Sur quelques problèmes de reconstruction en imagerie MA-TIRF et en optimisation parcimonieuse par relaxation continue exacte de critères pénalisés en norme-l0

Emmanuel Soubies

Université
Nice
Sophia Antipolis

THÈSE

En vue de l'obtention du titre de

DOCTEUR EN SCIENCES

Délivré par : *l'Université Nice Sophia Antipolis (UNS)*

Présentée et soutenue le *14/10/2016* **par :**

EMMANUEL SOUBIES

Sur quelques problèmes de reconstruction en imagerie MA-TIRF et en optimisation parcimonieuse par relaxation continue exacte de critères pénalisés en norme-ℓ_0

JURY

GILLES AUBERT	Professeur des Universités	Co-encadrant
LAURE BLANC-FÉRAUD	Directrice de Recherche	Directrice
MILA NIKOLOVA	Directrice de Recherche	Examinatrice
JEAN-CHRISTOPHE OLIVO-MARIN	Directeur de Recherche	Examinateur
JEAN-CHRISTOPHE PESQUET	Professeur des Universités	Rapporteur
GABRIEL PEYRÉ	Directeur de Recherche	Rapporteur
SÉBASTIEN SCHAUB	Ingénieur de Recherche	Co-encadrant
ELLEN VAN OBBERGHEN-SCHILLING	Directrice de Recherche	Examinatrice

---

École doctorale et spécialité :
STIC : Automatique, Traitement du Signal et des Images
Unité de Recherche :
I3S/INRIA/iBV, Projet MORPHEME
Directrice de Thèse :
Laure BLANC-FÉRAUD
Co-encadrants :
Gilles AUBERT et *Sébastien SCHAUB*
Rapporteurs :
Jean-Christophe PESQUET et *Gabriel PEYRÉ*

REMERCIEMENTS

Il est maintenant temps d'écrire les dernières lignes de ce manuscrit (bien qu'elles le débutent) afin de remercier toutes les personnes que j'ai eu la chance de rencontrer au cours de cette thèse mais aussi celles qui sont présentes depuis bien des années et qui ont largement contribué à rendre ces trois années aussi agréables qu'elles l'ont été.

Mes premiers remerciements s'adressent tout naturellement à mes encadrants, Laure, Gilles et Sébastien, pour m'avoir accompagné et guidé dans la bonne humeur tout au long de cette thèse. Je les remercie également pour tous les conseils précieux qu'ils m'ont apportés scientifiquement mais aussi pour la suite de l'aventure (post-thèse). Par ailleurs, vous avez toujours répondu à mes questions avec enthousiasme, même lorsque je débarquais à l'improviste dans le bureau au milieu de la journée (et Laure sait bien si c'était très souvent le cas ;-)). Vous avez également su orienter mes recherches tout en me laissant la liberté d'explorer les pistes que je souhaitais rendant le travail d'autant plus passionnant et je vous en suis très reconnaissant. Merci pour tout ! ! !

Je souhaite ensuite remercier chaleureusement Jean-Christophe Pesquet et Gabriel Peyré d'avoir accepté spontanément d'être rapporteurs de cette thèse. J'en suis très honoré. Je remercie aussi Gabriel Peyré de m'avoir accueilli une semaine au Ceremade et pour les échanges très intéressants que nous avons eu lors de cette visite.

Mes sincères remerciements vont également à Mila Nikolova, Ellen Van Obberghen-Schilling et Jean-Christophe Olivo-Marin, pour avoir acceptés de faire partie du jury. Je remercie aussi Ellen pour la patience dont elle a fait part afin de m'expliquer simplement le processus d'adhésion cellulaire. Bien que les détails restent trop complexes pour moi, ces explications me furent très bénéfiques pour la partie applicative de la première moitié de ce mémoire. Enfin, je suis heureux de la présence de Mila Nikolova dans ce jury dont les travaux sont à l'origine de la deuxième partie de cette thèse. Je tiens par ailleurs à la remercier pour tous les échanges instructifs que nous avons eu cette dernière année autour des thématiques présentées dans le chapitre 10.

Un grand merci aux autres permanents de l'équipe Morpheme : Xavier, Éric et Grégoire pour les échanges scientifiques, rugbystiques ou encore culturels que nous avons pu avoir durant cette thèse. Je les remercie aussi pour le cadre convivial qu'ils ont su créer dans l'équipe Morpheme notamment lors des repas du midi, des conférences ou encore du pique-nique annuel sur la plage ! Enfin, merci à Christine, Micheline, Nadia, Frédéric et Jane (et ses mails où toute l'information est contenue dans l'objet ;-)) pour leur grande aide lors de la préparation des missions tout comme pour l'organisation de la soutenance.

Qu'auraient été ces trois années de thèse sans les collègues/amis Niçois, Antibois ou encore Vallauriens ! Je pourrais simplement résumer mes remerciements les concernant en un simple «HUIII» [1] dont les concernés en comprendraient tout le sens mais je vais tout de même m'étendre un peu plus et commencer par remercier mes collègues de bureau, Alexis (Alpha) et Gaël (La Miche), pour leur bonne humeur et la bonne ambiance qui régnait dans ce bureau 127 faisant aussi office de laboratoire expérimental, café des sports, exposition d'œuvres (artistiques) et qui pouvait même parfois être assimilé au QG du premier étage du bâtiment Euclide ;-). Que de discussions et débats en tout genres, «piratage» informatique et autres perles (dont on pourrait en écrire un livre !) ont marqué toutes ces journées. En parlant de cet étage, mon épisode Sophipolitain n'aurait pas été le même sans les discussions (scientifiques) matinales, pouvant s'éterniser des heures, avec Arnaud (Cé-

1. Celui là est bien volontaire, ce n'est pas Gaël qui l'a ajouté lors d'un moment d'inattention de ma part !

dric), les rires de Lola (Lol), les débriefs Top Chef avec Agus (Hibou), les «shame on you» de Froso (La Frosse), les cours d'origami de Marine (La Divine), ou encore sans l'enthousiasme des autres doctorants, stagiaires et post-docs que j'ai eu la chance de rencontrer pendant cette aventure triennale : Alejandro, Sylvain, Rita, Yasmine, Manon, Emma (Pouliche), Simone, Sen, Wei, Yaqun, Nico, Djampa, Raphaël et enfin Simon avec qui j'ai eu le plaisir de travailler sur la reconstruction PALM qui est l'une des applications du chapitre 11 de ce manuscrit. Bien sûr, je n'oublie pas Mr le président de l'ADSTIC Benjamin et ses acolytes du deuxième étage, Ophélie, Jonathan, Emilien et nos sorties footing qui ont ponctuées mes derniers mois à Sophia, entraînez vous bien je reviendrai !

Merci à Adilson Chinatto pour s'être levé très tôt au Brésil lors de nos rendez-vous skype et pour sa patience à répondre à tous mes emails. Je remercie également les autres personnes qui ont été impliquées dans ce projet : Cynthia Junqueira, Pascal Larzabal, Jean-Pierre Barbot et João M. T. Romano, m'ayant permis de découvrir les problématiques rencontrées en traitement d'antennes.

Je tiens aussi à exprimer mes remerciements à Lionel Fillatre et Luc Deneire pour m'avoir fait confiance pour le monitorat et m'avoir donné une certaine liberté ayant rendu l'expérience d'autant plus enrichissante.

Je suis également très reconnaissant à Pierre Weiss pour m'avoir donné goût à la recherche lors de mon stage de M1, m'avoir encouragé à poursuivre dans le monde académique (au grand dam de mon père alors qu'une brillante carrière viticole me tendait les bras ;-)) et m'avoir toujours été de très bon conseil. Un grand merci à lui !

Merciiii aux coupaings du Creew : Coco, Fitia, Flo, Gégé, Grugru, KK, Lisa, Lulu, Paul et Toto pour toutes les fois où j'ai squatté à Paris, Toulouse, Alénya ou encore Leamington Spa ! «C'est sympa» (il fallait bien que je la case celle là ;-)) et ces weekends passés ensembles sont toujours de super moments. Maintenant, il faut prendre date pour Baltimore et Lausanne ! :-)

Je ne pourrais pas terminer ce paragraphe sans mentionner les amis de longe date, Valérie (et Alex), Clément (et Lucy) et Romain (et Pétia) qui sont toujours présents et ne manquent jamais, malgré l'éloignement géographique, d'organiser un petit restau lorsque nous nous retrouvons tous à Lisle sur Tarn :-). Merci aussi d'être venus me rendre visite à Nice et de m'avoir accueilli à Bormes-les-Mimosas ou encore en Bulgarie !

J'ai également une pensée pour les amis du Judo Club Lislois dont je ne vais pas commencer à faire une énumération au risque d'en oublier certains par mégarde. Bien que mon assiduité fût plus que «parcimonieuse» (voir chapitre 7 pour une définition) ces trois dernières années, c'est toujours un grand plaisir de passer vous voir à la Salle pour me dérouiller sur les tatamis puis partager un bon repas dans une excellente ambiance pour enfin terminer par une partie de belote ou de pétanque sur les tapis. Merci à tous et en particulier à Daniel pour m'avoir accompagné dans ce sport, mais aussi à l'extérieur, depuis mes plus jeunes années.

Pour finir, un grand merci à mes parents pour m'avoir toujours soutenu (jusqu'à la préparation du pot de thèse) et pour leur patience, s'étant souvent contentés d'un simple «ça va» les jours où je n'étais pas disposé à discuter de mon travail de thèse. Merci aussi à toute la famille pour leur soutien et leurs encouragements en cette période de fin de thèse.

TABLE DES MATIÈRES

LISTE DES ACRONYMES

MICROSCOPIE

BFP plan focal arrière, *Back Focal Plane* en anglais

EMCCD Electron Multiplying Charge-Coupled Device

HELM Harmonic Excitation Light Microscopy

MA-TIRF Multi-Angle Total Internal Reflection Fluorescence

PALM Photo Activation Localization Microscopy

PSF fonction d'étalement du point, *Point Spread Function* en anglais

STED STimulated Emission Depletion microscopy

STORM STochastic Optical Reconstruction Microscopy

TIRF Total Internal Reflection Fluorescence

ALGORITHMES D'OPTIMISATION

BP Basis Pursuit

BPDN Basis Pursuit De-Noising

CoSaMP Compressive Sampling Matching Pursuit

CSBR Continuation Single Best Replacement

DC Différence de fonctions Convexes

CP Chambolle-Pock

ESPRIT Estimation of Signal Parameters via Rotational Invariance Technique

FBS Forward-Backward Splitting

FISTA Fast Iterative Shrinkage Thresholding Algorithm

FoBa Algorithme glouton Forward-Backward de Zhang (2011)

FOCUSS FOCal Underdetermined System Solver

GIST General Iterative Shrinkage Thresholding

GNC Graduated Non Convexity

GNCelo Graduated Non convexity Celo

GNCeloRP GNCelo Regularization Path

gOMP Generalized Orthogonal Matching Pursuit

GSS Greedy Sparse Simplex

HTP Hard Thresholding Pursuit

IHT Iterative Hard Thresholding

IRL1 Iteratively Reweighted ℓ_1

IRLS Iteratively Reweighted Least Squares

ISTA Iterative Shrinkage Thresholding Algorithm

LARS Least Angle Regression
L0-PD ℓ_0-Path Descent
MM Majorisation-Minimisation
MP Matching Pursuit
MUSIC MUltiple SIgnal Classification
OLS Orthogonal Least Squares
OMP Orthogonal Matching Pursuit
OMPR Orthogonal Matching Pursuit with Replacement
PPXA Parallel ProXimal Algorithm
RL Richardson-Lucy
ROMP Regularized Orthogonal Matching Pursuit
SBR Single Best Replacement
SC Stochastic Continuation
SIM Structured Illumination Microscopy
SL0 Smoothed-ℓ_0
SMLR Single Most Likely Replacement
SP Subspace Pursuit
StOMP Stagewise Orthogonal Matching Pursuit

APPROXIMATIONS CONTINUES NON-CONVEXES DE LA NORME-ℓ_0

CEL0 Continuous Exact ℓ_0
MCP Minimax Concave Penalty
SCAD Smoothly Clipped Absolute Deviation

DIVERS

CW Coordinate-Wise
DoA Direction of Arrivals
EQM Erreur Quadratique Moyenne
ERC Exact Recovery Coefficient
iBV Institut de Biologie Valrose
i.i.d. indépendant et identiquement distribué
KL Kurdyka-Lojasiewicz
MAP Maximum A Posteriori
NSP Null Space Property
PMNE Programmation Mixte en Nombres Entiers
RIP Restricted Isometry Property
SNR Rapport Signal sur Bruit, *Signal to Noise Ratio* en anglais
s.c.i. semi-continue inférieurement
s.c. sous contrainte
ULA Uniform Linear Array

LISTE DES NOTATIONS

DÉFINITIONS ET NOTATIONS GÉNÉRALES

$\bar{\mathbb{R}}$	droite réelle achevée $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$,
I_d	matrice identité,
$e_k \in \mathbb{R}^N$	k-ème vecteur de la base canonique de $\mathbb{R}^N$,
$\sharp S$	cardinalité (i. e. nombre d'éléments) de l'ensemble S,
$\mathbb{I}_N$	ensemble des $N \in \mathbb{N}_+$ premiers indices $\mathbb{I}_N := \{1, \dots, N\}$,
ω^c	complémentaire de l'ensemble ω,

ALGÈBRE LINÉAIRE

$\langle \cdot, \cdot \rangle$	produit scalaire euclidien,
$\| \cdot \|_p$	norme-ℓ_p pour $p > 0$ définie par $\|x\|_p := (\sum_i \|x_i\|^p)^{1/p}$,
$\| \cdot \|_\infty$	norme-ℓ_∞ définie par $\|x\|_\infty := \max_i \ \|x_i\|$,
$\| \cdot \|$	lorsque l'indice n'est pas spécifié, désigne la norme euclidienne (ℓ_2) pour un vecteur et spectrale pour une matrice,
$\| \cdot \|_{2,1}$	norme mixte «2-1» définie pour $x \in \mathbb{R}^{N \times M}$ par $\|x\|_{2,1} = \sum_{i \in \mathbb{I}_N} \|(x_{i1}, \dots, x_{iM})\|$,
$\| \cdot \|_0$	pseudo norme[1]-ℓ_0 définie par $\|x\|_0 := \sharp\{x_i,\ i = 1, \dots, N : x_i \neq 0\}$,
$\| \cdot \|_0$	fonction «1-0» (norme-ℓ_0 dans $\mathbb{R}$), $\|u\|_0 = \{0$ si $u = 0,\ 1$ sinon$\}$,
$\mathrm{rank}(A)$	rang de la matrice A,
$\ker(A)$	noyau de A : $\ker(A) := \{x : Ax = 0\}$,
$\mathrm{span}(A)$	espace vectoriel engendré par les colonnes de A,
$A^\dagger$	pseudo-inverse de la matrice A,
$\mathcal{B}_p(c, \rho)$	boule ouverte ℓ_p ($p \in \{1, 2, \infty\}$) de centre c et de rayon ρ, $\mathcal{B}_p(x, \rho) := \{x : \|x - c\|_p < \rho\}$,
$\bar{\mathcal{B}}_p(c, \rho)$	boule fermée ℓ_p ($p \in \{1, 2, \infty\}$) de centre c et de rayon ρ, $\mathcal{B}_p(x, \rho) := \{x : \|x - c\|_p \leqslant \rho\}$,

1. Dans la suite du manuscrit, nous utiliserons la dénomination «norme-ℓ_0» même si ce n'est pas une norme car elle ne vérifie pas l'hypothèse d'homogénéité : $\forall (\lambda, x) \in \mathbb{R} \setminus \{-1, 1\} \times \mathbb{R} \setminus \{0\},\ \|\lambda x\|_0 = \|x\|_0 \neq |\lambda| \|x\|_0$.

DÉFINITIONS ET OPÉRATIONS SUR DES FONCTIONS

$\mathbb{1}_{\{x\in\mathcal{E}\}}$	indicatrice «0/1» sur l'ensemble $\mathcal{E}$ définie par $\mathbb{1}_{\{x\in\mathcal{E}\}} := \{1 \text{ si } x \in \mathcal{E},\ 0 \text{ sinon}\}$,
i_C	indicatrice «0/∞» sur l'ensemble C définie par $i_C(x) := \{0 \text{ si } x \in C,\ +\infty \text{ sinon}\}$,
$F^i(t;x^{(i)})$	restriction de la fonction F à la i-ème variable au point x, $F^i(t;x^{(i)}) = F(x^{(i)} + e_i t)$ avec $x^{(i)}$ défini ci-dessous,
cl F	fermeture de F (i. e. $\text{epi}(\text{cl}\, F) = \text{cl}(\text{epi}\, F)$),
co F	enveloppe convexe de F (définition 9.1 page 95),
$F^\star(x^\star)$	transformée de Legendre-Fenchel de la fonction F (fonction conjuguée, définition 9.2 page 96),
$F^{\star\star}(x)$	enveloppe convexe fermée de la fonction F (fonction biconjuguée),
$\partial F(x)$	gradient généralisé de la fonction F en x (définition 9.8 page 99),
$\text{prox}_{\gamma F}(\cdot)$	opérateur proximal de F défini en (3.22) (page 34) et en (8.5) (page 77),
$(\cdot)_+$	fonction $(u)_+ = \max(0, u)$,

NOTATIONS ASSOCIÉES AU PROBLÈME MA-TIRF

$\Omega \subseteq \mathbb{R}^2$	domaine image 2D (x, y),
$u := (x, y) \in \Omega$	position du domaine image Ω,
n_i	indice de réfraction du milieu incident ou inférieur (verre),
n_t	indice de réfraction du milieu transmit ou supérieur,
U_{gv}	tension appliquée au miroir galvanométrique,
NA	ouverture numérique de l'objectif : $NA = n \sin(\alpha)$ où n représente l'indice du milieu et α le demi angle d'ouverture de la lentille frontale,
F_{obj}	distance focale de l'objectif,
F_1	distance focale de la lentille 1 (voir figure 4 page 14),
α_c	angle critique $\alpha_c = \arcsin(n_t/n_i)$,
α_{max}	angle maximal du système $\alpha_{max} = \arcsin(NA/ni)$,
$\mathcal{A} = \{\alpha_1, \ldots, \alpha_L\}$	angles MA-TIRF tels que $\forall \alpha \in \mathcal{A}\ \alpha \in (\alpha_c, \alpha_{max}]$,
$p(\alpha)$	inverse de la profondeur de pénétration de l'onde évanescente définie en (1.3) page 12,
$I_0(\alpha)$	intensité à l'interface donnée par (1.7) page 15,
b_i	signal de fond au pixel $i \in \mathbb{I}_N$ défini en (1.8) page 15,
Q_e	efficacité quantique définie en (1.9) page 16,
$f : \Omega \times \mathbb{R}_+ \to \mathbb{R}_+$	densité de fluorophores,
$f \in \mathbb{R}^{N\times M}$	version discrète de f où N est le nombre de pixels discrétisant Ω et M est le nombre de points de discrétisation selon z,
$s(\alpha) \in \mathbb{R}^N$	acquisition TIRF pour l'angle $\alpha \in (\alpha_c, \alpha_{max}]$,
$S := \{s(\alpha) : \alpha \in \mathcal{A}\}$	acquisition MA-TIRF,

DÉFINITIONS ET NOTATIONS ASSOCIÉES AU PROBLÈME ℓ_2-ℓ_0

$A \in \mathbb{R}^{M \times N}$	matrice d'observation du signal ou dictionnaire,
$d \in \mathbb{R}^M$	signal observé,
$\lambda \in \mathbb{R}_+$	hyperparamètre du problème régularisé ℓ_2-ℓ_0,
(C_k)	problème $\arg\min_{x \in \mathbb{R}^N} \|Ax - d\|_2^2$ s.c. $\|x\|_0 \leqslant k$,
(C_ε)	problème $\arg\min_{x \in \mathbb{R}^N} \|x\|_0$ s.c. $\|Ax - d\|_2^2 \leqslant \varepsilon$,
(P_λ)	problème $\arg\min_{x \in \mathbb{R}^N} G_{\ell_0}(x) := \frac{1}{2}\|Ax - d\|_2^2 + \lambda\|x\|_0$,
$a_i \in \mathbb{R}^M$	i-ème colonne de $A \in \mathbb{R}^{M \times N}$. On considère $a_i \neq 0_{\mathbb{R}^M}$,
$G_{\ell_0}(x)$	fonction objectif ℓ_2-ℓ_0 définie en (P_λ),
$\phi(a, \lambda; u)$	pénalité CEL0 1D définie en (9.5) page 96,
$\Phi_{\text{CEL0}}(x)$	pénalité CEL0 ND définie par $\Phi_{\text{CEL0}}(x) := \sum_i \phi_{\text{CEL0}}(\|a_i\|, \lambda; x_i)$,
$G_{\text{CEL0}}(x)$	fonction objectif ℓ_2-CEL0 définie en (9.9) page 98,
$A_\omega = (a_{\omega[1]}, \ldots, a_{\omega[\sharp\omega]})$	restriction de $A \in \mathbb{R}^{M \times N}$ aux colonnes indexées par $\omega \subseteq \mathbb{I}_N$,
$x_\omega = (x_{\omega[1]}, \ldots, x_{\omega[\sharp\omega]})$	restriction de $x \in \mathbb{R}^N$ aux entrées indexées par $\omega \subseteq \mathbb{I}_N$,
$x^{(i)} \in \mathbb{R}^N$	définit le vecteur $x^{(i)} := (x_1, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_N)$,
$\sigma(x) \subseteq \mathbb{I}_N$	support de $x \in \mathbb{R}^N$ défini par $\sigma(x) := \{i \in \mathbb{I}_N ; x_i \neq 0\}$,
$\sigma^-(x) \subseteq \sigma(x)$	sous ensemble du support de $x \in \mathbb{R}^N$ défini en (9.21) page 101,
$\sigma^+(x) \supseteq \sigma^-(x)$	ensemble défini en (9.22) page 101 (pas forcément inclus dans $\sigma(x)$),
$\sigma_n^- = \sigma^-(x^n)$	notation utilisée dans le macro-algo du chapitre 10,
(P1)	propriété $\arg\min_{x \in \mathbb{R}^N} \tilde{G}(x) = \arg\min_{x \in \mathbb{R}^N} G_{\ell_0}(x)$ utilisée dans le chapitre 12 (page 165),
(P2)	propriété ($\hat{x}$ minimiseur (local) de $\tilde{G} \implies \hat{x}$ minimiseur (local) de G_{ℓ_0}) utilisée dans le chapitre 12 (page 165),

CONTEXTE GÉNÉRAL DE LA THÈSE ET CONTRIBUTIONS

CONTEXTE GÉNÉRAL

Cette thèse s'articule autour de deux parties pouvant être lues de manière indépendante. La première est dédiée à un problème de reconstruction rencontré en microscopie par réflexion totale interne (TIRF en anglais). Ce travail s'inscrit dans le contexte d'un système TIRF expérimental développé à l'Institut de Biologie Valrose (iBV) (Nice) par Sébastien Schaub, pour lequel il est possible de faire varier rapidement l'angle d'incidence de la lumière d'excitation (grâce à une électronique dédiée) permettant ainsi d'obtenir des acquisitions dites *multi-angles* (MA-TIRF). Notons qu'il n'existe pas encore de tel système dans le commerce. Les acquisitions ainsi obtenues ont le potentiel, grâce au développement d'algorithmes de reconstructions dédiés, de révéler une information super-résolue (de l'ordre de quelques dizaines de nanomètres) concernant la position axiale des structures observées sur une fine couche ($\approx$ 500 nm) adjacente à la lamelle de verre. L'objectif premier des travaux réalisés dans ce contexte concerne donc la mise en place d'une méthode de reconstruction à partir de données acquises par le système MA-TIRF expérimental. Par ailleurs, une question fondamentale concerne l'étalonnage du système ainsi que la simplification du modèle décrivant la physique de l'acquisition et la validation de ce dernier sur des expériences réelles. Ce travail s'inscrit donc également dans une collaboration avec l'équipe de biologistes d'Ellen Van Obberghen-Schilling afin de se confronter aux difficultés rencontrées sur des données réelles.

Le problème de reconstruction MA-TIRF étant mal posé, il est nécessaire de le régulariser en ajoutant de l'information *a priori*. En s'intéressant à la reconstruction de molécules isolées et de la membrane cellulaire, dont l'épaisseur est de l'ordre de 10 nm, nous nous sommes orientés vers les modèles parcimonieux ce qui fut le point de départ de la deuxième partie de ce manuscrit. Dans cette deuxième partie, nous nous intéressons à la minimisation du critère (NP-Difficile) des moindres carrés pénalisé en norme-ℓ_0. Plus précisément, nos travaux rentrent dans le contexte des reformulations continues exactes du problème. L'objectif consiste en la définition de relaxations continues de la fonctionnelle initiale préservant ses minimiseurs globaux. Par ailleurs, nous posons également la question de la relation entre les minimiseurs locaux des deux fonctionnelles et recherchons des relaxations pour lesquelles de tout minimiseur local on peut définir un minimiseur local du critère initial par une opération simple (e. g. seuillage). L'idée d'une telle relaxation n'est pas nouvelle mais les résultats théoriques concernant la relation entre les minimiseurs des fonctionnelles initiale et relaxée n'étaient jusqu'ici que partiels. La deuxième partie de cette thèse est donc principalement consacrée à l'étude de telles relaxations pour lesquelles nous nous efforçons de fournir des résultats complets sur les liens entre les minimiseurs des deux fonctionnelles. Bien que toujours non-convexes, ces relaxations ont l'avantage d'être continues permettant l'utilisation d'un certain nombre d'algorithmes récemment proposés pour l'optimisation non-convexe. Ainsi, nous terminons ce travail par une étude des performances de quelques uns de ces algorithmes dans le cadre de différentes applications en traitement du signal et des images.

CONTRIBUTIONS

Reconstruction MA-TIRF

Suite à une introduction du problème et une revue des méthodes de reconstruction de la littérature dans les chapitres 1 et 2, les contributions concernant la reconstruction MA-TIRF sont présentées dans les chapitres suivants et peuvent être classées en deux catégories :

Étude du problème inverse et résolution numérique : en considérant une étude numérique en dimension 1, nous montrons dans le chapitre 3 que la régularisation du problème doit être choisie avec précaution.

En effet, bien qu'imposer la positivité de la solution soit complètement naturel pour ce type de problème inverse, nous montrons qu'une telle contrainte nécessite de prendre en compte précisément le signal de fond présent sur les images afin d'être en mesure de reconstruire correctement les objets imagés. Pour ce faire, nous proposons de réaliser une estimation jointe de la densité de fluorophores et du signal de fond, tous deux soumis à une contrainte de positivité.

D'autre part, l'atténuation du contraste provoquée par une régularisation de variation totale (due à la norme-ℓ_1) s'avère être problématique. En particulier, nous mettons en évidence que ce phénomène entraine une perte de résolution axiale (pour les objets reconstruits) ce qui est contraire à l'objectif recherché. Nous préconisons alors d'utiliser une telle régularisation ou bien de manière mesurée ou bien de ne l'appliquer que dans les directions latérales x et y.

Enfin, toujours dans le cadre d'une étude numérique en dimension 1, nous montrons que la qualité des reconstructions n'est pas influencée par la considération d'une vraisemblance gaussienne ou poissoniene lorsque les données sont dégradées par un bruit mixte poissonien-gaussien comme cela est généralement le cas en microscopie.

Étalonnage du système et validation du modèle : afin de contrôler avec précision l'angle d'incidence (qui est un paramètre du modèle) du laser d'excitation sur le spécimen, nous proposons une adaptation simplifiée d'une méthode de la littérature fondée sur l'observation du plan focal arrière de l'objectif (chapitre 4). Ainsi, nous sommes en mesure d'étalonner précisément le système allant même jusqu'à détecter des différences en fonction de la longueur d'onde d'excitation utilisée dues à l'achromaticité non parfaite des éléments optiques du système.

D'autre part, toujours dans le chapitre 4, nous montrons que nous sommes capable de reconstruire précisément un échantillon de géométrie connue,[2] à partir du modèle TIRF pour lequel nous avons fait des simplifications (chapitre 1). En particulier, l'objet imagé est correctement reconstruit, avec une précision de l'ordre de 20 nm, sur une épaisseur de 400 nm. Cela est à mettre en perspective avec d'autres expériences similaires publiées dans la littérature ne présentant pas de résultats au-delà de 200 nm. Cette expérience est confortée par une expérience de co-localisation à deux couleurs où nous proposons de marquer une même structure biologique avec deux molécules fluorescentes différentes. Nous montrons alors que les reconstructions obtenues avec le modèle simplifié co-localisent avec une précision de l'ordre de 20-40 nm.

Pour finir, des phénomènes biologiques connus ont pu être observés à l'aide de la méthode proposée (chapitre 5) ce qui conforte encore sa validité et ouvre d'intéressantes perspectives pour l'étude des interactions entre différentes molécules au niveau de la membrane cellulaire.

2. Construit simplement à l'aide d'une lentille de verre et d'une solution fluorescente homogène.

Relaxations continues exactes du critère ℓ_2-ℓ_0

La deuxième partie du document concerne la minimisation de la fonctionnelle :

$$G_{\ell_0}(x) := \frac{1}{2}\|Ax - d\|^2 + \lambda\|x\|_0,$$

où $A \in \mathbb{R}^{M\times N}$, $d \in \mathbb{R}^M$, $\lambda > 0$ et $\|x\|_0$ compte le nombre de composantes non-nulles de $x \in \mathbb{R}^N$. Une introduction du problème ainsi qu'un état de l'art sur l'optimisation parcimonieuse sont les objets respectivement des chapitres 7 et 8. La suite de la deuxième partie du manuscrit est quant à elle dédiée aux différentes contributions apportées dans ce contexte :

La pénalité CEL0 : dans le chapitre 9 nous proposons la pénalité $\Phi_{\text{CEL0}} : \mathbb{R}^N \to \mathbb{R}$ (Continuous Exact ℓ_0) pour laquelle la fonctionnelle continue relaxée non-convexe,

$$G_{\text{CEL0}}(x) := \frac{1}{2}\|Ax - d\|^2 + \Phi_{\text{CEL0}}(x),$$

vérifie les propriétés principales suivantes :
- $\arg\min_{x\in\mathbb{R}^N} G_{\ell_0}(x) \subseteq \arg\min_{x\in\mathbb{R}^N} G_{\text{CEL0}}(x)$;
- de tout minimiseur (local) de G_{CEL0} on peut déduire un minimiseur (local) de G_{ℓ_0} par un simple seuillage ;
- la réciproque de la propriété précédente n'est pas vraie et G_{CEL0} élimine des minimiseurs locaux (non-globaux) de G_{ℓ_0}.

Une étude complète entre les minimiseurs des fonctionnelles G_{CEL0} et G_{ℓ_0} est donc proposée et une caractérisation des minimiseurs de G_{CEL0} est également détaillée. Ces résultats montrent que la pénalité CEL0 est une alternative continue à la pseudo norme-ℓ_0 idéale. Notons qu'à notre connaissance c'est la première fois qu'une telle pénalité, avec toutes les propriétés mentionnées ci-dessus assurant une reformulation exacte du problème, est proposée.

Algorithmes pour la minimisation de G_{CEL0} : la continuité de la fonctionnelle G_{CEL0} nous donne accès à un éventail d'algorithmes récents d'optimisation non-convexes adaptés à sa minimisation et donc indirectement à celle de G_{ℓ_0}. À partir de tels algorithmes, nous proposons dans le chapitre 10 :
- un Macro-Algo assurant la convergence vers un point qui est à la fois point critique de G_{CEL0} et minimiseur local de G_{ℓ_0} ;
- une heuristique inspirée des méthodes Graduated Non Convexity permettant d'améliorer la minimisation de la fonctionnelle ;
- une règle permettant de déterminer un «chemin de régularisation» pour G_{CEL0} afin de s'affranchir du choix du paramètre λ.

Applications en traitement du signal et des images : les algorithmes du paragraphe précédent (ou de la littérature) sont utilisés dans le cadre de diverses applications :
- déconvolution de trains d'impulsions ;
- estimation de canal et des directions d'arrivées en traitement d'antennes ;
- reconstruction en microscopie de super-résolution PALM.

Pour chacune de ces applications, nous montrons dans le chapitre 11 l'intérêt de minimiser G_{CEL0} plutôt que de minimiser directement G_{ℓ_0}. En particulier, le fait que G_{CEL0} élimine des minimiseurs locaux (non-globaux) de G_{ℓ_0} rend les algorithmes minimisant cette fonctionnelle plus robustes aux minimiseurs locaux que les algorithmes traitant directement G_{ℓ_0}. Ainsi, nous montrons que minimiser G_{CEL0} permet d'obtenir de meilleurs résultats no-

tamment en terme de reconstruction du support de la solution dans les différents contextes d'applications considérés.

Une vue unifiée : nous proposons dans le chapitre 12 une vue unifiée des pénalités continues approchant la pseudo norme-ℓ_0, notées Φ, dans le contexte des relaxations continues exactes de G_{ℓ_0}. Pour ce faire, nous dérivons cinq conditions sur la pénalité Φ *nécessaires et suffisantes* pour que la fonctionnelle relaxée $\tilde{G}$ associée vérifie les mêmes propriétés que G_{CEL0}. Plus précisément, trois de ces conditions sont nécessaires et suffisantes pour que la fonctionnelle relaxée $\tilde{G}$ préserve les minimiseurs globaux de G_{ℓ_0} alors que l'ensemble des cinq conditions sont nécessaires est suffisantes pour préserver les minimiseurs globaux et ne pas ajouter de minimiseurs locaux.

Par ailleurs, nous montrons que la pénalité CEL0 n'est autre que la limite inférieure de la classe de pénalités définie par les cinq conditions précédentes et que c'est celle qui peut potentiellement éliminer le plus de minimiseurs locaux de G_{ℓ_0} (ce qui la rend d'autant plus intéressante).

Enfin, nous proposons une analyse de différentes pénalités de la littérature dans ce contexte de reformulations continues exactes. Nous montrons que pour certaines d'entre elles, il existe des valeurs pour les paramètres les définissant permettant de vérifier les conditions précédemment établies.

PUBLICATIONS

Les travaux réalisés dans cette thèse peuvent également être trouvés dans les publications et rapports suivant(e)s :

Articles de revue soumis et acceptés

- **Structured CEL0 Relaxation for ℓ_0 optimization in channel and DOA estimation**. *Soumis à IEEE Transaction on Signal Processing, 2016.*
  Adilson Chinatto, Emmanuel Soubies, Cynthia Junqueira, João M. T. Romano, Pascal Larzabal, Jean-Pierre Barbot and Laure Blanc-Féraud.
- **A unified view of exact continuous penalties for ℓ_2-ℓ_0 minimization**. *Soumis à SIAM Journal on Optimization, 2016.*
  Emmanuel Soubies, Laure Blanc-Féraud and Gilles Aubert.
- **A Continuous Exact ℓ_0 penalty (CEL0) for least squares regularized problem**. *SIAM Journal on Imaging Science 8-3, pp. 1574-1606, 2015 (Erratum : SIIMS 9-1 pp. 490-494).*
  Emmanuel Soubies, Laure Blanc-Féraud and Gilles Aubert.

Actes de conférences

- **A framework for Multi-Angle TIRF microscope calibration**. *International Symposium on Biomedical Imaging (ISBI), 2016.*
  Emmanuel Soubies, Sébastien Schaub, Agata Radwanska, Ellen Van Obberghen-Schilling, Laure Blanc-Féraud and Gilles Aubert.
- **L_0-optimization for Channel and DOA sparse estimation**. *International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2015.*

Adilson Chinatto, Emmanuel Soubies, Cynthia Junqueira, João M. T. Romano, Pascal Larzabal, Jean-Pierre Barbot and Laure Blanc-Féraud.

— **Seuillage CEL0 pour la minimisation ℓ_2-ℓ_0 : comparaisons avec IHT**. *GRETSI, 2015*.
Emmanuel Soubies, Laure Blanc-Féraud and Gilles Aubert.

— **CEL0 : a continuous alternative to ℓ_0 penalty**. *Signal Processing with Adaptive Sparse Structured Representations (SPARS), 2015*.
Emmanuel Soubies, Laure Blanc-Féraud and Gilles Aubert.

— **Sparse reconstruction from Multi-Angle Total Internal Reflection Fluorescence Microscopy**. *International Conference on Image Processing (ICIP), 2014*.
Emmanuel Soubies, Laure Blanc-Féraud, Sébastien Schaub and Gilles Aubert.

— **A 3D model with shape prior information for biological structures reconstruction using Multiple-Angle Total Internal Reflection Fluorescence Microscopy.** *International Symposium on Biomedical Imaging (ISBI), 2014*.
Emmanuel Soubies, Laure Blanc-Féraud, Sébastien Schaub and Gilles Aubert.

Première partie

MICROSCOPIE MA-TIRF : DE L'ÉTALONNAGE DU SYSTÈME À LA RECONSTRUCTION D'IMAGES SUPER-RÉSOLUES

AVANT-PROPOS

La microscopie de fluorescence est devenue aujourd'hui une technique majeure pour l'observation de structures biologiques. En effet, le marquage fluorescent utilisé dans ce contexte permet de cibler les structures d'intérêt et de s'affranchir de la visualisation du reste de l'échantillon. Cependant, les techniques conventionnelles (épifluorescence, confocale, biphotonique...) sont limitées en résolution par la diffraction. En microscopie classique à épifluorescence, cette limite est connue d'après les travaux de Ernst Abbe, réalisés au cours du XIX$^{\text{ème}}$ siècle, comme étant $\lambda/(2 \times \text{NA})$ où λ représente la longueur d'onde de la lumière d'excitation et NA l'ouverture numérique de l'objectif. En microscopie confocale, inventée par Minsky en 1955 [3], cette limite est légèrement inférieure permettant d'obtenir des images mieux résolues. Il a fallu attendre la fin du XX$^{\text{ème}}$ et le début du XXI$^{\text{ème}}$ siècle pour voir émerger de nouvelles modalités d'acquisition permettant de dépasser la limite de diffraction (STED (Hell et Wichmann, 1994), SIM (Gustafsson, 2000), STORM (Rust et al., 2006), PALM (Betzig et al., 2006) parmi d'autres) atteignant pour certaines une résolution latérale de l'ordre du nanomètre. Des méthodes 3D-PALM et 3D-STED ont également été développées, en utilisant des PSF spécifiques dont la forme dépend de la profondeur, permettant aussi une amélioration de la résolution axiale (Huang et al., 2008). Cependant, ces méthodes nécessitent d'une part des marquages particuliers qui peuvent être lourds à mettre en œuvre et d'autre part un nombre très important d'acquisitions pour réaliser la reconstruction super-résolue du volume, ce qui peut être limitant pour certaines applications *invivo*.

Dans cette partie du manuscrit, nous nous intéressons à la microscopie par réflexion totale interne (Total Internal Reflection Fluorescence (TIRF) en anglais) capable de limiter le champ d'excitation à une fine couche, inférieure à la longueur d'onde, adjacente à la lamelle de verre. Cette propriété de sélectivité axiale en fait une technique idéale pour la visualisation des processus cellulaires se produisant dans le voisinage de la membrane plasmique comme par exemple l'étude des échanges entre la cellule et le milieu extérieur. De nombreuses applications peuvent tirer bénéfice d'une telle modalité d'acquisition et nous renvoyons le lecteur vers la review de Axelrod (2001), auteur à l'origine de cette technique de microscopie (Axelrod, 1981), pour plus de détails sur ces applications (voir aussi (Axelrod, 2008)). Il est également à noter que cette technique ne requiert pas de préparation particulière de l'échantillon.

Le chapitre 1 commence par présenter cette technique de microscopie ainsi que le système avec lequel nous travaillerons dans la suite de cette thèse. Par ailleurs, le principe du Multi-Angle Total Internal Reflection Fluorescence (MA-TIRF) de même que le problème de reconstruction 3D associé, permettant une amélioration significative de la résolution axiale, sont également présentés dans ce chapitre. Un état de l'art sur les méthodes de reconstruction est effectué dans le chapitre 2 et le chapitre 3 présente les travaux que nous avons réalisés dans ce contexte. Un aspect crucial de la reconstruction MA-TIRF concerne l'étalonnage du système ainsi que la validation du modèle simplifié utilisé pour la résolution du problème. Le chapitre 4 est donc dédié à ces questions. Enfin, des applications en biologie sont traitées dans le chapitre 5.

3. Pour l'histoire, voir le mémoire (Minsky, 1988).

1

DU TIRF AU MULTI-ANGLE TIRF

Sommaire

1.1 Formation d'une image en microscopie TIRF

1.1.1 *Lois de Snell-Descartes et angle critique*

Nous commençons ce chapitre par quelques rappels d'optique classique, nécessaires à la compréhension du principe de la microscopie TIRF. D'après les lois de Willebrord Snell et René Descartes nous savons que lorsqu'un rayon incident rencontre une interface entre un milieu d'indice de réfraction n_i et une autre milieu d'indice de réfraction $n_t < n_i$, une partie de la lumière est réfléchie dans le milieu incident alors que l'autre partie est transmise (réfractée) dans le second milieu (figure 1 gauche). Plus précisément, nous avons les deux lois suivantes :

— *Loi de la réflexion* : $\alpha_r = \alpha$;
— *Loi de la réfraction* : $n_i \sin(\alpha) = n_t \sin(\alpha_t)$;

où α, α_r et α_t définissent respectivement les angles incident, réfléchi et transmis. Le cas limite, avant réflexion totale (figure 1 centre), est obtenu pour $\alpha_t = 90°$ et correspond à un angle incident égal à *l'angle critique* α_c vérifiant :

$$\sin(\alpha_c) = \frac{n_t}{n_i}. \tag{1.1}$$

1.1.2 *Principe de la microscopie TIRF*

Le principe du TIRF repose sur le phénomène de réflexion totale interne d'une source lumineuse sur une interface diélectrique comme décrit dans la section précédente. Ainsi, en TIRF, l'angle incident α de la lumière d'excitation vérifie $\alpha > \alpha_c$ afin d'être dans le régime de réflexion totale (figure 1 droite). Dans un tel régime, bien que toute la lumière incidente soit réfléchie vers le milieu inférieur (verre), il y a création d'une onde évanescente se

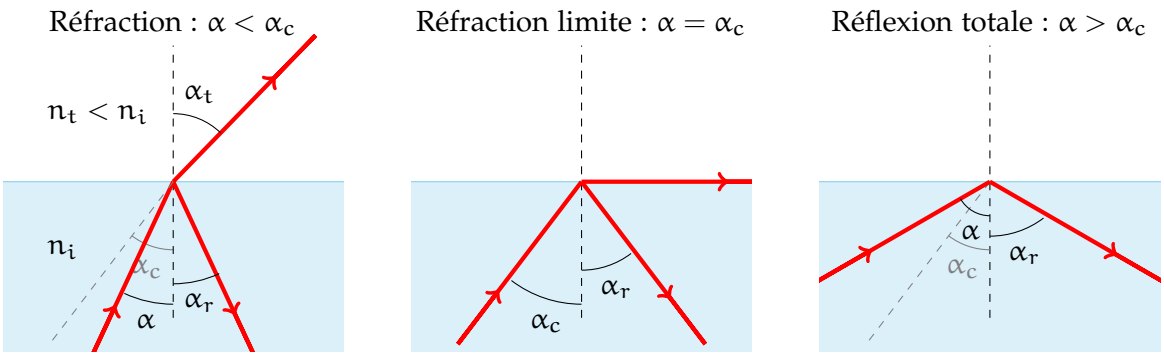


FIGURE 1 – Illustration des lois de Snell-Descartes [Schéma extrait et modifié depuis http://femto-physique.fr/optique/opt_C1.php]. Milieu incident en bleu et rayon lumineux en rouge.

propageant parallèlement au dioptre (AXELROD, 2008) permettant d'exciter les molécules fluorescentes (fluorophores) présentes dans une fine couche immédiatement adjacente à l'interface (voir figure 2).

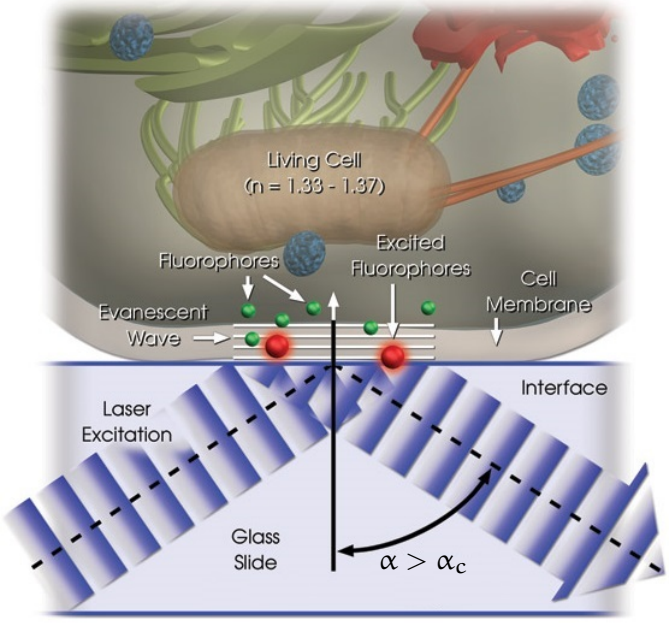


FIGURE 2 – Principe de la microscopie TIRF [Image extraite de http://www.photonics.com/Article.aspx?AID=33691].

Cette sélectivité axiale est due à la décroissance exponentielle de l'onde évanescente dans la direction (Oz), i. e. normale à l'interface. Plus précisément, l'onde évanescente décroît avec la profondeur z selon (AXELROD, 1981, 2001, 2008 ; MARTIN-FERNANDEZ et al., 2013)

$$I(z,\alpha) := I_0(\alpha)\exp\left(-zp(\alpha)\right), \tag{1.2}$$

où p est caractéristique de l'inverse de la profondeur de pénétration de l'onde évanescente :

$$p(\alpha) := \frac{4\pi n_i}{\lambda}\left(\sin^2(\alpha) - \sin^2(\alpha_c)\right), \tag{1.3}$$

avec λ la longueur d'onde de la lumière incidente. Concernant l'intensité à l'interface I_0 (i. e. pour $z = 0$), elle dépend de la polarisation du laser incident. Il existe deux types de polarisations : la polarisation dite «s» pour laquelle la direction du champ électrique est normale au plan d'incidence et la polarisation dite «p» pour laquelle la direction du champ électrique est contenue dans le plan d'incidence. À partir des lois de Snell-Descartes ainsi que des équations de Fresnel (BORN et WOLF, 2000), il est possible d'obtenir une expression analytique des intensités I_0^s et I_0^p :

$$I_0^s(\alpha) := \frac{4\cos^2(\alpha)}{(1-n^2)}, \tag{1.4}$$

$$I_0^p(\alpha) := \frac{4\cos^2(\alpha)(2\sin^2(\alpha)-n^2)}{n^4\cos^2(\alpha)+\sin^2(\alpha)-n^2}, \tag{1.5}$$

où $n = \frac{n_t}{n_i}$. Pour plus de détails, nous renvoyons le lecteur vers l'article très éclaircissant de MARTIN-FERNANDEZ et al. (2013). Enfin, l'intensité I_0 pondérant la décroissance (1.2) est obtenue par combinaison linéaire des intensités I_0^s et I_0^p dont les coefficients dépendent de la géométrie du système (voir paragraphe 1.1.3).

L'imagerie TIRF, de part sa sélectivité dans la direction axiale, permet donc de s'affranchir des plans «hors-champ» et les images résultantes ont ainsi peu de fluorescence de fond ainsi qu'un meilleur Rapport Signal sur Bruit, *Signal to Noise Ratio* en anglais (SNR), en comparaison avec d'autres techniques de microscopie (e. g. champ large, confocal...). À titre d'illustration, la figure 3 compare deux acquisitions d'un même échantillon : la première résultant de la microscopie à épifluorescence (champ large), intégrant l'échantillon sur toute sa profondeur, et la deuxième ayant été obtenue par microscopie TIRF limitant l'excitation à une fine couche adjacente à l'interface.

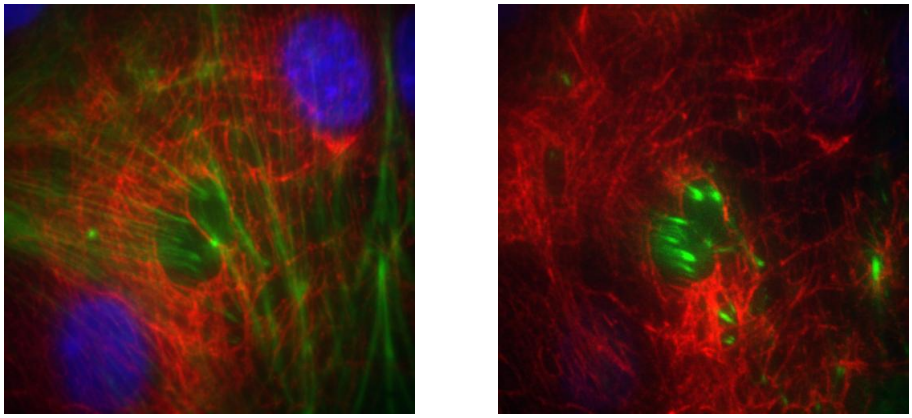


FIGURE 3 – Comparaison d'une acquisition par microscopie à épifluorescence (gauche) avec une acquisition du même échantillon en TIRF (droite). Différentes structures sont imagées : actine (vert), fibronectine (rouge) et le noyau (bleu). On voit clairement la « disparition » des structures les plus éloignées sur l'acquisition TIRF.

1.1.3 *Description physique du système*

Nous décrivons ici le système TIRF développé à l'iBV par Sébastien Schaub sur lequel nous travaillerons par la suite. Un schéma simplifié du système est présenté sur la figure 4.

L'élément central est le miroir galvanométrique (2) permettant de contrôler et de modifier rapidement l'angle d'incidence du rayon lumineux d'excitation.

Afin d'obtenir un rayon incident sur l'échantillon avec un angle α donné, le miroir (2) est positionné de manière à focaliser la source lumineuse en un point du plan focal arrière, *Back Focal Plane* en anglais (BFP), de l'objectif. En effet, la distance entre l'axe optique et un point de focalisation sur le BFP est caractéristique de l'angle avec lequel ce rayon incident arrive sur l'échantillon (voir chapitre 4).

L'émission des fluorophores générée par leur excitation est ensuite collectée par une caméra Electron Multiplying Charge-Coupled Device (EMCCD) focalisée sur le plan focal (avant) de l'objectif. Notons que sur le chemin optique de la figure 4, le miroir dichroïque (4) a la particularité de réfléchir la lumière incidente (bleue) et d'être «transparent» pour la lumière émise (verte). En ce qui concerne le miroir (8), il est amovible et permet, lorsqu'il est utilisé, de récupérer l'image du BFP de l'objectif avec une autre caméra focalisée sur ce dernier.

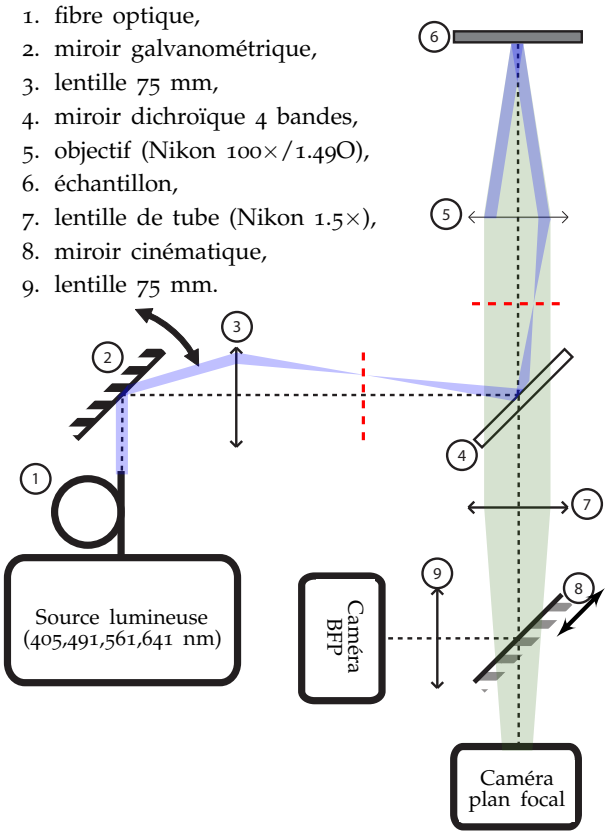


FIGURE 4 – Système TIRF. Le rayon d'excitation est représentée en bleu et l'émission des fluorophores en vert. Les tirets rouge représentent respectivement le BFP de la lentille (3) et de l'objectif (5).

Par ailleurs, notons qu'il existe deux types de montages TIRF (AXELROD, 2008) :

— montage avec prisme ;
— montage avec objectif.

Dans notre cas, nous utilisons un montage avec objectif, c'est-à-dire que l'objectif est utilisé pour dévier le faisceau incident afin de former l'angle d'illumination souhaité. L'objectif se trouve donc sous l'échantillon. Au contraire avec l'autre type de montage, la lumière incidente est déviée à l'aide d'un prisme et dans ce cas, l'objectif se trouve au dessus de l'échantillon. Le montage avec objectif impose une limite maximale sur l'angle incident que

l'on peut produire. Cette limite est dépendante de l'ouverture numérique de l'objectif NA [1] et est donnée par

$$\alpha_{max} := \sin^{-1}(NA/n_i). \tag{1.6}$$

Pour terminer la description du système, nous devons ajouter qu'une rotation azimutale du rayon incident est effectuée. Nous n'avons donc pas un point de focalisation du rayon incident sur le BFP de l'objectif, mais un ensemble de points décrivant un cercle (voir chapitre 4). Cette méthode a été développée par plusieurs auteurs (MATTHEYSES et al., 2006 ; FIOLKA et al., 2008a ; FIOLKA, 2009) dans le but d'homogénéiser le champ d'excitation et de réduire certains artefacts produits par la diffusion et la diffraction de la lumière cohérente (laser) par les éléments optiques qu'elle traverse. D'autre part, cette rotation «mélange» les polarisation s et p de la lumière incidente et l'intensité de l'onde évanescente à l'interface ($z = 0$) est alors donnée par (BOULANGER et al., 2014) :

$$I_0(\alpha) = \frac{3}{4} I_0^s(\alpha) + \frac{1}{4} I_0^p(\alpha), \tag{1.7}$$

où I_0^s et I_0^p sont respectivement définies par (1.4) et (1.5).

1.1.4 *Le modèle TIRF*

Nous sommes maintenant en mesure de définir le modèle décrivant la formation d'une image à travers le microscope TIRF. Notons $f : \Omega \times \mathbb{R}_+ \to \mathbb{R}_+$ la densité de fluorophores de l'échantillon 3D, où $\Omega \subset \mathbb{R}^2$ est le domaine d'observation 2D, i. e. correspondant aux variables latérales $(x, y) \in \Omega$, et $z \in \mathbb{R}_+$ est la variable axiale (de profondeur). Afin d'alléger les notations, on notera par la suite $u = (x, y)$. Pour un angle incident $\alpha \in]\alpha_c, \alpha_{max}]$ donné, nous obtenons l'acquisition échantillonnée $s(\alpha) \in \mathbb{R}^N$ définie par [2] : $\forall i \in \mathbb{I}_N$,

$$s(\alpha)_i := \iint_{u \in A_i} \left(Q_e \iint_{u' \in \Omega} h(u - u') \left(\int_{z=0}^{+\infty} Q(z) I(z, \alpha) f(u', z)\, dz \right) du' \right) du + b_i, \tag{1.8}$$

où $A_i \subset \Omega$ représente la région de Ω correspondant au i-ème pixel, h est la fonction d'étalement du point, *Point Spread Function* en anglais (PSF), qui est considérée 2D ici (i. e. constante en z), Q_e est l'efficacité quantique (*quantum efficiency* en anglais), $Q(z)$ représente l'efficacité de collection (*collection efficiency* en anglais), I modélise la décroissance de l'onde évanescente dans la direction axiale et est donnée par (1.2) et enfin b_i modélise un signal de fond (bruit de fond) présent au pixel $i \in \mathbb{I}_N$ et indépendant de l'angle incident α. Ce signal de fond est dû à des sources lumineuses parasites présentes dans l'environnement observé qui peuvent être la conséquence de la réflexion de la lumière, par exemple, ou bien de la diffusion de la fluorescence dans l'échantillon. Dans les paragraphes suivants, nous donnons des détails supplémentaires sur les termes composant le modèle (1.8).

EFFICACITÉ QUANTIQUE Q_e Lorsqu'ils sont excités par l'onde évanescente, les fluorophores émettent des photons qui sont ensuite collectés par une caméra EMCCD afin de les convertir en électrons puis d'encoder le signal résultant pour former une image. L'efficacité quantique Q_e (électrons/photons) représente le rapport entre le nombre de charges élec-

1. $NA = n \sin(\theta)$, où n est l'indice de l'huile dans laquelle l'objectif est immergé et θ et le plus grand angle que l'on puisse obtenir entre l'axe optique et un rayon entrant dans l'objectif depuis le foyer.

2. Ici $N \in \mathbb{N}^*$ représente le nombre de pixels de l'image acquise $s(\alpha)$. De plus, chaque pixel de cette image sera indexé par un seul indice $i \in \mathbb{I}_N = \{1, \ldots, N\}$.

triques générées par le capteur EMCCD (avant l'étape d'amplification d'électrons utilisée dans ce type de capteurs) et le nombre de photons incidents. En notant respectivement N_e et N_p le nombre d'électrons générés et le nombre de photons incidents pour un pixel, on définit :

$$Q_e := \frac{N_e}{N_p}. \tag{1.9}$$

L'efficacité quantique n'affecte donc les images résultantes que par un facteur multiplicatif.

EFFICACITÉ DE COLLECTION Q Au vu de la décroissance exponentielle de l'intensité de l'onde évanescente (excitation) donnée par l'équation (1.2), il est naturel de considérer que l'émission d'une molécule fluorescente excitée suit également une décroissance exponentielle avec la profondeur z. Cependant, cela n'est pas exactement correct. En effet, l'émission d'un fluorophore à proximité d'une interface diélectrique est perturbée par cette dernière (HELLEN et AXELROD, 1987). Une description qualitative de ce phénomène est également donnée par AXELROD (2008).

Le modèle d'émission d'un fluorophore est donc dépendant de sa distance à l'interface d'une manière assez complexe. Cette dépendance peut être modélisée par un facteur $Q(z)$ dont le rôle principal est de «déformer» le profil exponentiel lorsque z est proche de 0. Notons qu'il n'existe pas d'expression analytique pour $Q(z)$ qui dépend aussi d'autres facteurs comme l'ouverture numérique de l'objectif, la polarisation de l'excitation, ou encore l'orientation des fluorophores (HELLEN et AXELROD, 1987 ; AXELROD, 2008). Certains auteurs (ROHRBACH, 2000 ; MATTHEYSES et AXELROD, 2006 ; ÖLVECZKY et al., 1997) ont proposé des méthodes permettant de mesurer le profil de décroissance de l'onde évanescente prenant ainsi en compte l'efficacité de collection $Q(z)$.

FONCTION D'ÉTALEMENT DU POINT Étant donné l'épaisseur très fine que l'on observe (entre 100 et 500 nm), la PSF du système peut être considérée constante en z. Ainsi, son effet sur les acquisitions peut être modélisé via une convolution 2D comme cela est le cas dans le modèle (1.8). Avec un tel modèle, h est alors une fonction 2D représentant la tâche de diffraction (ou tâche d'Airy) qui dépend de la longueur d'onde de la lumière incidente ainsi que de l'ouverture numérique de l'objectif comme cela est présenté par la figure 5. Une formulation théorique de la PSF peut être trouvée dans (BORN et WOLF, 2000), et dans certain cas, le lobe central peut être bien approché par une gaussienne (ZHANG et al., 2007).

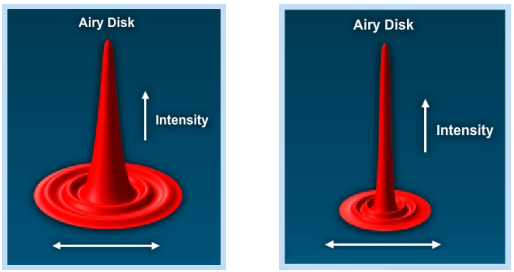


FIGURE 5 – Tâche d'Airy pour deux valeurs différentes de l'ouverture numérique de l'objectif : NA = 0.32 (gauche) et NA = 1.16 (droite). Ces images ont été réalisées avec les illustrations interactives des tutoriels Zeiss [http://www.zeiss.com/microscopy/en_de/solutions/reference/all-tutorials/].

NOTE SUR LES MODÈLES MULTI-INDICES Lors d'acquisitions *in vivo*, la simple interface entre la lamelle de verre et la solution aqueuse contenant l'échantillon considérée précédemment n'est pas complètement réaliste. En effet, un modèle plus complet devrait prendre en compte différentes interfaces séparant les différentes couches de l'échantillon (e. g. verre, eau, membrane, cytoplasme...), chacune ayant un indice de réfraction différent. Un tel modèle, considérant 4 couches, a été proposé par GINGELL et al. (1987). Cependant ce modèle est très complexe et nous garderons, dans la suite, l'hypothèse d'une unique interface entre deux milieux d'indices différents. Notons que ÖLVECZKY et al. (1997) ont étudié en simulations l'influence d'une telle simplification. Leurs conclusions indiquent qu'un léger biais sur l'estimation de la profondeur des structures biologiques est introduit par le modèle simplifié et qu'il peut être réduit en augmentant l'indice de la solution aqueuse de sorte à le rapprocher de celui du cytoplasme et de la membrane cellulaire. Aussi, BOULANGER et al. (2014) ont proposé un modèle où l'indice du milieu observé, bien que considéré constant selon z, peut varier dans le plan Ω. Des exemples numériques sur des données réelles montrent alors qu'estimer à la fois l'indice variable et la densité de fluorophores permet une meilleure localisation axiale des structures en comparaison à la seule estimation de la densité de fluorophores pour un indice fixé.

SIMPLIFICATION DU MODÈLE Dans la suite de ce manuscrit nous utiliserons une version simplifiée de (1.8) faisant abstraction de l'efficacité quantique $\mathcal{Q}_e$ (facteur multiplicatif), de l'efficacité de collection $Q(z)$ et de l'effet de la PSF. Ce modèle simplifié s'écrit : $\forall i \in \mathbb{I}_N$,

$$s(\alpha)_i := \iint_{u \in A_i} \left(I_0(\alpha) \int_{z=0}^{+\infty} \exp(-zp(\alpha)) f(u,z)\, dz \right) du + b_i. \tag{1.10}$$

Notons qu'avec un tel modèle, f représente la densité de fluorophores de l'échantillon convoluée par la PSF. Nous verrons au chapitre 4 qu'une telle simplification du modèle reste tout a fait représentative du système utilisé.

1.2 VARIER L'ANGLE D'INCIDENCE : UN ACCÈS À UNE INFORMATION TRIDIMENSIONNELLE

1.2.1 *Variation de la décroissance de l'intensité de l'onde évanescente avec l'angle d'incidence*

Nous nous intéressons maintenant aux acquisitions dites MA-TIRF. Considérons un ensemble de $L \in \mathbb{N}^\star$ angles incidents :

$$\mathcal{A} := \{\alpha_l, l \in \mathbb{I}_L : \alpha_c < \alpha_1 < \cdots < \alpha_L \leqslant \alpha_{max}\}. \tag{1.11}$$

Une acquisition MA-TIRF est alors définie comme étant l'ensemble des acquisitions TIRF obtenues pour les angles incidents appartenant à $\mathcal{A}$. On note :

$$S := (s(\alpha_1), \ldots, s(\alpha_L)) \in \mathbb{R}^{N \times L}, \tag{1.12}$$

où $s(\alpha)$ est défini par le modèle (1.8).

La profondeur de pénétration de l'onde évanescente étant directement liée à l'angle incident par la relation (1.3), la variation de ce dernier entraîne une modification de la décroissance du champ d'excitation. En particulier, plus l'angle incident est grand, plus la décroissance de l'onde évanescente est rapide comme cela est illustré sur la figure 6. Il

en résulte que les fluorophores les plus distants de l'interface ne sont plus excités pour des angles incidents éloignés de l'angle critique. Cela suggère qu'à partir d'acquisitions MA-TIRF, il est possible d'extraire une information tridimensionnelle sur les structures biologiques observées avec le potentiel de repousser les limites de résolution (axiale) rencontrées en microscopie conventionnelle.

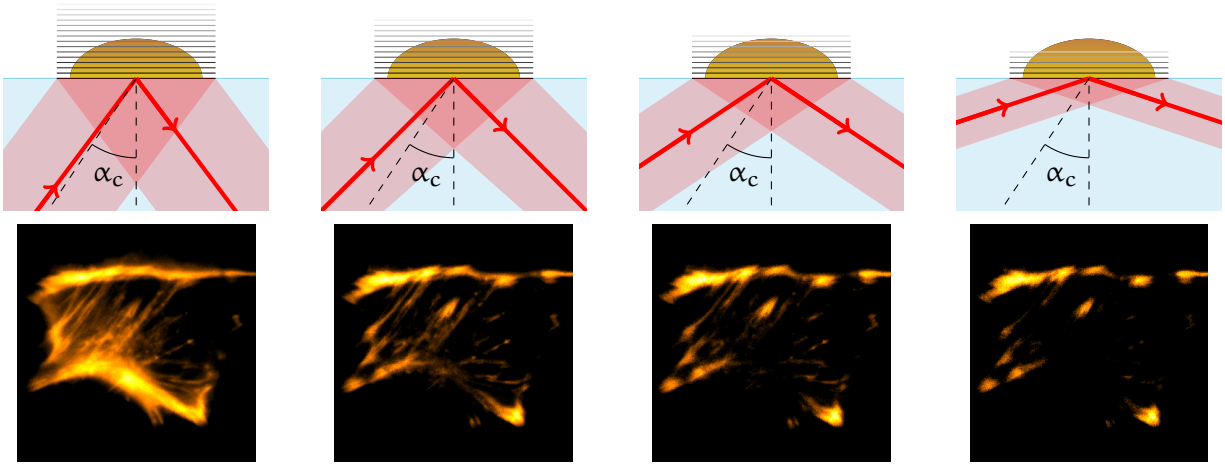


FIGURE 6 – MA-TIRF. La ligne du haut schématise l'évolution de l'onde évanescente lorsque l'angle incident augmente. La ligne du bas montre des exemples d'acquisitions d'un échantillon pour différents angles d'incidence évoluant comme sur la ligne supérieure.

À titre d'exemple, alors que la résolution axiale est d'environ 500 nm pour un microscope confocal ou de l'ordre de 250 nm pour un microscope à feuille de lumière (Selective Plane Illumination Microscopy), une simple acquisition TIRF pour un angle incident proche de l'angle maximal α_{max} est capable de limiter l'excitation sur une épaisseur de l'ordre de 100 nm au niveau de l'interface. Une reconstruction à partir d'acquisitions MA-TIRF peut quant à elle fournir des images ayant une résolution axiale de quelques dizaines de nanomètres sur une épaisseur allant de 500 à 800 nm (BOULANGER et al., 2014 ; DOS SANTOS et al., 2014). Cependant, ces approches ne permettent qu'une amélioration de la résolution axiale et la résolution latérale reste identique à celle des microscopes conventionnels. Notons tout de même qu'il est envisageable de combiner le MA-TIRF avec des techniques de microscopie permettant une amélioration de la résolution latérale. Par exemple FIOLKA et al. (2008b) combinent le TIRF avec la Harmonic Excitation Light Microscopy (HELM) qui est une technique utilisant une illumination structurée pour améliorer la résolution latérale des images (FROHN et al., 2000).

1.2.2 *Un problème inverse mal posé*

Afin d'atteindre une telle résolution axiale, tout l'enjeu est de résoudre le problème inverse. C'est-à-dire, estimer la position axiale des structures biologiques d'intérêt (inaccessible par une mesure directe) à partir d'une acquisition MA-TIRF constituée, comme nous venons de le voir, d'un ensemble d'acquisitions TIRF pour différents angles d'incidence. Tout d'abord, un premier travail consiste à étudier si le problème est *bien posé*. Un problème est dit *bien posé* au sens de HADAMARD (1902) si il vérifie les trois propriétés suivantes :

1. une solution existe ;
2. la solution est unique ;
3. la solution est continue par rapport aux données.

Pour étudier brièvement le caractère mal-posé de notre problème de reconstruction MA-TIRF, commençons par remarquer son lien avec la transformée de Laplace. En effet, en ignorant l'intégration sur A_i, le modèle simplifié (1.10) n'est autre qu'une transformée Laplace pondérée par le facteur I_0. Pour rappel, la transformée de Laplace d'une fonction f réelle continue définie sur $\mathbb{R}_+$ est donnée par :

$$\mathcal{L}f(p) := \int_0^{+\infty} f(x)\exp(-px)dx, \tag{1.13}$$

où $p \in \mathbb{C}$ est une variable complexe[3]. Étant donné une fonction $F(p)$, $p \in \mathbb{C}$, vérifiant certaines conditions (Bellman et Roth, 1984 ; Schiff, 2013), une formule d'inversion de la transformée de Laplace existe permettant de déterminer une fonction f telle que $\mathcal{L}f = F$. Par ailleurs, le théorème de Lerch (1903) nous assure que si deux fonctions f_1 et f_2 sont telles que $\mathcal{L}f_1 = \mathcal{L}f_2$, alors f_1 et f_2 sont égales sauf au plus en des points isolés. Ainsi, en se restreignant aux fonctions continues, l'inverse de la transformée de Laplace est uniquement définie.

Cependant pour notre problème, nous avons accès qu'à un nombre fini de mesures d'une fonction $F(p)$ pour des valeurs de $p \in [p_0, +\infty)$ ($p_0 \in \mathbb{R}_+$). L'inversion numérique à partir de telles données ne peut donc pas être réalisée à partir de méthodes numériques basées sur la formule d'inversion de la transformée de Laplace qui nécessite de connaître $F(p)$ sur une droite du plan complexe parallèle à l'axe imaginaire (Schiff, 2013, chapitre 4). Par ailleurs, le problème d'inversion de la transformée de Laplace à partir de valeurs sur la droite réelle[4] est connu pour être difficile et très instable (Epstein et Schotland, 2008).

Prenons l'exemple proposé par Bellman et Roth (1984) où ils considèrent la fonction $\varepsilon_\omega(t) = \sin(\omega t)$ dont la transformée de Laplace est donnée par

$$\mathcal{L}\varepsilon_\omega(p) = \frac{\omega}{p^2 + \omega^2}. \tag{1.14}$$

Considérons une suite $(\omega_n)_{n\in\mathbb{N}}$ telle que $\lim_{n\to+\infty} \omega_n = +\infty$. Alors on a,

$$\lim_{n\to+\infty} \mathcal{L}\varepsilon_{\omega_n} = 0, \tag{1.15}$$

ce qui montre que la condition de continuité de la solution par rapport aux données n'est pas vérifiée. En effet, nous avons

$$\lim_{n\to+\infty} \mathcal{L}(f + \varepsilon_{\omega_n}) = \mathcal{L}f, \tag{1.16}$$

mais

$$\lim_{n\to+\infty} f + \varepsilon_{\omega_n} \neq f. \tag{1.17}$$

Une autre façon de voir le caractère mal-posé du problème inverse MA-TIRF est de s'intéresser à sa version discrète qui est finalement celle que nous allons réellement utiliser en pratique. En considérant le cas unidimensionnel où $f \in \mathbb{R}^M$ est un vecteur représentant une

3. Notons que cette intégrale peut ne pas exister. En particulier, des conditions suffisante pour qu'elle existe sont que f soit intégrable sur $[0, T]$ pour tout $T \in \mathbb{R}_+$ et qu'elle vérifie $|f(x)| \leqslant ae^{bt}$ pour certaines constantes a et b. Sous de telles conditions, la transformée de Laplace existe pour tout p appartenant au demi plan complexe $\mathfrak{Re}(b) > 0$ (Bellman et Roth, 1984 ; Schiff, 2013).

4. Problème qui se pose généralement pour les phénomènes physiques réels.

discrétisation de f selon l'axe z pour une certaine position $u \in \Omega$, le modèle linéaire (1.10) se réécrit :

$$s = Hf + b, \tag{1.18}$$

où $s \in \mathbb{R}^L$ est un vecteur contenant les mesures pour les L angles incidents, $b \in \mathbb{R}$ est une constante modélisant le signal de fond et $H \in \mathbb{R}^{L\times M}$ est la matrice de l'opérateur discrétisé. Clairement, le problème inverse consistant à retrouver f à partir des données s admet une unique solution si et seulement si H est inversible. Cela impose d'une part d'avoir $L = M$, et d'autre part que $\mathrm{Ker}(H) = \{0_{\mathbb{R}^M}\}$. Généralement, l'une ou l'autre de ces deux conditions n'est pas vérifiée pour des configurations (i. e. L, M choix des discrétisations...) intéressantes en pratique. De plus, lorsque l'on se place dans un cadre favorable pour assurer l'inversibilité de H, son conditionnement reste très mauvais et le problème mal-posé.

Estimer la position axiale de structures biologiques à partir d'acquisitions MA-TIRF est donc un problème inverse mal posé et il sera nécessaire de le régulariser afin de le résoudre numériquement. Ces points seront abordés dans le chapitre 3 où nous verrons en particulier qu'imposer la positivité de la solution, qui est une contrainte complètement justifiée par la physique du problème, se révèle être très importante.

1.3 LES DIFFÉRENTS TYPES DE BRUITS

Nous nous intéressons dans cette section aux différents types de bruits venant dégrader la qualité des images acquises. Nous distinguerons deux types de bruit : ceux qui sont intrinsèques au signal reçu et ceux provenant du capteur.

1.3.1 *Bruit intrinsèque au signal reçu : le bruit de photons*

Le *bruit de photons*[5] (ou encore *bruit de grenaille* ou *bruit quantique*) est dû à la nature quantique de la lumière. Son origine vient du fait que les fluorophores émettent aléatoirement des photons, ce qui fait qu'entre t et $t + \tau_e$ (où τ_e représente le temps d'exposition), un nombre variable de photons ont été émis, et donc également un nombre variable de photons ont été collectés par le capteur. Ainsi l'amplitude du signal électrique généré dans le capteur (en chaque pixel) fluctue selon une statistique de Poisson dont la variance est donnée par

$$\sigma^2_{\mathrm{pht}} = \mathcal{Q}_e \Phi_{\mathrm{pht}} \tau_e, \tag{1.19}$$

où Φ_{pht} (photons/second/pixel) représente le flux de photons incident et $\mathcal{Q}_e$ l'efficacité quantique (1.9). Ce bruit est d'autant plus perceptible que l'émission de photons par les fluorophores excités est faible.

1.3.2 *Bruits émanant de la caméra*

En plus du bruit de photons, le capteur lui même introduit d'autres types de bruits (Robbins et Hadwen, 2003 ; Dussault et Hoess, 2004 ; Fellers et Davidson, 2004). Commençons tout d'abord par rappeler le fonctionnement des capteurs EMCCD. La toute première fonction d'un tel capteur consiste en la création d'une charge électrique (électron) en réac-

5. Shot noise en Anglais.

tion à une illumination (photon). Cette étape est réalisée par effet photoélectrique[6] au sein d'un semi-conducteur. La surface semi-conductrice du capteur est organisée sous forme de matrice où chaque élément correspond à un pixel. Chaque pixel est ainsi constitué d'un puits capable de collecter les électrons générés par effet photoélectrique lors de l'exposition. La quantité d'électrons retenue dans un puits est ainsi proportionnelle à la quantité de lumière reçue par ce pixel. Ensuite, une fois l'exposition terminée, les paquets d'électrons sont transférés lignes par lignes puis colonnes par colonnes vers un convertisseur en tension. Notons que les caméras EMCCD possèdent un amplificateur d'électrons placé avant le convertisseur de sortie. Enfin, la tension résultante est alors quantifiée afin d'encoder numériquement l'image observée.

BRUIT GÉNÉRÉ PAR LE COURANT D'OBSCURITÉ Le *courant d'obscurité*[7] (ou encore *bruit thermique*) est un phénomène dû à l'agitation naturelle des électrons. Cette agitation naturelle est suffisante pour éjecter occasionnellement un électron de la bande de valence vers la bande de conduction sans effet photoélectrique. Il y a donc création de charges supplémentaires qui viennent perturber le signal. Tout comme pour l'arrivée des photons sur le détecteur, le nombre d'électrons générés par agitation thermique suit une distribution de Poisson dont la variance est donnée par

$$\sigma_{obs}^2 = D_c \tau_e, \tag{1.20}$$

où D_c (électrons/secondes/pixels) représente le courant d'obscurité. Une expression de D_c fonction (entre autres) de la température peut être trouvée dans (DUSSAULT et HOESS, 2004). L'agitation des électrons augmentant avec la température, il est possible de réduire nettement ce bruit par refroidissement. Les cameras EMCCD actuelles sont généralement suffisamment refroidies ($\approx -70°$) pour que ce phénomène soit négligeable.

BRUIT D'AMPLIFICATION ET DE LECTURE Ce bruit est produit par le circuit électronique réalisant l'amplification et la conversion des paquets d'électrons en tension. C'est une combinaison de plusieurs bruits. En particulier nous pouvons dissocier :

— le *bruit de Johnson-Nyquist* (ou bruit thermique) produit par l'agitation thermique des électrons au sein d'une résistance de l'amplificateur de sortie ;
— le *bruit de reset* (reset noise ou encore KTC noise), lié aux fluctuations de la tension de référence d'une capacité impliquée dans la conversion charge/tension ;
— le *bruit de filcker* (ou bruit $1/f$) qui est inversement proportionnel à la fréquence de lecture des pixels ;
— le *bruit de quantification*.

Nous renvoyons le lecteur à (FELLERS et DAVIDSON, 2004) pour plus de détails. Généralement ce bruit de lecture est considéré gaussien et on notera sa variance σ_{lec}^2.

6. Passage d'un électron de la bande de valence à la bande de conduction lorsqu'on lui fournit une énergie suffisante. La particularité des semi-conducteurs, comparés aux isolants, est que le «gap» entre ces deux bandes est suffisamment faible pour permettre d'arracher un électron à la bande de valence et de créer une paire électron-trou.
7. Dark noise en Anglais.

1.3.3 *Quel(s) bruit(s) considérer dans le modèle ?*

D'après ce qui précède, et en incluant quelques simplifications[8], le bruit présent sur une acquisition peut être modélisé par un bruit mixte poissonien-gaussien :

$$\mathcal{P}(\sigma^2_{\mathtt{pht}}) + \mathcal{N}(0, \sigma^2_{\mathtt{lec}}), \tag{1.21}$$

où $\sigma^2_{\mathtt{pht}}$, la quantité d'électrons générée en réponse aux photons incidents, est dans le cas du TIRF donnée par le modèle (1.8). Ces bruits seront à considérer pour la résolution numérique du problème inverse qui sera présentée dans le chapitre 3.

8. Notamment en supposant la caméra suffisamment refroidie pour négliger le courant d'obscurité qui, d'après Zhang et Chen (2009), représente le bruit le plus important apporté par la caméra en dehors du bruit de lecture.

2

ÉTAT DE L'ART SUR LA RECONSTRUCTION MA-TIRF

Sommaire

Dans ce chapitre, nous nous intéressons aux différentes méthodes numériques de la littérature dédiées à la reconstruction MA-TIRF. Nous en distinguons plusieurs catégories. D'une part celles basées sur un *a priori* de forme des structures biologiques observées. Dans ce contexte, l'objectif consiste alors en l'estimation des paramètres d'un certain modèle de forme défini à partir des connaissances disponibles sur l'échantillon observé. D'autre part, nous verrons aussi que certains auteurs se sont penchés sur des méthodes variationnelles avec régularisation afin d'aborder le problème inverse. Enfin, nous présenterons des travaux récents où les auteurs proposent une méthode originale d'acquisition permettant de s'affranchir du problème de reconstruction.

2.1 Les méthodes basées sur un *a priori* de forme

2.1.1 *Modèles unidimensionnels*

Déterminer la profondeur de la membrane cellulaire est sans doute une des premières applications à avoir été étudiée à partir d'acquisitions TIRF/MA-TIRF. Une idée assez naturelle pour résoudre ce problème est basée sur les trois points suivant :

— définir un modèle paramétrique simple pour l'inconnue $f(u,\cdot)$, où au moins l'un des paramètres est caractéristique de la profondeur de la membrane pour la position latérale $u \in \Omega$;

— réaliser l'intégration TIRF simplifiée pour le modèle $f(u,\cdot)$ précédent :

$$s(\alpha,u) = I_0(\alpha)\int_0^{+\infty} \exp(-zp(\alpha))f(u,z)dz, \tag{2.1}$$

afin d'obtenir une expression exacte (fonction des paramètres introduits précédemment) des observations TIRF pour ce modèle ;

— estimer, pour chaque position discrète $u_i \in \Omega$, $i \in \mathbb{I}_N$, les paramètres du modèle par ajustement aux données TIRF/MA-TIRF $s(\alpha,u_i)$ (mesurées pour différents $\alpha > \alpha_c$).

Modèles proposés Trois modèles pour $f(u,\cdot)$ reviennent dans plusieurs travaux de la littérature. En notant respectivement $C_u \in \mathbb{R}_+$ et $z_u \in \mathbb{R}_+$ la concentration en fluorophores et la position axiale de la membrane pour la position latérale $u \in \Omega$, ces modèles sont les suivants :

— *le modèle Dirac*, utilisé dans le cas où la membrane contient le marquage fluorescent, est défini par (Reichert et Truskey, 1990 ; Truskey et al., 1992 ; Burmeister et al., 1994 ; Ölveczky et al., 1997 ; Dos Santos et al., 2014 ; Dos Santos et al., 2016) :

$$f(u,z) = C_u \delta_0(z - z_u). \tag{2.2}$$

Pour ce modèle, l'équation (2.1) devient :

$$s(\alpha,u) = I_0(\alpha) C_u \exp(-z_u p(\alpha)). \tag{2.3}$$

— *le modèle Top-Hat*, adapté pour la configuration où le marquage est réalisé sur la solution englobant la cellule, est donné par (Ölveczky et al., 1997) :

$$f(u,z) = C_u \mathbb{1}_{\{z \leqslant z_u\}}, \tag{2.4}$$

et conduit à

$$s(\alpha,u) = I_0(\alpha) \frac{C_u}{p(\alpha)} \Big(1 - \exp(-z_u p(\alpha))\Big). \tag{2.5}$$

— *le modèle shifted step function*, proposé pour le cas où les marqueurs fluorescents sont injectés dans le cytoplasme, s'exprime (Ölveczky et al., 1997) :

$$f(u,z) = C_u \mathbb{1}_{\{z \geqslant z_u\}}. \tag{2.6}$$

Pour ce dernier modèle, (2.1) devient

$$s(\alpha,u) = I_0(\alpha) \frac{C_u}{p(\alpha)} \exp(-z_u p(\alpha)). \tag{2.7}$$

ESTIMATION DES PARAMÈTRES Là encore, plusieurs méthodes d'estimation des paramètres C_u et z_u ont été utilisées par différents auteurs. La plus intuitive est certainement l'ajustement par moindres carrés des modèles (2.3), (2.5) et (2.7) aux données mesurées $s(\alpha, u_i)$. Ölveczky et al. (1997) réalisent cela avec l'algorithme de Levemberg-Marquardt (Marquardt, 1963).

Une autre méthode, utilisée par Dos Santos et al. (2014), Dos Santos et al. (2016) et Safarian et Kirchhausen (2008) consiste à normaliser une acquisition TIRF avec une image du même échantillon acquise par épifluorescence. Ce ratio a la particularité d'éliminer l'inconnue C_u du problème. Ainsi, seule la dépendance en l'inconnue z_u est conservée et une expression directe de cette dernière est dérivée. Notons que cette méthode nécessite un pré-traitement dû à la dégradation de l'image acquise en épifluorescence par un signal hors champ. Similairement, Stabley et al. (2015) utilisent le ratio entre deux acquisitions TIRF obtenues pour un même angle incident mais avec des longueurs d'ondes d'excitation différentes afin d'estimer la position axiale de molécules possédant un double marquage.

La principale limitation de ce type d'approches vient d'une part de la forte contrainte de forme imposée qui doit rester une hypothèse physiquement acceptable et d'autre part du fait que le bruit présent sur les images n'est pas pris en compte (en général les auteurs réalisent un filtrage des données avant la reconstruction pour diminuer le bruit mais cela a l'inconvénient de modifier également les acquisitions).

2.1.2 *Modèles 3D pour les vésicules de sécrétion*

Des méthodes très similaires à celles présentées dans le paragraphe précédent ont été proposées par ROHRBACH (2000) ou encore LOERKE et al. (2002) afin d'estimer la position et le diamètre de vésicules de sécrétion. Les auteurs modélisent de telles structures biologiques par des volumes sphériques (ROHRBACH, 2000) ou cubiques (LOERKE et al., 2002) et réalisent ensuite l'estimation (moindres carrés) des paramètres définissant ces modèles à partir d'acquisitions MA-TIRF. Notons que de telles méthodes ne fonctionnent que pour des vésicules isolées préalablement détectées sur l'une des acquisitions MA-TIRF.

Dans (SOUBIES et al., 2014a), nous avons également utilisé un modèle de sphères pour représenter des vésicules de sécrétion. Dans un contexte bayésien, nous avions formulé le problème comme la minimisation d'un terme d'attache aux données, prenant en considération la statistique de Poisson du bruit de photons, plus un terme *a priori* imposant en particulier une contrainte de non-recouvrement des objets. Ensuite, étant donné que les inconnues d'un tel problème sont à la fois le nombre d'objets et leurs paramètres, l'estimation était réalisée par Processus Ponctuels Marqués (DESCOMBES, 2011 ; VAN LIESHOUT, 2000).

Enfin, LIANG et al. (2012) ont proposé une méthode bayésienne pour l'estimation des caractéristiques (position axiale, rayon, intensité...) de particules subcellulaires. Considérant des particules de l'ordre de grandeur de la résolution latérale des pixels des images MA-TIRF, et prenant en considération l'effet de la PSF, les auteurs proposent de modéliser la distribution spatiale des fluorophores dans la plan Ω d'une image du stack MA-TIRF par une somme de gaussiennes dont les paramètres (positions, variance, intensité...) sont reliées aux caractéristiques des particules observées.

Après une détection des particules sur la moyenne du stack MA-TIRF (afin de réduire le bruit), les paramètres des gaussiennes sont estimés par Maximum A Posteriori (MAP). Le problème de maximisation résultant étant trop compliqué pour le résoudre directement, les auteurs proposent une alternative consistant en l'estimation successive des différents paramètres. Nous renvoyons le lecteur vers (LIANG et al., 2012) pour plus de détails. Enfin, bien que les rayons des particules dans le plan Ω soient estimés, ces dernières sont considérées ponctuelles (diracs) selon la direction axiale.

2.1.3 *Structures curvilignes*

Dans sa thèse, YANG (2010) (voir aussi (YANG et al., 2011)) s'intéresse à la reconstruction 3D de microtubules[1] à partir de données MA-TIRF. La méthode développée exploite la structure curviligne des microtubules et fonctionne en deux étapes :

1. segmentation des microtubules (dans Ω) par une méthode de plus court chemin (COHEN et KIMMEL, 1997) en ayant la connaissance des deux extrémités de ces derniers (indiquées manuellement) ;
2. estimation de la position axiale le long de chaque microbutule segmenté en excluant les points où deux microtubules se croisent afin d'éviter toute ambiguïté. Cette estimation est réalisée par MAP en considérant une statistique de bruit poissonienne prenant ainsi en compte le bruit de photons. Un a priori géométrique est également

1. Petits cylindres creux ($\simeq$ 25 nm) qui ont deux fonctions principales au sein de la cellule. D'une part ils permettent de transporter les vésicules ainsi que d'autres composants vers la membrane ou vers le corps cellulaire. D'autre part ils sont également impliqués dans le mécanisme de division cellulaire (mitose) où ils jouent un rôle très important en permettant le déplacement des chromosomes.

utilisé permettant de contrôler «l'élasticité» et la «rigidité» des microtubules. Enfin, l'optimisation du critère résultant est effectuée par un algorithme de descente de gradient.

Soulignons que Yang (2010) ne considère pas le modèle théorique (1.2) mais propose d'estimer le profil de décroissance de l'onde évanescente à partir d'une grande bille marquée sur sa surface par des molécules fluorescentes s'inspirant de la méthode proposée par Mattheyses et Axelrod (2006). Aussi, une méthode d'estimation de l'intensité à l'interface I_0 (donnée théoriquement par (1.4) et (1.5) suivant la polarisation) est proposée à partir d'une acquisition de petites billes fluorescentes placées à l'interface entre la lamelle de verre et l'échantillon.

2.2 APPROCHES VARIATIONNELLES RÉGULARISÉES

Boulanger et al. (2014) ont proposé une approche variationnelle pour estimer la densité tridimensionnelle de fluorophores à partir d'acquisitions MA-TIRF. À notre connaissance, ces travaux sont les premiers et les seuls à ce jour à utiliser une telle approche basée sur l'inversion de l'opérateur TIRF. Étant donné que le problème est mal posé (cf. chapitre 1), les auteurs imposent une contrainte de positivité et de régularité spatiale à la solution recherchée. Plus précisément, ils se proposent de résoudre le problème suivant :

$$\hat{f} \in \arg\min_{f} \ \|Hf - s\|^2 + \lambda \|f\|_{TV} + \frac{1}{2} d^2_{f \geqslant 0}, \tag{2.8}$$

où H et f sont des versions discrètes respectivement de l'opérateur TIRF et de la densité de fluorophores, s représente les acquisitions MA-TIRF, $\|\cdot\|_{TV}$ est la norme de variation totale définie en (3.18) (page 33) et enfin $d^2_{f \geqslant 0}$ définit la distance au carré de f à l'ensemble des positifs :

$$d^2_{f \geqslant 0} = \sum_{\substack{i \in \mathbb{I}_N \\ f_i < 0}} f_i^2. \tag{2.9}$$

Cette dernière contrainte est volontairement plus souple que la contrainte de positivité classique impliquant la fonction indicatrice. En effet, le terme $d^2_{f \geqslant 0}$ est un moyen de prendre en compte l'incertitude sur l'estimation du *background* (signal de fond) réalisée par Boulanger et al. (2014) à partir d'une image supplémentaire, dite *dark image*, qui est lissée et soustraite à l'acquisition. Enfin, l'optimisation de cette fonctionnelle est réalisée avec le Parallel ProXimal Algorithm (PPXA) (Combettes et Pesquet, 2008).

Afin de prendre en compte les variations de l'indice de réfraction dans l'échantillon, les auteurs proposent de sélectionner parmi un ensemble d'indices préalablement définis celui minimisant l'erreur de reconstruction (pour chaque pixel du plan Ω).

Pour finir, une dernière méthode de reconstruction aveugle est proposée. Dans ce cas, l'opérateur H ainsi que la densité de fluorophores f sont estimés de manière jointe. De plus, l'opérateur recherché est contraint à avoir une norme unitaire ($\|H\| = 1$) et à être positif ($H \geqslant 0$).

Ces différentes méthodes ont été utilisées pour observer des filaments d'actine[2], des microtubules ainsi que le phénomène d'exocytose[3] de vésicules de sécrétion. Une résolution

2. Plus de détails sur les filaments d'actine seront donnés dans le chapitre 5.

3. Processus de fusion d'une vésicule intra-cellulaire à la membrane permettant de libérer son contenu à l'extérieur de la cellule.

axiale de l'ordre de $40 - 50$ nm est atteinte sur une épaisseur de 800 nm. Notons que l'approche aveugle a permis de révéler plus de détails quant à la position axiale des structures observées en comparaison avec la méthode non-aveugle.

2.3 UNE TECHNIQUE D'ACQUISITION DIRECTE DU VOLUME 3D

Une méthode originale a récemment été proposée par Fu et al. (2016) pour extraire une information tridimensionnelle à partir d'images MA-TIRF. L'idée est relativement simple et consiste en une succession d'acquisitions et de photoblanchiment en partant de l'angle maximal vers l'angle critique comme cela est illustré par la figure 7. Pour le plus grand angle incident admissible, la profondeur de pénétration de l'onde évanescente est la plus courte. La première acquisition (dite «prebleach») vient donc imager une couche très fine adjacente à la lamelle de verre. Cette acquisition est suivie d'un photoblanchiment réalisé avec le même angle d'incidence ce qui a pour effet d'entraîner l'extinction de la fluorescence des molécules présentes dans cette couche. Une nouvelle acquisition dite «postbleach» est alors réalisée et la soustraction de l'image «prebleach» par l'image «postbleach» donne accès à une section axiale de l'échantillon. En répétant l'opération avec un angle incident plus faible, les auteurs ont maintenant directement accès à un plan plus profond de l'échantillon. Cette méthode permet donc de construire un volume 3D de l'échantillon dès l'acquisition sans avoir recours à des méthodes de reconstruction.

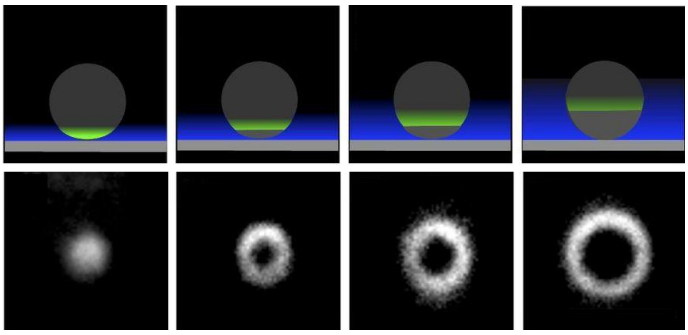


FIGURE 7 – Illustration schématique du processus séquentiel d'acquisition et photoblanchiment (ligne du haut). Exemple réel sur une bille marquée sur sa surface (ligne du bas). [Image extraite de (Fu et al., 2016)]

Les auteurs affirment atteindre avec cette méthode une résolution axiale de l'ordre de 20 nm sur une épaisseur de 200 nm. Notons que cette approche reste limitée pour des acquisitions *invivo* à cause du photoblanchiment qui est réalisé.

3

MÉTHODES NUMÉRIQUES POUR LA RÉSOLUTION DU PROBLÈME INVERSE

Sommaire

Ce chapitre présente l'approche que nous avons utilisée pour résoudre le problème inverse rencontré dans le contexte de reconstruction d'images biologiques à partir de données MA-TIRF. Après avoir discrétisé le modèle, nous formulons le problème de reconstruction en termes de minimisation d'une fonctionnelle, définie dans un cadre bayésien, à laquelle nous ajoutons un terme de régularisation afin de stabiliser la solution. Une étude en dimension 1 est ensuite réalisée pour comprendre l'effet des différents termes de régularisation ou encore de la modélisation du signal de fond.

3.1 Approche bayésienne du problème inverse

3.1.1 *Du modèle continu au modèle discret*

Afin de résoudre le problème inverse numériquement, nous devons tout d'abord discrétiser le modèle TIRF présenté dans le chapitre 1, et en particulier le modèle simplifié (1.10) dont nous rappelons l'expression : pour $\alpha \in]\alpha_c, \alpha_{max}]$, $\forall i \in \mathbb{I}_N$,

$$s(\alpha)_i := \iint_{u \in A_i} \left(I_0(\alpha) \int_{z=0}^{+\infty} \exp(-zp(\alpha)) f(u,z)\, dz \right) du + b_i, \tag{3.1}$$

où $s(\alpha) \in \mathbb{R}^N$ définit l'acquisition obtenue pour l'angle incident α, $A_i \subset \Omega$ représente la région de Ω correspondant au i-ème pixel[1], $f : \Omega \times \mathbb{R}_+ \to \mathbb{R}_+$ est la densité de fluorophores de l'échantillon observé et b_i modélise un signal de fond présent au pixel $i \in \mathbb{I}_N$ de l'acquisition $s(\alpha)$.

Nous commençons donc par discrétiser l'inconnue f du problème inverse. La discrétisation dans le plan Ω étant définie par l'acquisition $s(\alpha)$ (i. e. par le capteur EMCCD),

1. Où $\Omega \subset \mathbb{R}^2$ est le domaine d'observation 2D.

nous nous focalisons sur la discrétisation dans la direction axiale (Oz). Soit $z_{max} \in \mathbb{R}_+$, une borne supérieure à partir de laquelle nous considérons que le système TIRF n'est plus capable d'imager l'échantillon (typiquement $500-600$ nm). On se donne alors une discrétisation $\{z_1,\dots,z_M\}$, $M \in \mathbb{N}^\star$ de $[0, z_{max}]$ correspondant aux centres de cellules de taille $\{\delta_{z_1},\dots,\delta_{z_M}\}$ (i.e. la discrétisation n'est pas nécessairement régulière). Puis on considère que,

$$\forall i \in \mathbb{I}_N, \forall j \in \mathbb{I}_M, \quad f(u,z) = f_{ij} \in \mathbb{R} \quad \forall (u,z) \in A_i \times \left[z_j - \frac{\delta_{z_j}}{2}, z_j + \frac{\delta_{z_j}}{2} \right[. \tag{3.2}$$

Autrement dit, on approche f par une fonction constante par morceaux dont les coefficients dans la base $(\phi_{ij})_{i\in\mathbb{I}_N, j\in\mathbb{I}_M}$, définie par

$$\forall (u,z) \in \Omega \times [0, z_{max}],\ \phi_{ij}(u,z) = \begin{cases} 1 & \text{si } (u,z) \in A_i \times \left[z_j - \frac{\delta_{z_j}}{2}, z_j + \frac{\delta_{z_j}}{2} \right[, \\ 0 & \text{sinon}, \end{cases} \tag{3.3}$$

sont donnés par $f \in \mathbb{R}^{N\times M}$ (i.e. $f_{ij} = \langle f, \phi_{ij}\rangle$). Notons que nous aurions également pu représenter f dans une autre base (e.g. une base de B-Splines).

Considérons maintenant $S = (s(\alpha_1),\dots,s(\alpha_L)) \in \mathbb{R}^{N\times L}$, une acquisition MA-TIRF obtenue pour L angles incidents différents appartenant à $\mathcal{A}$ (défini en (1.11) page 17). Alors nous avons $\forall i \in \mathbb{I}_N$,

$$S_{i\cdot} = \tilde{H} f_{i\cdot} + b_i, \tag{3.4}$$

où $b_i \in \mathbb{R}$ est une constante[2], $S_{i\cdot} = [s(\alpha_1)_i,\dots,s(\alpha_L)_i]^T$, $f_{i\cdot} = [f_{i1},\dots,f_{iM}]^T$ et $\tilde{H} \in \mathbb{R}^{L\times M}$ définit la matrice d'acquisition dont les coefficients sont donnés par : $\forall (l,j) \in \mathbb{I}_L \times \mathbb{I}_M$,

$$\tilde{H}_{lj} = I_0(\alpha_l) \int_{z_j^-}^{z_j^+} \exp\left(-zp(\alpha_l)\right) dz = \frac{I_0(\alpha_l)}{p(\alpha_l)} \left[\exp\left(-z_j^- p(\alpha_l)\right) - \exp\left(-z_j^+ p(\alpha_l)\right) \right], \tag{3.5}$$

avec $z_j^- = z_j - \frac{\delta_{z_j}}{2}$ et $z_j^+ = z_j + \frac{\delta_{z_j}}{2}$. Un exemple d'une telle matrice est présenté sur la figure 8. On remarque clairement que les colonnes correspondants aux z grands sont très corrélées (i.e. presque identiques) ce qui suggère que la qualité de la reconstruction sera moindre pour des objets éloignés de l'interface. D'autre part, le conditionnement de cette matrice est très mauvais ($5e^{+17}$ pour l'exemple de la figure 8), montrant une fois de plus le caractère mal posé du problème de reconstruction associé (voir aussi la figure 10). Notons tout de même que ce conditionnement dépend du choix de la discrétisation axiale et de l'échantillonnage des angles incidents, mais reste toujours très mauvais pour des discrétisations intéressantes en pratique (i.e. $\delta_z \leqslant 50$ nm et un nombre d'angles L suffisamment petit, e.g. $10-20$, pour rester en mesure d'imager des dynamiques rapides).

Dans la suite, afin d'alléger les notations, nous définissons l'opérateur linéaire suivant :

$$\begin{aligned} H : \mathbb{R}^{N\times(M+1)} &\longrightarrow \mathbb{R}^{N\times L} \\ (f,b) &\longmapsto S \ \text{ tel que } \ S_{il} = (\tilde{H} f_{i\cdot})_l + b_i \end{aligned} \tag{3.6}$$

On a donc $S = H(f,b)$ et on notera $H^\star(S)$ l'opérateur adjoint de H vérifiant :

$$\langle H(f,b), S\rangle_{\mathbb{R}^{N\times L}} = \langle H^\star(S), [f,b]\rangle_{\mathbb{R}^{N\times(M+1)}}. \tag{3.7}$$

2. En effet, on considère que le signal de fond est indépendant du phénomène TIRF et donc indépendant de l'angle incident. Cependant, il peut varier dans le plan latéral Ω.

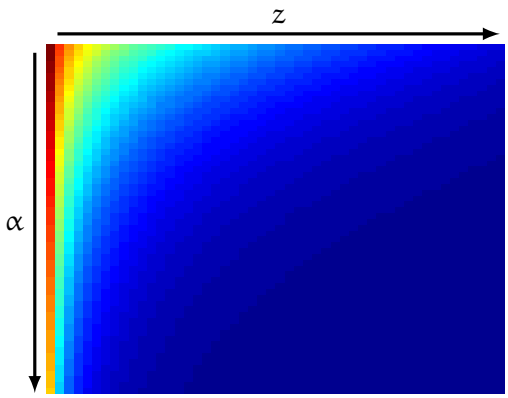


FIGURE 8 – Exemple d'une matrice d'acquisition TIRF.

