q_j}, \quad r_0(z) = 1. \tag{8.11}$$

On constate donc que le poids joue un rôle différent dans les procédures 8.10 et 8.11, ce qui jouera un rôle important dans ce qui suit.

Pour les paramètres adaptés à E l'enveloppe convexe de E_N, on choisit le poids $w(z) = 1$, et le choix des paramètres q_m pour $m = 1, 2, ..., [tN]$, où le paramètre $t \in (0, 1)$ a été choisi auparavant.

L'utilisation des points de Léja-Bagby dans le contexte de la méthode ADI a été suggérée entre autres dans [LeRe93].

Bien entendu, pour les calculs pratiques, on ne choisit pas l'intervalle réel $E = \mathrm{conv}(E_N)$, mais on prend comme ensemble E une discrétisation très fine de celui-ci, par exemple les points de Chebyshev décalés d'ordre $10N$ pour cet intervalle, points qui devraient donner sensiblement la même asymptotique.

On considère dorénavant une contrainte symétrique σ dont la partie positive est supportée sur $[0, B]$, vérifiant toutes les hypothèses de régularité évoquées dans le chapitre 5, 5.1, jusque 5.5, et avec les notations précédentes, on suppose que $supp\ (\sigma_1 - \mu_1^t) = [a_t, B]$.

On considère ici les points définis ci-dessus en 8.10 pour $E = [a_t, B]$ avec le poids

$$w(z) := \exp\left(-U^{\sigma|_{[-a_t,a_t]}}\right) \tag{8.12}$$

où σ est la mesure contrainte associée à notre problème.

Le lemme suivant montre qu'asymptotiquement, on ne perd rien à remplacer la contrainte σ par $+\infty$ sur la partie libre de la contrainte.

Lemme 8.5.2. *On définit la mesure $\tilde{\sigma}$ dont la restriction à $[-a_t, a_t]$ est égale à σ, et qui prend la valeur $+\infty$ sur $(a_t, B]$ et $-\infty$ sur $[-B, -a_t)$ au sens défini dans le chapitre 4.*

On note μ^t et $\tilde{\mu}^t$ les solutions respectives aux problèmes de minimisation d'énergie

$$\textit{Trouver } \mu^t \textit{ telle que} \qquad I(\mu^t) = \inf\left\{I(\mu),\ \mu \in \mathcal{M}_\sigma^t\right\}$$

et

$$\textit{Trouver } \tilde{\mu}^t \textit{ telle que} \qquad I(\tilde{\mu}^t) = \inf\left\{I(\tilde{\mu}),\ \tilde{\mu} \in \mathcal{M}_{\tilde{\sigma}}^t,\ \mathrm{supp}\,(\tilde{\mu}) \subset [-B, B]\right\},$$

où les ensembles de mesures candidates $\mathcal{M}_\sigma^t$ et $\mathcal{M}_{\tilde{\sigma}}^t$ ont été définis en 4.1.14.

Alors, $\mu^t = \tilde{\mu}^t$.

PREUVE : Il suffit de comparer les conditions d'équilibre vérifiées par μ^t et $\tilde{\mu}^t$ énoncées théorème 4.3.1 pour conclure, ces conditions permettent en effet d'après ce même théorème de caractériser les mesures extrémales. □

Voici enfin un lemme dont la démonstration repose sur l'identité des conditions d'équilibre vérifiées par des mesures extrémales.

Lemme 8.5.3. *Notons $s := t - \sigma([0, a_t])$.*

On définit le champ extérieur

$$Q := U^{\sigma|_{[-a_t,a_t]}}.$$

Alors, la solution au problème de minimisation d'énergie suivant, sans contrainte, avec champ extérieur Q sur l'ensemble

$$\{\mu := \mu_1 - \mu_2,\ \mu_j \textit{ mesure},\ \mu_j(\mathbb{C}) = s,\ \mathrm{supp}\,(\mu_1) \subset [a_t, B],\ \mathrm{supp}\,(\mu_2) \subset [-B, -a_t]\}$$

est égale à $\mu^t|_{[-B,-a_t]\cup[a_t,B]}$.

