Algorithmes d'optimisation de critères pénalisés pour la restauration d'images. Application à la déconvolution de trains d'impulsions en imagerie ultrasonore.

Christian Labat

Ecole Centrale de Nantes ***Université de Nantes***

ÉCOLE DOCTORALE

SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DES MATÉRIAUX

Année 2006

Thèse de Doctorat

Diplôme délivré conjointement par
L'École Centrale de Nantes et l'Université de Nantes

Spécialité : Automatique, Robotique, Traitement du Signal et Informatique Appliquée

Présentée et soutenue publiquement par:

CHRISTIAN LABAT

le 11 décembre 2006
à l'Ecole Centrale de Nantes

TITRE

ALGORITHMES D'OPTIMISATION DE CRITÈRES PÉNALISÉS POUR LA RESTAURATION D'IMAGES. APPLICATION À LA DÉCONVOLUTION DE TRAINS D'IMPULSIONS EN IMAGERIE ULTRASONORE.

JURY

Président :	Jean Pierre Guedon	Professeur à l'Université de Nantes
Rapporteurs :	Thierry Chonavel	Professeur à Telecom Bretagne (Brest)
	Mila Nikolova	Chargée de recherche CNRS au CMLA (ENS de Cachan)
Examinateurs :	Pierre Charbonnier	Directeur de recherche LCPC (Strasbourg)
	Yves Goussard	Professeur à l'École Polytechnique de Montréal (Canada)
	Jérôme Idier	Chargé de recherche CNRS à l'IRCCyN (Nantes)

Directeur de thèse : Jérôme Idier
Laboratoire : Institut de Recherche en Communications et Cybernétique de Nantes (IRCCyN)

N° ED 0366 – 292

A mes parents

"J'ouvre une parenthèse. Si vous avez un peu trop d'air, je la fermerai tout de suite."
Alphonse Allais

Table des matières

Table des figures

Liste des tableaux

Notations

$\mathbb{R}$	:	corps des réels
$\mathbb{C}$	:	corps des complexes
s	:	les scalaires sont notés par des minuscules
$\boldsymbol{v}$	:	les vecteurs sont notés par des minuscules en gras
$\boldsymbol{v}^T$	:	transposé du vecteur $\boldsymbol{v}$
$\mathbf{M}$	:	les matrices sont notées par des majuscules en gras
$\mathbf{M}^T$	:	transposée de la matrice $\mathbf{M}$
$\mathbf{M}^*$	:	transposée conjuguée de la matrice $\mathbf{M}$
$\mathbf{M}^{-1}$	:	matrice inverse de la matrice carrée $\mathbf{M}$
$\mathbf{M}^\dagger$	:	inverse généralisée de la matrice $\mathbf{M}$
Diag $\{\boldsymbol{v}\}$	:	matrice diagonale dont la diagonale est le vecteur $\boldsymbol{v}$
$m\,[i,j]$	:	$(ij)^{ème}$ élément de la matrice $\mathbf{M}$
$\mathbf{I}_n$	:	matrice identité de dimension $n \times n$
$\ker(\mathbf{M})$	:	noyau de la matrice $\mathbf{M}$
$\mathcal{C}^1$	:	continûment différentiable
$\mathcal{C}^2$	:	deux fois continûment différentiable
$f'(x)$	:	dérivée première de la fonction f par rapport à x
$\nabla\mathcal{J}(\boldsymbol{x})$	:	gradient du critère $\mathcal{J}$ par rapport à $\boldsymbol{x}$
$\nabla^2\mathcal{J}(\boldsymbol{x})$	:	Hessien du critère $\mathcal{J}$ par rapport à $\boldsymbol{x}$
$\lvert s\rvert$	:	module du scalaire s
$\lVert\boldsymbol{v}\rVert$	:	norme euclidienne du vecteur $\boldsymbol{v}$
$\lVert\mathbf{M}\rVert$	:	2-norme de la matrice $\mathbf{M}$ définie par $\lVert\mathbf{M}\rVert = \sup_{\boldsymbol{x}\neq 0} \frac{\lVert\mathbf{M}\boldsymbol{x}\rVert}{\lVert\boldsymbol{x}\rVert}$
$\lVert\boldsymbol{v}\rVert_{\mathbf{Q}}$	:	$\mathbf{Q}$-norme du vecteur $\boldsymbol{v}$
$p(\cdot)$	:	densité de probabilité
$f \propto g$	:	f est proportionnelle à g
$x \sim p(.)$	:	x suit une loi de probabilité $p(.)$
$x\lvert y$	:	x conditionnelement à y
$\mathcal{N}(\mu, \sigma^2)$	:	loi normale de moyenne μ et variance σ^2
$\star_1$	:	convolution 1D discrète
$\star_2$	:	convolution 2D discrète

Abréviations

SDP	:	symétrique défini positif
SQ	:	semi-quadratiques
GR	:	Geman et Reynolds
GY	:	Geman et Yang
GC	:	gradient conjugué
GCP	:	GC préconditionné
SQ+GCP	:	famille d'algorithmes SQ avec direction tronquée par algorithme GCP
SQ1D	:	forme scalaire des algorithmes SQ
GR1D	:	forme scalaire de l'algorithme de GR
GY1D	:	forme scalaire de l'algorithme de GY
GCNL	:	GC non linéaire
GCNL+SQ1D	:	famille d'algorithmes GCNL avec recherche du pas selon SQ1D
GCNL+GR1D	:	famille d'algorithmes GCNL avec recherche du pas selon GR1D
GCNL+GY1D	:	famille d'algorithmes GCNL avec recherche du pas selon GY1D
GCPPR	:	forme GCNL de Polak-Ribiere
GCPFR	:	forme GCNL de Fletcher-Reeves
GCPHS	:	forme GCNL de Hestenes-Stiefel
GCPLS	:	forme GCNL de Liu-Storey
GCPPR+SQ1D	:	algorithme GCPPR avec recherche du pas selon SQ1D
GCPPR+GR1D	:	algorithme GCPPR avec recherche du pas selon GR1D
GCPPR+GY1D	:	algorithme GCPPR avec recherche du pas selon GY1D
RSB	:	rapport signal sur bruit
CND	:	contrôle non destructif

Chapitre I

INTRODUCTION

Dans les domaines de traitement du signal et des images, de nombreux problèmes se ramènent à celui de l'optimisation d'un critère pénalisé permettant d'obtenir la solution la plus plausible, compte tenu des données et des informations sur l'application considérée. La résolution numérique de ce problème d'optimisation entraîne un coût de calcul qui devient d'autant plus important que sa taille est grande. Or, le traitement des images a pour spécificité d'être presque systématiquement confronté à des problèmes de grande taille. En effet, le nombre d'inconnues à estimer correspond au nombre de pixels de l'image qui est très souvent conséquent. Un exemple de domaine généralement confronté à des images de grande taille est celui de l'imagerie biomédicale.

La taille actuelle de nombreux problèmes de restauration et de reconstruction d'images est déjà importante, mais elle est certainement amenée à augmenter dans les prochaines années. Ceci est une conséquence directe de l'augmentation de la résolution des capteurs d'une part, et d'autre part, du développement de l'imagerie 3D, voire de l'imagerie dynamique "2D+t" ou "3D+t" qui en accroît d'autant plus la taille. La véritable difficulté réside dans l'augmentation de la taille des données plus rapide que celle de la puissance de calcul des ordinateurs. Ainsi, des avancées sur les méthodes d'optimisation dédiées aux problèmes de grande taille sont nécessaires pour compenser ce décalage.

Le contexte général de ce travail de thèse est d'étudier les différents aspects de méthodes d'optimisation dédiées à la résolution de problèmes inverses en traitement d'images. Plus précisément, nous proposons des méthodes d'optimisation des critères pénalisés qui répondent de manière pertinente à la restauration et à la reconstruction d'images pour un coût de calcul acceptable.

I.1 Position du problème

La solution de nombreux problèmes de restauration et de reconstruction d'images peut être définie comme le minimiseur $\hat{\boldsymbol{x}}$ d'un critère pénalisé $\mathcal{J} : \mathbb{R}^N \mapsto \mathbb{R}$ qui prend en compte conjointement les données observées et les informations préalables sur la solution. La solution $\hat{\boldsymbol{x}}$ ne peut généralement pas s'exprimer sous une forme analytique et un algorithme de minimisation doit être mis en œuvre pour en fournir un estimé. Bien que cette approche pénalisée fournisse des solutions de qualité satisfaisante, son implémentation actuelle a souvent pour inconvénient d'avoir une charge calculatoire trop importante pour les problèmes de grande taille.

Une résolution pertinente de ce problème d'optimisation doit tirer profit de la structure du critère pénalisé, ce qui est le cas des algorithmes semi-quadratiques auxquels nous prêterons une attention particulière tout au long de cette thèse.

[A] Algorithmes semi-quadratiques

Il est possible de tirer parti de la structure des critères pénalisés pour la mise en œuvre algorithmique du problème de minimisation de $\mathcal{J}$. Ainsi, des algorithmes d'optimisation spécifiques semi-quadratiques (SQ) exploitant la forme analytique des critères pénalisés ont été utilisés pour la restauration d'image préservant les discontinuités [Charbonnier *et al.*, 1997; Nikolova et Ng, 2005; Allain *et al.*, 2006]. Ces algorithmes SQ découlent des formulations de Geman et Yang (GY) et de Geman et Reynolds (GR) initialement proposées dans un cadre stochastique [Geman et Reynolds, 1992; Geman et Yang, 1995]. Néanmoins, la difficulté d'implémenter les algorithmes de GR et de GY pour les problèmes de grande taille a été soulignée dès leur introduction. Dans [Charbonnier, 1994] il est précisé que les algorithmes de GR et de GY "impliquent la résolution de systèmes linéaires de grande taille, lourds à manipuler d'un point de vue informatique".

[B] Algorithmes semi-quadratiques approchés

Afin de pallier cette difficulté d'inversion d'un système de grande taille, des versions approchées des algorithmes semi-quadratiques ont été adoptées dont le coût d'implémentation est bien plus faible. Comme précisé dans [Charbonnier *et al.*, 1997], il est possible d'utiliser l'algorithme du gradient conjugué pour résoudre de manière approchée le système linéaire des algorithmes de GR et de GY. Cette approche a été mise en œuvre dans [Nikolova et Ng, 2001, 2005] pour l'algorithme de GY et dans [Belge *et al.*, 2000] et [Nikolova et Ng, 2005] pour l'algorithme de GR. La famille d'algorithmes semi-quadratiques (GR ou GY) avec direction tronquée par l'algorithme du gradient conjugué préconditionné (GCP) est désignée par SQ+GCP. L'inconvénient de ces formes approchées est que les résultats de convergence valables pour les formes exactes de GR et de GY ne s'appliquent plus aux formes approchées, jusqu'à présent. La question de la convergence des formes SQ+GCP se pose. Ce premier point fera l'objet d'une analyse dans le cadre de ce travail.

[C] Algorithmes du gradient conjugué non linéaire

Les algorithmes de GR et de GY peuvent être interprétés comme des algorithmes de type quasi-Newton à pas fixe [Allain *et al.*, 2006]. Or, une alternative classique aux algorithmes de type quasi-Newton est le recours à des algorithmes de type gradient conjugué non linéaire (GCNL) dont l'intérêt est un plus faible encombrement mémoire et un coût de calcul par itération inférieur [Bertsekas, 1999, p. 155]. L'utilisation des algorithmes de type GCNL pour la minimisation des critères pénalisés non quadratiques a été proposée dans la communauté du traitement de l'image, mais sans assurance de la convergence [Fessler et Booth, 1999; Rivera et Marroquin, 2003]. Cette proposition consiste à utiliser la forme scalaire des algorithmes SQ (SQ1D) pour la recherche du pas nécessitée par les algorithmes GCNL. La famille d'algorithmes de minimisation qui en résulte est appelée GCNL+SQ1D.

Pourtant, il existe dans la littérature de l'optimisation des résultats de convergence des algorithmes GCNL. Ces résultats de convergence sont malheureusement peu connus en traitement du signal et des images. Néanmoins, l'utilisation des résultats de convergence des algorithmes GCNL de la littérature de l'optimisation, comme [Nocedal et Wright, 1999, Theorem 5.8], a pour principal inconvénient de manquer de simplicité.

La question est donc de savoir s'il est possible d'obtenir la convergence des algorithmes GCNL+SQ1D tels que déjà utilisés en traitement des images, sans avoir recours aux conditions à vérifier nécessitées par les résultats existants de la littérature de l'optimisation. L'intérêt étant une mise en œuvre largement simplifiée. Ce deuxième point fera aussi l'objet d'une analyse dans le cadre de ce travail.

I.2 Contributions

Les travaux menés pendant cette thèse concernent à la fois les aspects algorithmiques liés à la minimisation des critères pénalisés et également l'application de cette approche à un problème de déconvolution d'images en contrôle non destructif par ultrasons.

[A] Aspects algorithmiques

Cette thèse se place dans le prolongement des travaux [Idier, 2001; Allain *et al.*, 2006] pour les aspects algorithmiques. La difficulté de l'utilisation des critères pénalisés non quadratiques réside dans la recherche d'un algorithme de minimisation à la fois simple, efficace et convergent pour les problèmes de grande taille. La question est donc de savoir s'il est possible de trouver un tel algorithme. Jusqu'à présent, aucun algorithme ne répond entièrement à ces exigences. La convergence d'un algorithme de minimisation n'est pas une considération purement théorique. Il s'agit tout simplement de s'assurer que l'algorithme considéré effectue bien ce qu'on attend de lui. C'est donc un point essentiel. La convergence des algorithmes SQ s'appuie sur les propriétés de convexité[1] et de coercivité[2] [Allain *et al.*, 2006]. Nous montrons dans un effort de généralité mathématique qu'il est possible de relaxer l'hypothèse de convexité pour les algorithmes étudiés. En revanche, l'utilisation d'un critère non convexe ne permet pas d'exclure la présence de minima locaux. On n'est alors pas assuré d'obtenir la solution globale $\hat{\boldsymbol{x}}$. C'est pourquoi nous nous sommes limité en pratique à des critères convexes.

Nos contributions sur les aspects algorithmiques de la minimisation des critères pénalisés non quadratiques consistent à :

- Démontrer la convergence des familles d'algorithmes SQ+GCP et GCNL+SQ1D pour un nombre quelconque de sous-itérations de GCP ou de SQ1D. Les sous-itérations des deux familles précédentes correspondent respectivement à la recherche d'une direction de descente (vecteur) et à la recherche d'un pas (scalaire).
- Etablir des liens forts entre les familles d'algorithmes SQ+GCP et GCNL+SQ1D. Nous montrons que les algorithmes GCNL+SQ1D constituent une famille d'algorithmes de minimisation contenant les algorithmes exacts de GR et de GY. En réinterprétant les algorithmes de GR et de GY comme étant des choix particuliers de préconditionnement, certaines alternatives se démarquent par leur coût d'implémentation plus faible. De plus, la recherche du pas SQ1D peut être interprétée comme une généralisation naturelle du pas constant.
- Illustrer expérimentalement en déconvolution d'images le fait que les versions les plus efficaces des familles d'algorithmes SQ+GCP et GCNL+SQ1D consistent à effectuer peu de sous-itérations. Autrement dit, il est préférable de considérer des versions qui s'éloignent substantiellement des algorithmes SQ exacts.

Le travail présenté dans ce document permet donc de montrer que les familles d'algorithmes SQ+GCP et GCNL+SQ1D sont à la fois simples, efficaces et convergentes pour les problèmes de grande taille.

[1] Le critère $\mathcal{J} : \mathbb{R}^N \to \mathbb{R}$ est dit *convexe* si

$$\mathcal{J}(t\boldsymbol{x} + (1-t)\boldsymbol{y}) \leq t\mathcal{J}(\boldsymbol{x}) + (1-t)\mathcal{J}(\boldsymbol{y}), \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N, \forall t \in [0,1].$$

Il est dit *strictement* convexe si l'inéquation précédente est stricte pour tout $\boldsymbol{x} \neq \boldsymbol{y}$ et pour tout $t \in (0,1)$.

[2] Le critère $\mathcal{J} : \mathbb{R}^N \to \mathbb{R}$ est dit *coercif* si $\lim_{\|\boldsymbol{x}\| \to \infty} \mathcal{J}(\boldsymbol{x}) = \infty$.

[B] Contrôle non destructif par ultrasons

Au cours de notre thèse, nous avons participé à un contrat de recherche entre l'Institut de Recherche en Communication et Cybernétique de Nantes (IRCCyN, CNRS-UMR 6597) et le Groupe Systèmes Dynamiques et Traitement de l'Information (GSDTI) du département STEP (Simulation et Traitement de l'information pour l'Exploitation des systèmes de Production) à la direction des Etudes et Recherches d'Electricité de France (EDF). Les membres de cette collaboration étaient Jérôme Idier et moi-même (IRCCyN) ainsi que Benoît Richard et Laurence Chatellier (GSDTI). L'objectif de cette collaboration était d'appliquer l'approche par minimisation de critère pénalisé pour le problème du contrôle non destructif par ultrasons et de développer une extension 2D de la méthode de déconvolution 1D de [Gautier *et al.*, 2001]. Cette extension ne peut pas se faire directement sur les données réelles et nous avons proposé une procédure de recalage pour en assurer la robustesse. Nous illustrons les bonnes performances de la méthode proposée alliant recalage et déconvolution 2D, tant au niveau de la qualité de restauration que de l'efficacité numérique.

I.3 Organisation du document

L'exposé débute par une présentation du cadre méthodologique associé au traitement des images préservant les discontinuités. Le chapitre II traite de la régularisation dans un cadre déterministe (au sens de Tikhonov généralisé). On y présente l'approche pénalisée qui sera au cœur du document. Le chapitre III traite de la régularisation dans un cadre probabiliste (approche bayésienne).

Le chapitre IV expose la nécessité de recourir à un algorithme d'optimisation. Nous effectuons un tour d'horizon des algorithmes itératifs employés pour la minimisation des critères pénalisés différentiables. L'objet de ce chapitre est d'introduire les algorithmes semi-quadratiques (SQ) spécifiquement développés pour la minimisation des critères pénalisés et de les replacer par rapport aux méthodes générales d'optimisation. Nous indiquons alors que les algorithmes SQ peuvent recevoir plusieurs interprétations possibles. En particulier, les algorithmes SQ peuvent être interprétés comme minimisant une approximation quadratique majorant le critère pénalisé. Ce point de vue nous sera particulièrement utile pour établir les résultats de convergence des deux prochains chapitres.

Le chapitre V indique que les algorithmes SQ sont généralement trop coûteux pour les problèmes de grande taille. Il a donc été naturellement proposé d'approcher les algorithmes SQ pour en réduire leur coût. Ce type d'approche consiste à employer une méthode itérative tronquée, dans le sens d'un faible nombre d'itérations vis-à-vis de la dimension du problème, afin de résoudre de manière approchée le système linéaire induit par les algorithmes SQ. L'algorithme du gradient conjugué (GC), éventuellement préconditionné (GCP) est tout indiqué pour résoudre de manière approchée ce système linéaire. Les algorithmes qui en résultent sont notés SQ+GCP. La convergence de ces algorithmes n'est pas assurée en utilisant les résultats des formes SQ exacts. A notre connaissance, la convergence des algorithmes SQ+GCP est restée un problème ouvert, jusqu'à présent. La contribution de ce chapitre consiste précisément à démontrer la convergence des algorithmes SQ+GCP pour un nombre quelconque d'itérations de l'algorithme GCP. Les principaux résultats de ce chapitre se trouvent dans le rapport technique [2].

Le chapitre VI s'intéresse aux algorithmes du gradient conjugué non linéaire (GCNL) qui sont une alternative possible aux algorithmes SQ. Les algorithmes GCNL nécessitent la mise en œuvre d'une recherche du pas. Nous proposons d'utiliser la forme scalaire tronquée des algorithmes SQ pour cette recherche du pas. Les algorithmes qui en résultent sont notés GCNL+SQ1D. La

première contribution de ce chapitre consiste à établir des liens forts entre les algorithmes SQ et les algorithmes GCP+SQ1D. Nous montrons que les algorithmes GCP+SQ1D constituent une famille d'algorithmes de minimisation contenant les algorithmes SQ exacts. En réinterprétant les algorithmes SQ comme étant des choix particuliers de préconditionnement, certaines alternatives se démarquent par leur coût d'implémentation plus faible. La seconde contribution de ce chapitre est d'établir la convergence des algorithmes GCNL+SQ1D pour un nombre quelconque de sous-itérations de recherche du pas SQ1D. Ce résultat ne fait pas appel aux résultats de convergence des algorithmes GCNL présents dans la littérature de l'optimisation qui manquent de simplicité. Ce manque de simplicité est exposé dans l'annexe A, page 139. L'intérêt des résultats de convergence de ce chapitre pour les algorithmes GCNL+SQ1D consiste à apporter une nette simplification par rapport aux résultats de l'optimisation. Les principaux résultats de ce chapitre se trouvent dans les publications [2,5,6].