3.1.2 *Vraisemblance des observations*

Dans le but de résoudre le problème inverse, une méthode classique consiste à maximiser la vraisemblance des observations sachant le modèle, i. e.

$$(\hat{f}, \hat{b}) \in \arg \max_{\substack{f \in \mathbb{R}^{N\times M} \\ b \in \mathbb{R}^{N}}} \mathbb{P}(S|f, b). \tag{3.8}$$

Généralement, on s'intéressera plutôt à la minimisation de l'anti-logarithme de $\mathbb{P}(S|f, b)$ (ou encore *neg-log* vraisemblance) :

$$(\hat{f}, \hat{b}) \in \arg \min_{\substack{f \in \mathbb{R}^{N\times M} \\ b \in \mathbb{R}^{N}}} -\log\big(\mathbb{P}(S|f, b)\big). \tag{3.9}$$

Bien que le signal de fond b puisse être estimé à partir d'une zone de l'image ne contenant pas de structures biologiques, nous nous plaçons dans le contexte de l'estimation jointe de f et b. Nous reviendrons sur l'intérêt d'une telle estimation dans la section 3.2.4.

Dans les deux prochains paragraphes, nous explicitons la fonction de vraisemblance dans le cas où le bruit est considéré respectivement gaussien et poissonien. Le principe général est le suivant :

1. on détermine la probabilité conditionnelle $\mathbb{P}(S_{il}|H(f, b)_{il})$ qui n'est autre que la probabilité du bruit étant donné que l'on est sous l'hypothèse : $H(f, b)_{il}$ connu ;
2. l'indépendance du bruit ($\forall i \in \mathbb{I}_N$), et donc des probabilités conditionnelles, permet ensuite de déterminer la loi jointe conditionnelle (i. e. la vraisemblance) comme il suit :

$$\mathbb{P}(S|f, b) = \prod_{(i,l)\in\mathbb{I}_N\times\mathbb{I}_L} \mathbb{P}(S_{il}|H(f, b)_{il}). \tag{3.10}$$

CAS GAUSSIEN Lorsque le bruit est considéré gaussien de variance σ_g connue, la probabilité conditionnelle $\mathbb{P}(S_{il}|H(f, b)_{il})$ est donnée par

$$\mathbb{P}(S_{il}|H(f, b)_{il}) = \frac{1}{\sqrt{2\pi}\sigma_g} \exp\left(-\frac{(S_{il} - H(f, b)_{il})^2}{2\sigma_g^2}\right), \tag{3.11}$$

et l'estimation de f et b à partir des observations S nécessite alors la résolution du problème de moindres carrés (minimisation de la neg-log vraisemblance),

$$(\hat{f}, \hat{b}) \in \arg \min_{\substack{f \in \mathbb{R}^{N \times M} \\ b \in \mathbb{R}^{N}}} \|H(f, b) - S\|^2. \tag{3.12}$$

CAS POISSONIEN Dans le cas d'un bruit poissonien, nous avons

$$\mathbb{P}(S_{il}|H(f, b)_{il}) = \frac{[H(f, b)_{il}]^{S_{il}} \exp(-H(f, b)_{il})}{S_{il}!}, \tag{3.13}$$

où S_{il} représente ici un comptage de photons. Cette expression conduit au problème de minimisation suivant :

$$(\hat{f}, \hat{b}) \in \arg \min_{\substack{f \in \mathbb{R}^{N \times M} \\ b \in \mathbb{R}^{N}}} \sum_{(i,l) \in \mathbb{I}_N \times \mathbb{I}_L} -S_{il} \log(H(f, b)_{il} + \beta) + H(f, b)_{il}, \tag{3.14}$$

où nous avons introduit un paramètre $\beta \in \mathbb{R}_+^\star$ afin d'éviter la singularité de la fonctionnelle lorsque f et b sont nuls.

Notons que les deux problèmes (3.12) et (3.14) sont convexes et peuvent être résolus par de nombreux algorithmes (voir section 3.1.4).

CAS POISSONIEN-GAUSSIEN Comme nous l'avons vu dans la section 1.3, une description plus précise du bruit présent sur les données consiste à prendre en compte un bruit mixte poissonien-gaussien. Dans ce cas, la vraisemblance est une fonction plus complexe résultant d'une convolution entre les deux types de distributions (BENVENUTO et al., 2008). L'optimisation de la fonctionnelle neg-log résultante est alors un problème plus difficile que dans le cas où seul le bruit gaussien (resp. poissonien) est considéré. C'est pourquoi, il est très souvent considéré que seulement l'un de ces deux types de bruit est présent. Cependant, plusieurs auteurs ont montré la (stricte) convexité et l'existence de minimiseurs pour la neg-log associée au cas poissonien-gaussien (BENVENUTO et al., 2008). Par ailleurs, il a aussi été montré que cette fonctionnelle était à gradient-Lipschitz (JEZIERSKA, 2013 ; CHOUZENOUX et al., 2015) et des algorithmes d'optimisation itératifs ont été proposés.

L'existence de ces travaux est donnée à titre informatif et nous ferons, dans la suite, l'hypothèse qu'un seul type de bruit dégrade les acquisitions (gaussien ou poissonien).

3.1.3 *Régularisation*

Étant donné que le problème inverse auquel nous nous intéressons est mal-posé, il est nécessaire de le régulariser en ajoutant des a priori sur la solution recherchée. Ainsi, nous sommes amenés à minimiser un critère «régularisé» de la forme :

$$J(f, b) = J_d(f, b) + \lambda_r J_r(f, b), \tag{3.15}$$

où J_d est le terme dit *d'attache aux données*, correspondant généralement à la neg-log vraisemblance présentée dans le paragraphe précédent et J_r est le terme de *régularisation* traduisant les informations a priori que l'on souhaite imposer à la solution. Enfin, $\lambda_r \in \mathbb{R}_+$ est un paramètre de régularisation permettant de contrôler l'effet de cette dernière.

POSITIVITÉ DE LA SOLUTION L'inconnue f représentant une concentration en fluorophores, il est naturel de s'intéresser aux solutions du problème qui sont positives. Cette contrainte vaut aussi pour le signal de fond b. Ainsi, nous considérons le terme

$$J_r(f,b) = i_{\geqslant 0}([f,b]) := \begin{cases} 0 & \text{si } (f,b) \in \mathbb{R}_+^{N\times(M+1)}, \\ +\infty & \text{sinon.} \end{cases} \tag{3.16}$$

RÉGULARISATION SPATIALE Il est aussi courant d'ajouter un terme de régularisation spatiale permettant de limiter les petites variations de l'image et ainsi d'être plus robuste au bruit. Pour notre problème, nous appliquerons une telle régularisation uniquement sur la variable f étant donné que nous ne sommes pas réellement intéressés par l'estimation du signal de fond b qui peut donc permettre de collecter une partie du bruit[3]. Par exemple, il est possible d'utiliser une régularisation de Tikhonov d'ordre 1 :

$$J_r(f,b) := \|\nabla f\|_2^2, \tag{3.17}$$

où $\nabla : \mathbb{R}^{N\times M} \to (\mathbb{R}^{N\times M})^3$ représente l'opérateur gradient discret défini par différences finies décentrées. Cependant, une telle régularisation lisse de manière importante les contours des objets reconstruits. Ainsi, afin de préserver au mieux ces contours, la régularisation par variation totale est souvent préférée :

$$J_r(f,b) := \|f\|_{TV} = \|\nabla f\|_{2,1} = \sum_{(i,j)\in\mathbb{I}_N\times\mathbb{I}_M} \sqrt{(\nabla f)_{ij1}^2 + (\nabla f)_{ij2}^2 + (\nabla f)_{ij3}^2}, \tag{3.18}$$

où $(\nabla f)_{ij1}$ (resp. $(\nabla f)_{ij2}$ et $(\nabla f)_{ij3}$) correspond au gradient de f selon la direction (Ox) (resp. (Oy) et (Oz)) au point $(i,j) \in \mathbb{I}_N \times \mathbb{I}_M$. Ce terme de régularisation a l'inconvénient de ne pas être différentiable en zéro[4] et il est alors possible de considérer l'alternative différentiable suivante :

$$J_r(f,b) := \|f\|_{TV_\varepsilon} = \sum_{(i,j)\in\mathbb{I}_N\times\mathbb{I}_M} \sqrt{(\nabla f)_{ij1}^2 + (\nabla f)_{ij2}^2 + (\nabla f)_{ij3}^2 + \varepsilon}, \tag{3.19}$$

où $\varepsilon \in \mathbb{R}_+^\star$ est un paramètre permettant de régulariser la fonctionnelle au voisinage de zéro. Dans la suite, les deux problèmes régularisés suivants seront considérés :

— unique contrainte de positivité,

$$(\hat{f},\hat{b}) \in \arg\min_{\substack{f\in\mathbb{R}^{N\times M}\\ b\in\mathbb{R}^N}} J_d(f,b) + i_{\geqslant 0}([f,b]); \tag{3.20}$$

— positivité et variation totale

$$(\hat{f},\hat{b}) \in \arg\min_{\substack{f\in\mathbb{R}^{N\times M}\\ b\in\mathbb{R}^N}} J_d(f,b) + \lambda_r\|f\|_{TV} + i_{\geqslant 0}([f,b]), \tag{3.21}$$

ou encore le cas où $\|f\|_{TV}$ est remplacée par sa version différentiable $\|f\|_{TV_\varepsilon}$.

3. En effet, nous verrons dans la section 3.2.4 que la modélisation du signal de fond b est surtout importante pour éviter de biaiser f lorsque l'on impose également une contrainte de positivité comme celle de l'équation (3.16).

4. Notons que c'est aussi cette particularité qui favorise des solutions dont le gradient est parcimonieux, permettant ainsi de mieux préserver les contours.

3.1.4 *Algorithmes pour l'optimisation du critère régularisé*

Afin de résoudre les problèmes convexes (3.20) et (3.21), plusieurs possibilités s'offrent à nous et sont présentées dans la suite de cette section.

3.1.4.1 *Forward-Backward Splitting*

L'algorithme Forward-Backward Splitting (FBS) est un algorithme itératif permettant de minimiser des fonctionnelles du type $J(x) = J_1(x) + J_2(x)$ où J_1 est différentiable et où nous sommes en mesure de calculer efficacement l'opérateur proximal (Moreau, 1962) de J_2 défini par :

$$\text{prox}_{J_2}(y) = \arg\min_x \frac{1}{2}\|x-y\|^2 + J_2(x). \tag{3.22}$$

L'algorithme FBS itère entre une étape de descente de gradient sur le terme J_1 et le calcul du proximal de J_2. Ainsi, pour le problème (3.20), on a $J_1 = J_d$, $J_2 = i_{\geqslant 0}$ et FBS réalise les itérés suivants :

$$(f^{n+1}, b^{n+1}) = \text{prox}_{\gamma i_{\geqslant 0}}\Big([f^n, b^n] - \gamma \nabla J_d(f^n, b^n) \Big). \tag{3.23}$$

Dans le cas du problème (3.21) avec le terme $\lambda_r \|f\|_{TV_\varepsilon}$, la différentiabilité de ce dernier permet de l'introduire dans J_1 et on obtient :

$$(f^{n+1}, b^{n+1}) = \text{prox}_{\gamma i_{\geqslant 0}}\Big([f^n, b^n] - \gamma\Big(\nabla J_d(f^n, b^n) + \big[\lambda_r \nabla \|f^n\|_{TV_\varepsilon}, 0_{\mathbb{R}^N}\big]\Big)\Big). \tag{3.24}$$

La convergence de l'algorithme FBS vers un minimiseur global du critère J est démontrée dans (Combettes et Wajs, 2005 ; Combettes et Pesquet, 2011) lorsque les fonctions J_1 et J_2 sont convexes et lorsque $\gamma \in \left]0, \frac{2}{L}\right[$, où L est la constante de Lipschitz du gradient de J_1 :

$$\|\nabla J_1(x) - \nabla J_1(y)\| \leqslant L\|x-y\|, \forall (x,y). \tag{3.25}$$

Afin de mettre en œuvre cet algorithme pour les problèmes (3.20) et (3.21), nous avons besoin des expressions suivantes :

— **opérateur proximal de** $\gamma i_{\geqslant 0}$: pour $y \in \mathbb{R}^N$,

$$\text{prox}_{\gamma i_{\geqslant 0}}(y) = \Big(y_i \mathbb{1}_{\{y_i \geqslant 0\}}\Big)_{i \in \mathbb{I}_N} \tag{3.26}$$

— **gradient de** J_d : Lorsque $J_d(f,b) = \|H(f,b) - S\|^2$, on a

$$\nabla J_d(f,b) = 2H^\star\big(H(f,b) - S\big), \tag{3.27}$$

où $H^\star$ définit l'adjoint de H. La constante de Lipschitz du gradient est alors donnée par

$$L \leqslant \|H\| = \sup_{\|[f,b]\|=1} \|H(f,b)\|. \tag{3.28}$$

Pour le cas poissonien, où $J_d(f,b) = \sum_{(i,l)\in \mathbb{I}_N\times\mathbb{I}_L} -S_{il}\log\left(H(f,b)_{il}+\beta\right)+H(f,b)_{il}$, il vient :

$$\nabla J_d(f,b) = H^\star(\mathbb{1}) - H^\star\left(\frac{S}{H(f,b)+\beta}\right), \tag{3.29}$$

où $\mathbb{1}$ représente l'élément de $\mathbb{R}^{N\times L}$ dont toutes les composantes sont égales à 1. Ce gradient est également L-Lipschitz (Harmany et al., 2012, lemme 1) avec

$$L \leqslant \frac{\max(S)}{\beta^2}\|H\|^2; \tag{3.30}$$

— **gradient de** $\|\cdot\|_{TV_\varepsilon}$:

$$\nabla\|f\|_{TV_\varepsilon} = -\mathrm{div}\left(\frac{\nabla f}{\sqrt{\|(\nabla f)_{ij\cdot}\|^2+\varepsilon}}\right), \tag{3.31}$$

où div est l'opérateur divergence tel que $\nabla^* = -\mathrm{div}$ et $\|(\nabla f)_{ij\cdot}\|^2 = (\nabla f)^2_{ij1} + (\nabla f)^2_{ij2} + (\nabla f)^2_{ij3}$.

3.1.4.2 *Algorithme de Chambolle-Pock*

Une alternative pour résoudre le problème (3.20), lorsque $J_d = \frac{1}{2}\|H(\cdot)-S\|^2$, est l'algorithme proposé par Chambolle et Pock (2011). Cet algorithme primal-dual du premier ordre permet de minimiser des fonctionnelles du type $J(x) = J_1(Kx)+J_2(x)$ où K est un opérateur linéaire et J_1 et $J_2^\star$ sont propres, convexes et semi-continues inférieures s.c.i., avec $J_2^\star$ la fonction conjuguée de J_2 définie par la transformée de Legendre-Fenchel (Fenchel, 1949) :

$$J_2^\star(x^\star) = \sup_x \langle x^\star, x\rangle - J_2(x). \tag{3.32}$$

Pour le problème (3.20), on prendra $J_1 = \frac{1}{2}\|\cdot - S\|^2$, $K = H$, $J_2 = i_{\geqslant 0}$ et l'algorithme de Chambolle-Pock (CP) réalise alors les itérations suivantes :

— $g^{n+1} = \mathrm{prox}_{\sigma J_1^\star}\left(g^n + \sigma H([\bar{f}^n, \bar{b}^n])\right)$;

— $[f^{n+1}, b^{n+1}] = \mathrm{prox}_{\tau J_2}\left([f^n, b^n] - \tau H^\star(g^{n+1})\right)$;

— $[\bar{f}^{n+1}, \bar{b}^{n+1}] = [f^{n+1}, b^{n+1}] + \theta([f^{n+1}, b^{n+1}] - [f^n, b^n])$,

dans le but de calculer un point selle $([\hat{f}, \hat{b}], \hat{g}) \in \mathbb{R}^{N\times(M+1)} \times \mathbb{R}^{N\times L}$ du problème primal-dual associé. L'algorithme converge vers un tel point lorsque les paramètres vérifient $\theta = 1$ et $\sigma\tau\|H\|^2 < 1$ (Chambolle et Pock, 2011, théorème 1). Enfin, pour mettre en œuvre cet algorithme, nous avons besoin de l'opérateur proximal de $i_{\geqslant 0}$, donné en (3.26), mais aussi du proximal de la fonction conjuguée de $J_1 = \frac{1}{2}\|\cdot - S\|^2$ qui s'exprime comme il suit :

$$\mathrm{prox}_{\sigma J_1^\star}(g) = \frac{g-\sigma S}{\sigma+1}. \tag{3.33}$$

3.1.4.3 *Algorithme de Richardson-Lucy*

Dans le cas où le terme d'attache aux données (J_d) est celui associé à la statistique de Poisson (problème (3.14)), il est possible d'utiliser l'algorithme de Richardson-Lucy (RL) (Ri-

CHARDSON, 1972 ; LUCY, 1974). Dans sa version multiplicative, une itération est donnée par

$$[f^{n+1}, b^{n+1}] = \frac{[f^n, b^n]}{H^\star(\mathbb{1})} \odot H^\star\left(\frac{S}{H(f^n, b^n) + \beta}\right), \tag{3.34}$$

où $\odot$ définit le produit d'Hadamard (i. e. produit composantes par composantes). Les divisions sont aussi à considérer composantes par composantes. Un des intérêts de cet algorithme est qu'il préserve la positivité des itérés (f^n, b^n) dès lors que l'initialisation est positive, $(f^0, b^0) \in \mathbb{R}_+^{N\times(M+1)}$. Il existe également une version de cet algorithme pour le critère régularisé avec le terme $\lambda_r\|\cdot\|_{TV_\varepsilon}$ proposée par DEY et al. (2006). Dans ce cas, le schéma (3.34) est modifié selon :

$$[f^{n+1}, b^{n+1}] = \frac{[f^n, b^n]}{H^\star(\mathbb{1}) - \left[\lambda_r \operatorname{div}\left(\frac{\nabla f^n}{\sqrt{\|(\nabla f)_{ij\cdot}\|^2+\varepsilon}}\right), 0_{\mathbb{R}^N}\right]} \odot H^\star\left(\frac{S}{H(f^n, b^n) + \beta}\right). \tag{3.35}$$

Notons qu'ici la positivité des itérés (f^n, b^n) n'est plus assurée pour toutes les valeurs de λ_r, paramètre qui est donc à choisir avec précaution.

3.1.4.4 *Parallel ProXimal Algorithm*

Enfin, lorsque nous considérons le problème (3.21) avec le terme $\lambda_r\|f\|_{TV}$ ainsi que l'attache aux données quadratique, le critère objectif est alors composé de trois termes dont deux non-différentiables. L'algorithme PPXA (COMBETTES et PESQUET, 2008) permet de minimiser de telles fonctionnelles composées d'une somme de fonction convexes pour lesquelles nous sommes en mesure de calculer l'opérateur proximal associé[5]. Le schéma de PPXA est présenté dans l'algorithme 1. Pour le mettre en œuvre, nous avons besoin du proximal de $i_{\geqslant 0}$ qui est donné en (3.26) ainsi que ceux des termes $\frac{1}{2}\|H(\cdot) - S\|^2$ et $\lambda_r\|\cdot\|_{TV}$. Pour le premier, nous avons,

$$\operatorname{prox}_{\frac{\gamma}{2}\|H(\cdot)-S\|^2}([f, b]) = (\mathrm{Id} + \gamma H^\star H)^{-1}([f, b] + \gamma H^\star(S)), \tag{3.36}$$

où Id représente l'opérateur identité. En ce qui concerne la norme TV, on doit résoudre

$$\operatorname{prox}_{\gamma\|\cdot\|_{TV}}(g) = \arg\min_{f\in\mathbb{R}^{N\times M}} \frac{1}{2\gamma}\|f - g\|^2 + \lambda_r\|\nabla f\|_{2,1}, \tag{3.37}$$

ce qui peut être réalisé avec l'algorithme CP présenté précédemment en prenant $K = \nabla$, $J_1 = \lambda_r\|\cdot\|_{2,1}$ et $J_2 = \frac{1}{2\gamma}\|\cdot - g\|$. Notons que nous avons pour $h \in \mathbb{R}^{N\times M}$,

$$\operatorname{prox}_{\frac{\tau}{2\gamma}\|\cdot-g\|}(h) = \frac{\gamma h + \tau g}{\gamma + \tau}, \tag{3.38}$$

et, pour $h \in (\mathbb{R}^{N\times M})^3$,

$$\operatorname{prox}_{\sigma(\lambda_r\|\cdot\|_{2,1})^\star}(h) = \left(\frac{\lambda_r h_{ijk}}{\max\left(\lambda_r, \|h_{ij\cdot}\|\right)}\right)_{(i,j,k)\in\mathbb{I}_N\times\mathbb{I}_M\times\{1,2,3\}}. \tag{3.39}$$

5. Il est aussi possible d'appliquer cet algorithme dans le cas où l'attache aux données est celle associée au bruit de Poisson mais le calcul du proximal de cette dernière n'est pas direct et nécessitera la résolution d'un sous-problème comme nous le faisons ici pour la norme TV.

Algorithme 1 : PPXA pour le problème (3.21)

Entrées : $\lambda_r \in \mathbb{R}_+$, $\gamma \in \mathbb{R}_+^\star$, $\left(u_{j,0}\right)_{1\leqslant j\leqslant 3} \in (\mathbb{R}^{N\times(M+1)})^3$

Poser $\omega_1 = \omega_3 = \frac{1}{2+\lambda_r}$ et $\omega_2 = \frac{\lambda_r}{2+\lambda_r}$ /* Ainsi $\sum_{j=1}^{3} \omega_j = 1$ */

Définir $[f^0, b^0] = \sum_{j=0}^{3} \omega_j u_{j,0}$;

répéter

  $p_{1,n} = \mathrm{prox}_{\frac{\gamma}{\omega_1} J_d}(u_{1,n})$;

  $p_{2,n} = \mathrm{prox}_{\frac{\gamma}{\omega_2}\|\cdot\|_{TV}}(u_{2,n})$/* ici le prox s'applique uniquement sur la partie de $u_{2,n}$ correspondant à f (on ignore b) */

  $p_{3,n} = \mathrm{prox}_{\frac{\gamma}{\omega_3} i_{\geqslant 0}}(u_{3,n})$;

  $p_n = \sum_{j=1}^{3} \omega_j p_{j,n}$;

  Choisir $\lambda_n \in]0, 2[$;

  pour *j=1,2,3* **faire**

    $u_{j,n+1} = u_{j,n} + \lambda_n(2p_n - [f^n, b^n] - p_{j,n})$;

  $[f^{n+1}, b^{n+1}] = [f^n, b^n] + \lambda_n(p_n - [f^n, b^n])$;

jusqu'à *convergence*;

Sorties : $[f^n, b^n]$

3.2 ÉTUDE UNIDIMENSIONNELLE

Dans cette section, nous réalisons une étude numérique pour le cas unidimensionnel. L'objectif est d'une part d'observer l'effet, sur la solution, des différentes fonctionnelles présentées précédemment, et d'autre part de mettre en avant l'importance de la modélisation et l'estimation du signal de fond b. Bien qu'en pratique, le problème concerne la reconstruction d'un volume 3D, l'étude réalisée ici permet de mettre en évidence et de comprendre certains phénomènes plus facilement. Par ailleurs, dans ce contexte de reconstruction MA-TIRF, la représentation de reconstructions 1D est plus démonstrative que la visualisation d'un volume 3D dont les dimensions n'ont pas du tout le même ordre de grandeur (la profondeur z est très inférieure au champ d'observation latéral Ω).

3.2.1 *Simulation des données*

Tout d'abord, nous générons un ensemble «d'objets» pour différentes profondeurs comme cela est représenté sur la figure 9 (gauche). On peut par exemple considérer que ce sont des coupes de vésicules à l'intérieur desquelles la densité de fluorophores est homogène. Notons que chaque objet (i. e. correspondant à une certaine profondeur) sera traité indépendamment dans la suite de cette section (on ne considèrera donc pas de multiples objets alignés en z ici). La figure 9 (droite) présente les acquisitions MA-TIRF (non-bruitées et sans ajout d'un signal de fond) obtenues pour ces objets et pour les paramètres mentionnés en légende. En notant $f^\star \in \mathbb{R}^M$ le profil d'un l'objet, l'acquisition $s^\star \in \mathbb{R}^L$ correspondante sur la figure 9 (droite) est calculée selon,

$$s^\star = \tilde{H} f^\star, \tag{3.40}$$

où $\tilde{H}$ est la matrice d'acquisition, définie en (3.5), et représentée sur la figure 8.

Remarquons avec la figure 9 (droite) que les réponses, à travers le système MA-TIRF, des objets les plus éloignés sont très similaires, illustrant la difficulté à différencier de telles structures.

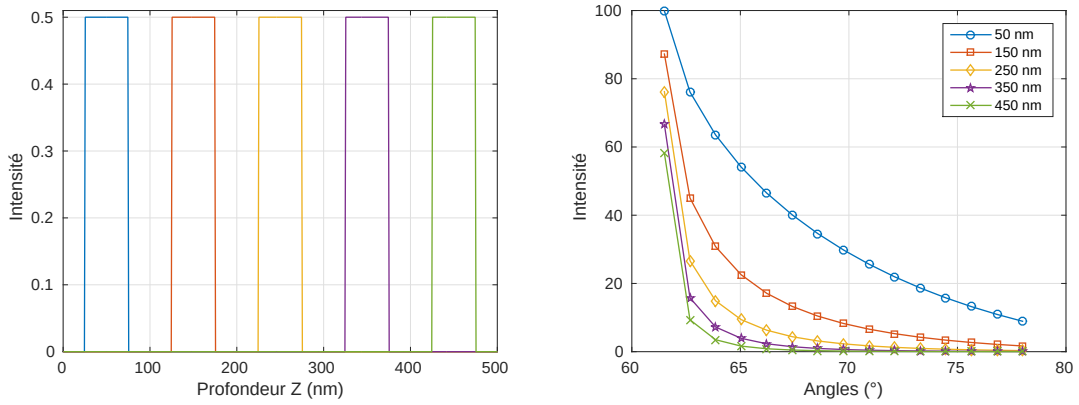


FIGURE 9 – Simulation d'objets 1D (gauche) d'épaisseur 50 nm et d'amplitude 0.5 pour différentes profondeurs $z \in \{50, 150, 250, 350, 450\}$ (nm), et acquisitions MA-TIRF, pour chaque objet considéré séparément, non-bruitées correspondantes (droite) pour $n_i = 1.518$, $n_t = 1.333$, $\lambda = 491$ nm et $L = 15$ angles incidents différents.

3.2.2 *Positivité de la solution*

Afin de visualiser l'effet et l'importance d'une contrainte de positivité, nous avons calculé la solution du système (3.40) au sens des moindres carrés pour une discrétisation axiale régulière, de pas $\delta_z = 1$ nm, donnée par $\{0.5, 1, \ldots, 499.5\}$(nm). Le système est alors extrêmement sous-déterminé ($\tilde{H} \in \mathbb{R}^{15\times500}$) et admet une infinité de solutions. La figure 10 (gauche) présente la solution de norme minimale (donnée par la pseudo-inverse de $\tilde{H}$) obtenue pour chacun des objets simulés. On voit clairement apparaître des oscillations indésirables montrant une fois de plus le caractère mal posé de ce problème de reconstruction. Au contraire, comme nous pouvons le constater sur la figure 10 (droite), contraindre la solution à être positive permet d'obtenir un résultat bien plus intéressant. Enfin, notons que la précision de la localisation décroît avec la profondeur des objets.

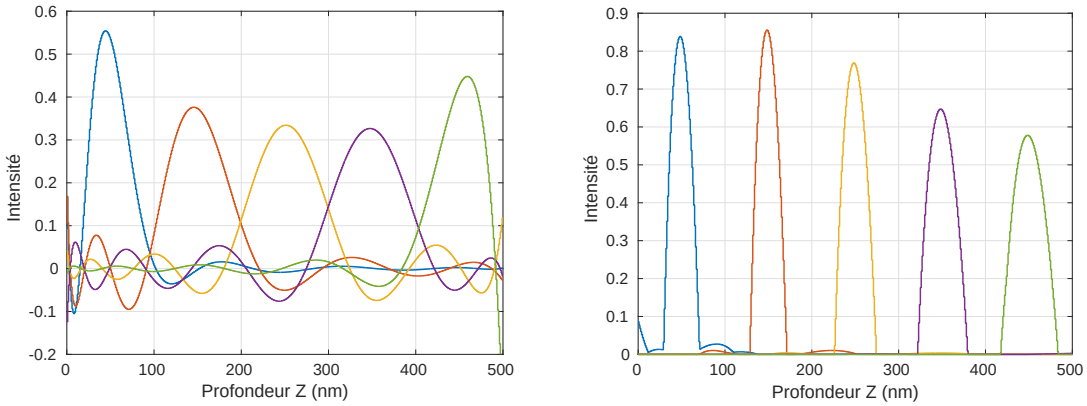


FIGURE 10 – Solutions obtenues pour une discrétisation axiale régulière, de pas $\delta_z = 1$ nm, donnée par $\{0.5, 1, \ldots, 499.5\}$(nm), avec : la pseudo inverse de $\tilde{H}$ (gauche) ou la résolution du problème de moindres carrés avec contrainte de positivité (i. e. problème (3.20) où $J_d = \frac{1}{2}\|\tilde{H}\cdot - s^\star\|^2$) en utilisant l'algorithme CP (droite).

3.2.3 *Effet de la régularisation TV*

En pratique, les données étant dégradées par différents types de bruits (voir section 1.3 page 20), il est classique d'ajouter un terme de régularisation au problème afin d'imposer un a priori sur la régularité de la solution (en plus de la positivité). Nous avons entre autres évoqué précédemment la régularisation par variation totale (TV) définie par (3.18) (ou (3.19) pour sa version différentiable). Une telle régularisation a pour but de promouvoir les solutions constantes par morceaux, c'est-à-dire constituées de zones homogènes séparées par des contours bien marqués. La norme TV étant définie par la norme-ℓ_1 de l'amplitude du gradient, elle favorise la parcimonie des discontinuités (ce qui est recherché) mais atténue également leurs amplitudes et donc l'intensité des zones homogènes.

Ce phénomène est bien connu (Strong et Chan, 2003) et est généralement peu problématique pour des applications telles que la segmentation, le débruitage ou encore la déconvolution par exemple, résultant simplement en une atténuation de l'intensité de l'image segmentée (restaurée). Cependant, ce n'est pas le cas du problème de reconstruction MA-TIRF auquel nous nous intéressons. En effet, l'illustration présentée sur la figure 11 (voir aussi la table 1) suggère qu'une régularisation TV pour le problème MA-TIRF doit être utilisée avec précaution.

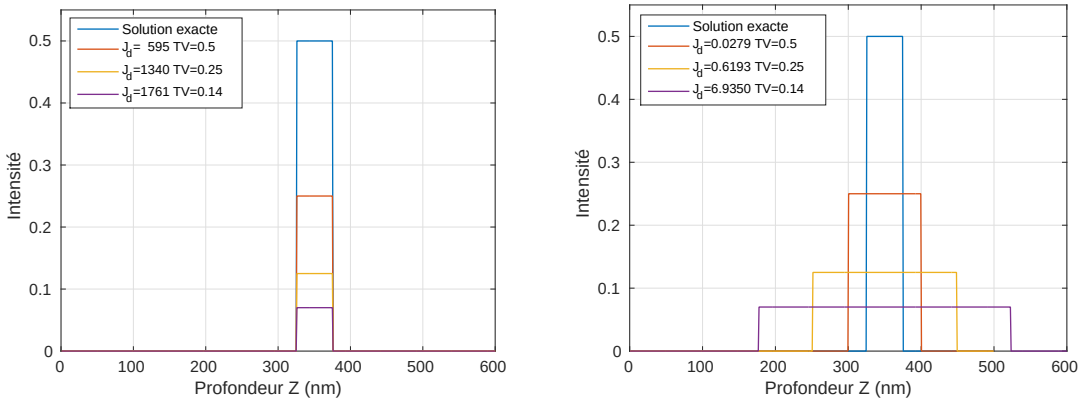


Figure 11 – Différentes solutions pour la reconstruction (non-bruitée) de l'objet en bleu positionné à 350 nm de l'interface. Pour chacune de ces solutions, la valeur du terme d'attache aux données $J_d = \frac{1}{2}\|\tilde{H} \cdot - s^\star\|^2$, ainsi que celle de la régularisation TV, sont mentionnées. La figure de gauche représente différents objets de même largeur (50 nm) mais ayant des intensités différentes. La figure de droite correspond aux objets minimisant J_d pour une valeur de TV fixée.

La figure 11, ainsi que la table 1, présentent les valeurs de $J_d = \frac{1}{2}\|\tilde{H} \cdot - s^\star\|^2$ et $\|\cdot\|_{TV}$ pour plusieurs objets différant de part leurs intensités et leurs «largeurs», mais étant tous centrés en $z = 350$ nm. Notons que le terme J_d est défini avec $s^\star \in \mathbb{R}^L$ correspondant à l'acquisition de l'objet de la figure 9 positionné à 350 nm (en bleu, également représenté sur la figure 11). Alors que sur la figure de gauche, les objets ne diffèrent que de part leurs intensités, la largeur des objets de la figure de droite a été déterminée par minimisation de J_d pour une intensité fixée (et donc la valeur de TV fixée). Cette illustration permet de constater que lorsque TV diminue (et donc l'intensité de l'objet diminue), il est possible, sans changer la valeur de TV, de décroître J_d en augmentant l'épaisseur (en z) de l'objet.

Ainsi, pour le problème de reconstruction MA-TIRF, l'atténuation des zones homogènes provoquée par TV entraine une *dégradation de la localisation axiale des objets reconstruits*, ce qui est en opposition avec l'objectif premier d'une telle reconstruction visant à localiser

précisément les structures biologiques dans la direction axiale. Cet effet est d'autant plus important que la position axiale de l'objet en question est éloignée de l'interface.

Notons que nous avions également observé ce phénomène dans (Soubies et al., 2014b) où nous nous intéressions à la reconstruction de membranes par microscopie MA-TIRF en considérant que de telles structures étaient parcimonieuses selon (Oz). Dans le cadre de ce travail, nous avions constaté qu'utiliser une norme-ℓ_1 sur les coefficients de la solution, non seulement les atténuait, mais introduisait un biais sur la localisation axiale de la membrane pour les parties les plus éloignées de l'interface . La position axiale de la membrane dans ces zones était alors sous-estimée (trop proche de l'interface). D'autre part, nous avions montré que ce problème pouvait était limité avec une méthode basée sur une approximation de la pseudo norme-ℓ_0 remplaçant la norme-ℓ_1.

TV \ LO (nm)	50	100	200	≈ 357
0.5	595	**0.0279**	—	—
0.25	1340	—	**0.6193**	—
0.145	1761	—	—	**6.935**

Table 1 – Résumé des valeurs de J_d présentées sur la figure 11 où LO représente la *Largeur de l'Objet* centré en $z = 350$ nm.

Cependant, TV reste intéressant en 3D pour régulariser dans les directions latérales (Ox) et (Oy), notamment lorsque nous sommes en présence d'un bruit important. Dans ce cas, il est important de ne pas prendre un valeur trop grande de λ_r affectée au terme TV pour ne pas trop dégrader la localisation axiale des structures. On peut aussi envisager de mettre TV uniquement en (x,y). Dans les applications réelles des chapitres 4 et 5, nous utiliserons une régularisation TV que dans les cas très bruités. Pour les images où le bruit n'est pas très important, nous nous contenterons de la positivité avec le problème (3.20). [6]

3.2.4 *Importance de la modélisation du signal de fond*

Afin d'étudier l'effet de la modélisation du signal de fond, une constante $b_g \in \mathbb{R}_+^\star$ a été ajoutée aux acquisitions de la figure 9 (droite), puis nous avons réalisé une reconstruction avec contrainte de positivité mais sans prendre en compte ce signal de fond dans le modèle. Les résultats obtenus avec et sans le terme TV, et pour deux valeurs de b_g différentes (1 et 5), sont présentés sur la figure 12. On observe sur cette figure la présence d'un fort signal en $z = 0$ ainsi qu'une erreur sur la position axiale des structures reconstruites. L'amplitude du «pic» en $z = 0$, aussi bien que l'erreur de localisation axiale, augmentent avec la constante b_g ajoutée aux acquisitions. Ce phénomène est expliqué par la proposition suivante.

Proposition 3.1. *Soit $\mathcal{H} : \mathbb{L}_2(\mathbb{R}_+) \to \mathbb{R}^L$, l'opérateur TIRF 1D ne modélisant pas de signal de fond :*

$$\forall f \in \mathbb{L}_2(\mathbb{R}_+),\ \mathcal{H}(f)(\alpha) = I_0(\alpha) \int_0^{+\infty} f(z) \exp\left(-zp(\alpha)\right) dz. \tag{3.41}$$

6. Si le bruit est visible sur la solution obtenue par minimisation de (3.20), alors c'est qu'il est prférable de considérer le terme TV.

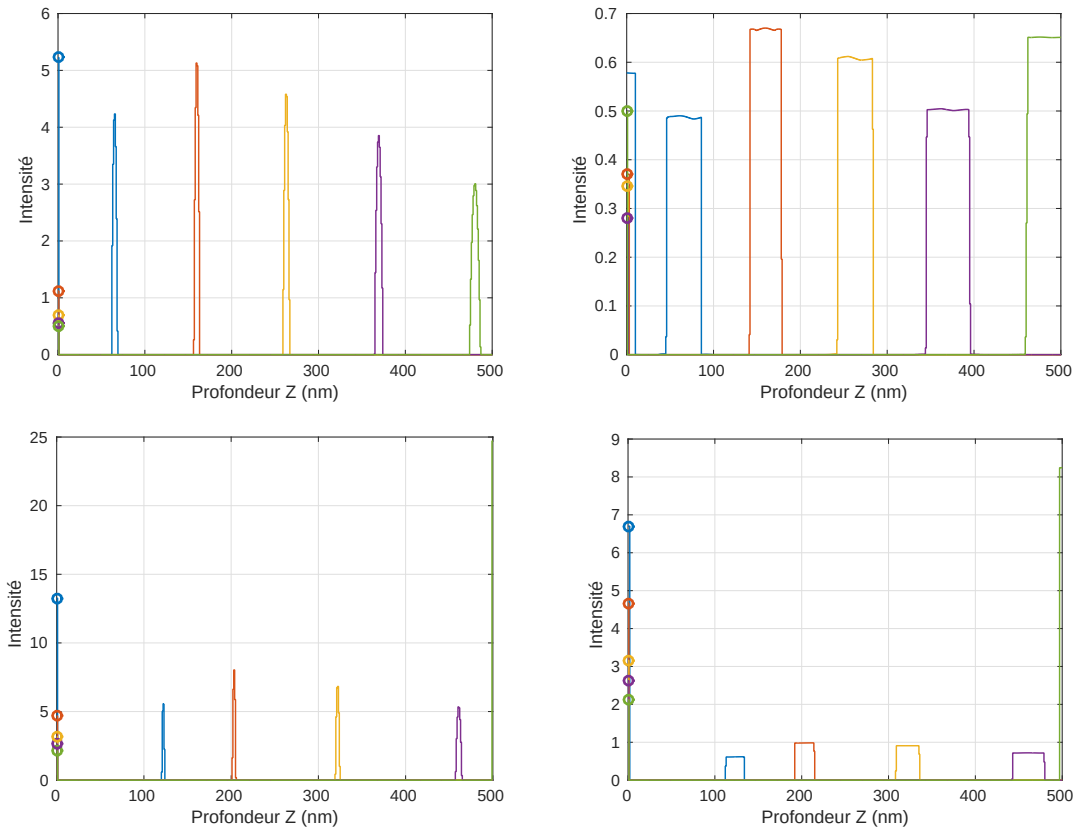


FIGURE 12 – Reconstructions obtenues avec les données de la figure 9 (droite) auxquelles une constante positive $b_g = 1$ (ligne du haut) et $b_g = 5$ (ligne du bas) a été ajoutée. Les reconstructions sont déterminées par minimisation de (3.20) (gauche) et (3.21) (pour $\lambda_r = 0.01$, droite) avec J_d le critère moindre carrés et respectivement les algorithmes CP et PPXA. Lors de la reconstruction, seule la variable f est optimisée et le signal de fond n'est pas estimé et est considéré nul.

Considérons un signal (non-bruitée) constant $s(\alpha) = b_g \in \mathbb{R}_+^\star \ \forall \alpha \in \mathcal{A}$. *Alors, la meilleure approximation* $f^\star \in \mathbb{L}_2(\mathbb{R}_+)$ *au sens des moindres carrés avec contrainte de positivité, telle que* $s \approx \mathcal{H}(f^\star)$, *est donnée par :*

$$f^\star(z) = K^\star \delta_0(z), \ \forall z \in \mathbb{R}_+, \tag{3.42}$$

où δ_0 *est la fonction de Dirac et,*

$$K^\star = \frac{\sum_{\alpha\in\mathcal{A}} I_0(\alpha)}{\sum_{\alpha\in\mathcal{A}} \left(I_0(\alpha)\right)^2} b_g. \tag{3.43}$$

Démonstration. On s'intéresse à la solution de

$$f^\star \in \arg \min_{f\in\mathbb{L}^2(\mathbb{R}_+), f\geqslant 0} J(f) := \frac{1}{2}\|\mathcal{H}(f) - s\|^2 = \frac{1}{2}\sum_{\alpha\in\mathcal{A}} \left(\mathcal{H}(f)(\alpha) - b_g\right)^2. \tag{3.44}$$

Soit $f_K(z) = K\delta_0(z)$ pour $K \geqslant 0$, alors on a

$$J(f_K) = \frac{1}{2}\sum_{\alpha\in\mathcal{A}} \left(I_0(\alpha)K - b_g\right)^2, \tag{3.45}$$

et $\forall K \geqslant 0$, $J(f_K) \geqslant J(f_{K^\star})$ où $K^\star$ est donné par

$$\sum_{\alpha\in\mathcal{A}} I_0^2(\alpha)K^\star - I_0(\alpha)b_g = 0 \iff K^\star = \frac{\sum_{\alpha\in\mathcal{A}} I_0(\alpha)}{\sum_{\alpha\in\mathcal{A}} I_0^2(\alpha)} b_g. \tag{3.46}$$

Considérons maintenant $g \in \mathbb{L}^2(\mathbb{R}_+), g \geqslant 0$ et posons $G(\alpha) = \mathcal{H}(g)(\alpha)/I_0(\alpha)$. G est une fonction décroissante et positive et il en est de même pour I_0. Ainsi, $\alpha \mapsto I_0(\alpha)G(\alpha)$ est également positive décroissante. Nous rappelons que les angles incidents dans $\mathcal{A}$ sont ordonnés : $\alpha_1 < \cdots < \alpha_L$. Nous distinguons trois cas :

— si $I_0(\alpha_1)G(\alpha_1) \leqslant b_g$ alors,

$$\begin{aligned}
&\forall \alpha \in \mathcal{A},\ I_0(\alpha)G(\alpha) \leqslant I_0(\alpha)G(\alpha_1) \leqslant I_0(\alpha_1)G(\alpha_1) \leqslant b_g,\\
\iff &\forall \alpha \in \mathcal{A},\ I_0(\alpha)G(\alpha) - b_g \leqslant I_0(\alpha)G(\alpha_1) - b_g \leqslant 0,\\
\iff &\forall \alpha \in \mathcal{A},\ |I_0(\alpha)G(\alpha) - b_g| \geqslant |I_0(\alpha)G(\alpha_1) - b_g|,\\
\implies &J(g) \geqslant J(f_{G(\alpha_1)}) \geqslant J(f_{K^\star}).
\end{aligned}$$

— si $I_0(\alpha_L)G(\alpha_L) \geqslant b_g$ alors,

$$\begin{aligned}
&\forall \alpha \in \mathcal{A},\ I_0(\alpha)G(\alpha) \geqslant I_0(\alpha)G(\alpha_L) \geqslant I_0(\alpha_L)G(\alpha_L) \geqslant b_g,\\
\iff &\forall \alpha \in \mathcal{A},\ I_0(\alpha)G(\alpha) - b_g \geqslant I_0(\alpha)G(\alpha_L) - b_g \geqslant 0,\\
\implies &J(g) \geqslant J(f_{G(\alpha_L)}) \geqslant J(f_{K^\star}).
\end{aligned}$$

— enfin, s'il existe $l \in \{2, \ldots, L-1\}$ tel que $I_0(\alpha_l)G(\alpha_l) \geqslant b_g \geqslant I_0(\alpha_{l+1})G(\alpha_{l+1})$ alors par continuité de $\alpha \mapsto I_0(\alpha)G(\alpha)$ il existe $\alpha^\star \in [\alpha_l, \alpha_{l+1}]$ tel que $I_0(\alpha^\star)G(\alpha^\star) = b_g$ (théorème des valeurs intermédiaires) et,

$$\begin{aligned}
&\forall \alpha \in \mathcal{A},\ \begin{cases} \alpha \leqslant \alpha^\star \implies I_0(\alpha)G(\alpha) \geqslant I_0(\alpha)G(\alpha^\star) \geqslant b_g,\\ \alpha \geqslant \alpha^\star \implies I_0(\alpha)G(\alpha) \leqslant I_0(\alpha)G(\alpha^\star) \leqslant b_g, \end{cases}\\
\iff &\forall \alpha \in \mathcal{A},\ \begin{cases} \alpha \leqslant \alpha^\star \implies I_0(\alpha)G(\alpha) - b_g \geqslant I_0(\alpha)G(\alpha^\star) - b_g \geqslant 0,\\ \alpha \geqslant \alpha^\star \implies I_0(\alpha)G(\alpha) - b_g \leqslant I_0(\alpha)G(\alpha^\star) - b_g \leqslant 0, \end{cases}\\
\iff &\forall \alpha \in \mathcal{A},\ |I_0(\alpha)G(\alpha) - b_g| \geqslant |I_0(\alpha)G(\alpha^\star) - b_g|,\\
\implies &J(g) \geqslant J(f_{G(\alpha^\star)}) \geqslant J(f_{K^\star}).
\end{aligned}$$

On a donc toujours $J(g) \geqslant J(f_{K^\star})$ ce qui montre que $f_{K^\star}$ est solution de (3.44) et termine la démonstration. □

Ainsi, la meilleure interprétation (au sens des moindres carrés positifs) d'un signal constant par le modèle TIRF (sans prise en compte d'un signal de fond) est un Dirac en $z = 0$. Bien qu'en pratique nous n'ayons pas uniquement un signal constant mais la somme du signal émis par les structures d'intérêt et d'un signal de fond (considéré constant en fonction des angles d'incidence), les résultats de reconstruction ne présentent pas un Dirac en $z = 0$ plus les objets correctement reconstruits. En effet, une partie du signal de fond a bien généré un Dirac en $z = 0$ mais une autre partie de ce signal est venu dégrader la localisation axiale des structures reconstruites.

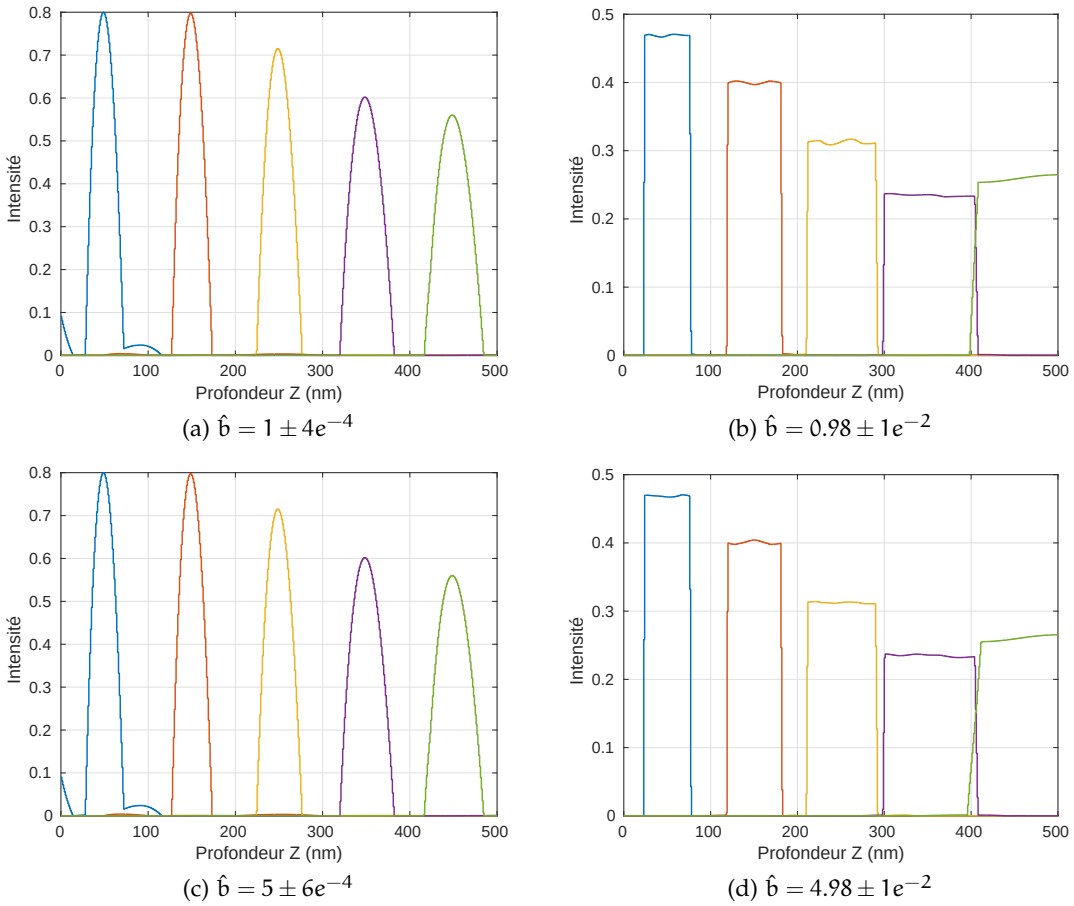
Afin de pallier ce problème, deux options sont envisageables :

1. estimer au préalable le signal de fond puis le soustraire aux acquisitions. Boulanger et al. (2014) utilisent une telle approche où l'estimation du signal de fond est réalisée à partir d'un acquisition supplémentaire (dite *dark image*) lissée puis soustraite

à toutes les acquisitions. L'incertitude d'une telle estimation est prise en compte en relâchant la contrainte de positivité stricte par la distance à l'ensemble des positifs (voir section 2.2, page 26) ;

2. estimer conjointement la densité f et le background b en imposant leur positivité ;

Dans ce manuscrit, nous avons choisi la deuxième approche et les résultats, pour les mêmes configurations que ceux de la figure 12, sont présentés sur la figure 13. On peut constater que la modélisation du signal de fond permet d'obtenir des reconstructions d'une qualité similaire à celles de la figure 10 (droite) correspondant au cas où les données ne sont pas dégradées par un signal de fond. Par ailleurs, remarquons que le signal de fond est bien estimé. Enfin, notons que l'effet de la régularisation TV présenté dans le paragraphe précédent est visible pour les objets les plus éloignés.



(a) $\hat{b} = 1 \pm 4e^{-4}$ (b) $\hat{b} = 0.98 \pm 1e^{-2}$

(c) $\hat{b} = 5 \pm 6e^{-4}$ (d) $\hat{b} = 4.98 \pm 1e^{-2}$

FIGURE 13 – Même expérience que sur la figure 12 à la différence qu'ici une estimation simultanée de f et du signal de fond b est réalisée. La ligne du haut correspond au cas où une constante $b_g = 1$ a été ajoutée aux acquisitions et celle du bas au cas $b_g = 5$. La colonne de gauche présente les résultats obtenus par moindres carrés en imposant une contrainte de positivité sur f et b (problème (3.20)) et celle de droite présente le cas où une régularisation TV ($\lambda_r = 0.01$) sur f est également considérée en plus de la positivité de f et b (problème (3.21)). Les valeurs estimées pour le signal de fond b sont données en légende de chaque sous figure.

3.2.5 *Comparaison poissonien/gaussien*

Nous avons évoqué précédemment le fait que les acquisitions étaient entachées d'un bruit mixte poissonien-gaussien mais que pour des raisons de simplicité, nous nous restrei-

gnons à un modèle purement gaussien ou purement poissonien. Dans ce contexte, la question de savoir si l'un de ces deux modèles est préférable se pose. Afin de répondre à cette question, nous nous proposons de réaliser une expérience visant à comparer les reconstructions obtenues par minimisation du critère imposant la positivité de la solution (3.20) en utilisant :

— l'algorithme CP lorsque $J_d(f,b) = \frac{1}{2}\|H(f,b) - s\|^2$ (modèle gaussien) ;

— l'algorithme RL lorsque $J_d(f,b) = \sum_{l=1}^{L} -s_l \log\left(H(f,b)_l + \beta\right) + H(f,b)_l$ (modèle poissonien).

Les données bruitées ont été générées selon :

$$\tilde{s} = \underbrace{\frac{1}{\gamma}\mathcal{P}\left(\gamma(\overbrace{\tilde{H}f^\star + b_g}^{=s^\star})\right)}_{=s^{\mathcal{P}}} + \mathcal{N}\left(0, \sigma_g^2\right), \tag{3.47}$$

où $\gamma \in \mathbb{R}_+^\star$ est un paramètre permettant de contrôler la variance du bruit de Poisson, $b_g \in \mathbb{R}$ est le signal de fond et σ_g^2 représente la variance du bruit gaussien donnée par

$$\sigma_g^2 = \frac{\|s^\star\|^2}{L \times 10^{SNR_{\mathcal{P}}/10}}, \tag{3.48}$$

avec $SNR_{\mathcal{P}}$ le SNR associé au bruit de Poisson :

$$SNR_{\mathcal{P}} = 10\log_{10}\left(\frac{\|s^\star\|^2}{\|s^\star - s^{\mathcal{P}}\|^2}\right). \tag{3.49}$$

L'idée ici est d'avoir approximativement une contribution égale des deux bruits gaussien et poissonien. Le paramètre γ est donc modifié pour obtenir des SNR globaux (i.e. équation (3.49) en remplaçant $s^{\mathcal{P}}$ par $\tilde{s}$) différents.

Afin d'être en mesure d'évaluer la qualité des reconstructions, il convient de se donner une mesure d'erreur. En notant $f^\star$ la solution recherchée (i.e. l'objet simulé), on peut par exemple calculer l'Erreur Quadratique Moyenne (EQM) entre l'estimation $\hat{f}$ et $f^\star$,

$$EQM(\hat{f}, f^\star) = \frac{1}{M}\|\hat{f} - f^\star\|^2, \tag{3.50}$$

où on rapelle que M est le nombre de points de la discrétisation axiale (i.e. $f \in \mathbb{R}^M$). Cependant, comme nous l'avons déjà évoqué, nous sommes principalement intéressés par la localisation axiale des structures, ce qui n'est pas reflété par le critère EQM. En notant $\hat{f}_{\epsilon_z}$ une version translatée (selon (Oz)) de $\hat{f}$ d'un pas $\epsilon_z \in \mathbb{R}$, nous définissons l'erreur de localisation suivante :

$$\epsilon_z(\hat{f}, f^\star) = \arg\min_{\epsilon_z \in \mathbb{R}} EQM(\hat{f}_{\epsilon_z}, f^\star). \tag{3.51}$$

Autrement dit ce critère met en avant la translation (selon (Oz)) minimisant l'EQM.

La figure 14 présente l'évolution de cette erreur de localisation en fonction du SNR pour deux objets respectivement positionnés en $z = 50$ nm et $z = 350$ nm. Ces courbes ont été obtenues pour un signal de fond $b_g = 2$ et moyennées sur 50 réalisations de bruit poissonien-gaussien. Les reconstructions $\hat{f}$ sont réalisées avec une discrétisation axiale $\delta_z = 20$ nm et

par résolution du problème avec contrainte de positivité (3.20) en utilisant l'algorithme CP, lorsque le terme d'attache aux données J_d est celui associé au bruit gaussien (quadratique), et l'algorithme RL, lorsque J_d correspond à la vraisemblance associée au bruit poissonien. Étant donné que la simulation $f^\star$ est quant à elle générée sur une grille de pas $\delta_z = 1$ nm, cette dernière est projetée sur les fonctions de base correspondant à $\delta_z = 20$ nm afin de calculer l'erreur (3.51).

On peut constater, d'après les courbes de la figure 14, que les résultats obtenus en considérant un bruit purement gaussien ou purement poissonien sont très similaires. Dans la suite, nous considèrerons donc principalement la modélisation gaussienne. Notons cependant que dans cette expérience nous avons simulé un bruit mixte poissonien-gaussien dont les contributions de chacun des deux bruits sont du même ordre de grandeur. C'est une hypothèse acceptable pour les applications réelles présentées dans les chapitres suivants étant donné que le signal des acquisitions est généralement (notamment pour les angles incidents les plus proches de l'angle critique) assez fort.

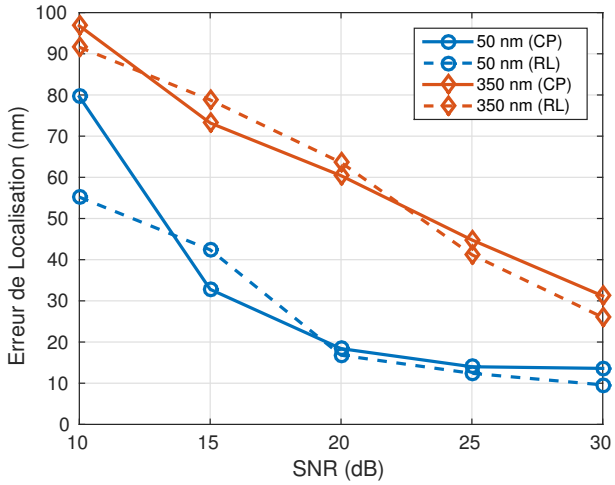


FIGURE 14 – Erreur de localisation moyenne en fonction du SNR pour deux objets localisés respectivement en $z = 50$ nm et $z = 350$ nm. Pour chaque valeur de SNR, 50 réalisations de bruit poissonien-gaussien sont générées et les reconstructions sont effectuées par minimisation du critère avec contrainte de positivité (3.20) en utilisant l'algorithme CP, lorsque J_d correspond à la vraisemblance pour une statistique gaussienne, et l'algorithme RL, lorsque J_d correspond au cas poissonien. Pour chacune des reconstructions ainsi obtenue, l'erreur de localisation (3.51) est calculée puis moyennée sur les 50 réalisations.

3.3 CONCLUSION

Dans ce chapitre, nous avons proposé de résoudre le problème inverse MA-TIRF avec une méthode variationnelle fondée sur une formulation bayésienne du problème. Après avoir discrétisé le modèle simplifié (2.1), nous avons présenté différents algorithmes qui sont à notre disposition pour résoudre les problèmes d'optimisation obtenus. Ensuite, nous nous sommes consacré à une étude 1D du problème afin de mettre en évidence l'effet que peuvent avoir les différents termes de régularisation considérés. Tout d'abord, nous avons montré qu'imposer la positivité de la solution nécessitait de prendre en considération le signal de fond présent sur les acquisitions. En effet, la meilleure interprétation d'un signal constant au sens des moindres carrés positifs avec le modèle MA-TIRF est un Dirac en zéro (proposition 3.1). Ceci se révèle être très problématique pour les reconstructions des objets simulés lorsque ce signal de fond n'est pas pris en compte. Nous avons alors proposé de

réaliser une estimation jointe de la densité de fluorophores f et du signal de fond b en imposant la positivité de ces deux inconnues.

Par ailleurs, une étude du terme de régularisation TV a montré que l'atténuation du contraste de la solution provoquée par ce dernier se traduisait par une perte de résolution axiale allant à l'encontre de l'objectif premier d'une telle reconstruction. Il est donc important, pour ce problème, de considérer avec précaution ce type de régularisation.

Enfin, nous avons comparé la minimisation des vraisemblances poissoniennes et gaussiennes sur des simulations où un bruit mixte poissonien-gaussien était ajouté aux acquisitions afin d'essayer autant que possible de simuler le type de bruit présent sur les acquisitions réelles. Les résultats obtenus ont montré qu'il y avait peu de différence entre les deux vraisemblances considérées pour la reconstruction.

4

ÉTALONNAGE DU SYSTÈME MA-TIRF

Sommaire

Comme nous avons pu le voir dans le chapitre 1, l'angle d'incidence α du laser d'excitation est un paramètre clé du modèle TIRF (1.8). En effet, la profondeur de pénétration de l'onde évanescente (1.3) est directement contrôlée par cet angle. D'autre part, la question de la validité du modèle simplifié (1.10) se pose. Dans le présent chapitre, nous proposons plusieurs expériences permettant à la fois d'étalonner le système, afin d'avoir un contrôle précis sur l'angle d'incidence, et de valider le modèle TIRF simplifié (1.10). Ce travail à fait l'objet d'une communication (Soubies et al., 2016a).

4.1 MESURER L'ANGLE D'INCIDENCE

L'angle d'incidence du laser est contrôlé par un miroir galvanométrique (voir section 1.1.3, page 13). Plus précisément, la position de ce dernier est fonction de la tension, notée U_{gv}, qui lui est appliquée ainsi que de certaines caractéristiques des éléments optiques constituants le système. À partir du schéma du microscope présenté sur la figure 4 (page 14), quelques calculs d'optique nous amènent à la relation

$$\alpha = \arcsin\left(\frac{\sin(KU_{gv})}{F_{obj}n_i}F_1\right), \tag{4.1}$$

où $F_{obj} = 2$ mm et $F_1 = 75$ mm sont respectivement les distances focales de l'objectif et de la lentille 3 (voir figure 4), $n_i = 1.518$ est l'indice de réfraction du milieu incident (verre) et K est une constante caractéristique du miroir galvanométrique spécifiée par le constructeur comme étant égale à $2°.V^{-1}$. Afin d'être en mesure de produire des reconstructions de qualité, il est donc essentiel de s'assurer que la relation (4.1) est bien vérifiée en pratique et si besoin d'ajuster certains paramètres (e. g. K) en fonction du système physique utilisé.

4.1.1 *Une figure caractéristique sur le plan focal arrière*

Sur le BFP de l'objectif, nous pouvons observer une configuration caractéristique qui est illustrée sur la figure 15. On observe ainsi deux spots lumineux correspondants res-

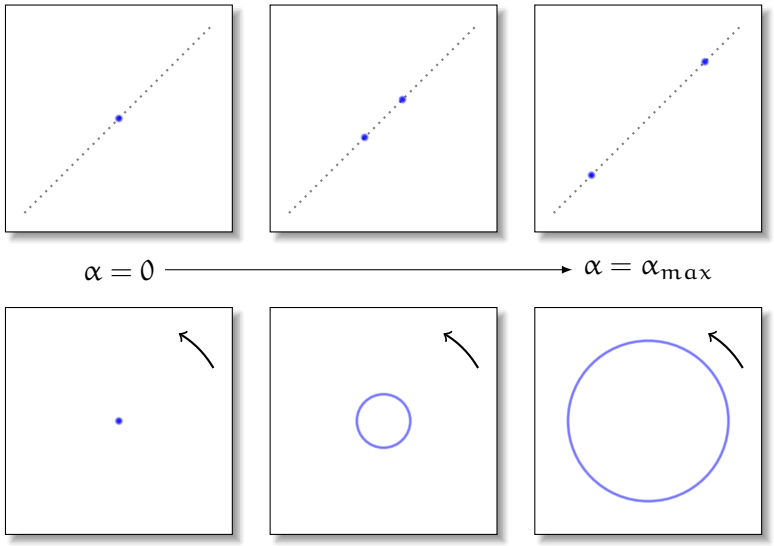


FIGURE 15 – Figure observée sur le plan focal arrière de l'objectif lorsque l'angle d'incidence α augmente. Haut : sans rotation azimutale, on voit les spots incident et réfléchi s'éloigner le long d'une droite symétriquement par rapport à leur position centrale. Bas : avec rotation azimutale, on observe un cercle dont le rayon augmente.

pectivement aux faisceaux incident et réfléchi du laser d'excitation. Ces deux spots sont superposés lorsque l'angle d'incidence est nul ($\alpha = 0°$) et s'éloignent le long d'une droite, symétriquement par rapport à cette position centrale, lorsque α augmente (ligne du haut sur la figure 15). Cependant, étant donné que notre système effectue une rotation azimutale (cf. section 1.1.3, page 13) lors de l'acquisition, nous observons un anneau dont le rayon croît avec l'angle d'incidence comme cela est présenté sur la figure 15 (bas). Dans ce qui suit, l'idée est donc d'extraire l'information contenue dans le BFP de l'objectif dans le but d'étalonner la relation tension/angle (4.1).

4.1.2 *Varier l'indice du milieu supérieur pour étalonner l'angle incident*

DOS SANTOS et al. (2014) ont récemment utilisé une stratégie originale afin de mesurer l'angle d'incidence. Leur idée consiste dans un premier temps à déposer une très fine couche fluorescente, ≈ 10 nm, de polyméthacrylate de méthyle dopé avec des points quantiques [1] sur la lamelle de verre, et ensuite à imager le BFP de l'objectif. Nous pouvons alors observer un anneau lumineux caractéristique de l'angle critique pour l'interface considérée. Or, à partir de l'indice du milieu supérieur n_t, nous savons que l'angle critique associé à cet anneau est donné par $\alpha_c = \arcsin(n_t/n_i)$. Ainsi, DOS SANTOS et al. (2014) proposent d'immerger cette fine couche fluorescente dans plusieurs solutions ayant des indices de réfraction n_t différents (air, eau, butanol), correspondant donc à différents angles critiques. Les acquisitions du BFP obtenues avec ces différentes configurations leur permettent ensuite de contrôler précisément l'angle d'incidence du laser sur l'échantillon.

Le dépôt de la fine couche fluorescente est réalisé grâce à la technique d'enduction par centrifugation (*spin coating* en anglais) nécessitant un peu de préparation. Dans la suite, nous proposons un protocole simplifié pour lequel uniquement des solutions ayant des indices de réfraction différents sont utilisées. La figure 16 présente l'évolution du BFP de l'objectif lorsque l'angle d'incidence croît pour une interface verre-air. On retrouve bien la configuration illustrée sur la figure 15.

1. *Quantum dots* en anglais utilisés en biologie comme des marqueurs fluorescents.

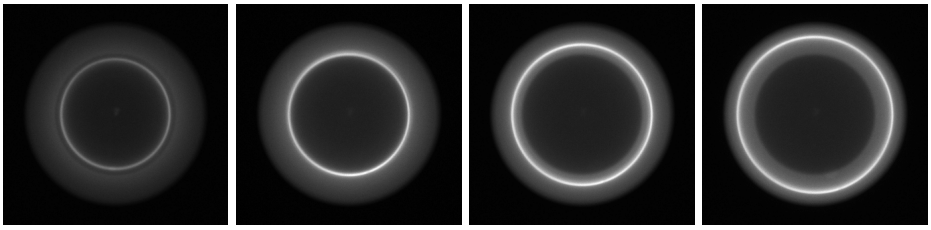


FIGURE 16 – Acquisitions du plan focal arrière de l'objectif pour différents angles d'incidence α pour une interface verre-air. De gauche à droite : angle croissant.

Par ailleurs, notons que l'intensité du cercle observé augmente nettement lorsque l'angle incident est supérieur à l'angle critique étant donné qu'il y a réflexion totale de la lumière incidente. Ainsi, moyenner les différentes acquisitions du BFP obtenues pour différents angles d'incidence $\alpha \in \mathcal{A}$ permet de mettre en évidence le cercle associé à l'angle critique comme nous pouvons le voir sur la figure 17. Sur cette figure, on distingue trois cercles caractéristiques associés respectivement à l'angle maximal α_{max} (vert), l'angle critique α_c (rouge) et le premier angle utilisé lors de l'acquisition [2] (bleu).

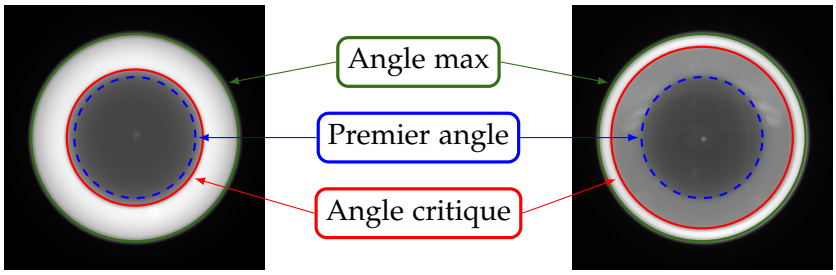


FIGURE 17 – Moyenne des images du plan focal arrière acquises pour différents $\alpha \in \mathcal{A}$. Gauche : interface verre-air ($\alpha_c = 41.2°$). Droite : interface verre-eau ($\alpha_c = 61.18°$).

4.1.2.1 *Extraction des caractéristiques*

L'idée consiste donc à extraire le rayon, noté r_c, du cercle correspondant à l'angle critique α_c pour différentes interfaces verre-milieu [3] afin d'obtenir une correspondance angle/rayon. La détection du cercle sur les images moyennées de la figure 17 est réalisée grâce à la transformation de Hough (DUDA et HART, 1972). Ensuite, nous recherchons dans le stack d'images du BFP (e. g. figure 16), pour un milieu supérieur donné, l'acquisition pour laquelle le rayon du cercle observé est le plus proche du rayon critique r_c précédemment déterminé. Cette acquisition ayant été obtenue pour une certaine tension U^c_{gv} appliquée au miroir galvanométrique, nous obtenons un couple (α_c, U^c_{gv}). Afin de déterminer le cercle observé sur chaque acquisition, nous employons encore une fois une méthode très simple basée sur la localisation des maxima locaux de l'image le long de droites radiales (passant par le centre du cercle correspondant à l'angle critique précédemment déterminé) comme sur la figure 18 (gauche). Ces maxima locaux correspondent à l'intersection des droites radiales avec le cercle recherché. Nous pouvons ainsi déterminer une estimation du rayon, pour chaque image du stack, par moyenne des rayons obtenus le long de chacune des droites utilisées. Quelques résultats de détection sont présentés sur la figure 18.

2. Qui n'est pas nécessairement $\alpha = 0°$.
3. Nous utilisons comme milieu supérieur : l'air, l'eau, et des solutions contenant différentes concentrations en sucrose.

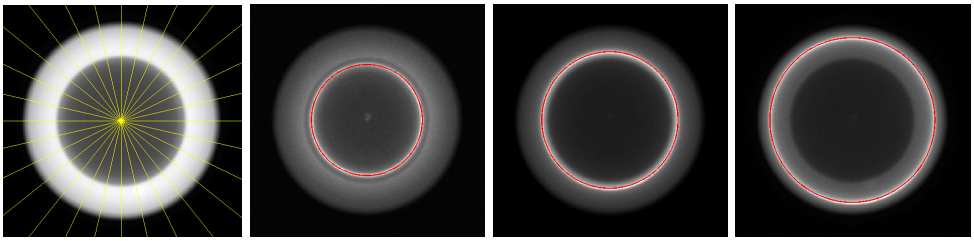


FIGURE 18 – Gauche : droites radiales utilisées pour détecter le cercle. Droite : résultats de détection pour quelques angles d'incidence.

4.1.2.2 *Ajustement du modèle*

Finalement, nous avons déterminé un ensemble de points $\{(\alpha_c, U^c_{gv})\}$, où chaque couple angle/tension correspond à une interface verre-milieu pour des milieux d'indices différents. La figure 19 présente les points obtenus pour deux longueur d'ondes, 546 nm (croix rouges) et 488 nm (cercles verts), du laser d'excitation ainsi que la courbe théorique (4.1) (courbe noire). Tout d'abord, notons que les points obtenus pour les deux longueurs d'ondes ne coïncident pas ce qui s'explique par le fait que l'achromaticité des éléments optiques présents dans le système n'est pas parfaite. Ensuite, il est clair que le modèle théorique (i. e. (4.1) avec $K = 2°.V^{-1}$) n'est pas adapté à notre système. Nous avons donc réalisé un ajustement du modèle (4.1) par rapport au paramètre K par moindres carrés. Les valeurs de K ainsi estimées, et qui seront utilisées dans la suite, sont respectivement $K = 1.967°.V^{-1}$ et $K = 1.952°.V^{-1}$ pour les longueurs d'ondes du laser incident 546 nm et 488 nm. Nous avons également réalisé des ajustements par rapport aux autres paramètres du modèle (4.1) (et aussi par rapport à plusieurs paramètres simultanément) et il s'est avéré que le paramètre K était le plus influent. C'est pourquoi nous ne présentons que l'ajustement du modèle par rapport à K. Enfin, notons que le manque de données entre les deux premiers points (de gauche à droite) sur la figure 19 est dû au fait que nous sommes dans l'impossibilité d'avoir un milieu dont l'indice est entre celui de l'air ($\simeq 1.0002$) et celui de l'eau ($\simeq 1.333$).

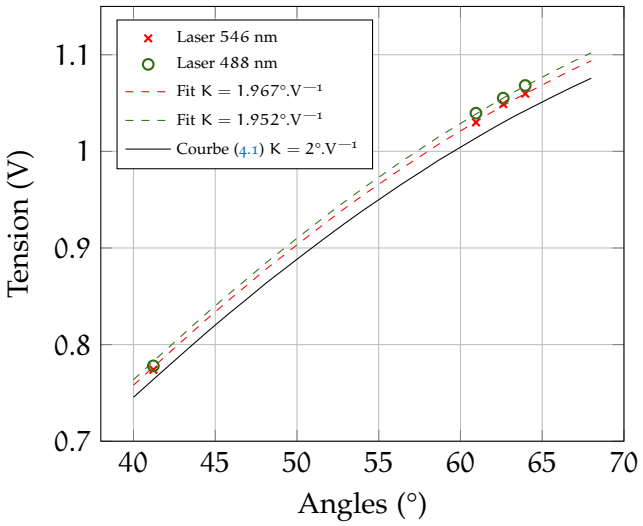


FIGURE 19 – Couples de points $\{(\alpha_c, U^c_{gv})\}$ calculés à partir de différentes interfaces verre-milieu pour deux longueurs d'ondes du laser incident : 546 nm (croix rouges) et 488 nm (cercles verts). Courbes théoriques (4.1) pour $K = 2°.V^{-1}$ (spécifications constructeur) en noir, $K = 1.967°.V^{-1}$ (ajustement du modèle aux données 546 nm) en tirets rouges et $K = 1.952°.V^{-1}$ (ajustement du modèle aux données 488 nm) en tirets verts.

4.2 VALIDATION DU PROFIL DE DÉCROISSANCE

Outre la nécessité de mesurer précisément l'angle d'incidence du laser, un autre point essentiel concerne la validation du modèle simplifié (1.10). En effet, la réponse d'un spécimen biologique au travers d'un système MA-TIRF est extrêmement liée à la décroissance de l'onde évanescente. Ainsi, la qualité des reconstructions numériques sera directement impactée par la capacité du modèle (1.10) (résultant de nombreuses simplifications) à décrire notre système MA-TIRF. Afin de valider ce modèle, l'une des problématiques principales concerne la conception d'un échantillon dont la géométrie est connue (échantillon dit *phantom*). Avec un tel échantillon, nous sommes en mesure d'évaluer la qualité des reconstructions réalisées. Plusieurs auteurs se sont penchés sur la question. Par exemple, Fiolka et al. (2008a) ou encore Boulanger et al. (2014) ont utilisé de petites billes fluorescentes disposées le long d'un plan incliné dont la pente est connue. Ainsi, connaissant la position latérale (i. e. dans le plan Ω) d'une bille, il est aisé d'en déduire sa profondeur afin d'étalonner/valider le modèle à partir des acquisitions multi-angles. Cependant, la mise en place d'une telle expérience est minutieuse, notamment pour bien maitriser la géométrie de l'échantillon. Dans la suite, nous proposons une méthode, similaire à (Ölveczky et al., 1997), pour laquelle la construction de l'échantillon *phantom* ne requiert qu'une lentille ainsi qu'une solution fluorescente homogène.

4.2.1 *Protocole expérimental*

Afin de construire notre échantillon de référence, nous avons immergé une lentille de rayon de courbure $R_c = 288.2$ mm et de diamètre $\varnothing = 25.4$ mm dans une solution fluorescente homogène comme cela est schématisé sur la figure 20. Étant donné que l'ordre de grandeur du diamètre de la lentille est très supérieur à la taille de la zone observée (en rouge sur la figure 20), nous considérons que le profil de la lentille est linéaire (cf. figure 20 centre). Des exemples d'acquisitions d'un tel échantillon pour différents angles d'incidence sont présentés sur la figure 20 (droite). Enfin, notons que certains auteurs (Ölveczky et al., 1997) ont également utilisé une approche similaire où la lentille est remplacée par une bille. Cependant, les expériences que nous avons pu mener avec un tel dispositif n'étaient pas convaincantes. En effet, les acquisitions étaient dégradées par des phénomènes de réflexion dus à la géométrie de la bille.

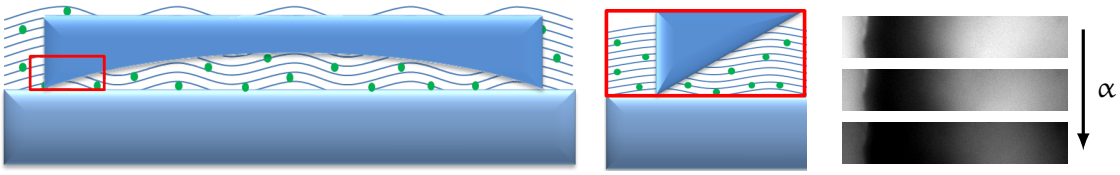


Figure 20 – Échantillon *phantom* construit à partir d'une lentille et d'une solution fluorescente homogène. La zone rouge définie la région observée par le microscope. La figure du centre illustre le fait que sur la région observée, le profil de la lentille est considéré linéaire. Des exemples d'acquisitions pour différents angles d'incidence sont présentés à droite.

4.2.2 *Méthode de reconstruction*

Étant donné la géométrie particulière de l'échantillon présenté dans le paragraphe précédent, nous allons utiliser un a priori de forme pour reconstruire cet échantillon à partir

d'acquisitions MA-TIRF. Pour chaque pixel $i \in \mathbb{I}_N$, on considère que la couche fluorescente $f(u_i, \cdot)$, où $u_i \in \Omega$ correspond au centre du i-ème pixel, peut être modélisée par une fonction Top-Hat par rapport à la variable z :

$$\forall z \in \mathbb{R}_+,\ f(u_i, z) = C_i \mathbb{1}_{\{z \leqslant \bar{Z}_i\}}, \tag{4.2}$$

où $C_i \in \mathbb{R}_+$ et $\bar{Z}_i \in \mathbb{R}_+$ définissent respectivement la «concentration» en fluorophores et l'épaisseur de solution comprise entre la lamelle de verre et la lentille pour le pixel $i \in \mathbb{I}_N$. La réponse du microscope (1.10) pour une telle densité de fluorophores est alors donnée par :

$$\forall i \in \mathbb{I}_N,\ \forall \alpha \in \mathcal{A},\ s(\alpha)_i = \frac{C_i}{p(\alpha)} I_0(\alpha) \left[1 - e^{-\bar{Z}_i p(\alpha)}\right] + b_i. \tag{4.3}$$

Ainsi, pour chaque pixel $i \in \mathbb{I}_N$ des acquisitions, une estimation de C_i, $\bar{Z}_i$ et b_i est réalisée par moindres carrés non-linéaires sous contrainte de positivité en utilisant une version de l'algorithme Levemberg-Marquardt adaptée à cette contrainte (Kanzow et al., 2004).

4.2.3 *Résultats*

Nous présentons dans cette section les résultats de reconstruction. Tout d'abord, observons l'ajustement du modèle (4.3) aux données mesurées. La figure 21 (ligne du haut) montre les ajustements obtenus pour certains pixels $i \in \mathbb{I}_N$. Au vu de ces figures, le modèle semble donc être particulièrement bien adapté pour représenter les données issues du système MA-TIRF que nous utilisons. Cependant, l'objectif principal étant la détermination de la position axiale de la lentille, il convient de s'intéresser plus précisément aux valeur $\bar{Z}_i$ estimées.

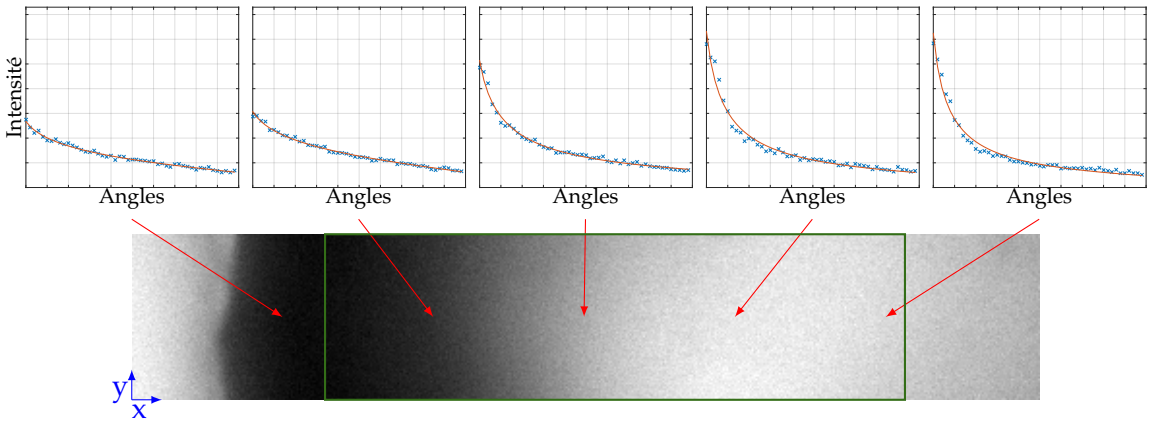


Figure 21 – Haut : Ajustement du modèle (4.3) aux données mesurées pour différents pixels $i \in \mathbb{I}_N$. Les croix bleues correspondent aux données et la courbe rouge au modèle ajusté (intensité en fonction des angles). Bas : une acquisition du *phantom* (figure 20) pour un angle incident $\alpha \in \mathcal{A}$.

Pour ce faire, nous avons calculé la valeur moyenne ainsi que l'écart-type des estimations de $\bar{Z}_i$ ($i \in \mathbb{I}_N$) le long de la direction Y dans la zone verte de la figure 21. En effet, de par la géométrie de la lentille, pour x fixé (sur la figure 21), la profondeur de la lentille est considérée constante par rapport à la variable y du plan image Ω.

Ces moyennes et écarts-types sont présentées sur la figure 22 où une droite a été ajustée en utilisant les points présents entre les deux droites verticales. Nous restreignons le «fit»

à ces points étant donné que la précision de la reconstruction diminue lorsque la distance lamelle-lentille augmente. En effet, on voit clairement qu'à partir d'une certaine profondeur, le profil estimé n'est plus linéaire. Par ailleurs, nous ne sommes pas non plus assuré du comportement linéaire de la lentille au voisinage de l'interface (i. e. $z = 0$) étant donné que le bord de la lentille est poli lors de la fabrication pour éviter d'éventuelles blessures lors des manipulations. [4]

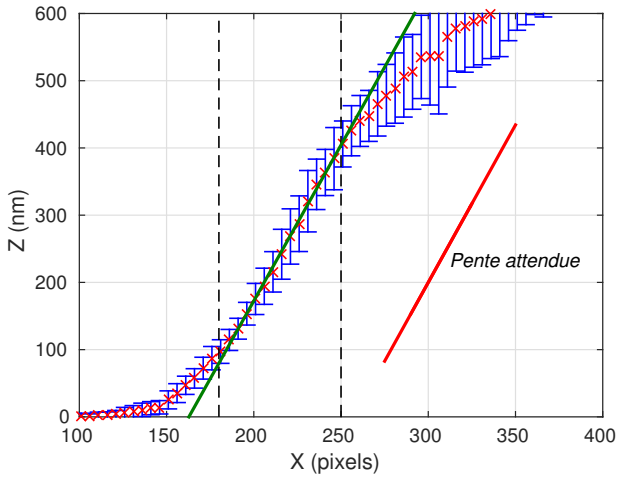


FIGURE 22 – Moyennes et écarts-types (le long de la direction Y) des estimés $\bar{Z}_i$ correspondants à la zone verte de la figure 21. La droite verte a été ajustée aux points présents entre les deux droites verticales. La pente de la lentille attendue, calculée à partir de la géométrie de cette dernière, est représentée en rouge.

Connaissant les caractéristiques R_c et $\varnothing$ de la lentille, nous estimons la pente que nous devrions reconstruire à 2.53° (en utilisant le fait que 1 pixel = 106 nm selon la direction X). Cette pente est en accord avec celle de la droite verte ajustée sur les données de la figure 22 déterminée à 2.51°. Ainsi le profil reconstruit suit précisément la pente calculée analytiquement d'après les caractéristiques de la lentille, jusqu'à une limite d'environ 400 nm. Notons que dans les expériences similaires réalisées par le passé, les résultats étaient généralement présentés seulement sur les 200 premiers nanomètres (BOULANGER et al., 2014 ; ÖLVECZKY et al., 1997). Cela illustre bien la limitation de ce type d'approches (reconstructions MA-TIRF) à localiser précisément les structures biologiques les plus éloignées. Enfin, comme évoqué précédemment, on observe une perte de précision au voisinage de l'interface qui, en plus des irrégularités de la lentille dues au polissage des bords, peut provenir de la non-uniformité de l'émission d'un fluorophore à proximité d'une interface diélectrique (HELLEN et AXELROD, 1987, et références associées) caractérisée par l'efficacité de collection [5] qui n'est pas prise en compte dans le modèle simplifié (1.10).

4.3 CO-LOCALISATION À DEUX COULEURS

Afin de compléter la validation du modèle et l'étalonnage du système, nous avons réalisé une dernière expérience visant à faire co-localiser deux reconstructions issues d'acquisitions indépendantes d'un même spécimen.

4. On voit très bien les irrégularités du bord de la lentille sur l'acquisition présentée sur la figure 21.
5. Cette terminologie est présentée en 1.1.4, page 15.

4.3.1 *Protocole expérimental*

Étant donné que le système TIRF, décrit dans la section 1.1.3 (page 13), permet d'utiliser plusieurs longueurs d'onde d'excitation et que le modèle (1.10) dépend de cette longueur d'onde, l'idée consiste à imager un échantillon pour lequel les structures d'intérêt ont été marquées par deux protéines fluorescentes, chacune sensible à une longueur d'onde particulière, et émettant respectivement dans deux couleurs différentes. Ainsi, nous obtenons deux acquisitions indépendantes des mêmes structures biologiques pour lesquelles les reconstructions doivent co-localiser. Le schéma présenté sur la figure 23 résume le principe de l'expérience réalisée dans cette section.

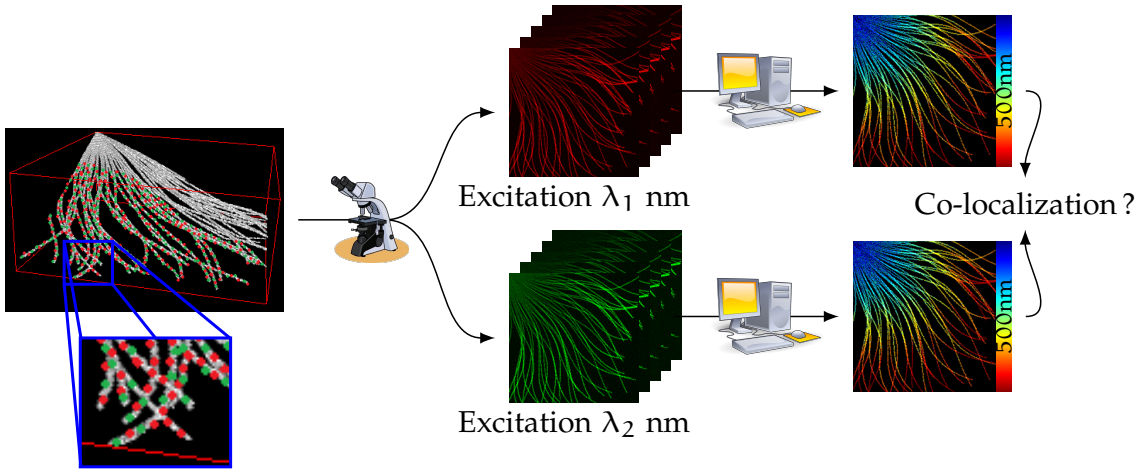


FIGURE 23 – Schéma du processus d'acquisition et de reconstruction pour l'expérience de co-localisation à deux couleurs. De gauche à droite : structures d'intérêt (tridimensionnelles) marquées par deux molécules fluorescentes différentes, chacune sensible à une longueur d'onde d'excitation particulière ; acquisitions pour les deux longueurs d'onde en question ; reconstruction numérique.

Pour cette expérience, les filaments d'actine de cellules endothéliales ont été marqués avec de la phalloïdine (toxine se liant à l'actine) couplée à deux types de fluorophores différents : Alexa Fluor 488 (fluorophores émettant dans le vert) et Alexa Fluor 546 (fluorophores émettant dans le jaune).

En ce qui concerne la méthode de reconstruction, étant donné que le signal sur les images MA-TIRF acquises est assez fort et très peu bruité, nous n'avons pas utilisé de régularisation TV et simplement minimisé le critère (3.20) dans le cas d'un modèle de bruit gaussien avec l'algorithme CP (voir section 3.1.4 page 34).

4.3.2 *Résultats*

Des résultats de co-localisation sont présentés sur la figure 24. Les reconstructions ont été réalisées pour une discrétisation axiale régulière de pas $\delta_z = 20$ nm et en considérant que l'indice de réfraction de l'échantillon (supposé être celui de l'eau [6]) est différent en fonction de la longueur d'onde d'excitation. En particulier, nous avons choisi $n_t = 1.34$ pour Alexa Fluor 488 et $n_t = 1.335$ pour Alexa Fluor 546 d'après (SEGELSTEIN, 2011). D'autre part, rappelons que nous utilisons des valeurs de K (voir équation (4.1)) différentes en fonction de la longueur d'onde d'après l'étalonnage de l'angle incident réalisé dans la section 4.1. Afin de visualiser les volumes reconstruits dont la dimension axiale est très inférieure aux

6. Nous sommes sur des échantillons fixés dont le cytoplasme a été remplacé par de l'eau.

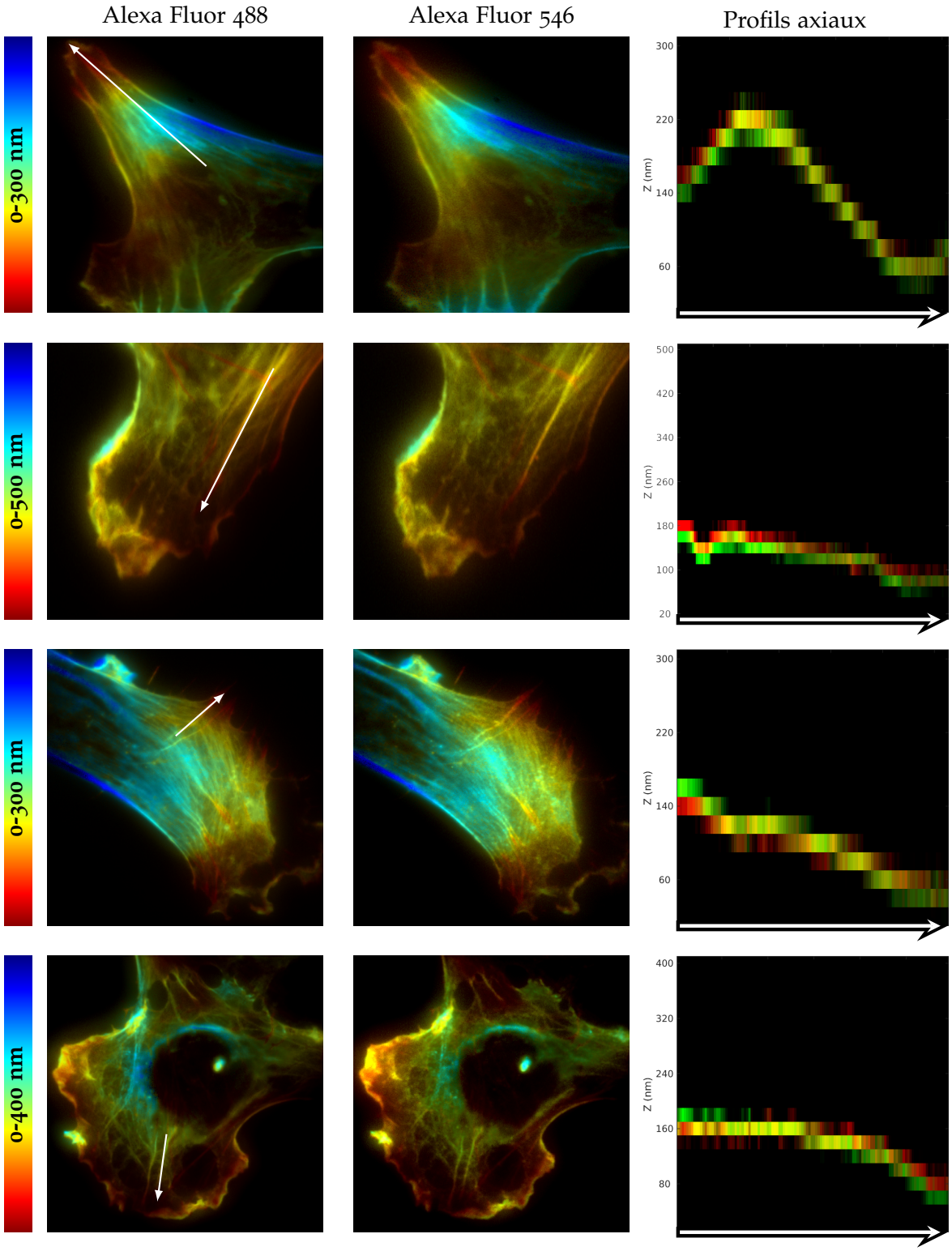


FIGURE 24 – Résultats de co-localisation. La colonne de gauche présente les reconstructions pour les acquisitions Alexa Fluor 488 alors que la colonne du milieu correspond à Alexa Fluor 546, les deux molécules marquant les mêmes structures d'intérêt (actine). Le code couleur caractérise la profondeur des objets (selon (Oz)) en utilisant la colorbar affichée (rouge $=0$ nm). Enfin, la colonne de droite montre la superposition des profils axiaux, extraits le long des flèches, pour les deux reconstructions (Alexa 488 est représenté en vert et Alexa 546 en rouge). Les reconstructions ont été réalisées par minimisation de (3.20) avec l'algorithme CP à partir de $L = 18$ angles incidents et en utilisant une discrétisation axiale régulière de pas $\delta_z = 20$ nm.

deux autres dimensions latérales, une méthode assez répandue consiste à représenter une carte de profondeur où la couleur est caractéristique de la distance des structures à l'interface. Une telle représentation est utilisée sur la figure 24 et nous avons également extrait quelques coupes axiales des différents volumes reconstruits le long de segments représentés par des flèches sur la figure 24.

Nous pouvons voir à travers ces résultats que nous obtenons bien l'effet escompté, à savoir la co-localisation des deux reconstructions obtenues à partir de deux acquisitions *indépendantes*. D'autre part, gardons à l'esprit que le modèle (1.10) dépend de la longueur d'onde d'excitation qui est différente pour les deux acquisitions réalisées dans cette expérience. Ainsi la concordance des reconstructions confirme la validité du modèle simplifié (1.10) pour décrire notre système MA-TIRF. Enfin, les profils extraits révèlent une co-localisation avec une précision d'environ 20-40 nm (correspondant à l'ordre de grandeur des fibres observées) sur une épaisseur variant de 300 à 500 nm selon les échantillons.

4.4 CONCLUSION

Dans ce chapitre, nous nous sommes intéressés à l'étalonnage et la validation du système qui sont deux étapes fondamentales en vue de la réalisation de reconstructions à partir d'acquisitions réelles. Nous avons ainsi proposé 3 expériences simples à mettre en œuvre en comparaisons avec des expériences similaires proposées dans la littérature :

1. étalonnage de l'angle incident à partir d'acquisitions du BFP de l'objectif pour plusieurs solutions aqueuses ayant des indices de réfraction différents. En particulier, cette expérience nous a permis d'ajuster la relation entre la tension appliquée au miroir galvanométrique et l'angle incident, nous permettant de contrôler précisément ce dernier ;
2. validation du modèle à partir d'un échantillon dont la géométrie est connue (construit avec une lentille de verre et une solution fluorescente homogène). La précision de la reconstruction obtenue pour un tel échantillon, sur une épaisseur de 400 nm adjacente à la lamelle de verre, a montré que le modèle simplifié (1.10) était suffisant pour décrire notre système MA-TIRF. À titre de comparaison, les expériences du même type que l'on peut trouver dans la littérature ne présentent que des résultats sur une couche de 200 nm d'épaisseur ;
3. confirmation de la validité du modèle avec une expérience de co-localisation qui a montré la concordance (avec une précision de l'ordre de 20-40 nm) entre deux reconstructions obtenues à partir de deux acquisitions indépendantes d'un même spécimen marqué par deux molécules fluorescentes sensible à différentes longueurs d'onde d'excitation.

5 APPLICATIONS EN BIOLOGIE

Sommaire

Dans ce chapitre nous présentons quelques résultats obtenus sur des données expérimentales pour l'étude du phénomène d'adhésion cellulaire et de l'assemblage de la matrice extracellulaire dans le contexte de l'angiogenèse [1] tumorale. Ce travail a été réalisé en collaboration avec l'équipe d'Ellen Van Obberghen-Schilling à l'iBV (Nice) qui s'intéresse à cette problématique. En particulier, la préparation des échantillons utilisés dans la suite [2] a été réalisée par Agata Radwanska et Dominique Grall.

5.1 UNE EXPLICATION SYNTHÉTIQUE DU PHÉNOMÈNE D'ADHÉSION CELLULAIRE

Le mécanisme d'adhésion cellulaire, pour les cellules endothéliales [3] considérées ici, est résumé de manière synthétique sur la figure 25 et suit les deux processus suivants :

1. *liaison de la cellule à la matrice extracellulaire* : cette liaison est réalisée à l'aide de protéines transmembranaires, les *intégrines*, dont une extrémité est capable de se lier à la matrice extracellulaire et l'autre extrémité peut se lier aux filaments d'actine constituant, avec d'autres structures, le cytosquelette de la cellule. Par ailleurs, les intégrines jouent aussi un rôle dans la mise en forme de la matrice extracellulaire, notamment pour son organisation sous forme de fibres ;
2. *assemblage de la fibronectine (fibrogénèse de la fibronectine)* : les cellules endothéliales sécrètent par exocytose une protéine, la *fibronectine*, qui est ensuite assemblée sous l'action de certaines intégrines en un maillage de fibres, appelé matrice extracellulaire, sur lequel la cellule va pouvoir «s'accrocher».

Notons que ces deux processus ne sont pas séquentiels, la cellule est en permanence en train de produire et de sécréter de la fibronectine, de l'assembler sous forme de fibres pour constituer la matrice extracellulaire, et de créer (renforcer) des adhésions.

Enfin, la formation de la matrice extracellulaire est encore aujourd'hui un phénomène qui n'est pas complètement compris. En particulier, plusieurs types d'intégrines sont connus, mais leur rôle précis dans la mise en forme de la matrice extracellulaire est quant à lui toujours incompris pour certaines d'entre elles.

1. Processus de croissance de nouveaux vaisseaux sanguins à partir des vaisseaux existants.
2. Et également ceux pour l'expérience de co-localisation présentée dans le chapitre 4.
3. Cellules qui tapissent les vaisseaux sanguins.

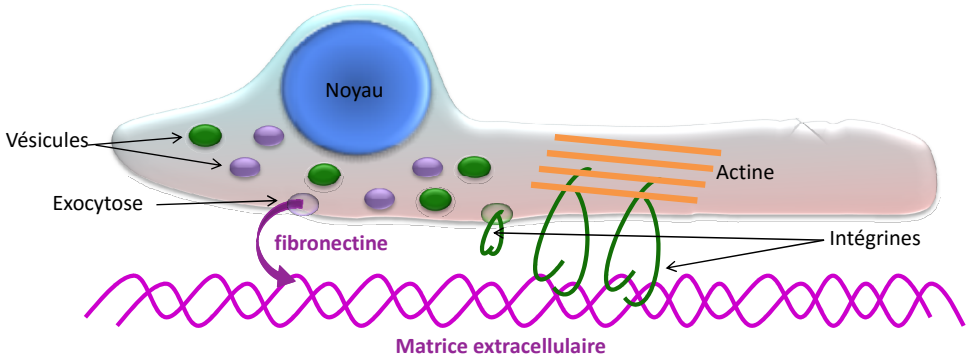


FIGURE 25 – Schématisation du processus d'adhésion cellulaire et d'assemblage de la matrice extracellulaire pour les cellules endothéliales.

5.2 EXPÉRIENCE RÉALISÉE

Nous avons vu dans le paragraphe précédent qu'il y avait plusieurs structures impliquées dans le mécanisme d'adhésion cellulaire. En particulier, nous avons mis l'accent sur la fibronectine, présente dans la matrice extracellulaire, les intégrines, localisées au niveau de la membrane cellulaire, et enfin les filaments d'actine, constituant en partie le cytosquelette de la cellule.

