UNIVERSITÉ DE LIMOGES

ÉCOLE DOCTORALE Sciences et Ingénierie pour l'Information
FACULTÉ des SCIENCES et TECHNIQUES
Département de Mathématiques et Informatique
Laboratoire XLIM (UMR 6172)

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE LIMOGES

Discipline : Mathématiques et ses applications

présentée et soutenue publiquement par

Elsa BOUSQUET

le 18 Décembre 2009

Optimisation non linéaire et application au réglage d'un réseau de télescopes

Thèse codirigée par Paul ARMAND et Joël BENOIST

JURY

Rapporteurs :

P. MARÉCHAL	Professeur, Université Paul Sabatier de Toulouse
D. ORBAN	Professeur, École Polytechnique de Montréal, Canada

Examinateurs :

P. ARMAND	Maître de Conférences-HDR, Université de Limoges
J. BENOIST	Maître de Conférences-HDR, Université de Limoges
M. HADDOU	Maître de Conférences, Université d'Orléans
M. THÉRA	Professeur, Université de Limoges

Invité :

F. REYNAUD	Professeur, Université de Limoges

Remerciements

Je tiens tout d'abord à remercier *Paul Armand* et *Joël Benoist* sans qui ce travail n'aurait pas vu le jour. Je les remercie d'avoir dirigé mes recherches ainsi que de leurs conseils durant ces années de thèse.

Je suis très honorée que *Dominique Orban* et *Pierre Maréchal* aient accepté de rapporter mon travail de thèse. Je les remercie pour leurs conseils et leurs précieuses remarques.

Je suis très reconnaissante envers *Michel Théra* pour ses encouragements pendant toutes ces années. Cela m'a fait très plaisir qu'il participe à mon jury de thèse.

Je remercie vivement *Mounir Haddou* de sa participation à ma soutenance de thèse.

Je remercie très chaleureusement *Laurent Delage* et *François Reynaud* pour m'avoir permis de travailler sur un sujet passionnant liant astronomie et optimisation. Travailler ensemble a été pour moi un réel plaisir.

Je remercie également *Serge Olivier* pour tous les renseignements complémentaires qu'il m'a apportés en astronomie et pour sa bonne humeur constante.

Je tiens aussi à remercier tous les enseignants-chercheurs du Département de Mathématiques et Informatique (DMI) de l'Université de Limoges. Tout d'abord, merci à *Anne Bellido* pour ses conseils lors de mes premières années à la Faculté des Sciences et Techniques de Limoges. Plus particulièrement, un grand merci à *Samir Adly* et *Jean-Marie Guillerme* pour m'avoir initiée dès la Licence à l'optimisation et pour leurs conseils. Également, je suis très reconnaissante envers mes anciens responsables de Master : *Jacques-Arthur Weil* et *Thierry Berger*. Enfin, merci à *Moulay Barkatou* pour son soutien et ses encouragements toutes ces années.

De même, je remercie les secrétaires du DMI : *Yolande Vieceli, Patricia Vareille* et *Nadia Rossi* pour leurs aides quand cela était nécessaire. Plus particulièrement, merci à *Yolande* d'avoir assuré la logistique de ma soutenance quand je n'étais pas sur Limoges. Également, merci à *Sylvie Laval* pour sa disponibilité et ses conseils.

Je tiens aussi à remercier *Henri Massias* pour son aide apportée en informatique et sa disponibilité pour répondre à mes questions.

Je remercie la Région Limousin pour le financement qu'elle m'a accordé pour faire cette thèse.

À présent, je remercie tous les doctorants, post-doctorants et ATER que j'ai rencontrés : *Samuel Maffre, Laurent Dubreuil, Nicolas Le Roux, Julien Angeli, Aurore*

Bernard, Sandrine Jean, Adrien Poteaux, Delphine Savary et *Romain Validire*. Je remercie également tous mes anciens et actuels collègues de bureau : *Guilhem Castagnos, Ahmed Aït Mokhtar, Sinaly Traore, Pierre-Louis Cayrel, Ainhoa Aparicio Monforte, Benjamin Pousse*. Je remercie également mes anciens camarades de Master : *Christophe Chabot* et *Daouda Niang Diatta*. Enfin, un petit clin d'œil à *Carole El Bacha* et *Hassan Saoud*. Je les remercie tous pour les bons moments que nous avons passés.

Une pensée aux personnes que j'ai croisées dans des conférences : *Aude Rondepierre, Laetitia Thevenet* et *Didier Auroux*. Également, merci à *Olivier Prot* pour ces longues discussions que nous avons eues.

Enfin, un immense merci à ma famille : ma mère pour m'avoir donné goût aux mathématiques et pour son soutien, mon père pour m'avoir poussée et motivée toutes ces années, ma grand-mère pour m'avoir permis de m'évader avec tous nos voyages et enfin mon frère *Amaury* pour sa jeunesse et sa détermination à qui je souhaite bonne chance pour sa thèse. Pour finir, un grand merci à *Vincent* pour son soutien et sa compréhension de tous les jours.

À ma famille

Table des matières

Remerciements i

Dédicace iii

Introduction générale 1

I Optimisation des paramètres d'un réseau de télescopes optiques 3

1 Problème physique et modélisation mathématique 5
- 1.1 Introduction . . . 5
- 1.2 Champ optique . . . 7
  - 1.2.1 Champ optique dans le plan pupille . . . 8
  - 1.2.2 Champ optique dans le plan image . . . 9
- 1.3 Réponse impulsionnelle . . . 9
  - 1.3.1 Réponse impulsionnelle pour un télescope . . . 9
  - 1.3.2 Réponse impulsionnelle pour n télescopes . . . 12
- 1.4 Réseau linéaire de télescopes . . . 14
- 1.5 Modélisation mathématique . . . 16
  - 1.5.1 Méthode des coefficients de Fourier . . . 16
  - 1.5.2 Méthode de l'optimisation de la dynamique . . . 18

2 Résultats numériques 21
- 2.1 Critères qualitatifs de la réponse impulsionnelle . . . 21
- 2.2 Résultats obtenus avec la méthode des coefficients de Fourier . . . 22
- 2.3 Résultats obtenus avec la méthode de l'optimisation de la dynamique 27
  - 2.3.1 Optimisation des amplitudes des champs . . . 27
  - 2.3.2 Optimisation des positions des pupilles . . . 30
  - 2.3.3 Conclusions sur les deux optimisations . . . 33
  - 2.3.4 Optimisation des amplitudes des champs et des positions des pupilles . . . 34
- 2.4 Étude de la sensibilité et de la robustesse de la configuration optimale 41
- 2.5 Conclusions . . . 43

3 Résultats théoriques **45**
3.1 Position du problème avec les normes $\|\cdot\|_2$ et $\|\cdot\|_\infty$ 45
3.1.1 Réponse impulsionnelle temporelle 46
3.1.2 Modèles pour l'optimisation de la dynamique 47
3.2 Conditions d'optimalité du problème écrit avec la norme $\|\cdot\|_2$ 48
3.2.1 Notations 48
3.2.2 Conditions d'optimalité pour des positions de pupilles fixées . 49
3.2.3 Conditions d'optimalité pour des amplitudes de champs fixées 51
3.2.4 Conditions d'optimalité pour des positions de pupilles et des amplitudes de champs non fixées 52
3.3 À propos de l'existence de la solution du problème écrit avec la norme $\|\cdot\|_2$ 53
3.3.1 Graphe de ϕ_1 55
3.3.2 Limites de ϕ_1 55
3.3.3 Dérivée de ϕ_1 58
3.4 À propos de l'existence de la solution du problème écrit avec la norme $\|\cdot\|_\infty$ 60
3.4.1 Rappels sur les polynômes de Tchebytchev 60
3.4.2 Notations 61
3.4.3 Énoncé du théorème 61
3.4.4 Preuve du théorème 61
3.4.5 Propriétés des amplitudes optimales des champs 70
3.4.6 Exemple illustratif 72

4 Conclusions et perspectives **73**

II Méthode primale-duale pour l'optimisation avec contraintes d'égalité **75**

1 Rappels sur les méthodes de pénalisation quadratique et SQP **77**
1.1 Introduction 77
1.2 Définitions, notations et hypothèses générales 77
1.3 Méthode de pénalisation quadratique 79
1.3.1 Algorithme local 80
1.3.2 Théorèmes de convergence 80
1.3.3 Mauvais conditionnement et reformulation 83
1.4 Méthode SQP 85
1.4.1 Description de la méthode 85
1.4.2 Algorithmes local et global 86
1.4.3 Théorèmes de convergence 90
1.4.4 Implémentation de la méthode 90

2 Présentation de la méthode primale-duale 93
2.1 Introduction . . . 93
2.2 Résultats préliminaires . . . 95
2.3 Principe général de l'algorithme . . . 98
2.4 Algorithme local . . . 100
2.4.1 Description de l'algorithme . . . 100
2.4.2 Théorème de convergence locale . . . 100
2.4.3 Exemple illustratif . . . 102
2.5 Algorithme global . . . 104
2.5.1 Description de l'algorithme . . . 104
2.5.2 Théorème de convergence globale . . . 107
2.5.3 Analyse asymptotique . . . 108

3 Résultats numériques 111
3.1 Liste des problèmes testés . . . 111
3.2 Valeurs des paramètres . . . 111
3.3 Régularisation de la jacobienne . . . 113
3.4 Décroissance du paramètre de pénalité . . . 114
3.5 Résultats . . . 115
3.6 Comparaison de la méthode primale-duale à la méthode SQP . . . 117

4 Conclusions et perspectives 119

A Coefficients d'apodisation 121

B Modèles pour l'optimisation de la dynamique 125
B.1 Optimisation des amplitudes des champs . . . 125
B.1.1 Norme $\|\cdot\|_2$ . . . 125
B.1.2 Norme $\|\cdot\|_\infty$ . . . 126
B.2 Optimisation des positions des pupilles . . . 127
B.2.1 Norme $\|\cdot\|_2$ . . . 127
B.2.2 Norme $\|\cdot\|_\infty$ . . . 128
B.3 Optimisation des amplitudes des champs et des positions des pupilles 128
B.3.1 Norme $\|\cdot\|_2$ . . . 128
B.3.2 Norme $\|\cdot\|_\infty$ . . . 129

C Optimization of a one dimensional hypertelescope for a direct imaging in astronomy 131
C.1 Introduction . . . 131
C.2 Densified pupil and point spread function . . . 133
C.3 Linear array of telescopes . . . 134
C.4 Optimization model . . . 134
C.5 Starting point strategy . . . 136
C.6 Numerical experiments . . . 138
C.7 Conclusion . . . 140

D Résultats numériques obtenus avec la méthode primale-duale 145

E Résultats numériques obtenus avec la méthode SQP 155

Table des figures

1.1 Télescope et réseau de télescopes. 6
1.2 Hypertélescope 6
1.3 Plan pupille et plan image de l'hypertélescope. 8
1.4 Figures d'Airy 10
1.5 PSF normalisée pour une pupille. 11
1.6 PSF normalisée pour quatre pupilles. 13
1.7 PSF normalisée pour huit pupilles alignées. 14
1.8 PSF normalisée pour huit pupilles alignées (échelle linéaire en bleu et échelle semi-logarithmique en vert). 15
1.9 PSF normalisée obtenue avec une fonction porte. 17

2.1 Critères qualitatifs de la PSF normalisée. 22
2.2 Représentation des fonctions d'apodisation. 24
2.3 Courbes bi-critères (D, ρ) obtenues avec la fonction porte pour $T = 1$ et $T = 2/3$. 25
2.4 PSF normalisée optimale obtenue avec la fonction porte pour $T = 1$ (a) et $T = 2/3$ (b) avec $D = 500$. 26
2.5 Courbes bi-critères (D, ρ) pour toutes les fonctions d'apodisation et PSF normalisée optimale obtenue avec la fonction de classe $\mathcal{C}^\infty$ pour $n = 9$ et $T = 1$. 26
2.6 Configuration optimale des quatre pupilles ($m = 4$) à droite de l'origine. La valeur en orange correspond à l'écart entre chaque pupille et la valeur en bleu correspond à l'amplitude des champs reçue par chacune des pupilles. Représentation de la PSF normalisée optimale (échelle linéaire en bleu et échelle semi-logarithmique en vert) dans cette configuration pour l'optimisation des amplitudes des champs. . . 28
2.7 Configuration optimale des quatre pupilles ($m = 4$) à droite de l'origine et représentation de la PSF normalisée optimale dans cette configuration pour l'optimisation des amplitudes des champs. 30
2.8 Configuration optimale des quatre pupilles ($m = 4$) à droite de l'origine et représentation de la PSF normalisée optimale dans cette configuration pour l'optimisation des positions des pupilles. 31
2.9 Configuration optimale des quatre pupilles ($m = 4$) à droite de l'origine et représentation de la PSF normalisée optimale dans cette configuration pour l'optimisation des positions des pupilles. 33

2.10 Configuration optimale des quatre pupilles ($m = 4$) à droite de l'origine et représentation de la PSF normalisée optimale dans cette configuration pour l'optimisation des amplitudes des champs et des positions des pupilles. 35
2.11 Comparaison des 10^4 solutions du problème (2.3) en termes de flux (F) et de dynamique (D) pour différents points de départ. 36
2.12 Configuration optimale des pupilles et représentation de la PSF normalisée pour la solution optimale trouvée par les 10^4 procédures d'optimisation. 37
2.13 Valeurs de D, F et R pour différents CLF en fonction de $\Delta\alpha$ (a) et de α_{min} (pour $\Delta\alpha = 0.40$) (b). 38
2.14 Configuration optimale des pupilles et représentation de la PSF normalisée optimale pour $\Delta\alpha = 0.20$ (haut) et pour $\Delta\alpha = 0.70$ (bas) avec $\alpha_{moy} = 0.50$. 39
2.15 Configuration optimale des pupilles et représentation de la PSF normalisée optimale pour $\alpha_{min} = 0.05$ (haut) et pour $\alpha_{min} = 0.30$ (bas) avec $\Delta\alpha = 0.40$. 40
2.16 Configuration optimale des pupilles et représentation de la PSF normalisée optimale pour quatre pupilles ($m = 2$) avec $CLF = [0.25, 0.75]$ (haut) et $D = 10^6$ (bas). 41
2.17 Configuration optimale des pupilles et représentation de la PSF normalisée optimale pour vingt-quatre pupilles ($m = 12$) avec $CLF = [0.25, 0.75]$ (haut) et $D = 10^6$ (bas). 42
2.18 PSF normalisée obtenue avec les paramètres optimaux perturbés pour $\varrho = 10^{-3}$ (gauche) et $\varrho = 10^{-2}$ (droite). 43

3.1 Hypertélescope temporel THT 46
3.2 PSF temporelle normalisée pour huit pupilles alignées. 47
3.3 Représentation de la fonction objectif ϕ_1 définie par la formule (3.27) pour plusieurs valeurs de n avec $I = [\pi/2, 3\pi/2]$. Pour $n = 1, 3, 7(a)$, l'échelle est linéaire et pour $n = 7(b), 9, 13$, elle est semi-logarithmique. 56
3.4 PSF temporelle optimale pour huit pupilles ($n = 7$) alignées réparties périodiquement avec $I = [\pi/2, 3\pi/2]$. 72

1.1 Contours de Q pour différentes valeurs de μ. 84

2.1 Exemple de trajectoire $\mu \mapsto x(\mu)$. 98
2.2 (a) et (b) : la suite du paramètre de pénalité $\{\mu_k\}$ converge superlinéairement vers zéro. La suite des itérés dans l'espace primal, est asymptotiquement tangente à la trajectoire. Au cours des itérations, $||w(\mu_k) - w_k||/\mu_k$ tend vers zéro. (c) et (d) : la suite du paramètre de pénalité $\{\mu_k\}$ converge quadratiquement vers zéro. La suite des itérés dans l'espace primal, n'est pas asymptotiquement tangente à la trajectoire. Au cours des itérations, $||w(\mu_k) - w_k||/\mu_k$ tend vers l'infini. 103

3.1 Profils de performance [20] des trois adaptations de la décroissance du paramètre de pénalité μ. Chaque courbe représente la proportion de problèmes qui sont résolus avec un nombre d'évaluations inférieur à 2^x fois le nombre d'évaluations de la meilleure méthode. Plus la courbe est au-dessus des autres, meilleure est la performance. 116
3.2 Décroissance de μ et de $||F(w, \mu)||$ au cours des itérations de l'algorithme E lorsque le problème ELEC ($n = 450, m = 150$) est résolu selon le choix dynamique ou non de μ. 117
3.3 Profils de performance des deux méthodes : méthode primale-duale et méthode SQP. 118

C.1 Structure of a hypertelescope. 132
C.2 Normalized PSF with four pupils. 133
C.3 Comparison of 10^4 solutions of problem (C.6) with different starting points. 137
C.4 Best solution found amongst 10^4 occurrences of problem (C.6) when $[\alpha_{\min}, \alpha_{\max}] = [0.25, 0.75]$. On the left figure, each rectangle represents the modulus of a pupil, the numbers on top are the distance between two pupils, the numbers in the middle are the coefficients a_k. The right figure shows the graph of the normalized PSF in linear and logarithmic scale. 137
C.5 Values of dynamic, central flux and number of resels according to a variation of $\Delta\alpha$ (on left), of $\alpha_{\min}$ but with $\Delta\alpha = 0.4$ (on right). . . . 139
C.6 Optimal solution for $[\alpha_{\min}, \alpha_{\max}] = [0.4, 0.6]$. 140
C.7 Optimal solution for $[\alpha_{\min}, \alpha_{\max}] = [0.15, 0.85]$. 140
C.8 Optimal solution for $[\alpha_{\min}, \alpha_{\max}] = [0.05, 0.45]$. 141
C.9 Optimal solution for $[\alpha_{\min}, \alpha_{\max}] = [0.3, 0.7]$. 141
C.10 Optimal solution for 24 pupils and $[\alpha_{\min}, \alpha_{\max}] = [0.25, 0.75]$. 142
C.11 Optimal solution for 24 pupils and $[\alpha_{\min}, \alpha_{\max}] = [0.09, 0.83]$. 142
C.12 Normalized PSF with perturbed optimal parameters ($\tau = 10^{-3}$ on left and $\tau = 10^{-2}$ on right). Compare with the optimal configuration shows in Figure C.4. 143

Liste des tableaux

1.1 Zéros et extréma de $\alpha \mapsto \frac{2}{\pi\alpha} J_1(\pi\alpha)$. 11
1.2 Positions des quatre pupilles à droite de l'origine et amplitudes des champs reçues par chacune d'elles. 15

2.1 Fonctions d'apodisation. 23
2.2 Coefficients de Fourier des fonctions d'apodisation. 25
2.3 Valeurs des amplitudes optimales des champs pour des positions de pupilles fixées. 28
2.4 Valeurs des critères optimaux de la PSF pour l'optimisation des amplitudes des champs. 29
2.5 Valeurs des amplitudes optimales des champs pour des positions de pupilles fixées. 29
2.6 Valeurs des critères optimaux de la PSF pour l'optimisation des amplitudes des champs. 29
2.7 Valeurs des positions optimales des pupilles pour des amplitudes de champs fixées. La configuration étant symétrique, la position u_{-1} est égale à l'opposé de la position u_1. 31
2.8 Valeurs des critères optimaux de la PSF pour l'optimisation des positions des pupilles. 32
2.9 Valeurs des positions des pupilles optimales pour des amplitudes de champs fixées. La configuration étant symétrique, la position u_{-1} est égale à l'opposé de la position u_1. 32
2.10 Valeurs des critères optimaux de la PSF pour l'optimisation des positions des pupilles. 33
2.11 Valeurs des amplitudes des champs et des positions des pupilles optimales pour l'optimisation simultanée des deux variables. La configuration étant symétrique, la position u_{-1} est égale à l'opposé de la position u_1. 34
2.12 Valeurs des critères optimaux de la PSF pour l'optimisation des amplitudes des champs et des positions des pupilles. 35
2.13 Amplitudes des champs et positions des pupilles pour la meilleure solution des 10^4 procédures d'optimisation. La configuration étant symétrique, la position u_{-1} est égale à l'opposé de la position u_1. . . . 37
2.14 Valeurs de D, R et F en fonction de $\Delta\alpha$ (a) et de $\alpha_{\min}$ (pour $\Delta\alpha = 0.40$) (b). 38
2.15 Valeurs de D, R et F en fonction du nombre de pupilles n avec $CLF = [0.25, 0.75]$ (a) et $D \simeq 10^6$ (b). 40

2.16 Valeurs de D, R et F pour des perturbations d'amplitudes de champs et/ou de positions de pupilles pour huit pupilles ($m = 4$) avec $CLF = [0.25, 0.75]$. 43

3.1 Valeurs de ℓ^* et de $\phi_1(\ell^*)$ pour différentes valeurs de n avec $I = [\pi/2, 3\pi/2]$. 55
3.2 Valeurs des limites de ϕ_1 en zéro et l'infini pour plusieurs valeurs de n. 58
3.3 Valeurs des amplitudes optimales des champs normalisées pour huit pupilles ($n = 7$) alignées réparties périodiquement avec $I = [\pi/2, 3\pi/2]$. 72

1.1 Valeurs du minimum x^*_μ de Q et du conditionnement de $\nabla^2_{xx} Q(x^*_\mu, \mu)$ pour différentes valeurs de μ. 84

3.1 Liste des 113 problèmes testés avec les algorithmes C et E provenant des librairies COPS 3.0 [21] et CUTEr [33]. 113

A.1 Amplitudes optimales des champs normalisées pour la fonction Porte. 121
A.2 Amplitudes optimales des champs normalisées pour la fonction Triangle. 122
A.3 Amplitudes optimales des champs normalisées pour la fonction $\mathcal{C}^\infty$. . 122
A.4 Amplitudes optimales des champs normalisées pour la fonction de Hanning. 123
A.5 Amplitudes optimales des champs normalisées pour la fonction de Hamming. 123
A.6 Amplitudes optimales des champs normalisées pour la fonction de Blackman. 124

C.1 Optimal values for the best solution. The value u_0 is set to the position of the pupil symmetric to the first one with respect to zero. 138
C.2 Values of dynamic, number of resels and central flux according to the variation of $\Delta\alpha$ (a) and $\alpha_{\min}$ with $\Delta\alpha = 0.4$ (b). 139
C.3 Values of dynamic, number of resels and central flux according to a variation of the number of pupils. 142
C.4 Values of dynamic, number of resels and central flux according to a perturbation of a_k and/or u_k with 8 pupils and CLF $= [0.25, 0.75]$. . . 142

D.1 Résulats numériques obtenus avec l'algorithme E en considérant la définition (3.2) pour la valeur de μ_0 ainsi que les expressions (3.3) et (3.4) pour le choix dynamique de μ. La décroissance de μ est en plus adaptée à la décroissance de $\|F(w, \mu)\|$. Dans la boucle qui permet d'adapter les deux décroissances, la valeur de F n'est pas recalculée chaque fois que la valeur de μ est modifiée. 148
D.2 Résulats numériques obtenus avec l'algorithme E en considérant la définition (3.2) pour la valeur de μ_0 ainsi que les expressions (3.3) et (3.4) pour le choix dynamique de μ. La décroissance de μ est en plus adaptée à la décroissance de $\|F(w, \mu)\|$. Dans la boucle qui permet d'adapter les deux décroissances, la valeur de F est recalculée chaque fois que la valeur de μ est modifiée. 151

D.3 Résulats numériques obtenus avec l'algorithme E en considérant la définition (3.2) pour la valeur de μ_0 ainsi que les expressions (3.3) et (3.4) pour le choix dynamique de μ. La décroissance de μ est en plus adaptée à la décroissance de $\|\nabla f(x) + \nabla c(x)\lambda\|$. 154

E.1 Résulats numériques obtenus avec l'algorithme C. 158

Introduction générale

Dans ce manuscrit, nous présentons des résultats théoriques et numériques d'optimisation non linéaire. Ces résultats ont été obtenus durant mes trois années de thèse effectuées au sein du Département de Mathématiques et Informatique (DMI) du laboratoire XLIM de l'Université de Limoges. Ces travaux de recherche correspondent

- à l'étude d'un problème issu de la physique. Cette application concerne l'optimisation de la configuration d'un réseau de télescopes ;
- à la mise en œuvre et à l'analyse d'une méthode numérique afin de résoudre un problème d'optimisation non linéaire.

Ces résultats ont fait l'objet d'un article paru [9] et de plusieurs communications orales (3 posters et 5 exposés). Deux articles sont soumis pour publication et un autre est en cours de rédaction.

Ce mémoire est divisé en deux parties.

- **Optimisation des paramètres d'un réseau de télescopes optiques**
- **Méthode primale-duale pour l'optimisation avec contraintes d'égalité**

Dans la première partie, nous présentons le travail que nous avons effectué en collaboration avec le département Photonique du laboratoire XLIM au sein du projet transverse Imagerie Radar et Optique (IRO). L'objectif de ce travail était d'optimiser les paramètres d'entrée d'un instrument optique constitué d'un réseau de télescopes [41] afin de pouvoir visualiser des objets de faibles luminosités et de très petites tailles tels que les planètes extra-solaires [43]. Pour la résolution du problème, nous avons considéré un instrument avec plusieurs télescopes alignés. Les paramètres de l'instrument à optimiser sont les positions des télescopes et les amplitudes des champs reçues par chacun d'eux. Pour les déterminer, nous avons proposé plusieurs modèles mathématiques sous la forme de problèmes d'optimisation non linéaires. Cette partie est divisée en quatre chapitres. Dans le chapitre 1, nous présentons le principe général de l'instrument optique, le problème physique étudié ainsi que sa modélisation mathématique. Les résultats numériques sont analysés dans le chapitre 2. Dans le chapitre 3, nous donnons les différents résultats théoriques obtenus dont le théorème d'existence de la solution du problème posé lorsque les télescopes sont positionnés avec une configuration périodique. Enfin, nous concluons dans le chapitre 4 sur les différents résultats obtenus ainsi que sur les perspectives de cette application.

Dans la deuxième partie, nous présentons un nouvel algorithme pour résoudre des problèmes d'optimisation non linéaires avec contraintes d'égalité. Dans les années 70, une des premières méthodes utilisées pour résoudre un problème avec contraintes était de remplacer le problème initial par un problème sans contrainte où la fonction à minimiser est une fonction de pénalisation [23]. Une autre méthode classique et très efficace pour résoudre des problèmes d'optimisation non linéaires avec contraintes est la méthode de programmation quadratique successive (SQP) [10, 68]. Elle transforme le problème initial en une suite de sous-problèmes quadratiques plus simples à résoudre. L'idée générale de notre méthode est de résoudre un système primal-dual qui peut s'interpréter comme les conditions d'optimalité perturbées du problème initial ou comme les conditions d'optimalité du problème pénalisé, avec une méthode Newtonnienne. La globalisation de la méthode est obtenue avec une technique de recherche linéaire. L'avantage de cette nouvelle approche est d'éviter le mauvais conditionnement présent avec les méthodes de pénalisation. Cette partie est divisée en quatre chapitres. Dans le chapitre 1, nous définissons le problème général que nous souhaitons résoudre et faisons des rappels sur les méthodes de pénalisation quadratique et SQP. Le principe de la méthode primale-duale est présenté dans le chapitre 2 ainsi que les résultats théoriques de convergence locale, globale et de l'analyse asymptotique. Dans le chapitre 3, nous analysons les résultats numériques obtenus avec notre méthode. De plus, une étude comparative est présentée entre les méthodes primale-duale et SQP. Enfin, nous concluons dans le chapitre 4 sur les différents résultats obtenus ainsi que sur les perspectives possibles de la méthode.

Première partie

Optimisation des paramètres d'un réseau de télescopes optiques

Chapitre 1

Problème physique et modélisation mathématique

1.1 Introduction

Précurseur du télescope, la lunette d'approche a été conçue en Italie au XVI$^{\text{ème}}$ siècle. En 1609, Galilée présenta la première lunette astronomique. Kepler en perfectionna le principe en proposant une formule optique à deux lentilles. Isaac Newton construisit une première version du télescope en 1671. Un télescope est un instrument optique constitué de plusieurs miroirs permettant d'augmenter la luminosité et la taille apparente des objets éloignés ou des astres à observer.

Le télescope est un des instruments optiques les plus utilisés. Le but d'une observation dans l'espace est d'obtenir une image caractérisant l'objet observé. Henry Draper, astronome américain et pionnier de l'astrophotographie, fut le premier à photographier le spectre stellaire de Vega en 1872 et de la nébuleuse d'Orion en 1880. Le télescope de 2.54 mètres de diamètre de l'observatoire du Mont Wilson en Californie est célèbre pour les travaux de l'astronome américain Edwin Hubble. Au début du siècle, les observations d'Hubble ont permis de montrer que les nébuleuses observées auparavant avec des télescopes moins puissants ne font pas partie de notre galaxie, mais elles constituent d'autres galaxies éloignées. De nos jours, les plus grands télescopes ont un diamètre d'une dizaine de mètres. Cependant, ces télescopes ne sont pas suffisamment grands pour détecter des objets de petites tailles et de faibles luminosités. Pour remédier à ce problème, l'astronomie utilise des techniques de haute résolution angulaire pour révéler les détails les plus petits des objets observés. La résolution spatiale correspond au plus petit détail perceptible sur l'image, appelé élément résolu ou résel. Le pouvoir de résolution d'un instrument optique est lié à la diffraction de la lumière et il augmente avec le diamètre de la lentille utilisée.

Une très haute résolution angulaire est obtenue grâce à l'interférométrie. Les astronomes considèrent un interféromètre constitué de plusieurs télescopes (voir figure 1.1). L'image observée correspond au mélange de lumière provenant de chaque télescope et le pouvoir de résolution de l'instrument correspond dans ce cas à la distance séparant les deux télescopes les plus éloignés. Ainsi, ce genre d'instrument permet d'obtenir une résolution bien meilleure que celle obtenue avec un seul télesco-

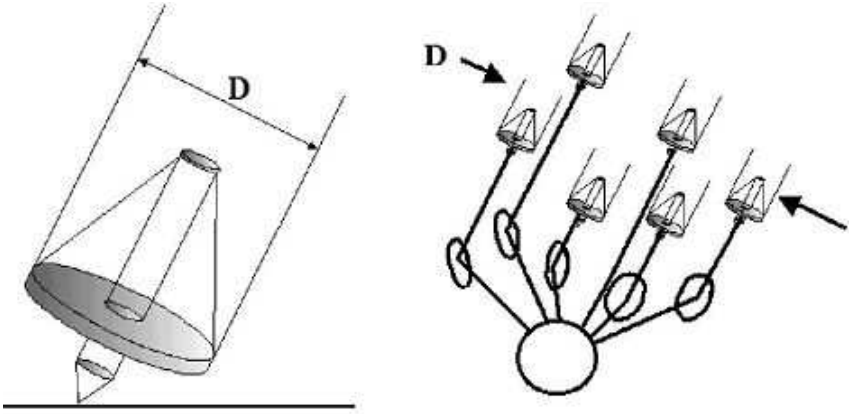


FIG. 1.1 – Télescope et réseau de télescopes.

pe. Actuellement, les deux télescopes de l'observatoire W. M. Keck sont les plus grands télescopes optiques des observatoires du mont Mauna Kea de l'île d'Hawaï. Ils ont chacun un miroir de 10 mètres de diamètre. Ils peuvent fonctionner indépendamment ou ensemble par le biais de l'interférométrie optique [57]. Leur résolution angulaire est équivalente à celle d'un miroir de 85 mètres de diamètre. De même, le Very Large Telescope (VLT) de l'observatoire du Cerro Paranal, situé dans le désert d'Atacama au nord du Chili, est constitué de quatre télescopes possédant chacun un miroir de 8.20 mètres de diamètre. Les télescopes fonctionnant ensemble sont équivalent à un télescope de 200 mètres de diamètre.

En 1996, Antoine Labeyrie [41] propose un instrument composé d'un réseau de télescopes fortement espacés permettant une imagerie directe à très haute résolution angulaire (voir figure 1.2). Ce concept instrumental est appelé hypertélescope.

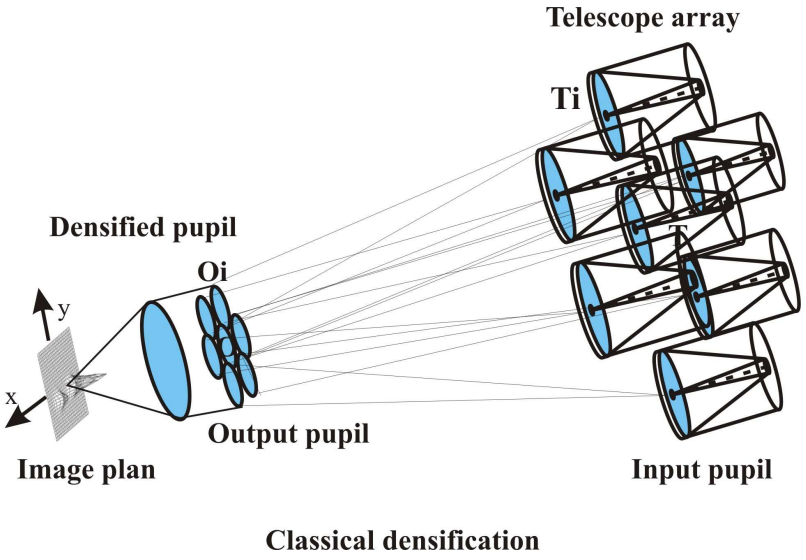


FIG. 1.2 – Hypertélescope

Le réseau de télescopes forme la pupille d'entrée du dispositif. Celle-ci est ensuite reconfigurée au niveau d'une pupille de sortie dite densifiée. Les positions des sous-pupilles sont conservées de manière homothétique dans le plan de la pupille densifiée et leur diamètre est augmenté pour obtenir une pupille compacte. L'intensité lumineuse correspondant au mélange de lumière provenant de chacun des télescopes est visualisée dans le plan image. La densification permet de concentrer l'intensité lumineuse collectée en un seul point dans le plan image. La configuration de l'Interferometric Remapped Array Nulling (IRAN), proposée par l'équipe de F. Vakili [63] en 2004, est une version dérivée de ce concept où l'équipe propose une densification dans le plan image. En 2007, F. Reynaud et L. Delage [61] ont proposé une version temporelle d'un hypertélescope.

Les propriétés de l'image observée dans le plan image vont dépendre des positions des sous-pupilles en entrée et des amplitudes complexes des champs optiques traversant chaque ouverture. Ces différents paramètres doivent être choisis et contrôlés de manière optimale afin que le pouvoir de résolution de l'instrument soit maximal tout en assurant une grande dynamique de la mesure des intensités. Un des objectifs de ce type d'instrument est de réaliser des images d'exoplanètes [43]. Ces planètes orbitent autour d'une autre étoile que le soleil. Pendant longtemps, l'existence de planètes extrasolaires n'a pas pu être prouvée par une observation directe. En effet, la distance et le manque de luminosité de ces objets célestes si petits par rapport aux étoiles autour desquelles ils orbitent, ont rendu leur détection directe difficile. La première planète extrasolaire (51 Pegasi b) a été découverte de manière indirecte en 1995. Aujourd'hui, plus de 300 exoplanètes ont été détectées. Toutes ces planètes présentent une masse supérieure à celle de la Terre. En novembre 2008, la première exoplanète découverte par visualisation directe (Formalhaut b) a été détectée sur une photographie coronographique provenant du télescope spatial Hubble. Au cours de cette même année, les télescopes Keck et Gemini d'Hawaï ont trouvé à l'aide d'une technique d'imagerie directe un système de trois planètes HR8799.

Le but de ce chapitre est de présenter le problème physique qui nous a été posé ainsi que sa modélisation mathématique. Dans la section 1.2, un modèle mathématique de l'hypertélescope est présenté et le champ optique est défini. L'intensité lumineuse visible dans le plan image pour un instrument composé d'un ou de plusieurs télescopes est ensuite caractérisée dans le paragraphe 1.3. Dans la partie 1.4, l'intensité lumineuse est définie pour un réseau linéaire de télescopes car c'est la configuration de l'instrument que nous avons supposée pour résoudre le problème. Enfin, les deux méthodes d'optimisation mises en place pour la résolution du problème sont décrites dans la section 1.5.

1.2 Champ optique

Dans le modèle mathématique considéré, nous nous intéressons uniquement au plan de la pupille densifiée et au plan image de l'hypertélescope qui sont représentés sur la figure 1.3. Le plan de coordonnées (u, v) est défini comme le plan pupille. Il correspond au plan de la pupille densifiée de l'hypertélescope dans la description

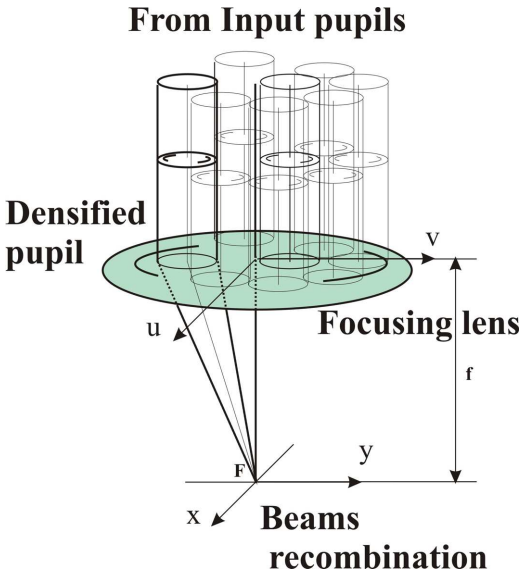


FIG. 1.3 – Plan pupille et plan image de l'hypertélescope.

précédente. Le plan de coordonnées (x, y) est le plan où se forme l'image. L'intensité lumineuse est recombinée dans ce plan.

1.2.1 Champ optique dans le plan pupille

Dans le plan pupille, n ouvertures de centres (u_k, v_k), $k = 1, \ldots, n$ de même diamètre d sont dénombrées. Nous supposons que les ouvertures ne se superposent pas.

Le champ optique reçu par une ouverture k est décrit par une onde plane monochromatique, caractérisée par le nombre complexe

$$a_k e^{i\varphi_k},$$

où $a_k \geq 0$ est l'amplitude du champ et $\varphi_k \in [0, 2\pi[$ sa phase. Pour la suite, les ondes sont supposées en phases, i.e. $\varphi_k = 0$ pour tout k.

Le champ optique dans le plan pupille est défini par

$$g_n(u, v) := \sum_{k=1}^{n} a_k \mathbf{1}_{B_k}(u, v),$$

où $\mathbf{1}_{B_k}$ est la fonction caractéristique de la boule fermée de centre (u_k, v_k) de diamètre d :

$$\mathbf{1}_{B_k}(u, v) := \begin{cases} 1 & \text{si } (u - u_k)^2 + (v - v_k)^2 \leq \left(\frac{d}{2}\right)^2 \\ 0 & \text{sinon.} \end{cases}$$

1.2.2 Champ optique dans le plan image

Dans le plan image, le champ optique correspond à la transformée de Fourier du champ optique du plan pupille, soit

$$\hat{g_n}(x,y) := \iint g_n(u,v) e^{-i\frac{2\pi}{\lambda f}(xu+yv)} \ \mathrm{d}u\mathrm{d}v,$$

où λ est la longueur d'onde du signal et f la distance focale de la lentille de focalisation.

En utilisant la linéarité de l'intégrale, la fonction $\hat{g}_n$ peut s'écrire sous la forme :

$$\hat{g_n}(x,y) = \hat{g}(x,y) \sum_{k=1}^{n} a_k e^{-i\frac{2\pi}{\lambda f}(xu_k+yv_k)},$$

avec $\hat{g}$ la transformée de Fourier associée à une seule ouverture circulaire centrée en l'origine de diamètre d telle que $a_1 = 1$. Elle est définie par

$$\hat{g}(x,y) := \iint e^{-i\frac{2\pi}{\lambda f}(xu+yv)} \mathbf{1}_B(u,v) \ \mathrm{d}u\mathrm{d}v,$$

où B est la boule fermée de centre 0 et de diamètre d.

La valeur de $\hat{g}(0,0) = \dfrac{\pi d^2}{4}$ représente la surface d'une lentille de diamètre d. Celle-ci sera notée S pour la suite.

1.3 Réponse impulsionnelle

La répartition de l'intensité lumineuse dans le plan image est appelée réponse impulsionnelle ou Point Spread Function (PSF). Elle est définie par

$$(x,y) \mapsto |\hat{g_n}(x,y)|^2.$$

Nous considérons uniquement la PSF normalisée définie par

$$\Psi_n(x,y) := \frac{|\hat{g_n}(x,y)|^2}{|\hat{g_n}(0,0)|^2}. \tag{1.1}$$

Dans les deux sections qui suivent, la réponse impulsionnelle normalisée est explicitée pour un instrument composé d'un ou de plusieurs télescopes.

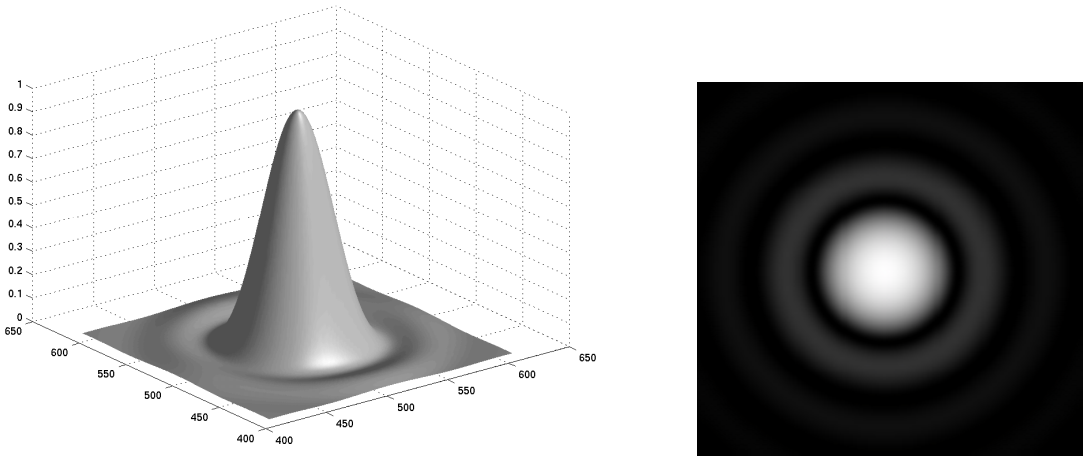
1.3.1 Réponse impulsionnelle pour un télescope

Notons Ψ la réponse impulsionnelle normalisée associée à un point objet lorsque l'instrument d'observation est un télescope. D'après l'optique géométrique, les instruments optiques sont stigmatiques, i.e. à un point objet correspond un point image. Une étoile située à l'infini constitue un point objet. Même si l'instrument est stigmatique, l'image de l'étoile formée par l'objectif dans son plan focal n'est pas un point. En effet, la lumière peut se diffracter lors de son passage à travers l'objectif à cause

de sa nature ondulatoire. Ainsi, au lieu d'avoir un point image dans le plan focal, il se forme une tache de diffraction. Celle-ci est également appelée tache d'Airy. Elle est définie par

$$\Psi(x,y) := \frac{|\hat{g}(x,y)|^2}{S^2}. \tag{1.2}$$

La figure 1.4 donne un exemple de tache d'Airy (vue normalement pour la figure (a) et vue de dessus pour la figure (b)). D'après cette figure, la réponse impulsionnelle normalisée pour une seule pupille est constituée d'un lobe central d'une hauteur égale à un. D'après la vue de dessus, nous pouvons observer un disque central entouré d'anneaux concentriques plus faiblement lumineux séparés par des intervalles obscurs.



a : Vue normale b : Vue de dessus

FIG. 1.4 – Figures d'Airy

Pour la suite, modifions l'écriture (1.2) de la réponse impulsionnelle normalisée associée à une seule ouverture. En considérant les coordonnées polaires

– dans le plan pupille :

$$u = r\cos\theta \text{ et } v = r\sin\theta,$$

– dans le plan image :

$$x = \rho\cos\gamma \text{ et } y = \rho\sin\gamma,$$

tel que $r^2 = u^2 + v^2$ et $\rho^2 = x^2 + y^2$, nous avons

$$\hat{g}(\rho\cos\gamma, \rho\sin\gamma) = \int_0^{\frac{d}{2}}\int_0^{2\pi} e^{-i\frac{2\pi}{\lambda f}r\rho\cos(\theta-\gamma)} r \ \mathrm{d}\theta\mathrm{d}r.$$

La fonction cos étant 2π périodique, $\hat{g}(x,y)$ est alors indépendante de γ. Afin de simplifier les formules, nous noterons pour la suite $\hat{g}(\rho)$ la valeur de la fonction précédente exprimée en coordonnées polaires.

En introduisant les fonctions de Bessel [1, 3] de première espèce d'ordres 0 et 1 :

$$J_0(\rho) = \frac{1}{2\pi}\int_0^{2\pi} e^{-i\rho\cos\theta} \ \mathrm{d}\theta,$$

$$J_1(\rho) = \frac{1}{i\pi} \int_0^{2\pi} e^{-i\rho\cos\theta} cos\theta \ \ \mathrm{d}\theta,$$

et sachant que la relation entre les deux fonctions est

$$\gamma J_1(\gamma) = \int_0^{\gamma} J_0(\omega)\omega \ \ \mathrm{d}\omega,$$

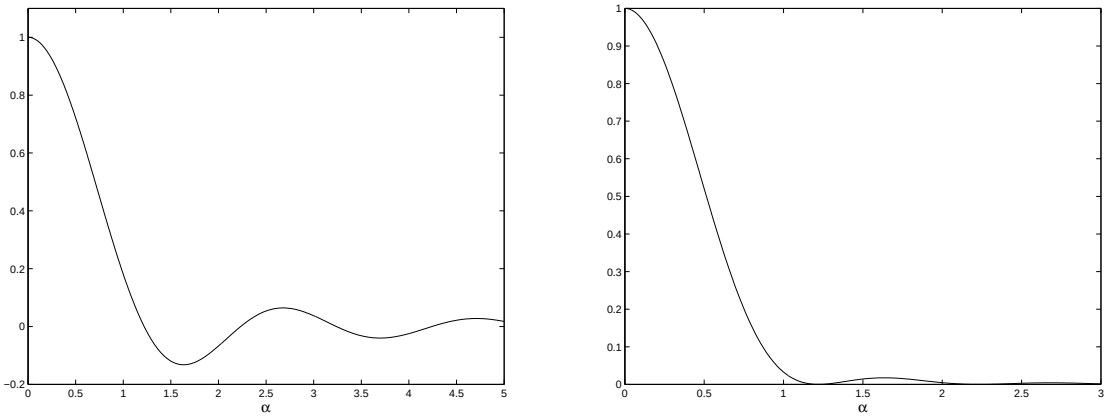
nous obtenons

$$\hat{g}(\rho) = 2\pi \int_0^{\frac{d}{2}} J_0\left(\frac{2\pi}{\lambda f} r\rho\right) r \ dr = \frac{2\lambda f}{\pi d} \frac{S}{\rho} J_1\left(\frac{\pi d}{\lambda f}\rho\right).$$

Via le changement d'échelle $\alpha = \dfrac{d}{\lambda f}\rho$, nous définissons une nouvelle fonction ψ :

$$\psi(\alpha) := \Psi(\frac{\lambda f \alpha}{d}, 0) = \frac{|\hat{g}(\rho)|^2}{S^2} = \left[\frac{2}{\pi\alpha} J_1(\pi\alpha)\right]^2, \tag{1.3}$$

qui sera aussi appelée pour la suite, réponse impulsionnelle normalisée pour une pupille ou tache d'Airy.



a : $\alpha \mapsto \dfrac{2}{\pi\alpha} J_1(\pi\alpha)$ b : $\alpha \mapsto \left[\dfrac{2}{\pi\alpha} J_1(\pi\alpha)\right]^2$

FIG. 1.5 – PSF normalisée pour une pupille.

La figure 1.5 représente le graphe de la fonction $\alpha \mapsto \dfrac{2}{\pi\alpha} J_1(\pi\alpha)$ et de son carré. Les premiers zéros et extréma de cette fonction sont donnés dans le tableau 1.1.

zéros	1.22	2.23	3.24	4.24
extréma	1.63	2.68	3.70	4.71

TAB. 1.1 – Zéros et extréma de $\alpha \mapsto \dfrac{2}{\pi\alpha} J_1(\pi\alpha)$.

Le pouvoir de résolution d'un système optique désigne la capacité à distinguer des détails fins. Il correspond à la distance angulaire minimale entre deux éléments d'un objet qui permet d'en obtenir deux images séparées. La diffraction limite le pouvoir de résolution des instruments optiques. Si deux objets ponctuels sont trop proches, les taches de diffraction vont se chevaucher et il est alors impossible d'obtenir des images séparées. La résolution est donc liée à la largeur du lobe central de la réponse impulsionnelle. D'après les formules précédentes et le tableau 1.1, pour arriver à distinguer deux taches, il faut que la distance qui les sépare soit au moins égale à 1.22. La résolution est alors définie pour

$$\alpha := \frac{d}{\lambda f}\rho = 1.22.$$

Avec une distance focale d'un mètre, la résolution est

$$\rho = 1.22\frac{\lambda}{d}.$$

Le pouvoir de résolution d'un télescope de dix mètres de diamètre ($d = 10$) avec une longueur d'onde $\lambda = 400 \times 10^{-9}$ mètres est d'environ 4.88×10^{-8} radians, i.e. 0.01 seconde d'arc. Cette grandeur signifie qu'un objet de dix-sept mètres situé sur la Lune est observable depuis la Terre.

L'idéal étant d'avoir un pouvoir de résolution élevé, le diamètre d du télescope doit donc être grand. Comme la construction de grands télescopes est difficile, les astronomes favorisent l'utilisation de l'interférométrie à l'aide de plusieurs télescopes. La résolution est alors définie comme la distance séparant les deux télescopes les plus éloignés du réseau.

1.3.2 Réponse impulsionnelle pour n télescopes

À présent, considérons un hypertélescope constitué de n télescopes ($n > 1$). En utilisant la définition générale (1.1) de la réponse impulsionnelle normalisée et les notations précédentes, nous avons

$$\Psi_n(x,y) := \frac{|\hat{g}(x,y)|^2}{S^2}\frac{\left|\sum_{k=1}^n a_k e^{-i\frac{2\pi}{\lambda f}(xu_k+yv_k)}\right|^2}{\left(\sum_{k=1}^n a_k\right)^2}.$$

Comme $\Psi_n(x,y)$ est homogène par rapport au vecteur des amplitudes $(a_1,\ldots,a_n)$, nous supposons pour la suite

$$\sum_{k=1}^n a_k = 1.$$

La PSF normalisée pour n télescopes s'écrit donc de la manière suivante :

$$\Psi_n(x,y) = \frac{|\hat{g}(x,y)|^2}{S^2}\left|\sum_{k=1}^n a_k e^{-i\frac{2\pi}{\lambda f}(xu_k+yv_k)}\right|^2. \tag{1.4}$$

Le premier terme de Ψ_n correspond à l'enveloppe de diffraction et le second au module au carré de la fonction d'interférence.

La figure 1.6 est un exemple de réponse impulsionnelle normalisée obtenue avec quatre pupilles ($n = 4$) de même diamètre ($d = 1$). Le graphe est constitué d'un lobe central d'une hauteur égale à un et de plusieurs lobes secondaires plus ou moins hauts. Il est important de noter que le graphe n'est pas nécessairement symétrique.

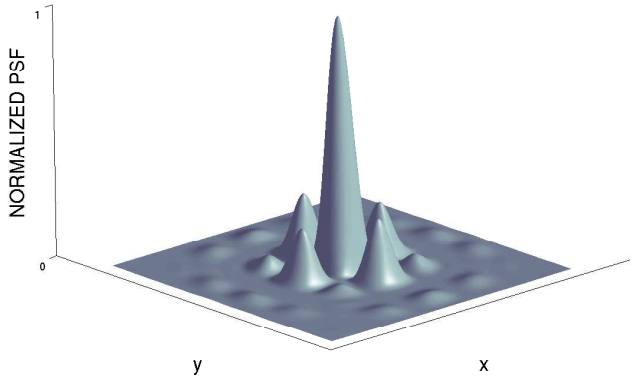


FIG. 1.6 – PSF normalisée pour quatre pupilles.

L'objectif est de trouver les valeurs des positions des pupilles (u_k, v_k) et des champs optiques a_k pour $k = 1, \ldots, n$ de sorte que le graphe de la PSF normalisée présente

- un lobe central le plus étroit possible pour avoir un grand pouvoir de résolution,
- des lobes secondaires les plus bas possibles sur un domaine prédéfini pour avoir une grande valeur de dynamique.

Afin de visualiser des exoplanètes, le pouvoir de résolution de l'instrument doit être de l'ordre de quelque micro-arc-seconde [46] et la dynamique de l'ordre de 10^6. En effet, pour distinguer le lobe central de la réponse impulsionnelle de l'exoplanète dans le domaine prédéfini, il faut que les lobes secondaires de la réponse impulsionnelle de l'étoile soient les plus bas possibles sur ce domaine. Ainsi, le problème posé est de nature bi-critère car il faut trouver un compromis optimal entre dynamique et résolution.

L'optimisation des instruments optiques a déjà fait l'objet de travaux de recherche, notamment dans les domaines de la microscopie [49] et de l'astronomie [40]. Pour détecter des exoplanètes, R.J. Vanderbei [64] a utilisé la technique de coronographie et a résolu un problème d'optimisation dans le but de déterminer la forme optimale que doit avoir un masque afin de l'appliquer sur la lentille d'un seul télescope. Le problème est ici différent puisque nous considérons un réseau de plusieurs télescopes. C. Aime et R. Soummer [2] ont utilisé des techniques d'apodisation. F. Patru et al. [44, 55, 56] combinent des techniques d'imagerie directe avec des méthodes de cophasage et de densification. Des travaux ont également été réalisés sur des radiotélescopes [39, 51, 66], télescopes utilisés en radioastronomie,

pour capter des ondes radioélectriques. Pour ce type d'instrument, l'optimisation est réalisée sur les positions des télescopes [12, 13].

1.4 Réseau linéaire de télescopes

La résolution du problème qui nous a été posé a été effectuée avec un hypertélescope constitué de n télescopes alignés. Cette hypothèse se justifie par le fait que les premières expérimentations ont été effectuées avec un hypertélescope constitué de huit pupilles alignées [8, 53]. Ainsi, nous supposons que $v_k = 0$ pour tout $k = 1, \ldots, n$. La réponse impulsionnelle normalisée (1.4) s'écrit alors :

$$\Psi_n(x, y) := \frac{|\hat{g}(x, y)|^2}{S^2} \left| \sum_{k=1}^{n} a_k e^{-i\frac{2\pi}{\lambda f} x u_k} \right|^2 .$$

La fonction d'interférence contient les paramètres d'optimisation, à savoir les positions u_k des ouvertures et les amplitudes des champs a_k pour $k = 1, \ldots, n$. Ce terme ne dépendant pas de y, une valeur arbitraire de y peut alors être considérée, notamment $y = 0$. La figure 1.7 est un exemple de réponse impulsionnelle normalisée obtenue avec huit pupilles alignées ($n = 8$) de même diamètre ($d = 1$).

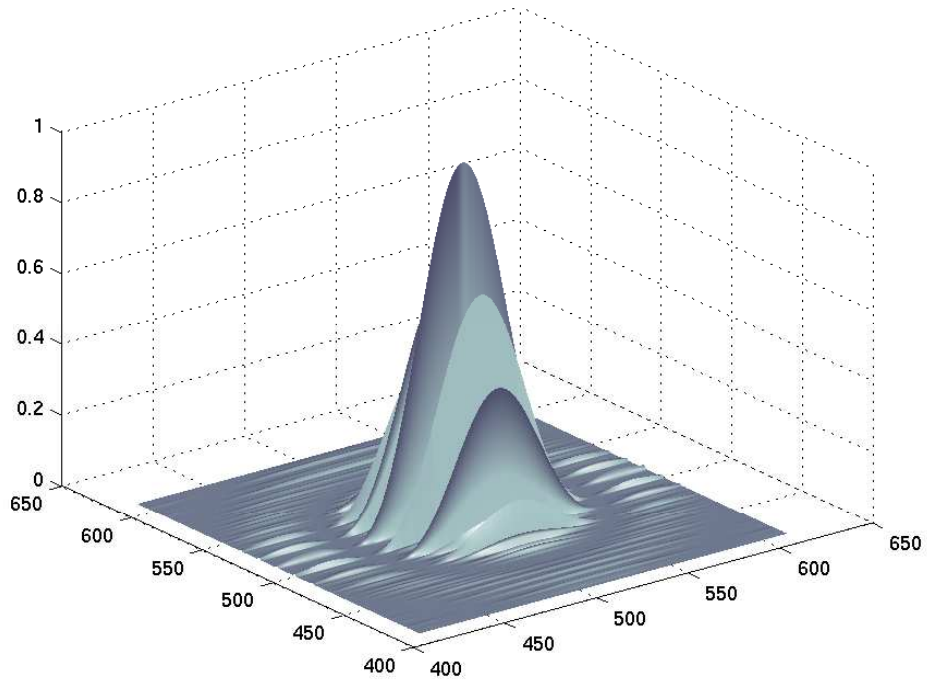


FIG. 1.7 – PSF normalisée pour huit pupilles alignées.

En utilisant la formule (1.3) et le changement de variable $\alpha = \frac{d}{\lambda f} x$, la nouvelle PSF normalisée pour n télescopes alignés s'écrit :

$$\psi_n(\alpha) := \Psi_n(\frac{\lambda f \alpha}{d}, 0) = \left[\frac{2}{\pi\alpha} J_1(\pi\alpha)\right]^2 \left|\sum_{k=1}^{n} a_k e^{-i\frac{2\pi u_k}{d}\alpha}\right|^2 = \psi(\alpha)\left|\sum_{k=1}^{n} a_k e^{-i\frac{2\pi u_k}{d}\alpha}\right|^2. \tag{1.5}$$

La figure 1.8 représente un exemple de réponse impulsionnelle normalisée obtenue avec huit pupilles alignées ($n = 8$) de même diamètre ($d = 1$) et disposées symétriquement autour de l'origine. Le graphe étant symétrique par rapport à l'origine, nous avons représenté uniquement la partie à droite de zéro de la PSF. Les positions des quatre pupilles se trouvant à droite de l'origine et les amplitudes des champs reçues par chacune d'elles sont reportées dans le tableau 1.2. Pour la présentation des résultats, les valeurs des amplitudes des champs sont normalisées par rapport à l'amplitude maximale. Cette considération sera utilisée pour la suite.

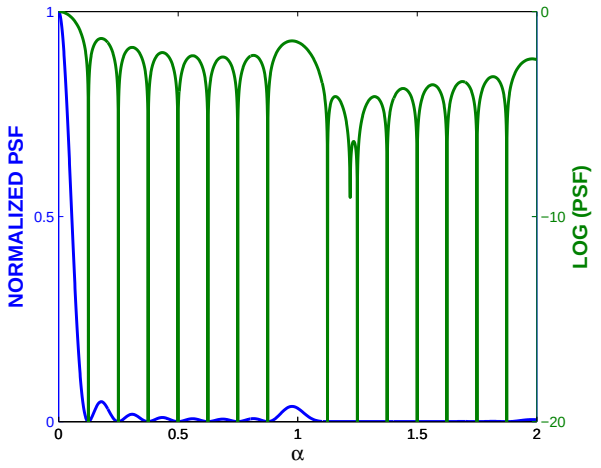


FIG. 1.8 – PSF normalisée pour huit pupilles alignées (échelle linéaire en bleu et échelle semi-logarithmique en vert).

k	u_k	a_k
1	0.50	1.00
2	1.50	1.00
3	2.50	1.00
4	3.50	1.00

TAB. 1.2 – Positions des quatre pupilles à droite de l'origine et amplitudes des champs reçues par chacune d'elles.