Le chapitre VII mène une comparaison des vitesses de convergence expérimentale des algorithmes SQ+GCP et GCNL+SQ1D sur un problème de déconvolution d'image de grande taille. Nous montrons expérimentalement qu'il est largement préférable pour les familles d'algorithmes SQ+GCP et GCNL+SQ1D de ne pas effectuer un nombre trop grand de sous-itérations. De plus, les algorithmes SQ+GCP ne devraient pas être vus comme des versions approchées des algorithmes SQ, mais bien comme des algorithmes propres, avec leur paramètre de réglage. Si les familles d'algorithmes SQ+GCP et GCNL+SQ1D ont une vitesse de convergence expérimentale de même ordre, elles se distinguent néanmoins par leur simplicité de réglage. En effet, il apparaît expérimentalement que le réglage optimum de la famille d'algorithmes GCNL+SQ1D est plus simple que celui de la famille d'algorithmes SQ+GCP.

Le chapitre VIII traite de l'approche pénalisée appliquée au contrôle non destructif par ultrasons. Le travail présenté dans ce chapitre a été initié dans le cadre d'un contrat de recherche entre l'IRCCyN et le GSDTI d'EDF. La méthode proposée consiste en une extension 2D de la méthode de déconvolution 1D de [Gautier *et al.*, 2001]. Cependant, cette extension n'est pas directe et il est nécessaire de mettre en œuvre une procédure de recalage pour en assurer la robustesse. La méthode proposée, illustrée sur des données réelles, se révèle simple et robuste. Elle permet d'une part d'augmenter la résolution temporelle et latérale des images ultrasonores et d'en faciliter ainsi l'interprétation par un expert. D'autre part, elle répond de manière très satisfaisante au problème crucial de l'estimation de la hauteur de défauts de petite taille. Les principaux résultats de ce chapitre se trouvent dans les publications [3,7,9,10].

L'annexe A est une étude bibliographique sur les résultats de convergence des GCNL présents dans la littérature de l'optimisation. Nous montrons que la mise en œuvre de ces résultats de convergence manque de simplicité.

Les démonstrations des résultats des chapitres IV, V et VI sont reportées dans l'annexe B.

L'annexe C est constituée de [1] à paraître dans *Journal of Optimization Theory and Applications*. Les résultats de cette annexe sont utilisés pour le chapitre VI.

L'annexe D contient les résultats expérimentaux de l'influence de certains paramètres des algorithmes SQ+GCP et GCNL+SQ1D.

L'annexe E est constituée de la publication [4]. Il s'agit de l'illustration d'outils propres au cadre probabiliste.

Publications

Article accepté

[1] C. Labat and J. Idier, **Convergence of conjugate gradient methods with a closed-form stepsize formula**, à paraître, *Journal of Optimization Theory and Applications*, 2007.
Comments from the Reviewers : "This paper is an excellent paper on the advancement of Conjugate Gradient Methods, the referee would like to strongly suggest that it could be published in JOTA asap."

Articles soumis ou en préparation

[2] C. Labat, J. Idier, **Convergence of truncated half-quadratic algorithms using conjugate gradient**, *Tech. Rep.,* IRCCyN, 2006.

[3] C. Labat, J. Idier, B. Richard and L. Chatellier, **Ultrasonic nondestructive evaluation based on 2D high resolution deconvolution**, *Tech. Rep.,* IRCCyN, 2006.

Conférences

[4] C. Labat and J. Idier, **Sparse blind deconvolution accounting for time-shift ambiguity,** *in Proc. IEEE ICASSP*, Toulouse, France, mai 2006.
Comments from the Reviewers : "The contribution of this paper is solid." "A very nice paper that describes Bayesian techniques for resolving time delay and amplitude ambiguities in blind deconvolution."

[5] C. Labat and J. Idier, **Preconditioned conjugate gradient without linesearch : a comparison with the half-quadratic approach for edge-preserving image restoration,** *in IS&T/SPIE Symposium on Electronic Imaging*, San Jose, CA, janvier 2006.

[6] C. Labat and J. Idier, **Comparaison entre les algorithmes semi-quadratiques et gradients conjugués préconditionnés pour la restauration d'image préservant les bords,** *in Actes 20e coll. GRETSI*, Louvain-la-Neuve, Belgium, septembre 2005.

[7] B. Richard, and L. Chatellier, C. Labat, O. Dupond, J. Idier, **Caractérisation de défauts plans par contrôle ultrasonore et traitement de signal adapté**, *in Journées COFREND*, Beaune, France, mai 2005.

[8] C. Labat, J. Idier, and Y. Goussard, **Comparison between half-quadratic and preconditioned conjugate gradient algorithms for MRI reconstruction**, *in PSIP'2005 : Physics in signal and Image processing*, Toulouse, France, janvier 2005.

[9] C. Labat, J. Idier, B. Richard, and L. Chatellier, **Ultrasonic nondestructive testing based on 2D deconvolution**, *in PSIP'2005*, Toulouse, France, janvier 2005.

Rapport de contrat de recherche

[10] C. Labat, J. Idier, **Prise en compte de la largeur du faisceau en déconvolution pour le CND par ultrasons**, *Rapport de contrat EDF*, IRCCyN, juillet 2004.

Prix

ICASSP 2006 : Finaliste du "Student Paper Contest" de la thématique "Signal Processing Theory and Methods topic" pour le papier intitulé "Sparse blind deconvolution accounting for time-shift ambiguity" [4].

Bibliographie

[Allain *et al.*, 2006] M. Allain, J. Idier et Y. Goussard. On global and local convergence of half-quadratic algorithms. *IEEE Trans. Image Processing*, 15 (5) : 1130–1142, mai 2006.

[Belge *et al.*, 2000] M. Belge, M. Kilmer et E. Miller. Wavelet domain image restoration with adaptive edge-preserving regularization. *IEEE Trans. Image Processing*, 9 (4) : 597–608, avril 2000.

[Bertsekas, 1999] D. P. Bertsekas. *Nonlinear programming*. Athena Scientific, Belmont, MA, USA, 2ème edition, 1999.

[Charbonnier, 1994] P. Charbonnier. *Reconstruction d'image : régularisation avec prise en compte des discontinuités*. thèse de doctorat, Université de Nice-Sophia Antipolis, Nice, septembre 1994.

[Charbonnier *et al.*, 1997] P. Charbonnier, L. Blanc-Féraud, G. Aubert et M. Barlaud. Deterministic edge-preserving regularization in computed imaging. *IEEE Trans. Image Processing*, 6 (2) : 298–311, février 1997.

[Fessler et Booth, 1999] J. A. Fessler et S. D. Booth. Conjugate-gradient preconditionning methods for shift-variant PET image reconstruction. *IEEE Trans. Image Processing*, 8 (5) : 668–699, mai 1999.

[Gautier *et al.*, 2001] S. Gautier, J. Idier, F. Champagnat et D. Villard. Restoring separate discontinuities from ultrasonic data. In *Review of Progress in Quantitative Nondestructive Evaluation, AIP Conf. Proc. Vol 615(1)*, pages 686–690, Brunswick, ME, USA, juillet 2001.

[Geman et Reynolds, 1992] D. Geman et G. Reynolds. Constrained restoration and the recovery of discontinuities. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14 (3) : 367–383, mars 1992.

[Geman et Yang, 1995] D. Geman et C. Yang. Nonlinear image recovery with half-quadratic regularization. *IEEE Trans. Image Processing*, 4 (7) : 932–946, juillet 1995.

[Idier, 2001] J. Idier. Convex half-quadratic criteria and interacting auxiliary variables for image restoration. *IEEE Trans. Image Processing*, 10 (7) : 1001–1009, juillet 2001.

[Nikolova et Ng, 2001] M. Nikolova et M. Ng. Fast image reconstruction algorithms combining half-quadratic regularization and preconditioning. In *Proc. IEEE ICIP*, pages 277–280, Thessaloniki, Grèce, octobre 2001.

[Nikolova et Ng, 2005] M. Nikolova et M. Ng. Analysis of half-quadratic minimization methods for signal and image recovery. *SIAM J. Sci. Comput.*, 27 : 937–966, 2005.

[Nocedal et Wright, 1999] J. Nocedal et S. J. Wright. *Numerical optimization*. Springer Texts in Operations Research. Springer-Verlag, New York, NY, USA, 1999.

[Rivera et Marroquin, 2003] M. Rivera et J. Marroquin. Efficient half-quadratic regularization with granularity control. *Image and Vision Comp.*, 21 (4) : 345–357, avril 2003.

Chapitre II

INVERSION PAR APPROCHES PÉNALISÉES

II.1 Difficultés de l'inversion

On est parfois confronté au besoin d'obtenir une quantité physique indirectement disponible, mais reliée par des lois physiques à la mesure. La notion de *problème inverse* consiste à inverser les lois physiques reliant la quantité désirée à la mesure.

De nombreux problèmes inverses consistant à estimer un vecteur inconnu $\boldsymbol{x} \in \mathbb{R}^N$ à partir d'un vecteur observation $\boldsymbol{y} \in \mathbb{R}^M$ peuvent être modélisés sous la forme linéaire suivante

$$\boldsymbol{y} = \mathbf{H}\boldsymbol{x} + \boldsymbol{\epsilon} \tag{II.1}$$

avec $\mathbf{H}$ matrice de $\mathbb{R}^{M\times N}$ décrivant le phénomène physique et $\boldsymbol{\epsilon} \in \mathbb{R}^M$ vecteur de bruit représentant à la fois le bruit de mesure et les erreurs de modélisation. La matrice $\mathbf{H}$ est appelée *matrice d'observation* et rend compte du lien physique entre l'inconnue $\boldsymbol{x}$ et l'observation $\boldsymbol{y}$. Il est à noter que le modèle (II.1) n'implique pas nécessairement que le problème est de nature 1D. Un problème de nature 2D ou même 3D peut être mis sous une forme vectorielle via une réindexation. Le modèle (II.1) considéré est discret. Nous avons pris le parti d'illustrer la méthodologie problème inverse à travers une modélisation discrète pour deux raisons. D'une part ce cas met en jeu des outils mathématiques plus simples que le cas continu, ce qui permet de simplifier l'exposé de la méthodologie. D'autre part, un problème inverse doit en définitive être implémenté dans une architecture informatique discrète. Cette étape de discrétisation du modèle est donc toujours présente.

L'équation (II.1) décrivant l'origine des observations est appelée le *modèle direct*. Par la suite, nous supposerons que la partie modélisation du problème, c'est-à-dire la détermination

de la matrice d'observation, a été préalablement effectuée. La matrice d'observation $\mathbf{H}$ peut typiquement être une matrice de convolution ou de projection. L'inversion consiste alors à estimer le vecteur inconnu $\boldsymbol{x}$ connaissant l'observation $\boldsymbol{y}$ et la matrice d'observation $\mathbf{H}$. Cette inversion est rendue difficile par la présence inévitable du bruit représenté par le vecteur $\boldsymbol{\epsilon}$.

II.1.1 Problèmes mal posés

La résolution des problèmes d'inversion issus de la physique est rendue difficile par leur caractère *mal posé* en général. Au début du XX siècle, Hadamard a défini les trois conditions suivantes pour qu'un problème soit *bien posé* [Hadamard, 1901] :

- pour chaque donnée $\boldsymbol{y}$ dans une classe définie $\boldsymbol{\mathcal{Y}}$, il existe une solution $\boldsymbol{x}$ dans une classe prescrite $\boldsymbol{\mathcal{X}}$ (*existence*) ;
- la solution est unique dans $\boldsymbol{\mathcal{X}}$ (*unicité*) ;
- la dépendance de $\boldsymbol{x}$ par rapport à $\boldsymbol{y}$ est continue : lorsque l'erreur $\delta_{\boldsymbol{y}}$ sur la donnée $\boldsymbol{y}$ tend vers zéro, l'erreur $\delta_{\boldsymbol{x}}$ induite sur la solution $\boldsymbol{x}$ tend aussi vers zéro (*continuité*).

A contrario, un problème qui ne satisfait pas une de ces trois conditions est dit *mal posé*.

L'existence et l'unicité sont les deux conditions classiques pour la résolution d'équation. L'exigence de continuité est reliée à celle de stabilité de la solution. Cependant elle n'est pas suffisante, à elle seule, pour assurer la robustesse de la solution comme nous le verrons dans la section II.1.3.

Lorsque la matrice d'observation $\mathbf{H}$ est carrée ($N = M$) et inversible ($\ker \mathbf{H} = \{\mathbf{0}\}$), une inversion directe, au sens de la solution calculée par $\hat{\boldsymbol{x}} = \mathbf{H}^{-1}\boldsymbol{y}$, peut être envisagée. Dans tous les autres cas, cette inversion directe est inapplicable. La nécessité de définir des solutions bien posées utilisables dans des situations générales est à l'origine des développements de théories alternatives à l'inversion directe, comme l'*inversion généralisée*, et surtout celle de la *régularisation*.

II.1.2 Inversion généralisée

Une approche naturelle pour résoudre le problème inverse consiste à utiliser les moindres carrés, c'est-à-dire considérer l'ensemble

$$S_{\mathrm{MC}} = \{x \in \mathbb{R}^N, \min \|\mathbf{H}\boldsymbol{x} - \boldsymbol{y}\|\}. \tag{II.2}$$

Les éléments de l'ensemble S_{MC} peuvent être exprimés comme solution de l'équation normale

$$\mathbf{H}^T\mathbf{H}\boldsymbol{x} = \mathbf{H}^{\mathrm{t}}\boldsymbol{y}. \tag{II.3}$$

Si la matrice $\mathbf{H}^T\mathbf{H}$ n'est pas inversible, alors il n'y a pas unicité de l'équation (II.3) ; le problème reste mal posé. Dans le cadre d'un problème de déconvolution, on peut relier le caractère inversible de $\mathbf{H}^T\mathbf{H}$ à la différence de dimension entre l'observation $\boldsymbol{y} \in \mathbb{R}^M$ et l'inconnue $\boldsymbol{x} \in \mathbb{R}^N$. Lorsque $\mathbf{H}$ est une matrice de taille $M \times N$ Toeplitz, on peut distinguer les trois cas suivants :

- $M > N$, cas redondant : $\mathbf{H}^T\mathbf{H}$ est Toeplitz et peut être inversible,
- $M = N$, cas carré : $\mathbf{H}^T\mathbf{H}$ est inversible si $\mathbf{H}$ l'est,
- $M < N$, cas déficient : $\mathbf{H}\mathbf{H}^T$ est Toeplitz mais $\mathbf{H}^T\mathbf{H}$ n'est pas inversible.

Ainsi, pour un problème de déconvolution et au moins dans le cas déficient (moins de données que d'inconnues), la solution basée sur les moindres carrés ne satisfait jamais au caractère bien posé.

Pour circonvenir au problème de la non unicité de la solution des moindres carrés la *solution généralisée* $\boldsymbol{x}^\dagger$ est définie comme étant la solution de norme minimale de l'ensemble des solutions

du problème des moindres carrés :

$$\boldsymbol{x}^\dagger = \arg\min_{\boldsymbol{x}} \|\boldsymbol{x}\| \quad \text{s.c. } \boldsymbol{x} \in S_{\text{MC}}. \tag{II.4}$$

L'existence et l'unicité de $\boldsymbol{x}^\dagger$ étant assurées par le fait que l'ensemble S_{MC} est fermé et convexe [Rockafellar, 1970, p. 263]. L'*inverse généralisé* $\mathbf{H}^\dagger$ de $\mathbf{H}$ est défini par

$$\mathbf{H}^\dagger \boldsymbol{y} = \boldsymbol{x}^\dagger. \tag{II.5}$$

II.1.3 Manque de robustesse de l'inversion généralisée

La solution inverse généralisée $\boldsymbol{x}^\dagger$ satisfait au caractère bien posé, au sens de Hadamard. Cependant, en pratique, cette solution n'est pas satisfaisante en général. Elle peut souffrir d'un manque de robustesse. La robustesse étant définie de telle façon qu'une petite perturbation sur les observations n'entraîne qu'une petite erreur sur la solution.

Illustrons ce manque de robustesse. Soit $\delta\boldsymbol{y}$ une erreur sur l'observation $\boldsymbol{y}$ et soit $\delta\boldsymbol{x}^\dagger$ l'erreur induite sur la solution inverse généralisée $\boldsymbol{x}^\dagger$. En utilisant la linéarité de (II.5) ainsi que celle de $\boldsymbol{y} = \mathbf{H}\boldsymbol{x}$, on peut obtenir l'inégalité suivante :

$$\frac{\|\delta\boldsymbol{x}^\dagger\|}{\|\boldsymbol{x}^\dagger\|} \leq c\frac{\|\delta\boldsymbol{y}\|}{\|\boldsymbol{y}\|} \tag{II.6}$$

avec $c = \|\mathbf{H}\|\|\mathbf{H}^\dagger\| \geq 1$, où $\|\mathbf{A}\|$ est la norme matricielle définie par $\|\mathbf{A}\| = \sup_{\boldsymbol{u}\neq\mathbf{0}} \|\mathbf{A}\boldsymbol{u}\|/\|\boldsymbol{u}\|$. La quantité c est appelée *nombre de condition* du problème. Lorsque c est proche de l'unité, le problème est dit *bien conditionné*, sinon il est dit *mal conditionné*. Il est à noter que l'inégalité (II.6) peut devenir une égalité pour certains couples $(\boldsymbol{y}, \delta\boldsymbol{y})$.

Le nombre de condition est intimement lié à la stabilité de la solution inverse généralisée $\boldsymbol{x}^\dagger$. Plus il sera grand, et plus l'influence de l'erreur $\delta\boldsymbol{y}$ sur l'observation sera marquée sur la solution inverse généralisée $\boldsymbol{x}^\dagger$. Le bruit en sera alors d'autant plus amplifié sur la solution ainsi estimée.

Ceci nous amène au constat que les conditions de Hadamard sont insuffisantes pour obtenir une solution de qualité. En effet, si la condition de Hadamard de continuité est nécessaire pour assurer la robustesse, elle n'est pas suffisante. Comme exposé précédemment, un problème bien posé peut être mal conditionné, ce qui rend alors sa solution non robuste.

Pour finir, la solution inverse généralisée $\boldsymbol{x}^\dagger$ peut ne pas être suffisante même dans le cas d'un problème bien conditionné. Pour un problème de débruitage (*i.e.,* $\mathbf{H} = \mathbf{I}$), la solution inverse généralisée associée est trivialement à la fois bien posée et bien conditionnée. Dans ce cadre, l'estimation de l'inconnue se trouve être exactement l'observation, ce qui est évidement loin d'être satisfaisant.

II.2 Régularisation par approches pénalisées

Les problèmes inverses mal conditionnés sont intrinsèquement instables et leur inversion naïve conduit systématiquement à des difficultés. L'exemple précédent a montré que tenter de définir une inversion à partir de la seule connaissance de l'observation $\boldsymbol{y}$ est voué à l'échec. Le principe de *régularisation* [Nashed, 1981, p. 223], fournit un cadre préliminaire permettant de définir une solution stable.

II.2.1 Principe de la régularisation

Définition II.2.1. *Un régularisateur de l'équation* $\boldsymbol{y} = \mathbf{H}\boldsymbol{x}$ *est une famille d'opérateurs* $\{R_\lambda; \lambda \in \Lambda\}$ *telle que :*

$$\begin{cases} \forall \lambda \in \Lambda, \quad R_\lambda : \mathbb{R}^M \mapsto \mathbb{R}^N \text{ est continu} & \text{(II.7a)} \\ \forall \boldsymbol{y} \in \mathbb{R}^M, \quad \lim\limits_{\lambda \to 0} R_\lambda(\boldsymbol{y}) = \mathbf{H}^\dagger \boldsymbol{y}. & \text{(II.7b)} \end{cases}$$

Dans cette définition, $\mathbf{H}^\dagger$ est l'inverse généralisé de $\mathbf{H}$ défini précédemment en section II.1.2 et λ est le *paramètre de régularisation* à valeur dans Λ. Un régularisateur ainsi défini contient comme cas limite ($\lambda \to 0$) la solution inverse généralisée, qui est, rappelons le, insatisfaisante en général. L'intérêt d'un régularisateur est d'offrir à l'utilisateur des degrés de liberté pour construire une solution plus satisfaisante que la solution inverse généralisée.