Plusieurs échantillons de cellules endothéliales ont donc été préparés en marquant la fibronectine, l'intégrine et l'actine avec trois fluorophores différents. Ces échantillons ont ensuite été imagés avec le système TIRF présenté dans le chapitre 1 étant donné qu'il permet l'utilisation de différentes longueurs d'onde d'excitation (voir section 1.1.3, page 13). Par ailleurs, l'un des objectifs étant la comparaison de différentes intégrines dans le but de comprendre leur rôle sur l'assemblage de la matrice extracellulaire, l'intégrine $\alpha_5\beta_1$ a été marquée pour certains echantillons alors que pour d'autres le marquage a été réalisé sur l'intégrine $\alpha_V\beta_3$.

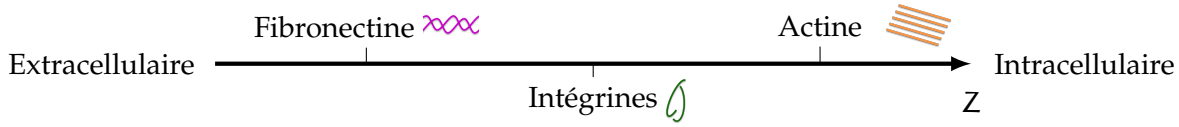
À partir des acquisitions MA-TIRF obtenues, les volumes 3D des échantillons ont été reconstruits, en considérant une discrétisation axiale $\delta_z = 20$ nm, avec les méthodes présentées précédemment. Nous précisons juste qu'ici une régularisation TV (avec un paramètre $\lambda_r = 1e^{-6}$) a été utilisée étant donné que le bruit présent sur les acquisitions était plus important que celui des acquisitions réalisées pour l'expérience de co-localisation, ou plutôt que le signal émis par les fluorophores était plus faible.

5.3 RÉSULTATS ET ANALYSE

Au moment de l'écriture de ce manuscrit, les reconstructions MA-TIRF sont encore en cours d'analyse afin de voir si les expériences réalisées permettent d'analyser le comportement des différents types d'intégrine et d'en définir le rôle. Cependant, quelques observations peuvent déjà être faites et constituent des résultats préliminaires qu'il faudra confirmer par la suite avec une analyse plus approfondie des reconstructions (e. g. extraction de données sur les différentes reconstructions pour faire des statistiques) ainsi que la réalisation d'expériences supplémentaires.

5.3.1 *Positions relatives entre fibronectine, intégrine et actine*

Une des premières observations que l'on peut faire sur de telles reconstructions, permettant de renforcer la validation de la méthode de reconstruction effectuée dans le chapitre 4, concerne la position relative des différentes structures imagées. En effet, on doit avoir (cf. figure 25) :



Les figures 26, 27 et 28, présentent des diagrammes décrivant les positions relatives d'un type de structures par rapport à un autre. Ces diagrammes ont été construits de la manière suivante. Soit $R^1 \in \mathbb{R}^{N\times M}$ et $R^2 \in \mathbb{R}^{N\times M}$ deux reconstructions obtenues pour deux types de structures différentes (e. g. actine et fibronectine). On définit alors les cartes de profondeur moyenne (en pixels) $CP^1 \in \mathbb{R}^N$ et $CP^2 \in \mathbb{R}^N$ par :

$$\forall i \in \mathbb{I}_N,\ CP^1_i = \left[\frac{1}{\sum_{j\in\mathbb{I}_M} R^1_{ij}} \sum_{j\in\mathbb{I}_M} j \times R^1_{ij} \right] \tag{5.1}$$

(respectivement pour CP^2 avec R^2), où $[\cdot]$ définit l'arrondi au plus proche entier. Autrement dit, pour chaque pixel $i \in \mathbb{I}_N$ du plan latéral des reconstructions, on calcule la profondeur moyenne (en pixels) des objets reconstruits. À partir de telles cartes de profondeur, nous pouvons maintenant définir un diagramme de position relative $D \in \mathbb{R}^{M\times M}$, entre les deux types de structures. Un tel diagramme doit être lu en sachant que l'axe des abscisses représente la profondeur du premier type de structures observées et l'axe des ordonnées représente la profondeur du deuxième type de structures (provenant du même échantillon mais imagées avec deux couleurs différentes).

Afin de construire D, nous parcourons les pixels $i \in \mathbb{I}_N$ des deux cartes CP^1 et CP^2 et en réalisons l'incrément :

$$D(CP^1_i, CP^2_i) = D(CP^1_i, CP^2_i) + \sqrt{\left(\sum_{j\in\mathbb{I}_M} R^1_{ij}\right) \times \left(\sum_{j\in\mathbb{I}_M} R^2_{ij}\right)}. \tag{5.2}$$

En d'autre termes, ce diagramme reflète le positionnement axial de R^1 par rapport à R^2 en tenant compte de l'intensité de ces derniers (si le signal est faible pour un pixel $i \in \mathbb{I}_N$, il contribuera peu au diagramme). La localisation des points les plus intenses de ce diagramme par rapport à la diagonale montre alors comment les deux type de structures se positionnent l'une par rapport à l'autre (selon (Oz)).

D'après les figures 26, 27 et 28 on peut constater que l'on a bien toujours la fibronectine plus proche de l'interface que l'intégrine qui est elle même moins profonde que les filaments d'actine. Notons également qu'entre ces différentes configurations, les fluorophores utilisés pour marquer les structures d'intérêt ont été intervertis. C'est-à-dire que pour certaines expériences, l'actine a été marquée par un fluorophore F_1, l'intégrine par F_2 et la fibronectine par F_3, alors que pour d'autres échantillons F_1 était utilisé pour la fibronectine et F_3 pour l'actine par exemple. Cela permet de montrer la robustesse de la méthode par rapport à la longueur d'onde d'excitation utilisée et vient conforter les expériences menées

dans le chapitre 4.

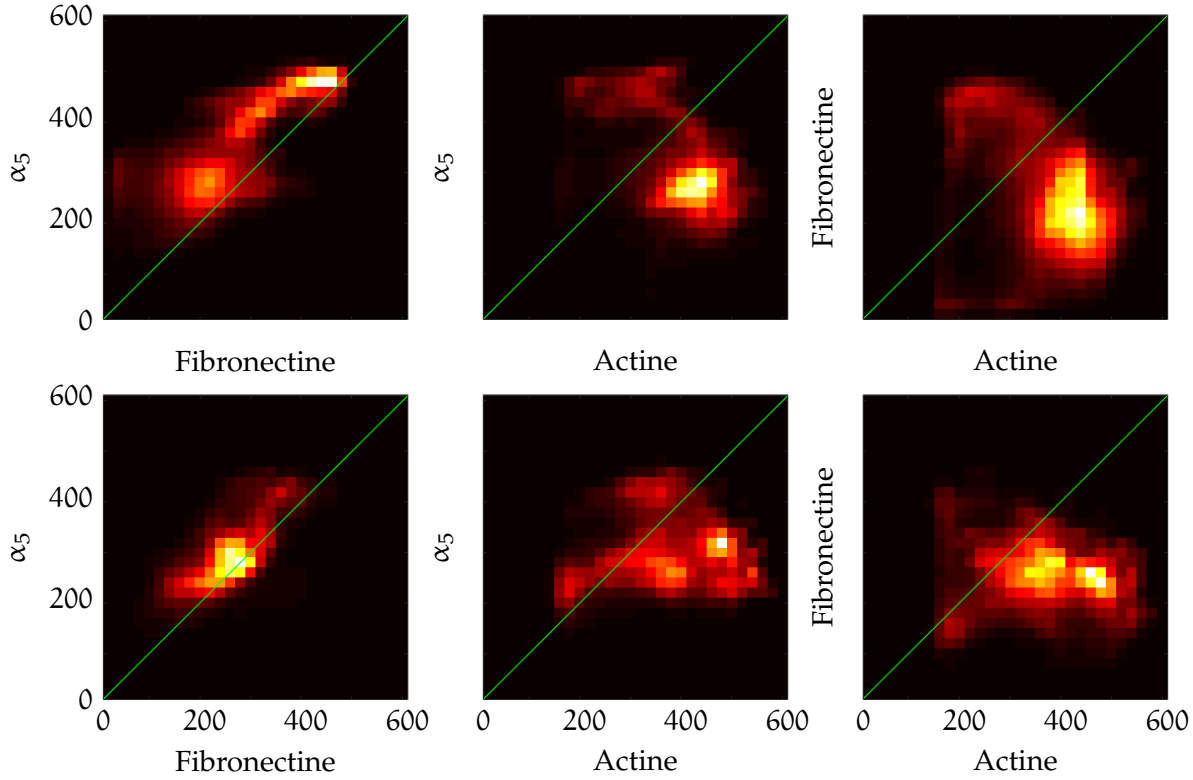


FIGURE 26 – Diagrammes des positions relatives entre l'actine, la fibronectine et l'intégrine $\alpha_5\beta_1$ après 6h de culture cellulaire. Chaque ligne représente un échantillon particulier.

Enfin, toujours sur les diagrammes des figures 26 et 27, nous pouvons remarquer que les structures observées semblent s'être rapprochées de l'interface (après 48h de culture), en particulier pour la fibronectine. En se rappelant que la fibronectine est sécrétée par la cellule elle-même, on peut donc expliquer ce phénomène par le fait qu'entre 6h et 48h de culture, la cellule a eu le temps de sécréter et d'assembler la fibronectine au niveau de la lamelle de verre (i. e. $z = 0$). Ainsi à 6h, la majeure partie de la fibronectine se trouve à l'intérieur de la cellule sous forme de vésicules alors qu'à 48h, une partie importante de celle-ci constitue la matrice extracellulaire au niveau de l'interface (voir par exemple les images sur la figure 29).

5.3.2 *Différences entre les intégrines $\alpha_5\beta_1$ et $\alpha_V\beta_3$*

Pour finir, nous nous intéressons plus particulièrement aux intégrines $\alpha_5\beta_1$ et $\alpha_V\beta_3$ dont le rôle (simplifié) est résumé ci-dessous (GEIGER et al., 2001) :

— $\alpha_V\beta_3$ permet à la cellule de réaliser des adhésions sur le substrat lorsque la fibronectine n'est pas encore assemblée. Ce sont des adhésions focales (visibles sous forme de tâches ovales) ;
— $\alpha_5\beta_1$ est responsable d'une part de l'assemblage de la fibronectine sous forme de fibre et d'autre part de la création d'adhésions fibrillaires stables (en comparaison avec celles produites par $\alpha_V\beta_3$) en réalisant un lien entre la matrice extracellulaire et les filaments d'actine à l'intérieur de la cellule. Cette intégrine est donc localisée le

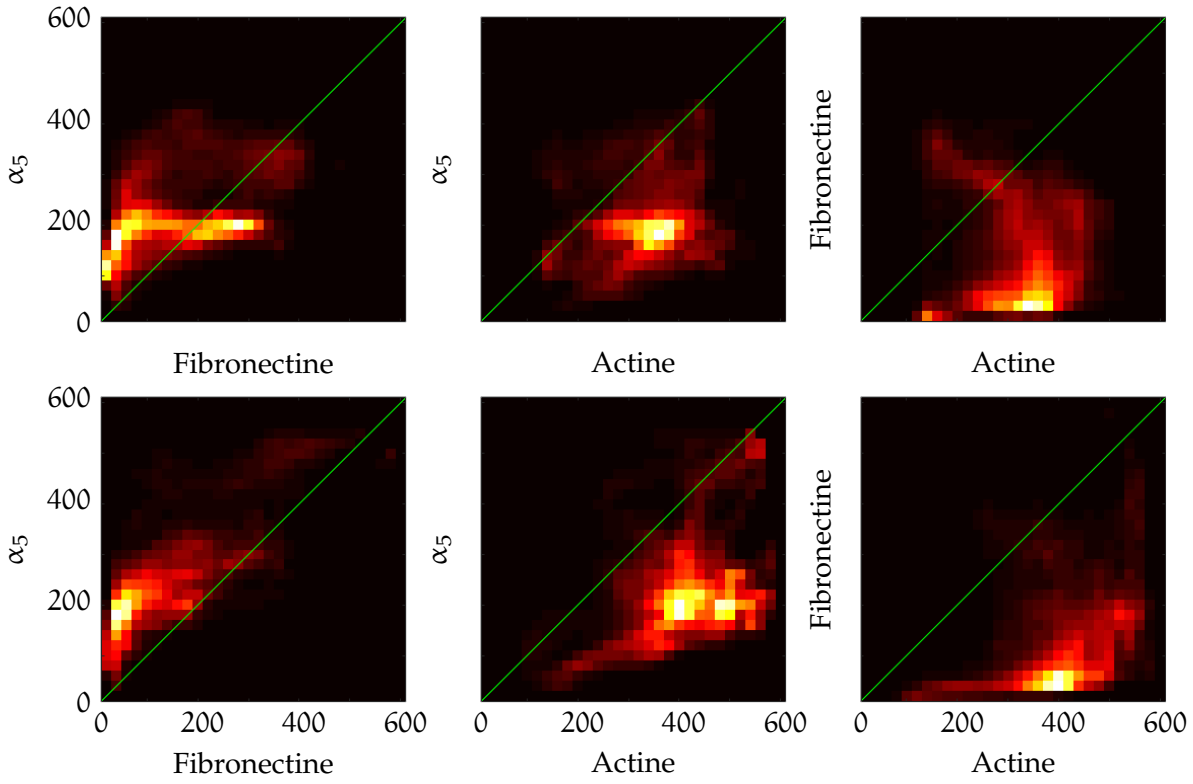


FIGURE 27 – Diagrammes des positions relatives entre l'actine, la fibronectine et l'intégrine $\alpha_5\beta_1$ après 48h de culture cellulaire. Chaque ligne représente un échantillon particulier.

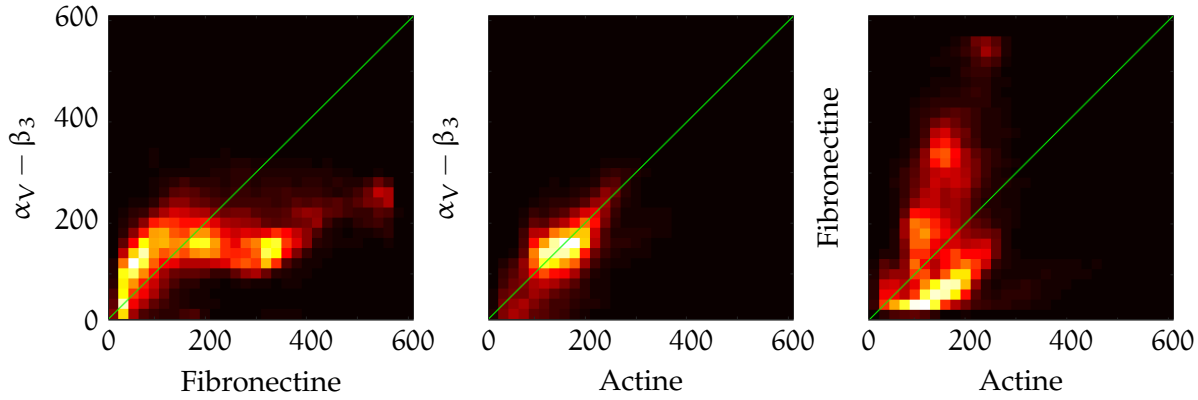


FIGURE 28 – Diagrammes des positions relatives entre l'actine, la fibronectine et l'intégrine $\alpha_V\beta_3$ après 48h de culture cellulaire.

long des fibres de fibronectine au dessus de celles-ci.

La figure 29 présente les quatre premières images du stack reconstruit pour les acquisitions à 48h où l'intégrine marquée est soit $\alpha_5\beta_1$ (image du haut), soit $\alpha_V\beta_3$ (image du bas). On peut alors observer que $\alpha_V\beta_3$ est présent sous forme de spots au niveau de l'interface alors que $\alpha_5\beta_1$ est localisée plus en profondeur, au dessus de la fibronectine, et suit la structure fibrillaire de cette dernière. Ces observations sont en accord avec la connaissance que nous avons du rôle de ces intégrines (décrit ci-dessus).

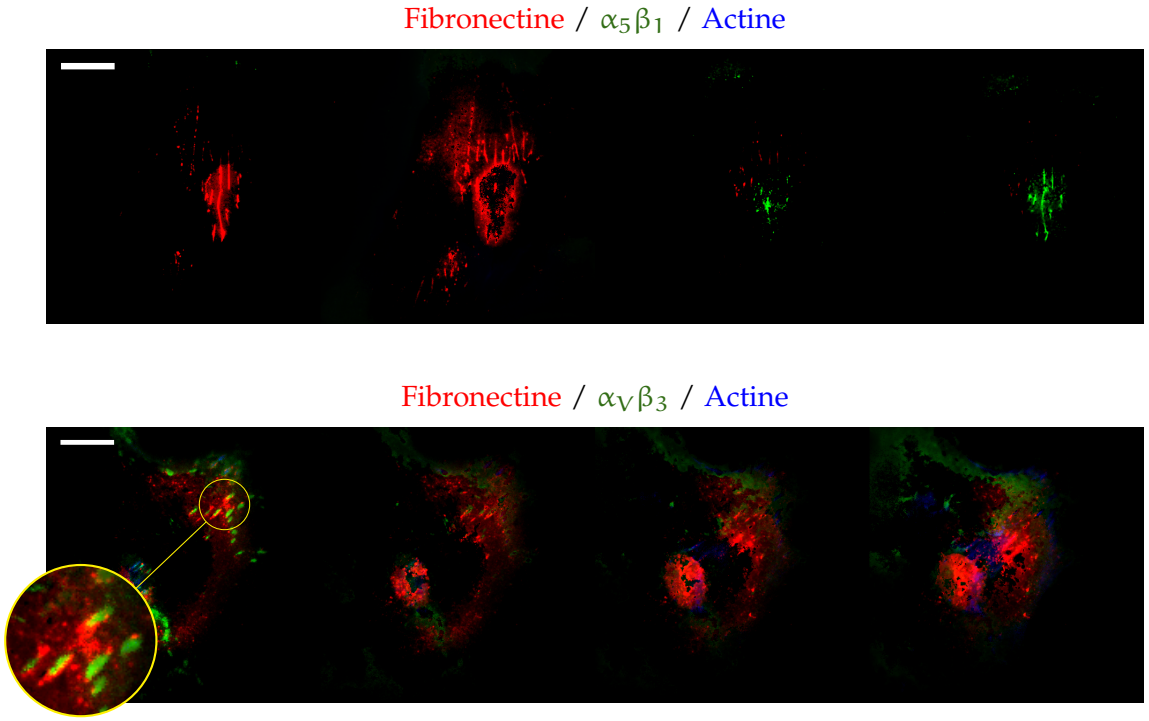


FIGURE 29 – Les quatre premières images du stack (de gauche à droite) reconstruit pour les échantillons à 48h où le marquage de l'intégrine est porté sur $\alpha_5\beta_1$ (haut) ou $\alpha_V\beta_3$ (bas). Sur ces images, l'actine est représentée en bleu, l'intégrine en vert et enfin la fibronectine en rouge. La barre blanche représente 10 µm.

5.4 CONCLUSION

Nous avons présenté dans ce chapitre une étude expérimentale préliminaire utilisant la reconstruction MA-TIRF dans le but de répondre à certaines questions biologiques concernant le mécanisme d'adhésion cellulaire. Nous avons montré d'une part que la position axiale relative des structures reconstruites les unes par rapport aux autres est cohérente avec leurs caractéristiques biologiques. En effet, les protéines constituant la matrice extracellulaire sont bien reconstruites en dessous (i. e. plus proche de l'interface) que les protéines situées au niveau de la membrane, elles mêmes sous les structures intracellulaires.

D'autre part, nous avons pu observer des comportements connus concernant les intégrines $\alpha_5\beta_1$ et $\alpha_V\beta_3$. Ces observations viennent compléter d'autres expériences menées précédemment par l'équipe d'Ellen Van Obberghen-Schilling pour lesquelles la troisième dimension (selon (Oz)) n'était pas disponible. Cette information sur la position axiale des structures a été rendu possible avec la méthode basée sur la microscopie MA-TIRF développée dans cette thèse. Ces résultats montrent que la méthode est capable de délivrer des informations tridimensionnelles précises avec trois couleurs différentes ce qui ouvre des possibilités intéressantes pour les biologistes cherchant à étudier les interactions entre différentes molécules à proximité de la membrane cellulaire.

6

CONCLUSIONS ET PERSPECTIVES SUR LA RECONSTRUCTION MA-TIRF

CONCLUSION

Cette partie de la thèse a été consacrée au développement d'une méthode de reconstruction tridimensionnelle à partir d'acquisitions MA-TIRF issues d'un système expérimental développé à l'iBV. Dans ce contexte, diverses contributions ont été apportées. Tout d'abord, après s'être placé dans un cadre bayésien et avoir formulé le problème de reconstruction comme la minimisation d'une fonctionnelle, nous avons proposé une étude 1D visant à mettre en avant les effets des différents termes de régularisation utilisés pour la résolution de ce problème inverse mal-posé. Cette étude a permis de révéler plusieurs comportements intéressants qu'il est important de prendre en compte lors de la reconstruction numérique des structures biologiques observées :

— *modélisation du signal de fond* : nous avons montré qu'un signal constant pour les différents angles d'incidence était vu par la reconstruction moindres carrés (positifs) à travers le modèle MA-TIRF comme un Dirac au niveau de l'interface (proposition 3.1). En partant de ce fait, il est alors indispensable de prendre en compte le signal de fond présent sur les acquisitions dans la méthode de reconstruction. Nous avons alors proposé de réaliser l'estimation jointe de la densité de fluorophores et du signal de fond. Des simulations numériques ont souligné le gain apporté par une telle modélisation.
— *régularisation TV* : une régularisation classique pour les problèmes inverses en signal/image est la régularisation de variation totale. Elle permet de favoriser les solutions constantes par morceaux et de préserver les contours. Cependant, elle est également connue pour atténuer le contraste de la solution. Dans le contexte de la reconstruction MA-TIRF, ce phénomène n'est pas sans conséquences et nous avons montré qu'il entrainait une perte de résolution axiale des objets reconstruits. L'objectif de ce problème étant de faire de la super-résolution axiale, nous en concluons qu'une telle régularisation doit être utilisée avec précaution dans la direction axiale.
— *Poisson Vs gaussien* : nous nous sommes également intéressés, dans cette étude 1D, aux différences entre l'utilisation d'une vraisemblance poissoniene ou gaussienne dans un contexte de bruit mixte poissonien-gaussien. Une expérience numérique en simulation a alors montré qu'il n'y avait pas de différence majeure à considérer une fonctionnelle plutôt que l'autre. Notons tout de même que dans ces simulations, nous nous sommes placés dans un contexte de bruit mixte poissonien-gaussien pour lequel les contributions des deux types de bruits sont du même ordre.

Dans un second temps, nous nous sommes intéressés à l'étalonnage du système et la validation du modèle qui sont deux étapes fondamentales dans le but d'effectuer des reconstructions sur des acquisitions réelles. Pour ce faire, nous avons proposé et mis en place plusieurs expériences. La première concerne l'étalonnage de l'angle incident du laser sur l'échantillon et est réalisée à partir d'acquisitions du plan focal arrière de l'objectif. À partir de telles acquisitions, nous sommes alors en mesure d'extraire les données nécessaires à l'ajustement d'un modèle théorique décrivant la relation entre l'angle incident et la tension appliquée au miroir galvanométrique contrôlant ce dernier. Concernant la validation du

modèle simplifié, nous avons réalisé l'acquisition d'un échantillon de géométrie connue pour lequel nous avons obtenus des résultats de reconstruction (en utilisant un a priori de forme) d'une précision de l'ordre de 20 nm sur un épaisseur de 400 nm. Ces résultats sont à mettre en perspective avec des expériences similaires de la littérature où les reconstructions ne sont pas présentées au-delà de 200 nm. Enfin, une expérience de co-localisation à deux couleurs a permis de conforter la validation du modèle ainsi que la méthode de reconstruction en étant capable de faire co-localiser deux reconstructions issues de deux acquisitions indépendantes des mêmes structures contenant un double marquage fluorescent.

Pour terminer, nous avons aussi confronté notre méthode à d'autres données expérimentales biologiques dans le contexte de l'adhésion cellulaire. Les reconstructions obtenues ont permis d'observer des phénomènes connus confortant une fois de plus la validité du modèle et de la méthode de reconstruction. Ces résultats ouvrent d'intéressantes perspectives pour d'autres applications en biologie nécessitant une telle modalité d'acquisition afin de visualiser différents phénomènes au niveau de la membrane cellulaire.

PERSPECTIVES

Outre l'utilisation de la méthode pour d'autres applications en biologie, une première perspective concerne l'accélération de l'algorithme de reconstruction. Les reconstructions présentées dans ce manuscrit sont réalisées avec Matlab en un temps de calcul d'environ 30 min pour des images de taille $(512 \times 512 \times 30)$. Vue l'expression du modèle TIRF (1.10), il est possible d'effectuer une importante parallélisation du code pour le calcul du modèle direct (et adjoint) étant donné que chaque pixel du plan Ω peut être calculé de manière indépendante ce qui suggère la possibilité d'améliorer de façon significative les temps de calculs en utilisant par exemple une implémentation GPU.

Comme nous l'avons déjà mentionné dans certains chapitres, une autre perspective concerne la prise en compte de la PSF dans le modèle simplifié utilisé pour la reconstruction. Pour que cela soit efficace, il sera alors nécessaire de modéliser précisément la PSF du système utilisé.

Enfin, la technique MA-TIRF utilisée ici ne permet qu'une amélioration de la résolution axiale. Afin d'obtenir des reconstructions avec une résolution isotrope dans toutes les directions, une idée consisterait alors à combiner les techniques MA-TIRF et PALM [1]. Une telle reconstruction nécessiterait de localiser les molécules activées dans le volume reconstruit à l'aide de méthodes d'optimisation parcimonieuses telles que celles qui font l'objet de la deuxième partie de ce manuscrit.

1. Voir la section 11.3 page 154 pour une description de la microscopie PALM.

Deuxième partie

RELAXATIONS CONTINUES EXACTES DU CRITÈRE MOINDRES CARRÉS PÉNALISÉ EN NORME ℓ_0

AVANT-PROPOS

L'approximation parcimonieuse est aujourd'hui un domaine de recherche en plein essor. Ces dernières années, de nombreuses contributions ont été apportées à ce problème qui est encore aujourd'hui un sujet de recherche très actif. Cet engouement pour cette problématique dans le domaine du traitement du signal et des images est en grande partie la conséquence des nombreuses applications que l'on y trouve de part la nature même des signaux qui nous entourent qui sont soit parcimonieux en tant que tels soit compressibles permettant ainsi de les représenter de manière parcimonieuse dans une base bien choisie. L'origine du travail sur ce sujet réalisé dans cette thèse étant le problème MA-TIRF étudié dans la première partie du manuscrit, nous nous sommes naturellement orientés vers le cas des problèmes où l'opérateur modélisant l'acquisition des données est complètement caractérisé par la physique du problème. Il est important de bien distinguer ce type de problèmes avec ceux rencontrés dans le contexte très répandu de l'échantillonnage compressé où la conception de la matrice d'échantillonnage (d'acquisition) concerne également une partie majeure du problème permettant alors de vérifier certaines hypothèses ayant conduit à d'important résultats théoriques qui ont valu à ce principe la popularité qu'on lui connaît.

Le problème central de cette partie concerne donc la minimisation du critère moindres carrés pénalisé en norme-ℓ_0 (critère mixte ℓ_2-ℓ_0) connu pour être NP-Difficile. Dans ce contexte, nous nous intéressons à des relaxations *continues* de la fonctionnelle et plus précisément aux liens qui existent entre les minimiseurs d'une telle relaxation et ceux de la fonctionnelle initiale. Ainsi, le travail réalisé se focalise sur l'étude et l'analyse de relaxations continues dites *exactes* dans le sens où elles ne changent pas le problème initial (mêmes minimiseurs globaux) et le rendent plus accessible de part la continuité apportée et l'élimination d'un certain nombre de minimiseurs locaux.

Après une introduction générale sur les signaux parcimonieux et les problèmes d'optimisation associés, ainsi qu'une une revue des méthodes existantes dans la littérature (chapitres 7 et 8), nous introduisons la pénalité CEL0 (Continuous Exact ℓ_0) dans le chapitre 9. Cette pénalité, déterminée dans le cas 1D et le cas ND orthogonal en calculant l'enveloppe convexe du critère ℓ_2-ℓ_0, est démontrée être une alternative continue *exacte* (au sens défini ci-dessus) de la nome-ℓ_0. La continuité du critère moindre carrés pénalisé résultant permet alors l'utilisation d'un certain nombre d'algorithmes récents d'optimisation non-convexe dont certains sont présentés dans le chapitre 10. Par ailleurs, nous proposons dans ce même chapitre des méthodes itératives basées sur des appels successifs de tels algorithmes afin d'améliorer la minimisation de la fonctionnelle ou encore de s'affranchir du choix du paramètre de régularisation. Le chapitre 11 est dédié à l'application de ces algorithmes à différents problèmes rencontrés en traitement du signal et des images. En particulier, nous mettons en évidence l'intérêt de minimiser la relaxation continue proposée en comparaison avec une minimisation directe du critère ℓ_2-ℓ_0. Enfin, une vue unifiée des pénalités continues approchant la norme-ℓ_0 est étudiée dans ce contexte de reformulation exacte (chapitre 12). Nous établissons alors des conditions *nécessaires et suffisantes* sur les pénalités afin d'assurer le caractère *exact* du critère pénalisé relaxé résultant.

7 INTRODUCTION

Sommaire

7.1 LES SIGNAUX PARCIMONIEUX

De nombreux phénomènes physiques ne sont pas directement observables et nécessitent la résolution d'un problème inverse afin de reconstruire le signal d'intérêt. Généralement, de part la nature même du phénomène, de tels problèmes sont mal posés [1] et peuvent ne pas admettre de solution unique mais aussi être très sensibles à de petites perturbations (e. g. bruits). Une manière d'aborder ce genre de problèmes consiste alors à ajouter une information *a priori*, i. e. à régulariser le problème. En particulier, une méthode très répandue vise à imposer la parcimonie de la solution. C'est le cas par exemple pour le problème de déconvolution d'impulsions (*spikes deconvolution* en anglais) que l'on rencontre en microscopie, en astronomie, mais aussi en géophysique. La séparation de sources en est un autre exemple qui se présente dans plusieurs domaines comme les télécommunications, le médical ou encore pour l'analyse de signaux audio. D'autre part, l'approximation parcimonieuse est largement utilisée en statistique pour la sélection de variables et l'apprentissage (*Machine learning*). Dans toutes ces applications, le signal recherché est parcimonieux en tant que tel. Cependant, la problématique de l'approximation parcimonieuse va au delà de ce type de signaux vraiment parcimonieux avec le constat que la plupart des autres signaux qui nous entourent sont généralement compressibles et peuvent alors être bien représentés par des signaux parcimonieux dans une base bien choisie. En traitement du signal et des images, ce concept de représentation parcimonieuse a pris de l'ampleur avec l'arrivée de l'échantillonnage compressé (*compressed sensing*, *compressive sampling* ou encore *compressive sensing* en anglais).

Dans ce contexte, l'idée consiste à compresser un signal dès son acquisition. Ainsi, plutôt que d'acquérir toute l'information contenue dans le signal pour le compresser par la suite, le principe repose sur l'acquisition d'un nombre limité de mesures tout en s'assurant d'être capable de reconstruire le signal original. Une motivation évidente est la réduction du temps d'acquisition. Le problème ici pose donc à la fois la question de l'échantillonnage (i. e. de la matrice d'échantillonnage) et celle de la reconstruction du signal, contrairement aux autres problèmes inverses présentés au début du paragraphe où l'opérateur à inverser est complètement déterminé par la physique du problème (e. g. déconvolution). Cette différence entre ces deux types de problèmes se révèle être très importante pour les algorithmes d'optimisation. En effet, en échantillonnage compressé la conception de la matrice d'acquisition est un aspect majeur qui peut être réalisé dans l'objectif de vérifier des propriétés/-hypothèses (voir section 8.1) assurant une reconstruction exacte du signal parcimonieux mesuré avec certains algorithmes. Malheureusement, pour les problèmes où la physique régit le modèle, les hypothèses utilisées en échantillonnage compressé sont généralement

1. Voir chapitre 1 pour une définition.

trop restrictives et d'autres algorithmes de reconstruction sont alors préférés. Dans la suite de ce manuscrit, nous nous plaçons dans le contexte de cette deuxième catégories de problèmes.

Nous considèrerons le cas d'observations linéaires suivant :

$$d = Ax^\star + \eta, \tag{7.1}$$

où $A \in \mathbb{R}^{M\times N}$ représente un dictionnaire d'atomes ou bien la matrice d'observation du signal (définie par la physique du problème), $d \in \mathbb{R}^M$ est un vecteur contenant les mesures (données), $x^\star \in \mathbb{R}^N$ définit le signal parcimonieux recherché et enfin $\eta \in \mathbb{R}^M$ est un bruit additif.

7.2 APPROXIMATION PARCIMONIEUSE : DIFFÉRENTES FORMULATION

Dans le but d'estimer le signal parcimonieux $x^\star \in \mathbb{R}^N$ à partir des données bruitées $d \in \mathbb{R}^M$, et étant donné que le système linéaire est généralement sous-déterminé ($M \ll N$), une méthode classique consiste à imposer une contrainte définie à partir de la pseudo-norme[2]-ℓ_0 :

$$\|x\|_0 := \sharp\{x_i, i \in \mathbb{I}_N : x_i \neq 0\}, \tag{7.2}$$

où $\sharp$ définit la cardinalité de l'ensemble et $\mathbb{I}_N := \{1, \dots, N\}$. En d'autre termes, cette norme compte le nombre de coefficients non nuls de x ce qui en fait une mesure de parcimonie que l'on peut qualifier «d'idéale». Lorsque que l'on se restreint à la droite réelle, on notera $|\cdot|_0$ la fonction dite «0-1» et définie par :

$$\forall u \in \mathbb{R},\ |u|_0 = \begin{cases} 0 & \text{si } u = 0, \\ 1 & \text{si } u \neq 0. \end{cases} \tag{7.3}$$

On a alors $\|x\|_0 = \sum_{i\in\mathbb{I}_N} |x_i|_0$.

On distingue plusieurs formulations du problème d'approximation parcimonieuse : les formes contraintes, utilisées lorsque l'on a une information sur la parcimonie du signal recherché ou bien sur la variance du bruit, ainsi que la forme pénalisée, utilisée lorsque nous n'avons pas accès à de telles informations. Par exemple, si l'on souhaite imposer une contrainte du type $\|x\|_0 \leqslant k \in \mathbb{N}$ sur la parcimonie de la solution, le problème s'écrit :

$$\hat{x} \in \arg\min_{x\in\mathbb{R}^N} \|Ax - d\|_2^2 \ \text{ s.c. } \ \|x\|_0 \leqslant k, \tag{C_k}$$

alors que si nous avons une connaissance sur le bruit nous allons plutôt préférer une formulation de la forme suivante :

$$\hat{x} \in \arg\min_{x\in\mathbb{R}^N} \|x\|_0 \ \text{ s.c. } \ \|Ax - d\|_2^2 \leqslant \epsilon, \tag{C_ϵ}$$

où $\varepsilon \in \mathbb{R}_+$ est déterminé par la variance du bruit η. Enfin, si nous n'avons aucune information de ce type, la forme pénalisée

$$\hat{x} \in \arg\min_{x\in\mathbb{R}^N} G_{\ell_0}(x) := \frac{1}{2}\|Ax - d\|_2^2 + \lambda\|x\|_0, \tag{P_λ}$$

2. Par abus de langage, nous parlerons dans la suite de norme-ℓ_0.

est généralement considérée. Ici, $\lambda \in \mathbb{R}_+$ est un paramètre de régularisation permettant de donner plus ou moins d'importance à la pénalisation ℓ_0. Enfin, on notera, pour un support donné $\omega \subset \mathbb{I}_N$, l'erreur quadratique minimale par

$$\mathcal{L}_2(\omega) := \min_{t \in \mathbb{R}^{\sharp\omega}} \|A_\omega t - d\|_2^2, \tag{7.4}$$

où $A_\omega := (a_{\omega[1]}, \ldots, a_{\omega[\sharp\omega]})$ définit la restriction de A aux colonnes indexées par les éléments de ω.

Tous ces problèmes sont connus pour être NP-Difficiles (Natarajan, 1995 ; Davis et al., 1997). Afin de préciser un peu plus cette notion, nous rappelons les différentes classes de complexité[3] (Cormen et al., 2009, Chapitre 34) :

— P : classe des problèmes pouvant être résolus en un temps polynomial ;
— NP : classe des problèmes dont une solution peut être vérifiée (mais pas forcément trouvée) en un temps polynomial. En particulier on a $P \subset NP$;
— $NP-Complet$ (NPC) : sous-classe de NP contenant les problèmes «les plus difficiles» de cette classe. Plus précisément, $\mathcal{P} \in NPC$ si
  1. $\mathcal{P} \in NP$,
  2. pour tout $\mathcal{P}' \in NP$, $\mathcal{P}'$ est réductible en temps polynomial[4] à $\mathcal{P}$;
— $NP-Difficile$: classe des problèmes «au moins aussi difficiles» que ceux de $NP-Complet$. C'est à dire les problèmes qui vérifient le deuxième point ci-dessus mais pas forcément le premier. En particulier, pour ces problèmes, une solution peut ne pas être vérifiable en un temps polynomial.

La figure 30 illustre les relations entre ces classes de complexité.

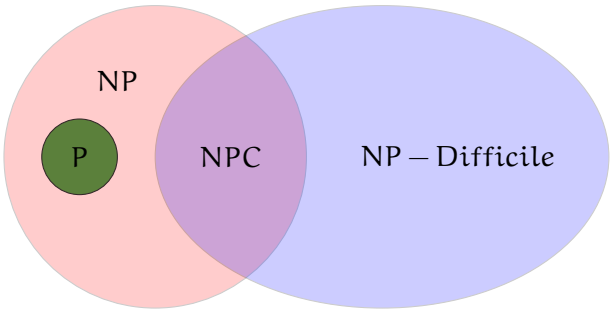


Figure 30 – Diagramme d'Euler pour les classes de complexité P, NP, $NP-Complet$ (NPC) et $NP-Difficile$.

Les problèmes (C_k), (C_ϵ) et (P_λ) sont donc extrêmement difficiles et à ce jour il n'existe pas d'algorithmes permettant de les résoudre efficacement sans restrictions particulières sur la matrice A. De plus, toujours dans un cadre général, il n'est même pas possible de vérifier la validité d'une solution de ces problèmes. Bien que la question de savoir si $P = NP$ est toujours ouverte, la majorité des chercheurs considèrent que de tels problèmes ne sont pas solvables.

> *Most theoretical computer scientists believe that the NP-complete problems are intractable, since given the wide range of NP-complete problems that have been studied to*

3. Initialement ces classes de complexité ont été introduites pour des problèmes de décision. Cependant, il est commun de les utiliser également pour décrire les problèmes d'optimisation étant donné les liens qu'il existe entre ces deux types de problèmes (Cormen et al., 2009, Chapitre 34).
4. Voir (Cormen et al., 2009) pour une définition.

date —— without anyone having discovered a polynomial time solution to any of them —— it would be truly astounding if all of them could be solved in polynomial time.

(Cormen et al., 2009, Chapitre 34)

Par conséquent, une multitude de travaux ont été consacrés d'une part à la détermination de conditions sur A pour lesquelles il existe des algorithmes efficaces ayant des garanties d'optimalité pour les problèmes (C_k), (C_ϵ) et (P_λ), et d'autre part au développement d'algorithmes (sous-optimaux) permettant d'approcher une solution de ces problèmes lorsque de telles conditions ne sont pas vérifiées. Une revue de la littérature à ce propos sera effectuée dans le chapitre 8.

Par ailleurs, de part la non-convexité de la norme-ℓ_0, (P_λ) n'est pas équivalent à (C_k) et (C_ϵ) (Soussen et al., 2015 ; Nikolova, 2016). En particulier, pour un vecteur $d \in \mathbb{R}^M$ donné, il peut exister des valeurs de k pour lesquelles un minimiseur global de (C_k) n'est pas un minimiseur global de (P_λ) pour tout $\lambda \in \mathbb{R}_+^\star$. On peut également relier les problèmes (C_k), (C_ϵ) et (P_λ) à l'optimisation bi-objectif où l'on cherche à minimiser à la fois l'erreur quadratique et la taille du support de la solution. Dans ce contexte, on s'intéresse généralement à la frontière de Pareto définie par les points pour lesquels il n'existe pas d'autres points permettant de décroître les deux critères simultanément (Marler et Arora, 2004). Dans le cas ℓ_2-ℓ_0 qui nous intéresse ici, nous pouvons voir que les solutions de (C_k) (resp. (C_ϵ)) pour différentes valeurs de $k \in \mathbb{N}$ (resp. $\varepsilon \in \mathbb{R}_+$) constituent la frontière de Pareto, pouvant être non convexe, alors que les solutions de (P_λ) appartiennent uniquement à l'enveloppe convexe de cette frontière (Soussen et al., 2015).

7.3 Quelques propriétés utilisées pour l'étude des garanties d'optimalité

Afin d'étudier les garanties d'optimalité de certains algorithmes, plusieurs propriétés/-hypothèses sur la matrice A ont été introduites par différents auteurs. Nous les répertorions ici à titre informatif.

Définition 7.1 (Cohérence de A (Donoho et Huo, 2001))**.** Pour une matrice $A \in \mathbb{R}^{M\times N}$ dont les colonnes sont normalisées, on appelle *cohérence* de A la quantité,

$$\mu(A) = \max_{1\leqslant i,j\leqslant N, i\neq j} |\langle a_i, a_j\rangle|, \tag{7.5}$$

où a_i ($i \in \mathbb{I}_N$) définit la i-ème colonne de la matrice A.

La cohérence reflète donc la dépendance entre les colonnes de la matrice.

Définition 7.2 (Spark (Donoho et Elad, 2003))**.** On appelle *spark* de la matrice $A \in \mathbb{R}^{M\times N}$ le plus petit nombre de colonnes de A linéairement dépendantes.

$$\text{spark}(A) = \min_{x\neq 0_{\mathbb{R}^N}} \|x\|_0 \ \text{ s.c. } \ Ax = 0. \tag{7.6}$$

Définition 7.3 (Null Space Property (NSP) d'ordre $k \in \mathbb{N}$ (GRIBONVAL et NIELSEN, 2003 ; COHEN et al., 2009)). Une matrice $A \in \mathbb{R}^{M \times N}$ est dite vérifier la NSP d'ordre $k \in \mathbb{N}$ avec la constante $\gamma \in (0, 1)$ si pour tous les sous-ensembles $\omega \subset \mathbb{I}_N$ tels que $\sharp\omega = k$ on a :

$$\|x_\omega\|_1 < \gamma \|x_{\omega^c}\|_1 \ \forall x \in \ker A \setminus \{0\}. \tag{7.7}$$

où $x_\omega = (x_{\omega[1]}, \ldots, x_{\omega[\sharp\omega]})$ et $\ker(A) := \{x : Ax = 0\}$.

Définition 7.4 (Restricted Isometry Property (RIP) d'ordre $k \in \mathbb{N}$ (CANDÈS et TAO, 2005)). Une matrice $A \in \mathbb{R}^{M \times N}$ vérifie la propriété RIP d'ordre $k \in \mathbb{N}$ si il existe une constante $\delta_k \in (0, 1)$ telle que

$$(1 - \delta_k)\|x\|^2 \leqslant \|Ax\|^2 \leqslant (1 + \delta_k)\|x\|^2, \quad \forall x \in \mathbb{R}^N \text{ t.q. } \|x\|_0 \leqslant k. \tag{7.8}$$

Cette condition assure que les sous-matrices de A formées en retenant k colonnes sont bien conditionnées et qu'elles se comportent comme un système orthogonal. Notons qu'il existe des relations entre ces différentes propriétés. En particulier, la condition RIP implique la NSP (CANDÈS, 2008, lemme 2.2) et la constante RIP est bornée par la cohérence μ (TROPP, 2006, proposition 21).

8

ÉTAT DE L'ART SUR L'OPTIMISATION PARCIMONIEUSE

Sommaire

Dans ce chapitre nous passons en revue différentes méthodes proposées dans la littérature pour l'approximation parcimonieuse. Étant donné l'étendue des travaux réalisés sur le sujet depuis plusieurs années, l'état de l'art effectué ici est loin d'être une revue exhaustive et nous renvoyons également le lecteur aux références des travaux cités. Nous distinguons cinq familles de méthodes :

1. relaxation convexe ℓ_1,
2. algorithmes gloutons,
3. algorithmes de seuillage itératif,
4. relaxations continues non-convexes (i. e. pénalités continues approchant la norme-ℓ_0),
5. reformulations «exactes».

Ces catégories constituent une certaine vision des algorithmes/méthodes existants et ne sont pas forcément toutes établies en tant que telles dans la littérature mais permettent, entre autres, d'introduire certaines idées qui seront développées dans les chapitres suivants. Notons aussi que certaines de ces classes peuvent se recouper.

8.1 RELAXATION CONVEXE ℓ_1

Une alternative à la norme-ℓ_0 consiste à considérer la relaxation convexe ℓ_1 des problèmes combinatoires présentés précédemment. Une telle relaxation a initialement été introduite sous le nom de Basis Pursuit (BP) (CHEN et al., 1998) définissant le problème

$$\arg\min_{x\in\mathbb{R}^N} \|x\|_1 \text{ s.c. } Ax = d. \tag{8.1}$$

Des versions prenant en compte des données bruitées ont également été proposées :

$$\arg\min_{x\in\mathbb{R}^N} \|x\|_1 \text{ s.c. } \|Ax-d\| \leqslant \varepsilon \quad \text{et} \quad \arg\min_{x\in\mathbb{R}^N} \frac{1}{2}\|Ax-d\|^2 + \lambda\|x\|_1, \tag{8.2}$$

où $\varepsilon \in \mathbb{R}_+$ et $\lambda \in \mathbb{R}_+$. Ces dernières sont également connues sous le nom de Basis Pursuit De-Noising (BPDN) ou LASSO en statistique (TIBSHIRANI, 1996).

Dans un contexte non bruité, afin de trouver la solution la plus parcimonieuse d'un système linéaire $Ax = d$, il est naturel de s'intéresser au problème

$$\arg\min_{x\in\mathbb{R}^N} \|x\|_0 \text{ s.c. } Ax = d. \tag{8.3}$$

En effet, si A vérifie la condition RIP d'ordre $2k$ avec la constante $\delta_{2k} < 1$, alors le problème (8.3) admet une *unique* solution k-parcimonieuse[1]. Ainsi, sous ces conditions, tout vecteur k-parcimonieux peut être exactement reconstruit par (8.3) (CANDÈS et TAO, 2005). Une autre condition pour qu'un signal k-parcimonieux x puisse être exactement reconstruit par (8.3) est que $2\|x\|_0 \leqslant \operatorname{spark}(A)$ (DONOHO et ELAD, 2003).

Par ailleurs, plusieurs auteurs ont montré que, sous certaines hypothèses sur A (plus restrictives que les précédentes), la résolution de BP permettait également de trouver la solution la plus parcimonieuse du système linéaire. Dans ce cas, BP et (8.3) sont équivalents.

Par exemple, DONOHO et ELAD (2003) et GRIBONVAL et NIELSEN (2003) ont montré que tout signal x^* vérifiant $\|x^*\|_0 < (1+\mu^{-1}(A))/2$ pouvait être exactement retrouvé par BP. Aussi, GRIBONVAL et NIELSEN (2003) ont utilisé la NSP afin de montrer l'optimalité de BP pour des signaux k-parcimonieux. Par ailleurs, CANDÈS (2008) montre le même résultat si A vérifie la condition RIP d'ordre $2k$ avec la constante $\delta_{2k} < \sqrt{2}-1$. Enfin, des résultats similaires impliquant la condition RIP ont également été montrés dans le cas de données bruitées, c'est-à-dire pour les problèmes relaxés (8.2) (CANDÈS et al., 2006 ; CANDÈS, 2008). Pour plus de détails sur le sujet, nous renvoyons le lecteur vers la vaste littérature comme par exemple les articles introductifs (CANDÈS et WAKIN, 2008 ; TROPP et WRIGHT, 2010 ; ELAD, 2010 ; FOUCART et RAUHUT, 2013) ou encore le site web `http://www.compressedsensing.com/`.

De nombreux algorithmes issus de l'optimisation convexe peuvent alors permettre de résoudre les problèmes relaxés (8.1) et (8.2). En particulier, ces problèmes peuvent être reformulés sous la forme de problèmes de programmation linéaire pouvant être ensuite résolus par l'algorithme du simplexe (DANTZIG, 1998), ou encore par les méthodes de points intérieurs (BOYD et VANDENBERGHE, 2004). Aussi, il est possible d'attaquer ces problèmes convexes avec l'algorithme FBS (COMBETTES et WAJS, 2005 ; COMBETTES et PESQUET, 2011).

1. Dont le nombre de coefficients non nuls est $k \in \mathbb{N}$.

Cet algorithme, également connu sous le nom de gradient proximal, est dédié aux problèmes du type

$$\hat{x} \in \arg\min_{x \in \mathbb{R}^N} J(x) := f(x) + g(x), \tag{8.4}$$

où f est une fonction différentiable et g est «simple» dans le sens où l'on peut obtenir une forme analytique de l'opérateur proximal défini par (Moreau, 1962) :

$$\text{prox}_g(y) = \arg\min_{x \in \mathbb{R}^N} \frac{1}{2}\|x-y\|^2 + g(x). \tag{8.5}$$

L'algorithme 2 présente les itérations de FBS.

Algorithme 2 : Forward-Backward Splitting (FBS) ou Gradient Proximal

Entrées : $\gamma \in \left]0, \frac{2}{L}\right[$, $x^0 \in \mathbb{R}^N$

1 **répéter**

2 $\quad x^{n+1} \in \text{prox}_{\gamma g}\left(x^n - \gamma \nabla f(x^n)\right)$;

3 **jusqu'à** *convergence*;

Sorties : x^n

Notons que plusieurs interprétations peuvent être données à ce schéma. Par exemple, il peut être vu comme un algorithme de Majorisation-Minimisation (MM) ou encore une méthode de point fixe (Parikh et Boyd, 2014). Lorsque le gradient de f est L-Lipschitz [2] et que les fonctions f et g sont convexes, FBS converges vers un minimiseur global du critère J dès lors que $\gamma \in \left]0, \frac{2}{L}\right[$ (Combettes et Wajs, 2005 ; Combettes et Pesquet, 2011). Dans le cas du problème (8.2), où $f(x) = \frac{1}{2}\|Ax-d\|^2$ et $g(x) = \lambda\|x\|_1$, l'algorithme est également connu sous de nom de Iterative Shrinkage Thresholding Algorithm (ISTA) (Figueiredo et Nowak, 2003 ; Daubechies et al., 2004 ; Bect et al., 2004) et on a :

$$\nabla f(x) = A^T(Ax-d), \tag{8.6}$$

et

$$\text{prox}_{\lambda\gamma\|x\|_1}(x) := \left(\text{sign}(x_i)(|x_i| - \lambda\gamma)_+\right)_{i \in \mathbb{I}_N}, \tag{8.7}$$

définissant le seuillage doux avec $(u)_+ = \max(0, u)$. Bien que cet algorithme soit très simple à mettre en œuvre, son principal inconvénient est que sa convergence vers un minimiseur $x^\star$ de J est lente. En effet, il a été montré que $J(x^n) - J(x^\star) \approx O\left(\frac{1}{n}\right)$. Des versions accélérées de cet algorithme ont alors été proposées par Beck et Teboulle (2009) avec le Fast Iterative Shrinkage Thresholding Algorithm (FISTA), ou encore par Nesterov (2013). Ces versions atteignent un taux de convergence de $O\left(\frac{1}{n^2}\right)$ en terme de décroissance de la fonction objectif. Cependant, la convergence de la séquence $(x^n)_{n\in\mathbb{N}}$ générée par ces algorithmes (plus particulièrement FISTA) n'a été démontrée que récemment (Chambolle et Dossal, 2015). L'algorithme 3 présente une version générale de FISTA.

Pour terminer, mentionnons que Efron et al. (2004) ont proposé la méthode Least Angle Regression (LARS) permettant de calculer un «chemin de régularisation» (i. e. l'ensemble des solutions pour un continuum de valeur de $\lambda \in \mathbb{R}_+$) pour le problème pénalisé (8.2).

2. I. e. il existe $L \in \mathbb{R}_+^\star$ tel que $\|\nabla f(x) - \nabla f(y)\| \leqslant L\|x-y\|$, $\forall (x,y) \in \mathbb{R}^N \times \mathbb{R}^N$.

Algorithme 3 : Fast Iterative Shrinkage Algorithm (FISTA)

Entrées : $\gamma \in \left]0, \frac{1}{L}\right[$, $x^0 \in \mathbb{R}^N$ et $(t_n)_{n\in\mathbb{N}}$ une séquence de réels positifs

1 **répéter**

2 $\quad x^n \in \text{prox}_{\gamma g}\left(y^n - \gamma \nabla f(y^n)\right)$;

3 $\quad y^{n+1} = x^n + \dfrac{t_n - 1}{t_{n-1}}(x^n - x^{n-1})$;

4 **jusqu'à** *convergence*;

Sorties : x^n

8.2 LES ALGORITHMES GLOUTONS

Dans cette section nous nous intéressons à une autre classe d'algorithmes très répandus pour l'optimisation parcimonieuse. Ce sont les algorithmes dits «gloutons» (*greedy* en anglais) dont l'idée est de mettre à jour, itérativement, un ensemble de composantes actives (i. e. correspondant au support de l'inconnue x) par ajout ou suppression de certains éléments. [3]

8.2.1 *L'algorithme Matching Pursuit (MP)*

L'algorithme Matching Pursuit (MP), proposé par MALLAT et ZHANG (1993), est sans doute la méthode gloutonne la plus simple et la moins coûteuse en ressources numériques parmi toutes les méthodes gloutonnes que nous présenterons dans la suite. Le problème étant de représenter un signal $d \in \mathbb{R}^M$ par une combinaison linéaire d'atomes d'un dictionnaire $A \in \mathbb{R}^{M\times N}$ (les atomes sont les colonnes a_i, $i \in \mathbb{I}_N$, de A et sont normalisées $\|a_i\| = 1$) :

$$d = \sum_{i\in\mathbb{I}_N} x_i a_i \iff d = Ax, \tag{8.8}$$

MALLAT et ZHANG (1993) proposent un algorithme itératif dont le but est de réduire au maximum l'erreur d'approximation à chaque itération en modifiant uniquement une seule composante de la représentation actuelle x^n.

Considérons, à l'itération n, que l'on a la représentation x^n et le résidu $R^n = d - Ax^n$. Alors, pour $i \in \mathbb{I}_N$, la réduction maximale du résidu en modifiant uniquement la composante x_i^n est donnée par

$$\min_{\alpha\in\mathbb{R}} \|R^n - \alpha a_i\|^2 = \|R^n - \langle a_i, R^n\rangle a_i\|^2 = \|R^n\|^2 - \langle a_i, R^n\rangle^2. \tag{8.9}$$

La composante i_n générant la plus grande décroissance du résidu est donc :

$$i_n \in \arg\max_{j\in\mathbb{I}_N} |\langle a_j, R^n\rangle|, \tag{8.10}$$

et la nouvelle représentation du signal d est donnée par

$$x^{n+1} = x^n + \langle a_{i_n}, R^n\rangle e_{i_n}, \tag{8.11}$$

3. Notons qu'à l'origine le terme *glouton* désigne des algorithmes ajoutant des atomes un à un au signal estimé. Cependant, par la suite, des algorithmes permettant la sélection comme la dé-sélection d'atomes ont été proposés et la dénomination «*glouton*» a été préservée.

où e_{i_n} représente le i_n-ème vecteur de la base canonique de $\mathbb{R}^N$. MP est résumé dans l'algorithme 4.

Algorithme 4 : Matching Pursuit (MP)

Entrées : dictionnaire $A \in \mathbb{R}^{M \times N}$, signal $d \in \mathbb{R}^M$, $\epsilon > 0$

1 $x^0 = 0_{\mathbb{R}^N}$;
2 $R^0 = d$;
3 **tant que** $\|R^n\| > \epsilon$ **faire**
4 $\quad i^\star \in \arg\max_{j \in \mathbb{I}_N} |\langle a_j, R^n \rangle|$;
5 $\quad x^{n+1} = x^n + \langle a_{i^\star}, R^n \rangle e_{i^\star}$;
6 $\quad R^{n+1} = R^n - \langle a_{i^\star}, R^n \rangle a_{i^\star}$;

Sorties : représentation x^n

Les auteurs ont montré que la norme du résidu $\|R^n\|$ converge exponentiellement vers 0 (Mallat et Zhang, 1993, lemme 2). Il est cependant à noter que la représentation x^n ne donne pas la meilleure approximation (Ax^n) du signal d que l'on peut obtenir par combinaison linéaire des éléments $(a_i)_{i \in \sigma(x^n)}$, où $\sigma(x^n)$ représente le support de x^n. En effet, Mallat et Zhang (1993) ont montré qu'une rétro-projection (*back-projection* en anglais), consistant à re-projeter le résidu R^n sur l'espace vectoriel engendré par les atomes $(a_i)_{i \in \sigma(x^n)}$, permettait de réduire l'erreur d'approximation.

Enfin, il est possible de définir, pour l'étape de sélection (8.10), d'autres fonctions de corrélation différentes du produit scalaire usuel $\langle \cdot, \cdot \rangle$. Par exemple, cela a été réalisé dans (Gribonval et al., 1996) où les auteurs montrent les bénéfices d'une nouvelle fonction de corrélation dans le contexte de la représentation temps-fréquence de signaux sonores.

8.2.2 *Les algorithmes Orthogonal Matching Pursuit (OMP), Orthogonal Least Squares (OLS) et leurs variantes*

ORTHOGONAL MATCHING PURSUIT Comme nous l'avons mentionné précédemment, un des principaux défauts de MP est qu'à chaque itération de l'algorithme, la représentation x^n obtenue est sous-optimale au sens où le résidu peut être diminué en re-projetant le signal d sur l'ensemble vectoriel généré par les atomes sélectionnés. Pati et al. (1993), tout comme Davis et al. (1994), ont exploité cette idée dans une version modifiée de MP connue sous le nom de Orthogonal Matching Pursuit (OMP)[4]. Le principe de OMP consiste donc à projeter, à chaque itération, le signal d sur l'ensemble vectoriel engendré par les atomes sélectionnés. Autrement dit, les lignes 5 et 6 de l'algorithme 4 deviennent respectivement

$$x^{n+1} \in \arg\min_{x \in \mathbb{R}^N} \|Ax - d\|^2 \quad \text{s.c.} \quad x_i = 0 \; \forall i \notin \omega := \sigma(x^n) \cup \{i^\star\}, \tag{8.12}$$

où $\sigma(x^n)$ représente le support de x^n, et

$$R^{n+1} = d - Ax^{n+1}, \tag{8.13}$$

Ainsi, à chaque itération, la représentation x^n fournie par OMP est optimale au sens où le résidu appartient toujours à l'espace supplémentaire orthogonal de l'espace engendré par les atomes sélectionnés. L'approximation (Ax^n) de d résultante est donc la meilleure que l'on puisse obtenir pour cette sélection d'atomes. Il s'en suit qu'un nouvel atome est alors sélectionné à chaque itération pour OMP contrairement à MP qui peut potentiellement

4. Ou encore *greedy algorithm* et *stepwise regression* dans d'autres domaines.

(re)sélectionner un atome déjà actif et faire une mise à jour de la composante de x^n correspondante. Par conséquent, OMP converge en au plus N itérations.

Pour une même itération $n \in \mathbb{N}$, OMP produit une meilleure approximation x^n de d (résidu plus faible) que MP, au prix d'un coût calculatoire plus important dû à la projection (8.12). Enfin, des garanties d'optimalité similaires à celle obtenues dans le cas de la relaxation convexe ℓ_1 ont été démontrées par Tropp (2004).

ORTHOGONAL LEAST SQUARES Cet algorithme, proposé par Chen et al. (1989) est très similaire à OMP. La différence réside dans le choix de la composante à ajouter au support de la solution courante. Alors que OMP utilise la même stratégie que MP, c'est à dire la sélection de l'atome du dictionnaire le plus corrélé avec le résidu courant (ligne 4 de l'algorithme 4), Orthogonal Least Squares (OLS) recherche l'atome maximisant la décroissance du résidu. En d'autres termes, à chaque itération, OLS sélectionne l'atome $a_{i^\star}$ tel que

$$i^\star \in \arg \min_{j \in \mathbb{I}_N \setminus \sigma^n} \mathcal{L}_2(\sigma^n \cup \{j\}), \tag{8.14}$$

où $\sigma^n = \sigma(x^n)$ correspond au support de x^n et $\mathcal{L}_2$ est définie en (7.4). Ainsi, à chaque itération, OLS effectue $N - n$ projections orthogonales de d sur les sous-espaces $\operatorname{span}(A_{\sigma^n \cup \{j\}})$ pour $j \in \mathbb{I}_N \setminus \sigma^n$. Notons que les auteurs dans (Chen et al., 1989) utilisent des implémentations astucieuses — basées sur des décompositions (Choleski, orthogonale, valeurs singulières...) de la matrice A — afin de résoudre efficacement ces problèmes de moindres carrés (8.14). Malgré cela, OLS reste bien plus coûteux que OMP mais fournit souvent une meilleure approximation x^n de d (pour une même valeur de $\|x^n\|_0$) que OMP. Une comparaison intéressante entre OMP et OLS ainsi qu'une revue des confusions que l'on peut rencontrer dans la littérature entre ces deux algorithmes peuvent être trouvées dans (Blumensath et Davies, 2007). Enfin, Soussen et al. (2013) ont étendu les travaux de Tropp (2004), concernant les garanties d'optimalité de OMP, au cas de OLS.

VARIANTES Tout d'abord, mentionnons qu'il existe de nombreuses variantes modifiant la règle de sélection de l'atome à insérer au support. Nous renvoyons par exemple le lecteur vers l'article de Temlyakov (2008).

D'autre part, afin d'améliorer les temps de calcul sur des problèmes de grande taille, Donoho et al. (2012) ont proposé l'algorithme Stagewise Orthogonal Matching Pursuit (StOMP) basé sur le même principe que OMP mais permettant de sélectionner plusieurs atomes à chaque itération. En effet, la mise à jour du support est effectuée par

$$\omega = \sigma(x^n) \cup \{i \in \mathbb{I}_N : |(A^T R^n)_i| > s\}, \tag{8.15}$$

où $s \in \mathbb{R}_+^\star$ est un seuil prédéfini. Ensuite x^{n+1} est calculé par (8.12) en remplaçant ω par la version (8.15). À partir de diagrammes de phase [5], Donoho et al. (2012) montrent que les performances de StOMP sont aussi bonnes, voire meilleures, que celles de BP et OMP pour un moindre coût calculatoire. Notons que d'autres variantes permettant la sélection de plusieurs atomes à chaque itération ont été proposées comme l'algorithme Generalized Orthogonal Matching Pursuit (gOMP) (Wang et al., 2012) ou encore l'algorithme Regularized Orthogonal Matching Pursuit (ROMP) (Needell et Vershynin, 2009).

Tous ces algorithmes du type OMP et OLS sont des algorithmes gloutons dits «forward», c'est à dire qu'ils ajoutent itérativement des composantes non-nulles à la solution. Inversement, bien que cela semble *a priori* moins intuitif étant donné que l'on recherche des solutions parcimonieuses, on pourrait imaginer un algorithme itératif «backward» fixant

5. Voir la section 10.5.3 page 135 pour une définition.

itérativement à zéro certaines composantes non nulles de la solution à partir d'une initialisation non parcimonieuse (i.e. $x_i^0 \neq 0 \ \forall i \in \mathbb{I}_N$). COUVREUR et BRESLER (2000) ont exploré cette idée et proposé le Backward Greedy Algorithm dont le principe consiste en la dé-sélection, à chaque itération, de l'atome du dictionnaire minimisant la croissance du résidu. En comparaison avec OLS, (8.14) devient,

$$i^\star \in \arg\min_{j \in \sigma^n} \mathcal{L}_2(\sigma^n \setminus \{j\}), \tag{8.16}$$

et l'atome $a_{i^\star}$ est alors dé-sélectionné. Dans le cas où la matrice $A \in \mathbb{R}^{M \times N}$ est surdéterminée ($M > N$), les auteurs montrent que le Backward Greedy Algorithm permet de retrouver exactement le support du signal $x^\star$ (problème (7.1)) lorsque le bruit est suffisamment faible, i.e. $\|\eta\| \leqslant \delta$. Notons que la borne δ est explicité dans (COUVREUR et BRESLER, 2000) et est fonction de la matrice A. En particulier elle décroît lorsque le conditionnement de A augmente.

8.2.3 *Extensions «forward-backward» de OMP et OLS*

Les algorithmes gloutons *forward* présentés précédemment peuvent être vus comme des algorithmes de descente pour le problème contraint (C_k) (SOUSSEN, 2013). En ce qui concerne la minimisation de G_{ℓ_0} (i.e. problème (P_λ)), des extensions bidirectionnelles (*forward-backward*) de OMP et OLS ont été proposées.

En particulier, dans un contexte bayésien, il a été montré par plusieurs auteurs (SOUSSEN et al., 2011 ; HERZET et DRÉMEAU, 2010) que le problème pénalisé (P_λ) pouvait être vu comme un cas limite de l'estimation MAP d'un modèle Bernoulli-gaussien. En effet, modéliser un signal parcimonieux $x \in \mathbb{R}^N$ par un processus Bernoulli-gaussien revient à définir :

- un vecteur aléatoire Bernoulli de paramètre ρ, $q \sim \mathcal{B}(\rho)$, représentant les composantes du support de x ;
- un vecteur aléatoire gaussien centré de variance σ_x^2, $r \sim \mathcal{N}(0, \sigma_x^2 I_N)$, définissant l'amplitude des entrées non nulles de x.

Ainsi, nous avons $x_i = q_i r_i$ pour tout $i \in \mathbb{I}_N$. Avec une telle modélisation du signal $x = (q, r) \in \{0,1\}^N \times \mathbb{R}^N$, l'estimation MAP pour le problème (7.1) revient à minimiser la fonction objectif suivante (KORMYLO et MENDEL, 1982) :

$$\mathcal{L}(q, r) := \|d - A\Delta_q r\|^2 + \frac{\sigma_\eta^2}{\sigma_x^2}\|r\|^2 + 2\sigma_\eta^2 \log(1/\rho - 1)\|q\|_0, \tag{8.17}$$

où σ_η^2 représente la variance du bruit (considéré gaussien) et Δ_q est une matrice diagonale ($N \times N$) dont les entrées sur la diagonale sont les q_i. Parallèlement, HERZET et DRÉMEAU (2010) ainsi que SOUSSEN et al. (2011) ont montré que le problème limite de (8.17), lorsque $\sigma_x^2 \to +\infty$ — i.e. aucune contrainte sur l'amplitude des composantes de x — était équivalent au problème (P_λ). Cela les a conduit à proposer des extensions bidirectionnelles respectivement de OMP et OLS que nous présentons dans les deux prochains paragraphes.

ALGORITHMES DE POURSUITE BAYÉSIENS Dans ce contexte bayésien, HERZET et DRÉMEAU (2010) proposent quatre algorithmes gloutons dans le but de maximiser la probabilité a posteriori du modèle Bernoulli-gaussien. [6] Ces algorithmes sont présentés comme des versions bayésiennes de MP, OMP, StOMP et CoSaMP et sont construits, comme les méthodes gloutonnes, selon deux étapes :

6. Ou de manière équivalente minimiser (8.17).

1. mise à jour du support ;
2. mise à jour des coefficients de la solution.

Ces deux étapes sont détaillées dans (HERZET et DRÉMEAU, 2010) ainsi que dans le rapport technique (HERZET et DRÉMEAU, 2014).

Une particularité des versions bayésiennes et qu'elles permettent la dé-sélection d'atomes, ce qui n'est pas toujours le cas des méthodes initiales «non-bayésiennes». Par ailleurs, il est à noter que les algorithmes proposés ne se placent pas dans le cas limite $\sigma_x \to +\infty$ comme présenté précédemment et permettent ainsi de prendre en compte un *a priori* sur l'amplitude des coefficients de la solution. Enfin, une autre flexibilité de cette approche est de pouvoir considérer un paramètre de Bernoulli ρ_i différent pour chacune des variables x_i ($i \in \mathbb{I}_N$) permettant d'imposer un *a priori* sur la parcimonie de la solution.

SINGLE BEST REPLACEMENT (SBR) Cet algorithme, proposé par SOUSSEN et al. (2011), est basé sur l'algorithme Single Most Likely Replacement (SMLR) initialement introduit par KORMYLO et MENDEL (1982) pour la minimisation de la fonctionnelle $\mathcal{L}$ définie en (8.17). Au vu de la relation entre (8.17) et (P_λ) détaillée précédemment, une adaptation naturelle de l'algorithme SMLR au problème (P_λ) a donc été proposée sous le nom de Single Best Replacement (SBR) et s'apparente à une extension «forward-backward» de OLS (en particulier lorsque $\lambda = 0$, les deux algorithmes coïncident).

Plus précisément, à chaque itération, SBR recherche le remplacement élémentaire — ajout ou retrait d'un atome du support courant — produisant la décroissance du critère objectif G_{ℓ_0} du problème (P_λ) (voir page 70) la plus importante. Cette recherche est faite de manière exhaustive en résolvant les N problèmes moindres carrés associés. Notons encore une fois qu'il existe des implémentations efficaces pour réaliser ces inversions récursivement dans ce contexte où seulement une modification élémentaire est effectuée (SOUSSEN et al., 2011). Enfin, des expériences numériques sur des problèmes de déconvolution et de détection de discontinuités ont montré la supériorité de SBR par rapport à OMP et OLS au prix d'un coût calculatoire plus important.

D'autres extensions *forward-backward* des algorithmes OMP et OLS ont également été proposées dans la littérature. Cependant ces méthodes n'ont pas été développées en tant qu'algorithmes de descente pour G_{ℓ_0} contrairement à SBR et aux algorithmes de poursuite bayésiens présentés précédemment. Par exemple, ZHANG (2011) a proposé l'algorithme FoBa pouvant être vu comme une modification forward-backward de OMP où soit une insertion soit un retrait est effectué à chaque itération en utilisant des règles de sélection similaires à OMP. Nous pouvons également mentionner l'algorithme Orthogonal Matching Pursuit with Replacement (OMPR) proposé par JAIN et al. (2011) basé sur le remplacement d'un élément du support par un nouvel élément à chaque itération. Il maintient ainsi toujours un support de taille $K \in \mathbb{N}$ et l'initialisation doit vérifier cette contrainte. Les auteurs des deux algorithmes FoBa et OMPR ont également montré des résultats d'optimalité sous la condition RIP. Enfin HAUGLAND (2007) a proposé une extension bidirectionnelle de OLS similaire à SBR, la différence résidant dans les stratégies de recherches et d'implémentations.

8.2.4 *Les algorithmes Continuation Single Best Replacement (CSBR) et ℓ_0-Path Descent (L0-PD)*

Nous avons vu que certains des algorithmes bidirectionnels (forward-backward) présentés dans la section précédente pouvaient être vus comme des algorithmes de descente pour G_{ℓ_0}. En pratique, pour un problème donné, une question fondamentale se pose : comment choisir la valeur de l'hyperparamètre λ ? Cela est loin d'être évident *a priori*. Une idée

consiste alors à estimer un «chemin de régularisation» pour G_{ℓ_0}, c'est à dire à rechercher un ensemble de solutions pour un continuum de valeurs de λ dans l'esprit de l'algorithme LARS proposé par EFRON et al. (2004) pour le problème ℓ_2-ℓ_1.

À partir du constat que l'ensemble des solutions de (P_λ) est constant par morceaux en fonction de λ, SOUSSEN et al. (2015) ont proposé deux heuristiques de recherche sous-optimales afin de minimiser G_{ℓ_0} pour un continuum de valeurs de λ. La première méthode est principalement une extension de l'algorithme SBR alors que la deuxième exploite le fait que la «courbe-ℓ_0», i. e. $\{\min G_{\ell_0}(x)\}$ fonction de λ, est affine par morceaux et concave. Nous présentons ces deux algorithmes dans les deux prochains paragraphes.

CONTINUATION SINGLE BEST REPLACEMENT Considérons $\lambda_n > 0$ et $\hat{\sigma}_n$ le support estimé par l'algorithme SBR pour cette valeur de l'hyperparamètre. Alors, en notant $\hat{\sigma}_n \pm \{i\}$ le remplacement élémentaire impliquant l'élément a_i (ajout si $i \notin \hat{\sigma}_n$ et retrait sinon), $\hat{\sigma}_n$ vérifie la condition d'optimalité suivante (SOUSSEN et al., 2015) :

$$\forall i \in \mathbb{I}_N,\ \mathcal{L}_2(\hat{\sigma}_n \pm \{i\}) + \lambda_n \sharp(\hat{\sigma}_n \pm \{i\}) \geqslant \mathcal{L}_2(\hat{\sigma}_n) + \lambda_n \sharp \hat{\sigma}_n. \tag{8.18}$$

En d'autres termes, tout remplacement élémentaire conduit à un accroissement de la fonction objectif. Ensuite, en remarquant que cette condition est également vérifiée pour tout $\lambda \in [\lambda^-, \lambda^+]$, où λ^- (resp. λ^+) est défini par

$$\lambda^- := \max_{i \notin \hat{\sigma}_n} \mathcal{L}_2(\hat{\sigma}_n) - \mathcal{L}_2(\hat{\sigma}_n \cup \{i\}),\ \left(\text{resp. } \lambda^+ := \min_{i \in \hat{\sigma}_n} \mathcal{L}_2(\hat{\sigma}_n \setminus \{i\}) - \mathcal{L}_2(\hat{\sigma}_n)\right), \tag{8.19}$$

SOUSSEN et al. (2015) définissent $\lambda_{n+1} = \lambda^- \leqslant \lambda_n$ et proposent d'exécuter à nouveau l'algorithme SBR initialisé avec $\hat{\sigma}_n \cup \{i^\star\}$, où $i^\star$ est l'argument du max défini en (8.19). Étant donné que $\mathcal{L}_2(\hat{\sigma}_n) + \lambda_{n+1}\sharp\hat{\sigma}_n = \mathcal{L}_2(\hat{\sigma}_n \cup \{i^\star\}) + \lambda_{n+1}(\sharp\hat{\sigma}_n + 1)$, les auteurs interdisent la dé-sélection de $i^\star$ lors de la première itération du nouvel appel à SBR. La génération de cette séquence décroissance $(\lambda_n)_{n\in\mathbb{N}}$ et des supports associés $(\hat{\sigma}_n)_{n\in\mathbb{N}}$ par appels successifs à l'algorithme SBR constitue l'algorithme Continuation Single Best Replacement (CSBR), initialisé par $\lambda_0 = +\infty$ et $\hat{\sigma}_0 = \emptyset$.

ℓ_0-PATH DESCENT Le second algorithme est motivé par le fait que la «courbe-ℓ_0» que l'on pourrait construire en sortie de CSBR (sous-optimale) n'est pas continue et concave alors que de telles conditions sont nécessaires pour l'optimalité du chemin de régularisation (SOUSSEN et al., 2015). L'idée n'est donc plus de déterminer successivement des solutions pour des valeurs de λ décroissantes mais de mettre à jour, itérativement, une liste de supports $S = \{\sigma_n\}$ associés à des valeurs critiques $\Lambda = \{\lambda_n\}$ (caractérisant un changement de support de la solution) allant de $+\infty$ à 0, de façon à décroître le polygone concave associé. [7]

Concrètement, à chaque itération de l'algorithme ℓ_0-Path Descent (L0-PD), un nouveau support candidat σ_{new} est proposé et intégré à l'ensemble S uniquement s'il permet de décroître le polygone concave. En d'autres termes, σ_{new} est intégré à S si la droite affine qu'il génère dans le plan $\{\min G_{\ell_0}(x)\}$ versus λ intersecte le polygone concave actuel. Dans ce cas, S ainsi que la liste des valeurs critiques Λ associée sont mis à jour. Concernant le choix du support σ_{new} à proposer, nous renvoyons le lecteur à (SOUSSEN et al., 2015). Enfin, il est

7. Dans le plan $\{\min G_{\ell_0}(x)\}$ versus λ, un support $\sigma \subseteq \mathbb{I}_N$ se traduit par une droite affine de pente $\sharp\sigma$. Ainsi, le polygone concave associé à S est défini comme l'enveloppe concave des droites affines supportées par les éléments de S (SOUSSEN et al., 2015). Les sommets de ce polygone sont les valeurs critiques de Λ pour lesquelles le support de la solution est modifié. Enfin, le polygone minimal (ou «courbe-ℓ_0») est donc l'enveloppe concave des droites affines supportées par l'ensemble des supports $\sigma \subseteq \mathbb{I}_N$.

à noter que L0-PD tire bénéfice des deux bornes λ^- et λ^+ définies en (8.19) contrairement à CSBR qui n'exploite que la borne λ^-. Ainsi L0-PD peut potentiellement «corriger» n'importe quelle solution du chemin de régularisation courant à chaque itération alors que CSBR calcule uniquement de nouvelles solutions pour des λ plus faibles. Cependant, la complexité de L0-PD est plus importante que celle de CSBR.

8.2.5 *L'algorithme Greedy Sparse Simplex (GSS)*

Pour terminer cette section sur les algorithmes gloutons, nous pouvons évoquer les travaux de Beck et Eldar (2013) pour le problème (C_k). Après avoir introduit différentes conditions d'optimalité :

— *basic feasibility*, généralisant la condition d'annulation du gradient des problèmes non-contraints différentiables ;
— *L-stationarity*, inspirée des conditions de stationnarité pour les problèmes contraints sur un ensemble convexe ;
— *optimalité Coordinate-Wise (CW)* (Voir section 9.3.1.4 page 108),

et étudié leurs relations, les auteurs en déduisent deux algorithmes.

Le premier n'est autre que le Iterative Hard Thresholding (IHT), qui sera présenté dans la prochaine section, pour lequel la convergence vers des points L-stationnaires est démontrée (Beck et Eldar, 2013, théorème 3.2).

Le second repose principalement sur une minimisation coordonnée par coordonnée,[8] modifiant ainsi un seul élément de la solution à chaque itération, ce qui lui a valu le nom de Greedy Sparse Simplex (GSS) par analogie avec l'algorithme du simplexe. Les auteurs montrent alors que les points d'accumulation de l'algorithme GSS sont des minimiseurs CW du problème (C_k) (Beck et Eldar, 2013, théorème 3.3). Cette condition d'optimalité étant la plus sévère des trois conditions proposées, GSS est donc potentiellement capable d'éviter des minima locaux (ceux qui ne sont pas CW). Cela est illustré par des exemples numériques de petite dimension.

8.3 LES ALGORITHMES DE SEUILLAGE ITÉRATIF

8.3.1 *L'algorithme Iterative Hard Thresholding (IHT)*

L'algorithme IHT a été introduit par Blumensath et Davies (2008). Il itère entre une étape de descente de gradient et une étape de seuillage dur comme cela est présenté dans le schéma général de l'algorithme 5 où le seuillage S est défini ci-après (équations (8.20) et (8.21)).

Algorithme 5 : Iterative Hard Thresholding (forme générale)

Entrées : $A \in \mathbb{R}^{M\times N}$, $d \in \mathbb{R}^M$, $\gamma > 0$, $x^0 \in \mathbb{R}^N$
1 **répéter**
2 $\quad x^{n+1} = S\left(x^n - \gamma A^T(Ax^n - d)\right)$;
3 **jusqu'à** *convergence*;
Sorties : x^n

8. Voir section 10.1.4 page 121.

Dans les travaux initiaux (BLUMENSATH et DAVIES, 2008), les auteurs proposent l'algorithme 5 avec $\gamma = 1$ ainsi que le seuillage

$$S_\lambda(x) := \left(x_i \mathbb{1}_{\{|x_i| > \sqrt{2\lambda}\}}\right)_{i \in \mathbb{I}_N} \tag{8.20}$$

pour la minimisation de (P_λ) et le seuillage

$$S_k(x) := \left(x_i \mathbb{1}_{\{|x_i| > \lambda_x^k\}}\right)_{i \in \mathbb{I}_N} \tag{8.21}$$

pour la minimisation de (C_k), où λ_x^k correspond à l'amplitude de la k-ème plus grande composante de x. La version utilisant le seuillage (8.20) peut également être trouvée dans des travaux antérieurs (HERRITY et al., 2006 ; BLUMENSATH et al., 2007).

La convergence de ces deux versions de l'algorithme vers des minimiseurs locaux respectivement de (P_λ) et (C_k) a été établie par BLUMENSATH et DAVIES (2008) lorsque $\|A\| < 1$.

Pour la version associée au problème (P_λ), cette contrainte sur l'opérateur A a récemment été relâchée par ATTOUCH et al. (2013) en considérant un pas de descente $\gamma \in]0, 1/\|A\|^2[$. Dans ce contexte, l'algorithme IHT n'est autre que l'algorithme FBS (présenté dans l'algorithme 2 page 77) appliqué à G_{ℓ_0} où :

$$\operatorname{prox}_{\gamma\lambda\|\cdot\|_0}(y) = \left(\operatorname{prox}_{\gamma\lambda|\cdot|_0}(y_i)\right)_{i \in \mathbb{I}_N} \quad \text{et} \quad \operatorname{prox}_{\gamma\lambda|\cdot|_0}(u) = \begin{cases} 0 & \text{si } |u| < \sqrt{2\gamma\lambda}\,, \\ \{0, u\} & \text{si } |u| = \sqrt{2\gamma\lambda}\,, \\ u & \text{sinon}\,. \end{cases} \tag{8.22}$$

La preuve de convergence repose sur le fait que G_{ℓ_0} vérifie l'inégalité de Kurdyka-Lojasiewicz (KL) [9]. Par une autre approche, et sans la nécessité pour G_{ℓ_0} de vérifier l'inégalité KL, KOWALSKI (2014) montre la convergence de cette version de l'algorithme pour n'importe quelle *règle de seuillage* S [10]. Aussi, MARJANOVIC et al. (2015) ont récemment proposé, et montré la convergence, d'une version accélérée de l'algorithme IHT pour le problème pénalisé (P_λ). Cette méthode s'apparente à une extension de la version accélérée FISTA (BECK et TEBOULLE, 2009) initialement proposée pour le cas convexe.

Enfin, l'algorithme 5 pour le problème (C_k) (i. e. utilisant le seuillage (8.21)) avec un paramètre γ différent de 1, ou même variable, a été étudié par BLUMENSATH et DAVIES (2010) et GARG et KHANDEKAR (2009). Là encore, la considération d'un pas de descente γ permet d'assurer la convergence de l'algorithme sans la nécessité d'avoir $\|A\| < 1$.

8.3.2 *Compressive Sampling Matching Pursuit, Subspace Pursuit et Hard Thresholding Pursuit*

Tout d'abord, mentionnons que les algorithmes présentés ici sont très liés à OMP, introduit dans la section précédente. Nous avons cependant choisi de ne pas les classer parmi les méthodes gloutonnes étant donné qu'ils ne partagent pas la structure d'ajouter (ou de retirer), à chaque itération, des composantes au support de la solution. En d'autres termes, à l'itération $n \in \mathbb{N}$, nous n'avons pas $\sigma^{n-1} \subset \sigma^n$ ou bien $\sigma^n \subset \sigma^{n-1}$, avec σ^n le support de la solution. Par ailleurs, ces algorithmes contiennent une étape de seuillage très similaire

9. Voir (ATTOUCH et al., 2013, Section 2.2) et les références associées pour plus de détails sur cette inégalité. La preuve de convergence proposée par ces auteurs rentre en fait dans un cadre plus large dont l'algorithme IHT en est un exemple.

10. Également des seuillages associés à d'autres pénalités (Voir section 8.4).

à celle de l'algorithme IHT et nous les classons donc dans la famille des algorithmes de seuillage itératif [11].

L'algorithme Compressive Sampling Matching Pursuit (CoSaMP), proposé par Needell et Tropp (2009), est devenu très populaire notamment pour les applications en échantillonnage compressé. Il est présenté dans l'algorithme 6. On peut voir qu'il a une structure très similaire à celle de OMP mais diffère selon deux aspects :

— plusieurs composantes sont sélectionnées à chaque itération (ligne 4) et sont intégrées au support de la solution courante (ligne 5) ;
— après projection orthogonale du signal d sur $\mathrm{span}(A_\omega)$ (ligne 6), x^{n+1} est défini en préservant les k coefficients de plus forte amplitude (ligne 7).

Par ailleurs, notons que les étapes 4 et 7 de l'algorithme sont définies par un seuillage dur, c'est pourquoi nous classons CoSaMP parmi les méthodes de seuillage itératif. Enfin, les auteurs montrent que CoSaMP possède des garanties d'optimalité du même ordre que BP mais à un moindre coût calculatoire, propriétés qui lui ont valu son succès.

Algorithme 6 : Compressive Sampling Matching Pursuit (CoSaMP)

Entrées : dictionnaire $A \in \mathbb{R}^{M\times N}$, signal $d \in \mathbb{R}^M$, niveau de parcimonie $k \in \mathbb{N}$

1 $x^0 = 0_{\mathbb{R}^N}$;
2 $R^0 = d$;
3 **répéter**
4 $\quad \tilde{\omega} = S_{2k}(|A^T R^n|)$ `/* où S_2k est défini par (8.21) et préserve les 2k composantes de plus forte amplitude. */`
5 $\quad \omega = \tilde{\omega} \cup \sigma(x^n)$;
6 $\quad y \in \arg\min_{x\in\mathbb{R}^N} \|Ax - d\|^2$ s.c. $x_i = 0\ \forall i \notin \omega$;
7 $\quad x^{n+1} = S_k(y)$;
8 $\quad R^{n+1} = d - Ax^{n+1}$;
9 **jusqu'à** *convergence*;

Sorties : x^n

Parallèlement, Dai et Milenkovic (2009) ont proposé un algorithme très similaire nommé Subspace Pursuit (SP). La principale différence avec CoSaMP réside dans le fait que SP sélectionne uniquement les k plus grands coefficients de $|A^T R^n|$ au lieu des $2k$ pour CoSaMP (ligne 4 de l'algorithme 6). Une analyse théorique avec des résultats similaires à ceux obtenus pour CoSaMP est aussi proposée par Dai et Milenkovic (2009).

Enfin, un peu plus récemment, Foucart (2011) a proposé l'algorithme Hard Thresholding Pursuit (HTP), une combinaison des algorithmes CoSaMP (ou SP) et IHT. Au lieu de sélectionner les k plus grandes entrées de $|A^T R^n|$, HTP sélectionne les k plus grands coefficients de $|x^n + \gamma A^T(d - Ax^n)| = |x^n + \gamma A^T R^n|$, pour $\gamma > 0$, en s'inspirant des itérés de IHT. Ainsi, HTP itère entre les deux étapes suivantes :

— $\omega = S_k\left(|x^n + \gamma(d - Ax^n)|\right)$;
— $x^{n+1} \in \arg\min_{x\in\mathbb{R}^N} \|Ax - d\|^2$ s.c. $x_i = 0\ \forall i \notin \omega$.

Nous renvoyons le lecteur à Foucart (2011) pour une étude de la convergence de l'algorithme ainsi que des garanties d'optimalité en terme de reconstruction exacte de signaux parcimonieux.

11. C'est un point de vue et ces algorithmes sont également souvent considérés parmi les méthodes gloutonnes dans la littérature.

8.4 RELAXATIONS CONTINUES NON-CONVEXES

Outre la non-convexité de la norme-ℓ_0, sa discontinuité en 0 est une difficulté supplémentaire dans la conception d'algorithmes pour minimiser les problèmes (C_ϵ), (C_k) ou (P_λ). Nous avons vu précédemment que la relaxation ℓ_1 pouvait être une très bonne alternative. Cependant, cette dernière introduit également un biais sur les coefficients estimés qui peut s'avérer problématique pour certaines applications (FAN et LI, 2001 ; ZOU, 2006) et les conditions (incohérence, NSP, RIP...) assurant l'optimalité de la solution ℓ_1 sont généralement trop restrictives pour de nombreux problèmes inverses mal posés. Divers travaux ont alors été consacrés à la définition et l'étude de pénalités *continues* approchant mieux la norme-ℓ_0 que ne le fait la norme-ℓ_1 au détriment de la convexité. À l'origine, de telles approximations ont principalement été étudiées en statistiques dans le contexte de la sélection de variables mais se sont rapidement répandues à d'autres domaines d'application, notamment en traitement du signal et des images. Dans la suite, nous présentons un certain nombre de ces pénalités que l'on peut trouver dans la littérature. Nous nous restreignons ici uniquement à la description des pénalités étant donné que les différents algorithmes permettant de minimiser les fonctionnelles sous-jacentes seront abordés dans le chapitre 10. Ces pénalités sont généralement séparables et s'écrivent sous la forme

$$\Phi(x) := \sum_{i \in \mathbb{I}_N} \phi(p_i; x_i), \tag{8.23}$$

où p_i représente un ou plusieurs paramètres pouvant être différents pour chaque composante et $\phi : \mathbb{R} \to \mathbb{R}$.

8.4.1 *Adaptative LASSO , NonNenegative Garrote et pénalité Log-sum*

Afin de réduire le biais introduit par la norme-ℓ_1, plusieurs auteurs ont proposé de pondérer différemment chacun des termes de cette pénalité. Autrement dit, ils définissent une version pondérée de ℓ_1 donnée par (8.23) avec

$$\phi(w; u) := w|u|, \tag{8.24}$$

où $w \in \mathbb{R}_+^\star$. Le problème est toujours convexe mais nous incluons cette méthode dans cette section étant donné que l'objectif de cette approche est de réduire le biais introduit par le LASSO. Ainsi ZOU (2006) propose l'Adaptative LASSO où les poids w_i ($i \in \mathbb{I}_N$) sont définis à partir d'une solution initiale $\hat{x}$ (par exemple l'estimation par moindres carrés) et d'un paramètre $\gamma > 0$ selon,

$$w_i := \frac{1}{|\hat{x}_i|^\gamma}. \tag{8.25}$$

L'idée ici est de pénaliser plus fortement les petits coefficients de $\hat{x}$. Notons, qu'en prenant $\gamma = 1$, $\hat{x}^{LS}$ la solution moindres carrés et en ajoutant une contrainte du type $x_i \hat{x}_i^{LS} \geqslant 0$, l'Adaptative Lasso n'est autre que le NonNegative Garrote précedemment proposé par BREIMAN (1995).

Par ailleurs, Candès et al. (2008) ont proposé un algorithme basé sur la résolution d'une séquence de problèmes ℓ_1 pondérés, connu sous le nom de Iteratively Reweighted ℓ_1 (IRL1), dont les poids à l'itération $n \in \mathbb{N}$ sont donnés par :

$$w_i^n = \frac{1}{|x_i^{n-1}| + \epsilon}. \tag{8.26}$$

Le gain en performance apporté par la (re)pondération par rapport à la relaxation ℓ_1 classique a été mis en évidence sur plusieurs exemples d'applications en signal/image. Les auteurs justifient le choix des poids (8.26) en montrant qu'utiliser une telle approche revient à minimiser la pénalité dite «Log-Sum», ayant la propriété de mieux approcher la norme-ℓ_0 que la relaxation ℓ_1 (voir figure 31), et qui est définie par

$$\phi_{Log}(\theta; u) := \log(1 + |u|\theta), \tag{8.27}$$

avec $\theta \in \mathbb{R}_+^\star$. Cette pénalité avait également été proposée quelques années auparavant par Weston et al. (2003) comme alternative continue à la norme-ℓ_0 dans le contexte de l'apprentissage statistique.

8.4.2 *Smoothly Clipped Absolute Deviation*

Selon Fan et Li (2001), une «bonne» fonction de pénalisation doit fournir une solution :

1. *non-biaisée* lorsque la variable inconnue prend de grandes valeurs ;
2. *parcimonieuse*, c'est à dire que la solution est donnée par un seuillage qui met directement les petits coefficients à zéro et simplifie ainsi la complexité du modèle ;
3. *continue* par rapport aux données afin d'avoir un modèle stable.

En se plaçant dans le cas orthonormé (i. e. $A^T A = I_d$, où I_d est la matrice identité) et en considérant une pénalité de la forme $\Phi(x) = \sum_{i \in \mathbb{I}_N} \phi(x_i)$, Fan et Li (2001) établissent trois conditions sur ϕ permettant de vérifier les propriétés précédentes :

1. $\phi'(u) = 0$ pour $|u|$ grand [suffisante pour avoir une solution (presque) non-biaisée] ;
2. $0 < \min_u \{|u| + \phi'(u)\}$ [nécessaire et suffisante pour que la solution résulte d'un seuillage] ;
3. $0 = \arg\min_u \{|u| + \phi'(u)\}$ [nécessaire et suffisante pour avoir la condition de continuité].

Notons que les deux dernières conditions imposent la non-différentiabilité de la pénalité ϕ à l'origine alors que la première condition impose à la pénalité d'être constante pour $|u|$ grand. La norme-ℓ_1, associée au seuillage doux (8.7), ne vérifie pas la première condition (biais pour les grands coefficients) contrairement à la norme-ℓ_0, associée au seuillage dur (8.20), qui ne vérifie pas la dernière condition (discontinuité du seuillage). Afin de vérifier les trois conditions simultanément, les auteurs proposent la pénalité Smoothly Clipped Absolute Deviation (SCAD) dont l'expression est donnée par :

$$\phi_{SCAD}(\gamma, \tilde{\lambda}; u) = \tilde{\lambda} \left\{ |u| \mathbb{1}_{\{|u| \leqslant \tilde{\lambda}\}} - \frac{\tilde{\lambda}^2 - 2\gamma\tilde{\lambda}|u| + u^2}{2(\gamma - 1)\tilde{\lambda}} \mathbb{1}_{\{\tilde{\lambda} < |u| \leqslant \gamma\tilde{\lambda}\}} + \frac{(\gamma + 1)\tilde{\lambda}}{2} \mathbb{1}_{\{|u| > \gamma\tilde{\lambda}\}} \right\}, \tag{8.28}$$

où $\gamma \in]2, +\infty[$ et $\tilde{\lambda} \in \mathbb{R}_+^\star$. Un exemple de graphe de ϕ_{SCAD} est représenté sur la figure 31.

8.4.3 *Minimax Concave Penalty*

Dans le même esprit, ZHANG (2010) a proposé la Minimax Concave Penalty (MCP) définie par :

$$\phi_{\text{MCP}}(\gamma, \tilde{\lambda}; u) := \tilde{\lambda} \int_0^{|u|} \left(1 - x/(\gamma\tilde{\lambda})\right)_+ dx \tag{8.29}$$

$$= \tilde{\lambda} \left(\frac{\gamma\tilde{\lambda}}{2} \mathbb{1}_{\{|u| > \gamma\tilde{\lambda}\}} + \left(|u| - \frac{u^2}{2\gamma\tilde{\lambda}} \right) \mathbb{1}_{\{|u| \leqslant \gamma\tilde{\lambda}\}} \right), \tag{8.30}$$

où $\gamma \in \mathbb{R}_+^\star$, $\tilde{\lambda} \in \mathbb{R}_+^\star$ (voir aussi la version «résumée» (ZHANG, 2008)). L'allure de ϕ_{MCP} est présentée sur la figure 31. Cette pénalité est celle minimisant le maximum de concavité κ défini par :

$$\kappa(\phi) = \max_{u \in \mathbb{R}_+^\star} -\phi''(u) \tag{8.31}$$

sous les contraintes,

$$\phi'(u) = 0 \quad \forall u \geqslant \gamma\tilde{\lambda}, \tag{8.32}$$

$$\phi'(0^+) = \tilde{\lambda}. \tag{8.33}$$

La contrainte (8.32) donne à la fonction de pénalisation ϕ la caractéristique nécessaire pour obtenir une solution *non-biaisée* alors que la contrainte (8.33) procure à la pénalité ϕ la caractéristique permettant d'obtenir une solution *parcimonieuse*. Afin de justifier l'intérêt de la pénalité MCP, ZHANG (2010) introduit la notion de «*sparse convexity*».

Définition 8.1 (Sparse convexity)**.** La fonctionnelle $x \mapsto \frac{1}{2}\|Ax - d\|^2 + \sum_{i \in \mathbb{I}_N} \phi(x_i)$ est dite *sparse convex* de rang $r^\star$ si :

$$\kappa(\phi) < \min_{\omega \subseteq \mathbb{I}_N, \sharp\omega \leqslant r^\star} \lambda_{\min}\left((A_\omega)^T A_\omega\right) \tag{8.34}$$

où $\lambda_{\min}((A_\omega)^T A_\omega)$ représente la plus petite valeur propre de la matrice $(A_\omega)^T A_\omega$.

En d'autres termes, la fonctionnelle est *sparse convex* de rang $r^\star$ si la convexité du terme d'attache aux données réduit ($\frac{1}{2}\|(A_\omega)^T A_\omega \cdot - d\|^2$) est supérieure à la concavité de la pénalité ϕ pour tout support $\omega \in \mathbb{I}_N$ dont la taille est inférieure à $r^\star$. En effet, pour un support $\omega \subseteq \mathbb{I}_N$ tel que $\sharp\omega \leqslant r^\star$, la fonctionnelle $x \mapsto \frac{1}{2}\|(A_\omega)^T A_\omega x - d\|^2 + \sum_{i \in \mathbb{I}_{\sharp\omega}} \phi(x_i)$ est convexe dès lors que la matrice $(A_\omega)^T A_\omega - \kappa(\phi) I_d$ est définie positive. On peut alors voir que les valeurs propres de cette dernière sont données par $\{\nu_j - \kappa(\phi)\}_{j \in \mathbb{I}_{\sharp\omega}}$, où les $\{\nu_j\}_{j \in \mathbb{I}_{\sharp\omega}}$ représentent les valeurs propres de $(A_\omega)^T A_\omega$. Ainsi, la condition (8.34) assure la convexité de $x \mapsto \frac{1}{2}\|(A_\omega)^T A_\omega x - d\|^2 + \sum_{i \in \mathbb{I}_{\sharp\omega}} \phi(x_i)$ pour tout $\omega \in \mathbb{I}_N$ tel que $\sharp\omega \leqslant r^\star$.

Par définition, la pénalité MCP minimise le maximum de concavité (8.31), c'est donc celle qui permet d'obtenir le rang de *sparse convexity* le plus grand parmi les pénalités vérifiant les conditions (8.32) et (8.33) ce qui motive son utilisation. Enfin, notons que si $\gamma \to +\infty$ dans (8.30) on retrouve la norme-ℓ_1 et lorsque $\gamma \to 0$ la pénalité MCP tend vers la norme-ℓ_0.

8.4.4 *Et bien d'autres !*

Nous pouvons citer un certain nombre d'autres pénalités continues non-convexes introduites pour approcher la norme-ℓ_0. L'une des premières remonte certainement aux travaux

de Mangasarian (1996) qui propose une approximation exponentielle du type (8.23) avec

$$\phi_{\text{exp}}(\theta; u) := 1 - \exp\left(-\theta|u|\right), \tag{8.35}$$

pour $\theta \in \mathbb{R}_+^\star$. Le graphe de cette pénalité est représenté sur la figure 31. Nous pouvons aussi mentionner la pénalité dite «Capped-ℓ_1» (ou ℓ_1 tronquée), définie par (8.23) avec

$$\phi_{\text{Cap}}(\theta; u) := \min\{\theta|u|, 1\}, \tag{8.36}$$

pour $\theta \in \mathbb{R}_+^\star$. Cette pénalité, représentée sur la figure 31, est connue pour être une bonne alternative à la norme-ℓ_0 (Fan, 1997 ; Zhang, 2009 ; Le Thi et al., 2014).

Considérer les normes-ℓ_p pour $p \in]0, 1[$ semble tout naturel en vue de trouver un compromis entre ℓ_0 et ℓ_1. Dans ce cas, les pénalités utilisées sont définies par (8.23) avec

$$\phi(p, \theta; u) := \theta|u|^p, \tag{8.37}$$

où $\theta \in \mathbb{R}_+^\star$ et $p \in]0, 1[$. Plusieurs auteurs se sont intéressés à l'étude de ces normes-ℓ_p en terme de reconstruction exacte comme prolongement du cas $p = 1$ très largement étudié (voir section 8.1). Par exemple, Chartrand (2007) aborde le problème du point de vue de la condition RIP sur la matrice A et une autre version de ces résultats peut être trouvée dans (Foucart et Lai, 2009).

Mohimani et al. (2009) ont utilisé des approximations non-convexes différentiables définies par

$$\phi(\theta; u) := 1 - f_\theta(u), \tag{8.38}$$

où f_θ appartient à une famille de fonctions approchant la fonction «0-1» lorsque θ tend vers 0. Par exemple, les auteurs utilisent une approximation exponentielle $f_\theta(u) := \exp(-u^2/(2\theta^2))$ ou encore $f_\theta(u) := \theta^2/(u^2 + \theta^2)$ ainsi que la fonction quadratique tronquée $f_\theta(u) := 1 - \min(u^2/\theta^2, 1)$. Ces approximations sont ensuite utilisées dans un schéma du type Graduated Non Convexity (GNC) détaillé dans le chapitre 10.

Enfin, notons que certains auteurs ont également utilisé le ratio ℓ_1/ℓ_2 comme mesure de parcimonie (Repetti et al., 2015, et références associées).

Nous renvoyons aussi le lecteur à (Antoniadis et al., 2011) pour d'autres pénalités (non)-convexes, (non)-différentiables en zéro et continues.