1.5 Modélisation mathématique

L'objectif est d'optimiser les amplitudes des champs $a_1, \ldots, a_n$ et les positions des pupilles $u_1, \ldots, u_n$ pour que le graphe de la réponse impulsionnelle ψ_n définie en (1.5) ait un lobe central le plus étroit possible et des lobes secondaires les plus bas possibles sur un intervalle choisi. Pour obtenir un pouvoir de résolution élevé et une grande valeur de dynamique, deux méthodes d'optimisation ont été considérées.

- La méthode des coefficients de Fourier, décrite dans la section 1.5.1, suppose des pupilles régulièrement espacées. Cette configuration permet d'interpréter la fonction d'interférence associée à la réponse impulsionnelle normalisée définie en (1.5) comme la série de Fourier tronquée d'une fonction gabarit qui permettrait d'obtenir une PSF normalisée idéale. Les amplitudes optimales sont alors obtenues à partir du calcul des coefficients de Fourier de la fonction gabarit. En faisant varier un paramètre de cette fonction, des courbes bi-critères (dynamique, résolution) sont obtenues sur lesquelles une solution de compromis est choisie.
- La méthode de l'optimisation de la dynamique, décrite dans la section 1.5.2, consiste à optimiser les positions des pupilles et/ou les amplitudes des champs de sorte que la PSF normalisée présente des lobes secondaires les plus bas possibles sur un intervalle fixé à priori. Le problème est modélisé sous la forme d'un problème d'optimisation.

1.5.1 Méthode des coefficients de Fourier

Avec cette première approche, les pupilles sont supposées régulièrement espacées d'une distance $\ell \geq d$ entre deux centres successifs u_k et u_{k+1}, pour $k = 1, \ldots, n-1$. Pour un nombre n pair de pupilles, les positions sont

$$\ldots, -\frac{3}{2}\ell, -\frac{1}{2}\ell, \frac{1}{2}\ell, \frac{3}{2}\ell, \ldots$$

et pour un nombre impair,

$$\ldots, -2\ell, -\ell, 0, \ell, 2\ell, \ldots$$

Supposons que le nombre d'ouvertures soit impair, soit $n = 2m + 1$ pour $m \in \mathbb{N}$. D'après la formule (1.5) et l'expression des positions, la réponse impulsionnelle normalisée s'écrit :

$$\psi_{2m+1}(\alpha) = \psi(\alpha) \left| \sum_{k=-m}^{m} a_k e^{-i\frac{2\pi\ell k}{d}\alpha} \right|^2. \tag{1.6}$$

L'idée générale de la méthode est de trouver les valeurs des amplitudes a_k et l'écart ℓ entre chaque pupille pour que le terme associé à la somme dans la formule (1.6) de la réponse impulsionnelle s'approche au mieux d'une fonction périodique réelle f bien choisie vérifiant la condition $f(0) = 1$ et de période

$$T := \frac{d}{\ell} \leq 1.$$

Cette fonction peut être interprétée comme un gabarit qui permet d'obtenir une réponse impulsionnelle idéale. La fonction f est aussi appelée fonction d'apodisation [36, 62]. Pour cela, la bonne approximation sera la série de Fourier tronquée à l'ordre m de f normalisée. Celle-ci sera notée $\hat{f}_m$ et vérifiera la condition

$$\hat{f}_m(0) = \sum_{k=-m}^{m} a_k = 1. \tag{1.7}$$

Dans ce cas, la réponse impulsionnelle normalisée s'écrit :

$$\psi_{2m+1}(\alpha) = \psi(\alpha) \left| \hat{f}_m(\alpha) \right|^2. \tag{1.8}$$

La figure 1.9 est un exemple de représentation de $\psi \times f^2$ lorsque la fonction gabarit f est une fonction porte périodique. L'objectif est que le graphe de ψ_n se rapproche au mieux du graphe de $\psi \times f^2$ en présentant un lobe central le plus étroit possible et des lobes secondaires les plus bas possibles sur un intervalle prédéfini.

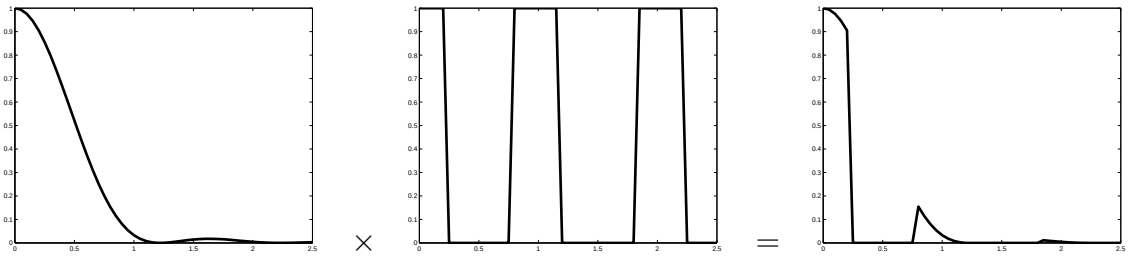


FIG. 1.9 – PSF normalisée obtenue avec une fonction porte.

La somme partielle d'ordre m de la série de Fourier notée $S_m(f)$ est une approximation de f au sens suivant. Désignons par L^2_T ($T = \dfrac{d}{\ell}$) l'ensemble des fonctions $f : \left]-\dfrac{T}{2}, \dfrac{T}{2}\right[\to \mathbb{C}$ de carré intégrable, muni du produit scalaire

$$\langle f, g \rangle := \int_{-\frac{T}{2}}^{\frac{T}{2}} f(t)\overline{g(t)} \ \mathrm{d}t,$$

et de la norme associée

$$\|f\| := \langle f, f \rangle^{\frac{1}{2}} = \left(\int_{-\frac{T}{2}}^{\frac{T}{2}} |f(t)|^2 \ \mathrm{d}t \right)^{\frac{1}{2}}.$$

En notant $\omega := \dfrac{2\pi}{T}$, $e_k(t) := e^{ik\omega t}$ pour $t \in \left]-\dfrac{T}{2}, \dfrac{T}{2}\right[$ et $E_m := \mathrm{vect}\{e_k : k \in [\![-m, m]\!]\}$ l'espace vectoriel engendré par les fonctions e_k, la fonction $S_m(f)$ est alors la meilleure approximation au sens L^2_T de f sur E_m, i.e. la projection orthogonale de f sur E_m. En effet,

$$||f - S_m(f)|| = \min_{g \in E_m} ||f - g||.$$

Pour $k \in [\![-m, m]\!]$, notons $c_k(f)$ le k-ième coefficient de Fourier de f tel que

$$c_k(f) := \frac{\omega}{2\pi} \int_{-\frac{T}{2}}^{\frac{T}{2}} f(t) e^{-ik\omega t} \ \mathrm{d}t.$$

Pour la suite, nous supposons que la fonction d'apodisation est paire. Les coefficients de Fourier obtenus sont alors symétriques, i.e. $c_{-k}(f) = c_k(f)$ pour tout $k \in \mathbb{Z}$, donc réels (car on a toujours $c_k(f) = \overline{c_{-k}(f)}$) et vérifient pour $k \in [\![0, m]\!]$,

$$c_k(f) = \frac{\omega}{2\pi} \int_{-\frac{T}{2}}^{\frac{T}{2}} f(t) \cos(\omega k t) \ \mathrm{d}t.$$

Dans ce cas, $S_m(f)$ vérifie

$$S_m(f)(\alpha) = \sum_{k=-m}^{m} c_k(f) e^{ik\omega\alpha} = c_0(f) + 2\sum_{k=1}^{m} c_k(f) \cos(\omega k \alpha).$$

Pour prendre en compte la normalisation (1.7), nous choisissons comme valeurs pour les amplitudes des champs

$$a_k = \frac{c_k(f)}{\sum_{k=-m}^{m} c_k(f)}$$

pour $k \in [\![-m, m]\!]$.
D'après le Théorème de Dirichlet [38], nous avons pour $\alpha = 0$

$$f(0) = 1 = \sum_{k=-\infty}^{+\infty} c_k(f).$$

Donc pour m assez grand, ceci entraîne que a_k est proche de $c_k(f)$.

1.5.2 Méthode de l'optimisation de la dynamique

La deuxième approche que nous proposons consiste à modéliser le problème physique sous la forme d'un problème de minimisation. L'objectif est de minimiser la hauteur des lobes secondaires de la réponse impulsionnelle sur un intervalle de valeurs de α fixé à priori, noté $[\alpha_{min}, \alpha_{max}] \subset]0, +\infty[$ avec $\alpha_{min} < \alpha_{max}$. Dans cet intervalle, le lobe central de la réponse impulsionnelle associée à l'exoplanète doit être visualisé. En définissant

$$\sigma_n(\alpha) := \frac{2}{\pi\alpha} J_1(\pi\alpha) \sum_{k=1}^{n} a_k e^{-i\frac{2\pi u_k}{d}\alpha},$$

la réponse impulsionnelle normalisée pour n télescopes peut s'écrire sous la forme :

$$\psi_n(\alpha) = |\sigma_n(\alpha)|^2.$$

Les variables d'optimisation sont les amplitudes des champs a_k et les positions u_k des pupilles pour tout k. Des contraintes sont imposées sur les deux paramètres.

– Pour éviter que les pupilles se superposent, nous imposons

$$u_{k+1} - u_k \geq d,$$

pour $k = 1, \ldots, n-1$.

– Comme la réponse impulsionnelle est homogène par rapport au vecteur des amplitudes, nous supposons

$$\sum_{k=1}^{n} a_k = 1,$$

et

$$a_k \geq 0,$$

pour tout k.

Notons les ensembles définissant les contraintes admissibles sur les positions des pupilles et les amplitudes des champs,

$$\mathcal{U}_+ := \{u \in \mathbb{R}^n : u_{k+1} - u_k \geq d, k = 1, \ldots, n-1\},$$

et

$$\mathcal{A}_+ := \left\{ a \in \mathbb{R}^n_+ : \sum_{k=1}^{n} a_k = 1 \right\}.$$

Le modèle général associé à la technique de l'optimisation de la dynamique peut s'écrire :

$$\begin{array}{ll} \text{minimiser}_{(a,u)\in\mathbb{R}^n\times\mathbb{R}^n} & ||\sigma_n||_p\,, \\ \text{sous contrainte} & (a,u) \in \mathcal{A}_+ \times \mathcal{U}_+. \end{array}$$

Deux normes $p = 2$ et $p = \infty$ ont été choisies pour la modélisation du problème. Elles sont définies par

$$\|\sigma\|_2 := \left(\frac{1}{\alpha_{max} - \alpha_{min}} \int_{\alpha_{min}}^{\alpha_{max}} |\sigma(\alpha)|^2 \ \mathrm{d}\alpha \right)^{\frac{1}{2}},$$

et

$$\|\sigma\|_\infty := \max_{\alpha\in[\alpha_{min},\alpha_{max}]} |\sigma(\alpha)|.$$

Pour ces deux normes, les deux problèmes d'optimisation s'écrivent sous les formes suivantes.

Modèle avec la norme $\|\cdot\|_2$

En tenant compte des contraintes et en utilisant la définition de la réponse impulsionnelle, nous écrivons le problème sous la forme :

$$\begin{array}{ll} \text{minimiser}_{(a,u)\in\mathbb{R}^n\times\mathbb{R}^n} & \dfrac{1}{(\alpha_{max} - \alpha_{min})^2} \displaystyle\int_{\alpha_{min}}^{\alpha_{max}} \left[\frac{2}{\pi\alpha} J_1(\pi\alpha)\right]^2 \left| \sum_{k=1}^{n} a_k e^{-i\frac{2\pi u_k}{d}\alpha} \right|^2 \ \mathrm{d}\alpha \\ \text{sous contraintes} & u_{k+1} - u_k \geq d, \ \text{ pour } \ k = 1, \ldots, n-1, \\ & \displaystyle\sum_{k=1}^{n} a_k = 1, \\ & a_k \geq 0, \ \text{ pour } \ k = 1, \ldots, n. \end{array} \tag{1.9}$$

Il s'agit d'un problème d'optimisation non linéaire avec des contraintes convexes.

Modèle avec la norme $\|\cdot\|_\infty$

En tenant compte des contraintes et en utilisant la définition de la réponse impulsionnelle, nous écrivons le problème sous la forme :

$$\begin{array}{ll}
\text{minimiser}_{(a,u)\in\mathbb{R}^n\times\mathbb{R}^n} & \max\left\{\left|\dfrac{2}{\pi\alpha}J_1(\pi\alpha)\displaystyle\sum_{k=1}^{n}a_k e^{-i\frac{2\pi u_k}{d}\alpha}\right| : \alpha\in[\alpha_{min},\alpha_{max}]\right\}\\
\text{sous contraintes} & u_{k+1}-u_k\geq d, \quad \text{pour} \quad k=1,\dots,n-1,\\
& \displaystyle\sum_{k=1}^{n}a_k=1,\\
& a_k\geq 0, \quad \text{pour} \quad k=1,\dots,n.
\end{array}$$

Ce problème peut être reformulé de la façon suivante :

$$\begin{array}{lll}
\text{minimiser}_{(a,u,t)\in\mathbb{R}^n\times\mathbb{R}^n\times\mathbb{R}} & t & \\
\text{sous contraintes} & \left|\dfrac{2}{\pi\alpha}J_1(\pi\alpha)\displaystyle\sum_{k=1}^{n}a_k e^{-i\frac{2\pi u_k}{d}\alpha}\right|\leq t, & \text{pour tout } \alpha\in[\alpha_{\min},\alpha_{\max}],\\
& u_{k+1}-u_k\geq d, & \text{pour } k=1,\dots,n-1,\\
& \displaystyle\sum_{k=1}^{n}a_k=1, & \\
& a_k\geq 0, & \text{pour } k=1,\dots,n,
\end{array} \tag{1.10}$$

où t est un palier donnant la valeur maximale de la dynamique sur l'intervalle d'optimisation $[\alpha_{\min},\alpha_{\max}]$. Il s'agit d'un problème d'optimisation non linéaire semi-infini [47]. Ce problème a donc un nombre fini de variables et un nombre infini de contraintes.

Chapitre 2

Résultats numériques

Dans ce chapitre, nous présentons les résultats numériques obtenus avec les deux stratégies d'optimisation décrites dans le chapitre précédent. Pour pouvoir comparer ces résultats, des critères qualitatifs de la réponse impulsionnelle normalisée sont définis dans la section 2.1. Dans la partie 2.2, les résultats trouvés à l'aide de la méthode des coefficients de Fourier sont décrits pour un nombre impair de télescopes. Les résultats obtenus grâce à la méthode de l'optimisation de la dynamique sont ensuite analysés dans le paragraphe 2.3 pour un nombre pair de télescopes. Dans la section 2.4, la sensibilité et la robustesse des configurations optimales sont étudiées en introduisant des perturbations dans le modèle. Enfin, la configuration optimale de l'instrument est caractérisée dans la partie 2.5.

2.1 Critères qualitatifs de la réponse impulsionnelle

Afin de qualifier quantitativement la réponse impulsionnelle normalisée et de comparer les résultats obtenus, cinq critères ont été définis.

- *La résolution du dispositif imageur* est définie comme la largeur à mi-hauteur du lobe principal de la réponse impulsionnelle normalisée dans le plan image. Elle est notée ρ, où $\rho \in [0,1]$ est telle que

$$\psi_n(\rho) := \min\left\{\alpha \in [0,1] : \psi_n(\alpha) = \frac{1}{2}\right\}. \tag{2.1}$$

- *Le champ de vue utile* est l'intervalle d'optimisation. Il est aussi appelé intervalle spatial libre ou champ propre (CLean Field of view). Il est noté $CLF = [\alpha_{\min}, \alpha_{\max}]$. Le lobe principal de la réponse impulsionnelle normalisée associée à l'exoplanète doit être observable dans cet intervalle.
- *Le nombre de résels* est défini comme le nombre de plus petits éléments résolus sur le CLF. Il est noté R avec

$$R := \frac{\alpha_{\max} - \alpha_{\min}}{2\rho}.$$

– *La dynamique en intensité* est définie comme le rapport entre la hauteur du lobe principal et la hauteur maximale des lobes secondaires de la réponse impulsionnelle normalisée sur le CLF, soit

$$D := \frac{1}{\max\{\psi_n(\alpha) : \alpha_{\min} \leq \alpha \leq \alpha_{\max}\}}.$$

– *Le flux* est défini comme la quantité relative d'énergie comprise dans le lobe principal de la réponse impulsionnelle normalisée. Il est noté F avec

$$F := \frac{\displaystyle\int_0^{\alpha_{\min}} \psi_n(\alpha) \ \mathrm{d}\alpha}{\displaystyle\int_0^{+\infty} \psi_n(\alpha) \ \mathrm{d}\alpha}.$$

La figure 2.1 représente l'ensemble des critères définis précédemment.

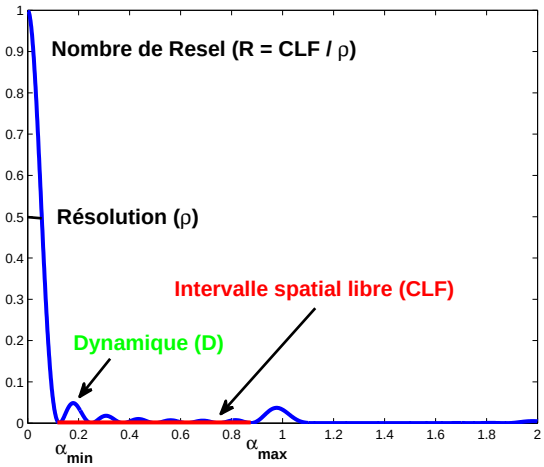


FIG. 2.1 – Critères qualitatifs de la PSF normalisée.

L'objectif est de trouver les amplitudes des champs et les positions des pupilles optimales pour que les valeurs de la dynamique D, du flux F et du nombre de résels R soient suffisamment élevées et que la valeur de la résolution ρ soit la plus petite possible.

2.2 Résultats obtenus avec la méthode des coefficients de Fourier

Pour présenter les résultats numériques obtenus avec la première stratégie d'optimisation, un instrument présentant un nombre impair de télescopes ($n = 2m + 1$) alignés et répartis périodiquement, a été considéré. Les variables d'optimisation sont les amplitudes des champs a_k ($k = 0, \ldots, m$) et l'écart ℓ entre chaque télescope. Les résultats obtenus sont comparés à l'aide de deux critères, à savoir la résolution ρ

et la dynamique D. La résolution est calculée grâce à la formule (2.1). La hauteur du premier lobe secondaire de la PSF normalisée est utilisée pour le calcul de la dynamique.

Rappelons que la PSF normalisée avec cette méthode s'écrit sous la forme :

$$\psi_{2m+1}(\alpha) = \psi(\alpha) \left| \hat{f}_m(\alpha) \right|^2 ,$$

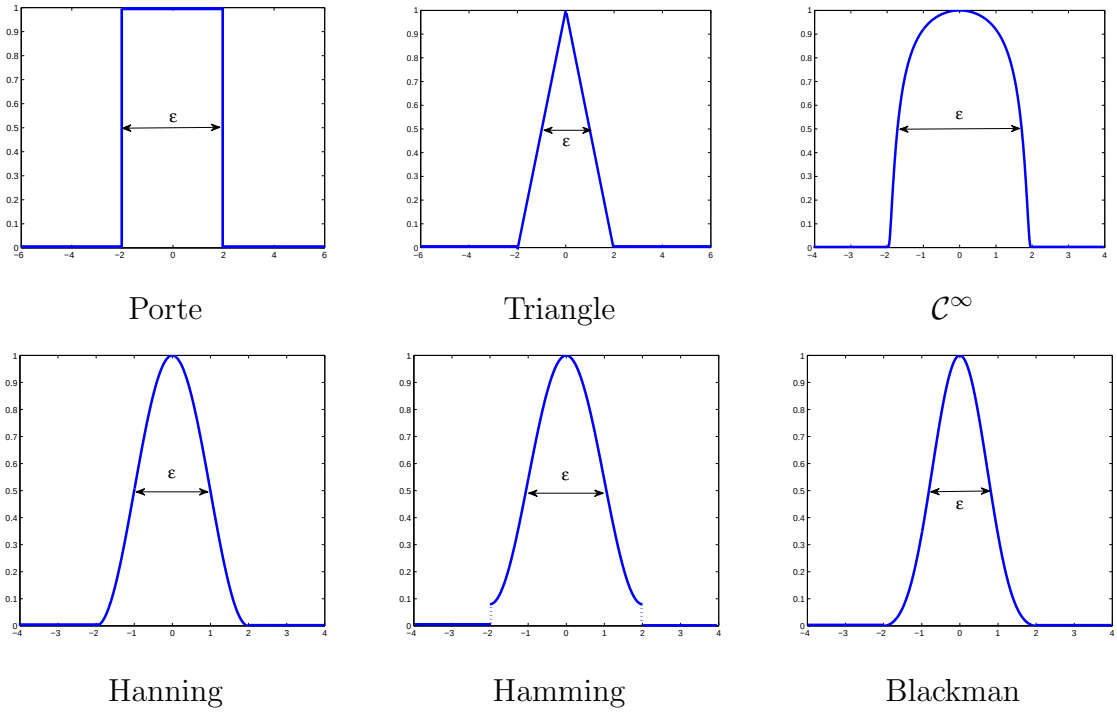
où $\hat{f}_m$ est la série de Fourier tronquée à l'ordre m d'une fonction d'apodisation f normalisée. Il existe plusieurs fonctions d'apodisation. Le tableau 2.1 et la figure 2.2 donnent des exemples de ces fonctions et de leurs représentations. La valeur de ε définit la largeur à mi-hauteur du gabarit se répétant avec une période $T := \dfrac{d}{\ell}$.

Porte	$f(t) = \begin{cases} 1 & \text{si } \lvert t \rvert \leq \dfrac{\varepsilon}{2}, \\ 0 & \text{sinon.} \end{cases}$
Triangle	$f(t) = \begin{cases} -\dfrac{2}{\varepsilon}t + 1 & \text{si } 0 < t < \dfrac{\varepsilon}{2}, \\ \dfrac{2}{\varepsilon}t + 1 & \text{si } -\dfrac{\varepsilon}{2} < t < 0, \\ 0 & \text{sinon.} \end{cases}$
$\mathcal{C}^\infty$	$f(t) = \begin{cases} \exp\left(\dfrac{1}{\varepsilon^2} + \dfrac{1}{t^2 - \varepsilon^2}\right) & \text{si } \lvert t \rvert \leq \varepsilon, \\ 0 & \text{sinon.} \end{cases}$
Hanning	$f(t) = \begin{cases} \cos^2\left(\dfrac{\pi t}{2\varepsilon}\right) & \text{si } \lvert t \rvert \leq \varepsilon, \\ 0 & \text{sinon.} \end{cases}$
Hamming	$f(t) = \begin{cases} 0.54 + 0.46 \cos\left(\dfrac{\pi t}{\varepsilon}\right) & \text{si } \lvert t \rvert \leq \varepsilon, \\ 0 & \text{sinon.} \end{cases}$
Blackman	$f(t) = \begin{cases} 0.42 + 0.50 \cos\left(\dfrac{\pi t}{\varepsilon}\right) + 0.08 \cos\left(\dfrac{2\pi t}{\varepsilon}\right) & \text{si } \lvert t \rvert \leq \varepsilon, \\ 0 & \text{sinon.} \end{cases}$

TAB. 2.1 – Fonctions d'apodisation.

Le tableau 2.2 donne l'expression des coefficients de Fourier de chacune des fonctions d'apodisation. Les coefficients de la fonction de classe $\mathcal{C}^\infty$ sont calculés par intégration numérique. Les amplitudes optimales des champs sont déterminées à partir de ces coefficients.

Pour chaque fonction gabarit, une courbe bi-critère a été tracée donnant la variation de la résolution ρ en fonction de la dynamique D pour différentes valeurs de ε comprises dans un intervalle noté $[\varepsilon_{min}, \varepsilon_{max}]$. La valeur de ε_{max} est choisie de telle sorte que les valeurs des amplitudes des champs soient positives. L'intervalle est discrétisé en un nombre fini q de valeurs de ε et la borne inférieure ε_{min} est



Porte Triangle $\mathcal{C}^\infty$

Hanning Hamming Blackman

FIG. 2.2 – Représentation des fonctions d'apodisation.

définie par $\varepsilon_{min} := \dfrac{\varepsilon_{max}}{q}$. L'objectif est de déterminer sur ces courbes une solution de compromis entre les deux critères D et ρ.

Deux configurations concernant la répartition des télescopes ont été supposées pour effectuer les tests.

- Configuration 1 : les pupilles sont bord à bord, i.e. $T = 1$.
- Configuration 2 : l'écart entre les pupilles est non nul, i.e. $T < 1$.

Le but de l'étude numérique est de comparer les deux configurations afin d'obtenir les meilleures valeurs des critères. Nous ne présentons pas ici une étude détaillée des différents tests effectués car des résultats plus précis sont dans [14].

La figure 2.3 représente l'évolution de la courbe bi-critère (D, ρ) pour neuf pupilles $(n = 9)$ lorsque la fonction gabarit est une fonction porte pour les deux configurations : $T = 1$ et $T = \dfrac{2}{3}$. Dans cet exemple, cinquante points de discrétisation $(q = 50)$ ont été considérés avec $\varepsilon_{max} = \dfrac{T}{4}$. Le fait de diminuer la période T permet d'améliorer la résolution ρ sans dégrader la valeur de la dynamique D. Ceci semble logique puisque l'écart entre les deux ouvertures les plus éloignées de l'hypertélescope augmente et assure donc un pouvoir de résolution élevé à l'instrument. La figure 2.4 représente la réponse impulsionnelle normalisée obtenue avec la fonction porte pour les deux périodes précédentes avec une même valeur de dynamique $(D = 500)$. La résolution est meilleure lorsque les pupilles sont écartées avec une période T petite. Les mêmes tests ont été effectués pour les fonctions gabarits du tableau 2.1. Des résultats identiques ont été obtenus.

Gabarits	$c_0(f)$	$c_k(f)$, pour $k \geq 1$
Porte	$\frac{\varepsilon}{T}$	$\frac{1}{k\pi}\sin(\frac{k\pi\varepsilon}{T})$
Triangle	$\frac{\varepsilon}{2T}$	$\frac{T(1-\cos(\frac{k\pi\varepsilon}{T}))}{\varepsilon(\pi k)^2}$
De Hanning	$\frac{\varepsilon}{T}$	$\frac{\pi^2\sin(\omega k\varepsilon)}{\omega kT(\pi^2-(\omega k\varepsilon)^2)}$
De Hamming	$\frac{1.08\varepsilon}{T}$	$\frac{(1.08\pi^2-0.16(\omega k\varepsilon)^2)\sin(\omega k\varepsilon)}{\omega kT(\pi^2-(\omega k\varepsilon)^2)}$
De Blackman	$\frac{0.84\varepsilon}{T}$	$\frac{-3.36\pi^4+0.36(\omega k\varepsilon\pi)^2}{\omega kT(\pi^2-(\omega k\varepsilon)^2)((\omega k\varepsilon)^2-4\pi^2)}$

TAB. 2.2 – Coefficients de Fourier des fonctions d'apodisation.

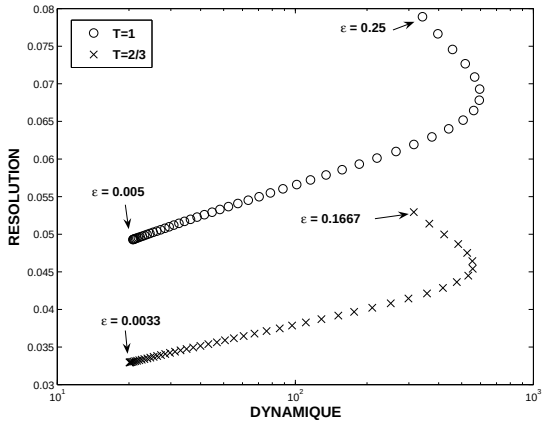
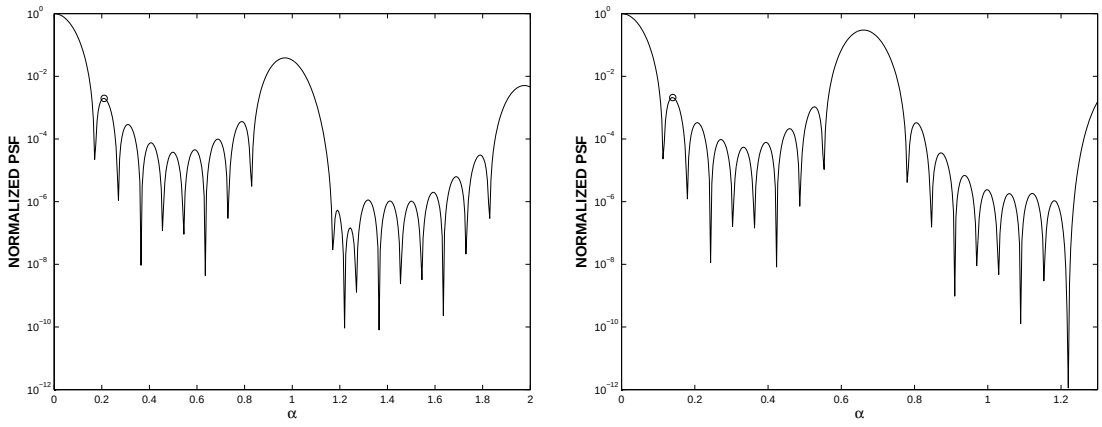


FIG. 2.3 – Courbes bi-critères (D, ρ) obtenues avec la fonction porte pour $T = 1$ et $T = 2/3$.

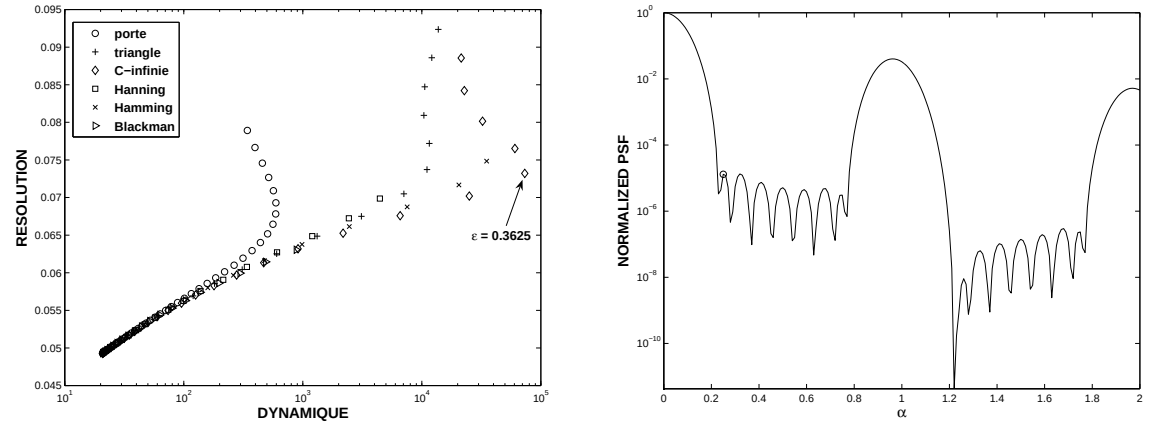
L'inconvénient de cette méthode réside dans la difficulté à choisir une fonction gabarit afin d'obtenir des valeurs de critères optimales. La figure 2.5 (a) permet de comparer les courbes bi-critères (D, ρ) obtenues pour chaque fonction d'apodisation du tableau 2.1 lorsque neuf pupilles bord à bord $(n = 9, T = 1)$ sont utilisées. Les meilleurs résultats en terme de dynamique D sont obtenus pour la fonction de classe $\mathcal{C}^\infty$ avec une largeur de gabarit ε évaluée à 0.3625. La figure 2.5 (b) représente la PSF normalisée optimale obtenue avec ce gabarit pour cette valeur de ε. Les valeurs de la résolution ρ et de la dynamique D sont correctes. Des tests identiques ont été effectués avec d'autres valeurs de période. Dans tous les cas, la fonction de classe $\mathcal{C}^\infty$ est le meilleur gabarit.

Les tableaux de l'annexe A donnent les valeurs des amplitudes optimales des champs pour chaque fonction d'apodisation en fonction de ε dans l'exemple où neuf pupilles bord à bord $(n = 9, T = 1)$ sont utilisées. La configuration étant symétrique, seules les amplitudes des champs des pupilles qui se trouvent à droite de l'origine $(k = 0, \ldots, 4)$ sont données. Nous remarquons que les amplitudes des champs sont apodisées (i.e. différentes pour chacune des pupilles) et que plus la valeur de ε est



a : $T = 1$, $\rho = 0.0652$ et $D = 500$ b : $T = \frac{2}{3}$, $\rho = 0.0436$ et $D = 500$

FIG. 2.4 – PSF normalisée optimale obtenue avec la fonction porte pour $T = 1$ (a) et $T = 2/3$ (b) avec $D = 500$.



(a) : Comparaison des courbes bi-critères (b) : $\varepsilon = 0.3625$, $\rho = 0.0741$ et $D = 76923$

FIG. 2.5 – Courbes bi-critères (D, ρ) pour toutes les fonctions d'apodisation et PSF normalisée optimale obtenue avec la fonction de classe $\mathcal{C}^\infty$ pour $n = 9$ et $T = 1$.

grande, plus l'apodisation des pupilles est importante.

Selon cette méthode, un bon compromis entre les deux critères D et ρ est obtenu lorsque les positions des télescopes sont écartées et lorsque les amplitudes des champs sont apodisées. En revanche, celle-ci présente deux inconvénients.

- Elle dépend indirectement de la fonction gabarit choisie, ce qui a une influence sur l'optimalité de la solution. Il est donc difficile de faire un tel choix à priori.
- Elle impose un positionnement périodique des lentilles qui ne conduit pas obligatoirement à la meilleure solution.

Pour ces raisons, une deuxième approche a été envisagée afin de minimiser la hauteur des lobes secondaires sur le CLF en optimisant les amplitudes des champs et/ou les positions des pupilles.

2.3 Résultats obtenus avec la méthode de l'optimisation de la dynamique

Reprenons le modèle de la dynamique écrit avec les normes $\|\cdot\|_2$ et $\|\cdot\|_\infty$ du chapitre 1. Pour la résolution des problèmes (1.9) et (1.10), un nombre pair de télescopes ($n = 2m$) est supposé avec une configuration symétrique ($a_{-k} = a_k$ et $u_{-k} = -u_k$, pour $k = 1, \ldots, m$). L'intervalle d'optimisation CLF est discrétisé en un nombre fini q de valeurs de α. Afin de comparer les résultats obtenus, tous les critères de la partie 2.1 ont été considérés. Tous les modèles présentés dans cette section ont été écrits avec le langage de modélisation AMPL [27] (voir annexe B). La valeur de la fonction de Bessel [1, 3] a été obtenue avec la bibliothèque GSL (GNU Scientific Library). La visualisation graphique des résultats est réalisée avec le logiciel MATLAB. Pour la présentation des résultats, nous considérons l'exemple où l'instrument est composé de huit pupilles ($m = 4$) de même diamètre ($d = 1$). L'intervalle d'optimisation est $CLF = [0.25, 0.75]$. Il est discrétisé en 80 points ($q = 80$). Les résultats présentés dans cette section ont fait l'objet d'un article [9] (voir annexe C).

2.3.1 Optimisation des amplitudes des champs

Les positions des pupilles $u_1, \ldots, u_m$ sont d'abord fixées à priori afin d'optimiser les amplitudes des champs $a_1, \ldots, a_m$.

Modèle avec la norme $\|\cdot\|_2$

Après la discrétisation de l'intervalle spatial libre, le problème (1.9) devient :

$$\begin{array}{ll} \text{minimiser}_{a\in\mathbb{R}^m} & \displaystyle\sum_{j=1}^{q}\left(\frac{4}{\pi\alpha_j}J_1(\pi\alpha_j)\sum_{k=1}^{m}a_k\cos\left(\frac{2\pi}{d}u_k\alpha_j\right)\right)^2 \\ \text{sous contraintes} & \displaystyle\sum_{k=1}^{m}a_k = \frac{1}{2}, \\ & a_k \geq 0, \quad \text{pour } k = 1, \ldots, m. \end{array}$$

Il s'agit d'un problème d'optimisation quadratique convexe avec des contraintes linéaires en a_k pour tout k. La solution calculée est alors un optimum global du problème. Le solveur LOQO, code de points intérieurs de R.J. Vanderbei [65], a été utilisé pour résoudre le problème avec une tolérance de convergence égale à 10^{-6}.

Dans l'exemple, les pupilles sont fixées bord à bord. Comme point de départ, les amplitudes des champs sont toutes égales à un. Le tableau 2.3 donne les valeurs des amplitudes optimales des champs trouvées par le solveur après 7 itérations et les positions des quatre pupilles à droite de l'origine. Les valeurs des amplitudes optimales sont apodisées.

La figure 2.6 représente les positions des pupilles à droite de l'origine avec les valeurs des amplitudes optimales du signal reçu par chacune d'elles et la réponse impulsionnelle normalisée correspondante. Le tableau 2.4 donne l'ensemble des critères op-

k	a_k	u_k
1	1.00	0.5
2	0.72	1.5
3	0.35	2.5
4	0.09	3.5

TAB. 2.3 – Valeurs des amplitudes optimales des champs pour des positions de pupilles fixées.

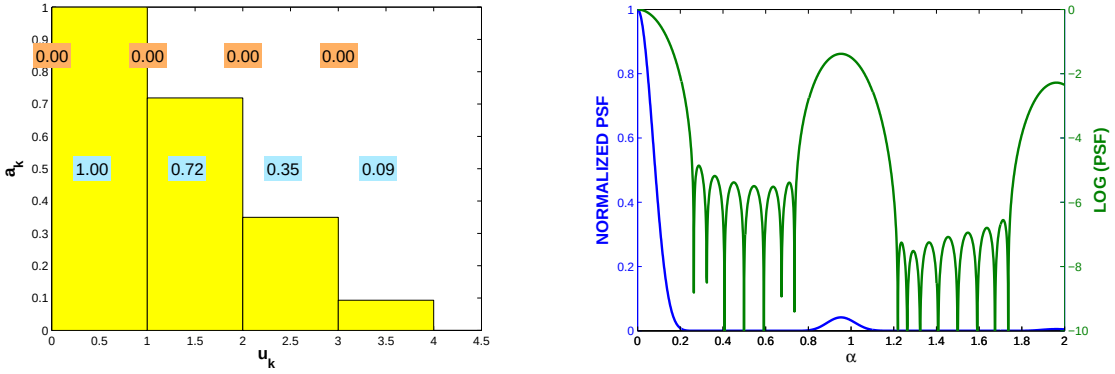


FIG. 2.6 – Configuration optimale des quatre pupilles ($m = 4$) à droite de l'origine. La valeur en orange correspond à l'écart entre chaque pupille et la valeur en bleu correspond à l'amplitude des champs reçue par chacune des pupilles. Représentation de la PSF normalisée optimale (échelle linéaire en bleu et échelle semi-logarithmique en vert) dans cette configuration pour l'optimisation des amplitudes des champs.

timaux de la PSF qui sont obtenus dans cette configuration. Les valeurs des différents critères sont correctes. La valeur de la résolution ρ est moins bonne que celle obtenue avec la technique des coefficients de Fourier. La valeur du flux F signifie que 92% de l'énergie est répartie dans le lobe principal de la réponse impulsionnelle, i.e. que seulement 8% de l'énergie est dispersée dans les lobes secondaires. Trois pixels sont visualisés dans le plan image.

Modèle avec la norme $\|\cdot\|_\infty$

Après la discrétisation de l'intervalle spatial libre, le problème (1.10) devient :

$$\begin{aligned} &\text{minimiser}_{(a,t)\in\mathbb{R}^m\times\mathbb{R}} \quad t \\ &\text{sous contraintes} \quad \left|\frac{2}{\pi\alpha_j}J_1(\pi\alpha_j)\sum_{k=1}^{m}a_k\cos\left(\frac{2\pi}{d}u_k\alpha_j\right)\right| \le t, \quad j=1,\dots,q, \\ &\qquad\qquad \sum_{k=1}^{m}a_k = \frac{1}{2}, \\ &\qquad\qquad a_k \ge 0, \quad \text{pour } k=1,\dots,m. \end{aligned} \tag{2.2}$$

D	18089
ρ	0.17
R	3.01
F	0.92

TAB. 2.4 – Valeurs des critères optimaux de la PSF pour l'optimisation des amplitudes des champs.

Il s'agit d'un problème de programmation linéaire car la fonction objectif et l'ensemble des contraintes sont linéaires en a_k pour tout k. La solution calculée est un optimum global du problème. Le solveur d'optimisation LOQO [65] a été utilisé pour la résolution du problème avec une tolérance de convergence égale à 10^{-6}.

Dans l'exemple, les pupilles sont fixées bord à bord. Comme point de départ, les amplitudes des champs sont toutes égales à un. Le tableau 2.5 donne les valeurs des amplitudes optimales trouvées par le solveur après 15 itérations et les positions des quatre pupilles à droite de l'origine. Les valeurs des amplitudes optimales des champs sont aussi apodisées.

k	a_k	u_k
1	1.00	0.5
2	0.73	1.5
3	0.37	2.5
4	0.11	3.5

TAB. 2.5 – Valeurs des amplitudes optimales des champs pour des positions de pupilles fixées.

La figure 2.7 représente les positions des pupilles avec les amplitudes optimales du signal reçu par chacune d'elles et la réponse impulsionnelle normalisée correspondante. Le tableau 2.6 donne l'ensemble des critères optimaux de la PSF qui sont obtenus dans cette configuration.

D	75090
ρ	0.16
R	3.05
F	0.92

TAB. 2.6 – Valeurs des critères optimaux de la PSF pour l'optimisation des amplitudes des champs.

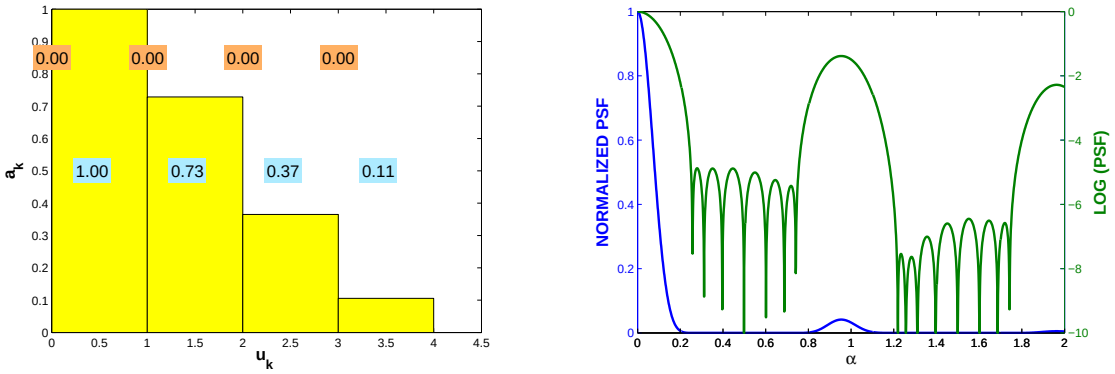


FIG. 2.7 – Configuration optimale des quatre pupilles ($m = 4$) à droite de l'origine et représentation de la PSF normalisée optimale dans cette configuration pour l'optimisation des amplitudes des champs.

Comparaison des deux normes

La solution optimale correspond dans les deux cas à l'apodisation des amplitudes des champs. Les valeurs des amplitudes optimales obtenues avec les deux normes ne sont pas identiques. Le solveur utilise plus d'itérations pour la résolution du problème écrit avec la norme $\|\cdot\|_\infty$. Concernant les valeurs des critères optimaux qui qualifient la réponse impulsionnelle, une meilleure valeur de dynamique D est obtenue avec la norme $\|\cdot\|_\infty$. Les valeurs des autres critères sont très proches pour les deux normes. La valeur de la dynamique optimale est insuffisante pour détecter des exoplanètes.

2.3.2 Optimisation des positions des pupilles

À présent, les amplitudes des champs $a_1, \ldots, a_m$ sont fixées pour optimiser uniquement les positions des pupilles $u_1, \ldots, u_m$. En comparaison à l'optimisation des amplitudes, une étude sur l'optimisation du positionnement des pupilles a été effectuée. Les valeurs des critères obtenues sont alors comparées pour les deux méthodes d'optimisation.

Modèle avec la norme $\|\cdot\|_2$

Après la discrétisation de l'intervalle spatial libre, le problème (1.9) devient :

$$\begin{aligned} \text{minimiser}_{u\in\mathbb{R}^m} \quad & \sum_{j=1}^{q}\left(\frac{4}{\pi\alpha_j}J_1(\pi\alpha_j)\sum_{k=1}^{m}a_k\cos\left(\frac{2\pi}{d}u_k\alpha_j\right)\right)^2 \\ \text{sous contraintes} \quad & u_1 \geq \frac{d}{2} \\ & u_{k+1} - u_k \geq d, \quad \text{pour } k = 1, \ldots, m-1. \end{aligned}$$

Ce problème d'optimisation est non linéaire. L'optimum calculé est une solution locale du problème qui n'est pas obligatoirement globale. La solution optimale trouvée par le solveur dépend essentiellement des conditions initiales fixées par l'utilisateur. Comme point de départ, les pupilles sont positionnées bord à bord. Le solveur

d'optimisation non linéaire IPOPT [67] d'A. Wächter a été utilisé pour résoudre le problème avec une tolérance de convergence égale à 10^{-8}.

Dans l'exemple, les valeurs des amplitudes des champs sont égales à un. Le tableau 2.7 donne les positions optimales des pupilles trouvées par le solveur après 8 itérations et l'écart entre deux pupilles successives. Les trois premières pupilles sont bord à bord mais la dernière est écartée.

k	u_k	$u_k - u_{k-1}$
1	0.51	1.00
2	1.51	1.00
3	2.51	1.00
4	3.52	1.01

TAB. 2.7 – Valeurs des positions optimales des pupilles pour des amplitudes de champs fixées. La configuration étant symétrique, la position u_{-1} est égale à l'opposé de la position u_1.

La figure 2.8 représente les positions optimales des pupilles avec les amplitudes du signal reçu par chacune d'elles et la réponse impulsionnelle normalisée correspondante.

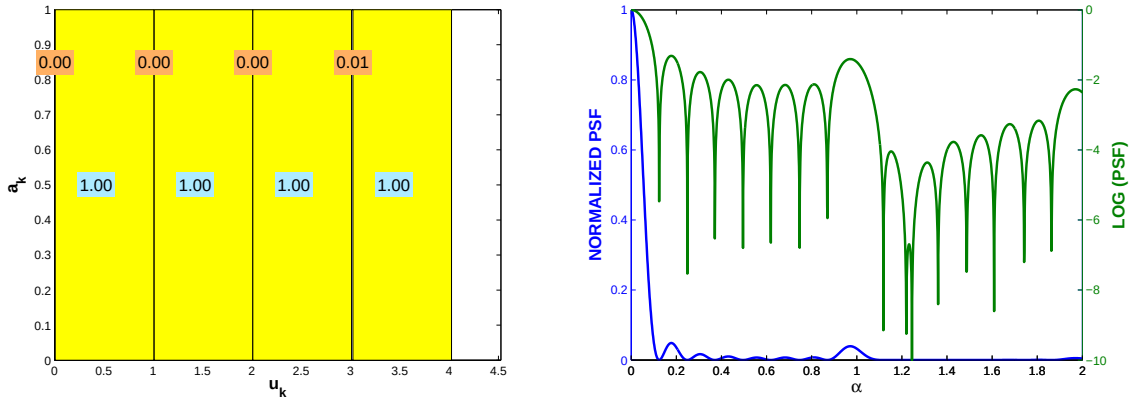


FIG. 2.8 – Configuration optimale des quatre pupilles ($m = 4$) à droite de l'origine et représentation de la PSF normalisée optimale dans cette configuration pour l'optimisation des positions des pupilles.

Le tableau 2.8 donne l'ensemble des critères optimaux de la PSF. Des améliorations des valeurs de la résolution ρ et du nombre de résels R sont obtenues au détriment des deux autres critères, à savoir la dynamique D et le flux F. En effet, la valeur de la dynamique D est très mauvaise et la dispersion d'énergie est ici plus importante dans les lobes secondaires que lors de l'optimisation des amplitudes des champs.

D	59.59
ρ	0.11
R	4.53
F	0.88

TAB. 2.8 – Valeurs des critères optimaux de la PSF pour l'optimisation des positions des pupilles.

Modèle avec la norme $\|\cdot\|_\infty$

Après la discrétisation de l'intervalle spatial libre, le problème (1.10) devient :

$$\begin{array}{ll} \text{minimiser}_{(u,t)\in\mathbb{R}^m\times\mathbb{R}} & t \\ \text{sous contraintes} & \left|\dfrac{2}{\pi\alpha_j}J_1(\pi\alpha_j)\displaystyle\sum_{k=1}^{m}a_k\cos\left(\frac{2\pi}{d}u_k\alpha_j\right)\right| \leq t, \quad j=1,\ldots,q, \\ & u_1 \geq \dfrac{d}{2} \\ & u_{k+1}-u_k \geq d, \ \text{ pour } k=1,\ldots,m-1. \end{array}$$

Ce problème d'optimisation est également non linéaire. L'optimum calculé est une solution locale du problème qui n'est pas nécessairement globale. Comme point de départ, les pupilles sont supposées bord à bord. Le solveur IPOPT [67] a été utilisé pour résoudre le problème avec une tolérance de convergence égale à 10^{-8}.

Dans l'exemple, les valeurs des amplitudes des champs sont égales à un. Le tableau 2.9 reporte les positions optimales des pupilles trouvées par le solveur après 16 itérations et l'écart entre deux pupilles successives. Les pupilles sont écartées et l'écart entre deux pupilles successives est différent.

k	u_k	$u_k - u_{k-1}$
1	0.55	1.10
2	1.55	1.00
3	2.55	1.00
4	3.60	1.05

TAB. 2.9 – Valeurs des positions des pupilles optimales pour des amplitudes de champs fixées. La configuration étant symétrique, la position u_{-1} est égale à l'opposé de la position u_1.

La figure 2.9 représente les positions optimales des pupilles avec les amplitudes du signal reçu par chacune d'elles et la réponse impulsionnelle correspondante. Le tableau 2.10 donne l'ensemble des critères optimaux de la PSF qui sont obtenus dans cette configuration.

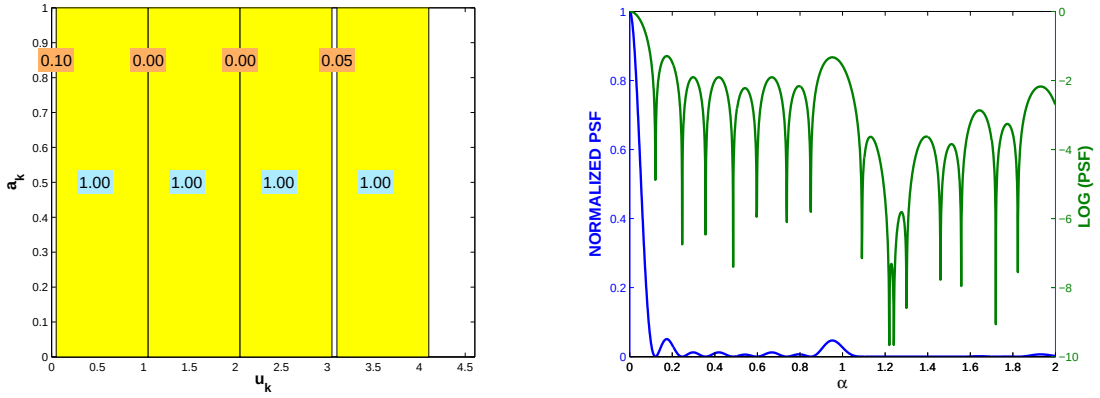


FIG. 2.9 – Configuration optimale des quatre pupilles ($m = 4$) à droite de l'origine et représentation de la PSF normalisée optimale dans cette configuration pour l'optimisation des positions des pupilles.

D	80.31
ρ	0.10
R	4.63
F	0.86

TAB. 2.10 – Valeurs des critères optimaux de la PSF pour l'optimisation des positions des pupilles.

Comparaison des deux normes

Pour les deux normes, la solution optimale correspond à celle où les pupilles sont écartées les unes des autres. Concernant les valeurs des critères, les meilleurs résultats en terme de dynamique D, résolution ρ et nombre de résels R sont obtenus avec la norme $\|\cdot\|_\infty$. En revanche, la valeur du flux F est meilleure avec la norme $\|\cdot\|_2$. La valeur de la dynamique D est loin d'être suffisante pour arriver à détecter des exoplanètes dans les deux cas.

2.3.3 Conclusions sur les deux optimisations

Ces deux approches mettent en évidence la nécessité d'optimiser simultanément les deux paramètres, à savoir les amplitudes des champs et les positions des pupilles pour obtenir le meilleur compromis possible entre les différents critères qualifiant la réponse impulsionnelle. En effet, l'écart entre les pupilles améliore les valeurs de ρ et de R au détriment des deux autres critères D et F. L'apodisation des pupilles assure de meilleures valeurs de D et de F. Les résultats obtenus dépendent de la norme considérée et sont meilleurs avec la norme $\|\cdot\|_\infty$.

2.3.4 Optimisation des amplitudes des champs et des positions des pupilles

À présent, les deux variables sont optimisées simultanément. L'exemple considéré jusque-là est présenté avec la norme $\|\cdot\|_2$. Une analyse plus précise des résultats est effectuée avec la norme $\|\cdot\|_\infty$ car cette norme assure les meilleurs résultats.

Modèle avec la norme $\|\cdot\|_2$

Nous écrivons le problème (1.9) sous la forme :

$$\begin{aligned}
\text{minimiser}_{(a,u)\in\mathbb{R}^m\times\mathbb{R}^m} \quad & \sum_{j=1}^{q}\left(\frac{4}{\pi\alpha_j}J_1(\pi\alpha_j)\sum_{k=1}^{m}a_k\cos\left(\frac{2\pi}{d}u_k\alpha_j\right)\right)^2 \\
\text{sous contraintes} \quad & u_1 \geq \frac{d}{2} \\
& u_{k+1}-u_k \geq d, \quad \text{pour } k=1,\dots,m-1, \\
& \sum_{k=1}^{m}a_k = \frac{1}{2}, \\
& a_k \geq 0, \quad \text{pour } k=1,\dots,m.
\end{aligned}$$

Il s'agit d'un problème d'optimisation non linéaire. La solution optimale trouvée par le solveur est un optimum local du problème qui n'est pas nécessairement global. Comme point de départ, les amplitudes des champs sont égales à un et les pupilles sont fixées bord à bord. Le solveur d'optimisation IPOPT [67] utilisé pour résoudre le problème a une tolérance de convergence égale à 10^{-8}.

Le tableau 2.11 reporte les amplitudes des champs et les positions optimales des pupilles trouvées par le solveur après 11 itérations. Les amplitudes optimales des champs sont apodisées et les positions optimales des pupilles sont réparties de manière quasi-périodique puisque les trois premières pupilles sont bord à bord et la dernière est écartée.

k	a_k	u_k	$u_k - u_{k-1}$
1	1.00	0.50	1.00
2	0.72	1.50	1.00
3	0.35	2.50	1.00
4	0.09	3.51	1.01

TAB. 2.11 – Valeurs des amplitudes des champs et des positions des pupilles optimales pour l'optimisation simultanée des deux variables. La configuration étant symétrique, la position u_{-1} est égale à l'opposé de la position u_1.

Le tableau 2.12 donne l'ensemble des critères optimaux de la PSF obtenus dans cette configuration. L'optimisation simultanée des deux variables améliore l'ensemble des critères. La figure 2.10 représente les positions optimales des pupilles avec les amplitudes du signal reçu par chacune d'elles et la réponse impulsionnelle correspondante.

D	19693
ρ	0.17
R	3.01
F	0.92

TAB. 2.12 – Valeurs des critères optimaux de la PSF pour l'optimisation des amplitudes des champs et des positions des pupilles.

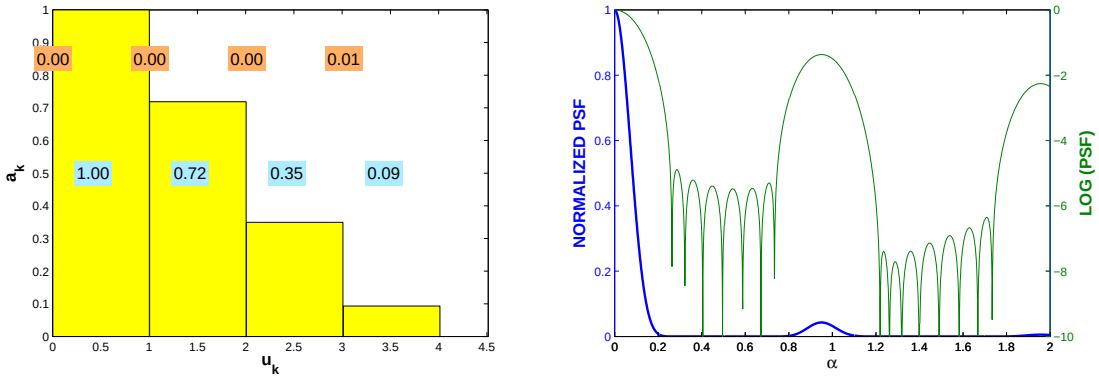


FIG. 2.10 – Configuration optimale des quatre pupilles ($m = 4$) à droite de l'origine et représentation de la PSF normalisée optimale dans cette configuration pour l'optimisation des amplitudes des champs et des positions des pupilles.

Modèle avec la norme $\|\cdot\|_\infty$

Nous écrivons le problème (1.10) sous la forme :

$$
\begin{aligned}
&\text{minimiser}_{(a,u,t)\in\mathbb{R}^m\times\mathbb{R}^m\times\mathbb{R}} \quad t \\
&\text{sous contraintes} \quad \left|\frac{2}{\pi\alpha_j}J_1(\pi\alpha_j)\sum_{k=1}^{m}a_k\cos\left(\frac{2\pi}{d}u_k\alpha_j\right)\right| \le t, \quad j=1,\dots,q, \\
&\qquad\qquad u_1 \ge \frac{d}{2} \\
&\qquad\qquad u_{k+1}-u_k \ge d, \quad \text{pour } k=1,\dots,m-1, \\
&\qquad\qquad \sum_{k=1}^{m}a_k = \frac{1}{2}, \\
&\qquad\qquad a_k \ge 0, \quad \text{pour } k=1,\dots,m.
\end{aligned}
\tag{2.3}
$$

Le problème d'optimisation est non linéaire et sa résolution est effectuée à l'aide du solveur d'optimisation non linéaire IPOPT [67] avec une tolérance de convergence égale à 10^{-10}. L'inconvénient du problème réside dans le fait que la solution calculée est locale. Ainsi, l'optimum obtenu n'est pas nécessairement global. La solution optimale trouvée par le solveur dépendant essentiellement des conditions initiales fixées par l'utilisateur, une stratégie a été mise en place pour définir les valeurs initiales des amplitudes des champs et des positions des pupilles.