Le théorème [SaTo97, Théorème VIII.3.5] donne alors après renormalisation un résultat concernant l'optimalité asymptotique des points de Léja-Bagby selon la procédure décrite ci-dessus en 8.11.

En effet, si l'on pose avec les notations du lemme 8.5.3 $Q^\star := \frac{1}{s}Q$ et $\mu^\star = \mu_1^\star - \mu_2^\star$ la solution au problème de minimisation d'énergie, sans contrainte, avec champ extérieur $Q^\star$ sur l'ensemble

$$\{\mu := \mu_1 - \mu_2,\ \mu_j \text{ mesure},\ \mu_j(\mathbb{C}) = 1,\ supp\,(\mu_1) \subset [a_t, B],\ supp\,(\mu_2) \subset [-B, -a_t]\},$$

on a d'après les conditions d'équilibre vérifiées par chacune de ces mesures

$$\mu^\star = \frac{1}{s}\mu^t|_{[-B,-a_t]\cup[a_t,B]}.$$

Nous pouvons alors appliquer la procédure 8.11 avec le poids $\exp(-Q^\star)$ pour calculer les points de Léja-Bagby rationnels d'ordre n notés $(q_\ell)_{0\leqslant\ell\leqslant n} \subset [a_t, B]$.

On remarque que la condition imposée dans l'énoncé de [SaTo97, Théorème VIII.3.5] sur les ensembles E et F égaux dans notre contexte à $[a_t, B]$ et $[-B, -a_t]$ n'est pas nécessaire grâce à la remarque [SaTo97, Remarque VIII.3.6] et aux conditions d'équilibre vérifiées par $\mu^\star$.

D'après [SaTo97, Théorème VIII.3.5], la mesure de comptage normalisée de cette famille d'ensembles converge faiblement vers $\mu_1^\star$. Cependant, cela ne permet pas d'établir un résultat asymptotique concernant les points définis en 8.10 du fait que le poids joue un rôle différent dans les procédures 8.10 et 8.11.

On conjecture néanmoins un résultat asymptotique similaire au sujet des points définis en 8.10, mais nous ne sommes pas en mesure de prouver ce résultat pour le moment.

Cela dit, d'après les lemmes 8.5.2 et 8.5.3, la théorie du potentiel nous apprend donc que pour une approche judicieuse des problèmes numériques discrets, on ne perd rien -au moins asymptotiquement- à remplacer une partie de l'ensemble discret par son enveloppe convexe, alors qu'il est souhaitable de tenir compte de la nature discrète des ensembles considérés là où la contrainte est atteinte. C'est enfin la solution au problème de minimisation d'énergie sous contrainte, et plus précisément les équations intégrales obtenues dans le chapitre 7 qui permettent de choisir la partie du spectre que l'on peut raisonnablement remplacer par son enveloppe convexe.

Grâce à la théorie développée dans ce travail, nous savons ainsi que les points de l'ensemble discret E_N en dehors de $supp\,(\sigma_1 - \mu_1^t)$ jouent un rôle déterminant dans l'asymptotique du problème, on retrouve ici des considérations étudiées dans [BeKu01b] pour le cas de l'algorithme du gradient conjugué.

Ainsi, pour le cas du Laplacien $4D$, le fait d'isoler des points proches de l'origine n'est probablement pas le choix optimal pour cet exemple pour lequel on conjecture que $supp\ (\sigma_1 - \mu_1^t)$ est un intervalle centré en 4.

Bien sûr, le poids défini en (8.12) ne correspond pas à la réalité numérique, où l'on utilise un poids construit à partir d'un ensemble de points discrets qui est par conséquent une fraction rationnelle.

Pour la procédure concrète, pour imiter de façon précise la structure de E_N, on choisit $a \in (A, B)$ suffisamment grand (dans le but d'affectuer $[tN]$ itérations avec $a = a_t$), et on calcule ensuite toutes les valeurs propres de la matrice A_N situées dans $[0, a]$, que l'on

note $q_1, ..., q_M$. On remarque que l'entier M dépend implicitement de N, et que l'on a à la limite la relation

$$\lim_{N\to+\infty} \frac{M}{N} = \frac{\sigma([0, a_t])}{\sigma([0, B])}.$$

Il nous reste alors à déterminer $[tN] - M$ paramètres supplémentaires $q_{M+1}, ..., q_{[tN]}$ en appliquant la procédure usuelle pour déterminer les points de Léja-Bagby rationnels décrite ci-dessus sur $E = [a, B]$ à partir de l'indice $M + 1$, ou en d'autres termes avec le poids $w = r_M$.