Toutes ces pénalités sont généralement associées à des seuillages obtenus en étudiant le cas orthogonal (i. e. $A^T A$ diagonale). Comme il est parfois plus instructif de visualiser de tels seuillages plutôt que les pénalités elles mêmes, quelques expressions de ces derniers sont données dans la table 2 et les graphes associés sur la figure 32. Remarquons que pour certaines valeurs des paramètres, les pénalités Capped-ℓ_1 et MCP produisent un seuillage dur comme pour la pénalité ℓ_0.

8.5 REFORMULATIONS «EXACTES»

Nous avons vu dans le paragraphe précédent que de nombreuses relaxations continues des problèmes (C_k), (C_ϵ) et (P_λ) avaient été proposées principalement à partir d'approximations continues non-convexes de la norme-ℓ_0. Cependant, le choix d'une relaxation par rapport à une autre reste obscur. En particulier, une question importante concerne l'étude de la consistance entre les minimiseurs de la fonctionnelle initiale et ceux de la relaxation. En d'autres termes :

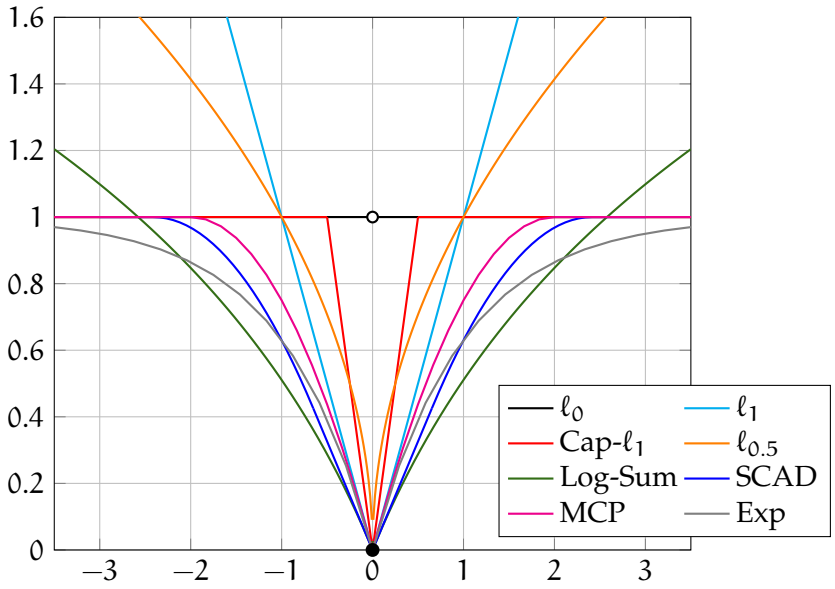


FIGURE 31 – Pénalités ℓ_0, ℓ_1, Capped-ℓ_1, $\ell_{0.5}$, Log-Sum, Exp, SCAD et MCP. Le paramètre λ est fixé à 1. Notons que pour SCAD et MCP, les paramètres γ et $\tilde{\lambda}$ ont été choisis afin d'avoir une pénalité constante égale à $\lambda(=1)$ pour $|u|$ grand.

Pénalité	$\phi(u)$	Seuillage $S(v)$
ℓ_0	$\lambda\|u\|_0$	$v\mathbb{1}_{\{\|v\|>\sqrt{2\lambda}\}} + \{0,v\}\mathbb{1}_{\{\|v\|=\sqrt{2\lambda}\}}$
ℓ_1	$\lambda\|u\|$	$\mathrm{sign}(v)(\|v\|-\lambda)_+$
Cap-ℓ_1	$\lambda\min(\theta\|u\|,1)$	$\begin{cases} v\mathbb{1}_{\{\|v\|>\sqrt{2\lambda}\}} + \{0,v\}\mathbb{1}_{\{\|v\|=\sqrt{2\lambda}\}} & \text{si } \lambda\theta^2 \geqslant 2, \\ \mathrm{sign}(v)(\|v\|-\lambda\theta)\mathbb{1}_{\{\lambda\theta<\|v\|<\frac{1}{\theta}+\frac{\lambda\theta}{2}\}} + d\mathbb{1}_{\{\|v\|\geqslant\frac{1}{\theta}+\frac{\lambda\theta}{2}\}} & \text{si } \lambda\theta^2 < 2, \end{cases}$
MCP	(8.29)	$\begin{cases} v\mathbb{1}_{\{\|v\|>\sqrt{\gamma\tilde{\lambda}^2}\}} + \{0,v\}\mathbb{1}_{\{\|v\|=\sqrt{\gamma\tilde{\lambda}^2}\}} & \text{si } \gamma \leqslant 1, \\ \mathrm{sign}(v)\min\left(\frac{\gamma(\|v\|-\tilde{\lambda})_+}{\gamma-1}, \|v\|\right) & \text{si } \gamma > 1, \end{cases}$
SCAD	(8.28)	$\begin{cases} \mathrm{sign}(d)(\|v\|-\tilde{\lambda})_+ & \text{si } \|v\| \leqslant 2\tilde{\lambda}, \\ ((\gamma-1)v - \mathrm{sign}(v)\gamma\tilde{\lambda})/(\gamma-2) & \text{si } 2\tilde{\lambda} < \|v\| \leqslant \gamma\tilde{\lambda}, \\ v & \text{sinon}, \end{cases}$

TABLE 2 – Expressions des seuillages associés aux pénalités ℓ_0, ℓ_1, Capped-ℓ_1, MCP et SCAD. Ils sont déterminés selon : $S(v) := \arg\min_{u\in\mathbb{R}} \frac{1}{2}(u-v)^2 + \phi(u)$.

— la relaxation continue préserve-t-elle les minimiseurs globaux de la fonctionnelle initiale ?
— ajoute-t-elle de nouveaux minimiseurs locaux ?

Dans le cas où la réponse est affirmative pour la première question et négative pour la seconde, nous parlerons alors de reformulation (relaxation) exacte. Plusieurs auteurs se sont intéressés à ces questions et nous présentons leurs travaux dans la suite de ce paragraphe.

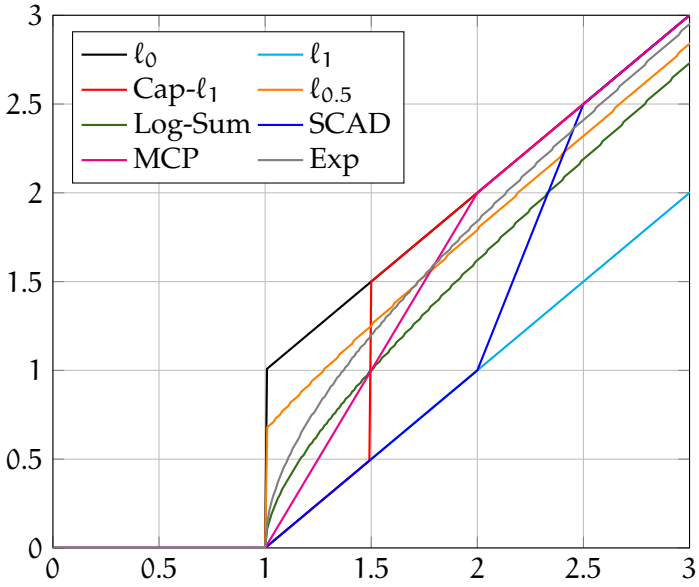


FIGURE 32 – Seuillages (sur $\mathbb{R}_+$) associés aux pénalités ℓ_0, ℓ_1, Capped-ℓ_1, $\ell_{0.5}$, Log-Sum, Exp, SCAD et MCP. Le paramètre λ n'est pas le même pour tous les seuillages afin d'avoir l'intervalle $[0, 1]$ nul pour chacune des représentations. Notons que les seuillages associés aux pénalités $\ell_{0.5}$, Log-Sum et Exp ont été calculés numériquement.

8.5.1 *Une classe de pénalités non-convexes et différentiables*

Une classe de pénalités continues et différentiables a été proposée par CHOUZENOUX et al. (2013) afin d'approcher la norme-ℓ_0. Plus précisément, les auteurs utilisent des pénalités continues $\phi_\delta : \mathbb{R} \to \mathbb{R}$ ($\delta \in \mathbb{R}_+^\star$) vérifiant

— $\forall (\delta_1, \delta_2) \in (0, +\infty)^2,\ \delta_1 \leqslant \delta_2 \ \Rightarrow \forall t \in \mathbb{R},\ \phi_{\delta_1}(t) \geqslant \phi_{\delta_2}(t)$,

— $\exists \lambda \in \mathbb{R}$ tel que $\forall t \in \mathbb{R},\ \lim\limits_{\substack{\delta \to 0 \\ >}} \phi_\delta(t) = \lambda |t|_0$.

Considérant un critère objectif régularisé par un terme du type $\lambda \|x\|_0 + \|x\|^2$ [12], noté F, les auteurs montrent des connections asymptotiques entre les minimiseurs de F et ceux de F_δ (où la norme-ℓ_0 est remplacée par ϕ_δ) lorsque $\delta \to 0$ (CHOUZENOUX et al., 2013, proposition 2). Un minimiseur de F peut alors être bien approché par la minimisation de F_δ pour δ suffisamment petit. Notons par ailleurs qu'une approche du type GNC pourrait être envisagée avec une telle classe de fonctions (voir chapitre 10).

8.5.2 *Pénalités ℓ_p et approximation exponentielle*

Considérant le problème d'estimation de la solution la plus parcimonieuse d'un système d'égalités et d'inégalités linéaires, exprimé sous la forme :

$$\min_{x \in \mathbb{R}^N} \|x\|_0 \ \text{ s.c. } \ Ax = a,\ Bx \geqslant b,\ \|x\|_\infty \leqslant 1, \tag{8.39}$$

FUNG et MANGASARIAN (2011) montrent que ce dernier est équivalent au problème

$$\min_{x \in \mathbb{R}^N} \|x\|_p \ \text{ s.c. } \ Ax = a,\ Bx \geqslant b,\ \|x\|_\infty \leqslant 1, \tag{8.40}$$

pour $p \leqslant 1$ suffisamment petit. Plus précisément, après avoir reformulé ces deux problèmes comme la minimisation d'une fonction objectif non-convexe sur un ensemble polyédral

12. Le terme ℓ_2 permet d'assurer certaines propriétés concernant les minimiseurs, notamment l'existence.

borné [13], les auteurs prouvent l'existence d'un sommet de cet ensemble qui est solution des deux problèmes pour certains $p \leqslant 1$ (FUNG et MANGASARIAN, 2011, proposition 3.1).

La preuve de ce résultat est inspirée de travaux plus anciens (BRADLEY et al., 1998, théorème 2.1) où les auteurs montrent l'équivalence entre les problèmes :

$$\min_{x\in S} f(x) + \lambda \sum_{i\in\mathbb{I}_N} w_i |x_i|_0 \quad \text{et} \quad \min_{x\in S} f(x) + \lambda \sum_{i\in\mathbb{I}_N} w_i \phi_{\text{exp}}(\theta; x_i), \tag{8.41}$$

pour $\theta \in \mathbb{R}_+^\star$ suffisamment grand, ϕ_{exp} l'approximation exponentielle de la norme-ℓ_0 donnée par (8.35), f concave sur $\mathbb{R}^N$ et bornée inférieurement sur l'ensemble polyédral S et enfin $w \in \mathbb{R}_+^N$. Ici l'équivalence est à comprendre au sens où l'intersection de l'ensemble des solutions des deux problèmes est non-vide. Notons qu'une vision unifiée de ces résultats peut être trouvée dans (RINALDI et al., 2010), incluant par ailleurs l'approximation Log-Sum (8.27).

8.5.3 *Programmation Mixte en Nombres Entiers*

BOURGUIGNON et al. (2015, 2016) ont récemment proposé de reformuler les problèmes (C_k), (C_ϵ) et (P_λ) sous la forme de problèmes de Programmation Mixte en Nombres Entiers (PMNE) mêlant variables continues et variables entières (BIXBY, 2012). Sous l'hypothèse que la solution du problème est bornée selon $\|x\|_\infty < M \in \mathbb{R}_+$, les auteurs reformulent la norme-ℓ_0 en introduisant des variables binaires comme il suit :

$$\|x\|_0 \leqslant K \;\Leftrightarrow\; \exists b \in \{0,1\}^N \;\text{ t.q. }\; \begin{cases} \sum_{i\in\mathbb{I}_N} b_i \leqslant K, \\ -Mb \leqslant x \leqslant Mb, \end{cases} \tag{8.42}$$

$$\min_{x\in\mathbb{R}^N} \|x\|_0 \;\Leftrightarrow\; \min_{x\in\mathbb{R}^N,\, b\in\{0,1\}^N} \sum_{i\in\mathbb{I}_N} b_i \;\text{ t.q. }\; -Mb \leqslant x \leqslant Mb. \tag{8.43}$$

À partir de (8.42) et (8.43), (C_k), (C_ϵ) et (P_λ) sont reformulés sous la forme PMNE (BOURGUIGNON et al., 2016, Table 1). Bien que les problèmes ainsi obtenus soient également NP-difficiles, les progrès concernant d'une part la puissance de calcul des ordinateurs et d'autre part les méthodes numériques, permettent de résoudre *exactement* de tels problèmes lorsque leur taille n'excède pas quelques centaines de variables. Les auteurs optent pour l'algorithme CPLEX combinant la méthode de séparation et évaluation (*branch and bound* en anglais) et celle des plans sécants (*cutting plan method* en anglais). L'idée derrière ces méthodes est d'éliminer d'importantes parties du domaine de recherche afin de converger plus rapidement vers la solution globale du problème.

À notre connaissance, l'approche proposée par BOURGUIGNON et al. (2015, 2016) est à ce jour la seule (outre la recherche exhaustive) à assurer la convergence vers un *minimiseur global* des problèmes (C_k) (C_ϵ) et (P_λ) sans aucune restrictions sur la matrice A. Cependant, de part sa complexité de calcul, une telle approche reste tout de même limitée à des problèmes de taille modérée, de l'ordre de quelques centaines de variables.

8.5.4 *Approximations DC de la norme-ℓ_0*

Enfin, les récents travaux de LE THI et al. (2015) sont consacrés à l'étude de la consistance entre les minimiseurs de la fonctionnelle

$$F(x,y) := f(x,y) + \lambda\|x\|_0, \;\; (x,y) \in K, \tag{8.44}$$

13. Qui est le même pour les problèmes (8.39) et (8.40)

où $K \in \mathbb{R}^N \times \mathbb{R}^M$ est un ensemble convexe, $\lambda \in \mathbb{R}_+^\star$ et f admet une décomposition DC (Difference of Convex functions : $f = g - h$ où g et h sont des fonctions convexes), et ceux de l'approximation

$$F_\theta(x, y) := f(x, y) + \lambda \sum_{i \in IN} \phi(\theta; x_i),\ (x, y) \in K, \tag{8.45}$$

avec $\phi(\theta; \cdot)$ une pénalité DC vérifiant un certain nombre de propriétés dont, entre autres, la parité et la convergence simple vers $|\cdot|_0$ lorsque θ tend vers $+\infty$ (Le Thi et al., 2015, Hypothèse 1). Dans ce contexte, les auteurs montrent plusieurs résultats :

- pour θ suffisamment grand, tout minimiseur global de F_θ est dans un ε-voisinage d'un minimiseur global de F (Le Thi et al., 2015, théorème 1) ;
- le point précédent est également vrai pour les minimiseurs locaux (Le Thi et al., 2015, théorème 2) ;
- si f est concave et bornée inférieurement sur K, alors quelque soit la valeur du paramètre θ, les minimiseurs globaux de F_θ sont aussi des minimiseurs globaux de F (Le Thi et al., 2015, corolaire 1) ;
- pour le cas des pénalités Capped-ℓ_1 et SCAD, et pour un choix du paramètre θ approprié, les minimiseurs globaux des deux fonctionnelles coïncident (Le Thi et al., 2015, propositions 3, 4 et 5) (voir aussi (Le Thi et al., 2014)) ;

Notons cependant que les deux derniers résultats ne sont pas valables pour les minimiseurs locaux comme nous le verrons dans le chapitre 12 pour le problème (P_λ).

9 LA PÉNALITÉ CEL0

Sommaire

Dans ce chapitre, nous nous intéressons au problème pénalisé (P_λ). À partir du calcul de l'enveloppe convexe de la fonctionnelle dans le cas unidimensionnel et dans le cas orthogonal en dimension quelconque (sections 9.1 et 9.2), nous proposons une relaxation (non-convexe) *continue* de la norme-ℓ_0 menant à une reformulation continue exacte du problème (P_λ) sans aucune restriction sur les données du problème (i. e. $A \in \mathbb{R}^{M\times N}$ et $d \in \mathbb{R}^M$). Le caractère exact de la relaxation proposée est mis en évidence par une étude des liens entre les minimiseurs de la fonctionnelle initiale (fonction objectif de (P_λ)) et ceux de la fonctionnelle relaxée, menée dans la section 9.3.

Le travail présenté dans ce chapitre peut être trouvé dans l'article (Soubies et al., 2015a) ainsi que dans les communications (Soubies et al., 2015b,c).

9.1 L'enveloppe convexe dans le cas unidimensionnel

Commençons par considérer le cas $N = 1$ (1D). Nous pouvons alors réécrire (P_λ) comme il suit :

$$\hat{u} = \arg\min_{u\in\mathbb{R}} g_0(u) := \frac{1}{2}(au - d)^2 + \lambda|u|_0, \tag{9.1}$$

où $a \in \mathbb{R}_+^\star$ et $d \in \mathbb{R}$.

Notons que la restriction $a \in \mathbb{R}_+^\star$ ne fait perdre en aucun cas la généralité des développements qui vont suivre. En effet, nous pouvons aisément voir que g_0 reste inchangée en multipliant a et d par -1.

Une notion importante en optimisation non-convexe est celle d'enveloppe convexe dont la définition est donné ci-dessous.

Définition 9.1 (Enveloppe convexe (Rockafellar et Wets, 2009))**.** L'enveloppe convexe d'une fonction $f : \mathbb{R}^N \to \bar{\mathbb{R}}$ est la plus grande fonction convexe $\operatorname{co} f : \mathbb{R}^N \to \bar{\mathbb{R}}$ minorant f, c'est à dire :

$$\forall x \in \mathbb{R}^N,\ (\operatorname{co} f)(x) := \sup\left\{h(x) : h(x') \leqslant f(x')\ \forall x' \in \mathbb{R}^N,\ h \text{ convexe}\right\}. \tag{9.2}$$

L'enveloppe convexe d'une fonction possède des propriétés particulièrement intéressantes pour l'optimisation. En effet, les minimiseurs globaux d'une fonction sont inclus

dans ceux de son enveloppe convexe et les valeurs minimales de ces deux fonctions coïncident [1]. De fait, si nous sommes en mesure de calculer l'enveloppe convexe d'une fonctionnelle non-convexe, la minimisation de cette dernière peut alors être simplifiée à la minimisation d'une fonction convexe. Malheureusement, le calcul de l'enveloppe convexe d'une fonction non-convexe est généralement un problème difficile.

Cependant, pour le problème unidimensionnel (7.3), une forme analytique de l'enveloppe convexe fermée de g_0 peut aisément être calculée comme nous nous proposons de le faire dans la suite. Un moyen d'accéder à l'enveloppe convexe fermée d'une fonction est de calculer sa biconjuguée, c'est-à-dire d'appliquer deux fois la transformée de Legendre-Fenchel définie comme il suit.

Définition 9.2 (Transformée de Legendre-Fenchel (Fenchel, 1949))**.** La transformée de Legendre-Fenchel (ou fonction conjuguée) de $f : \mathbb{R}^N \to \bar{\mathbb{R}}$ est donnée par :

$$f^\star(x^\star) = \sup_{x \in \mathbb{R}^N} \langle x^\star, x \rangle - f(x). \tag{9.3}$$

On a $f^{\star\star} = (f^\star)^\star = \mathrm{cl}(\mathrm{co}\, f)$ où $\mathrm{cl}(\cdot)$ défini la fermeture de f. La proposition suivante donne l'expression de la fonction biconjuguée (enveloppe convexe fermée) $g_0^{\star\star}$ de g_0.

Proposition 9.3 (Enveloppe convexe de g_0)**.** *L'enveloppe convexe (ou biconjuguée) de* g_0*, notée* $g_0^{\star\star}$*, est donnée par*

$$\forall u \in \mathbb{R},\ g_0^{\star\star}(u) = \frac{1}{2}(au - d)^2 + \phi(a, \lambda; u), \tag{9.4}$$

où pour $a \in \mathbb{R}_+^\star$ *et* $\lambda \in \mathbb{R}_+$*,*

$$\forall u \in \mathbb{R},\ \phi(a, \lambda; u) = \lambda - \frac{a^2}{2}\left(|u| - \frac{\sqrt{2\lambda}}{a}\right)^2 \mathbb{1}_{\left\{|u| \leqslant \frac{\sqrt{2\lambda}}{a}\right\}}. \tag{9.5}$$

Démonstration. La preuve est donnée en annexe A.1.1 (page 187). Notons que le calcul de l'enveloppe convexe dans le cas unidimensionel peut aussi être trouvé dans (Dinh et Le Thi, 2014 ; Jojic et al., 2011). À titre informatif, des résultats similaires existent aussi pour l'approximation de matrices de rang faible (Larsson et al., 2014). □

Au regard de la proposition 9.3, l'enveloppe convexe de g_0 est donc obtenue en remplaçant le terme de régularisation ℓ_0, dans (9.1), par la pénalité ϕ donnée en (9.5). Remarquons que cette pénalité est non-convexe, non-différentiable en 0 mais *continue* (cf. figure 34). La figure 33 présente trois exemples montrant g_0 et son enveloppe convexe $g_0^{\star\star}$ où les minima globaux sont respectivement 0 (gauche), $\{0, \frac{d}{a}\}$ (centre, dans ce cas tout l'intervalle $[0, \frac{d}{a}]$ est minimiseur global pour $g_0^{\star\star}$) et $\frac{d}{a}$ (droite).

Sur la figure 34, nous avons tracé la pénalité (9.5) ainsi que la fonction «0-1». Il est à noter que cette pénalité vérifie les conditions proposées par Fan et Li (2001) à savoir être non-différentiable en 0 pour favoriser la parcimonie ainsi qu'être constante pour de grandes

1. Par définition, $\mathrm{co}\, f \leqslant f$. Supposons que f admette au moins un minimiseur global et que $\inf \mathrm{co}\, f < \min f$. Alors, étant donné que $\mathrm{co}\, f$ est convexe, la fonction $h(x) := \max\{\min f, \mathrm{co}\, f(x)\}$ est également convexe, vérifie $\mathrm{co}\, f \leqslant h \leqslant f$ et $\exists x \in \mathbb{R}^N$ tel que $\mathrm{co}\, f(x) < h(x)$ ce qui est en contradiction avec la définition de l'enveloppe convexe d'une fonction. Donc $\mathrm{co}\, f$ admet aussi des minimiseurs globaux et $\min \mathrm{co}\, f = \min f$. De plus, l'inégalité $\mathrm{co}\, f \leqslant f$ montre que $\arg\min\ f \subseteq \arg\min\ \mathrm{co}\, f$.

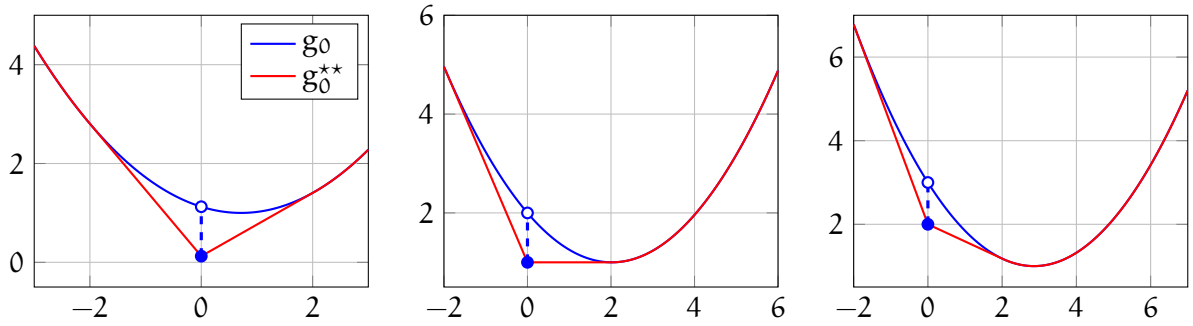


FIGURE 33 – Fonction g_0 (bleu) et son enveloppe convexe $g_0^{\star\star}$ (rouge) pour $a = 0.7$, $\lambda = 1$ et $d = 0.5$ (gauche), $d = \sqrt{2}$ (centre) ou $d = 2$ (droite).

valeur de $|u|$ afin d'obtenir une solution non-biaisée (dont l'amplitude des coefficients non-nuls n'est pas atténuée par rapport à ceux de la solution recherchée).

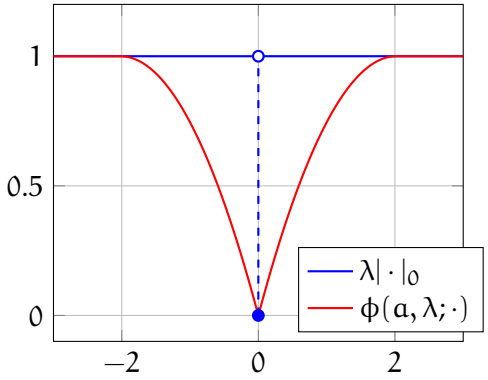


FIGURE 34 – Graphe de $\lambda|\cdot|_0$ (bleu) et $\phi(a,\lambda;\cdot)$ pour $a = 0.7$ et $\lambda = 1$.

9.2 EXTENSION AU CAS MULTIDIMENSIONNEL ORTHOGONAL

Étendre les résultats de la section précédente à la dimension $N \in \mathbb{N}$ sans aucune restriction sur la matrice $A \in \mathbb{R}^{M\times N}$ n'est pas envisageable. En effet, la fonction conjuguée de $G_{\ell_0} : \mathbb{R}^N \to \mathbb{R}$, définie en ($P_\lambda$), est donnée par

$$G^\star_{\ell_0}(x^\star) = \sup_{x\in\mathbb{R}^N} \left\{ \langle x^\star, x\rangle_{\mathbb{R}^N} - \frac{1}{2}\|Ax-d\|^2 - \lambda\|x\|_0 \right\}, \tag{9.6}$$

qui est un problème combinatoire aussi difficile que l'original (P_λ). Cependant, dans le cas particulier où les colonnes de A sont deux à deux orthogonales (i. e. A^TA est diagonale) et non nulles, le problème (9.6) peut être résolu analytiquement et nous avons le résultat suivant.

Proposition 9.4. *Lorsque $A \in \mathbb{R}^{M\times N}$ est une matrice dont les colonnes sont deux à deux orthogonales (i. e. A^TA est diagonale) et non nulles ($\|a_i\| > 0\ \forall i \in \mathbb{I}_N$), l'enveloppe convexe de G_{ℓ_0}, notée $G^{\star\star}_{\ell_0}$, est donnée par*

$$\forall x \in \mathbb{R}^N,\ G^{\star\star}_{\ell_0}(x) = \frac{1}{2}\|Ax-d\|^2 + \sum_{i\in\mathbb{I}_N} \phi(\|a_i\|,\lambda;x_i), \tag{9.7}$$

où, pour $a \in \mathbb{R}_+^\star$ *et* $\lambda \in \mathbb{R}_+$, $\phi(a, \lambda; \cdot)$ *est définie en* (9.5) *et* $\|a_i\|$ *représente la norme de la* i*-ème colonne de* A.

Démonstration. La preuve est donnée en annexe A.1.2 (page 189). □

Ainsi, comme dans le cas unidimensionnel, l'enveloppe convexe de G_{ℓ_0} lorsque les colonnes de la matrice A sont orthogonales et non nulles est obtenue en remplaçant norme ℓ_0 par une pénalité également non-convexe mais *continue*, nommée Continuous Exact ℓ_0 (CEL0) et définie par

$$\Phi_{\text{CEL0}}(x) := \sum_{i \in \mathbb{I}_N} \phi(\|a_i\|, \lambda, x_i) = \sum_{i \in \mathbb{I}_N} \lambda - \frac{\|a_i\|^2}{2} \left(|x_i| - \frac{\sqrt{2\lambda}}{\|a_i\|} \right)^2 \mathbb{1}_{\left\{ |x_i| \leqslant \frac{\sqrt{2\lambda}}{\|a_i\|} \right\}}, \tag{9.8}$$

pour $\lambda \in \mathbb{R}_+$.

Dans la section suivante, nous étudions les propriétés de cette pénalité lorsque $A \in \mathbb{R}^{M \times N}$ est arbitraire (pas forcément $A^T A$ diagonale).

9.3 ETUDE DU CAS GÉNÉRAL MULTIDIMENSIONNEL

Nous considérons maintenant une matrice $A \in \mathbb{R}^{M \times N}$ quelconque et nous nous intéressons à la fonctionnelle

$$G_{\text{CEL0}}(x) := \frac{1}{2}\|Ax - d\|^2 + \Phi_{\text{CEL0}}(x), \tag{9.9}$$

qui est une relaxation (non-convexe) *continue* de G_{ℓ_0}. Étant donné qu'elle est définie à partir de la pénalité CEL0, donnée en (9.8), nous l'appellerons la fonctionnelle CEL0. Notons que le terme CEL0 est dû aux deux principaux résultats, concernant les liens entre les minimiseurs de G_{ℓ_0} et G_{CEL0}, qui sont présentés dans la suite de cette section.

Le premier, établi par le théorème 9.16, assure que l'ensemble des minimiseurs globaux de G_{ℓ_0} est inclus dans l'ensemble des minimiseurs globaux de G_{CEL0} et qu'à partir de tout minimiseur global de G_{CEL0} on peut facilement extraire un minimiseur global de G_{ℓ_0} par un simple seuillage.

Le second résultat, donné par le théorème 9.21, étend partiellement le théorème 9.16 aux minimiseurs locaux : de tout minimiseur local de G_{CEL0}, on extrait simplement un minimiseur local de G_{ℓ_0}. De plus, il est montré que certains minimiseurs locaux de G_{ℓ_0} ne sont pas des points critiques de G_{CEL0}. Cela est illustré numériquement sur des exemples montrant qu'un nombre non-négligeable de minimiseurs locaux de G_{ℓ_0} sont ainsi éliminés par G_{CEL0}.

Remarque 9.5. On peut aisément voir que G_{CEL0} minore G_{ℓ_0}. En effet, pour $\lambda > 0$, nous avons $0 \leqslant \phi(\|a_i\|, \lambda; u) \leqslant 1 \ \forall u \in \mathbb{R}, \forall i \in \mathbb{I}_N$. Ensuite, comme $\phi(\|a_i\|, \lambda; 0) = 0$, il vient que $\phi(\|a_i\|, \lambda; u) \leqslant |u|_0, \forall u \in \mathbb{R}, \forall i \in \mathbb{I}_N$ ce qui prouve la précédente affirmation.

Une première propriété de G_{CEL0} qui découle directement de l'étude 1D faite dans la section 9.1 est donnée par la proposition suivante.

Proposition 9.6. *Pour* $x \in \mathbb{R}^N$ *et* $i \in \mathbb{I}_N$, *la restriction* $G^i_{\text{CEL0}}(\cdot; x^{(i)})$ *de* G_{CEL0} *à la* i*-ème variable au point* x *s'écrit :* $\forall t \in \mathbb{R}$,

$$G^i_{\text{CEL0}}(t; x^{(i)}) = C_i + \begin{cases} t\langle a_i, Ax^{(i)} - d\rangle + |t| \, \|a_i\|\sqrt{2\lambda} & \text{si } \ 0 \leqslant |t| \leqslant \frac{\sqrt{2\lambda}}{\|a_i\|}, \\ \frac{1}{2}\|a_i\|^2 t^2 + t\langle a_i, Ax^{(i)} - d\rangle + \lambda & \text{si } \ |t| \geqslant \frac{\sqrt{2\lambda}}{\|a_i\|} \end{cases} \tag{9.10}$$

où C_i est une constante indépendante de t donnée par

$$C_i = \frac{1}{2}\|Ax^{(i)} - d\|^2 + \sum_{j\neq i} \phi(\|a_j\|, \lambda; x_j), \tag{9.11}$$

$$x^{(i)} = (x_1, \cdots, x_{i-1}, 0, x_{i+1}, \cdots, x_N). \tag{9.12}$$

De plus $G^i_{\text{CEL0}}(\cdot; x^{(i)})$ est convexe et plus précisément elle est affine sur $\left]-\frac{\sqrt{2\lambda}}{\|a_i\|}, 0\right[$ et $\left]0, \frac{\sqrt{2\lambda}}{\|a_i\|}\right[$ et strictement convexe au delà.

Démonstration. Prenons $x \in \mathbb{R}^N$ et $i \in \mathbb{I}_N$. Alors $G^i_{\text{CEL0}}(\cdot; x^{(i)})$, la restriction de G_{CEL0} à la i-*ème* variable au point x s'écrit : $\forall t \in \mathbb{R}$

$$\begin{aligned}
G^i_{\text{CEL0}}(t; x^{(i)}) &= G_{\text{CEL0}}(x^{(i)} + te_i), \\
&= \frac{1}{2}\|Ax^{(i)} + a_i t - d\|^2 + \sum_{j\in\mathbb{I}_N\setminus\{i\}} \phi(\|a_j\|, \lambda; x_j) + \phi(\|a_i\|, \lambda; t), \\
&= \frac{1}{2}\|Ax^{(i)} - d\|^2 + t\langle a_i, Ax^{(i)} - d\rangle + \frac{1}{2}\|a_i\|^2 t^2 \\
&\qquad + \sum_{j\in\mathbb{I}_N\setminus\{i\}} \phi(\|a_j\|, \lambda; x_j) + \phi(\|a_i\|, \lambda; t), \\
&\underset{(9.5)}{=} C_i + t\langle a_i, Ax^{(i)} - d\rangle + \begin{cases} |t|\,\|a_i\|\sqrt{2\lambda} & \text{si } |t| \leqslant \frac{\sqrt{2\lambda}}{\|a_i\|}, \\ \frac{1}{2}\|a_i\|^2 t^2 + \lambda & \text{si } |t| \geqslant \frac{\sqrt{2\lambda}}{\|a_i\|}. \end{cases}
\end{aligned}$$

où C_i est donnée par (9.11). Enfin, la dernière assertion de la proposition est évidente avec l'expression précédente. □

9.3.1 *Résultats théoriques*

9.3.1.1 *Les points critiques de G_{CEL0}*

Afin d'étudier les minimiseurs de G_{CEL0}, il convient tout d'abord de s'intéresser à ses points critiques. Étant donné que G_{CEL0} est une fonction non-convexe, nous utiliserons la notion de *gradient généralisé* introduit par Clarke (1990) qui étend la notion de sous-différentiel pour les fonctions convexes aux fonctions non-convexes. Avant de donner la définition du gradient généralisé, nous avons besoin d'introduire la définition suivante :

Définition 9.7 (Fonction localement Lipschitz)**.** Une fonction f est dite être *localement Lipschitz* au point x si,

$$\exists \varepsilon > 0,\ \forall (y, y') \in \mathcal{B}(x, \varepsilon)^2\ |f(y) - f(y')| \leqslant K\|y - y'\|, \tag{9.13}$$

où $K \in \mathbb{R}_+$ et $\mathcal{B}(x, \varepsilon)$ est un ε-voisinage de x.

Nous pouvons maintenant définir le gradient généralisé d'une fonction localement Lipschitz en un point x.

Définition 9.8 (Gradient généralisé (Clarke, 1990))**.** Le *gradient généralisé* d'une fonction $f : \mathbb{R}^N \to \mathbb{R}$ localement Lipschitz en x, noté $\partial f(x)$, est défini par

$$\partial f(x) := \left\{\xi \in \mathbb{R}^N :\ f^0(x, v) \geqslant \langle v, \xi\rangle\ \forall v \in \mathbb{R}^N\right\}, \tag{9.14}$$

où $f^0(x, v)$ représente la dérivée directionnelle généralisée de f en x dans la direction v,

$$f^0(x, v) = \limsup_{\substack{y \to x \\ \eta \downarrow 0}} \frac{f(y + \eta v) - f(y)}{\eta}. \tag{9.15}$$

Il est à noter que lorsque f est continument différentiable, $\partial f(x)$ est réduit au singleton $\{\nabla f(x)\}$ (Clarke, 1990, proposition 2.2.4 et corolaires associés) et, dans le cas convexe, le gradient généralisé de Clarke coïncide avec la définition classique du sous-différentiel (Clarke, 1990, proposition 2.2.7). La proposition suivante montre l'importance de cette notion de gradient généralisé lorsque l'on souhaite étudier les minimiseurs d'une fonction.

Proposition 9.9 (Extrema local (Clarke, 1990))**.** *Si f atteint un minimum ou maximum local en x, alors $0 \in \partial f(x)$.*

La condition $0 \in \partial f(x)$ est donc *nécessaire* pour que x soit un minimiseur (ou maximiseur) local de f. On appelle *points critiques* de f les points $x \in \mathbb{R}^N$ vérifiant $0 \in \partial f(x)$.

À partir de la définition 9.8 nous pouvons calculer $\partial\phi(a, \lambda; \cdot)$ où $\phi(a, \lambda; \cdot)$ est définie en (9.5). Étant donné que pour tout $u \neq 0$, $\phi(a, \lambda; \cdot)$ est différentiable, nous avons pour $a \in \mathbb{R}_+^\star$ et $\lambda \in \mathbb{R}_+$,

$$\forall u \neq 0,\ \partial\phi(a, \lambda; u) = \begin{cases} \operatorname{sign}(u)\sqrt{2\lambda}a - a^2 u & \text{si } 0 < |u| \leqslant \frac{\sqrt{2\lambda}}{a}, \\ 0 & \text{si } |u| \geqslant \frac{\sqrt{2\lambda}}{a}. \end{cases} \tag{9.16}$$

Il reste à traiter le cas $u = 0$. Tout d'abord, commençons par calculer la dérivée directionnelle généralisée (9.15) en $u = 0$,

$$\begin{aligned} \forall v \in \mathbb{R},\ f^0(0, v) &= \limsup_{\substack{y \to 0 \\ \eta \downarrow 0}} \frac{\phi(y + \eta v) - \phi(y)}{\eta}, \\ &= \limsup_{\substack{y \to 0 \\ \eta \downarrow 0}} \frac{1}{2\eta} \left[\left(a|y| - \sqrt{2\lambda} \right)^2 - \left(a|y + \eta v| - \sqrt{2\lambda} \right)^2 \right]. \end{aligned}$$

En développant les termes dans l'expression précédente, on montre facilement que,

$$f^0(0, v) = a|v|\sqrt{2\lambda}. \tag{9.17}$$

Ainsi, d'après (9.14), $\partial\phi(0)$ contient tous les $\xi \in \mathbb{R}$ vérifiant $a|v|\sqrt{2\lambda} \geqslant \xi v\ \forall v \in \mathbb{R}$, c'est à dire

$$\partial\phi(a, \lambda; 0) = \left[-a\sqrt{2\lambda}, a\sqrt{2\lambda} \right]. \tag{9.18}$$

Nous pouvons à présent déduire du gradient généralisé de la pénalité 1D $\phi(a, \lambda; \cdot)$ celui de la pénalité CEL0 Φ_{CEL0}. Pour $\lambda \in \mathbb{R}_+$,

$$\begin{aligned} \partial\Phi_{\text{CEL0}}(x) = \prod_{i \in \mathbb{I}_N} &\left[-\sqrt{2\lambda}\|a_i\|, \sqrt{2\lambda}\|a_i\| \right] \mathbb{1}_{\{x_i = 0\}} \\ &+ \|a_i\| \left\{ \operatorname{sign}(x_i)\sqrt{2\lambda} - \|a_i\| x_i \right\} \mathbb{1}_{\left\{ 0 < |x_i| \leqslant \frac{\sqrt{2\lambda}}{\|a_i\|} \right\}}. \end{aligned} \tag{9.19}$$

Le lemme suivant caractérise les points critiques de G_{CEL0} qui, d'après ce qui précède, sont les points $x \in \mathbb{R}^N$ vérifiant $0_{\mathbb{R}^N} \in \partial G_{\text{CEL0}}(x)$

Lemme 9.10 (Points critiques de G_{CEL0}). *Soit* G_{CEL0} *définie en* (9.9). *Posons* $s_i = \operatorname{sign}(\langle a_i, A\hat{x}^{(i)} - d\rangle)$. *Alors* $\hat{x} \in \mathbb{R}^N$ *est un point critique de* G_{CEL0} *(i. e.* $0_{\mathbb{R}^N} \in \partial G_{\text{CEL0}}(\hat{x})$*)* si et seulement si

$$\forall i \in \mathbb{I}_N \begin{cases} \hat{x}_i = 0 & \text{ssi } |\langle a_i, A\hat{x}^{(i)} - d\rangle| \leqslant \sqrt{2\lambda}\|a_i\|, \\ \hat{x}_i = -s_i t,\ t \in \left[0, \dfrac{\sqrt{2\lambda}}{\|a_i\|}\right] & \text{ssi } |\langle a_i, A\hat{x}^{(i)} - d\rangle| = \sqrt{2\lambda}\|a_i\|, \\ \hat{x}_i = -\dfrac{\langle a_i, A\hat{x}^{(i)} - d\rangle}{\|a_i\|^2} & \text{ssi } |\langle a_i, A\hat{x}^{(i)} - d\rangle| \geqslant \sqrt{2\lambda}\|a_i\|. \end{cases} \tag{9.20}$$

où $x^{(i)}$ *est définie en* (9.12).

Démonstration. La preuve est donnée en annexe A.1.3 (page 190). □

Pour la suite, nous introduisons les deux sous ensembles d'indices suivants :

$$\forall x \in \mathbb{R}^N,\ \sigma^-(x) := \left\{ i \in \mathbb{I}_N : 0 < |x_i| < \frac{\sqrt{2\lambda}}{\|a_i\|} \right\} \subseteq \sigma(x), \tag{9.21}$$

et pour un point critique $\hat{x} \in \mathbb{R}^N$ de G_{CEL0},

$$\begin{aligned} \sigma^+(\hat{x}) &:= \left\{ i : |\langle a_i, A\hat{x}^{(i)} - d\rangle| = \sqrt{2\lambda}\|a_i\| \right\} \\ &\underset{(9.20)}{=} \left\{ i : \hat{x}_i = 0 \text{ and } |\langle a_i, A\hat{x}^{(i)} - d\rangle| = \sqrt{2\lambda}\|a_i\| \right\} \cup \left\{ i : 0 < |x_i| \leqslant \tfrac{\sqrt{2\lambda}}{\|a_i\|} \right\}, \end{aligned} \tag{9.22}$$

où nous rappelons que $\sigma(x)$ représente le support de x. Notons que $\sigma^+(\hat{x})$ n'est pas nécessairement inclus $\sigma(\hat{x})$ et en particulier, pour un point critique $\hat{x} \in \mathbb{R}^N$ de G_{CEL0}, on a $\sigma^-(\hat{x}) = \sigma^+(\hat{x}) \cap \sigma(\hat{x})$.

9.3.1.2 *Sur les minimiseurs de* G_{CEL0}

Nous commençons par rapeller deux résultats démontrés par Nikolova (2013). Le premier fournit une borne inférieure sur l'amplitude des coefficients non nuls des minimiseurs globaux de G_{ℓ_0} alors que le second caractérise les minimiseurs (locaux) [2] de G_{ℓ_0}.

Proposition 9.11 (Nikolova, 2013). *Si* G_{ℓ_0} *admet un minimum global en* $\hat{x} \in \mathbb{R}^N$. *Alors,*

$$i \in \sigma(\hat{x}) \implies |\hat{x}_i| \geqslant \frac{\sqrt{2\lambda}}{\|a_i\|}, \tag{9.23}$$

Démonstration. La preuve est donnée dans (Nikolova, 2013, annexe 8.2). Ce résultat est aussi connu de (Nikolova, 2005, proposition 3.4) dans un contexte plus général. □

2. Dans la suite de ce manuscrit, la notation «(local)» avec les parenthèses réfèrera à tous les minimiseurs de la fonctionnelle considérée. Lorsque l'on s'intéressera au minimiseurs globaux cela sera spécifié et on parlera aussi, au contraire, de minimiseurs locaux (non globaux).

Corolaire 9.12 (Nikolova, 2013). *Soit $\hat{x} \in \mathbb{R}^N$ un minimiseur (local) de G_{ℓ_0}. Posons $\hat{\sigma} = \sigma(\hat{x})$. Alors,*

$$\hat{x}_{\hat{\sigma}} \text{ est solution du système } (A_{\hat{\sigma}})^T A_{\hat{\sigma}} \hat{x}_{\hat{\sigma}} = (A_{\hat{\sigma}})^T d. \tag{9.24}$$

Réciproquement, si $\hat{x} \in \mathbb{R}^N$ vérifie (9.24) pour $\hat{\sigma} = \sigma(\hat{x})$, alors $\hat{x}$ est un minimiseur (local) de G_{ℓ_0}.

Démonstration. La preuve se déduit directement de (Nikolova, 2013, lemme 2.4) pour (9.24) et (Nikolova, 2013, proposition 2.3) pour la réciproque. □

Pour la fonctionnelle G_{CEL0}, le lemme suivant fournit un résultat similaire à celui établi par la proposition 9.11 pour G_{ℓ_0}.

Lemme 9.13. *Soit $\hat{x} \in \mathbb{R}^N$ un minimiseur (local) de G_{CEL0}. Posons $s_i = \operatorname{sign}(\langle a_i, A\hat{x}^{(i)} - d \rangle)$ et $\hat{\sigma}^+ := \sigma^+(\hat{x})$. Alors,*

(i) $\forall i \in \hat{\sigma}^+$, $\exists T_i \subseteq \left[0, \frac{\sqrt{2\lambda}}{\|a_i\|}\right]$, un intervalle non-dégénéré (i. e. non réduit à un singleton) de $\mathbb{R}$, tel que $|\hat{x}_i| \in T_i$ et $\forall t \in T_i$,

$$\bar{x} = \hat{x}^{(i)} - s_i e_i t \text{ est aussi un minimiseur (local) de } G_{\text{CEL0}}. \tag{9.25}$$

(ii) si $\hat{x}$ est un minimiseur global, alors $\forall i \in \hat{\sigma}^+$, $T_i = \left[0, \frac{\sqrt{2\lambda}}{\|a_i\|}\right]$ et $\bar{x}$ est aussi global.

Démonstration. La preuve est donnée en annexe A.1.4 (page 191). □

Remarque 9.14. Au regard du lemme 9.13 (et de sa preuve), nous pouvons interpréter $\sigma^+(\hat{x})$ (pour un point critique $\hat{x}$ de G_{CEL0}) comme l'ensemble des indices pour lesquels $t \mapsto G_{\text{CEL0}}(\hat{x}^{(i)} - s_i e_i t)$ est une fonction constante sur $\left[0, \frac{\sqrt{2\lambda}}{\|a_i\|}\right]$.

Le corolaire suivant découle du lemme 9.13.

Corolaire 9.15. *Tous les minimiseurs stricts de G_{CEL0}— i.e. les points $\hat{x} \in \mathbb{R}^N$ tels qu'il existe un voisinage $\mathcal{V} \subset \mathbb{R}^N$ contenant $\hat{x}$ pour lequel $\forall y \in \mathcal{V} \setminus \{\hat{x}\}$, $G_{\text{CEL0}}(\hat{x}) < G_{\text{CEL0}}(y)$ — vérifient $\sigma^+(\hat{x}) = \emptyset$.*

Démonstration. Supposons qu'un minimiseur strict de G_{CEL0}, noté $\hat{x}$, est tel que $\sigma^+(\hat{x}) \neq \emptyset$. Alors le lemme 9.13 nous assure que $\forall i \in \sigma^+(\hat{x})$, il existe $T_i \subseteq [0, \sqrt{2\lambda}/\|a_i\|]$, un intervalle non-dégénéré de $\mathbb{R}$ contenant $|\hat{x}_i|$, tel que $\forall t \in T_i \setminus \{|\hat{x}_i|\}$, $\bar{x} = \hat{x}^{(i)} - s_i e_i t$ est aussi un minimiseur de G_{CEL0} ce qui est en contradiction avec le fait que $\hat{x}$ est strict et termine la démonstration. □

Pour un minimiseur global non-strict $\hat{x} \in \mathbb{R}^N$ de G_{CEL0}, fixer toutes ses composantes non nulles indexées par $\sigma^-(\hat{x})$ à zéro résulte, d'après le lemme 9.13 (ii), en un autre minimiseur global défini par

$$\forall i \in \mathbb{I}_N,\ \hat{x}_i^0 := \begin{cases} \hat{x}_i & \text{si } i \notin \sigma^-(\hat{x}) \\ 0 & \text{si } i \in \sigma^-(\hat{x}) \end{cases} = \hat{x}_i \mathbb{1}_{\left\{|\hat{x}_i| \geqslant \frac{\sqrt{2\lambda}}{\|a_i\|}\right\}}, \tag{9.26}$$

et qui vérifie

$$G_{\text{CEL0}}(\hat{x}^0) = G_{\text{CEL0}}(\hat{x}). \tag{9.27}$$

Basé sur le lemme 9.13, le théorème suivant établit une relation entre les minimiseurs *globaux* de G_{ℓ_0} et G_{CEL0}.

Theorème 9.16 (Lien entre les minimiseurs globaux de G_{ℓ_0} et G_{CEL0})**.**

(i) L'ensemble des minimiseurs globaux de G_{ℓ_0} est inclus dans l'ensemble des minimiseurs globaux de G_{CEL0},

$$\arg\min_{x\in\mathbb{R}^N} G_{\ell_0}(x) \subseteq \arg\min_{x\in\mathbb{R}^N} G_{\text{CEL0}}(x), \tag{9.28}$$

(ii) Réciproquement, si $\hat{x} \in \mathbb{R}^N$ est un minimiseur global de G_{CEL0}, alors $\hat{x}^0$, défini par (9.26), est un minimiseur global de G_{ℓ_0} et

$$G_{\text{CEL0}}(\hat{x}) = G_{\text{CEL0}}(\hat{x}^0) = G_{\ell_0}(\hat{x}^0). \tag{9.29}$$

Démonstration. La preuve est détaillée en annexe A.1.5 (page 192). □

Remarque 9.17. Comme $\hat{x}^0$ est un minimiseur global de G_{ℓ_0}, il est strict (Nikolova, 2013, théorème 4.4 (ii)) et nous pouvons conclure d'après (Nikolova, 2013, théorème 3.2) que $A_{\sigma(\hat{x}^0)}$ est de rang plein (i. e. $\operatorname{rank}(A_{\sigma(\hat{x}^0)}) = \sharp\sigma(\hat{x}^0)$). Ainsi, en notant $\hat{\sigma}^0 = \sigma(\hat{x}^0)$, nous avons

$$\hat{x}^0_{\hat{\sigma}^0} = \left((A_{\hat{\sigma}^0})^T A_{\hat{\sigma}^0}\right)^{-1} (A_{\hat{\sigma}^0})^T d \quad \text{et} \quad \hat{x}^0_{\mathbb{I}_N \setminus \hat{\sigma}^0} = 0. \tag{9.30}$$

Proposition 9.18 (Existence de minimiseurs globaux pour G_{CEL0})**.** *L'ensemble des minimiseurs globaux de G_{CEL0} est non-vide.*

Démonstration. (Nikolova, 2013, théorème 4.4 (i)) nous assure que l'ensemble des minimiseurs globaux de G_{ℓ_0} est non-vide ce qui, avec (9.28), entraine le résultat. □

Sous certaines conditions, les minimiseurs globaux des deux fonctionnelles coïncident exactement.

Corolaire 9.19.

$$\arg\min_{x\in\mathbb{R}^N} G_{\ell_0}(x) = \arg\min_{x\in\mathbb{R}^N} G_{\text{CEL0}}(x), \tag{9.31}$$

si et seulement si *pour tout couple $(\hat{x}^1, \hat{x}^2)$ de minimiseurs globaux de G_{ℓ_0} ($\hat{x}^1 \neq \hat{x}^2$),*

$$\|\hat{x}_1 - \hat{x}_2\|_0 > 1. \tag{9.32}$$

En particulier, c'est le cas lorsque G_{ℓ_0} admet un unique minimiseur global (voir Nikolova, 2013, pour l'unicité concernant G_{ℓ_0}).

Démonstration. La preuve est donnée en annexe A.1.6 (page 193). □

Il existe également un résultat similaire au théorème 9.16 pour les minimiseurs *locaux* (non globaux). Afin de démontrer ce résultat, nous avons besoin du lemme suivant.

Lemme 9.20. *Si* G_{CEL0} *admet un minimum en* $\hat{x} \in \mathbb{R}^N$*, alors*

$$\forall i \in \sigma^+(\hat{x}),\ \forall j \in \sigma(\hat{x})\backslash\{i\},\ \langle a_i, a_j \rangle = 0. \tag{9.33}$$

Démonstration. La preuve est détaillée en annexe A.1.7 (page 194). □

Theorème 9.21 (Lien entre les minimiseurs locaux de G_{ℓ_0} et G_{CEL0})**.** *Si* G_{CEL0} *admet un minimiseur local (non global) en* $\hat{x} \in \mathbb{R}^N$*. Alors* $\hat{x}^0$*, défini par* (9.26)*, est un minimiseur local (non global) de* G_{ℓ_0} *et* (9.29) *est vérifiée.*

Démonstration. La preuve est détaillée en annexe A.1.8 (page 195). □

Remarque 9.22. Bien que, pour un minimiseur global $\hat{x}$ de G_{CEL0}, $\hat{x}^0$ soit un minimiseur strict pour G_{ℓ_0} (remarque 9.17), ce n'est pas toujours le cas pour un minimiseur local (non global) de G_{CEL0}. En effet, soit $\hat{x}$ un minimiseur local (non global) de G_{CEL0} tel que $\operatorname{rank}(A_{\hat{\sigma}^0}) \leqslant \sharp\hat{\sigma}^0 - 1$ où $\hat{\sigma}^0 = \sigma(\hat{x}^0)$ et $\hat{x}^0$ est défini par (9.26) (on considère $\hat{\sigma}^0 \neq \emptyset$). Alors, d'après le théorème 9.21, $\hat{x}^0$ est un minimiseur local de G_{ℓ_0} mais n'est pas strict d'après (Nikolova, 2013, théorème 3.2). De plus, le point $\hat{x}^0$ défini par le théorème 9.21 comme un minimiseur local (non global) de G_{ℓ_0}, n'est pas assuré d'être un point critique (et donc possiblement un minimiseur) de G_{CEL0} contrairement au cas des minimiseurs globaux énoncé par le théorème 9.16.

Au regard du théorème 9.21 et de la remarque 9.22, de tout minimiseur local (non global) de G_{CEL0}, on peut construire par un simple seuillage un minimiseur local (non global) de G_{ℓ_0}. Cependant, il est essentiel de vérifier que ce minimiseur soit un point critique pour G_{CEL0} (ce qui n'est pas assuré par le théorème 9.21).

Remarque 9.23. Pour un minimiseur (local) $\hat{x}$ de G_{CEL0}, on peut fixer $\hat{x}_i$, $\forall i \in \hat{\sigma}^- = \sigma^-(\hat{x})$, à 0 ou à $-s_i \frac{\sqrt{2\lambda}}{\|a_i\|}$ afin obtenir un minimiseur (local), noté $\tilde{x}$, de G_{ℓ_0}. En effet, considérons $\{\omega_-, \omega_+\}$, une partition de $\hat{\sigma}^-$ (*i.e.* $\omega_- \subseteq \hat{\sigma}^-$, $\omega_+ \subseteq \hat{\sigma}^-$ tels que $\omega_- \cup \omega_+ = \hat{\sigma}^-$ et $\omega_- \cap \omega_+ = \emptyset$) et définissons

$$\forall i \in \mathbb{I}_N,\ \tilde{x}_i = \begin{cases} \hat{x}_i & \text{si } i \notin (\omega_- \cup \omega_+), \\ 0 & \text{si } i \in \omega_-, \\ -s_i \frac{\sqrt{2\lambda}}{\|a_i\|} & \text{si } i \in \omega_+. \end{cases} \tag{9.34}$$

où $s_i = \text{sign}(\langle a_i, A\hat{x}^{(i)} - d\rangle)$. Étant donné que $\hat{x}$ est un point critique de G_{CEL0}, le lemme 9.10 nous dit que $\forall i \in \hat{\sigma}^-$, $|\langle a_i, A\hat{x}^{(i)} - d\rangle| = \sqrt{2\lambda}\|a_i\|$ et $\forall i \in \hat{\sigma} \setminus \hat{\sigma}^-$, $\hat{x}_i = -\langle a_i, A\hat{x}^{(i)} - d\rangle / \|a_i\|^2$. Ainsi,

$$
\begin{aligned}
&\forall i \in \sigma(\tilde{x}),\ \tilde{x}_i = -\frac{\langle a_i, A\hat{x}^{(i)} - d\rangle}{\|a_i\|^2}, \\
&\underset{(9.33)\,\&\,(9.34)}{\iff} \tilde{x}_i = -\frac{\langle a_i, A\tilde{x}^{(i)} - d\rangle}{\|a_i\|^2} - \frac{1}{\|a_i\|^2} \sum_{j \in \omega_-} \underbrace{\langle a_i, a_j\rangle}_{=0} \hat{x}_j \\
&\qquad\qquad - \frac{1}{\|a_i\|^2} \sum_{\substack{j \in \omega_+ \\ j \neq i}} \underbrace{\langle a_i, a_j\rangle}_{=0} \left(\hat{x}_j + s_j \frac{\sqrt{2\lambda}}{\|a_j\|} \right), \\
&\iff \quad \langle a_i, A_{\sigma(\tilde{x})} \tilde{x}_{\sigma(\tilde{x})} - d\rangle = 0,
\end{aligned}
$$

qui, avec (Nikolova, 2013, Corollary 2.5), assure que $\tilde{x}$ est un minimiseur (local) de G_{ℓ_0}. Il existe $2^{\sharp\hat{\sigma}^-}$ minimiseurs pouvant être définis de la sorte. Parmi ces minimiseurs, $\hat{x}^0$ est le plus parcimonieux. Il est à noter que cette remarque peut être étendue aux points $\tilde{x}$ définis par (9.34) avec $\{\omega_-, \omega_+\}$ une partition de $\sigma^+(\hat{x})$ telle que $\forall (i,j) \in (\omega_+ \setminus \sigma(\hat{x}))^2$, $\langle a_i, a_j\rangle = 0$.

En d'autres termes, le théorème 9.21 est le pendant du théorème 9.16 pour les minimiseurs locaux (non globaux). En particulier, il affirme que le deuxième point du théorème 9.16 est également valable pour les minimiseurs locaux (non globaux) de G_{CEL0}. Cependant, la réciproque n'est pas vérifiée pour tous les minimiseurs locaux (non globaux) de G_{ℓ_0}. En effet, considérons le cas $N = 1$ (*i.e.* le problème 9.1) avec $a = 1$ et $d > \sqrt{2\lambda}$. Dans un tel cas, g_0 atteint un minimum global en $u = d$ et un minimum local en $u = 0$ alors que $g_0^{\star\star}$ admet un unique minimum global en $u = d$. Ainsi, tous les minimiseurs locaux de g_0 ne sont pas des minimiseurs locaux de $g_0^{\star\star}$. Cette remarque s'étend facilement au cas ND orthogonal étant donné que, dans ce cas, remplacer la norme-ℓ_0 par la pénalité CEL0 donne l'enveloppe convexe de G_{ℓ_0} (voir la section 9.2) qui n'admet pas de minimiseurs locaux (non globaux).

Plus généralement, pour toute matrice $A \in \mathbb{R}^{M \times N}$, considérons $\hat{x} \in \mathbb{R}^N$ un minimiseur local (non global) de G_{ℓ_0}. Alors, d'après le corolaire 9.12, $\hat{x}_{\hat{\sigma}}$ est solution du système suivant :

$$
(A_{\hat{\sigma}})^T A_{\hat{\sigma}} \hat{x}_{\hat{\sigma}} = (A_{\hat{\sigma}})^T d, \tag{9.35}
$$

où $\hat{\sigma} = \sigma(\hat{x})$. Afin d'être également un point critique de G_{CEL0}, $\hat{x}$ doit vérifier les conditions (9.20). On en déduit que $\hat{x}$ est un point critique de G_{CEL0} si et seulement si, en plus d'être solution de (9.35), il vérifie les conditions suivantes :

$$
\begin{cases}
|\langle a_i, A\hat{x}^{(i)} - d\rangle| \leqslant \sqrt{2\lambda}\|a_i\| & \forall i \notin \sigma(\hat{x}), \\
|\langle a_i, A\hat{x}^{(i)} - d\rangle| = \sqrt{2\lambda}\|a_i\| & \forall i \in \sigma^-(\hat{x}), \\
|\langle a_i, A\hat{x}^{(i)} - d\rangle| \geqslant \sqrt{2\lambda}\|a_i\| & \forall i \in \sigma(\hat{x}) \setminus \sigma^-(\hat{x}).
\end{cases} \tag{9.36}
$$

Par conséquent si $\hat{x}$, un minimiseur local (non global) de G_{ℓ_0}, ne vérifie pas (9.36), alors il n'est pas point critique de G_{CEL0} et donc G_{CEL0} n'admet pas de minimum local (non global) en $\hat{x}$.

En particulier, si $\hat{x}$ est solution de (9.35) alors $\forall i \in \sigma(\hat{x})\ \hat{x}_i = -\langle a_i, A\hat{x}^{(i)} - d\rangle / \|a_i\|^2$ et on obtient,

$$\forall i \in \sigma(\hat{x}),\ \begin{cases} i \in \sigma(\hat{x}) \backslash \sigma^-(\hat{x}) \ \left(\text{i. e. } |\hat{x}_i| \geqslant \frac{\sqrt{2\lambda}}{\|a_i\|}\right) & \Rightarrow \quad |\langle a_i, A\hat{x}^{(i)} - d\rangle| \geqslant \sqrt{2\lambda}\|a_i\|, \\ i \in \sigma^-(\hat{x}) \ \left(\text{i. e. } 0 < |\hat{x}_i| < \frac{\sqrt{2\lambda}}{\|a_i\|}\right) & \Rightarrow \quad 0 < |\langle a_i, A\hat{x}^{(i)} - d\rangle| < \sqrt{2\lambda}\|a_i\|. \end{cases}$$

De fait, un minimiseur local (non global) $\hat{x}$ de G_{ℓ_0}, qui est nécessairement solution de (9.35), ne peut pas vérifier la seconde ligne de (9.36) dès lors que $\sigma^-(\hat{x}) \neq \emptyset$. De plus, rien ne permet d'assurer qu'il vérifie la première ligne de (9.36). Ainsi, tous les minimiseurs locaux (non globaux) de G_{ℓ_0} ne sont pas des points critiques de G_{CEL0}.

On en déduit que G_{CEL0} *élimine les minimiseurs locaux (non globaux)* $\hat{x}$ *de* G_{ℓ_0} *tels que* $\sigma^-(\hat{x}) \neq \emptyset$. Nous résumons ce résultat dans la proposition suivante.

Proposition 9.24. *Soit* $\hat{x} \in \mathbb{R}^N$ *minimiseur de* G_{ℓ_0}. *Si* $\hat{x}$ *est aussi un point critique de* G_{CEL0}, *alors* $\sigma^-(\hat{x}) = \emptyset$.

Notons que les minimiseurs globaux de G_{ℓ_0} vérifient nécessairement $\sigma^-(\hat{x}) = \emptyset$ d'après la proposition 9.11 ce qui est en accord avec le fait qu'il sont préservés par G_{CEL0} (théorème 9.16).

Enfin, une conséquence du corolaire 9.15 et des théorèmes 9.16 et 9.21 est donnée par le corolaire suivant.

Corolaire 9.25. *Soit* $\hat{x}$ *un minimiseur (local) strict de* G_{CEL0}. *Alors* $\hat{x}$ *est un minimiseur (local) strict de* G_{ℓ_0}. *De plus,* $A_{\sigma(\hat{x})}$ *est de rang plein.*

Démonstration. Étant donné que $\hat{x}$ est un minimiseur (local) strict de G_{CEL0}, le corolaire 9.15 nous assure que $\sigma^-(\hat{x}) = \emptyset$. Il s'en suit, d'après les théorèmes 9.16 et 9.21, que $\hat{x}$ est un minimiseur (local) de G_{ℓ_0} et que $G_{\text{CEL0}}(\hat{x}) = G_{\ell_0}(\hat{x})$.

Supposons maintenant que $\hat{x}$ n'est pas strict pour G_{ℓ_0}. Alors, il existe un voisinage $\mathcal{V}_0 \subset \mathbb{R}^N$ de $\hat{x}$ tel que pour tout $\mathcal{V} \subset \mathcal{V}_0$ contenant $\hat{x}$, il existe $\bar{x} \in \mathcal{V}$ vérifiant $G_{\ell_0}(\bar{x}) = G_{\ell_0}(\hat{x})$. Comme par définition G_{CEL0} minore G_{ℓ_0} (remarque 9.5), nous avons

$$\forall\, \mathcal{V} \subset \mathcal{V}_0,\ \text{t.q. } \hat{x} \in \mathcal{V}, \exists \bar{x} \in \mathcal{V},\ G_{\text{CEL0}}(\bar{x}) \leqslant G_{\ell_0}(\bar{x}) = G_{\ell_0}(\hat{x}) = G_{\text{CEL0}}(\hat{x}), \tag{9.37}$$

ce qui est en contradiction avec le fait que $\hat{x}$ est un minimiseur (local) *strict* de G_{CEL0}. Par conséquent, $\hat{x}$ est un minimiseur (local) *strict* de G_{ℓ_0}. Enfin, (Nikolova, 2013, Theorem 3.2) montre que $\operatorname{rank}(A_{\sigma(\hat{x})}) = \sharp\sigma(\hat{x})$ ce qui termine la preuve. □

Nous venons de caractériser les liens entre les minimiseurs des fonctionnelles G_{ℓ_0} et G_{CEL0}. Cependant, nous ne disposons que d'une caractérisation des points critiques de G_{CEL0} (lemme 9.10). Il serait donc intéressant d'être en mesure de discerner, parmi les points critiques de G_{CEL0}, ceux qui sont des minimiseurs. Le théorème suivant caractérise les minimiseurs stricts de G_{CEL0}.

Theorème 9.26 (Minimiseurs (locaux) stricts de G_{CEL0}). *Soit* $\hat{x} \in \mathbb{R}^N$, *un point critique de* G_{CEL0}. *Alors l'équivalence suivante est vérifiée :*

$$\hat{x} \textit{ est un minimiseur (local) strict de } G_{\text{CEL0}} \iff \sigma^+(\hat{x}) = \emptyset \ \textit{ et } \ \operatorname{rank}(A_{\sigma(\hat{x})}) = \sharp\sigma(\hat{x}). \tag{9.38}$$

Démonstration. La preuve est donnée en annexe A.1.9 (page 195). □

Une telle caractéristique des minimiseurs stricts de G_{CEL0} est importante notamment lorsque nous sommes dans le cadre du corolaire 9.19 où les minimiseurs globaux de G_{CEL0} sont stricts. En revanche, en ce qui concerne les minimiseurs (locaux) non-stricts, nous n'avons que la condition suffisante suivante.

Theorème 9.27 (Condition suffisante pour être minimiseur (local) de G_{CEL0})**.** *Soit* $\hat{x} \in \mathbb{R}^N$ *un point critique de* G_{CEL0}*. Alors,*

$$\sigma^+(\hat{x}) = \emptyset \implies \hat{x} \text{ est un minimiseur (local) de } G_{CEL0}. \tag{9.39}$$

Démonstration. On considère $\rho > 0$ défini comme dans la preuve du théorème 9.26. Alors, d'après les équations (A.56) et (A.57) de cette même preuve, on a

$$\forall \varepsilon \in \mathcal{B}_\infty(0_{\mathbb{R}^N}, \rho) \setminus \{0_{\mathbb{R}^N}\},\ G_{CEL0}(\hat{x} + \varepsilon) \geqslant G_{CEL0}(\hat{x}), \tag{9.40}$$

ce qui termine la démonstration. □

Cependant, sous certaines conditions sur la matrice $A \in \mathbb{R}^{M\times N}$, on peut avoir une caractérisation complète (conditions nécessaires et suffisantes) des minimiseurs (locaux) non-stricts de G_{CEL0}.

Corolaire 9.28. *Supposons que la matrice* $A \in \mathbb{R}^{M\times N}$ *soit telle que*

$$\forall i \in \mathbb{I}_N,\ \forall j \in \mathbb{I}_N \setminus \{i\},\ \langle a_i, a_j \rangle \neq 0. \tag{9.41}$$

Alors, pour $\hat{x} \in \mathbb{R}^N$ *point critique de* G_{CEL0} *tel que* $\|\hat{x}\|_0 > 1$*, l'équivalence suivante est vérifiée :*

$$\hat{x} \text{ est un minimiseur (local) de } G_{CEL0} \iff \sigma^+(\hat{x}) = \emptyset. \tag{9.42}$$

Notons que dans ce cas $\hat{x}$ *est aussi un minimiseur (local) de* G_{ℓ_0} *d'après le théorème 9.21.*

Démonstration. $\Longleftarrow$ est direct d'après le théorème 9.27. Afin de démontrer la réciproque ($\Longrightarrow$), considérons $\hat{x} \in \mathbb{R}^N$ un minimiseur (local) de G_{CEL0} vérifiant $\|\hat{x}\|_0 > 1$. Supposons que $\sigma^+(\hat{x}) \neq \emptyset$. Le lemme 9.20 nous assure alors que

$$\forall i \in \sigma^+(\hat{x}),\ \forall j \in \sigma(\hat{x}) \setminus \{i\},\ \langle a_i, a_j \rangle = 0. \tag{9.43}$$

Le fait que $\|\hat{x}\|_0 > 1$ (i. e. $\sharp\sigma(\hat{x}) > 1$) et $\sharp\sigma^+(\hat{x}) \geqslant 1$ implique que

$$\exists (i, j) \in \sigma^+(\hat{x}) \times \sigma(\hat{x})\ \text{ t.q. }\ i \neq j\ \text{ et } \langle a_i, a_j \rangle = 0, \tag{9.44}$$

ce qui est en contradiction avec l'hypothèse faite sur A et termine la preuve. □

Sous les conditions du corolaire 9.28, qui peuvent être vues comme l'extrême opposé du cas orthogonal où toutes les colonnes sont orthogonales deux à deux, il est possible de conclure sur le fait qu'un point critique est minimiseur (local) ou non dès lors qu'il admet au moins deux composantes non-nulles. Notons qu'une telle condition sur A est généralement vérifiée pour des problèmes où les colonnes de A sont très corrélées (e. g. déconvolution).

9.3.1.3 *Retour sur le cas orthogonal*

Comme cela a été montré dans la section 9.2, lorsque les colonnes de la matrice A sont deux à deux orthogonales et non nulles, la fonctionnelle G_{CEL0} définit l'enveloppe convexe de G_{ℓ_0}. Dans ce cas, G_{CEL0} est alors convexe et tous ses points critiques sont des minimiseurs globaux à partir desquels le théorème 9.16 nous permet d'extraire des minimiseurs globaux de G_{ℓ_0}. En utilisant l'orthogonalité de A, la caractérisation des points critiques de G_{CEL0} donnée par le lemme 9.10 se réécrit

$$\forall i \in \mathbb{I}_N, \begin{cases} \hat{x}_i = 0 & \text{si } |\langle a_i, d\rangle| \leqslant \sqrt{2\lambda}\|a_i\|, \\ \hat{x}_i = -s_i t,\ t \in \left[0, \dfrac{\sqrt{2\lambda}}{\|a_i\|}\right] & \text{si } |\langle a_i, d\rangle| = \sqrt{2\lambda}\|a_i\|, \\ \hat{x}_i = \dfrac{\langle a_i, d\rangle}{\|a_i\|^2} & \text{si } |\langle a_i, d\rangle| \geqslant \sqrt{2\lambda}\|a_i\|. \end{cases} \tag{9.45}$$

Posons $z = A^T d$, alors le minimiseur global de G_{CEL0} le plus parcimonieux, noté $\hat{x}^0$, est donné par la règle de seuillage suivante :

$$\forall i \in \mathbb{I}_N,\ \hat{x}_i^0 = \frac{z_i}{\|a_i\|^2} \mathbb{1}_{\{|z_i| > \sqrt{2\lambda}\|a_i\|\}}. \tag{9.46}$$

Finalement, lorsque $A^T A$ est diagonale, remplacer la norme-ℓ_0 par la pénalité CEL0 transforme le problème initial en un problème convexe (son enveloppe convexe) pour lequel le minimiseur global le plus parcimonieux s'obtient par un simple seuillage dur. Nous retrouvons ici un résultat bien connu. Notons que si les colonnes de A ne sont pas normalisées (i. e. $\|a_i\| = 1,\ \forall i \in \mathbb{I}_N$), le seuillage en (9.46) est alors différent pour chacune des composantes de x.

9.3.1.4 *Étude des minimiseurs dits «Coordinate-Wise» de G_{CEL0} et G_{ℓ_0}*

Nous nous intéressons maintenant à un type particulier de minimiseurs de G_{ℓ_0} et G_{CEL0} : les minimiseurs Coordinate-Wise (CW). Cette notion est sans doute apparue pour la première fois dans les travaux de Geman et Reynolds (1992) et a récemment été utilisée par Beck et Eldar (2013) dans le contexte du problème (C_k).

Définition 9.29. (Minimiseur «Coordinate-Wise») Soit $F : \mathbb{R}^N \to \mathbb{R}$, alors $\hat{x} \in \mathbb{R}^N$ est appelé minimiseur CW de F si

$$\forall i \in \mathbb{I}_N,\ F(\hat{x}) = \min_{t \in \mathbb{R}} F(\hat{x} + t e_i) \tag{9.47}$$

Notons que pour une fonction F générale, un minimiseur CW peut ne pas être un minimiseur de F (e. g. nous pouvons avoir $F(\hat{x}) > F(\hat{x} + t(e_i + e_j))$ pour tout t dans un ouvert contenant zéro). Cependant, nous avons aussi le résultat suivant.

Lemme 9.30. *Tout minimiseur global d'une fonction $F : \mathbb{R}^N \to \mathbb{R}$ est CW.*

Démonstration. La preuve découle directement de (9.47) et de la définition d'un minimiseur global. □

Ainsi, être un minimiseur CW est une condition nécessaire d'optimalité globale. De plus, certains minimiseurs locaux (non globaux) peuvent ne pas être CW (il suffit de regarder le

problème ℓ_2-ℓ_0 en 1D). Cette notion est donc intéressante dans le sens où elle permet de réduire le nombre de «candidats» (parmi les minimiseurs d'une fonction non-convexe) qui pourraient être des minimiseurs globaux (Beck et Eldar, 2013).

Nous avons vu précédemment que certains minimiseurs locaux (non globaux) de G_{ℓ_0} étaient éliminés par G_{CEL0} (proposition 9.24). La notion de minimiseur CW va nous permettre de préciser cette propriété de la fonctionnelle CEL0. Commençons par caractériser les minimiseurs CW de G_{CEL0} avec le lemme suivant.

Lemme 9.31. *$\hat{x} \in \mathbb{R}^N$ est un minimiseur CW de G_{CEL0} si et seulement si c'est un point critique de G_{CEL0}*

Démonstration.

$\Longrightarrow$ Soit $\hat{x} \in \mathbb{R}^N$ un minimiseur CW de G_{CEL0}. Alors, par définition on a

$$\forall i \in \mathbb{I}_N,\ 0 \in \partial G^i_{\text{CEL0}}(\hat{x}_i; \hat{x}^{(i)}), \tag{9.48}$$

et, comme $\partial G_{\text{CEL0}} = \prod_i \partial G^i_{\text{CEL0}}$, $0_{\mathbb{R}^N} \in \partial G_{\text{CEL0}}(\hat{x})$ et $\hat{x}$ est un point critique de G_{CEL0}.

$\Longleftarrow$ Soit $\hat{x} \in \mathbb{R}^N$ un point critique de G_{CEL0}. Alors $0_{\mathbb{R}^N} \in \partial G_{\text{CEL0}}(\hat{x})$ et (9.48) est vérifié. Enfin, la convexité de $G^i_{\text{CEL0}}(\cdot; x^{(i)})$, $\forall i \in \mathbb{I}_N$ (proposition 9.6) termine la démonstration. □

Theorème 9.32. *Les deux assertions suivantes sont équivalentes :*

(i) $\hat{x}$ est un minimiseur CW de G_{ℓ_0} ,

(ii) $\hat{x}$ est un minimiseur CW de G_{CEL0} et $\sigma^-(\hat{x}) = \emptyset$.

Démonstration. La preuve est donnée en annexe A.1.10 (page 197). □

Nous pouvons déduire de ce résultat le corolaire suivant.

Corolaire 9.33. *Tout minimiseur CW de G_{ℓ_0} est minimiseur (local) de G_{ℓ_0}.*

Démonstration. Soit $\hat{x}$ un minimiseur CW de G_{ℓ_0}. Alors, d'après le lemme 9.31 et le théorème 9.32, $\hat{x}$ est un point critique de G_{CEL0} tel que $\sigma^-(\hat{x}) = \emptyset$ et le lemme 10.1 (voir chapitre 10) stipule que $\hat{x}$ est un minimiseur (local) de G_{ℓ_0}. □

Finalement, les résultats qui précèdent montrent que la fonctionnelle CEL0 élimine en particulier tous les minimiseurs locaux de G_{ℓ_0} qui ne sont pas CW. Cela sera illustré dans la section suivante. Par ailleurs, obtenir un minimiseur CW de G_{ℓ_0} peut être réalisé en assurant la convergence d'un algorithme vers un point critique de G_{CEL0} pour lequel $\sigma^-(\hat{x}) = \emptyset$. Dans le chapitre 10, nous présenterons une méthode permettant d'assurer un telle convergence.

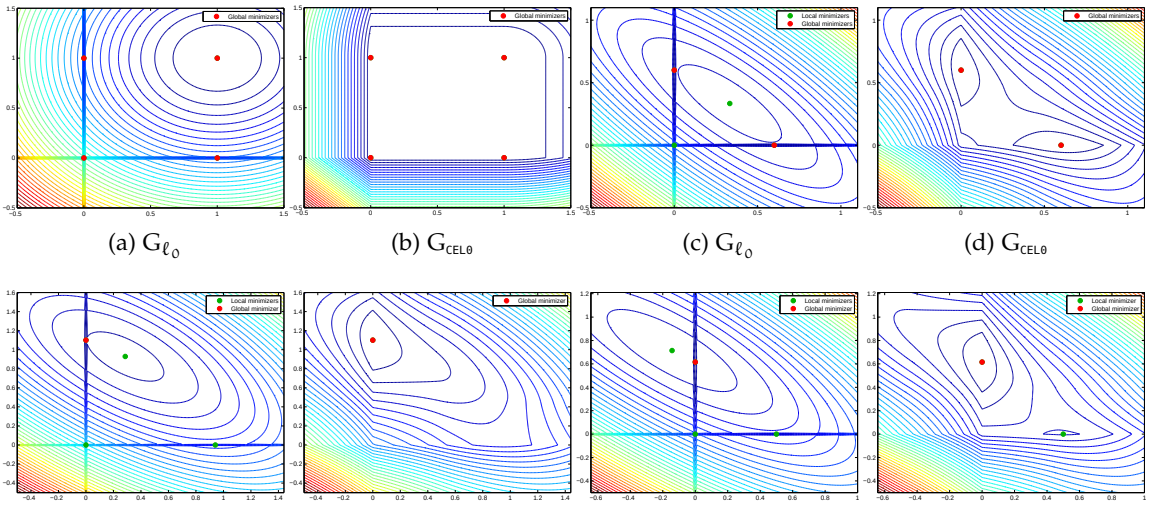
9.3.2 *Illustrations numériques*

Dans cette section, nous illustrons numériquement, sur des problèmes de petite taille, le fait que G_{CEL0} «élimine» des minimiseurs locaux (en particulier les CW) de G_{ℓ_0}.

9.3.2.1 *Exemples en dimension 2*

Nous considérons le cas $N = M = 2$. Sur la figure 35, les lignes de niveau des deux fonctionnelles G_{ℓ_0} et G_{CEL0} sont représentées pour différents $A \in \mathbb{R}^{2\times 2}$, $d \in \mathbb{R}^2$ et $\lambda > 0$. Dans tous les cas, les minimiseurs globaux de G_{ℓ_0} le sont aussi pour G_{CEL0} comme cela est affirmé par le théorème 9.16. Concernant l'exemple proposé sur les figures 35a et 35b, $\forall x^\star \in [0,1]^2$, $x^\star$ est un minimiseur global de G_{CEL0}. Soit $\hat{x} \in]0,1[^2$ un de ces optima globaux de G_{CEL0}, alors il est clair que, pour cet exemple, $\hat{x}^0$ défini par (9.26) est un minimiseur global pour les deux fonctionnelles G_{CEL0} et G_{ℓ_0} comme énoncé par le second point du théorème 9.16. Notons que cet exemple illustre également la remarque 9.23.

Les autres exemples proposés sur la figure 35 illustrent le fait que G_{CEL0} admet en général moins de minimiseurs locaux (non globaux) que G_{ℓ_0}. Lorsque $A \in \mathbb{R}^{2\times 2}$, G_{ℓ_0} admet toujours quatre minimiseurs (locaux et globaux confondus). Pour l'exemple des figures 35c et 35d nous pouvons voir que G_{CEL0}, au contraire, admet uniquement deux minimiseurs globaux et pour l'exemple correspondant aux figures 35e et 35f, G_{CEL0} n'a qu'un unique minimum global. Dans ces deux cas, l'utilisation de la pénalité CEL0 permet *d'éliminer tous les minima locaux (non globaux) de* G_{ℓ_0} tout en préservant les minima globaux. Cependant, ce n'est pas toujours le cas comme nous pouvons le constater avec l'exemple des figures 35g et 35h où G_{CEL0} a un minimiseur global et un local (non global). Pour cet exemple G_{CEL0} admet deux minimiseurs locaux (non globaux) de moins que G_{ℓ_0}.



(a) G_{ℓ_0} (b) G_{CEL0} (c) G_{ℓ_0} (d) G_{CEL0}

(e) G_{ℓ_0} (f) G_{CEL0} (g) G_{ℓ_0} (h) G_{CEL0}

FIGURE 35 – Lignes de niveau de G_{ℓ_0} et G_{CEL0} pour quatre exemples où $N = M = 2$. (a)-(b) $A = [1, 0; 0, 1]$, $b = [1; 1]$ et $\lambda = 0.5$ (c)-(d) $A = [1, 2; 2, 1]$, $b = [1; 1]$ et $\lambda = 0.5$ (e)-(f) $A = [0.5, 2; 2, 1]$, $b = [2; 1.5]$ et $\lambda = 0.5$ (g)-(h) $A = [3, 2; 1, 3]$, $b = [1; 2]$ et $\lambda = 1$. Les minimiseurs locaux (non globaux) sont représentés en vert et les globaux en rouge.

9.3.2.2 *Exemples en plus grande dimension*

Nous considérons maintenant des exemples en plus grande dimension tout en restant suffisamment faible pour envisager une recherche exhaustive. En suivant les simulations

numériques proposées dans (Nikolova, 2013, §6.2), nous définissons G_{ℓ_0} avec $M = 5$, $N = 10$,

$$A = \begin{pmatrix} 7 & 2 & 4 & 9 & 0 & 3 & 3 & 6 & 6 & 7 \\ 3 & 4 & 9 & 3 & 3 & 9 & 1 & 3 & 1 & 5 \\ 5 & 4 & 2 & 4 & 0 & 7 & 1 & 9 & 2 & 9 \\ 8 & 4 & 0 & 9 & 6 & 0 & 4 & 2 & 3 & 7 \\ 6 & 3 & 6 & 5 & 0 & 9 & 0 & 0 & 3 & 8 \end{pmatrix}, \tag{9.49}$$

et

$$d = Ax^\star \quad \text{avec} \quad x^\star = (0, 1, 8, 0, 3, 0, 0, 0, 0, 9)^T. \tag{9.50}$$

Pour ce problème, nous allons nous intéresser aux minimiseurs (locaux) stricts de G_{ℓ_0} et G_{CEL0}. Comme cela est énoncé par Nikolova (2013, théorème 4.4), les minimiseurs globaux de G_{ℓ_0} sont stricts. Cela donne un intérêt tout particulier aux minimiseurs stricts. Toujours d'après Nikolova (2013, corolaire 3.3), un minimiseur (local) strict de G_{ℓ_0} peut être facilement obtenu en choisissant un support $\omega \in \Omega_{max}$, où Ω_{max} est défini par (Nikolova, 2013, définition 3.1)

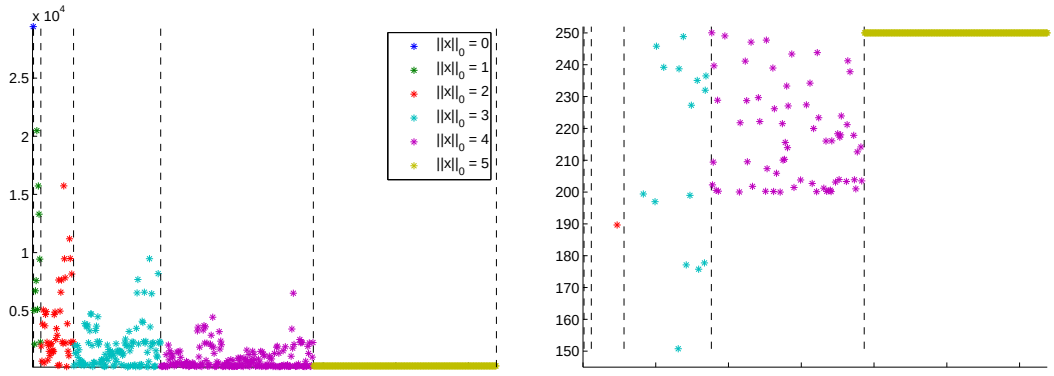
$$\Omega_{max} = \bigcup_{r=0}^{M} \Omega_r \quad \text{où} \quad \Omega_r = \{\omega \subset \mathbb{I}_N \ : \ \#\omega = r = \operatorname{rank}(A_\omega)\}, \tag{9.51}$$

et en résolvant le système des équations normales restreintes au support choisi (comme en (9.24)). En d'autres termes, les minimiseurs (locaux) stricts de G_{ℓ_0} sont ceux dont le support appartient à Ω_{max}.

Ainsi, nous pouvons calculer tous les minimiseurs (locaux) stricts de G_{ℓ_0} en résolvant les systèmes d'équations normales restreintes à tous les supports de Ω_{max} qui est un ensemble fini. Cela a été réalisé pour G_{ℓ_0} défini avec (9.49) et (9.50). Les résultats sont présentés sur la figure 36a en respectant la même convention de représentation que celle utilisée dans (Nikolova, 2013) où l'axe des abscisses liste tous les minimiseurs (locaux) stricts en fonction de la taille de leur support et l'axe des ordonnées représente la valeur de la fonctionnelle G_{ℓ_0} associée.

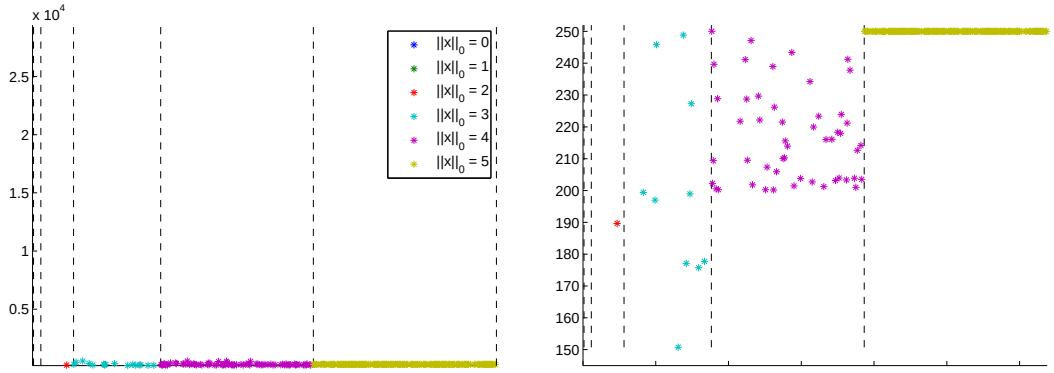
Parmi les 638 minimiseurs (locaux) stricts de G_{ℓ_0}, nous pouvons nous intéresser à ceux qui sont des points critiques de G_{CEL0} et qui vérifient donc les conditions (9.36). Ces points sont représentés sur la figure 36c. On remarque que seulement 283 des 638 minimiseurs (locaux) stricts de G_{ℓ_0} sont ainsi préservés. En vertu du corolaire 9.25, ces 283 points critiques de G_{CEL0} contiennent tous les minimiseurs (locaux) stricts de G_{CEL0}. En effet, supposons que $\hat{x} \in \mathbb{R}^N$, un minimiseur (local) strict de G_{CEL0}, n'est pas représenté sur la figure 36c. Alors, par construction, il n'est pas non plus représenté sur la figure 36a. Or, d'après le corolaire 9.25, c'est également un minimiseur (local) strict de G_{ℓ_0} ce qui contredit le fait que la figure 36a contient tous les minimiseurs (locaux) stricts de G_{ℓ_0} et prouve l'affirmation précédente.

Cependant, les minimiseurs (locaux) stricts de G_{ℓ_0} qui sont aussi des points critiques de G_{CEL0} (figure 36c) ne sont pas nécessairement des minimiseurs de G_{CEL0}. En effet, A vérifiant les hypothèses du corolaire 9.28 et en remarquant sur la figure 36c que tous les minimiseurs stricts de G_{ℓ_0} tels que $\|\hat{x}\|_0 \leqslant 1$ ne sont pas des points critiques de G_{CEL0} (pour cet exemple), une condition nécessaire et suffisante pour que ceux qui sont des points critiques de G_{CEL0} soient des minimiseurs est donnée par $\sigma^+(\hat{x}) = \emptyset$. Or, nous savons uniquement que les



(a) Les **638** minimiseurs (locaux) stricts de G_{ℓ_0}

(b) Zoom de 36a le long de l'axe y



(c) Les **283** qui sont points critiques de G_{CEL0}

(d) Zoom de 36c le long de l'axe y

FIGURE 36 – (a)-(b) Minimiseurs (locaux) stricts de G_{ℓ_0} définie avec les équations (9.49) et (9.50) pour $\lambda = 50$. (c)-(d) Ceux vérifiant les conditions (9.36) et qui sont donc des points critiques de G_{CEL0}.

points de la figure 36c vérifient $\sigma^-(\hat{x}) = \emptyset$ (proposition 9.24) et $\sigma^-(\hat{x}) \subseteq \sigma^+(\hat{x})$ ne nous permet pas de conclure.

Par contre, les points critiques de la figure 36c qui sont des minimiseurs de G_{CEL0} sont nécessairement stricts (théorème 9.26) étant donné qu'ils sont stricts pour G_{ℓ_0} impliquant $\text{rank}(A_{\sigma(\hat{x})}) = \sharp\sigma(\hat{x})$ par définition de Ω_{max} en (9.51).

Par conséquent, utiliser la pénalité CEL0 au lieu de la norme-ℓ_0 semble *éliminer un nombre non négligeable de minimiseurs locaux (non globaux) stricts de* G_{ℓ_0}. Cette propriété est intéressante étant donné que les algorithmes utilisés pour la minimisation de G_{CEL0} auront une «plus faible probabilité» de converger vers un minimiseur local non optimal. Nous pouvons aussi observer, sur les zooms 36b and 36d, que l'unique minimiseur global de G_{ℓ_0} (voir (NIKOLOVA, 2013) pour l'unicité sur cet exemple) est préservé par G_{CEL0} illustrant ainsi le théorème 9.16. Finalement, il est à noter que parmi les minimiseurs locaux (non globaux) stricts de G_{ℓ_0}, G_{CEL0} semble préserver ceux pour lesquels la valeur de la fonction objectif G_{ℓ_0} ou G_{CEL0} (nous avons égalité des deux fonctionnelles pour ces points) est la plus faible.

La table 3 dénombre les minimiseurs (locaux) stricts de G_{ℓ_0} pour différentes valeurs de M et N. Pour chaque couple (M, N), 1000 matrices $A \in \mathbb{R}^{M\times N}$ et $d \in \mathbb{R}^M$ ont été générées aléatoirement selon une distribution uniforme et le nombre de minimiseurs (locaux) stricts de G_{ℓ_0} préservés par G_{CEL0} (on notera ce nombre P) a été calculé. Nous reportons dans la table 3 la valeur minimale, maximale et moyenne de P sur les 1000 réalisations. Ces expériences sont en accord avec le fait que G_{CEL0} admet moins de minimiseurs locaux (non

globaux)[3] stricts que G_{ℓ_0}. Nous pouvons aussi noter, à travers les résultats de la table 3, que le nombre de minimiseurs stricts éliminés par G_{CEL0} est plus important pour de grandes valeurs de λ.

TABLE 3 – Nombre de minimiseurs (locaux) stricts ($\#\Omega_{max}$) de G_{ℓ_0} ainsi que le nombre P d'entre eux qui sont des points critiques de G_{CEL0}. Ces valeurs sont calculées à partir de 1000 matrices générées aléatoirement selon une distribution uniforme. P_{min}, P_{max} et $\bar{P}$ correspondent respectivement aux valeurs minimale, maximale et moyenne de P sur les 1000 réalisations. Cette expérience est répétée pour différentes tailles de matrices et deux valeurs différentes de λ.

		$\#\Omega_{max}$	P_{min}	P_{max}	$\bar{P}$
$\lambda = 0.5$	$M = 3, N = 5$	26	1	20	9
	$M = 5, N = 10$	638	39	347	206
	$M = 7, N = 15$	16384	2237	7374	5037
$\lambda = 0.1$	$M = 3, N = 5$	26	4	23	13
	$M = 5, N = 10$	638	173	444	306
	$M = 7, N = 15$	16384	5370	9431	7570

9.4 CONCLUSION

Dans ce chapitre, nous avons introduit la pénalité CEL0 et proposé une analyse théorique des liens qui existent entre les minimiseurs de la fonctionnelle CEL0 sous-jacente et ceux du critère ℓ_2-ℓ_0. Cette relaxation de la pseudo norme-ℓ_0 apparaît naturellement lorsqu'on s'intéresse au calcul de l'enveloppe convexe du problème ℓ_2-ℓ_0 dans les cas unidimensionnel et multidimensionnel orthogonal.

Bien que cette propriété d'enveloppe convexe ne soit plus vérifiée en dehors de ces deux cas particuliers, la relaxation continue G_{CEL0} de G_{ℓ_0} obtenue en remplaçant la norme-ℓ_0 par la pénalité CEL0 n'en reste pas moins intéressante dans un cadre général. En effet, les minimiseurs globaux de G_{CEL0} contiennent ceux de G_{ℓ_0} et réciproquement, de tout minimiseur global de G_{CEL0} on peut facilement en déduire un minimiseur global de G_{ℓ_0} (théorème 9.16). Et même, sous certaines conditions, les ensembles des minimiseurs globaux des deux fonctionnelles coïncident (corolaire 9.19). Ensuite nous avons aussi montré, avec le théorème 9.21, que ce résultat s'étend partiellement aux minimiseurs locaux (non globaux). Tout minimiseur local de G_{CEL0} est associé à un minimiseur local de G_{ℓ_0} au travers d'une simple opération de seuillage. Cependant, contrairement aux minimiseurs globaux, la réciproque n'est pas toujours vérifiée pour de tels minimiseurs locaux de G_{ℓ_0}. Plus précisément, G_{CEL0} élimine les minimiseurs locaux $\hat{x}$ de G_{ℓ_0} pour lesquels $\sigma^-(\hat{x}) \neq \emptyset$ (proposition 9.24) ou encore les minimiseurs de G_{ℓ_0} qui ne sont pas CW. Cela a été illustré numériquement sur des exemples en petite dimension montrant qu'un nombre non négligeable de minimiseurs locaux (non globaux) de G_{ℓ_0} n'étaient pas préservés par G_{CEL0}.

De nombreuses perspectives s'ouvrent alors autour de cette pénalité et de ses liens avec le problème ℓ_2-ℓ_0 initial.

— comment minimiser la relaxation continue G_{CEL0} ? Bien que ce problème soit toujours non convexe, il est continu et les liens établis entre les fonctionnelles G_{CEL0} et G_{ℓ_0} permettent d'envisager l'utilisation d'algorithmes récents d'optimisation non-convexe ne pouvant pas s'appliquer directement à G_{ℓ_0}, de part les discontinuités de la norme-ℓ_0,

3. Le théorème 9.16 assure que les minimiseurs éliminés ne sont pas globaux.

mais pouvant parfaitement être utilisés pour la minimisation de G_{CEL0}. Cette problématique sera discutée dans le Chapitre 10.

— le présent chapitre présente la pénalité CEL0 et montre les propriétés intéressantes qui existent entre les minimiseurs des deux fonctionnelles G_{CEL0} et G_{ℓ_0}. Cependant, il n'explique pas pourquoi cela fonctionne. En d'autres termes, quelle(s) particularité(s) possède la pénalité CEL0 par rapport aux nombreuses autres pénalités proposées dans la littérature pour approcher continument la norme-ℓ_0 ? Il est alors naturel de se poser la question de l'existence d'autres pénalités continues conduisant à des relaxations de G_{ℓ_0} possédant le même genre de propriétés que G_{CEL0}. Cette question sera adressée dans le Chapitre 12.

— enfin, une extension tout à fait intéressante concerne le cas où le terme d'attache aux données n'est pas quadratique. Existe-t-il une pénalité analogue à CEL0 pour de tels problèmes ? Cette question est importante étant donné que de nombreux problèmes, notamment en apprentissage, utilisent des termes d'attache aux données non quadratiques. Nous discuterons d'une telle extension dans les perspectives de cette partie du manuscrit.

10

ALGORITHMES POUR LA MINIMISATION DU CRITÈRE ℓ_2-CEL0

Sommaire

L'étude menée dans le chapitre précédent montre que le problème (P_λ) peut être abordé par la minimisation de la fonctionnelle continue G_{CEL0} de manière équivalente (au sens des minimiseurs). Nous nous intéressons donc, dans ce chapitre, à la minimisation de G_{CEL0}. L'optimisation non-convexe et non-différentiable est un domaine en plein essor où de nombreux algorithmes (sous-optimaux), pouvant s'appliquer directement à G_{CEL0}, ont été proposés très récemment. Nous commençons donc par présenter quelques uns de ces algorithmes avant de se focaliser plus précisément sur certains d'entre eux.

10.1 REVUE DES ALGORITHMES DITS «NONSMOOTH-NONCONVEX»

Dans cette section nous présentons cinq familles d'algorithmes d'optimisation non-convexe pouvant être utilisés pour la minimisation de G_{CEL0}.

10.1.1 *Forward-Backward Splitting*

L'algorithme FBS, évoqué dans le chapitre 8 (algorithme 2, page 77), est également applicable pour des fonctionnelles non-convexes. En effet, les récents travaux de Attouch et al. (2013) assurent la convergence de la séquence $(x^k)_{k\in\mathbb{N}}$, générée par

$$x^{n+1} \in \operatorname{prox}_{\gamma g}\left(x^n - \gamma \nabla f(x^n)\right), \tag{10.1}$$

vers un point critique de la fonctionnelle $J(x) := f(x) + g(x)$ [1]. Cette convergence est établie pour J vérifiant l'inégalité de Kurdyka-Lojasiewicz (KL) et pour $\gamma \in \left]0, \frac{1}{L}\right[$, où L est la constante de Lipschitz du gradient de f (Attouch et al., 2013, théorème 5.1). Comme nous l'avons vu dans la section 8.3.1, ces résultats permettent notamment de montrer la convergence de l'algorithme IHT pour (P_λ) sans condition sur A contrairement aux travaux initiaux de Blumensath et Davies (2008). Dans ce cas, nous rappelons que l'opérateur proximal de la norme-ℓ_0 n'est autre que le seuillage dur

$$\operatorname{prox}_{\gamma\lambda\|\cdot\|_0}(y) = \left(\operatorname{prox}_{\gamma\lambda|\cdot|_0}(y_i)\right)_{i\in\mathbb{I}_N}, \tag{10.2}$$

avec

$$\operatorname{prox}_{\gamma\lambda|\cdot|_0}(u) = \begin{cases} 0 & \text{si } |u| < \sqrt{2\gamma\lambda}, \\ \{0, u\} & \text{si } |u| = \sqrt{2\gamma\lambda}, \\ u & \text{sinon}. \end{cases} \tag{10.3}$$

Pour G_{CEL0}, en suivant (Attouch et al., 2013), étant donné que $x \mapsto \frac{1}{2}\|Ax - d\|^2$ est une fonction polynomiale et que Φ_{CEL0} est polynomiale par morceaux, G_{CEL0} est semi-algébrique et vérifie donc l'inégalité KL. Ainsi, la séquence $(x^k)_{k\in\mathbb{N}}$ générée par (10.1) avec

$$\operatorname{prox}_{\gamma\Phi_{\text{CEL0}}}(y) = \left(\operatorname{prox}_{\gamma\phi(\|a_i\|,\lambda;\cdot)}(y_i)\right)_{i\in\mathbb{I}_N}, \tag{10.4}$$

et, pour $a \in \mathbb{R}_+^\star$,

$$\operatorname{prox}_{\gamma\phi(a,\lambda;\cdot)}(u) = \begin{cases} \operatorname{sign}(u)\min\left(|u|, \dfrac{(|u| - \sqrt{2\lambda}\gamma a)_+}{1 - a^2\gamma}\right) & \text{si } a^2\gamma < 1, \quad (10.5a) \\ u\mathbb{1}_{\{|u|>\sqrt{2\gamma\lambda}\}} + \{0, u\}\mathbb{1}_{\{|u|=\sqrt{2\gamma\lambda}\}} & \text{si } a^2\gamma \geqslant 1, \quad (10.5b) \end{cases}$$

est assurée de converger vers un point critique de G_{CEL0}. Lorsque $a^2\gamma < 1$, on obtient ainsi un seuillage *continu* (10.5a) qui est présenté sur la figure 37 en bleu pour $a = 0.5$ et $\lambda = \gamma = 1$. Dans le cas contraire, on retrouve le seuillage dur (10.5b). Cependant, pour G_{CEL0}, la constante de Lipschitz du gradient de $x \mapsto \frac{1}{2}\|Ax - d\|^2$ est donnée par $L = \|A\|^2$ et on a

$$\|A\| = \sup_{\|x\|=1} \|Ax\| \geqslant \max_{i\in\mathbb{I}_N} \|a_i\|, \tag{10.6}$$

puisque pour le vecteur e_i ($i \in \mathbb{I}_N$), $\|e_i\| = 1$ et $\|Ae_i\| = \|a_i\|$. La condition $\gamma \in \left]0, \frac{1}{L}\right[$ implique alors que

$$\forall i \in \mathbb{I}_N,\ \|a_i\|^2\gamma < 1, \tag{10.7}$$

et $\operatorname{prox}_{\gamma\Phi_{\text{CEL0}}}$ est réduit à la partie *continue* (10.5a) pour chacune des variables. Notons que $\operatorname{prox}_{\gamma\Phi_{\text{CEL0}}}$ définit un seuillage (comme ceux présentés dans la table 3 pour d'autres pénalités continues non-convexes approchant la norme-ℓ_0). On retrouve ici le seuillage FIRM introduit par Bruce et Gao (1995) comme un compromis entre les seuillages doux et dur. Ce seuillage est défini par deux paramètres $\{\lambda_1, \lambda_2\}$ et on peut identifier à partir de (10.5a) que le seuillage CEL0 correspond à $\lambda_1 = \sqrt{2\lambda}\gamma a$ et $\lambda_2 = \sqrt{2\lambda}/a$ où $a = \|a_i\|$ est

1. f différentiable à gradient Lipschitz et g admettant une forme explicite de l'opérateur proximal. Par ailleurs $J = f + g$ doit être s.c.i. et bornée inférieurement.

ici différent pour chacune des composantes du signal $i \in \mathbb{I}_N$ (si les colonnes de A ne sont pas normalisées).