Stratégie du point de départ

La démarche a consisté à réaliser un grand nombre de simulations en supposant des positions d'ouvertures aléatoires comme conditions initiales.
À titre d'exemple, considérons un hypertélescope avec huit pupilles ($m = 4$) de même diamètre ($d = 1$) pour optimiser la valeur de la dynamique D de la réponse impulsionnelle sur l'intervalle d'optimisation $CLF = [0.25, 0.75]$. Ce problème a été résolu 10^4 fois. La condition initiale sur le positionnement des pupilles est choisie par un tirage aléatoire de probabilité uniforme avec les deux contraintes suivantes :

- $\frac{7}{2}d \leq u_4 \leq 14d$,
- $u_{k+1} - u_k \geq d$ pour $k = 1, \ldots, m-1$,

pour éviter que les pupilles se superposent.

Les valeurs de la dynamique D et du flux F sont calculées pour chaque optimisation afin de comparer les solutions optimales trouvées. La figure 2.11 montre l'évolution de F en fonction de D pour les 10^4 procédures d'optimisation effectuées avec les conditions initiales précédentes. La solution optimale correspond au point où les valeurs des deux critères D et F sont les plus grandes. Les autres solutions, représentées par le nuage de points, peuvent être considérées comme des optima locaux du problème.

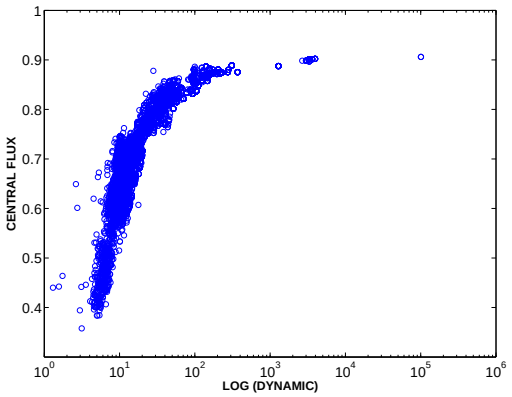


FIG. 2.11 – Comparaison des 10^4 solutions du problème (2.3) en termes de flux (F) et de dynamique (D) pour différents points de départ.

La figure 2.12 représente la configuration optimale des pupilles et la réponse impulsionnelle correspondant à la solution trouvée par le processus d'optimisation. Le tableau 2.13 reporte les valeurs des amplitudes a_k et des positions u_k pour $k = 1, \ldots, 4$ correspondant à cette solution optimale. La solution du problème qui paraît globale impose une quasi-périodicité dans le positionnement des ouvertures dans le plan pupille ainsi qu'une apodisation des amplitudes des champs collectés.

Un grand nombre d'expériences identiques a été réalisé en faisant varier le nombre n de pupilles et l'intervalle d'optimisation $[\alpha_{min}, \alpha_{max}]$. Pour chacune de ces expériences, la solution optimale au problème de la dynamique correspond à un positionnement quasi-périodique des ouvertures et à une apodisation des amplitudes des champs.

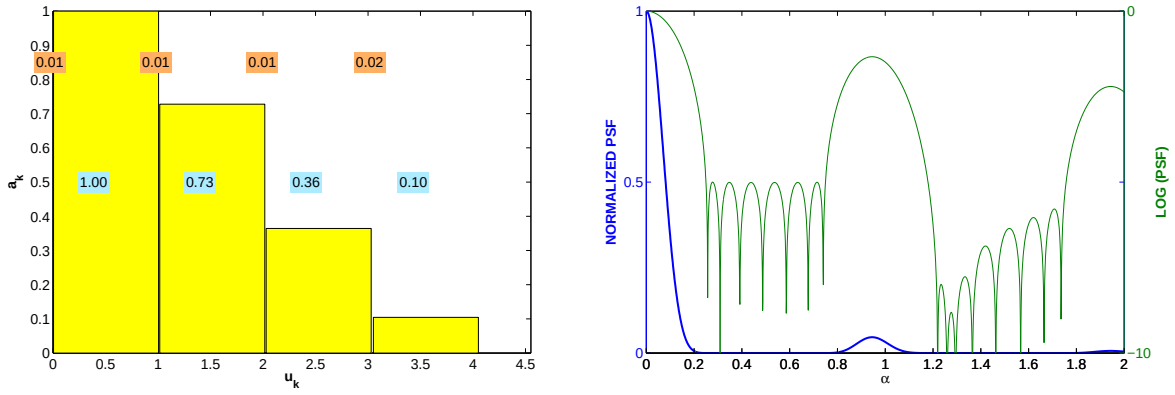


FIG. 2.12 – Configuration optimale des pupilles et représentation de la PSF normalisée pour la solution optimale trouvée par les 10^4 procédures d'optimisation.

k	a_k	u_k	$u_k - u_{k-1}$
1	1.00	0.50	1.01
2	0.73	1.51	1.01
3	0.36	2.52	1.01
4	0.10	3.55	1.02

TAB. 2.13 – Amplitudes des champs et positions des pupilles pour la meilleure solution des 10^4 procédures d'optimisation. La configuration étant symétrique, la position u_{-1} est égale à l'opposé de la position u_1.

Partant de ce constat et afin de limiter le temps de calcul de résolution du problème (2.3), la recherche de la solution globale s'effectue en trois étapes.

- Les positions des ouvertures dans le plan de la pupille densifiée sont calculées en fonction de l'intervalle spatial libre CLF pour que la valeur de la dynamique D soit maximale. Ces positions sont calculées avec la formule heuristique suivante :
$$u_k = \left(k - \frac{1}{2}\right) \frac{d}{\alpha_{\min} + \alpha_{\max}}, \quad k = 1, \ldots, m.$$
Ce choix impose à l'intervalle spatial libre CLF d'être centré entre deux entiers consécutifs afin de fixer la périodicité des positions des ouvertures.
- La deuxième étape consiste à résoudre le problème d'optimisation (2.2). Les positions des pupilles sont fixées avec la relation précédente et seules les amplitudes des champs sont optimisées. La solution obtenue est un optimum global du problème car le problème d'optimisation est linéaire.
- La dernière étape consiste à résoudre le problème (2.3) en supposant comme conditions initiales, les positions des pupilles et les amplitudes des champs déterminées dans les deux étapes précédentes.

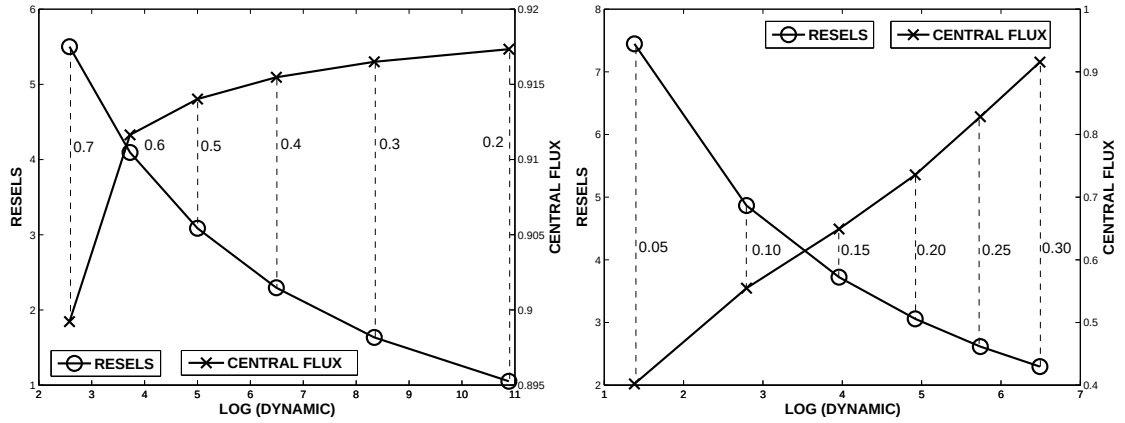
En suivant cette démarche, la solution optimale trouvée correspond systématiquement à celle obtenue avec le processus statistique de résolution du problème d'op-

timisation. En pratique, cette stratégie est très efficace. En effet, le gain en temps de calcul est considérable puisque le solveur IPOPT [67] calcule la solution après une dizaine d'itérations et la solution trouvée semble être le minimum global du problème.

Analyse des résultats

Le choix du point de départ étant fait, une analyse des résultats numériques obtenus est présentée. L'influence de l'intervalle d'optimisation CLF sur les valeurs des critères qualifiant la réponse impulsionnelle en terme de dynamique D, nombre de résels R et flux F est étudiée.

Décrivons les résultats numériques obtenus lorsque l'hypertélescope a huit pupilles ($m = 4$) de même diamètre ($d = 1$). La figure 2.13 (a) présente l'évolution des valeurs de la dynamique D, du nombre de résels R et du flux F pour différentes valeurs de $\Delta\alpha$ (où $\Delta\alpha := \alpha_{max} - \alpha_{min}$) lorsque l'intervalle est centré autour de $\alpha_{moy} = 0.50$. Les valeurs numériques des différents critères sont données dans le tableau 2.14 (a).



a b

FIG. 2.13 – Valeurs de D, F et R pour différents CLF en fonction de $\Delta\alpha$ (a) et de α_{min} (pour $\Delta\alpha = 0.40$) (b).

$\alpha_{\min}$	$\alpha_{\max}$	$\Delta\alpha$	D	R	F
0.40	0.60	0.2	7.7e10	1.1	0.917
0.35	0.65	0.3	2.2e8	1.6	0.916
0.30	0.70	0.4	3.1e6	2.3	0.915
0.25	0.75	0.5	1.0e5	3.1	0.914
0.20	0.80	0.6	5.3e3	4.1	0.912
0.15	0.85	0.7	3.8e2	5.5	0.899

a

$\alpha_{\min}$	$\alpha_{\max}$	D	R	F
0.05	0.45	2.4e1	7.4	0.402
0.10	0.50	6.2e2	4.9	0.555
0.15	0.55	9.1e3	3.7	0.649
0.20	0.60	8.3e4	3.1	0.736
0.25	0.65	5.5e5	2.6	0.828
0.30	0.70	3.1e6	2.3	0.915

b

TAB. 2.14 – Valeurs de D, R et F en fonction de $\Delta\alpha$ (a) et de $\alpha_{\min}$ (pour $\Delta\alpha = 0.40$) (b).

Plus le CLF est étroit, meilleure est la valeur de la dynamique D puisque celle-ci peut atteindre une valeur de 10^{10}. Cependant, le gain en dynamique D se fait au détriment du nombre de résels R qui est moins bon pour un CLF étroit. La valeur du flux F reste quasi-constante puisqu'elle varie peu ($F = 0.90$).

La figure 2.14 représente les pupilles ainsi que la réponse impulsionnelle optimale obtenue pour les deux cas extrêmes $\Delta\alpha = 0.20$ et $\Delta\alpha = 0.70$. Pour $\Delta\alpha = 0.20$, les positions des pupilles sont périodiques et pour $\Delta\alpha = 0.70$, elles sont quasi-périodiques. Dans les deux cas, les amplitudes des champs sont apodisées et l'apodisation est d'autant plus importante que l'intervalle d'optimisation est étroit.

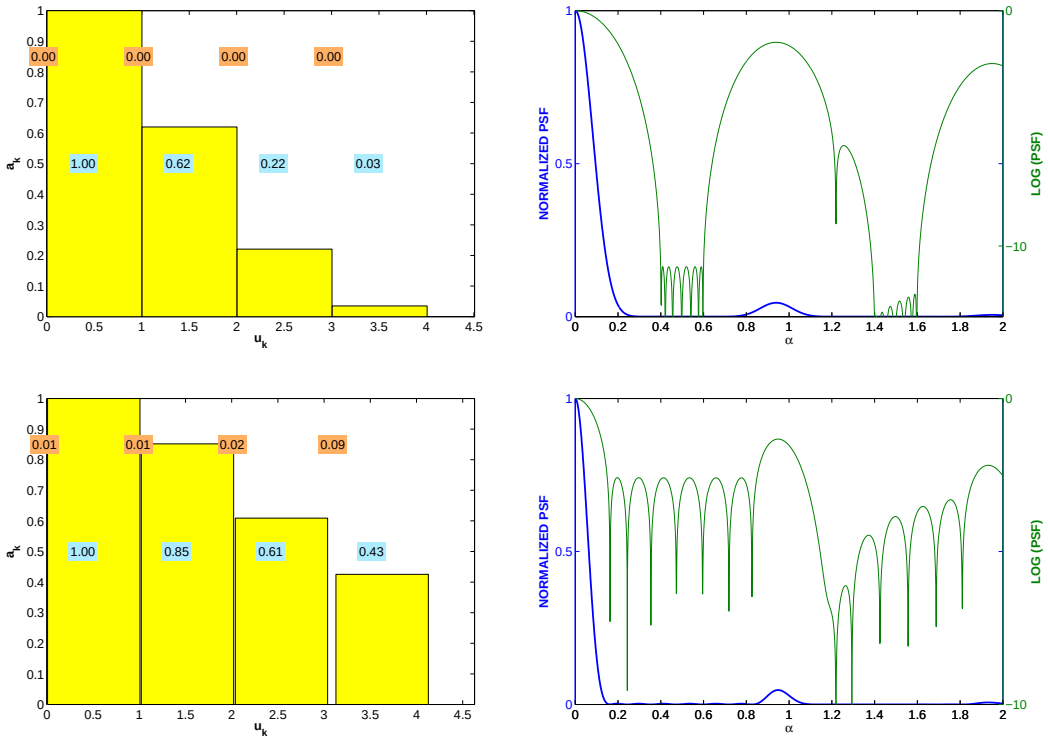


FIG. 2.14 – Configuration optimale des pupilles et représentation de la PSF normalisée optimale pour $\Delta\alpha = 0.20$ (haut) et pour $\Delta\alpha = 0.70$ (bas) avec $\alpha_{moy} = 0.50$.

La figure 2.13 (b) représente les valeurs des critères pour un CLF de largeur constante ($\Delta\alpha = 0.40$) où seulement α_{min} évolue. Le tableau 2.14 (b) donne l'évolution des trois critères en fonction de α_{min}. Plus la borne α_{min} est proche du lobe central de la PSF ($\alpha = 0$), plus il est difficile d'obtenir une grande valeur de dynamique D. De plus, la valeur du flux F est d'autant plus faible que la valeur de α_{min} est petite. L'énergie est donc dispersée dans les lobes secondaires. En revanche, le phénomène inverse se produit pour le nombre de résels R car il est d'autant plus grand que la valeur de α_{min} est petite. La figure 2.15 représente les pupilles ainsi que la réponse impulsionnelle optimale obtenue pour les deux cas extrêmes à savoir $\alpha_{min} = 0.05$ (haut) et $\alpha_{min} = 0.30$ (bas). Le positionnement des pupilles est quasi-périodique pour une valeur de α_{min} petite et il est périodique pour une valeur de α_{min} plus grande. Dans les deux cas, les amplitudes des champs sont apodisées. Toutefois, la solution trouvée pour $\alpha_{min} = 0.05$ ne semble pas correcte.

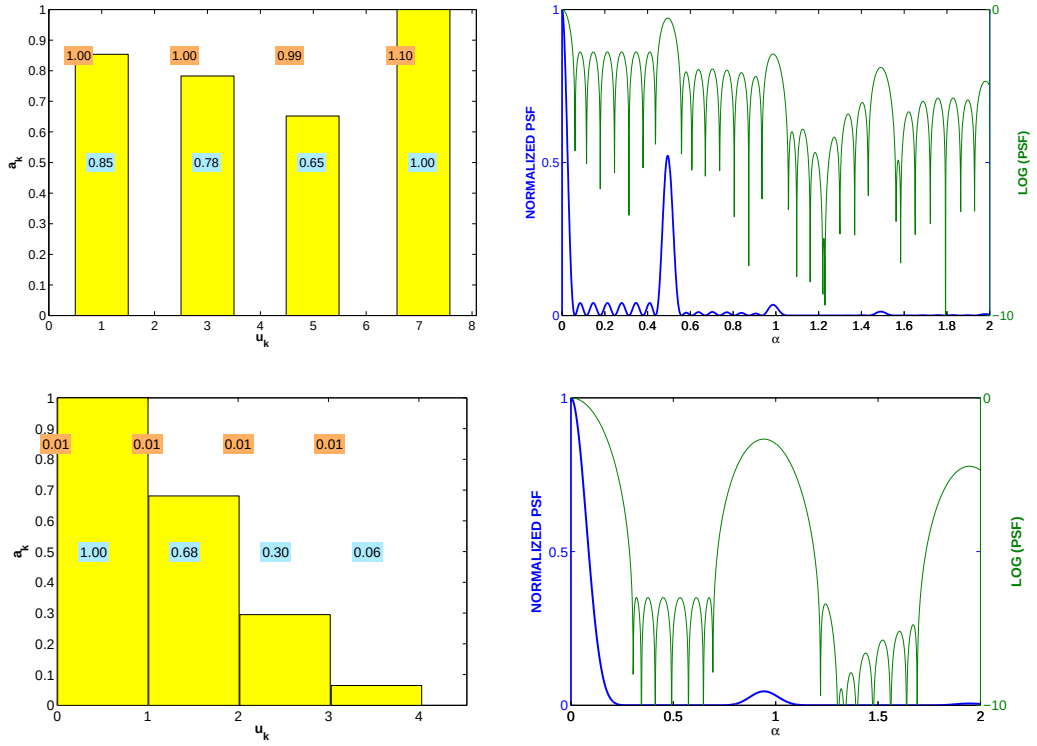


FIG. 2.15 – Configuration optimale des pupilles et représentation de la PSF normalisée optimale pour $\alpha_{min} = 0.05$ (haut) et pour $\alpha_{min} = 0.30$ (bas) avec $\Delta\alpha = 0.40$.

Présentons les solutions optimales obtenues pour un nombre n de pupilles variant de 4 à 24 afin de mettre en évidence l'influence du nombre de pupilles sur les valeurs des critères. Le tableau 2.15 (a) donne les valeurs des trois critères D, R et F selon le nombre de pupilles pour l'intervalle d'optimisation fixé $CLF = [0.25, 0.75]$. Les valeurs des trois critères augmentent avec le nombre de pupilles. Cette augmentation est très significative en terme de dynamique.

n	D	R	F
4	9.1e1	2.1	0.8728
8	1.0e5	3.1	0.9140
16	1.3e11	4.4	0.9241
24	5.9e16	5.4	0.9286

a : $[\alpha_{\min}, \alpha_{\max}] = [0.25, 0.75]$

n	D	R	F	$[\alpha_{min}, \alpha_{max}]$
4	5.9e5	0.4	0.5623	[0.27,0.34]
8	9.1e5	2.5	0.8734	[0.27,0.68]
16	1.0e6	7.7	0.7317	[0.12,0.67]
24	5.9e5	13.5	0.8562	[0.09,0.83]

b : $D \simeq 10^6$

TAB. 2.15 – Valeurs de D, R et F en fonction du nombre de pupilles n avec $CLF = [0.25, 0.75]$ (a) et $D \simeq 10^6$ (b).

Le tableau 2.15 (b) donne l'évolution de R et de F en fonction du nombre d'ouvertures n pour une dynamique D fixée à 10^6. Dans tous les cas, l'intervalle d'optimisation est centré autour de $\alpha_{moy} = 0.50$. Plus le nombre de pupilles est important, meilleures sont les valeurs du nombre de résels R et du flux F pour une valeur de dynamique fixée.

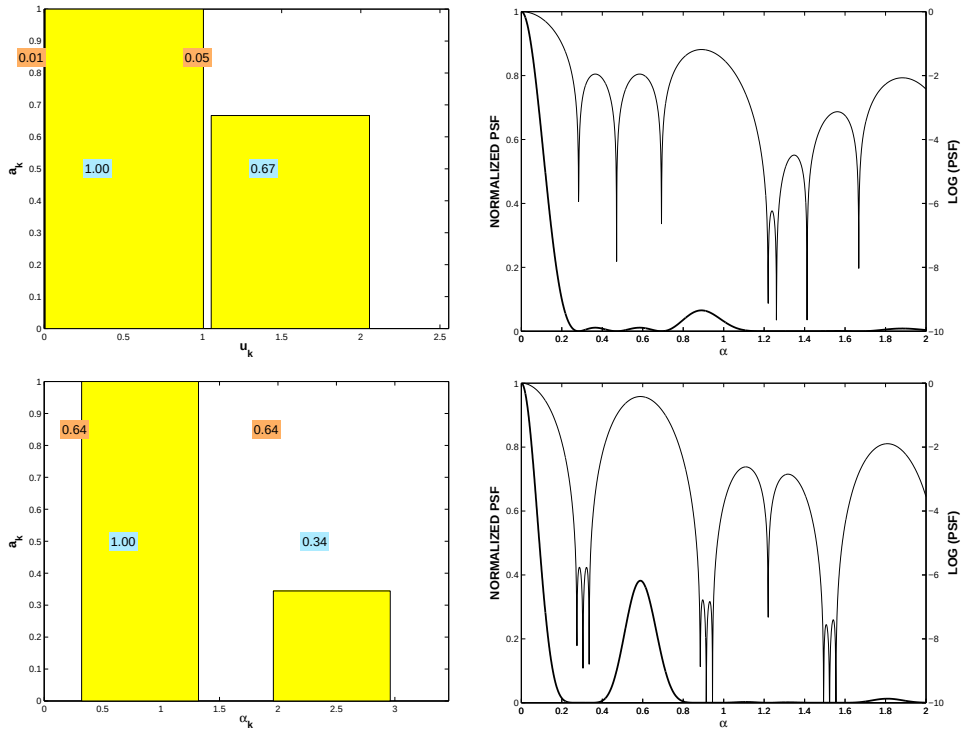


FIG. 2.16 – Configuration optimale des pupilles et représentation de la PSF normalisée optimale pour quatre pupilles ($m = 2$) avec $CLF = [0.25, 0.75]$ (haut) et $D = 10^6$ (bas).

Les figures 2.16 et 2.17 représentent les pupilles et la réponse impulsionnelle optimale obtenues pour $n = 4$ et $n = 24$. Ces expérimentations montrent que le nombre de pupilles a une influence directe sur les valeurs des différents critères. Un nombre restreint de pupilles permet tout de même d'obtenir une dynamique de l'ordre de 10^6. Une apodisation des amplitudes des champs et un positionnement périodique ou quasi-périodique des pupilles sont toujours obtenus comme solution optimale.

Cette approche met en évidence l'importance de considérer simultanément les variables d'optimisation, à savoir les amplitudes des champs et les positions des pupilles. Une valeur élevée de dynamique est obtenue avec une configuration quasi-périodique des positions des ouvertures et une apodisation des amplitudes des champs.

2.4 Étude de la sensibilité et de la robustesse de la configuration optimale

Dans cette partie, la sensibilité et la robustesse des configurations optimales trouvées à l'aide du modèle de l'optimisation de la dynamique écrit avec la norme $\|\cdot\|_\infty$ sont étudiées lorsque les amplitudes des champs et les positions des pupilles sont optimisées simultanément. L'objectif est de mettre en évidence l'effet de petites perturbations des amplitudes des champs et/ou des positions des pupilles sur les valeurs des trois critères à savoir la dynamique D, le nombre de résels R et le flux F. En effet, les tests numériques effectués précédemment supposent des conditions

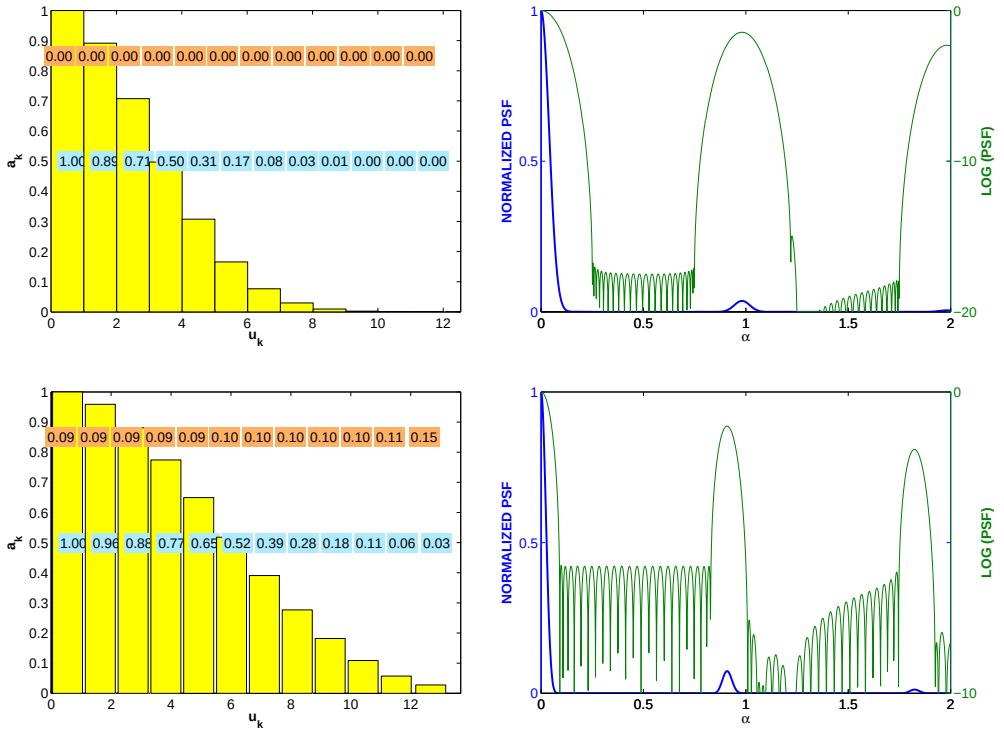


FIG. 2.17 – Configuration optimale des pupilles et représentation de la PSF normalisée optimale pour vingt-quatre pupilles ($m = 12$) avec $CLF = [0.25, 0.75]$ (haut) et $D = 10^6$ (bas).

parfaites pour réaliser les expérimentations mais dans la réalité des perturbations de nature thermique ou mécanique apparaissent au cours des expériences.

Reprenons l'exemple où l'hypertélescope présente huit ouvertures ($m = 4$) de même diamètre ($d = 1$) avec l'intervalle d'optimisation $CLF = [0.25, 0.75]$. La configuration optimale trouvée par le solveur (partie 2.3) correspond à des amplitudes de champs apodisées pour un positionnement quasi-périodique des pupilles. La réponse impulsionnelle optimale et les valeurs de D, R et F sont plus particulièrement étudiées lorsque les amplitudes des champs et les positions des pupilles sont perturbées. Les amplitudes et les positions perturbées sont respectivement notées a'_k et u'_k telles que pour $k = 1, \ldots, 4$,

$$a'_k := a_k(1 + t_k),$$

$$u'_k - u'_{k-1} := \max\{1, u_k - u_{k-1}(1 + t'_k)\},$$

où t_k et t'_k sont choisis par un tirage aléatoire de probabilité uniforme dans $[-\varrho, \varrho]$, avec a_k et u_k les valeurs des amplitudes et des positions optimales trouvées par le solveur (tableau 2.13). Une centaine d'expérimentations a été effectuée avec $\varrho = 10^{-2}$ et $\varrho = 10^{-3}$. Pour comparer les solutions trouvées, le pire des cas a été considéré en prenant les plus petites valeurs des trois critères D, R et F. Les résultats obtenus sont reportés dans le tableau 2.16. Les amplitudes des champs et/ou les positions des pupilles ont été perturbées pour chaque valeur de ϱ considérée. Seule la valeur de

ϱ	perturbation	D	R	F
10^{-3}	amplitudes	$8.0e4$	3.085	0.914
	positions	$7.0e4$	3.084	0.914
	amplitudes et positions	$6.6e4$	3.083	0.913
10^{-2}	amplitudes	$2.2e4$	3.074	0.914
	positions	$8.9e3$	3.069	0.908
	amplitudes et positions	$9.0e3$	3.068	0.908

TAB. 2.16 – Valeurs de D, R et F pour des perturbations d'amplitudes de champs et/ou de positions de pupilles pour huit pupilles ($m = 4$) avec $CLF = [0.25, 0.75]$.

la dynamique D est véritablement sensible à la perturbation. En effet, le nombre de résels R et la valeur du flux F ne varient pas ou varient de façon non significative. La valeur de la dynamique D varie de façon plus importante quand les positions des pupilles sont perturbées. Une perturbation de 1% sur les valeurs des positions implique une décroissance non négligeable de la dynamique d'un facteur 10. La figure 2.18 représente la réponse impulsionnelle lorsque les paramètres optimaux sont perturbés pour $\varrho = 10^{-3}$ et $\varrho = 10^{-2}$.

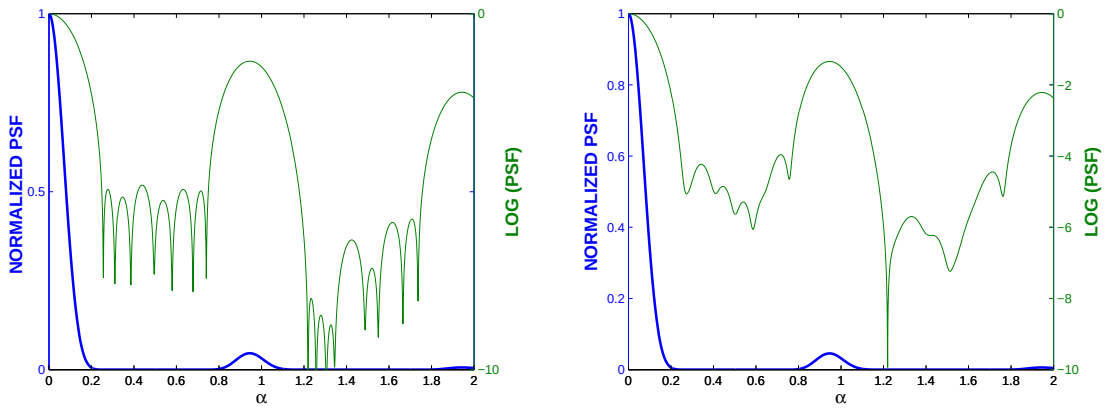


FIG. 2.18 – PSF normalisée obtenue avec les paramètres optimaux perturbés pour $\varrho = 10^{-3}$ (gauche) et $\varrho = 10^{-2}$ (droite).

La comparaison de ces courbes à la figure 2.12 représentant la réponse impulsionnelle sans perturbation montre que les critères sont moins bons lorque les paramètres optimisés sont perturbés.

2.5 Conclusions

La résolution du problème physique a été effectuée à l'aide de deux méthodes. La première approche, qui impose un positionnement périodique des pupilles, a montré que les amplitudes optimales des champs doivent être apodisées. Cette méthode a deux inconvénients. En effet, elle dépend indirectement de la fonction gabarit choisie

et elle impose un positionnement particulier des pupilles. Une deuxième approche a été envisagée. Elle consiste à optimiser la dynamique de la réponse impulsionnelle sur un certain intervalle. Cette méthode met en évidence la nécessité d'optimiser simultanément les deux variables à savoir les amplitudes des champs et les positions des pupilles pour obtenir les meilleures valeurs de critères. La configuration optimale de la pupille densifiée de l'hypertélescope est une répartition quasi-périodique des pupilles avec des amplitudes de champs apodisées. Pour tenir compte des perturbations qui ont lieu au cours des expérimentations, les valeurs des amplitudes des champs et/ou des positions des pupilles sont perturbées de manière à caractériser leurs influences sur les critères de la réponse impulsionnelle. Au final, seule la dynamique est modifiée.

Chapitre 3

Résultats théoriques

Après avoir résolu le problème de manière numérique, nous nous sommes intéressés à sa résolution théorique lorsque la configuration de l'instrument est temporelle. Celle-ci correspond au concept d'hypertélescope proposé par F. Reynaud et L. Delage [61]. Dans la section 3.1, le principe de l'hypertélescope temporel est présenté ainsi que sa modélisation mathématique écrite à l'aide des normes $\|\cdot\|_2$ et $\|\cdot\|_\infty$. Remarquons que l'étude numérique présentée dans le chapitre précédent pour la configuration spatiale de l'instrument, a aussi été réalisée lorsque la configuration de l'instrument est temporelle afin d'obtenir les valeurs des positions des pupilles et des amplitudes des champs optimales. Cette étude étant très similaire à celle présentée précédemment, nous ne l'avons pas développée dans ce manuscrit. Celle-ci a montré que la solution optimale semblait correspondre à un positionnement périodique des pupilles. C'est pourquoi dans ce chapitre, la plupart des résultats théoriques, trop complexes pour un positionnement quelconque des pupilles, se place dans le cadre du positionnement périodique de celles-ci. Dans les paragraphes 3.2, 3.3 et 3.4, nous nous intéressons principalement à l'existence d'une solution du problème. Pour la norme $\|\cdot\|_2$, le résultat principal se trouve dans la proposition 3.3.1. Cette proposition permet de ramener l'étude de l'existence d'une solution du problème à l'étude d'une fonction présentée dans la partie 3.3. Pour la norme $\|\cdot\|_\infty$, le résultat décrit dans le théorème 3.4.1 est inattendu et remarquable puisqu'une formule explicite de la solution est obtenue à l'aide des polynômes de Tchebytchev.

3.1 Position du problème avec les normes $\|\cdot\|_2$ et $\|\cdot\|_\infty$

En 2007, F. Reynaud et L. Delage [61] ont proposé un nouveau concept d'hypertélescope, nommé Temporal HyperTelescope (THT) réalisant un affichage temporel de l'image observée dans le plan image. La figure 3.1 est un exemple d'architecture possible pour une configuration temporelle d'un hypertélescope. L'instrument est composé d'une pupille d'entrée constituée d'un réseau de plusieurs télescopes, d'un système de recombinaison optique et d'un plan image dans lequel la réponse impulsionnelle est observée temporellement. La particularité de cet instrument est qu'il est composé de fibres optiques unimodales permettant le transport des champs

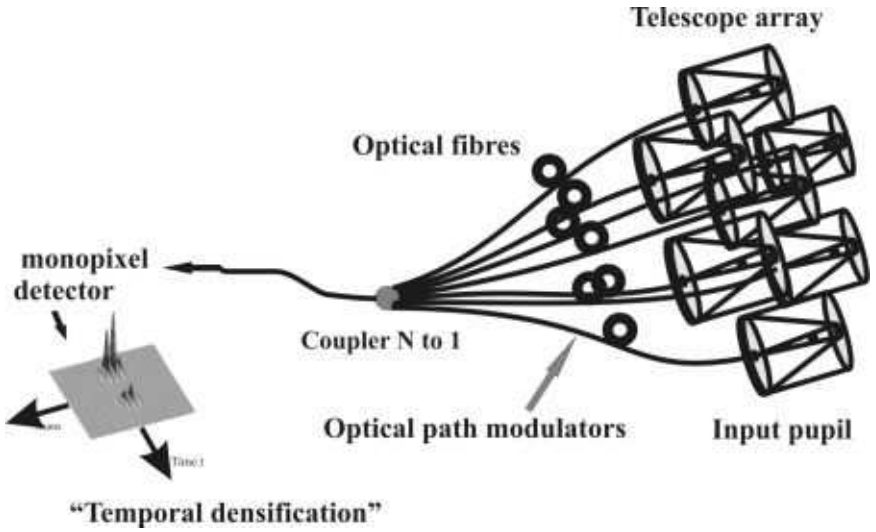


FIG. 3.1 – Hypertélescope temporel THT

optiques entre la pupille d'entrée et le système de recombinaison. La notion générale d'hypertélescope temporel n'impose pas nécessairement l'utilisation de ces fibres. Cependant, celles-ci ont l'avantage de simplifier la mise en œuvre expérimentale du banc test réalisé en laboratoire [53].

3.1.1 Réponse impulsionnelle temporelle

Dans le cas temporel, la réponse impulsionnelle normalisée à une dimension pour $n+1$ télescopes est réduite au module au carré de la fonction d'interférence. Le plan pupille n'existant plus dans la configuration temporelle de l'instrument, le terme correspondant à l'enveloppe de diffraction dans la définition de la réponse impulsionnelle associée à la configuration spatiale de l'instrument, n'apparait plus dans la définition de la réponse impulsionnelle temporelle. En utilisant les notations du chapitre 1, la réponse impulsionnelle temporelle normalisée pour $n+1$ télescopes alignés peut s'écrire :

$$t \mapsto \left| \sum_{k=0}^{n} a_k e^{itu_k} \right|^2, \tag{3.1}$$

où t est la variable temporelle d'observation. Comme dans le cas spatial, nous supposons

$$\sum_{k=0}^{n} a_k = 1. \tag{3.2}$$

La figure 3.2 est un exemple de réponse impulsionnelle temporelle normalisée obtenue pour huit pupilles alignées ($n = 8$) de même diamètre ($d = 1$) et disposées symétriquement autour de l'origine.

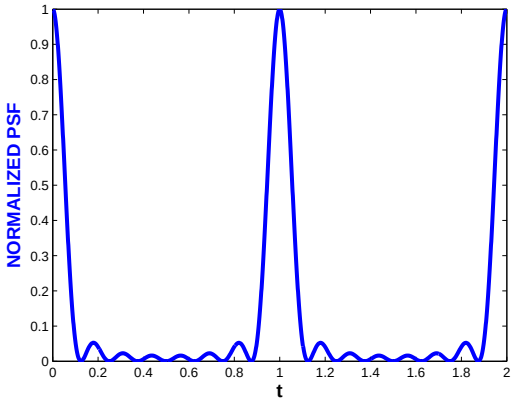


FIG. 3.2 – PSF temporelle normalisée pour huit pupilles alignées.

3.1.2 Modèles pour l'optimisation de la dynamique

Comme dans le cas spatial, l'objectif est que le graphe de la réponse impulsionnelle temporelle définie en (3.1) présente un lobe central le plus étroit possible pour avoir une grande résolution et des lobes secondaires les plus bas possibles sur un intervalle prédéfini pour avoir une grande valeur de dynamique. La technique de l'optimisation de la dynamique a été considérée pour la modélisation du problème. Les variables d'optimisation sont les amplitudes des champs $a_0, \ldots, a_n$ et les positions des pupilles $u_0, \ldots, u_n$. Pour les résultats théoriques, aucune contrainte n'est imposée sur les positions des pupilles. La contrainte (3.2) sur les amplitudes des champs est supposée. En revanche, nous n'imposons plus la contrainte de positivité des amplitudes. L'ensemble définissant les contraintes admissibles sur les amplitudes des champs est noté

$$\mathcal{A} := \{a \in \mathbb{R}^{n+1} : \sum_{k=0}^{n} a_k = 1\}.$$

La réponse impulsionnelle temporelle doit présenter des lobes secondaires les plus bas possibles sur un intervalle noté $I = [t_{min}, t_{max}] \subset]0, +\infty[$ $(0 < t_{min} < t_{max})$. En utilisant les définitions des deux normes suivantes, pour $f : \mathbb{R} \to \mathbb{C}$,

$$\|f\|_2 := \left(\frac{1}{t_{max} - t_{min}} \int_I |f(t)|^2 \ \mathrm{d}t\right)^{\frac{1}{2}},$$

et

$$\|f\|_\infty := \max_{t \in I} |f(t)|,$$

le problème de l'optimisation de la dynamique s'écrit sous les formes suivantes.

Modèle avec la norme $\|\cdot\|_2$

$$\text{minimiser}_{(a,u)\in\mathcal{A}\times\mathbb{R}^{n+1}} \quad \frac{1}{2(t_{max} - t_{min})} \int_I \left|\sum_{k=0}^{n} a_k e^{itu_k}\right|^2 \ \mathrm{d}t \tag{3.3}$$

Modèle avec la norme $\|\cdot\|_\infty$

$$\text{minimiser}_{(a,u)\in\mathcal{A}\times\mathbb{R}^{n+1}} \quad \max_{t\in I}\left|\sum_{k=0}^{n} a_k e^{itu_k}\right| \tag{3.4}$$

Ces deux problèmes d'optimisation sont non linéaires. L'existence d'une solution de ces deux problèmes va être étudiée par la suite. Comme il l'a été justifié au début de ce chapitre, les problèmes (3.3) et (3.4) ont été résolus pour des positions de pupilles redondantes. Dans ce cas, les variables d'optimisation sont les amplitudes des champs et l'écart entre chaque pupille. L'existence d'une solution de ces deux problèmes a été analysée de deux manières différentes. Les conditions d'optimalité du problème (3.3) ont été écrites pour exprimer les positions optimales des pupilles et les amplitudes optimales des champs. Ce problème s'écrit alors de manière plus simple avec ces formules. La courbe de la fonction objectif du problème a été tracée et son graphe analysé. L'existence d'une solution n'a pas été prouvée mais les résultats obtenus sont tout de même présentés. L'existence d'une solution du problème (3.4) a été démontrée. Ce résultat fait intervenir les polynômes de Tchebytchev. Les résultats obtenus dans ces deux études sont présentés dans les sections suivantes.

3.2 Conditions d'optimalité du problème écrit avec la norme $\|\cdot\|_2$

Dans cette partie, les conditions d'optimalité du problème (3.3) sont écrites. Les variables d'optimisation sont les amplitudes des champs et/ou les positions des pupilles. L'objectif est d'exprimer les amplitudes des champs et les positions des pupilles optimales avec une formule explicite.

3.2.1 Notations

- L'application $g_u : I \to \mathbb{C}$ est définie, pour tout $t \in I$, par
$$g_u(t) := e^{itu}.$$
- Le produit scalaire $\langle\cdot,\cdot\rangle$ associé à la norme $\|\cdot\|_2$ est défini, pour $f, g : \mathbb{R} \to \mathbb{C}$, par
$$\langle f, g\rangle := \frac{1}{t_{max} - t_{min}} \int_I f(t)\overline{g(t)}\ \mathrm{d}t.$$
- Soit $e := (1, \ldots, 1)^\top$, le vecteur de taille $(n+1) \times 1$.
- L'ensemble des matrices à coefficients dans $\mathbb{R}$ (respectivement dans $\mathbb{C}$) de dimension $(n+1) \times (n+1)$ est noté $\mathcal{M}(\mathbb{R})$ (respectivement $\mathcal{M}(\mathbb{C})$).
- La partie réelle et la partie imaginaire d'un nombre $z \in \mathbb{C}$ sont notées $\Re(z)$ et $\Im(z)$.
- Si la matrice $M \in \mathcal{M}(\mathbb{C})$, alors nous avons $M := \Re(M) + i\Im(M)$ où $\Re(M)$ est la partie réelle de la matrice M et $\Im(M)$ sa partie imaginaire.

3.2.2 Conditions d'optimalité pour des positions de pupilles fixées

La fonction objectif du problème (3.3) peut s'écrire :

$$\frac{1}{2}\left\|\sum_{k=0}^{n} a_k g_{u_k}\right\|_2^2 = \frac{1}{2}a^\top G a, \tag{3.5}$$

où la matrice $G \in \mathcal{M}(\mathbb{C})$ est définie par $G_{kk'} := \langle g_{u_k}, g_{u_{k'}}\rangle$ pour $k, k' = 0, \ldots, n$.

En fixant les positions des pupilles $u_0, \ldots, u_n$ et en utilisant les notations précédentes, le problème (3.3) devient :

$$\text{minimiser}_{a\in\mathcal{A}} \quad \frac{1}{2}a^\top G a. \tag{3.6}$$

Comme G est une matrice hermitienne, pour tout vecteur $a \in \mathbb{R}^{n+1}$ nous avons

$$a^\top G a = a^\top H a, \tag{3.7}$$

où $H := \Re(G)$ est symétrique.

Le problème (3.6) peut ainsi se reformuler :

$$\text{minimiser}_{a\in\mathcal{A}} \quad \frac{1}{2}a^\top H a. \tag{3.8}$$

Ce problème est quadratique convexe car la fonction objectif est quadratique, son hessien H est semi-défini positif (d'après les égalités (3.5) et (3.7)) et la contrainte d'égalité est linéaire. Les conditions nécessaires d'optimalité du premier ordre sont donc suffisantes pour l'obtention d'un minimum global.

Proposition 3.2.1 *Le vecteur des amplitudes $a \in \mathbb{R}^{n+1}$ est solution du problème (3.8) si et seulement si il existe $\lambda \in \mathbb{R}$ tel que*

$$\begin{cases} Ha + \lambda e = 0 \\ e^\top a = 1. \end{cases} \tag{3.9}$$

Preuve. Soit $\mathcal{L} : \mathbb{R}^{n+1} \times \mathbb{R} \to \mathbb{R}$ le lagrangien associé au problème (3.8) défini par

$$\mathcal{L}(a, \lambda) := \frac{1}{2}a^\top H a + \lambda(e^\top a - 1),$$

où $\lambda \in \mathbb{R}$ est le vecteur de multiplicateurs de lagrange associé à la contrainte du problème. Les conditions d'optimalité nécessaires et suffisantes du problème sont

$$\begin{cases} \nabla_a \mathcal{L}(a, \lambda) = 0 \\ e^\top a - 1 = 0, \end{cases}$$

soient

$$\begin{cases} Ha + \lambda e = 0 \\ e^\top a = 1. \end{cases} \tag{3.10}$$

□

Lemme 3.2.2 *Pour des positions de pupilles $u_0, \ldots, u_n$ deux à deux distinctes, la matrice H est inversible.*

Preuve. Soit $a \in \mathbb{R}^{n+1}$. D'après (3.7), nous avons

$$a^\top H a = \left\| \sum_{k=0}^{n} a_k g_{u_k} \right\|_2^2 .$$

La matrice H est alors semi-définie positive. Comme les positions des pupilles sont deux à deux distinctes, la famille $\{g_{u_k} : k = 0, \ldots, n\}$ forme une base. Ainsi, $a^\top H a$ s'annule uniquement si $a \in \mathbb{R}^{n+1}$ est le vecteur nul. Donc, la matrice H est définie positive et le résultat est obtenu.

□

Proposition 3.2.3 *Si la matrice H est inversible, alors le vecteur des amplitudes, solution du problème (3.6) est unique et s'écrit :*

$$a = \frac{H^{-1}e}{e^\top H^{-1}e}. \tag{3.11}$$

Le vecteur des multiplicateurs de lagrange $\lambda \in \mathbb{R}$, est donné par les formules :

$$\lambda = -a^\top H a = -\frac{1}{e^\top H^{-1}e}. \tag{3.12}$$

Preuve. D'après la première équation de (3.9) et comme H est inversible, nous avons

$$a = -\lambda H^{-1}e, \tag{3.13}$$

En multipliant cette égalité par e et en utilisant la deuxième équation de (3.9), nous obtenons

$$\lambda = -\frac{1}{e^\top H^{-1}e}.$$

En remplaçant λ par son expression dans (3.13), nous obtenons la formule (3.11).

En multipliant la première équation du système (3.9) par a et en utilisant l'égalité $e^\top a = 1$, il en résulte

$$\lambda = -a^\top H a.$$

□

Proposition 3.2.4 *Si la matrice H est inversible, alors*

$$\min_{a \in \mathcal{A}} \quad \frac{1}{2} a^\top H a \ = \frac{1}{2e^\top H^{-1}e}.$$

Preuve. D'après (3.7) et (3.12), le résultat est obtenu.

□

Remarque 3.2.5 La valeur du minimum du problème (3.6) peut être interprétée comme la distance entre le vecteur nul de l'espace des fonctions continues de $\mathbb{R}$ dans $\mathbb{C}$ et l'espace affine engendré par la famille $\{g_{u_k} : k = 0, \ldots, n\}$ noté aff $(g_{u_0}, \ldots, g_{u_n})$. Cette distance est notée d$(0, \text{aff}\,(g_{u_0}, \ldots, g_{u_n}))$. C'est pourquoi, le problème (3.6) admet toujours une solution. Elle n'est pas unique lorsque deux positions de pupilles sont confondues. Lorsque les positions des pupilles $u_0, \ldots, u_n$ sont deux à deux distinctes, d'après la relation (3.7) et la proposition 3.2.4, nous avons

$$\mathrm{d}(0, \text{aff}\,(g_{u_0}, \ldots, g_{u_n})) = \frac{1}{e^\top H^{-1} e}. \tag{3.14}$$

3.2.3 Conditions d'optimalité pour des amplitudes de champs fixées

En fixant les valeurs des amplitudes des champs $a_0, \ldots, a_n$, le problème (3.3) devient un problème de minimisation sans contraintes de la forme :

$$\text{minimiser}_{u \in \mathbb{R}^{n+1}} \quad \frac{1}{2} \left\| \sum_{k=0}^{n} a_k g_{u_k} \right\|_2^2. \tag{3.15}$$

Proposition 3.2.6 *Si $u \in \mathbb{R}^{n+1}$ est solution du problème (3.15), alors les conditions d'optimalité du problème s'écrivent :*

$$W(u)a = 0, \tag{3.16}$$

où la matrice $W(u) \in \mathcal{M}(\mathbb{R})$ est définie par $W_{kk'}(u) := a_k \Im(\langle g_{u_{k'}}, h_{u_k} \rangle)$ pour $k, k' = 0, \ldots, n$ avec $h_{u_k}(t) := t g_{u_k}(t) = t e^{i t u_k}$.

Preuve. Notons $\chi : \mathbb{R}^{n+1} \to \mathbb{R}$, l'application définie pour tout $u \in \mathbb{R}^{n+1}$ par

$$\chi(u) := \frac{1}{2} \left\| \sum_{k=0}^{n} a_k g_{u_k} \right\|_2^2.$$

Les conditions du Théorème de Leibniz [50] étant vérifiées, nous avons

$$\begin{aligned}
\frac{\partial \chi}{\partial u_{k'}}(t) &= \frac{1}{2} i \int_I a_{k'} \left(\sum_{k=0}^{n} a_k \left(h_{u_{k'}}(t) \overline{g_{u_k}(t)} - \overline{h_{u_{k'}}(t)} g_{u_k}(t) \right) \right) \mathrm{d}t \\
&= -a_{k'} \sum_{k=0}^{n} a_k \int_I \frac{h_{u_{k'}}(t) \overline{g_{u_k}(t)} - \overline{h_{u_{k'}}(t)} g_{u_k}(t)}{2i} \mathrm{d}t \\
&= a_{k'} \sum_{k=0}^{n} a_k \Im(\langle g_{u_k}(t), h_{u_{k'}}(t) \rangle).
\end{aligned}$$

Le problème n'ayant pas de contrainte, les conditions d'optimalité pour $u \in \mathbb{R}^{n+1}$ s'écrivent :

$$\nabla \chi(u) = 0,$$

où $\nabla \chi(u) = W(u)a$.

□

3.2.4 Conditions d'optimalité pour des positions de pupilles et des amplitudes de champs non fixées

En utilisant les notations des sections précédentes, le problème d'optimisation non linéaire (3.3) s'écrit sous la forme :

$$\text{minimiser}_{(a,u)\in\mathcal{A}\times\mathbb{R}^{n+1}} \quad \frac{1}{2}a^\top H(u)a, \tag{3.17}$$

où la matrice $H(u) \in \mathcal{M}(\mathbb{R})$ est définie par $H(u) := \Re(G(u))$ avec $G_{kk'}(u) := \langle g_{u_k}, g_{u_{k'}}\rangle$ pour $k, k' = 0, \ldots, n$.

Proposition 3.2.7 *Si $(a, u) \in \mathbb{R}^{n+1} \times \mathbb{R}^{n+1}$ est solution du problème (3.17), alors il existe $\lambda \in \mathbb{R}$ tel que*

$$\begin{cases} H(u)a + \lambda e = 0 \\ W(u)a = 0 \\ e^\top a = 1. \end{cases} \tag{3.18}$$

Preuve. Soit $\mathcal{L} : \mathbb{R}^{n+1} \times \mathbb{R}^{n+1} \times \mathbb{R} \to \mathbb{R}$ le lagrangien associé au problème (3.17) défini par

$$\mathcal{L}(a, u, \lambda) := \frac{1}{2}a^\top H(u)a + \lambda(e^\top a - 1),$$

où $\lambda \in \mathbb{R}$ est le vecteur de multiplicateurs de lagrange associé à la contrainte du problème. D'après les preuves des propositions 3.2.1 et 3.2.6 pour $(a, u) \in \mathbb{R}^{n+1} \times \mathbb{R}^{n+1}$, les conditions d'optimalité du problème (3.17) s'écrivent :

$$\begin{cases} H(u)a + \lambda e = 0 \\ W(u)a = 0 \\ e^\top a = 1. \end{cases}$$

□

Lemme 3.2.8 *Pour des positions de pupilles $u_0, \ldots, u_n$ deux à deux distinctes, la matrice $H(u)$ est inversible.*

Preuve. La preuve est identique à celle du lemme 3.2.2 en remplaçant la matrice H par la matrice $H(u)$. □

Proposition 3.2.9 *Si $(a, u) \in \mathbb{R}^{n+1} \times \mathbb{R}^{n+1}$ est solution du problème (3.17) et si la matrice $H(u)$ est inversible, alors le vecteur des amplitudes optimales s'écrit :*

$$a = \frac{H(u)^{-1}e}{e^\top H(u)^{-1}e}. \tag{3.19}$$

Le vecteur des multiplicateurs de lagrange $\lambda \in \mathbb{R}$ est donné par les formules :

$$\lambda = -a^\top H(u)a = -\frac{1}{e^\top H(u)^{-1}e}. \tag{3.20}$$

Preuve. La preuve est identique à celle de la proposition 3.2.3 en remplaçant la matrice H par la matrice $H(u)$.

□

3.3 À propos de l'existence de la solution du problème écrit avec la norme $\|\cdot\|_2$

La proposition suivante montre que la variable $a \in \mathbb{R}^{n+1}$ peut être supprimée dans le problème (3.3).

Proposition 3.3.1 *En notant* $\mathcal{U} := \{u \in \mathbb{R}^{n+1} : u_0 < u_1 < \ldots < u_n\}$, *nous avons*

$$\inf_{(a,u)\in\mathcal{A}\times\mathbb{R}^{n+1}} \frac{1}{2}\left\|\sum_{k=0}^{n} a_k g_{u_k}\right\|_2^2 = \inf_{u\in\mathcal{U}} \frac{1}{2e^\top H(u)^{-1}e}.$$

Preuve. Notons $\zeta : \mathbb{R}^{n+1}\times\mathbb{R}^{n+1} \to \mathbb{R}$, l'application définie par

$$\zeta(a,u) := \frac{1}{2}\left\|\sum_{k=0}^{n} a_k g_{u_k}\right\|_2^2.$$

Notons V et W les ensembles définis par

$$V := \zeta(\mathcal{A}\times\mathcal{U}) \quad \text{et} \quad W := \zeta(\mathcal{A}\times\mathbb{R}^{n+1}).$$

Il est clair que $V \subset W$. Montrons que $V = W$. Soit $t \in W$, il existe $(a,u) \in \mathcal{A}\times\mathbb{R}^{n+1}$ tel que $t = \zeta(a,u)$. Il existe $v_0 < v_1 < \ldots < v_p$ tels que $\{v_0,\ldots,v_p\} = \{u_0,\ldots,u_n\}$. Des réels $v_{p+1},\ldots,v_n$ peuvent être rajoutés tels que $(v_0,\ldots,v_n) \in \mathcal{U}$. Nous pouvons écrire :

$$\sum_{i=0}^{n} a_i g_{u_i} = \sum_{i=0}^{p} b_i g_{v_i},$$

avec pour $i = 0,\ldots,p$, $b_i = \sum_{\{k:u_k=v_i\}} a_k$. En posant $b_i = 0$ pour $i > p$, l'égalité devient :

$$\sum_{i=0}^{n} a_i g_{u_i} = \sum_{i=0}^{n} b_i g_{v_i} = \zeta(b,v),$$

et

$$\sum_{i=0}^{n} b_i = \sum_{i=0}^{n} a_i = 1.$$

Donc $(b,v) \in \mathcal{A}\times\mathcal{U}$, ce qui implique que $t \in V$. L'égalité $V = W$ a été démontrée. Ainsi, $\inf V = \inf W$. Donc, grâce à la proposition 3.2.4,

$$\begin{aligned}\inf_{(a,u)\in\mathcal{A}\times\mathbb{R}^{n+1}} \zeta(a,u) &= \inf_{(a,u)\in\mathcal{A}\times\mathcal{U}} \zeta(a,u)\\ &= \inf_{u\in\mathcal{U}} \inf_{a\in\mathcal{A}} \zeta(a,u)\\ &= \inf_{u\in\mathcal{U}} \frac{1}{2e^\top H(u)^{-1}e}.\end{aligned}$$

□

Étudier le problème (3.3) revient donc à étudier le problème :

$$\text{minimiser}_{u\in\mathcal{U}} \frac{1}{2e^\top H(u)^{-1}e}. \tag{3.21}$$

Le problème (3.21) étant difficile à résoudre et l'étude numérique ayant montrée que le positionnement optimal des pupilles était périodique, nous avons étudié un problème plus simple où les positions des pupilles $u_0, \dots, u_n$ sont redondantes, i.e. définies par

$$u_k := k \times \ell, \tag{3.22}$$

pour $k = 0, \dots, n$ où $\ell \in \mathbb{R}$ représente la distance séparant deux pupilles successives. Avec cette hypothèse simplificatrice, l'analogue du problème (3.3) s'écrit :

$$\text{minimiser}_{(a,\ell)\in\mathcal{A}\times\mathbb{R}} \quad \frac{1}{2}\left\|\sum_{k=0}^{n} a_k g_{u_0+k\ell}\right\|_2^2. \tag{3.23}$$

Or, pour tout u_0, ℓ et $t \in \mathbb{R}$, nous avons l'égalité

$$\left|\sum_{k=0}^{n} a_k e^{i(u_0+k\ell)t}\right| = \left|\sum_{k=0}^{n} a_k e^{ik\ell t}\right|. \tag{3.24}$$

Donc, le problème simplifié (3.23) s'écrit sous la forme :

$$\text{minimiser}_{(a,\ell)\in\mathcal{A}\times\mathbb{R}_+} \quad \frac{1}{2}\left\|\sum_{k=0}^{n} a_k g_{k\ell}\right\|_2^2. \tag{3.25}$$

Dans cette section, nous allons nous intéresser à l'étude de l'existence d'une solution du problème (3.25). En commençant à prendre le minimum de ce problème par rapport à la variable $a \in \mathcal{A}$, le problème (3.25) devient :

$$\text{minimiser}_{\ell\in\mathbb{R}_+} \phi_1(\ell), \tag{3.26}$$

où la fonction $\phi_1 : \mathbb{R} \to \mathbb{R}$ est telle que $\phi_1(0) := 1$ et pour $\ell \neq 0$,

$$\phi_1(\ell) := \min_{a\in\mathcal{A}} \frac{1}{2}\left\|\sum_{k=0}^{n} a_k e^{ik\ell t}\right\|_2^2 = \frac{1}{2e^\top H(\ell)^{-1}e}, \tag{3.27}$$

où la matrice $H(\ell) \in \mathcal{M}(\mathbb{R})$ est définie par $H(\ell)_{kk'} := \Re(\langle g_{k\ell}, g_{k'\ell}\rangle)$ pour $k, k' = 0, \dots, n$ avec $g_{k\ell}(t) := e^{ik\ell t}$. Montrer l'existence d'une solution du problème (3.25) revient à montrer que ϕ_1 admet un minimum. C'est pourquoi, nous étudions les propriétés de cette fonction dans les paragraphes suivants.