Par la suite, nous expliciterons différents régularisateurs, tous basés sur la même structure générale. En somme, ce cadre abstrait définissant un régularisateur illustre en fait la méthodologie mise en œuvre en pratique pour déterminer une solution au problème inverse. Cette méthodologie peut être vue comme une procédure en deux temps. La structure analytique du régularisateur est d'abord choisie, sur la base de connaissances *a priori* sur la solution recherchée. Puis, le paramètre de régularisation λ est alors ajusté afin de fournir une solution jugée satisfaisante.

Pour les images, les propriétés *a priori* de la solution se caractérisent très souvent par une certaine *régularité locale*. Un moyen simple de prise en compte de cet *a priori* consiste à pénaliser les candidats qui s'écartent de cette notion de régularité tout en privilégiant les candidats respectant cette régularité souhaitée.

II.2.2 Régularisation au sens de Tikhonov

Les travaux précurseurs de Tikhonov [Tikhonov, 1963] initialement dans un cadre continu, consistent dans un cadre discret à définir la solution du problème inverse comme étant le minimiseur d'un critère des moindres carrés pénalisé par un terme quadratique

$$\mathcal{J}(\boldsymbol{x}) = \|\boldsymbol{y} - \mathbf{H}\boldsymbol{x}\|^2 + \lambda \mathcal{P}(\boldsymbol{x})$$

avec $\lambda > 0$. Dans le cadre de la définition II.2.1 d'un régularisateur, l'ensemble Λ se trouve être $\mathbb{R}^+$.

Dans ses premiers travaux, Tikhonov propose de pénaliser les trop fortes valeurs de la solution par le biais de l'utilisation de

$$\mathcal{P}(\boldsymbol{x}) = \|\boldsymbol{x}\|^2.$$

On peut noter que la solution ainsi définie est reliée à la solution généralisée (II.4). La solution définie par Tikhonov peut être interprétée comme la relaxation de la contrainte d'être exactement une solution des moindres carrés. Il s'agit d'un compromis entre être proche de la solution des moindres carrés et assurer à la solution d'avoir une norme minimale. Il est connu que cette pénalisation entraîne un effet de rappel vers zéro de la solution.

Plutôt que de pénaliser les fortes valeurs, un modèle d'image plus fin peut être envisagé sur une base très similaire. Les travaux ultérieurs de Tikhonov [Tikhonov et Arsénine, 1976, p.60] ont consisté à pénaliser les *irrégularités locales* par l'intermédiaire du terme de pénalisation défini par

$$\mathcal{P}(\boldsymbol{x}) = \|\mathbf{D}\boldsymbol{x}\|^2 \tag{II.8}$$

où $\mathbf{D}$ est une matrice de différenciation d'ordre d. Le choix le plus répandu consiste à utiliser la dérivée première ($d = 1$), ce qui a pour effet de favoriser l'apparition de *zones uniformes* au sein de l'image.

Lorsque la matrice $\mathbf{H}^{\mathrm{t}}\mathbf{H} + \lambda\mathbf{D}^{\mathrm{t}}\mathbf{D}$ est inversible, ce qui est en général le cas car la condition $\ker \mathbf{H} \cap \ker \mathbf{D} = \{\mathbf{0}\}$ est le plus souvent vérifiée, l'équation normale admet une unique solution définie par

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \mathcal{J}(\boldsymbol{x}) = (\mathbf{H}^{\mathrm{t}}\mathbf{H} + \lambda\mathbf{D}^{\mathrm{t}}\mathbf{D})^{-1}\mathbf{H}^{\mathrm{t}}\boldsymbol{y}.$$

Cette solution satisfait au caractère bien posé, au sens de Hadamard. De plus, le paramètre de régularisation est en général choisi de telle sorte que le conditionnement de la matrice $\mathbf{H}^{\mathrm{t}}\mathbf{H} + \lambda\mathbf{D}^{\mathrm{t}}\mathbf{D}$ soit meilleur (plus proche de l'unité) que celui de $\mathbf{H}^{\mathrm{t}}\mathbf{H}$. Ainsi, par rapport à la solution généralisée, la robustesse est généralement augmentée.

L'utilisation de critère pénalisé avec pénalisation quadratique s'avère limitée dans sa capacité à restituer les discontinuités des images. Un effet de lissage systématique des frontières de l'image est observé. Il en résulte un flou dont il est impossible de se débarrasser quel que soit le réglage du paramètre de régularisation λ. Cette perte de résolution introduite par la pénalisation quadratique est soulignée depuis maintenant près de deux décennies par la communauté du traitement des images ; voir par exemple [Geman et Reynolds, 1992, Sec. I.B]. Pour faire face à cette perte de résolution, des approches restant dans l'esprit de la méthode de Tikhonov mais s'appuyant sur des pénalisations non quadratiques préservant les discontinuités ont été proposées en restauration d'images.

II.2.3 Pénalisations non quadratiques

L'échec de la méthode de Tikhonov pour restaurer les discontinuités des images peut être expliqué simplement en considérant le cas où la matrice $\mathbf{D}$ est celle des différences du premier ordre. La pénalisation quadratique (II.8) pénalise "sans distinction" les différences inter-pixels et conduit ainsi à une solution "douce". L'approche par préservation des discontinuités consiste à "adapter" la pénalisation en fonction de l'amplitude des variations. En particulier, il s'agit de pénaliser moins fortement qu'une quadratique les fortes valeurs correspondant aux discontinuités afin que ces caractéristiques recherchées puissent se retrouver dans l'image restaurée.

Ainsi, à la place de la pénalisation quadratique (II.8) est substituée la pénalisation non quadratique suivante

$$\mathcal{P}(\boldsymbol{x}) = \sum_{c=1}^{C} \phi([\mathbf{D}\boldsymbol{x}]_c) \tag{II.9}$$

où $\phi : \mathbb{R} \mapsto \mathbb{R}$ est appelée *fonction de régularisation*. Notons que lorsque la fonction de régularisation est la parabole $\phi(t) = t^2$ alors la pénalisation (II.9) s'identifie à celle de Tikhonov (II.8). L'intérêt de l'approche par pénalisation non quadratique est illustré dans la figure II.1 page 34 pour un problème de restauration d'image en IRM [Labat *et al.*, 2005]. La figure II.1(b) qui est la version restaurée de la figure II.1(a) montre que cette approche permet bien de préserver les discontinuités de l'image.

De nombreux candidats pour le choix de la fonction de régularisation ϕ ont été proposés. Le principe de toutes ces fonctions de régularisation consiste à remplacer la pénalisation quadratique de Tikhonov par une fonction mieux adaptée à la préservation des discontinuités, c'est-à-dire permettant ponctuellement des variations importantes de l'estimée. Il est à noter que le choix de la fonction de régularisation et les arguments rattachés sont très similaires à l'utilisation des *normes robustes* en statistique [Black et Rangarajan, 1996].

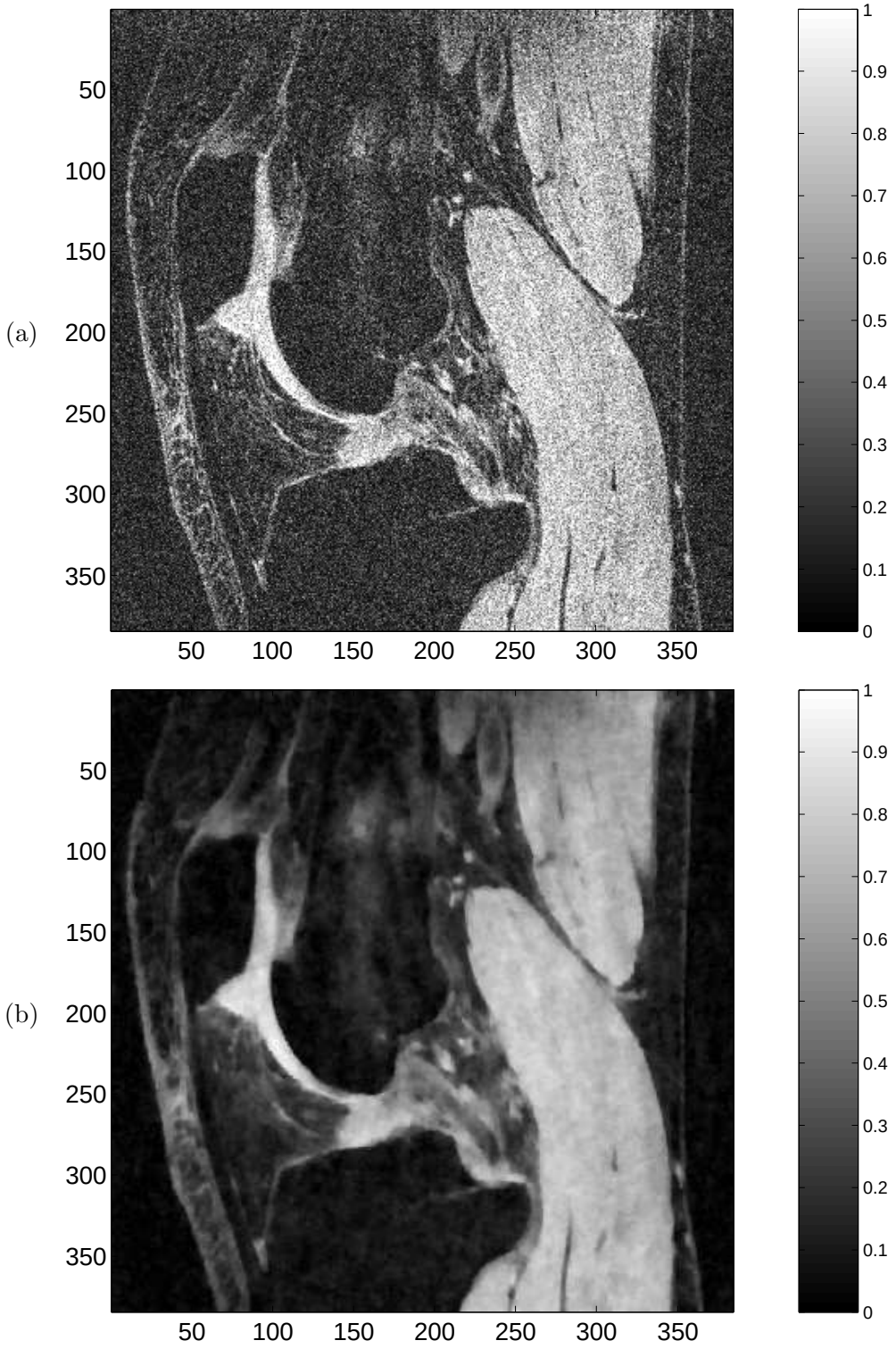


FIG. II.1: Restauration d'une image IRM de genou (GRBB, Ecole Polytechnique de Montréal). (a) Image originale, (b) Image restaurée par approche pénalisée préservant les discontinuités (fonction de régularisation hyperbolique (II.12) page 36 avec pour matrice $\mathbf{D}$ (II.9) celle des différences du premier ordre).

Plutôt que de répertorier toutes les fonctions de régularisation préservant les discontinuités que nous avons rencontrées dans la littérature, nous préférons établir ce qui nous a paru être le plus petit dénominateur commun. A notre connaissance, toutes les fonctions de régularisation ϕ préservant les discontinuités utilisées en restauration d'images présentent les propriétés suivantes :

$$\begin{cases} \text{positive : } \phi(t) \geq 0, \forall t \in \mathbb{R}, & \text{(II.10a)} \\ \text{paire : } \phi(t) = \phi(-t), & \text{(II.10b)} \\ \text{non décroissante sur } \mathbb{R}^+, & \text{(II.10c)} \\ \phi(t) = 0 \text{ ssi } t = 0 & \text{(II.10d)} \\ \text{croît moins vite qu'une parabole vers l'infini : } \lim\limits_{t \to \infty} \phi(t)/t^2 = 0, & \text{(II.10e)} \\ \text{dépend d'un seul paramètre de réglage } \delta. & \text{(II.10f)} \end{cases}$$

Le paramètre δ des fonctions de régularisation non quadratiques préservant les discontinuités, utilisé en restauration d'images, représente un "écart" par rapport au caractère quadratique. En général, plus δ est grand et plus la fonction de régularisation ainsi définie est proche d'une quadratique.

Précisons que nous distinguons ici les conditions de régularisation assurant la préservation des discontinuités de la mise en œuvre algorithmique. Les fonctions de régularisation seront supposées différentiables dans les prochains chapitres portant sur les aspects algorithmiques. Indiquons que les termes $\phi'(t)/t$ et $\phi''(t)$ peuvent être interprétés géométriquement comme des termes de lissages dans le cadre d'une pénalisation portant sur le gradient de l'image [Blanc-Féraud *et al.*, 1995]. Ainsi, le choix d'une fonction de régularisation quadratique induit un lissage isotrope tandis qu'une fonction non quadratique induit un lissage anisotrope permettant de restituer les discontinuités de l'image. Notons que si certaines fonctions de régularisation ϕ définies dans la littérature ne vérifient pas directement $\phi(0) = 0$, il est toujours possible d'effectuer un changement d'origine en les remplaçant par

$$\phi(t) := \phi(t) - \phi(0)$$

car la pénalisation résultante est inchangée à une constante près. D'ailleurs, pour des raisons de stabilité numérique, il est souvent préférable d'utiliser une fonction de régularisation s'annulant en zéro.

Il est à remarquer que les fonctions de régularisation non quadratiques dépendent d'un paramètre δ, contrairement à la fonction quadratique de Tikhonov. Ainsi, les fonctions de régularisation préservant les discontinuités répondent à l'exigence de préservation de discontinuités, mais au prix d'une certaine augmentation de complexité du réglage par rapport à l'approche par pénalisation quadratique. Dans le cadre de la définition II.2.1 d'un régularisateur, le couple de paramètres (λ, δ) de régularisation se trouve dans l'ensemble $\Lambda = \mathbb{R}^+ \times \mathbb{R}^+$ dans l'immense majorité des fonctions de régularisation, à l'exception notable de la famille ℓ_p présentée ci-dessous où $\Lambda = \mathbb{R}^+ \times [1, 2[$.

Si les fonctions de régularisation considérées dans le cadre de la restauration d'images sont paires, il est à noter que ce n'est pas systématiquement le cas en traitement du signal. Par exemple, il est fait usage en spectroscopie de fonctions de régularisation asymétriques, définies par une partie quadratique en $\mathbb{R}^-$ et avec une partie préservant les discontinuités en $\mathbb{R}^+$ [Mazet *et al.*, 2005].

Bien que toutes les fonctions de régularisation préservant les discontinuités utilisées en restauration d'images vérifient les propriétés (II.10a)-(II.10f), une démarcation importante peut être effectuée selon leur nature convexe ou non. Cette distinction n'a rien de formel puisque qu'elle conditionne la convexité du critère pénalisé $\mathcal{J}$ qui elle-même a un impact très sensible sur la solution ainsi estimée comme expliqué ci-après.

[A] Pénalisations convexes

Parmi les pénalisations convexes, on peut distinguer deux familles qui satisfont bien aux caractéristiques de préservation des discontinuités (II.10a)-(II.10f), si on effectue un changement à l'origine.

(i) Les fonctions de régularisation $\ell_2\ell_1$:

Elles sont convexes, C^1, C^2 en 0 et asymptotiquement linéaires. La désignation $\ell_2\ell_1$ fait explicitement référence au régime ℓ_2 (quadratique) près de 0 et au régime ℓ_1 (droite de pente non nulle) vers l'infini. Les fonctions de régularisation $\ell_2\ell_1$ les plus connues sont la *fonction de Huber*, C^1, introduite initialement en statistique robuste [Huber, 1981] :

$$\phi(t) = \begin{cases} t^2 & \text{pour } |t| < \delta, \\ 2\delta\,|t| - \delta^2 & \text{sinon} \end{cases} \tag{II.11}$$

représentée à la figure II.2(a) et la *fonction hyperbolique*, C^2, utilisée notamment par [Vogel et Oman, 1996; Charbonnier *et al.*, 1997] :

$$\phi(t) = \sqrt{t^2 + \delta^2} \tag{II.12}$$

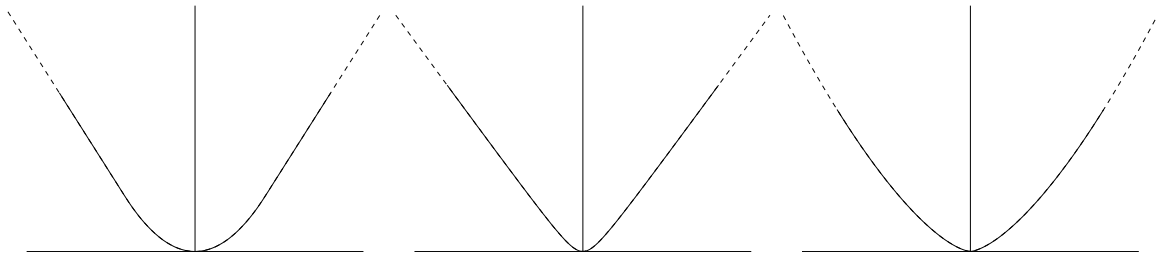
représentée à la figure II.2(b). Le paramètre δ est à valeur dans $\mathbb{R}^+$ pour ces deux fonctions.

(ii) Les fonctions de régularisation ℓ_p :

Elles sont utilisées notamment par [Bouman et Sauer, 1993] et sont définies par

$$\phi(t) = |t|^\delta$$

avec $\delta \in [1, 2)$. Elles sont C^1 pour $\delta > 1$ et non différentiables en zéro pour $\delta = 1$. Un représentant est donné à la figure II.2(c).



(a) (b) (c)

Fig. II.2: Fonctions : (a) de Huber : (II.11), (b) hyperbolique : $\sqrt{\delta^2 + t^2}$, (c) ℓ_p : $|t|^\delta$.

[B] Pénalisations non convexes

Les fonctions de régularisation $\ell_2\ell_0$ sont des représentantes classiques des fonctions de régularisation non convexes. Elles satisfont aux caractéristiques de préservation des discontinuités (II.10a)-(II.10f) et sont asymptotiquement constantes (caractère ℓ_0). La *quadratique tronquée* introduite par [Blake et Zisserman, 1987] est un représentant des fonctions de régularisation $\ell_2\ell_0$:

$$\phi(t) = \min\left\{t^2, \delta^2\right\} \tag{II.13}$$

avec $\delta > 0$. Cette fonction n'est pas différentiable en $t = \pm\delta$. Un représentant est donné à la figure II.3(a).

D'autres fonctions de régularisation $\ell_2\ell_0$ plus régulières sont introduites dans [Geman et McClure, 1987] comme

$$\phi(t) = \frac{t^2}{t^2 + \delta^2} \tag{II.14}$$

avec $\delta > 0$ et qui est cette fois C^2. Un représentant est donné à la figure II.3(b).

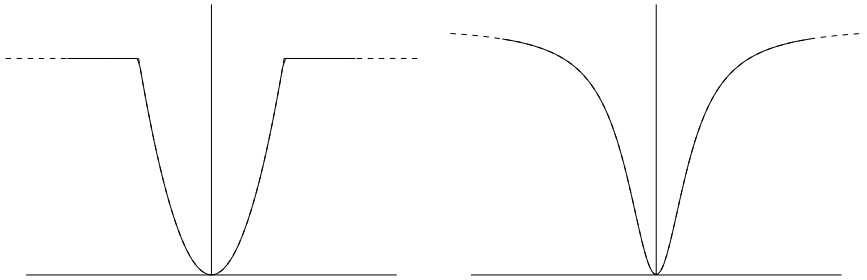


Fig. II.3: Fonctions : (a) quadratique tronquée : $\min\left\{t^2, \delta^2\right\}$, (b) : $\dfrac{t^2}{\delta^2 + t^2}$.

Les fonctions de régularisation $\ell_2\ell_0$ sont non convexes et non coercives. Notons qu'il existe des fonctions de régularisation non convexes mais coercives comme la fonction suivante proposée dans [Hebert et Leahy, 1989]

$$\phi(t) = \log\left(t^2 + \delta^2\right).$$

Cette fonction est présentée comme un compromis entre la quadratique et la fonction de Geman et McClure (II.14). Un représentant est donné à la figure II.4.