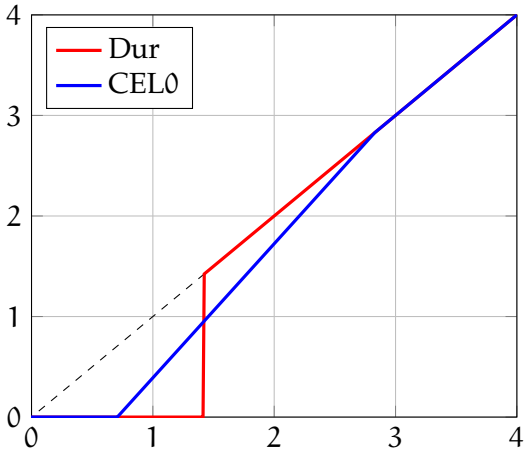


FIGURE 37 – Seuillages dur (rouge) et CEL0 (bleu) pour $a = 0.5$ et $\lambda = \gamma = 1$.

Enfin, là encore, il existe des versions accélérées de FBS dans le cas non-convexe. Par exemple, l'algorithme General Iterative Shrinkage Thresholding (GIST) proposé par GONG et al. (2013) inclue une étape de recherche linéaire afin de déterminer un bon pas de descente à chaque itération. D'autres accélérations ont été proposées par OCHS et al. (2014) ou encore LIANG et al. (2016) en suivant l'idée de la méthode dite *heavy ball* proposée initialement par POLYAK (1964). Le principe de telles méthodes repose sur l'ajout d'un terme combinant les itérés précédents, comme c'est le cas par exemple pour FISTA présenté dans l'algorithme 3 (page 78).

10.1.2 *Majorisation-Minimisation*

Une autre classe d'algorithmes couramment employés en optimisation est la classe des algorithmes MM (rappelons que FBS peut aussi être vu comme un algorithme MM (PARIKH et BOYD, 2014)). Le principe de ces méthodes est de générer et de minimiser une suite de fonctionnelles convexes majorantes et égales à la fonction objectif initiale au point courant (voir figure 38). L'idée sous-jacente est de transformer la résolution d'un problème difficile en la résolution d'une succession de problème plus simples. Le schéma général de ce type d'algorithme, pour la minimisation d'une fonction $F : \mathbb{R}^N \to \bar{\mathbb{R}}$ semi-continue inférieurement (s.c.i.) et propre, est le suivant.

Algorithme 7 : Schéma général des algorithmes MM

Entrées : $x^0 \in \mathbb{R}^N$

répéter

Construire une fonction majorante $M_{x^n} : \mathbb{R}^N \to \mathbb{R}$ telle que :

1. $\forall x \in \mathbb{R}^N$, $F(x) \leqslant M_{x^n}(x)$;
2. $F(x^n) = M_{x^n}(x^n)$;
3. M_{x^n} s.c.i., propre et convexe ;

$x^{n+1} \in \arg\min_{x \in \mathbb{R}^N} M_{x^n}(x)$, ;

jusqu'à *convergence*;

Sorties : x^n

Toute la difficulté se trouve donc dans la construction de fonctions majorantes M_{x^n} qui puissent être minimisées efficacement par des algorithmes d'optimisation convexe.

Une famille importante d'algorithmes MM est basée sur des majorisations quadratiques de la fonction objectif. De telles majorisations ont vu le jour dans le contexte de la *régularisation semi-quadratique* (Geman et Reynolds, 1992 ; Charbonnier et al., 1997) où les auteurs reformulent certaines fonctions de régularisation comme l'infimum d'une famille de fonctions quadratiques. Par ailleurs, dans le contexte de l'approximation parcimonieuse pour la minimisation de critères ℓ_p ($p \leqslant 1$), un algorithme MM très répandu est l'algorithme Iteratively Reweighted Least Squares (IRLS) (Daubechies et al., 2010), aussi connu sous le nom de FOCal Underdetermined System Solver (FOCUSS) (Gorodnitsky et Rao, 1997 ; Rao et al., 2003). La méthode consiste en la résolution d'une séquence de problèmes impliquant une norme-ℓ_2 pondérée (i. e. $\sum_i w_i x_i^2$), où les poids w_i sont calculés à partir de la solution à l'itération précédente.

Cependant, IRLS est adapté à des fonctionnelles suffisamment lisses pouvant être bien approchées par des majorants quadratiques ce qui n'est pas le cas de la pénalité Φ_{CEL0} (ainsi que la plupart des pénalités introduites dans la section 8.4) qui est non-différentiable à l'origine. Une alternative «nonsmooth» pour la minimisation de critères du type

$$J(x) := \frac{1}{2}\|Ax - d\|^2 + \sum_{i \in \mathbb{I}_N} \phi_i(|x_i|), \tag{10.8}$$

où $\phi_i : \mathbb{R}_+ \to \mathbb{R}$ ($i \in \mathbb{I}_N$) est concave et non-décroissante, est donnée par l'algorithme Iteratively Reweighted ℓ_1 (IRL1) minimisant une séquence de problèmes ℓ_1 pondérés (Ochs et al., 2015) qui sont des majorants de la fonction objectif J vérifiant les conditions détaillées dans l'algorithme 7.

Algorithme 8 : Iteratively Reweighted ℓ_1 (IRL1)

Entrées : $x^0 \in \mathbb{R}^N$

1 **répéter**

2 $\quad w_i^{x^n} \in \partial^+ \phi_i(|x_i^n|)\ \forall i \in \mathbb{I}_N$;

3 $\quad x^{n+1} \in \arg\min_{x \in \mathbb{R}^N} \frac{1}{2}\|Ax - d\|^2 + \sum_{i \in \mathbb{I}_N} w_i^{x^n} |x_i|$;

4 **jusqu'à** *convergence*;

Sorties : x^n

Dans l'algorithme 8, $\partial^+ \phi_i(|x^n|) = -\partial^-[-\phi_i(|x^n|)]$ où ∂^- définit le «limiting-subgradient» (Rockafellar et Wets, 2009, définition 8.3 page 301). De part la concavité de ϕ_i sur $\mathbb{R}_+$, $u \mapsto w_i^{x^n}|u|$ est un majorant de ϕ_i (à une constante près). En effet, si on considère ϕ_i différentiable sur $[0^+, +\infty)$, ϕ_i est majorée par sa tangente en $|x_i^n|$ donnée par

$$u \in \mathbb{R}_+ \longmapsto w_i^{x^n} u - w_i^{x^n}|x_i^n| + \phi_i(|x_i^n|) = w_i^{x^n} u + C_{x_i^n}, \tag{10.9}$$

où $C_{x_i^n} \in \mathbb{R}$. La justification pour le cas non-différentiable est similaire en considérant les demi-tangentes. Des illustrations sont présentées sur la figure 38.

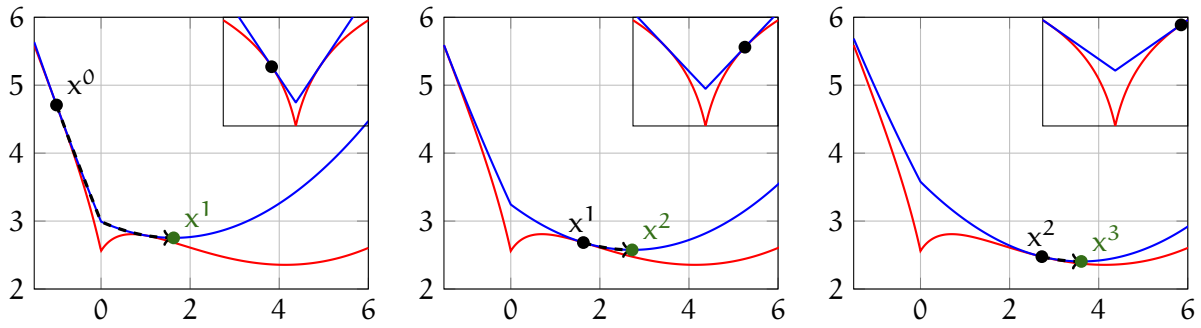


FIGURE 38 – Trois itérations de l'algorithme IRL1 pour minimiser la fonctionnelle $x \mapsto \frac{1}{2}(0.3x-1.6)^2 + \log(1+2|x|)$ (rouge). La courbe bleue représente la fonction majorante $x \mapsto \frac{1}{2}(0.3x-1.6)^2 + w^{x^n}|x| + C_{x^n}$ où $w^{x^n} = 2/(1+2|x^n|)$ et $C_{x^n} \in \mathbb{R}$ est une constante (car la fonction majorante de l'algorithme 8 est définie à une constante près). Le sous graphe en haut à droite de chaque figure montre la majorisation de la pénalité $\log(1+2|x|)$ correspondante. Notons que nous n'avons pas utilisé la pénalité CEL0 pour cet exemple puisque, dans le cas unidimensionnel, la fonctionnelle CEL0 résultante est convexe.

Dans le cadre de la minimisation de G_{CEL0} avec l'algorithme IRL1, les poids utilisés dans l'algorithme 8 sont donnés par : $\forall i \in \mathbb{I}_N$,

$$w_i^{x^n} = \begin{cases} \sqrt{2\lambda}\|a_i\| - \|a_i\|^2|x_i^n| & \text{si } 0 \leqslant |x_i^n| < \sqrt{2\lambda}/\|a_i\|, \\ 0 & \text{si } |x_i^n| \geqslant \sqrt{2\lambda}/\|a_i\|. \end{cases} \tag{10.10}$$

Ensuite le problème convexe de la ligne 3 de l'algorithme 8 peut être traité avec FISTA par exemple. La convergence de l'algorithme 8 est démontrée par OCHS et al. (2015) sous la condition que la fonction objectif J dans (10.8) vérifie la propriété KL (ce qui est le cas lorsque $\phi_i(u) = \phi(\|a_i\|, \lambda; u)$, $u \in \mathbb{R}_+$). Notons que OCHS et al. (2015) s'intéressent à une classe plus large d'algorithmes de majorisation convexe applicables à des fonctionnelles plus générales que (10.8). Cependant, afin de simplifier la lecture au cas qui nous intéresse, IRL1 est présenté dans l'algorithme 8 pour le cas particulier des problèmes du type (10.8). Une illustration des itérés de l'algorithme est proposée sur la figure 38.

10.1.3 *Programmation DC*

Décomposer une fonction objectif $J : \mathbb{R}^N \to \mathbb{R}$ (non-convexe) comme une Différence de fonctions Convexes (DC) :

$$J(x) = J_1(x) - J_2(x), \tag{10.11}$$

où $J_1 : \mathbb{R}^N \to \mathbb{R}$ et $J_2 : \mathbb{R}^N \to \mathbb{R}$ sont s.c.i., propres et convexes, est également une méthode permettant de transformer un problème d'optimisation difficile en une séquence de problèmes plus simples. Cette méthode est connue sous le nom de programmation DC et est couramment employée en optimisation non-convexe (HORST et THOAI, 1999 ; TAO et al., 2005). L'idée des algorithmes DC est alors de générer deux séquences $(x^n)_{n\in\mathbb{N}}$ et $(y^n)_{n\in\mathbb{N}}$ convergeant respectivement vers un point critique $\hat{x}$ du problème primal

$$\min_{x\in\mathbb{R}^N} J_1(x) - J_2(x), \tag{10.12}$$

et un point critique $\hat{y}$ du problème dual (Tao et al., 2005, section 2.2)

$$\min_{y\in\mathbb{R}^N} J_2^\star(y) - J_1^\star(y), \tag{10.13}$$

où $J_1^\star$ (resp. $J_2^\star$) définit la fonction conjuguée [2] de J_1 (resp. J_2). Ces séquences sont générées selon le schéma présenté dans l'algorithme 9. À chaque itération de l'algorithme, la mise à jour de y^n est effectuée par minimisation d'une approximation (convexe) de (10.13) définie par linéarisation du terme $J_1^\star$ au point y^{n-1} (sachant $x^n \in \partial J_1^\star(y^{n-1})$) :

$$y^n \in \arg\min_{y\in\mathbb{R}^N} J_2^\star(y) - J_1^\star(y^{n-1}) - \langle y - y^{n-1}, x^n\rangle, \tag{10.14}$$

$$= \arg\min_{y\in\mathbb{R}^N} J_2^\star(y) - \langle y, x^n\rangle, \tag{10.15}$$

ce qui est équivalent à $y^n \in \partial J_2(x^n)$ (voir Rockafellar et Wets, 2009, Proposition 11.3 page 476). De manière similaire, x^{n+1} est calculé par minimisation de l'approximation de (10.12) suivante (sachant $y^n \in \partial J_2(x^n)$) :

$$x^{n+1} \in \arg\min_{x\in\mathbb{R}^N} J_1(x) - J_2(x^n) - \langle x - x^n, y^n\rangle, \tag{10.16}$$

$$= \arg\min_{x\in\mathbb{R}^N} J_1(x) - \langle x, y^n\rangle, \tag{10.17}$$

qui n'est autre que $x^{n+1} \in \partial J_1^\star(y^n)$.

Algorithme 9 : Algorithme DC

Entrées : $x^0 \in \mathbb{R}^N$

répéter

  $y^n \in \partial J_2(x^n)$;

  $x^{n+1} \in \partial J_1^\star(y^n)$;

jusqu'à *convergence*;

Sorties : x^n, y^n

En suivant l'idée proposée par Gasso et al. (2009) nous pouvons décomposer G_{CEL0} comme il suit :

$$G_{\text{CEL0}}(x) = \underbrace{\frac{1}{2}\|A(x^+ - x^-) - d\|^2 + \sum_{i\in\mathbb{I}_N} \sqrt{2\lambda}\|a_i\|(x_i^+ + x_i^-)}_{J_1(x^+,x^-)} - \underbrace{\sum_{i\in\mathbb{I}_N} h_i(x_i^+ + x_i^-)}_{J_2(x^+,x^-)}, \tag{10.18}$$

avec $x = x^+ - x^-$ pour $(x^+, x^-) \in \mathbb{R}_+^{2N}$ et où les $h_i : \mathbb{R}_+ \to \mathbb{R}$ ($i \in \mathbb{I}_N$) sont définies par

$$\forall u \in \mathbb{R}_+,\ h_i(u) := \begin{cases} \dfrac{\|a_i\|}{2}u^2 & \text{si } u \leqslant \dfrac{\sqrt{2\lambda}}{\|a_i\|}, \\ \sqrt{2\lambda}\|a_i\|u - \lambda & \text{sinon.} \end{cases} \tag{10.19}$$

Notons que (10.18) provient uniquement de la décomposition de $\phi_i(a, \lambda; \cdot)$ ($a \in \mathbb{R}_+$, $\lambda \in \mathbb{R}_+$) comme la différence des fonctions $\sqrt{2\lambda}a|\cdot|$ et $h_i(|\cdot|)$ dont un exemple est présenté sur la figure 39.

2. Voir définition 9.2.

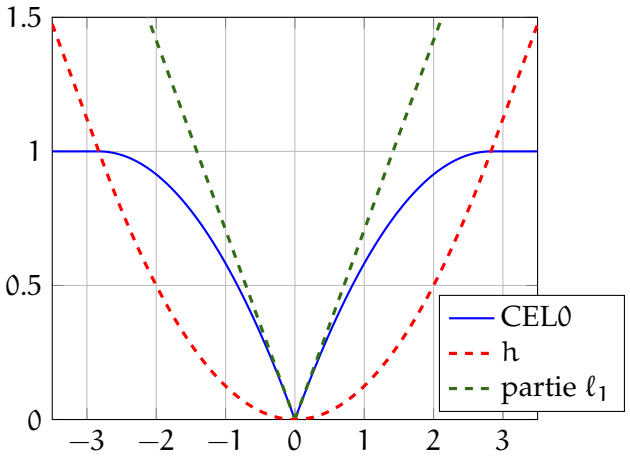


FIGURE 39 – Décomposition de la pénalité CEL0 donnée par l'équation (10.18) pour $a = 0.5$ et $\lambda = 1$.

Avec une telle décomposition (10.18), l'étape 2 de l'algorithme 9 est simplement donnée par $\forall i \in \mathbb{I}_N$,

$$y_i^n = h_i'((x_i^+)^n + (x_i^-)^n) \quad \text{avec} \quad \forall u \in \mathbb{R}_+,\ h_i'(u) = \begin{cases} \|a_i\|^2 u & \text{si } u \leqslant \dfrac{\sqrt{2\lambda}}{\|a_i\|}, \\ \sqrt{2\lambda}\|a_i\| & \text{sinon.} \end{cases} \tag{10.20}$$

Ensuite l'étape 3 de l'algorithme revient, d'après (10.17), à résoudre le problème

$$x^{n+1} \in \arg \min_{(x^+,x^-)\in\mathbb{R}_+^{2N}} J_1(x^+, x^-) - \langle x^+ + x^-, y^n \rangle, \tag{10.21}$$

qui peut se réécrire comme le problème ℓ_1 pondéré suivant :

$$x^{n+1} \in \arg \min_{x\in\mathbb{R}^N} \frac{1}{2}\|Ax - d\|^2 + \sum_{i\in\mathbb{I}_N} w_i |x_i|, \tag{10.22}$$

avec $\forall i \in \mathbb{I}_N$, w_i donné par (10.10). On retrouve donc ici exactement le schéma de l'algorithme 8 avec les poids (10.10). En effet, la programmation DC peut être également interprétée comme une méthode MM. Pour la convergence de l'algorithme 9 vers un point critique de la fonctionnelle DC nous renvoyons à (GASSO et al., 2009, théorème 1 et références associées). Enfin, notons qu'avec une décomposition différente de celle proposée en (10.18), nous aurions eu un schéma différent de celui présenté dans la section précédente.

10.1.4 *Minimisation Coordinate-Wise*

Nous évoquons ici très brièvement les algorithmes dits «Coordinate-Wise» dont le principe est, à chaque itération, de minimiser la fonction objectif $J : \mathbb{R}^N \to \mathbb{R}$ par rapport à une variable en fixant les autres comme cela est présenté par l'algorithme 10 [3].

L'étape de sélection à la ligne 2 de l'algorithme 10 peut par exemple être cyclique (on incrémente i à chaque itération et on revient à 1 lorsque l'on atteint N) ou encore correspondre à la composante générant la décroissance la plus importante du critère.

De telles méthodes ont été développées dans le contexte de l'optimisation parcimonieuse. Par exemple, nous trouvons l'algorithme GSS (BECK et ELDAR, 2013), pour le problème (C_k),

3. Nous rappelons que $u \mapsto J^i(u; (x^k)^{(i)})$ définit la restriction de J à la i-ème variable au point x^k

Algorithme 10 : Minimisation Coordinate-Wise

Entrées : $x^0 \in \mathbb{R}^N$

répéter

  choisir $i \in \mathbb{I}_N$;

  $x_i^{k+1} \in \arg\min_{u \in \mathbb{R}} J^i(u; (x^k)^{(i)})$;

jusqu'à *convergence*;

Sorties : x^k

que nous avions déjà évoqué dans la chapitre 2. Aussi, des auteurs ont proposé des algorithmes de descente par coordonnées pour la minimisation de critères moindres carrés régularisés par des pénalités non-convexes comme MCP (Mazumder et al., 2012 ; Breheny et Huang, 2011).

Cependant, nous nous intéresserons pas à ce genre de méthodes pour la minimisation de G_{CEL0}. Une première raison est que la restriction de G_{CEL0} à une variable étant l'enveloppe convexe (à une constante près) de G_{ℓ_0} restreinte à la même variable, appliquer l'algorithme 10 sur G_{CEL0} ou directement sur G_{ℓ_0} revient au même. Une autre raison est que nous avons pu constater sur des exemples numériques que cet algorithme était souvent moins performant que les méthodes présentées précédemment.

10.1.5 *Graduated Non Convexity*

Nous présentons pour finir une dernière classe de méthodes pouvant être utilisées pour la minimisation de G_{CEL0}. Il s'agit de la classe des méthodes dites GNC (également connues sous le nom de méthodes de continuation) initialement introduites par Blake et Zisserman (1987). Encore une fois, l'idée consiste en la résolution d'une séquence de problèmes permettant d'approcher la solution d'un problème plus complexe. Contrairement aux méthodes MM présentées précédemment qui sont basées sur la résolution d'un problème convexe à chaque itération, le GNC commence par la minimisation d'une fonction convexe, puis déforme cette dernière en introduisant progressivement la non-convexité de la fonction objectif finale. La non-convexité des problèmes ainsi générés par la méthode est donc de plus en plus importante au fil des itérations.

Plus formellement, considérons un critère objectif $J : \mathbb{R}^N \to \mathbb{R}$ non-convexe ainsi qu'une approximation paramétrée J_δ ($\delta \in \mathbb{R}_+$) telle que :

$$\exists \delta_0 \in \mathbb{R}_+ \text{ tel que } J_{\delta_0} \text{ est convexe,} \tag{10.23}$$

$$\lim_{\substack{\delta \to 0 \\ > }} J_\delta = J. \tag{10.24}$$

Le principe de la méthode GNC est alors détaillé dans l'algorithme 11 et illustré sur la figure 40.

Afin d'assurer la convergence d'une telle approche vers un minimiseur global du critère J, deux ingrédients doivent être vérifiés :

1. atteindre un minimiseur global de la fonctionnelle initiale convexe J_{δ_0} ;
2. pour tout $k \in \{1, \dots, K\}$ et tout minimiseur global $\hat{x}$ de J_{δ_k}
   — il existe une région $\mathcal{V} \ni \hat{x}$ sur laquelle J_{δ_k} est localement convexe ;
   — $\mathcal{V} \cap \arg\min_{x \in \mathbb{R}^N} J_{\delta_{k-1}} \neq \emptyset$.

Algorithme 11 : Méthode Graduated Non Convexity (GNC)

Entrées : $x^0 \in \mathbb{R}^N$, $(\delta_0, \ldots, \delta_K)$ une séquence décroissante

1 **pour** $k \in \{0, \ldots, K\}$ **faire**

2 $\quad x^{k+1} \longleftarrow$ Minimiser(J_{δ_k}, x^k) où Minimiser représente n'importe quel algorithme (sous-optimal) minimisant J_{δ_k} initialisé par x^k;

Sorties : x^K

Bien que le premier point ci-dessus soit facilement réalisable, le deuxième est bien plus compliqué et il n'est généralement possible d'assurer l'optimalité de la méthode que pour des problèmes particuliers (Blake et Zisserman, 1987). Pour une analyse théorique de la méthode nous renvoyons le lecteur vers les récents travaux de Mobahi et Fisher III (2015a,b).

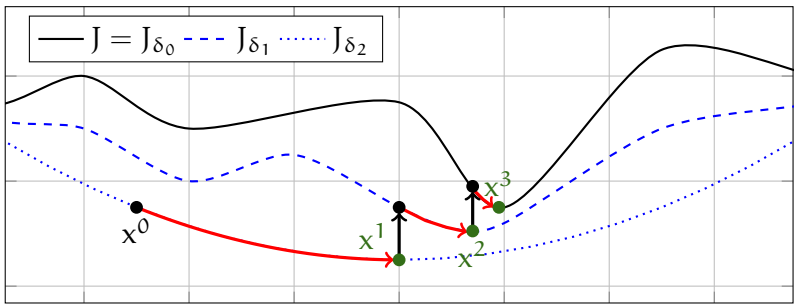


Figure 40 – Illustration de la méthode GNC pour minimiser la fonctionnelle J (noir, trait plein). Trois itérations sont effectuées.

Dans le contexte de l'optimisation parcimonieuse, plusieurs auteurs ont employé une telle approche. Par exemple Leclerc (1989) proposa une méthode du type GNC pour minimiser un critère pénalisé en norme-ℓ_0 dans le cadre de la segmentation d'images. D'autre part, une généralisation de ces méthodes GNC a été proposé ultérieurement par Nikolova (1999) pour des applications en reconstruction d'images. Ce travail considère un ensemble de pénalités (fonctions de potentiel), incluant la norme-ℓ_0, et propose une étude sur la construction et l'initialisation de la séquence de fonctionnelles approchantes. Notons aussi les progrès apportés par Robini et al. (2007) (voir aussi (Robini et Magnin, 2010)) avec la méthode dite Stochastic Continuation (SC) combinant recuit simulé[4] et GNC en tirant bénéfice des avantages des deux approches comme les garanties de convergence globale du recuit simulé et l'efficacité (en terme de coût de calcul) du GNC.

Enfin, Mohimani et al. (2009) ont utilisé une approche GNC pour le problème

$$\hat{x} \in \arg\min_{x \in \mathbb{R}^N} \|x\|_0 \quad \text{s.c. } Ax = b, \tag{10.25}$$

Pour ce faire, ils introduisent une famille de fonctions continues et différentiable approchant $|\cdot|_0$ et dont un exemple est donné par

$$\phi_\delta(u) := 1 - \exp\left(-\frac{u^2}{2\delta^2}\right). \tag{10.26}$$

Les auteurs mettent en place un schéma similaire à celui présenté par l'algorithme 11 (voir Mohimani et al., 2009, figure 1) qu'ils nomment Smoothed-ℓ_0 (SL0). Sous certaines

4. Voir par exemple les travaux de Geman et Geman (1984).

conditions définies en terme de constante RIP asymétriques[5] et de $\|A\|$, la convergence de l'algorithme vers la solution la plus parcimonieuse du système linéaire $Ax = b$ est partiellement montrée dans (Mohimani et al., 2009) puis complétée dans (Mohimani et al., 2010).

10.2 UN MACRO-ALGO ASSURANT LA CONVERGENCE VERS UN MINIMISEUR LOCAL DE G_{ℓ_0}

Tous les algorithmes présentés précédemment nous permettent de minimiser G_{CEL0} mais ne nous assurent généralement qu'une convergence vers des points critiques de la fonctionnelle. Au contraire, les théorèmes 9.16 et 9.21 établissent les relations entre les minimiseurs des fonctionnelles G_{CEL0} et G_{ℓ_0}. Bien que nous sachions partiellement (resp. complètement) caractériser les points critiques de G_{CEL0} qui sont des minimiseurs (resp. des minimiseurs stricts) grâce au théorème 9.27 (resp. théorème 9.26), comment peut-on assurer la convergence de l'algorithme vers de tels points ? En particulier, une condition que l'on retrouve dans les théorèmes 9.27 et 9.26 est que le point critique en question, disons $\hat{x}$, vérifie $\sigma^+(\hat{x}) = \emptyset$. Avant de voir comment nous pouvons répondre à la question que nous venons de soulever, nous introduisons le résultat suivant.

Lemme 10.1 (Lien entre les points critiques de G_{CEL0} et les minimiseurs de G_{ℓ_0})**.** *Soit $\hat{x} \in \mathbb{R}^N$ un point critique de G_{CEL0} tel que $\sigma^-(\hat{x}) = \emptyset$. Alors, $\hat{x}$ est un minimiseur (local) de G_{ℓ_0} et*

$$G_{\text{CEL0}}(\hat{x}) = G_{\ell_0}(\hat{x}). \tag{10.27}$$

Démonstration. D'après le lemme 9.10, nous savons que $\hat{x}$, qui est un point critique de G_{CEL0} tel que $\sigma^-(\hat{x}) = \emptyset$ (i.e. $\forall i \in \sigma(\hat{x}),\ |\hat{x}_i| \geqslant \frac{\sqrt{2\lambda}}{\|a_i\|}$), vérifie

$$\forall i \in \sigma(\hat{x}),\ \hat{x}_i = -\frac{\langle a_i, A\hat{x}^{(i)} - d\rangle}{\|a_i\|^2} \iff \langle a_i, A\hat{x} - d\rangle = 0. \tag{10.28}$$

Notons $\hat{\sigma} = \sigma(\hat{x})$, alors nous avons

$$(A_{\hat{\sigma}})^T(A\hat{x} - d) = 0 \iff (A_{\hat{\sigma}})^T(A_{\hat{\sigma}}\hat{x}_{\hat{\sigma}} + A_{\hat{\sigma}^c}\underbrace{\hat{x}_{\hat{\sigma}^c}}_{=0} - d) = 0, \tag{10.29}$$

$$\iff (A_{\hat{\sigma}})^T A_{\hat{\sigma}}\hat{x}_{\hat{\sigma}} = (A_{\hat{\sigma}})^T d. \tag{10.30}$$

Il s'en suit, d'après le corolaire 9.12, que $\hat{x}$ est un minimiseur (local) de G_{ℓ_0}. Enfin, le fait que $G_{\text{CEL0}}(\hat{x}) = G_{\ell_0}(\hat{x})$ provient des mêmes arguments que ceux utilisés dans la preuve du théorème 9.16. □

Ce lemme nous donne donc une condition nécessaire pour qu'un point critique de G_{CEL0} soit un minimiseur de G_{ℓ_0}. Cette condition est plus faible que $\sigma^+(\hat{x}) = \emptyset$ mais ne nous assure pas qu'un tel point soit aussi un minimiseur local de G_{CEL0} qui, d'après le théorème 9.16 est une condition nécessaire d'optimalité globale pour les deux fonctionnelles G_{CEL0} et G_{ℓ_0}. Cependant, ce résultat va nous servir à garantir une convergence vers un point qui est à la fois point critique de G_{CEL0} et minimiseur local de G_{ℓ_0} mais n'est pas

5. On appelle constante RIP asymétrique d'ordre k, les plus petites constantes positives δ_k^{min} et δ_k^{max} telles que $(1-\delta_k^{min})\|x\|^2 \leqslant \|Ax\|^2 \leqslant (1+\delta_k^{max})\|x\|^2$ pour tout $x \in \mathbb{R}^N$ vérifiant $\|x\|_0 \leqslant k$.

forcément un minimiseur pour G_{CEL0} (voir remarque 9.22).

10.2.1 *Hypothèses de travail et description de l'algorithme*

Dans la suite, nous considérons un algorithme, noté $\mathsf{Alg}(x^{init}, \lambda)$, initialisé par $x^{init} \in \mathbb{R}^N$ et produisant une séquence d'itérés $(x^n)_{n\in\mathbb{N}}$ vérifiant les deux hypothèses suivantes :

H1. convergence vers un point critique de G_{CEL0},

H2. décroissance suffisante de la fonction objectif

$$\forall n \in \mathbb{N},\ G_{CEL0}(x^{n+1}) \leqslant G_{CEL0}(x^n) - \beta\|x^{n+1} - x^n\|^2\,, \tag{10.31}$$

où $\beta \in \mathbb{R}_+^\star$.

L'hypothèse H2 impose à l'algorithme Alg de décroître suffisamment le critère à chaque itération. Cette condition est généralement à la base des preuves de convergence des algorithmes de descente et est donc vérifiée pour de nombreux algorithmes comme ceux présentés précédemment. À partir de maintenant, nous utiliserons la notation $\sigma_n^- = \sigma^-(x^n)$.

Soit $x^n \in \mathbb{R}^N$ un point critique de G_{CEL0} déterminé par Alg vérifiant les hypothèses H1 et H2. Supposons que ce point vérifie $\sharp\sigma_n^- \geqslant 1$. Alors, d'après le lemme 9.10 et la proposition 9.6, nous pouvons voir que, pour $i \in \sigma_n^-$, la restriction $G_{CEL0}^i(\cdot; (x^n)^{(i)})$ est constante sur l'intervalle [6]

$$\left[0, \frac{\sqrt{2\lambda}}{\|a_i\|}\right] \quad \text{si} \quad \operatorname{sign}(x_i^n) > 0, \tag{10.32}$$

$$\left[-\frac{\sqrt{2\lambda}}{\|a_i\|}, 0\right] \quad \text{si} \quad \operatorname{sign}(x_i^n) < 0. \tag{10.33}$$

Par conséquent, pour $i \in \sigma_n^-$, l'égalité suivante est vérifiée,

$$G_{CEL0}((x^n)^{(i)}) = G_{CEL0}(x^n), \tag{10.34}$$

où nous rappelons que $(x^n)^{(i)} = \{x_1^n, \ldots, x_{i-1}^n, 0, x_{i+1}^n, \ldots, x_N^n\}$. Posons $x^{temp} = (x^n)^{(i)}$, deux configurations se présentent :

1. soit x^{temp} *est un point critique* de G_{CEL0} et on peut alors poser $x^{n+1} = x^{temp}$ qui vérifiera $\sharp\sigma_{n+1}^- = \sharp\sigma_n^- - 1$;
2. soit x^{temp} *n'est pas un point critique* de G_{CEL0} et on peut alors faire un nouvel appel à Alg pour définir un nouveau point :

   $$x^{n+1} \leftarrow \mathsf{Alg}(x^{temp}, \lambda).$$

   Ainsi, x^{n+1} sera un nouveau point critique de G_{CEL0} (par H1) tel que $G_{CEL0}(x^{n+1}) < G_{CEL0}(x^n)$ (par H2).

Dans les deux cas, x^{n+1} est un point critique de G_{CEL0} pour lequel, si $\sharp\sigma^{n+1} \geqslant 1$, les deux configurations présentées ci dessus se posent à nouveau et le processus précédent peut être répété avec un nouvel indice $i' \in \sigma_{n+1}^-$. Dans le cas où $\sigma^{n+1} = \emptyset$, x^{n+1} vérifie les hypothèses du lemme 10.1 et est donc un minimiseur local de G_{ℓ_0}.

6. Voir preuve du lemme 9.13 en annexe A.1.4 page 191.

Le raisonnement précédent est la base du *Macro-Algo* que nous proposons ci-après (algorithme 12).

Algorithme 12 : Macro-Algo pour la détermination d'un point critique de G_{CEL0} qui soit aussi un minimiseur (local) de G_{ℓ_0}.

Entrées : Alg vérifiant H1 et H2, $x^{init} \in \mathbb{R}^N$, $\lambda > 0$

$x^1 \leftarrow \text{Alg}(x^{init}, \lambda)$;
Calculer σ_1^-;
tant que $\sigma_n^- \neq \emptyset$ **faire**
  $x^{temp} = x^n$;
  tant que x^{temp} *est un point critique de* G_{CEL0} *et* $\sigma_n^- \neq \emptyset$ **faire**
    Sélectionner $i \in \sigma_n^-$;
    $x_i^{temp} = 0$;
    $\sigma_n^- = \sigma_n^- \backslash \{i\}$;
  $x^{n+1} \leftarrow \text{Alg}(x^{temp}, \lambda)$;
  Calculer σ_n^-;

Sorties : x^n

Pour résumer, à partir d'un algorithme Alg vérifiant les hypothèses H1 et H2 (par exemple un de ceux présentés dans la section 10.1), le Macro-Algo itère d'un point critique de G_{CEL0} à un autre, tout en assurant la décroissance de la fonction objectif, jusqu'à atteindre un point critique $\hat{x}$ vérifiant $\sigma^-(\hat{x}) = \emptyset$. D'après le lemme 10.1, un tel point critique de G_{CEL0} est aussi un minimiseur (local) de G_{ℓ_0}.

Remarque 10.2. Le fait que la restriction $G^i_{\text{CEL0}}(\cdot; x_i^n)$ soit constante sur l'intervalle défini par (10.32) et (10.33) est également vrai pour $i \in \sigma^+(x^n)$. Ainsi, nous pouvons définir $\sigma_n^- := \sigma^-(x^n) \cup \left\{ i \in \mathbb{I}_N : |x_i^n| = \frac{\sqrt{2\lambda}}{\|a_i\|} \right\}$. Cela permet de mettre plus de composantes de x^n à zéro.

10.2.2 *Résultat de convergence*

Pour avoir convergence du Macro-Algo, il faut d'abord s'assurer que l'ensemble des points critiques de G_{CEL0} vérifiant $\sigma^-(\hat{x}) = \emptyset$ est non-vide. Cela est vérifié car pour tout minimiseur global $\hat{x} \in \mathbb{R}^N$ de G_{CEL0} (il en existe au moins un d'après la proposition 9.18), le théorème 9.16 permet de définir un autre minimiseur global $\hat{x}^0$ (défini par (9.26)) tel que $\sigma^-(\hat{x}^0) = \emptyset$. Ensuite le théorème suivant nous permet de conclure sur la convergence de l'algorithme en un nombre fini d'étapes.

> **Theorème 10.3** (Convergence du Macro-Algo)**.** *Soit* $(x^n)_{n\in\mathbb{N}}$ *la séquence générée par le Macro-Algo (algorithme 12). Alors, il existe* $n^\star \in \mathbb{N}$ *tel que* $\sigma_{n^\star}^- = \emptyset$ *et* $x^{n^\star}$ *est un minimiseur (local) de* G_{ℓ_0}.

Démonstration. La preuve est détaillée en annexe A.2.1 page 198. □

Remarque 10.4. La preuve du théorème 10.3 repose sur le fait que l'ensemble image des points critiques G_{CEL0} (au sens de Clarke) par G_{CEL0} est fini. Le nombre d'éléments de cet ensemble peut être très important et ainsi $n^\star \in \mathbb{N}$ très grand. Cependant, pour les illustrations numériques qui seront présentées dans la section 10.5, le Macro-Algo s'arrête généralement après une ou deux itérations.

En effet, chaque nouvelle itération de la boucle externe de l'algorithme 12 réalisée signifie que Alg avait atteint, à l'itération précédente, un point critique que l'on qualifiera de «instable» (i. e. point critique qui n'est pas un minimiseur). Bien que Alg puisse converger vers de tels points, cela est moins fréquent que de s'enfermer dans un minimiseur (local) de la fonctionnelle de part l'instabilité de ces points.

Remarque 10.5. L'étape de sélection du Macro-Algo à la ligne 6 de l'algorithme 12 est arbitraire et la question de savoir s'il existe une sélection «optimale» (dans un sens qu'il faut également définir) reste à étudier.

10.2.3 *Illustrations numériques*

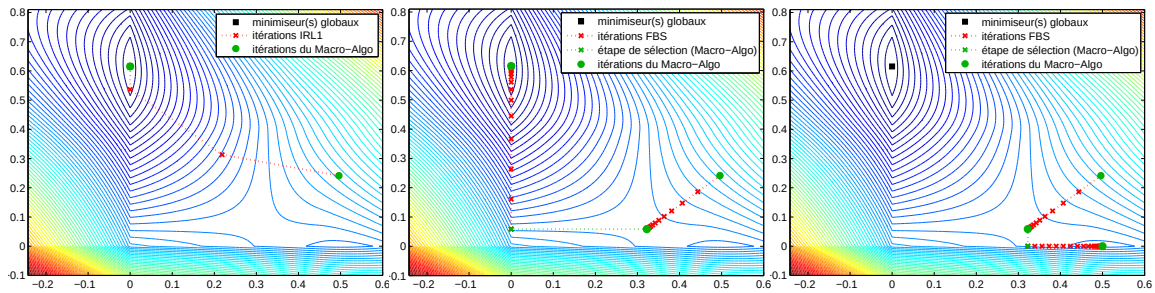
Nous présentons ici un petit exemple en dimension 2 montrant le comportement du Macro-Algo en pratique. La figure 41 représente l'évolution des itérés du Macro-Algo (points verts) ainsi que ceux de l'algorithme interne Alg (croix rouges). D'autre part, les croix vertes correspondent à la variable temporaire x^{temp} utilisée dans l'algorithme 12.

Étant donné que le Macro-Algo minimise $G_{\texttt{CEL0}}$, nous représentons à la fois les itérés et les lignes de niveau de la fonctionnelle sur le même graphe. En ce qui concerne le choix de Alg, deux options ont été considérées :

— IRL1 (voir section 10.1.2) correspondant à la figure 41a ;
— FBS (voir section 10.1.1) correspondant aux figures 41b et 41c, la différence résidant dans l'étape de sélection (ligne 6) de l'algorithme 12.

Lorsque l'algorithme IRL1 est utilisé, le Macro-Algo (croix vertes) converge en une itération vers le minimiseur global de $G_{\texttt{CEL0}}$. Au contraire, avec le même point initial, l'algorithme FBS converge vers un point critique «instable» tel que $\sigma^-(\hat{x}) = \{1, 2\}$ (i. e. les deux composantes de $\hat{x}$ appartiennent à $\sigma^-(\hat{x})$) et le Macro-Algo doit alors réaliser une deuxième itération. Dans cette situation, l'algorithme doit mettre une des deux composantes à zéro (line 6 de l'algorithme 12).

On voit clairement sur les figures 41b et 41c que ce choix est déterminant pour la convergence de l'algorithme vers le minimiseur global de $G_{\texttt{CEL0}}$. Cette observation reflète la remarque 10.5 sur l'éventuelle possibilité de définir une règle de sélection «optimale» permettant d'assurer la convergence vers le «meilleur» (en terme de minimisation de la fonctionnelle) minimiseur (local) que l'algorithme peut atteindre pour une configuration initiale donnée.



(a) Macro-Algo avec IRL1 (b) Macro-Algo avec FBS (c) Macro-Algo avec FBS

FIGURE 41 – Évolution des itérés du Macro-Algo (algorithme 12) sur l'exemple défini par la figure 35h (i. e. $A = [3, 2; 1, 3]$, $b = [1, 2]$ et $\lambda = 1$). La différence entre (b) et (c) réside dans le choix de la composante à la ligne 6 de l'algorithme 12.

Il est à noter que sur les exemples de la figure 41, le point initial x^{init} a été défini afin que le Macro-Algo réalise deux itérations dans sa boucle externe lorsque qu'il est combiné avec l'algorithme FBS. Ce point initial a été déterminé avec une précision de l'ordre de grandeur de celle du critère d'arrêt (i.e. $\|x^n - x^{n-1}\| \leqslant 10^{-5}$). Si on diminue la tolérance de ce critère d'arrêt sans recalculer plus précisément le point initial pour avoir convergence, lors du premier appel à FBS, vers le point critique instable, ce dernier va être «évité» et FBS convergera directement vers un minimiseur (local) de G_{CEL0}. Ces observations illustrent la remarque 10.4 étant donné qu'il est nécessaire de définir un point initial très spécifique pour atteindre un des points critiques que nous qualifions d'instables. Ainsi, nous pouvons prétendre qu'en général, l'algorithme 12 va converger en une ou très peu d'itérations (pour la boucle externe).

10.3 UNE MÉTHODE INSPIRÉE GNC POUR LA MINIMISATION DE G_{CEL0}

Dans cette section, nous nous intéressons à une approche du type GNC pour la minimisation de G_{CEL0}. L'algorithme proposé dans la suite est plus une preuve de concept de l'intérêt que peut apporter une telle approche pour la minimisation de G_{CEL0}, qu'une méthode bien établie et étudiée théoriquement. C'est donc un travail préliminaire qui demande à être affiné par la suite.

10.3.1 *Paramétrisation de la fonctionnelle*

Tout d'abord, nous devons définir une fonction G_δ vérifiant les conditions (10.23) et (10.24). Nous avons fait le choix de paramétrer la pénalité Φ_{CEL0} comme il suit :

$$\Phi_\delta(x) := \sum_{i \in \mathbb{I}_N} \phi_{\text{CEL0}}(\|a_i\|, \lambda; \delta x_i), \tag{10.35}$$

$$= N\lambda - \sum_{i \in \mathbb{I}_N} \frac{\|a_i\|^2}{2} \left(\delta |x_i| - \frac{\sqrt{2\lambda}}{\|a_i\|} \right)^2 \mathbb{1}_{\left\{ \delta |x_i| \leqslant \frac{\sqrt{2\lambda}}{\|a_i\|} \right\}}, \tag{10.36}$$

pour $\delta \in [0, 1]$. Notons qu'ici $G_\delta \to G_{\text{CEL0}}$ lorsque $\delta \to 1$. La figure 42 illustre cette pénalité pour différentes valeurs de δ. La fonctionnelle G_δ est donc définie par :

$$G_\delta(x) := \frac{1}{2}\|Ax - d\|^2 + \Phi_\delta(x). \tag{10.37}$$

La proposition suivante montre que G_δ vérifie les conditions (10.23) et (10.24).

> **Proposition 10.6.** *Lorsque* $\delta = 0$, $G_\delta(x) = \frac{1}{2}\|Ax - d\|^2$ *est convexe et lorsque* $\delta = 1$, $G_\delta = G_{\text{CEL0}}$.

Démonstration. La preuve est directe d'après la définition de Φ_δ en (10.35). □

10.3.2 *Une heuristique pour la variation du paramètre*

À partir de la paramétrisation G_δ de la fonctionnelle CEL0, l'objectif est maintenant de définir une séquence croissante $(\delta_k)_{k \in \mathbb{I}_K}$ telle que $\delta_K = 1$ puis de minimiser successivement les fonctionnelles G_{δ_k} (en utilisant n'importe quel algorithme assurant la convergence vers

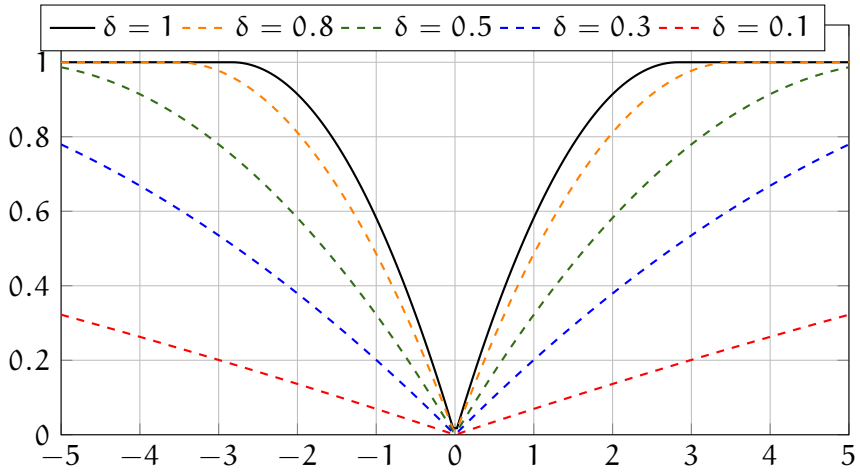


FIGURE 42 – Graphe 1D de la pénalité Φ_δ, définie en (10.35), pour différentes valeurs de $\delta \in [0,1]$ et $a = \lambda = 1$.

un point critique de cette dernière) tout en tirant bénéfice d'une initialisation «à chaud», donnée par la solution $\hat{x}^{k-1}$ obtenue pour le paramètre δ_{k-1} précédent (voir figure 40, page 123, pour un exemple illustratif).

Considérons une valeur initiale $\delta_0 \in [0,1]$ ainsi que la sortie $\hat{x}^0$ de l'algorithme IRL1, par exemple, utilisé pour minimiser G_{δ_0}. Comment peut-on alors déterminer la prochaine valeur $\delta_1 \in]\delta_0, 1]$? Plus généralement, à partir de $\hat{x}^k$ et δ_k à l'itération k de la procédure GNC, comment sélectionner la valeur $\delta_{k+1} \in]\delta_k, 1]$? Dans la suite, nous proposons une heuristique simple pour répondre à cette question.

Soit $\hat{x}^k$ la solution après k itérations (i. e. un point critique de G_{δ_k}), alors les composantes de $\hat{x}^k$ peuvent être divisées en trois ensembles :

— $i \notin \sigma(\hat{x}^k)$, i. e. $x_i^k = 0$,

— $i \in \sigma^-(\hat{x}^k)$ où σ^- est défini pour G_δ ($\delta \in (0,1]$) de manière similaire au cas correspondant à G_{CEL0}, i. e.

$$\sigma^-(x) := \left\{ i \in \sigma(x) : 0 < |x_i| < \frac{\sqrt{2\lambda}}{\|a_i\|\delta} \right\}. \tag{10.38}$$

Les composantes de $\hat{x}^k$ appartenant à $\sigma^-(\hat{x}^k)$ sont appelées «*composantes pénalisées*» étant donné qu'une petite perturbation de ces dernières entraine une modification de la valeur de la pénalité, i. e. $\exists \mu > 0$ tel que $\Phi_{\delta_k}(\hat{x}^k + \varepsilon e_i) \neq \Phi_{\delta_k}(\hat{x}^k)$ pour tout $i \in \sigma^-(\hat{x}^k)$ et tout $|\varepsilon| \leqslant \mu$.

— par opposition, $i \in \sigma(\hat{x}^k) \setminus \sigma^-(\hat{x}^k)$ sont les composantes dites «*non-pénalisées*» puisque $\exists \mu > 0$ tel que $\Phi_{\delta_k}(\hat{x}^k + \varepsilon e_i) = \Phi_{\delta_k}(\hat{x}^k)$ pour tout $i \in \sigma(\hat{x}^k) \setminus \sigma^-(\hat{x}^k)$ et $|\varepsilon| < \mu$ (excepté pour $\hat{x}_i^k = \frac{\sqrt{2\lambda}}{\|a_i\|\delta}$ où l'égalité précédente n'est vraie que pour ε du même signe que $\hat{x}_i^k$).

L'idée consiste alors à choisir δ_{k+1} de sorte que les ensembles décrits ci-dessus restent inchangés pour la nouvelle valeur δ_{k+1} (avant minimisation). Ce choix est motivé par le fait que nous pouvons voir l'ensemble des composantes *non-pénalisées* comme un ensemble de composantes actives candidates à faire partie du support de la solution finale. La procédure GNC est alors vue comme une méthode sélectionnant et/ou dé-sélectionnant des

composantes à chaque itération pour terminer, lorsque $\delta = 1$, avec une solution $\hat{x}$ contenant uniquement des composantes *nulles* et *non-pénalisées* (i. e. $\sigma^-(\hat{x}) = \emptyset$).

En suivant cette heuristique, les ensembles définis ci-dessus restent inchangés pour tout $\delta \in [\delta_k, \delta_\star[$ où

$$\delta_\star = \min_{i \in \sigma^-(\hat{x}^k)} \frac{\sqrt{2\lambda}}{\|a_i\||\hat{x}_i^k|}. \tag{10.39}$$

L'objectif étant de définir une séquence croissante $(\delta_k)_{k \in \mathbb{I}_K}$, nous considérons

$$\delta_{k+1} := \delta_\star \times \delta_{acc} > \delta_k, \tag{10.40}$$

où $\delta_{acc} > 1$ est un coefficient d'accroissement pour δ donné en entrée de l'algorithme. Notons que ce coefficient est en contradiction avec la règle que nous nous sommes fixé précédemment, à savoir préserver les trois ensembles que nous avons décrits ci-dessus, mais il est utile en pratique afin d'éviter de faire des «pas» trop petits et permet d'accélérer la convergence (typiquement, on prendra $\delta_{acc} = 1.01$). L'algorithme 13 résume la procédure inspirée GNC proposée, nommée Graduated Non convexity Celo (GNCelo).

Algorithme 13 : GNCelo pour la minimisation de G_{CEL0}.

Entrées : Algo, $\delta_0 \in [0, 1[$, $\delta_{acc} > 1$, $x^{init} \in \mathbb{R}^N$

1 /* Algo$(\tilde{x}, \delta, \lambda)$ est un algorithme, initialisé par $\tilde{x}$, convergeant vers un point critique de G_δ pour le λ donné (et tel que $\sigma^-(\hat{x}) = \emptyset$ lorsque $\delta = 1$). */

2 $\hat{x}^0 \leftarrow$ Algo$(x^{init}, \delta_0, \lambda)$;

3 **tant que** $\delta_k < 1$ **faire**

4 $\quad \delta_{k+1} = \min\left(1, \delta_{acc} \times \min_{i \in \sigma^-(\hat{x}^k)} \frac{\sqrt{2\lambda}}{\|a_i\||\hat{x}_i^k|}\right)$;

5 $\quad \hat{x}^{k+1} \leftarrow$ Algo$(\hat{x}^k, \delta_{k+1}, \lambda)$;

Sorties : x^k

La proposition suivante assure la terminaison de l'algorithme GNCelo.

Proposition 10.7. *(Terminaison de l'algorithme GNCelo) La séquence $(\delta_k)_{k \in \mathbb{I}_K}$ générée par l'algorithme GNCelo est strictement croissante. De plus, si $\delta_0 > 0$, l'algorithme est assuré de terminer en moins de*

$$K^\star = -\frac{\ln(\delta_0)}{\ln(\delta_{acc})} \tag{10.41}$$

itérations.

Démonstration. Soit $\hat{x}^k$ la solution à l'itération $k \in \mathbb{N}$ associée au paramètre δ_k. Par définition de $\sigma^-(\hat{x}^k)$ en (10.38), nous avons

$$\forall i \in \sigma^-(\hat{x}^k),\ |\hat{x}_i^k| < \frac{\sqrt{2\lambda}}{\|a_i\|\delta_k} \tag{10.42}$$

$$\implies \forall i \in \sigma^-(\hat{x}^k),\ \delta_k < \frac{\sqrt{2\lambda}}{\|a_i\||\hat{x}_i^k|} \tag{10.43}$$

$$\implies \delta_k < \min_{i\in\sigma^-(\hat{x}^k)} \frac{\sqrt{2\lambda}}{\|a_i\||\hat{x}_i^k|} = \delta_\star \tag{10.44}$$

$$\implies \delta_{k+1} = \delta_\star \times \delta_{acc} > \delta_\star > \delta_k.. \tag{10.45}$$

Ainsi, la séquence $(\delta_k)_{k\in\mathbb{I}_k}$ générée par l'algorithme est strictement croissante. Par ailleurs, des inégalités précédentes on tire

$$\delta_\star > \delta_k \implies \delta_{k+1} = \delta_\star \times \delta_{acc} > \delta_k \times \delta_{acc} \implies \delta_{k+1} > \delta_0 \times (\delta_{acc})^{k+1}, \tag{10.46}$$

La suite géométrique $(\delta_0 \times (\delta_{acc})^{k+1})_{k\in\mathbb{N}}$ divergeant vers $+\infty$ (car $\delta_{acc} \geqslant 1$) assure qu'il existe $K^\star \in \mathbb{N}$ tel que $\delta_{K^\star} \geqslant 1$ et l'algorithme s'arrête. En particulier on a

$$\delta_0 \times (\delta_{acc})^{K^\star} \geqslant 1 \iff K^\star \geqslant -\frac{\ln(\delta_0)}{\ln(\delta_{acc})}, \tag{10.47}$$

ce qui termine la preuve. □

Remarque 10.8. La borne (10.41) donnée par la proposition 10.7 est très pessimiste (la majoration dans la preuve est grossière). En pratique, l'algorithme s'arrête après un nombre d'itérations bien moins important que celui suggéré par cette borne. Par ailleurs, la condition $\delta_0 > 0$ est utilisée pour alléger la preuve mais on peut également montrer que le résultat de la proposition est valable lorsque $\delta_0 = 0$ (en modifiant légèrement la borne (10.41)). Cependant, nous verrons qu'en pratique, il est préférable de ne pas choisir δ_0 trop faible.

10.4 CHEMIN DE RÉGULARISATION POUR G_{CEL0}

Lorsque l'on travaille avec des problèmes pénalisés du type (P_λ), la question du choix du paramètre λ est toujours délicate mais cruciale en pratique. Une idée consiste alors à chercher un ensemble de solutions pour un continuum de valeurs décroissantes de λ, allant de $+\infty$ à 0, plutôt que de résoudre le problème pour $\lambda \in \mathbb{R}_+^\star$ fixé. On appelle ainsi l'ensemble des solutions obtenues *chemin de régularisation* pour la fonctionnelle considérée (*regularization path* en anglais). Une telle approche a été utilisée par Efron et al. (2004) pour le problème de moindres carrés pénalisé en norme ℓ_1 (BPDN) ou encore par Mazumder et al. (2012) lorsque la pénalité MCP est substituée à la norme-ℓ_0 dans (P_λ) (notons que dans ces travaux, la séquence $+\infty = \lambda^0 < \lambda^1 < \cdots < \lambda^K = 0$ est fixée par l'utilisateur). Enfin, les algorithmes CSBR et L0-PD, proposés par Soussen et al. (2015), ont été développés dans le but de déterminer un chemin de régularisation (sous-optimal) pour le problème (P_λ) où la séquence $(\lambda^k)_{k\in\mathbb{N}}$ est déterminée par l'algorithme comme pour le LARS (Efron et al., 2004).

Afin de proposer une méthode capable de construire à la fois une séquence $(\lambda^k)_{k\in\mathbb{N}}$ ainsi qu'une séquence $(\hat{x}_{\lambda^k})_{k\in\mathbb{N}}$ de points critiques de G_{CEL0} associés vérifiant $\sigma^-(\hat{x}_{\lambda^k}) = \emptyset$ (qui sont donc aussi des minimiseurs locaux de G_{ℓ_0} d'après le lemme 10.1), nous introduisons le résultat suivant.

Proposition 10.9. *Soit* $\hat{x} \in \mathbb{R}^N$ *un point critique de* G_{CEL0} *pour* $\lambda \in \mathbb{R}_+$ *tel que* $\hat{\sigma}^- = \sigma^-(\hat{x}) = \emptyset$. *Notons* $\hat{\sigma} = \sigma(\hat{x})$. *Alors, il existe*

$$\underline{\lambda} := \max_{i \in \hat{\sigma}^c} \frac{\langle a_i, A\hat{x} - d\rangle^2}{2\|a_i\|^2} \quad \textit{et} \quad \bar{\lambda} := \min_{i \in \hat{\sigma}} \frac{\hat{x}_i^2 \|a_i\|^2}{2}, \tag{10.48}$$

tels que $\hat{x}$ *est également un point critique de* G_{CEL0} *pour tout* $\lambda' \in [\underline{\lambda}, \bar{\lambda}]$. *De plus si* $\sigma(\hat{x}) = \mathbb{I}_N$ *(resp.* $\sigma(\hat{x}) = \emptyset$*), alors* $\underline{\lambda} = 0$ *(resp.* $\bar{\lambda} = +\infty$*).*

Démonstration. Puisque que $\hat{x}$ est un point critique de G_{CEL0} tel que $\hat{\sigma}^- = \emptyset$ il vérifie, d'après le lemme 9.10 (page 101),

$$\forall i \in \hat{\sigma},\ \hat{x}_i \|a_i\|^2 = -\langle a_i, A\hat{x}^{(i)} - d\rangle, \tag{10.49}$$

qui est indépendant de λ, ainsi que les inégalités suivantes :

$$\begin{cases} |\langle a_i, A\hat{x}^{(i)} - d\rangle| \leqslant \sqrt{2\lambda}\|a_i\| & \forall i \in \hat{\sigma}^c, \\ |\langle a_i, A\hat{x}^{(i)} - d\rangle| \geqslant \sqrt{2\lambda}\|a_i\| & \forall i \in \hat{\sigma}, \end{cases} \tag{10.50}$$

$$\underset{(10.49)}{\iff} \begin{cases} \langle a_i, A\hat{x} - d\rangle^2/(2\|a_i\|^2) \leqslant \lambda & \forall i \in \hat{\sigma}^c, \\ \hat{x}_i^2\|a_i\|^2/2 \geqslant \lambda & \forall i \in \hat{\sigma}, \end{cases} \tag{10.51}$$

Ainsi $\hat{x}$ vérifie toujours $\hat{\sigma}^- = \emptyset$, (10.49) et (10.51) pour tout $\lambda \in [\underline{\lambda}, \bar{\lambda}]$ où $\underline{\lambda}$ et $\bar{\lambda}$ sont définis par (10.48) ce qui termine la preuve. □

La proposition 10.9 suggère donc l'algorithme 14 dans le but de déterminer un ensemble de points critiques de G_{CEL0} pour un continuum de valeurs décroissantes de λ allant de $+\infty$ à 0. Notons que pour $\lambda = +\infty$ le minimiseur global de G_{CEL0} (et G_{ℓ_0}) est $\hat{x} = 0$ et lorsque $\lambda = 0$, on considèrera la solution de norme minimale donnée par la pseudo inverse $A^\dagger$ avec $\hat{x} = A^\dagger d$.

Algorithme 14 : «Chemin de régularisation» pour G_{CEL0}.

Entrées : Algo, $\lambda_{dec} \in]0,1[$, $\lambda_{stop} \in \mathbb{R}_+$

/* Algo$(\tilde{x}, \lambda)$ est un algorithme, initialisé par $\tilde{x}$, convergeant vers un point critique $\hat{x}$ de G_{CEL0} (pour le λ donné) tel que $\sigma^-(\hat{x}) = \emptyset$ (par exemple le Macro-Algo, ou le GNCelo). */

$\lambda^0 = +\infty$;

$x^0 = 0_{\mathbb{R}^N}$;

tant que $\lambda^k > \lambda_{stop}$ **faire**

  $\underline{\lambda} = \max_{i \in (\sigma(x^k))^c} \frac{\langle a_i, Ax^k - d\rangle^2}{2\|a_i\|^2}$;

  $\lambda^{k+1} = \underline{\lambda} \times \lambda_{dec}$;

  $x^{k+1} \leftarrow \text{Algo}(x^k, \lambda^{k+1})$;

$\lambda^{k+1} = 0$;

$x^{k+1} = A^\dagger d$;

Sorties : $(x^k)_{k \in \mathbb{I}_K}, (\lambda^k)_{k \in \mathbb{I}_K}$

10.5 QUELQUES COMPARAISONS NUMÉRIQUES

Dans cette section, nous présentons quelques expériences numériques en petite dimension utilisant les méthodes décrites dans le présent chapitre pour la minimisation de G_{CEL0}. En particulier, nous montrerons que minimiser CEL0 plutôt que minimiser directement G_{ℓ_0} (avec l'algorithme IHT par exemple), permet d'atteindre de «meilleurs» minima locaux en terme de valeur de la fonction objectif G_{ℓ_0}. Pour des résultats numériques en plus grande dimension, nous renvoyons le lecteur au chapitre 11 dédié à différentes applications que l'on peut rencontrer, entre autres, en microscopie et traitement d'antennes.

10.5.1 *Capacité à minimiser la fonctionnelle CEL0*

Afin d'étudier la capacité des algorithmes à minimiser G_{CEL0} (et donc G_{ℓ_0} d'après les résultats théoriques du chapitre 9), nous considérons des problèmes de petite taille (e. g. $M = 7$ et $N = 15$) pour lesquels nous sommes en mesure d'effectuer une recherche exhaustive de la valeur minimale de la fonctionnelle.

La comparaison entre l'algorithme IHT (minimisant directement G_{ℓ_0}) et le Macro-Algo combiné avec l'algorithme FBS ou IRL1 (minimisant G_{ℓ_0} par l'intermédiaire de G_{CEL0}), sera effectuée en terme de minimisation de la fonction objectif G_{ℓ_0}. En d'autres termes, on dira qu'un algorithme est meilleur qu'un autre s'il converge vers une valeur plus faible de la fonction objectif G_{ℓ_0}. Pour ce faire, pour chaque problème considéré, une recherche exhaustive sera réalisée afin de déterminer un minimiseur global $x^\star$ ainsi que la valeur de la fonctionnelle associée $G_{\ell_0}(x^\star)$. Nous pourrons alors considérer l'erreur

$$\epsilon(\hat{x}, x^\star) := |G_{\ell_0}(\hat{x}) - G_{\ell_0}(x^\star)|, \tag{10.52}$$

comme mesure de performance. Plus cette erreur est faible, meilleure est la solution estimée $\hat{x} \in \mathbb{R}^N$.

DESCRIPTION DE L'EXPÉRIENCE Dans la suite, nous réalisons l'expérience suivante :

1. génération aléatoire de 1000 problèmes (P_λ) de taille (7×15) où les entrées de $A \in \mathbb{R}^{7\times 15}$ et de $d \in \mathbb{R}^7$ sont tirées selon une distribution uniforme.
2. pour chacun de ces problèmes, on exécute l'algorithme IHT ainsi que le Macro-Algo combiné avec FBS et IRL1 à partir de deux initialisations différentes :
   — $x^{init} = A^T d$;
   — $x^{init} = \hat{x}_{\ell_1}$, la solution du problème relaxé ℓ_1.
3. pour chaque solution $\hat{x}$ estimée par l'un des algorithmes mentionnées dans l'étape 2, nous calculons l'erreur $\epsilon(\hat{x}, x^\star)$ (équation (10.52)), où $x^\star$ est une solution globale du problème (P_λ) déterminée par recherche exhaustive.

Le pas de descente utilisé dans les algorithmes IHT, FBS et FISTA[7] est fixé à $\frac{0.99}{L}$ où $L = \|A\|^2$ correspond à la constante de Lipschitz du gradient du terme quadratique de la fonction objectif de (P_λ).

RÉSULTATS Sur la figure 43, nous pouvons observer les histogrammes cumulés normalisés des erreurs $\epsilon(\hat{x}, x^\star)$ obtenues pour les 1000 générations de problèmes, les deux initialisations mentionnées précédemment et enfin trois valeurs différentes de λ (0.1, 0.5 et

7. FISTA (BECK et TEBOULLE, 2009) est utilisé dans la boucle interne de l'algorithme IRL1.

1). Ces courbes représentent donc la proportion des problèmes générés pour lesquels l'algorithme a convergé vers une solution $\hat{x}$ dont l'erreur $\epsilon(\hat{x}, x^\star)$ est inférieure à la valeur correspondante sur l'axe des abscisses. Ainsi,

— l'ordonnée à l'origine de ces courbes correspond à la proportion, sur les 1000 réalisations, où l'algorithme a atteint un minimum global du problème ;
— la plus petite valeur de l'erreur ϵ pour laquelle la courbe est égale à 1 nous donne la plus grande erreur obtenue parmi toutes les générations de problèmes ;
— enfin, la rapidité de la courbe à tendre vers 1 est caractéristique du bon comportement de l'algorithme.

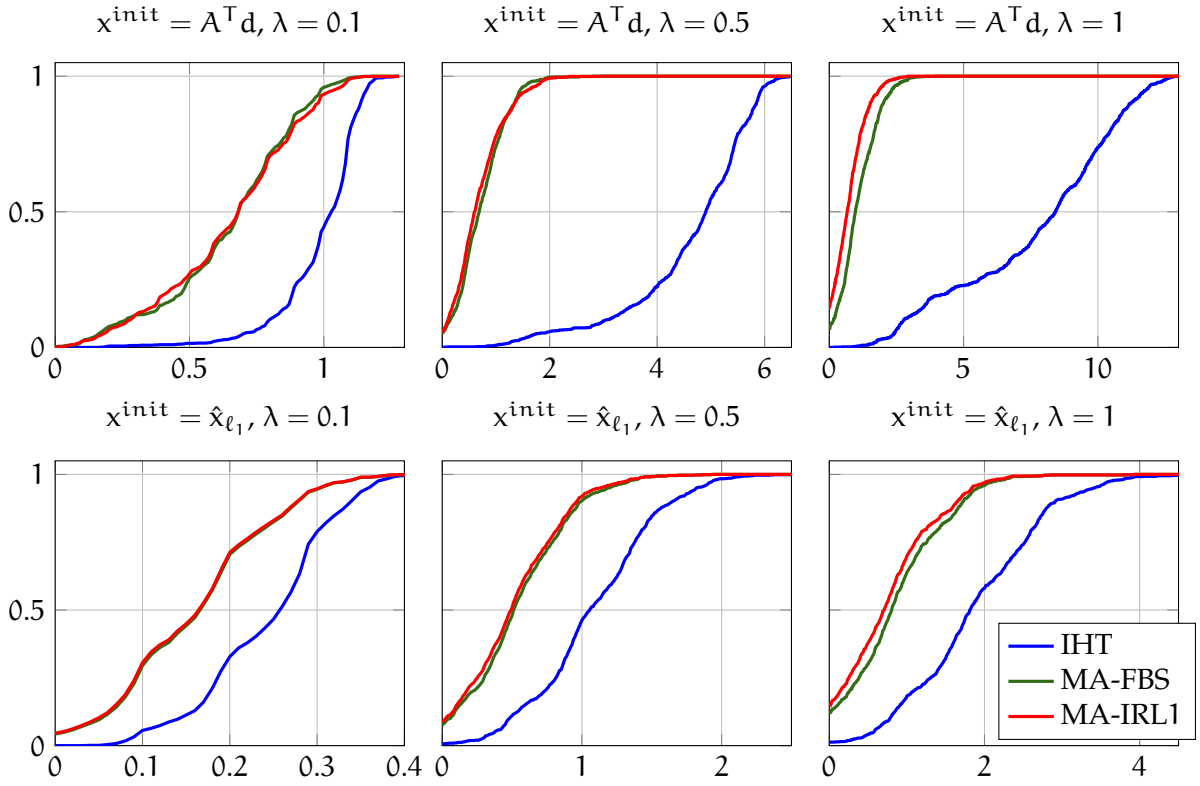


FIGURE 43 – Histogrammes cumulés normalisés des erreurs $|G_{\ell_0}(\hat{x}) - G_{\ell_0}(x^\star)|$, où $\hat{x}$ est la solution estimée par l'algorithme et $x^\star$ est un minimiseur global de G_{ℓ_0}. Ces histogrammes sont déterminés à partir de 1000 générations aléatoires de $A \in \mathbb{R}^{7\times 15}$ et $d \in \mathbb{R}^7$ selon une distribution uniforme. Deux initialisations différentes sont considérées : $x^{init} = A^T d$ et $x^{init} = \hat{x}_{\ell_1}$ la solution du problème relaxé ℓ_1. L'expérience est répétée pour trois valeurs de λ (0.1, 0.5 and 1). Pour chaque configuration, l'estimation d'une solution est réalisée avec l'algorithme IHT (bleu) ainsi que le Macro-Algo combiné avec IRL1 (rouge) ou FBS (vert).

Comme nous pouvons l'observer sur la figure 43, le Macro-Algo produit de meilleurs résultats que IHT quelque soit la configuration. Cette observation est en accord avec le fait que G_{CEL0} élimine des minimiseurs locaux de G_{ℓ_0} ce qui rend le Macro-Algo plus apte à «éviter» des minimiseurs locaux de G_{ℓ_0}. Nous pouvons aussi remarquer une légère différence entre l'utilisation de l'algorithme FBS ou IRL1 à l'intérieur du Macro-Algo. Cela illustre un point souligné par OCHS et al. (2014) concernant la capacité de l'algorithme IRL1 à éviter des minimiseurs locaux (voir aussi l'exemple 1D de la figure 38 page 119).

Concernant l'initialisation, elle affecte grandement l'efficacité de l'algorithme IHT alors que le Macro-Algo semble être moins sensible à cette dernière. En effet, initialiser avec la solution de la relaxation convexe ℓ_1 améliore clairement les résultats de IHT ce qui est

moins le cas pour le Macro-Algo lorsque $\lambda = 0.5$ ou $\lambda = 1$. Pour le cas $\lambda = 0.1$, on observe aussi une amélioration des performances du Macro-Algo lorsque il est initialisé par $\hat{x}_{\ell_1}$. Ce phénomène peut s'expliquer avec les résultats présentés dans la table 3 (page 113) où l'on observe qu'un nombre plus important de minimiseurs stricts de G_{ℓ_0} sont éliminés par G_{CEL0} lorsque λ prend de grandes valeurs.

Finalement, les résultats de la figure 43 montrent un comportement très intéressant du Macro-Algo et suggère qu'il est préférable d'attaquer le problème (P_λ) par l'intermédiaire de la minimisation de la fonctionnelle G_{CEL0}.

10.5.2 *Gain apporté par la méthode inspirée GNC*

Dans la section 10.3, nous avons proposé la méthode GNCelo, utilisant les idées des méthodes GNC pour minimiser G_{CEL0} (voir algorithme 13). Afin de voir le gain apporté par cette méthode en comparaison avec une minimisation directe de G_{CEL0}, nous reprenons l'expérience du paragraphe précédent (figure 43).

La figure 44 présente les résultats obtenus avec l'algorithme GNCelo (combiné avec le Macro-Algo utilisant IRL1) pour différentes valeurs du paramètre initial δ_0. Clairement, la stratégie GNC employée par l'algorithme GNCelo permet de mieux minimiser G_{CEL0} par rapport à une minimisation directe.

Cependant, nous pouvons remarquer que la performance de la méthode proposée est dépendante du paramètre initial δ_0. En effet, on peut voir sur la figure 44 que pour $\lambda = 1$, une valeur de δ_0 trop petite (par exemple 10^{-6}) produit des résultats moins bons que la minimisation directe. Au contraire, pour des λ plus petits, comme par exemple $\lambda = 0.1$, les résultats obtenus avec $\delta_0 = 10^{-6}$ s'améliorent alors que ceux correspondants à $\delta_0 = 0.5$ se dégradent (en comparaisons avec ceux obtenus pour $\delta_0 = 0.1$ et $\delta_0 = 10^{-3}$). On peut expliquer ce comportement par le fait que, plus δ_0 est faible, moins la première solution obtenue $\hat{x}_{\delta_0}$ est parcimonieuse et ainsi :

— pour de grandes valeurs de λ, les minimiseurs globaux de G_{CEL0} étant très parcimonieux, prendre δ_0 trop faible (tel que $\hat{x}_{\delta_0}$ ne soit pas parcimonieux) entraîne l'algorithme vers un «mauvais» minimiseur local.

— inversement lorsque λ décroît, les minimiseurs globaux de G_{CEL0} sont de moins en moins parcimonieux et une première solution $\hat{x}_{\delta_0}$ peu parcimonieuse peut alors devenir une bonne initialisation pour les minimisations suivantes.

Notons aussi que sur la figure 44, l'algorithme GNCelo est initialisé par $x^{init} = 0_{\mathbb{R}^N}$ ce qui peut expliquer les légères différences que l'on observe pour le cas $\delta_0 = 1$ avec la courbe correspondant à MA-IRL1 sur la figure 43.

Au vu de ces résultats, l'algorithme GNCelo doit être plutôt vu comme une stratégie d'initialisation par déformation de la fonctionnelle G_{CEL0} plutôt que comme une méthode GNC à proprement parler étant donné qu'il est nécessaire de l'initialiser avec une fonctionnelle G_{δ_0} non-convexe. Dans la suite (en particulier la section 11.1 du chapitre 11), nous considèrerons généralement $\delta_0 = 0.1$ qui s'est révélé être efficace en pratique.

10.5.3 *Reconstruction exacte (Exact Recovery)*

Une expérience classique dans le contexte de l'optimisation parcimonieuse, et plus particulièrement en échantillonnage compressé, s'intéresse à la capacité des algorithmes à reconstruire exactement un signal parcimonieux observé à travers une matrice d'acquisition.

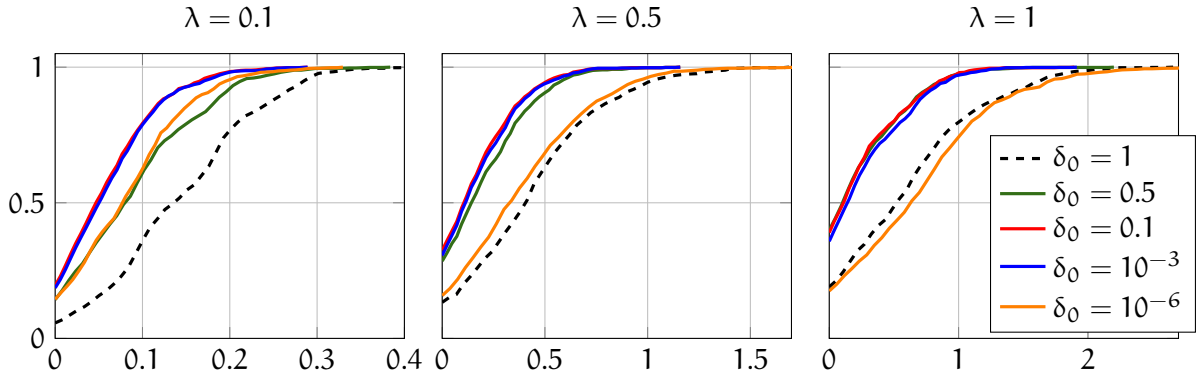


FIGURE 44 – . Histogrammes cumulés des erreurs $|G_{\ell_0}(\hat{x}) - G_{\ell_0}(x^\star)|$, calculées comme sur la figure 43, pour l'algorithme GNCelo initialisé par $x^{init} = 0_{\mathbb{R}^N}$ et différentes valeurs du paramètre initial δ_0

Considérons une matrice $A \in \mathbb{R}^{M\times N}$ dont les entrées sont générées selon une distribution gaussienne et un signal $x^\star$ composé de $K \leqslant N$ entrées non-nulles générées comme les composantes de A. Les données non-bruitées sont alors définies par $y = Ax^\star$.

Plusieurs réalisations de A et de $x^\star$ sont ainsi effectuées et pour chacune d'entre elles, le signal $\hat{x}$, obtenu avec l'algorithme étudié, est dit reconstruire exactement $x^\star$ dès lors que

$$\|\hat{x} - x^\star\|_\infty \leqslant 10^{-3}. \tag{10.53}$$

On peut alors définir une figure, appelée *Diagramme de Phase*, qui reflète la probabilité qu'a l'algorithme de reconstruire exactement un signal parcimonieux en fonction de M et K.

La figure 45 présente les résultats obtenus pour $M \in [50 : 5 : 130]$, $N = 256$ et $K \in [20 : 5 : 100]$ avec l'algorithme IRL1 de CANDÈS et al. (2008)[8] et GNCelo ($\delta_0 = 0.1$). On peut voir sur cette figure que l'algorithme GNCelo produit de meilleurs résultats que IRL1 qui est déjà plus performant que la simple relaxation convexe ℓ_1.

10.6 CONCLUSION

Dans ce chapitre, nous avons présenté différents algorithmes «nonsmooth-nonconvex» pouvant être utilisés pour la minimisation de G_{CEL0}. Notons que ces algorithmes ont été proposés, pour la plupart, ces dernières années et ce domaine de recherche est actuellement très actif. Cela suggère que des améliorations seront encore faites dans les prochaines années ce qui sera sans doute bénéfique pour la minimisation de G_{CEL0}.

Étant donné que de tels algorithmes ne nous assurent généralement qu'une convergence vers des points critiques de la fonctionnelle, nous avons proposé un Macro-Algo permettant d'ajouter une boucle externe à n'importe quel algorithme de l'état de l'art (pourvu qu'il converge vers un point critique de G_{CEL0} et qu'il vérifie une condition de décroissance suffisante du critère) assurant la convergence de la séquence générée vers un point critique de G_{CEL0} qui est aussi minimiseur (local) de G_{ℓ_0} (théorème 10.3). Des illustrations numériques ont alors montré l'intérêt de minimiser la fonctionnelle CEL0 en comparaison avec une minimisation directe de G_{ℓ_0}.

8. Nous utilisons le logiciel ℓ_1-MAGIC www.l1-magic.org développé entre autres par les auteurs (CANDÈS et al., 2008).

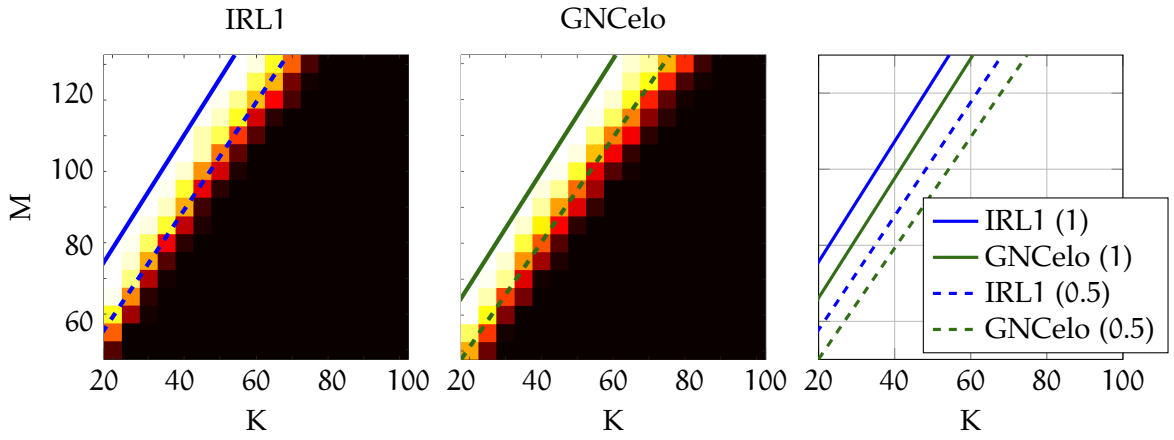


FIGURE 45 – Diagrammes de Phase pour les algorithmes IRL1 et GNCelo (pour $\delta_0 = 0.1$). Pour chaque valeur de $M \in [50 : 5 : 130]$ et de $K \in [20 : 5 : 100]$, 100 matrices $A^{M\times 256}$ et vecteurs K-parcimonieux $x^\star \in \mathbb{R}^{256}$ sont générés selon une distribution gaussienne. À partir des données non-bruitées $y = Ax^\star$, les deux algorithmes sont exécutés et l'estimation $\hat{x}$ obtenue est considérée exacte lorsque (10.53) est vérifiée. Les diagrammes de cette figure représentent la probabilité de l'algorithme a exactement reconstruire $x^\star$ (blanc = 1 et noir = 0). Les droites représentent les transitions où la probabilité de reconstruction vaut 1 (traits pleins) et 0.5 (tirets).

D'autre part, nous nous sommes inspiré des méthodes GNC pour proposer l'algorithme GNCelo (algorithme 13) dont les performances se sont révélées être meilleures qu'une minimisation directe du critère G_{CEL0} avec un algorithme de l'état de l'art. GNCelo résout une séquence de problèmes pour lesquels la fonction objectif est une déformation de G_{CEL0} en tirant bénéfice, pour chaque problème de la séquence, d'une initialisation «à chaud» donnée par la solution obtenue pour le problème précédent.

Afin de s'affranchir de la sélection du paramètre λ, nous avons également proposé une méthode (algorithme 14) permettant de déterminer un «chemin de régularisation» (sous-optimal) pour G_{CEL0}. Cette méthode utilise la caractérisation des points critiques de G_{CEL0} dans le but de générer une séquence de problèmes pour différentes valeurs de λ décroissantes allant de $+\infty$ à 0. Des résultats intéressants ont alors été montrés dans le contexte du problème de reconstruction exacte de signaux. Notons que cette stratégie sera également utilisée dans la section 11.1 du chapitre 11 sur un problème de déconvolution d'impulsions.

Les résultats obtenus ici, ainsi que ceux que nous présenterons dans le chapitre 11, sont prometteurs et montrent que G_{CEL0} est une très bonne alternative (exacte) à G_{ℓ_0}.

11

APPLICATIONS EN TRAITEMENT DU SIGNAL ET DES IMAGES

Sommaire

Ce chapitre est dédié à l'application des algorithmes présentés précédemment pour la minimisation de G_{CEL0} à différentes types de problèmes rencontrés en traitement du signal et des images. Nous aborderons en particulier les problèmes de déconvolution de trains d'impulsions, d'estimation de canal et de directions d'arrivées en traitement d'antennes ou encore le problème de reconstruction PALM pour la super-résolution en microscopie de fluorescence.

11.1 DÉCONVOLUTION DE TRAINS D'IMPULSIONS

11.1.1 *Présentation du problème*

Nous commençons avec le problème classique de déconvolution parcimonieuse d'impulsions (*sparse spikes deconvolution* en anglais) que l'on rencontre dans de nombreux domaines comme par exemple la sismologie, l'astronomie ou encore le contrôle non-destructif (Idier, 2013, chapitre 5). Le problème s'écrit :

$$y = h * x^\star + n, \tag{11.1}$$

où $x^\star$ représente le signal parcimonieux recherché, h la réponse impulsionnelle du système d'acquisition (noyau de convolution) et n est un bruit gaussien. Étant donné que ce problème est linéaire il peut évidemment être réécrit sous la forme matricielle suivante :

$$y = Hx^\star + n, \tag{11.2}$$

avec $y \in \mathbb{R}^{N+M-1}$, $H \in \mathbb{R}^{(N+M-1)\times N}$ une matrice de Toeplitz dont les colonnes sont des versions translatées du noyau h et $n \sim \mathcal{N}(0, \sigma_n^2 I_{N+M-1})$. Dans la formulation précédente, le support de h est considéré de taille $M \in \mathbb{N}$ et la matrice H n'est pas carrée afin d'intégrer des conditions aux bords nulles. Une illustration de ce problème est donnée par la figure 46.

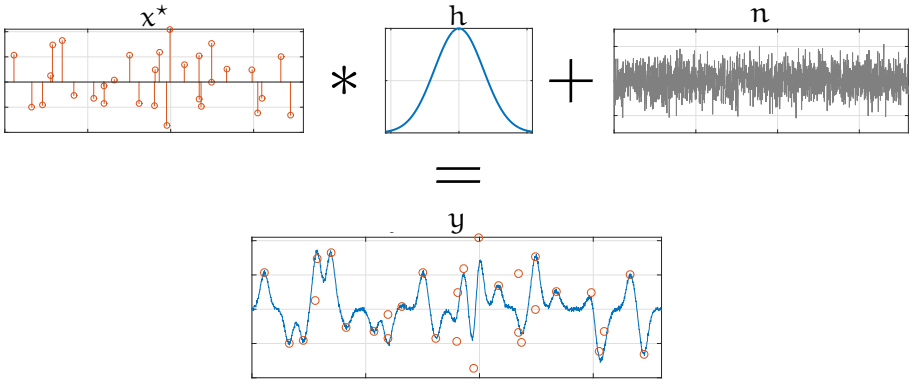


FIGURE 46 – Problème de déconvolution de train d'impulsions.

11.1.2 *Génération des données, algorithmes et critères de performance*

Dans la suite, nous reprenons en partie certaines expériences menées dans (SOUSSEN et al., 2015).

GÉNÉRATION DES DONNÉES Les données utilisées pour l'évaluation des algorithmes sont générées en considérant :

- un noyau h construit en tronquant une gaussienne de variance σ_g^2 sur $[-3\sigma_g, 3\sigma_g]$;
- les $K \in \mathbb{N}$ entrées non-nulles de $x^\star$ sélectionnées uniformément et leurs valeurs générées de manière i.i.d. selon une distribution gaussienne ;
- la variance du bruit déterminée par :

$$\sigma_n^2 = \frac{\|Hx^\star\|^2}{(M+N-1)10^{-\mathrm{SNR}/10}}, \tag{11.3}$$

pour un SNR (dB) fixé ;

Trois configurations $(N, \sigma_g, K, \mathrm{SNR})$, présentées dans la table 4, seront utilisées dans la suite et pour chacune d'entre elles 50 réalisations de problèmes (i. e. $x^\star$ et le bruit) seront effectuées afin de moyenner les résultats.

Configuration	$N+M-1$	N	K	σ_g	SNR (dB)
C_1	300	282	30	3	25
C_2	900	756	10	24	25
C_3	1800	1692	30	18	25

TABLE 4 – Configurations utilisées pour générer les données.

ALGORITHMES Nous souhaitons comparer des algorithmes de la littérature avec la méthode proposée dans le chapitre 10 permettant la détermination d'un «chemin de régu-

larisation» pour G_{CEL0}, combinée avec l'algorithme GNCelo [1]. On notera cette combinaison GNCelo Regularization Path (GNCeloRP) dans la suite. Pour ce faire, nous avons choisi les algorithmes CSBR et L0-PD, proposés par Soussen et al. (2015) (voir section 8.2.4), comme référence étant donné que :

1. ces deux algorithmes minimisent directement G_{ℓ_0} et calculent un ensemble de solutions (sous-optimales) pour un continuum de valeurs du paramètre λ. Ils sont donc directement comparables avec GNCeloRP ;
2. ils ont été comparés récemment dans (Soussen et al., 2015) à plusieurs algorithmes de l'état de l'art et se sont révélés être plus efficaces en terme de minimisation de la fonctionnelle G_{ℓ_0}.

Les trois algorithmes sont stoppés une fois que les solutions pour $\lambda \in [1e^{-5}, +\infty)$ aient été déterminées. Par ailleurs, pour GNCeloRP nous fixons $\lambda_{dec} = 0.9$, $\delta_0 = 0.1$, $\delta_{acc} = 1.1$ et l'algorithme interne IRL1 (exécuté à chaque itération de GNCelo) est stoppé lorsque les deux conditions suivantes sont vérifiées :

$$\frac{\|x^n - x^{n-1}\|}{\|x^{n-1}\| + \varepsilon} < 1e^{-5} \quad \text{et} \quad \frac{G_{\ell_0}(x^n) - G_{\ell_0}(x^{n-1})}{G_{\ell_0}(x^{n-1}) + \varepsilon} < 1e^{-5}, \tag{11.4}$$

pour $\varepsilon \in \mathbb{R}_+^\star$ petit. Pour les algorithmes CSBR et L0-PD, nous utilisons les implémentations fournies sur la page personnelle des auteurs : http://w3.cran.univ-lorraine.fr/perso/charles.soussen/software.html.

Enfin dans GNCeloRP, pour chaque nouvelle valeur λ^{k+1} déterminée, l'appel à GNCelo est initialisé par $0_{\mathbb{R}^N}$ au lieu de x^k comme cela est le cas sur la ligne 7 de l'algorithme 14. En effet, nous avons constaté qu'une initialisation avec la solution précédente lorsque les données étaient bruitées conduisait souvent l'algorithme à converger vers de «mauvais» minima locaux et que de meilleurs résultats étaient obtenus en initialisant toujours par $0_{\mathbb{R}^N}$. Cependant, nous préservons la stratégie de calcul automatique de la séquence décroissante de λ.

Remarque 11.1. Le noyau h étant normalisé, les normes des colonnes de l'opérateur H, nécessaires à la définition de G_{CEL0}, sont égales à 1.

CRITÈRES DE PERFORMANCE Afin de comparer les performances des algorithmes, nous utiliserons trois types de représentations à partir des solutions estimées $\hat{x}_\lambda$:

— tracé du résidu normalisé $\|H\hat{x}_\lambda - y\|^2/\|Hx^\star\|^2$ en fonction de la cardinalité $\|\hat{x}_\lambda\|_0$. Sur une telle représentation, plus la courbe obtenue est basse, meilleurs sont les résultats ;
— tracé de la valeur objectif normalisée $G_{\ell_0}(\hat{x}_\lambda)/\|d\|^2$ en fonction de λ, reflétant la capacité de l'algorithme à minimiser G_{ℓ_0}. Là encore, plus la courbe est basse, meilleurs sont les résultats ;
— tracé des Détections Correctes (DC) en fonction des Fausses Alarmes (FA) pour le support de la solution estimée $\sigma(\hat{x}_\lambda)$. Nous introduisons un paramètre $\Delta > 0$ définissant une tolérance comme cela est représenté sur la figure 47.

11.1.3 *Résultats numériques*

Nous présentons maintenant les résultats obtenus, moyennés sur 50 réalisations de bruit, pour les trois configurations de la table 4. La figure 48 présente les courbes du résidu normalisé en fonction de la cardinalité alors que les figures 49 et 50 montrent respectivement les tracés de la valeur objectif normalisée en fonction de λ et des DC en fonction des FA.

1. Voir l'algorithme 13 page 130 pour GNCelo et l'algorithme 14 page 132 pour la construction de la séquence de λ.

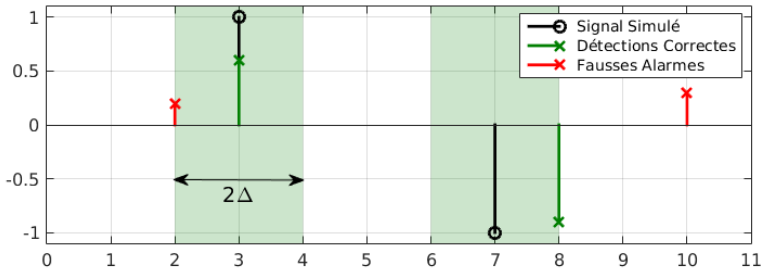


FIGURE 47 – Détections Correctes (DC) et Fausses Alarmes (FA) avec une tolérance $\Delta > 0$. La détection en rouge la plus à gauche n'est pas considérée comme correcte, bien qu'elle soit dans la zone de tolérance, étant donné que le «pic» exact au centre de cette zone est déjà associé à une meilleure détection.

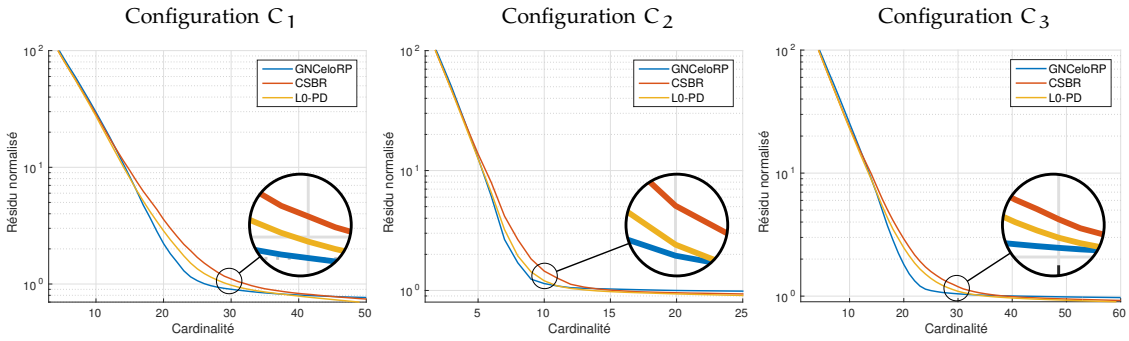


FIGURE 48 – Résidu normalisé $\|H\hat{x}_\lambda - y\|^2/\|Hx^\star\|^2$ en fonction de la cardinalité $\|\hat{x}_\lambda\|_0$ moyenné sur les solutions obtenues pour 50 réalisations de bruit. Chaque courbe correspond à un algorithme et chaque graphe à une configuration de la table 4.

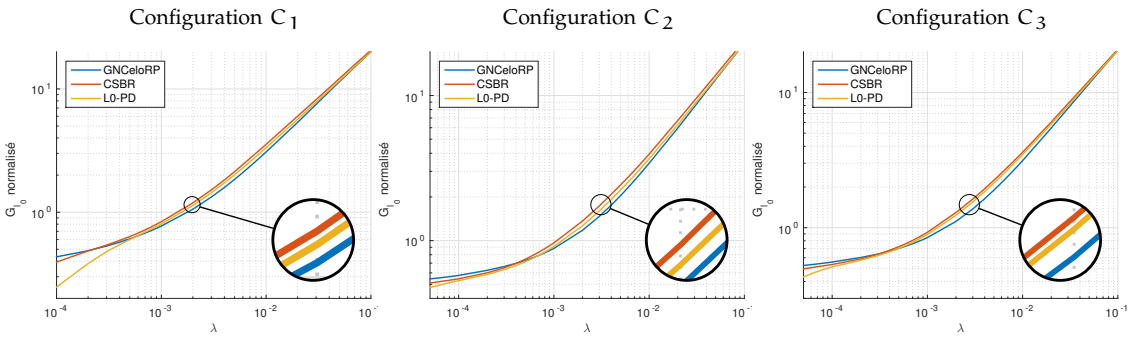


FIGURE 49 – Valeur objectif normalisée $G_{\ell_0}(\hat{x}_\lambda)/\|d\|^2$ en fonction de λ moyennée sur les solutions obtenues pour 50 réalisations de bruit. Chaque courbe correspond à un algorithme et chaque graphe à une configuration de la table 4.

Tout d'abord, nous pouvons remarquer sur les courbes de la figure 48 que pour un niveau de parcimonie donné $K = \|\hat{x}\|_0$, l'algorithme GNCeloRP a convergé vers une solution pour laquelle le résidu est plus faible. En d'autre termes, GNCeloRP permet de déterminer une meilleure K-approximation du signal $x^\star$ que les algorithmes CSBR et L0-PD. Cependant, ce n'est plus vrai lorsque le niveau de parcimonie augmente. Cela s'explique en partie avec les résultats de la figure 49 où nous pouvons voir que GNCeloRP minimise mieux le critère G_{ℓ_0} pour des valeurs de λ généralement supérieures à $5e^{-4}$ (sur ces exemples), alors que pour des λ inférieurs à cette limite les solutions déterminées par GNCeloRP sont nettement moins bonnes que celles obtenues par L0-PD. Ainsi, la cardinalité de la solution augmentant

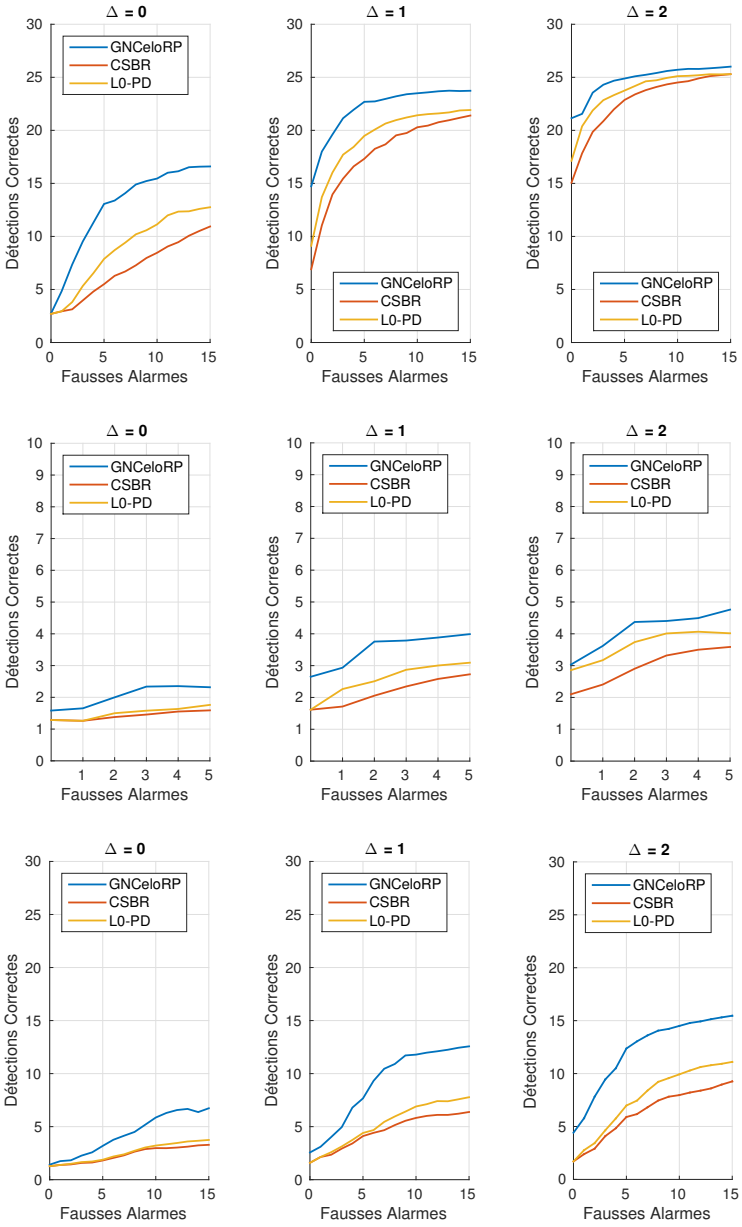


FIGURE 50 – Détections correctes en fonction des fausses alarmes, moyennées sur les solutions obtenues pour 50 réalisations de bruit, pour différentes tolérances $\Delta > 0$ (de gauche à droite $\Delta = 0, 1, 2$). Chaque courbe correspond à un algorithme et chaque ligne correspond à une configuration de la table 4 (de haut en bas C_1, C_2 et C_3).

avec la diminution du paramètre λ, les résultats de la figure 49 correspondant aux valeurs de λ les plus faibles expliquent que les algorithmes L0-PD et CSBR fournissent un meilleur résidu pour les cardinalités les plus grandes de la figure 48.

Notons tout de même que l'on peut améliorer les résultats de GNCeloRP lorsque λ diminue en considérant un paramètre initial δ_0 plus petit. Étant donné que de telles valeurs de λ ne sont pas celles qui nous intéressent ici (elles produisent des solutions pas assez parcimonieuses pour les problèmes générés), nous ne présentons que les résultats obtenus avec $\delta_0 = 0.1$.