En pratique, on prend à nouveau pour ensemble E un ensemble de points de Tchebychev décalés d'ordre élevé, par exemple, $10N$ dans $[a, B]$.

Cependant, les paramètres obtenus sont rangés dans un ordre ne permettant pas d'obtenir une erreur petite pour toutes les itérations de la méthode ADI pour $n = 1, 2, ..., [tN]$. On applique alors la procédure de Léja-Bagby une nouvelle fois avec le poids $w = 1$ sur l'ensemble $E = \{q_1, ..., q_{[tN]}\}$ de tous les paramètres obtenus, et ces paramètres permutés semblent donner une erreur petite pour toutes les itérations de la méthode.

Nous ne disposons pas pour le moment de justification théorique de cette méthode dont nous sommes réduits à constater l'efficacité numérique. L'intuition tend à indiquer que le choix de paramètres ordonnés de façon monotone ne soit pas optimal, ce qui nous a poussé à effectuer une deuxième fois cet algorithme pour réordonner les points.

Le fait d'isoler quelques valeurs propres proches de zéro pour améliorer le taux de convergence de la méthode est une technique souvent utilisée en algèbre linéaire numérique, on peut finalement voir notre description du choix des paramètres comme une justification théorique de cette pratique numérique de bon sens. Nos résultats en théorie du potentiel logarithmique permettent enfin de déterminer le seuil à partir duquel la nature discrète ou continue des ensembles considérés ne modifie plus de façon sensible l'asymptotique du problème.

Conclusion

Nous avons présenté dans le chapitre 1 le lien entre la quantité de Zolotarev pour des ensembles discrets et certains problèmes d'algèbre linéaire ou d'approximation, mais d'autres pistes dans cette direction semblent s'offrir à nous, en particulier pour l'estimations des valeurs de Ritz rationnelles d'une matrice hermitienne de grande dimension, ou de certaines fonctions matricielles.

Les valeurs de Ritz d'ordre n sont données par les valeurs propres de la matrice $V_n^\star A V_n$ où les colonnes de $V_n \in \mathbb{C}^{n\times N}$ forment un système orthonormal, ce qui revient à projeter A sur l'espace engendré par les colonnes de V_n.

On définit un espace de Krylov rationnel par

$$\mathcal{K}_n^{rat}(A,b) := q_{n-1}(A)^{-1}\,\mathrm{vect}\{b, Ab, \ldots, A^{n-1}b\},\ q_{n-1}(z) := \prod_{\ell=1}^{n-1}(z-\chi_\ell),$$

et les éléments $(\chi_\ell)_{\ell=0}^{n-1} \in \mathbb{R}\setminus\Lambda(A)$ sont les pôles de l'espace de Krylov rationnel considéré.

Dans le cas d'un espace de Krylov classique

$$\mathcal{K}_n(A,b) := \mathrm{vect}\{b, Ab, \ldots, A^{n-1}b\},$$

les valeurs de Ritz correspondantes sont appelées *valeurs de Ritz polynômiales*, et permettent typiquement d'approcher les valeurs extrémales du spectre de A. Dans les applications, n est petit devant N, et le choix des pôles de l'espace de Krylov rationnel considéré permet alors d'amplifier certaines parties du spectre de A.

Dans [BeGuVa09], on quantifie la convergence des valeurs de Ritz rationnelles vers les valeurs propres de A en utilisant des techniques de théorie du potentiel logarithmique semblables à celles développées dans le chapitre 4.

Dans [DrKnZa08], on cherche en lien avec des questions de détection d'hydrocarbures dans les profondeurs à estimer la fonction

$$u(t) := \mathrm{e}^{-tA}\,\phi,\ \text{où } A \in \mathbb{R}^{N\times N},\ \phi \in \mathbb{R}^N,\ ||\phi|| = 1$$

où A est également hermitienne de grande dimension, ce qui fait intervenir un problème de Zolotarev sur le spectre de A et sur certaines valeurs de Ritz rationnelles de A.