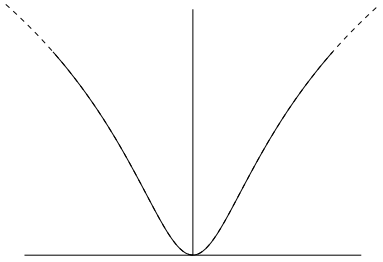


Fig. II.4: Fonction $\log\left(t^2 + \delta^2\right)$.

[C] Avantages de la pénalisation convexe

La convexité du critère pénalisé $\mathcal{J}$ est conditionnée par le caractère convexe de la fonction de régularisation non quadratique. Si un critère convexe a pour propriété d'être *unimodal*, un critère non convexe est en général *multimodal*.

L'emploi de fonctions de régularisation $\ell_2\ell_0$ conduit systématiquement à un critère $\mathcal{J}$ multimodal [Li, 1995, Sec. IV.B], [Blake, 1989, p. 3]. Ce caractère multimodal qu'entraîne l'usage de fonctions de régularisation non convexes a pour conséquence un manque de robustesse de la solution. En effet, le critère non convexe, de par ses multiples minima locaux présente une forme d'*instabilité* de la solution $\hat{\boldsymbol{x}}$ par rapport aux données et aux paramètres de réglage. Cette instabilité est soulignée notamment dans [Bouman et Sauer, 1993, Sec. I.B]. Indiquons que cette instabilité est liée à un aspect *décision* induit par les fonctions de régularisation non convexes [Idier et Blanc-Féraud, 2001, Sec. 6.3.2].

Ce manque de stabilité introduit par les fonctions de régularisation non convexes tient au fait que $\hat{\boldsymbol{x}}$ n'est pas une fonction continue des données ni des paramètres de réglage. En d'autres termes, la troisième condition de Hadamard n'est pas satisfaite par $\hat{\boldsymbol{x}}$, donc le problème est encore mal posé malgré la pénalisation. Par contre, l'usage de fonctions de régularisation convexes ne présente pas ces faiblesses et conduit à un problème bien posé.

Même s'il est possible de trouver un exemple en *simulation* où l'utilisation de fonctions de régularisation non convexes permet d'obtenir une solution plus proche de l'original que l'utilisation de fonctions de régularisation convexes, face à des *données réelles*, le réglage des hyperparamètres est rendu difficile par l'instabilité de la solution. Ainsi, il peut être préférable d'utiliser dans un cadre réel des fonctions de régularisation convexes pour des raisons de robustesse. Il est intéressant de noter que si initialement, les fonctions de régularisation non convexes introduites par [Geman et Geman, 1984] ont été largement utilisées avant les fonctions de régularisation convexes, un revirement est observé faisant du recours aux fonctions de régularisation convexes l'approche prépondérante, à la suite de travaux tels que [Bouman et Sauer, 1993].

En outre, la minimisation d'un critère multimodal nécessite la mise en œuvre d'algorithmes d'optimisation autrement plus complexes et donc d'un coût calculatoire bien plus élevé que la minimisation d'un critère unimodal.

II.2.4 Critère pénalisé généralisé

La pénalisation (II.9) ne permet pas de prendre en compte certains modèles. C'est le cas par exemple d'une pénalisation souhaitée sur le module d'une inconnue complexe. Afin de pouvoir tenir compte de ce type de situations, il est nécessaire de généraliser la pénalisation (II.9). C'est pourquoi nous considérons à présent le critère des moindres carrés

$$\mathcal{J}(\boldsymbol{x}) = \|\boldsymbol{y} - \mathbf{H}\boldsymbol{x}\|^2 + \lambda\Phi(\boldsymbol{x}) \tag{II.15}$$

avec la forme pénalisée généralisée suivante [Aubert et Kornprobst, 2002, Sec. 3.2], [Nikolova, 2002]

$$\Phi(\boldsymbol{x}) = \sum_{c=1}^{C} \phi(\|\mathbf{V}_c\boldsymbol{x} - \boldsymbol{\omega}_c\|) \tag{II.16}$$

où $\mathbf{V}_c \in \mathbb{R}^{P\times N}$, $\boldsymbol{\omega}_c \in \mathbb{R}^P$, pour $c = 1, \ldots, C$, $\|.\|$ est la norme euclidienne et $\phi : \mathbb{R}^+ \to \mathbb{R}$ est une fonction de régularisation préservant les discontinuités. Le vecteur $\boldsymbol{\omega}_c$ joue le rôle d'un terme de rappel. La pénalisation (II.16) est bien une forme généralisée de la pénalisation précédente puisqu'en considérant $\boldsymbol{\omega}_c = \mathbf{0}$ et $P = 1$ on retrouve bien (II.9).

Illustrons le fait que la pénalisation (II.16) permet de traiter naturellement le cas complexe en considérant $P = 2$. Rattacher le cas complexe au cas réel est un procédé d'usage courant afin de pouvoir exploiter directement les résultats existants exprimés le plus souvent pour le cas réel. Soit $\tilde{\boldsymbol{x}} \in \mathbb{C}^N$ un vecteur à complexe. Supposons qu'on souhaite pénaliser le module de $\tilde{\boldsymbol{x}}$: $|\tilde{\boldsymbol{x}}|$. En séparant $\tilde{\boldsymbol{x}}$ en partie réelle $\boldsymbol{r} \in \mathbb{R}^N$ et en partie imaginaire $\boldsymbol{s} \in \mathbb{R}^N$, on définit le vecteur réel $\boldsymbol{x} \in \mathbb{R}^{2N}$ en concaténant les parties réelle et imaginaire du vecteur complexe $\tilde{\boldsymbol{x}}$: $\boldsymbol{x} = [\boldsymbol{r}\ \boldsymbol{s}]^{\mathrm{t}}$. La pénalisation sur le module du vecteur complexe $\tilde{\boldsymbol{x}}$ peut alors s'écrire

$$\Phi(\tilde{\boldsymbol{x}}) = \sum_{n=1}^{N} \phi(|\tilde{x_n}|) = \sum_{n=1}^{N} \phi(|r_n + is_n|) = \sum_{n=1}^{N} \phi(\|\mathbf{V}_n \boldsymbol{x}\|)$$

où $\mathbf{V}_n$ est la matrice de taille $2 \times 2N$ définie par

$$\mathbf{V}_n = \begin{bmatrix} \mathbf{I}(n)\ \mathbf{0}_{1N} \\ \mathbf{0}_{1N}\ \mathbf{I}(n) \end{bmatrix}$$

où $\mathbf{I}(n)$ est la matrice de taille $1 \times N$ nulle à l'exception du $n^{\text{ème}}$ coefficient valant un. Ce traitement du cas complexe sera utilisé dans le chapitre VIII. Par la suite, lorsque nous évoquerons un critère pénalisé généralisé, nous ferons référence à la forme (II.15)-(II.16).

II.2.5 Interprétation intuitive du critère pénalisé

Différentes interprétations du critère pénalisé (II.15)-(II.16) peuvent être avancées. Il nous semble qu'une des interprétations les plus intuitives consiste à voir le critère pénalisé $\mathcal{J}$ préservant les discontinuités comme une somme de pseudo-distances.

On rappelle que les fonctions de régularisation préservant les discontinuités rencontrées dans la littérature satisfont en particulier les propriétés (II.10a)-(II.10d). En posant $d_c(\mathbf{u}, \mathbf{v}) = \lambda\phi(\|\mathbf{u} - \mathbf{v}\|)$ pour $c = 1, \ldots, C$ et $d_0(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|^2$ le critère pénalisé (II.15)-(II.16) peut être vu sous la forme

$$\mathcal{J}(\boldsymbol{x}) = \sum_{c=0}^{C} d_c(\mathbf{V}_c \boldsymbol{x}, \boldsymbol{\omega}_c)$$

où $\mathbf{V}_0 = \mathbf{H}$, $\boldsymbol{\omega}_0 = \boldsymbol{y}$ et $d_c(\boldsymbol{u}, \boldsymbol{v}) : \mathbb{R}^{P_c} \times \mathbb{R}^{P_c} \mapsto \mathbb{R}$ est une pseudo-distance, $P_c \in \mathbb{N}$. La pseudo-distance est ici entendue au sens que les axiomes de non négativité, de séparation et de symétrie sont vérifiés :

- $\forall \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^{P_c}, \quad d_c(\boldsymbol{u}, \boldsymbol{v}) \geq 0$
- $\forall \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^{P_c}, \quad d_c(\boldsymbol{u}, \mathbf{v}) = 0$ ssi $\boldsymbol{u} = \mathbf{v}$
- $\forall \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^{P_c}, \quad d_c(\boldsymbol{u}, \boldsymbol{v}) = d_c(\boldsymbol{v}, \boldsymbol{u})$

à l'exception de l'inégalité triangulaire qui permettrait d'obtenir pleinement une distance :

- $\forall \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^{P_c}, \quad d_c(\boldsymbol{u}, \boldsymbol{w}) \leq d_c(\boldsymbol{u}, \boldsymbol{v}) + d_c(\boldsymbol{v}, \boldsymbol{w})$.

Il est à noter que $\sqrt{d_0(\mathbf{u}, \mathbf{v})} = \|\boldsymbol{y} - \mathbf{H}\boldsymbol{x}\|$ est pleinement une distance (contrairement à $d_0(\mathbf{u}, \mathbf{v})$), mais l'inégalité triangulaire n'est en général pas satisfaite par $\sqrt{d_c(\mathbf{u}, \mathbf{v})}$ pour $c = 1, \ldots, C$.

Ainsi, la minimisation du critère pénalisé (II.15)-(II.16) peut être interprétée comme la minimisation des pseudo-distances entre les vecteurs $\mathbf{V}_c \boldsymbol{x}$ dépendant de la solution et des vecteurs fixes $\boldsymbol{\omega}_c$ se comportant ainsi comme un terme de rappel.

II.3 Conclusion

On a vu qu'une régularisation est nécessaire pour les problèmes mal posés, ce qui est la majorité des cas des problèmes inverses. Cette régularisation peut s'effectuer de manière séduisante par l'utilisation de critères pénalisés non quadratiques généralisant l'approche de Tikhonov. Le chapitre III fournit une interprétation probabiliste des critères pénalisés dans un cadre bayésien. L'approche pénalisée nécessite la mise en œuvre d'un algorithme de minimisation, c'est l'objet du chapitre IV.

Bibliographie

[Aubert et Kornprobst, 2002] G. Aubert et P. Kornprobst. *Mathematical problems in images processing*. Springer-Verlag, Berlin, 2002.

[Black et Rangarajan, 1996] M. J. Black et A. Rangarajan. On the unification of line processes, outlier rejection, and robust statistics with applications in early vision. *Int. J. Computer Vision*, 19 (1) : 57–91, 1996.

[Blake, 1989] A. Blake. Comparison of the efficiency of deterministic and stochastic algorithms for visual reconstruction. *IEEE Trans. Pattern Anal. Mach. Intell.*, PAMI-11 (1) : 2–12, janvier 1989.

[Blake et Zisserman, 1987] A. Blake et A. Zisserman. *Visual reconstruction*. The MIT Press, Cambridge, MA, USA, 1987.

[Blanc-Féraud *et al.*, 1995] L. Blanc-Féraud, P. Charbonnier, G. Aubert et M. Barlaud. Nonlinear image processing : Modelling and fast algorithm for regularisation with edge detection. In *Proc. IEEE ICIP*, volume 2, pages 474–477, 1995.

[Bouman et Sauer, 1993] C. A. Bouman et K. D. Sauer. A generalized Gaussian image model for edge-preserving MAP estimation. *IEEE Trans. Image Processing*, 2 (3) : 296–310, juillet 1993.

[Charbonnier *et al.*, 1997] P. Charbonnier, L. Blanc-Féraud, G. Aubert et M. Barlaud. Deterministic edge-preserving regularization in computed imaging. *IEEE Trans. Image Processing*, 6 (2) : 298–311, février 1997.

[Geman et Reynolds, 1992] D. Geman et G. Reynolds. Constrained restoration and the recovery of discontinuities. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14 (3) : 367–383, mars 1992.

[Geman et Geman, 1984] S. Geman et D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, PAMI-6 (6) : 721–741, novembre 1984.

[Geman et McClure, 1987] S. Geman et D. McClure. Statistical methods for tomographic image reconstruction. In *Proceedings of the 46th Session of the ICI, Bulletin of the ICI*, volume 52, pages 5–21, 1987.

[Hadamard, 1901] J. Hadamard. Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton Univ. Bull.*, 13, 1901.

[Hebert et Leahy, 1989] T. Hebert et R. Leahy. A generalized EM algorithm for 3-D Bayesian reconstruction from Poisson data using Gibbs priors. *IEEE Trans. Medical Imaging*, 8 (2) : 194–202, juin 1989.

[Huber, 1981] P. J. Huber. *Robust Statistics*. John Wiley, New York, NY, USA, 1981.

[Idier et Blanc-Féraud, 2001] J. Idier et L. Blanc-Féraud. *Déconvolution en imagerie*, chapitre 6, pages 139–165. Traité IC2, Série traitement du signal et de l'image, Hermès, Paris, novembre 2001.

[Labat *et al.*, 2005] C. Labat, J. Idier et Y. Goussard. Comparison between half-quadratic and preconditioned conjugate gradient algorithms for mri reconstruction. In *PSIP'2005 : Physics in signal and Image processing*, Toulouse, janvier 2005.

[Li, 1995] S. Z. Li. On discontinuity-adaptive smoothness priors in computer vision. *IEEE Trans. Pattern Anal. Mach. Intell.*, PAMI-17 (6) : 576–586, juin 1995.

[Mazet *et al.*, 2005] V. Mazet, C. Carteret, D. Brie, J. Idier et B. Humbert. Background removal from spectra by designing and minimising a non-quadratic cost function. *Chemometrics and Intelligent Laboratory Systems*, 76 : 121–133, 2005.

[Nashed, 1981] M. Z. Nashed. Operator-theoretic and computational approaches to ill-posed problems with applications to antenna theory. *IEEE Trans. Ant. Propag.*, 29 : 220–231, 1981.

[Nikolova, 2002] M. Nikolova. Minimizers of cost-functions involving non-smooth data-fidelity terms. Application to the processing of outliers. *SIAM J. Num. Anal.*, 40 (3) : 965–994, 2002.

[Rockafellar, 1970] R. T. Rockafellar. *Convex Analysis*. Princeton Univ. Press, 1970.

[Tikhonov, 1963] A. Tikhonov. Regularization of incorrectly posed problems. *Soviet. Math. Dokl.*, 4 : 1624–1627, 1963.

[Tikhonov et Arsénine, 1976] A. Tikhonov et V. Arsénine. *Méthodes de résolution de problèmes mal posés*. Éditions MIR, Moscou, Russie, 1976.

[Vogel et Oman, 1996] R. V. Vogel et M. E. Oman. Iterative methods for total variation denoising. *SIAM J. Sci. Comput.*, 17 (1) : 227–238, janvier 1996.

Chapitre III

INTERPRÉTATION PROBABILISTE ET INFÉRENCE BAYÉSIENNE

L'objet de ce chapitre est de montrer que l'inversion pénalisée présentée dans le chapitre précédent peut être interprétée dans un cadre probabiliste par l'intermédiaire de l'inférence bayésienne. Si ce cadre probabiliste permet de donner un éclairage différent de l'approche pénalisée déterministe, l'intérêt majeur réside principalement dans l'introduction de certains outils propres au cadre probabiliste. Ces outils sont illustrés dans l'annexe E, page 175.

III.1 Vraisemblance et adéquation aux données

Rappelons le modèle d'observation linéaire présenté au début du chapitre précédent

$$\boldsymbol{y} = \mathbf{H}\boldsymbol{x} + \boldsymbol{\epsilon}. \tag{III.1}$$

Jusqu'à présent, rien n'avait été précisé concernant la composante additive de bruit. On suppose maintenant que $\boldsymbol{\epsilon}$ représente une *réalisation d'un vecteur aléatoire* $\boldsymbol{\varepsilon}$ qui admet une *loi à densité* $f_{\boldsymbol{\varepsilon}}(\boldsymbol{\epsilon})$. L'image $\boldsymbol{x}$ inconnue étant considérée déterministe, on en déduit que le vecteur des observées $\boldsymbol{y}$ constitue une réalisation d'un vecteur aléatoire Y dont la densité de probabilité se déduit de celle du bruit

$$f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{x}) = f_{\boldsymbol{\varepsilon}}(\boldsymbol{y} - \mathbf{H}\boldsymbol{x})$$

Une situation courante consiste à considérer que la loi du bruit $\boldsymbol{\varepsilon}$ est une densité gaussienne centrée

$$f_{\boldsymbol{\varepsilon}}(\boldsymbol{\epsilon}) = \frac{1}{Z}\exp\left\{-\frac{1}{2}\boldsymbol{\epsilon}^{\mathrm{t}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\epsilon}\right\}, \quad \forall \boldsymbol{\epsilon} \in \mathbb{R}^M$$

où Z et $\mathbf{\Sigma} \in \mathbb{R}^{M\times M}$ sont respectivement, la constante de normalisation et la matrice de covariance symétrique définie positive. Dans ce cas, le vecteur aléatoire des observations $\boldsymbol{Y}$ admet une loi à densité gaussienne de même covariance que $\boldsymbol{\varepsilon}$ et de moyenne $\mathbf{H}\boldsymbol{x}$:

$$f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{x}) = \frac{1}{Z}\exp\left\{-\frac{1}{2}(\boldsymbol{y}-\mathbf{H}\boldsymbol{x})^{\mathrm{t}}\mathbf{\Sigma}^{-1}(\boldsymbol{y}-\mathbf{H}\boldsymbol{x})\right\} \tag{III.2}$$

De manière générale, on identifie *la fonction de vraisemblance des données* $\boldsymbol{y}$ notée $\mathcal{V}(\boldsymbol{y};\boldsymbol{x})$ à la densité de probabilité des observations paramétrées par l'image, *i.e.,*

$$\mathcal{V}(\boldsymbol{y};\boldsymbol{x}) = f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{x})$$

Cette fonction de vraisemblance joue un rôle central dans la construction d'estimateurs. En effet, elle résume à elle seule toute l'information contenue dans les données sur l'objet déterministe $\boldsymbol{x}$ inconnu.

III.1.1 Estimation au sens du maximum de vraisemblance

Une fois obtenu le modèle probabiliste générateur des données, il reste à déterminer l'estimateur de l'image. En statistique classique, l'estimateur du maximum de vraisemblance (MV)

$$\boldsymbol{x}^{MV} = \arg\max_{\boldsymbol{x}} \mathcal{V}(\boldsymbol{y};\boldsymbol{x})$$

est largement utilisé en particulier pour ses bonnes propriétés asymptotiques – biais nul et variance minimale ; voir par exemple [Fourgeaud et Fuchs, 1972, Chap. 14]. L'estimateur du MV fait alors intervenir un problème d'optimisation qu'il est courant d'aborder via une transformation monotone logarithmique

$$\boldsymbol{x}^{MV} = \arg\min_{\boldsymbol{x}} -\log \mathcal{V}(\boldsymbol{y};\boldsymbol{x})$$

cette approche simplifiant généralement les expressions puisque la fonction exponentielle intervient couramment dans l'expression des densités de probabilité.

[A] Fonction d'adéquation aux données

De manière générale, on peut associer une vraisemblance à une *fonction d'adéquation aux données* $\mathcal{Q}(\boldsymbol{y};\boldsymbol{x})$ pour peu que la loi de probabilité ci-dessous ait effectivement un sens

$$f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{x}) = \frac{1}{Z_1}\exp\left\{-\mathcal{Q}(\boldsymbol{y};\boldsymbol{x})/T_1\right\}$$

avec Z_1 la constante de normalisation et T_1 un paramètre de "température". Réciproquement, on associera une adéquation aux données $\mathcal{Q}(\boldsymbol{y};\boldsymbol{x})$ à toute vraisemblance. L'existence d'une fonction d'adéquation aux données est généralement vérifiée en pratique ; c'est en particulier le cas de la vraisemblance gaussienne. Ainsi, les solutions du maximum de vraisemblance correspondent aux solutions *non régularisées* minimisant le terme d'adéquation aux données tel que

$$-\log \mathcal{V}(\boldsymbol{y};\boldsymbol{x}) = \mathcal{Q}(\boldsymbol{y};\boldsymbol{x})/T_1.$$

[B] Hypothèses gaussiennes et moindres carrés pondérés

Dans le cas d'un problème d'inversion (III.1) et sous l'hypothèse d'un modèle gaussien du bruit, l'opposé du logarithme de la vraisemblance définie par (III.2) s'écrit

$$-\log \mathcal{V}(\boldsymbol{y};\boldsymbol{x}) = \frac{1}{2}(\boldsymbol{y}-\mathbf{H}\boldsymbol{x})^{\mathrm{t}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\mathbf{H}\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{y}-\mathbf{H}\boldsymbol{x})\|^2$$

L'ensemble des solutions du MV s'écrit alors

$$S_{\mathrm{MV}} = \{x \in \mathbb{R}^N, \min \|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{y}-\mathbf{H}\boldsymbol{x})\|\}. \tag{III.3}$$

Ces solutions correspondent aux solutions des moindres carrés pondérés. Ces solutions sont à rapprocher des solutions des moindres carrés, qui correspondent au cas particulier $\boldsymbol{\Sigma} = \mathbf{I}$, c'est-à-dire où les composantes du bruit sont indépendantes. L'ensemble des solutions des moindres carrés pondérés n'est en général pas unique. Il s'agit donc d'un problème mal posé.