3.3.1 Graphe de ϕ_1

La figure 3.3 donne des exemples de graphes de ϕ_1 réalisés avec le logiciel MATLAB pour différentes valeurs de n. Pour $k, k' = 0, \ldots, n$, les éléments de la matrice $H(\ell)$ sont

$$H(\ell)_{kk'} = \begin{cases} \Re\left(\dfrac{e^{i\ell(k'-k)t_{max}} - e^{i\ell(k'-k)t_{min}}}{i\ell(k'-k)(t_{max}-t_{min})}\right) & \text{pour } k \neq k', \\ 1 & \text{sinon.} \end{cases} \tag{3.28}$$

Une méthode numérique a été utilisée afin de calculer l'inverse de la matrice $H(\ell)$. Dans l'exemple, l'intervalle d'optimisation est $I = [\pi/2, 3\pi/2]$. À partir de sept pupilles ($n = 7$), une échelle semi-logarithmique a été utilisée pour tracer la courbe. En effet, il est difficile de visualiser le minimum avec une échelle linéaire. De plus, le calcul de la fonction objectif autour de zéro est délicat pour de grandes valeurs de n car la matrice $H(\ell)$ devient mal conditionnée. L'allure de la courbe est la même quelque soit le nombre de pupilles. Nous conjecturons que le premier minimum de chacune des courbes est l'optimum global du problème. Le tableau 3.1 donne les valeurs de ce minimum, noté ℓ^* ainsi que la valeur de l'objectif en ce point, noté $\phi_1(\ell^*)$. Dans cet exemple, nous conjecturons que ℓ^* converge vers 1.

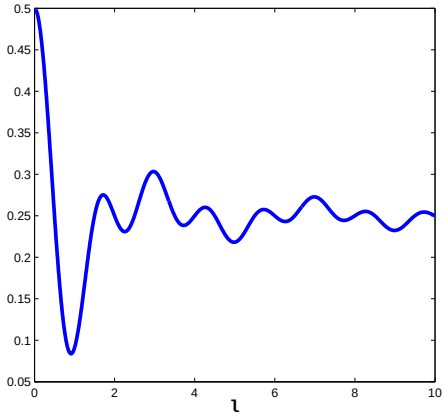
n	ℓ^*	$\phi_1(\ell^*)$
1	0.9100	0.0835
3	0.9800	0.0029
7	0.9919	$2.75.10^{-6}$
9	0.9945	$8.28.10^{-8}$
13	0.9945	$7.77.10^{-11}$

TAB. 3.1 – Valeurs de ℓ^* et de $\phi_1(\ell^*)$ pour différentes valeurs de n avec $I = [\pi/2, 3\pi/2]$.

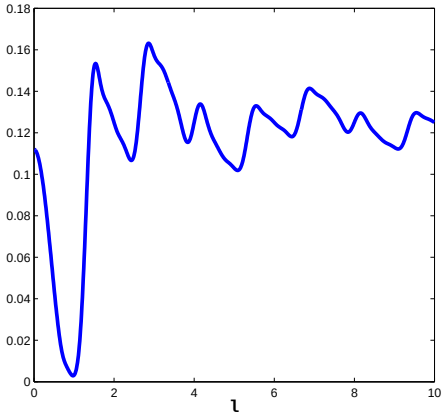
Pour prouver que ϕ_1 admet un minimum, il faut bien comprendre le comportement de cette fonction aux voisinages de 0^+ et de $+\infty$. En effet, en montrant que la fonction est décroissante dans un voisinage de 0^+ et que $\lim_{\ell\to 0^+} \phi_1(\ell) \leq \lim_{\ell\to+\infty} \phi_1(\ell)$, un argument de compacité permettrait de conclure sur l'existence d'un minimum de la fonction continue ϕ_1. Ces propriétés, étudiées dans les paragraphes suivants, n'ont en fait été démontrées que partiellement. Bien que visuellement l'existence d'un minimum est certaine, une preuve complète reste à faire.

3.3.2 Limites de ϕ_1

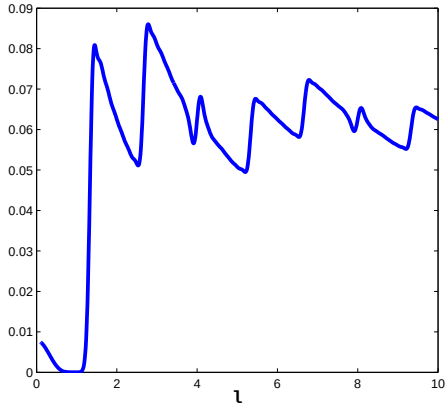
D'après les courbes de la figure 3.3, nous pouvons conjecturer que la fonction ϕ_1 admet une limite finie lorsque ℓ tend vers 0^+ (respectivement vers $+\infty$). Les valeurs de ces limites sont présentées dans cette partie.



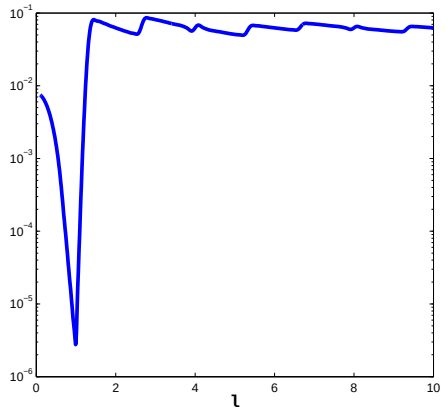
$n = 1$



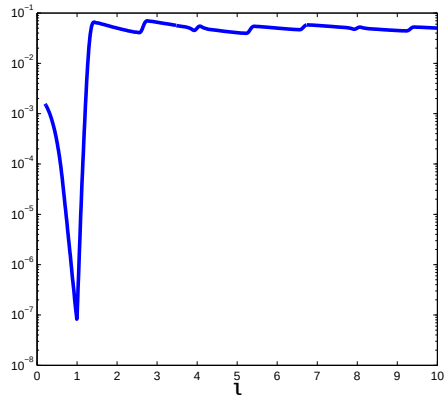
$n = 3$



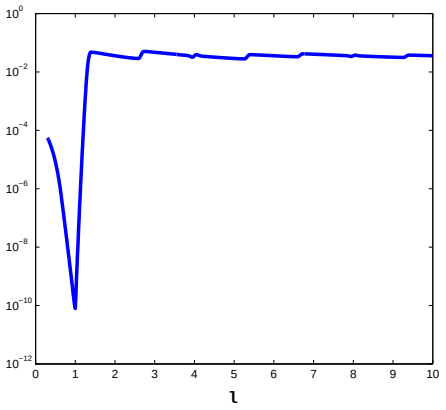
$n = 7\ (a)$



$n = 7\ (b)$



$n = 9$



$n = 13$

FIG. 3.3 – Représentation de la fonction objectif ϕ_1 définie par la formule (3.27) pour plusieurs valeurs de n avec $I = [\pi/2, 3\pi/2]$. Pour $n = 1, 3, 7(a)$, l'échelle est linéaire et pour $n = 7(b), 9, 13$, elle est semi-logarithmique.

Proposition 3.3.2 *La limite de ϕ_1 lorsque ℓ tend vers l'infini existe et est égale à :*

$$\lim_{\ell\to+\infty}\phi_1(\ell)=\frac{1}{2(n+1)}.$$

Preuve. D'après la formule (3.28), nous avons pour $k,k'\in\{0,\dots,n\}$

$$\lim_{\ell\to+\infty}H(\ell)_{kk'}=1,\ \text{ si } k=k',$$

et

$$\lim_{\ell\to+\infty}H(\ell)_{kk'}=0,\ \text{ si } k\neq k'.$$

Nous en déduisons

$$\lim_{\ell\to+\infty}H(\ell)=Id,$$

où Id est la matrice identité de dimension $(n+1)\times(n+1)$. D'après la formule (3.27), nous obtenons

$$\lim_{\ell\to+\infty}\phi_1(\ell)=\frac{1}{2e^\top Ide}=\frac{1}{2e^\top e}=\frac{1}{2(n+1)}.$$

□

Proposition 3.3.3 *Notons pour $k=0,\dots,n$, f_k l'application définie, pour tout $t\in\mathbb{R}$, par $f_k(t):=1+(it)^k$. La limite de ϕ_1 lorsque ℓ tend vers zéro existe et est égale à :*

$$\lim_{\ell\to0^+}\phi_1(\ell)=\frac{1}{2e^\top M^{-1}e},$$

où la matrice $M\in\mathcal{M}(\mathbb{R})$ est définie par $M_{kk'}:=\Re(\langle f_k,f_{k'}\rangle)$ pour $k,k'=0,\dots,n$.

Plan de la preuve. L'espace vectoriel $\mathbb{H}=L^2(I,\mathbb{C})$ muni de la norme $||\cdot||_2$ est un espace de Banach. Soit l'application

$$\begin{array}{rccc}\Delta: & \mathbb{R} & \to & \mathbb{H}\\ & u & \mapsto & g_u=e^{iut}.\end{array}$$

Δ est de classe $\mathcal{C}^\infty$. Elle est aussi analytique, i.e. pour tout $u\in\mathbb{R}$,

$$\Delta(u)=\sum_{k=0}^{\infty}u^k\beta_k,$$

avec pour tout k, $\beta_k\in\mathbb{H}$ défini par $\beta_k(t):=\dfrac{(it)^k}{k!}$. De plus, pour tout k, nous avons

$$\Delta^k(0):t\mapsto(it)^k,$$

et la famille $(\Delta^k(0))_{k\in\mathbb{N}}$ est libre dans $\mathbb{H}$.
D'après la définition (3.27) de ϕ_1 et la remarque 3.2.5, nous avons

$$\begin{array}{rcl}2\phi_1(\ell) & = & \mathrm{d}(0,\mathrm{aff}(g_0,g_\ell,\dots,g_{n\ell})),\\ & = & \mathrm{d}(0,\mathrm{aff}(\Delta(0),\Delta(\ell),\dots,\Delta(n\ell))).\end{array}$$

Lorsque ℓ tend vers zéro, un résultat connu de géométrie algébrique montre que l'espace affine aff $(\Delta(0), \Delta(\ell), \ldots, \Delta(n\ell))$ « converge » vers l'espace affine $\Delta(0) +$ vect $(\Delta'(0), \ldots, \Delta^{(n)}(0))$. Par continuité de la fonction distance (du vecteur nul à un sous-espace affine de dimension n), nous en déduisons

$$\begin{aligned}\lim_{\ell \to 0^+} 2\phi_1(\ell) &= \mathrm{d}(0, \Delta(0) + \text{vect } (\Delta'(0), \ldots, \Delta^{(n)}(0))), \\ &= \mathrm{d}(0, \text{aff } (f_0, f_1, \ldots, f_n)).\end{aligned}$$

En appliquant la formule (3.14) au nouvel espace affine, nous obtenons

$$\lim_{\ell \to 0^+} \phi_1(\ell) = \frac{1}{2e^\top M^{-1} e},$$

où la matrice $M \in \mathcal{M}(\mathbb{R})$ est telle que pour $k, k' = 0, \ldots, n$,

$$\begin{aligned}M_{kk'} &:= \Re(\langle f_k, f_{k'} \rangle) \\ &= \Re(\langle 1 + (it)^k, 1 + (it)^{k'} \rangle), \\ &= \Re(\langle 1, 1 \rangle + \langle 1, (it)^{k'} \rangle + \langle (it)^k, 1 \rangle + \langle (it)^k, (it)^{k'} \rangle) \\ &= \Re(1 + (i)^{k-1} \frac{t_{max}^k - t_{min}^k}{k(t_{max} - t_{min})} + (-i)^{k'-1} \frac{t_{max}^{k'} - t_{min}^{k'}}{k'(t_{max} - t_{min})} \\ &\quad + (i)^{k-k'} \frac{t_{max}^{k+k'-1} - t_{min}^{k+k'-1}}{(k + k' - 1)(t_{max} - t_{min})}).\end{aligned}$$

□

Le tableau 3.2 donne les valeurs des limites de ϕ_1 lorsque ℓ tend vers zéro et l'infini pour les exemples de la figure 3.3.

n	limite en 0^+	limite en $+\infty$
1	0.5000	0.2500
3	0.1120	0.1250
7	0.0079	0.0625
9	0.0021	0.0500
13	$1.33.10^{-4}$	0.0357

TAB. 3.2 – Valeurs des limites de ϕ_1 en zéro et l'infini pour plusieurs valeurs de n.

Nous conjecturons que pour $n \geq 3$, la limite de ϕ_1 en l'infini est supérieure à celle en zéro. Si de plus ϕ_1 était décroissante au voisinage de zéro, cela permettrait par un argument de compacité de prouver l'existence de la solution du problème (3.25). C'est pourquoi, dans le paragraphe suivant, nous étudions le sens de variation de la fonction ϕ_1.

3.3.3 Dérivée de ϕ_1

Dans cette partie, la dérivée de la fonction objectif ϕ_1 est calculée. Le but était de montrer que la fonction est décroissante au voisinage de zéro à partir de la dérivée.

Proposition 3.3.4 *La dérivée de ϕ_1 est*

$$\phi_1'(\ell) = \frac{e^\top H(\ell)^{-1} H'(\ell) H(\ell)^{-1} e}{2(e^\top H(\ell)^{-1} e)^2},$$

où la matrice $H'(\ell)$ est la dérivée terme à terme de la matrice $H(\ell)$.

Preuve. Soit l'application $\varphi : \mathbb{R}_+ \to \mathbb{R}$ définie, pour tout $\ell > 0$, par

$$\varphi(\ell) := e^\top H(\ell)^{-1} e.$$

Cette application peut s'écrire comme la composée des trois applications suivantes :

$$\begin{array}{rccccccccccc} f: & \mathbb{R} & \to & \mathcal{M}(\mathbb{R}) & g: & \mathcal{M}(\mathbb{R}) & \to & \mathcal{M}(\mathbb{R}) & h: & \mathcal{M}(\mathbb{R}) & \to & \mathbb{R} \\ & \ell & \mapsto & H(\ell) & & H & \mapsto & H^{-1} & & H^{-1} & \mapsto & e^\top H^{-1} e, \end{array}$$

i.e. $\varphi = h \circ (g \circ f)$. Soient $\ell > 0$ et $t \in \mathbb{R}$, la formule de la différentiabilité de la composée est

$$\varphi'(\ell)(t) = h'((g \circ f)(\ell))((g \circ f)'(\ell)(t)),$$

où

$$\begin{aligned} (g \circ f)'(\ell)(t) &= g'(f(\ell))(f'(\ell).t) \\ &= g'(f(\ell))(t.H'(\ell)) \\ &= g'(H(\ell))(t.H'(\ell)) \\ &= -t.H(\ell)^{-1} H'(\ell) H(\ell)^{-1}. \end{aligned}$$

Cette dernière égalité est donnée dans le Théorème 2.4.4 de [16]. Nous obtenons alors $\varphi'(\ell)(t) = h'(H(\ell)^{-1})(-t.H(\ell)^{-1} H'(\ell) H(\ell)^{-1})$. Comme h est linéaire, nous avons

$$\begin{aligned} \varphi'(\ell)(t) &= h'(e^\top(-t.H(\ell)^{-1} H'(\ell) H(\ell)^{-1})e) \\ &= e^\top(-t.H(\ell)^{-1} H'(\ell) H(\ell)^{-1})e \\ &= -t.e^\top H(\ell)^{-1} H'(\ell) H(\ell)^{-1} e. \end{aligned}$$

Cette dernière égalité étant vraie pour tout ℓ, nous obtenons

$$\varphi'(\ell) = -e^\top H(\ell)^{-1} H'(\ell) H(\ell)^{-1} e.$$

Comme $\phi_1(\ell) = \dfrac{1}{2\varphi(\ell)}$, il en résulte

$$\phi_1'(\ell) = -\frac{\varphi'(\ell)}{2\varphi(\ell)^2}.$$

En remplaçant $\varphi(\ell)$ et $\varphi'(\ell)$ par leurs expressions dans l'égalité précédente, le résultat est obtenu. □

L'expression de la dérivée de ϕ_1 étant compliquée, son signe ne peut être étudié que numériquement. En utilisant l'expression (3.27) de ϕ_1, il est évident que ϕ_1 est paire. Donc, si ϕ_1 est dérivable en 0, alors $\phi_1'(0) = 0$. Ceci implique que $\phi_1'(0)$ si elle existe, ne peut pas être strictement négative. On ne peut donc pas conclure sur la décroissance de ϕ_1 au voisinage de zéro. N'aboutissant pas au résultat souhaité, l'étude n'a pas été menée plus loin.

3.4 À propos de l'existence de la solution du problème écrit avec la norme $\|\cdot\|_\infty$

En utilisant les notations précédentes, le problème (3.4) s'écrit sous la forme :

$$\text{minimiser}_{(a,u)\in\mathcal{A}\times\mathbb{R}^{n+1}} \quad \max_{t\in I}\left|\sum_{k=0}^{n} a_k g_{u_k}(t)\right|. \tag{3.29}$$

Pour résoudre ce problème et pour les mêmes raisons qu'avec la norme $||\cdot||_2$, les positions des pupilles sont supposées régulièrement espacées, définies par la relation (3.22). L'égalité (3.24) étant aussi vérifiée, le problème (3.29) devient :

$$\text{minimiser}_{(a,\ell)\in\mathcal{A}\times\mathbb{R}_+} \quad \max_{t\in I}\left|\sum_{k=0}^{n} a_k g_{k\ell}(t)\right|. \tag{3.30}$$

En commençant par prendre l'infimum de ce problème par rapport à la variable $a \in \mathcal{A}$, le problème se reformule :

$$\inf_{\ell\in\mathbb{R}_+} \quad \phi_2(\ell), \tag{3.31}$$

où la fonction $\phi_2 : \mathbb{R} \to \mathbb{R}$ est telle que $\phi_2(0) := 1$ et pour $\ell \neq 0$,

$$\phi_2(\ell) := \inf_{a\in\mathcal{A}} \quad \max_{t\in I}\left|\sum_{k=0}^{n} a_k g_{k\ell}(t)\right|. \tag{3.32}$$

Dans cette partie, l'existence d'une solution optimale du problème (3.31) est démontrée. Cette solution, notée $(a^*, \ell^*) \in \mathcal{A}\times\mathbb{R}_+$, est liée aux polynômes de Tchebytchev. Les résultats présentés dans cette section font l'objet d'un article soumis pour publication. Avant d'énoncer le théorème d'existence de la solution, quelques rappels sur les polynômes de Tchebytchev sont présentés ainsi que les notations utiles pour la suite.

3.4.1 Rappels sur les polynômes de Tchebytchev

Pour tout entier $n \in \mathbb{N}$, le polynôme de Tchebytchev T_n de degré n est défini par la formule trigonométrique

$$T_n(\cos X) = \cos(nX),$$

pour tout $X \in \mathbb{R}$.

Pour tout entier n, le polynôme T_n a n racines simples dans l'intervalle ouvert $]-1, 1[$. Les polynômes de Tchebytchev sont aussi définis par la relation de récurrence

$$T_{n+1}(X) + T_{n-1}(X) = 2XT_n(X), \tag{3.33}$$

pour tout entier $n \geq 1$ où $T_0(X) = 1$ et $T_1(X) = X$.
Plus de propriétés sur la suite des polynômes de Tchebytchev sont décrites dans [1, 19].

3.4.2 Notations

- La norme $\|\cdot\|_\infty$ est notée sous la forme $\|\cdot\|_I$ où I est l'intervalle d'optimisation.
- Un nombre pair $n = 2m$ ou un nombre impair $n = 2m + 1$ de télescopes est supposé. Pour la suite, nous posons $m := \lfloor \frac{n}{2} \rfloor$.
- La matrice Id est la matrice identité.
- L'espace $\mathbb{E} := \mathbb{R}_n[X]$ est l'espace vectoriel des polynômes de degré inférieur ou égal à n.
- D'après la section 3.4.1, nous pouvons écrire :

$$T_n(X) = \sum_{j=0}^{m} t_j X^{n-2j}, \tag{3.34}$$

et définir dans $\mathbb{E}$, le polynôme

$$T_n^*(X) := \frac{1}{T_n(\frac{1}{c})} \sum_{j=0}^{m} t_j X^j \left(\frac{1+X}{2c} \right)^{n-2j}, \tag{3.35}$$

où $c := \cos\left(\dfrac{t_{min}\pi}{t_{min} + t_{max}} \right) \in]0, 1[$.

Dans les deux sections suivantes, le théorème d'existence de la solution du problème (3.31) est énonçé et nous en donnons sa preuve.

3.4.3 Énoncé du théorème

Théorème 3.4.1 *L'infimum du problème (3.31) est atteint en* $\ell^* := \dfrac{2\pi}{t_{min} + t_{max}}$. *De plus,*

$$\phi_2(\ell^*) = \left\| \sum_{k=0}^{n} a_k^* g_{k\ell^*} \right\|_I = \frac{1}{T_n(\frac{1}{c})},$$

où $a^* = (a_0^*, \ldots, a_n^*)$ *est le vecteur des coefficients de* T_n^*, *tel que*

$$T_n^*(X) = \sum_{k=0}^{n} a_k^* X^k.$$

3.4.4 Preuve du théorème

Avant de donner la preuve du théorème 3.4.1, quelques notations et un lemme sont énoncés.
Pour tout $P \in \mathbb{E}$, nous utilisons la notation

$$P := \sum_{k=0}^{n} p_k X^k.$$

Notons $\mathbb{F}$ et $\mathbb{F}'$ les sous-ensembles des polynômes pairs et impairs définis par

$$\mathbb{F} := \{P \in \mathbb{E} : P(X) = P(-X)\} \text{ et } \mathbb{F}' := \{P \in \mathbb{E} : P(X) = -P(-X)\}.$$

Nous avons $\mathbb{F} \oplus \mathbb{F}' = \mathbb{E}$. Notons $p_{\mathbb{F}} : \mathbb{E} \to \mathbb{F}$ la projection sur $\mathbb{F}$ par rapport à $\mathbb{F}'$ telle que

$$p_{\mathbb{F}}(\sum_{k=0}^{n} p_k X^k) = \sum_{j=0}^{m} p_{2j} X^{2j}.$$

La notation suivante est aussi utilisée :

$$\mathbb{G} := \{P \in \mathbb{E} : P(X) = (-1)^n P(-X)\}.$$

Si n est pair, alors $\mathbb{G} = \mathbb{F}$ sinon $\mathbb{G} = \mathbb{F}'$.
Trois isomorphismes linéaires sont définis. Le premier isomorphisme est

$$\begin{array}{rccc} h: & \mathbb{R}_m[X] & \to & \mathbb{G} \\ & P & \mapsto & X^n P\left(\dfrac{1}{X^2}\right). \end{array}$$

Il s'écrit encore :

$$h(\sum_{j=0}^{m} p_j X^j) = \sum_{j=0}^{m} p_j X^{n-2j}. \tag{3.36}$$

Le deuxième isomorphisme est

$$\begin{array}{rccc} f: & \mathbb{E} & \to & \mathbb{E} \\ & A = \displaystyle\sum_{k=0}^{n} a_k X^k & \mapsto & B = \displaystyle\sum_{k=0}^{n} b_k X^k, \end{array}$$

défini par

$$f(A) = (1-X)^n A\left(\frac{1+X}{1-X}\right). \tag{3.37}$$

Il s'écrit ainsi :

$$B = \sum_{k=0}^{n} a_k (1+X)^k (1-X)^{n-k}.$$

D'après l'expression (3.37) de f, son inverse vérifie

$$f^{-1}(B) = \left(\frac{X+1}{2}\right)^n B\left(\frac{X-1}{X+1}\right). \tag{3.38}$$

Il s'écrit encore :

$$A = \sum_{k=0}^{n} b_k \left(\frac{X+1}{2}\right)^{n-k} \left(\frac{X-1}{2}\right)^k. \tag{3.39}$$

Le troisième isomorphisme est

$$\begin{array}{rccc} g: & \mathbb{F} & \to & \mathbb{G} \\ & C = \displaystyle\sum_{j=0}^{m} c_j X^{2j} & \mapsto & D = \displaystyle\sum_{j=0}^{m} d_j X^{n-2j}, \end{array}$$

défini par

$$g(C) = X^n C\left(\frac{\sqrt{X^2-1}}{X}\right). \tag{3.40}$$

Il s'écrit aussi :

$$D = \sum_{j=0}^{m} c_j (X^2-1)^j X^{n-2j}. \tag{3.41}$$

En faisant le changement de variable $Y = \dfrac{1}{X^2}$ dans la dernière égalité, nous avons

$$\sum_{j=0}^{m} d_j Y^j = \sum_{j=0}^{m} c_j (1-Y)^j.$$

Cette égalité permet de mettre en évidence le fait que l'inverse de l'isomorphisme de g satisfait la relation suivante :

$$g^{-1}(D) = h^{-1}(D)(1-X^2). \tag{3.42}$$

Nous pouvons ainsi définir $q := g \circ p_{\mathbb{F}} \circ f$ et $q' := f^{-1} \circ g^{-1}$. Nous avons alors

$$q \circ q' = Id_{\mathbb{G}}. \tag{3.43}$$

Un lemme qui est utile pour la suite est présenté ici.

Lemme 3.4.2 *Soient* $\mathbb{E}_1 = \{A \in \mathbb{E} : A(1) = 1\}$ *et* $\mathbb{G}_1 = \{D \in \mathbb{G} : D(1) = 1\}$.
- *Pour tout* $A \in \mathbb{E}$, *nous avons* $A(1) = q(A)(1)$.
- *Nous avons l'inclusion* $q'(\mathbb{G}_1) \subset \mathbb{E}_1$ *et l'égalité* $q(\mathbb{E}_1) = \mathbb{G}_1$.

Preuve.
- Soit $A \in \mathbb{E}$. D'après l'expression (3.37) de f, le polynôme $B := f(A)$ vérifie

  $$B(0) = A(1).$$

  D'après la définition de $p_{\mathbb{F}}$, le polynôme $C := p_{\mathbb{F}}(B)$ vérifie

  $$C(0) = B(0).$$

  D'après l'expression (3.40) de g, le polynôme $D := g(C) = q(A)$ vérifie

  $$D(1) = C(0).$$

  En combinant ces trois égalités, nous en concluons que $A(1) = q(A)(1)$.
- Soit $A \in q'(\mathbb{G}_1)$, il existe $D \in \mathbb{G}_1$ tel que $A = q(D)$. En utilisant le résultat précédent et la formule (3.43), nous avons

  $$A(1) = q(A)(1) = (q \circ q')(D)(1) = D(1) = 1,$$

  ce qui implique que $A \in \mathbb{E}_1$. En composant l'inclusion avec q, nous obtenons $\mathbb{G}_1 \subset q(\mathbb{E}_1)$. L'inclusion inverse est une conséquence de $A(1) = q(A)(1)$.

□

La preuve du théorème 3.4.1 est basée sur plusieurs lemmes.

Preuve. Dans le premier lemme, des reformulations de la fonction ϕ_2 définie par la relation (3.32) et du problème (3.30) sont données.

Lemme 3.4.3 *Pour tout $\ell \geq 0$,*

$$\phi_2(\ell) = \inf_{A \in \mathbb{E}_1} \|A \circ g_1\|_{I(\ell)}, \tag{3.44}$$

où $I(\ell) := [\ell t_{min}, \ell t_{max}]$. Le problème (3.30) peut s'écrire :

$$\inf_{(A,\ell) \in \mathbb{E}_1 \times \mathbb{R}_+} \|A \circ g_1\|_{I(\ell)}. \tag{3.45}$$

Preuve. Soit $\tau : \mathcal{A} \to \mathbb{E}_1$ définie pour tout $a \in \mathcal{A}$ par

$$\tau(a) := \sum_{k=0}^{n} a_k X^k.$$

Alors, pour tout $(a, \ell) \in \mathcal{A} \times \mathbb{R}_+$, en définissant $A := \tau(a)$, nous avons

$$\begin{aligned} \left\| \sum_{k=0}^{n} a_k g_{k\ell} \right\|_I &= \sup_{t \in I} \left| \sum_{k=0}^{n} a_k e^{ik\ell t} \right| \\ &= \sup_{t \in I} \left| A\left(e^{it\ell}\right) \right| \\ &= \sup_{\theta \in I(\ell)} \left| A\left(e^{i\theta}\right) \right| \\ &= \|A \circ g_1\|_{I(\ell)}. \end{aligned}$$

D'après l'expression (3.32) de ϕ_2, la relation (3.44) est obtenue. Avec le changement de variable τ, le problème général (3.30) peut se reformuler comme le problème (3.45).

□

Dans le prochain lemme, une nouvelle formulation de la fonction ϕ_2 est proposée.

Lemme 3.4.4 *Soit $\ell \in]0, \dfrac{2\pi}{t_{max}}]$ et considérons l'intervalle*

$$J(\ell) := [\gamma(\ell), \delta(\ell)] := [\cos(\frac{\ell t_{max}}{2}), \cos(\frac{\ell t_{min}}{2})]$$

inclus dans $]-1, 1[$ tel que $\gamma(\ell) < \delta(\ell)$.

- *Pour tout $(\theta, A) \in \mathbb{R} \times \mathbb{E}$, nous avons $|A(e^{i\theta})| \geq |q(A)(\cos\dfrac{\theta}{2})|$ avec égalité lorsque $A \in f^{-1}(\mathbb{F})$.*
- *Pour tout $A \in \mathbb{E}$, nous avons $\|A \circ g_1\|_{I(\ell)} \geq \|q(A)\|_{J(\ell)}$ avec égalité lorsque $A \in f^{-1}(\mathbb{F})$.*
- *Nous avons*

$$\phi_2(\ell) = \inf_{D \in \mathbb{G}_1} \|D\|_{J(\ell)}. \tag{3.46}$$

De plus, si $\phi_2(\ell) = \|D^*\|_{J(\ell)}$ pour $D^* \in \mathbb{G}_1$, alors

$$\phi_2(\ell) = \|q'(D^*) \circ g_1\|_{I(\ell)}.$$

Preuve.

– Soit $A = \sum_{k=0}^{n} a_k X^k \in \mathbb{E}$, $\theta \in \mathbb{R}$ et posons $\theta' = \dfrac{\theta}{2}$. Définissons

$$B := f(A) = \sum_{k=0}^{n} b_k X^k, \;\; C := p_{\mathbb{F}}(B) = \sum_{j=0}^{m} b_{2j} X^{2j} \;\text{ et }\; D := g(C).$$

Nous obtenons $D = q(A)$. En utilisant l'expression (3.39) de A, nous écrivons

$$\begin{aligned} A(e^{i\theta}) &= \sum_{k=0}^{n} b_k \left(\frac{e^{i\theta}+1}{2}\right)^{n-k} \left(\frac{e^{i\theta}-1}{2}\right)^{k} \\ &= e^{in\theta'} \sum_{k=0}^{n} b_k i^k \left(\cos\theta'\right)^{n-k} \left(\sin\theta'\right)^{k}. \end{aligned}$$

En prenant le module des deux côtés de l'égalité et en utilisant la relation $|z| \geq |\Re(z)|$ pour tout nombre complexe z, il en résulte

$$|A(e^{i\theta})| \geq \left| \sum_{j=0}^{m} b_{2j} (-1)^j \left(\sin\theta'\right)^{2j} \left(\cos\theta'\right)^{n-2j} \right|, \tag{3.47}$$

avec égalité lorsque $B \in \mathbb{F}$ ou de manière équivalente lorsque $A \in f^{-1}(\mathbb{F})$. En utilisant l'expression (3.41) de D, nous en déduisons

$$D\left(\cos\theta'\right) = \sum_{j=0}^{m} b_{2j} \left(\cos^2\theta' - 1\right)^j \left(\cos\theta'\right)^{n-2j}. \tag{3.48}$$

En combinant les relations (3.47) et (3.48), nous en concluons

$$|A(e^{i\theta})| \geq \left|D\left(\cos\theta'\right)\right|,$$

avec égalité lorsque $A \in f^{-1}(\mathbb{F})$.

– Il suffit de prendre le suprémum des deux côtés de l'inégalité précédente.

– En utilisant l'expression (3.44) de ϕ_2 et l'inégalité précédente, nous avons

$$\phi_2(\ell) \geq \inf_{A \in \mathbb{E}_1} \|q(A)\|_{J(\ell)}.$$

D'après le lemme 3.4.2, nous obtenons

$$\phi_2(\ell) \geq \inf_{D \in \mathbb{G}_1} \|D\|_{J(\ell)}. \tag{3.49}$$

Soit $D_1 \in \mathbb{G}_1$ et posons $A_1 := q'(D_1)$. En utilisant la définition de q', nous constatons que $A_1 = f^{-1}(g^{-1}(D_1)) \in f^{-1}(\mathbb{F})$. Le lemme 3.4.2 implique

$$A_1 \in \mathbb{E}_1. \tag{3.50}$$

De plus, selon la relation (3.43),

$$q(A_1) = D_1. \tag{3.51}$$

En utilisant les propriétés (3.50) et (3.51), nous avons

$$\|D_1\|_{J(\ell)} = \|q(A_1)\|_{J(\ell)} = \|A \circ g_1\|_{I(\ell)} \geq \inf_{A \in \mathbb{E}_1} \|A \circ g_1\|_{I(\ell)} = \phi_2(\ell).$$

En prenant l'infimum des deux côtés de l'inégalité, nous obtenons

$$\inf_{D_1 \in \mathbb{G}_1} \|D_1\|_{J(\ell)} \geq \phi_2(\ell).$$

En combinant les relations précédentes, l'égalité demandée est obtenue. En supposant qu'il existe $D^* \in \mathbb{G}_1$ tel que $\phi_2(\ell) = \|D^*\|_{J(\ell)}$, nous avons

$$\phi_2(\ell) = \|A^* \circ g_1\|_{I(\ell)},$$

où $A^* := q'(D^*)$.

□

Dans le prochain lemme, une autre définition de ϕ_2 est donnée. Celle-ci est utile pour établir la valeur de ϕ_2 en son minimum.

Lemme 3.4.5 *Soit* $\ell \in]0, \dfrac{2\pi}{t_{max}}]$ *et notons*

$$\phi_3(\ell) := \inf_{D \in \mathbb{E}_1} \|D\|_{J(\ell)}. \tag{3.52}$$

Alors, le polynôme

$$D_\ell := \frac{T_n\left(\frac{2X-\delta(\ell)-\gamma(\ell)}{\delta(\ell)-\gamma(\ell)}\right)}{T_n\left(\frac{2-\delta(\ell)-\gamma(\ell)}{\delta(\ell)-\gamma(\ell)}\right)}$$

de $\mathbb{E}_1$ *satisfait les égalités suivantes :*

$$\phi_3(\ell) = \|D_l\|_{J(\ell)} = \frac{1}{T_n\left(\frac{2-\delta(\ell)-\gamma(\ell)}{\delta(\ell)-\gamma(\ell)}\right)}.$$

Preuve. Posons $\mathbb{E}_0 = \{Q \in \mathbb{E} : Q(1) = 0\}$. En faisant le changement de variable $D = X^n - Q$ dans le problème (3.52), nous avons

$$\phi_3(\ell) = \inf_{Q \in \mathbb{E}_0} \|X^n - Q\|_{J(\ell)}.$$

Tout élément non nul $Q \in \mathbb{E}$ satisfait $Q(1) = 0$ et a au plus $n-1$ racines dans l'intervalle $J(\ell)$ ce qui signifie que $\mathbb{E}_0$ satisfait la condition de Haar (voir Définition 3.4.1 dans [45]).
Soit la transformation affine $\Gamma : J(\ell) = [\gamma(\ell), \delta(\ell)] \to [-1, 1]$ définie par

$$\Gamma(X) := \frac{2X - \delta(\ell) - \gamma(\ell)}{\delta(\ell) - \gamma(\ell)},$$

pour tout $X \in J(\ell)$. Nous avons alors $\Gamma(\gamma(\ell)) = -1$ et $\Gamma(\delta(\ell)) = 1$.

Comme le polynôme de Tchebytchev T_n équioscille $n+1$ fois dans l'intervalle $[-1,1]$ et que $\|T_n\|_{[-1,1]}=1$, alors $T_n\circ\Gamma$ équioscille également $n+1$ fois dans $J(\ell)$ et

$$\|T_n\circ\Gamma\|_{[-1,1]}=1.$$

De plus, comme $J(\ell)\subset]-1,1[$ et que Γ n'est pas décroissante, nous avons

$$\Gamma(1)=\frac{2-\delta(\ell)-\gamma(\ell)}{\delta(\ell)-\gamma(\ell)}>1,$$

ce qui implique que $T_n\circ\Gamma(1)>0$ car l'intervalle $]-1,1[$ contient toutes les racines de T_n et $T_n(1)=1$. Donc, le polynôme D_ℓ peut s'écrire

$$D_\ell=\frac{T_n\circ\Gamma}{T_n\circ\Gamma(1)}.$$

Ce polynôme équioscille $n+1$ fois dans $J(\ell)$ et

$$\|D_\ell\|_{J(\ell)}=\frac{1}{T_n\left(\frac{2-\delta(\ell)-\gamma(\ell)}{\delta(\ell)-\gamma(\ell)}\right)}.$$

En utilisant le Théorème alterné de Tchebytchev (voir Théorème 3.5.2 dans [45]), nous en concluons que le polynôme $Q_\ell=X^n-D_\ell$ qui appartient à $\mathbb{E}_0$, est l'unique projection de X^n sur $\mathbb{E}_0$ et que $\phi_3(\ell)=\|D_\ell\|_{J(\ell)}$.

□

À présent, donnons un dernier lemme qui permet de finir la preuve du théorème 3.4.1.

Lemme 3.4.6 *En rappelant que* $\ell^*:=\dfrac{2\pi}{t_{min}+t_{max}}$ *et* $c:=\cos\left(\dfrac{t_{min}\pi}{t_{min}+t_{max}}\right)$, *nous avons les quatre propriétés suivantes.*

- *Pour tout* $\ell\in]0,\ell^*]$, $\phi_2(\ell^*)=\phi_3(\ell^*)\leq\phi_3(\ell)\leq\phi_2(\ell)$.
- *Pour tout* $\ell\in[\ell^*,+\infty[$, $\phi_2(\ell^*)\leq\phi_2(\ell)$.
- *Nous avons*
$$\inf_{\ell\in\mathbb{R}_+}\phi_2(\ell)=\phi_2(\ell^*)=\|D_{\ell^*}\|_{J(\ell^*)}=\frac{1}{T_n(\frac{1}{c})}.$$
- *Nous avons*
$$||D_{\ell^*}||_{J(\ell^*)}=\|T_n^*\circ g_1\|_{I(\ell^*)}.$$

Preuve.

- L'inégalité la plus à droite est dûe aux expressions (3.46) et (3.52) de ϕ_2 et de ϕ_3 ainsi qu'à l'inclusion $\mathbb{G}_1\subset\mathbb{E}_1$. D'après les définitions de $\gamma(\ell)$ et de $\delta(\ell)$ du lemme 3.4.4, nous avons $\delta(\ell^*)=-\gamma(\ell^*)=c$. Il en résulte
$$\frac{2X-\delta(\ell^*)-\gamma(\ell^*)}{\delta(\ell^*)-\gamma(\ell^*)}=\frac{X}{c},\tag{3.53}$$
et donc le polynôme D_ℓ défini dans le lemme 3.4.5, satisfait la relation :
$$D_{\ell^*}=\frac{T_n(\frac{X}{c})}{T_n(\frac{1}{c})}.\tag{3.54}$$

En utilisant le lemme 3.4.5, que $D_{\ell^*} \in \mathbb{G}_1$ et l'expression (3.46) de ϕ_2, nous avons

$$\begin{aligned}\phi_3(\ell^*) &= \|D_{\ell^*}\|_{J(\ell^*)} \\ &\geq \inf_{D\in\mathbb{G}_1} \|D\|_{J(\ell^*)} \\ &= \phi_2(\ell^*) \\ &\geq \phi_3(\ell^*).\end{aligned}$$

Toutes les inégalités sont des égalités et donc $\phi_2(\ell^*) = \phi_3(\ell^*)$.
Pour conclure, il suffit de vérifier que $\phi_3(\ell^*) \leq \phi_3(\ell)$ pour tout $\ell \in]0, \ell^*]$. Soit $\ell \in]0, \ell^*]$, posons $s := \dfrac{t_{max}}{t_{min}} > 1$ et $\kappa := \dfrac{\ell t_{min}}{4} \in]0, \dfrac{\pi}{2(1+s)}[$.
La dérivée première de la fonction $t \mapsto \dfrac{\sin st}{\sin t}$ est négative sur $]0, \dfrac{\pi}{2s}[$ si et seulement si $s\tan t < \tan st$ pour tout $t \in]0, \dfrac{\pi}{2s}[$. Cette dernière inégalité est une conséquence de la stricte convexité de la fonction tangente sur $]0, \dfrac{\pi}{2}[$. Ceci implique que la fonction $t \mapsto \dfrac{\sin st}{\sin t}$ est décroissante sur $]0, \dfrac{\pi}{2(1+s)}[$. Nous en déduisons les inégalités suivantes :

$$\frac{\sin s\kappa}{\sin\kappa} \geq \frac{\sin\frac{\pi s}{2(1+s)}}{\sin\frac{\pi}{2(1+s)}} = \cot\frac{\pi}{2(1+s)} \geq 1$$

et

$$1 + \frac{2}{(\frac{\sin s\kappa}{\sin\kappa})^2 - 1} \leq 1 + \frac{2}{\cot^2\frac{\pi}{2(1+s)}} = \frac{1}{\cos\frac{\pi}{1+s}}.$$

Nous obtenons alors

$$\begin{aligned}\frac{2-\delta(\ell)-\gamma(\ell)}{\delta(\ell)-\gamma(\ell)} &= \frac{(1-\cos\frac{\ell t_{max}}{2}) + (1-\cos\frac{\ell t_{min}}{2})}{(1-\cos\frac{\ell t_{max}}{2}) - (1-\cos\frac{\ell t_{min}}{2})} \\ &= \frac{\sin^2\frac{\ell t_{max}}{4} + \sin^2\frac{\ell t_{min}}{4}}{\sin^2\frac{\ell t_{max}}{4} - \sin^2\frac{\ell t_{min}}{4}} \\ &= 1 + \frac{2}{(\frac{\sin s\kappa}{\sin\kappa})^2 - 1} \\ &\leq \frac{1}{\cos\frac{\pi}{1+s}} \\ &= \frac{1}{\cos\frac{\pi t_{min}}{t_{min}+t_{max}}} \\ &= \frac{1}{c}.\end{aligned}$$

L'intervalle $]-1, 1[$ contient toutes les racines du polynôme T_n. Le Théorème de Rolle montre que les propriétés sont similaires pour le polynôme dérivé T_n'.

Comme $T_n(1) = 1$, le polynôme T_n est alors croissant sur l'intervalle $[1, +\infty[$. L'inégalité précédente implique

$$T_n\left(\frac{2-\delta(\ell)-\gamma(\ell)}{\delta(\ell)-\gamma(\ell)}\right) \leq T_n\left(\frac{1}{c}\right).$$

En utilisant le lemme 3.4.5 et la relation (3.53), la dernière inégalité peut se réécrire $\phi_2(\ell^*) \leq \phi_2(\ell)$.

– Soit $\ell > \ell^*$. Nous allons montrer que $\phi_2(\ell) \geq \phi_2(\ell^*)$ en étudiant deux cas.

Premier cas : Supposons que $\ell t_{max} < 2\pi$.

En posant $\ell_1 := \dfrac{2\pi - \ell t_{max}}{t_{min}}$ et en rappelant que $\ell > \ell^* = \dfrac{2\pi}{t_{min}+t_{max}}$, nous avons $\ell_1 \in]0, \ell^*[$ et $I(\ell_1) \subset [2\pi - \ell t_{max}, 2\pi - \ell t_{min}]$. Ainsi, pour tout $A \in \mathbb{E}_1$,

$$\begin{aligned}\|A \circ g_1\|_{I(\ell)} &= \|A \circ g_1\|_{[2\pi-\ell t_{max}, 2\pi-\ell t_{min}]} \\ &\geq \|A \circ g_1\|_{I(\ell_1)}.\end{aligned}$$

En prenant l'infimum dans l'inégalité précédente pour tout $A \in \mathbb{E}_1$ et en utilisant l'expression (3.44) de ϕ_2, nous en déduisons que $\phi_2(\ell) \geq \phi_2(\ell_1) \geq \phi_2(\ell^*)$.

Deuxième cas : Supposons que $\ell t_{max} \geq 2\pi$.

Il existe $n \in \mathbb{N}^*$ tel que $2n\pi \leq \ell t_{max} \leq 2n\pi + 2\pi$. Si $\ell t_{min} \leq 2n\pi$, alors pour tout $A \in \mathbb{E}_1$, nous avons $\|A \circ g_1\|_{I(\ell)} \geq 1$ ce qui implique la conclusion souhaitée à savoir $\phi_2(\ell) \geq 1 = \phi_2(0) \geq \phi_2(\ell^*)$. Nous pouvons donc supposer que $\ell t_{min} > 2n\pi$ ce qui signifie que $[\ell t_{min}, \ell t_{max}] \subset]2n\pi, 2(n+1)\pi[$.

En posant $\ell_2 := \ell - \dfrac{2n\pi}{t_{max}}$, nous avons $\ell_2 t_{max} \in]0, 2\pi[$ et $I(\ell_2) \subset [\ell t_{min} - 2n\pi, \ell t_{max} - 2n\pi]$. Alors, pour tout $A \in \mathbb{E}_1$,

$$\begin{aligned}\|A \circ g_1\|_{I(\ell)} &= \|A \circ g_1\|_{[\ell t_{min}-2n\pi, \ell t_{max}-2n\pi]} \\ &\geq \|A \circ g_1\|_{I(\ell_2)}.\end{aligned}$$

En prenant l'infimum dans l'inégalité précédente pour tout $A \in \mathbb{E}_1$ et en utilisant le premier résultat, nous en déduisons que $\phi_2(\ell) \geq \phi_2(\ell_2) \geq \phi_2(\ell^*)$.

– La première égalité est une conséquence des deux résultats démontrés précédemment et de l'inégalité $\phi_2(0) = 1 \geq \phi_2(\ell^*)$. Les autres égalités sont des conséquences du lemme 3.4.4 et des égalités $\phi_2(\ell^*) = \phi_3(\ell^*)$ et $J(\ell^*) = [-c, c]$.

– Il suffit de montrer que $q'(D_{\ell^*}) = T_n^*$ pour prouver l'égalité. D'après la relation (3.54) de D_{ℓ^*} et l'expression (3.34) du polynôme de Tchebytchev T_n, nous avons

$$D_{\ell^*}(X) = \frac{1}{T_n(\frac{1}{c})} \sum_{j=0}^{m} \frac{t_j}{c^{n-2j}} X^{n-2j}.$$

En utilisant l'expression (3.36) de h, nous obtenons

$$h^{-1}(D_{\ell^*})(X) = \frac{1}{T_n(\frac{1}{c})} \sum_{j=0}^{m} \frac{t_j}{c^{n-2j}} X^{j}.$$

En utilisant l'expression (3.42) de g^{-1}, nous avons

$$g^{-1}(D_{\ell^*})(X) = \frac{1}{T_n(\frac{1}{c})} \sum_{j=0}^{m} \frac{t_j}{c^{n-2j}} (1-X^2)^j,$$

et en utilisant l'expression (3.38) de f^{-1}, nous pouvons conclure

$$\begin{aligned} q'(D_{\ell^*})(X) &= \frac{(\frac{X+1}{2})^n}{T_n(\frac{1}{c})} \sum_{j=0}^{m} \frac{t_j}{c^{n-2j}} \left(1 - \left(\frac{X-1}{X+1}\right)^2\right)^j \\ &= T_n^*(X). \end{aligned}$$

□

□

3.4.5 Propriétés des amplitudes optimales des champs

Proposition 3.4.7 *Les coefficients du polynôme* $T_n^*(X) = \sum_{k=0}^{n} a_k^* X^k$*, défini par la relation (3.35), sont symétriques positifs :*

$$a_k^* = a_{n-k}^* \quad et \quad a_k^* \geq 0.$$

Preuve. Tout d'abord, notons que le polynôme T_n^* défini par la relation (3.35), peut s'écrire sous la forme :

$$T_n^*(X) := \frac{1}{T_n(\frac{1}{c})} \sqrt{X}^n T_n\left(\frac{1}{2c}\left(\sqrt{X} + \frac{1}{\sqrt{X}}\right)\right). \tag{3.55}$$

D'après la formule (3.55), nous avons $X^n T_n^*\left(\frac{1}{X}\right) = T_n^*(X)$ ce qui implique que les coefficients de T_n^* sont symétriques.

Maintenant, prouvons que les coefficients sont positifs. Définissons le polynôme

$$Q_n := T_n\left(\frac{1}{c}\right) T_n^*.$$

D'après les relations (3.33) et (3.55), la suite des polynômes $\{Q_n\}_{n\in\mathbb{N}}$ satisfait la relation de récurrence

$$Q_{n+1}(X) = \frac{1}{c}(X+1)Q_n - XQ_{n-1} \quad \text{pour tout } n \geq 1, \tag{3.56}$$

avec $Q_0(X) = 1$ et $Q_1(X) = \frac{1}{2c}(1+X)$. En particulier, nous avons

$$Q_2(X) = \frac{1}{2c^2}(1 + 2(1-c^2)X + X^2).$$

Nous allons prouver que pour tout entier $k = 0, \ldots, n$, la dérivée $k$ième de Q_n en zéro, notée $Q_n^{(k)}(0)$ est positive. Pour cela, montrons par récurrence pour $n \in \mathbb{N}$ que

$$\forall k \in \{0, \ldots, n-2\} \;\; Q_n^{(k)}(0) \geq \frac{1}{c} Q_{n-1}^{(k)}(0) \quad \text{et} \quad \forall k \in \{0, \ldots, n\} \;\; Q_n^{(k)}(0) \geq 0. \tag{3.57}$$

D'après les relations précédentes, les propriétés sont vraies pour $n = 0, n = 1$ et $n = 2$. Supposons que les inégalités définies par (3.57) sont vérifiées pour tout entier $m \leq n$ où $n \geq 2$ et montrons qu'elles le sont aussi au rang $n+1$.

Tout d'abord, d'après (3.56) et l'hypothèse de récurrence, nous avons

$$Q_{n+1}^{(0)}(0) = \frac{1}{c} Q_n^{(0)}(0) \geq 0.$$

Soit $k \in \{1, \ldots, n-1\}$. En utilisant la relation de récurrence (3.56), nous avons que la $k$ième dérivée de Q_{n+1} est de la forme :

$$Q_{n+1}^{(k)} = \frac{1}{c} Q_n^{(k)} + \frac{k}{c}(Q_n^{(k-1)} - c Q_{n-1}^{(k-1)}) + \frac{X}{c}(Q_n^{(k)} - c Q_{n-1}^{(k)}).$$

Nous en déduisons alors

$$\begin{aligned}
Q_{n+1}^{(k)}(0) &= \frac{1}{c} Q_n^{(k)}(0) + \frac{k}{c}(Q_n^{(k-1)}(0) - c Q_{n-1}^{(k-1)}(0)) \\
&\geq \frac{1}{c} Q_n^{(k)}(0) + \frac{k}{c}(\frac{1}{c} Q_{n-1}^{(k-1)}(0) - c Q_{n-1}^{(k-1)}(0)) \\
&= \frac{1}{c} Q_n^{(k)}(0) + \frac{k(1-c^2)}{c^2} Q_{n-1}^{(k-1)}(0) \\
&\geq \frac{1}{c} Q_n^{(k)}(0) \\
&\geq 0,
\end{aligned}$$

où chaque inégalité provient de l'hypothèse de récurrence. De plus, comme les coefficients de Q_{n+1} sont symétriques et comme nous avons montré que $Q_{n+1}^{(1)}(0) \geq 0$ et $Q_{n+1}^{(0)}(0) \geq 0$, nous avons aussi $Q_{n+1}^{(n)}(0) \geq 0$ et $Q_{n+1}^{(n+1)}(0) \geq 0$. Donc, les propriétés sont vraies au rang $n+1$ et ainsi pour tout $n \in \mathbb{N}$.

□

Proposition 3.4.8 *Pour tout $t \in \mathbb{R}$, nous avons*

$$\sum_{k=0}^{n} a_k^* e^{2ik\ell^* t} = e^{in\ell^* t} \, \frac{T_n\left(\frac{1}{c}\cos(\ell^* t)\right)}{T_n\left(\frac{1}{c}\right)}.$$

Preuve. D'après le théorème 3.4.1 et l'expression (3.55) de T_n^*, nous avons pour tout $t \in \mathbb{R}$,

$$\sum_{k=0}^{n} a_k^* e^{2ik\ell^* t} = T_n^*(e^{2i\ell^* t}) = \frac{1}{T_n(\frac{1}{c})} e^{in\ell^* t} \, T_n\left(\frac{1}{c}\cos(\ell^* t)\right).$$

□

3.4.6 Exemple illustratif

Considérons un instrument avec huit pupilles ($n = 7$) alignées réparties périodiquement et optimisons la dynamique de la réponse impulsionnelle normalisée sur l'intervalle d'optimisation $I = [\pi/2, 3\pi/2]$. Nous obtenons un écart optimal égal à un ($\ell^* = 1$) et les amplitudes optimales des champs normalisées du tableau 3.3.

k	a_k^*
0	0.10
1	0.36
2	0.73
3	1.00
4	1.00
5	0.73
6	0.36
7	0.10

TAB. 3.3 – Valeurs des amplitudes optimales des champs normalisées pour huit pupilles ($n = 7$) alignées réparties périodiquement avec $I = [\pi/2, 3\pi/2]$.

Ces valeurs sont positives et symétriques, i.e. $a_0 = a_7, a_1 = a_6, \ldots$. Elles sont identiques à celles du tableau 2.13 du chapitre 2 lorsque les amplitudes des champs et les positions des pupilles sont optimisées avec le problème de l'optimisation de la dynamique écrit avec la norme $\|\cdot\|_\infty$ pour la configuration spatiale de l'instrument. La figure 3.4 représente la réponse impulsionnelle normalisée optimale associée à cet exemple. La valeur de $\phi_2(\ell^*)$ est ici égale à $1.75.10^{-5}$.

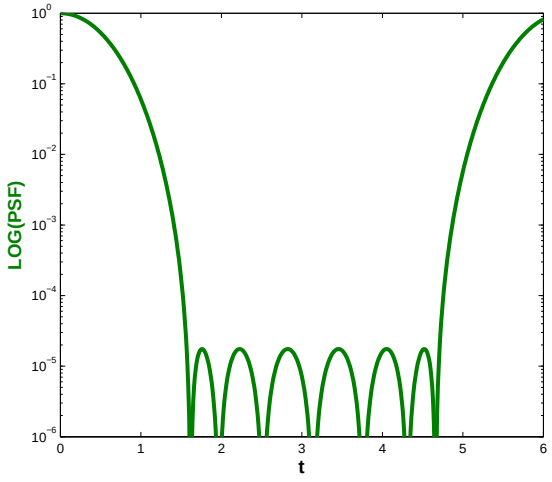


FIG. 3.4 – PSF temporelle optimale pour huit pupilles ($n = 7$) alignées réparties périodiquement avec $I = [\pi/2, 3\pi/2]$.

Chapitre 4

Conclusions et perspectives

Dans cette partie, une application en astronomie a été présentée. L'instrument optique considéré est un hypertélescope [41], i.e. un réseau de plusieurs télescopes. Un des objectifs de cet appareil d'observation est de réaliser des images d'exoplanètes [43]. Celles-ci sont très difficiles à détecter car leurs intensités lumineuses sont très faibles par rapport aux intensités lumineuses des étoiles autour desquelles elles gravitent. Pour arriver à détecter ce type d'objet, il est donc indispensable de maximiser l'efficacité de l'instrument optique. Pour cela, il faut optimiser les positions des télescopes et les amplitudes des champs reçues par chacun d'eux afin que la répartition de l'intensité lumineuse (aussi appelée réponse impulsionnelle) de l'étoile dans le plan image présente un grand pouvoir de résolution et une grande valeur de dynamique.

Pour résoudre le problème, nous avons considéré un réseau aligné de télescopes. Numériquement, deux méthodes d'optimisation ont été proposées. L'approche la plus efficace a consisté à modéliser le problème sous la forme d'un problème d'optimisation non linéaire. Le but était de minimiser la hauteur des lobes secondaires de la réponse impulsionnelle de l'étoile sur un intervalle précis. En effet, le lobe central de la réponse impulsionnelle de l'exoplanète doit être observable sur cet intervalle. Les valeurs de la dynamique et de la résolution ont alors été optimisées. Le meilleur compromis entre les deux critères a été obtenu pour une optimisation simultanée des positions des télescopes et des amplitudes des champs reçues par chacun d'eux. La configuration optimale de l'instrument correspond à des amplitudes de champs apodisées et à un positionnement quasi-périodique des ouvertures. Des perturbations ont été par la suite ajoutées dans le modèle mathématique afin de rendre l'étude réaliste et de tenir compte des conditions expérimentales. Ces perturbations ont essentiellement une influence sur la valeur de la dynamique. D'après l'étude numérique, l'usage d'un nombre restreint de télescopes est envisageable pour détecter des exoplanètes. Théoriquement, nous avons étudié l'existence d'une solution du problème de l'optimisation de la dynamique écrit avec les normes $\|\cdot\|_2$ et $\|\cdot\|_\infty$ lorsque les positions des télescopes sont réparties de manière périodique. Pour la norme $\|\cdot\|_2$, étudier l'existence d'une solution du problème revient à étudier une fonction. L'existence d'une solution n'a pas été démontrée mais les résultats obtenus ont tout de même été présentés. Pour la norme $\|\cdot\|_\infty$, une formule explicite de la solution optimale a été obtenue. L'écart optimal entre chaque télescope dépend de

l'intervalle d'optimisation et les amplitudes optimales des champs sont apodisées.

Concernant ce projet de recherche, plusieurs perspectives sont à envisager :

- comparer les résultats numériques obtenus avec les résultats expérimentaux ;
- étudier le cas bidimensionnel ;
- optimiser une configuration donnée pour des projets ou instruments existants ;
- définir des critères d'optimisation dépendant du type d'objets à observer ;
- démontrer l'existence de la solution du problème de l'optimisation de la dynamique modélisé avec la norme $\|\cdot\|_2$;
- écrire les conditions d'optimalité du problème de l'optimisation de la dynamique modélisé avec la norme $\|\cdot\|_\infty$;
- démontrer l'existence de la solution du problème de l'optimisation de la dynamique pour un positionnement quelconque des télescopes.

Deuxième partie

Méthode primale-duale pour l'optimisation avec contraintes d'égalité

Chapitre 1

Rappels sur les méthodes de pénalisation quadratique et SQP

1.1 Introduction

Nous nous intéressons à la résolution d'un problème d'optimisation non linéaire de la forme :

$$\begin{array}{ll} \text{minimiser}_{x\in\mathbb{R}^n} & f(x) \\ \text{sous contrainte} & c(x) = 0, \end{array} \tag{1.1}$$

où $f : \mathbb{R}^n \to \mathbb{R}$ et $c : \mathbb{R}^n \to \mathbb{R}^m$ $(1 \leq m \leq n)$ sont des fonctions deux fois continument différentiables.

Dans les années 70, une des premières méthodes utilisées pour résoudre des problèmes d'optimisation avec contraintes était de remplacer le problème initial (1.1) par un problème sans contraintes de la forme :

$$\text{minimiser}_{x\in\mathbb{R}^n} \quad f(x) + \pi(x),$$

où $\pi : \mathbb{R}^n \to \mathbb{R}$ est une fonction pénalisant la violation des contraintes. Le nouveau critère à minimiser est appelé fonction de pénalité ou fonction de pénalisation [23].

Une autre méthode classique et très efficace pour la résolution des problèmes d'optimisation non linéaires avec contraintes est la méthode de programmation quadratique successive (SQP) [10, 68]. L'idée générale de la méthode est de linéariser les conditions d'optimalité du problème initial (1.1) et d'exprimer le système linéaire qui en résulte sous une forme propice au calcul. L'intérêt d'une telle linéarisation est d'avoir un algorithme de convergence rapide. La méthode SQP transforme ainsi le problème (1.1) en une suite de sous-problèmes quadratiques plus simples à résoudre.