Les liens entre nos travaux concernant le problème de Zolotarev pour des ensembles discrets, les techniques de théorie du potentiel logarithmique développées dans ce but et les deux travaux cités ci-dessus semblent prometteurs sur un plan à la fois théorique et numérique.

On semble en outre pouvoir obtenir un lien entre le problème de Zolotarev sur des ensembles discrets et la minimisation des produits de Blaschke sur des ensembles discrets

dans le cas particulier d'ensembles vérifiant $F = 1/E$ dans le but d'étendre les résultats de [FiSa99] aux ensembles discrets.

Concernant les applications de nos travaux à l'étude de la méthode ADI, l'étude effectuée ici gagnerait à être étendue à l'étude asymptotique du problème de Zolotarev généralisé où l'ensemble des fractions rationnelles candidates est donné par $\mathcal{R}_{n,m}$. Cela permettrait sans doute d'améliorer les résultats de [LeRe93] pour l'étude du taux de convergence de la méthode ADI généralisée.

Dans le chapitre 6, la condition suffisante pour que la partie libre de la contrainte soit donnée par l'union de deux intervalles n'a été formulée que dans le cas d'une contrainte symétrique, et l'établissement d'autres conditions plus générales de ce type nous semble constituer un progrès théorique intéressant.

Les investigations numériques que nous avons menées soulèvent également quelques interrogations : tout notre travail est construit sous l'hypothèse que le ratio

$$\frac{\text{nombre d'itérations effectuées}}{\text{taille des matrices considérées}} = \frac{n}{N}$$

admette une limite finie strictement positive, ce qu'on pourrait appeler un choix d'échelle.

On peut alors se demander si ce choix -qui semble le plus naturel- est le plus judicieux, et ce qu'il adviendrait des expériences numériques du chapitre 8 pour un autre choix d'échelle, par exemple en supposant que le ratio $\frac{n^2}{N}$ admette une limite finie, ce qui pourrait sembler plus approprié à des objets de plus grande taille.

Sur le plan théorique, nous n'avons pu pour le moment qu'émettre des conjectures concernant l'optimalité asymptotique des points de Léja-Bagby rationnels discrets dans notre cas, la preuve de ce résultat apporterait la confirmation de l'intuition développée au vu de nos expériences numériques. Il serait de même intéressant de comprendre pourquoi le fait de réordonner les points une seconde fois selon l'ordre de Léja-Bagby améliore la convergence. Ces points sont utilisés en pratique pour les expérience numériques, et tout résultat théorique à leur sujet permettrait de progresser vers la compréhension plus profonde du choix des paramètres adaptés à la méthode ADI sur des ensembles discrets.

Enfin, nos résultats asymptotiques ne sont valables qu'au sens de la racine N-ème, et les résultats numériques du chapitre 8 nous incitent très largement à conjecturer l'existence de ces résultats dans un cadre plus fort. Des recherches dans cette direction semblent pertinentes, mais nécessitent probablement des outils théoriques plus fins que ceux que nous avons utilisés au sein de cette thèse.

Table des figures

Index

Bibliographie

[AbSt64] Abramowitz M. and Stegun I., *Handbook of mathematical functions, National Bureau of Standards Applied Mathematics Series* **55**, U.S. Government Printing Office, Washington, D.C. (1964).

[Ac56] N.I. Achieser, *Theory of approximation, F. Ungar Publishing Co* (1956).

[Ak90] N.I. Akhieser, *Elements of the Theory of Elliptic Functions, Transl. of Math. Monographs* **79**, AMS, Providence RI (1990).

[An94] J.E. Andersson, *Best Rational Approximation to Markov Functions, J. of Approx Theory* **76** (1994), 219–232.

[An98] A. C. Antoulas, *Approximation of linear dynamical systems, Wiley Encyclopedia of Electrical and Electronics Engineering*, ed. J.G. Webster, vol. **11**, 403-422 (1999).

[An05] A. C. Antoulas, *Approximation of large-scale dynamical systems, SIAM* **76** Philadelphia (2005).