[C] Conclusion

Pour un problème inverse numériquement instable comme la déconvolution, l'estimateur du MV n'apporte donc pas de solution pertinente. L'inférence bayésienne, présentée ci-dessous, fournit des alternatives intéressantes.

III.2 Inférence bayésienne

L'inférence bayésienne se distingue de l'inférence classique par l'apport d'une connaissance *a priori* sur la grandeur à estimer. Dans un cadre bayésien, cette information sur l'objet prend la forme d'une loi à densité donnée *a priori* $f_{\boldsymbol{X}}(\boldsymbol{x})$: l'image inconnue $\boldsymbol{x}$ est alors considérée comme une réalisation d'un vecteur aléatoire $\boldsymbol{X}$.

III.2.1 Vraisemblance *a posteriori*

Introduisons la notation $f_{\boldsymbol{A}|\boldsymbol{B}}(\boldsymbol{a}|\boldsymbol{b})$ pour désigner la loi de probabilité à densité du vecteur aléatoire $\boldsymbol{A}$ conditionnellement à l'événement $\boldsymbol{B} = \boldsymbol{b}$. Le cadre bayésien conduit à modifier la notation de la densité des observations étant donné l'image ; celle-ci sera notée désormais

$$f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}) = f_{\boldsymbol{\varepsilon}}(\boldsymbol{y}-\mathbf{H}\boldsymbol{x})$$

cette densité correspondant toujours à la vraisemblance des données notée $\mathcal{V}(\boldsymbol{y};\boldsymbol{x})$. La règle de Bayes fournit alors le lien existant entre les lois *a priori* et *a posteriori* :

$$f_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y}) = \mathcal{V}(\boldsymbol{y};\boldsymbol{x})\frac{f_{\boldsymbol{X}}(\boldsymbol{x})}{f_{\boldsymbol{Y}}(\boldsymbol{y})}$$

Le terme au numérateur $f_{\boldsymbol{Y}}(\boldsymbol{y})$ est la densité marginale

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \int f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x})d\boldsymbol{x}$$

qui assure la normalisation de la loi *a posteriori* ; la loi *a posteriori* s'écrit donc à un facteur multiplicatif près :

$$f_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y}) \propto f_{\boldsymbol{X}}(\boldsymbol{x})\mathcal{V}(\boldsymbol{y};\boldsymbol{x}).$$

Le membre de droite de cette relation est appelé fonction de vraisemblance *a posteriori* et est noté

$$\mathcal{V}_P(\boldsymbol{y};\boldsymbol{x}) = f_{\boldsymbol{X}}(\boldsymbol{x})\mathcal{V}(\boldsymbol{y};\boldsymbol{x})$$

La vraisemblance *a posteriori* résume toute l'information disponible sur l'image, à savoir à la fois l'adéquation aux données induite par la vraisemblance $\mathcal{V}(\boldsymbol{y};\boldsymbol{x})$ et l'information *a priori* sur l'image contenue dans $f_{\boldsymbol{X}}(\boldsymbol{x})$. A partir de la vraisemblance *a posteriori* un certain nombre d'estimateurs aux caractéristiques diverses peuvent être définis. L'estimateur du maximum *a posteriori* (MAP), développé plus bas, est certainement le plus connu. Néanmoins, d'autres estimateurs comme la moyenne *a posteriori*, l'estimateur de la médiane et le MAP marginal sont parfois utilisés.

III.2.2 Maximum *a posteriori* et approches pénalisées

En pratique, on se tourne souvent vers l'estimateur du *maximum a posteriori* pour résoudre un problème d'inférence dans le cadre bayésien. De manière équivalente, on est alors amené à considérer l'ensemble des minimiseurs de l'inverse de la log vraisemblance *a posteriori*

$$S_{\mathrm{MAP}} = \{x \in \mathcal{X}, \min \mathcal{J}_{\mathrm{MAP}}(\boldsymbol{x})\} \tag{III.4}$$

où

$$\begin{aligned}\mathcal{J}_{\mathrm{MAP}}(\boldsymbol{x}) &= -\log \mathcal{V}_P(\boldsymbol{y};\boldsymbol{x}) \\ &= -\log \mathcal{V}(\boldsymbol{y};\boldsymbol{x}) - \log f_{\boldsymbol{X}}(\boldsymbol{x}).\end{aligned}$$

Dans le cas du modèle (III.1) où la loi de densité du bruit est une gaussienne avec matrice de covariance unitaire, le critère $\mathcal{J}_{\mathrm{MAP}}$ est de type moindres carrés pénalisés

$$\mathcal{J}_{\mathrm{MAP}}(\boldsymbol{x}) = \frac{1}{2}(\boldsymbol{y} - \mathbf{H}\boldsymbol{x})^2 - \log f_{\boldsymbol{X}}(\boldsymbol{x}).$$

Comme pour l'estimation au sens du MV, un lien fort peut généralement être établi entre l'estimation du MAP et la régularisation par pénalisation dans un cadre déterministe : sous réserve que l'on puisse établir des liens tels que

$$\begin{aligned}-\log \mathcal{V}(\boldsymbol{y};\boldsymbol{x}) \quad &\longleftrightarrow \quad \mathcal{Q}(\boldsymbol{y};\boldsymbol{x}) \\ -\log f_{\boldsymbol{X}}(\boldsymbol{x}) \quad &\longleftrightarrow \quad \lambda\mathcal{P}(\boldsymbol{x}), \quad \lambda > 0\end{aligned}$$

alors l'ensemble des solutions S_{MAP} correspond à l'ensemble des solutions produites par la régularisation de Tikhonov généralisée.

Plus précisément, l'implication correspondant au passage de l'estimation du MAP à une régularisation par pénalisation peut toujours se faire. Par contre, la réciproque n'est pas toujours vraie. En effet, si le terme d'adéquation Q découle généralement d'une loi de vraisemblance normalisable, un certain nombre de cas pratiques conduisent à des lois *a priori* qui sont non normalisables. C'est par exemple ce qui se produit pour le choix très répandu d'un *a pirori* pénalisant la différence entre paire de pixels

$$\mathcal{P}(\boldsymbol{x}) = \sum_{r\sim s} \phi(x_r - x_s)$$

où $r \sim s$ décrit un ensemble de couples de sites "spatialement voisins". Lorsque $\boldsymbol{x} \in \mathbb{R}$, la loi $\exp\{-\mathcal{P}(\boldsymbol{x})/T\}$ est non normalisable [Besag *et al.*, 1995, Sec. 3],[Idier, 2001, Sec. 7.3.1].

III.3 Adéquation robuste aux données

Jusqu'ici nous avons considéré une approche pénalisée où le terme d'adéquation aux données est quadratique

$$\mathcal{Q}(\boldsymbol{y};\boldsymbol{x}) = \|\boldsymbol{y} - \mathbf{H}\boldsymbol{x}\|^2.$$

Comme nous l'avons vu, ceci correspond à une interprétation probabiliste où la loi de densité du bruit est une gaussienne. Bien que ce modèle soit extrêmement courant, dans certaines situations ce modèle simple n'est pas totalement vérifié. Il peut alors être avantageux de changer de modèle et donc par conséquent la nature de l'adéquation aux données.

C'est en particulier le cas quand une loi de densité du bruit poissonnienne se révèle être plus adaptée qu'une gaussienne, en tomographie par exemple [Koulibaly *et al.*, 1996]. C'est également le cas quand un certain nombre de données aberrantes ne suivent pas le modèle probabiliste initial. Ce type de problème est précisément étudié en *statistique robuste* [Huber, 1981]. Cette approche propose l'utilisation d'une adéquation robuste aux données aberrantes sous la forme suivante

$$\mathcal{Q}(\boldsymbol{y};\boldsymbol{x}) = \sum_{i=1}^{M} \phi\left(\boldsymbol{y}_i - \boldsymbol{h}_i\boldsymbol{x}\right)$$

où $\{\boldsymbol{h}_i\}$ sont les lignes de la matrice d'observation $\mathbf{H}$. La fonction ϕ est une fonction *robuste*. Le lecteur notera que nous utilisons la même notation ϕ que pour la fonction de régularisation préservant les discontinuités introduite dans le chapitre II. Ce choix de notation n'est pas innocent car il indique une certaine proximité entre ces deux usages. Si l'objectif pratique n'est pas le même, le fonctionnement sous-jacent est bien le même. Il s'agit de ne pas trop pénaliser les fortes valeurs ; qu'elles correspondent à des données aberrantes, ou bien à de fortes différences entre pixels indiquant des discontinuités à préserver. Il est donc tout à fait possible d'envisager d'utiliser la même fonction ϕ à la fois pour l'adéquation aux données et pour la préservation des discontinuités en traitement de l'image.

Il y a finalement assez peu de contributions en traitement de l'image ou du signal qui font appel à une adéquation robuste des données. On peut citer [Mazet *et al.*, 2005] qui dans le cadre de la spectroscopie utilisent un critère composé uniquement d'un terme d'adéquation aux données, sans terme de pénalisation car le problème est bien posé. En image, [Nikolova, 2002] considère une adéquation robuste aux données avec un terme de pénalisation préservant les discontinuités.

Notons que le terme d'adéquation robuste aux données peut être mathématiquement incorporé dans le terme de pénalisation en annulant la partie d'adéquation quadratique. Ainsi, tous les résultats que nous obtiendrons dans les chapitres suivants sur le critère pénalisé de la forme (II.15)-(II.16) pourront être utilisés pour un critère avec une adéquation robuste aux données.

III.4 Modèles *a priori* sur les images

L'objectif du choix de la densité $f_{\boldsymbol{X}}(\boldsymbol{x})$ est d'employer une classe de modèles pertinente vis-à-vis de l'application et de charge calculatoire acceptable. Nous présentons ci-dessous deux classes de modèles qui satisfont souvent en pratique à ces deux impératifs souvent contradictoires.

III.4.1 Champ de Gibbs-Markov

[A] Champ de Markov

Un champ de Markov (MRF) est un champ aléatoire dont les propriétés sont régies par des interactions locales. Plus précisément, la probabilité conditionnelle d'un point connaissant tous les autres points ne dépend que des valeurs des points *voisins*. En supposant que le support de l'objet est un ensemble de sites $\mathcal{S} = \{1, \ldots, S\}$, on peut donner la définition suivante [Brémaud, 1999, Sec. 7.1].

Définition III.4.1 (Champ aléatoire de Markov)**.** *On appelle "champ aléatoire de Markov" associé à $\mathcal{S}$ et à un système de voisinage η tout champ $\mathcal{X}$ de support $\mathcal{S}$ tel que les densités de probabilité conditionnelles de ses éléments $\boldsymbol{X}_i$ de coordonnées i vérifient la relation suivante*

$$f(x_i|x_j, j \in \Omega) = f(x_i|x_j, j \in \eta_i)$$

pour tout sous-ensemble Ω de $\mathcal{S}$ contenant le voisinage η_i de i et ne contenant pas i.

Notons que les MRF sont particulièrement utiles si le nombre de voisins est restreint car, dans ce cas, la description locale des interactions permet d'appliquer des méthodes de résolution à charge calculatoire réduite.

[B] Théorème de Hammersley-Clifford et champ de Gibbs

L'intérêt d'une formulation à base de MRF serait finalement très réduite si on ne pouvait écrire la probabilité *a priori* $f_{\boldsymbol{X}}(\boldsymbol{x})$ de l'objet sous une forme explicite et simple. Ceci est possible en faisant intervenir les potentiels de Gibbs : on introduit l'ensemble $\mathcal{C}$ constitué de C sous-ensembles de $\mathcal{S}$, chaque élément de C étant appelé une *clique* ; on donne alors la définition suivante [Brémaud, 1998, Sec. 7.2].

Définition III.4.2 (Champ aléatoire de Gibbs)**.** *Sur un support fini $\mathcal{S}$, on appelle "champ aléatoire de Gibbs" (GRF) associé à l'ensemble de clique $\mathcal{C}$, un champ $\mathcal{X}$ dont la densité de probabilité est de la forme*

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{Z_2} \exp -\mathcal{U}(\boldsymbol{x})/T_2 \tag{III.5}$$

où, Z_2, T_2 sont, respectivement, les constantes de normalisation et de température. $\mathcal{U}(\boldsymbol{x})$ est appelée la fonction d'énergie et s'écrit

$$\mathcal{U}(\boldsymbol{x}) = \sum_{c \in \mathcal{C}} \mathbf{U}_c(\boldsymbol{x})$$

avec $\mathbf{U}_c$ une fonction potentiel associée à la clique c.

Une formulation par champ de Gibbs présente le gros avantage de donner directement accès à la densité *a priori*, $f_{\boldsymbol{X}}(\boldsymbol{x})$, par l'intermédiaire de (III.5). Sous des hypothèses réalistes en traitement d'image, le théorème de Hammersley-Clifford [Winkler, 1995, Th. 3.3] établit néanmoins l'équivalence entre champ de Markov et champ de Gibbs. En pratique, on définit alors souvent un MRF par sa densité de Gibbs équivalente, l'ensemble des cliques C découlant directement du système de voisinage considéré par le MRF.

[C] Conclusion

L'intérêt des MRF réside dans leur simplicité. Néanmoins, certains résultats liés à l'estimation des hyperparamètres laissent entendre que les MRF ne modélisent pas très fidèlement les images issues du monde réel [Descombes *et al.*, 1999]. En particulier, ils ne sont pas toujours capables de modéliser correctement les images contenant des textures complexes. Par contre, une approche par ondelettes semble assez bien répondre à cette exigence d'un modèle plus riche sans pour autant en accroître considérablement la complexité [Moulin et Liu, 1999; Belge *et al.*, 2000].

III.4.2 Ondelettes

Les modèles *a priori* à base de MRF sont utilisés dans le domaine des pixels de l'image. Il est possible d'utiliser un domaine différent tel que celui des coefficients d'ondelettes de l'image. L'*a priori* considéré est le même pour ces deux approches, à savoir que l'image est supposée être constituée de zones homogènes séparées par quelques discontinuités. La différence entre ces deux modèles réside dans la façon de tenir compte de cet *a priori* sur les images. Les approches par MRF appliquent cet *a priori* dans le domaine des pixels de l'image alors que les approches par ondelettes l'appliquent dans le domaine des coefficients d'ondelettes de l'image.

Selon certains auteurs [Belge *et al.*, 2000; Jalobeanu *et al.*, 2004], l'intérêt de travailler dans le domaine des coefficients d'ondelettes de l'image plutôt que dans le domaine des pixels de l'image résiderait dans la qualité accrue de l'image restaurée. En particulier, l'approche par ondelettes permettrait une meilleure restauration des textures dans le cas d'images complexes.

L'approche par ondelettes d'un problème de restauration d'image (déconvolution) modélisé par (III.1) peut se faire en se donnant un modèle probabiliste de la répartition des coefficients d'ondelettes de l'image [Belge *et al.*, 2000; Jalobeanu *et al.*, 2004]. Soit $\boldsymbol{x}$ l'image à restaurer de taille $m \times n$ avec $1 \leq m, n \leq 2^J$. Les coefficients de la transformée d'ondelettes 2D sont désignés par $\boldsymbol{x}_j$ de taille $2^j \times 2^j$ avec $j \in \{0, \ldots, J-1\}$ qui est le niveau de décomposition d'ondelette.

Une modélisation classique consite à considérer l'image comme une réalisation aléatoire dont les coefficients de la transformée d'ondelettes $x_j(m,n)$ sont distribués selon une loi gaussienne généralisée (GG) de la forme [Mallat, 1989; Antonini *et al.*, 1992; Moulin et Liu, 1999; Belge *et al.*, 2000]

$$f(x_j(m,n)|p,k_j) \propto \exp\left\{-\frac{1}{p}\left|\frac{x_j(m,n)}{k_j}\right|^p\right\}$$

où $0 < p \leq 2$ est un paramètre déterminant la forme de la distribution et k_j est un paramètre d'échelle. Cette modélisation est basée sur le fait que la GG, ayant une queue plus lourde que la gaussienne classique, décrit mieux la distribution des coefficients de la transformée d'ondelette. En effet, les coefficients d'ondelettes sont principalement distribués vers zéro, pour ce qui est de la contribution des zones homogènes, et ont une queue lourde pour ce qui est de la contribution des discontinuités.

En considérant classiquement un modèle de bruit blanc gaussien, [Belge *et al.*, 2000] propose d'utiliser l'estimateur du MAP qui peut s'écrire comme le minimiseur du critère des moindres carrés pénalisés suivant

$$\mathcal{J}(\boldsymbol{x}) = \|\boldsymbol{y} - \mathbf{H}\boldsymbol{x}\|_2^2 + \sum_{j=1}^{M} \lambda_j \|\mathbf{R}_j \boldsymbol{x}\|_p^p \tag{III.6}$$

où $\|.\|_p$ est la norme ℓ_p, λ_j, $j = 1, \ldots, M$ sont les paramètres de régularisation et $\mathbf{R}_j$ sont les matrices de régularisation associées. Ici, $\boldsymbol{x}$, $\boldsymbol{y}$ et $\mathbf{H}$ sont dans le domaine d'ondelettes et $\mathbf{R}_j$ sont

des opérateurs qui extraient les parties désirées de la transformée d'ondelettes de l'image. Dans [Belge *et al.*, 2000] plusieurs modèles sont considérés donnant lieu à différentes matrices $\mathbf{R}_j$. Dans [Wang *et al.*, 1995, Sec. III.C], les opérateurs $\mathbf{R}_j$ de (III.6) consistent en l'extraction des coefficients de la transformée d'ondelettes de l'image de hautes fréquences. Dans [Antoniadis et Fan, 2001], les matrices identité sont considérées pour les opérateurs $\mathbf{R}_j$ de (III.6).

Le critère (III.6) ainsi obtenu est structurellement un cas particulier du critère pénalisé généralisé (II.15)-(II.16). On voit donc qu'une approche par ondelettes peut être aussi vue comme une approche pénalisée. Nous n'avons pas utilisé en pratique ces approches par ondelettes car la qualité des images restaurées en travaillant dans le domaine des pixels des images nous semble satisfaisante pour les problèmes que nous avons considérés, tel que le contrôle non destructif par ultrasons traité dans le chapitre VIII. Par contre, cette approche par ondelettes vue comme critère pénalisé avec un choix adapté des matrices de régularisation, peut être intéressante pour des images de complexité plus importante. Notre objectif en mettant l'accent sur ce lien est d'affirmer que les résultats de convergence pour la minimisation des critères pénalisés présentés dans les chapitres suivants peuvent être directement appliqués pour un certain type d'approche par ondelettes. D'ailleurs, pour résoudre le problème de minimisation du critère (III.6) dans le domaine en ondelettes il est proposé dans [Belge *et al.*, 2000] de recourir à un algorithme développé pour les approches pénalisées préservant les discontinuités [Charbonnier *et al.*, 1997].

Bibliographie

[Antoniadis et Fan, 2001] A. Antoniadis et J. Fan. Regularization of wavelet approximations. *J. Amer. Statist. Assoc.*, 96 (455) : 939–967, 2001.

[Antonini *et al.*, 1992] M. Antonini, M. Barlaud, P. Mathieu et I. Daubechies. Image coding using wavelet transform. *IEEE Trans. Image Processing*, 1 (2) : 205–220, 1992.

[Belge *et al.*, 2000] M. Belge, M. Kilmer et E. Miller. Wavelet domain image restoration with adaptive edge-preserving regularization. *IEEE Trans. Image Processing*, 9 (4) : 597–608, avril 2000.