Si on s'intéresse maintenant à la figure 50, on peut observer la meilleure capacité de GNCeloRP à localiser les impulsions avec peu de fausses détections bien que l'on ai jamais reconstruction exacte pour les configurations de problèmes considérées ici. Un exemple de

reconstruction pour la configuration C_1 avec les algorithmes GNCeloRP et L0-PD est présenté sur la figure 51. Les différences entre les solutions des deux algorithmes, pour une même cardinalité, sont mises en avant sur les zooms. On peut ainsi voir la meilleure localisation des impulsions produite par GNCeloRP qui permet généralement de bien estimer le support de la solution correspondant aux impulsions dont l'intensité est suffisamment importante.

Pour finir, nous pouvons mentionner que nous avons également testé des configurations avec un SNR plus faible (e. g. 10 dB) pour lesquelles nous n'avons pas noté de grande différence entre GNCelo et L0-PD (excepté bien sûr pour les λ les plus faibles comme nous venons de l'évoquer).

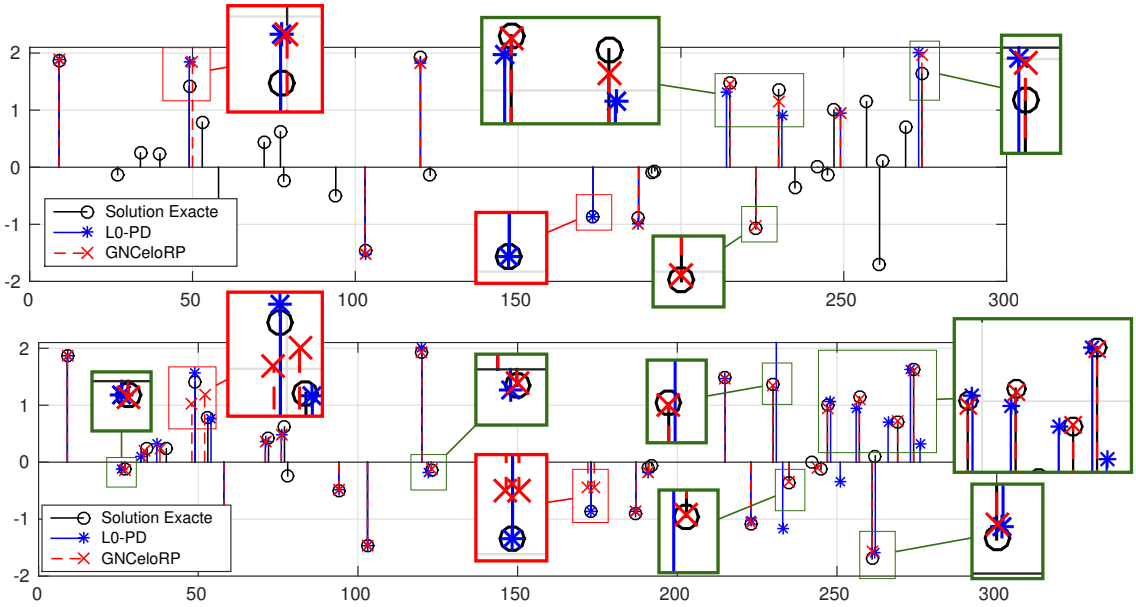


FIGURE 51 – Exemples de reconstructions obtenues avec les algorithmes L0-PD et GNCeloRP pour la configuration C_1. Nous présentons deux solutions, parmi l'ensemble des solutions obtenues pour différentes valeurs de λ, correspondant à des cardinalités $\|\hat{x}\|_0 = 10$ (haut) et $\|\hat{x}\|_0 = 28$ (bas). Les zooms rouges montrent des cas où L0-PD a localisé correctement une impulsion alors que GNCeloRP a échoué et les zooms verts mettent en avant l'inverse.

11.2 TRAITEMENT D'ANTENNES : ESTIMATION DE CANAL ET DE DIRECTIONS D'ARRIVÉES

Nous nous intéressons maintenant à deux autres problèmes rencontrés dans le domaine du traitement d'antennes. Ce travail a été réalisé en collaboration principalement avec Adilson Chinatto de l'université de Campinas (Brésil) dont une partie a été présentée lors du workshop CAMSAP 2015 (CHINATTO et al., 2015). Avant de rentrer dans la présentation de ces deux problèmes et des résultats de simulation, nous devons étendre la pénalité CEL0 au cas complexe et au cas d'une parcimonie structurée par ligne.

11.2.1 *Extension de la pénalité CEL0 au cas complexe et à la parcimonie structurée par ligne*

Le problème d'estimation des directions d'arrivées, présenté dans la suite, nécessite la minimisation de la fonctionnelle

$$X \in \arg\min_{X \in \mathbb{C}^{N \times L}} G_{\ell_{2,0}}(X) := \frac{1}{2}\|AX - Y\|^2 + \lambda \|X\|_{2,0}, \tag{11.5}$$

où $A \in \mathbb{C}^{M\times N}$ définit la matrice d'acquisition, $Y \in \mathbb{C}^{M\times L}$ représente les données mesurées bruitées et $\|\cdot\|_{2,0}$ est la norme mixe $\ell_{2,0}$ donnée par,

$$\|X\|_{2,0} = \sharp\{i \in \mathbb{I}_N : \|X_{i\cdot}\| > 0\} = \sum_{i=1}^{N} |\,\|X_{i\cdot}\|\,|_0, \tag{11.6}$$

avec $X_{i\cdot}$ la i-ème ligne de la matrice $X \in \mathbb{C}^{N\times L}$. La solution recherchée ici est donc une matrice ayant seulement quelques lignes non-nulles. On parlera de parcimonie structurée par ligne. Par ailleurs, les variables sont à valeurs complexes. La question est alors de savoir si les résultats obtenus dans le cas réel vectoriel (voir chapitre 9) sont aussi valables pour le problème (11.5), et dans ce cas de déterminer l'expression de la pénalité CEL0. À la vue de (11.5), il semble qu'il n'y ait pas vraiment de problème à étendre les résultats précédents étant donné que chaque ligne de $X \in \mathbb{C}^{N\times L}$ est associée à une colonne de $A \in \mathbb{C}^{M\times N}$. En particulier, en se restreignant au cas «1D» de (11.5) défini par

$$\hat{x} \in \arg\min_{x\in\mathbb{C}^{1\times L}} g_{2,0}(x) := \frac{1}{2}\|ax - y\|^2 + \lambda\,|\,\|x\|\,|_0, \tag{11.7}$$

où $a \in \mathbb{C}$ et $y \in \mathbb{C}^{1\times L}$ sont les versions «1D»[2] de la matrice A et des données Y, on peut montrer le résultat suivant.

Proposition 11.2. *L'enveloppe convexe de* $g_{2,0}$ *définie en* (11.7) *est donnée par*

$$g^{\star\star}_{2,0}(x) = \frac{1}{2}\|ax - y\|^2 + \phi(a,\lambda;x), \tag{11.8}$$

où $\phi(a,\lambda;\cdot)$ *est définie par :* $\forall x \in \mathbb{C}^{1\times L}$

$$\phi(a,\lambda;x) = \lambda - \frac{|a|^2}{2}\left(\|x\| - \frac{\sqrt{2\lambda}}{|a|}\right)^2 \mathbb{1}_{\left\{\|x\|\leqslant\frac{\sqrt{2\lambda}}{|a|}\right\}}. \tag{11.9}$$

avec $|a|$ *le module du nombre complexe* $a \in \mathbb{C}$ *et* $\|x\|$ *la norme euclidienne de* $x \in \mathbb{C}^{1\times L}$.

Démonstration. La preuve est simplement une réécriture de celle de la Proposition 9.3. □

Nous retrouvons donc la même pénalité CEL0 que pour le cas réel vectoriel excepté la considération du module de $a \in \mathbb{C}$ et de la norme $\|x\|$. Nous pouvons ensuite étendre ce résultat au cas où A^HA est diagonale en suivant les mêmes arguments que ceux utilisés dans la section 9.2. Enfin, en adaptant les résultats de (Nikolova, 2013) et ceux du chapitre 9 au cas du problème (11.5), tous les résultats démontrés pour CEL0 sont également valables pour le problème (11.5) avec la pénalité CEL0 définie par la proposition 11.2. Enfin, on notera la fonction objectif associée par $G_{CEL_{20}}$.

11.2.2 *Problème d'estimation de canal*

Un problème majeur dans la transmission de l'information concerne la détermination de la fonction de transfert du milieu dans lequel le signal se propage. La connaissance de cette fonction est importante afin de diminuer les erreurs de transmission et d'optimiser

2. On entend par «1D» pour x et y que ce sont des vecteurs ligne, étant donné qu'on utilise la norme mixte $\ell_{2,0}$.

la vitesse du transfert. Cependant, elle est généralement inconnue et il est nécessaire de l'estimer. On parle alors *d'estimation de canal* étant donné que l'on recherche les propriétés du canal de transmission. Une manière d'aborder ce problème consiste, par exemple, à utiliser un signal pilote connu à la fois de l'émetteur et du récepteur (Tang et al., 2007) comme cela est décrit dans le paragraphe suivant.

11.2.2.1 *Modélisation*

Lorsqu'un émetteur envoie un signal vers un récepteur, ce dernier reçoit plusieurs versions retardées et atténuées du signal initial. Cela est dû aux multiples réflexions du signal émis sur l'environnement dans lequel il se propage (canal de transmission) comme cela est représenté sur la figure 52. La réponse impulsionnelle du canal de transmission peut alors être modélisée comme un ensemble fini de K délais $\{\tau_1, \dots, \tau_K\}$ et atténuations $\{a_1, \dots, a_K\}$. Formellement, en notant h cette réponse impulsionnelle, on a :

$$h(t) = \sum_{i=1}^{K} a_i \delta_0(t - \tau_i), \tag{11.10}$$

où δ_0 est la fonction de Dirac. Ainsi, pour un signal s émis, le signal reçu au niveau du récepteur est donné par

$$y(t) = (h * s)(t) + b(t), \tag{11.11}$$

où b est un bruit additif.

Considérons maintenant que ces K délais peuvent être représentés sur une grille de taille $N \gg K$ dont les échantillons sont séparés de T_S secondes et que les signaux s envoyés par l'antenne émettrice peuvent être échantillonnés (émis) avec une fréquence $f_S = 1/T_S$. Prenons alors un signal $s \in \mathbb{C}^P$ ($P \geqslant N$) échantillonné et supposé connu à la fois de l'émetteur et du récepteur (*signal pilote* ou *séquence d'apprentissage*). Au vu de ce qui précède, le signal $\tilde{y} \in \mathbb{C}^{N+P-1}$ au niveau du récepteur est donné par

$$\tilde{y} = \tilde{S}h + b, \tag{11.12}$$

où $h \in \mathbb{C}^N$ représente la version discrète de la réponse impulsionnelle h, $b \in \mathbb{C}^{N+P-1}$ est un vecteur de bruit et

$$\tilde{S} = \begin{bmatrix} s_0 & & \star \\ \vdots & \ddots & \\ s_{N-1} & & s_0 \\ \vdots & \ddots & \vdots \\ s_{P-1} & & s_{P-N} \\ & \ddots & \vdots \\ \star & & s_{P-1} \end{bmatrix} \in \mathbb{C}^{(N+P-1)\times N} \tag{11.13}$$

est une matrice dont les colonnes sont des versions décalées du pilote s. Les éléments notés par $\star$ dans $\tilde{S}$ correspondent soit à des zéros lorsque la séquence s est précédée et suivie d'un intervalle de garde de taille $N-1$ (durée pendant laquelle aucune information n'est transmise), soit à des éléments inconnus. Nous nous plaçons dans le deuxième cas et réduisons alors le système à la partie connue (en rouge) de $\tilde{S}$, notée $\tilde{S}_R \in \mathbb{C}^{(P-N+1)\times N}$.

Les futures générations de systèmes de communication (5G) (PIRINEN, 2014) vont demander des fréquences f_S de plus en plus importantes augmentant ainsi considérablement la complexité, le coût et la consommation d'énergie au niveau du récepteur. Afin de s'affranchir de ces inconvénients, une idée consiste à diminuer la fréquence d'échantillonnage au niveau du récepteur tout en assurant la même qualité d'estimation du canal. En considérant un échantillonnage de fréquence $f_R < f_S$ au niveau du récepteur, le signal reçu $y \in \mathbb{C}^M$ (avec $M < P - N + 1$) est maintenant donné par :

$$y = Sh + b, \quad \text{avec} \quad S = U\tilde{S}_R, \tag{11.14}$$

où ici $b \in \mathbb{C}^M$ et $U \in \mathbb{C}^{M\times(P-N+1)}$ est une matrice modélisant l'opération d'échantillonnage au niveau du récepteur. Par exemple, si on a $f_R = f_S/3$, la matrice U s'écrit :

$$U = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{11.15}$$

Cette diminution de la fréquence d'échantillonnage au niveau du récepteur entraine généralement un violation de la condition de Shannon-Nyquist. Cependant, étant donné que $N \gg K$, il est possible de tirer bénéfice de la parcimonie du canal afin de préserver la même résolution pour l'estimation de ce dernier que lorsque Shannon-Nyquist est respecté. Dans la suite, nous traiterons le problème d'estimation parcimonieuse résultant par minimisation de G_{CEL0}.

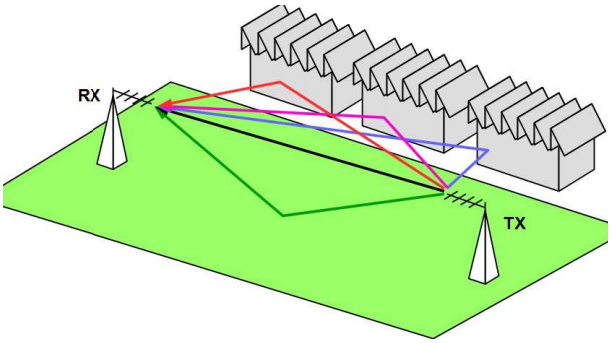


FIGURE 52 – Transmission d'un signal depuis un émetteur (TX) vers un récepteur (RX) avec les différentes réflexions de ce signal sur l'environnement.

11.2.2.2 *Génération des données, algorithmes et critères de performance*

SIMULATION DES DONNÉES On considère un modèle de canal de transmission dit *Extended Vehicular A* (EVA), défini par le 3rd Generation Partnership Project (3GPP) pour les télécommunications (3GPP, Mar. 2010). La table 5 présente les délais considérés dans les simulations qui vont suivre.

Ensuite, un signal pilote de largeur de bande $f_{LB} = 50$ MHz (i. e. $f_{max} = 25$ MHz), suréchantillonné à $f_S = 100$ MHz (afin d'améliorer la résolution de l'estimation des délais qui seront ainsi calculés sur une grille de pas $1/f_S = 10$ ns), est transmis à travers le canal de communication considéré et est échantillonné au niveau du récepteur à une fréquence

Délais (ns)	0	30	150	310	370	710	1090	1730	2510
Puissance relative (dB)	0.0	-1.5	-1.4	-3.6	-0.6	-9.1	-7.0	-12.0	-16.9

TABLE 5 – Modèle de canal EVA utilisé dans les simulations.

$f_R = f_S/3 \approx 33.3$ MHz $< f_{LB}$ (on est donc inférieur à la fréquence de Nyquist). Dans ce cas, la matrice U est définie exactement par (11.15) et en considérant que les délais à estimer ne dépassent pas 6000 ns, on prend $N = 600$. Enfin, M est choisi égal à 200 et la taille du pilote s échantillonné est déterminée en conséquence (i. e. $P = 3M + 1 + N - 1 = 1200$).

En ce qui concerne le bruit, plusieurs valeurs de SNR sont considérées ($[0 : 2.5 : 22.5]$ dB) et 100 générations de bruit sont réalisées pour chaque SNR afin de moyenner les résultats.

ALGORITHMES L'estimation est effectuée par minimisation de G_{CEL0} avec les algorithmes FBS et IRL1 (voir chapitre 10) en comparaison avec l'algorithme IHT (voir section 8.3.1) minimisant directement G_{ℓ_0}. Pour chacun des algorithmes et chaque instance du problème, plusieurs valeurs de λ sont testées de 0.1 à 10 par pas de 0.1. Les algorithmes sont initialisés avec le vecteur nul et limités à 10000 itérations.

CRITÈRES DE PERFORMANCE Les résultats des algorithmes sont évalués en terme d'EQM (plus précisément la racine de l'EQM), définie pour un SNR donné par :

$$\text{REQM} = \sqrt{\sum_{i=1}^{100} \frac{1}{100} \|\hat{h}^i - h^\star\|^2}, \tag{11.16}$$

où $\hat{h}^i \in \mathbb{C}^N$ est le vecteur estimé[3] pour la i-ème réalisation de bruit et $h^\star \in \mathbb{C}^N$ le vecteur simulé. On peut ainsi tracer des courbes montrant l'évolution de la racine de l'EQM en fonction du SNR pour chaque algorithme testé. Nous représentons également la racine de l'EQM de l'estimateur oracle, donnée pour un SNR fixé par (CANDES et TAO, 2007) :

$$\text{REQM}_{\text{Oracle}} := \sqrt{\sigma_n^2 \text{tr}\{(S_{\sigma(h^\star)}^H S_{\sigma(h^\star)})^{-1}\}}, \tag{11.17}$$

où $S_{\sigma(h^\star)}$ est la sous-matrice de S composée des colonnes indexées par les éléments du support $\sigma(h^\star)$ de $h^\star$, σ_n^2 représente la variance du bruit gaussien pour le SNR considéré et tr{} définit l'opérateur trace. Cette racine de l'EQM de l'estimateur oracle fournit une borne inférieure théorique aux résultats numériques obtenus.

Étant donné que l'objectif premier de ce problème concerne l'estimation des retards τ_k $k \in \mathbb{I}_K$ (i. e. le support de $h^\star$), il convient d'utiliser une autre mesure de performance reflétant la capacité de l'algorithme à estimer correctement ce support (ce qui n'est pas le cas du critère REQM). Nous avons donc tout simplement choisi de tracer la probabilité d'estimation exacte des délais en fonction du SNR. Pour ce faire on compte le nombre de fois (sur les 100 simulations à SNR fixé) où l'on a :

$$\sigma(\hat{h}) = \sigma(h^\star), \tag{11.18}$$

(on rappelle que $\sigma(\cdot)$ est la fonction retournant le support de la solution) puis on divise le résultat par le nombre de simulations (100).

3. On ne retient que la solution produisant la plus faible EQM parmi toutes les solutions obtenues avec les différents λ testés.

11.2.2.3 *Résultats numériques*

La figure 53 présente les résultats obtenus en termes d'EQM (gauche) et en termes de probabilité d'estimation exacte du support (droite). On voit clairement le gain apporté par la minimisation de G_{CEL0} en comparaison avec la minimisation directe de G_{ℓ_0}. En effet, les résultats obtenus par minimisation de G_{CEL0} en utilisant FBS et IRL1 sont nettement meilleurs que ceux obtenus avec l'algorithme IHT (i. e. FBS pour minimiser G_{ℓ_0}) que ce soit en termes d'EQM ou en termes d'estimation exacte du support. En particulier, pour des SNR forts ($\geqslant$ 15 dB) les performances des algorithmes minimisant G_{CEL0} tendent vers la référence donnée par l'estimateur oracle. Rappelons que le calcul de l'estimateur oracle nécessite la connaissance du support de la solution ce qui n'est pas envisageable pour des situations réelles. Ici, nous montrons qu'avec un choix du paramètre λ approprié, la méthode proposée est capable d'atteindre cette performance «idéale» de l'estimateur oracle pour des SNR au dessus de 15 dB. Les mêmes conclusions peuvent être tirées des courbes de reconstruction exactes. Au contraire, les deux critères révèlent des performances beaucoup moins bonnes pour IHT qui ne parvient pas à estimer correctement le support et ce quelque soit le SNR. Par ailleurs, on notera encore une fois que l'algorithme IRL1 se comporte mieux que FBS ce qui peux s'expliquer par sa plus grande capacité à «éviter» des minima locaux de la fonctionnelle G_{CEL0}. Pour finir, notons qu'une minimisation ℓ_1 pour ce problème produit des résultats encore moins bons que ceux obtenus avec l'algorithme IHT (voir CHINATTO et al., 2015).

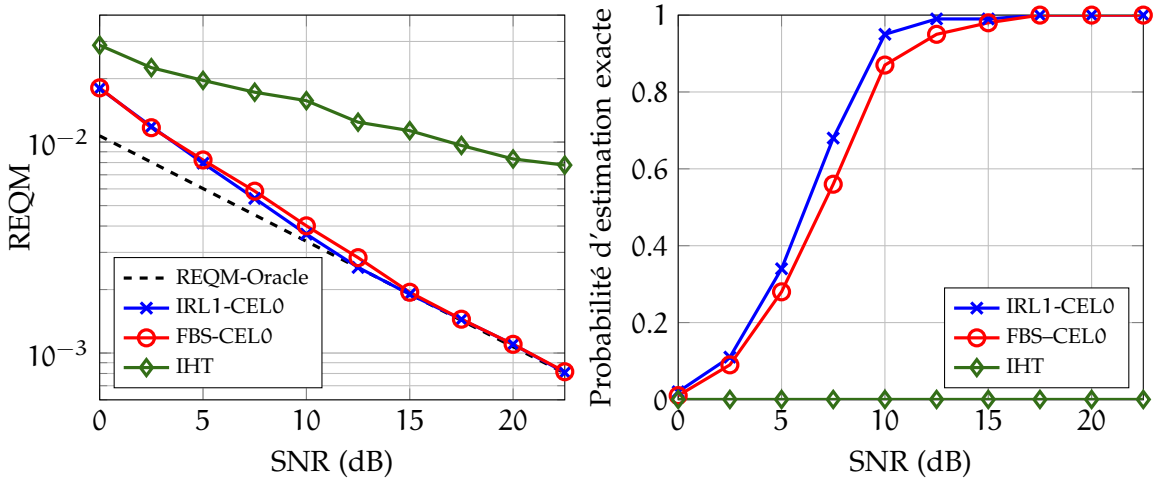


FIGURE 53 – Racine de l'EQM (gauche) et probabilité d'estimation exacte du support de la solution (droite) en fonction du SNR pour les algorithmes FBS et IRL1 minimisant G_{CEL0} et IHT minimisant G_{ℓ_0}. Les courbes sont déterminées en moyennant les résultats obtenus pour 100 réalisations de bruit. Enfin, plusieurs valeurs de λ sont testées pour chaque réalisation de bruit et seule la solution produisant la plus faible REQM est alors utilisée pour le calcul des moyennes.

Enfin, un point essentiel dans ce type de problèmes concerne le choix du paramètre λ. Étant donné que pour chaque simulation nous avons considéré la solution minimisant la REQM parmi les solutions obtenues pour différentes valeurs de λ, nous pouvons tracer l'évolution de la valeur du λ sélectionné (en moyenne) en fonction du SNR. La figure 54 présente les courbes obtenues. On peut voir que plus le SNR est faible, plus il faut donner de l'importance au terme de régularisation. En d'autres termes, plus les données sont bruitées plus l'estimation requiert d'information *a priori*. Par ailleurs, alors que les courbes pour FBS et IRL1 sont similaires, IHT nécessite de prendre des valeurs de λ beaucoup plus grandes. Étant donné que les minimiseurs locaux de G_{ℓ_0} ne dépendent pas de λ (NIKOLOVA,

2013, remarque 5), le fait de devoir considérer des λ plus grands pour IHT est une manière d'éviter de «mauvais» minima locaux.

Les courbes de la figure 54 peuvent également être utilisées afin de choisir la valeur de λ dans le cas de données réelles pour lesquelles une estimation du SNR est disponible.

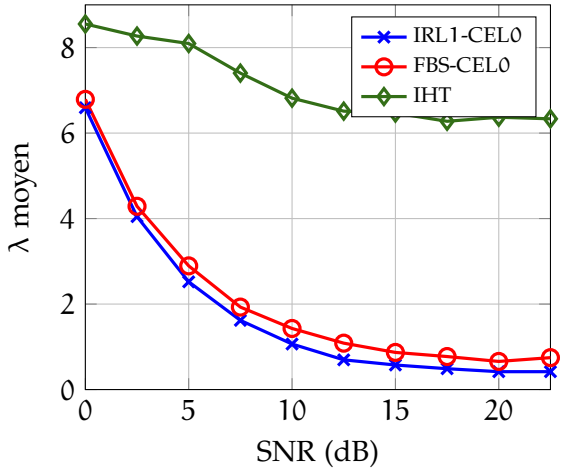


FIGURE 54 – Valeur moyenne (sur les 100 simulations de problèmes) du paramètre λ sélectionné (comme étant celui résultant en la solution minimisant la REQM parmi l'ensemble des valeurs testées) en fonction du SNR pour les algorithmes FBS, IRL1 (minimisation de G_{CEL0}) et IHT (minimisation de G_{ℓ_0}).

11.2.3 *Problème d'estimation des directions d'arrivées*

Un autre problème classique en traitement d'antennes, et plus précisément en *array processing*,[4] concerne l'estimation des directions d'arrivées[5] d'un signal, généré par un ensemble fini de sources, sur un ensemble d'antennes (array) de géométrie connue. Ce problème est rencontré dans de nombreuses applications à commencer par le développement de systèmes radars/sonars ou encore dans le contexte des communications sans fils.

11.2.3.1 *Modélisation*

Dans ce travail, nous considérons les systèmes radio (utilisant des ondes radio) ainsi que le cas d'un ensemble de M antennes omnidirectionnelles[6] équidistantes d'une demi-longueur d'onde (de l'onde radio considérée) le long d'un axe. Ce type de configuration est appelé en anglais Uniform Linear Array (ULA) et nous utiliserons cette notation dans la suite. Soient K signaux incidents s_k ($k \in \mathbb{I}_K$) à bande limitée, arrivant sur le ULA avec des directions d'arrivées $\phi_k \in [-90°, +90°]$ ($k \in \mathbb{I}_K$) calculées depuis la normale à l'axe contenant les antennes comme cela est représenté sur la figure 55. À chaque instant t, le signal reçu par l'antenne $m \in \mathbb{I}_M$ est donné par (TUNCER et FRIEDLANDER, 2009) :

$$y_m(t) = \sum_{k \in \mathbb{I}_K} s_k(t) e^{j\pi(m-1)\sin(\phi_k)} + b_m(t), \tag{11.19}$$

où b_m ($m \in \mathbb{I}_M$) est un bruit additif associé à la m-ème antenne.

4. Thématique de recherche s'intéressant à l'estimation des caractéristiques de signaux à partir de données mesurées par un ensemble d'antennes de géométrie connue.
5. Direction of Arrivals (DoA) en anglais.
6. i. e. qui rayonnent uniformément dans toutes les directions du plan horizontal.

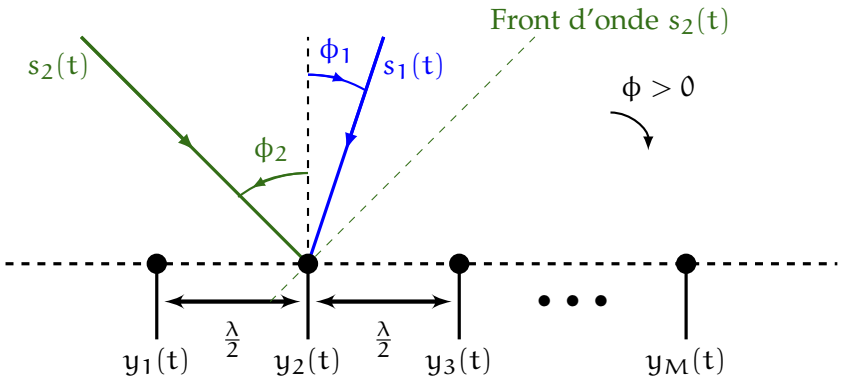


FIGURE 55 – Représentation de deux signaux $s_1(t)$ et $s_2(t)$ arrivant sur un ULA constitué de M antennes omnidirectionnelles espacées d'une demi longueur d'onde.

En notant

$$a_k(\phi_k) = \left[1, e^{j\pi\sin(\phi_k)}, e^{j2\pi\sin(\phi_k)}, \ldots, e^{j\pi(M-1)\sin(\phi_k)}\right]^T, \tag{11.20}$$

le vecteur de direction (*steering vector*) du k-ème signal, le modèle à l'instant t pour l'ensemble des M antennes est donné par :

$$y(t) = A_\phi s(t) + b(t), \tag{11.21}$$

avec les variables $y(t) = [y_1(t), \ldots, y_M(t)]^T \in \mathbb{C}^M$, $b(t) = [b_1(t), \ldots, b_M(t)]^T \in \mathbb{C}^M$, $s(t) = [s_1(t), \ldots, s_K(t)]^T \in \mathbb{C}^K$ et la matrice $A_\phi = [a_1(\phi_1), \ldots, a_K(\phi_K)] \in \mathbb{C}^{M\times K}$. En prenant des mesures pour différents temps t_l, $l \in \mathbb{I}_L$, on a

$$Y = A_\phi S + B, \tag{11.22}$$

où Y, S et B sont maintenant des matrices contenant les vecteurs de (11.21) pour les différents temps t_l : $Y = [y(t_1), \ldots, y(t_L)] \in \mathbb{C}^{M\times L}$, $B = [b(t_1), \ldots, b(t_L)] \in \mathbb{C}^{M\times L}$ et $S = [s(t_1), \ldots, s(t_L)] \in \mathbb{C}^{K\times L}$. Dans la littérature les modèles (11.21) et (11.22) sont souvent nommés respectivement Single Measurement Vector (SMV) et Multiple Measurement Vector (MMV).

L'objectif étant d'estimer les directions d'arrivées ϕ_k ($k \in \mathbb{I}_K$), le modèle (11.22) est donc non linéaire par rapport aux variables d'intérêt. Une idée consiste alors à linéariser le problème en découpant l'ensemble des angles possibles $[-90°, +90°]$ en N intervalles de taille $\Delta_\phi°$ (voir figure 56 gauche). On obtient de la sorte N *steering vectors* candidats a_n (voir équation (11.20)), permettant de définir une nouvelle matrice $A_N \in \mathbb{C}^{M\times N}$. En prenant Δ_ϕ petit, nous avons $N \gg K$ et le problème est alors réduit à l'estimation d'une matrice $S_N \in \mathbb{C}^{N\times L}$ parcimonieuse selon les lignes avec uniquement K lignes non-nulles (voir figure 56 droite). Les positions des lignes non-nulles étant alors caractéristiques des directions d'arrivées. Le modèle devient :

$$Y = A_N S_N + B. \tag{11.23}$$

Nous sommes donc exactement dans le cadre de la parcimonie structurée par ligne introduite dans la section 11.2.1 et nous pouvons donc utiliser l'extension de la pénalité CEL0 associée (11.9).

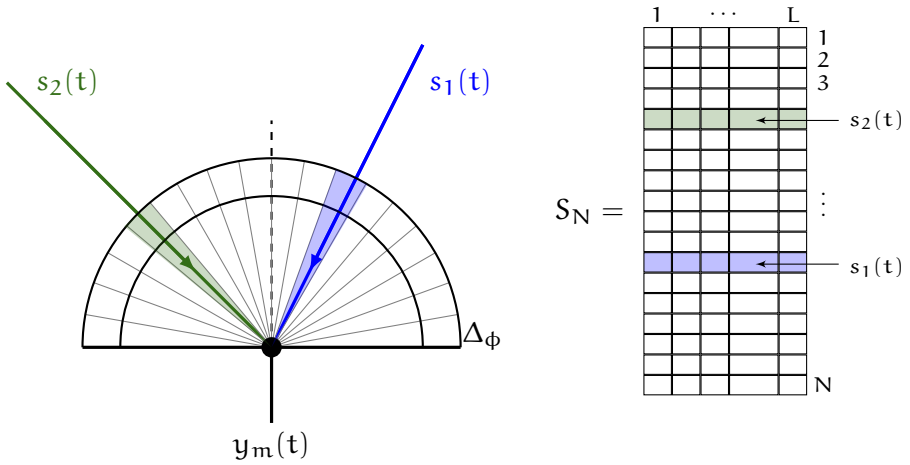


FIGURE 56 – Linéarisation du modèle DoA par discrétisation des directions d'arrivées (gauche) et matrice S_N pour deux signaux incidents $s_1(t)$ et $s_2(t)$ (droite).

11.2.3.2 *Génération des données, algorithmes et critères de performance*

SIMULATIONS DES DONNÉES Nous nous proposons de réaliser des simulations en considérant :

- un ULA constitué de $M = 10$ antennes omnidirectionnelles espacées de la moitié de la longueur d'onde utilisée ;
- deux signaux incidents ($K = 2$) à bande étroite, arrivant sur le ULA respectivement avec les directions $\phi_1 = 0°$ et $\phi_2 = 5°$. La proximité de ces deux directions d'arrivées rend le problème plus difficile et permet de tester les limites des algorithmes ;
- $L = 10$ pas de temps ;

Enfin, comme précédemment, plusieurs valeurs de SNR sont considérées ($[0 : 5 : 40]$ dB) et 100 générations de bruit sont réalisées pour chaque SNR afin de moyenner les résultats.

ALGORITHMES Au vu du problème, nous considérons les versions structurées (selon les lignes) $G_{\ell_{2,0}}$ et $G_{CEL_{20}}$ présentées dans la section 11.2.1. La minimisation de ces fonctionnelles est quant à elle toujours effectuée avec les algorithmes FBS et IRL1 pour $G_{CEL_{20}}$ et IHT (FBS) pour $G_{\ell_{2,0}}$. Notons que les opérateurs proximaux nécessaires pour la mise en œuvre de ces algorithmes sont définis par les mêmes expressions que dans le cas vectoriel réel en remplaçant les valeurs absolues par la norme euclidienne $\|\cdot\|$ et $\text{sign}(x)$ par

$$\begin{cases} \dfrac{x}{\|x\|} & \text{si } x \neq 0, \\ 0 & \text{si } x = 0. \end{cases} \tag{11.24}$$

Afin de déterminer les estimés $\hat{\phi}_1$ et $\hat{\phi}_2$, nous considérons une discrétisation de l'ensemble $[-45°, +45°]$ de pas $\Delta_\phi = 1°$ résultant en une matrice $A_N \in \mathbb{C}^{10\times 91}$. Étant donné que le nombre d'antennes est $M = 10$ et que les deux angles incidents sont espacés de 5°, les sources ne sont pas séparées par beamforming puisque les deux signaux sont dans la région (3 dB) du lobe principal qui est de l'ordre de 10°, ce qui rend le problème d'estimation difficile. Enfin, comme pour le cas de l'estimation de canal, plusieurs valeurs du paramètres λ sont testées de 0.5 à 20 par pas de 0.5 et les algorithmes sont initialisés par une solution nulle et limités à 10000 itérations.

Pour finir, il existe deux algorithmes très répandus pour ce type de problème qui sont les algorithmes MUltiple SIgnal Classification (MUSIC) (SCHMIDT, 1986) et Estimation of Signal Parameters via Rotational Invariance Technique (ESPRIT) (ROY et KAILATH, 1989). Dans la

suite, nous utiliserons l'algorithme MUSIC à titre de comparaison. Notons que ce dernier n'étant pas basé sur une grille discrète d'angles, la solution obtenue est projetée sur la grille utilisée par les autres algorithmes. Plus précisément, à partir des deux angles estimés par MUSIC (il utilise le fait que deux angles sont recherchés), nous construisons le support ω (de taille 2) correspondant sur notre grille d'angles. Ensuite, nous obtenons $\hat{S}_N$ avec (11.26) (voir ci-après) qui est maintenant comparable aux sorties des autres algorithmes considérés.

CRITÈRES DE PERFORMANCES Tout comme pour le problème d'estimation de canal, nous nous intéressons à la racine de l'EQM définie ici, pour un SNR fixé, par :

$$\text{REQM} = \sqrt{\frac{1}{100 \times L} \sum_{i=1}^{100} \|S_N - \hat{S}_N^i\|^2}, \tag{11.25}$$

où S_N est la matrice recherchée composée de deux lignes non nulles dont les positions correspondent aux deux angles ϕ_1 et ϕ_2, et $\hat{S}_N^i$ est la solution estimée lors de la i-ème génération de bruit[7] pour le SNR considéré. Notons que la principale difficulté étant l'estimation du support, la matrice $\hat{S}_N^i$ dans (11.25) est considérée comme étant la solution moindres carrés sur le support estimé. Autrement dit, en notant $\omega \subseteq \{1, \dots, N\}$ le support estimé, nous considérons

$$\hat{S}_N^i = \begin{cases} (A_N)_\omega^\dagger Y & \text{sur } \omega, \\ 0 & \text{sinon,} \end{cases} \tag{11.26}$$

où $(A_N)_\omega^\dagger$ est la pseudo inverse de la restriction de A_N aux colonnes indexées par les éléments de ω.

La racine de l'EQM de l'estimateur oracle, pour un SNR fixé, est donnée ici par :

$$\text{REQM}_{\text{Oracle}} = \sqrt{\sigma_n^2 \text{tr}\{[(A_N)_{\omega^*}^H (A_N)_{\omega^*}]^{-1}\}}, \tag{11.27}$$

où ω^* correspond au support exact sur notre grille d'angles. Les résultats seront donc présentés en comparaison avec les performances de l'oracle.

La racine de l'EQM ne reflétant ni la parcimonie de la solution, ni la capacité de l'algorithme à déterminer le support exact de cette dernière, nous présenterons conjointement aux courbes REQM, les probabilités d'estimation exacte du support.

11.2.3.3 *Résultats numériques*

La figure 57 (gauche) présente l'évolution de la racine de l'EQM en fonction du SNR pour les différents algorithmes considérés. Comme on pouvait s'y attendre, les erreurs augmentent lorsque le SNR diminue. Pour des niveaux de bruit faibles (SNR forts), les performances des algorithmes minimisant la fonctionnelle CEL_{20} ainsi que l'algorithme MUSIC atteignent les performances de l'oracle. Cependant, pour des SNR en dessous de 25 dB, IRL1-CEL_{20} et FBS-CEL_{20} produisent de meilleurs résultats que les deux autres algorithmes. De plus, on retrouve encore le meilleur comportement de IRL1 en comparaison avec FBS pour la minimisation de $G_{\text{CEL}_{20}}$. Notons aussi la très mauvaise performance de IHT qui s'explique par le fait que cet algorithme échoue à trouver une solution parcimonieuse pour la configuration étudiée. En effet, pour l'ensemble des simulations effectuées, les solutions estimées avec IHT ont toujours un support étalé autour des deux positions exactes.

7. Une fois de plus, nous retenons la solution pour laquelle l'erreur est la plus faible parmi les solutions obtenues avec les différents λ considérés.

Étant donné que MUSIC utilise le fait qu'il y a deux signaux incidents à estimer, nous avons introduit cette connaissance dans le calcul de la REQM pour les trois autres algorithmes. Pour ce faire, pour chaque solution déterminée, nous considérons que le support estimé contient seulement les indices des deux lignes de plus grande norme. Ensuite $\hat{S}_N^i$ est calculé avec (11.26) pour ce support de taille 2. Les résultats sont présentés sur la figure 57 (droite).

Alors que cette modification affecte très peu les résultats de IRL1-CEL$_{20}$ et FBS-CEL$_{20}$, les performances de IHT sont considérablement améliorées. Ainsi, en utilisant le même *a priori* que MUSIC, IHT obtient de meilleures performances que ce dernier en régime de faible SNR. Cependant, les méthodes fondées sur la minimisation de la fonctionnelle CEL$_{20}$ restent les plus performantes dans un tel régime de bruit.

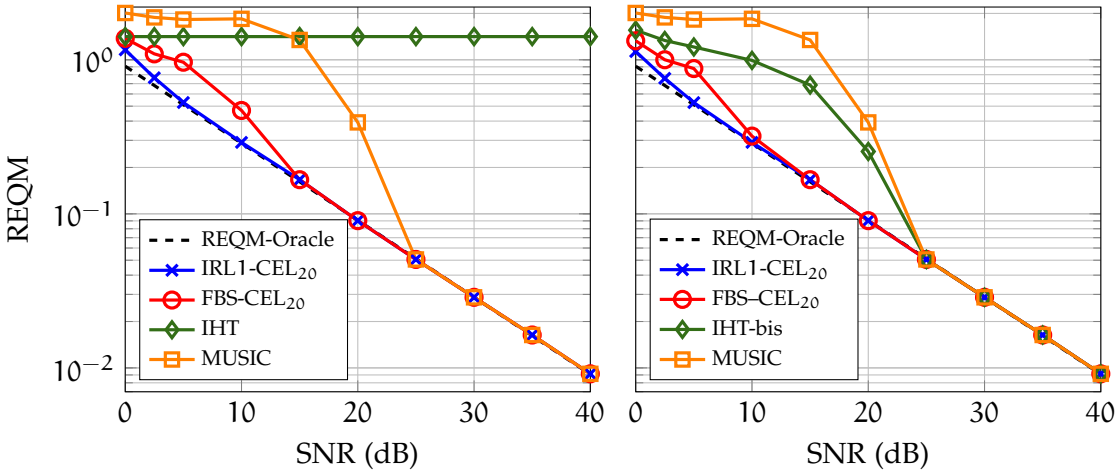


FIGURE 57 – Racine de l'erreur quadratique moyenne en fonction du SNR pour les algorithmes FBS et IRL1 minimisant la fonctionnelle $G_{CEL_{20}}$, IHT minimisant $G_{\ell_{2,0}}$ et l'algorithme MUSIC. Les courbes sont déterminées en moyennant les résultats obtenus pour 100 réalisations de bruit. Enfin, plusieurs valeurs de λ sont testées pour chaque réalisation de bruit et seule la solution produisant la plus faible REQM est alors utilisée pour le calcul des moyennes. La figure de gauche calcule la REQM en considérant le support estimé tout entier alors que celle de droite réduit le support estimé aux deux lignes de plus grande norme.

Enfin, les probabilités d'estimation exacte du support de la solution présentées sur la figure 58 viennent renforcer les observations faites à partir de la racine de l'EQM.

11.3 MICROSCOPIE PALM/STORM ET SUPER-RÉSOLUTION

Ce dernier exemple d'application pour lequel minimiser la fonctionnelle G_{CEL_0} s'avère pertinent concerne la microscopie PALM/STORM. Ce travail préliminaire, et toujours en cours au moment de l'écriture de ce manuscrit, a été effectué avec Simon Gazagnes lors de son stage de master au sein de l'équipe MORPHEME.

11.3.1 *Principe de la microscopie PALM/STORM*

Les techniques PALM (BETZIG et al., 2006) et STORM (RUST et al., 2006) sont fondées sur la localisation de molécules isolées afin de «contourner» la limite de résolution de la microscopie conventionnelle due à la diffraction de la lumière. Le principe repose sur l'utilisation de protéines fluorescentes spécifiques dites *photoactivables*. De telles molécules sont caractérisées par deux états : un état non-fluorescent et un état fluorescent pour lequel elles peuvent alors émettre lorsque qu'elles sont excitées avec une longueur d'onde appropriée.

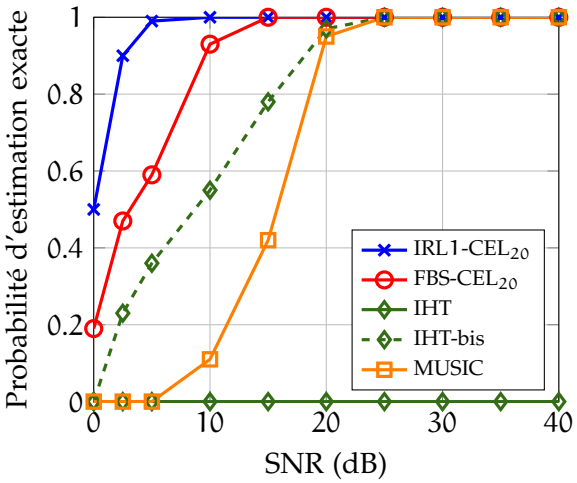


FIGURE 58 – Probabilité d'estimation exacte du support de la solution en fonction du SNR pour les algorithmes FBS et IRL1 minimisant la fonctionnelle $G_{CEL_{20}}$, IHT minimisant $G_{\ell_{2,0}}$ et l'algorithme MUSIC. Les courbes sont déterminées en moyennant les résultats obtenus pour 100 réalisations de bruit. Enfin, plusieurs valeurs de λ sont testées pour chaque réalisation de bruit et seule la solution produisant la plus faible REQM est alors utilisée pour le calcul des moyennes. IHT-bis correspond au cas où l'on restreint le support estimé aux deux lignes de plus grande norme (cf. figure 57 droite).

Afin de pouvoir les imager, il est donc d'abord nécessaire de les activer pour les faire passer dans leur état fluorescent (généralement avec une excitation dans l'ultraviolet) puis ce n'est qu'une fois qu'elles sont dans cet état qu'il est possible de réaliser l'acquisition comme en microscopie classique. Notons qu'ils est aussi courant d'utiliser des molécules *photoconvertibles*, possédant deux états de fluorescence sensibles à des longueurs d'onde différentes. L'idée est donc de limiter le nombre de molécules qui émettent simultanément afin d'être en mesure de les localiser précisément. Le processus d'acquisition, illustré par la figure 59, consiste alors à répéter les trois étapes suivantes :

1. photoactivation de quelques molécules de l'échantillon ;
2. excitation des molécules photoactivées et acquisition d'une image par microscopie classique ;
3. localisation des molécules (isolées) à partir de l'image basse résolution obtenue précédemment.

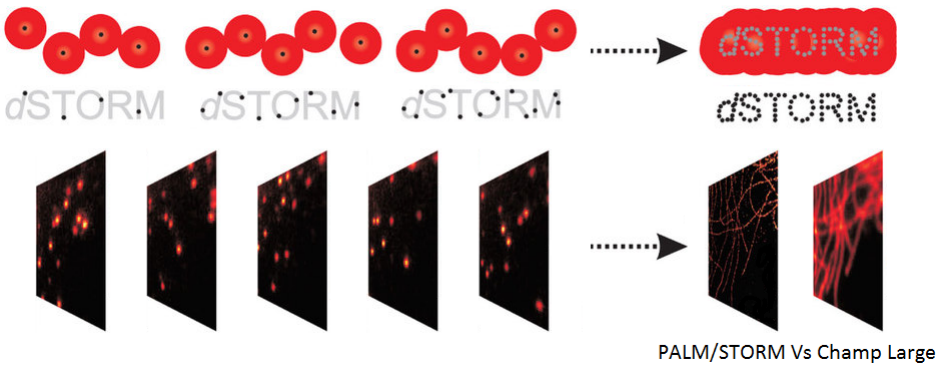


FIGURE 59 – Illustration du principe de la microscopie PALM/STORM. Image extraite et adaptée de (LINDE et al., 2011).

La partie qui nous intéresse ici concerne la localisation des molécules qui peut être formulée comme un problème d'optimisation parcimonieuse que nous présentons dans le

paragraphe suivant. Notons que de nombreuses méthodes ont été proposées pour réaliser une telle localisation et nous renvoyons le lecteur vers l'article de Sage et al. (2015) comparant les performances de nombreux algorithmes proposés ces dernières années.

Enfin, il est important de souligner que cette technique de microscopie nécessite l'acquisition d'un très grand nombre d'images (quelques centaines voire quelques milliers) pour bien représenter l'objet biologique observé. Cela pose clairement un problème en ce qui concerne la résolution temporelle que l'on peut espérer pour des acquisitions *invivo*. En effet, si le déplacement des objets observés est trop rapide par rapport au temps nécessaire pour l'acquisition de toutes les images utilisées pour la localisation, la méthode n'est pas applicable. Cependant, ce temps peut être réduit en activant un nombre plus important de molécules simultanément et ainsi en acquérant un nombre moins important d'images à défaut de rendre le problème de localisation plus difficile avec le recouvrement des «spots» présents sur les acquisitions. Il est alors important de s'intéresser à des méthodes de localisation pouvant être précises lorsque la densité de fluorophores photoactivés est importante.

11.3.2 *Un problème de reconstruction parcimonieuse*

Soit $Y \in \mathbb{R}^{N\times N}$ une image correspondant à une acquisition obtenue suite à l'activation de certaines molécules. Le but étant de localiser le plus précisément possible les molécules à l'origine des «spots» sur l'image Y, nous cherchons la position de ces molécules sur une grille L^2 fois plus fine ($L \in \mathbb{N}$) définie en discrétisant chaque pixel de la grille de Y en $L \times L$ pixels. Notre inconnue est donc une image $X \in \mathbb{R}^{NL\times NL}$ et le modèle linéaire (non-bruité) de formation de l'image Y est donné par :

$$Y = M_L\big(H(X)\big), \tag{11.28}$$

où $H : \mathbb{R}^{NL\times NL} \longrightarrow \mathbb{R}^{NL\times NL}$ représente l'opération de convolution par le noyau H_{ker} et $M_L : \mathbb{R}^{NL\times NL} \longrightarrow \mathbb{R}^{N\times N}$ est l'opérateur moyennant les groupes de pixels $L \times L$ permettant de passer de la grille haute résolution à la grille basse résolution initiale. Plus précisément, pour $X \in \mathbb{R}^{NL\times NL}$,

$$M_L(X) = MXM^T, \tag{11.29}$$

où $M \in \mathbb{R}^{N\times NL}$ est définie par

$$M = \begin{bmatrix} 1 & \cdots & 1 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 & & & & & & & \vdots \\ \vdots & & & & & & \ddots & & & & & & \vdots \\ \vdots & & & & & & & 1 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & 1 & \cdots & 1 \end{bmatrix} \tag{11.30}$$

avec chaque séquence $1 \cdots 1$ de taille L. Étant donné qu'en pratique les données sont dégradées par du bruit,[8] nous considèrerons le problème de moindres carrés pénalisé en norme-ℓ_0 afin d'imposer des solutions parcimonieuses :

$$\hat{X} \in \arg\min_{X\in\mathbb{R}^{NL\times NL}} \frac{1}{2}\|M_L\big(H(X)\big) - Y\|^2 + \lambda\|X\|_0 + i_{\geqslant 0}(X), \tag{11.31}$$

8. Les mêmes types de bruits que ceux présentés dans la section 1.3 page 20 sont présents ici.

avec $\lambda \in \mathbb{R}_+^\star$. Notons que ce problème est à résoudre pour chacune des multiples acquisitions nécessaires à la reconstruction PALM/STORM et la somme de toutes les solutions obtenues nous donnera l'image super-résolue recherchée.

Nous aborderons le problème (11.31) par minimisation de G_{CEL0} (en ajoutant également le terme $i_{\geqslant 0}(X)$, assurant la positivité de la solution, qui ne change en rien les résultats des chapitres précédents) avec l'algorithme IRL1 étant donné que nous avons pu constater sur les applications précédentes qu'il était plus efficace que l'algorithme FBS.

En ce qui concerne la norme des colonnes de l'opérateur $M_L\big(H(\cdot)\big)$, nécessaires à la définition de la pénalité CEL0, nous n'en avons pas une expression exacte. Si il n'y avait que l'opérateur de convolution, alors toutes les colonnes auraient la même norme qui serait égale à la norme du noyau H_{ker}. Cependant, du fait de l'opérateur M_L moyennant les pixels L par L, cela n'est plus vrai ici. Une façon de procéder consiste alors à considérer tous les vecteurs $e_i \in \mathbb{R}^{NL\times NL}$ de la base canonique de $\mathbb{R}^{NL\times NL}$ et à calculer $\|M_L\big(H(e_i)\big)\|$ correspondant à la norme de la i-ème colonne de l'opérateur linéaire. Notons qu'en procédant de la sorte, les valeurs obtenues vont se répéter périodiquement et il n'y a en fait que L^2 coefficients à déterminer.

11.3.3 *Capacité de l'algorithme à séparer trois points sources*

Nous nous sommes tout d'abord intéressés à la capacité de l'algorithme à séparer trois points sources. Pour ce faire, nous avons généré des images de taille 100×100 sur lesquelles nous avons placé trois molécules sur un cercle de diamètre D centré sur l'image. La variation du diamètre du cercle nous permet ainsi de modifier la distance entre les points afin de tester les limites de la méthode. Des exemples de simulations sont présentés sur la ligne du haut de la figure 61.

Ensuite, l'acquisition PALM (i. e. qui en pratique correspondrait à une seule activation de molécules) est réalisée en considérant une PSF gaussienne donnée par :

$$H_{ker}(x,y) = \frac{1}{\sigma_h\sqrt{2\pi}} \exp\left(-\frac{x^2+y^2}{2\sigma_h^2}\right), \tag{11.32}$$

avec $\sigma_h = 15$ px, et en utilisant des grilles d'acquisition de taille 50×50 ($N = 50$ et $L = 2$), 25×25 ($N = 25$ et $L = 4$) et 10×10 ($N = 10$ et $L = 10$). La figure 60 présente la PSF utilisée (sur la grille 100×100) et la figure 61 montre les acquisitions obtenues avec les trois grilles basse résolution considérées.

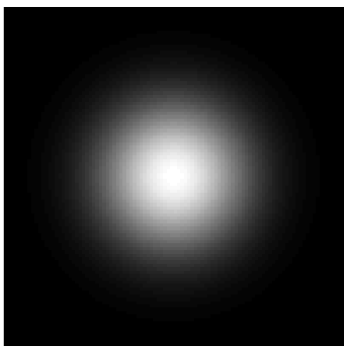


FIGURE 60 – PSF gaussienne $\sigma = 15$ px utilisée pour les expériences menées dans cette section.

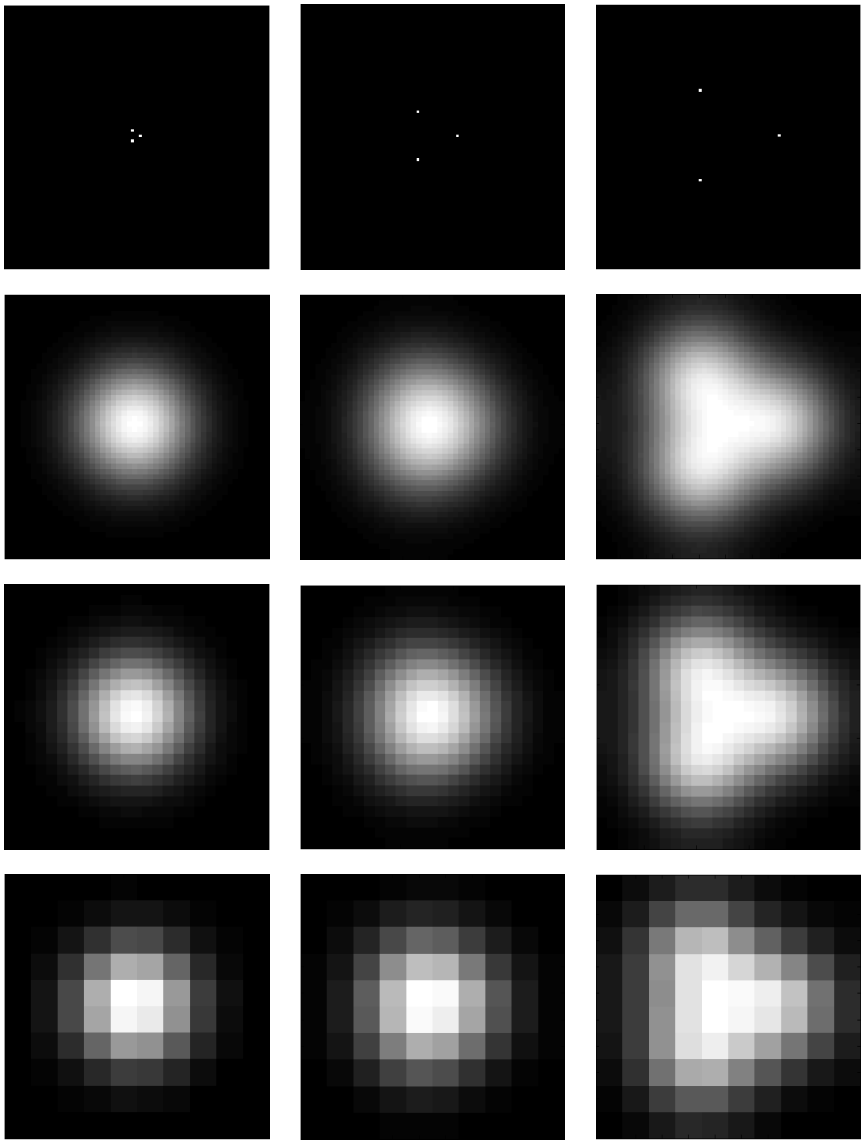


FIGURE 61 – Images simulées avec trois molécules localisées sur un cercle de diamètre 4 px, 20 px et 40 px (ligne du haut de gauche à droite). Les autres images correspondent aux acquisitions associées, utilisant le noyau de convolution de la figure 60 et les grilles grossières de tailles (de haut en bas) 50×50, 25×25 et 10×10.

Les résultats obtenus avec l'algorithme IRL1 sont reportés dans la table 6 pour différents diamètres D, pour les trois grilles basse résolution mentionnées précédemment et enfin pour différentes valeurs du paramètre de régularisation λ. Ces résultats sont présentés en terme de nombre de molécules estimées (i. e. taille du support de $X \in \mathbb{R}^{100\times100}$) et les cas où la position des molécules a été exactement déterminée sont en bleu.

On peut voir qu'au delà d'un diamètre de 12 px (pour lequel l'intersection des «taches» sur les acquisitions est déjà très importante au vu des exemples de la figure 61), il y a toujours des valeurs de λ pour lesquelles on retrouve exactement la position des molécules. Notons par ailleurs qu'il y a généralement une plage de valeurs de λ pour lesquelles nous avons une reconstruction exacte et que cette plage varie peu en fonction du diamètre (donc en fonction de la distance entre les points) et de la taille de la grille basse résolution. Dans ce cas idéal non bruité, il est donc possible d'améliorer la résolution d'un facteur 10 tant que les molécules activées ne sont pas trop proches. Cependant, elles peuvent être tout de même suffisamment proches pour ne pas être distinguables sur les acquisitions.

Taille Acqui	D (px)	$\lambda \rightarrow$	10000	1000	500	100	50	10	1	0.1	0.01
50×50	4		1	1	1	1	1	8	32	39	41
	8		1	1	1	2	2	2	55	83	90
	12		1	3	3	3	3	9	39	53	62
	20		1	3	3	3	3	19	26	31	31
	40		0	3	3	3	3	15	15	15	15
25×25	4		1	1	1	5	10	31	37	41	41
	8		1	1	1	1	1	3	78	103	110
	12		1	3	3	3	3	33	54	69	70
	20		1	3	3	5	19	24	31	31	32
	40		3	3	3	9	15	15	15	15	15
10×10	4		1	1	9	27	31	33	41	41	41
	8		1	1	1	3	7	49	105	117	121
	12		3	3	3	23	37	52	68	73	77
	20		3	3	9	23	24	31	33	33	33
	40		3	5	15	15	15	15	15	17	17

TABLE 6 – Nombre de molécules reconstruites pour les différentes configurations de la figure 61 (différents diamètres D du cercle sur lequel sont générées les molécules et différentes grilles d'acquisition). Les reconstructions sont effectuées sur une grille de taille 100×100. Plusieurs valeurs du paramètre de régularisation λ sont reportées et les reconstructions exactes sont représentées en bleu.

Afin de raffiner les résultats présentés dans la table 6, nous avons procédé à une recherche dichotomique sur λ afin d'obtenir une reconstruction exacte pour les diamètres $D = 4$ px et $D = 8$ px. Les résultats sont présentés dans la table 7. Nous pouvons voir que pour un diamètre $D = 8$px, nous avons toujours trouvé une valeur de λ permettant de reconstruire exactement les molécules simulées. Notons que la recherche dichotomique est stoppée dès qu'une reconstruction exacte a été réalisée et les valeurs de λ présentées dans la table 7 ne sont donc sans doute pas les seules permettant un tel résultat. En ce qui concerne le cas $D = 4$ px, nous n'avons jamais trouvé de λ permettant une localisation exacte et nous atteignons ici les limites de la méthode utilisée.

D (px)	50×50	25×25	10×10
4	X	X	X
8	$\lambda = 3$	$\lambda = 8.5$	$\lambda = 47.5$

TABLE 7 – Recherche dichotomique du paramètre λ pour avoir reconstruction exacte. Les croix rouges signifient que l'intervalle de recherche sur λ a été réduit à une taille de 10^{-5} sans jamais obtenir de reconstruction exacte.

11.3.4 *Résultats numériques dans le cas bruité en fonction de la densité de molécules*

Nous nous intéressons maintenant à l'étude des performances de la méthode pour des données bruitées contenant différentes densités de molécules. Ces dernières sont générées sur une image de taille 512×512 où les pixels sont considérés de taille 25 nm. Cinq densités

de molécules sont utilisées : 1, 2.5, 5, 7,5 et 10 μm^{-2}. Ensuite, nous prenons $\sigma_h = 150$ nm et $L = 4$ produisant ainsi des données de taille 128×128 pour lesquelles la taille du pixel est de 100 nm. Des exemples de telles acquisitions pour les différentes densités considérées sont présentés sur la figure 62. Enfin, un bruit poissonnien est ajouté à ces données afin d'avoir un SNR de 10 dB ou 20 dB avec :

$$\text{SNR} = 10 \log_{10} \left(\frac{\|Y^\star\|^2}{\|Y^\star - Y_b\|^2} \right), \tag{11.33}$$

où $Y^\star$ et Y_b représentent respectivement les données non-bruitées et bruitées.

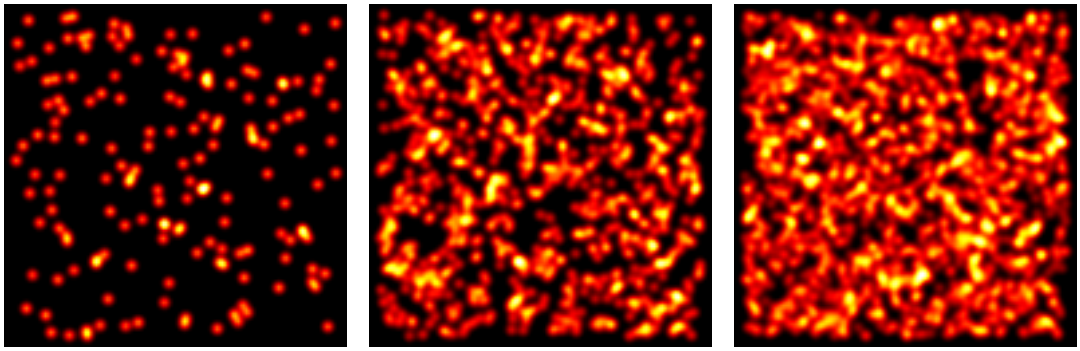


FIGURE 62 – Exemples d'acquisitions 128×128 (taille pixel 100 nm) avec des densités de molécules égales à (de gauche à droite) : 1 μm^{-2}, 5 μm^{-2} et 10 μm^{-2}.

Pour chaque densité et chaque SNR considéré, nous avons généré 20 réalisations de bruit puis minimisé G_{CEL0} avec l'algorithme IRL1 afin de reconstruire les molécules simulées sur la grille haute résolution 512×512. Différentes valeurs du paramètre λ sont utilisées et pour chaque solution ainsi obtenue, nous calculons le nombre de molécules correctement détectées (CD), le nombre de fausses alarmes (FA) et le nombre de non détections (ND). Afin de déterminer ces quantités, un graphe biparti est construit entre les molécules estimées et celles de la vérité de terrain (VT) permettant ainsi de réaliser un appariement entre ces molécules de sorte à minimiser la somme des distances entre les molécules de chaque appariement (SAGE et al., 2015). De plus, une molécule estimée peut être appariée avec une molécule VT uniquement si la distance les séparant est inférieure à une tolérance Δ comme cela est schématisé sur la figure 63. Toutes les molécules appariées sont ainsi classées comme CD, les molécules estimées restantes comme FA et les molécules VT non appariées comme ND. Finalement, les performances seront évaluées en termes d'indice de Jaccard défini par :

$$\text{Jacc}\ (\%) := \frac{\text{CD}}{\text{CD} + \text{FA} + \text{ND}} \times 100. \tag{11.34}$$

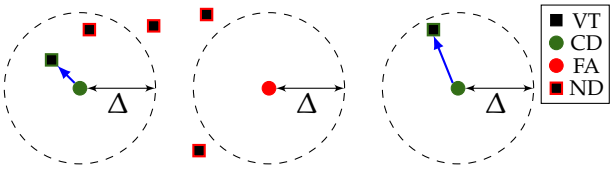


FIGURE 63 – Exemple d'appariements (flèches bleues) entre des molécules de la vérité de terrain (VT) et des molécules estimées. Les disques de tolérance de rayon Δ sont représentés ainsi que les molécules CD, FA et ND.

D'autre part, nous utilisons comme référence l'algorithme DAOSTORM (HOLDEN et al., 2011) qui s'est avéré être bien approprié pour des données haute densité (SAGE et al., 2015). Tout comme l'algorithme MUSIC pour l'estimation des direction d'arrivées, l'algorithme DAOSTORM retourne des positions absolues. Afin d'être en mesure de comparer les résultats, nous avons donc projeté les molécules estimées par DAOSTORM sur la grille utilisée par notre méthode.

La figure 64 présente l'évolution de l'indice de Jaccard (moyenné sur les 20 réalisations de bruit)[9] en fonction de la densité de molécules simulées pour deux SNR différents. Ici le calcul des DC, FA et ND est réalisé à partir d'une tolérance de $\Delta = 100$ nm ce qui correspond à un tiers de la largeur à mi-hauteur de la PSF utilisée ($\approx 2.355\sigma_h = 350.25$ nm). Clairement, on peut observer que l'indice de Jaccard décroît avec la densité de molécules ce qui est en accord avec l'augmentation de la difficulté du problème de localisation. Par ailleurs, nous pouvons voir que la minimisation de CEL0 avec l'algorithme IRL1 fournit de meilleurs résultats que l'algorithme DAOSTORM. De plus, l'écart entre les performances des deux méthodes augmente avec la densité.

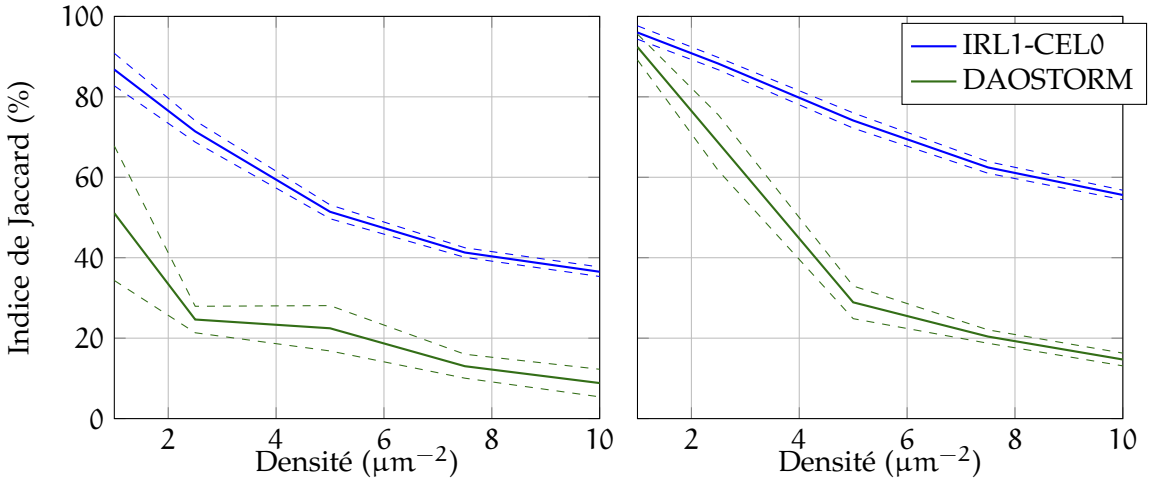


FIGURE 64 – Évolution de l'indice de Jaccard (moyenné sur les 20 réalisations de bruit) en fonction de la densité des molécules simulées. Les courbes discontinues représentent l'écart type. Gauche : SNR 10 dB. Droite : SNR 20 dB.

11.3.5 *Un exemple sur des données réelles*

Pour terminer, nous donnons un exemple sur des données réelles qui ont été proposées pour le challenge ISBI 2013 dédié aux méthodes de localisation en microscopie de super-résolution[10]. Quelques acquisitions, parmi l'ensemble des 500 images acquises pour cet échantillon de microtubules, sont présentées sur la figure 65. Remarquons que la densité de molécules est très élevée. Ces images ont une taille de 128×128 et la taille du pixel est de 100 nm. La PSF est considérée gaussienne avec $\sigma_h = 151$ nm d'après la mesure expérimentale réalisée par CHAHID (2014).

Les résultats, pour $L = 4$ (donc la taille de l'image haute résolution est 512×512 avec une taille de pixel de 25 nm), sont présentés sur la figure 66. On peut apprécier le gain en résolution obtenu et notamment la meilleure séparation des différentes fibres produite par

9. Pour chaque SNR et chaque réalisation de bruit, nous retenons l'estimation ayant le meilleur indice de Jaccard parmi les solutions obtenues avec les différentes valeurs des paramètres testées.

10. http://bigwww.epfl.ch/smlm/challenge2013/index.html

IRL1-CEL0 en comparaison avec DAOSTORM. Notons également qu'un nombre important de molécules ont été reconstruites permettant de bien représenter les structures sur l'image haute résolution, ce qui n'est pas toujours le cas avec des acquisitions ayant une telle densité de molécules (voir par exemple les résultats de certains algorithmes sur ce même échantillon qui sont présentés dans (CHAHID, 2014)).

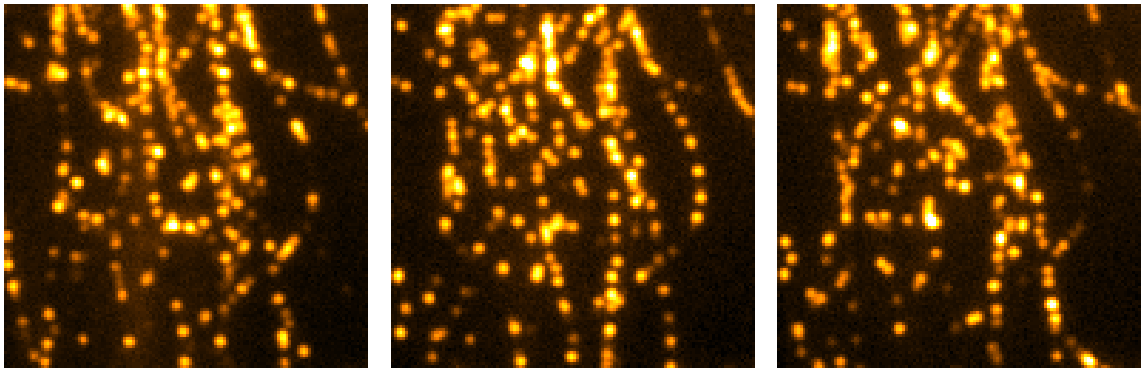


FIGURE 65 – Quelques acquisitions réelles parmi les 500 images acquises pour cet échantillon. La densité de molécules activées sur ces acquisitions est forte.

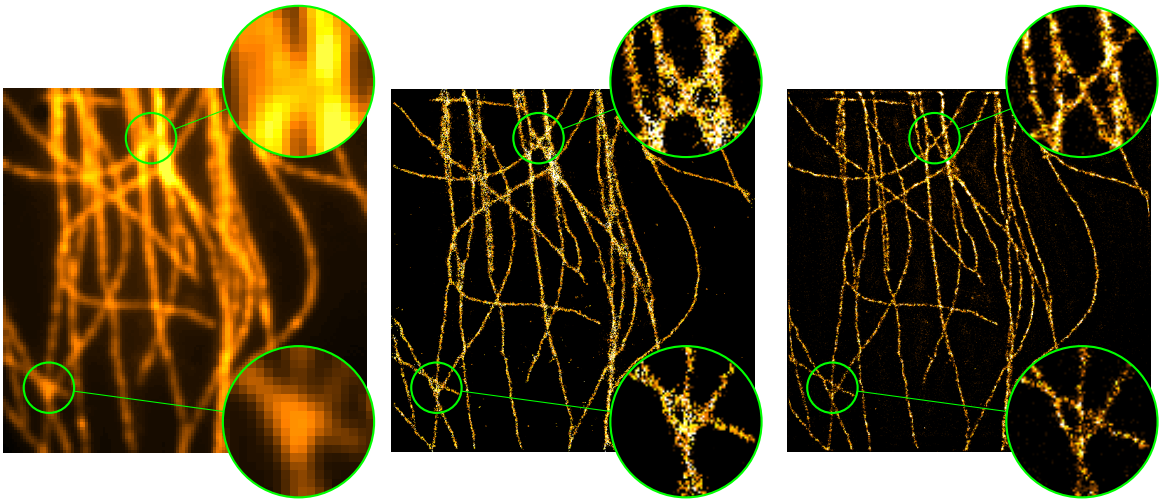


FIGURE 66 – Somme des 500 acquisitions PALM dont certaines sont présentées sur la figure 65 (gauche). Image haute résolution reconstruite par DAOSTORM (centre) et par IRL1-CEL0 (droite). Les images haute résolution sont reconstruites pour $L = 4$.

11.4 CONCLUSION

Dans ce chapitre, nous avons présenté plusieurs applications en traitement du signal et des images pour lesquelles la minimisation de G_{CEL0}, avec des algorithmes de l'état de l'art mais aussi avec la stratégie GNC ainsi que la méthode calculant un «chemin de régularisation» présentées dans le chapitre 10, est pertinente. Pour tous ces problèmes, la solution est directement parcimonieuse et nous n'en cherchons pas une représentation parcimonieuse dans une certaine base. Ainsi, imposer une parcimonie «stricte» avec la norme-ℓ_0 est complètement justifié. Par ailleurs, nous avons également vu avec le problème DoA que les propriétés obtenues pour la pénalité CEL0 dans les chapitres précédents sont également valables pour une parcimonie structurée par ligne.

Les résultats obtenus pour ces différentes applications se sont révélés être plus précis (ou au moins aussi bons dans certains cas), en terme de localisation du support, que d'autres algorithmes de la littérature proposés pour de tels problèmes. Tous ces résultats sont donc très encourageants et motivants pour la conception ou l'amélioration d'algorithmes permettant de mieux minimiser $G_{\texttt{CEL0}}$.

Enfin, il est à noter que si nous avons utilisé l'algorithme GNCeloRP uniquement pour la déconvolution de trains d'impulsions, cela est purement chronologique, le développement de la méthode GNCeloRP ayant été effectuée en fin de thèse. Nous prévoyons par la suite de tester cette méthode sur les autres problèmes considérés dans ce chapitre.

12

PÉNALITÉS EXACTES POUR LE PROBLÈME ℓ_2-ℓ_0 : UNE VUE UNIFIÉE

Sommaire

Dans la lignée des résultats du chapitre 9 mais aussi par rapport aux idées développées dans (Le Thi et al., 2015), il semble tout à fait pertinent de s'intéresser à une vue unifiée des approximations continues de la norme-ℓ_0 dans le contexte des relaxations continues exactes de G_{ℓ_0} dont nous rappelons l'expression

$$G_{\ell_0}(x) = \frac{1}{2}\|Ax - d\|^2 + \lambda\|x\|_0, \tag{12.1}$$

où $A \in \mathbb{R}^{M\times N}$, $d \in \mathbb{R}^M$ et $\lambda \in \mathbb{R}_+^\star$. Dans la suite, nous considérons des approximations continues de la norme-ℓ_0 de la forme :

$$\Phi(x) = \sum_{i=1}^{N} \phi_i(x_i), \tag{12.2}$$

où les ϕ_i sont des pénalités continues 1D approchant $\lambda|\cdot|_0$. Nous obtenons ainsi des relaxations continues de G_{ℓ_0} définies par

$$\tilde{G}(x) = \frac{1}{2}\|Ax - d\|^2 + \Phi(x). \tag{12.3}$$

Considérons maintenant une matrice (de manière équivalente un opérateur linéaire) $A \in \mathbb{R}^{M\times N}$. L'objectif de ce chapitre concerne alors la détermination de conditions *nécessaires et suffisantes* sur ϕ_i (conditions qui peuvent dépendre des éléments de A mais ne requièrent aucune hypothèse sur A) assurant les deux propriétés suivantes pour tout $d \in \mathbb{R}^M$:

$$\arg\min_{x\in\mathbb{R}^N} \tilde{G}(x) = \arg\min_{x\in\mathbb{R}^N} G_{\ell_0}(x), \tag{P1}$$

$$\hat{x} \text{ minimiseur (local) de } \tilde{G} \implies \hat{x} \text{ minimiseur (local) de } G_{\ell_0}. \tag{P2}$$

En d'autres termes, nous souhaitons déterminer une classe de relaxations continues de G_{ℓ_0}, préservant tous ses minimiseurs globaux et pour lesquelles tout minimiseur local

est aussi un minimiseur local de la fonctionnelle initiale G_{ℓ_0}. Notons que d'après (P2), $\tilde{G}$ peut potentiellement éliminer des minimiseurs locaux (non-globaux) de G_{ℓ_0} ce qui est une propriété intéressante pour une telle fonctionnelle non-convexe. D'autre part, les approximations 1D ϕ_i dépendent de λ, peuvent dépendre de A mais *ne doivent pas* dépendre de d. En effet, l'approximation recherchée peut dépendre du problème (défini par la matrice A) mais doit rester indépendant des données dans le but de pouvoir utiliser la même relaxation pour toutes les données acquises par un même système.

Ce travail a été soumis à SIAM Journal on Optimization (SIOPT) (Soubies et al., 2016b).

12.1 ÉTUDE UNIDIMENSIONNELLE

Nous commençons encore une fois avec une étude en dimension 1 où nous considérons le problème

$$\hat{u} \in \arg\min_{u\in\mathbb{R}} g_0(u) := \frac{1}{2}(au-d)^2 + \lambda|u|_0, \tag{12.4}$$

avec $a \in \mathbb{R}_+^\star$, $\lambda \in \mathbb{R}_+^\star$ et $d \in \mathbb{R}$. Nous sommes intéressés par une relaxation *continue* de $\lambda|\cdot|_0$, notée ϕ, associée à la relaxation de g_0 suivante :

$$\tilde{g}(u) := \frac{1}{2}(au-d)^2 + \phi(u). \tag{12.5}$$

L'objectif est alors de dériver des conditions *nécessaires et suffisantes* sur cette pénalité ϕ assurant la validité des propriétés (P1) et (P2) pour $\tilde{g}$. Afin d'éviter de définir ϕ à une constante additive près, nous considérons la condition supplémentaire (naturelle) suivante :

$$\min_{u\in\mathbb{R}} g_0(u) = \min_{u\in\mathbb{R}} \tilde{g}(u). \tag{12.6}$$

D'autre part, nous considérons uniquement les pénalités ϕ indépendantes de d (i. e. les conditions que nous recherchons ne doivent pas dépendre de d) afin d'être en mesure d'étendre les résultats obtenus en dimension 1 au cas d'une dimension quelconque N mais aussi, comme nous l'avons mentionné précédemment, dans le but de pouvoir utiliser l'approximation définie quelque soit les observations pour un problème donné (i. e. pour a donné).

Par ailleurs, tout au long du chapitre, nous considèrerons les deux hypothèses suivantes sur ϕ.

Hypothèse 12.1. ϕ est deux fois continument différentiable (C^2) sur $\mathbb{R}$ sauf en un nombre fini de points de $\mathbb{R}$ appartenant à un ensemble que l'on notera B. De plus, pour tout $v \in B$, les limites

$$\lim_{\substack{u\to v\\ u<v}} \phi'(u) \quad \text{et} \quad \lim_{\substack{u\to v\\ u>v}} \phi'(u) \tag{12.7}$$

existent et sont différentes (i. e. ϕ n'est pas différentiable sur B).

Hypothèse 12.2. ϕ est localement Lipschitz, c'est à dire $\forall u \in \mathbb{R}$,

$$\exists \varepsilon > 0,\ \forall (v,v') \in]u-\varepsilon, u+\varepsilon[^2,\ |\phi(v)-\phi(v')| \leqslant K_u|v-v'|, \tag{12.8}$$

pour K_u positif.

Tout d'abord, nous rappelons la caractérisation des minimiseurs globaux de g_0 déjà utilisée dans le chapitre 9.

Proposition 12.3 (Minimiseurs globaux de g_0)**.** *Soient* $a \in \mathbb{R}_+^\star$, $\lambda \in \mathbb{R}_+^\star$ *et* $d \in \mathbb{R}$. *Alors, un minimiseur global,* $u^\star \in \mathbb{R}$, *de* g_0 *vérifie*

$$u^\star = \begin{cases} 0 & \text{ssi } |d| \leqslant \sqrt{2\lambda}, \\ \frac{d}{a} & \text{ssi } |d| \geqslant \sqrt{2\lambda}. \end{cases} \tag{12.9}$$

Démonstration. Il est clair que g_0 admet toujours deux minimiseurs (locaux) $u_1 = 0$ et $u_2 = \frac{d}{a}$. Alors, en remarquant que

$$g_0(0) = \frac{d^2}{2} \quad \text{and} \quad g_0\left(\frac{d}{a}\right) = \lambda, \tag{12.10}$$

la preuve est terminée. □

Ensuite, dans le but d'assurer la préservation des minimiseurs globaux de g_0 par $\tilde{g}$, il est nécessaire dans un premier temps de caractériser les points critiques de $\tilde{g}$ (i. e. les points $\hat{u} \in \mathbb{R}$ tels que $0 \in \partial\tilde{g}(\hat{u})$ où $\partial\tilde{g}$ définit le gradient généralisé de $\tilde{g}$ (Clarke, 1990)) et ensuite d'assurer que les minimiseurs globaux de g_0 sont des points critiques de $\tilde{g}$.

Nous rappelons tout d'abord la définition de la dérivée directionnelle généralisée de Clarke pour ϕ au point $u \in \mathbb{R}$ dans la direction $w \in \mathbb{R}$:

$$\phi^\circ(u; w) := \limsup_{\substack{v \to u \\ t \downarrow 0}} \frac{\phi(v + tw) - \phi(v)}{t}, \tag{12.11}$$

à partir de laquelle nous pouvons définir le gradient généralisé de Clarke (1990) :

$$\partial\phi(u) := \{\xi \in \mathbb{R} : \phi^\circ(u; w) \geqslant \xi w, \ \forall w \in \mathbb{R}\}. \tag{12.12}$$

Ces définitions sont données sous l'hypothèse 12.2. La proposition suivante caractérise les points critiques de $\tilde{g}$.

Proposition 12.4 (Points critiques de $\tilde{g}$)**.** *Soient* $a \in \mathbb{R}_+^\star$, $\lambda \in \mathbb{R}_+^\star$ *et* $d \in \mathbb{R}$. *Alors,* $\hat{u} \in \mathbb{R}$ *est un point critique de* $\tilde{g}$ *(i. e.* $0 \in \partial\tilde{g}(\hat{u})$*) si et seulement si*

$$\begin{cases} ad - a^2\hat{u} \in \left[\underline{\delta}^{\hat{u}}, \bar{\delta}^{\hat{u}}\right] & \text{ssi } \hat{u} \in B, \\ a^2\hat{u} - ad + \phi'(\hat{u}) = 0 & \text{ssi } \hat{u} \in \mathbb{R} \setminus B, \end{cases} \tag{12.13}$$

où

$$\forall v \in B, \ \underline{\delta}^v = \min\{l_v^-, l_v^+\} \ \text{ et } \ \bar{\delta}^v = \max\{l_v^-, l_v^+\}, \tag{12.14}$$

avec

$$l_v^- = \lim_{\substack{u \to v \\ u < v}} \phi'(u) \ \text{ et } \ l_v^+ = \lim_{\substack{u \to v \\ u > v}} \phi'(u). \tag{12.15}$$

Démonstration. Considérons le cas $u \in \mathbb{R} \setminus B$. D'après (Clarke, 1990, corollaire de la proposition 2.1.2), $\partial\phi(u)$ est réduit au singleton $\{\phi'(u)\}$ si et seulement si ϕ est continument différentiable dans un voisinage de u. Ainsi, $\partial\phi(u) = \{\phi'(u)\}$ pour tout $u \in \mathbb{R} \setminus B$ par l'hypothèse 12.1.

Considérons maintenant le cas $u \in B$. Alors nous avons d'après (Clarke, 1990, corolaire du théorème 2.5.1) que

$$\phi^\circ(u; w) = \limsup_{v \to u} \{\phi'(v)w : v \notin S \cup B\}, \tag{12.16}$$

où S est un ensemble de mesure nulle. Cette égalité avec la définition donnée en (12.12) entraîne

$$\begin{aligned}
\xi \in \partial\phi(u) &\iff \forall w \in \mathbb{R},\ \limsup_{v \to u} \{\phi'(v)w : v \notin S \cup B\} \geqslant \xi w \\
&\iff \begin{cases} \forall w \in \mathbb{R}_+,\ w \times \limsup_{v \to u} \{\phi'(v) : v \notin S \cup B\} \geqslant \xi w \\ \forall w \in \mathbb{R}_-,\ w \times \liminf_{v \to u} \{\phi'(v) : v \notin S \cup B\} \geqslant \xi w \end{cases} \\
&\iff \liminf_{v \to u} \{\phi'(v) : v \notin S \cup B\} \leqslant \xi \leqslant \limsup_{v \to u} \{\phi'(v) : v \notin S \cup B\} \\
&\iff \min\{l_v^-, l_v^+\} \leqslant \xi \leqslant \max\{l_v^-, l_v^+\},
\end{aligned}$$

pour l_v^- et l_v^+ donnés par (12.15) (notons que ces limites existent par l'hypothèse 12.1). Ainsi, d'après ce qui précède, nous avons

$$\partial\phi(u) = \begin{cases} [\underline{\delta}^u, \bar{\delta}^u] & \text{si } u \in B, \\ \{\phi'(u)\} & \text{si } u \in \mathbb{R} \setminus B. \end{cases} \tag{12.17}$$

Finalement, la différentiabilité du terme quadratique de (12.5) entraine

$$\partial\tilde{g}(u) = a(au - d) + \partial\phi(u), \tag{12.18}$$

ce qui termine la preuve. □

D'après les propositions 12.3 et 12.4, nous sommes maintenant en mesure de définir des conditions sur ϕ nécessaires et suffisantes pour que les minimiseurs globaux $u^\star \in \{0, \frac{d}{a}\}$ de g_0 soient des points critiques de $\tilde{g}$.

Lemme 12.5 (Les minimiseurs globaux de g_0 sont des points critiques de $\tilde{g}$)**.** *Soient $a \in \mathbb{R}_+^\star$ et $\lambda \in \mathbb{R}_+^\star$, alors les minimiseurs globaux de g_0 sont des points critiques de $\tilde{g}$ pour tout $d \in \mathbb{R}$ si et seulement si ϕ vérifie les deux conditions suivantes :*

$$0 \in B \ \text{ et } \ \underline{\delta}^0 \leqslant -\sqrt{2\lambda}a \ \text{ et } \ \bar{\delta}^0 \geqslant \sqrt{2\lambda}a, \tag{12.19a}$$

$$B \subset [-\sqrt{2\lambda}/a, \sqrt{2\lambda}/a] \ \text{ et } \ \forall u \in \mathbb{R} \setminus [-\sqrt{2\lambda}/a, \sqrt{2\lambda}/a],\ \phi'(u) = 0, \tag{12.19b}$$

Démonstration. La preuve est détaillée en annexe A.3.1 page 200. □

Notons que lorsque $d = \pm\sqrt{2\lambda}$, 0 et $\frac{d}{a}$ sont tous deux des minimiseurs globaux de g_0 et, sous les conditions (12.19), sont aussi des points critiques de $\tilde{g}$. La figure 67 illustre les conditions (12.19).

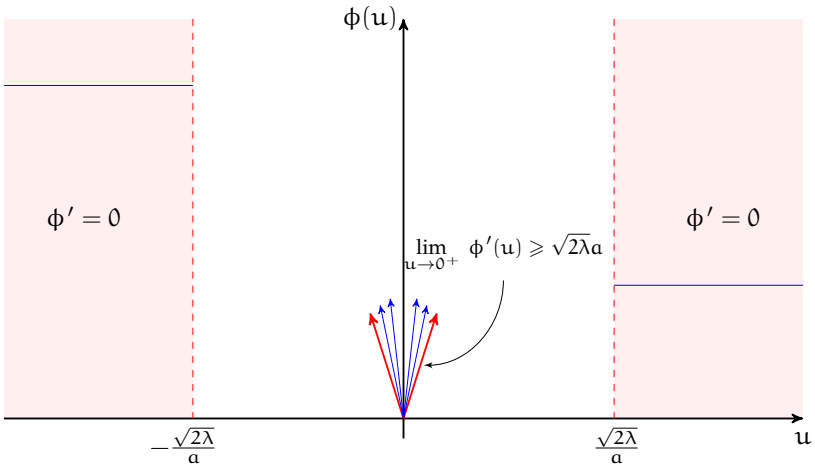


FIGURE 67 – Illustration des conditions sur ϕ données par le lemme 12.5. Les flèches représentent la condition (12.19a) (en rouge) avec des demi-tangentes admissibles pour ϕ en 0 (en bleu). La zone rose sur laquelle ϕ doit être constante illustre la condition (12.19b).

Lemme 12.6. *Soient* $a \in \mathbb{R}_+^\star$ *et* $\lambda \in \mathbb{R}_+^\star$*, alors* $\tilde{g}$ n'admet aucun *minimiseur global sur* $]-\frac{\sqrt{2\lambda}}{a}, 0[\cup]0, \frac{\sqrt{2\lambda}}{a}[$ *pour tout* $d \in \mathbb{R}$ si et seulement si ϕ *vérifie :*

$$\forall u \in]-\sqrt{2\lambda}/a, 0[\cup]0, \sqrt{2\lambda}/a[,\ \phi(u) > \phi_{\text{CEL0}}(a, \lambda; u), \tag{12.20}$$

où ϕ_{CEL0} *est la pénalité* CEL0 *(1D) donnée en* (9.5) *(page 96).*

Démonstration. La preuve est donnée en annexe A.3.2 page 202. □

À partir du lemme 12.6, nous sommes en mesure de démontrer le théorème suivant donnant des conditions *nécessaires et suffisantes* sur ϕ afin d'avoir la propriété (P1) pour $\tilde{g}$.

Theorème 12.7 (Conditions nécessaires et suffisantes pour (P1))**.** *Soient* $a \in \mathbb{R}_+^\star$ *et* $\lambda \in \mathbb{R}_+^\star$*, alors* $\tilde{g}$ *a la propriété* (P1) *(et* (12.6)*) pour tout* $d \in \mathbb{R}$ si et seulement si ϕ *vérifie les trois conditions suivantes :*

$$\phi(0) = 0, \tag{12.21a}$$
$$\forall u \in \mathbb{R}\backslash]-\sqrt{2\lambda}/a, \sqrt{2\lambda}/a[,\ \phi(u) = \lambda |u|_0 = \lambda, \tag{12.21b}$$
$$\forall u \in]-\sqrt{2\lambda}/a, \sqrt{2\lambda}/a[\backslash\{0\},\ \phi(u) > \phi_{\text{CEL0}}(a, \lambda; u), \tag{12.21c}$$

Démonstration. La preuve est présentée en annexe A.3.3 page 202. □

Remarque 12.8. Il est à noter que (12.21c) impose à la pénalité ϕ d'être singulière à l'origine. Nous retrouvons une fois de plus le fait que pour obtenir des solutions parcimonieuses, la pénalité doit être non-différentiable en 0. De plus, la condition (12.21b) impose à la pénalité d'être constante pour $|u|$ grand. Cette propriété était connue pour être une condition permettant d'obtenir des solutions non-biaisées (FAN et LI, 2001).

Une illustration des conditions (12.21) est présentée sur la figure 68. Nous pouvons voir qu'à l'intérieur de la zone grise, la pénalité peut être complètement arbitraire dès lors qu'elle est continue, égale à 0 à l'origine et égale à λ en $\pm\sqrt{2\lambda}/a$. Enfin, pour $|u| \geqslant \sqrt{2\lambda}/a$,

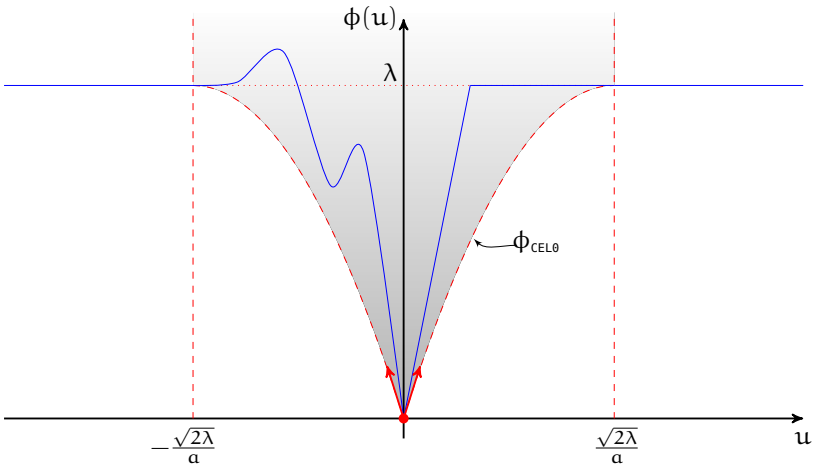


FIGURE 68 – Illustration des conditions données par le théorème 12.7. Les courbes rouges avec le point rouge en 0 représentent les conditions (12.21). La zone grise représente la région du plan admissible pour le graphe de ϕ. Un exemple de pénalité vérifiant les conditions (12.21) est tracé en bleu.

la pénalité doit être constante égale à λ.

Maintenant que nous avons des conditions nécessaires et suffisantes sous lesquelles la relaxation $\tilde{g}$ préserve tous les minimiseurs globaux de la fonctionnelle initiale g_0 (i. e. elle vérifie (P1)), nous pouvons nous intéresser au cas des minimiseurs locaux (non-globaux). Le théorème suivant établit des conditions *nécessaires et suffisantes* sur ϕ afin que les deux propriétés (P1) and (P2) soient vérifiées pour la relaxation $\tilde{g}$.

Theorème 12.9 (Conditions nécessaires et suffisantes pour (P1) et (P2))**.** *Soient* $a \in \mathbb{R}_+^\star$ *et* $\lambda \in \mathbb{R}_+^\star$*, alors* $\tilde{g}$ *a les deux propriétés* (P1) *et* (P2) *pour tout* $d \in \mathbb{R}$ si et seulement si*, en plus des conditions* (12.21) *données par le théorème* 12.7*,* ϕ *vérifie des deux conditions suivantes :*

$$\forall u \in B \setminus \{0\},\ \lim_{\substack{v \to u \\ v<u}} \phi'(v) > \lim_{\substack{v \to u \\ v>u}} \phi'(v), \tag{12.22a}$$

$$\forall u \in]\beta^-, \beta^+[\setminus B,\ \begin{cases} \phi''(u) \leqslant -a^2 \text{ et il existe un intervalle } \mathcal{V} \subset \mathbb{R} \setminus B \\ \text{t.q. } u \in \mathcal{V} \text{ et } \forall v \in \mathcal{V} \setminus \{u\},\ \phi''(v) < -a^2 \end{cases} \tag{12.22b}$$

pour $\beta^- \in \left[-\sqrt{2\lambda}/a, 0\right[$ *(resp.* $\beta^+ \in \left]0, \sqrt{2\lambda}/a\right]$*) défini comme le plus grand (resp. le plus petit) réel pour lequel* ϕ *est constante sur l'intervalle* $]-\infty, \beta^-]$ *(resp.* $[\beta^+, +\infty[$*). Avec cette définition on a* $B \subset [\beta^-, \beta^+]$*.*

Démonstration. La preuve est donnée en annexe A.3.4 (page 203). □

Remarque 12.10. Pour la pénalité CEL0 (9.5), la condition (12.22b) n'est pas vérifiée. En effet, nous avons,

$$\forall\, 0 < |u| < \frac{\sqrt{2\lambda}}{a},\ \phi''_{\text{CEL0}}(a, \lambda; u) = -a^2. \tag{12.23}$$

Cependant, dans ce cas, lorsque $\phi'_{\text{CEL0}}(a, \lambda; u) = ad - a^2 u \Leftrightarrow |ad| = \sqrt{2\lambda} a$ (par définition de CEL0), tous les points de l'intervalle $[0, \frac{\sqrt{2\lambda}}{a}]$ (resp. $[-\frac{\sqrt{2\lambda}}{a}, 0]$ suivant le signe de la quantité ad) sont des minimiseurs de $\tilde{g}$ et on peut aisément en déduire un minimiseur de g_0 par un

simple seuillage (théorème 9.21 page 104). De plus, la condition (12.21c) n'est également pas vérifiée par ϕ_{CEL0}. Au final, CEL0 peut être vu comme la limite inférieure de la classe de pénalités résultant des conditions données par les théorèmes 12.7 et 12.9.

Remarque 12.11. Sous les conditions des théorèmes 12.7 et 12.9, ϕ est strictement concave décroissante sur $[\beta^-, 0]$ (resp. strictement concave croissante sur $[0, \beta^+]$). Par conséquent, d'après (12.21c), si $\beta^- = -\frac{\sqrt{2\lambda}}{a}$ (resp. $\beta^+ = \frac{\sqrt{2\lambda}}{a}$) alors nécessairement $\beta^- \notin B$ (resp. $\beta^+ \notin B$) et $\phi'(\beta^-) = 0$ (resp. $\phi'(\beta^+) = 0$). Voir la figure 69 pour une illustration.

Finalement, les théorèmes 12.7 et 12.9 nous donnent des conditions *nécessaires et suffisantes* sur la pénalité ϕ afin que la relaxation continue $\tilde{g}$ associée vérifie les propriétés (P1) et (P2). Ces conditions sont illustrées sur la figure 69.

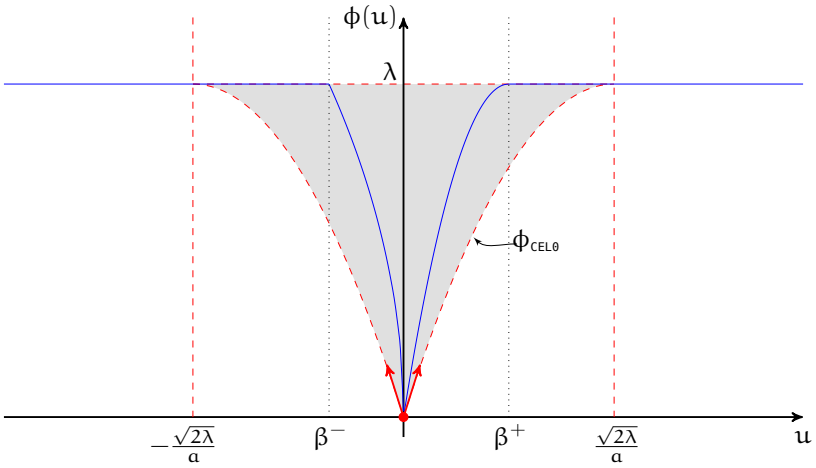


FIGURE 69 – Illustrations de l'ensemble des conditions établies par les théorèmes 12.7 et 12.9 qui sont nécessaires et suffisantes pour avoir (P1) et (P2). Les courbes rouges et le point rouge en 0 représentent les conditions alors que les courbes bleu correspondent à un exemple d'une pénalité vérifiant de telles conditions. Entre β^- et β^+, la pénalité doit rester dans la zone grise et vérifier les conditions de concavité (12.22).

12.2 LE CAS OÙ LES COLONNES DE LA MATRICE A SONT ORTHOGONALES

Tout comme nous avions déduit la pénalité CEL0 dans le cas où les colonnes de la matrice A sont deux à deux orthogonales à partir de l'étude en dimension 1 (voir section 9.2 page 97), nous pouvons déduire des résultats précédents des conditions nécessaires et suffisantes pour avoir (P1) et (P2) dans le cas où A^TA est diagonale. Notons que dans ce cas nous avons nécessairement $M \geqslant N$. Rappelons qu'en posant $\hat{d} = AD^{-2}A^Td$ et $\tilde{z} = D^{-1}A^Td$, où $D \in \mathbb{R}^{N\times N}$ est une matrice diagonale dont les éléments diagonaux sont donnés par $d_i = \|a_i\|\ \forall i \in \mathbb{I}_N$ (D^{-1} est bien définie puisque nous considérons que $\|a_i\| > 0$ pour tout $i \in \mathbb{I}_N$), nous avons

$$\frac{1}{2}\|Ax - d\|^2 = \frac{1}{2}\|d - \hat{d}\|^2 + \frac{1}{2}\|Dx - \tilde{z}\|^2, \tag{12.24}$$

qui nous permet de réécrire G_{ℓ_0} (donnée en (12.1)) selon

$$G_{\ell_0}(x) = \frac{1}{2}\|d - \hat{d}\|^2 + \sum_{i\in\mathbb{I}_N} \frac{1}{2}(\|a_i\|^2 x_i - \tilde{z}_i)^2 + \lambda|x_i|_0. \tag{12.25}$$

Ainsi, la minimisation de G_{ℓ_0} est réduite à la minimisation de N fonctionnelles 1D. D'après les résultats des théorèmes 12.7 et 12.9, on en déduit facilement des conditions

nécessaires et suffisantes sur Φ afin d'assurer la validité de (P1) et (P2) pour $\tilde{G}$ définie en (12.3). Ces conditions sont les suivantes : $\forall i \in \mathbb{I}_N$,

$$\phi_i(0) = 0, \tag{12.26a}$$

$$\exists \beta^{i-} \in \left[-\tfrac{\sqrt{2\lambda}}{\|a_i\|}, 0\right[\text{ et } \beta^{i+} \in \left]0, \tfrac{\sqrt{2\lambda}}{\|a_i\|}\right] \text{ t.q. } \forall u \in \mathbb{R}\backslash]\beta^{i-}, \beta^{i+}[,\ \phi_i(u) = \lambda, \tag{12.26b}$$

$$\forall u \in]\beta^{i-}, \beta^{i+}[\backslash\{0\},\ \phi_i(u) > \phi_{\text{CEL0}}(\|a_i\|, \lambda; u), \tag{12.26c}$$

$$\forall u \in B^i \setminus \{0\},\ \lim_{\substack{v \to u \\ v < u}} \phi_i'(v) > \lim_{\substack{v \to u \\ v > u}} \phi_i'(v), \tag{12.26d}$$

$$\forall u \in]\beta^{i-}, \beta^{i+}[\backslash B^i,\ \begin{cases} \phi''(u) \leqslant -\|a_i\|^2 \text{ et il existe un intervalle } \mathcal{V} \subset \mathbb{R} \setminus B^i \\ \text{t.q. } u \in \mathcal{V} \text{ et } \forall v \in \mathcal{V} \setminus \{u\},\ \phi''(v) < -\|a_i\|^2 \end{cases} \tag{12.26e}$$

où $B^i \ni 0$ est un sous-ensemble de $[\beta^{i-}, \beta^{i+}]$ contenant un nombre fini de points pour lesquels ϕ_i n'est pas différentiable. Nous rappelons que la pénalité ND Φ est donnée par,

$$\Phi(x) = \sum_{i \in \mathbb{I}_N} \phi_i(x_i). \tag{12.27}$$

et est associée à la fonctionnelle relaxée :

$$\tilde{G}(x) = \frac{1}{2}\|Ax - d\|^2 + \Phi(x) \tag{12.28}$$

La proposition suivante établie une relation entre les conditions (12.26) qui peut s'avérer utile en pratique pour définir des pénalités vérifiant les cinq conditions (12.26).