[AnIoRo08] A. C. Antoulas, R. Ionitiu and J. Rommes *Passivity-preserving model reduction using dominant spectral zero interpolation, IEEE transactions of computer aided design of integrated circuits and systems* vol. **27**, no. **12** (2008)

[AnSo02] A. C. Antoulas and D.C. Sorensen, *The Sylvester equation and approximate balanced reduction, Linear Algebra Appl* **351/352** (2002), 671–700.

[BaTh00] Le Bailly, B. Thiran, *Optimal rational functions for the generalized Zolotarev problem in the complex plane, SIAM J. Numer. Anal. 38* , no. **5**, 1409–1424 (2000).

[Ba87] L. Baratchart, *Sur l'approximation rationnelle L^2 pour les systèmes dynamiques inéaires, Thèse présentée à l'Université de Nice* (1987).

[BaSt72] R. H. Bartels, G. W. Stewart, *Solution of the matrix equation $AX + XB = C$, Communications of the ACM*, no. **9**, 820–826 (1972).

[Be00] B. Beckermann, *On a conjecture of E.A. Rakhmanov, Constrx.approx* **16** (2000) 427–448.

[Be04] B. Beckermann, *Singular values of small displacement rank matrices*, Talk at conference *Structured Numerical Linear Algebra Problems : Algorithms and Applications*, Cortona, 2004.

[Be06] B. Beckermann, *Discrete orthogonal polynomials and superlinear convergence of Krylov subspace methods in numerical linear algebra in Orthogonal Polynomial and Special Functions*, F. Marcellan, W. Van Assche (Eds.), *Lecture Notes in Mathematics* **1883**, Springer Verlag (2006), 119–185.

[BeGuVa09] B. Beckermann, S. Güttel and R. Vanderbril, *On the convergence of rational Ritz values*, *preprint* (2009).

[BeKu01a] B. Beckermann and A.B.J. Kuijlaars, *Superlinear convergence of conjugate gradients*, *SINUM* **39** (2001) 300–329.

[BeKu01b] B. Beckermann and A.B.J. Kuijlaars, *On the sharpness of an asymptotic error estimate for Conjugate Gradients*, *BIT* **41** (2001), 856–867.

[BeKu02] B. Beckermann and A.B.J. Kuijlaars, *Superlinear CG convergence for special right-hand sides*, *Electr. Trans. Num. Anal.* **14** (2002) 1-19.

[BeSC07] B. Beckermann and S. Serra-Capizzano, *On the asymptotic spectrum of finite element matrix sequences*, *SIAM J. Numer. Anal.* **45** (2007), 746–769.

[BeLiTr08] P. Benner, R-C Li and N. Truhar, *On the ADI method for Sylvester equations*, preprint (2008).

[BeQu02] P. Benner, G. and E. Quintana-Orti, *Numerical solution of discrete stable linear matrix equations on multicomputers*, *Parallel Alg. Appl.* **17** (2006), 127-146.

[BeQu06] P. Benner, G. and E. Quintana-Orti, *Solving stable Sylvester equations via rational iterative schemes*, *J. Sci. Comput.* **1** (2006), 51-83.

[BiVa59] G. Birkhoff and R. S. Varga, *Implicit alternating direction methods*, *Trans. Amer. Math. Soc.* **92** (1959) 13-24.

[BöSi97] A. Böttcher, B. Silbermann, *Introduction to large truncated Toeplitz matrices*, *Springer Universitext* (1997).

[Br80] D. Braess, *Nonlinear Approximation Theory*, *Springer series in computational mathematics* **7** (1980).

[Br87] D. Braess, *Rational Approximation of Stieltjes Functions by the Carathéodory-Fejer Method*, *Constrx.approx.* **3** (1987), 43–50.

[BuRa99] V. Buyarov and E.A. Rakhmanov *Families of equilibrium measures with external field on the real axis*, *Sb.Maths* **190** (1999) 791–802.

[CaLeRe97] D. Calvetti, N. Levenberg, L. Reichel *Iterative methods for* $X - AXB = C$, *J. of Comp. and Appl. Maths* **86** (1997) 73–101.