[Besag *et al.*, 1995] J. Besag, P. Green, D. Higdon et K. Mengersen. Bayesian computation and stochastic systems. *Statistical Science*, 10 : 3–66, 1995.

[Brémaud, 1998] P. Brémaud. *Markov chains. Gibbs fields and Monte Carlo.* Cours ENSTA, Paris, 1998.

[Brémaud, 1999] P. Brémaud. *Markov Chains. Gibbs fields, Monte Carlo Simulation, and Queues.* Texts in Applied Mathematics 31. Springer, New York, NY, USA, 1999.

[Charbonnier *et al.*, 1997] P. Charbonnier, L. Blanc-Féraud, G. Aubert et M. Barlaud. Deterministic edge-preserving regularization in computed imaging. *IEEE Trans. Image Processing*, 6 (2) : 298–311, février 1997.

[Descombes *et al.*, 1999] X. Descombes, M. Sigelle et F. Preteux. Estimating Gaussian Markov random field parameters in a nonstationnary framework : Application to remote sensing imaging. *IEEE Trans. Image Processing*, 8 : 490–503, avril 1999.

[Fourgeaud et Fuchs, 1972] C. Fourgeaud et A. Fuchs. *Statistique.* Dunod, Paris, 2ème edition, 1972.

[Huber, 1981] P. J. Huber. *Robust Statistics.* John Wiley, New York, NY, USA, 1981.

[Idier, 2001] J. Idier, éditeur. *Approche bayésienne pour les problèmes inverses.* Traité IC2, Série traitement du signal et de l'image, Hermès, Paris, novembre 2001.

[Jalobeanu *et al.*, 2004] A. Jalobeanu, L. Blanc-Feraud et J. Zerubia. An adaptive gaussian model for satellite image deblurring. *IEEE Trans. Image Processing*, 13 (4) : 613–621, avril 2004.

[Koulibaly *et al.*, 1996] P. Koulibaly, P. Charbonnier, L. Blanc Feraud, I. Laurette, J. Darcourt et M. Barlaud. Poisson statistic and half-quadratic regularization for emission tomography reconstruction algorithm. In *Proc. IEEE ICIP*, pages 729–732, 1996.

[Mallat, 1989] S. G. Mallat. A theory for multiresolution signal decomposition : The wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 11 (7) : 674–693, juillet 1989.

[Mazet *et al.*, 2005] V. Mazet, C. Carteret, D. Brie, J. Idier et B. Humbert. Background removal from spectra by designing and minimising a non-quadratic cost function. *Chemometrics and Intelligent Laboratory Systems*, 76 : 121–133, 2005.

[Moulin et Liu, 1999] P. Moulin et J. Liu. Analysis of multiresolution image denoising schemes using generalized - gaussian and complexity priors. *IEEE Trans. Inf. Theory*, 45 (3) : 909–919, avril 1999.

[Nikolova, 2002] M. Nikolova. Minimizers of cost-functions involving non-smooth data-fidelity terms. Application to the processing of outliers. *SIAM J. Num. Anal.*, 40 (3) : 965–994, 2002.

[Wang *et al.*, 1995] G. Wang, J. Zhang et G.-W. Pan. Solution of inverse problems in image processing by wavelet expansion. *IEEE Trans. Image Processing*, 4 (5) : 579–593, mai 1995.

[Winkler, 1995] G. Winkler. *Image Analysis, Random Fields and Dynamic Monte Carlo Methods.* Springer Verlag, Berlin, Allemagne, 1995.

Chapitre IV

MINIMISATION DES CRITÈRES PÉNALISÉS

Nous avons vu dans le chapitre II qu'un problème inverse peut être résolu à travers la minimisation d'un critère pénalisé. Il s'agit alors de mettre en œuvre une méthode d'optimisation. Le recours à une méthode d'optimisation se rencontre très fréquemment en traitement du signal et des images. Il nous semble donc tout à fait justifié d'investir sur cet aspect. L'optimisation est un domaine des mathématiques appliquées très riche et très actif. De très nombreuses méthodes d'optimisation existent pour différents types de problèmes. Les méthodes rencontrées dans le domaine de l'optimisation sont valables pour des critères très généraux et n'exigent donc pas du critère une structure analytique spécifique.

Ce constat fait, deux points de vue peuvent être considérés. Le premier point de vue consiste à appliquer une méthode d'optimisation générale, sans tenir compte de la spécificité du problème traité. Le deuxième point de vue consiste, au contraire, à tenir compte au maximum de la spécificité du problème et en particulier de la structure analytique particulière du critère. Ceci amène à envisager des méthodes dont le champ d'application est beaucoup plus réduit que les méthodes générales d'optimisation. L'intérêt réside bien dans l'espoir d'une efficacité meilleure en prenant en compte plus d'informations sur le problème à traiter. En somme, il s'agit de perdre

en généralité pour gagner en efficacité. Cette considération n'est pas nouvelle en traitement du signal et des images. L'objet de ce chapitre est de présenter ces algorithmes d'optimisation spécifiquement développés pour la minimisation de critère pénalisé et de les replacer par rapport à des méthodes générales d'optimisation.

IV.1 Le problème d'optimisation

IV.1.1 Formulation du problème

Rappelons le problème considéré. Il consiste à déterminer un minimiseur $\hat{\boldsymbol{x}}$ du critère des moindres carrés pénalisés $\mathcal{J}$ défini par

$$\mathcal{J}(\boldsymbol{x}) = \|\boldsymbol{y} - \mathbf{H}\boldsymbol{x}\|^2 + \lambda\Phi(\boldsymbol{x}) \tag{IV.1}$$

avec pour pénalisation généralisée

$$\Phi(\boldsymbol{x}) = \sum_{c=1}^{C} \phi(\|\mathbf{V}_c\boldsymbol{x} - \boldsymbol{\omega}_c\|) \tag{IV.2}$$

où $\mathbf{V}_c \in \mathbb{R}^{P\times N}$, $\boldsymbol{\omega}_c \in \mathbb{R}^P$, pour $c = 1, \ldots, C$, $\|.\|$ est la norme euclidienne et $\phi : \mathbb{R}^+ \to \mathbb{R}$ est une fonction de régularisation préservant les discontinuités. Soit $\mathbf{V}$ la matrice de taille $CP \times N$ obtenue par concaténation des C matrices $\mathbf{V}_c$:

$$\mathbf{V}^{\mathrm{t}} = \left[\mathbf{V}_1^{\mathrm{t}} | \ldots | \mathbf{V}_C^{\mathrm{t}}\right]. \tag{IV.3}$$

Autrement dit, on recherche un élément $\hat{\boldsymbol{x}}$ de l'ensemble S_{CP} défini par :

$$S_{\mathrm{CP}} = \{\boldsymbol{x} \in \mathbb{R}^N, \min \mathcal{J}(\boldsymbol{x})\}.$$

Il s'agit d'un problème de minimisation *sans contrainte*. Dans certains cas, on peut être amené à considérer un problème de minimisation *avec contrainte*. Au lieu de considérer $\boldsymbol{x} \in \mathbb{R}^N$, on considère alors $\boldsymbol{x} \in \mathcal{X}$, où $\mathcal{X}$ est l'ensemble de contraintes. La prise en compte de contraintes ne sera pas abordée mais nous indiquons que les méthodes présentées dans ce chapitre peuvent être en partie adaptées pour tenir compte de contraintes séparables, telle que celle de non négativité [Kaufman, 1993].

IV.1.2 Différentiabilité du critère pénalisé généralisé

L'étude des propriétés du critère à minimiser est une étape essentielle pour résoudre un problème d'optimisation. En effet, ce sont bien les propriétés du critère qui conditionnent le type de méthode employée pour répondre au problème de la minimisation du critère. Parmi les propriétés à étudier, une des plus importantes est la *différentiabilité* du critère.

De nombreuses fonctions de régularisation ϕ préservant les discontinuités utilisés en restauration d'images sont $\mathcal{C}^1$, comme celles présentées dans les parties II.2.3.[A] et II.2.3.[B] à l'exception de la quadratique tronquée. Par la suite, nous supposerons systématiquement la différentiabilité de la fonction de régularisation ϕ. Il a déjà été établi que la différentiabilité de la fonction de régularisation ϕ entraîne la différentiabilité du critère pénalisé non généralisé [Allain *et al.*, 2006]. La différentiabilité du critère pénalisé généralisé (IV.1)-(IV.2) est à présent établie.

Le gradient $\nabla\mathcal{J}(\boldsymbol{x})$ du critère pénalisé généralisé $\mathcal{J}$ au point $\boldsymbol{x}$ s'écrit

$$\nabla\mathcal{J}(\boldsymbol{x}) = 2\mathbf{H}^{\mathrm{t}}\left(\mathbf{H}\boldsymbol{x} - \boldsymbol{y}\right) + \lambda\nabla\Phi(\boldsymbol{x})$$

où le gradient de la pénalisation s'écrit

$$\nabla\Phi(\boldsymbol{x}) = \sum_{c=1}^{C} \frac{\phi'(\|\mathbf{V}_c\boldsymbol{x} - \boldsymbol{\omega}_c\|)}{\|\mathbf{V}_c\boldsymbol{x} - \boldsymbol{\omega}_c\|}\mathbf{V}_c^{\mathrm{t}}(\mathbf{V}_c\boldsymbol{x} - \boldsymbol{\omega}_c). \tag{IV.4}$$

En remarquant que la fonction $\boldsymbol{u} \mapsto \mathbf{V}_c^{\mathrm{t}}\boldsymbol{u}/\left\|\boldsymbol{u}\right\|$ est bornée sur $\mathbb{R}^N$, le gradient de la pénalisation est bien défini sur $\mathbb{R}^N$, ce qui assure la différentiabilité du critère pénalisé généralisé $\mathcal{J}$. En supposant que ϕ est C^1 et que $\phi'(0) = 0$, le gradient de la pénalisation est alors continu ce qui assure le caractère C^1 du critère pénalisé généralisé $\mathcal{J}$ sur $\mathbb{R}^N$. Contrairement au critère pénalisé non généralisé, le caractère C^1 de la fonction de régularisation ϕ ne suffit pas pour assurer le caractère C^1 du critère pénalisé généralisé. Il faut aussi supposer l'annulation de la dérivée de la fonction de régularisation en zéro. Notons que cette dernière propriété est systématiquement vérifiée par les fonctions de régularisation C^1 paires.

Il est important de souligner que le gradient du critère pénalisé généralisé s'obtient sous une forme analytique simple. De plus, la complexité du calcul du gradient est du même ordre de grandeur que l'évaluation du critère. A présent, nous présentons les conditions classiques d'optimalité, s'appuyant sur la différentiabilité du critère pénalisé généralisé, qui définissent de manière pratique les solutions du problème de minimisation.

IV.1.3 Conditions d'optimalité

L'objectif est de trouver au moins un élément de l'ensemble S_{CP}. On rappelle les définitions d'un minimum local et global du critère $\mathcal{J}$ [Bertsekas, 1999, Sec. 1.1.1]. Un vecteur $\boldsymbol{x}^*$ est un *mininum local* de $\mathcal{J}$ s'il existe $\epsilon > 0$ tel que

$$\mathcal{J}(\boldsymbol{x}^*) \leq \mathcal{J}(\boldsymbol{x}), \quad \forall\boldsymbol{x} \in \mathbb{R}^N \text{ tel que } \|\boldsymbol{x}^* - \boldsymbol{x}\| < \epsilon.$$

Autrement dit, un minimum local ne fait pas moins bien que ses voisins.

Le vecteur $\boldsymbol{x}^*$ est un *minimum global* de $\mathcal{J}$ si

$$\mathcal{J}(\boldsymbol{x}^*) \leq \mathcal{J}(\boldsymbol{x}), \quad \forall\boldsymbol{x} \in \mathbb{R}^N.$$

Lorsque les deux inégalités ci-dessus sont strictes pour $\boldsymbol{x} \neq \boldsymbol{x}^*$ alors le minimum local ou global est dit *strict*.

Pour qu'un vecteur $\boldsymbol{x}^*$ soit un minimum local du critère différentiable $\mathcal{J}$, il est nécessaire que la condition suivante soit vérifiée

$$\nabla\mathcal{J}(\boldsymbol{x}^*) = \mathbf{0}. \tag{IV.5}$$

Un vecteur $\boldsymbol{x}^*$ satisfaisant cette condition est appelé *point stationnaire*. Cependant, cette condition n'est pas suffisante dans le cas général pour assurer que $\boldsymbol{x}^*$ est un minimum local. Lorsque le critère $\mathcal{J}$ est C^2, il est nécessaire que le Hessien en $\boldsymbol{x}^*$ soit défini non négatif. Par contre, il suffit que le Hessien en $\boldsymbol{x}^*$ soit défini strictement positif pour assurer que $\boldsymbol{x}^*$ est un minimum local [Bertsekas, 1999, Prop. 1.1.1 et 1.1.3]. On voit ainsi qu'il n'existe pas de condition nécessaire et suffisante dans le cas général assurant qu'un point stationnaire est un minimum local.

[A] Cas convexe

La convexité de la fonction de régularisation ϕ entraîne celle du critère $\mathcal{J}$. Plus précisément, si ϕ est convexe alors $\mathcal{J}$ l'est aussi. De plus, si ϕ est strictement convexe, alors le critère pénalisé $\mathcal{J}$ l'est aussi si et seulement si la condition $\ker(\mathbf{H}^{\mathrm{t}}\mathbf{H}) \cap \ker(\mathbf{V}^{\mathrm{t}}\mathbf{V}) = \{\mathbf{0}\}$ est satisfaite [Delaney et Bresler, 1998, Théorème 2, p. 209]. La convexité du critère $\mathcal{J}$ est de toute première importance car elle entraîne des propriétés fortes d'optimalité. D'une part, si $\mathcal{J}$ est convexe alors la condition d'annulation du gradient (IV.5) est une condition nécessaire et suffisante assurant que le point $\boldsymbol{x}^*$ est un minimum local. De plus, tous les minima locaux du critère $\mathcal{J}$ sont des minima globaux. Finalement si le critère $\mathcal{J}$ est strictement convexe alors il n'existe qu'un unique minimum global [Bertsekas, 1999, Prop. 1.1.2]. Notons que ces propriétés sont valables même si le critère $\mathcal{J}$ est convexe sans être pour autant C^2, ce qui est le cas par exemple avec la fonction de régularisation de Huber définie par (II.11).

La convexité du critère $\mathcal{J}$ est une condition suffisante pour obtenir ces propriétés. Bien que toutes les fonctions de régularisation préservant les discontinuités utilisées en restauration d'images sont *unimodales* ; *i.e.,* ont un unique minimum local, ce caractère n'est pas suffisant à lui seul pour assurer l'unimodalité du critère $\mathcal{J}$. Si la convexité stricte du critère est une propriété plus forte que l'unimodalité, elle est néanmoins vérifiable bien plus simplement en pratique. Ceci explique pourquoi l'accent est plutôt mis sur la propriété de convexité du critère plutôt que sur l'unimodalité.

Ainsi, ces propriétés fortes liées à la convexité indiquent que le problème de minimisation d'un critère convexe n'est pas de nature comparable à celui d'un critère non convexe. En somme, il s'agit d'un problème considérablement plus simple puisqu'une caractéristique *globale* (le minimiseur de $\mathcal{J}$) est directement associée à une caractéristique *locale*. Résoudre le problème global passe alors simplement par résoudre le problème au niveau local dans le cas convexe. Il s'ensuit que la mise en œuvre d'une méthode d'optimisation locale résout entièrement un problème convexe.

[B] Cas non convexe

La recherche d'un minimum global du critère $\mathcal{J}$ non convexe ne se résume pas, en général, à la recherche d'un minimum local, contrairement au cas convexe. Un critère non convexe est en général *multimodal*, c'est-à-dire qu'il possède plusieurs minima locaux. Il est donc nécessaire de mettre en œuvre une approche par optimisation globale au coût calculatoire bien plus élevé que pour le cas convexe. Deux types d'approches peuvent être distingués ; d'une part les méthodes déterministes comme la *non convexité graduelle* (NCG) [Blake et Zisserman, 1987; Nikolova *et al.*, 1998] qui consiste à générer une suite de critères pour lesquels une minimisation est effectuée à l'aide d'une méthode de descente (locale), en commençant par un critère convexe pour terminer par le critère non convexe. Précisons qu'il existe des contre-exemples montrant que cette approche ne converge pas toujours [Nikolova, 1999, p. 1209]. D'autre part, les méthodes pseudo-aléatoires comme *le recuit simulé* [Geman et Geman, 1984] exploitent la génération d'un grand nombre de points pour explorer l'ensemble des modes du critère. Le recuit simulé converge en probabilité. Il est possible de considérer une version hybride pour les méthodes pseudo-aléatoires en effectuant une étape d'optimisation locale pour les points générés aléatoirement.

Plutôt que d'opposer optimisation globale et locale notons que ces deux approches sont fortement couplées car les méthodes modernes d'optimisation globale font appel à une ou plusieurs étapes d'optimisation locale. Ceci montre l'importance de l'optimisation locale, même pour un critère non convexe.

IV.2 Méthodes itératives de minimisation

IV.2.1 Recherche itérative de la solution

Lorsqu'une pénalisation quadratique de Tikhonov est considérée, la solution $\hat{\boldsymbol{x}}$ minimiseur du critère pénalisé peut s'exprimer de manière analytique. Cependant, le calcul numérique de la solution de Tikhonov ne peut généralement pas s'effectuer de manière directe, car trop coûteux. Lorsqu'une pénalisation non quadratique est considérée, la solution $\hat{\boldsymbol{x}}$ ne peut pas se mettre sous une forme analytique. Dans ces deux situations, un *algorithme itératif* est alors mis en œuvre afin d'estimer la solution $\hat{\boldsymbol{x}}$. A partir d'un point initial $\boldsymbol{x}_0$, cet algorithme génère une suite d'itérées $\{\boldsymbol{x}_k\}$ convergeant vers la solution $\hat{\boldsymbol{x}}$, lorsque les conditions de convergence associées sont vérifiées.

L'objectif fixé pour les algorithmes itératifs est donc plus faible que celui consistant à trouver exactement un minimum local. Il s'agit plutôt de s'approcher suffisamment près d'un minimum local du critère pénalisé $\mathcal{J}$. Les algorithmes itératifs considérés dans ce chapitre auront au moins les propriétés suivantes

1) La nouvelle itérée dépend exclusivement de la précédente :

$$\boldsymbol{x}_{k+1} = f(\boldsymbol{x}_k)$$

2) L'application $f : \mathbb{R}^N \mapsto \mathbb{R}^N$ est déterministe
3) Les itérées $\{\boldsymbol{x}_k\}$ vérifient la condition de descente :

$$\mathcal{J}(\boldsymbol{x}_{k+1}) \leq \mathcal{J}(\boldsymbol{x}_k)$$

4) La suite des gradients converge vers le vecteur nul :

$$\lim_{k\to\infty} \nabla\mathcal{J}(\boldsymbol{x}_k) = \mathbf{0}.$$

Plus précisément, les deux dernières propriétés seront vérifiées si les hypothèses associées sont satisfaites. La troisième propriété est insuffisante à elle seule pour assurer la convergence du gradient du critère pénalisé vers zéro. Si la descente est insuffisante à chaque itération, la suite des gradients du critère pénalisé peut ne pas s'annuler. On verra par la suite qu'il faut remplacer cette condition de descente par des conditions de *descente suffisante*. Par la suite, lorsqu'on parlera de convergence d'un algorithme de minimisation itératif, on fera référence à la quatrième propriété. On rappelle que cette convergence entraîne la convergence vers le minimum global si le critère pénalisé est strictement convexe.