Proposition 12.12. *Soient $i \in \mathbb{I}_N$ et $B^i \subseteq \{\beta^{i-}, 0, \beta^{i+}\}$, alors on a :*

$$\{(12.26a),(12.26b),(12.26e)\} \Longrightarrow (12.26c).$$

Démonstration. Soit ϕ_i vérifiant les conditions (12.26a), (12.26b), (12.26e) et

$$f = \phi_i - \phi_{\text{CEL0}}(\|a_i\|, \lambda; \cdot).$$

Alors nous avons $f(0) = 0$ et $f(\beta^{i+}) = \lambda - \phi_{\text{CEL0}}(\|a_i\|, \lambda; \beta^{i+}) \geqslant 0$. De plus, par hypothèse sur B^i, f est deux fois différentiable sur $]0, \beta^{i+}[$ et $\forall u \in]0, \beta^{i+}[$,

$$\begin{cases} f''(u) = \phi_i''(u) + \|a_i\|^2 \leqslant 0 \text{ et il existe un intervalle } \mathcal{V} \subset \mathbb{R} \setminus B^i \\ \text{t.q. } u \in \mathcal{V} \text{ et } \forall v \in \mathcal{V} \setminus \{u\},\ f''(v) = \phi_i''(v) + \|a_i\|^2 < 0 \end{cases}, \tag{12.29}$$

montrant que $f''(u) = 0$ uniquement pour des points isolés dans $]0, \beta^{i+}[$. Ainsi f est strictement concave sur $]0, \beta^{i+}[$ impliquant que $\forall u \in]0, \beta^{i+}[\ \phi_i(u) > \phi_{\text{CEL0}}(\|a_i\|, \lambda; u)$. Le même raisonnement peut être suivi pour montrer le résultat sur $]\beta^{i-}, 0[$. □

D'après la section 12.1, alors que les cinq conditions (12.26) sont nécessaires et suffisantes pour avoir {(P1),(P2)} (théorème 12.9), seules les conditions (12.26a), (12.26b) et (12.26c) sont nécessaires et suffisantes pour avoir seulement (P1) (théorème 12.7). La principale question est maintenant de savoir si ces conditions sont également valables pour une matrice $A \in \mathbb{R}^{M \times N}$ quelconque.

12.3 EXTENSION AU CAS GÉNÉRAL ND

Dans cette section, nous montrons que les conditions (12.26) sont également *nécessaires et suffisantes* pour que $\tilde{G}$, défini en (12.3), ait les propriétés (P1) et (P2) pour une matrice $A \in \mathbb{R}^{M\times N}$ quelconque (avec $\|a_i\| > 0$ pour tout i). Tout d'abord, étant donné que le cas où les colonnes de A sont deux à deux orthogonales est un cas particulier de $A \in \mathbb{R}^{M\times N}$, nous déduisons de la section précédente que si nous ne considérons pas d'hypothèses particulières sur $A \in \mathbb{R}^{M\times N}$,

— les conditions (12.26) sont *nécessaires* pour avoir {(P1),(P2)},
— seules les conditions (12.26a), (12.26b) et (12.26c) sont *nécessaires* pour avoir (P1).

Notons qu'il est sans doute possible de déterminer des conditions plus faibles pour une matrice spécifique $A \in \mathbb{R}^{M\times N}$ et un vecteur $d \in \mathbb{R}^M$ particulier. Cependant, dans ce travail nous sommes intéressés par des conditions valides pour n'importe quel vecteur $d \in \mathbb{R}^M$ et qui ne requièrent pas d'hypothèse sur la matrice $A \in \mathbb{R}^{M\times N}$ (mais les conditions peuvent dépendre des éléments de A). Le but de cette section est donc de montrer que les conditions (12.26) sont également *suffisantes* dans ce contexte général.

Theorème 12.13 (Liens entre les minimiseurs globaux de $\tilde{G}$ et G_{ℓ_0})**.** *Soient* $A \in \mathbb{R}^{M\times N}$, $\lambda \in \mathbb{R}_+^\star$ *et* $\tilde{G}$ *définie avec* Φ *vérifiant les conditions* (12.26a), (12.26b) *et* (12.26c)*. Alors,* $\forall d \in \mathbb{R}^M$

$$\arg\min_{x\in\mathbb{R}^N} G_{\ell_0}(x) = \arg\min_{x\in\mathbb{R}^N} \tilde{G}(x), \tag{12.30}$$

et,

$$\min_{x\in\mathbb{R}^N} G_{\ell_0}(x) = \min_{x\in\mathbb{R}^N} \tilde{G}(x). \tag{12.31}$$

Démonstration. D'après (12.26a), (12.26b) et (12.26c), nous avons

$$\forall x \in \mathbb{R}^N,\ G_{\text{CEL0}}(x) \leqslant \tilde{G}(x). \tag{12.32}$$

Prenons $\hat{x} \in \mathbb{R}^N$ un minimiseur global de G_{ℓ_0}, alors d'après (Nikolova, 2013, proposition 4.1) il vient,

$$\forall i \in \sigma(\hat{x}),\ \hat{x}_i \in \mathbb{R} \setminus \left]-\frac{\sqrt{2\lambda}}{\|a_i\|}, \frac{\sqrt{2\lambda}}{\|a_i\|}\right[\tag{12.33}$$

ce qui implique, d'après (12.26a) et (12.26b),

$$G_{\ell_0}(\hat{x}) = G_{\text{CEL0}}(\hat{x}) = \tilde{G}(\hat{x}). \tag{12.34}$$

Ensuite, le théorème 9.16 affirme que $\hat{x}$ est aussi un minimiseur global de G_{CEL0} qui, avec (12.32) et (12.34), prouve l'inclusion $\subseteq$ de l'égalité (12.30).

Il existe donc au moins un point, que nous notons $x^\star \in \mathbb{R}^N$, qui est un minimiseur global des trois fonctionnelles (nous rappelons que l'existence de minimiseurs pour G_{ℓ_0} a été établie dans (Nikolova, 2013, théorème 4.4 (i))). Considérons maintenant un minimiseur global $\hat{x} \in \mathbb{R}^N$ de $\tilde{G}$. Il est clair que $\tilde{G}(\hat{x}) = \tilde{G}(x^\star) = G_{\text{CEL0}}(x^\star)$ et puisque $G_{\text{CEL0}} \leqslant \tilde{G}$

(eq. (12.32)), $\hat{x}$ est également un minimiseur global de G_{CEL0} et $\tilde{G}(\hat{x}) = G_{\text{CEL0}}(\hat{x})$. En combinant cette dernière égalité avec (12.26c) on obtient

$$\forall i \in \sigma(\hat{x}),\ \hat{x}_i \notin]\beta^{i-}, \beta^{i+}[, \tag{12.35}$$

et nous avons alors avec (12.26a) et (12.26b) que $G_{\ell_0}(\hat{x}) = \tilde{G}(\hat{x}) = \tilde{G}(x^\star) = G_{\ell_0}(x^\star)$. Ainsi, $\hat{x}$ est un minimiseur global de G_{ℓ_0} ce qui prouve l'inclusion $\supseteq$ dans (12.30). Enfin, l'égalité (12.31) est triviale d'après ce qui précède. □

Le théorème 12.13 montre donc que les conditions (12.26a), (12.26b) et (12.26c) sont suffisantes pour avoir (P1). De plus, nous pouvons déduire de ce résultat, l'existence de minimiseurs pour $\tilde{G}$ comme cela est énoncé par la proposition suivante.

> **Proposition 12.14** (Existence de minimiseurs globaux pour $\tilde{G}$)**.** *Soit $\tilde{G}$ définie comme dans le théorème 12.13. Alors, l'ensemble des minimiseurs globaux de $\tilde{G}$ est non-vide.*

Démonstration. D'après (Nikolova, 2013, théorème 4.4 (i)), l'ensemble des minimiseurs globaux de G_{ℓ_0} est non-vide et (12.30) termine la preuve. □

Afin d'analyser maintenant les liens entre les minimiseurs (locaux) des fonctionnelles $\tilde{G}$ et G_{ℓ_0}, nous commençons par démontrer deux résultats préliminaires.

> **Proposition 12.15.** *Soient $A \in \mathbb{R}^{M\times N}$, $\lambda \in \mathbb{R}_+^\star$ et $\tilde{G}$ définie avec Φ vérifiant les conditions (12.26b), (12.26d) et (12.26e). Alors $\forall i \in \mathbb{I}_N$, $\tilde{G}^i(\cdot; x^{(i)})$, la restriction de $\tilde{G}$ à la i-ème variable au point $x \in \mathbb{R}^N$, est strictement concave sur $]\beta^{i-}, 0[$ et sur $]0, \beta^{i+}[$ et strictement convexe au delà.*

Démonstration. Soit $i \in \mathbb{I}_N$ et considérons la restriction de $\tilde{G}$ à la i-ème variable au point $x \in \mathbb{R}^N$,

$$f(t) = \tilde{G}^i(t; x^{(i)}) = \frac{\|a_i\|^2}{2}t^2 + t\langle a_i, Ax^{(i)} - d\rangle + \phi_i(t) + C, \tag{12.36}$$

où $C = \frac{1}{2}\|Ax^{(i)} - d\|^2 + \sum_{j\in\mathbb{I}_N\setminus\{i\}} \phi_j(x_j)$ est une constante indépendante de t. Alors, d'après (12.26e), nous pouvons voir que

$$\forall t \in]\beta^{i-}, \beta^{i+}[\setminus B^i, \begin{cases} f''(t) = \phi_i''(t) + \|a_i\|^2 \leqslant 0 \text{ et } \exists \text{ un intervalle } \mathcal{V} \subset \mathbb{R}\setminus B^i \\ \text{t.q. } u \in \mathcal{V} \text{ et } \forall v \in \mathcal{V}\setminus\{u\},\ f''(t) = \phi_i''(t) + \|a_i\|^2 < 0 \end{cases}$$

De plus (12.26d) entraine

$$\forall t \in B^i \setminus \{0\},\ \lim_{\substack{u\to t\\ u<t}} f'(u) > \lim_{\substack{u\to t\\ u>t}} f'(u). \tag{12.37}$$

Ces deux résultats prouvent la stricte concavité de f sur $]\beta^{i-}, 0[$ et sur $]0, \beta^{i+}[$. Ensuite d'après (12.26b) nous avons $\forall t \in]-\infty, \beta^{i-}[\cup]\beta^{i+}, +\infty[$ $f''(t) = \|a_i\|^2 > 0$ ce qui termine la démonstration. □

Une conséquence de la proposition 12.15 est donnée par la proposition suivante.

Proposition 12.16. *Soit $\tilde{G}$ définie comme dans la proposition 12.15. Si $\tilde{G}$ admet un minimiseur (local) en $\hat{x} \in \mathbb{R}^N$, alors*

$$\forall i \in \sigma(\hat{x}), \quad \hat{x}_i \in]-\infty, \beta^{i-}] \cup [\beta^{i+}, +\infty[. \tag{12.38}$$

De plus, si $\beta^{i-} \in B$ (resp. $\beta^{i+} \in B$), alors l'intervalle $]-\infty, \beta^{i-}]$ (resp. $[\beta^{i+}, +\infty[$) dans (12.38) peut être réduit à $]-\infty, \beta^{i-}[$ (resp. $]\beta^{i+}, +\infty[$).

Démonstration. La preuve de ce résultat est aisée à partir de la proposition 12.15 qui affirme que la restriction de $\tilde{G}$ à la i-ème variable au point $x \in \mathbb{R}^N$ est strictement concave sur $]\beta^{i-}, 0[$ et sur $]0, \beta^{i+}[$. Ensuite, le fait que l'on puisse considérer des intervalles ouverts lorsque $\beta^{i-} \in B$ (resp. $\beta^{i+} \in B$) provient d'arguments similaires à ceux utilisés dans la preuve de la proposition 12.15 pour les points appartenant à B^i. □

Remarque 12.17. Nous avons vu dans le chapitre 9 (proposition 9.24 page 106) que les minimiseurs locaux (non-globaux) de G_{ℓ_0} pour lesquels $\exists i \in \sigma(\hat{x})$ tel que $|\hat{x}_i| < \frac{\sqrt{2\lambda}}{\|a_i\|}$ sont éliminés par la fonctionnelle CEL0. La proposition 12.16 étend ce résultat à $\tilde{G}$ (si les conditions (12.26b), (12.26d) et (12.26e) sont vérifiées) et montre qu'une telle fonctionnelle $\tilde{G}$ élimine les minimiseurs $\hat{x}$ de G_{ℓ_0} pour lesquels

$$\exists i \in \sigma(\hat{x}) \text{ tel que } \hat{x}_i \in]\beta^{i-}, \beta^{i+}[. \tag{12.39}$$

Ainsi, d'après (12.26b), $\tilde{G}$ va potentiellement éliminer «moins» de minimiseurs locaux (non-globaux) de G_{ℓ_0} que G_{CEL0}.

Nous sommes maintenant en mesure de démontrer le résultat suivant concernant les liens entre les minimiseurs (locaux) de $\tilde{G}$ et G_{ℓ_0}.

Theorème 12.18 (Liens entre les minimiseurs (locaux) de $\tilde{G}$ et G_{ℓ_0})**.** *Soient $A \in \mathbb{R}^{M\times N}$, $\lambda \in \mathbb{R}_+^\star$ et $\tilde{G}$ définie avec Φ vérifiant les conditions (12.26). Alors, $\forall d \in \mathbb{R}^M$ les deux assertions suivantes sont vérifiées,*

1. *$\hat{x}$ est un minimiseur (local) de $\tilde{G} \implies$ $\hat{x}$ est un minimiseur (local) de G_{ℓ_0},* (12.40)
2. $\tilde{G}(\hat{x}) = G_{\ell_0}(\hat{x})$. (12.41)

Démonstration. Soit $\hat{x} \in \mathbb{R}^N$ un minimiseur (local) de $\tilde{G}$ et posons $\hat{\sigma} = \sigma(\hat{x})$. Alors nous avons $0_{\mathbb{R}^N} \in \partial\tilde{G}(\hat{x})$ qui est une condition nécessaire pour qu'un point soit un optimum local de $\tilde{G}$. Étant donné que le terme quadratique dans $\tilde{G}$ est différentiable on a,

$$\forall x \in \mathbb{R}^N, \ \partial\tilde{G}(x) = A^T(Ax-d) + \partial\Phi(x) = \prod_{i\in\mathbb{I}_N} \left[A^T(Ax-d)\right]_i + \partial\phi_i(x_i). \tag{12.42}$$

D'après la proposition 12.16, $\hat{x}$ vérifie (12.38) qui, combiné avec les conditions (12.26b) et (12.26c) entraine

$$0_{\mathbb{R}^N} \in \partial\tilde{G}(\hat{x}) \iff \forall i \in \mathbb{I}_N, \begin{cases} \langle a_i, d - A\hat{x}\rangle \in \left[\underline{\delta}_i^0, \bar{\delta}_i^0\right] & \text{si } i \notin \hat{\sigma}, \\ \langle a_i, d - A\hat{x}\rangle = 0 & \text{si } i \in \hat{\sigma}. \end{cases} \tag{12.43}$$

où $\underline{\delta}_i^0$ et $\bar{\delta}_i^0$ sont définis comme en (12.14) pour ϕ_i. La seconde ligne de (12.43) peut être réécrite comme il suit :

$$(A_{\hat{\sigma}})^T A_{\hat{\sigma}} \hat{x}_{\hat{\sigma}} = (A_{\hat{\sigma}})^T d, \tag{12.44}$$

montrant que $\hat{x}$ est un minimiseur (local) de G_{ℓ_0} (Nikolova, 2013, corolaire 2.5). Enfin, l'égalité (12.41) provient du résultat de la proposition 12.16 et des conditions (12.26a) et (12.26b). □

Le théorème 12.18 est donc le pendant du théorème 12.13 pour la propriété (P2) montrant que les conditions (12.26) sont suffisantes pour que $\tilde{G}$ ait la propriété (P2).

Finalement, les conditions dérivées dans les cas unidimensionnel et orthogonal sont en fait valables pour n'importe quelle matrice $A \in \mathbb{R}^{M\times N}$. En résumé :

— (12.26a),(12.26b) (12.26c) sont nécessaires et suffisantes pour avoir (P1) ;
— l'ensemble des conditions (12.26) sont nécessaires et suffisantes pour avoir (P1) et (P2).

12.4 ANALYSE DE QUELQUES PÉNALITÉS DE L'ÉTAT DE L'ART

Cette section est dédiée à l'analyse de pénalités qui ont été proposées dans la littérature (voir section 8.4 page 87) dans le contexte des reformulations continues exactes étudié en première partie de ce chapitre. Grâce aux conditions (12.26) précédemment déterminées, nous sommes en mesure de calculer des bornes sur les paramètres définissant les pénalités analysées afin d'assurer la validité des propriétés (P1) ou {(P1),(P2)} pour la fonctionnelle relaxée $\tilde{G}$ associée.

12.4.1 ℓ_1 *tronquée (Capped-ℓ_1)*

La pénalité Capped-ℓ_1 (ou ℓ_1 tronquée) (Zhang et Chen, 2009) est définie par

$$\Phi_{\text{Cap}}(x) := \sum_{i\in\mathbb{I}_N} \lambda \min\{\theta_i|x_i|, 1\}, \tag{12.45}$$

où $\theta_i \in \mathbb{R}^\star_+$ pour tout $i \in \mathbb{I}_N$.

Comme cela est énoncé par la proposition suivante, avec un choix approprié des paramètres θ_i ($i \in \mathbb{I}_N$) la fonctionnelle relaxée associée, notée G_{Cap}, vérifie la propriété (P1).

Proposition 12.19. *G_{Cap} a la propriété (P1) si et seulement si*

$$\forall i \in \mathbb{I}_N,\ \lambda\theta_i \geqslant \sqrt{2\lambda}\|a_i\|. \tag{12.46}$$

D'autre part, G_{Cap} ne peut jamais vérifier (P2).

Démonstration. Par définition de Φ_{Cap}, la condition (12.26a) est vérifiée pour tout $\theta_i \in \mathbb{R}^\star_+$. Ensuite, nous pouvons voir que pour la pénalité Capped-ℓ_1, nous avons

$$\forall i \in \mathbb{I}_N,\ \beta^{i-} = -\frac{1}{\theta_i} \text{ et } \beta^{i+} = \frac{1}{\theta_i}. \tag{12.47}$$

Alors,

$$(12.46) \Longrightarrow \forall i \in \mathbb{I}_N,\ \frac{1}{\theta_i} \leqslant \frac{\sqrt{\lambda}}{\sqrt{2}\|a_i\|} \leqslant \frac{\sqrt{2\lambda}}{\|a_i\|} \underset{(12.47)}{\Longrightarrow} (12.26b). \tag{12.48}$$

De plus,

$$\forall i \in \mathbb{I}_N, \lim_{u \to 0^+} \phi'_{\text{Cap}}(\theta_i, \lambda; u) = \lambda\theta_i \text{ and } \lim_{u \to 0^-} \phi'_{\text{Cap}}(\theta_i, \lambda; u) = -\lambda\theta_i, \tag{12.49}$$

où $\phi_{\text{Cap}}(\theta, \lambda; u) = \lambda \min\{\theta|u|, 1\}$ pour $u \in \mathbb{R}$, $\theta \in \mathbb{R}^\star_+$ et $\lambda \in \mathbb{R}^\star_+$.

Le fait que $\phi'_{\text{CEL0}}(\|a_i\|, \lambda; u) \to \sqrt{2\lambda}\|a_i\|$ lorsque $u \to 0^+$ (resp. $-\sqrt{2\lambda}\|a_i\|$ lorsque $u \to 0^-$) combiné avec (12.49) et le fait que sur $[\beta^-, 0]$ et $[0, \beta^+]$, ϕ_{Cap} est linéaire et ϕ_{CEL0} strictement concave, montre que (12.26c) $\iff$ (12.46).

Enfin, il est facile de voir que (12.26e) ne peut pas être vérifiée étant donné que ϕ_{Cap} est linéaire sur $[\beta^-, 0]$ et sur $[0, \beta^+]$. □

La figure 70 présente la pénalité Capped-ℓ_1 pour différents paramètres θ choisis en suivant le résultat de la proposition 12.19. Nous pouvons voir sur le graphe de droite que le minimiseur global est préservé et cela est vrai pour n'importe quelle valeur de $d \in \mathbb{R}$ étant donné que (P1) est vérifiée pour les valeurs de θ utilisées. Cependant, sur ces exemples, un minimiseur local de G_{Cap} qui n'est pas un minimiseur pour G_{ℓ_0} existe lorsque $\lambda\theta = \sqrt{2\lambda}a$. Ce minimiseur local $\hat{u} \in]0, 1/\theta[$ vérifie

$$\phi'_{\text{Cap}}(\theta, \lambda; \hat{u}) = ad - a^2\hat{u} \iff \lambda\theta = ad - a^2\hat{u} \iff \hat{u} = \frac{ad - \lambda\theta}{a^2}. \tag{12.50}$$

En fait, pour tout $u_0 \in]0, 1/\theta[$, il existe $d_0 = (\lambda\theta + a^2u_0)/a \in \mathbb{R}$ pour lequel u_0 est un minimiseur local de la fonctionnelle relaxée G_{Cap} mais n'en est pas un pour G_{ℓ_0}. Cela illustre le fait que (P2) ne peux jamais être vérifiée avec la pénalité Capped-ℓ_1.

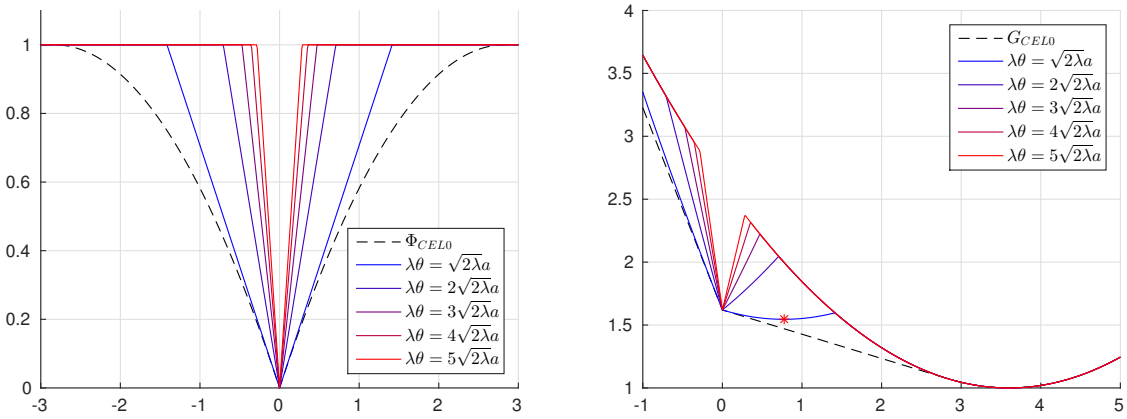


FIGURE 70 – Pénalité Capped-ℓ_1 (gauche) et relaxation continue G_{Cap} associée (droite) pour différents paramètres θ choisis en suivant le résultat de la proposition 12.19 afin de vérifier (P1). Les courbes sont tracées pour $a = 0.5$, $\lambda = 1$ et $d = 1.8$. Le point rouge représente un minimiseur local de G_{Cap} qui n'est pas un minimiseur pour G_{ℓ_0}.

Notons que des résultats similaires (concernant les minimiseurs globaux) pour la pénalité Capped-ℓ_1 ont été montrés dans (LE THI et al., 2014, 2015).

12.4.2 *Smoothly Clipped Absolute Deviation*

La pénalité SCAD, qui peut être vue comme une version «lissée» de la pénalité Capped-ℓ_1, a été proposée par FAN et LI (2001) et son expression est donnée par :

$$\Phi_{\text{SCAD}}(x) := \sum_{i \in \mathbb{I}_N} \phi_{\text{SCAD}}(\gamma_i, \tilde{\lambda}_i; x_i), \tag{12.51}$$

où $\gamma_i \in]2, +\infty[$, $\tilde{\lambda}_i \in \mathbb{R}_+^\star$ pour tout $i \in \mathbb{I}_N$ et ϕ_{SCAD} est donnée par : $\forall u \in \mathbb{R}$,

$$\phi_{\text{SCAD}}(\gamma, \tilde{\lambda}; u) := \tilde{\lambda}\left(|u|\mathbb{1}_{\{|u|\leqslant\tilde{\lambda}\}} - \frac{\tilde{\lambda}^2 - 2\gamma\tilde{\lambda}|u| + u^2}{2(\gamma-1)\tilde{\lambda}}\mathbb{1}_{\{\tilde{\lambda}<|u|\leqslant\gamma\tilde{\lambda}\}} + \frac{(\gamma+1)\tilde{\lambda}}{2}\mathbb{1}_{\{|u|>\gamma\tilde{\lambda}\}}\right). \quad (12.52)$$

Pour cette pénalité nous avons un résultat similaire à celui de la proposition 12.19 établi pour la pénalité Capped-ℓ_1.

Proposition 12.20. *Considérons $\|a_i\| < 1/\sqrt{3}$ pour tout $i \in \mathbb{I}_N$. Alors, G_{SCAD} a la propriété* (P1) si et seulement si

$$\forall i \in \mathbb{I}_N,\ \frac{(\gamma_i+1)\tilde{\lambda}_i^2}{2} = \lambda \ \textit{et}\ \ 2 < \gamma_i \leqslant \frac{1}{\|a_i\|^2} - 1. \quad (12.53)$$

D'autre part, G_{SCAD} ne peut jamais vérifier (P2).

Démonstration. La preuve est détaillée en annexe A.3.5 page 205. □

L'hypothèse $\|a_i\| < 1/\sqrt{3}$ ($i \in \mathbb{I}_N$) dans la proposition 12.20 peut toujours être vérifiée en normalisant les colonnes de A puis en multipliant la matrice par un réel $\zeta < 1/\sqrt{3}$ car cela ne change pas le problème (changement de variable dans G_{ℓ_0}). La figure 71 présente la pénalité SCAD (gauche) ainsi que la relaxation continue G_{SCAD} associée (droite) pour différents paramètres γ choisis selon le résultat de la proposition 12.20 ($\tilde{\lambda}$ est donc également déterminé par (12.53)). Les mêmes conclusions que pour la pénalité Capped-ℓ_1 (figure 70) peuvent être faites.

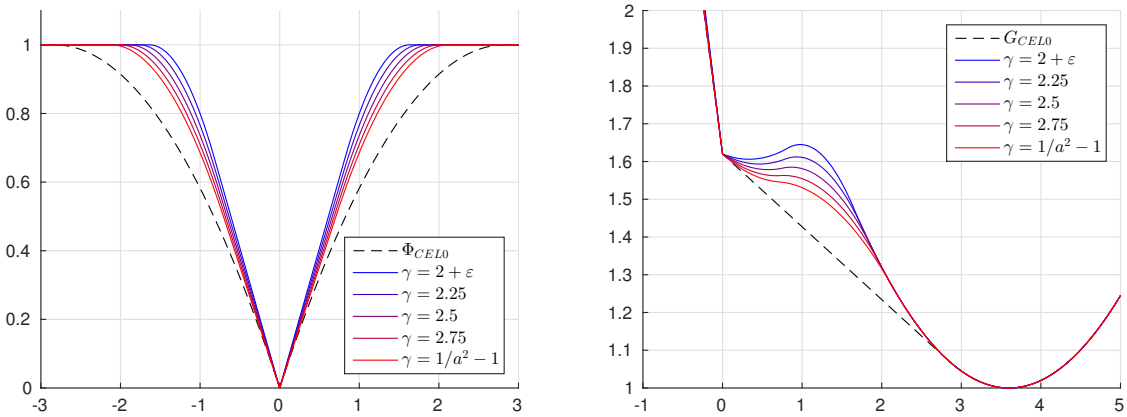


FIGURE 71 – Pénalité SCAD (gauche) et relaxation continue G_{SCAD} associée (droite) pour différents paramètres γ choisis en suivant le résultat de la proposition 12.20 afin de vérifier (P1) ($\tilde{\lambda}$ est également déterminé par (12.53)). Les courbes sont tracées pour $a = 0.5 < 1/\sqrt{3}$, $\lambda = 1$ et $d = 1.8$.

12.4.3 *Minimax Concave Penalty*

La pénalité MCP, proposée par ZHANG (2010), est définie par

$$\Phi_{\text{MCP}}(x) := \sum_{i\in\mathbb{I}_N} \phi_{\text{MCP}}(\gamma_i, \tilde{\lambda}_i; x_i), \quad (12.54)$$

où $\gamma_i \in \mathbb{R}_+^\star$, $\tilde{\lambda}_i \in \mathbb{R}_+^\star$ pour tout $i \in \mathbb{I}_N$ et ϕ_{MCP} est donnée par : $\forall u \in \mathbb{R}$,

$$\phi_{\text{MCP}}(\gamma, \tilde{\lambda}; u) := \tilde{\lambda} \int_0^{|u|} \left(1 - x/(\gamma\tilde{\lambda})\right)_+ dx = \tilde{\lambda} \left(\frac{\gamma\tilde{\lambda}}{2} \mathbb{1}_{\{|u| > \gamma\tilde{\lambda}\}} + \left(|u| - \frac{u^2}{2\gamma\tilde{\lambda}} \right) \mathbb{1}_{\{|u| \leqslant \gamma\tilde{\lambda}\}} \right). \quad (12.55)$$

D'après les conditions (12.26) établies dans la section précédente, nous pouvons montrer qu'avec un choix judicieux des paramètres γ_i et $\tilde{\lambda}_i$ ($i \in \mathbb{I}_N$), il est possible d'avoir les propriétés (P1) et (P2) pour la fonctionnelle relaxée G_{MCP}.

Proposition 12.21. *G_{MCP} a les deux propriétés (P1) et (P2) si et seulement si*

$$\forall i \in \mathbb{I}_N, \ \frac{\gamma_i \tilde{\lambda}_i^2}{2} = \lambda \ \textit{et} \ \ \gamma_i < \frac{1}{\|a_i\|^2}. \quad (12.56)$$

Démonstration. Par définition de Φ_{MCP}, les conditions (12.26a) et (12.26d) sont vérifiées pour tout $\gamma_i \in \mathbb{R}_+^\star$ et $\tilde{\lambda}_i \in \mathbb{R}_+^\star$ ($i \in \mathbb{I}_N$). Alors, la preuve consiste maintenant à montrer que (12.56) est équivalente aux trois conditions (12.26b), (12.26c) et (12.26e). Avec la définition de Φ_{MCP}, nous avons (par symétrie nous restreignons la preuve à $\mathbb{R}_+$),

$$(12.26b) \iff \forall i \in \mathbb{I}_N, \ \frac{\gamma_i \tilde{\lambda}_i^2}{2} = \lambda \ \text{ et } \ \gamma_i \tilde{\lambda}_i = \beta^{i+} \leqslant \frac{\sqrt{2\lambda}}{\|a_i\|} \quad (12.57)$$

$$\iff \forall i \in \mathbb{I}_N, \ \frac{\gamma_i \tilde{\lambda}_i^2}{2} = \lambda \ \text{ et } \ \gamma_i \leqslant \frac{1}{\|a_i\|^2} \quad (12.58)$$

$$(12.26e) \iff \forall i \in \mathbb{I}_N, \forall u \in]-\gamma_i\tilde{\lambda}_i, \gamma_i\tilde{\lambda}_i[\setminus\{0\}, \ \phi''_{\text{MCP}}(u) = -\frac{1}{\gamma_i} < -\|a_i\|^2 \quad (12.59)$$

$$\iff \forall i \in \mathbb{I}_N, \ \gamma_i < \frac{1}{\|a_i\|^2} \quad (12.60)$$

D'après les équivalences précédentes il est clair que (12.56) est équivalente à l'ensemble des conditions {(12.26b),(12.26e)}. Il ne reste alors plus qu'à montrer que (12.56) $\implies$ (12.26c) ce qui est direct avec le résultat de la proposition 12.12 et termine la preuve. □

Prenons Φ_{MCP} définie en suivant les conditions de la proposition 12.21, alors on peut réécrire l'expression de cette pénalité en utilisant seulement le paramètre γ_i (i. e. on enlève la dépendance en $\tilde{\lambda}_i$ grâce à la condition $\frac{\gamma_i\tilde{\lambda}_i^2}{2} = \lambda$),

$$\Phi_{\text{MCP}}(x) = \sum_{i \in \mathbb{I}_N} \lambda - \frac{1}{2\gamma_i} \left(|x_i| - \sqrt{2\lambda\gamma_i} \right)^2 \mathbb{1}_{\{|x_i| \leqslant \sqrt{2\lambda\gamma_i}\}}. \quad (12.61)$$

Ainsi, $\forall i \in \mathbb{I}_N, \ \forall \gamma_i < \frac{1}{\|a_i\|^2}$, la fonctionnelle relaxée G_{MCP} a les propriétés (P1) et (P2). Cela définit une sous-famille de MCP qui sont des approximations continues exactes de la norme-ℓ_0. Finalement, nous pouvons également noter que la limite inférieure de cette sous-famille est donnée par

$$\forall x \in \mathbb{R}^N, \ \lim_{\substack{\gamma_i \to 1/\|a_i\|^2 \\ \forall i \in \mathbb{I}_N}} \Phi_{\text{MCP}}(x) = \Phi_{\text{CEL0}}(x). \quad (12.62)$$

La figure 72 présente la pénalité MCP et la relaxation G_{MCP} associée pour différentes valeurs de γ vérifiant les conditions de la proposition 12.21 afin d'avoir les propriétés (P1) et (P2) pour G_{MCP}.

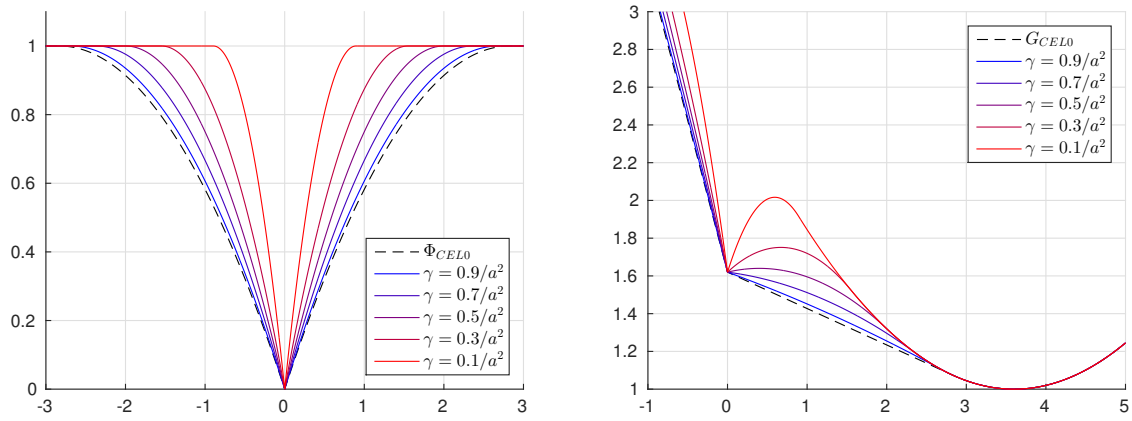


FIGURE 72 – Pénalité MCP (gauche) et relaxation continue G_{MCP} associée (droite) pour différents paramètres γ choisis en suivant la proposition 12.21 afin d'avoir les propriétés (P1) et (P2) pour G_{MCP}. Les courbes sont tracées pour $a = 0.5$, $\lambda = 1$ et $d = 1.8$.

12.4.4 ℓ_p *tronquée* ($p \in]0,1[$)

Il est évident que les approximations de la norme-ℓ_0 telles que la pénalité log-sum ou encore les normes-ℓ_p ($0 < p < 1$) ne peuvent pas permettre d'avoir (P1) ni (P2) étant donné qu'elle ne vérifient pas la condition (12.26b). Cependant, nous pouvons définir des versions tronquées de ces pénalités dans le même esprit que Capped-ℓ_1 étudié dans le paragraphe 12.4.1. Dans la suite, nous analysons les pénalités ℓ_p-tronquées définies par :

$$\Phi_{TLp}(x) := \sum_{i \in \mathbb{I}_N} \lambda \min\left\{(\theta_i |x_i| + \varepsilon_i)^{p_i}, 1 + \varepsilon_i^{p_i}\right\} - \lambda \varepsilon_i^{p_i}, \tag{12.63}$$

où $\theta_i \in \mathbb{R}_+^\star$, $\varepsilon_i \in \mathbb{R}_+^\star$ et $p_i \in]0,1[$ pour tout $i \in \mathbb{I}_N$. Notons que les paramètres ε_i rendent la pénalité localement Lipschitz en 0 afin de satisfaire l'hypothèse 12.2.

Tout comme pour MCP, pour p_i et ε_i donnés ($i \in \mathbb{I}_N$), il existe des paramètres θ_i pour lesquels la relaxation continue G_{TLp} a les deux propriétés (P1) et (P2).

Proposition 12.22. *G_{TLp} a les deux propriétés (P1) et (P2) si et seulement si*

$$\forall i \in \mathbb{I}_N,\ \theta_i \geqslant \theta_0^i := \|a_i\| \max\left\{\frac{(1+\varepsilon_i^{p_i})^{1/p_i - 0.5}}{\sqrt{p_i(1-p_i)\lambda}}, \frac{(1+\varepsilon_i^{p_i})^{1/p_i} - \varepsilon_i}{\sqrt{2\lambda}}\right\}. \tag{12.64}$$

Démonstration. De manière évidente, les conditions (12.26a) et (12.26d) sont vérifiées, par définition de Φ_{TLp}, pour tout $\theta_i > 0$, $p_i \in]0,1[$ et $\varepsilon_i \in \mathbb{R}_+^\star$ ($i \in \mathbb{I}_N$). Ensuite, (12.26e) est équivalente à (nous restreignons la preuve à $\mathbb{R}_+$ par symétrie) : $\forall i \in \mathbb{I}_N$, $\forall u \in \left]0, \frac{1}{\theta_i}\left((1+\varepsilon_i^{p_i})^{1/p_i} - \varepsilon_i\right)\right[$,

$$\begin{cases} \phi''_{TLp}(\theta_i, p_i, \varepsilon_i, \lambda; u) = p_i(p_i - 1)\lambda\theta_i^2(\theta_i u + \varepsilon_i)^{p_i - 2} \leqslant -\|a_i\|^2, \\ \text{et il existe un intervalle } \mathcal{V} \subseteq \mathbb{R} \\ \text{t.q. } u \in \mathcal{V} \text{ et } \forall v \in \mathcal{V} \setminus \{u\},\ \phi''_{TLp}(\theta_i, p_i, \varepsilon_i, \lambda; v) < -\|a_i\|^2 \end{cases} \tag{12.65}$$

où $\phi_{\text{TLp}}(\theta, p, \varepsilon, \lambda; u) := \lambda \min\{(\theta|u| + \varepsilon)^p, 1 + \varepsilon^p\} - \lambda\varepsilon^p$ pour $u \in \mathbb{R}$, $\theta \in \mathbb{R}_+^\star$, $p \in]0, 1[$, $\varepsilon \in \mathbb{R}_+^\star$ et $\lambda \in \mathbb{R}_+^\star$. Comme $\phi''_{\text{TLp}}(\theta_i, p_i, \varepsilon_i, \lambda; \cdot)$ est strictement croissante, (12.65) se réduit à

$$\forall i \in \mathbb{I}_N,\ \phi''_{\text{TLp}}(\theta_i, p_i, \varepsilon_i, \lambda; \beta^{i+}) = p_i(p_i - 1)\lambda\theta_i^2(1 + \varepsilon_i^{p_i})^{1-2/p_i} \leqslant -\|a_i\|^2, \tag{12.66}$$

$$\iff \theta_i \geqslant \frac{\|a_i\|}{\sqrt{p_i(1-p_i)\lambda}}(1 + \varepsilon_i^{p_i})^{1/p_i - 0.5}, \tag{12.67}$$

où $\beta^{i+} = \frac{1}{\theta_i}\left((1 + \varepsilon_i^{p_i})^{1/p_i} - \varepsilon_i\right)$. Ensuite, nous pouvons voir que

$$(12.26b) \iff \forall i \in \mathbb{I}_N,\ \beta^{i+} = \frac{1}{\theta_i}\left((1 + \varepsilon_i^{p_i})^{1/p_i} - \varepsilon_i\right) \leqslant \frac{\sqrt{2\lambda}}{\|a_i\|}, \tag{12.68}$$

$$\iff \theta_i \geqslant \frac{\|a_i\|}{\sqrt{2\lambda}}\left((1 + \varepsilon_i^{p_i})^{1/p_i} - \varepsilon_i\right). \tag{12.69}$$

Ainsi (12.64) $\iff$ {(12.26b),(12.26e)}. Finalement ce qui précède avec la proposition 12.12 montre que (12.64) $\implies$ (12.26c) et complète la démonstration. □

Une analyse similaire pourrait être effectuée avec une version tronquée de la pénalité log-sum ou n'importe quelle pénalité n'étant pas constante pour des grandes valeurs de $|u|$. Enfin, la figure 73 présente la pénalité ℓ_p-tronquée (gauche) et la relaxation continue G_{TLp} associée (droite) pour différents paramètres θ et $p = 0.5$ choisis en accord avec la proposition 12.22 afin d'avoir les propriétés (P1) et (P2) pour G_{TLp}.

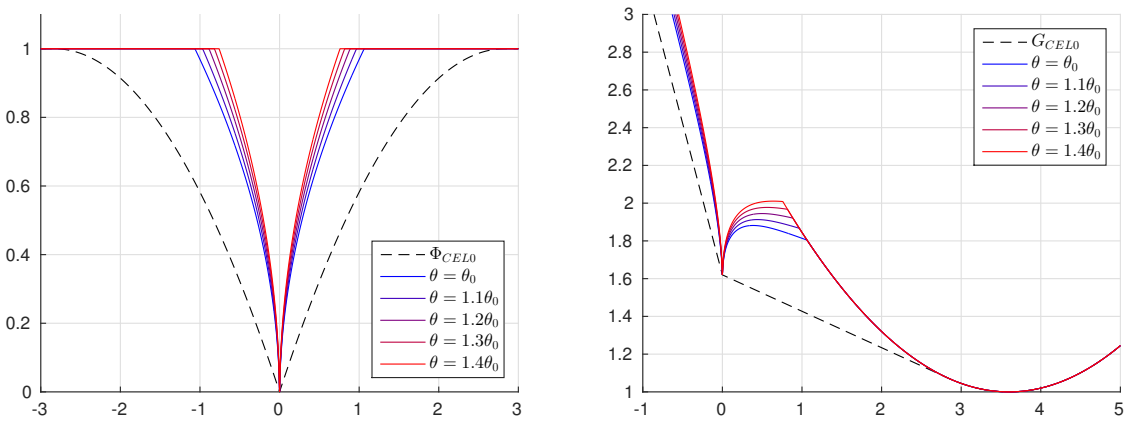


FIGURE 73 – Pénalité ℓ_p-tronquée (gauche) et relaxation continue G_{TLp} associée (droite) pour différents paramètres θ et $p = 0.5$ choisis en suivant la proposition 12.22 afin d'avoir les propriétés (P1) et (P2) pour G_{TLp}. Les courbes sont tracées pour $a = 0.5$, $\lambda = 1$ et $d = 1.8$.

La table 8 résume l'ensemble des résultats obtenus dans cette section pour les différentes pénalités de l'état de l'art qui ont été analysées.

12.5 CONCLUSION

Dans ce chapitre, nous avons proposé une vue unifiée des pénalités continues approchant la norme-ℓ_0 dans le contexte des *reformulations continues exactes* du problème de moindres carrés pénalisé en norme-ℓ_0. Plus précisément, nous avons établi cinq conditions *nécessaires et suffisantes* pour qu'une pénalité continue approchant la norme-ℓ_0 résulte en une relaxation continue $\tilde{G}$ préservant tous les minimiseurs globaux de la fonctionnelle initiale G_{ℓ_0} (théorème 12.13), et dont les minimiseurs (locaux) sont également des minimiseurs locaux de G_{ℓ_0} (théorème 12.18). Bien que la relaxation résultante soit toujours non-convexe, un

Pénalité	Définition ϕ_i	(P1)	(P2)	Conditions
Cap-ℓ_1	$\lambda \min\{\theta_i \lvert x_i\rvert, 1\}$	✔	✘	$\lambda\theta_i \geqslant \sqrt{2\lambda}\lVert a_i\rVert$
SCAD	$\begin{cases} \tilde{\lambda}_i \lvert u\rvert & \text{if } \lvert u\rvert \leqslant \tilde{\lambda}_i, \\ \frac{2\gamma_i\tilde{\lambda}_i\lvert u\rvert - \tilde{\lambda}_i^2 - u^2}{2(\gamma_i-1)} & \text{if } \tilde{\lambda}_i < \lvert u\rvert \leqslant \gamma_i\tilde{\lambda}_i, \\ \frac{(\gamma_i+1)\tilde{\lambda}_i^2}{2} & \text{if } \lvert u\rvert > \gamma_i\tilde{\lambda}_i \end{cases}$	✔	✘	$\frac{(\gamma_i+1)\tilde{\lambda}_i^2}{2} = \lambda$ $2 < \gamma_i \leqslant \frac{1}{\lVert a_i\rVert^2} - 1$
MCP	$\begin{cases} \frac{\gamma_i\tilde{\lambda}_i^2}{2} & \text{if } \lvert u\rvert > \gamma_i\tilde{\lambda}_i \\ \left(\tilde{\lambda}_i\lvert u\rvert - \frac{u^2}{2\gamma_i}\right) & \text{if } \lvert u\rvert \leqslant \gamma_i\tilde{\lambda}_i \end{cases}$	✔	✔	$\frac{\gamma_i\tilde{\lambda}_i^2}{2} = \lambda$ $\gamma_i < \frac{1}{\lVert a_i\rVert^2}$
ℓ_p-Tronq	$\lambda \min\left\{(\theta_i\lvert x_i\rvert + \varepsilon_i)^{p_i}, 1 + \varepsilon_i^{p_i}\right\} - \lambda\varepsilon_i^{p_i}$	✔	✔	(12.64)

TABLE 8 – Récapitulatif des conditions sur les paramètres des pénalités Capped-ℓ_1, SCAD, MCP et ℓ_p-tronquée permettant d'avoir la propriété (P1) ou les deux propriétés {(P1),(P2)}

point intéressant est que certains minimiseurs locaux (non-globaux) de la fonctionnelle initiale sont éliminés par une telle relaxation continue (remarque 12.17).

L'unification proposée donne accès à une nouvelle façon de comparer les pénalités approchant la norme-ℓ_0 pour G_{ℓ_0}. Nous avons ainsi montré, pour différentes pénalités de la littérature, qu'un choix judicieux des paramètres les définissant résultait en une relaxation continue exacte du critère G_{ℓ_0} vérifiant les propriétés (P1) et (P2) (pour MCP et ℓ_p-tronquée) alors qu'il n'est possible de vérifier que (P1) pour d'autres pénalités comme Capped-ℓ_1 ou SCAD. Par ailleurs, la pénalité CEL0 est la limite inférieure de la classe de pénalités déterminée dans le présent chapitre (remarque 12.17). De plus c'est celle qui potentiellement élimine le plus de minimiseurs locaux de G_{ℓ_0}. Cette observation, combinée à sa convexité par rapport à chacune des variables de la base canonique de $\mathbb{R}^N$, et au fait qu'elle résulte en l'enveloppe convexe de G_{ℓ_0} dans le cas orthogonal, en font l'alternative à la norme-ℓ_0 la plus intéressante.

13

CONCLUSIONS ET PERSPECTIVES SUR L'OPTIMISATION DE CRITÈRES PÉNALISÉS EN NORME ZÉRO

CONCLUSION

Dans cette partie, des contributions ont été apportées tant sur le plan théorique que sur des applications dans le contexte de la minimisation du critère des moindres carrés pénalisé en norme-ℓ_0.

Les contributions les plus importantes de ce travail concernent la définition et l'analyse de relaxations continues *exactes* du problème. Bien que l'intérêt envers de telles relaxations ne soit pas nouveau, c'est à notre connaissance la première fois qu'une étude complète concernant les liens entre les minimiseurs locaux et globaux des deux fonctionnelles est proposée. Nous avons ainsi mis en évidence qu'il était possible de s'attaquer au problème ℓ_2-ℓ_0 (minimisation de G_{ℓ_0}) par la minimisation d'une fonctionnelle continue (non-convexe) G_{CEL0} telle que

— $\arg\min_x G_{\ell_0}(x) \subseteq \arg\min_x G_{\text{CEL0}}(x)$;
— $\hat{x} \in \mathbb{R}^N$ minimiseur (local) de $G_{\text{CEL0}} \Longrightarrow \hat{x}^0$ minimiseur (local) de G_{ℓ_0}, où $\hat{x}^0$ est déterminé à partir de $\hat{x}$ par une simple opération de seuillage ;
— certains minimiseurs locaux de G_{ℓ_0} ne sont pas des points critiques de G_{CEL0}.

La continuité de cette relaxation G_{CEL0} nous a ensuite permis d'adresser le problème de minimisation avec différents algorithmes d'optimisation non-convexe ne pouvant pas s'appliquer directement sur G_{ℓ_0}. En s'appuyant sur de tels algorithmes de la littérature, nous avons également proposé différentes stratégies pour améliorer les résultats de minimisation de la fonctionnelle non-convexe mais aussi pour s'affranchir de la sélection du paramètre de régularisation en calculant un «chemin de régularisation» pour G_{CEL0}.

Ces méthodes, ainsi que certains algorithmes de l'état de l'art ont été utilisés ensuite pour la minimisation de G_{CEL0} dans le contexte de divers problèmes en traitement du signal et des images. À travers ces expériences, nous avons mis en évidence l'intérêt de minimiser la fonctionnelle relaxée plutôt que de travailler directement sur la fonctionnelle initiale G_{ℓ_0}. En particulier, la propriété qu'a G_{CEL0} d'éliminer des minimiseurs locaux de G_{ℓ_0} est en entre autres l'élément permettant aux algorithmes optimisant G_{CEL0} d'être plus performants, en évitant ces minimiseurs locaux, que ceux traitant directement avec G_{ℓ_0}.

Nous avons finalement proposé une vue unifiée des pénalités continues approchant la norme-ℓ_0 dans ce contexte de relaxations exactes. Ce travail, constituant une autre contribution importante de ce manuscrit, offre une nouvelle vision de ces pénalités et donne de nouveaux critères pour les comparer. En effet, nous avons dérivé cinq conditions *nécessaires et suffisantes* que doit vérifier une pénalité continue approchant la norme-ℓ_0 dans l'objectif que le critère moindres carrés pénalité $\tilde{G}$ associé vérifie les deux propriétés suivantes :

— $\arg\min_x G_{\ell_0}(x) = \arg\min_x \tilde{G}(x)$;
— $\hat{x} \in \mathbb{R}^N$ minimiseur (local) de $\tilde{G} \Longrightarrow \hat{x}$ minimiseur (local) de G_{ℓ_0}.

Nous avons ainsi pu identifier des paramètres pour certaines pénalités de la littérature permettant de vérifier uniquement la première (Capped-ℓ_1 et SCAD) ou les deux (MCP et

ℓ_p-tronquée, $p \in]0, 1[$) propriétés décrites ci-dessus. Par ailleurs, cette étude nous a conduit une fois de plus à la pénalité CEL0 qui est la limite inférieure de la classe de pénalités définie par les conditions obtenues. Par ailleurs, CEL0 est aussi la pénalité qui peut potentiellement éliminer le plus de minimiseurs locaux mais également la seule à être convexe par rapport à chacune des variables et à résulter en l'enveloppe convexe de G_{ℓ_0} dans le cas orthogonal. Ces propriétés font de CEL0 la pénalité qui semble être la mieux appropriée dans le but d'obtenir une relaxation continue exacte de G_{ℓ_0}.

PERSPECTIVES

Au delà des perspectives qui ont pu être mentionnées en conclusion des différents chapitres, deux directions principales de recherche se dégagent suite au travail réalisé dans cette partie de la thèse.

Minimisation de la fonctionnelle G_{CEL0}

La première concerne la minimisation de G_{CEL0}. Plusieurs algorithmes permettant de minimiser G_{CEL0} ont été présentés dans le chapitre 10 et une heuristique inspirée des méthodes GNC a été proposée. Les résultats obtenus sur le problème de déconvolution de trains d'impulsions dans le chapitre 11 sont d'une certaine manière un preuve de concept qu'une telle stratégie peut être intéressante pour la minimisation de G_{CEL0}. Plusieurs pistes pour prolonger le travail réalisé dans cette thèse sont envisageables :

— comment déformer la fonctionnelle G_{CEL0} dans un cadre GNC ? Cette question est loin d'être triviale et nécessite d'être analysée plus précisément.
— convergence globale sous certaines conditions ?

Par ailleurs, nous avons présenté dans l'état de l'art du chapitre 2 des algorithmes gloutons bidirectionnels comme SBR construit dans l'objectif de minimiser la fonctionnelle G_{ℓ_0}. Ainsi à chaque itération de l'algorithme la solution est un minimiseur local de G_{ℓ_0}. Étant donné que G_{CEL0} élimine des minimiseurs locaux de G_{ℓ_0}, il semble intéressant de voir si ces algorithmes construisent également une séquence de minimiseurs pour G_{CEL0}. Si ce n'est pas le cas, est-ce que cela peut être assuré en modifiant par exemple la règle de sélection de la composante à ajouter ou à retirer ? (par exemple en se basant sur la caractérisation des points critiques de G_{CEL0}).

Généralisation des reformulations continues exactes à d'autres problèmes

Dans l'objectif d'étendre les travaux réalisés dans cette thèse à d'autre fonctionnelles non-convexes, une première étape naturelle serait de considérer un critère non-quadratique pénalisé en norme-ℓ_0 ce qui correspondrait à la prise en compte d'autres types de bruits que le bruit gaussien. En s'inspirant de la généralisation proposée dans le chapitre 12 pour le cas ℓ_2-ℓ_0, une idée consisterait alors à définir un ensemble de conditions (au moins suffisantes) permettant d'assurer les propriétés que nous recherchons concernant les minimiseurs de la fonctionnelle relaxée. Une telle étude a commencé à être explorée sur la fin de cette thèse. La difficulté majeure n'est alors pas tant la définition de conditions, mais la définition de conditions qui *soient calculables et utilisables en pratique*. En effet, dans le cas quadratique, la définition de CEL0 ne nécessite que le calcul de la norme des colonnes de l'opérateur et tout est séparable ce qui facilite la manipulation. Malheureusement, ce n'est en général pas le

cas avec des fonctionnelles non-quadratiques pour lesquelles le calculs des conjuguées et biconjuguées (même en 1D) est plus complexe et ne conduit pas toujours à des solutions explicites.

La pénalité CEL0 a été introduite dans le chapitre 9 à partir de l'enveloppe convexe dans le cas orthogonal. On peut alors se poser la question de savoir s'il n'existe pas une transformation plus directe permettant de passer de G_{ℓ_0} à G_{CEL0} sans la nécessité d'étudier le cas orthogonal. Aussi, cela laisse entrevoir qu'une telle transformation pourrait être valide pour d'autres fonctionnelles ce qui serait une forme de généralisation des travaux présentés dans cette thèse. Nous avons commencé à explorer cette piste et nous avons été amenés à définir une transformation que l'on appelle *Recursive Coordinate Convex Hull* qui s'exprime comme suit :

Soit $f : \mathbb{R}^N \to \bar{\mathbb{R}}$ une fonction non-convexe, propre admettant une minorante affine. Nous appelons *Recursive Coordinate Convex Hull* de f, la fonction $f_J^{\star\star r}$ définie par

$$f_J^{\star\star r} = ((f^{\star\star(j_1)})^{\star\star(j_2)} \dots)^{\star\star(j_N)}, \tag{13.1}$$

avec $J := \{j_1, \dots, j_N\} \in \mathcal{P}_{\mathbb{I}_N} := \{I \in \mathcal{P}(\mathbb{I}_N) : |I| = N\}$ une permutation donnée de $\mathbb{I}_N$ et $f^{\star\star(j)}$ la fonctionnelle obtenue par application de l'enveloppe convexe (1D) dans la direction $j \in \mathbb{I}_N$:

$$\forall x \in \mathbb{R}^N,\ (f^{\star\star(j)})_x^j = (f_x^j)^{\star\star}. \tag{13.2}$$

où la notation (f_x^j) est utilisée pour mentionner la restriction de f à la j-ème variable au point $x \in \mathbb{R}^N$. Notons que dans le cas du problème ℓ_2-ℓ_0, appliquer une telle transformation résulte exactement en la fonctionnelle CEL0 avec les propriétés qu'on lui connaît et ce indépendamment de l'ordre des composantes utilisé pour le calcul des enveloppes convexes récurcives. Dans le cas d'une fonction f quelconque nous avons montré quelques résultats préliminaires :

- $\arg\min_{x\in\mathbb{R}^N} f(x) \subseteq \arg\min_{x\in\mathbb{R}^N} f_J^{\star\star r}(x)$ pour tout $J \in \mathcal{P}_{\mathbb{I}_N}$;
- de tout minimiseur global $\hat{x}$ de $f_J^{\star\star r}(x)$, on peut déterminer un minimiseur global $\hat{x}^0$ de f en modifiant les composantes de $\hat{x}$ une à une le long de segments de $\mathbb{R}^N$ sur lesquels G_{CEL0} est constante.

Ces premiers résultats sont encourageants et plusieurs questions nécessitent d'être étudiées :

- qu'en est t-il pour les minimiseurs locaux ?
- est-ce que des minimiseurs locaux de la fonctionnelle initiale sont éliminés par $f_J^{\star\star r}$ et ce quelque soit f ?
- quel est l'influence de l'ordre de la récursion (i. e. le choix de $J \in \mathcal{P}_{\mathbb{I}_N}$) ?
- dans quels cas peut-on calculer une telle transformation pour être capable de l'utiliser en pratique ?

Ainsi, si nous sommes en mesure de montrer que cette transformation permet d'obtenir les mêmes propriétés que CEL0 et ce, pour toute fonction f (ou certaines classes de fonctions), de nombreuses applications nécessitant la résolution de problèmes non-convexes pourraient alors bénéficier d'une reformulation continue exacte telle que CEL0 pour le cas ℓ_2-ℓ_0 si tant est qu'on sache calculer explicitement les biconjuguées (1D) récursives. Cela motive la réalisation d'une étude approfondie de cette transformation.

A

DÉMONSTRATIONS

Sommaire

A.1 DÉMONSTRATIONS DU CHAPITRE 9

A.1.1 *Preuve de la proposition 9.3*

Nous commençons par le calcul de la conjuguée (transformée de Legendre-Fenchel) de g_0 qui est donné par

$$\forall u^\star \in \mathbb{R},\ g_0^\star(u^\star) = \sup_{u\in\mathbb{R}}\ h_{u^\star}(u) := u^\star u - g_0(u)\,. \tag{A.1}$$

D'après la définition de $|u|_0$ en (7.3), il est évident que $h_{u^\star}$ admet au plus deux minima (locaux) sur $\mathbb{R}$. Le premier est atteint en $u_0 = 0$ et correspond à

$$h_{u^\star}(u_0) = -\frac{d^2}{2}\,. \tag{A.2}$$

Le deuxième est quant à lui atteint en $u_1 \neq 0$ et correspond à la partie continue de $h_{u^\star}$ (i. e. lorsque $|u|_0 = 1$). Il est donc solution de

$$h'_{u^\star}(u_1) = 0,\ u_1 \neq 0 \iff u^\star - a^2 u_1 + ad = 0,\ u_1 \neq 0\,, \tag{A.3}$$

qui est encore équivalent à

$$u_1 = \frac{1}{a^2}(u^\star + ad),\ u^\star \neq -ad\,, \tag{A.4}$$

et nous avons

$$h_{u^\star}(u_1) = \frac{(u^\star)^2}{2a^2} + \frac{u^\star d}{a} - \lambda = \frac{1}{2a^2}(u^\star + ad)^2 - \frac{d^2}{2} - \lambda,\ u^\star \neq -ad\,. \tag{A.5}$$

Il convient donc maintenant de comparer les valeurs données en (A.2) et (A.5) afin de déterminer qui de u_0 ou u_1 maximise $h_{u^\star}$.

Si $u^\star = -ad$ alors $u_0 = u_1 = 0$ et $g_0^\star(u^\star) = h_{u^\star}(0) = -\frac{d^2}{2}$. Sinon, le supremum de $h_{u^\star}$ est atteint en u_1 si et seulement si

$$\frac{1}{2a^2}(u^\star + ad)^2 - \frac{d^2}{2} - \lambda + \frac{d^2}{2} \geqslant 0 \iff |u^\star + ad| \geqslant \sqrt{2\lambda}a, \tag{A.6}$$

et il est atteint en u_0 si et seulement si $|u^\star + ad| \leqslant \sqrt{2\lambda}a$.

On en déduit ainsi l'expression de $g_0^\star$:

$$\forall u^\star \in \mathbb{R},\ g_0^\star(u^\star) = \left\{\frac{1}{2a^2}(u^\star + ad)^2 - \lambda\right\} \mathbb{1}_{\{|u^\star + ad| \geqslant \sqrt{2\lambda}a\}} - \frac{d^2}{2}. \tag{A.7}$$

Nous pouvons maintenant calculer la biconjuguée de g_0 an appliquant la transformée de Legendre-Fenchel à $g_0^\star$,

$$\forall u \in \mathbb{R},\ g_0^{\star\star}(u) = \sup_{v \in \mathbb{R}} h_u^\star(v) := uv - g_0^\star(v). \tag{A.8}$$

Deux cas sont alors à distinguer :

1. Si $|v + ad| < \sqrt{2\lambda}a$ alors

$$h_u^\star(v) = uv + \frac{d^2}{2}. \tag{A.9}$$

Dans ce cas $h_u^\star$ est linéaire et on cherche le supremum sur l'intervalle

$$\left[-ad - \sqrt{2\lambda}a, -ad + \sqrt{2\lambda}a\right].$$

Il s'en suit que le supremum de $h_u^\star$ est atteint en tout point de cet intervalle lorsque $u = 0$ et en l'une des deux bornes de ce dernier si $u \neq 0$ (dépendant du signe de u). On peut donc en déduire que

$$v_1 = -ad + \operatorname{sign}(u)\sqrt{2\lambda}a, \tag{A.10}$$

maximise $h_u^\star$ et nous avons

$$h_u^\star(v_1) = -adu + \operatorname{sign}(u)u\sqrt{2\lambda}a + \frac{d^2}{2} = -adu + |u|\sqrt{2\lambda}a + \frac{d^2}{2}. \tag{A.11}$$

2. Si $|v + ad| \geqslant \sqrt{2\lambda}a$ alors

$$h_u^\star(v) = uv - \frac{1}{2a^2}(v + ad)^2 + \lambda + \frac{d^2}{2}, \tag{A.12}$$

qui atteint son supremum en v_2 vérifiant

$$(h_u^\star)'(v_2) = 0 \iff u - \frac{1}{a^2}(v_2 + ad) = 0 \iff v_2 = a^2u - ad. \tag{A.13}$$

Ensuite, nous devons vérifier que

$$|v_2 + ad| \geqslant \sqrt{2\lambda}a \iff |u| \geqslant \frac{\sqrt{2\lambda}}{a}. \tag{A.14}$$

Finalement, si u vérifie la condition (A.14) alors $v_2 = a^2u - ad$ maximise $h_u^\star$ sur $\mathbb{R} \setminus [-ad - \sqrt{2\lambda}a, -ad + \sqrt{2\lambda}a]$ et

$$h_u^\star(v_2) = \frac{a^2u^2}{2} - aud + \frac{d^2}{2} + \lambda = \frac{1}{2}(au - d)^2 + \lambda. \tag{A.15}$$

Sinon le supremum de $h_u^\star$ sur $\mathbb{R} \setminus [-ad - \sqrt{2\lambda}a, -ad + \sqrt{2\lambda}a]$ est donné par (A.11).

D'après les deux cas précédents et en remarquant que les expressions (A.11) et (A.15) sont égales pour $|u| = \frac{\sqrt{2\lambda}}{a}$, il est évident que

$$g_0^{\star\star}(u) = \begin{cases} -adu + |u|\sqrt{2\lambda}a + \dfrac{d^2}{2} & \text{si } |u| \leqslant \dfrac{\sqrt{2\lambda}}{a}, \\ \dfrac{1}{2}(au - d)^2 + \lambda & \text{si } |u| \geqslant \dfrac{\sqrt{2\lambda}}{a}, \end{cases} \tag{A.16}$$

qui se réécrit encore

$$g_0^{\star\star}(u) = \frac{1}{2}(au - d)^2 + \lambda - \frac{a^2}{2}\left(|u| - \frac{\sqrt{2\lambda}}{a}\right)^2 \mathbb{1}_{\left\{|u| \leqslant \frac{\sqrt{2\lambda}}{a}\right\}}. \tag{A.17}$$

et complète la preuve.

A.1.2 *Preuve de la proposition 9.4*

Soit $D \in \mathbb{R}^{N\times N}$ une matrice diagonale dont les entrées diagonales sont définies par $d_i = \|a_i\|\ \forall i \in \mathbb{I}_N$. Notons que D^{-1} est bien définie par hypothèse que $\|a_i\| \neq 0\ \forall i \in \mathbb{I}_N$. Posons $\hat{d} = AD^{-2}A^Td$ et $\tilde{z} = D^{-1}A^Td$. Étant donné que A^TA est diagonale, le terme quadratique dans (9.6) peut être réécrit comme il suit :

$$\frac{1}{2}\|Ax - d\|^2 = \frac{1}{2}\|d - \hat{d}\|^2 + \frac{1}{2}\|Dx - \tilde{z}\|^2. \tag{A.18}$$

En combinant (9.6) et (A.18) on obtient,

$$\begin{aligned} G_{\ell_0}^\star(x^\star) &= \sup_{x\in\mathbb{R}^N} \langle x^\star, x\rangle_{\mathbb{R}^N} - \frac{1}{2}\|d - \hat{d}\|^2 - \frac{1}{2}\|Dx - \tilde{z}\|^2 - \lambda\|x\|_0, \\ &= -\frac{1}{2}\|d - \hat{d}\|^2 + \sup_{x\in\mathbb{R}^N} \sum_{i\in\mathbb{I}_N} x_i^\star x_i - \frac{1}{2}(\|a_i\|x_i - \tilde{z}_i)^2 - \lambda|x_i|_0, \\ &= -\frac{1}{2}\|d - \hat{d}\|^2 + \sum_{i\in\mathbb{I}_N} \sup_{x_i\in\mathbb{R}} x_i^\star x_i - \frac{1}{2}(\|a_i\|x_i - \tilde{z}_i)^2 - \lambda|x_i|_0. \end{aligned} \tag{A.19}$$

Finalement, lorsque les colonnes de A sont deux à deux orthogonales et non nulles, résoudre (9.6) est équivalent à résoudre N problèmes 1D indépendants. En utilisant la

fonction conjuguée obtenue dans le cas 1D (voir preuve proposition 9.3), donnée en (A.7), il vient

$$\begin{aligned} G^{\star}_{\ell_0}(x^{\star}) = &-\frac{1}{2}\|d-\hat{d}\|^2 \\ &+\sum_{i\in\mathbb{I}_N}\left\{\frac{1}{2\|a_i\|^2}\left(x_i^{\star}+\|a_i\|\tilde{z}_i\right)^2-\lambda\right\}\mathbb{1}_{\{|x_i^{\star}+\|a_i\|\tilde{z}_i|\geqslant\sqrt{2\lambda}\|a_i\|\}}-\frac{\tilde{z}_i^2}{2}. \end{aligned} \tag{A.20}$$

Enfin, le calcul de la biconjuguée se réduit également à la résolution de N problèmes 1D indépendants. En utilisant l'expression de l'enveloppe convexe 1D $g_0^{\star\star}$, donnée par la proposition 9.3 (eq. 9.4), on obtient

$$\begin{aligned} G^{\star\star}_{\ell_0}(x) &= \frac{1}{2}\|d-\hat{d}\|^2+\sum_{i\in\mathbb{I}_N}\frac{1}{2}\left(\|a_i\|x_i-\tilde{z}_i\right)^2+\phi(\|a_i\|,\lambda;x_i), \\ &= \frac{1}{2}\|d-\hat{d}\|^2+\frac{1}{2}\|Dx-\tilde{z}\|^2+\sum_{i\in\mathbb{I}_N}\phi(\|a_i\|,\lambda;x_i), \\ &\underset{(A.18)}{=} \frac{1}{2}\|Ax-d\|^2+\sum_{i\in\mathbb{I}_N}\phi(\|a_i\|,\lambda;x_i), \end{aligned} \tag{A.21}$$

ce qui complète la preuve.

A.1.3 *Preuve du lemme 9.10*

Tout d'abord, rappelons que par hypothèse nous avons

$$\forall i\in\mathbb{I}_N,\ \|a_i\|>0, \tag{A.22}$$

et les conditions (9.20) sont alors bien définies. Ensuite, étant donné que le terme d'attache aux données ($\frac{1}{2}\|Ax-d\|^2$) dans G_{CEL0} est différentiable, on a

$$\forall x\in\mathbb{R}^N,\ \partial G_{\text{CEL0}}(x)=A^T(Ax-d)+\partial\Phi_{\text{CEL0}}(x). \tag{A.23}$$

Soit $\hat{x}\in\mathbb{R}^N$ un point critique de G_{CEL0}, alors

$$0_{\mathbb{R}^N}\in\partial G_{\text{CEL0}}(\hat{x}), \tag{A.24}$$

qui, en suivant (A.23), se réécrit comme il suit :

$$0_{\mathbb{R}^N}\in\prod_{i\in\mathbb{I}_N}\left[A^T(A\hat{x}-d)\right]_i+\partial\phi(\|a_i\|,\lambda;\hat{x}_i).$$

En injectant l'expression du sous-différentiel $\partial\phi(\|a_i\|,\lambda;\hat{x}_i)$, donné par les équations (9.16) et (9.18), l'équation précédente devient, $\forall i\in\mathbb{I}_N$

$$\begin{cases} 0\in[\langle a_i,A\hat{x}-d\rangle-\sqrt{2\lambda}\|a_i\|,\langle a_i,A\hat{x}-d\rangle+\sqrt{2\lambda}\|a_i\|] & \text{ssi } \hat{x}_i=0, \\ 0=\langle a_i,A\hat{x}-d\rangle-\|a_i\|^2\hat{x}_i+\text{sign}(\hat{x}_i)\sqrt{2\lambda}\|a_i\| & \text{ssi } 0<|\hat{x}_i|\leqslant\dfrac{\sqrt{2\lambda}}{\|a_i\|}, \\ 0=\langle a_i,A\hat{x}-d\rangle & \text{ssi } |\hat{x}_i|\geqslant\dfrac{\sqrt{2\lambda}}{\|a_i\|}, \end{cases}$$

$$\iff \begin{cases} |\langle a_i, A\hat{x}^{(i)} - d\rangle| \leqslant \sqrt{2\lambda}\|a_i\| & \text{ssi } \hat{x}_i = 0, \\ 0 = \langle a_i, A\hat{x}^{(i)} - d\rangle + \operatorname{sign}(\hat{x}_i)\sqrt{2\lambda}\|a_i\| & \text{ssi } 0 < |\hat{x}_i| \leqslant \dfrac{\sqrt{2\lambda}}{\|a_i\|}, \\ 0 = \langle a_i, A\hat{x}^{(i)} - d\rangle + \|a_i\|^2\hat{x}_i & \text{ssi } |\hat{x}_i| \geqslant \dfrac{\sqrt{2\lambda}}{\|a_i\|}, \end{cases}$$

$$\iff \begin{cases} \hat{x}_i = 0 & \text{ssi } |\langle a_i, A\hat{x}^{(i)} - d\rangle| \leqslant \sqrt{2\lambda}\|a_i\|, \\ \hat{x}_i = -s_i t,\ t \in \left[0, \dfrac{\sqrt{2\lambda}}{\|a_i\|}\right] & \text{ssi } |\langle a_i, A\hat{x}^{(i)} - d\rangle| = \sqrt{2\lambda}\|a_i\|, \\ \hat{x}_i = -\dfrac{\langle a_i, A\hat{x}^{(i)} - d\rangle}{\|a_i\|^2} & \text{ssi } |\langle a_i, A\hat{x}^{(i)} - d\rangle| \geqslant \sqrt{2\lambda}\|a_i\|, \end{cases}$$

où $\hat{x}^{(i)} = (\hat{x}_1, \cdots, \hat{x}_{i-1}, 0, \hat{x}_{i+1}, \cdots, \hat{x}_N)$ et $s_i = \operatorname{sign}(\langle a_i, A\hat{x}^{(i)} - d\rangle) = -\operatorname{sign}(\hat{x}_i)$, ce qui termine la preuve.

A.1.4 *Preuve du lemme 9.13*

Soit $\hat{x} \in \mathbb{R}^N$ un minimiseur (local) de G_{CEL0}. À partir du résultat de la proposition 9.6, nous pouvons voir que

$$\forall i \in \sigma^+(\hat{x}),\ \forall t \in \left[0, \frac{\sqrt{2\lambda}}{\|a_i\|}\right],\ G^i_{\text{CEL0}}(-s_i t; \hat{x}^{(i)}) = G_{\text{CEL0}}(\hat{x}^{(i)} - s_i e_i t) = C, \tag{A.25}$$

où C est une constante de $\mathbb{R}_+$. En effet, étant donné que $\hat{x}$ est un minimiseur (local) de G_{CEL0}, il doit vérifier les conditions données en (9.20) (point critique) desquelles on déduit que

$$\forall i \in \sigma^+(\hat{x}),\ \langle a_i, A\hat{x}^{(i)} - d\rangle = s_i\sqrt{2\lambda}\|a_i\|. \tag{A.26}$$

Puisque que $s_i = \operatorname{sign}(\langle a_i, A\hat{x}^{(i)} - d\rangle)$ (voir preuve lemme 9.10).

Il s'en suit, d'après la proposition 9.6, que pour tout $i \in \sigma^+(\hat{x})$,

$$\forall t \in \left[0, \frac{\sqrt{2\lambda}}{\|a_i\|}\right],\ G^i_{\text{CEL0}}(-s_i t; \hat{x}^{(i)}) = -s_i t\langle a_i, Ax^{(i)} - d\rangle + t\,\|a_i\|\sqrt{2\lambda} + C_i \underset{(A.26)}{=} C_i, \tag{A.27}$$

où C_i est une constante, donnée par (9.11), indépendante de t ce qui prouve (A.25). Nous pouvons maintenant démontrer les points (i) et (ii) du lemme.

(i) Étant donné que $\hat{x}$ est un minimiseur (local) de G_{CEL0}, il existe $\varepsilon > 0$ tel que

$$\forall x \in \mathcal{B}_2(\hat{x}, \varepsilon),\ G_{\text{CEL0}}(\hat{x}) \leqslant G_{\text{CEL0}}(x). \tag{A.28}$$

où $\mathcal{B}_2(\hat{x}, \varepsilon)$ dénote la boule ouverte ℓ_2 de centre $\hat{x}$ et de rayon ε. Ainsi, à partir des équations (A.25) and (A.28), on obtient le résultat donné en (9.25). En effet, $\forall i \in \sigma^+(\hat{x})$, posons

$$T_i = \left\{ t \in \left[0, \frac{\sqrt{2\lambda}}{\|a_i\|}\right] : x = \hat{x}^{(i)} - s_i e_i t \in \mathcal{B}_2(\hat{x}, \varepsilon) \right\}, \tag{A.29}$$

Il est clair que comme $\varepsilon > 0$ et $|\hat{x}_i| \in \left[0, \frac{\sqrt{2\lambda}}{\|a_i\|}\right]$, T_i est un intervalle non-dégénéré de $\mathbb{R}$. Il s'en suit que

$$\forall t \in T_i,\ \exists \varepsilon' \in (0, \varepsilon),\ \text{ tel que } \mathcal{B}_2(\bar{x}, \varepsilon') \subset \mathcal{B}_2(\hat{x}, \varepsilon), \tag{A.30}$$

où $\bar{x} = \hat{x}^{(i)} - s_i e_i t$ et on a

$$\forall x \in \mathcal{B}_2(\bar{x}, \varepsilon'),\ G_{\text{CEL0}}(\bar{x}) \underset{\text{(A.25)}}{=} G_{\text{CEL0}}(\hat{x}) \underset{\text{(A.28)\&(A.30)}}{\leqslant} G_{\text{CEL0}}(x), \tag{A.31}$$

ce qui prouve l'assertion (i) du lemme.

(ii) En utilisant le fait que $\hat{x}$ est un minimiseur global de G_{CEL0}, (A.25) termine la preuve.

A.1.5 *Preuve du théorème 9.16*

Tout d'abord, notons que par définition de G_{CEL0} en (9.9), nous avons

$$G_{\text{CEL0}}(x) = G_{\ell_0}(x),\ \forall x \in \mathbb{R}^N \setminus S \ \text{ où } S := \left\{x \in \mathbb{R}^N \,:\, \sigma^-(x) \neq \emptyset\right\}. \tag{A.32}$$

(i) Soit $\hat{x} \in \mathbb{R}^N$ un minimiseur global de G_{ℓ_0} (il en existe au moins un d'après (Nikolova, 2013, Theorem 4.4)). La proposition 9.11 avec (A.32) nous assure alors que

$$G_{\text{CEL0}}(\hat{x}) = G_{\ell_0}(\hat{x}). \tag{A.33}$$

Supposons maintenant qu'il existe $\bar{x} \in \mathbb{R}^N$ tel que

$$G_{\text{CEL0}}(\bar{x}) < G_{\text{CEL0}}(\hat{x}) = G_{\ell_0}(\hat{x}). \tag{A.34}$$

— si $\bar{x} \in \mathbb{R}^N \setminus S$ alors, d'après (A.32),

$$G_{\ell_0}(\bar{x}) = G_{\text{CEL0}}(\bar{x}) < G_{\text{CEL0}}(\hat{x}) = G_{\ell_0}(\hat{x}), \tag{A.35}$$

qui est en contradiction avec le fait que $\hat{x}$ est un minimiseur global de G_{ℓ_0}.

— si $\bar{x} \in S$, prenons $i \in \sigma^-(\bar{x})$ et notons $G^i_{\text{CEL0}}(\cdot; \bar{x}^{(i)})$ la restriction de G_{CEL0} à la i-*ème* variable au point $\bar{x}$. La proposition 9.6 nous montre alors que $G^i_{\text{CEL0}}(\cdot; \bar{x}^{(i)})$ est convexe et on déduit facilement de l'expression (9.10) que

$$\forall t \in \mathbb{R},\ G^i_{\text{CEL0}}(t; \bar{x}^{(i)}) = \frac{1}{2}(\|a_i\| t - \tilde{d})^2 + \phi(\|a_i\|, \lambda; t) + C_i - \frac{\tilde{d}^2}{2}, \tag{A.36}$$

où C_i est une constante donnée en (9.11) et $\tilde{d} = \langle a_i, d - A\bar{x}^{(i)} \rangle / \|a_i\| \in \mathbb{R}$. Posons

$$\hat{t} \ \in \ \arg\min_{t \in \mathbb{R}} G^i_{\text{CEL0}}(t). \tag{A.37}$$

D'après le lemme 9.10 et (A.36), un tel $\hat{t}$ vérifie,

$$\hat{t} = \begin{cases} 0 & \text{si } |\tilde{d}| \leqslant \sqrt{2\lambda}, \\ \text{sign}(\tilde{d})t,\ t \in \left[0, \frac{\tilde{d}}{\|a_i\|}\right] & \text{si } |\tilde{d}| = \sqrt{2\lambda}, \\ \frac{\tilde{d}}{\|a_i\|} & \text{si } |\tilde{d}| \geqslant \sqrt{2\lambda}. \end{cases} \tag{A.38}$$

Définissons,

$$\bar{x}^1 = \bar{x}^{(i)} + e_i \frac{\tilde{d}}{\|a_i\|} \mathbb{1}_{\{|\tilde{d}|>\sqrt{2\lambda}\}}. \tag{A.39}$$

Clairement, $|\bar{x}_i^1| \in \{0, \tilde{d}/\|a_i\|\}$ et d'après (A.37) et (A.38) $G_{\text{CEL0}}(\bar{x}^1) \leqslant G_{\text{CEL0}}(\bar{x})$. De plus $\sharp\sigma^-(\bar{x}^1) = \sharp\sigma^-(\bar{x}) - 1$. Ce processus peut alors être répété pour un autre $i \in \sigma^-(\bar{x}^1) \subset \sigma^-(\bar{x})$ et nous pouvons ainsi construire une séquence $(\bar{x}^k)_{k\in\{1\cdots K\}}$ (où $K = \sharp\sigma^-(\bar{x})$) telle que,

$$G_{\text{CEL0}}(\bar{x}^K) \leqslant G_{\text{CEL0}}(\bar{x}^{K-1}) \leqslant \cdots \leqslant G_{\text{CEL0}}(\bar{x}^1) \leqslant G_{\text{CEL0}}(\bar{x}), \tag{A.40}$$

$$\sharp\sigma^-(\bar{x}^K) = \sharp\sigma^-(\bar{x}^{K-1}) - 1 = \cdots = \sharp\sigma^-(\bar{x}) - \sharp\sigma^-(\bar{x}) = 0. \tag{A.41}$$

Ensuite, d'après (A.32), (A.34), (A.40) et (A.41) on obtient

$$G_{\ell_0}(\bar{x}^K) = G_{\text{CEL0}}(\bar{x}^K) \leqslant G_{\text{CEL0}}(\bar{x}) < G_{\text{CEL0}}(\hat{x}) = G_{\ell_0}(\hat{x}), \tag{A.42}$$

qui est en contradiction avec le fait que $\hat{x}$ est un minimiseur global de G_{ℓ_0} et complète la preuve de l'assertion (i).

(ii) Soit $\hat{x} \in \mathbb{R}^N$ un minimiseur global de G_{CEL0}. D'après le lemme 9.13 (ii), $\hat{x}^0$, défini par (9.26), est aussi un minimiseur global de G_{CEL0} tel que $\sigma^-(\hat{x}^0) = \emptyset$ et, avec (A.32), nous avons

$$G_{\text{CEL0}}(\hat{x}) = G_{\text{CEL0}}(\hat{x}^0) = G_{\ell_0}(\hat{x}^0). \tag{A.43}$$

Étant donné que G_{CEL0} minore G_{ℓ_0} (remarque 9.5), $\hat{x}^0$ est un minimiseur global de G_{ℓ_0} ce qui termine la démonstration.

A.1.6 *Preuve du corolaire 9.19*

On note $S_0 = \arg\min_{x\in\mathbb{R}^N} G_{\ell_0}(x)$ et $S_{\text{CEL0}} = \arg\min_{x\in\mathbb{R}^N} G_{\text{CEL0}}(x)$.

$\Longrightarrow$ Tout d'abord, étant donné que les minimiseurs globaux de G_{ℓ_0} sont stricts (Nikolova, 2013, théorème 4.4 (ii)), l'égalité $S_0 = S_{\text{CEL0}}$ implique que ces minimiseurs globaux sont également stricts pour G_{CEL0}. Alors, pour $\hat{x} \in S_{\text{CEL0}} = S_0$, $\sigma^+(\hat{x}) = \emptyset$ d'après le corolaire 9.15.

Supposons maintenant qu'il existe deux minimiseurs $(\hat{x}^1, \hat{x}^2) \in S_{\text{CEL0}}^2$ tels que $\|\hat{x}^1 - \hat{x}^2\|_0 \leqslant 1$. Deux cas sont alors possibles, $\|\hat{x}^1 - \hat{x}^2\|_0 = 1$ et $\|\hat{x}^1 - \hat{x}^2\|_0 = 0$. Cependant seul le premier cas est intéressant étant que les minimiseurs sont stricts et donc que $\|\hat{x}^1 - \hat{x}^2\|_0 = 0 \Rightarrow \hat{x}^1 = \hat{x}^2$. Soit $i \in \mathbb{I}_N$ l'indice de la composante pour laquelle les solutions $\hat{x}^1$ et $\hat{x}^2$ diffèrent. Par définition, on a $\hat{x}_i^1 \neq \hat{x}_i^2$ et tout deux sont des minimiseurs globaux de la restriction convexe (proposition 9.6) $G_{\text{CEL0}}^i(\cdot; \hat{x}^{(i)})$. Il s'en suit que l'intervalle

$$\left[\min(\hat{x}_i^1, \hat{x}_i^2), \max(\hat{x}_i^1, \hat{x}_i^2)\right]$$

tout entier minimise $G_{\text{CEL0}}^i(\cdot; \hat{x}^{(i)})$ ce qui contredit le fait que $\hat{x}^1$ et $\hat{x}^2$ sont stricts pour G_{CEL0}. Ainsi, (9.32) est nécessairement vérifiée.

$\Longleftarrow$ Nous avons déjà l'inclusion $S_0 \subseteq S_{\text{CEL0}}$ avec le théorème 9.16 (i). Supposons maintenant qu'il existe $\hat{x} \in S_{\text{CEL0}}$ tel que $\hat{x} \notin S_0$. Alors, d'après le théorème 9.16 (ii), $\sigma^-(\hat{x}) \neq \emptyset$ et $\hat{x}^0 \in S_0 \cap S_{\text{CEL0}}$. D'autre part, pour $i \in \sigma^-(\hat{x})$, le lemme 9.13 nous assure que $\tilde{x} := \hat{x}^0 - s_i e_i \frac{\sqrt{2\lambda}}{\|a_i\|}$ est un minimiseur global de G_{CEL0} qui est donc aussi

un minimiseur global de G_{ℓ_0} d'après le théorème 9.16 (car $\sigma^-(\tilde{x}) = \emptyset$ par définition). On a donc $\tilde{x} \in S_0 \cap S_{\text{CEL0}}$ et

$$\|\hat{x}^0 - \tilde{x}\|_0 = 1,$$

ce qui est en contradiction avec (9.32). Ainsi, $\hat{x} \in S_0$ et termine la démonstration.

A.1.7 *Preuve du lemme 9.20*

Notons $\hat{\sigma} = \sigma(\hat{x})$, $\hat{\sigma}^- = \sigma^-(\hat{x})$ et $\hat{\sigma}^+ = \sigma^+(\hat{x})$. Prenons $i \in \hat{\sigma}^+$, alors comme $\hat{x}$ est un minimiseur (local) de G_{CEL0}, le lemme 9.13 nous assure l'existence d'un intervalle non-dégénéré de $\mathbb{R}$, noté T_i, tel que $|\hat{x}_i| \in T_i$ et tel que pour tout $t \in T_i$, $\bar{x} = \hat{x}^{(i)} - s_i e_i t$ est aussi un minimiseur (local) de G_{CEL0}. Soit $\bar{t} \in T_i \setminus \{|\hat{x}_i|\}$, alors $\bar{x} := \hat{x}^{(i)} - s_i e_i \bar{t}$ vérifie les conditions du lemme 9.10 (points critiques). Par construction,

$$\forall j \in \mathbb{I}_N \backslash \{i\},\ \hat{x}_j = \bar{x}_j. \tag{A.44}$$

Ensuite, les conditions du lemme 9.10 entrainent

$$\forall j \in \hat{\sigma}\backslash\{i\},\ \begin{cases} |\langle a_j, A\bar{x}^{(j)} - d\rangle| = \sqrt{2\lambda}\|a_j\| & \text{si } j \in \hat{\sigma}^-, \\ \bar{x}_j = -\frac{\langle a_j, A\bar{x}^{(j)} - d\rangle}{\|a_j\|^2} & \text{si } j \in \hat{\sigma}\backslash\hat{\sigma}^-, \end{cases} \tag{A.45}$$

D'après (A.44), $\forall j \in \mathbb{I}_N \backslash \{i\}$, $\bar{x}^{(j)} = \hat{x}^{(j)} - \hat{x}_i e_i + \bar{x}_i e_i$ et (A.45) peut se réécrire comme il suit :

$$\forall j \in \hat{\sigma}\backslash\{i\},\ \begin{cases} |\langle a_j, A\hat{x}^{(j)} - d\rangle + \langle a_j, a_i\rangle(\bar{x}_i - \hat{x}_i)| = \sqrt{2\lambda}\|a_j\| & \text{si } j \in \hat{\sigma}^-, \\ \bar{x}_j = -\frac{\langle a_j, A\hat{x}^{(j)} - d\rangle}{\|a_j\|^2} + \frac{\langle a_j, a_i\rangle}{\|a_j\|^2}(\hat{x}_i - \bar{x}_i) & \text{si } j \in \hat{\sigma}\backslash\hat{\sigma}^-. \end{cases} \tag{A.46}$$

Comme $\hat{x}$ est un point critique de G_{CEL0} on obtient, $\forall j \in \hat{\sigma}\backslash\{i\}$

$$\begin{cases} |s_j\sqrt{2\lambda}\|a_j\| + \langle a_j, a_i\rangle(\bar{x}_i - \hat{x}_i)| = \sqrt{2\lambda}\|a_j\| & \text{si } j \in \hat{\sigma}^-, \\ \bar{x}_j = \hat{x}_j + \frac{\langle a_j, a_i\rangle}{\|a_j\|^2}(\hat{x}_i - \bar{x}_i) & \text{si } j \in \hat{\sigma}\backslash\hat{\sigma}^-, \end{cases}$$

$$\overset{\Longrightarrow}{\hat{x}_i \neq \bar{x}_i \ \&\ (A.44)} \begin{cases} \langle a_j, a_i\rangle = 0 & \text{si } j \in \hat{\sigma}^-, \\ \langle a_j, a_i\rangle = 0 & \text{si } j \in \hat{\sigma}\backslash\hat{\sigma}^-, \end{cases} \tag{A.47}$$

qui, avec le fait que $\hat{\sigma}^- \subseteq \hat{\sigma}$, termine la démonstration.

A.1.8 *Preuve du théorème 9.21*

Posons $\hat{\sigma} = \sigma(\hat{x})$, $\hat{\sigma}^- = \sigma^-(\hat{x})$ et $\hat{\sigma}^0 = \sigma(\hat{x}^0)$. D'après le lemme 9.20, étant donné que $\hat{x}$ est un minimiseur de G_{CEL0}, nous avons (9.33). De plus, $\hat{x}$ est un point critique de G_{CEL0} et d'après le lemme 9.10 on obtient

$$\begin{aligned}
\forall i \in \hat{\sigma}^0 = \hat{\sigma} \setminus \hat{\sigma}^-,\ \hat{x}_i = \hat{x}_i^0 &= -\frac{\langle a_i, A\hat{x}^{(i)} - d\rangle}{\|a_i\|^2}, \\
&\iff \hat{x}_i^0 = -\frac{\langle a_i, A_{\hat{\sigma}^0}(\hat{x}^0_{\hat{\sigma}^0})^{(i)} + \sum_{j\in\hat{\sigma}^-} a_j\hat{x}_j - d\rangle}{\|a_i\|^2}, \\
&\underset{(9.33)}{\iff} \hat{x}_i^0 = -\frac{\langle a_i, A_{\hat{\sigma}^0}(\hat{x}^0_{\hat{\sigma}^0})^{(i)} - d\rangle}{\|a_i\|^2} - \frac{1}{\|a_i\|^2}\sum_{j\in\hat{\sigma}^-}\underbrace{\langle a_i, a_j\rangle}_{=0}\hat{x}_j, \\
&\iff \langle a_i, A_{\hat{\sigma}^0}\hat{x}^0_{\hat{\sigma}^0} - d\rangle = 0.
\end{aligned}$$

Ainsi, nous avons $(A_{\hat{\sigma}^0})^T A_{\hat{\sigma}^0}\hat{x}^0_{\hat{\sigma}^0} = (A_{\hat{\sigma}^0})^T d$ qui, avec (Nikolova, 2013, corolaire 2.5), assure que $\hat{x}^0$ est un minimiseur local de G_{ℓ_0}. Le fait que $\hat{x}^0$ ne soit pas global pour G_{ℓ_0} vient du fait qu'il n'est pas global pour G_{CEL0} (en vertu du théorème 9.16 (i)). Ensuite,

$$\begin{aligned}
G_{\text{CEL0}}(\hat{x}) &= \frac{1}{2}\|A_{(\hat{\sigma}^-)^c}\hat{x}_{(\hat{\sigma}^-)^c} + A_{\hat{\sigma}^-}\hat{x}_{\hat{\sigma}^-} - d\|^2 + \sum_{i\in\mathbb{I}_N}\phi(\|a_i\|,\lambda;\hat{x}_i), \\
&= \frac{1}{2}\|A_{(\hat{\sigma}^-)^c}\hat{x}_{(\hat{\sigma}^-)^c} - d\|^2 + \langle A_{\hat{\sigma}^-}\hat{x}_{\hat{\sigma}^-}, A_{(\hat{\sigma}^-)^c}\hat{x}_{(\hat{\sigma}^-)^c} - d\rangle \\
&\qquad + \frac{1}{2}\|A_{\hat{\sigma}^-}\hat{x}_{\hat{\sigma}^-}\|^2 + \sum_{i\in\mathbb{I}_N}\phi(\|a_i\|,\lambda;\hat{x}_i), \\
&= G_{\text{CEL0}}(\hat{x}^0) + \sum_{i\in\hat{\sigma}^-}\left(\phi(\|a_i\|,\lambda;\hat{x}_i) + \hat{x}_i\langle a_i, A_{(\hat{\sigma}^-)^c}\hat{x}_{(\hat{\sigma}^-)^c} - d\rangle + \frac{\hat{x}_i}{2}\langle a_i, A_{\hat{\sigma}^-}\hat{x}_{\hat{\sigma}^-}\rangle\right), \\
&\underset{(9.33)}{=} G_{\text{CEL0}}(\hat{x}^0) + \sum_{i\in\hat{\sigma}^-}\left(\phi(\|a_i\|,\lambda;\hat{x}_i) + \hat{x}_i\langle a_i, A\hat{x}^{(i)} - d\rangle + \frac{\|a_i\|^2}{2}\hat{x}_i^2\right),
\end{aligned}$$

et, comme $\hat{x}$ est un point critique de G_{CEL0}, nous avons

$$\begin{aligned}
G_{\text{CEL0}}(\hat{x}) = G_{\text{CEL0}}(\hat{x}^0) + \sum_{i\in\hat{\sigma}^-}\left(\phi(\|a_i\|,\lambda;\hat{x}_i) - \operatorname{sign}(\hat{x}_i)\sqrt{2\lambda}\|a_i\|\hat{x}_i + \frac{\|a_i\|^2}{2}\hat{x}_i^2\right) \\
= G_{\text{CEL0}}(\hat{x}^0). \quad \text{(A.48)}
\end{aligned}$$

Ici, nous avons utilisé la définition de ϕ en (9.8) entrainant

$$\forall i \in \hat{\sigma}^-,\ \phi(\|a_i\|,\lambda;\hat{x}_i) - \operatorname{sign}(\hat{x}_i)\sqrt{2\lambda}\|a_i\|\hat{x}_i + \frac{\|a_i\|^2}{2}\hat{x}_i^2 = 0, \quad \text{(A.49)}$$

Enfin, $G_{\text{CEL0}}(\hat{x}^0) = G_{\ell_0}(\hat{x}^0)$ provient des mêmes arguments que ceux utilisés dans la preuve du théorème 9.16 et termine la démonstration.

A.1.9 *Preuve du théorème 9.26*

Posons $\hat{\sigma} = \sigma(\hat{x})$ et $\hat{\sigma}^+ = \sigma^+(\hat{x})$. On procède alors en montrant les deux implications.

$\implies$ La preuve est directe d'après les corolaires 9.15 et 9.25.

$\Longleftarrow$ Soit $\hat{o}^+ = \emptyset$ et $\mathrm{rank}(A_{\hat{o}}) = \sharp\hat{o}$. On définit

$$\rho = \min\left\{ \min_{i\in\hat{o}} \left(|\hat{x}_i| - \frac{\sqrt{2\lambda}}{\|a_i\|} \right), \min_{i\in\hat{o}^c} \frac{\sqrt{2\lambda}}{\|a_i\|}, \min_{i\in\hat{o}^c} 2\frac{\sqrt{2\lambda}\|a_i\| - |\langle a_i, A\hat{x} - d\rangle|}{\|a_i\|^2} \right\}.$$

Clairement, $\rho > 0$ étant donné que,

$$\hat{o}^+ = \emptyset \Rightarrow (\forall i \in \hat{o},\ |\hat{x}_i| > \sqrt{2\lambda}/\|a_i\| \text{ et } \forall i \in \hat{o}^c,\ |\langle a_i, A\hat{x} - d\rangle| < \sqrt{2\lambda}\|a_i\|).$$

Pour tout $\varepsilon \in \mathcal{B}_\infty(0_{\mathbb{R}^N}, \rho)$ (boule ouverte ℓ_∞ de centre $0_{\mathbb{R}^N}$ et de rayon ρ), on a par définition de ρ,

$$\begin{aligned} \forall i \in \hat{o},\ |\hat{x}_i + \varepsilon_i| &> |\hat{x}_i| - \rho \\ &\geqslant |\hat{x}_i| - \min_{j\in\hat{o}} \left(|\hat{x}_j| - \frac{\sqrt{2\lambda}}{\|a_j\|} \right) \\ &\geqslant |\hat{x}_i| - |\hat{x}_i| + \frac{\sqrt{2\lambda}}{\|a_i\|} = \frac{\sqrt{2\lambda}}{\|a_i\|}, \end{aligned} \tag{A.50}$$

et

$$\forall i \in \hat{o}^c,\ |\hat{x}_i + \varepsilon_i| = |\varepsilon_i| < \rho \leqslant \min_{j\in\hat{o}^c} \frac{\sqrt{2\lambda}}{\|a_j\|} \leqslant \frac{\sqrt{2\lambda}}{\|a_i\|}. \tag{A.51}$$

Ainsi, d'après (A.50) et la définition de ϕ en (9.5), on a

$$\forall i \in \hat{o},\ \phi(\|a_i\|, \lambda; \hat{x}_i + \varepsilon_i) = \phi(\|a_i\|, \lambda; \hat{x}_i), \tag{A.52}$$

et il vient

$$\Phi_{\text{CEL0}}(\hat{x} + \varepsilon) = \sum_{i\in\hat{o}} \phi(\|a_i\|, \lambda; \hat{x}_i) + \sum_{i\in\hat{o}^c} \phi(\|a_i\|, \lambda; \varepsilon_i). \tag{A.53}$$

Dans la suite on utilisera la notation $\phi_i(\cdot) = \phi(\|a_i\|, \lambda; \cdot)$.
Prenons $\varepsilon \in \mathcal{B}_\infty(0_{\mathbb{R}^N}, \rho)\backslash\{0_{\mathbb{R}^N}\}$ et intéressons nous à la quantité $G_{\text{CEL0}}(\hat{x} + \varepsilon)$,

$$\begin{aligned} G_{\text{CEL0}}(\hat{x} + \varepsilon) &\underset{(A.53)}{=} \frac{1}{2}\|A(\hat{x} + \varepsilon) - d\|^2 + \sum_{i\in\hat{o}} \phi_i(\hat{x}_i) + \sum_{i\in\hat{o}^c} \phi_i(\varepsilon_i), \\ &= \frac{1}{2}\|A\hat{x} - d\|^2 + \frac{1}{2}\|A\varepsilon\|^2 + \langle A\varepsilon, A\hat{x} - d\rangle + \sum_{i\in\hat{o}} \phi_i(\hat{x}_i) + \sum_{i\in\hat{o}^c} \phi_i(\varepsilon_i), \\ &= G_{\text{CEL0}}(\hat{x}) + \frac{1}{2}\|A\varepsilon\|^2 + \langle \varepsilon_{\hat{o}}, (A_{\hat{o}})^T(A\hat{x} - d)\rangle \\ &\qquad + \langle \varepsilon_{\hat{o}^c}, (A_{\hat{o}^c})^T(A\hat{x} - d)\rangle + \sum_{i\in\hat{o}^c} \phi_i(\varepsilon_i). \end{aligned}$$

Étant donné que $\hat{x}$ est un point critique de G_{CEL0} et que $\hat{o}^+ = \emptyset$, le lemme 9.10 nous assure que $(A_{\hat{o}})^T(A\hat{x} - d) = 0_{\mathbb{R}^{\sharp\hat{o}}}$ et on a alors

$$G_{\text{CEL0}}(\hat{x} + \varepsilon) = G_{\text{CEL0}}(\hat{x}) + \frac{1}{2}\|A\varepsilon\|^2 + \sum_{i\in\hat{o}^c} \varepsilon_i\langle a_i, A\hat{x} - d\rangle + \phi_i(\varepsilon_i), \tag{A.54}$$

$$\geqslant G_{\text{CEL0}}(\hat{x}) + \frac{1}{2}\|A\varepsilon\|^2 + \sum_{i\in\hat{o}^c} \phi_i(\varepsilon_i) - |\varepsilon_i|\,|\langle a_i, A\hat{x} - d\rangle|. \tag{A.55}$$

Or d'après (A.51) nous savons que $\forall \varepsilon \in \mathcal{B}_\infty(0_{\mathbb{R}^N}, \rho)\backslash\{0_{\mathbb{R}^N}\},\ 0 \leqslant |\varepsilon_i| < \sqrt{2\lambda}/\|a_i\|\ \forall i \in \hat{\sigma}^c$ et par définition de $\phi_i(\cdot)$, (A.55) devient

$$G_{\text{CEL0}}(\hat{x}+\varepsilon) \geqslant G_{\text{CEL0}}(\hat{x}) + \frac{1}{2}\|A\varepsilon\|^2 + \sum_{i\in\hat{\sigma}^c} |\varepsilon_i| \left(\sqrt{2\lambda}\|a_i\| - \frac{\|a_i\|^2}{2}|\varepsilon_i| - |\langle a_i, A\hat{x}-d\rangle|\right). \quad \text{(A.56)}$$

De plus,

$$\forall i \in \hat{\sigma}^c,\ \sqrt{2\lambda}\|a_i\| - \frac{\|a_i\|^2}{2}|\varepsilon_i| - |\langle a_i, A\hat{x}-d\rangle| > 0, \quad \text{(A.57)}$$

$$\iff \forall i \in \hat{\sigma}^c,\ |\varepsilon_i| < \frac{2}{\|a_i\|^2}\left(\sqrt{2\lambda}\|a_i\| - |\langle a_i, A\hat{x}-d\rangle|\right). \quad \text{(A.58)}$$

ce qui est vrai par définition de ρ.