Les deux méthodes citées précédemment sont rappelées dans ce chapitre. Elles sont largement décrites dans la littérature, notamment dans [10, 11, 23, 52].

1.2 Définitions, notations et hypothèses générales

Le produit scalaire euclidien de deux vecteurs $x, y \in \mathbb{R}^n$ est défini par

$$x^\top y := \sum_{i=1}^{n} x_i y_i,$$

où x_i est la $i^{\text{ème}}$ composante du vecteur x. La norme associée est définie par

$$||x|| := (x^\top x)^{\frac{1}{2}}. \tag{1.2}$$

La boule ouverte centrée en $x \in \mathbb{R}^n$ de rayon $r > 0$ est définie par

$$B(x, r) := \{y \in \mathbb{R}^n : ||x - y|| < r\}.$$

Le lagrangien du problème (1.1) est la fonction $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ définie par

$$\mathcal{L}(x, \lambda) := f(x) + \lambda^\top c(x), \tag{1.3}$$

où $\lambda \in \mathbb{R}^m$ est le vecteur des multiplicateurs de lagrange associé aux contraintes d'égalité. Nous noterons $w := (x, \lambda) \in \mathbb{R}^{n+m}$, le vecteur des variables primales-duales du problème (1.1).

Notons X l'ensemble des solutions admissibles du problème (1.1) :

$$X := \{x \in \mathbb{R}^n : c(x) = 0\}.$$

Un minimum global du problème (1.1) est un point $x^* \in X$, minimisant f sur l'ensemble admissible X :

$$\forall x \in X, \quad f(x^*) \leq f(x).$$

Un minimum local de (1.1) est un point x^* admissible, minimisant f localement sur l'ensemble admissible X :

$$\exists\, r > 0, \forall\, x \in B(x^*, r) \cap X, \quad f(x^*) \leq f(x).$$

Les conditions nécessaires du premier ordre du problème (1.1) sont

$$\begin{cases} \nabla f(x) + A(x)^\top \lambda = 0 \\ c(x) = 0, \end{cases} \tag{1.4}$$

où $A(x) := \nabla c(x)^\top$ est la jacobienne des contraintes de dimension $m \times n$.
Un couple $w^* := (x^*, \lambda^*) \in \mathbb{R}^n \times \mathbb{R}^m$ qui vérifie le système (1.4) est appelé solution primale-duale du problème (1.1) et $x^* \in \mathbb{R}^n$ est dit stationnaire.

La jacobienne de (1.4) par rapport à $w \in \mathbb{R}^{n+m}$ est définie par

$$\begin{pmatrix} \nabla^2_{xx}\mathcal{L}(w) & A(x)^\top \\ A(x) & 0 \end{pmatrix}, \tag{1.5}$$

où $\nabla^2_{xx}\mathcal{L}(w)$ est le hessien du lagrangien.

Soient deux suites de nombres positifs $\{a_k\}$ et $\{b_k\}$ qui convergent vers zéro.

- On dit que a_k est un petit o de b_k et on note $a_k := o(b_k)$ si $\lim\limits_{k \to +\infty} \dfrac{a_k}{b_k} = 0$.
- On dit que a_k est un grand O de b_k et on note $a_k := O(b_k)$ s'il existe une constante $c > 0$ telle que $a_k \leq c b_k$ pour tout k suffisamment grand.

Soit $\mathcal{M}(\mathbb{R})$ l'ensemble des matrices carrées symétriques à coefficients réels. Soient A et $B \in \mathcal{M}(\mathbb{R})$. La matrice identité est notée Id. On appelle inertie de la matrice A le triplet $i(A) := (i_+(A), i_-(A), i_0(A))$ où $i_+(A)$ est le nombre de valeurs propres positives de A, $i_-(A)$ est le nombre de valeurs propres négatives de A et $i_0(A)$ est le nombre de valeurs propres nulles de A.

Deux matrices A et B sont dites congruentes s'il existe une matrice $P \in \mathcal{M}(\mathbb{R})$ inversible telle que $A = P^\top BP$.

D'après la loi d'inertie de Sylvester (Théorème 4.5.8 dans [37]), deux matrices sont congruentes si et seulement si elles ont la même inertie.

D'après le Théorème 16.3. dans [52], en notant $K(w)$ la matrice définie par (1.5) et en supposant que la jacobienne des contraintes $A(x)$ est de plein rang, nous avons

$$i(K(w)) = i(Z(x)^\top \nabla^2_{xx}\mathcal{L}(w)Z(x)) + (m, m, 0),$$

où $Z(x)$ est une matrice de dimension $n \times (n-m)$ dont les colonnes forment une base du noyau de $A(x)$. La matrice $Z(x)^\top \nabla^2_{xx}\mathcal{L}(w)Z(x)$ est définie positive si et seulement si $i(K(w)) = (n, m, 0)$.

Pour la suite, nous considérons les hypothèses suivantes.

Hypothèse 1.2.1 *Le problème (1.1) admet un minimum local $x^* \in \mathbb{R}^n$.*

Hypothèse 1.2.2 *Les fonctions f et c sont deux fois continument différentiables dans un voisinage de x^* et leurs dérivées secondes sont continument lipschitziennes.*

Hypothèse 1.2.3 *La jacobienne $A(x^*)$ est de rang m.*

Ces hypothèses impliquent qu'il existe un unique vecteur des multiplicateurs de lagrange $\lambda^* \in \mathbb{R}^m$ tel que $w^* := (x^*, \lambda^*) \in \mathbb{R}^{n+m}$ est solution du système (1.4).

Hypothèse 1.2.4 *Les conditions du second ordre sont satisfaites au point w^*, i.e. $u^\top \nabla^2_{xx}\mathcal{L}(w^*)u > 0$ pour tout $u \neq 0$ tel que $A(x^*)u = 0$.*

1.3 Méthode de pénalisation quadratique

Dans ce manuscrit, nous présentons la méthode de pénalisation quadratique introduite pour la pemière fois par R. Courant [18]. La fonction de pénalisation quadratique est définie par

$$Q(x, \mu) := f(x) + \frac{1}{2\mu}||c(x)||^2, \tag{1.6}$$

où $\mu > 0$ est le paramètre de pénalité.

La méthode de pénalisation quadratique consiste à résoudre une suite de problèmes sans contraintes de la forme :

$$\text{minimiser}_{x\in\mathbb{R}^n} \quad Q(x, \mu), \tag{1.7}$$

pour des valeurs de μ qui tendent vers zéro. En faisant tendre le paramètre de pénalité vers zéro, la violation des contraintes est pénalisée sévèrement ce qui force le minimum du problème (1.7) à converger vers le minimum du problème (1.1).

Les conditions nécessaires du premier ordre du problème (1.7) sont

$$\nabla_x Q(x, \mu) = 0, \tag{1.8}$$

où $\nabla_x Q(x, \mu) := \nabla f(x) + \frac{1}{\mu} A(x)^\top c(x)$.

1.3.1 Algorithme local

Présentons l'algorithme local utilisé pour résoudre le problème (1.7). Soient $\{\mu_k\}$ et $\{\varepsilon_k\}$ deux suites de nombres positifs qui tendent vers zéro. Soit $x_0^d \in \mathbb{R}^n$, un point de départ de l'algorithme.

Algorithme A : méthode de pénalisation quadratique

Pour $k = 0, 1, 2, \ldots$
Trouver un minimum approché x_k de $Q(., \mu_k)$ en partant du point de départ x_k^d et s'arrêter lorsque

$$||\nabla_x Q(x, \mu_k)|| \leq \varepsilon_k.$$

Si un test final de convergence est satisfait
alors s'arrêter à la solution approchée x_k,
Fin
Choisir un nouveau paramètre de pénalité $\mu_{k+1} < \mu_k$.
Choisir un nouveau point de départ x_{k+1}^d.
Fin

La suite du paramètre de pénalité $\{\mu_k\}$ est adaptée suivant la difficulté à minimiser la fonction de pénalité à chaque itération. Numériquement, si l'algorithme A a besoin de beaucoup d'itérations pour minimiser la fonction de pénalisation alors il faut choisir une faible décroissance de μ et inversement pour un nombre restreint d'itérations. Pour globaliser l'algorithme A, une méthode de recherche linéaire ou de régions de confiance peut être utilisée.

1.3.2 Théorèmes de convergence

Dans cette partie, les propriétés de convergence de la méthode de pénalisation quadratique sont rappelées à l'aide de deux théorèmes. Dans le premier théorème, nous supposons que la fonction de pénalité $Q(x, \mu_k)$ admet un minimum pour chaque valeur de μ_k.

Théorème 1.3.1 (Théorème 17.1. dans [52])
Supposons que le problème (1.1) admette une solution, que chaque solution x_k soit le minimum global de $Q(\cdot, \mu_k)$ dans l'algorithme A et que la suite $\{\mu_k\}$ tende vers zéro. Alors, tout point limite $x^* \in \mathbb{R}^n$ de la suite $\{x_k\}$ est un minimum global du problème (1.1).

Preuve. Soit $\bar{x} \in \mathbb{R}^n$ un minimum global du problème (1.1). Comme x_k minimise $Q(., \mu_k)$ pour chaque k, nous avons

$$Q(x_k, \mu_k) \leq Q(\bar{x}, \mu_k),$$

ce qui implique l'inégalité

$$f(x_k) + \frac{1}{2\mu_k}||c(x_k)||^2 \leq f(\bar{x}) + \frac{1}{2\mu_k}||c(\bar{x})||^2 = f(\bar{x}). \tag{1.9}$$

En arrangeant cette expression, nous obtenons

$$||c(x_k)||^2 \leq 2\mu_k(f(\bar{x}) - f(x_k)). \tag{1.10}$$

Supposons que $x^* \in \mathbb{R}^n$ est un point limite de $\{x_k\}$. Il existe alors un sous-ensemble $\mathcal{K} \subset \mathbb{N}$ tel que

$$\lim_{\substack{k \to +\infty \\ k \in \mathcal{K}}} x_k = x^*.$$

En prenant la limite lorsque $k \to +\infty$, $k \in \mathcal{K}$, dans l'inégalité (1.10) nous avons

$$||c(x^*)||^2 = \lim_{\substack{k \to +\infty \\ k \in \mathcal{K}}} ||c(x_k)||^2 \leq \lim_{\substack{k \to +\infty \\ k \in \mathcal{K}}} 2\mu_k(f(\bar{x}) - f(x_k)) = 0.$$

Nous avons alors $c(x^*) = 0$, i.e. x^* est un point admissible. D'après les positivités de μ_k et de $||c(x_k)||^2$, nous obtenons

$$f(x_k) \leq f(x_k) + \frac{1}{2\mu_k}||c(x_k)||^2.$$

D'après l'inégalité (1.9), nous en déduisons que $f(x_k) \leq f(\bar{x})$ et en passant à la limite lorsque $k \to +\infty$ pour $k \in \mathcal{K}$, nous avons

$$f(x^*) \leq f(\bar{x}).$$

Donc, x^* est aussi un minimum global du problème (1.1).

□

Comme ce résultat exige de trouver le minimum global de chaque sous-problème, la propriété de convergence vers la solution globale du problème (1.1) est difficile à obtenir. Le prochain résultat concerne les propriétés de convergence de la suite $\{x_k\}$. À la différence du théorème 1.3.1, ce théorème montre que la suite des itérés peut converger vers des points non admissibles ou vers des points stationnaires. Il montre également que le rapport $c_i(x_k)/\mu_k$ peut être utilisé pour estimer les multiplicateurs de lagrange λ_i^* pour $i = 1, \ldots, m$ dans certaines circonstances. Dans ce théorème, nous supposons que le test d'arrêt $||\nabla_x Q(x, \mu_k)|| \leq \varepsilon_k$ de l'algorithme A est satisfait pour tout k.

Théorème 1.3.2 (Théorème 17.2. dans [52])
Soient $\{\varepsilon_k\}$ et $\{\mu_k\}$ deux suites de nombres positifs qui convergent vers zéro. Tout point limite de la suite $\{x_k\}$ est un point stationnaire de la mesure d'admissibilité

$||c(\cdot)||^2$. *De plus, si* $A(x^*)$ *est de plein rang, alors* x^* *est admissible. Pour ces points et pour tout ensemble* $\mathcal{K} \subset \mathbb{N}$ *tel que* $\lim\limits_{\substack{k\to+\infty \\ k\in\mathcal{K}}} x_k = x^*$, *nous avons pour tout* $i \in \{1, \ldots, m\}$

$$\lim_{\substack{k\to+\infty \\ k\in\mathcal{K}}} \frac{1}{\mu_k} c_i(x_k) = \lambda_i^*, \tag{1.11}$$

où $\lambda^* \in \mathbb{R}^m$ *est le vecteur des multiplicateurs de lagrange.*

Preuve. En utilisant l'égalité (1.8) et le test d'arrêt de l'algorithme A, nous avons

$$\left|\left|\nabla f(x_k) + \frac{1}{\mu_k} A(x_k)^\top c(x_k)\right|\right| \leq \varepsilon_k. \tag{1.12}$$

En utilisant l'inégalité $||a|| - ||b|| \leq ||a + b||$, nous obtenons

$$\left|\left|A(x_k)^\top c(x_k)\right|\right| \leq \mu_k(\varepsilon_k + ||\nabla f(x_k)||). \tag{1.13}$$

Soit $x^* \in \mathbb{R}^n$ un point limite de la suite des itérés $\{x_k\}$. Il existe alors un sous-ensemble $\mathcal{K} \subset \mathbb{N}$ tel que $\lim\limits_{\substack{k\to+\infty \\ k\in\mathcal{K}}} x_k = x^*$. En prenant la limite lorsque $k \to +\infty$ pour $k \in \mathcal{K}$ dans l'inégalité (1.13), comme μ_k et ε_k tendent vers zéro, le terme de droite de l'inégalité tend lui aussi vers zéro. Il en résulte

$$A(x^*)^\top c(x^*) = 0, \tag{1.14}$$

donc x^* est un point stationnaire de la fonction $||c(\cdot)||^2$.
Si $A(x^*)$ est de plein rang, alors $c(x^*) = 0$, donc x^* est un point admissible. Ceci implique que la deuxième équation de (1.4) est vérifiée. Regardons si la première équation de ce système l'est aussi. En introduisant

$$\lambda_k := \frac{1}{\mu_k} c(x_k),$$

d'après l'égalité (1.8), nous avons

$$A(x_k)^\top \lambda_k = \nabla_x Q(x_k, \mu_k) - \nabla f(x_k). \tag{1.15}$$

Pour tout $k \in \mathcal{K}$ suffisamment grand, la matrice $A(x_k)$ est de plein rang. Donc, $A(x_k)A(x_k)^\top$ est non singulière. En multipliant les membres de l'égalité (1.15) par $A(x_k)$ et en arrangeant l'expression, nous avons

$$\lambda_k = [A(x_k)A(x_k)^\top]^{-1} A(x_k)[\nabla_x Q(x_k, \mu_k) - \nabla f(x_k)].$$

Ainsi, en utilisant l'inégalité (1.12) et en prenant la limite lorsque $k \to +\infty$, pour $k \in \mathcal{K}$, nous obtenons

$$\lim_{\substack{k\to+\infty \\ k\in\mathcal{K}}} \lambda_k = \lambda^* = -[A(x^*)A(x^*)^\top]^{-1} A(x^*)\nabla f(x^*).$$

En prenant la limite dans l'inégalité (1.12), il en résulte

$$\nabla f(x^*) + A(x^*)^\top \lambda^* = 0,$$

i.e. λ^* satisfait la première équation du système (1.4). Ainsi, (x^*, λ^*) vérifie les conditions nécessaires du premier ordre du problème (1.1).

□

Ce résultat montre que si un point limite $x^* \in \mathbb{R}^n$ n'est pas admissible, alors il est au moins un point stationnaire de la fonction $||c(\cdot)||^2$. Si le problème non linéaire (1.1) n'est pas réalisable, alors la méthode de pénalisation quadratique peut converger vers des points stationnaires ou des minima de $||c(\cdot)||^2$.

1.3.3 Mauvais conditionnement et reformulation

Numériquement, la méthode de pénalisation quadratique présente un inconvénient. En effet, la minimisation de $Q(\cdot, \mu)$ devient en général délicate à partir du moment où la valeur du paramètre de pénalité μ est proche de zéro. Son hessien défini par

$$\nabla^2_{xx} Q(x, \mu) := \nabla^2 f(x) + \frac{1}{\mu} \sum_{i=1}^{m} c_i(x) \nabla^2 c_i(x) + \frac{1}{\mu} A(x)^\top A(x), \tag{1.16}$$

devient mal conditionné proche de la solution et peut alors entraîner des difficultés numériques pour le calcul du pas de Newton obtenu en résolvant le système

$$\nabla^2_{xx} Q(x, \mu) d^x = -\nabla_x Q(x, \mu). \tag{1.17}$$

Supposons que x soit proche du minimum de $Q(., \mu)$, que μ soit proche de zéro et que les conditions du théorème 1.3.2 soient satisfaites. D'après la formule (1.11), nous avons

$$\nabla^2_{xx} Q(x, \mu) \approx \nabla^2_{xx} \mathcal{L}(x, \lambda^*) + \frac{1}{\mu} A(x)^\top A(x). \tag{1.18}$$

Cette expression montre que certaines valeurs propres du hessien $\nabla^2_{xx} Q(x, \mu)$ sont de l'ordre de $1/\mu$ ce qui entraîne un mauvais conditionnement de la matrice lorsque μ tend vers zéro. Illustrons ce phénomène par un exemple.

Exemple 1.3.3 Soient $x := (x_1, x_2) \in \mathbb{R}^2$ et le problème

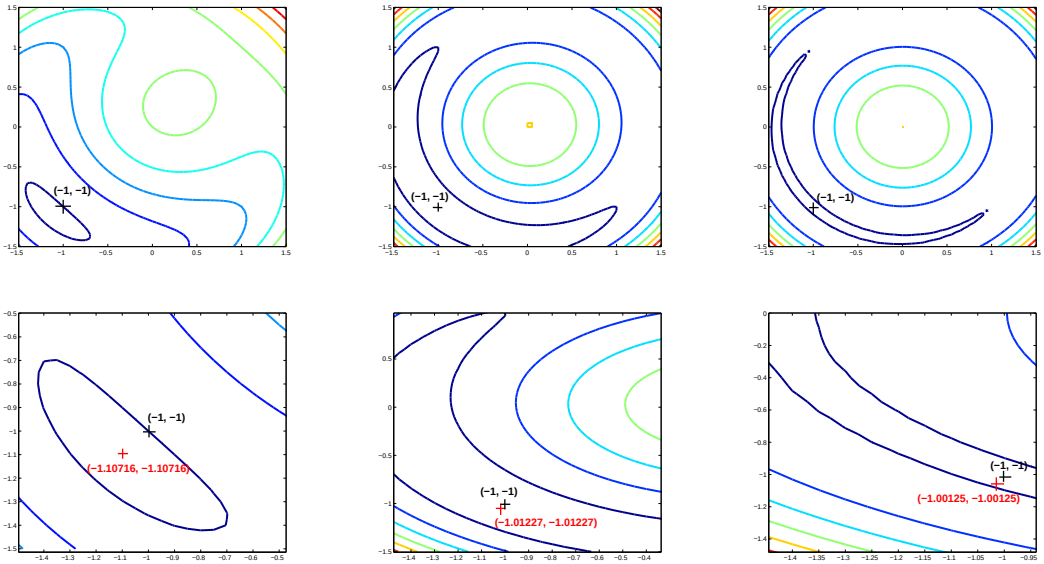
$$\begin{array}{ll} \text{minimiser}_{x \in \mathbb{R}^2} & x_1 + x_2 \\ \text{sous contrainte} & x_1^2 + x_2^2 - 2 = 0. \end{array} \tag{1.19}$$

La solution du problème (1.19) est $x^* = (-1, -1)^\top$. La fonction de pénalisation quadratique associée au problème est définie par

$$Q(x, \mu) := x_1 + x_2 + \frac{1}{2\mu}(x_1^2 + x_2^2 - 2)^2,$$

pour des valeurs de μ positives.

Les trois graphes du dessus de la figure 1.1 représentent le minimum x^* du problème (1.19) et les contours de Q pour trois valeurs du paramètre de pénalité : $\mu = 1, \mu = 10^{-1}$ et $\mu = 10^{-2}$. Plus la valeur de μ diminue, plus les courbes sont aplaties.



$\mu = 1$ $\mu = 10^{-1}$ $\mu = 10^{-2}$

FIG. 1.1 – Contours de Q pour différentes valeurs de μ.

Cela signifie que le problème est mal conditionné. Les trois graphes du dessous de la figure 1.1 représentent ces mêmes contours mais agrandis autour du minimum x^*. Notons que la fonction de pénalisation admet deux points stationnaires : un minimum x^*_μ en rouge et un maximum proche de $(0,0)^\top$. Lorsque μ tend vers zéro, x^*_μ converge vers x^* et le maximum converge vers $(0,0)^\top$ qui est un point stationnaire de la mesure d'admissibilité. Nous remarquons d'après le tableau 1.1 que x^*_μ converge vers x^* avec une vitesse de convergence égale à la vitesse de décroissance de μ, qui est ici linéaire d'un facteur 10.

μ	x^*_μ	Conditionnement de $\nabla^2_{xx}Q(x^*_\mu,\mu)$
1	$(-1.10716, -1.10716)$	11.9
10^{-1}	$(-1.01227, -1.01227)$	84.0
10^{-2}	$(-1.00125, -1.00125)$	802.5

TAB. 1.1 – Valeurs du minimum x^*_μ de Q et du conditionnement de $\nabla^2_{xx}Q(x^*_\mu,\mu)$ pour différentes valeurs de μ.

Pour chaque valeur de μ, le conditionnement de la matrice hessienne de Q évalué en x^*_μ a été calculé. Il est reporté dans le tableau 1.1. Plus le paramètre de pénalité diminue, plus le conditionnement de la matrice hessienne est mauvais.

Des formules alternatives [31, 52] existent afin d'éviter ces problèmes de mauvais conditionnement. En introduisant

$$\zeta := \frac{1}{\mu} A(x) d^x,$$

la direction $d^x \in \mathbb{R}^n$, solution du système (1.17) est aussi solution du système

$$\begin{pmatrix} \nabla^2 f(x) + \frac{1}{\mu}\sum_{i=1}^{m} c_i(x)\nabla^2 c_i(x) & A(x)^\top \\ A(x) & -\mu I \end{pmatrix} \begin{pmatrix} d^x \\ \zeta \end{pmatrix} = \begin{pmatrix} -\nabla_x Q(x,\mu) \\ 0 \end{pmatrix}. \quad (1.20)$$

Lorsque $x \in \mathbb{R}^n$ est proche de la solution $x^* \in \mathbb{R}^n$ et μ est proche de zéro, la matrice du système ne présente pas de valeurs singulières élevées (de l'ordre de $1/\mu$). Le système (1.20) est donc une reformulation du système (1.17) avec un bon conditionnement.

Dans la méthode que nous proposons, le système linéaire qui est résolu à chaque itération a la même structure que le système (1.20). Il en est cependant différent sur trois points : le hessien de la fonction de pénalisation est remplacé par le hessien du lagrangien, dans le second membre le gradient du lagrangien remplace le gradient de la fonction de pénalisation et le zéro est remplacé par $c(x) - \mu\lambda$. Tous les détails sont donnés dans le chapitre suivant.

1.4 Méthode SQP

La méthode de programmation quadratique successive (SQP) est une technique générale pour résoudre des problèmes d'optimisation non linéaires avec contraintes. Le principe de la méthode est de linéariser les conditions d'optimalité (1.4) du problème (1.1) afin d'exprimer le système linéaire sous une forme propice aux calculs. L'intérêt de la linéarisation réside dans le fait que l'algorithme obtenu présente une convergence locale rapide. La méthode SQP transforme ainsi un problème d'optimisation non linéaire en une suite de problèmes quadratiques. Cette méthode a été proposée en 1963 dans la thèse de R.B Wilson [68]. Dans les années 70, la méthode a largement été développée notamment par U.M. Garcia-Palomares et O.L. Mangasarian [28], S.P. Han [34, 35] et M.J.D. Powell [58, 59, 60]. La recherche sur la méthode SQP se poursuit encore aujourd'hui surtout sur son utilisation dans la résolution des problèmes de grande taille. Elle intervient aussi comme un outil dans les méthodes de points intérieurs en programmation non linéaire.

1.4.1 Description de la méthode

La méthode SQP peut s'interpréter de deux manières.

– Interprétation 1 : Pour résoudre le problème (1.1), la méthode de Newton peut être appliquée sur les conditions d'optimalité (1.4). Le pas de Newton à l'itéré $w_k := (x_k, \lambda_k)$ est alors donné par la relation

$$\begin{pmatrix} x_{k+1} \\ \lambda_{k+1} \end{pmatrix} = \begin{pmatrix} x_k \\ \lambda_k \end{pmatrix} + \begin{pmatrix} d_k^x \\ d_k^\lambda \end{pmatrix}, \quad (1.21)$$

où d_k^x et d_k^λ sont les solutions du système

$$\begin{pmatrix} \nabla^2_{xx}\mathcal{L}(w_k) & A(x_k)^\top \\ A(x_k) & 0 \end{pmatrix} \begin{pmatrix} d_k^x \\ d_k^\lambda \end{pmatrix} = \begin{pmatrix} -\nabla f(x_k) - A(x_k)^\top \lambda_k \\ -c(x_k) \end{pmatrix}. \quad (1.22)$$

L'itération de Newton est bien définie lorsque la jacobienne du système (1.22) est inversible. Cette propriété est vraie si w_k est proche de la solution $w^* \in \mathbb{R}^{n+m}$ du problème (1.1) et si les hypothèses 1.2.3 et 1.2.4 sont vérifiées.

– Interprétation 2 : Le système de Newton (1.22) peut s'interpréter comme les conditions d'optimalité d'un sous-problème quadratique. À l'itéré (x_k, λ_k), la méthode SQP transforme le problème (1.1) en une suite de problèmes quadratiques de la forme :

$$\begin{array}{ll} \text{minimiser}_{d^x \in \mathbb{R}^n} & \nabla f(x_k)^\top d^x + \dfrac{1}{2}(d^x)^\top \nabla^2_{xx}\mathcal{L}(w_k)d^x \\ \text{sous contrainte} & A(x_k)d^x + c(x_k) = 0. \end{array} \tag{1.23}$$

Si le problème (1.23) admet une solution d^x_k, alors il existe un multiplicateur δ_k tel que le couple (d^x_k, δ_k) soit solution des conditions d'optimalité du premier ordre du problème (1.23) définies par

$$\left\{ \begin{array}{l} \nabla^2_{xx}\mathcal{L}(w_k)d^x_k + \nabla f(x_k) + A(x_k)^\top \delta_k = 0, \\ A(x_k)d^x_k + c(x_k) = 0. \end{array} \right.$$

Les vecteurs d^x_k et δ_k se définissent aussi à partir des solutions du système (1.22) qui peut s'écrire sous la forme :

$$\begin{pmatrix} \nabla^2_{xx}\mathcal{L}(w_k) & A(x_k)^\top \\ A(x_k) & 0 \end{pmatrix} \begin{pmatrix} d^x_k \\ \delta_k \end{pmatrix} = \begin{pmatrix} -\nabla f(x_k) \\ -c(x_k) \end{pmatrix}, \tag{1.24}$$

où $\delta_k = \lambda_k + d^\lambda_k = \lambda_{k+1}$. Le nouvel itéré (d^x_k, λ_{k+1}) est donc la solution du problème quadratique (1.23).

La résolution du problème quadratique (1.23) est préférable à la résolution du système (1.22) car en résolvant ce système, il est possible de converger vers un point stationnaire de (1.1) qui n'est pas un minimum. Il y a équivalence entre les deux interprétations si l'hypothèse 1.2.4 est satisfaite. En effet, l'application $d^x \mapsto \nabla f(x_k)^\top d^x + \dfrac{1}{2}(d^x)^\top \nabla^2_{xx}\mathcal{L}(w_k)d^x$ est quadratique strictement convexe sur l'espace affine $\{d^x : A(x_k)d^x + c(x_k) = 0\}$. Donc, le problème (1.23) admet une solution unique qui est solution du système (1.22). En revanche, si l'hypothèse 1.2.4 n'est pas vérifiée alors le problème de minimisation (1.23) peut avoir un point stationnaire (vérifiant le système (1.22)) mais pas de minimum.

1.4.2 Algorithmes local et global

Présentons les algorithmes local et global associés à la méthode SQP pour résoudre le problème d'optimisation non linéaire (1.1). Soit $w_0 := (x_0, \lambda_0) \in \mathbb{R}^{n+m}$ un point de départ de l'algorithme.

Algorithme B : algorithme SQP local

Evaluer $f(x_0), \nabla f(x_0), \nabla^2_{xx}\mathcal{L}(w_0), c(x_0)$ et $A(x_0)$.
Répéter jusqu'à ce qu'un test de convergence soit satisfait
Résoudre le problème (1.23) pour obtenir d_k^x et δ_k.
Faire $x_{k+1} = x_k + d_k^x$ et $\lambda_{k+1} = \delta_k$.
Évaluer $f(x_{k+1}), \nabla f(x_{k+1}), \nabla^2_{xx}\mathcal{L}(w_{k+1}), c(x_{k+1})$ et $A(x_{k+1})$.
Fin

Une technique de recherche linéaire ou de régions de confiance peut être utilisée pour globaliser l'algorithme B. Nous rappelons dans ce manuscrit uniquement la globalisation de la méthode avec une technique de recherche linéaire en utilisant la fonction de mérite

$$\theta_\sigma(x) := f(x) + \sigma||c(x)||_1, \tag{1.25}$$

appelée fonction de pénalisation de Han [34, 35] où $\sigma > 0$ est le paramètre de pénalité. Remarquons qu'une autre norme pourrait être utilisée dans la définition de la fonction de mérite.

Dans les méthodes de recherche linéaire, les itérés sont générés dans le cas le plus simple par la relation de récurrence

$$x_{k+1} := x_k + \alpha_k d_k^x,$$

où d_k^x est une direction de $\mathbb{R}^n$ et $\alpha_k > 0$ est un pas choisi de manière à faire décroître la fonction de mérite. Le pas α_k est accepté si la condition suivante est satisfaite :

$$\theta_{\sigma_k}(x_k + \alpha_k d_k^x) \leq \theta_{\sigma_k}(x_k) + \eta\alpha_k\theta'_\sigma(x_k; d_k^x),$$

où $\eta \in (0, 1)$ et $\theta'_\sigma(x_k; d_k^x)$ est la dérivée directionnelle de θ_σ en x_k dans la direction d_k^x. Cette inégalité correspond à la condition d'Armijo appliquée à la fonction θ_σ. À chaque itération, la direction d_k^x doit être une direction de descente de la fonction de mérite, i.e. $\theta'_\sigma(x_k; d_k^x) < 0$. Le théorème suivant donne l'expression de θ'_σ ainsi qu'une majoration qui permettra de déterminer les valeurs de σ pour lesquelles d_k^x est une direction de descente de θ_σ.

Théorème 1.4.1 (Théorème 18.2. dans [52])
Soit (d_k^x, λ_{k+1}) une solution du système (1.24). La dérivée directionnelle de θ_σ en x_k dans la direction d_k^x satisfait

$$\theta'_\sigma(x_k; d_k^x) = \nabla f(x_k)^\top d_k^x - \sigma||c(x_k)||_1. \tag{1.26}$$

De plus,

$$\theta'_\sigma(x_k; d_k^x) \leq -(d_k^x)^\top \nabla^2_{xx}\mathcal{L}(x_k, \lambda_k) d_k^x - (\sigma - ||\lambda_{k+1}||_\infty)||c(x_k)||_1. \tag{1.27}$$

Preuve. Soit $p(x) := ||c(x)||_1 := (n \circ c)(x)$ où l'application $n : \mathbb{R}^m \to \mathbb{R}$ est définie pour $y \in \mathbb{R}^m$, par $n(y) := ||y||_1$. La fonction de mérite θ_σ s'écrit alors :

$$\theta_\sigma(x) = f(x) + \sigma p(x).$$

La dérivée directionnelle de θ_σ en x_k dans la direction d_k^x est alors :

$$\theta'_\sigma(x_k; d_k^x) = \nabla f(x_k)^\top d_k^x + \sigma p'(x_k; \sigma_k). \tag{1.28}$$

D'après la définition de p et le Lemme 9.1 de [11], la dérivée directionnelle de p en x_k dans la direction d_k^x est

$$p'(x_k; d_k^x) = n'(c(x_k); A(x_k).d_k^x).$$

Si d_k^x vérifie l'égalité $A(x_k)d_k^x + c(x_k) = 0$, alors nous obtenons

$$\begin{aligned} p'(x_k; d_k^x) &= n'(c(x_k); -c(x_k)) \\ &= \lim_{t\to 0} \frac{n(c(x_k) - tc(x_k)) - n(c(x_k))}{t} \\ &= \lim_{t\to 0} \frac{(1-t)||c(x_k)||_1 - ||c(x_k)||_1}{t} \\ &= \lim_{t\to 0} -||c(x_k)||_1 \\ &= -||c(x_k)||_1. \end{aligned}$$

L'égalité (1.28) s'écrit alors :

$$\theta'_\sigma(x_k; d_k^x) = \nabla f(x_k)^\top d_k^x - \sigma||c(x_k)||_1,$$

soit comme l'égalité (1.26). Comme d_k^x vérifie la première équation du système (1.24), nous avons

$$\theta'_\sigma(x_k; d_k^x) = -(d_k^x)^\top \nabla^2_{xx}\mathcal{L}(w_k)d_k^x + (d_k^x)^\top A(x_k)^\top \lambda_{k+1} - \sigma||c(x_k)||_1.$$

D'après la deuxième équation du système (1.24), le terme $A(x_k)d_k^x$ peut être remplacé par $-c(x_k)$. En faisant la substitution dans l'expression précédente et en utilisant l'inégalité de Cauchy-Schwartz, nous avons

$$-c(x_k)^\top \lambda_{k+1} \leq ||c(x_k)||_1||\lambda_{k+1}||_\infty,$$

soit l'inégalité (1.27).

□

L'inégalité (1.27) montre que d_k^x est une direction de descente de θ_σ si $\nabla^2_{xx}\mathcal{L}(w_k)$ est définie positive, $d_k^x \neq 0$ et si

$$\sigma > ||\lambda_{k+1}||_\infty. \tag{1.29}$$

Afin de déterminer la nouvelle valeur du paramètre de pénalité σ dans θ_σ à chaque itération, une stratégie possible est d'augmenter sa valeur précédente pour que l'inégalité (1.29) soit satisfaite. En pratique, nous n'avons pas programmé cette technique parce que la matrice hessienne $\nabla^2_{xx}\mathcal{L}(w_k)$ n'est pas définie positive. Nous avons utilisé une technique décrite dans [52] qui consiste à étudier l'effet du pas sur le modèle quadratique de la fonction de mérite. Le modèle quadratique de θ_σ est défini par

$$q_\sigma(d^x) := f(x_k) + \nabla f(x_k)^\top d^x + \frac{\rho}{2}(d^x)^\top \nabla^2_{xx}\mathcal{L}(w_k)d^x + \sigma m(d^x), \tag{1.30}$$

où

$$m(p) := ||c(x_k) + A(x_k)p||_1,$$

et où ρ est défini par

$$\rho := \begin{cases} 1 & \text{si } (d^x)^\top \nabla^2_{xx}\mathcal{L}(w_k)d^x > 0, \\ 0 & \text{sinon.} \end{cases}$$

Après avoir calculé la direction d_k^x, le paramètre de pénalité σ est choisi tel que

$$q_\sigma(0) - q_\sigma(d_k^x) \geq \varsigma\sigma(m(0) - m(d_k^x)), \tag{1.31}$$

où $\varsigma \in (0,1)$. D'après les formules (1.30) et (1.23), l'inégalité (1.31) est satisfaite lorsque

$$\sigma \geq \frac{\nabla f(x_k)^\top d_k^x + \frac{\rho}{2}(d_k^x)^\top \nabla^2_{xx}\mathcal{L}(w_k)d_k^x}{(1-\varsigma)||c(x_k)||_1}. \tag{1.32}$$

Si à l'itération k, la valeur de σ satisfait l'inégalité (1.32), alors sa valeur n'est pas modifiée. Sinon, elle est augmentée jusqu'à ce que l'inégalité soit satisfaite. Il est facile de vérifier que si σ satisfait l'inégalité (1.32), alors le choix de ρ assure que $\theta'_\sigma(x_k; d_k^x) \leq -\varsigma\sigma||c(x_k)||_1$, soit que d_k^x est une direction de descente de θ_σ.

À présent, donnons l'algorithme global associé à la méthode SQP. Soient $\eta \in (0, 0.5), \varsigma \in (0,1), \sigma_0 = 1$ et un point de départ $w_0 \in \mathbb{R}^{n+m}$.

Algorithme C : algorithme SQP global

Evaluer $f(x_0), \nabla f(x_0), c(x_0)$ et $A(x_0)$.

Répéter jusqu'à ce qu'un test de convergence soit satisfait

Calculer d_k^x en résolvant le problème (1.23) avec $\hat{\lambda}$ le multiplicateur associé.

Faire $d_k^\lambda = \hat{\lambda} - \lambda_k$.

Choisir $\sigma_k \geq \sigma_{k-1}$ tel que

$$\sigma_k \geq \frac{\nabla f(x_k)^\top d_k^x + \frac{\rho}{2}(d_k^x)^\top \nabla^2_{xx}\mathcal{L}(w_k)d_k^x}{(1-\varsigma)||c(x_k)||_1}.$$

Faire $\alpha_k = 1$.

Tant que $\theta_{\sigma_k}(x_k + \alpha_k d_k^x) > \theta_{\sigma_k}(x_k) + \eta\alpha_k\theta'_{\sigma_k}(x_k; d_k^x)$

Faire $\alpha_{k+1} = \dfrac{\alpha_k}{2}$.

Fin

Faire $x_{k+1} = x_k + \alpha_k d_k^x$ et $\lambda_{k+1} = \lambda_k + \alpha_k d_k^\lambda$.

Évaluer $f(x_{k+1}), \nabla f(x_{k+1}), \nabla^2_{xx}\mathcal{L}(w_{k+1}), c(x_{k+1})$ et $A(x_{k+1})$.

Fin

Remarque 1.4.2 Pour que la méthode de globalisation soit efficace, il faut que $\alpha_k = 1$ proche d'une solution du problème (1.1) afin de retrouver la convergence quadratique de la méthode locale. L'admissibilité asymptotique du pas unité n'est

pas toujours vérifiée. En effet, lorsque la direction d_k^x est solution du problème quadratique (1.23), il est possible d'avoir

$$\theta_\sigma(x_k + d_k^x) > \theta_\sigma(x_k),$$

même lorsque x_k est proche de x^*. La raison de ce comportement est l'inadéquation entre la fonction de mérite θ_σ et le modèle quadratique utilisé pour calculer la direction d_k^x. Ceci est connu sous le nom d'effet Maratos [48]. Numériquement, cette difficulté peut être évitée en faisant des corrections du second ordre [24]. Dans les tests numériques que nous avons effectués, nous n'avons pas considéré de corrections de ce type.

1.4.3 Théorèmes de convergence

Tout d'abord, décrivons les conditions qui garantissent la convergence locale de la méthode SQP.

Théorème 1.4.3 (Théorème 18.4. dans [52])
Supposons que les hypothèses 1.2.1-1.2.4 soient vérifiées. Il existe un voisinage de $x^* \in \mathbb{R}^n$ tel que si $w_0 \in \mathbb{R}^{n+m}$ est dans un voisinage de $w^* \in \mathbb{R}^{n+m}$, alors l'itéré w_k généré par l'algorithme B converge quadratiquement vers w^*.

Ce théorème se démontre facilement car l'algorithme B est équivalent à la méthode de Newton appliquée au système non linéaire (1.4).

À présent, donnons le résultat de convergence associé à l'algorithme global de la méthode SQP. Pour ce résultat, nous supposons que le problème quadratique (1.23) est réalisable et que le paramètre de pénalité σ est fixé pour tout k suffisamment grand.

Théorème 1.4.4
Supposons que les suites $\{x_k\}$ et $\{x_k + d_k^x\}$ soient contenues dans un sous-ensemble convexe fermé borné de $\mathbb{R}^n$ pour lesquelles f et c sont deux fois continument différentiables. Supposons que le hessien du lagrangien et les multiplicateurs de lagrange soient bornés et que σ satisfait l'inégalité $\sigma \geq ||\lambda_k||_\infty + \upsilon$ pour tout k où υ est une constante positive. Alors, les points limites de la suite $\{x_k\}$ vérifient les conditions nécessaires du premier ordre (1.4) du problème non linéaire (1.1).

La conclusion du théorème est très intéressante mais les hypothèses exigées sont plutôt restrictives. Des résultats de convergence globale avec des hypothèses plus réalistes ont été établis notamment par A.R. Conn, N.I.M. Gould et P.L. Toint [17].

1.4.4 Implémentation de la méthode

En pratique, la méthode SQP est une méthode efficace. De nombreux codes utilisent cette méthode, notamment SNOPT [30], FILTERSQP [25] et KNITRO/ACTIVE [15]. Ces solveurs peuvent s'utiliser avec le langage de modélisation AMPL [27]. Pour comparer de manière objective l'algorithme primal-dual décrit dans le chapitre 2 à la méthode SQP, nous avons programmé sous MATLAB une variante de

l'algorithme C. Nous considérons l'interprétation 1 de la méthode, i.e. la résolution du système linéaire (1.22) avec la méthode de Newton à la place de la résolution du problème quadratique (1.23). Comme précisé dans [52], nous contrôlons l'inertie de la jacobienne de (1.4) dans le système (1.5) de sorte que la matrice hessienne réduite soit définie positive dans le plan tangent des contraintes. D'après le Théorème 16.3. dans [52], l'inertie de la jacobienne de (1.4) est égale à $(n, m, 0)$ si et seulement si le hessien réduit du lagrangien est défini positif. Pour la globalisation de l'algorithme, nous avons considéré comme dans l'algorithme C une technique de recherche linéaire en utilisant la fonction de mérite de Han définie en (1.25). Le réglage du paramètre de pénalisation est assuré par la formule (1.32). Une simple méthode de backtracking a été utilisée pour le calcul de la suite des pas α. On divise le pas par deux. Des tests numériques ont été effectués sur des problèmes provenant des librairies COPS 3.0 [21] et CUTEr [33]. Une étude détaillée de ces tests est présentée dans le chapitre 3. Nous précisons ci-dessous les valeurs des paramètres utilisées dans l'algorithme.

Valeurs des paramètres dans l'algorithme programmé

Le test d'arrêt de l'algorithme est

$$\left\| \begin{pmatrix} \nabla f(x) + A(x)^\top \lambda \\ c(x) \end{pmatrix} \right\| \leq 10^{-8}.$$

Pour le réglage du paramètre de pénalisation dans (1.32), nous considérons $\sigma_0 = 1$ et $\varsigma = 0.9$. Dans la globalisation de la méthode, nous supposons $\eta = 10^{-2}$. Comme point de départ $w_0 := (x_0, \lambda_0) \in \mathbb{R}^{n+m}$ de l'algorithme,

- x_0 est donné par le problème,
- λ_0 est solution au sens des moindres carrés du problème

$$\text{minimiser}_{\lambda \in \mathbb{R}^m} \quad ||\nabla f(x_0) + A(x_0)^\top \lambda||.$$

Chapitre 2

Présentation de la méthode primale-duale

2.1 Introduction

Comme nous l'avons vu dans le chapitre 1, une des premières méthodes utilisées pour résoudre le problème non linéaire (1.1) était de résoudre une suite de problèmes pénalisés de la forme (1.7) pour des valeurs du paramètre de pénalité qui tendent vers zéro. L'inconvénient de ces méthodes est le mauvais conditionnement du problème lorsque le paramètre de pénalité se rapproche de zéro.

Pour contourner ces problèmes de mauvais conditionnement, nous proposons une nouvelle approche qui consiste à résoudre le système primal-dual

$$\left\{ \begin{array}{l} \nabla f(x) + A(x)^\top \lambda = 0 \\ c(x) - \mu\lambda = 0, \end{array} \right. \tag{2.1}$$

avec une méthode Newtonnienne. En posant

$$\lambda := \frac{c(x)}{\mu}, \tag{2.2}$$

pour $\mu > 0$, le système (2.1) peut s'interpréter comme les conditions d'optimalité perturbées du problème non linéaire (1.1) ou comme les conditions d'optimalité du problème pénalisé (1.7). En notant $w := (x, \lambda) \in \mathbb{R}^{n+m}$ le couple des variables primales-duales du problème non linéaire (1.1) et $F : \mathbb{R}^{n+m+1} \to \mathbb{R}^{n+m}$ l'application définie par

$$F(w, \mu) = \begin{pmatrix} \nabla f(x) + A(x)^\top \lambda \\ c(x) - \mu\lambda \end{pmatrix},$$

le système (2.1) revient à résoudre

$$F(w, \mu) = 0. \tag{2.3}$$

La jacobienne de F par rapport à w est définie par

$$F'_w(w, \mu) := \begin{pmatrix} \nabla^2_{xx}\mathcal{L}(w) & A(x)^\top \\ A(x) & -\mu Id \end{pmatrix}. \tag{2.4}$$

L'algorithme que nous proposons consiste à appliquer la méthode de Newton au système (2.3) en w pour des valeurs de μ positives qui décroissent vers zéro. Sous certaines hypothèses et pour $\mu > 0$ suffisamment petit, la solution de (2.3) définit une courbe $\mu \mapsto w(\mu)$ appelée trajectoire ou chemin. Le point final de cette courbe noté $w^* := w(0)$ est solution du système (2.3), i.e. $F(w^*, 0) = 0$. L'inversibilité de $F'_w(w^*, 0)$ permet de montrer qu'il existe une boule $B(w^*, r^*)$ et un nombre $\mu^* > 0$ tels que la méthode de Newton converge pour toute solution de départ $(w_0, \mu_0) \in B(w^*, r^*) \times]0, \mu^*[$.

Les méthodes classiques héritées de la méthode SUMT (Sequential Unconstrained Minimization Techniques) de Fiacco et McCormick [23] résolvent de manière approchée les conditions d'optimalité perturbées (2.3) pour une suite décroissante de valeurs du paramètre de pénalité tendant vers zéro. Dans ces méthodes, les itérés doivent satisfaire une mesure de proximité du type

$$\|F(w_k, \mu_k)\| \leq \varepsilon_k,$$

où ε_k est du même ordre que μ_k. Ces méthodes sont très souvent utilisées notamment dans [32]. L'analyse de convergence proposée par P. Armand et J. Benoist dans [5] et par P. Armand, J. Benoist et D. Orban dans [6] pour la résolution des problèmes avec contraintes d'égalité et d'inégalité est différente des analyses classiques dans le sens où ils ne supposent pas que ε_k est du même ordre que μ_k. En effet, ils veulent éviter la résolution d'un problème barrière (même approximativement) car cela peut être très coûteux pour les premières valeurs de μ et les solutions approchées peuvent être assez loin de la solution du problème initial. Les théorèmes 1 dans [5] et 4.2 dans [6] pour des problèmes d'optimisation non linéaires avec contraintes d'égalité et d'inégalité sont transposables à notre étude. Ceci a une importance pratique car si un itéré est dans la boule de convergence de la méthode de Newton appliquée à l'équation $F(w, \mu) = 0$, alors la suite $\{w_k\}$ converge vers w^* et la suite est asymptotiquement tangente à la trajectoire pour une convergence de μ vers zéro au plus superlinéaire. Ce résultat nous a permis d'envisager une globalisation de la méthode primale-duale qui tient compte de ce comportement asymptotique. L'algorithme local est donc globalisé avec une technique de recherche linéaire en utilisant une fonction de mérite primale-duale. La technique de globalisation considérée dans la méthode primale-duale est similaire à celle proposée par P. Armand, J. Benoist et D. Orban dans [7] pour la résolution des problèmes avec contraintes d'égalité et d'inégalité. L'analyse de convergence est plus simple dans notre cas car il n'y a pas besoin de montrer que le pas unité est accepté et les itérés ne doivent pas satisfaire des contraintes d'inégalité. Il suffit de vérifier que le test d'arrêt dans la globalisation est satisfait. Cette nouvelle approche permet d'éviter le mauvais conditionnement et l'effet Maratos [48] présents dans les méthodes de pénalisation et SQP.

A. Forsgren et P.E. Gill dans [26] se sont également intéressés aux méthodes primales-duales pour résoudre des problèmes d'optimisation non linéaires avec contraintes d'égalité et d'inégalité. Leur méthode consiste à résoudre un sous-problème sans contraintes où le critère à minimiser est une fonction de pénalité qui contient à la fois les variables primales et duales du problème. Ils utilisent une méthode Newtonnienne et la globalisent avec une méthode de recherche linéaire. Ils utilisent

également un contrôle de l'inertie de leur matrice primale-duale. E.M. Gertz et P.E. Gill dans [29] considèrent la même fonction de pénalité que celle dans [26]. Mais, pour la globalisation de leur méthode, ils utilisent une technique de régions de confiance.

Dans la section 2.2, des résultats préliminaires associés à la méthode primale-duale sont décrits. Le principe général de l'algorithme est expliqué dans la partie 2.3. Dans le paragraphe 2.4, l'algorithme local et le théorème de convergence associé sont présentés. Dans la section 2.5, la globalisation de la méthode, le théorème de convergence globale et le résultat de l'analyse asymptotique sont explicités.

2.2 Résultats préliminaires

Dans cette partie, plusieurs résultats caractérisant la méthode primale-duale sont présentés.

Lemme 2.2.1 *Sous les hypothèses 1.2.1-1.2.4, la jacobienne de F par rapport à w en $(w^*, 0)$ définie par*

$$F'_w(w^*, 0) := \begin{pmatrix} \nabla^2_{xx}\mathcal{L}(w^*) & A(x^*)^\top \\ A(x^*) & 0 \end{pmatrix} \tag{2.5}$$

est inversible.

Preuve. Soient $u \in \mathbb{R}^n$ et $v \in \mathbb{R}^m$. Nous avons

$$\begin{aligned} F'_w(w^*, 0)\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} &\Leftrightarrow \begin{pmatrix} \nabla^2_{xx}\mathcal{L}(w^*) & A(x^*)^\top \\ A(x^*) & 0 \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \\ &\Leftrightarrow \begin{cases} \nabla^2_{xx}\mathcal{L}(w^*)u + A(x^*)^\top v = 0 \quad (\triangle) \\ A(x^*)u = 0 \end{cases} \\ &\Rightarrow \begin{cases} u^\top \nabla^2_{xx}\mathcal{L}(w^*)u + (A(x^*)u)^\top v = 0 \\ A(x^*)u = 0 \end{cases} \\ &\Rightarrow \begin{cases} u^\top \nabla^2_{xx}\mathcal{L}(w^*)u = 0 \\ A(x^*)u = 0. \end{cases} \end{aligned}$$

D'après l'hypothèse 1.2.4, le dernier système implique que $u = 0$. Donc d'après l'égalité $(\triangle)$, nous avons $A(x^*)^\top v = 0$. Cette égalité et l'hypothèse 1.2.3 impliquent que $v = 0$. Donc, la matrice $F'_w(w^*, 0)$ est inversible.

□

Lemme 2.2.2 *Il existe $\tilde{r} > 0, \tilde{\mu} > 0$ et $\eta > 0$ tel que pour tout $w \in B(w^*, \tilde{r})$ et $\mu \in]-\tilde{\mu}, \tilde{\mu}[$, la jacobienne $F'_w(w, \mu)$ définie en (2.4) est inversible et vérifie*

$$\|F'_w(w, \mu)^{-1}\| \leq \eta. \tag{2.6}$$

Preuve. D'après le lemme 2.2.1, la matrice $F'_w(w^*, 0)$ est inversible. D'après le Lemme 2.3.2. (Perturbation Lemma) de [54], la matrice $F'_w(w, \mu)$ est inversible dans un voisinage de w et de μ et il existe $\eta > 0$ tel que l'inégalité (2.6) soit vérifiée.

□

Lemme 2.2.3 *La jacobienne F'_w est localement lipschitzienne en $(w^*, 0)$,i.e. il existe $\hat{r} > 0, \hat{\mu} > 0$ et $\vartheta > 0$ tel que pour tout $((w, \mu), (w', \mu')) \in B(w^*, \hat{r})\times]-\hat{\mu}, \hat{\mu}[$,*

$$\|F'_w(w, \mu) - F'_w(w', \mu')\| \leq \vartheta(\|w - w'\| + |\mu - \mu'|).$$

Preuve. D'après l'hypothèse 1.2.2, chaque composante de la matrice F'_w donnée par la formule (2.4) est localement lipschitzienne en $(w^*, 0)$. Comme il y a un nombre fini de composantes, la matrice F'_w est localement lipschitzienne en $(w^*, 0)$.

□

Lemme 2.2.4 *Il existe $\bar{r} > 0, \bar{\mu} > 0$ et une fonction continument différentiable $w(\cdot) :]-\bar{\mu}, \bar{\mu}[\rightarrow \mathbb{R}^{n+m}$ telle que pour tout $(w, \mu) \in B(w^*, \bar{r})\times]-\bar{\mu}, \bar{\mu}[$*

$$F(w, \mu) = 0 \quad \textit{si et seulement si} \quad w = w(\mu).$$

Pour $\mu \in]-\bar{\mu}, \bar{\mu}[$, nous avons

$$w(\mu) = w^* + w'(0)\mu + o(\mu), \tag{2.7}$$

où

$$w'(0) = F'_w(w^*, 0)^{-1} \begin{pmatrix} 0 \\ \lambda^* \end{pmatrix}. \tag{2.8}$$

Il existe une constante $\iota > 0$ telle que pour tout $\mu, \mu' \in]-\bar{\mu}, \bar{\mu}[$

$$\|w(\mu) - w(\mu')\| \leq \iota|\mu - \mu'|. \tag{2.9}$$

Preuve. D'après le lemme 2.2.1, le théorème des fonctions implicites peut être appliqué à F. Il existe alors $\bar{r} > 0, \bar{\mu} > 0$ et une fonction $w(\cdot) :]-\bar{\mu}, \bar{\mu}[\rightarrow \mathbb{R}^{n+m}$ continument différentiable telle que pour tout $\mu \in]-\bar{\mu}, \bar{\mu}[$, $F(w, \mu) = 0$ si et seulement si $w = w(\mu)$.
En calculant un développement au premier ordre de w, nous avons pour $\mu \in]-\bar{\mu}, \bar{\mu}[$

$$w(\mu) = w(0) + w'(0)\mu + o(\mu).$$

Or, $w^* = w(0)$ donc nous obtenons l'égalité (2.7).
Pour tout $\mu \in]-\bar{\mu}, \bar{\mu}[$, $F(w(\mu), \mu) = 0$. En dérivant $F(w(\cdot), \cdot)$ par rapport à μ, nous avons

$$F'_w(w(\mu), \mu)w'(\mu) + F'_\mu(w(\mu), \mu) = 0.$$

En particulier, pour $\mu = 0$ nous obtenons

$$F'_w(w^*, 0)w'(0) + F'_\mu(w^*, 0) = 0.$$

D'après le lemme 2.2.1, nous avons

$$\begin{aligned} w'(0) &= -F'_w(w^*, 0)^{-1}F'_\mu(w^*, 0) \\ &= -F'_w(w^*, 0)^{-1} \begin{pmatrix} 0 \\ -\lambda^* \end{pmatrix}, \end{aligned}$$

d'où l'expression (2.8).

Comme w est continument différentiable, d'après l'inégalité des accroissements finis, nous avons

$$\|w(\mu) - w(\mu')\| \leq \sup_{\tau \in]-\bar{\mu},\bar{\mu}[} \|w'(\tau)\| \ \ |\mu - \mu'|.$$

On en déduit (2.9) en posant $\iota := \sup_{\tau \in]-\bar{\mu},\bar{\mu}[} \|w'(\tau)\|$.

□

L'application $\mu \mapsto w(\mu)$ est appelée trajectoire ou chemin. Notons que $w'(0) := (x'(0), \lambda'(0))$ est solution du système

$$\left\{ \begin{array}{l} \nabla^2_{xx}\mathcal{L}(w^*)x'(0) + A(x^*)^\top \lambda'(0) = 0 \\ A(x^*)x'(0) = \lambda^*. \end{array} \right. \tag{2.10}$$

Dans le cas particulier où $\lambda^* = 0$, la trajectoire est réduite à un point. En effet, nous avons

$$F(w^*, \mu) = 0 \iff \left(\begin{array}{c} \nabla f(x^*) \\ c(x^*) \end{array} \right) = 0 \iff w(\mu) = w^* = (x^*, 0) \ \ \forall\, \mu.$$

Lorsque $\lambda^* \neq 0$, nous pouvons interpréter $w'(0)$ comme la solution primale-duale du problème

$$\begin{array}{ll} \text{minimiser}_{\xi \in \mathbb{R}^n} & \dfrac{1}{2}\, \xi^\top \nabla^2_{xx}\mathcal{L}(w^*)\xi \\ \text{sous contrainte} & A(x^*)\xi = \lambda^*. \end{array}$$

Notons que dans ce cas $x'(0) \notin \ker(A(x^*))$ et donc dans l'espace primal la trajectoire n'est pas tangente à la variété des contraintes.

Voici un exemple de trajectoire.

Exemple 2.2.5 Soit le problème

$$\begin{array}{ll} \text{minimiser}_{x \in \mathbb{R}^2} & e^\top x \\ \text{sous contrainte} & \|x\|^2 - 2 = 0, \end{array}$$

où $e := (1, 1)^\top$. Le système primal-dual associé au problème est

$$F(w, \mu) = \left(\begin{array}{c} e + 2x\lambda \\ \|x\|^2 - 2 - \mu\lambda \end{array} \right) = 0.$$

La figure 2.1 représente dans l'espace primal les courbes de niveaux de $e^\top x = k$ pour $k \in \mathbb{R}$ et de $\|x\|^2 - 2 = 0$. Les points de tangence entre les courbes de niveaux de la fonction objectif et de la contrainte définissent le chemin $\mu \mapsto x(\mu)$. La trajectoire est ici une droite.

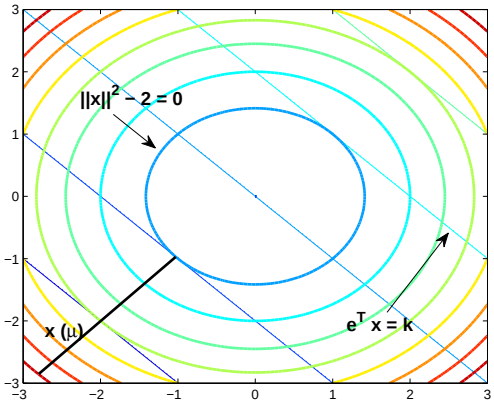


FIG. 2.1 – Exemple de trajectoire $\mu \mapsto x(\mu)$.