IV.2.2 Critère d'arrêt

Les algorithmes itératifs ne sont généralement pas convergents en un nombre fini d'itérations. Il faut donc se munir d'un critère d'arrêt afin de s'assurer qu'on se trouve suffisamment proche d'un point stationnaire. Le critère d'arrêt le plus courant est

$$\|\nabla\mathcal{J}(\boldsymbol{x}_k)\| \leq \epsilon \tag{IV.6}$$

où $\epsilon > 0$ est un seuil à déterminer. En général il faut s'assurer que ϵ soit suffisamment petit. Lorsque $\nabla^2\mathcal{J}(\boldsymbol{x})$ est définie positive, l'utilisation du critère d'arrêt (IV.6) entraîne une majoration de l'écart de l'itérée courante par rapport au minimum global $\boldsymbol{x}^*$ de $\mathcal{J}$, lorsque l'itérée courante est suffisamment proche de $\boldsymbol{x}^*$. Plus précisément, s'il existe $m > 0$ tel que pour tout $\boldsymbol{x}$ dans une sphère S centrée en $\boldsymbol{x}^*$ on a

$$m\|\boldsymbol{z}\|^2 \leq \boldsymbol{z}'\nabla^2\mathcal{J}(\boldsymbol{x})\boldsymbol{z}, \quad \forall \boldsymbol{z} \in \mathbb{R}^N$$

alors tout $\boldsymbol{x} \in S$ qui vérifie $\|\nabla \mathcal{J}(\boldsymbol{x})\| \leq \epsilon$ satisfait aussi [Bertsekas, 1999, p. 37]

$$\|\boldsymbol{x} - \boldsymbol{x}^*\| \leq \epsilon/m. \tag{IV.7}$$

Ainsi, ce résultat suggère comment déterminer ϵ pour obtenir la précision souhaitée.

Cependant, le résultat (IV.7) n'est en général pas intéressant en pratique car le calcul d'une valeur propre d'une matrice doit le plus souvent se faire de manière itérative, donc avec un coût important [Mackens et Voss, 2000]. En pratique, on utilisera donc le critère d'arrêt (IV.6).

IV.2.3 Qu'est-ce qu'un bon algorithme itératif ?

Plusieurs algorithmes itératifs peuvent être envisagés pour minimiser (localement) les critères pénalisés. Il s'agit alors de choisir un algorithme itératif qui soit le plus adapté possible au problème de minimisation considéré. On peut s'interroger sur les propriétés que doit vérifier un "bon" algorithme itératif. Un "bon" algorithme itératif peut, à notre sens, être caractérisé par les propriétés suivantes

- *Convergence* : avec hypothèses associées suffisamment larges.
- *Efficacité* : bon compromis entre taux de convergence et coût calculatoire par itération.
- *Simplicité* : faible nombre de paramètres et facilité de réglage.
- *Stabilité* face aux erreurs numériques.

Ici, nous ne considérons que des algorithmes itératifs convergents. Cette propriété indique simplement que l'algorithme itératif réalise bien ce qui est attendu de lui. L'efficacité, prise au sens de faible consommation des ressources informatiques, est un critère décisif pour choisir un algorithme itératif. Pour un algorithme itératif donné, à l'image du domaine de convergence, un domaine d'efficacité peut lui être associé. Ainsi, il faut bien avoir à l'esprit qu'un algorithme itératif plus efficace qu'un autre pour une situation donnée pourra être moins efficace pour une autre situation.

La simplicité nous semble être une caractéristique tout aussi importante. En pratique, elle est fortement liée à l'efficacité, car l'objectif du réglage des paramètres de l'algorithme itératif est bien d'obtenir au final un algorithme efficace pour un problème donné.

La stabilité face aux erreurs numériques liées à la précision finie de l'implémentation informatique ne sera pas traitée dans le cadre de ce travail. Une étude rigoureuse de la convergence impliquerait de tenir compte des erreurs numériques.

A présent nous allons passer en revue certains algorithmes itératifs couramment utilisés en optimisation. On les comparera selon leurs propriétés d'efficacité et de simplicité.

IV.3 Méthodes itératives génériques

L'optimisation est un domaine très riche des mathématiques appliquées. Cette section se propose de faire un tour d'horizon de certains algorithmes itératifs employés pour la minimisation des critères pénalisés différentiables. Ces algorithmes itératifs peuvent être regroupés selon deux familles. La première famille se compose des algorithmes de *relaxation*, qui consistent à fragmenter le problème de minimisation initial en une série de sous-problèmes de dimension réduite. La deuxième famille se compose des méthodes travaillant sur l'ensemble de l'espace solution, en utilisant le *gradient* du critère pénalisé.

IV.3.1 Algorithmes de relaxation

L'algorithme de relaxation coordonnée par coordonnée minimise le critère $\mathcal{J}$ en minimisant le critère *monovarié* $\mathcal{J}^{(n)}$ associé à la variable $x^{(n)}$ mis à jour à la $k+1$ itération selon

$$x_{k+1}^{(n)} = \arg\min_{u} \mathcal{J}^{(n)}(u) \tag{IV.8}$$

où

$$\mathcal{J}^{(n)}(u) = \mathcal{J}(x_{k+1}^{(1)}, \ldots, x_{k+1}^{(n-1)}, u, x_{k}^{(n+1)}, \ldots, x_{k}^{(N)})$$

une itération complète $k \rightarrow k+1$ est obtenue après un balayage complet des N composantes de $\boldsymbol{x}$. Ce schéma itératif coordonnée par coordonnée est connu sous le nom de *méthode de Gauss Seidel*. On peut citer [Bouman et Sauer, 1996; Erdogan et Fessler, 1999] pour l'emploi de ces algorithmes itératifs dans le domaine de la reconstruction d'images.

Notons que les méthodes de relaxation ne sont pas limitées à la mise à jour coordonnée par coordonnée, mais qu'elles peuvent mettre en jeu des *blocs de coordonnées*. Dans ce cas, chaque itération conduit à des sous problèmes d'optimisation multivariés. Une condition suffisante de convergence globale de ces algorithmes de relaxation est la convexité stricte du critère et le fait que l'ensemble des lignes de niveaux soit un compact [Tseng et Bertsekas, 1987, p. 306].

Un des inconvénients de ces algorithmes de relaxation réside dans leur faible taux de convergence. Le taux de convergence des algorithmes de relaxation peut se révéler bien plus faible que celui de l'algorithme de plus forte descente [Nocedal et Wright, 1999, p. 54]. Notons que l'ordre dans lequel la relaxation est mise en œuvre a une influence parfois sensible sur le taux de convergence.

IV.3.2 Algorithmes à directions de descente

Les algorithmes à directions de descente sont des schémas itératifs répandus en optimisation et largement employés pour la minimisation des critères pénalisés. Nous présentons à présent certaines des variantes les plus courantes.

[A] Forme générale

A l'itération courante k, la mise à jour s'écrit

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \theta_k \boldsymbol{d}_k$$

avec $\boldsymbol{d}_k$ et θ_k respectivement la direction de déplacement et le pas. Les méthodes à directions de descente assurent que la direction $\boldsymbol{d}_k$ fait décroître strictement le critère

$$\mathcal{J}(\boldsymbol{x}_{k+1}) < \mathcal{J}(\boldsymbol{x}_k)$$

si le pas $\theta_k > 0$ est choisi suffisamment petit. Cependant, garantir la décroissance stricte ne suffit pas à garantir la convergence globale de ces algorithmes.

Nous passons en revue les algorithmes à directions de descente les plus répandus dans le domaine du traitement de l'image, soit les algorithmes de plus forte descente, du gradient conjugué et de quasi-Newton. Ces algorithmes se distinguent par le type de direction de descente employée. L'efficacité varie sensiblement d'un algorithme à l'autre.

[B] Algorithme de plus forte descente

On commence par l'algorithme à directions de descente le plus simple. Cet algorithme produit une mise à jour courante suivant la plus forte pente locale

$$\boldsymbol{d}_k = -\nabla\mathcal{J}(\boldsymbol{x}_k).$$

L'algorithme de plus forte descente est connu pour son faible taux de convergence en comparaison du gradient conjugué ou d'une méthode de quasi-Newton. Ainsi, on lui préfère souvent un algorithme du gradient conjugué qui permet une convergence plus rapide au prix d'un encombrement mémoire et d'un coût de calcul par itération très légèrement supérieur.

[C] Algorithmes du gradient conjugué

L'algorithme du gradient conjugué est une méthode d'optimisation qui a été initialement proposé par [Hestenes et Stiefel, 1952] pour la minimisation de critères quadratiques dont le Hessien est symétrique défini positif. Cet algorithme a ensuite été étendu à des critères non quadratiques donnant naissance à de nombreuses variantes. Nous reviendrons de manière détaillée sur cette famille d'algorithmes dans le chapitre VI.

[D] Algorithmes de quasi-Newton

L'algorithme de Newton calcule une direction de déplacement à partir du Hessien $\nabla^2\mathcal{J}(\boldsymbol{x}_k)$ du critère $\mathcal{J}$ au point courant

$$\boldsymbol{d}_k = -(\nabla^2\mathcal{J}(\boldsymbol{x}_k))^{-1}\nabla\mathcal{J}(\boldsymbol{x}_k).$$

Cette méthode ne permet pas d'assurer dans le cas général que $\boldsymbol{d}_k$ est une direction de descente. En particulier, cet algorithme n'est pas défini lorsque le Hessien est singulier. L'algorithme de Newton a souvent un comportement pathologique et peut diverger ou cycler sans converger [Bertsekas, 1999, p. 92]. Il ne peut donc généralement pas être utilisé tel quel et doit être modifié. De plus, le coût en ressources informatiques (temps et stockage) du calcul de l'inverse du Hessien devient vite prohibitif pour les critères pénalisés en fonction de la taille du problème. Les formes de quasi-Newton ne calculant pas explicitement l'inverse du Hessien ont un coût bien plus faible. Ainsi, les algorithmes de quasi-Newton sont préférés à l'algorithme de Newton pour des raisons de convergence et de meilleure efficacité.

Les directions de descente des algorithmes de quasi-Newton ont la structure suivante [Nocedal et Wright, 1999, p. 194]

$$\boldsymbol{d}_k = -\mathbf{M}_k^{-1}\nabla\mathcal{J}(\boldsymbol{x}_k) \tag{IV.9}$$

où $\mathbf{M}_k$ est une matrice définie positive. La matrice $\mathbf{M}_k$ est choisie de telle sorte que la direction $\boldsymbol{d}_k$ résultante tende à approcher la direction de Newton. Le principe des algorithmes de quasi-Newton est d'établir un compromis entre l'efficacité de la méthode de Newton en terme de taux de convergence et le coût de calcul.

L'algorithme BFGS est un représentant classique des ces algorithmes de quasi-Newton [Nocedal et Wright, 1999, p. 194]. L'intérêt de cet algorithme est que le calcul de la direction ne fait pas intervenir d'inversion de matrice contrairement à l'algoritme de Newton [Nocedal et Wright, 1999, p. 198]. Cependant, l'algorithme BFGS semble peu utilisé pour les problèmes de traitement d'images. D'autres algorithmes de quasi-Newton sont utilisés à la place de l'algorithme BFGS en traitement de l'image. Il s'agit des algorithmes semi-quadratiques, présentés dans la section suivante.

IV.4 Algorithmes semi-quadratiques

Les algorithmes semi-quadratiques ont été spécifiquement développés pour l'optimisation des critères pénalisés. Leur principe consiste à substituer un problème d'optimisation équivalent à celui des critères pénalisés, mais plus simple. Il en résulte une famille d'algorithmes de formulation simple, qui sont globalement convergents sous des hypothèses peu contraignantes.

IV.4.1 Principe

Des algorithmes d'optimisation spécifiques semi-quadratiques (SQ) exploitant la forme analytique des critères pénalisés ont été utilisés dans la communauté du traitement de l'image pour la restauration d'image préservant les discontinuités [Charbonnier *et al.*, 1997; Nikolova et Ng, 2005; Allain *et al.*, 2006]. Ces algorithmes SQ découlent des formulations de Geman et Yang (GY) et celle de Geman et Reynolds (GR). Le principe de ces approches consiste à transformer le problème initial de la minimisation du critère pénalisé non quadratique $\mathcal{J}(\boldsymbol{x})$ en un problème équivalent plus simple. Il s'agit de déterminer un critère $\mathcal{K}(\boldsymbol{x}, \boldsymbol{b})$ qui est un *équivalent augmenté* du critère $\mathcal{J}(\boldsymbol{x})$

$$\min_{\boldsymbol{b}} \mathcal{K}(., \boldsymbol{b}) = \mathcal{J}(.)$$

au sens où le critère $\mathcal{K}$ dépend non seulement de $\boldsymbol{x}$, mais aussi des variables *auxiliaires* $\boldsymbol{b}$. Tout l'enjeu est alors d'expliciter le critère $\mathcal{K}$ afin que la minimisation soit plus simple que le problème initial.

L'intérêt des approches SQ réside dans le fait que le critère augmenté $\mathcal{K}$ présente des propriétés structurelles remarquables : il est *quadratique* selon $\boldsymbol{x}$ et sa minimisation vis-à-vis des variables auxiliaires $\boldsymbol{b}$ est explicite. Ces qualités compensent largement l'augmentation du nombre de variables. La minimisation du critère $\mathcal{K}$ se déroule alors par *relaxation*, c'est-à-dire en considérant le sous-problème de minimisation à $\boldsymbol{x}$ fixé, puis à $\boldsymbol{b}$ fixé, de manière alternée.

Notons que la minimisation du critère $\mathcal{K}$ a été initialement proposée dans un cadre stochastique [Geman et Reynolds, 1992; Geman et Yang, 1995]. A présent, ces approches SQ sont largement utilisées dans un cadre déterministe où le critère pénalisé est convexe [Charbonnier *et al.*, 1997; Idier, 2001]. Ce cadre convexe est particulièrement bien adapté à la démarche SQ. Dans [Charbonnier *et al.*, 1997], il est démontré que la procédure de descente *bloc de coordonnées par bloc de coordonnées*, alternativement à $\boldsymbol{x}$ fixé puis à $\boldsymbol{b}$ fixé, issue de la construction de GR converge dans le cas convexe vers l'unique minimum sous des conditions larges. Dans [Aubert et Vese, 1997], la convergence de la procédure de descente bloc de coordonnées par bloc de coordonnées issue de la construction de GY est démontrée sous des conditions similaires.

IV.4.2 Hypothèses sur la fonction de régularisation

Nous présentons ci-dessous les deux ensembles d'hypothèses sur la fonction de régularisation ϕ du critère pénalisé, associés respectivement à la construction de GY et à celle de GR. Ce sont les mêmes hypothèses que celles de [Allain *et al.*, 2006], à l'exception de la relaxation de l'hypothèse de convexité. L'ensemble d'hypothèses associées à la construction de GY fait appel au caractère gradient Lipschitz défini ci-dessous.

Définition IV.4.1. *Soit E un espace métrique. $\mathcal{F} : E \mapsto \mathbb{R}$ est dit μ-$\mathcal{LC}^1$ si $\mathcal{F}$ est différentiable en E et $\nabla\mathcal{F}$ est Lipschitz continue en E avec la constante de Lipschitz $\mu > 0$, c'est-à-dire*

$$\|\nabla\mathcal{F}(\boldsymbol{x}) - \nabla\mathcal{F}(\boldsymbol{x}')\| \leq \mu\|\boldsymbol{x} - \boldsymbol{x}'\|, \quad \forall \boldsymbol{x}, \boldsymbol{x}' \in E.$$

Hypothèse 1.

$$\begin{cases} \phi \text{ est coercive.} & \text{(IV.10a)} \\ \phi \text{ est } L\text{-}\mathcal{LC}^1 \text{ avec } L = 1/\hat{a}. & \text{(IV.10b)} \\ \phi'(0) = 0. & \text{(IV.10c)} \end{cases}$$

Hypothèse 2.

$$\begin{cases} \phi \text{ est } \mathcal{C}^1\text{, paire et coercive.} & \text{(IV.11a)} \\ \phi(\sqrt{.}) \text{ est concave sur } \mathbb{R}^+. & \text{(IV.11b)} \\ \phi'(t)/t < \infty, \quad \forall t \in \mathbb{R}. & \text{(IV.11c)} \end{cases}$$

Il est à noter que les Hypothèses 1 et 2 sur la fonction de régularisation du critère pénalisé sont des conditions techniques à respecter et n'assurent pas, à elles seules, le caractère préservant les discontinuités. En effet, la quadratique $\phi(t) = t^2$ vérifie ces deux hypothèses alors qu'elle ne permet pas de préserver les discontinuités. Par contre, la plupart des fonctions de régularisation préservant les discontinuités vérifient ces hypothèses, du moins celles qui sont suffisamment régulières (caractère $\mathcal{C}^1$). Par exemple, les fonctions de régularisation de Huber et hyperbolique définies dans la section II.2.3.[A] page 36 vérifient chacune les Hypothèses 1 et 2.

Dans les deux sections suivantes sont présentés les critères augmentés de GR et GY donnant lieu à des problèmes semi-quadratiques équivalents au problème initial du critère pénalisé. Ces constructions s'appuient sur des résultats de l'analyse convexe, principalement sur la notion de *fonction conjuguée* [Rockafellar, 1970].

Dans un souci de clarté, nous indiquons le critère pénalisé considéré dans cette partie

$$\mathcal{J}(\boldsymbol{x}) = \|\boldsymbol{y} - \mathbf{H}\boldsymbol{x}\|^2 + \sum_{c=1}^{C} \phi(\boldsymbol{v}_c^{\mathrm{t}}\boldsymbol{x} - \omega_c). \tag{IV.12}$$

ainsi que le critère pénalisé généralisé

$$\mathcal{J}(\boldsymbol{x}) = \|\boldsymbol{y} - \mathbf{H}\boldsymbol{x}\|^2 + \sum_{c=1}^{C} \phi(\|\mathbf{V}_c\boldsymbol{x} - \boldsymbol{\omega}_c\|) \tag{IV.13}$$

où le paramètre de régularisation λ est intégré à la fonction de régularisation ϕ afin de simplifier les notations.

IV.4.3 Construction de Geman et Reynolds

[A] Critère pénalisé

On considère le critère pénalisé (IV.12). Lorsque la fonction de régularisation ϕ vérifie l'hypothèse 2, le critère augmenté de Geman et Reynolds s'écrit [Idier, 2001, Sec. IV]

$$\mathcal{K}_{GR}(\boldsymbol{x}, \boldsymbol{b}) = \|\boldsymbol{y} - \mathbf{H}\boldsymbol{x}\|^2 + \sum_{c=1}^{C} \left(b_c(\boldsymbol{v}_c^{\mathrm{t}}\boldsymbol{x} - \omega_c)^2 + \psi(b_c) \right)$$

où la variable auxiliaire $\boldsymbol{b} = (b_1, \ldots, b_C)^{\mathrm{t}}$ et la fonction ψ dépendent de la fonction de régularisation ϕ. La fonction ψ est définie par

$$\psi(b) = \sup_{u \in \mathbb{R}} \left(\phi(u) - bu^2 \right).$$

Des résultats de l'analyse convexe permettent d'expliciter la valeur des variables auxiliaires minimisant $\mathcal{K}_{GR}$ à $\boldsymbol{x}$ fixé [Idier, 2001, Sec. IV]

$$b_c = \frac{\phi'(\boldsymbol{v}_c^{\mathrm{t}}\boldsymbol{x} - \omega_c)}{2(\boldsymbol{v}_c^{\mathrm{t}}\boldsymbol{x} - \omega_c)}, \quad c \in [1, C].$$

La forme quadratique en $\boldsymbol{x}$ à $\boldsymbol{b}$ fixé permet de déduire que le minimiseur de $\mathcal{K}_{GR}$ à $\boldsymbol{b}$ fixé vérifie l'équation normale

$$(\mathbf{H}^{\mathrm{t}}\mathbf{H} + \mathbf{V}^{\mathrm{t}}\mathbf{B}\mathbf{V})\boldsymbol{x} = \mathbf{V}^{\mathrm{t}}\mathbf{B}\boldsymbol{\omega} + \mathbf{H}^{\mathrm{t}}\boldsymbol{y} \tag{IV.14}$$

où $\mathbf{B} = \mathrm{Diag}\,\{b_1, \dots b_C\}$ et la matrice $\mathbf{V}$ de taille $C \times N$ définie par

$$\mathbf{V}^{\mathrm{t}} = (\boldsymbol{v}_1 | \dots | \boldsymbol{v}_C) \tag{IV.15}$$

correspond au cas $P = 1$ de (IV.3).

La matrice $\mathbf{H}^{\mathrm{t}}\mathbf{H} + \mathbf{V}^{\mathrm{t}}\mathbf{B}\mathbf{V}$ de l'équation normale (IV.14) dépend de l'itérée courante. Elle joue le rôle d'une approximation du Hessien du critère pénalisé. Notons aussi que la construction de GR ne fait pas intervenir de paramètre propre, elle est entièrement déterminée par le critère pénalisé. Les constructions suivantes mettent en jeu une approximation du Hessien constante, mais dépendant d'un paramètre de réglage propre, contrairement à la construction de GR.