[DrSa97] P.D. Dragnev and E.B. Saff, *Constrained energy problems with applications to orthogonal polynomials of a discrete variable*, *J. d'Analyse Math.* **72** (1997), 223–259.

[DrToTr98] T.A. Driscoll, K.-C. Toh and L.N. Trefethen, *From potential theory to matrix iteration in six steps*, *SIAM* Rev. **40** (1998), 547–578.

[DrKnZa08] V. Druskin, L. Knizhnerman and M. Zaslavsky, *Solution of large scale evolutionary problems using rational Krylov subspaces with optimized shifts*, preprint, (2008).

[EFLSV02] J. van den Eshof, A. Frommer, Th. Lippert, K. Schilling, H.A. van der Vorst, *Numerical Methods for the QCD Overlap Operator : I. Sign-Function and Error Bounds*, *Comp. Physics* Comm. **146** (2002), 203–224.

[FiSa99] S.D. Fisher and E.B. Saff *The Asymptotic Distribution of Minimal Blaschke Products*, *J. of Approx Theory* **98** (1999), 104–116.

[GoVLo96] Golub G. H., Van Loan F., *Matrix computations. Third edition.*, *Johns Hopkins Studies in the Mathematical Sciences.*, Johns Hopkins University Press, Baltimore, MD, (1996).

[Go78] A.A. Gonchar *On the speed of rational approximation of some analytic functions*, *Math.USSR-Sb.*, **34** (1978) 131–145.

[GoNaVL79] Golub, G. H., Nash, S., Van Loan, F., *A Hessenberg-Schur method for the problem AX + XB = C*, *IEEE Trans. Auto. Control.*, **24** 900–913 (1979).

[GrRy00] Gradshteyn, I. S. and Ryzhik, I. M., *Table of integrals, series, and products. Sixth edition.*, *Academic Press, Inc.*, San Diego, CA, (2000)

[GrSz58] U. Grenander, G Szegö, *Toeplitz forms and their applications*, *Univ. of California Press Berkeley* (1958)

[GuAn04] S. Gugercin and A. C. Antoulas, *A survey of model reduction by balanced truncation and some new results*, *International Journal of Control*, Volume : **77** Issue : **8**, pp. 748-766, (2004).

[Ham82] S. J. Hammarling *Numerical solution of the stable, non-negative, definite Lyapunov equation IMA J. Num. Anal.* **2** 303–323 (1982).

[Han04] H. Hancock, *Lectures on the theory of elliptic functions Dover Publications*, 2004.

[Hi08] H.J. Higham, *Functions of Matrices : Theory and Computation*, *SIAM*, 2008.

[IsTh95] Istace, M.-P., Thiran, J.-P. *On the third and fourth Zolotarev problems in the complex plane*, *SIAM J. Numer. Anal.* **32** (1995), 249–259.

[Ku00] A.B.J. Kuijlaars, *Which eigenvalues are found by the Lanczos method ? SIAM J. Matrix Anal. Appl.* **22** (2000), 306–321.

[Ku06] A.B.J. Kuijlaars, *Convergence analysis of Krylov subspace iterations with methods from potential theory*, *SIAM* Review **48** (2006), 3–40.

[KuDr99] A.B.J. Kuijlaars and P.D. Dragnev, *Equilibrium problems associated with fast decreasing polynomials*, *Proc. Amer. Math. Soc.* **127** (1999), 1065–1074.

[La06] M. A. Lapik *Support of the extremal measure in a vector equilibrium problem*, *Sb.Maths* **197** (2006), 1205–1219.

[LeLu01] A.L. Levin and D.S. Lubinsky, *Green equilibrium measures and representations of an external field*, *J. of Approx theory* **113** (2001), 298–323.

[LeRe93] N.Levenberg and L. Reichel, *A generalized ADI iterative method*, *Numer. Math.* **66** (1993) 215–233.

[LeSa01] A.L. Levin and E.B. Saff, *The Distribution of Zeros and Poles of Asymptotically Extremal Rational Functions for Zolotarev's Problem*, *J. of Approx theory* **110** (2001), 88–108.