2.3 Principe général de l'algorithme

Pour résoudre le système (2.3), nous appliquons la méthode de Newton en (w, μ). Celle-ci correspond à la linéarisation de l'équation (2.3) par rapport aux variables w et μ. En notant l'itéré courant (w, μ), l'itéré suivant est défini par

$$(w^+, \mu^+) := (w, \mu) + (d^w, d^\mu), \tag{2.11}$$

où (d^w, d^μ) est défini par le système

$$F'(w,\mu) \begin{pmatrix} d^w \\ d^\mu \end{pmatrix} + F(w,\mu) = 0.$$

Or,

$$\begin{aligned} F'(w,\mu) \begin{pmatrix} d^w \\ d^\mu \end{pmatrix} + F(w,\mu) &= F'_w(w,\mu)d^w + F'_\mu(w,\mu)d^\mu + F(w,\mu) \\ &= F'_w(w,\mu)d^w + \begin{pmatrix} 0 \\ -\lambda \end{pmatrix} d^\mu + \begin{pmatrix} \nabla f(x) + A(x)^\top \lambda \\ c(x) - \mu\lambda \end{pmatrix} \\ &= F'_w(w,\mu)d^w + \begin{pmatrix} \nabla f(x) + A(x)^\top \lambda \\ c(x) - \lambda(\mu + d^\mu) \end{pmatrix} \\ &= F'_w(w,\mu)d^w + F(w,\mu^+). \end{aligned}$$

Ainsi, l'itération est définie par le choix d'une valeur μ^+ et d'un nouvel itéré w^+ défini par $w^+ := w + d^w$ où d^w est solution du système

$$F'_w(w,\mu)d^w + F(w,\mu^+) = 0. \tag{2.12}$$

Donc lorsque $F'_w(w,\mu)$ est inversible, l'itéré de Newton w^+ vérifie

$$w^+ = w - F'_w(w,\mu)^{-1}F(w,\mu^+). \tag{2.13}$$

Avant de présenter l'algorithme associé à la méthode et les résultats de convergence, un lemme est énonçé. Il donne des propriétés sur l'itéré w^+. Choisissons des constantes $\tilde{\mu}, \hat{\mu}, \bar{\mu}, \tilde{r}, \hat{r}, \bar{r}, \eta, \vartheta$ et ι telles que les lemmes 2.2.2, 2.2.3 et 2.2.4 soient satisfaits. Définissons les deux constantes positives

$$r^* := \min\left\{\tilde{r}, \hat{r}, \bar{r}, \frac{1}{4\eta\vartheta}\right\} \text{ et } \mu^* := \min\left\{\tilde{\mu}, \hat{\mu}, \bar{\mu}, \frac{r^*}{1+\iota}(1+\sqrt{1+2\eta\vartheta r^*})^{-1}\right\}. \quad (2.14)$$

Lemme 2.3.1 *Pour tout $w \in B(w^*, r^*)$ et $\mu^+ \in [0, \mu^*[$, l'itéré w^+ défini par (2.13) vérifie*

$$\|w^+ - w(\mu^+)\| \leq \frac{\eta\vartheta}{2}\left(\|w - w(\mu^+)\|^2 + \|w - w(\mu^+)\||\mu - \mu^+|\right), \quad (2.15)$$

et

$$\|w^+ - w^*\| \leq \frac{1}{2}\|w - w^*\| + \frac{\mu^+}{\mu^*}\frac{r^*}{2}. \quad (2.16)$$

En particulier, $w^+ \in B(w^*, r^*)$.

Preuve. Soient $w \in B(w^*, r^*)$ et $\mu^+ \in [0, \mu^*[$. D'après le lemme 2.2.4, nous avons $F(w(\mu^+), \mu^+) = 0$. En notant pour $t \in [0, 1]$

$$\phi(t) := F((1-t)w(\mu^+) + tw, \mu^+),$$

et en utilisant $\phi(1) - \phi(0) = \displaystyle\int_0^1 \phi'(t)\ \mathrm{d}t$, nous avons

$$\begin{aligned}
&w^+ - w(\mu^+)\\
&= w - w(\mu^+) - F'_w(w,\mu)^{-1}F(w,\mu^+)\\
&= -F'_w(w,\mu)^{-1}[F(w,\mu^+) - F(w(\mu^+),\mu^+) - F'_w(w,\mu)(w - w(\mu^+))]\\
&= -F'_w(w,\mu)^{-1}[\phi(1) - \phi(0) - F'_w(w,\mu)(w - w(\mu^+))]\\
&= -F'_w(w,\mu)^{-1}\int_0^1 (F'_w((1-t)w(\mu^+) + tw, \mu^+) - F'_w(w,\mu))(w - w(\mu^+))\ \mathrm{d}t.
\end{aligned}$$

En prenant la norme des deux côtés de l'égalité précédente et d'après les lemmes 2.2.2 et 2.2.3, nous obtenons l'inégalité (2.15).
En utilisant l'inégalité (2.15) et $(a+b)^2 \leq 2(a^2+b^2)$, nous avons

$$\begin{aligned}
\|w^+ - w^*\| &\leq \|w^+ - w(\mu^+)\| + \|w(\mu^+) - w^*\|\\
&\leq \frac{\eta\vartheta}{2}\|w - w(\mu^+)\|^2 + \frac{\eta\vartheta}{2}\|w - w(\mu^+)\||\mu - \mu^+| + \iota\mu^+\\
&\leq \eta\vartheta\|w - w^*\|^2 + \eta\vartheta\|w^* - w(\mu^+)\|^2 + \frac{\eta\vartheta}{2}\|w - w^*\||\mu - \mu^+|\\
&\quad + \frac{\eta\vartheta}{2}\|w^* - w(\mu^+)\||\mu - \mu^+| + \iota\mu^+\\
&\leq \eta\vartheta(r^* + \mu^*)\|w - w^*\| + \iota\mu^+\big(\eta\vartheta\mu^*(\iota+1) + 1\big).
\end{aligned}$$

L'inégalité (2.16) est obtenue par les choix des valeurs r^* et μ^*.

□

Dans la section qui suit, nous présentons l'algorithme local associé à la méthode primale-duale utilisé pour résoudre le problème non linéaire (1.1) ainsi que le théorème de convergence locale.

2.4 Algorithme local

2.4.1 Description de l'algorithme

Soient $w_0 := (x_0, \lambda_0) \in \mathbb{R}^{n+m}$ un point de départ et $\mu_0 > 0$ une valeur initiale du paramètre de pénalisation.

Algorithme D : algorithme primal-dual local

Répéter pour $k = 0, 1, 2, \ldots$ jusqu'à ce qu'un test de convergence soit satisfait

Choisir μ_{k+1} tel que $\{\mu_k\} \to 0$.

Calculer la direction d_k^w en résolvant le système linéaire

$$F'_w(w_k, \mu_k) d_k^w + F(w_k, \mu_{k+1}) = 0.$$

Faire $w_{k+1} = w_k + d_k^w$

Fin

Dans l'algorithme D, le paramètre de pénalité μ_k vérifie l'hypothèse suivante.

Hypothèse 2.4.1 *La suite de nombres positifs $\{\mu_k\}$ converge vers zéro telle que pour tout entier k*

$$\beta \mu_k^{1+\kappa} \leq \mu_{k+1} \leq \gamma \mu_k, \tag{2.17}$$

pour des constantes $\beta > 0, \gamma > 0$ et $0 < \kappa < 1$.

L'inégalité de gauche de (2.17) signifie que le taux de convergence de la suite des itérés $\{\mu_k\}$ est au plus superlinéaire d'ordre $1 + \kappa$. En particulier, le taux de convergence de $\{\mu_k\}$ n'est pas quadratique et nous avons

$$\mu_k^2 = o(\mu_{k+1}). \tag{2.18}$$

2.4.2 Théorème de convergence locale

À présent, nous effectuons l'analyse de convergence locale associée à l'algorithme D. La technique de preuve du théorème qui suit utilise la même technique que la preuve du Théorème 4.2 dans [6]. Ce résultat montre que si l'itéré initial w_0 est suffisamment proche de la solution w^*, alors la suite $\{w_k\}$ générée par l'algorithme converge vers w^* et les itérés vont se rapprocher de la trajectoire $w(\cdot)$ de manière tangentielle.

Théorème 2.4.2 *Soient r^* et μ^* les constantes définies par (2.14). Soient $w_0 \in B(w^*, r^*)$ et $\mu_0 \in]0, \mu^*[$. La suite des itérés $\{w_k\}$ générée par l'algorithme D converge vers w^* et pour tout entier k*

$$w_k = w(\mu_k) + o(\mu_k). \tag{2.19}$$

Avant de faire la preuve du théorème 2.4.2, nous énonçons un lemme qui donne une borne de la distance qui sépare l'itéré de Newton à la trajectoire. D'après l'hypothèse 2.4.1 et en notant $\gamma' := \max\{1, \gamma - 1\}$, nous avons pour tout entier k

$$|\mu_{k+1} - \mu_k| \leq \gamma' \mu_k. \tag{2.20}$$

Lemme 2.4.3 *En notant* $\zeta := \frac{3}{2}\eta\vartheta \max\left\{1, \frac{\gamma'}{2}, (\iota\gamma')^2 + \frac{\iota\gamma'^2}{2}\right\}$, *nous avons pour tout entier* k

$$||w_{k+1} - w(\mu_{k+1})|| \leq \zeta(||w_k - w(\mu_k)||^2 + \mu_k^2). \tag{2.21}$$

Preuve. D'après les inégalités (2.15), (2.9), (2.20) et $(a+b)^2 \leq 2(a^2+b^2)$, nous avons

$$\begin{aligned}
\|w_{k+1} - w(\mu_{k+1})\| &\leq \frac{\eta\vartheta}{2}\|w_k - w(\mu_{k+1})\|^2 + \frac{\eta\vartheta}{2}\|w_k - w(\mu_{k+1})\||\mu_k - \mu_{k+1}| \\
&\leq \frac{\eta\vartheta}{2}\|w_k - w(\mu_k) + w(\mu_k) - w(\mu_{k+1})\|^2 \\
&+ \frac{\eta\vartheta}{2}\|w_k - w(\mu_k) + w(\mu_k) - w(\mu_{k+1})\||\mu_k - \mu_{k+1}| \\
&\leq \frac{\eta\vartheta}{2}(\|w_k - w(\mu_k)\| + \iota|\mu_k - \mu_{k+1}|)^2 \\
&+ \frac{\eta\vartheta}{2}\|w_k - w(\mu_k)\||\mu_k - \mu_{k+1}| + \frac{\eta\vartheta\iota}{2}|\mu_k - \mu_{k+1}|^2 \\
&\leq \eta\vartheta\|w_k - w(\mu_k)\|^2 + \frac{\eta\vartheta}{2}\gamma'\|w_k - w(\mu_k)\|\mu_k \\
&+ \mu_k^2(\eta\vartheta(\iota\gamma')^2 + \frac{\eta\vartheta\iota}{2}\gamma'^2) \\
&\leq \frac{2\zeta}{3}(\|w_k - w(\mu_k)\|^2 + \|w_k - w(\mu_k)\|\mu_k + \mu_k^2) \\
&\leq \zeta(\|w_k - w(\mu_k)\|^2 + \mu_k^2).
\end{aligned}$$

□

À présent, donnons la preuve du théorème 2.4.2.
Preuve. D'après l'inégalité (2.16) du lemme 2.3.1, nous avons

$$||w_{k+1} - w^*|| \leq \frac{1}{2}||w_k - w^*|| + \frac{r^*}{2\mu^*}\mu_{k+1}.$$

Cette inégalité signifie que l'itération est bien définie. De plus, lorsque $k \to +\infty$, $\{w_k\}$ converge vers w^*.

À présent, montrons l'égalité (2.19). Soit ζ la constante définie dans le lemme 2.4.3 et posons $s_k := \zeta||w_k - w(\mu_k)||$ et $\varepsilon_k := \zeta\mu_k$. D'après l'inégalité (2.21) du lemme 2.4.3, nous avons

$$s_{k+1} \leq s_k^2 + \varepsilon_k^2. \tag{2.22}$$

L'hypothèse 2.4.1 implique qu'il existe $\xi > 0$, $\varrho \in]0,1[$ et $\bar{k}$ tels que

$$\xi\varrho^{(1+\kappa)^k} \leq \mu_k \quad \text{pour} \quad k \geq \bar{k}. \tag{2.23}$$

D'après la définition de ε_k, nous avons $\varepsilon_k^2 = o(\varepsilon_{k+1})$. Il existe alors $\bar{k} \geq 0$ tel que

$$\varepsilon_k^2 \leq \frac{1}{2}\varepsilon_{k+1} \quad \text{pour} \quad k \geq \bar{k}.$$

Montrons qu'il existe $k_0 \geq \bar{k}$ tel que $s_{k_0} \leq \varepsilon_{k_0}$. Pour ceci, effectuons un raisonnement par l'absurde en supposant que $s_k > \varepsilon_k$ pour tout $k \geq \bar{k}$. L'inégalité (2.22) implique que $s_{k+1} \leq 2s_k^2$ ce qui implique que la suite $\{s_k\}$ converge quadratiquement vers zéro et donc que $s_k = O(\alpha^{2^k})$ pour un certain $\alpha > 0$. Mais, comme nous avons supposé que $s_k > \varepsilon_k$, nous avons aussi que $\varepsilon_k = O(\alpha^{2^k})$. Ce résultat est une contradiction avec l'inégalité (2.23). Donc, il existe $k_0 \geq \bar{k}$ tel que $s_{k_0} \leq \varepsilon_{k_0}$.

Montrons par récurrence que $s_k \leq \varepsilon_k$ pour $k \geq k_0$. La propriété est vraie pour $k = k_0$. Supposons que la propriété est vraie pour un entier $k \geq k_0$. D'après l'inégalité (2.22) et l'hypothèse de récurrence $s_k \leq \varepsilon_k$, nous avons

$$s_{k+1} \leq 2\varepsilon_k^2 \leq \varepsilon_{k+1}.$$

Donc, la propriété est aussi vraie au rang $k+1$. D'après l'inégalité (2.22), nous pouvons en déduire que pour tout $k \geq k_0$, $s_{k+1} \leq 2\varepsilon_k^2 = o(\varepsilon_{k+1})$ ce qui implique l'égalité (2.19).

□

2.4.3 Exemple illustratif

L'égalité (2.19) signifie que les itérés w_k se rapprochent de la trajectoire $w(\cdot)$ de manière tangentielle. De plus, d'après le lemme 2.2.4, nous avons

$$\lim_{k \to +\infty} \frac{w_k - w^*}{\mu_k} = \lim_{k \to +\infty} \frac{w(\mu_k) - w^*}{\mu_k} = w'(0).$$

Nous pouvons en déduire que

$$||w_k - w^*|| \sim ||w'(0)||\mu_k,$$

ce qui implique que la suite $\{w_k\}$ converge vers w^* avec le même taux de convergence que $\{\mu_k\}$ vers zéro. Nous montrons avec un exemple que si la suite du paramètre de pénalité $\{\mu_k\}$ converge quadratiquement vers zéro, alors les itérés w_k générés par l'algorithme D ne vont pas se rapprocher du chemin $w(\cdot)$ de manière tangentielle et donc le théorème 2.4.2 de convergence locale n'est plus vérifié.

Exemple 2.4.4 Considérons le problème

$$\begin{array}{ll} \text{minimiser}_{x \in \mathbb{R}^2} & e^\top x \\ \text{sous contrainte} & ||x||^2 - 1 = 0, \end{array}$$

où $e := (1,1)^\top$. Le système primal-dual associé au problème est

$$F(w,\mu) = \begin{pmatrix} e + 2x\lambda \\ ||x||^2 - 1 - \mu\lambda \end{pmatrix} = 0.$$

La solution de ce système est $w^* = \left(\dfrac{-\sqrt{2}}{2} e, \dfrac{1}{\sqrt{2}} \right)$.

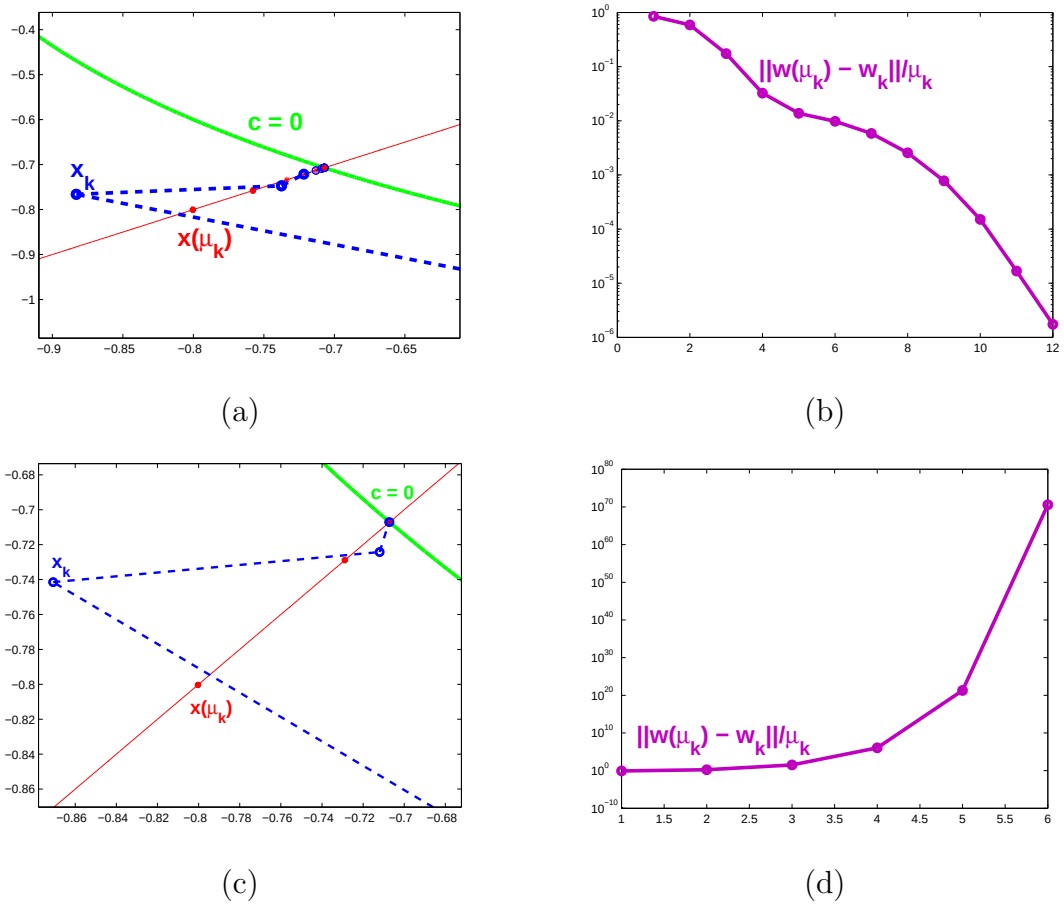


FIG. 2.2 – (a) et (b) : la suite du paramètre de pénalité $\{\mu_k\}$ converge superlinéairement vers zéro. La suite des itérés dans l'espace primal, est asymptotiquement tangente à la trajectoire. Au cours des itérations, $||w(\mu_k) - w_k||/\mu_k$ tend vers zéro. (c) et (d) : la suite du paramètre de pénalité $\{\mu_k\}$ converge quadratiquement vers zéro. La suite des itérés dans l'espace primal, n'est pas asymptotiquement tangente à la trajectoire. Au cours des itérations, $||w(\mu_k) - w_k||/\mu_k$ tend vers l'infini.

La figure 2.2 (a) montre dans l'espace primal comment l'itéré x_k converge au cours de l'algorithme vers la solution primale x^* lorsque la suite du paramètre de pénalité $\{\mu_k\}$ converge vers zéro de manière superlinéaire, i.e. $\mu_0 = 0$ et $\mu_{k+1} = \mu_k^{1.5}$. Il est clair que l'itéré x_k arrive de manière tangentielle sur la trajectoire $x(\mu_k)$.
La figure 2.2 (b) montre que pour ce taux de convergence, nous avons

$$\lim_{k \to +\infty} \frac{||w(\mu_k) - w_k||}{\mu_k} = 0.$$

Les figures 2.2 (c) et (d) représentent les mêmes courbes mais lorsque la suite du paramètre de pénalité $\{\mu_k\}$ converge quadratiquement vers zéro, i.e. $\mu_0 = 0$ et $\mu_{k+1} = \mu_k^2$. Dans ce cas, l'itéré x_k ne se rapproche pas de la trajectoire $x(\mu_k)$ de façon tangentielle et nous avons

$$\lim_{k \to +\infty} \frac{||w(\mu_k) - w_k||}{\mu_k} = +\infty.$$

Le théorème 2.4.2 nous permet d'envisager une globalisation de la méthode puisque les itérés vont se rapprocher naturellement de la trajectoire. La globalisation de l'algorithme D est présentée dans la partie suivante ainsi que le théorème de convergence globale et de l'analyse asymptotique.

2.5 Algorithme global

2.5.1 Description de l'algorithme

L'algorithme D décrit précédemment est globalisé avec une technique de recherche linéaire. La technique de globalisation que nous proposons est analogue à celle proposée par P. Armand, J. Benoist et D. Orban dans [7] pour les problèmes d'optimisation non linéaires avec contraintes d'inégalité. L'analyse de convergence de l'algorithme que nous proposons est plus simple que celle proposée dans [7] puisqu'il n'est plus nécessaire de montrer que le pas unité est accepté lorsqu'on applique la règle du déplacement maximal à la frontière (inégalité (1.11) dans [7]). Il suffit de démontrer que l'inégalité correspondant au test d'arrêt dans la globalisation est vérifiée. L'algorithme global associé à la méthode primale-duale utilisé pour résoudre le problème non linéaire (1.1) a été programmé sous MATLAB. Des tests numériques ont été effectués sur des problèmes provenant des librairies COPS 3.0 [21] et CUTEr [33]. Une étude détaillée de ces tests est présentée dans le chapitre 3.
Soient $w_0 := (x_0, \lambda_0) \in \mathbb{R}^{n+m}$ un point de départ, $\mu_0 > 0$ une valeur initiale du paramètre de pénalisation et $\rho \in]0, 1[$.

Algorithme E : algorithme primal-dual global

Répéter pour $k = 0, 1, 2, \ldots$ jusqu'à ce qu'un test de convergence soit satisfait

Choisir μ_{k+1} et $\varepsilon_k > 0$ avec $\{\mu_k\} \to 0$ et $\{\varepsilon_k\} \to 0$.

Calculer la direction d_k^w en résolvant le système linéaire

$$F'_w(w_k, \mu_k) d_k^w + F(w_k, \mu_{k+1}) = 0.$$

Si

$$\|F(w_k + d_k^w, \mu_{k+1})\| \leq \rho \|F(w_k, \mu_k)\| + \varepsilon_k,$$

alors faire $w_{k+1} = w_k + d_k^w$,

sinon appliquer une suite d'itérations internes avec un paramètre de pénalité μ_{k+1} fixé pour déterminer w_{k+1} tel que

$$\|F(w_{k+1}, \mu_{k+1})\| \leq \rho \|F(w_k, \mu_k)\| + \varepsilon_k.$$

Fin

Fin

Description du test d'arrêt de la globalisation

Le test d'arrêt de la globalisation est une condition de décroissance suffisante de la norme des conditions d'optimalité perturbées auquel nous ajoutons un terme de

relaxation ε_k. Ce dernier est important car il permet de relacher la condition si l'itéré courant est sur le chemin $w(\cdot)$. En effet, si la condition de décroissance classique

$$\|F(w_k + d_k^w, \mu_{k+1})\| \leq \rho \|F(w_k, \mu_k)\|, \tag{2.24}$$

est utilisée et si l'itéré w_k est sur le chemin, i.e. $w_k = w(\mu_k)$, alors le terme à droite de l'inégalité (2.24) est égal à zéro ce qui impose que le prochain itéré $w_k + d_k^w$ soit de nouveau sur le chemin $w(\cdot)$. Or, cette condition est trop exigeante et n'est pas nécessaire pour obtenir la convergence.

Description des itérations internes

Les itérations internes sont utilisées pour contrôler la décroissance de la norme de F au cours des itérations. Elles assurent la convergence globale de la méthode primale-duale quand le paramètre de pénalité μ est fixé.

Elles correspondent à la résolution du système $F(w, \mu) = 0$ avec la méthode de Newton lorsque le paramètre de pénalité μ est fixé. Pour globaliser la méthode, une technique de recherche linéaire est appliquée en utilisant la fonction de mérite primale-duale

$$\Theta_{\mu,\sigma}(w) := f(x) + \frac{1}{2\mu}\|c(x)\|^2 + \sigma\|c(x) - \mu\lambda\|^2, \tag{2.25}$$

où $\sigma > 0$. Les deux premiers termes de la fonction de mérite (2.25) correspondent à la fonction de pénalisation quadratique (1.6) et le dernier terme contrôle la réalisation de la contrainte perturbée.

Décrivons de manière plus précise les itérations internes. Soit $\eta \in]0, 1[$.

Itérations internes de l'algorithme E

Choisir $w_k^0 \in \mathbb{R}^{n+m}$ et $\sigma_{k+1} > 0$.

Répéter pour $i = 0, 1, 2, \ldots$

Calculer la direction $d_{k,i}^w$ en résolvant le système linéaire

$$F'_w(w_k^i, \mu_{k+1}) d_{k,i}^w + F(w_k^i, \mu_{k+1}) = 0.$$

Choisir le plus grand α dans $\left\{1, \frac{1}{2}, \frac{1}{4}, \ldots\right\}$ tel que

$$\Theta_{\mu_{k+1},\sigma_{k+1}}(w_k^i + \alpha d_{k,i}^w) \leq \Theta_{\mu_{k+1},\sigma_{k+1}}(w_k^i) + \alpha\eta\Theta'_{\mu_{k+1},\sigma_{k+1}}(w_k^i; d_{k,i}^w).$$

Faire $w_k^{i+1} = w_k^i + \alpha d_{k,i}^w$.

Si

$$\|F(w_k^{i+1}, \mu_{k+1})\| \leq \rho\|F(w_k, \mu_k)\| + \varepsilon_k,$$

alors faire $w_{k+1} = w_k^{i+1}$ et stop.

Fin

Fin

Directions de descente

Au cours des itérations internes, il faut vérifier que la direction $d^w_{k,i} := (d^x_{k,i}, d^\lambda_{k,i})$ est une direction de descente de la fonction de mérite $\Theta_{\mu,\sigma}$, i.e. $\Theta'_{\mu,\sigma}(w_k; d^w_k) < 0$. La proposition suivante donne l'expression de la dérivée directionnelle de $\Theta_{\mu,\sigma}$.

Proposition 2.5.1 *La dérivée directionnelle de $\Theta_{\mu,\sigma}$ en $w_k := (x_k, \lambda_k)$ dans la direction d^w_k satisfait*

$$\Theta'_{\mu,\sigma}(w_k; d^w_k) = -(d^x_k)^\top(\nabla^2_{xx}\mathcal{L}(w_k) + \frac{1}{\mu}A(x_k)^\top A(x_k))d^x_k - 2\sigma||c(x_k) - \mu\lambda_k||^2. \quad (2.26)$$

Preuve. D'après la définition (2.25) de $\Theta_{\mu,\sigma}$, sa dérivée directionnelle en w_k dans la direction d^w_k est

$$\begin{aligned}\Theta'_{\mu,\sigma}(w_k; d^w_k) &= \nabla f(x_k)^\top d^x_k + \frac{1}{\mu}c(x_k)^\top A(x_k)d^x_k + 2\sigma(c(x_k) - \mu\lambda_k)^\top A(x_k)d^x_k \\ &\quad - 2\mu\sigma(c(x_k) - \mu\lambda_k)^\top d^\lambda_k.\end{aligned}$$

À l'itération k, le système $F'_w(w_k, \mu)d^w_k + F(w_k, \mu) = 0$ s'écrit

$$\left\{\begin{array}{l} \nabla^2_{xx}\mathcal{L}(w_k)d^x_k + \nabla f(x_k) + A(x_k)^\top(\lambda_k + d^\lambda_k) = 0 \\ A(x_k)d^x_k + c(x_k) - \mu(\lambda_k + d^\lambda_k) = 0. \end{array}\right. \quad (2.27)$$

En utilisant le système (2.27), nous avons

$$\begin{aligned}&\nabla f(x_k)^\top d^x_k + \frac{1}{\mu}c(x_k)^\top A(x_k)d^x_k \\ &= -(d^x_k)^\top\nabla^2_{xx}\mathcal{L}(w_k)d^x_k - (\lambda_k + d^\lambda_k)^\top A(x_k)d^x_k + \frac{1}{\mu}c(x_k)^\top A(x_k)d^x_k \\ &= -(d^x_k)^\top\nabla^2_{xx}\mathcal{L}(w_k)d^x_k - \frac{1}{\mu}(A(x_k)d^x_k)^\top(A(x_k)d^x_k) \\ &= -(d^x_k)^\top(\nabla^2_{xx}\mathcal{L}(w_k) + \frac{1}{\mu}A(x_k)^\top A(x_k))d^x_k\end{aligned}$$

et

$$\begin{aligned}&2\sigma(c(x_k) - \mu\lambda_k)^\top A(x_k)d^x_k - 2\mu\sigma(c(x_k) - \mu\lambda_k)^\top d^\lambda_k \\ &= 2\sigma(c(x_k) - \mu\lambda_k)^\top(A(x_k)d^x_k - \mu d^\lambda_k) \\ &= -2\sigma(c(x_k) - \mu\lambda_k)^\top(c(x_k) - \mu\lambda_k) \\ &= -2\sigma||c(x_k) - \mu\lambda_k||^2.\end{aligned}$$

En utilisant ces égalités, l'expression (2.26) de la dérivée directionnelle de $\Theta_{\mu,\sigma}$ est obtenue.

□

Proposition 2.5.2 *La direction $d^w_k \neq 0$ est une direction de descente de la fonction de mérite $\Theta_{\mu,\sigma}$ si l'inertie de la matrice $F'_w(w_k, \mu)$ est égale à $(n, m, 0)$.*

Preuve. La matrice $F'_w(w_k, \mu)$ peut s'écrire sous la forme :

$$\begin{aligned} F'_w(w_k, \mu) &:= \begin{pmatrix} \nabla^2_{xx}\mathcal{L}(w_k) & A(x_k)^\top \\ A(x_k) & -\mu Id \end{pmatrix} \\ &= P_k B_k P_k^\top, \end{aligned}$$

où les matrices P_k et B_k sont définies par

$$P_k := \begin{pmatrix} I & -\frac{1}{\mu}A(x_k)^\top \\ 0 & Id \end{pmatrix} \text{ et } B_k := \begin{pmatrix} \nabla^2_{xx}\mathcal{L}(w_k) + \frac{1}{\mu}A(x_k)^\top A(x_k) & 0 \\ 0 & -\mu Id \end{pmatrix}.$$

Les matrices $F'_w(w_k, \mu)$ et B_k sont congruentes et d'après la loi d'inertie de Sylvester (Théorème 4.5.8 dans [37]), elles ont même inertie. Nous avons ainsi

$$\begin{aligned} i(F'_w(w_k, \mu)) &= i(B_k) \\ &= i(\nabla^2_{xx}\mathcal{L}(w_k) + \frac{1}{\mu}A(x_k)^\top A(x_k)) + (0, m, 0). \end{aligned}$$

On a donc $i(F'_w(w_k, \mu)) = (n, m, 0)$ si et seulement si la matrice

$$\nabla^2_{xx}\mathcal{L}(w_k) + \frac{1}{\mu}A(x_k)^\top A(x_k) \tag{2.28}$$

est définie positive.
Soit $d^w_k := (d^x_k, d^\lambda_k) \neq 0$. Si $d^x_k = 0$, alors $d^\lambda_k \neq 0$ et d'après (2.27), nous avons $c(x_k) - \mu\lambda_k = -\mu d^\lambda_k \neq 0$. Donc, d'après (2.26), $\Theta'_{\mu,\sigma}(w_k; d^w_k) < 0$. Si $d^x_k \neq 0$ et si $i(F'_w(w_k, \mu)) = (n, m, 0)$, alors d'après ce qui précède la matrice (2.28) est définie positive et nous avons $\Theta'_{\mu,\sigma}(w_k; d^w_k) < 0$.

□

Remarque 2.5.3 Numériquement, pour factoriser et calculer l'inertie de la matrice $F'_w(w_k, \mu)$, l'interface MA57 (Harwell Subroutine Library) [4, 22] a été utilisée dans notre implémentation. Si l'inertie n'est pas correcte, nous considérons $F'_w(w_k, \mu)$ sous la forme

$$\begin{pmatrix} \nabla^2_{xx}\mathcal{L}(w_k) + \delta Id & A(x_k)^\top \\ A(x_k) & -\mu Id \end{pmatrix},$$

où $\delta > 0$ est choisi suffisamment grand. Ce type de correction est souvent utilisé en pratique. Nous avons considéré le même algorithme de correction que celui utilisé dans le code de points intérieurs IPOPT [67] .

2.5.2 Théorème de convergence globale

À présent, donnons le théorème de convergence globale associé à l'algorithme E avec sa démonstration.

Théorème 2.5.4 *Si les suites $\{\mu_k\}$ et $\{\varepsilon_k\}$ convergent vers zéro et la suite $\{w_k\}$ existe, alors la suite $\{F(w_k, \mu_k)\}$ converge vers zéro.*

Preuve. D'après la définition de w_{k+1}, nous avons dans l'algorithme E pour $k \in \mathbb{N}$ et $\rho \in]0,1[$

$$\|F(w_{k+1},\mu_{k+1})\| \leq (1-\rho)\|F(w_k,\mu_k)\| + \varepsilon_k. \tag{2.29}$$

En notant $\bar{\varepsilon} := \sup_{k\in\mathbb{N}} \varepsilon_k$, nous avons

$$\|F(w_{k+1},\mu_{k+1})\| \leq (1-\rho)\|F(w_k,\mu_k)\| + \bar{\varepsilon},$$

et pour $k \in \mathbb{N}$,

$$\|F(w_{k+1},\mu_{k+1})\| - \frac{\bar{\varepsilon}}{\rho} \leq (1-\rho)\left(\|F(w_k,\mu_k)\| - \frac{\bar{\varepsilon}}{\rho}\right). \tag{2.30}$$

Nous pouvons donc en déduire que

$$l := \lim_{k\to+\infty} \sup \|F(w_k,\mu_k)\| \leq \frac{\bar{\varepsilon}}{\rho}.$$

En passant à la limite dans l'inégalité (2.29), nous avons

$$l \leq (1-\rho)\, l.$$

Comme $\rho \in]0,1[$, nous avons $l = 0$ et donc le résultat.

□

2.5.3 Analyse asymptotique

À présent, démontrons les propriétés de convergence locale que nous avons présentées dans le paragraphe 2.4. Pour cela, considérons l'hypothèse suivante.

Hypothèse 2.5.5 *La suite des nombres positifs $\{\varepsilon_k\}$ converge vers zéro et pour tout $k \in \mathbb{N}$, il existe $\nu > 0$ tel que*

$$\varepsilon_k \geq \nu\mu_{k+1}.$$

Théorème 2.5.6 *Supposons que les hypothèses 1.2.1-1.2.4, 2.4.1 et 2.5.5 soient vérifiées. Supposons que l'algorithme E génère une suite convergente $\{w_k\}$ vers la solution $w^* \in \mathbb{R}^{n+m}$. Alors, pour $k \in \mathbb{N}$ suffisamment grand, $w_{k+1} = w_k + d_k^w$ ce qui signifie*

$$||F(w_k + d_k^w, \mu_{k+1})|| \leq \rho||F(w_k,\mu_k)|| + \varepsilon_k.$$

En particulier, $w_k = w(\mu_k) + o(\mu_k)$ ce qui implique lorsque $w'(0)$ est non nul que le taux de convergence de $\{w_k\}$ vers w^* est le même que celui de $\{\mu_k\}$ vers zéro.

La preuve du théorème 2.5.6 repose sur les trois lemmes suivants. Comme $\{w_k\} \to w^*$ et $\{\mu_k\} \to 0$, nous supposons implicitement dans ces résultats que w est suffisamment proche de w^* et $\mu > 0$ est suffisamment petit.

Lemme 2.5.7 *Il existe une constante $\upsilon > 0$ telle que pour tout $k \in \mathbb{N}$,*

$$||w_{k+1} - w(\mu_{k+1})|| \leq \upsilon||w_k + d_k^w - w(\mu_{k+1})||.$$

Preuve. Supposons que $w_{k+1} = w_k + d_k^w$, alors l'inégalité est évidente pour $\upsilon = 1$. Supposons que w_{k+1} est généré par les itérations internes de l'algorithme E. Comme la jacobienne de F par rapport à w est uniformément inversible dans un voisinage de $(w^*, 0)$ (lemme 2.2.2), il existe $a > 0$ et $b > 0$ tels que pour tout w, w' suffisamment proches de w^* et $\mu > 0$ suffisamment petit on a

$$a||w - w'|| \leq ||F(w, \mu) - F(w', \mu)|| \leq b||w - w'||.$$

Pour $k \in \mathbb{N}$ grand, nous avons

$$\begin{aligned} b\|w_k + d_k^w - w(\mu_{k+1})\| &\geq \|F(w_k + d_k^w, \mu_{k+1}) - F(w(\mu_{k+1}), \mu_{k+1})\| \\ &= \|F(w_k + d_k^w, \mu_{k+1})\| \\ &> \rho\|F(w_k, \mu_k)\| + \varepsilon_k \\ &\geq \|F(w_{k+1}, \mu_{k+1})\| \\ &= \|F(w_{k+1}, \mu_{k+1}) - F(w(\mu_{k+1}), \mu_{k+1})\| \\ &\geq a\|w_{k+1} - w(\mu_{k+1})\|. \end{aligned}$$

Ainsi,

$$\|w_{k+1} - w(\mu_{k+1})\| \leq \frac{b}{a}\|w_k + d_k^w - w(\mu_{k+1})\|.$$

En prenant $\upsilon \geq \dfrac{b}{a}$, le résultat est obtenu.

□

Le prochain lemme montre une propriété fondamentale sur la distance qui sépare les itérés à la trajectoire. Cette propriété est la clé du taux d'analyse de convergence de $\{w_k\}$.

Lemme 2.5.8 *Pour tout $k \in \mathbb{N}$,*

$$||w_k - w(\mu_k)|| = o(\mu_k).$$

Preuve. D'après le lemme 2.5.7, nous avons pour $k \in \mathbb{N}$

$$||w_{k+1} - w(\mu_{k+1})|| \leq \upsilon||w_k + d_k^w - w(\mu_{k+1})||.$$

Par le lemme 2.4.3 où nous avons dans la formule (2.21) que $w_{k+1} = w_k + d_k^w$, nous pouvons en déduire

$$||w_{k+1} - w(\mu_{k+1})|| \leq \upsilon\zeta(||w_k - w(\mu_k)||^2 + \mu_k^2).$$

En notant $e_{k+1} := \upsilon\zeta||w_{k+1} - w(\mu_{k+1})||$, l'inégalité précédente s'écrit

$$e_{k+1} \leq e_k^2 + \varepsilon_k^2,$$

où $\varepsilon_k = \upsilon\zeta\mu_k$. En utilisant un raisonnement similaire à celui utilisé dans la preuve du théorème 2.4.2, nous avons

$$e_k = o(\mu_k),$$

et donc le résultat.

□

Lemme 2.5.9 *Pour $k \in \mathbb{N}$ suffisamment grand,*

$$w_{k+1} = w_k + d_k^w,$$

ce qui signifie qu'asymptotiquement l'algorithme E n'a plus besoin d'itérations internes.

Preuve. Il suffit de montrer que le test d'arrêt de la globalisation dans l'algorithme E est satisfait pour $k \in \mathbb{N}$ grand. Nous avons pour $k \in \mathbb{N}$ suffisamment grand

$$\begin{aligned} \|F(w_k + d_k^w, \mu_{k+1})\| &= \|F(w_k + d_k^w, \mu_{k+1}) - F(w(\mu_{k+1}), \mu_{k+1})\| \\ &\leq b\zeta(||w_k - w(\mu_k)||^2 + \mu_k^2) \\ &= o(\mu_k^2) + O(\mu_k^2) \\ &= O(\mu_k^2) \\ &= o(\mu_{k+1}). \end{aligned}$$

D'après l'hypothèse 2.5.5, nous avons $\|F(w_k + d_k^w, \mu_{k+1})\| \leq \varepsilon_k$ soit le résultat.

□

Avec ces différents lemmes, la preuve du théorème 2.5.6 est obtenue.

Preuve. D'après les lemmes 2.2.4 et 2.5.9, nous avons pour $k \in \mathbb{N}$ suffisamment grand

$$w_{k+1} = w_k + d_k^w \text{ et } w_k = w(\mu_k) + o(\mu_k).$$

D'après l'égalité (2.7), nous en déduisons que

$$w_k - w^* = \mu_k w'(0) + o(\mu_k).$$

En passant à la norme, nous obtenons

$$||w_k - w^*|| \sim \mu_k ||w'(0)||.$$

Donc lorsque $w'(0)$ est non nul, le taux de convergence de $\{w_k\}$ vers w^* est le même que celui de $\{\mu_k\}$ vers zéro.

□

Chapitre 3

Résultats numériques

Dans ce chapitre, nous présentons les résultats numériques que nous avons obtenus avec l'algorithme E décrit dans le paragraphe 2.5 du chapitre 2. Cet algorithme a été programmé sous MATLAB et testé sur des problèmes provenant des librairies COPS 3.0 [21] et CUTEr [33]. Pour savoir si la méthode primale-duale est robuste et fiable, nous l'avons comparée à la méthode SQP [10, 68]. Pour cela, nous avons programmé une variante de l'algorithme C décrit dans le chapitre 1. Pour comparer les deux méthodes, des profils de performance [20] ont été utilisés. Dans la section 3.1, nous donnons la liste des problèmes testés avec leurs caractéristiques. Les valeurs des paramètres de l'algorithme E sont précisées dans la partie 3.2. Dans le paragraphe 3.3, nous décrivons la méthode utilisée pour vérifier que l'inertie de la jacobienne de F par rapport à w est bien égale à $(n, m, 0)$. Dans la section 3.4, nous expliquons comment nous avons choisi la décroissance du paramètre de pénalité pour effectuer les tests. Dans la partie 3.5, nous présentons les résultats obtenus avec la méthode primale-duale et dans le paragraphe 3.6, nous les comparons aux résultats obtenus avec la méthode SQP [10, 68].

3.1 Liste des problèmes testés

Les algorithmes C et E ont été testés sur 113 problèmes non linéaires avec contraintes d'égalité provenant des librairies COPS 3.0 [21] et CUTEr [33]. Le tableau 3.1 donne la liste de ces problèmes avec pour chacun d'eux le nombre de variables n et le nombre de contraintes m. Nous avons uniquement considéré les problèmes où $n \neq m$ et $n > m$.

3.2 Valeurs des paramètres

Nous avons programmé sous MATLAB l'algorithme E correspondant à l'algorithme primal-dual global décrit dans le paragraphe 2.5 du chapitre 2. Les valeurs des paramètres dans l'algorithme programmé sont les suivantes.

Le test d'arrêt de l'algorithme est

$$||F(w, 0)|| \leq 10^{-8}.$$

Nom du problème	n	m	Nom du problème	n	m
AUG2D	212	96	HAGER1	10000	5000
AUG3DC	3873	1000	HAGER2	10000	5000
AUG3D	3873	1000	HAGER3	10000	5000
BT1	2	1	HS006	2	1
BT2	3	1	HS007	2	1
BT3	5	3	HS009	2	1
BT4	3	2	HS026	3	1
BT5	3	2	HS027	3	1
BT6	5	2	HS028	3	1
BT7	5	3	HS039	4	2
BT8	5	2	HS040	4	3
BT9	4	2	HS042	3	1
BT11	5	3	HS046	5	2
BT12	5	3	HS047	5	3
BYRDSPHR	3	2	HS048	5	2
CATENA	32	11	HS049	5	2
CHAIN1	599	450	HS050	5	3
CHAIN2	799	600	HS051	5	3
CHAIN3	999	750	HS052	5	3
CHNRSBNE	10	18	HS056	7	4
DECONVNE	61	40	HS061	3	2
DIXCHLNG	10	5	HS077	5	2
DTOC1NA	1485	990	HS078	5	3
DTOC1NB	1485	990	HS079	5	3
DTOC1NC	1485	990	HS100LNP	7	2
DTOC1ND	735	490	HS111LNP	10	3
DTOC2	5994	3996	LUKVLE1	100	98
DTOC5	9998	4999	LUKVLE3	100	2
DTOC6	10000	5000	LUKVLE6	99	49
ELEC1	360	120	LUKVLE7	100	4
ELEC2	450	150	LUKVLE9	100	6
ELEC3	600	200	LUKVLE11	98	64
EIGENA2	110	55	LUKVLE12	97	72
EIGENACO	110	55	LUKVLE13	98	64
EIGENCCO	30	15	LUKVLE14	98	64
EIGENBCO	110	55	LUKVLE15	97	72
EIGENB2	110	55	LUKVLE16	97	72
EIGENC2	110	55	LUKVLE17	97	72
GILBERT	1000	1	LUKVLE18	97	72
GENHS28	10	8	MARATOS	2	1
GRIDNETE	7564	3844	MWRIGHT	5	3
GRIDNETB	60	36	ORTHREGA	517	256
GRIDNETH	61	36	ORTHREGB	27	6

Nom du problème	n	m	Nom du problème	n	m
ORTHREGC	215	105	S321	2	1
ORTHRGDM	49	23	S322	2	1
ORTHREGD	209	103	S335	3	2
ORTHRDM2	8003	4000	S336	3	2
ORTHRGDS	155	76	S338	3	2
S216	2	1	S344	3	1
S219	4	2	S345	3	1
S235	3	1	S375	10	9
S252	3	1	S378	10	3
S269	5	3	S394	20	1
S316	2	1	S395	50	1
S317	2	1	SPMSQRT	4999	8329
S318	2	1			
S319	2	1			
S320	2	1			

TAB. 3.1 – Liste des 113 problèmes testés avec les algorithmes C et E provenant des librairies COPS 3.0 [21] et CUTEr [33].

Dans le test d'arrêt de la globalisation, nous avons considéré $\rho = 0.8$ et $\varepsilon_k = 10\mu_{k+1}$. Dans la fonction de mérite primale-duale, nous avons choisi $\sigma_{k+1} = 1$. Pour le calcul de la suite des pas α, une simple méthode de backtracking a été utilisée où le pas est divisé par deux et où $\eta = 10^{-2}$.

Comme point de départ $w_0 = (x_0, \lambda_0) \in \mathbb{R}^{n+m}$ de l'algorithme,

– x_0 est donné par le problème,

– λ_0 est solution au sens des moindres carrés du problème

$$\text{minimiser}_{\lambda \in \mathbb{R}^m} \quad ||\nabla f(x_0) + A(x_0)^\top \lambda||. \tag{3.1}$$

Ces valeurs sont assez similaires aux valeurs considérées dans l'algorithme C programmé (voir chapitre 1) de manière à comparer le plus objectivement possible la méthode primale-duale à la méthode SQP.

3.3 Régularisation de la jacobienne

Pour vérifier que la direction $d^w_{k,i}$ est une direction de descente de la fonction de mérite primale-duale $\Theta_{\mu,\sigma}$, nous calculons l'inertie de la matrice $F'_w(w_k, \mu)$ (voir proposition 2.5.2 dans le chapitre 2). Dans notre implémentation, nous avons utilisé l'interface MA57 (Harwell Subroutine Library) [4, 22] pour factoriser et calculer l'inertie de la matrice. Si l'inertie n'est pas correcte, i.e. différente de $(n, m, 0)$, alors la matrice $F'_w(w_k, \mu)$ est considérée sous la forme

$$\begin{pmatrix} \nabla^2_{xx}\mathcal{L}(w_k) + \delta Id & A(x_k)^\top \\ A(x_k) & -\mu Id \end{pmatrix},$$

où $\delta > 0$ est choisi suffisamment grand. Nous avons programmé le même algorithme de correction que celui utilisé dans IPOPT [67].

3.4 Décroissance du paramètre de pénalité

Une des difficultés principales dans l'algorithme E est de savoir comment faire décroître le paramètre de pénalité au cours des itérations. Pour que le résultat de convergence locale présenté dans le chapitre 2 (voir le théorème 2.4.2) soit vérifié, il faut que le paramètre de pénalité tende vers zéro avec une vitesse de convergence au plus superlinéaire sinon les itérés ne se rapprochent pas de la trajectoire de manière tangentielle (voir l'exemple dans le paragraphe 2.4.3). Pour effectuer les tests numériques, nous avons choisi une décroissance du paramètre de pénalité similaire à celle proposée par P. Armand, J. Benoist et D. Orban dans [6] où ils supposent un modèle dynamique de mise à jour du paramètre de pénalité. Cette technique permet d'avoir un meilleur ajustement entre la valeur du paramètre de pénalité et la valeur du résidu du système primal-dual au cours des itérations, ce qui améliore les performances en termes de nombre d'itérations et de robustesse. Pour effectuer les tests, nous avons donc considéré un modèle dynamique de mise à jour du paramètre de pénalité et nous avons en plus adapté la décroissance du paramètre à la décroissance de la norme des conditions d'optimalité perturbées du problème non linéaire (1.1) ou à la décroissance de la norme du lagrangien (1.3).

Comme valeur initiale du paramètre de pénalité, nous avons considéré

$$\mu_0 = \begin{cases} \max\left(0.9\ ;\ \dfrac{||F(w_0,0)||}{||\lambda_0||}\right) & \text{si } ||\lambda_0|| \neq 0, \\ 0.9 & \text{sinon.} \end{cases} \tag{3.2}$$

Comme choix dynamique du paramètre de pénalité, nous avons choisi

$$\mu_{k+1} = \max\left(\hat{\rho}_k \mu_k\ ;\ \min\left(\frac{\mu_k}{5}; \mu_k^{1.5}\right)\right), \tag{3.3}$$

où

$$\hat{\rho}_k = \min\left(\frac{||F(w_k,\mu_k)||}{||F(w_{k-1},\mu_{k-1})||}\ ;\ 0.8\right). \tag{3.4}$$

Pour éviter que le paramètre de pénalité décroisse trop vite au cours des itérations, nous avons contrôlé sa décroissance en l'adaptant soit à la décroissance de la norme des conditions d'optimalité perturbées (2.1) du problème non linéaire (1.1) soit à la décroissance de la norme du lagrangien (1.3). Cela permet de résoudre plus de problèmes et d'améliorer les performances. Nous avons testé trois possibilités d'adaptation du paramètre de pénalité.

– Cas 1 : la décroissance de μ est adaptée à la décroissance de $||F(w,\mu)||$. Dans la boucle qui permet d'adapter les deux décroissances, la valeur de F n'est pas recalculée chaque fois que la valeur de μ est modifiée.

- Cas 2 : la décroissance de μ est adaptée à la décroissance de $||F(w,\mu)||$. Dans la boucle qui permet d'adapter les deux décroissances, la valeur de F est ici recalculée chaque fois que la valeur de μ est modifiée.
- Cas 3 : la décroissance de μ est adaptée à la décroissance de $||\nabla f(x)+A(x)^\top\lambda||$.

Dans le paragraphe suivant, nous analysons les résultats numériques que nous avons obtenus avec l'algorithme E testé sur l'ensemble des problèmes présentés dans le tableau 3.1 pour les trois adaptations de la décroissance du paramètre de pénalité décrites auparavant.

3.5 Résultats

Afin de comparer les résultats numériques obtenus, nous avons évalué différentes grandeurs pour chacun des problèmes testés.

- nf représente le nombre d'évaluations de fonctions f et c,
- ng représente le nombre d'évaluations de gradients ∇f,
- nh représente le nombre d'évaluations de hessiens $\nabla^2 f$,
- nfact représente le nombre de factorisations de la matrice $F'_w(w,\mu)$,
- ffin représente la valeur finale de f au dernier point calculé par l'algorithme,
- $\|F\|$ représente la norme des conditions d'optimalité non perturbées en w, i.e. $\|F(w,0)\|$,
- t représente le temps CPU (en secondes) nécessaire à l'arrêt de l'algorithme,
- Info représente le mode d'arrêt de l'algorithme
  - Info = 0 lorsque l'algorithme s'arrête normalement,
  - Info = 1 lorsque le nombre maximal d'évaluations du nombre de fonctions et de contraintes est atteint,
  - Info = 2 lorsque le nombre maximal de factorisations de $F'_w(w,\mu)$ est atteint,
  - Info = 3 lorsque le nombre maximal de corrections de $F'_w(w,\mu)$ est atteint.

Les résultats numériques que nous avons obtenus avec l'algorithme E sont présentés dans l'annexe D pour les trois cas d'adaptation décrits dans le paragraphe précédent. Nous remarquons que l'information retournée est dans tous les cas 0 ou 1. Le choix de l'adaptation de la décroissance du paramètre de pénalité est important car les nombres d'évaluations nf, ng, nh et nfact varient selon le cas considéré. On peut noter que l'adaptation de la décroissance du paramètre de pénalité à la décroissance de la norme des conditions d'optimalité perturbées (2.1) du problème non linéaire (1.1) assure en général de meilleurs résultats. En revanche, adapter la décroissance du paramètre de pénalité à la norme du lagrangien (1.3) apporte quand même une amélioration des résultats pour plusieurs problèmes : BT1, CHAIN, ORTHREGD, S375. En effet, le nombre d'évaluations nf est plus petit ce qui signifie que l'algorithme E n'a pas eu besoin de beaucoup d'itérations internes pour résoudre le problème. Ces tests numériques sont assez satisfaisants et confirment l'efficacité de la méthode proposée. Pour comparer de manière plus précise les trois cas d'adaptation du paramètre de pénalité, nous avons représenté sur la figure 3.1 les profils

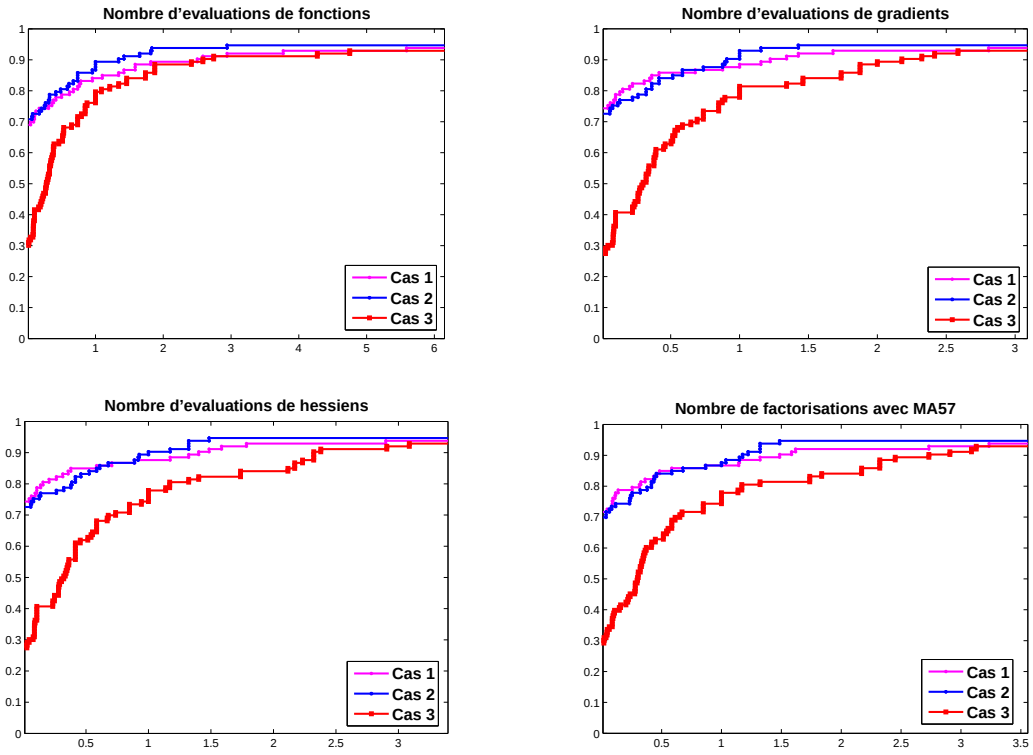
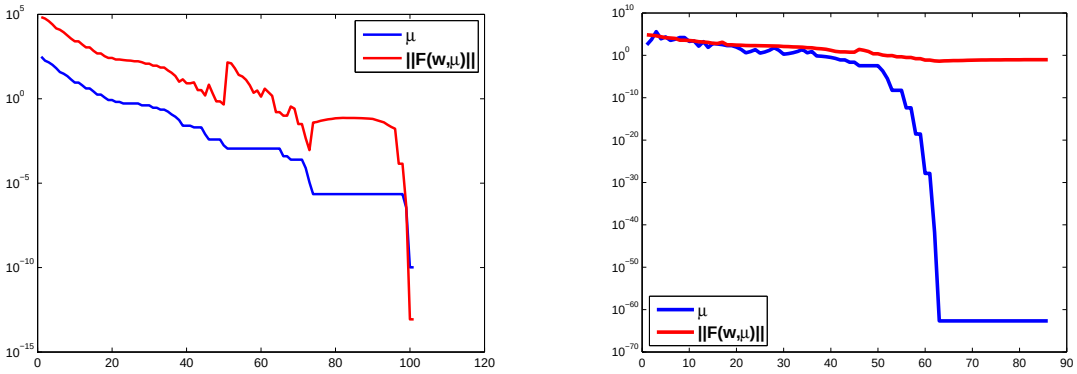


FIG. 3.1 – Profils de performance [20] des trois adaptations de la décroissance du paramètre de pénalité μ. Chaque courbe représente la proportion de problèmes qui sont résolus avec un nombre d'évaluations inférieur à 2^x fois le nombre d'évaluations de la meilleure méthode. Plus la courbe est au-dessus des autres, meilleure est la performance.

de performance [20] des différentes évaluations. Ces profils confirment que tous les problèmes testés n'ont pas été résolus. Dans tous les cas, l'adaptation la plus robuste correspond à celle de la décroissance de la norme des conditions d'optimalité perturbées (avec ou sans réévaluation de F dans la boucle). Elle est robuste avec une probabilité égale à 95%. D'après le profil associé au nombre d'évaluations de fonction, 70% des problèmes sont résolus lorsque la décroissance du paramètre de pénalité est adaptée à celle de la norme des conditions d'optimalité perturbées. Seulement 30% des problèmes le sont lorsque l'adaptation est faite par rapport à la norme du lagrangien. Ces probabilités restent les mêmes pour les autres évaluations. D'après les tableaux dans l'annexe D, le temps de calcul nécessaire à la résolution des problèmes est indépendant de l'adaptation considérée. Le temps le plus long est obtenu pour la résolution du problème ELEC de la librairie COPS 3.0 [21] lorsque $(n, m) = (450, 150)$ et $(n, m) = (600, 200)$. Ce problème consiste à optimiser la répartition des électrons sur une sphère. Il est difficile à résoudre car il a beaucoup de minima locaux. Le temps mis par l'algorithme E pour le résoudre est environ égal à 2 et 4 minutes.

Afin de montrer l'importance du choix dynamique du paramètre de pénalité dans l'algorithme E, nous avons résolu le problème ELEC lorsque $(n, m) = (450, 150)$ avec cet algorithme mais pour une décroissance du paramètre de pénalité non dynamique

et superlinéaire, i.e. lorsque $\mu_{k+1} = \mu_k^{1.5}$. La figure 3.2 représente la décroissance du paramètre de pénalité et de la norme des conditions d'optimalité perturbées au cours des itérations de l'algorithme E pour les deux choix du paramètre de pénalité.



choix dynamique de μ choix non dynamique de μ

FIG. 3.2 – Décroissance de μ et de $||F(w,\mu)||$ au cours des itérations de l'algorithme E lorsque le problème ELEC ($n = 450, m = 150$) est résolu selon le choix dynamique ou non de μ.