Finalement, étant donné que $\text{rank}(A_{\hat{\sigma}}) = \sharp\hat{\sigma}$ et $\varepsilon \neq 0_{\mathbb{R}^N}$, au moins une des deux assertions suivantes est vérifiée :

- $\exists i \in \hat{\sigma}$ tel que $|\varepsilon_i| > 0$ impliquant $\|A\varepsilon\| > 0$,
- $\exists i \in \hat{\sigma}^c$ tel que $|\varepsilon_i| > 0$ et la somme dans (A.56) est donc positive.

L'inégalité dans (A.56) est donc stricte :

$$G_{\text{CEL0}}(\hat{x}+\varepsilon) > G_{\text{CEL0}}(\hat{x}), \quad \text{(A.59)}$$

ce qui termine la preuve.

A.1.10 *Preuve du théorème 9.32*

Rappelons tout d'abord que pour $x \in \mathbb{R}^N$, nous avons (d'après (9.10)),

$$\forall i \in \mathbb{I}_N\ \forall t \in \mathbb{R},\ G^i_{\text{CEL0}}(t; x^{(i)}) = C_i + \frac{1}{2}(\|a_i\|t - \tilde{d}_i)^2 + \phi(\|a_i\|, \lambda; t), \quad \text{(A.60)}$$

où $\tilde{d}_i = \langle a_i, d - Ax^{(i)}\rangle/\|a_i\|$ et C_i est une constante indépendante de t donnée par (9.11). De manière similaire, il est facile de voir que

$$\forall i \in \mathbb{I}_N\ \forall t \in \mathbb{R},\ G^i_{\ell_0}(t; x^{(i)}) = C'_i + \frac{1}{2}(\|a_i\|t - \tilde{d}_i)^2 + \lambda|t|_0, \quad \text{(A.61)}$$

où $C'_i = \frac{1}{2}\|Ax^{(i)} - d\|^2 + \lambda\sum_{j\neq i}|x_j|_0$.

Ensuite, d'après la proposition 9.3, il vient que pour tout $i \in \mathbb{I}_N$ et $x \in \mathbb{R}^N$, $G^i_{\text{CEL0}}(\cdot; x^{(i)})$ n'est autre que l'enveloppe convexe de $G^i_{\ell_0}(\cdot; x^{(i)})$ à une constante $C = C'_i - C_i \geqslant 0$ près. Dans la suite de la preuve, la notation (EC), pour Enveloppe Convexe, réfèrera à cette propriété.

$\Longrightarrow$ Soit $\hat{x} \in \mathbb{R}^N$ un minimiseur CW de G_{ℓ_0}. Alors, d'après la définition 9.29, pour tout $i \in \mathbb{I}_N$, $\hat{x}_i$ est un minimiseur global de $G^i_{\ell_0}(\cdot; \hat{x}^{(i)})$. Il découle alors de (EC) que pour tout $i \in \mathbb{I}_N$, $\hat{x}_i$ est aussi un minimiseur global de $G^i_{\text{CEL0}}(\cdot; \hat{x}^{(i)})$ étant donné qu'un minimiseur global d'une fonction est également un minimiseur global de son enveloppe convexe. Par conséquent, $\hat{x}$ est un minimiseur CW de G_{CEL0}.

Supposons maintenant que $\sigma^-(\hat{x}) \neq \emptyset$ et prenons $j \in \sigma^-(\hat{x})$. Alors, par définition de σ^- et de la pénalité CEL0 (9.5), nous avons

$$G^j_{\text{CEL0}}(\hat{x}_j; \hat{x}^{(j)}) + C'_j - C_j < G^j_{\ell_0}(\hat{x}_j; \hat{x}^{(j)}), \tag{A.62}$$

ce qui est en contradiction [1] avec (EC) et prouve que $\sigma^-(\hat{x}) = \emptyset$.

$\Longleftarrow$ Soit $\hat{x} \in \mathbb{R}^N$ un minimiseur CW de G_{CEL0} tel que $\sigma^-(\hat{x}) = \emptyset$. Alors, par définition de σ^- et de Φ_{CEL0}, nous avons

$$G_{\text{CEL0}}(\hat{x}) = G_{\ell_0}(\hat{x}). \tag{A.63}$$

Ensuite, avec des arguments similaires à ceux utilisés pour la preuve de $\Longrightarrow$, (EC) termine la démonstration.

A.2 DÉMONSTRATIONS DU CHAPITRE 10

A.2.1 *Preuve du théorème 10.3*

Dans cette preuve, nous considérons la restriction de G_{CEL0} définie par :

$$G^{\mathcal{E}}_{\text{CEL0}} : \mathcal{E} \longrightarrow \mathbb{R}, \tag{A.64}$$

où $\mathcal{E} := \{x \in \mathbb{R}^N,\ G_{\text{CEL0}}(x) < G_{\text{CEL0}}(x^{init})\} \subset \mathbb{R}^N$. Notons que puisque le Macro-Algo est un algorithme de descente, ses itérés sont contraints à rester dans $\mathcal{E}$ et nous pouvons donc limiter la preuve à l'étude de la restriction $G^{\mathcal{E}}_{\text{CEL0}}$.

On note C l'ensemble des points critiques (au sens de Clarke) de $G^{\mathcal{E}}_{\text{CEL0}}$. Nous commençons par montrer que l'image de C par $G^{\mathcal{E}}_{\text{CEL0}}$, notée $G^{\mathcal{E}}_{\text{CEL0}}(C)$, est un ensemble fini. Soit $\hat{x} \in C$, $\hat{\sigma} = \sigma(\hat{x})$ et $\hat{\sigma}^- = \sigma^-(\hat{x})$. Alors, d'après la première accolade dans la preuve du lemme 9.10 (annexe A.1.3 page 190), $\hat{x}$ est solution de

$$\begin{aligned}
\forall i \in \hat{\sigma} \begin{cases} 0 = \langle a_i, A\hat{x} - d\rangle - \|a_i\|^2 \hat{x}_i + \operatorname{sign}(\hat{x}_i)\sqrt{2\lambda}\|a_i\| & \text{ssi } i \in \hat{\sigma}^-, \\ 0 = \langle a_i, A\hat{x} - d\rangle & \text{ssi } i \in \hat{\sigma} \setminus \hat{\sigma}^-. \end{cases} \\
\Longleftrightarrow \begin{cases} 0 = (A_{\hat{\sigma}^-})^T(A_{\hat{\sigma}}\hat{x}_{\hat{\sigma}} - d) - D_{\hat{\sigma}^-}\hat{x}_{\hat{\sigma}^-} + \sqrt{2\lambda} b_{\hat{\sigma}^-}, \\ 0 = (A_{\hat{\sigma}\setminus\hat{\sigma}^-})^T(A_{\hat{\sigma}}\hat{x}_{\hat{\sigma}} - d). \end{cases}
\end{aligned} \tag{A.65}$$

où $D \in \mathbb{R}^{N\times N}$ est une matrice diagonale telle que $D_{ii} = \|a_i\|^2 \mathbb{1}_{\{i\in\hat{\sigma}^-\}}$ et $b \in \mathbb{R}^N$ est un vecteur défini par $b_i = \operatorname{sign}(\hat{x}_i)\|a_i\|\mathbb{1}_{\{i\in\hat{\sigma}^-\}}$. Alors (A.65) se réécrit

$$[(A_{\hat{\sigma}})^T A_{\hat{\sigma}} - D_{\hat{\sigma}}]\hat{x}_{\hat{\sigma}} = (A_{\hat{\sigma}})^T d - \sqrt{2\lambda} b_{\hat{\sigma}}. \tag{A.66}$$

Ainsi, appartenir à $\mathcal{E}$ et être solution du système (A.66) est une condition *nécessaire* [2] pour être un point critique de $G^{\mathcal{E}}_{\text{CEL0}}$. Considérons maintenant $\omega \subset \mathbb{I}_N$, $\omega^- \subset \omega$ et $\bar{x} \in \mathbb{R}^N$ solution de

$$\left[(A_\omega)^T A_\omega - D_\omega\right]\bar{x}_\omega = (A_\omega)^T d - \sqrt{2\lambda} b_\omega, \tag{A.67}$$

1. Le min d'une fonction coïncide avec celui de son enveloppe convexe.
2. Mais pas suffisante étant donné que, d'après le lemme 9.10, un point critique de G_{CEL0} doit aussi vérifier $|\langle a_i, A\hat{x}^{(i)} - d\rangle| \leqslant \sqrt{2\lambda}\|a_i\|$, $\forall i \in \hat{\sigma}^c$, condition qui n'est pas prise en compte dans (A.66).

où $b_i = \pm\|a_i\|\mathbb{1}_{\{i\in\omega^-\}}$ et $D_{ii} = \|a_i\|^2\mathbb{1}_{\{i\in\omega^-\}}$. Alors un tel $\bar{x}$ est point critique de $G^{\mathcal{E}}_{\text{CEL0}}$ *si et seulement si* $\bar{x} \in \mathcal{E}$, est solution de (A.66) (i.e. $\sigma(\bar{x}) = \omega$, $\sigma^-(\bar{x}) = \omega^-$ et $\text{sign}(\bar{x}_i) = \text{sign}(b_i)$, $\forall i \in \omega^-$) et vérifie $|\langle a_i, A\bar{x}^{(i)} - d\rangle| \leqslant \sqrt{2\lambda}\|a_i\|$, $\forall i \in \sigma(\bar{x})^c$.

Un nombre fini de systèmes du type (A.67) peuvent être construits en choisissant, ω, ω^- ainsi que le signe des composantes non-nulles de b. En effet, le nombre de sous-ensembles $\omega \subset \mathbb{I}_N$ est égal à $\sum_{k=1}^{N}\binom{N}{k}$ et le nombre de sous-ensembles $\omega^- \subset \omega$ à $\sum_{l=1}^{\#\omega}\binom{\#\omega}{l}$. Enfin, il y a le choix du signe des entrées non-nulles de b, i.e. $2^{\#\omega^-}$ possibilités. Au final, il existe

$$\sum_{k=1}^{N}\binom{N}{k} \times \sum_{l=1}^{k}\binom{k}{l} \times (2^l) < +\infty, \tag{A.68}$$

différents systèmes du type (A.67). Ce nombre est extrêmement grand mais *fini*. Rappelons que les solutions, $\bar{x}$, de ces systèmes sont des points critiques de $G^{\mathcal{E}}_{\text{CEL0}}$ *si et seulement si* elles appartiennent à $\mathcal{E}$, elles sont solution de (A.66) et vérifient $|\langle a_i, A\bar{x}^{(i)} - d\rangle| \leqslant \sqrt{2\lambda}\|a_i\|$, $\forall i \in \sigma(\bar{x})^c$. De fait, le nombre de systèmes comme (A.67) résultants en des points critiques de $G^{\mathcal{E}}_{\text{CEL0}}$ est inférieur à (A.68).

De tels systèmes (A.67) peuvent admettre
— une unique solution si $\text{rank}((A_\omega)^T A_\omega - D_\omega) = \#\omega$;
— aucune ou une infinité de solutions si $\text{rank}((A_\omega)^T A_\omega - D_\omega) \leqslant \#\omega - 1$.
Dans le cas où le système admet une infinité de solutions, elles appartiennent à un sous-espace affine de $\mathbb{R}^N$ (translation du sous-espace vectoriel défini par les solutions de l'équation homogène). Parmi ces solutions, les points critiques de $G^{\mathcal{E}}_{\text{CEL0}}$ appartiennent à l'intersection entre le sous-espace affine des solutions, $\mathcal{E}$ et E défini par

$$E := I_0 \cap \prod_{i\in\mathbb{I}_N} I_i, \tag{A.69}$$

où,

$$I_i = \begin{cases} \{0\} & \text{si } i \in \omega^c, \\ \left[-\frac{\sqrt{2\lambda}}{\|a_i\|}, 0\right[, & \text{si } i \in \omega^- \text{ and } b_i < 0, \\ \left]0, \frac{\sqrt{2\lambda}}{\|a_i\|}\right], & \text{si } i \in \omega^- \text{ and } b_i > 0, \\ \left]-\infty, -\frac{\sqrt{2\lambda}}{\|a_i\|}\right] \cup \left[\frac{\sqrt{2\lambda}}{\|a_i\|}, +\infty\right[& \text{si } i \in \omega\backslash\omega^-, \end{cases} \tag{A.70}$$

et

$$I_0 = \bigcap_{i\in\omega^c} \left\{ x \in \mathbb{R}^N : |\langle a_i, Ax^{(i)} - d\rangle| \leqslant \sqrt{2\lambda}\|a_i\| \right\}, \tag{A.71}$$

On peut voir que E (eq. A.69) est composé d'un nombre fini de composantes connexes dans $\mathbb{R}^N$. En effet, comme $x \mapsto \langle a_i, Ax^{(i)} - d\rangle$ est linéaire, chaque ensemble de l'intersection définissant I_0 est une composante connexe dans $\mathbb{R}^N$. Par conséquent I_0 est aussi une composante connexe dans $\mathbb{R}^N$. Finalement, intersecter I_0 avec le produit cartésien des I_i ($i \in \mathbb{I}_N$) prouve l'affirmation précédente.

Un système du type (A.67) résulte (au plus) en un nombre *fini* de composantes connexes dans C. En combinant cela avec (A.68), on en déduit que C contient *un nombre fini de com-*

posantes connexes. Ensuite, en suivant (Attouch et al., 2013), étant donné que $\frac{1}{2}\|Ax-d\|^2$ est une fonction polynomiale et que Φ_{CEL0} a un graphe polynomial par morceaux, $G^{\mathcal{E}}_{\text{CEL0}}$ est *semi-algébrique* et donc *sous-analytique*. Nous avons donc toutes les conditions requises par (Bolte et al., 2005, théorème 7) stipulant que $G^{\mathcal{E}}_{\text{CEL0}}$ est alors constante sur chacune de ces composantes connexes de C. Par conséquent $G^{\mathcal{E}}_{\text{CEL0}}(C)$ est un *ensemble fini*.

Nous sommes maintenant en mesure de terminer la démonstration. D'après H1 et par construction de l'algorithme, $\forall n \in \mathbb{N},\ x^n \in C$. Alors, comme cela est démontré dans la preuve du lemme 9.13 (annexe A.1.4 page 191), pour $i \in \sigma_n^-$,

$$G^{\mathcal{E}}_{\text{CEL0}}((x^n)^{(i)}) = G^{\mathcal{E}}_{\text{CEL0}}(x^n). \tag{A.72}$$

et $G^{\mathcal{E}}_{\text{CEL0}}(x^{temp}) = G^{\mathcal{E}}_{\text{CEL0}}(x^n)$. De plus, avec les hypothèses H1 et H2 sur Alg, on a d'une part

$$\forall n \in \mathbb{N},\ G^{\mathcal{E}}_{\text{CEL0}}(x^{n+1}) \leqslant G^{\mathcal{E}}_{\text{CEL0}}(x^{temp}) - \beta\|x^{n+1}-x^{temp}\|^2, \tag{A.73}$$

et d'autre part que x^{n+1} est un point critique de $G^{\mathcal{E}}_{\text{CEL0}}$. Nous distinguons deux cas :

— si $x^{n+1} = x^{temp}$, alors x^{temp} était un point critique et la boucle interne du Macro-Algo a été stoppée par la condition $\sigma_n^- = \emptyset$. Dans ce cas, l'algorithme s'arrête.
— si $x^{n+1} \neq x^{temp}$ nous avons, d'après (A.73),

$$G^{\mathcal{E}}_{\text{CEL0}}(x^{n+1}) < G^{\mathcal{E}}_{\text{CEL0}}(x^{temp}) = G^{\mathcal{E}}_{\text{CEL0}}(x^n). \tag{A.74}$$

Enfin, pour tout minimiseur global $\hat{x} \in C$ (il en existe au moins un d'après la proposition 9.18), le lemme 9.13 assure que $\hat{x}^0$ est aussi un minimiseur global vérifiant $\sigma^-(\hat{x}^0) = \emptyset$. Ainsi, il existe au moins un élément $\bar{x} \in C$ vérifiant $\sigma^-(\bar{x}) = \emptyset$ et $G^{\mathcal{E}}_{\text{CEL0}}(\bar{x}) = \min\{G^{\mathcal{E}}_{\text{CEL0}}(C)\}$. Ce dernier point avec (A.74) et le fait que $G^{\mathcal{E}}_{\text{CEL0}}(C)$ est un ensemble fini permet de conclure qu'il existe $n^\star \in \mathbb{N}$ tel que $\sigma_{n^\star}^- = \emptyset$.

Finalement, d'après le lemme 10.1, $x^{n^\star}$ est un minimiseur (local) de G_{ℓ_0} ce qui termine la preuve.

A.3 Démonstrations du chapitre 12

A.3.1 *Preuve du lemme 12.5*

Soit $u_d^\star$ un minimiseur global de g_0 pour $d \in \mathbb{R}$. Alors la proposition 12.3 établie que

$$\forall d \in \mathbb{R}, \begin{cases} |d| \leqslant \sqrt{2\lambda} \implies u_d^\star = 0, \\ |d| \geqslant \sqrt{2\lambda} \implies u_d^\star = \frac{d}{a}. \end{cases} \tag{A.75}$$

De plus, d'après la proposition 12.4, nous avons les deux équivalences suivantes :

$$0 \text{ est un point critique de } \tilde{g} \iff \begin{cases} ad \in \left[\underline{\delta}^0, \bar{\delta}^0\right] & \text{si } 0 \in B, \\ \phi'(0) = ad & \text{si } 0 \in \mathbb{R}\setminus B, \end{cases} \tag{A.76}$$

$$\frac{d}{a} \text{ est un point critique de } \tilde{g} \iff \begin{cases} 0 \in \left[\underline{\delta}^{d/a}, \bar{\delta}^{d/a}\right] & \text{si } \frac{d}{a} \in B, \\ \phi'\left(\frac{d}{a}\right) = 0 & \text{si } \frac{d}{a} \in \mathbb{R} \setminus B. \end{cases} \tag{A.77}$$

Alors, il découle des deux équations précédentes que

$$\{\forall d \in \mathbb{R},\ u^\star \text{ est un minimiseur global de } g_0 \implies u^\star \text{ est un point critique de } \tilde{g}\} \tag{A.78}$$

est équivalent à : $\forall d \in \mathbb{R}$,

$$|d| \leqslant \sqrt{2\lambda} \Longrightarrow \begin{cases} ad \in \left[\underline{\delta}^0, \bar{\delta}^0\right] & \text{si } 0 \in B, \\ \phi'(0) = ad & \text{si } 0 \in \mathbb{R} \setminus B. \end{cases} \tag{A.79}$$

$$|d| \geqslant \sqrt{2\lambda} \Longrightarrow \begin{cases} 0 \in \left[\underline{\delta}^{d/a}, \bar{\delta}^{d/a}\right] & \text{si } \frac{d}{a} \in B, \\ \phi'\left(\frac{d}{a}\right) = 0 & \text{si } \frac{d}{a} \in \mathbb{R} \setminus B. \end{cases} \tag{A.80}$$

En gardant à l'esprit que (A.79) est vérifiée $\forall d \in \mathbb{R}$, nous pouvons la réécrire comme il suit,

$$\begin{cases} [-\sqrt{2\lambda}a, \sqrt{2\lambda}a] \subseteq \left[\underline{\delta}^0, \bar{\delta}^0\right] & \text{si } 0 \in B, \\ \phi'(0) = ad \quad \forall |d| \leqslant \sqrt{2\lambda} & \text{si } 0 \in \mathbb{R} \setminus B. \end{cases} \tag{A.81}$$

La seconde ligne de (A.81) est impossible pour ϕ fixé. Ainsi (A.79) est équivalent à (12.19a).

Similairement, nous pouvons réécrire (A.80) par

$$\forall |u| \geqslant \sqrt{2\lambda}/a, \begin{cases} 0 \in \left[\underline{\delta}^u, \bar{\delta}^u\right] & \text{si } u \in B, \\ \phi'(u) = 0 & \text{si } u \in \mathbb{R} \setminus B. \end{cases} \tag{A.82}$$

Le fait que B contienne un nombre fini de points de $\mathbb{R}$ et par continuité de ϕ, seulement la deuxième ligne de (A.82) peut être vérifiée pour $|u| > \sqrt{2\lambda}/a$. En effet, supposons qu'il existe $u \in B$ tel que $u > \sqrt{2\lambda}/a$ (on peut faire la même chose pour $u < -\sqrt{2\lambda}/a$). Alors étant donné que B est un ensemble fini de points de $\mathbb{R}$, il contient uniquement des points isolés, i. e.

$$\exists \varepsilon > 0, \quad \text{t.q.} \quad B \cap]u - \varepsilon, u + \varepsilon[= \{u\}. \tag{A.83}$$

Il s'en suit, d'après la deuxième ligne de (A.82), que

$$\forall v \in \left]\max\left(\sqrt{2\lambda}/a, u - \varepsilon\right), u + \varepsilon\right[\setminus \{u\},\ \phi'(v) = 0, \tag{A.84}$$

et que ϕ est constante sur $]\max(\sqrt{2\lambda}/a, u - \varepsilon), u[$ et sur $]u, u + \varepsilon[$. Alors par continuité, ϕ est constante sur $]\max(\sqrt{2\lambda}/a, u - \varepsilon), u + \varepsilon[$ tout entier ce qui est en contradiction avec le fait que $u \in B$ (i. e. que ϕ n'est pas différentiable en u). Ainsi, $B \subseteq [-\sqrt{2\lambda}/a, \sqrt{2\lambda}/a]$ et $\forall u \in \mathbb{R} \setminus [-\sqrt{2\lambda}/a, \sqrt{2\lambda}/a]$, $\phi'(u) = 0$ ce qui termine la preuve.

A.3.2 *Preuve du lemme 12.6*

Étant donné que nous considérons (12.6), «$\tilde{g}$ n'admet aucun minimiseur global sur $]-\frac{\sqrt{2\lambda}}{a},0[\cup]0,\frac{\sqrt{2\lambda}}{a}[$ pour tout $d\in\mathbb{R}$» est équivalent à,

$$\forall u\in]-\sqrt{2\lambda}/a,0[\cup]0,\sqrt{2\lambda}/a[,\ \tilde{g}(u)>\begin{cases} g_0(0)=\frac{d^2}{2} & \text{si } |d|\leqslant\sqrt{2\lambda},\\ g_0\left(\frac{d}{a}\right)=\lambda & \text{si } |d|\geqslant\sqrt{2\lambda},\end{cases}\quad\forall d\in\mathbb{R},\tag{A.85}$$

ce qui s'écrit encore,

$$\forall u\in]-\sqrt{2\lambda}/a,0[\cup]0,\sqrt{2\lambda}/a[,\ \phi(u)>\phi_{\min}(u):=\sup_{d\in\mathbb{R}} f_u(d),\tag{A.86}$$

où

$$f_u(d)=\begin{cases} -\frac{a^2u^2}{2}+aud & \text{si } |d|\leqslant\sqrt{2\lambda},\\ \lambda-\frac{1}{2}(au-d)^2 & \text{si } |d|\geqslant\sqrt{2\lambda}.\end{cases}\tag{A.87}$$

Alors, il est aisé de voir que

$$\sup_{|d|\leqslant\sqrt{2\lambda}} f_u(d)=-\frac{a^2u^2}{2}+\sqrt{2\lambda}a|u|,\tag{A.88}$$

et

$$\sup_{|d|\geqslant\sqrt{2\lambda}} f_u(d)=\begin{cases} \lambda & \text{if } |u|\geqslant\frac{\sqrt{2\lambda}}{a},\\ -\frac{a^2u^2}{2}+\sqrt{2\lambda}a|u| & \text{sinon},\end{cases}\tag{A.89}$$

Finalement nous avons

$$\forall u\in]-\sqrt{2\lambda}/a,0[\cup]0,\sqrt{2\lambda}/a[,\ \phi_{\min}(u)=-\frac{a^2u^2}{2}+\sqrt{2\lambda}a|u|=\phi_{\text{CEL0}}(a,\lambda;u),\tag{A.90}$$

où ϕ_{CEL0} est donné par (9.5) (page 96), ce qui termine la preuve.

A.3.3 *Preuve du théorème 12.7*

Nous démontrons chaque implication indépendamment.

$\Longleftarrow$ remarquons tout d'abord que (12.21) $\Longrightarrow$ (12.19). En effet, il est évident que (12.21b) $\Longrightarrow$ (12.19b) et, par définition de ϕ_{CEL0}, nous avons $\phi_{\text{CEL0}}(0)=0$ et

$$\lim_{\substack{u\to0\\u>0}}\phi'_{\text{CEL0}}(u)=\sqrt{2\lambda}a\ \text{ et }\ \lim_{\substack{u\to0\\u<0}}\phi'_{\text{CEL0}}(u)=-\sqrt{2\lambda}a.\tag{A.91}$$

Ainsi, (A.91) avec (12.21a) et (12.21c) (i.e. $\phi(0)=0$ et $\phi>\phi_{\text{CEL0}}$ sur $]-\sqrt{2\lambda}/a,\sqrt{2\lambda}/a[\backslash\{0\}$) impliquent

$$\lim_{\substack{u\to0\\u<0}}\phi'(u)\leqslant-\sqrt{2\lambda}a\ \text{ et }\ \lim_{\substack{u\to0\\u>0}}\phi'(u)\geqslant\sqrt{2\lambda}a,\tag{A.92}$$

qui peut être réécrit comme $[-\sqrt{2\lambda}a, \sqrt{2\lambda}a] \subseteq [\underline{\delta}^0, \bar{\delta}^0]$ et montre que (12.19a) est vérifiée (le fait que $0 \in B$ est une conséquence de $[-\sqrt{2\lambda}a, \sqrt{2\lambda}a] \subseteq [\underline{\delta}^0, \bar{\delta}^0]$).
Il s'en suit, d'après le lemme 12.5, que les minimiseurs globaux de g_0 sont des points critiques de $\tilde{g}$ et que $\tilde{g}$ a au moins un point critique puisque g_0 a toujours au moins un minimiseur global dans $\{0, d/a\}$ (proposition 12.3). Alors, parmi les points critiques de $\tilde{g}$ il en existe au moins un qui est un minimiseur global de $\tilde{g}$. En effet, de part la continuité de ϕ et d'après les conditions (12.21a) et (12.21b), ϕ est bornée. De plus, l'attache aux données quadratique de $\tilde{g}$ est coercive. Ainsi $\tilde{g}$ est coercive et sa contiuité assure l'existence d'un minimiseur global.
Maintenant, d'après (12.21a) et (12.21b) nous avons que

$$\forall u \notin]-\sqrt{2\lambda}/a, 0[\cup]0, \sqrt{2\lambda}/a[,\ \phi(u) = \lambda|u|_0. \tag{A.93}$$

Par ailleurs, (12.21c) avec le lemme 12.6 assurent que $\tilde{g}$ n'admet pas de minimiseurs globaux sur $]-\sqrt{2\lambda}/a, 0[\cup]0, \sqrt{2\lambda}/a[$ et la proposition 12.3 montre la même chose pour g_0. Ainsi, nous venons de montrer que g_0 et $\tilde{g}$ étaient égales en tout point qui est potentiellement un minimiseur global de l'une des deux fonctionnelles ce qui nous permet de conclure que (P1) est vérifiée (avec (12.6)).

$\Longrightarrow$ sous la condition (P1), tous les minimiseurs globaux de g_0 sont des points critiques de $\tilde{g}$ ce qui, d'après le lemme 12.5, est équivalent à (12.19). De plus, (12.6) entraine (12.21a) et permet de réduire (12.19b) à (12.21b).
Supposons que (12.21c) n'est pas vérifiée et qu'il existe $u_0 \in (0, \sqrt{2\lambda}/a)$ tel que $\phi(u_0) \leqslant \phi_{\text{CEL0}}(u_0)$. Alors, il est facile de déterminer un $d_0 = \sqrt{2\lambda}$ pour lequel g_0 admet deux minimiseurs globaux $\{0, \frac{d_0}{a}\}$ et pour lequel l'intervalle $[0, \frac{d_0}{a}] = [0, \frac{\sqrt{2\lambda}}{a}]$ tout entier minimise la fonctionnelle g_{CEL0} (définie par (12.5) avec $\phi = \phi_{\text{CEL0}}$) puisque g_{CEL0} est l'enveloppe convexe de g_0 (voir section 9.1 page 95). Ainsi nous avons $u_0 \notin \{0, \frac{d_0}{a}\}$ et néanmoins $\tilde{g}(u_0) \leqslant g_{\text{CEL0}}(u_0) = g_0(0) = g_0(\frac{d_0}{a})$ ce qui contredit (P1) et termine la preuve.

A.3.4 *Preuve du théorème 12.9*

Étant donné que g_0 a toujours deux minimiseurs (locaux), $\hat{u}_1 = 0$ et $\hat{u}_2 = \frac{d}{a}$, qui coïncident lorsque $d = 0$, nous pouvons voir que vérifier (P2) est équivalent à vérifier

$$\forall d \in \mathbb{R},\ \hat{u} \in C \setminus \left\{0, \frac{d}{a}\right\} \Longrightarrow \hat{u} \text{ n'est } \textbf{pas} \text{ un minimiseur (local) de } \tilde{g}, \tag{A.94}$$

où $C = \{u \in \mathbb{R} : 0 \in \partial\tilde{g}(u)\}$ est l'ensemble des points critiques de $\tilde{g}$. Ainsi, la démonstration consiste maintenant à montrer que, sous les conditions (12.21), (A.94) est équivalent à (12.22).

(12.22) $\Longrightarrow$ (A.94) Soit $\hat{u} \in C \setminus \{0, \frac{d}{a}\}$ pour $d \in \mathbb{R}$. Alors, d'après la caractérisation des points critiques de $\tilde{g}$ donnée par (12.13) et la définition des bornes β^- et β^+, nous en déduisons que $\hat{u} \in [\beta^-, \beta^+] \setminus \{0\}$. En effet, un point critique de $\tilde{g}$ sur la partie constante de ϕ vérifie nécessairement $\hat{u} = d/a$ d'après (12.13). De plus, d'après (12.22) on a :
— si $\hat{u} \in B$, alors

$$\lim_{\substack{v \to \hat{u} \\ v < \hat{u}}} \tilde{g}'(v) = a^2\hat{u} - ad + \lim_{\substack{v \to \hat{u} \\ v < \hat{u}}} \phi'(v) \underset{(12.22a)}{>} a^2\hat{u} - ad + \lim_{\substack{v \to \hat{u} \\ v > \hat{u}}} \phi'(v) = \lim_{\substack{v \to \hat{u} \\ v > \hat{u}}} \tilde{g}'(v),$$

montrant que $\hat{u}$ n'est pas un minimiseur (local) de $\tilde{g}$.

— si $\hat{u} \notin B$, alors puisque $\hat{u} \neq d/a$ nous avons $\phi'(\hat{u}) \neq 0$ et nécessairement $\hat{u} \in]\beta^-, \beta^+[\setminus B$ (en effet si $\hat{u} = \beta^-$, ou β^+, alors par (12.21) et étant donné que nous sommes dans le cas $\hat{u} \notin B$, nous obtenons $\phi'(\beta^-) = 0$ ce qui est incompatible avec ce qui précède impliquant $\hat{u} \neq \beta^-$ et β^+). Enfin, d'après (12.22b) on obtient

$$\begin{cases} \tilde{g}''(\hat{u}) = a^2 + \phi''(\hat{u}) \leqslant 0 \text{ et il existe un intervalle } \mathcal{V} \subset \mathbb{R} \setminus B \\ \text{t.q. } \hat{u} \in \mathcal{V} \text{ et } \forall v \in \mathcal{V} \setminus \{\hat{u}\},\ \tilde{g}''(v) = a^2 + \phi''(v) < 0, \end{cases}$$

montrant que $\hat{u}$ n'est pas un minimiseur (local) de $\tilde{g}$.

(12.22) $\Longleftarrow$ (A.94) D'après la caractérisation des points critiques (12.13), on peut réécrire (A.94) comme il suit :

$$\begin{aligned} &\forall d \in \mathbb{R},\ \hat{u} \in C \setminus \left\{0, \frac{d}{a}\right\} \\ &\Longrightarrow \begin{cases} \begin{cases} \tilde{g}''(\hat{u}) \leqslant 0 \text{ et il existe un intervalle } \mathcal{V} \subset \mathbb{R} \setminus B \\ \text{t.q. } \hat{u} \in \mathcal{V} \text{ et } \forall v \in \mathcal{V} \setminus \{\hat{u}\},\ \tilde{g}''(v) < 0 \end{cases} & \text{si } \hat{u} \notin B, \\ \lim\limits_{\substack{v \to \hat{u} \\ v < \hat{u}}} \tilde{g}'(v) > \lim\limits_{\substack{v \to \hat{u} \\ v > \hat{u}}} \tilde{g}'(v) & \text{si } \hat{u} \in B. \end{cases} \end{aligned} \tag{A.95}$$

Notons que la deuxième partie du «premier si» est utile pour discerner les minimiseurs (locaux) des autres points critiques lorsque $\tilde{g}''(\hat{u}) = 0$. En ce qui concerne le «deuxième si», puisque nous sommes dans le cas où $\hat{u} \in B$ est un point critique, les dérivées à gauche et à droite de $\tilde{g}$ au point $\hat{u}$ sont de signe différent et l'inégalité donnée est alors nécessaire pour assurer que $\hat{u}$ n'est pas un minimiseur (local).

Ensuite, nous pouvons voir que

$$\forall u \in \left[-\sqrt{2\lambda}/a, \sqrt{2\lambda}/a\right] \setminus \Big\{\{0\} \cup \{u \notin B : \phi'(u) = 0\}\Big\},$$

il existe $d_0 \in \mathbb{R}$ tel que $d_0 \neq au$ et

$$\begin{cases} a^2 u - a d_0 + \phi'(u) = 0 \text{ si } u \notin B, \\ a d_0 - a^2 u \in \left[\underline{\delta}^u, \bar{\delta}^u\right] \quad \text{si } u \in B. \end{cases}$$

impliquant, d'après (12.13), que $u \in C \setminus \left\{0, \frac{d_0}{a}\right\}$. En combinant cela avec (A.95) nous obtenons

$$
\begin{aligned}
& u \in \left[-\sqrt{2\lambda}/a, \sqrt{2\lambda}/a\right] \setminus \Big\{\{0\} \cup \{u \notin B : \phi'(u) = 0\}\Big\} \\
\implies & \begin{cases} \begin{cases} \tilde{g}''(u) \leqslant 0 \text{ et il existe un intervalle } \mathcal{V} \subset \mathbb{R} \setminus B \\ \text{t.q. } u \in \mathcal{V} \text{ et } \forall v \in \mathcal{V} \setminus \{u\},\ \tilde{g}''(v) < 0 \end{cases} & \text{si } u \notin B, \\ \lim\limits_{\substack{v \to u \\ v < u}} \tilde{g}'(v) > \lim\limits_{\substack{v \to u \\ v > u}} \tilde{g}'(v) & \text{si } u \in B, \end{cases} \\
\iff & \begin{cases} \begin{cases} \phi''(u) \leqslant -a^2 \text{ et il existe un intervalle } \mathcal{V} \subset \mathbb{R} \setminus B \\ \text{t.q. } u \in \mathcal{V} \text{ et } \forall v \in \mathcal{V} \setminus \{u\},\ \phi''(v) < -a^2 \end{cases} & \text{si } u \notin B, \\ \lim\limits_{\substack{v \to u \\ v < u}} \phi'(v) > \lim\limits_{\substack{v \to u \\ v > u}} \phi'(v) & \text{si } u \in B. \end{cases}
\end{aligned}
\tag{A.96}
$$

Ainsi, nous avons montré (12.22a). Pour (12.22b), nous devons nous intéresser aux points $u \in [-\sqrt{2\lambda}/a, \sqrt{2\lambda}/a] \setminus B$ tels que $\phi'(u) = 0$. Remarquons tout d'abord que d'après (12.21) nous avons

$$
\lim_{\substack{u \to 0 \\ u > 0}} \phi'(u) \geqslant \sqrt{2\lambda} a > 0 \quad \text{et} \quad \lim_{\substack{u \to \sqrt{2\lambda}/a \\ u > \sqrt{2\lambda}/a}} \phi'(u) = 0, \tag{A.97}
$$

et que, d'après (A.96), la «dérivée» de ϕ est une fonction discontinue (aux points dans B) et décroissante sur $[0, \sqrt{2\lambda}/a]$. Ainsi, s'il existe $\bar{u} \in [0, \sqrt{2\lambda}/a] \setminus B$ tel que $\phi'(\bar{u}) = 0$, alors $\phi'(u) = 0\ \forall u \in [\bar{u}, +\infty)$. Un raisonnement similaire peut être réalisé sur $\mathbb{R}_-$ et alors, en définissant β^- et β^+ comme dans l'énoncé du théorème, nous obtenons que $B \subset [\beta^-, \beta^+]$ et que $\forall u \in]\beta^-, \beta^+[\setminus B,\ \phi'(u) \neq 0$. De là, (A.96) devient : $\forall u \in]\beta^-, \beta^+[\setminus \{0\}$,

$$
\begin{cases} \begin{cases} \phi''(u) \leqslant -a^2 \text{ et il existe un intervalle } \mathcal{V} \subset \mathbb{R} \setminus B \\ \text{t.q. } u \in \mathcal{V} \text{ et } \forall v \in \mathcal{V} \setminus \{u\},\ \phi''(v) < -a^2 \end{cases} & \text{si } u \notin B, \\ \lim\limits_{\substack{v \to u \\ v < u}} \phi'(v) > \lim\limits_{\substack{v \to u \\ v > u}} \phi'(v) & \text{si } u \in B, \end{cases} \tag{A.98}
$$

ce qui termine la preuve.

A.3.5 *Preuve de la proposition 12.20*

Par définition de Φ_{SCAD}, la condition (12.26a) est vérifiée pour tout $\gamma_i > 2$ et $\tilde{\lambda}_i > 0$ ($i \in \mathbb{I}_N$). De plus, on obtient facilement que

$$
(12.26b) \iff \forall i \in \mathbb{I}_N,\ \frac{(\gamma_i + 1)\tilde{\lambda}_i^2}{2} = \lambda \text{ et } \gamma_i \tilde{\lambda}_i = \beta^{i+} \leqslant \frac{\sqrt{2\lambda}}{\|a_i\|}, \tag{A.99}
$$

$$
(12.26c) \implies \forall i \in \mathbb{I}_N,\ \tilde{\lambda}_i \geqslant \sqrt{2\lambda}\|a_i\|, \tag{A.100}
$$

où la dernière implication provient du même argument que celui utilisé dans la preuve de la proposition 12.19 (par symétrie $\beta^{i-} = -\beta^{i+}$). Regardons maintenant si les conditions ci-dessus sur γ_i and $\tilde{\lambda}_i$ peuvent être vérifiées simultanément. On a, $\forall i \in \mathbb{I}_N$

$$\tilde{\lambda}_i = \sqrt{\frac{2\lambda}{\gamma_i + 1}} \text{ et } \frac{\gamma_i}{\sqrt{\gamma_i + 1}} \leqslant \frac{1}{\|a_i\|} \text{ et } \frac{1}{\sqrt{\gamma_i + 1}} \geqslant \|a_i\|, \tag{A.101}$$

$$\underset{\gamma_i > 2}{\iff} \tilde{\lambda}_i = \sqrt{\frac{2\lambda}{\gamma_i + 1}} \text{ et } \|a_i\| \leqslant \frac{1}{\sqrt{\gamma_i + 1}} < \frac{\sqrt{\gamma_i + 1}}{\gamma_i} \text{ et } \|a_i\| < \frac{1}{\sqrt{3}}. \tag{A.102}$$

$$\iff \tilde{\lambda}_i = \sqrt{\frac{2\lambda}{\gamma_i + 1}} \text{ et } \gamma_i \leqslant \frac{1}{\|a_i\|^2} - 1 \text{ et } \|a_i\| < \frac{1}{\sqrt{3}}. \tag{A.103}$$

Afin de conclure la preuve, nous devons montrer que (12.53) $\implies$ (12.26c). Par symétrie on restreint la preuve à $\mathbb{R}_+$. En utilisant les mêmes arguments que dans la preuve de la proposition 12.19, nous avons sous la condition (12.53)

$$\forall i \in \mathbb{I}_N, \forall u \in]0, \tilde{\lambda}_i], \ \phi_{\text{CEL0}}(\|a_i\|, \lambda; u) < \phi_{\text{SCAD}}(\gamma_i, \tilde{\lambda}_i; u), \tag{A.104}$$

puisque (12.53) $\implies \tilde{\lambda}_i \geqslant \sqrt{2\lambda}\|a_i\| \ \forall i$. Ensuite, $\forall i \in \mathbb{I}_N$, $\forall u \in [\tilde{\lambda}_i, \gamma_i\tilde{\lambda}_i]$, nous avons

$$\phi_{\text{CEL0}}(\|a_i\|, \lambda; u) = P_1(u) = -\frac{\|a_i\|^2}{2}u^2 + \sqrt{2\lambda}\|a_i\|u,$$
$$\phi_{\text{SCAD}}(\gamma_i, \tilde{\lambda}_i; u) = P_2(u) = -\frac{\tilde{\lambda}_i^2 - 2\gamma_i\tilde{\lambda}_i|u| + u^2}{2(\gamma_i - 1)}$$

où $(P_1, P_2) \in (\mathbb{R}^2[X])^2$ sont deux polynomes d'ordre 2. Considérons le polynôme $Q = P_2 - P_1$. Alors, (A.99) et (A.104) nous assurent que $Q(\tilde{\lambda}_i) > 0$ et $Q(\gamma_i\tilde{\lambda}_i) \geqslant 0$ (étant donné que l'on est toujours sous la condition (12.53) et que (12.53) $\implies (\frac{(\gamma_i+1)\tilde{\lambda}_i^2}{2} = \lambda$ et $\gamma_i\tilde{\lambda}_i \leqslant \frac{\sqrt{2\lambda}}{\|a_i\|})$. De plus, nous pouvons voir que

$$\forall u \in [\tilde{\lambda}_i, \gamma_i\tilde{\lambda}_i], \ Q''(u) = \|a_i\|^2 - \frac{1}{\gamma_i - 1} \underset{(12.53)}{<} 0 \tag{A.105}$$

Ainsi, Q est strictement positif et concave sur $[\tilde{\lambda}_i, \gamma_i\tilde{\lambda}_i]$ ce qui, avec $Q(\tilde{\lambda}_i) > 0$ et $Q(\gamma_i\tilde{\lambda}_i) \geqslant 0$, implique que that $Q(u) > 0 \ \forall u \in [\tilde{\lambda}_i, \gamma_i\tilde{\lambda}_i[$. Cela montre que (12.53) $\implies$ (12.26c). Enfin, le fait que G_{SCAD} ne puisse jamais avoir la propriété (P2) provient des mêmes arguments que ceux utilisés dans la preuve de la proposition 12.19 pour la pénalité Capped-ℓ_1.

BIBLIOGRAPHIE

3GPP (Mar. 2010). « 3rd Generation Partnership Project ; Technical Specification Group Radio Access Network ; Evolved Universal Terrestrial Radio Access (E-ULTRA) ». In : *Physical Channels and Modulation* Release 9, V9.1.0 (cf. p. 147).

Antoniadis, Anestis, Irène Gijbels et Mila Nikolova (2011). « Penalized likelihood regression for generalized linear models with non-quadratic penalties ». In : *Annals of the Institute of Statistical Mathematics* 63.3, p. 585–615. doi : 10.1007/s10463-009-0242-4 (cf. p. 90).

Attouch, Hedy, Jérôme Bolte et Benar Fux Svaiter (2013). « Convergence of descent methods for semi-algebraic and tame problems : proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods ». In : *Mathematical Programming* 137.1-2, p. 91–129. doi : 10.1007/s10107-011-0484-9 (cf. p. 85, 115, 116, 200).

Axelrod, Daniel (1981). « Cell-substrate contacts illuminated by total internal reflection fluorescence. » In : *The Journal of cell biology* 89.1, p. 141–145. doi : 10.1083/jcb.89.1.141 (cf. p. 9, 12).

Axelrod, Daniel (2001). « Total internal reflection fluorescence microscopy in cell biology ». In : *Traffic* 2.11, p. 764–774. doi : 10.1034/j.1600-0854.2001.21104.x (cf. p. 9, 12).

Axelrod, Daniel (2008). « Total internal reflection fluorescence microscopy ». In : *Methods in cell biology* 89, p. 169–221. doi : 10.1016/S0091-679X(08)00607-9 (cf. p. 9, 12, 14, 16).

Beck, Amir et Yonina C Eldar (2013). « Sparsity constrained nonlinear optimization : Optimality conditions and algorithms ». In : *SIAM Journal on Optimization* 23.3, p. 1480–1509. doi : 10.1137/120869778 (cf. p. 84, 108, 109, 121).

Beck, Amir et Marc Teboulle (2009). « A fast iterative shrinkage-thresholding algorithm for linear inverse problems ». In : *SIAM Journal on Imaging Sciences* 2.1, p. 183–202. doi : 10.1137/080716542 (cf. p. 77, 85, 133).

Bect, Julien, Laure Blanc-Féraud, Gilles Aubert et Antonin Chambolle (2004). « A l1-unified variational framework for image restoration ». In : *European Conference on Computer Vision*. Springer, p. 1–13. doi : 10.1007/978-3-540-24673-2_1 (cf. p. 77).

Bellman, Richard et Robert S Roth (1984). *The laplace transform*. T. 3. World Scientific (cf. p. 19).

Benvenuto, F, A La Camera, C Theys, A Ferrari, H Lantéri et M Bertero (2008). « The study of an iterative method for the reconstruction of images corrupted by Poisson and Gaussian noise ». In : *Inverse Problems* 24.3, p. 035016. doi : 10.1088/0266-5611/24/3/035016 (cf. p. 32).

Betzig, Eric, George H Patterson, Rachid Sougrat, O Wolf Lindwasser, Scott Olenych, Juan S Bonifacino, Michael W Davidson, Jennifer Lippincott-Schwartz et Harald F Hess (2006). « Imaging intracellular fluorescent proteins at nanometer resolution ». In : *Science* 313.5793, p. 1642–1645. doi : 10.1126/science.1127344 (cf. p. 9, 154).

Bixby, Robert E (2012). « A brief history of linear and mixed-integer programming computation ». In : *Documenta Mathematica*, p. 107–121 (cf. p. 93).

Blake, Andrew et Andrew Zisserman (1987). *Visual reconstruction*. T. 2. MIT press Cambridge (cf. p. 122, 123).

Blumensath, Thomas et Mike E Davies (2007). *On the difference between orthogonal matching pursuit and orthogonal least squares*. Rapp. tech. url : http://www.personal.soton.ac.uk/tb1m08/papers/BDOMPvsOLS07.pdf (cf. p. 80).

Blumensath, Thomas et Mike E Davies (2008). « Iterative thresholding for sparse approximations ». In : *Journal of Fourier Analysis and Applications* 14.5-6, p. 629–654. doi : 10.1007/s00041-008-9035-z (cf. p. 84, 85, 116).

Blumensath, Thomas et Mike E Davies (2010). « Normalized iterative hard thresholding : Guaranteed stability and performance ». In : *Selected Topics in Signal Processing, IEEE Journal of* 4.2, p. 298–309. doi : 10.1109/JSTSP.2010.2042411 (cf. p. 85).

Blumensath, Thomas, Mehrdad Yaghoobi et Mike E Davies (2007). « Iterative hard thresholding and lo regularisation ». In : *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*. T. 3. IEEE, p. III–877. doi : 10.1109/ICASSP.2007.366820 (cf. p. 85).

Bolte, Jerôme, Aris Daniilidis, Adrian Lewis et Masahiro Shiota (2005). « Clarke critical values of subanalytic Lipschitz continuous functions ». In : *Ann. Polon. Math.(memorial issue for S. Lojasiewicz)* (cf. p. 200).

Born, Max et Emil Wolf (2000). *Principles of optics : electromagnetic theory of propagation, interference and diffraction of light*. CUP Archive (cf. p. 13, 16).

Boulanger, J., C. Gueudry, D. Münch, B. Cinquin, P. Paul-Gilloteaux, S. Bardin, C. Guérin, F. Senger, L. Blanchoin et J. Salamero (2014). « Fast high-resolution 3D total internal reflection fluorescence microscopy by incidence angle scanning and azimuthal averaging ». In : *Proceedings of the National Academy of Sciences* 111.48. doi : 10.1073/pnas.1414106111 (cf. p. 15, 17, 18, 26, 42, 51, 53).

Bourguignon, Sébastien, Jordan Ninin, Hervé Carfantan et Marcel Mongeau (2015). « Optimisation exacte de critères parcimonieux en norme lo par programmation mixte en nombres entiers ». In : *Colloque GRETSI* (cf. p. 93).

Bourguignon, Sébastien, Jordan Ninin, Hervé Carfantan et Marcel Mongeau (2016). « Exact Sparse Approximation Problems via Mixed-Integer Programming : Formulations and Computational Performance ». In : *IEEE Transactions on Signal Processing* 64.6, p. 1405–1419. doi : 10.1109/TSP.2015.2496367 (cf. p. 93).

Boyd, Stephen et Lieven Vandenberghe (2004). *Convex optimization*. Cambridge university press (cf. p. 76).

Bradley, PS, OL Mangasarian et JB Rosen (1998). « Parsimonious least norm approximation ». In : *Computational Optimization and Applications* 11.1, p. 5–21. doi : 10.1023/A:1018361916442 (cf. p. 93).

Breheny, Patrick et Jian Huang (2011). « Coordinate descent algorithms for nonconvex penalized regression with applications to biological feature selection ». In : *The annals of applied statistics* 5.1, p. 232. doi : 10.1214/10-AOAS388 (cf. p. 122).

Breiman, Leo (1995). « Better subset regression using the nonnegative garrote ». In : *Technometrics* 37.4, p. 373–384. doi : 10.1080/00401706.1995.10484371 (cf. p. 87).

Bruce, A.G. et H-Y. Gao (1995). « Waveshrink : shrinkage functions and thresholds ». In : *International Symposium on Optical Science, Engineering, and Instrumentation*. SPIE, p. 270–281. doi : 10.1117/12.217582 (cf. p. 116).

Burmeister, JS, George A Truskey et William M Reichert (1994). « Quantitative analysis of variable-angle total internal reflection fluorescence microscopy (VA-TIRFM) of cell/-substrate contacts ». In : *Journal of microscopy* 173.1, p. 39–51. doi : 10.1111/j.1365-2818.1994.tb03426.x (cf. p. 24).

Candès, Emmanuel J (2008). « The restricted isometry property and its implications for compressed sensing ». In : *Comptes Rendus Mathematique* 346.9, p. 589–592. doi : 10.1016/j.crma.2008.03.014 (cf. p. 73, 76).

Candès, Emmanuel J et Terence Tao (2005). « Decoding by linear programming ». In : *Information Theory, IEEE Transactions on* 51.12, p. 4203–4215. doi : 10.1109/TIT.2005.858979 (cf. p. 73, 76).

Candès, Emmanuel J et Michael B Wakin (2008). « An introduction to compressive sampling ». In : *Signal Processing Magazine, IEEE* 25.2, p. 21–30. doi : 10.1109/MSP.2007.914731 (cf. p. 76).

Candès, Emmanuel J, Justin K Romberg et Terence Tao (2006). « Stable signal recovery from incomplete and inaccurate measurements ». In : *Communications on pure and applied mathematics* 59.8, p. 1207–1223. doi : 10.1002/cpa.20124 (cf. p. 76).

Candès, Emmanuel J, Michael B Wakin et Stephen P Boyd (2008). « Enhancing sparsity by reweighted ℓ_1 minimization ». In : *Journal of Fourier analysis and applications* 14.5-6, p. 877–905. doi : 10.1007/s00041-008-9045-x (cf. p. 88, 136).

Candes, Emmanuel et Terence Tao (2007). « The Dantzig selector : Statistical estimation when p is much larger than n ». In : *The Annals of Statistics*, p. 2313–2351 (cf. p. 148).

Chahid, Makhlad (2014). « Echantillonnage compressif appliqué à la microscopie de fluorescence et à la microscopie de super résolution ». Thèse de doct. Bordeaux (cf. p. 161, 162).

Chambolle, Antonin et Charles Dossal (2015). « On the convergence of the iterates of the «fast iterative shrinkage/thresholding algorithm» ». In : *Journal of Optimization Theory and Applications* 166.3, p. 968–982. doi : 10.1007/s10957-015-0746-4 (cf. p. 77).

Chambolle, Antonin et Thomas Pock (2011). « A first-order primal-dual algorithm for convex problems with applications to imaging ». In : *Journal of Mathematical Imaging and Vision* 40.1, p. 120–145. doi : 10.1007/s10851-010-0251-1 (cf. p. 35).

Charbonnier, Pierre, Laure Blanc-Féraud, Gilles Aubert et Michel Barlaud (1997). « Deterministic edge-preserving regularization in computed imaging ». In : *Image Processing, IEEE Transactions on* 6.2, p. 298–311. doi : 10.1109/83.551699 (cf. p. 118).

Chartrand, Rick (2007). « Exact reconstruction of sparse signals via nonconvex minimization ». In : *Signal Processing Letters, IEEE* 14.10, p. 707–710. doi : 10.1109/LSP.2007.898300 (cf. p. 90).

Chen, Scott Shaobing, David L Donoho et Michael A Saunders (1998). « Atomic decomposition by basis pursuit ». In : *SIAM journal on scientific computing* 20.1, p. 33–61. doi : 10.1137/S003614450037906X (cf. p. 76).

Chen, Sheng, Stephen A Billings et Wan Luo (1989). « Orthogonal least squares methods and their application to non-linear system identification ». In : *International Journal of control* 50.5, p. 1873–1896. doi : 10.1080/00207178908953472 (cf. p. 80).

Chinatto, Adilson, Emmanuel Soubies, Cynthia Junqueira, Joao MT Romano, Pascal Larzabal, Jean-Pierre Barbot et Laure Blanc-Féraud (2015). « Lo-Optimization for Channel and DOA Sparse Estimation ». In : *International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. IEEE, p. 4. doi : 10.1109/CAMSAP.2015.7383797 (cf. p. 144, 149).

Chouzenoux, Emilie, Anna Jezierska, Jean-Christophe Pesquet et Hugues Talbot (2013). « A majorize-minimize subspace approach for ℓ_2-ℓ_0 image regularization ». In : *SIAM Journal on Imaging Sciences* 6.1, p. 563–591. doi : 10.1137/11085997X (cf. p. 92).

Chouzenoux, Emilie, Anna Jezierska, Jean-Christophe Pesquet et Hugues Talbot (2015). « A Convex Approach for Image Restoration with Exact Poisson–Gaussian Likelihood ». In : *SIAM Journal on Imaging Sciences* 8.4, p. 2662–2682. doi : 10.1137/15M1014395 (cf. p. 32).

Clarke, Frank H (1990). *Optimization and nonsmooth analysis*. T. 5. SIAM. isbn : 978-0-898712-56-8 (cf. p. 99, 100, 167, 168).

Cohen, Albert, Wolfgang Dahmen et Ronald DeVore (2009). « Compressed sensing and best ? ?-term approximation ». In : *Journal of the American mathematical society* 22.1, p. 211–231. doi : 10.1090/S0894-0347-08-00610-3 (cf. p. 73).

Cohen, Laurent D et Ron Kimmel (1997). « Global minimum for active contour models : A minimal path approach ». In : *International journal of computer vision* 24.1, p. 57–78. doi : 10.1023/A:1007922224810 (cf. p. 25).

Combettes, Patrick L et Jean-Christophe Pesquet (2008). « A proximal decomposition method for solving convex variational inverse problems ». In : *Inverse problems* 24.6, p. 065014. doi : 10.1088/0266-5611/24/6/065014 (cf. p. 26, 36).

Combettes, Patrick L et Jean-Christophe Pesquet (2011). « Proximal splitting methods in signal processing ». In : *Fixed-point algorithms for inverse problems in science and engineering*. Springer, p. 185–212. doi : 10.1007/978-1-4419-9569-8_10 (cf. p. 34, 76, 77).

Combettes, Patrick L et Valérie R Wajs (2005). « Signal recovery by proximal forward-backward splitting ». In : *Multiscale Modeling & Simulation* 4.4, p. 1168–1200. doi : 10.1137/050626090 (cf. p. 34, 76, 77).

Cormen, Thomas H, Charles E Leiserson, Ronald L Rivest et Clifford Stein (2009). *Introduction to algorithms* (cf. p. 71, 72).

Couvreur, Christophe et Yoram Bresler (2000). « On the optimality of the backward greedy algorithm for the subset selection problem ». In : *SIAM Journal on Matrix Analysis and Applications* 21.3, p. 797–808. doi : 10.1137/S0895479898332928 (cf. p. 81).

Dai, Wei et Olgica Milenkovic (2009). « Subspace pursuit for compressive sensing signal reconstruction ». In : *Information Theory, IEEE Transactions on* 55.5, p. 2230–2249. doi : 10.1109/TIT.2009.2016006 (cf. p. 86).

Dantzig, George Bernard (1998). *Linear programming and extensions*. Princeton university press (cf. p. 76).

Daubechies, Ingrid, Michel Defrise et Christine De Mol (2004). « An iterative thresholding algorithm for linear inverse problems with a sparsity constraint ». In : *Communications on Pure and Applied Mathematics* 57.11, p. 1413–1457. doi : 10.1002/cpa.20042 (cf. p. 77).

Daubechies, Ingrid, Ronald DeVore, Massimo Fornasier et C Sinan Güntürk (2010). « Iteratively reweighted least squares minimization for sparse recovery ». In : *Communications on Pure and Applied Mathematics* 63.1, p. 1–38. doi : 10.1002/cpa.20303 (cf. p. 118).

Davis, Geoff, Stephane Mallat et Marco Avellaneda (1997). « Adaptive greedy approximations ». In : *Constructive approximation* 13.1, p. 57–98. doi : 10.1007/BF02678430 (cf. p. 71).

Davis, Geoffrey M, Stephane G Mallat et Zhifeng Zhang (1994). « Adaptive time-frequency decompositions ». In : *Optical Engineering* 33.7, p. 2183–2191. doi : 10.1117/12.173207 (cf. p. 79).

Descombes, Xavier (2011). *Stochastic geometry for image analysis*. Wiley/Iste, x. descombes edition (cf. p. 25).

Dey, Nicolas, Laure Blanc-Feraud, Christophe Zimmer, Pascal Roux, Zvi Kam, Jean-Christophe Olivo-Marin et Josiane Zerubia (2006). « Richardson-Lucy algorithm with total variation regularization for 3D confocal microscope deconvolution ». In : *Microscopy research and technique* 69.4, p. 260–266. doi : 10.1002/jemt.20294 (cf. p. 36).

Dinh, Tao Pham et Hoai An Le Thi (2014). « Recent advances in DC programming and DCA ». In : *Transactions on Computational Intelligence XIII*. Springer, p. 1–37. doi : 10.1007/978-3-642-54455-2_1 (cf. p. 96).

Donoho, David L et Michael Elad (2003). « Optimally sparse representation in general (nonorthogonal) dictionaries via l1 minimization ». In : *Proceedings of the National Academy of Sciences* 100.5, p. 2197–2202. doi : 10.1073/pnas.0437847100 (cf. p. 72, 76).

Donoho, David L et Xiaoming Huo (2001). « Uncertainty principles and ideal atomic decomposition ». In : *Information Theory, IEEE Transactions on* 47.7, p. 2845–2862. doi : 10.1109/18.959265 (cf. p. 72).

Donoho, David L, Yaakov Tsaig, Iddo Drori et Jean-Luc Starck (2012). « Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit ». In : *Information Theory, IEEE Transactions on* 58.2, p. 1094–1121. doi : 10.1109/TIT.2011.2173241 (cf. p. 80).

Dos Santos, M. C., R. Déturche, C. Vézy et R. Jaffiol (2014). « Axial nanoscale localization by normalized total internal reflection fluorescence microscopy ». In : *Optics letters* 39.4. doi : 10.1364/OL.39.000869 (cf. p. 18, 24, 48).

Dos Santos, Marcelina Cardoso, Cyrille Vézy et Rodolphe Jaffiol (2016). « Nanoscale characterization of vesicle adhesion by normalized total internal reflection fluorescence microscopy ». In : *Biochimica et Biophysica Acta (BBA)-Biomembranes* 1858.6, p. 1244–1253. doi : 10.1016/j.bbamem.2016.03.008 (cf. p. 24).

Duda, R.O. et P.E. Hart (1972). « Use of the Hough Transformation to Detect Lines and Curves in Pictures ». In : *Com. ACM* 15.1. doi : 10.1145/361237.361242 (cf. p. 49).

Dussault, David et Paul Hoess (2004). « Noise performance comparison of ICCD with CCD and EMCCD cameras ». In : *Optical Science and Technology, the SPIE 49th Annual Meeting*. International Society for Optics et Photonics, p. 195–204. doi : 10.1117/12.561839 (cf. p. 20, 21).

Efron, Bradley, Trevor Hastie, Iain Johnstone et Robert Tibshirani (2004). « Least angle regression ». In : *Ann. Statist.* 32.2, p. 407–499. doi : 10.1214/009053604000000067 (cf. p. 77, 83, 131).

Elad, Michael (2010). *Sparse and Redundant Representations : From Theory to Applications in Signal and Image Processing*. Springer (cf. p. 76).

Epstein, Charles L et John Schotland (2008). « The bad truth about Laplace's transform ». In : *SIAM review* 50.3, p. 504–520. doi : 10.1137/060657273 (cf. p. 19).

Fan, Jianqing (1997). « Comments on «wavelets in statistics : A review» by a. antoniadis ». In : *Journal of the Italian Statistical Society* 6.2, p. 131–138. doi : 10.1007/BF03178906 (cf. p. 90).

Fan, Jianqing et Runze Li (2001). « Variable selection via nonconcave penalized likelihood and its oracle properties ». In : *Journal of the American Statistical Association* 96.456, p. 1348–1360. doi : 10.1198/016214501753382273 (cf. p. 87, 88, 96, 169, 177).

Fellers, TJ et MW Davidson (2004). « CCD noise sources and signal-to-noise ratio ». In : *Optical Microscopy Primer,(Molecular Expressions, Florida State Univ.)* url : http://micro.magnet.fsu.edu/primer/digitalimaging/concepts/ccdsnr.html (cf. p. 20, 21).

Fenchel, Werner (1949). « On conjugate convex functions ». In : *Canad. J. Math* 1.73-77 (cf. p. 35, 96).

Figueiredo, Mário AT et Robert D Nowak (2003). « An EM algorithm for wavelet-based image restoration ». In : *IEEE Transactions on Image Processing* 12.8, p. 906–916. doi : 10.1109/TIP.2003.814255 (cf. p. 77).

Fiolka, R, Y Belyaev, H Ewers et A Stemmer (2008a). « Even illumination in total internal reflection fluorescence microscopy using laser light ». In : *Microscopy research and technique* 71.1, p. 45–50. doi : 10.1002/jemt.20527 (cf. p. 15, 51).

Fiolka, Reto (2009). « Improving the resolution in total internal reflection fluorescence and phase microscopy ». Thèse de doct. Diss., Eidgenössische Technische Hochschule ETH Zürich, Nr. 18639, 2009 (cf. p. 15).

Fiolka, Reto, Markus Beck et Andreas Stemmer (2008b). « Structured illumination in total internal reflection fluorescence microscopy using a spatial light modulator ». In : *Opt. Lett.* 33.14, p. 1629–1631. doi : 10.1364/OL.33.001629 (cf. p. 18).

Foucart, Simon (2011). « Hard thresholding pursuit : an algorithm for compressive sensing ». In : *SIAM Journal on Numerical Analysis* 49.6, p. 2543–2563. doi : 10.1137/100806278 (cf. p. 86).

Foucart, Simon et Ming-Jun Lai (2009). « Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0 < q \leqslant 1$ ». In : *Applied and Computational Harmonic Analysis* 26.3, p. 395–407. doi : 10.1016/j.acha.2008.09.001 (cf. p. 90).

Foucart, Simon et Holger Rauhut (2013). *A mathematical introduction to compressive sensing*. T. 1. 3. Springer (cf. p. 76).

Frohn, Jan T, Helmut F Knapp et Andreas Stemmer (2000). « True optical resolution beyond the Rayleigh limit achieved by standing wave illumination ». In : *Proceedings of the National Academy of Sciences* 97.13, p. 7232–7236. doi : 10.1073/pnas.130181797 (cf. p. 18).

Fu, Yan, Peter W Winter, Raul Rojas, Victor Wang, Matthew McAuliffe et George H Patterson (2016). « Axial superresolution via multiangle TIRF microscopy with sequential imaging and photobleaching ». In : *Proceedings of the National Academy of Sciences* 113.16, p. 4368–4373. doi : 10.1073/pnas.1516715113 (cf. p. 27).

Fung, GM et OL Mangasarian (2011). « Equivalence of minimal ℓ_0- and ℓ_p-norm solutions of linear equalities, inequalities and linear programs for sufficiently small p ». In : *Journal of optimization theory and applications* 151.1, p. 1–10. doi : 10.1007/s10957-011-9871-x (cf. p. 92, 93).

Garg, Rahul et Rohit Khandekar (2009). « Gradient descent with sparsification : an iterative algorithm for sparse recovery with restricted isometry property ». In : *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, p. 337–344. doi : 10.1145/1553374.1553417 (cf. p. 85).

Gasso, Gilles, Alain Rakotomamonjy et Stéphane Canu (2009). « Recovering sparse signals with a certain family of nonconvex penalties and DC programming ». In : *Signal Processing, IEEE Transactions on* 57.12, p. 4686–4698. doi : 10.1109/TSP.2009.2026004 (cf. p. 120, 121).

Geiger, Benjamin, Alexander Bershadsky, Roumen Pankov et Kenneth M Yamada (2001). « Transmembrane crosstalk between the extracellular matrix and the cytoskeleton ». In : *Nature Reviews Molecular Cell Biology* 2.11, p. 793–805. doi : 10.1038/35099066 (cf. p. 60).

Geman, Donald et George Reynolds (1992). « Constrained restoration and the recovery of discontinuities ». In : *IEEE Transactions on Pattern Analysis & Machine Intelligence* 3, p. 367–383. doi : 10.1109/34.120331 (cf. p. 108, 118).

Geman, Stuart et Donald Geman (1984). « Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images ». In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6, p. 721–741. doi : 10.1109/TPAMI.1984.4767596 (cf. p. 123).

Gingell, D, OS Heavens et JS Mellor (1987). « General electromagnetic theory of total internal reflection fluorescence : the quantitative basis for mapping cell-substratum topography ». In : *Journal of cell science* 87.5, p. 677–693. issn : 1477-9137 (cf. p. 17).

Gong, Pinghua, Changshui Zhang, Zhaosong Lu, Jianhua Huang et Jieping Ye (2013). « A General Iterative Shrinkage and Thresholding Algorithm for Non-convex Regularized Optimization Problems ». In : *Proceedings of The 30th International Conference on Machine Learning*, p. 37–45 (cf. p. 117).

Gorodnitsky, Irina F et Bhaskar D Rao (1997). « Sparse signal reconstruction from limited data using FOCUSS : A re-weighted minimum norm algorithm ». In : *Signal Processing, IEEE Transactions on* 45.3, p. 600–616. doi : 10.1109/78.558475 (cf. p. 118).

Gribonval, Rémi et Morten Nielsen (2003). « Sparse representations in unions of bases ». In : *Information Theory, IEEE Transactions on* 49.12, p. 3320–3325. doi : 10.1109/TIT.2003.820031 (cf. p. 73, 76).

Gribonval, Rémi, Philippe Depalle, Xavier Rodet, Emmanuel Bacry et Stéphane Mallat (1996). « Sound signals decomposition using a high resolution matching pursuit ». In : *ICMC : International Computer Music Conference*, p. 293–296 (cf. p. 79).

Gustafsson, Mats GL (2000). « Surpassing the lateral resolution limit by a factor of two using structured illumination microscopy ». In : *Journal of microscopy* 198.2, p. 82–87. doi : 10.1046/j.1365-2818.2000.00710.x (cf. p. 9).

Hadamard, Jacques (1902). « Sur les problèmes aux dérivées partielles et leur signification physique ». In : *Princeton university bulletin* 13, p. 49–52 (cf. p. 18).

Harmany, Zachary T, Roummel F Marcia et Rebecca M Willett (2012). « This is SPIRAL-TAP : Sparse Poisson intensity reconstruction algorithms-theory and practice ». In : *IEEE Transactions on Image Processing* 21.3, p. 1084–1096. doi : 10.1109/TIP.2011.2168410 (cf. p. 35).

Haugland, Dag (2007). « A bidirectional greedy heuristic for the subspace selection problem ». In : *Engineering stochastic local search algorithms. Designing, implementing and analyzing effective heuristics*. Springer, p. 162–176. doi : 10.1007/978-3-540-74446-7_12 (cf. p. 82).

Hell, Stefan W et Jan Wichmann (1994). « Breaking the diffraction resolution limit by stimulated emission : stimulated-emission-depletion fluorescence microscopy ». In : *Optics letters* 19.11, p. 780–782. doi : 10.1364/OL.19.000780 (cf. p. 9).

Hellen, Edward H et Daniel Axelrod (1987). « Fluorescence emission at dielectric and metal-film interfaces ». In : *JOSA B* 4.3, p. 337–350. doi : 10.1364/JOSAB.4.000337 (cf. p. 16, 53).

Herrity, Kyle K, Anna C Gilbert et Joel A Tropp (2006). « Sparse approximation via iterative thresholding ». In : *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*. T. 3. IEEE, p. III–III. doi : 10.1109/ICASSP.2006.1660731 (cf. p. 85).

Herzet, Cédric et Angélique Drémeau (2014). *Bayesian pursuit algorithms*. Rapp. tech. url : https://hal.inria.fr/hal-00673801/ (cf. p. 82).

Herzet, Cedue et Angélique Drémeau (2010). « Bayesian pursuit algorithms ». In : *Signal Processing Conference, 2010 18th European*. IEEE, p. 1474–1478 (cf. p. 81, 82).

Holden, Seamus J, Stephan Uphoff et Achillefs N Kapanidis (2011). « DAOSTORM : an algorithm for high-density super-resolution microscopy ». In : *Nature Methods* 8.4, p. 279–280. doi : 10.1038/nmeth0411-279 (cf. p. 161).

Horst, Reiner et Nguyen V Thoai (1999). « DC programming : overview ». In : *Journal of Optimization Theory and Applications* 103.1, p. 1–43. doi : 10.1023/A:1021765131316 (cf. p. 119).

Huang, Bo, Wenqin Wang, Mark Bates et Xiaowei Zhuang (2008). « Three-dimensional super-resolution imaging by stochastic optical reconstruction microscopy ». In : *Science* 319.5864, p. 810–813. doi : 10.1126/science.1153529 (cf. p. 9).

Idier, Jérôme (2013). *Bayesian approach to inverse problems*. John Wiley & Sons (cf. p. 139).

Jain, Prateek, Ambuj Tewari et Inderjit S Dhillon (2011). « Orthogonal matching pursuit with replacement ». In : *Advances in Neural Information Processing Systems*, p. 1215–1223 (cf. p. 82).

Jezierska, Anna Maria (2013). « Image restoration in the presence of Poisson-Gaussian noise ». Thèse de doct. Paris Est (cf. p. 32).

Jojic, Vladimir, Suchi Saria et Daphne Koller (2011). « Convex envelopes of complexity controlling penalties : the case against premature envelopment. » In : *AISTATS*, p. 399–406 (cf. p. 96).

Kanzow, C., N. Yamashita et M. Fukushima (2004). « Levenberg–Marquardt methods with strong local convergence properties for solving nonlinear equations with convex constraints ». In : *J. of Comput. and App. Math.* 172.2. doi : 10.1016/j.cam.2004.02.013 (cf. p. 52).

Kormylo, John J et Jerry M Mendel (1982). « Maximum likelihood detection and estimation of Bernoulli-Gaussian processes ». In : *IEEE Transactions on Information Theory* 28.3, p. 482–488. doi : 10.1109/TIT.1982.1056496 (cf. p. 81, 82).

Kowalski, M. (2014). « Thresholding rules and iterative shrinkage/thresholding algorithm : A convergence study ». In : *ICIP*. IEEE. doi : 10.1109/ICIP.2014.7025843 (cf. p. 85).

Larsson, Viktor, Carl Olsson, Erik Bylow et Fredrik Kahl (2014). « Rank minimization with structured data patterns ». In : *European Conference on Computer Vision*. Springer, p. 250–265 (cf. p. 96).

Le Thi, Hoai An, Hoai Minh Le et Tao Pham Dinh (2014). « Feature selection in machine learning : an exact penalty approach using a Difference of Convex function Algorithm ». In : *Machine Learning*, p. 1–24. doi : 10.1007/s10994-014-5455-y (cf. p. 90, 94, 177).

Le Thi, Hoai An, T Pham Dinh, Hoai Minh Le et Xuan Thanh Vo (2015). « DC approximation approaches for sparse optimization ». In : *European Journal of Operational Research* 244.1, p. 26–46. doi : 10.1016/j.ejor.2014.11.031 (cf. p. 93, 94, 165, 177).

Leclerc, Yvan G (1989). « Constructing simple stable descriptions for image partitioning ». In : *International Journal of Computer Vision* 3.1, p. 73–102. doi : 10.1007/BF00054839 (cf. p. 123).

Lerch, M (1903). « Sur un point de la théorie des fonctions génératrices d'Abel ». In : *Acta mathematica* 27.1, p. 339–351 (cf. p. 19).

Liang, Jingwei, Jalal Fadili et Gabriel Peyré (2016). « A Multi-step Inertial Forward–Backward Splitting Method for Non-convex Optimization ». In : *arXiv preprint*. url : http://arxiv.org/abs/1606.02118 (cf. p. 117).

Liang, Liang, Hongying Shen, Yingke Xu, Pietro De Camilli, Derek K Toomre et James S Duncan (2012). « A Bayesian method for 3D estimation of subcellular particle features in multi-angle TIRF microscopy ». In : *Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on*. IEEE, p. 984–987. doi : 10.1109/ISBI.2012.6235722 (cf. p. 25).

Linde, Sebastian van de, Anna Löschberger, Teresa Klein, Meike Heidbreder, Steve Wolter, Mike Heilemann et Markus Sauer (2011). « Direct stochastic optical reconstruction microscopy with standard fluorescent probes ». In : *Nature protocols* 6.7, p. 991–1009. doi : 10.1038/nprot.2011.336 (cf. p. 155).

Loerke, Dinah, Walter Stühmer et Martin Oheim (2002). « Quantifying axial secretory-granule motion with variable-angle evanescent-field excitation. » In : *Journal of neuroscience methods* 119.1, p. 65. doi : 10.1016/S0165-0270(02)00178-4 (cf. p. 25).

Lucy, Leon B (1974). « An iterative technique for the rectification of observed distributions ». In : *The astronomical journal* 79, p. 745. doi : 10.1086/111605 (cf. p. 36).

Mallat, Stéphane G et Zhifeng Zhang (1993). « Matching pursuits with time-frequency dictionaries ». In : *IEEE Transactions on Signal Processing* 41.12, p. 3397–3415. doi : 10.1109/78.258082 (cf. p. 78, 79).

Mangasarian, OL (1996). « Machine learning via polyhedral concave minimization ». In : *Applied Mathematics and Parallel Computing*. Springer, p. 175–188. doi : 10.1007/978-3-642-99789-1_13 (cf. p. 90).

Marjanovic, Goran, Magnus O Ulfarsson et Alfred O Hero (2015). « MIST : Lo sparse linear regression with momentum ». In : *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, p. 3551–3555. doi : 10.1109/ICASSP.2015.7178632 (cf. p. 85).

Marler, R Timothy et Jasbir S Arora (2004). « Survey of multi-objective optimization methods for engineering ». In : *Structural and multidisciplinary optimization* 26.6, p. 369–395. doi : 10.1007/s00158-003-0368-6s (cf. p. 72).

Marquardt, Donald W (1963). « An algorithm for least-squares estimation of nonlinear parameters ». In : *Journal of the society for Industrial and Applied Mathematics* 11.2, p. 431–441. doi : 10.1137/0111030 (cf. p. 24).

Martin-Fernandez, ML, CJ Tynan et SED Webb (2013). « A "pocket guide" to total internal reflection fluorescence ». In : *Journal of microscopy* 252.1, p. 16–22. doi : 10.1111/jmi.12070 (cf. p. 12, 13).

Mattheyses, A. L. et D. Axelrod (2006). « Direct measurement of the evanescent field profile produced by objective-based total internal reflection fluorescence ». In : *J. of biomedical optics* 11.1. doi : 10.1117/1.2161018 (cf. p. 16, 26).

Mattheyses, Alexa L, Keith Shaw et Daniel Axelrod (2006). « Effective elimination of laser interference fringing in fluorescence microscopy by spinning azimuthal incidence angle ». In : *Microscopy research and technique* 69.8, p. 642–647. doi : 10.1002/jemt.20334 (cf. p. 15).

Mazumder, Rahul, Jerome H Friedman et Trevor Hastie (2012). « SparseNet : Coordinate descent with nonconvex penalties ». In : *Journal of the American Statistical Association*. doi : 10.1198/jasa.2011.tm09738 (cf. p. 122, 131).

Minsky, Marvin (1988). « Memoir on inventing the confocal scanning microscope ». In : *Scanning* 10.4, p. 128–138. doi : 10.1002/sca.4950100403 (cf. p. 9).

Mobahi, Hossein et John W Fisher III (2015a). « A theoretical analysis of optimization by gaussian continuation ». In : *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. AAAI Press, p. 1205–1211. isbn : 0-262-51129-0 (cf. p. 123).

Mobahi, Hossein et John W Fisher III (2015b). « On the link between gaussian homotopy continuation and convex envelopes ». In : *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*. Springer, p. 43–56. doi : 10.1007/978-3-319-14612-6_4 (cf. p. 123).

Mohimani, Hosein, Massoud Babaie-Zadeh et Christian Jutten (2009). « A fast approach for overcomplete sparse decomposition based on smoothed ℓ_0 norm ». In : *IEEE Transactions on Signal Processing* 57.1, p. 289–301. doi : 10.1109/TSP.2008.2007606 (cf. p. 90, 123, 124).

Mohimani, Hosein, Massoud Babaie-Zadeh, Irina Gorodnitsky et Christian Jutten (2010). « Sparse Recovery using Smoothed ℓ_0 (SL0) : Convergence Analysis ». In : *arXiv preprint arXiv :1001.5073* (cf. p. 124).

Moreau, Jean-Jacques (1962). « Fonctions convexes duales et points proximaux dans un espace hilbertien ». In : *CR Acad. Sci. Paris Sér. A Math* 255, p. 2897–2899 (cf. p. 34, 77).

Natarajan, Balas Kausik (1995). « Sparse approximate solutions to linear systems ». In : *SIAM journal on computing* 24.2, p. 227–234. doi : 10.1137/S0097539792240406 (cf. p. 71).

Needell, Deanna et Joel A Tropp (2009). « CoSaMP : Iterative signal recovery from incomplete and inaccurate samples ». In : *Applied and Computational Harmonic Analysis* 26.3, p. 301–321. doi : 10.1016/j.acha.2008.07.002 (cf. p. 86).

Needell, Deanna et Roman Vershynin (2009). « Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit ». In : *Foundations of computational mathematics* 9.3, p. 317–334. doi : 10.1007/s10208-008-9031-3 (cf. p. 80).

Nesterov, Yu (2013). « Gradient methods for minimizing composite functions ». In : *Mathematical Programming* 140.1, p. 125–161. doi : 10.1007/s10107-012-0629-5 (cf. p. 77).

Nikolova, Mila (1999). « Markovian reconstruction using a GNC approach ». In : *IEEE Transactions on Image Processing* 8.9, p. 1204–1220. doi : 10.1109/83.784433 (cf. p. 123).

Nikolova, Mila (2005). « Analysis of the recovery of edges in images and signals by minimizing nonconvex regularized least-squares ». In : *SIAM Journal on Multiscale Modeling & Simulation* 4.3, p. 960–991. doi : 10.1137/040619582 (cf. p. 101).

Nikolova, Mila (2013). « Description of the minimizers of least squares regularized with ℓ_0-norm. Uniqueness of the global minimizer ». In : *SIAM Journal on Imaging Sciences* 6.2, p. 904–937. doi : 10.1137/11085476X (cf. p. 101–106, 111, 112, 145, 149, 173, 174, 176, 192, 193, 195).

Nikolova, Mila (2016). « Relationship between the optimal solutions of least squares regularized with l 0-norm and constrained by k-sparsity ». In : *Applied and Computational Harmonic Analysis*. doi : 10.1016/j.acha.2015.10.010 (cf. p. 72).

Ochs, Peter, Yunjin Chen, Thomas Brox et Thomas Pock (2014). « iPiano : Inertial proximal algorithm for nonconvex optimization ». In : *SIAM Journal on Imaging Sciences* 7.2, p. 1388–1419. doi : 10.1137/130942954 (cf. p. 117, 134).

Ochs, Peter, Alexey Dosovitskiy, Thomas Brox et Thomas Pock (2015). « On iteratively reweighted algorithms for nonsmooth nonconvex optimization in computer vision ». In : *SIAM Journal on Imaging Sciences* 8.1, p. 331–372. doi : 10.1137/140971518 (cf. p. 118, 119).

Ölveczky, B. P., N. Periasamy et A. S. Verkman (1997). « Mapping fluorophore distributions in three dimensions by quantitative multiple angle-total internal reflection fluorescence microscopy. » In : *Biophysical journal* 73.5. doi : 10.1016/S0006-3495(97)78312-7 (cf. p. 16, 17, 24, 51, 53).

Parikh, Neal et Stephen P Boyd (2014). « Proximal Algorithms. » In : *Foundations and Trends in optimization* 1.3, p. 127–239. doi : 10.1561/2400000003 (cf. p. 77, 117).

Pati, Yagyensh Chandra, Ramin Rezaiifar et PS Krishnaprasad (1993). « Orthogonal matching pursuit : Recursive function approximation with applications to wavelet decomposition ». In : *Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on*. IEEE, p. 40–44. doi : 10.1109/ACSSC.1993.342465 (cf. p. 79).

Pirinen, Pekka (2014). « A brief overview of 5G research activities ». In : *5G for Ubiquitous Connectivity (5GU), 2014 1st International Conference on*. IEEE, p. 17–22 (cf. p. 147).

Polyak, Boris T (1964). « Some methods of speeding up the convergence of iteration methods ». In : *USSR Computational Mathematics and Mathematical Physics* 4.5, p. 1–17. doi : 10.1016/0041-5553(64)90137-5 (cf. p. 117).

Rao, Bhaskar D, Kjersti Engan, Shane F Cotter, Jason Palmer et Kenneth Kreutz-Delgado (2003). « Subset selection in noise based on diversity measure minimization ». In : *Signal Processing, IEEE Transactions on* 51.3, p. 760–770. doi : 10.1109/TSP.2002.808076 (cf. p. 118).

Reichert, WM et GA Truskey (1990). « Total internal reflection fluorescence microscopy (TIRFM). I Modelling cell contact region fluorescence ». In : *Journal of cell science* 96.2, p. 219–230. issn : 1477-9137 (cf. p. 24).

Repetti, Audrey, Mai Quyen Pham, Laurent Duval, Emilie Chouzenoux et J-C Pesquet (2015). « Euclid in a Taxicab : Sparse Blind Deconvolution with Smoothed Regularization ». In : *Signal Processing Letters, IEEE* 22.5, p. 539–543. doi : 10.1109/LSP.2014.2362861 (cf. p. 90).

Richardson, William Hadley (1972). « Bayesian-based iterative method of image restoration ». In : *JOSA* 62.1, p. 55–59. doi : 10.1364/JOSA.62.000055 (cf. p. 35).

Rinaldi, Francesco, Fabio Schoen et Marco Sciandrone (2010). « Concave programming for minimizing the zero-norm over polyhedral sets ». In : *Computational Optimization and Applications* 46.3, p. 467–486. doi : 10.1007/s10589-008-9202-9 (cf. p. 93).

Robbins, Mark Stanford et Benjamin James Hadwen (2003). « The noise performance of electron multiplying charge-coupled devices ». In : *Electron Devices, IEEE Transactions on* 50.5, p. 1227–1232. doi : 10.1109/TED.2003.813462 (cf. p. 20).

Robini, Marc C et Isabelle E Magnin (2010). « Optimization by stochastic continuation ». In : *SIAM journal on Imaging Sciences* 3.4, p. 1096–1121. doi : 10.1137/090756181 (cf. p. 123).

Robini, Marc C, Aimé Lachal et Isabelle E Magnin (2007). « A stochastic continuation approach to piecewise constant reconstruction ». In : *IEEE Transactions on Image Processing* 16.10, p. 2576–2589. doi : 10.1109/TIP.2007.904975 (cf. p. 123).

Rockafellar, R Tyrrell et Roger J-B Wets (2009). *Variational analysis*. T. 317. Springer Science & Business Media (cf. p. 95, 118, 120).

Rohrbach, Alexander (2000). « Observing secretory granules with a multiangle evanescent wave microscope. » In : *Biophysical journal* 78.5, p. 2641. doi : 10.1016/S0006-3495(00)76808-1 (cf. p. 16, 25).

Roy, Richard et Thomas Kailath (1989). « ESPRIT-estimation of signal parameters via rotational invariance techniques ». In : *IEEE Transactions on Acoustics, Speech, and Signal Processing* 37.7, p. 984–995. doi : 10.1109/29.32276 (cf. p. 152).

Rust, Michael J, Mark Bates et Xiaowei Zhuang (2006). « Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM) ». In : *Nature methods* 3.10, p. 793–796. doi : 10.1038/nmeth929 (cf. p. 9, 154).

Saffarian, Saveez et Tomas Kirchhausen (2008). « Differential evanescence nanometry : live-cell fluorescence measurements with 10-nm axial resolution on the plasma membrane ». In : *Biophysical journal* 94.6, p. 2333–2342. doi : 10.1529/biophysj.107.117234 (cf. p. 24).

Sage, Daniel, Hagai Kirshner, Thomas Pengo, Nico Stuurman, Junhong Min, Suliana Manley et Michael Unser (2015). « Quantitative evaluation of software packages for single-molecule localization microscopy ». In : *Nature methods* 12.8, p. 717–724. doi : 10.1038/nmeth.3442 (cf. p. 156, 160, 161).

Schiff, Joel L (2013). *The Laplace transform : theory and applications*. Springer Science & Business Media (cf. p. 19).

Schmidt, Ralph (1986). « Multiple emitter location and signal parameter estimation ». In : *IEEE transactions on antennas and propagation* 34.3, p. 276–280. doi : 10.1109/TAP.1986.1143830 (cf. p. 152).

Segelstein, David J (2011). « The complex refractive index of water ». Thèse de doct. University of Missouri–Kansas City (cf. p. 54).

Soubies, Emmanuel, Laure Blanc-Féraud, Sébastien Schaub et Gilles Aubert (2014a). « A 3D model with shape prior information for biological structures reconstruction using Multiple-Angle Total Internal Reflection Fluorescence Microscopy ». In : *Biomedical Imaging (ISBI), 2014 IEEE 11th International Symposium on*. IEEE, p. 608–611. doi : 10.1109/ISBI.2014.6867944 (cf. p. 25).

Soubies, Emmanuel, Laure Blanc-Féraud, Sébastien Schaub et Gilles Aubert (2014b). « Sparse reconstruction from Multiple-Angle Total Internal Reflection Fluorescence Microscopy ». In : *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, p. 2844–2848. doi : 10.1109/ICIP.2014.7025575 (cf. p. 40).

Soubies, Emmanuel, Laure Blanc-Féraud et Gilles Aubert (2015a). « A Continuous Exact lo Penalty (CELo) for Least Squares Regularized Problem ». In : *SIAM Journal on Imaging Sciences* 8.3, p. 1607–1639. doi : 10.1137/151003714 (cf. p. 95).

Soubies, Emmanuel, Laure Blanc-Féraud et Gilles Aubert (2015b). « CELo : a continuous alternative to lo penalty ». In : *Signal Processing with Adaptive Sparse Structured Representations (SPARS)* (cf. p. 95).

Soubies, Emmanuel, Laure Blanc Féraud et Gilles Aubert (2015c). « Seuillage CELo pour la minimisation l2-lo : comparaisons avec IHT ». In : *Colloque Gretsi* (cf. p. 95).

Soubies, Emmanuel, Sébastien Schaub, Agata Radwanska, Ellen Van Obberghen-Schilling, Laure Blanc-Féraud et Gilles Aubert (2016a). « A Framework for Multi-angle Tirf Microscope Calibration ». In : *ISBI-International Symposium on Biomedical Imaging*. IEEE, p. 4 (cf. p. 47).

Soubies, Emmanuel, Laure Blanc-Féraud et Gilles Aubert (2016b). « A unified view of exact continuous penalties for l2-lo minimization ». In : *(submitted)* (cf. p. 166).

Soussen, Charles (2013). « Sparse approximation algorithms inspired by Orthogonal Least Squares for inverse problems ». Habilitation à diriger des recherches. Université de Lorraine. url : https://tel.archives-ouvertes.fr/tel-00924578 (cf. p. 81).

Soussen, Charles, Jérôme Idier, David Brie et Junbo Duan (2011). « From Bernoulli–Gaussian deconvolution to sparse signal restoration ». In : *IEEE Transactions on Signal Processing* 59.10, p. 4572–4584. doi : 10.1109/TSP.2011.2160633 (cf. p. 81, 82).

Soussen, Charles, Rémi Gribonval, Jérôme Idier et Cédric Herzet (2013). « Joint k-step analysis of orthogonal matching pursuit and orthogonal least squares ». In : *Information Theory, IEEE Transactions on* 59.5, p. 3158–3174. doi : 10.1109/TIT.2013.2238606 (cf. p. 80).

Soussen, Charles, Jérome Idier, Junbo Duan et David Brie (2015). « Homotopy based algorithms for Lo-regularized least-squares ». In : *IEEE Transactions on Signal Processing* 63.13, p. 3301–3316. doi : 10.1109/TSP.2015.2421476 (cf. p. 72, 83, 131, 140, 141).

Stabley, Daniel R, Thomas Oh, Sanford M Simon, Alexa L Mattheyses et Khalid Salaita (2015). « Real-time fluorescence imaging with 20 nm axial resolution ». In : *Nature communications* 6. doi : 10.1038/ncomms9307 (cf. p. 24).

Strong, David et Tony Chan (2003). « Edge-preserving and scale-dependent properties of total variation regularization ». In : *Inverse problems* 19.6, S165. doi : 10.1088/0266-5611/19/6/059 (cf. p. 39).

Tang, Zijian, Rocco Claudio Cannizzaro, Geert Leus et Paolo Banelli (2007). « Pilot-assisted time-varying channel estimation for OFDM systems ». In : *IEEE Transactions on Signal Processing* 55.5, p. 2226–2238 (cf. p. 146).

Tao, Pham Dinh et al. (2005). « The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems ». In : *Annals of Operations Research* 133.1-4, p. 23–46. doi : 10.1007/s10479-004-5022-1 (cf. p. 119, 120).

Temlyakov, Vladimir N (2008). « Greedy approximation ». In : *Acta Numerica* 17, p. 235–409. doi : 10.1017/S0962492906380014 (cf. p. 80).

Tibshirani, Robert (1996). « Regression shrinkage and selection via the lasso ». In : *Journal of the Royal Statistical Society. Series B (Methodological)*, p. 267–288 (cf. p. 76).

Tropp, Joel A (2004). « Greed is good : Algorithmic results for sparse approximation ». In : *IEEE Transactions on Information Theory* 50.10, p. 2231–2242. doi : 10.1109/TIT.2004.834793 (cf. p. 80).

Tropp, Joel A (2006). « Just relax : Convex programming methods for identifying sparse signals in noise ». In : *Information Theory, IEEE Transactions on* 52.3, p. 1030–1051. doi : 10.1109/TIT.2005.864420 (cf. p. 73).

Tropp, Joel A et Stephen J Wright (2010). « Computational methods for sparse solution of linear inverse problems ». In : *Proceedings of the IEEE* 98.6, p. 948–958. doi : 10.1109/JPROC.2010.2044010 (cf. p. 76).

Truskey, GA, JS Burmeister, E Grapa et WM Reichert (1992). « Total internal reflection fluorescence microscopy (TIRFM). II. Topographical mapping of relative cell/substratum separation distances ». In : *Journal of cell science* 103.2, p. 491–499. issn : 1477-9137 (cf. p. 24).

Tuncer, T Engin et Benjamin Friedlander (2009). *Classical and modern direction-of-arrival estimation*. Academic Press (cf. p. 150).

Van Lieshout, MNM (2000). « Markov point processes ». In : *Markov point processes and Their Applications* (cf. p. 25).

Wang, Jian, Seokbeop Kwon et Byonghyo Shim (2012). « Generalized orthogonal matching pursuit ». In : *Signal Processing, IEEE Transactions on* 60.12, p. 6202–6216. doi : 10.1109/TSP.2012.2218810 (cf. p. 80).

Weston, Jason, André Elisseeff, Bernhard Schölkopf et Mike Tipping (2003). « Use of the zero norm with linear models and kernel methods ». In : *The Journal of Machine Learning Research* 3, p. 1439–1461. issn : 1532-4435 (cf. p. 88).

Yang, Qian (2010). « 3D Reconstruction and measurement of microtubules from multiple angle-total internal reflection fluorescence microscopy ». Thèse de doct. Yale University (cf. p. 25, 26).

Yang, Qian, Alexander Karpikov, Derek Toomre et James Duncan (2011). « 3D Reconstruction of microtubules from multiple-angle total internal reflection fluorescence microscopy using Bayesian framework ». In : *IEEE transactions on Image Processing*, p. 2248–2259. doi : 10.1109/TIP.2011.2114359 (cf. p. 25).

Zhang, Bo, Josiane Zerubia et Jean-Christophe Olivo-Marin (2007). « Gaussian approximations of fluorescence microscope point-spread function models ». In : *Applied Optics* 46.10, p. 1819–1829. doi : 10.1364/AO.46.001819 (cf. p. 16).

Zhang, Cun-Hui (2008). « Discussion : One-step sparse estimates in nonconcave penalized likelihood models ». In : *The Annals of Statistics* 36.4, p. 1553–1560. doi : 10.1214/07-AOS0316A (cf. p. 89).

Zhang, Cun-Hui (2010). « Nearly unbiased variable selection under minimax concave penalty ». In : *The Annals of Statistics* 38.2, p. 894–942. doi : 10.1214/09-AOS729 (cf. p. 89, 178).

Zhang, Tong (2009). « Multi-stage convex relaxation for learning with sparse regularization ». In : *Advances in Neural Information Processing Systems*, p. 1929–1936 (cf. p. 90).

Zhang, Tong (2011). « Adaptive forward-backward greedy algorithm for learning sparse representations ». In : *IEEE Transactions on Information Theory* 57.7, p. 4689–4708. doi : 10.1109/TIT.2011.2146690 (cf. p. ix, 82).

Zhang, Wenwen et Qian Chen (2009). « Signal-to-noise ratio performance comparison of electron multiplying CCD and intensified CCD detectors ». In : *Image Analysis and Signal Processing, 2009. IASP 2009. International Conference on*. IEEE, p. 337–341. doi : 10.1109/IASP.2009.5054588 (cf. p. 22, 176).

Zou, Hui (2006). « The adaptive lasso and its oracle properties ». In : *Journal of the American statistical association* 101.476, p. 1418–1429. doi : 10.1198/016214506000000735 (cf. p. 87).

Résumé :

Cette thèse s'intéresse à deux problèmes rencontrés en traitement du signal et des images. Le premier concerne la reconstruction 3D de structures biologiques à partir d'acquisitions multi-angles en microscopie par réflexion totale interne (MA-TIRF). Dans ce contexte, nous proposons de résoudre le problème inverse avec une approche variationnelle et étudions l'effet de la régularisation. Une batterie d'expériences, simples à mettre en œuvre, sont ensuite proposées pour étalonner le système et valider le modèle utilisé. La méthode proposée s'est montrée être en mesure de reconstruire avec précision un échantillon phantom de géométrie connue sur une épaisseur de 400 nm*, de co-localiser deux molécules fluorescentes marquant les mêmes structures biologiques et d'observer des phénomènes biologiques connus, le tout avec une résolution axiale de l'ordre de* 20 nm*. La deuxième partie de cette thèse considère plus précisément la régularisation ℓ_0 et la minimisation du critère moindres carrés pénalisé (ℓ_2-ℓ_0) dans le contexte des relaxations continues exactes de cette fonctionnelle. Nous proposons dans un premier temps la pénalité CEL0 (Continuous Exact ℓ_0) résultant en une relaxation de la fonctionnelle ℓ_2-ℓ_0 préservant ses minimiseurs globaux et pour laquelle de tout minimiseur local on peut définir un minimiseur local de ℓ_2-ℓ_0 par un simple seuillage. Par ailleurs, nous montrons que cette relaxation élimine des minimiseurs locaux de la fonctionnelle initiale. La minimisation de cette fonctionnelle avec des algorithmes d'optimisation non-convexe est ensuite utilisée pour différentes applications montrant l'intérêt de la minimisation de la relaxation par rapport à une minimisation directe du critère ℓ_2-ℓ_0. Enfin, une vue unifiée des pénalités continues de la littérature est proposée dans ce contexte de reformulation exacte du problème.*

Mots clés : *problèmes inverses, microscopie MA-TIRF, reconstruction 3D, étalonnage microscope, optimisation parcimonieuse, norme-ℓ_0, relaxations continues exactes, équivalence des minimiseurs.*

Abstract :

This thesis is devoted to two problems encountered in signal and image processing. The first one concerns the 3D reconstruction of biological structures from multi-angle total interval reflection fluorescence microscopy (MA-TIRF). Within this context, we propose to tackle the inverse problem by using a variational approach and we analyze the effect of the regularization. A set of simple experiments is then proposed to both calibrate the system and validate the used model. The proposed method has been shown to be able to reconstruct precisely a phantom sample of known geometry on a 400 nm *depth layer, to co-localize two fluorescent molecules used to mark the same biological structures and also to observe known biological phenomena, everything with an axial resolution of* 20 nm*. The second part of this thesis considers more precisely the ℓ_0 regularization and the minimization of the penalized least squares criteria (ℓ_2-ℓ_0) within the context of exact continuous relaxations of this functional. Firstly, we propose the Continuous Exact ℓ_0 (CEL0) penalty leading to a relaxation of the ℓ_2-ℓ_0 functional which preserves its global minimizers and for which from each local minimizer we can define a local minimizer of ℓ_2-ℓ_0 by a simple thresholding. Moreover, we show that this relaxed functional eliminates some local minimizers of the initial functional. The minimization of this functional with nonsmooth nonconvex algorithms is then used on various applications showing the interest of minimizing the relaxation in contrast to a direct minimization of the ℓ_2-ℓ_0 criteria. Finally we propose a unified view of continuous penalties of the literature within this exact problem reformulation framework.*

Key words : *inverse problems, MA-TIRF microscopy, 3D reconstruction, microscope calibration, sparse optimization, ℓ_0-norm, exact continuous relaxations, minimizers equivalence.*