[LeSa06] A.L. Levin and E.B. Saff, *Potential Theoretic Tools in Polynomial and Rational Approximation in Harmonic Analysis and Rational Approximation*, Vol. **327** (Fournier, Grimm, Leblond, Partington, Eds.), Springer, (2006), 71–94.

[MhSa85] H.N. Mhaskar and E.B. Saff *Where does the sup norm of a weighted polynomial live ? (A generalization of incomplete polynomials).*, *Constrx.approx.* **1** (1985), 71–91.

[Mo81] B.C. Moore. *Principal component analysis in linear systems : Controllability, observability and model reduction.*, *IEEE Trans. Auto. Control* **26** (1981), 17–32.

[Ni01a] N. K. Nikolski *Operators, functions and systems : an easy reading, Volume 1 : Hardy, Hankel and Toeplitz.*, *Mathematical surveys and monographs, AMS* **93** (2001).

[Ni01b] N. K. Nikolski *Operators, functions and systems : an easy reading, Volume 2 : model operators and systems.*, *Mathematical surveys and monographs, AMS* **93** (2001).

[NiSo91] E.M. Nikishin and V.N. Sorokin *Rational Approximations and Orthogonality*, *AMS, Translations of Mathematical monographs*, vol. **92** (1991).

[PeRa55] D. Peaceman, H. Rachford *The numerical solution of elliptic and parabolic differential equations*, *J. Soc. Indust. Apll. Math.*, vol. **3** 28–41 (1955).

[Pe00a] T. Penzl, *Eigenvalue decay bound for solutions of Lyapounov equations : the symmetric case*, *Systems and Control Letters* **40** (2000), 139–144.

[Pe00b] T. Penzl, *A cyclic low-rank Smith method for large sparse Lyapunov equations*, *SIAM J. Sci. Comp.* **21** 1401–1418 (2000), 139–144.

[PoWi98] J.W. Polderman and J.C. Willems, *Introduction to Mathematical Systems Theory : A Behavioral Approach*, *Springer Verlag* (1998).

[Po81] M. J. D. Powell, *Approximation theory and methods*, *Cambridge University Press* (1981).

[PrBrMa88] Prudnikov, A. P. Brychkov, Yu. A. Marichev, *Integrals and series. Vol. 3. More Special functions.*, *Second edition. Gordon and Breach Science Publishers, New York,* (1988).

[Ra96] E.A. Rakhmanov, *Equilibrium measure and the distribution of zeros of the extremals polynomials of a discrete variable*, *Math.Sbornik* **187** (1996), 1213–1228.

[Ra95] T. Ransford, *Potential theory in the complex plane*, *Cambridge university press*, London math. soc. student text **28** (1995).

[Ru75] W. Rudin, *Analyse réelle et complexe*, éditions Masson, 1975.

[Sa06] J. Sabino, *Solution of large-scale Lyapunov equations via the block modified Smith method*, *Phd thesis, Rice university* (2006).

[SaTo97] E.B. Saff and V. Totik, *Logarithmic potentials with external fields*, *A Series of comprehensive studies in mathematics*, Springer vol. **316**, (1997).

[Si67] R. A. Silverman, *Introductory complex analysis*, Dover publications, (1972). (1997).

[Sm68] R. A. Smith, *Matrix equation* $XA + BX = C^{\star}$, *SIAM J. Appl. Math.* **16** 198–201 (1968).

[Tr90] L.N. Trefethen, *Approximation theory and numerical linear algebra*, in : Algorithms for Approximation II, J.C. Mason and M.G. Cox, eds, Chapman & Hall, London, 1990.

[TrBa97] L.N. Trefethen, D. Bau III, *Numerical Linear Algebra*, SIAM, Philadelphia PA, 1997.

[Wa63] E. L. Wachspress, *Extended application of alternating direction implicit iteration model problem theory*, *J. Soc. Indust. Appl. Math.* **11** 994-1016 (1963)

[Wa69] E. L. Wachspress, *Solution of the generalized ADI minmax problem*, *Information Processing 68 (Proc. IFIP Congress Edinburgh 1968)* Vol. **1** : Mathematics Software 99-105 North-Holland Amsterdam (1969)

[Zo32] E.I. Zolotarev, *Collected work* **2**, USSR Acad. Sci, 1932 (in Russian).