Pour le choix dynamique de μ donné par les formules (3.3) et (3.4), la décroissance du paramètre de pénalité au cours des itérations est à peu près la même que celle de la norme des conditions d'optimalité perturbées. En revanche, pour le choix non dynamique de μ, ce n'est plus vrai car le paramètre de pénalité décroît trop vite par rapport à la norme des conditions d'optimalité perturbées. Ceci peut provoquer des difficultés numériques pour la résolution du problème. C'est pour cette raison que le choix dynamique du paramètre de pénalité, comme dans [6], a été favorisé dans l'algorithme E.

3.6 Comparaison de la méthode primale-duale à la méthode SQP

Pour savoir si la méthode primale-duale que nous proposons est robuste et efficace, nous l'avons comparée à la méthode SQP rappelée dans chapitre 1. Les valeurs des paramètres de l'algorithme C sont données dans le paragraphe 1.4.4. Cet algorithme a été programmé sous MATLAB et testé sur les problèmes du tableau 3.1. Les résultats numériques que nous avons obtenus avec l'algorithme C sont présentés dans l'annexe E. Pour comparer le plus objectivement possible les deux méthodes, nous avons utilisé des profils de performance [20]. La figure 3.3 représente les profils correspondant aux nombres d'évaluations de fonctions, de gradients, de hessiens et de factorisations nécessaires aux deux méthodes pour résoudre les problèmes testés. Dans la méthode primale-duale, la décroissance du paramètre de pénalité est adaptée à la décroissance de la norme des conditions d'optimalité perturbées sans réévaluer F dans la boucle.

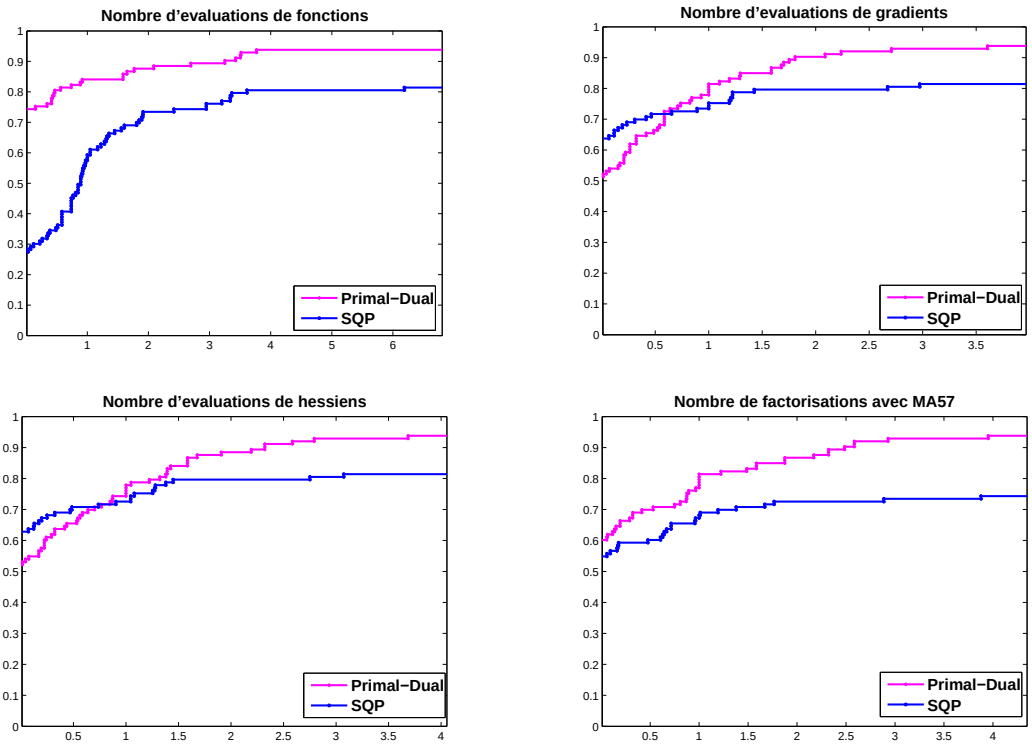


FIG. 3.3 – Profils de performance des deux méthodes : méthode primale-duale et méthode SQP.

Avec ces profils, nous remarquons que tous les problèmes testés n'ont pas été résolus. Le nombre d'évaluations de fonctions est meilleur pour la méthode primale-duale. Concernant les nombres d'évaluations de gradients et de hessiens, ils sont meilleurs pour la méthode SQP puisque 63% des problèmes sont résolus avec la méthode SQP contre 53% pour la méthode primale-duale. La supériorité apparente de la méthode SQP concernant le nombre d'évaluations de gradients est en partie due au fait que pour les problèmes quadratiques (fonction objectif quadratique et contraintes d'égalité linéaires), la méthode SQP converge en une seule itération. Concernant le nombre de factorisations effectuées avec MA57, il est clair que les résultats sont moins bons avec la méthode SQP. D'après l'annexe E, plusieurs problèmes ne sont pas résolus avec la méthode SQP car le nombre maximal de corrections de la matrice est atteint. Pour faire une comparaison plus précise et plus approfondie des deux méthodes, il faudrait programmer pour la méthode SQP des corrections du second ordre. Pour les deux approches, il faudrait améliorer la méthode utilisée pour le calcul de la suite des pas α. Ces résultats sont à approfondir mais ils montrent tout de même que la méthode primale-duale est robuste et efficace.

Chapitre 4

Conclusions et perspectives

Dans cette partie, un nouvel algorithme a été présenté pour résoudre des problèmes d'optimisation non linéaires avec contraintes d'égalité. Dans les années 70, une des premières méthodes utilisées pour résoudre ce type de problèmes était de remplacer le problème initial par un problème sans contrainte où le critère à minimiser est une fonction de pénalité [23]. Une autre méthode très efficace pour résoudre des problèmes d'optimisation non linéaires avec contraintes est la méthode de programmation quadratique successive (SQP) [10, 68]. Celle-ci transforme le problème initial en une suite de sous-problèmes quadratiques plus simples à résoudre.

La méthode que nous proposons consiste à résoudre un système primal-dual avec une méthode Newtonnienne. Ce système peut s'interpréter comme les conditions d'optimalité perturbées du problème initial ou comme les conditions d'optimalité du problème pénalisé. Sous certaines hypothèses, la solution du système primal-dual définit une trajectoire. Le point final de cette trajectoire est solution du système primal-dual non perturbé. L'analyse de convergence locale associée à notre méthode montre que les itérés générés par l'algorithme vont se rapprocher naturellement de la trajectoire. Ce résultat nous a permis d'envisager une globalisation de la méthode primale-duale qui tient compte de ce comportement asymptotique. L'algorithme local a donc été globalisé avec une technique de recherche linéaire en utilisant une fonction de mérite primale-duale. L'avantage de cette nouvelle approche est que nous évitons le mauvais conditionnement et l'effet Maratos [48] présents avec les méthodes de pénalisation et SQP. L'algorithme global associé à la méthode primale-duale a été programmé et testé sur plusieurs librairies de problèmes [21, 33]. Une étude comparative a été effectuée entre la méthode primale-duale et la méthode SQP. Les résultats obtenus numériquement sont satisfaisants et confirment l'efficacité de la méthode proposée.

Concernant ce projet de recherche, plusieurs perspectives sont à envisager :

- tester d'autres fonctions de mérite primales-duales ;
- utiliser une méthode de régions de confiance dans la globalisation de l'algorithme ;
- utiliser une méthode de Newton inexacte pour la résolution du système primal-dual ;

– étendre la méthode aux problèmes d'optimisation non linéaires avec contraintes d'inégalité.

Annexe A

Coefficients d'apodisation

Dans cette annexe, nous présentons les amplitudes optimales des champs trouvées avec la méthode des coefficients de Fourier pour chaque fonction d'apodisation testée et pour plusieurs largeurs ε de ces fonctions. Nous avons considéré un hypertélescope constitué de neuf pupilles ($m = 4$) alignées et bord à bord ($T = 1$). Comme la configuration de l'instrument est supposée symétrique par rapport à l'origine, seules les cinq amplitudes des champs à droite de zéro sont données. Pour la présentation des résultats, ces valeurs sont normalisées par rapport à l'amplitude maximale.

ε	a_0	a_1	a_2	a_3	a_4
0.01	1.0000	0.9998	0.9993	0.9985	0.9974
0.02	1.0000	0.9993	0.9974	0.9941	0.9895
0.03	1.0000	0.9985	0.9941	0.9867	0.9765
0.04	1.0000	0.9974	0.9895	0.9765	0.9584
0.05	1.0000	0.9959	0.9836	0.9634	0.9355
0.06	1.0000	0.9941	0.9765	0.9475	0.9079
0.07	1.0000	0.9920	0.9681	0.9290	0.8759
0.08	1.0000	0.9895	0.9584	0.9079	0.8399
0.09	1.0000	0.9867	0.9475	0.8843	0.8000
0.10	1.0000	0.9836	0.9355	0.8584	0.7568
0.11	1.0000	0.9802	0.9223	0.8303	0.7106
0.12	1.0000	0.9765	0.9079	0.8000	0.6618
0.13	1.0000	0.9724	0.8925	0.7679	0.6109
0.14	1.0000	0.9681	0.8759	0.7341	0.5583
0.15	1.0000	0.9634	0.8584	0.6986	0.5046
0.16	1.0000	0.9584	0.8399	0.6618	0.4500
0.17	1.0000	0.9531	0.8204	0.6238	0.3952
0.18	1.0000	0.9475	0.8000	0.5848	0.3406
0.19	1.0000	0.9417	0.7788	0.5450	0.2867
0.20	1.0000	0.9355	0.7568	0.5046	0.2339

Tab. A.1 – Amplitudes optimales des champs normalisées pour la fonction Porte.

ε	a_0	a_1	a_2	a_3	a_4
0.02	1.0000	0.9997	0.9987	0.9970	0.9947
0.04	1.0000	0.9987	0.9947	0.9882	0.9791
0.06	1.0000	0.9970	0.9882	0.9736	0.9535
0.08	1.0000	0.9947	0.9791	0.9535	0.9186
0.10	1.0000	0.9918	0.9675	0.9281	0.8751
0.12	1.0000	0.9882	0.9535	0.8979	0.8243
0.14	1.0000	0.9840	0.9372	0.8631	0.7673
0.16	1.0000	0.9791	0.9186	0.8243	0.7054
0.18	1.0000	0.9736	0.8979	0.7820	0.6401
0.20	1.0000	0.9675	0.8751	0.7368	0.5728
0.22	1.0000	0.9608	0.8506	0.6893	0.5050
0.24	1.0000	0.9535	0.8243	0.6401	0.4380
0.26	1.0000	0.9456	0.7965	0.5897	0.3732
0.28	1.0000	0.9372	0.7673	0.5389	0.3117
0.30	1.0000	0.9281	0.7368	0.4881	0.2546
0.32	1.0000	0.9186	0.7054	0.4380	0.2025
0.34	1.0000	0.9085	0.6731	0.3892	0.1562
0.36	1.0000	0.8979	0.6401	0.3420	0.1160
0.38	1.0000	0.8867	0.6066	0.2970	0.0822
0.40	1.0000	0.8751	0.5728	0.2546	0.0547

TAB. A.2 – Amplitudes optimales des champs normalisées pour la fonction Triangle.

ε	a_0	a_1	a_2	a_3	a_4
0.04	1.0000	1.0000	0.9999	0.9998	0.9996
0.05	1.0000	0.9999	0.9998	0.9994	0.9990
0.06	1.0000	0.9999	0.9995	0.9989	0.9980
0.07	1.0000	0.9998	0.9991	0.9979	0.9963
0.08	1.0000	0.9996	0.9984	0.9964	0.9937
0.09	1.0000	0.9994	0.9975	0.9944	0.9900
0.10	1.0000	0.9990	0.9962	0.9913	0.9846
0.11	1.0000	0.9986	0.9944	0.9875	0.9779
0.12	1.0000	0.9980	0.9922	0.9825	0.9691
0.13	1.0000	0.9973	0.9893	0.9761	0.9579
0.14	1.0000	0.9964	0.9858	0.9682	0.9442
0.15	1.0000	0.9953	0.9814	0.9587	0.9276
0.16	1.0000	0.9940	0.9762	0.9472	0.9080
0.17	1.0000	0.9924	0.9700	0.9337	0.8851
0.18	1.0000	0.9906	0.9628	0.9181	0.8588
0.19	1.0000	0.9884	0.9544	0.9002	0.8291
0.20	1.0000	0.9859	0.9448	0.8799	0.7959
0.21	1.0000	0.9831	0.9339	0.8571	0.7594
0.22	1.0000	0.9799	0.9217	0.8319	0.7197
0.23	1.0000	0.9763	0.9081	0.8042	0.6770
0.24	1.0000	0.9722	0.8930	0.7741	0.6318
0.25	1.0000	0.9677	0.8765	0.7416	0.5844

TAB. A.3 – Amplitudes optimales des champs normalisées pour la fonction $\mathcal{C}^\infty$.

ε	a_0	a_1	a_2	a_3	a_4
0.01	1.0000	0.9997	0.9990	0.9977	0.9959
0.02	1.0000	0.9990	0.9959	0.9907	0.9836
0.03	1.0000	0.9977	0.9907	0.9793	0.9634
0.04	1.0000	0.9959	0.9836	0.9634	0.9357
0.05	1.0000	0.9936	0.9745	0.9433	0.9010
0.06	1.0000	0.9907	0.9634	0.9192	0.8600
0.07	1.0000	0.9874	0.9505	0.8913	0.8134
0.08	1.0000	0.9836	0.9357	0.8600	0.7622
0.09	1.0000	0.9793	0.9192	0.8255	0.7073
0.10	1.0000	0.9745	0.9010	0.7884	0.6496
0.11	1.0000	0.9692	0.8812	0.7488	0.5902
0.12	1.0000	0.9634	0.8600	0.7073	0.5301
0.13	1.0000	0.9572	0.8373	0.6643	0.4701
0.14	1.0000	0.9505	0.8134	0.6201	0.4113
0.15	1.0000	0.9433	0.7884	0.5752	0.3544
0.16	1.0000	0.9357	0.7622	0.5301	0.3001
0.17	1.0000	0.9276	0.7352	0.4850	0.2493
0.18	1.0000	0.9192	0.7073	0.4405	0.2022
0.19	1.0000	0.9103	0.6788	0.3968	0.1595
0.20	1.0000	0.9010	0.6496	0.3544	0.1213

TAB. A.4 – Amplitudes optimales des champs normalisées pour la fonction de Hanning.

ε	a_0	a_1	a_2	a_3	a_4
0.01	1.0000	0.9997	0.9987	0.9971	0.9949
0.02	1.0000	0.9987	0.9949	0.9886	0.9799
0.03	1.0000	0.9971	0.9886	0.9746	0.9552
0.04	1.0000	0.9949	0.9799	0.9552	0.9215
0.05	1.0000	0.9921	0.9687	0.9307	0.8796
0.06	1.0000	0.9886	0.9552	0.9015	0.8306
0.07	1.0000	0.9846	0.9394	0.8680	0.7756
0.08	1.0000	0.9799	0.9215	0.8306	0.7160
0.09	1.0000	0.9746	0.9015	0.7899	0.6530
0.10	1.0000	0.9687	0.8796	0.7463	0.5880
0.11	1.0000	0.9622	0.8559	0.7005	0.5225
0.12	1.0000	0.9552	0.8306	0.6530	0.4577
0.13	1.0000	0.9476	0.8038	0.6044	0.3948
0.14	1.0000	0.9394	0.7756	0.5553	0.3348
0.15	1.0000	0.9307	0.7463	0.5062	0.2788
0.16	1.0000	0.9215	0.7160	0.4577	0.2273
0.17	1.0000	0.9118	0.6848	0.4103	0.1810
0.18	1.0000	0.9015	0.6530	0.3644	0.1401
0.19	1.0000	0.8908	0.6207	0.3204	0.1049
0.20	1.0000	0.8796	0.5880	0.2788	0.0753

TAB. A.5 – Amplitudes optimales des champs normalisées pour la fonction de Hamming.

ε	a_0	a_1	a_2	a_3	a_4
0.01	1.0000	0.9998	0.9992	0.9982	0.9968
0.02	1.0000	0.9992	0.9968	0.9928	0.9872
0.03	1.0000	0.9982	0.9928	0.9838	0.9714
0.04	1.0000	0.9968	0.9872	0.9714	0.9497
0.05	1.0000	0.9950	0.9801	0.9557	0.9224
0.06	1.0000	0.9928	0.9714	0.9368	0.8900
0.07	1.0000	0.9902	0.9613	0.9148	0.8530
0.08	1.0000	0.9872	0.9497	0.8900	0.8119
0.09	1.0000	0.9838	0.9368	0.8626	0.7675
0.10	1.0000	0.9801	0.9224	0.8329	0.7204
0.11	1.0000	0.9760	0.9068	0.8011	0.6712
0.12	1.0000	0.9714	0.8900	0.7675	0.6207
0.13	1.0000	0.9666	0.8720	0.7324	0.5697
0.14	1.0000	0.9613	0.8530	0.6960	0.5186
0.15	1.0000	0.9557	0.8329	0.6587	0.4683
0.16	1.0000	0.9497	0.8119	0.6207	0.4191
0.17	1.0000	0.9434	0.7901	0.5825	0.3718
0.18	1.0000	0.9368	0.7675	0.5441	0.3266
0.19	1.0000	0.9298	0.7442	0.5059	0.2841
0.20	1.0000	0.9224	0.7204	0.4683	0.2445

TAB. A.6 – Amplitudes optimales des champs normalisées pour la fonction de Blackman.

Annexe B

Modèles pour l'optimisation de la dynamique

Les modèles correspondant à la méthode de l'optimisation de la dynamique sont présentés dans cette annexe. Ils ont été écrits avec le langage de modélisation AMPL [27]. Les variables d'optimisation sont les amplitudes des champs et/ou les positions des pupilles. Pour modéliser les problèmes, nous avons utilisé les normes $\|\cdot\|_2$ et $\|\cdot\|_\infty$ définies dans le chapitre 1.

B.1 Optimisation des amplitudes des champs

B.1.1 Norme $\|\cdot\|_2$

```
function besselJ1;

param pi := 4*atan(1);
param n >= 2;
param p := n/2;
param d > 0;
param m > 0;
param umin > 0;
param umax > umin;
param u {j in 1..m} := pi*(umin+(j-1)*(umax-umin)/(m-1));
param x {1..p};

var a {1..p} >= 0;
var t;
var y {j in 1..m} = 4*besselJ1(u[j])*sum {k in 1..p} a[k]*
                    cos(2/d*u[j]*x[k]);

minimize dynamique : sum {j in 1..m} y[j]^2;
subject to asum : sum {k in 1..p} a[k] = 0.5;

data;
```

```
param n := 8;
param d := 1;
param m := 80;
param umin := 0.25;
param umax := 0.75;
let x[1] := 0.5;
let x[2] := 1.5;
let x[3] := 2.5;
let x[4] := 3.5;
let a{1..p} := 0.125;

solve;
```

B.1.2 Norme $\|\cdot\|_\infty$

```
function besselJ1;

param pi := 4*atan(1);
param n >= 2;
param p :=n/2;
param d > 0;
param m > 0;
param umin > 0;
param umax > umin;
param u {j in 1..m} := pi*(umin+(j-1)*(umax-umin)/(m-1);
param x {1..p};

var a {1..p} >= 0;
var t;
var y {j in 1..m} = 4*besselJ1(u[j])*sum {k in 1..p} a[k]*
                    cos(2/d*u[j]*x[k]);

minimize dynamique : t;
subject to dynup {j in 1..m} : y[j] <= t*u[j];
subject to dynlo {j in 1..m} : y[j] >= -t*u[j];
subject to asum : sum {k in 1..p} a[k] = 0.5;

data;
param n := 8;
param d := 1;
param m := 80;
param umin := 0.25;
param umax := 0.75;
let x[1] := 0.5;
let x[2] := 1.5;
let x[3] := 2.5;
```

```
let x[4] := 3.5;
let a{1..p} := 0.125;

solve;
```

B.2 Optimisation des positions des pupilles

B.2.1 Norme $\|\cdot\|_2$

```
function besselJ1;

param pi := 4*atan(1);
param n >= 2;
param p := n/2;
param d > 0;
param m > 0;
param umin > 0;
param umax > umin;
param u {j in 1..m} := pi*(umin+(j-1)*(umax-umin)/(m-1));
param a {1..p};

var x {1..p} >= 0;
var t;
var y {j in 1..m} = 4*besselJ1(u[j])*sum {k in 1..p} a[k]*
                    cos(2/d*u[j]*x[k]);

minimize dynamique : sum {j in 1..m} y[j]^2;
subject to xone : x[1] >= d/2;
subject to disjoint {k in {1..p-1}} : x[k+1] - x[k] >= d;

data;
param n := 8;
param d := 1;
param m := 80;
param umin := 0.25;
param umax := 0.75;
let x[1] := 0.5;
let x[2] := 1.5;
let x[3] := 2.5;
let x[4] := 3.5;
let a{1..p} := 0.125;

solve;
```

B.2.2 Norme $\|\cdot\|_\infty$

```
function besselJ1;

param pi := 4*atan(1);
param n >= 2;
param p := n/2;
param d > 0;
param m > 0;
param umin > 0;
param umax > umin;
param u {j in 1..m} := pi*(umin+(j-1)*(umax-umin)/(m-1));
param a {1..p};

var x {1..p} >= 0;
var t;
var y {j in 1..m} = 4*besselJ1(u[j])*sum {k in 1..p} a[k]*
                    cos(2/d*u[j]*x[k]);

minimize dynamique : t;
subject to dynup {j in 1..m} : y[j] <= t*u[j];
subject to dynlo {j in 1..m} : y[j] >= -t*u[j];
subject to xone : x[1] >= d/2;
subject to disjoint {k in {1..p-1}} : x[k+1] - x[k] >= d;

data;
param n := 8;
param d := 1;
param m := 80;
param umin := 0.25;
param umax := 0.75;
let x[1] := 0.5;
let x[2] := 1.5;
let x[3] := 2.5;
let x[4] := 3.5;
let a{1..p} := 0.125;

solve;
```

B.3 Optimisation des amplitudes des champs et des positions des pupilles

B.3.1 Norme $\|\cdot\|_2$

```
function besselJ1;
```

```
param pi := 4*atan(1);
param n >= 2;
param p := n/2;
param d > 0;
param m > 0;
param umin > 0;
param umax > umin;
param u {j in 1..m} := pi*(umin+(j-1)*(umax-umin)/(m-1));

var x {1..p};
var a {1..p};
var t;
var y {j in 1..m} = 4*besselJ1(u[j])*sum {k in 1..p} a[k]*
                    cos(2/d*u[j]*x[k]);

minimize dynamique : sum {j in 1..m} y[j]^2;
subject to xone : x[1] >= d/2;
subject to disjoint {k in {1..p-1}} : x[k+1] - x[k] >= d;
subject to asum : sum {k in 1..p} a[k] = 0.5;

data;
param n := 8;
param d := 1;
param m := 80;
param umin := 0.25;
param umax := 0.75;
let x[1] := 0.5;
let x[2] := 1.5;
let x[3] := 2.5;
let x[4] := 3.5;
let a{1..p} := 0.125;

solve;
```

B.3.2 Norme $\|\cdot\|_\infty$

```
function besselJ1;

param pi := 4*atan(1);}
param n >= 2;
param p := n/2;
param d > 0;
param m > 0;
param umin > 0;
param umax > umin;
param u {j in 1..m} := pi*(umin+(j-1)*(umax-umin)/(m-1));
```

```
var x {1..p};
var a {1..p};
var t;
var y {j in 1..m} = 4*besselJ1(u[j])*sum {k in 1..p} a[k]*
                    cos(2/d*u[j]*x[k]);

minimize dynamique : t;
subject to dynup {j in 1..m}: y[j] <= t*u[j];
subject to dynlo {j in 1..m}: y[j] >= -t*u[j];
subject to xone: x[1] >= d/2;
subject to disjoint {k in {1..p-1}}: x[k+1] - x[k] >= d;
subject to asum : sum {k in 1..p} a[k] = 0.5;

data;
param n := 8;
param d := 1;
param m := 80;
param umin := 0.25;
param umax := 0.75;
let x[1] := 0.5;
let x[2] := 1.5;
let x[3] := 2.5;
let x[4] := 3.5;
let a{1..p} := 0.125;

solve;
```

Annexe C

Optimization of a one dimensional hypertelescope for a direct imaging in astronomy

Paul Armand[1], Joël Benoist[1], Elsa Bousquet[1],
Laurent Delage[2], Serge Olivier[2] and François Reynaud[2]

Abstract. We describe an application of nonlinear optimization in interferometric optical astronomy. The aim is to find the relative positions of the output pupils and the modulus of the beams through each pupil of a linear array of telescopes in order to design an instrument capable of imaging exoplanets. The problem is modelized under the form of a semi-infinite nonlinear minimization problem. The model problem is transformed by a simple discretization into a minimization problem with a finite number of constraints, then it is solved by using a minimization solver. Numerical experiments are reported.

Key words. Nonlinear programming, Constrained optimization, High contrast imaging, Exoplanet, Hypertelescope, Astronomical technique, Optical systems

C.1 Introduction

The new generation of high resolution optical imaging system for astronomy is based on the linkage of a telescope array in order to reach micro or nanoradian resolution [46]. The principle of the image restoration uses an indirect method. The interferometric signals give partial information on the spectrum of the object spatial distribution. A numerical image reconstruction is necessary to post-process the data. Using this technique never allows to select the light of one pixel in the observed field for example to achieved a direct spectral analysis. The hypertelescope concepts proposed by Labeyrie [41], Vakili et al [63] and Reynaud-Delage [61] (see also the recent book [42]), answer this problem. By using a specific conditioning

[1]XLIM UMR 6172, Département Mathématiques et Informatique, CNRS et Université de Limoges (France) ; e-mail : paul.armand@xlim.fr
[2]XLIM, Département Photonique ; email francois.reynaud@xlim.fr

of the beam coming from the telescopes, it is possible to perform a direct imaging thanks to an accurate equalisation of the optical paths and a pupil densification, see figure C.1. The first operation so called cophasing has to be achieved with a

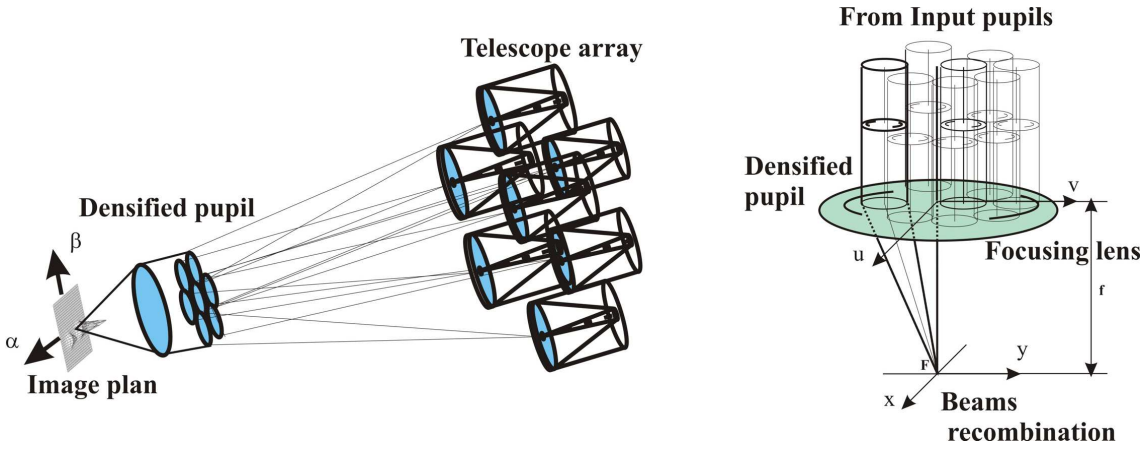


FIG. C.1 – Structure of a hypertelescope.

sub-micrometric accuracy. Over all this process, coherent properties of the beams have to be preserved taking care of polarization and dispersion differential effects. The pupil densification is a specific step of the hypertelescope design. The input pupil mapping is homothetically reduced and the resulting beams are expended in order to maximize the densified pupil coverage. The goal of such instruments is to design an optical instrument with a strong dynamic in the frame of a high resolution imaging. For example, the angular separation between a star and an exoplanet is expected to be in the range of nanoradian and the ratio of their relative intensities can be less than 10^{-6}. Of course, this tremendous result will be reached in a limited field and the number of pixels remains low as reported in [41, 61, 63]. In a general way, the optimization of optical instrument is adressed in various domains such as microscopy, see for example [49]. The purpose of this paper is to find the modulus of each beam and the relative positions of the output pupils to obtain high resolution and dynamic of the beams recombination. In this preliminary study we will consider a linear array of telescopes corresponding to pupils arranged along a straight line. Applied optimization to the design of an optical instrument capable to image exoplanets has been already investigated by Kasdin etal [40] and Vanderbei [64]. As in Vanderbei approach, we propose to formulate the design problem as a semi-infinite minimization problem. After discretization the model problem is transformed into a constrained nonlinear minimization problem. Then the problem is solved by using a nonlinear optimization solver. The paper is organized as follows. We first present in Section C.2 our general experimental setup, then we state the particular case of a linear array of telescopes in Section C.3. The optimization model is presented in Section C.4. A starting point strategy for the optimization solver is discussed in Section C.5, then numerical experiments are presented in Section C.6.

C.2 Densified pupil and point spread function

This section is focused on the optical field description in the densified pupil and to derive the corresponding point spread function.

The optical field of a monochromatic plane wave through a pupil is characterised by a modulus and a phase. We assume that all beams have the same phase. It follows that the optical field of n pupils centered at (u_k, v_k), $k = 1, \ldots, n$ and with the same diameter d is given by

$$g(u,v) = \sum_{k=1}^{n} a_k \mathbf{1}_{B_k}(u,v),$$

where a_k is the modulus of beam k and $\mathbf{1}_{B_k}$ is the characteristic function of the closed disk with center (u_k, v_k) and diameter d.

In the image plane, the optical field is the Fourier transform of the function g and is defined by

$$\hat{g}(x,y) = \iint g(u,v) e^{-\frac{2i\pi}{\lambda f}(xu+yv)} \, \mathrm{d}u \, \mathrm{d}v,$$

where λ is the wave length and f is the focal length. The image plane intensity is given by the point spread function (PSF) and is defined to be the square of the modulus of $\hat{g}$. We defined the normalized PSF by

$$\Psi(x,y) = \frac{|\hat{g}(x,y)|^2}{|\hat{g}(0,0)|^2}.$$

Figure C.2 shows the graph of a normalized PSF obtained with four pupils.Our aim is to find moduli a_k and positions (u_k, v_k) such that the main central lobe is as narrow as possible, to get a high resolution, and the secondary lobes are as low as possible, to get a strong dynamic.

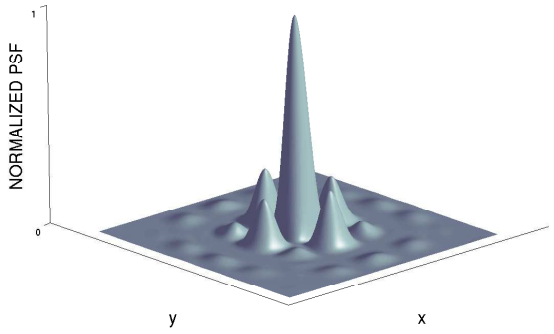


FIG. C.2 – Normalized PSF with four pupils.

Let B be the closed disk with center 0 and diameter d. Let s be the surface of one pupil of diameter d. Since all pupils have the same diameter, the normalized PSF can be written as

$$\Psi(x,y) = \frac{1}{s}|\hat{h}(x,y)|^2 \frac{\left|\sum_{k=1}^{n} a_k e^{-\frac{2i\pi}{\lambda f}(xu_k+yv_k)}\right|^2}{\left|\sum_{k=1}^{n} a_k\right|^2}, \tag{C.1}$$

where

$$\hat{h}(x,y) = \iint e^{-\frac{2i\pi}{\lambda f}(xu+yv)} \mathbf{1}_B(u,v)\, \mathrm{d}u\, \mathrm{d}v$$

is the PSF associated to only one pupil. Since the normalized PSF is homogeneous with respect to a_k, we can assume that

$$\sum_{k=1}^{n} a_k = 1. \tag{C.2}$$

C.3 Linear array of telescopes

In this first paper we consider the particular case of a linear array of telescopes, that is $v_k = 0$ for all $k = 1, \ldots, n$ in (C.1). Taking (C.2) into account, the normalized PSF becomes

$$\Psi(x,y) = \frac{1}{s^2}|\hat{h}(x,y)|^2 \left| \sum_{k=1}^{n} a_k e^{-\frac{2i\pi}{\lambda f} x u_k} \right|^2.$$

In the right hand side of the above formula, the term which contains the optimization parameters a_k and u_k does not depend on y. It follows that in our study we can then set y to any arbitrary value. Let us set $y = 0$. By introducing the polar coordinates $u = r\cos\theta$ and $v = r\sin\theta$, the value of $\hat{h}$ at $(x, 0)$ can be written

$$\hat{h}(x,0) = \int_0^{d/2} \int_0^{2\pi} e^{-\frac{2i\pi}{\lambda f} x r \cos\theta} r\, \mathrm{d}\theta\, \mathrm{d}r.$$

By using the change of variable $\alpha = \dfrac{d}{\lambda f} x$, we can then write $\hat{h}$ as a function of α, that is

$$\hat{h}(\alpha) = \frac{2s}{\pi\alpha} J_1(\pi\alpha)$$

where J_1 is the first order Bessel function of the first kind. The normalized PSF can then be written as a function of α and becomes

$$\Psi(\alpha) = \left| \frac{2}{\pi\alpha} J_1(\pi\alpha) \right|^2 \left| \sum_{k=1}^{n} a_k e^{-\frac{2i\pi u_k}{d}\alpha} \right|^2. \tag{C.3}$$

C.4 Optimization model

In the particular case of regularly spaced pupils, the second term in the right hand side of (C.3) can be interpreted as a truncated Fourier series of some periodic function. This function can be seen as a size function to obtain a given PSF. This technique is used for the design of antenna array in telecommunication network where such a function is called an apodization function. In this specific case the parameters a_k are then the Fourier coefficients of the apodization function (see e.g., [36]). In our case, the inconvenience of this approach is that the values of dynamic and resolution do not depend straightforwardly on the choice of the apodization function, so that

it is not clear how to choose such a function to get optimal a_k. Moreover, it is not sure that the choice of a regular spacing of the pupils leads to the best solution. We do not go further along this approach.

We formulate the problem as follows. We must find the moduli a_k and the positions u_k for which the PSF is as small as possible in a region close to the main central lobe. We define an interval $[\alpha_{\min}, \alpha_{\max}]$ of α values for which we want that $\Psi(\alpha)$ is as small as possible. It is in this interval that the main central lobe of a secondary source of light could be detected. We call it the *clean field of view* (CLF). Our optimization model is then

$$\begin{array}{ll} \text{minimize} & \max\{\Psi(\alpha) : \alpha_{\min} \leq \alpha \leq \alpha_{\max}\} \\ \text{subject to} & u_{k+1} - u_k \geq d, \quad \text{for } k = 1, \ldots, n-1 \\ & \displaystyle\sum_{k=1}^{n} a_k = 1, \\ & a_k \geq 0, \quad \text{for } k = 1, \ldots, n. \end{array} \tag{C.4}$$

The first constraint is those of non overlapping of the pupils and the second is the normalization constraint. To transform the problem into a computational form, by using (C.3) we rewrite (C.4) as

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & \left| \dfrac{J_1(\pi\alpha)}{\alpha} \displaystyle\sum_{k=1}^{n} a_k e^{-\frac{2i\pi u_k}{d}\alpha} \right| \leq t, \quad \alpha_{\min} \leq \alpha \leq \alpha_{\max} \\ & u_{k+1} - u_k \geq d, \quad \text{for } k = 1, \ldots, n-1 \\ & \displaystyle\sum_{k=1}^{n} a_k = 1, \\ & a_k \geq 0, \quad \text{for } k = 1, \ldots, n. \end{array} \tag{C.5}$$

This is a semi-infinite nonlinear optimization problem. There is only a finite number of optimization variables, but an infinite number of constraints. To solve the problem we simply use a discretization method.

Suppose now that there is an even number $n = 2m$ of pupils. The case of an odd number could be considered similarly. We assume also that the pupils are symmetrically placed around zero. By discretizing the interval $[\alpha_{\min}, \alpha_{\max}]$ in (C.5) with q points uniformly spaced, the optimization model becomes

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & -t \leq \dfrac{J_1(\pi\alpha_j)}{\alpha_j} \displaystyle\sum_{k=1}^{m} a_k \cos(\frac{2\pi}{d} u_k \alpha_j) \leq t \quad \text{for } j = 1, \ldots, q \\ & u_1 \geq \dfrac{d}{2}, \\ & u_{k+1} - u_k \geq d, \quad \text{for } k = 1, \ldots, m-1 \\ & \displaystyle\sum_{k=1}^{m} a_k = \frac{1}{2}, \\ & a_k \geq 0, \quad \text{for } k = 1, \ldots, m. \end{array} \tag{C.6}$$

This is a nonlinear optimization problem with a finite number of variables and constraints. In our experiments, the number of discretization points was set to $q = 10n$.

The solution of problem (C.6) requires the use of a nonlinear optimization solver. In our experiments we used the modelling language AMPL [27] to formulate the problem. The values of the Bessel function in (C.6) were computed by using a routine from the GNU Scientific Library. The problem has been solved by using the open source solver IPOPT [67]. It is an optimization package designed to compute local solutions of a nonlinear optimization problem of the form

$$\begin{array}{ll} \text{minimize} & f(x), \\ \text{subject to} & c_i(x) = \text{ or } \leq 0, \quad i = 1, \ldots, m \end{array}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $c : \mathbb{R}^n \to \mathbb{R}^m$ are objective and constrained functions. It uses an interior point algorithm to solve the first order optimality conditions of the minimization problem. Note that other approaches, such as sequential quadratic programming or augmented Lagrangian algorithm, could also be used to solve the problem. But here, an interior point method is well adapted because the problem possesses a great number of inequalities and these methods are well competitive in that case.

C.5 Starting point strategy

A difficulty when solving problem (C.6) is to avoid to be trapped by a locally optimal point which is not a global minimizer, a common situation which is hard to dealt with in nonlinear optimization. To cope with this difficulty we considered to solve several occurrences of the same problem but with a random starting point for the solver. We solved an instance of problem (C.6) with eight pupils ($m = 4$), $d = 1$, $\alpha_{\min} = 0.25$ and $\alpha_{\max} = 0.75$. The starting points were chosen following a uniform distribution such that $\frac{7}{2}d \leq u_4 \leq 14d$ and that the pupils do not overlap. To compare the computed solutions, we define two measures. The first one is the optimization criterion, which we call the *dynamic* of a configuration. It is defined as the inverse of the maximum value of the normalized PSF over the CLF, that is

$$D = \frac{1}{\max\{\Psi(\alpha) : \alpha_{\min} \leq \alpha \leq \alpha_{\max}\}}.$$

The second measure is on the relative amount of energy that lands in the interval $[0, \alpha_{\min}]$. We define the *central flux* by

$$F = \frac{\int_0^{\alpha_{\min}} \Psi(\alpha)\, \mathrm{d}\alpha}{\int_0^{\infty} \Psi(\alpha)\, \mathrm{d}\alpha}.$$

Figure C.3 shows the values of dynamic and central flux for a sample of 10^4 occurrences of minimization procedures. It is clear that there is a best solution amongst all of the others. Figure C.4 gives a representation of this optimal solution and the graph of the corresponding normalized PSF in linear and logarithmic scales.

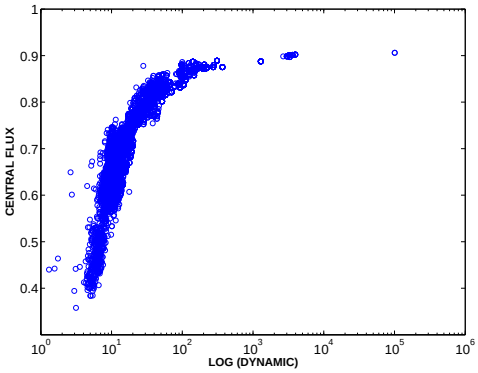


FIG. C.3 – Comparison of 10^4 solutions of problem (C.6) with different starting points.

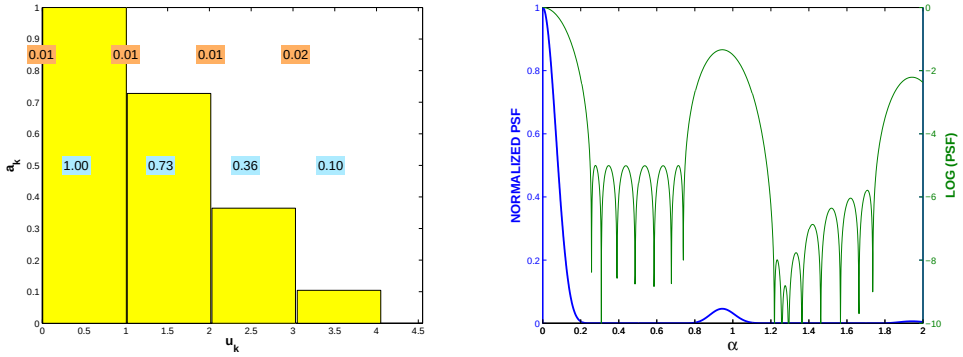


FIG. C.4 – Best solution found amongst 10^4 occurrences of problem (C.6) when $[\alpha_{\min}, \alpha_{\max}] = [0.25, 0.75]$. On the left figure, each rectangle represents the modulus of a pupil, the numbers on top are the distance between two pupils, the numbers in the middle are the coefficients a_k. The right figure shows the graph of the normalized PSF in linear and logarithmic scale.

Table C.1 reports the optimal values of a_k and u_k. Note that for the presentation of the numerical values and since the normalized PSF is homogeneous to a_k, the values of a_k are renormalized such that the greatest component (generally a_1) is equal to one. Note also that the accuracy of the optimal values a_k and u_k depends on the convergence tolerance of the optimization solver. In our experiments, we used a convergence tolerance of 10^{-10} (`tol` value in IPOPT package). This allows to obtain an absolute precison of at least 10^{-8} on the optimal values of a_k and u_k.

It is important to note that the optimal positions of the pupils are nearly periodic. By analysing some other similar experiments, we always observed that the best solution shows a quasi-periodic positioning of the pupils. We found that the initial positions defined by

$$u_k = (k - \frac{1}{2})\frac{d}{\alpha_{\min} + \alpha_{\max}}, \quad k = 1, \ldots, m \tag{C.7}$$

give a fairly good estimate of the optimal positions.

k	a_k	u_k	$u_k - u_{k-1}$
1	1.0000	0.5054	1.0108
2	0.7277	1.5167	1.0113
3	0.3644	2.5299	1.0132
4	0.1044	3.5510	1.0211

TAB. C.1 – Optimal values for the best solution. The value u_0 is set to the position of the pupil symmetric to the first one with respect to zero.

Our starting point strategy is as follows. Problem (C.6) is solved with the variables u_k fixed to the values defined by (C.7). Note that in that case, (C.6) is reduced to a linear programming problem and thus the global optimality of the computed solution can be guaranteed. Then, starting from these optimal values, the solver is rerun to solve (C.6) with respect to the whole set of variables. This strategy works fine in practice. On one hand, IPOPT is able to compute the optimal solution in some tens of iterations. On the other hand, the computed solution seems to be the global minimum, though we are not able to prove it, except in some trivial cases with two or three pupils.

C.6 Numerical experiments

We first analyse the effect of the choice of the CLF interval on an optimal configuration. Let us define a third measure of the quality of a given configuration. Let $\rho \in [0, 1]$ be such that

$$\Psi(\rho) = \min\{\alpha \in [0, 1] : \Psi(\alpha) = \frac{1}{2}\}.$$

This is the width of the PSF curve at half height of the main central lobe. The *number of resel* is the ratio

$$R = \frac{\alpha_{\max} - \alpha_{\min}}{\rho}.$$

This is the number of resolved elements in the clean field of view interval.

Let us observe the variations of the three parameters, dynamic (D), number of resels (R) and central flux (F) when the CLF interval varies. Figure C.5 (left) shows the variation of these measures for different values of the width $\Delta\alpha := \alpha_{\max} - \alpha_{\min}$. The experiments correspond to a configuration with $d = 1$ and $n = 8$. The numerical values are reported in Table C.2 (a).

We can observe that the number of resels increases with $\Delta\alpha$, while the dynamic goes to smaller values. The variation of the central flux is not really significant. Figures C.6 and C.7 show the solution of the two extreme values in Table C.2 (a).

Figure C.5 (right) shows the variation of D, R and F for a fixed width ($\Delta\alpha = 0.4$), but for different values of $\alpha_{\min}$. The corresponding numerical values are reported in Table C.2 (b). Figures C.8 and C.9 show the resulting optimal solution of the two extreme values in Table C.2 (b). We can observe that good values of dynamic and

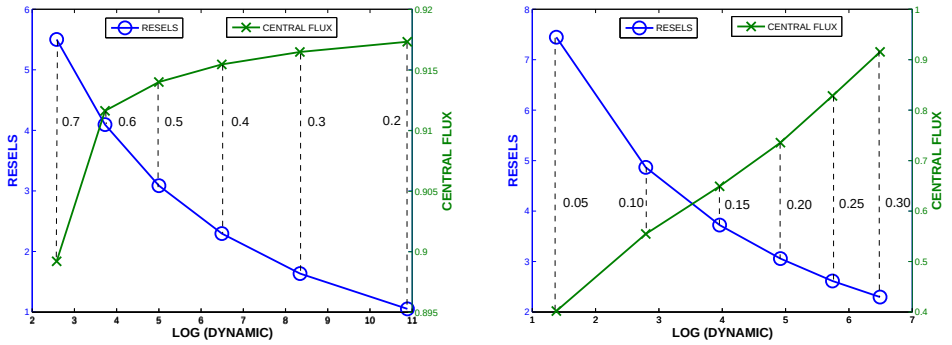


FIG. C.5 – Values of dynamic, central flux and number of resels according to a variation of $\Delta\alpha$ (on left), of $\alpha_{\min}$ but with $\Delta\alpha = 0.4$ (on right).

$\alpha_{\min}$	$\alpha_{\max}$	$\Delta\alpha$	D	R	F
0.40	0.60	0.2	7.7e10	1.1	0.917
0.35	0.65	0.3	2.2e8	1.6	0.916
0.30	0.70	0.4	3.1e6	2.3	0.915
0.25	0.75	0.5	1.0e5	3.1	0.914
0.20	0.80	0.6	5.3e3	4.1	0.912
0.15	0.85	0.7	3.8e2	5.5	0.899

a

$\alpha_{\min}$	$\alpha_{\max}$	D	R	F
0.05	0.45	2.4e1	7.4	0.402
0.10	0.50	6.2e2	4.9	0.555
0.15	0.55	9.1e3	3.7	0.649
0.20	0.60	8.3e4	3.1	0.736
0.25	0.65	5.5e5	2.6	0.828
0.30	0.70	3.1e6	2.3	0.915

b

TAB. C.2 – Values of dynamic, number of resels and central flux according to the variation of $\Delta\alpha$ (a) and $\alpha_{\min}$ with $\Delta\alpha = 0.4$ (b).

central flux are obtained when $\alpha_{\min}$ is far away from zero. On the contrary, the number of resels increases when $\alpha_{\min}$ goes to zero, which follows from the fact that the main central lobe becomes narrow.

We study the effect of the number of pupils on an optimal configuration. We first choose a given CLF interval. Problem (C.6) is solved with $d = 1$, $\alpha_{\min} = 0.25$ and $\alpha_{\max} = 0.75$ and for different number of pupils ($n = 4$, 8, 16 and 24). The results are reported in Table C.3 (a). We can see that the dynamic can go to very large values when the number of pupils increases. The number of resels increases slowly because the CLF interval is constant and the main central lobe becomes slightly more narrow. The values reported in Table C.3 (b) are obtained by choosing a CLF interval such that the value of dynamic is about 10^{-6}. We can observe that both the number of resels and the central flux increase. We can also observe that the number of resels is about half the number of pupils. Figures C.10 and C.11 show the PSF for 24 pupils.

We end this section by studying the sensitivity of an optimal configuration. We would like to show how some small perturbations on the optimal values of u_k and/or a_k have some effect on the measures of quality D, R and F. Starting form the optimal configuration with 8 pupils and a CLF interval equals to $[0.25, 0.75]$ (see Table C.1, Figure C.4 and Table C.2 (a) for optimal values), we computed the PSF function and the corresponding three measures D, R and F for perturbed values a'_k and u'_k

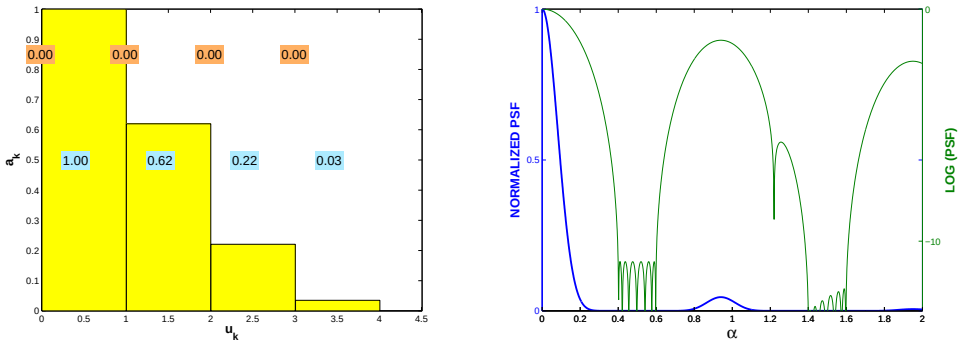


FIG. C.6 – Optimal solution for $[\alpha_{\min}, \alpha_{\max}] = [0.4, 0.6]$.

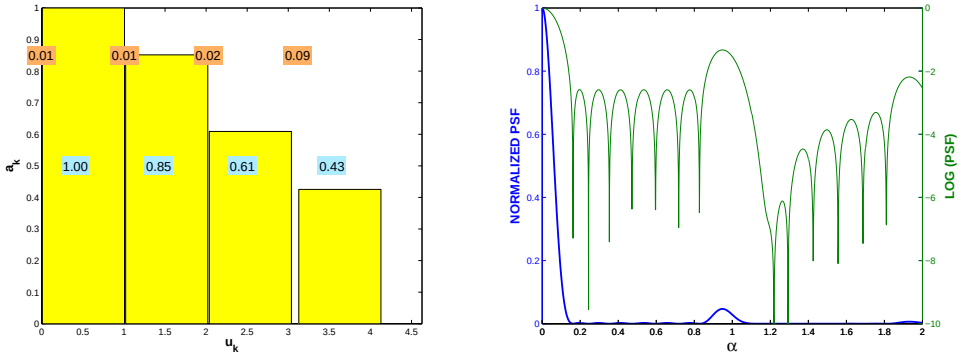


FIG. C.7 – Optimal solution for $[\alpha_{\min}, \alpha_{\max}] = [0.15, 0.85]$.

such that for $k = 1, \ldots, m$

$$a'_k = a_k(1 + t_k) \quad \text{and} \quad u'_k - u'_{k-1} = \max\{1, (u_k - u_{k-1})(1 + t'_k)\},$$

where t_k and t'_k are some random numbers following a uniform distribution on $[-\tau, \tau]$. The same perturbation was applied to the second half of pupils symmetric to zero. We performed a hundred of experiments with $\tau = 10^{-3}$ and with $\tau = 10^{-2}$. To compare the solutions, we considered only the worse case, that is the configurations which return the smallest value for the three measures D, R and F. The results are reported in Table C.4. For each value of τ, three kind of pertubation is done : beam moduli only, positions only, both moduli and positions. We observe first that only the dynamic (D) is sensitive to perturbation, the number of resels (R) and the central flux (F) are not changed or very slightly. We note also that the dynamic is more sensitive to a perturbation of positions than of beam moduli. A perturbation of 1% of the position implies a decrease of a factor 10 in dynamic. Figure C.12 shows the normalized PSF with perturbed optimal parameters.

C.7 Conclusion

Hypertelescopes appear to be good candidates for the next generation of high resolution and high dynamic imaging device for astronomy. The optimization of

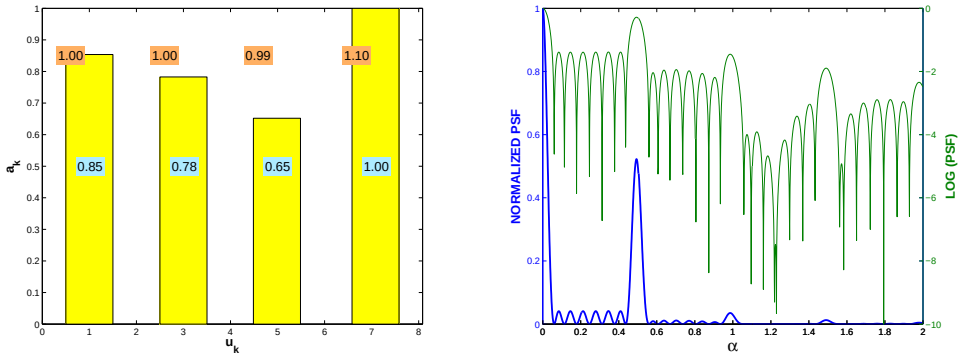


FIG. C.8 – Optimal solution for $[\alpha_{\min}, \alpha_{\max}] = [0.05, 0.45]$.

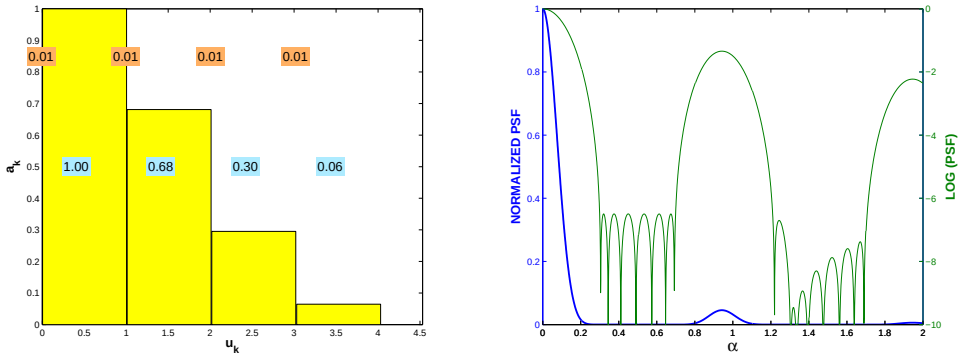


FIG. C.9 – Optimal solution for $[\alpha_{\min}, \alpha_{\max}] = [0.3, 0.7]$.

the point spread function is a crucial step to maximize the efficiency of these imaging systems. This paper has demonstrated the potential of nonlinear optimization technology to adjust the densified pupil configuration in order to address specific observation program (i.e. resolution, dynamic). As preliminary approach, this work has been done in the context of a linear array. We plane to extend this study to a two dimensional array and to generalize this method to the different configurations to be used with a hypertelescope.

Acknowledgements

This work is supported by the Centre National d'Étude Spatiale (CNES).

n	D	R	F
4	$9.1e1$	2.1	0.8728
8	$1.0e5$	3.1	0.9140
16	$1.3e11$	4.4	0.9241
24	$5.9e16$	5.4	0.9286

a : $[\alpha_{\min}, \alpha_{\max}] = [0.25, 0.75]$

n	D	R	F
4	$5.9e5$	0.4	0.5623
8	$9.1e5$	2.5	0.8734
16	$1.0e6$	7.7	0.7317
24	$5.9e5$	13.5	0.8562

b : $D \simeq 10^6$

TAB. C.3 – Values of dynamic, number of resels and central flux according to a variation of the number of pupils.

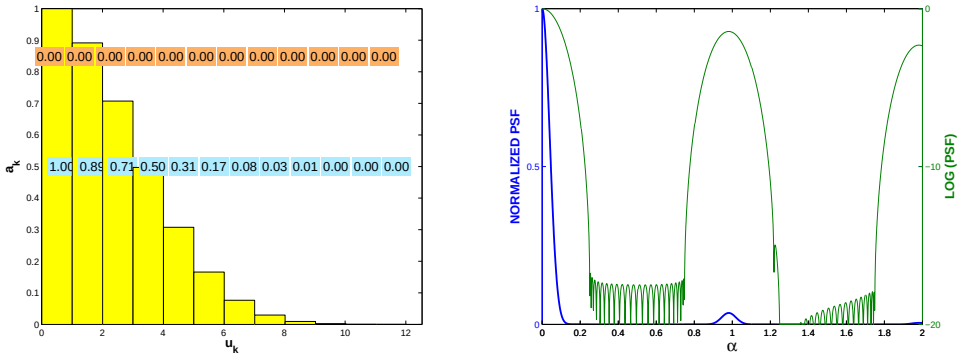


FIG. C.10 – Optimal solution for 24 pupils and $[\alpha_{\min}, \alpha_{\max}] = [0.25, 0.75]$.

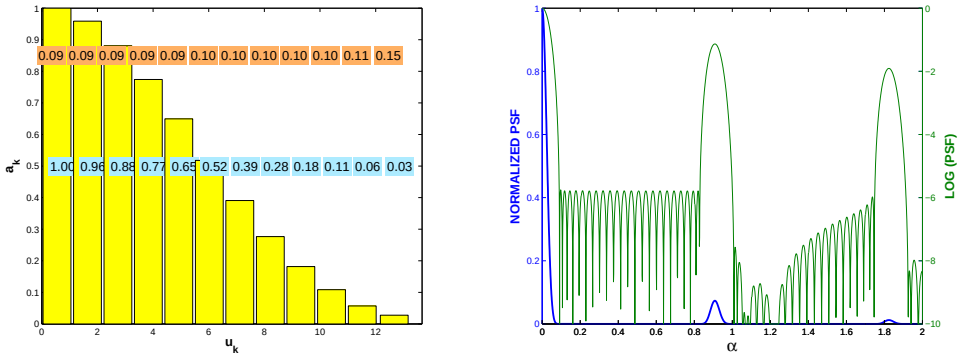


FIG. C.11 – Optimal solution for 24 pupils and $[\alpha_{\min}, \alpha_{\max}] = [0.09, 0.83]$.

τ	perturbation	D	R	F
10^{-3}	moduli	$8.0e4$	3.085	0.914
	positions	$7.0e4$	3.084	0.914
	both	$6.6e4$	3.083	0.913
10^{-2}	moduli	$2.2e4$	3.074	0.914
	positions	$8.9e3$	3.069	0.908
	both	$9.0e3$	3.068	0.908

TAB. C.4 – Values of dynamic, number of resels and central flux according to a perturbation of a_k and/or u_k with 8 pupils and CLF $= [0.25, 0.75]$.

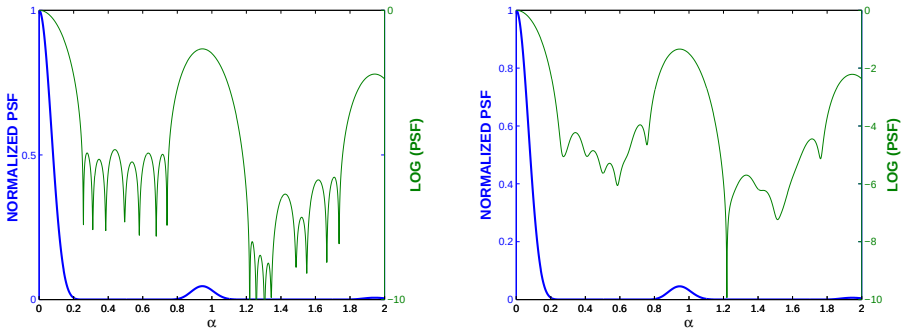


FIG. C.12 – Normalized PSF with perturbed optimal parameters ($\tau = 10^{-3}$ on left and $\tau = 10^{-2}$ on right). Compare with the optimal configuration shows in Figure C.4.

Annexe D

Résultats numériques obtenus avec la méthode primale-duale

Dans cette annexe, nous présentons les résultats numériques obtenus avec l'algorithme E testé sur l'ensemble des problèmes du tableau 3.1. Nous avons considéré un modèle dynamique de mise à jour du paramètre de pénalité et nous avons en plus adaptée la décroissance du paramètre à la décroissance de la norme des conditions d'optimalité perturbées du problème non linéaire (1.1) ou à la décroissance de la norme du lagrangien (1.3). Les tableaux suivants énumèrent ces résultats où

- nf représente le nombre d'évaluations de fonctions f et c,
- ng représente le nombre d'évaluations de gradients ∇f,
- nh représente le nombre d'évaluations de hessiens $\nabla^2 f$,
- nfact représente le nombre de factorisations de la matrice $F'_w(w, \mu)$,
- ffin représente la valeur finale de f au dernier point calculé par l'algorithme,
- $\|F\|$ représente la norme des conditions d'optimalité non perturbées en w, i.e. $\|F(w, 0)\|$,
- t représente le temps CPU (en secondes) nécessaire à l'arrêt de l'algorithme,
- Info représente le mode d'arrêt de l'algorithme
  - Info = 0 lorsque l'algorithme s'arrête normalement,
  - Info = 1 lorsque le nombre maximal d'évaluations du nombre de fonctions et de contraintes est atteint,
  - Info = 2 lorsque le nombre maximal de factorisations de $F'_w(w, \mu)$ est atteint,
  - Info = 3 lorsque le nombre maximal de corrections de $F'_w(w, \mu)$ est atteint.

Nom du problème	nf	ng	nh	nfact	ffin	$\|\|F\|\|$	t	Info
AUG2D	4	4	3	3	$1.1e+02$	$4.4e-16$	0.0	0
AUG3DC	3	3	2	2	$7.7e+02$	$1.3e-15$	1.9	0
AUG3D	4	4	3	6	$5.5e+02$	$5.3e-11$	6.4	0
BT1	300	22	21	44	$-1.0e+00$	$5.3e-11$	0.2	0
BT2	13	13	12	12	$3.3e-02$	$4.1e-12$	0.0	0
BT3	3	3	2	2	$4.1e+00$	$4.7e-15$	0.0	0
BT4	16	12	11	17	$-4.6e+01$	$2.1e-15$	0.1	0
BT5	9	9	8	8	$9.6e+02$	$6.0e-12$	0.0	0
BT6	14	14	13	13	$2.8e-01$	$1.1e-12$	0.1	0
BT7	39	32	31	41	$3.1e+02$	$8.9e-16$	0.1	0
BT8	38	38	37	93	$1.0e-00$	$4.0e-09$	0.1	0
BT9	15	15	14	17	$-1.0e+00$	$1.0e-12$	0.1	0
BT11	10	10	9	9	$8.2e-01$	$9.3e-14$	0.0	0
BT12	5	5	4	4	$6.2e+00$	$2.3e-13$	0.0	0
BYRDSPHR	19	19	18	26	$-4.7e+00$	$1.7e-15$	0.1	0
CATENA	34	33	32	32	$-2.3e+04$	$2.2e-12$	0.1	0
CHAIN1	110	52	51	109	$5.1e+00$	$9.9e-11$	3.6	0
CHAIN2	69	40	39	79	$5.1e+00$	$4.7e-09$	3.6	0
CHAIN3	215	64	63	112	$5.1e+00$	$3.5e-11$	6.9	0
CHNRSBNE	59	27	26	50	$0.0e+00$	$1.5e-12$	0.1	0
DECONVNE	3	3	2	4	$0.0e+00$	$1.1e-09$	0.1	0
DIXCHLNG	31	27	26	45	$2.8e-18$	$2.0e-10$	0.1	0
DTOC1NA	7	7	6	6	$1.3e+01$	$3.7e-11$	2.1	0
DTOC1NB	7	7	6	6	$1.6e+01$	$3.3e-13$	2.1	0
DTOC1NC	674	98	97	160	$2.5e+01$	$5.6e-11$	50.0	0
DTOC1ND	1000	227	227	572	$1.3e+01$	$6.1e-02$	60.1	1
DTOC2	27	18	17	28	$5.0e-01$	$6.4e-13$	14.7	0
DTOC5	6	6	5	5	$1.5e+00$	$3.4e-13$	13.3	0
DTOC6	23	23	22	22	$1.3e+05$	$1.8e-12$	19.8	0
ELEC1	55	55	54	126	$6.5e+03$	$1.1e-09$	40.1	0
ELEC2	208	103	102	225	$1.0e+04$	$8.5e-14$	125.4	0
ELEC3	153	105	104	243	$1.8e+04$	$5.0e-09$	231.1	0
EIGENA2	17	16	15	28	$7.1e-28$	$5.3e-13$	0.4	0
EIGENACO	17	17	16	31	$3.4e-28$	$5.0e-13$	0.8	0
EIGENCCO	13	13	12	23	$1.3e-25$	$5.8e-13$	0.1	0
EIGENBCO	1000	195	195	493	$2.2e-01$	$6.4e-07$	18.5	1
EIGENB2	1000	201	201	509	$4.5e-01$	$2.7e-07$	9.8	1
EIGENC2	314	207	206	511	$1.6e-20$	$1.3e-10$	105.5	0
GENHS28	3	3	2	2	$9.3e-01$	$2.2e-16$	0.0	0
GILBERT	21	21	20	23	$4.8e+02$	$5.4e-13$	0.9	0

Nom du problème	nf	ng	nh	nfact	ffin	$\|\|F\|\|$	t	Info
GRIDNETE	6	6	5	5	$2.1e+02$	$2.7e-11$	9.9	0
GRIDNETB	34	6	5	5	$3.8e+01$	$5.3e-16$	0.0	0
GRIDNETH	6	6	5	5	$4.0e+01$	$3.8e-11$	0.0	0
HAGER1	9	7	6	6	$8.8e-01$	$9.1e-13$	13.2	0
HAGER2	5	4	3	3	$4.3e-01$	$1.6e-12$	26.9	0
HAGER3	9	6	5	5	$1.4e-01$	$1.5e-12$	14.2	0
HS006	8	7	6	11	$0.0e+00$	$1.8e-15$	0.0	0
HS007	31	14	13	26	$-1.7e+00$	$2.8e-12$	0.1	0
HS009	7	6	5	11	$-5.0e-01$	$5.4e-12$	0.0	0
HS026	26	26	25	25	$1.4e-16$	$6.6e-09$	0.1	0
HS027	60	32	31	43	$4.0e-02$	$1.6e-12$	0.1	0
HS028	2	2	1	1	$4.9e-32$	$1.0e-15$	0.0	0
HS039	15	15	14	17	$-1.0e+00$	$1.0e-12$	0.0	0
HS040	5	5	4	4	$-2.5e-01$	$5.0e-16$	0.0	0
HS042	6	6	5	5	$1.4e+01$	$6.2e-15$	0.0	0
HS046	24	24	23	26	$6.4e-16$	$6.3e-09$	0.1	0
HS047	45	32	31	42	$9.3e-14$	$6.1e-09$	0.1	0
HS048	2	2	1	1	$2.0e-30$	$3.4e-15$	0.0	0
HS049	21	21	20	20	$2.1e-12$	$7.0e-09$	0.3	0
HS050	10	10	9	9	$5.9e-23$	$9.4e-12$	0.0	0
HS051	2	2	1	1	$6.2e-32$	$5.8e-16$	0.0	0
HS052	3	3	2	2	$5.3e+00$	$6.1e-15$	0.0	0
HS056	37	15	14	24	$-3.5e+00$	$4.3e-11$	0.1	0
HS061	16	16	15	19	$-1.4e+02$	$7.1e-15$	0.0	0
HS077	12	12	11	11	$2.4e-01$	$2.4e-09$	0.1	0
HS078	5	5	4	4	$-2.9e+00$	$7.7e-10$	0.0	0
HS079	5	5	4	4	$7.9e-02$	$3.5e-09$	0.0	0
HS100LNP	9	9	8	14	$6.8e+02$	$4.1e-10$	0.0	0
HS111LNP	17	17	16	25	$-4.8e+01$	$5.2e-12$	0.1	0
LUKVLE1	18	16	15	22	$6.2e+00$	$4.3e-13$	0.2	0
LUKVLE3	15	15	14	18	$2.8e+01$	$3.6e-14$	0.1	0
LUKVLE6	17	17	16	16	$6.0e+03$	$7.2e-10$	0.2	0
LUKVLE7	15	14	13	22	$-2.6e+01$	$2.1e-14$	0.1	0
LUKVLE9	30	21	20	30	$1.0e+01$	$4.3e-12$	0.1	0
LUKVLE11	26	23	22	31	$3.0e-11$	$8.1e-09$	0.2	0
LUKVLE12	115	69	68	108	$1.2e+03$	$2.4e-10$	0.7	0
LUKVLE13	39	24	23	54	$5.3e+02$	$9.8e-09$	0.7	0
LUKVLE14	58	42	41	46	$2.6e+06$	$8.4e-10$	0.3	0
LUKVLE15	151	57	56	91	$1.7e-11$	$8.4e-09$	0.5	0
LUKVLE16	20	20	19	23	$8.1e-12$	$5.7e-09$	0.1	0
LUKVLE17	1000	998	998	1002	$3.2e+02$	$1.4e-11$	6.0	1
LUKVLE18	1000	979	979	2427	$1.1e+02$	$1.3e-10$	21.1	1

Nom du problème	nf	ng	nh	nfact	ffin	$\|\|F\|\|$	t	Info
MARATOS	5	5	4	4	$-1.0e+00$	$6.2e-15$	0.0	0
MWRIGHT	18	18	17	28	$1.3e+00$	$5.9e-15$	0.1	0
ORTHREGA	115	74	73	158	$1.4e+03$	$2.8e-11$	7.5	0
ORTHREGB	3	3	2	4	$4.5e-20$	$2.1e-10$	0.0	0
ORTHREGC	10	10	9	13	$4.0e+00$	$5.0e-11$	0.2	0
ORTHRGDM	17	16	15	28	$7.1e+00$	$2.5e-09$	0.1	0
ORTHREGD	1000	99	99	127	$1.0e+02$	$1.2e-10$	2.0	1
ORTHRDM2	10	10	9	14	$3.1e+02$	$9.1e-10$	23.7	0
ORTHRGDS	1000	169	169	199	$1.3e+02$	$2.3e-10$	2.4	1
S216	19	15	14	14	$1.0e-00$	$1.6e-13$	0.0	0
S219	21	21	20	23	$-1.0e+00$	$7.7e-10$	0.1	0
S235	192	43	42	42	$4.0e-02$	$4.8e-16$	0.2	0
S252	16	13	12	17	$4.0e-02$	$4.1e-12$	0.0	0
S269	3	3	2	2	$4.1e+00$	$1.8e-15$	0.0	0
S316	15	15	14	14	$3.3e+02$	$5.5e-11$	0.0	0
S317	17	17	16	16	$3.7e+02$	$2.6e-10$	0.1	0
S318	17	17	16	16	$4.1e+02$	$7.1e-15$	0.0	0
S319	15	15	14	14	$4.5e+02$	$2.8e-14$	0.0	0
S320	18	18	17	17	$4.9e+02$	$7.1e-15$	0.1	0
S321	19	19	18	18	$5.0e+02$	$1.5e-12$	0.1	0
S322	24	24	23	23	$5.0e+02$	$2.7e-12$	0.1	0
S335	26	26	25	25	$-4.5e-03$	$1.5e-10$	0.1	0
S336	33	33	32	40	$-3.4e-01$	$2.1e-14$	0.1	0
S338	717	112	111	197	$-1.1e+01$	$4.4e-16$	0.5	0
S344	8	8	7	7	$3.3e-02$	$1.1e-11$	0.0	0
S345	14	14	13	14	$3.3e-02$	$1.0e-12$	0.1	0
S375	630	98	97	162	$-1.5e+01$	$1.8e-15$	0.5	0
S378	17	17	16	25	$-4.8e+01$	$5.2e-12$	0.1	0
S394	13	13	12	17	$1.9e+00$	$9.1e-10$	0.0	0
S395	15	15	14	20	$1.9e+00$	$2.8e-14$	0.4	0
SPMSQRT	7	7	6	6	$0.0e+00$	$3.0e-09$	14.1	0

TAB. D.1 – Résulats numériques obtenus avec l'algorithme E en considérant la définition (3.2) pour la valeur de μ_0 ainsi que les expressions (3.3) et (3.4) pour le choix dynamique de μ. La décroissance de μ est en plus adaptée à la décroissance de $\|F(w,\mu)\|$. Dans la boucle qui permet d'adapter les deux décroissances, la valeur de F n'est pas recalculée chaque fois que la valeur de μ est modifiée.

Nom du problème	nf	ng	nh	nfact	ffin	$\|\|F\|\|$	t	Info
AUG2D	5	5	4	4	$1.1e+02$	$2.0e-11$	0.1	0
AUG3DC	5	5	4	4	$7.7e+02$	$4.7e-15$	2.5	0
AUG3D	5	5	4	8	$5.5e+02$	$1.7e-11$	7.8	0
BT1	1000	24	24	24	$-1.0e+02$	$8.9e-01$	0.4	1
BT2	13	13	12	12	$3.3e-02$	$4.1e-12$	0.1	0
BT3	6	6	5	5	$4.1e+00$	$1.3e-15$	0.0	0
BT4	19	13	12	18	$-4.6e+01$	$1.8e-15$	0.1	0
BT5	9	9	8	8	$9.6e+02$	$7.0e-13$	0.0	0
BT6	14	14	13	13	$2.8e-01$	$1.1e-12$	0.1	0
BT7	39	32	31	41	$3.1e+02$	$1.5e-10$	0.1	0
BT8	15	15	14	14	$1.0e+00$	$4.0e-09$	0.1	0
BT9	14	14	13	16	$-1.0e+00$	$1.6e-09$	0.1	0
BT11	10	10	9	9	$8.2e-01$	$1.1e-13$	0.0	0
BT12	6	6	5	5	$6.2e+00$	$1.1e-13$	0.0	0
BYRDSPHR	19	19	18	26	$-4.7e+00$	$1.7e-15$	0.1	0
CATENA	33	32	31	31	$-2.3e+04$	$4.7e-12$	0.1	0
CHAIN1	201	60	59	107	$5.1e+00$	$1.2e-09$	3.8	0
CHAIN2	61	38	37	74	$5.1e+00$	$2.7e-10$	3.6	0
CHAIN3	598	120	119	253	$5.1e+00$	$2.8e-11$	15.4	0
CHNRSBNE	59	27	26	50	$0.0e+00$	$1.5e-12$	0.1	0
DECONVNE	3	3	2	4	$0.0e+00$	$1.1e-09$	0.1	0
DIXCHLNG	31	27	26	45	$2.8e-18$	$2.0e-10$	0.1	0
DTOC1NA	7	7	6	6	$1.3e+01$	$3.4e-11$	2.1	0
DTOC1NB	7	7	6	6	$1.6e+01$	$5.9e-13$	2.1	0
DTOC1NC	14	14	13	20	$2.5e+01$	$6.7e-10$	5.0	0
DTOC1ND	505	142	141	296	$1.2e+01$	$4.7e-13$	32.1	0
DTOC2	27	18	17	28	$5.0e-01$	$7.2e-13$	14.9	0
DTOC5	6	6	5	5	$1.5e+00$	$9.8e-11$	33.3	0
DTOC6	24	24	23	23	$1.3e+05$	$8.0e-11$	24.2	0
ELEC1	55	55	54	126	$6.5e+03$	$4.4e-09$	41.3	0
ELEC2	208	103	102	225	$1.0e+04$	$8.5e-14$	125.0	0
ELEC3	153	105	104	243	$1.8e+04$	$5.0e-09$	229.4	0
EIGENA2	17	16	15	28	$7.1e-28$	$5.3e-13$	0.4	0
EIGENACO	17	17	16	31	$3.4e-28$	$5.0e-13$	1.2	0
EIGENCCO	13	13	12	23	$1.3e-25$	$5.8e-13$	0.1	0
EIGENBCO	1000	195	195	493	$2.2e-01$	$6.4e-07$	18.9	1
EIGENB2	1000	202	202	511	$4.5e-01$	$1.4e-07$	9.5	1
EIGENC2	314	207	206	511	$1.6e-20$	$1.3e-10$	105.9	0
GENHS28	4	4	3	3	$9.3e-01$	$1.2e-10$	0.0	0
GILBERT	21	21	20	23	$4.8e+02$	$1.6e-12$	1.2	0

Nom du problème	nf	ng	nh	nfact	ffin	$\|\|F\|\|$	t	Info
GRIDNETE	9	9	8	8	$2.1e+02$	$4.9e-12$	11.4	0
GRIDNETB	6	6	5	5	$3.8e+01$	$1.8e-10$	0.0	0
GRIDNETH	6	6	5	5	$4.0e+01$	$1.1e-10$	0.0	0
HAGER1	3	3	2	2	$8.8e-01$	$2.9e-12$	15.2	0
HAGER2	3	3	2	2	$4.3e-01$	$4.2e-10$	21.8	0
HAGER3	3	3	2	2	$1.4e-01$	$5.4e-11$	12.4	0
HS006	10	7	6	13	$6.0e-31$	$1.8e-15$	0.1	0
HS007	31	14	13	26	$-1.7e+00$	$4.0e-12$	0.1	0
HS009	7	6	5	11	$-5.0e-01$	$5.4e-12$	0.0	0
HS026	26	26	25	25	$1.3e-16$	$6.4e-09$	0.1	0
HS027	10	10	9	14	$4.0e-02$	$7.5e-10$	0.1	0
HS028	2	2	1	1	$4.9e-32$	$1.0e-15$	0.1	0
HS039	14	14	13	16	$-1.0e+00$	$1.6e-09$	0.1	0
HS040	5	5	4	4	$-2.5e-01$	$2.1e-15$	0.1	0
HS042	6	6	5	5	$1.4e+01$	$2.1e-14$	0.0	0
HS046	24	24	23	26	$6.4e-16$	$6.3e-09$	0.1	0
HS047	32	30	29	39	$2.5e-14$	$2.6e-09$	0.1	0
HS048	2	2	1	1	$2.0e-30$	$3.4e-15$	0.0	0
HS049	21	21	20	20	$2.1e-12$	$7.0e-09$	0.1	0
HS050	10	10	9	9	$5.9e-23$	$9.4e-12$	0.0	0
HS051	2	2	1	1	$6.2e-32$	$5.8e-16$	0.0	0
HS052	6	6	5	5	$5.3e+00$	$7.7e-15$	0.0	0
HS056	34	15	14	24	$-3.5e+00$	$1.3e-15$	0.1	0
HS061	16	16	15	19	$-1.4e+02$	$2.7e-14$	0.1	0
HS077	12	12	11	11	$2.4e-01$	$2.5e-09$	0.2	0
HS078	5	5	4	4	$-2.9e+00$	$7.7e-10$	0.0	0
HS079	5	5	4	4	$7.9e-02$	$3.2e-09$	0.0	0
HS100LNP	15	12	11	19	$6.8e+02$	$4.0e-12$	0.1	0
HS111LNP	17	17	16	25	$-4.8e+01$	$5.1e-12$	0.1	0
LUKVLE1	18	16	15	22	$6.2e+00$	$1.2e-12$	0.2	0
LUKVLE3	15	15	14	14	$2.8e+01$	$2.1e-14$	0.1	0
LUKVLE6	17	17	16	16	$6.0e+03$	$7.2e-10$	0.2	0
LUKVLE7	21	18	17	26	$-2.6e+01$	$1.7e-11$	0.1	0
LUKVLE9	76	39	38	70	$1.1e+01$	$4.0e-14$	0.3	0
LUKVLE11	26	23	22	31	$3.0e-11$	$8.1e-09$	0.2	0
LUKVLE12	75	60	59	100	$1.2e+03$	$3.6e-09$	0.6	0
LUKVLE13	23	23	22	51	$5.3e+02$	$3.3e-09$	0.7	0
LUKVLE14	41	33	32	37	$2.6e+06$	$3.0e-09$	0.2	0
LUKVLE15	151	57	56	91	$1.7e-11$	$8.4e-09$	0.6	0
LUKVLE16	20	20	19	23	$8.1e-12$	$5.7e-09$	0.3	0
LUKVLE17	63	53	52	56	$3.2e+02$	$5.7e-09$	0.4	0
LUKVLE18	1000	965	965	2390	$1.1e+02$	$1.4e-10$	29.3	1

Nom du problème	nf	ng	nh	nfact	ffin	$\|\|F\|\|$	t	Info
MARATOS	5	5	4	4	$-1.0e+00$	$5.7e-15$	0.1	0
MWRIGHT	18	18	17	28	$1.3e+00$	$5.6e-15$	0.1	0
ORTHREGA	121	77	76	162	$1.4e+03$	$1.0e-11$	7.7	0
ORTHREGB	3	3	2	4	$4.5e-20$	$2.1e-10$	0.0	0
ORTHREGC	10	10	9	13	$4.0e+00$	$5.0e-11$	0.2	0
ORTHRGDM	17	16	15	28	$7.1e+00$	$2.5e-09$	0.1	0
ORTHREGD	1000	72	72	100	$1.0e+02$	$1.4e-04$	1.7	1
ORTHRDM2	10	10	9	14	$3.1e+02$	$9.1e-10$	23.8	0
ORTHRGDS	1000	118	118	128	$1.2e+02$	$4.1e-08$	1.7	1
S216	19	15	14	14	$1.0e-00$	$1.6e-13$	0.0	0
S219	21	21	20	23	$-1.0e+00$	$3.7e-10$	0.1	0
S235	192	43	42	42	$4.0e-02$	$4.2e-16$	0.2	0
S252	16	13	12	17	$4.0e-02$	$3.3e-12$	0.0	0
S269	6	6	5	5	$4.1e+00$	$5.9e-12$	0.0	0
S316	15	15	14	14	$3.3e+02$	$3.6e-15$	0.0	0
S317	17	17	16	16	$3.7e+02$	$1.0e-11$	0.0	0
S318	16	16	15	15	$4.1e+02$	$1.5e-12$	0.0	0
S319	15	15	14	14	$4.5e+02$	$5.8e-10$	0.0	0
S320	18	18	17	17	$4.9e+02$	$2.2e-12$	0.1	0
S321	19	19	18	18	$5.0e+02$	$6.4e-14$	0.1	0
S322	24	24	23	23	$5.0e+02$	$1.4e-14$	0.1	0
S335	26	26	25	25	$-4.5e-03$	$1.5e-10$	0.1	0
S336	33	33	32	41	$-3.4e-01$	$4.0e-12$	0.1	0
S338	717	112	111	197	$-1.1e+01$	$8.9e-16$	0.5	0
S344	8	8	7	7	$3.3e-02$	$1.1e-11$	0.0	0
S345	14	14	13	14	$3.3e-02$	$6.1e-13$	0.0	0
S375	630	98	97	162	$-1.5e+01$	$1.8e-15$	0.5	0
S378	17	17	16	25	$-4.8e+01$	$5.1e-12$	0.1	0
S394	13	13	12	17	$1.9e+00$	$9.1e-10$	0.0	0
S395	15	15	14	20	$1.9e+00$	$2.9e-14$	0.1	0
SPMSQRT	7	7	6	6	$0.0e+00$	$3.0e-09$	13.4	0

TAB. D.2 – Résulats numériques obtenus avec l'algorithme E en considérant la définition (3.2) pour la valeur de μ_0 ainsi que les expressions (3.3) et (3.4) pour le choix dynamique de μ. La décroissance de μ est en plus adaptée à la décroissance de $\|F(w,\mu)\|$. Dans la boucle qui permet d'adapter les deux décroissances, la valeur de F est recalculée chaque fois que la valeur de μ est modifiée.

Nom du problème	nf	ng	nh	nfact	ffin	$\|\|F\|\|$	t	Info
AUG2D	11	11	10	10	$1.1e+02$	$5.6e-16$	0.1	0
AUG3DC	10	10	9	9	$7.7e+02$	$1.2e-15$	3.0	0
AUG3D	11	11	10	20	$5.5e+02$	$1.4e-15$	16.3	0
BT1	22	22	21	32	$-1.0e+00$	$2.8e-14$	0.1	0
BT2	16	16	15	15	$3.3e-02$	$2.1e-15$	0.1	0
BT3	11	11	10	10	$4.1e+00$	$4.4e-16$	0.1	0
BT4	23	17	16	22	$-4.6e+01$	$7.1e-15$	0.1	0
BT5	13	13	12	12	$9.6e+02$	$4.3e-15$	0.1	0
BT6	15	15	14	14	$2.8e-01$	$3.7e-15$	0.1	0
BT7	1000	35	35	75	$1.0e+00$	$3.0e-01$	0.5	1
BT8	38	38	37	93	$1.0e-00$	$4.9e-09$	0.1	0
BT9	15	15	14	17	$-1.0e+00$	$1.0e-14$	0.1	0
BT11	13	13	12	12	$8.2e-01$	$7.1e-15$	0.1	0
BT12	10	10	9	9	$6.2e+00$	$1.1e-13$	0.0	0
BYRDSPHR	23	23	22	29	$-4.7e+00$	$2.9e-15$	0.1	0
CATENA	34	33	32	32	$-2.3e+04$	$1.8e-10$	0.1	0
CHAIN1	64	42	41	78	$5.1e+00$	$7.3e-11$	2.6	0
CHAIN2	53	36	35	63	$5.1e+00$	$7.9e-12$	2.9	0
CHAIN3	168	68	67	117	$5.1e+00$	$6.9e-11$	6.9	0
CHNRSBNE	37	21	20	37	$0.0e+00$	$9.5e-16$	0.1	0
DECONVNE	3	3	2	5	$0.0e+00$	$2.7e-10$	0.1	0
DIXCHLNG	31	27	26	45	$1.0e-23$	$2.0e-10$	0.1	0
DTOC1NA	10	10	9	9	$1.3e+01$	$9.7e-15$	3.1	0
DTOC1NB	10	10	9	9	$1.6e+01$	$1.2e-14$	3.0	0
DTOC1NC	14	14	13	17	$2.5e+01$	$1.4e-14$	4.5	0
DTOC1ND	1000	190	190	471	$1.3e+01$	$5.4e-02$	51.1	1
DTOC2	24	21	20	33	$5.0e-01$	$2.0e-10$	16.3	0
DTOC5	10	10	9	9	$1.5e+00$	$6.8e-13$	17.9	0
DTOC6	36	36	35	35	$1.3e+05$	$1.8e-12$	25.0	0
ELEC1	58	58	57	130	$6.5e+03$	$1.7e-12$	42.2	0
ELEC2	210	105	104	228	$1.0e+04$	$4.3e-13$	124.7	0
ELEC3	155	107	106	246	$1.8e+04$	$1.5e-11$	233.3	0
EIGENA2	17	16	15	28	$3.0e-28$	$3.4e-13$	0.4	0
EIGENACO	20	20	19	39	$4.2e-27$	$1.6e-12$	1.3	0
EIGENCCO	13	13	12	23	$1.2e-25$	$5.5e-13$	0.1	0
EIGENBCO	1000	195	195	493	$2.2e-01$	$5.6e-07$	18.6	1
EIGENB2	1000	202	202	511	$4.5e-01$	$1.4e-07$	9.5	1
EIGENC2	357	303	302	750	$2.7e-27$	$5.8e-14$	163.4	0
GENHS28	10	10	9	9	$9.3e-01$	$4.4e-16$	0.0	0
GILBERT	21	21	20	23	$4.8e+02$	$2.9e-13$	0.9	0

Nom du problème	nf	ng	nh	nfact	ffin	$\|\|F\|\|$	t	Info
GRIDNETE	11	11	10	10	$2.1e+02$	$2.4e-15$	12.9	0
GRIDNETB	10	10	9	9	$3.8e+01$	$8.5e-16$	0.1	0
GRIDNETH	12	12	11	11	$4.0e+01$	$1.1e-15$	0.1	0
HAGER1	6	6	5	5	$8.8e-01$	$2.2e-12$	14.9	0
HAGER2	18	18	17	17	$4.3e-01$	$1.7e-12$	59.6	0
HAGER3	16	16	15	15	$1.4e-01$	$1.7e-12$	18.7	0
HS006	6	6	5	10	$0.0e+00$	$0.0e+00$	0.1	0
HS007	16	15	14	24	$-1.7e+00$	$1.6e-15$	0.1	0
HS009	9	9	8	15	$-5.0e-01$	$4.9e-13$	0.0	0
HS026	26	26	25	25	$2.1e-16$	$8.1e-09$	0.1	0
HS027	67	40	39	50	$4.0e-02$	$1.1e-14$	0.1	0
HS028	2	2	1	1	$1.0e-30$	$8.9e-16$	0.0	0
HS039	15	15	14	17	$-1.0e+00$	$1.1e-14$	0.1	0
HS040	9	9	8	8	$-2.5e-01$	$1.7e-15$	0.0	0
HS042	10	10	9	9	$1.4e+01$	$8.9e-16$	0.0	0
HS046	29	29	28	33	$5.3e-16$	$5.8e-09$	0.1	0
HS047	21	20	19	19	$1.3e-13$	$7.7e-09$	0.1	0
HS048	2	2	1	1	$2.0e-30$	$3.4e-15$	0.0	0
HS049	21	21	20	20	$2.1e-12$	$7.0e-09$	0.1	0
HS050	10	10	9	9	$1.6e-30$	$3.4e-15$	0.0	0
HS051	2	2	1	1	$2.0e-31$	$1.0e-15$	0.0	0
HS052	11	11	10	10	$5.3e+00$	$8.9e-16$	0.0	0
HS056	38	19	18	28	$-3.5e+00$	$1.3e-15$	0.1	0
HS061	20	20	19	23	$-1.4e+02$	$1.8e-15$	0.3	0
HS077	15	15	14	14	$2.4e-01$	$1.8e-15$	0.0	0
HS078	10	10	9	9	$-2.9e+00$	$8.9e-16$	0.0	0
HS079	9	9	8	8	$7.9e-02$	$1.2e-15$	0.0	0
HS100LNP	13	13	12	18	$6.8e+02$	$2.1e-14$	0.1	0
HS111LNP	18	18	17	25	$-4.8e+01$	$3.6e-15$	0.1	0
LUKVLE1	23	21	20	27	$6.2e+00$	$6.3e-14$	0.2	0
LUKVLE3	19	19	18	22	$2.8e+01$	$1.4e-14$	0.1	0
LUKVLE6	17	17	16	16	$6.0e+03$	$7.2e-10$	0.2	0
LUKVLE7	26	26	25	32	$-2.6e+01$	$2.2e-13$	0.1	0
LUKVLE9	808	105	104	164	$1.0e+01$	$2.8e-14$	0.9	0
LUKVLE11	21	21	20	26	$1.0e-11$	$3.7e-09$	0.2	0
LUKVLE12	1000	125	125	155	$1.6e+05$	$1.0e-04$	1.3	1
LUKVLE13	23	23	22	51	$5.3e+02$	$3.3e-09$	0.5	0
LUKVLE14	47	39	38	43	$2.6e+06$	$1.0e-09$	0.3	0
LUKVLE15	151	57	56	91	$1.7e-11$	$8.4e-09$	0.5	0
LUKVLE16	20	20	19	23	$8.1e-12$	$5.7e-09$	0.1	0
LUKVLE17	1000	965	965	969	$3.2e+02$	$1.1e-10$	5.9	1
LUKVLE18	1000	958	958	2374	$1.1e+02$	$2.6e-11$	26.1	1

Nom du problème	nf	ng	nh	nfact	ffin	$\|\|F\|\|$	t	Info
MARATOS	9	9	8	8	$-1.0e+00$	$5.0e-16$	0.0	0
MWRIGHT	21	21	20	31	$1.3e+00$	$4.8e-16$	0.1	0
ORTHREGA	266	101	100	211	$1.4e+03$	$4.1e-10$	10.0	0
ORTHREGB	3	3	2	4	$4.5e-20$	$2.0e-10$	0.0	0
ORTHREGC	13	13	12	16	$4.0e+00$	$2.0e-14$	0.3	0
ORTHRGDM	21	21	20	40	$7.1e+00$	$1.3e-13$	0.1	0
ORTHREGD	27	25	24	49	$3.2e+01$	$1.2e-13$	0.6	0
ORTHRDM2	13	13	12	17	$3.1e+02$	$2.3e-13$	27.1	0
ORTHRGDS	1000	195	195	265	$1.8e+02$	$5.1e-10$	2.8	1
S216	22	18	17	17	$1.0e-00$	$5.4e-13$	0.1	0
S219	406	95	94	201	$-1.0e+00$	$5.6e-15$	0.4	0
S235	25	16	15	15	$4.0e-02$	$1.8e-15$	0.1	0
S252	34	21	20	26	$4.0e-02$	$7.1e-15$	0.2	0
S269	11	11	10	10	$4.1e+00$	$2.2e-16$	0.0	0
S316	15	15	14	14	$3.3e+02$	$3.0e-12$	0.0	0
S317	18	18	17	17	$3.7e+02$	$7.1e-15$	0.1	0
S318	17	17	16	16	$4.1e+02$	$7.1e-17$	0.0	0
S319	18	18	17	17	$4.5e+02$	$1.4e-14$	0.1	0
S320	22	22	21	21	$4.9e+02$	$3.6e-14$	0.1	0
S321	24	24	23	23	$5.0e+02$	$1.9e-16$	0.1	0
S322	30	30	29	29	$5.0e+02$	$1.4e-14$	0.1	0
S335	26	26	25	25	$-4.5e-03$	$8.8e-13$	0.1	0
S336	35	35	34	41	$-3.4e-01$	$5.7e-14$	0.1	0
S338	268	61	60	108	$-1.1e+01$	$8.9e-16$	0.2	0
S344	11	11	10	10	$3.3e-02$	$8.2e-16$	0.0	0
S345	15	15	14	15	$3.3e-02$	$3.1e-15$	0.1	0
S375	179	44	43	73	$-1.5e+01$	$2.7e-15$	0.2	0
S378	18	18	17	25	$-4.8e+01$	$7.1e-15$	0.1	0
S394	17	17	16	21	$1.9e+00$	$8.9e-16$	0.1	0
S395	18	18	17	23	$1.9e+00$	$4.4e-16$	0.1	0
SPMSQRT	7	7	6	6	$0.0e+00$	$5.3e-11$	13.5	0

Tab. D.3 – Résulats numériques obtenus avec l'algorithme E en considérant la définition (3.2) pour la valeur de μ_0 ainsi que les expressions (3.3) et (3.4) pour le choix dynamique de μ. La décroissance de μ est en plus adaptée à la décroissance de $\|\nabla f(x) + \nabla c(x)\lambda\|$.

Annexe E

Résultats numériques obtenus avec la méthode SQP

Dans cette annexe, nous présentons les résultats numériques obtenus avec l'algorithme C testé sur l'ensemble des problèmes du tableau 3.1. Le tableau suivant énumère les résultats où

- nf représente le nombre d'évaluations de fonctions f et c,
- ng représente le nombre d'évaluations de gradients ∇f,
- nh représente le nombre d'évaluations de hessiens $\nabla^2 f$,
- nfact représente le nombre de factorisations de la matrice $F'_w(w, \mu)$,
- ffin représente la valeur finale de f au dernier point calculé par l'algorithme,
- $\|F\|$ représente la norme des conditions d'optimalité non perturbées en w, i.e. $\|F(w, 0)\|$,
- t représente le temps CPU (en secondes) nécessaire à l'arrêt de l'algorithme,
- Info représente le mode d'arrêt de l'algorithme
  - Info = 0 lorsque l'algorithme s'arrête normalement,
  - Info = 1 lorsque le nombre maximal d'évaluations du nombre de fonctions et de contraintes est atteint,
  - Info = 2 lorsque le nombre maximal de factorisations de $F'_w(w, \mu)$ est atteint,
  - Info = 3 lorsque le nombre maximal de corrections de $F'_w(w, \mu)$ est atteint.

Nom du problème	nf	ng	nh	nfact	ffin	$\|\|F\|\|$	t	Info
AUG2D	3	2	1	1	$1.1e+02$	$1.6e+00$	0.0	0
AUG3D	5	3	2	4	$5.5e+02$	$3.0e-04$	5.1	0
AUG3DC	3	2	1	1	$7.7e+02$	$1.0e+00$	1.9	0
BT1	22	9	8	12	$-1.0e-00$	$1.3e-04$	0.0	0
BT2	25	13	12	12	$3.3e-02$	$1.9e-06$	0.1	0
BT3	3	2	1	1	$4.1e+00$	$8.0e+01$	0.0	0
BT4	32	13	12	19	$-4.6e+01$	$5.0e-06$	0.1	0
BT5	15	8	7	7	$9.6e+02$	$4.2e-07$	0.0	0
BT6	20	10	9	9	$2.8e-01$	$2.1e-05$	0.0	0
BT7	51	16	15	21	$3.1e+02$	$2.3e-07$	0.1	0
BT8	203	102	101	296	$1.0e+00$	$3.8e+06$	0.3	3
BT9	31	13	12	17	$-1.0e+00$	$5.6e-08$	0.1	0
BT11	17	9	8	8	$8.2e-01$	$3.3e-08$	0.0	0
BT12	9	5	4	4	$6.2e+00$	$5.5e-06$	0.0	0
BYRDSPHR	105	20	19	38	$-4.7e+00$	$2.2e-07$	0.1	0
CATENA	15	8	7	7	$-2.3e+04$	$1.3e-07$	0.0	0
CHAIN1	61	17	16	19	$5.1e+00$	$7.8e-06$	1.1	0
CHAIN2	83	22	21	26	$5.1e+00$	$4.3e-08$	2.0	0
CHAIN3	67	20	19	24	$5.1e+00$	$9.3e-07$	2.5	0
CHNRSBNE	1000	40	39	112	$0.0e+00$	$1.2e+09$	0.5	1
DECONVNE	1000	27	26	65	$0.0e+00$	$3.9e-04$	1.1	1
DIXCHLNG	21	11	10	10	$2.5e+03$	$4.6e-05$	0.0	0
DTOC1NA	13	7	6	6	$1.3e+01$	$3.4e-06$	2.9	0
DTOC1NB	13	7	6	6	$1.6e+01$	$2.6e-08$	3.0	0
DTOC1NC	59	15	14	21	$2.5e+01$	$3.9e-05$	7.8	0
DTOC1ND	794	102	101	161	$1.3e+01$	$8.8e-03$	28.2	3
DTOC2	1000	85	84	93	$5.0e-01$	$1.8e-07$	53.4	1
DTOC5	9	5	4	4	$1.5e+00$	$2.5e-08$	39.0	0
DTOC6	25	13	12	12	$1.3e+05$	$1.2e-07$	45.9	0
ELEC1	556	102	101	254	$6.5e+03$	$1.3e-01$	102.5	3
ELEC2	394	102	101	253	$1.0e+04$	$6.7e-01$	158.4	3
ELEC3	379	102	101	253	$1.8e+04$	$1.5e-01$	288.3	3
EIGENA2	9	5	4	5	$4.4e-30$	$1.2e-05$	0.2	0
EIGENACO	9	5	4	5	$5.7e-30$	$8.5e-05$	0.2	0
EIGENCCO	26	13	12	23	$8.4e-28$	$4.0e-07$	0.1	0
EIGENBCO	343	102	101	255	$2.2e-01$	$3.4e-03$	9.6	3
EIGENB2	284	102	101	256	$2.8e-01$	$2.0e-02$	4.8	3
EIGENC2	33	17	16	33	$3.9e-24$	$3.3e-06$	10.0	0
GENHS28	3	2	1	1	$9.3e-01$	$5.6e+00$	0.0	0
GILBERT	39	20	19	22	$4.8e+02$	$4.2e-05$	1.5	0

Nom du problème	nf	ng	nh	nfact	ffin	$\|\|F\|\|$	t	Info
GRIDNETE	14	14	13	13	$2.1e+02$	$2.4e-15$	0.0	0
GRIDNETB	3	2	1	1	$3.8e+01$	$1.0e+01$	0.0	0
GRIDNETH	15	8	7	17	$4.0e+01$	$4.9e-06$	0.1	0
HAGER1	3	2	1	1	$8.8e-01$	$5.0e+03$	16.5	0
HAGER2	3	2	1	1	$4.3e-01$	$5.0e+03$	19.7	0
HAGER3	3	2	1	1	$1.4e-01$	$5.0e+03$	15.0	0
HS006	30	11	10	17	$0.0e+00$	$2.0e-01$	0.0	0
HS007	23	9	8	14	$-1.7e+00$	$3.3e-05$	0.0	0
HS009	17	6	5	6	$-5.0e-01$	$6.8e-07$	0.0	0
HS026	51	26	25	25	$1.3e-16$	$1.4e-08$	0.1	0
HS027	1000	76	75	183	$1.1e-01$	$1.3e-02$	0.4	1
HS028	3	2	1	1	$3.1e-31$	$6.1e+00$	0.0	0
HS039	31	13	12	17	$-1.0e+00$	$5.6e-08$	0.0	0
HS040	7	4	3	3	$-2.5e-01$	$1.6e-04$	0.0	0
HS042	9	5	4	4	$1.4e+01$	$6.2e-08$	0.0	0
HS046	54	26	25	25	$5.0e-16$	$1.3e-08$	0.1	0
HS047	46	22	21	25	$3.2e-14$	$1.2e-08$	0.1	0
HS048	3	2	1	1	$3.9e-31$	$1.6e+01$	0.0	0
HS049	41	21	20	20	$2.1e-12$	$2.3e-08$	0.1	0
HS050	19	10	9	9	$1.2e-32$	$2.2e-06$	0.0	0
HS051	3	2	1	1	$0.0e+00$	$4.4e+00$	0.0	0
HS052	3	2	1	1	$5.3e+00$	$3.3e+01$	0.0	0
HS056	27	13	12	21	$-3.5e+00$	$9.8e-06$	0.0	0
HS061	1000	54	53	67	$9.7e+02$	$4.2e+02$	0.3	1
HS077	20	10	9	9	$2.4e-01$	$4.9e-07$	0.0	0
HS078	9	5	4	4	$-2.9e+00$	$1.0e-05$	0.0	0
HS079	9	5	4	4	$7.9e-02$	$2.5e-04$	0.0	0
HS100LNP	23	10	9	1	$46.8e+02$	$4.7e-05$	0.0	0
HS111LNP	131	34	33	41	$-4.8e+01$	$5.2e-08$	0.2	0
LUKVLE1	13	7	6	6	$6.2e+00$	$6.5e-03$	0.1	0
LUKVLE3	19	10	9	9	$2.8e+01$	$1.9e-04$	0.1	0
LUKVLE6	50	20	19	25	$6.0e+03$	$1.9e-06$	0.2	0
LUKVLE7	25	12	11	20	$-2.6e+01$	$5.6e-04$	0.1	0
LUKVLE9	108	26	25	32	$1.1e+01$	$3.4e-07$	0.2	0
LUKVLE11	1000	67	66	145	$7.7e+03$	$1.0e+15$	1.2	1
LUKVLE12	1000	114	113	142	$1.6e+05$	$1.0e-04$	1.2	1
LUKVLE13	50	24	23	54	$7.9e+02$	$1.4e-08$	0.3	0
LUKVLE14	215	48	47	52	$2.6e+06$	$1.1e-08$	0.4	0
LUKVLE15	1000	55	54	126	$8.8e+02$	$1.6e+02$	1.0	1
LUKVLE16	1000	54	53	57	$2.2e+06$	$2.2e+23$	0.7	1
LUKVLE17	1000	36	35	35	$3.0e+02$	$9.9e+04$	0.6	1
LUKVLE18	1000	20	19	28	$9.4e+01$	$3.2e+02$	0.6	1

Nom du problème	nf	ng	nh	nfact	ffin	$\|\|F\|\|$	t	Info
MARATOS	9	5	4	4	$-1.0e+00$	$3.9e-08$	0.0	0
MWRIGHT	21	11	10	16	$2.5e+01$	$2.9e-08$	0.0	0
ORTHREGA	350	102	101	251	$1.7e+03$	$3.9e-01$	11.7	3
ORTHREGB	5	3	2	4	$4.5e-20$	$1.2e-04$	0.1	0
ORTHREGC	17	9	8	12	$4.0e+00$	$1.7e-07$	0.2	0
ORTHRGDM	208	102	101	207	$3.6e+02$	$1.1e+02$	0.7	3
ORTHREGD	14	7	6	6	$3.2e+01$	$6.3e-06$	0.1	0
ORTHRDM2	12	6	5	5	$3.1e+02$	$2.3e-02$	22.7	0
ORTHRGDS	1000	65	64	117	$1.2e+03$	$1.3e+06$	1.4	1
S216	15	7	6	6	$1.0e-00$	$4.3e-04$	0.0	0
S219	79	22	21	32	$-1.0e+00$	$1.0e-05$	0.1	0
S235	104	24	23	23	$4.0e-02$	$2.1e-05$	0.1	0
S252	147	30	29	33	$4.0e-02$	$1.7e-07$	0.1	0
S269	3	2	1	1	$4.1e+00$	$8.0e+00$	0.0	0
S316	1000	21	20	62	$1.5e-01$	$3.8e+07$	0.3	1
S317	1000	21	20	62	$1.5e-01$	$5.5e+07$	0.3	1
S318	1000	22	21	64	$7.6e+04$	$9.2e+07$	0.3	1
S319	1000	22	21	64	$2.3e+00$	$2.1e+08$	0.3	1
S320	1000	22	21	64	$3.8e+01$	$5.3e+08$	0.3	1
S321	1000	22	21	64	$2.2e+02$	$9.1e+08$	0.3	1
S322	42	18	17	35	$5.0e+02$	$5.7e-04$	0.1	0
S335	91	26	25	25	$-4.5e-03$	$3.5e-07$	0.1	0
S336	90	23	22	36	$-3.4e-01$	$1.4e-07$	0.4	0
S338	1000	23	22	68	$-8.6e+01$	$1.6e+11$	0.3	1
S344	15	8	7	7	$3.3e-02$	$3.1e-06$	0.0	0
S345	144	32	31	32	$3.3e-02$	$1.8e-05$	0.1	0
S375	1000	21	20	54	$-9.9e+01$	$5.5e+03$	0.3	1
S378	131	34	33	41	$-4.8e+01$	$5.2e-08$	0.1	0
S394	947	102	101	250	$8.5e+00$	$4.4e+00$	0.6	3
S395	152	35	34	39	$1.9e+00$	$1.5e-07$	0.4	0
SPMSQRT	1000	190	189	75	$4.8e-09$	$5.3e-11$	16.3	1

Tab. E.1 – Résulats numériques obtenus avec l'algorithme C.

Bibliographie

[1] M. Abramowitz, I. A. Stegun, *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, vol. 55 of National Bureau of Standards Applied Mathematics Series, For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C., 1964.

[2] C. Aime, R. Soummer, *Introduction to stellar coronography with entrance pupil apodization*, Astronomy with High Contrast Imaging, EAS Publications Series, 8, 79, 2003.

[3] W. Appel, *Mathématiques pour la physique et les physiciens*, H et K Editions, 2005.

[4] M. Arioli, J.W. Demmel, I.S. Duff, *Solving sparse linear systems with sparse backward error*, SIAM Journal on Matrix Analysis and Applications, 10, 1989, pp. 165–190.

[5] P. Armand, J. Benoist, *A local convergence property of primal-dual methods for nonlinear programming*, Mathematical Programming, Series A, 115 (2008), pp. 199–222.

[6] P. Armand, J. Benoist, and D. Orban, *Dynamic updates of the barrier parameter in primal-dual methods for nonlinear programming*, Computational Optimization and Applications, 41 (2008), pp. 1–25.

[7] P. Armand, J. Benoist, and D. Orban, *Global convergence of primal-dual methods for nonlinear programming*, Rapport de recherche Xlim, 2008.

[8] P. Armand, J. Benoist, E. Bousquet, L. Delage, S. Olivier, and F. Reynaud, *Etude du principe de modulation temporelle dans un hypertélescope fibré*, Rapport CNES DCT/SI/OP/2005-108, XLIM, 2007.

[9] P. Armand, J. Benoist, E. Bousquet, L. Delage, S. Olivier, and F. Reynaud, *Optimization of a one dimensional hypertelescope for a direct imaging instrument in astronomy*, European Journal of Operational Research, 195 (2009), pp. 519–527.

[10] P.T. Boggs, J.W. Tolle, *Sequential Quadratic Programming*, Acta Numerica, 4 (1995), pp. 1–52.

[11] J.F. Bonnans, J.C. Gilbert, C. Lemaréchal, C. Sagastizabal, *Optimisation Numérique. Aspects théoriques et pratiques*, Mathématiques et Applications, Springer-Verlag, 1997.

[12] F. Boone, *Interferometric array design : optimizing the locations of the antenna pads*, Astronomy & Astrophysics, 377 (2001), pp. 368–376.

[13] F. Boone, *Interferometric array design : distribution of Fourier samples for imaging*, Astronomy & Astrophysics, 386 (2002), pp. 1160–1171.

[14] E. Bousquet, *Optimisation du positionnement des pupilles d'entrée d'un hypertélescope*, Mémoire de Master Recherche, 2006.

[15] R.H. Byrd, N.I.M. Gould, J. Nocedal, and R.A. Waltz, *An algorithm for nonlinear optimization using linear programming and equality constrained subproblems*, Mathematical Programming, Series B, 100 (2004), pp. 27–48.

[16] H. Cartan, *Cous de calcul différentiel*, Collection Méthodes, Hermann Paris, 1967.

[17] A.R. Conn, N.I.M. Gould, and P.L. Toint, *Trust-Region Methods*, MPS-SIAM Series on Optimization, SIAM, 2000.

[18] R. Courant, *Variational methods for the solution of problems with equilibrium and vibration*, Bulletin of the American Mathematical Society, 49 (1943), pp. 1–23.

[19] J.P. Demailly, *Analyse Numérique et équations différentielles*, Presses Universitaires de Grenoble, 1996.

[20] E.D. Dolan, J.J. Moré, *Benchmarking optimization software with performance profiles*, Mathematical Programming, 91 (2002), pp. 201–213.

[21] E.D. Dolan, J.J. Moré, and T.S. Munson, *Benchmarking optimization software with COPS 3.0*, Technical Report ANL/MCS-273, Argonne National Laboratory, 2004.

[22] Iain S. Duff, *MA57-A code for the solution of sparse symmetric definite and indefinite systems*, ACM Transactions on Mathematical Software, 30, pp. 118–144.

[23] A.V. Fiacco, and G.P. McCormick, *Nonlinear Programming : Sequential Unconstrained Minimization Techniques*, John Wiley and Sons, New York, 1968.

[24] R. Fletcher, *Second order corrections for non-differentiable optimization*, Numerical Analysis, D. Griffiths, ed., Springer Verlag, 1982, pp. 85–114.

[25] R. Fletcher, and S. Leyffer, *Nonlinear Programming without a penalty function*, Mathematical Programming, Series A, 91 (2002), pp. 239–269.

[26] A. Forsgren, P.E. Gill, *Primal-dual interior methods for nonconvex nonlinear programming*, SIAM Journal on Optimization, 8 (1998), pp. 1132–1152.

[27] R. Fourer, D. M. Gay, and B. W. Kernighan, *AMPL A modeling language for mathematical programming*, Duxbury Press Brooks Cole Publishing Co., second ed., 2003.

[28] U.M. Garcia-Palomares, O.L. Mangasarian, *Superlinearly convergent quasi-Newton methods for nonlinearly constrained optimization problems*, Mathematical Programming Studies, 16 (1982), pp. 18–44.

[29] E.M. Gertz, P.E. Gill, *A primal-dual trust region algorithm for nonlinear optimization*, Mathematical Programming, Series B, 100 (2004), pp. 49–94.

[30] P.E. Gill, W. Murray, and M.A. Saunders, *SNOPT : An SQP algorithm for large-scale constrained optimization*, SIAM Journal on Optimization, 12 (2002), pp. 979–1006.

[31] N.I.M. Gould, *On the accurate determination of search directions for simple differentiable penalty functions*, I.M.A. Journal on Numerical Analysis, 6 (1986), pp. 357–372.

[32] N.I.M. Gould, *On the convergence of a sequential penalty function method for constrained minimization*, Siam Journal Numerical Analysis, 26 (1989), pp. 107–126.

[33] N.I.M. Gould, D. Orban, P.L. Toint, *CUTEr (ans SifDec), a Constrained and Unconstrained Testing Environment, revisited*, ACM Transactions on Mathematical Software, 2003.

[34] S.P. Han, *Superlinearly convergent variable metric algorithms for general nonlinear programming problems*, Mathematical Programming, 11 (1976), pp. 263–282.

[35] S.P. Han, *A globally convergent method for nonlinear programming*, Journal of Optimization Theory and Applications, 22 (1977), pp. 297–309.

[36] E. Hecht, *Optics*, Addison-Wesly Publishing Company, second ed., 1987.

[37] Roger A. Horn, Charles R. Johnson, *Matrix Analysis*, Cambridge University Press, 1990.

[38] J.P. Kahane, P.G. Lemarié-Rieusset, *Séries de Fourier et ondelettes*, Nouvelle bibliothèque mathématique, Cassini, 1998.

[39] A. Karastergiou, R. Neri, M.A. Gurwell, *Adapting and expanding interferometric arrays*, The Astrophysical Journal Supplement Series, 164 (2006), pp. 552–558.

[40] N. J. Kasdin, R. J. Vanderbei, D. N. Spergel, and M. G. Littman, *Extrasolar planet finding via optimal apodized-pupil and shaped-pupil coronagraphs*, The Astrophysical Journal, 582 (2003), pp. 1147–1161.

[41] A. Labeyrie, *Resolved imaging of extra-solar planets with future 10-100 km optical interferometric arrays*, Astronomy & Astrophysics, 118 (1996), pp. 517–524.

[42] A. Labeyrie, S. G. Lipson, and P. Nisenson, *An Introduction to Optical Stellar Interferometry*, Cambridge University Press, 2006.

[43] A. Labeyrie, H. Le Coroller, J. Dejonghe, F. Martinache, V. Borkowski, G. Lardière, L. Koechlin, *Hypertelescope imaging : from exoplanets to neutron stars*, Proc. SPIE, Hawaï, 2002.

[44] O. Lardière, F. Martinache, F. Patru, *Direct imaging with highly diluted apertures. I. Field-of-view limitations*, Monthly Notices of the Royal Astronomical Society, 375, (2007), pp. 977–988.

[45] P.J. Laurent, *Approximation et optimisation*, Collection Enseignement des sciences, Hermann, 1972.

[46] P. R. Lawson, *Selected papers on long baseline stellar interferometry*, vol. MS 139, Spie Milestone Series, second ed., 1997.

[47] M. Lopez, G. Still, *Semi-infinite programming*, European Journal of Operational Research, 180 (2007), pp. 491–518.

[48] N. Maratos, *Exact penalty function algorithms for finite dimensional and control optimization problems*, Thèse de Doctorat, Imperial College, London, 1978.

[49] M. Martínez-Corrala, P. Andrésa, J. Ojeda-Castañedab, and G. Saavedraa, *Tunable axial superresolution by annular binary filters. application to confocal microscopy*, Optics Communications, 119 (1995), pp. 491–498.

[50] J. Mawhin, *Analyse, fondements, techniques, évolution*, 2ème édition, De Boeck Université, 1997.

[51] N.J. Miller, M.P. Dierking, B.D. Duncan, *Optical sparse aperture imaging*, Optical Society of America, 46 (2007), pp. 5933–5943.

[52] J. Nocedal, S.J. Wright, *Numerical Optimization*, Second Edition, Springer Series in Operations Research, 2006.

[53] S. Olivier, *Utilisation de dispositifs d'optique guidée pour des applications en imagerie haute résolution*, Thèse de Doctorat : Université de Limoges, 2007.

[54] J.M. Ortega, W.C. Rheinboldt, *Iterative solution of nonlinear equation in several variables*, Classics in Applied Mathematics, 2000.

[55] F. Patru, D. Mourard, D. Lardière, S. Lagarde, *Optimization of the direct imaging properties of an optical-fibred long baseline interferometer*, Monthly Notices of the Royal Astronomical Society, 376, (2007), pp. 1047–1053.

[56] F. Patru, N. Tarmoul, D. Mourard, O. Lardière, *Direct imaging with highly diluted apertures. II. Properties of the point spread function of a hypertelescope*, Monthly Notices of the Royal Astronomical Society, 395, (2009), pp. 2363–2372.

[57] G. Perrin, J. Woillez, O. Lai, J. Guérin, T. Kotani, P.L. Wizinowitch, D. Le Mignant, M. Hrynecych, J. Gathright, P. Léna, F. Chaffee, S. Vergnole, L. Delage, F. Reynaud, A.J. Adamson, C. Berthod, B. Brient, C. Collin, J. Crétenet, F. Dauny, C. Deléglise, P. Fédou, T. Goeltzenlichter, O. Guyon, R. Hulin, C. Marlot, M. Marteaud, B.T. Melse, J. Nishikawa, J.M. Reess, S.T. Ridgway, F. Rigaud, K. Roth, A.T. Tokunaga, D. Ziegler, *Interferometric Coupling of the Keck Telescopes with Single-Mode Fibers*, SCIENCE, 311, 2006.

[58] M.J.D. Powell, *A fast algorithm for nonlinearly constrained optimization calculations*, in Numerical Analysis Dundee 1977, G.A. Watson, ed., Springer Verlag, Berlin, 1977, pp. 144–157.

[59] M.J.D. Powell, *Algorithms for nonlinear constraints that use Lagrangian functions*, Mathematical Programming, 14 (1978), pp. 224–248.

[60] M.J.D. Powell, *The convergence of variable metric methods for nonlinearly constrained optimization calculations*, in Nonlinear Programming 3, Academic Press, New York and London, 1978, pp. 27–63.

[61] F. Reynaud and L. Delage, *Proposal for a temporal version of a hypertelescope*, Astronomy & Astrophysics, 465 (2007), pp. 1093–1097.

[62] D. Slepian, *Analytic solution of two apodization problems*, Journal of the optical society of america, 55 (1965), pp. 1110–1115.

[63] F. Vakili, E. Aristidi, L. Abe, and B. Lopez, *Interferometric remapped array nulling*, Astronomy & Astrophysics, 421 (2004), pp. 147–156.

[64] R. J. Vanderbei, *Extreme optics and the search for Earth-like planets*, Mathematical Programming, 112 (2008), pp. 255–272.

[65] R. J. Vanderbei, *LOQO : An interior point code for quadratic programming*, Optimization Methods and Software, 12, (1999), pp. 451–484.

[66] M. De Villiers, *Interferometric array layout design by tomographic projection*, Astronomy & Astrophysics, 469, (2007), pp. 793–797.

[67] A. Wächter and L. T. Biegler, *On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming*, Math. Program., 106 (2006), pp. 25–57.

[68] R.B. Wilson, *A simplicial algorithm for concave programming*, PhD Thesis, Graduate School of Business Administration, Harvard University, 1963.