[B] Critère pénalisé généralisé

On considère le critère pénalisé généralisé (IV.13). Lorsque la fonction de régularisation ϕ vérifie l'hypothèse 2, il est également possible de définir un critère augmenté de Geman et Reynolds [Nikolova et Ng, 2005, Sec. 4.1]

$$\mathcal{K}_{GR}(\boldsymbol{x}, \boldsymbol{b}) = \|\boldsymbol{y} - \mathbf{H}\boldsymbol{x}\|^2 + \sum_{c=1}^{C} \left(b_c \|\mathbf{V}_c\boldsymbol{x} - \boldsymbol{\omega}_c\|^2 + \psi(b_c) \right) \tag{IV.16}$$

avec

$$b_c = \frac{\phi'(\|\mathbf{V}_c\boldsymbol{x} - \boldsymbol{\omega}_c\|)}{2\|\mathbf{V}_c\boldsymbol{x} - \boldsymbol{\omega}_c\|}, \quad c \in [1, C].$$

IV.4.4 Construction de Geman et Yang

[A] Critère pénalisé

On considère le critère pénalisé (IV.12).

Lemme IV.4.1. *Supposons que l'Hypothèse 1 est satisfaite. Alors la fonction*

$$g(u) = u^2/2 - a\phi(u) \tag{IV.17}$$

est convexe pour $0 < a < \hat{a}$.

Preuve. Voir Annexe B.1.1, page 143. □

Lorsque la fonction de régularisation ϕ assure que la fonction g définie par (IV.17) est convexe, le critère augmenté de Geman et Yang s'écrit [Idier, 2001, Sec. III]

$$\mathcal{K}_{GY}(\boldsymbol{x}, \boldsymbol{b}) = \|\boldsymbol{y} - \mathbf{H}\boldsymbol{x}\|^2 + \frac{1}{a}\sum_{c=1}^{C} \left(\frac{1}{2}(\boldsymbol{v}_c^{\mathrm{t}}\boldsymbol{x} - \omega_c - b_c)^2 + \zeta(b_c) \right)$$

avec

$$\zeta(b) = \sup_{u \in \mathbb{R}} \left(a\phi(u) - (b-u)^2/2 \right).$$

Comme pour la construction de GR, la valeur des variables auxiliaires minimisant $\mathcal{K}_{GR}$ à $\boldsymbol{x}$ fixé peut être explicitée [Idier, 2001, Sec. III]

$$b_c = \boldsymbol{v}_c^{\mathrm{t}}\boldsymbol{x} - \omega_c - a\phi'(\boldsymbol{v}_c^{\mathrm{t}}\boldsymbol{x} - \omega_c), \quad c \in [1, C].$$

Comme pour la construction de GR, la forme du critère augmenté $\mathcal{K}_{GY}$ est quadratique en $\boldsymbol{x}$, ce qui permet de déduire que le minimiseur de $\mathcal{K}_{GY}$ à $\boldsymbol{b}$ fixé vérifie l'équation normale

$$(2\mathbf{H}^{\mathrm{t}}\mathbf{H} + \frac{1}{a}\mathbf{V}^{\mathrm{t}}\mathbf{V})\boldsymbol{x} = \frac{1}{a}\mathbf{V}^{\mathrm{t}}(\boldsymbol{b} + \boldsymbol{\omega}) + 2\mathbf{H}^{\mathrm{t}}\boldsymbol{y} \tag{IV.18}$$

où $\mathbf{V}$ est définie par (IV.15), et $a > 0$ est le facteur d'échelle. Notons que contrairement à la construction de GR, la matrice de gauche ne fait pas intervenir les variables auxiliaires $\boldsymbol{b}$. Cette caractéristique peut rendre parfois la construction de GY plus intéressante que celle de GR pour la mise en œuvre.

[B] Critère pénalisé généralisé

On considère le critère pénalisé généralisé (IV.13). Lorsque la fonction de régularisation ϕ vérifie l'hypothèse 1, il est également possible de définir un critère augmenté de Geman et Yang [Ciuciu et Idier, 2002]

$$\mathcal{K}(\boldsymbol{x}, \boldsymbol{b}) = \|\boldsymbol{y} - \mathbf{H}\boldsymbol{x}\|^2 + \frac{1}{a}\sum_{c=1}^{C} \left(\frac{1}{2}\|\mathbf{V}_c\boldsymbol{x} - \boldsymbol{\omega}_c - \boldsymbol{b}_c\|^2 + \zeta(\boldsymbol{b}_c) \right)$$

Notons que contrairement à la forme de Geman et Reynolds (IV.16), les variables auxiliaires $\boldsymbol{b}_c \in \mathbb{R}^P$ sont ici vectorielles.

IV.4.5 Algorithmes semi-quadratiques

[A] Principe

La structure des critères augmentés semi-quadratiques $\mathcal{K}_{GR}$ et $\mathcal{K}_{GY}$ suggère naturellement une mise en œuvre algorithmique via un schéma de relaxation, c'est-à-dire en alternant la minimisation selon des blocs de variables.

La forme la plus évidente consiste à suivre le partitionnement naturel en $\boldsymbol{x}$ et $\boldsymbol{b}$. Le schéma itératif ainsi défini de la mise à jour des itérées $(\boldsymbol{x}_k, \boldsymbol{b}_k)$ est de la forme

$$\boldsymbol{b}_{k+1} = \arg\min_{\boldsymbol{b}} \mathcal{K}(\boldsymbol{x}_k, \boldsymbol{b}) \tag{IV.19}$$

$$\boldsymbol{x}_{k+1} = \arg\min_{\boldsymbol{x}} \mathcal{K}(\boldsymbol{x}, \boldsymbol{b}_{k+1}). \tag{IV.20}$$

Ce schéma itératif est appelé LEGEND dans [Charbonnier *et al.*, 1994] lorsque $\mathcal{K} = \mathcal{K}_{GY}$. Il est appelé ARTUR dans [Charbonnier *et al.*, 1997] lorsque $\mathcal{K} = \mathcal{K}_{GR}$.

[B] Généralisation

Ce schéma peut être généralisé doublement. D'une part, en considérant un découpage en blocs quelconques des variables, sous réserve que la réunion des blocs contienne à la fois $\boldsymbol{x}$ et les variables auxiliaires $\boldsymbol{b}$. D'autre part en introduisant des coefficients de relaxation permettant d'obtenir des versions sous ou sur relaxées. Soit $\boldsymbol{y} = (\boldsymbol{x}, \boldsymbol{b})$ de taille $N + C$. Considérons le découpage en blocs de variables $\boldsymbol{y}^\ell$ de taille N_ℓ avec $\sum_{\ell=1}^{L} N_\ell = N + C$ tel que $\boldsymbol{y} = (\boldsymbol{y}^1, \ldots, \boldsymbol{y}^L)$. Ainsi, le schéma itératif (IV.19)-(IV.20) peut être généralisé sous la forme

$$\boldsymbol{y}^\ell_{k+1} = (1 - \theta_\ell)\boldsymbol{y}^\ell_k + \theta_\ell \arg\min_{\zeta} \mathcal{K}(\boldsymbol{y}^1_{k+1}, \ldots, \boldsymbol{y}^{\ell-1}_{k+1}, \zeta, \boldsymbol{y}^{\ell+1}_k, \ldots, \boldsymbol{y}^L_k) \tag{IV.21}$$

où $0 < \theta_\ell < 2$ si $\ell \leq N$ (sous ou sur-relaxation possible pour $\boldsymbol{x}$) et $0 < \theta_\ell \leq 1$ autrement (sous-relaxation possible pour les variables auxiliaires $\boldsymbol{b}$). Dans [Idier, 2001] des conditions larges portant sur la fonction de régularisation ϕ (dont la convexité) assurent la convexité des critères augmentés $\mathcal{K}$, d'où est tirée la convergence du schéma itératif (IV.21). Notons que la sur-relaxation ($\theta_\ell > 1$) des variables auxiliaires n'assure pas la convergence contrairement à celle des variables extraites de $\boldsymbol{x}$ [Idier, 2001].

L'intérêt de l'introduction de ces degrés de liberté, correspondant aux coefficients de relaxation θ_ℓ, réside dans le gain possible en terme de vitesse de convergence, par rapport à une version où ces coefficients sont fixés à un. Néanmoins, il a été constaté dans [Allain, 2002, p. 140] que la sous-relaxation des variables auxiliaires se révèle être d'un intérêt marginal. En pratique, la convergence est même ralentie. Ainsi, nous ne considérerons pas par la suite de sous-relaxation des variables auxiliaires : ces coefficients seront systématiquement fixés à un.

Les algorithmes ARTUR et LEGEND correspondent au cas où tous les coefficients de relaxation θ_ℓ sont fixés à un, y compris les coefficients correspondant aux variables $\boldsymbol{x}$. Nous désignerons par algorithmes de GR et de GY les formes étendues des algorithmes ARTUR et LEGEND, dans le sens où le coefficient de relaxation des variables $\boldsymbol{x}$ est un degré de liberté, non nécessairement fixé à un :

$$\begin{aligned}
\boldsymbol{b}_{k+1} &= \arg\min_{\boldsymbol{b}} \mathcal{K}(\boldsymbol{x}_k, \boldsymbol{b}) \\
\boldsymbol{x}_{k+1} &= (1 - \theta)\boldsymbol{x}_k + \theta \arg\min_{\boldsymbol{x}} \mathcal{K}(\boldsymbol{x}, \boldsymbol{b}_{k+1}).
\end{aligned}$$

IV.4.6 Interprétations des algorithmes semi-quadratiques

Les algorithmes semi-quadratiques de GR et de GY peuvent recevoir plusieurs interprétations possibles. Ces différentes interprétations laissent ainsi à penser que les algorithmes SQ sont des algorithmes profonds. Il est souligné dans [Idier, 2001] que les algorithmes SQ de GR et de GY sont aussi connus respectivement sous le nom de *Iterative reweighted least squares* (IRLS) et de *Residual steepest descent* (RSD) dans la communauté de la statistique robuste. Ainsi, ces différentes désignations recouvrent les mêmes structures algorithmiques, que nous pouvons résumer par "GR=IRLS=ARTUR" et "GY=RSD=LEGEND". Nous présentons ci-dessous deux interprétations possibles des algorithmes SQ. La première interprétation relie les algorithmes SQ à une famille classique d'algorithmes d'optimisation, les algorithmes de type quasi-Newton. La seconde interprétation rattache, dans un contexte probabiliste, les algorithmes SQ à la classe des algorithmes EM.

[A] Interprétation de type quasi-Newton des algorithmes SQ

Les algorithmes de GR et de GY peuvent être mis sous la forme d'algorithmes de quasi-Newton à pas fixe [Nikolova et Ng, 2005; Allain *et al.*, 2006]

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \theta \boldsymbol{d}_k \tag{IV.22}$$

$$\boldsymbol{d}_k = -\mathbf{A}(\boldsymbol{x}_k)^{-1} \nabla \mathcal{J}(\boldsymbol{x}_k) \tag{IV.23}$$

où la matrice normale définie positive de GR présente un caractère adaptatif par rapport à l'itérée courante et est définie par $\mathbf{A}(\boldsymbol{x}_k) := \mathbf{A}_{\mathrm{GR}}(\boldsymbol{x}_k)$ où

$$\mathbf{A}_{\mathrm{GR}}(\boldsymbol{x}) = 2\mathbf{H}^{\mathrm{t}}\mathbf{H} + \mathbf{V}^{\mathrm{t}}\mathbf{L}(\boldsymbol{x})\mathbf{V} \tag{IV.24}$$

$$\mathbf{L}(\boldsymbol{x}) = \mathrm{Diag}\{\phi'(\boldsymbol{v}_c^{\mathrm{t}}\boldsymbol{x} - \omega_c)/(\boldsymbol{v}_c^{\mathrm{t}}\boldsymbol{x} - \omega_c)\} \tag{IV.25}$$

et la matrice normale définie positive de GY est constante et définie par $\mathbf{A}(\boldsymbol{x}_k) := \mathbf{A}^a_{\mathrm{GY}}$ où

$$\mathbf{A}^a_{\mathrm{GY}} = 2\mathbf{H}^{\mathrm{t}}\mathbf{H} + \mathbf{V}^{\mathrm{t}}\mathbf{V}/a. \tag{IV.26}$$

Les matrices $\mathbf{A}_{\mathrm{GR}}(\boldsymbol{x})$ et $\mathbf{A}^a_{\mathrm{GY}}$ s'identifient avec les matrices des équations normales (IV.14) et (IV.18). Notons que les variables auxiliaires $\boldsymbol{b}$ n'apparaissent plus explicitement dans le schéma itératif (IV.22)-(IV.23). Le lecteur trouvera par exemple dans [Allain, 2002, p. 140] le détail des calculs amenant à la forme de type quasi-Newton à pas fixe en partant de la minimisation par relaxation groupée en $\boldsymbol{x}$ et $\boldsymbol{b}$ des critères augmentés $\mathcal{K}_{GR}$ et $\mathcal{K}_{GY}$.

Une définition forte des algorithmes de quasi-Newton impose que la matrice $\mathbf{M}_k$ (IV.9), page 60 vérifie l'équation sécante suivante [Nocedal et Wright, 1999, p. 195]

$$\mathbf{M}_{k+1}(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k) = \nabla \mathcal{J}(\boldsymbol{x}_{k+1}) - \nabla \mathcal{J}(\boldsymbol{x}_k). \tag{IV.27}$$

L'algorithme BFGS vérifie l'équation sécante (IV.27), lorsque le critère est strictement convexe. A notre connaissance, la question de savoir si les algorithmes SQ vérifient également l'équation de la sécante n'a pas été abordée dans la littérature. Nous avons constaté numériquement que les algorithmes SQ ne vérifient pas l'équation sécante pour un critère pénalisé strictement convexe non quadratique. Ceci indique que les algorithmes SQ ne sont pas quasi-Newton au sens fort. Néanmoins, ce résultat est à relativiser car les algorithmes de type BFGS à mémoire limitée (l-BFGS) [Nocedal et Wright, 1999, Sec. 9.1] ne vérifient pas non plus, à notre connaissance, l'équation sécante. Nous utilisons donc dans ce document la définition plus faible d'algorithme de quasi-Newton au sens où la direction tend à approcher la direction de Newton [Bertsekas, 1999, p. 149].

[B] Interprétation en terme d'algorithmes EM

Nous avons vu dans la section III.2.2 du chapitre III que les algorithmes SQ peuvent être vus comme maximiseurs du MAP dans un contexte probabiliste. Il est montré dans [Champagnat et Idier, 2004] que les algorithmes SQ peuvent être rattachés à la classe des algorithmes EM. Plus précisément, il est montré que l'estimateur du MAP avec une forme adaptée de l'algorithme EM revient à minimiser le critère pénalisé selon une approche SQ.

L'analyse des algorithmes EM peut être basée sur leur propriété de minimiseur d'un critère de substitution majorant le critère initial [Lange *et al.*, 2000]. Nous nous intéressons dans la partie suivante à la propriété de majoration du critère pénalisé des constructions SQ. Cette propriété sera une des clés de voûte des résultats de convergence établis dans les deux prochains chapitres.

IV.4.7 Approximation quadratique majorante

Le principe des algorithmes SQ consiste à substituer un problème plus simple au problème initial de la minimisation du critère pénalisé. Ce type d'approche s'appuie sur la connaissance de la structure analytique du critère pénalisé. Les algorithmes SQ ont été initialement introduits dans [Geman et Reynolds, 1992; Geman et Yang, 1995] comme minimiseur d'un critère équivalent augmenté $\mathcal{K}(\boldsymbol{x}, \boldsymbol{b})$ du critère pénalisé $\mathcal{J}(\boldsymbol{x})$. Cette approche fait intervenir des variables auxiliaires $\boldsymbol{b}$ en sus des variables $\boldsymbol{x}$. L'intérêt du critère équivalent augmenté étant que ce nouveau problème peut être résolu de manière analytique, contrairement au problème initial.

Dans cette section nous montrons qu'on peut interpréter les algorithmes SQ comme minimiseur d'un critère de substitution plus générique que le critère équivalent augmenté $\mathcal{K}(\boldsymbol{x}, \boldsymbol{b})$. Dans [Chan et Mulet, 1999] il est montré que l'algorithme GR peut être analysé comme un algorithme minimisant un critère de substitution quadratique. Un tel résultat est aussi obtenu pour l'algorithme GY [Allain *et al.*, 2006]. Cette approche par critère de substitution quadratique est d'une part plus générique que l'utilisation du critère équivalent augmenté $\mathcal{K}(\boldsymbol{x}, \boldsymbol{b})$. D'autre part, on la trouve dans les publications [Weiszfeld, 1937; Voss et Eckhardt, 1980] et [Huber, 1981] (dans le cadre de la régression robuste) qui sont antérieures à [Geman et Reynolds, 1992; Geman et Yang, 1995].

Les approximations quadratiques majorantes existantes dans la littérature sont définies pour les critères pénalisés, non généralisés. Notre contribution dans cette partie consiste à montrer qu'il est possible d'étendre la construction de telles approximations quadratiques majorantes aux critères pénalisés généralisés. L'intérêt de ces approximations quadratiques majorantes ne réside pas seulement dans une nouvelle interprétation des algorithmes SQ. Nous verrons par la suite qu'ils permettront d'établir les résultats de convergence des deux prochains chapitres.

[A] Préliminaires

Nous commençons par définir la notion d'approximation quadratique ainsi que celle d'opérateur à spectre uniformément positif borné.

Définition IV.4.2. *Soit $\mathcal{J} : \mathbb{R}^N \mapsto \mathbb{R}$ un critère C^1. On définit l'approximation quadratique de $\mathcal{J}$ en $\boldsymbol{x}$ par*

$$\hat{J}_{\mathbf{A}}(\boldsymbol{x}^+, \boldsymbol{x}) = \mathcal{J}(\boldsymbol{x}) + (\boldsymbol{x}^+ - \boldsymbol{x})^{\mathrm{t}} \nabla \mathcal{J}(\boldsymbol{x}) + (\boldsymbol{x}^+ - \boldsymbol{x})^{\mathrm{t}} \mathbf{A}(\boldsymbol{x}) (\boldsymbol{x}^+ - \boldsymbol{x})/2 \tag{IV.28}$$

où $\mathbf{A} : \mathbb{R}^N \mapsto \mathbb{R}^{N \times N}$ est un opérateur défini positif.

Définition IV.4.3. *Un opérateur défini positif $\mathbf{A} : \mathbb{R}^N \mapsto \mathbb{R}^{N \times N}$ est dit à spectre uniformément positif borné s'il existe $\nu_1, \nu_2 \in \mathbb{R}$ avec $\nu_2 \geq \nu_1 > 0$ tels que*

$$\nu_1 \|\boldsymbol{v}\|^2 \leq \boldsymbol{v}^{\mathrm{t}} \mathbf{A}(\boldsymbol{u}) \boldsymbol{v} \leq \nu_2 \|\boldsymbol{v}\|^2, \quad \forall \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^N. \tag{IV.29}$$

[B] Résultats pour les critères pénalisés

Cette partie récapitule certains résultats présentés dans [Allain *et al.*, 2006] qui restent valides lorsque l'hypothèse ϕ convexe et coercive est relaxée en ϕ coercive. Nous commençons à montrer que les opérateurs définis à partir des matrice normales SQ sont à spectre uniformément positif borné.

Lemme IV.4.2. *Supposons que l'Hypothèse 1 (resp. Hypothèse 2) est satisfaite. Supposons aussi que les matrices $\mathbf{H}$ et $\mathbf{V}$ sont telles que*

$$\ker(\mathbf{H}^{\mathrm{t}}\mathbf{H}) \cap \ker(\mathbf{V}^{\mathrm{t}}\mathbf{V}) = \{\mathbf{0}\}. \tag{IV.30}$$

Alors l'opérateur $\mathbf{A} := \{\mathbf{A}_{GY}^{a}\}$ *avec* $0 < a < \hat{a}$ *(resp.* $\mathbf{A} := \{\mathbf{A}_{GR}(.)\}$*) est à spectre uniformément positif borné.*

Preuve. D'après (IV.30), la preuve est immédiate pour $\mathbf{A} := \{\mathbf{A}_{\mathrm{GY}}^{a}\}$ avec $0 < a$, car $\mathbf{A}_{\mathrm{GY}}^{a}$ est alors une matrice symétrique définie positive.

D'après [Allain *et al.*, 2006, Prop. 8], $\mathbf{A} := \{\mathbf{A}_{\mathrm{GR}}(.)\}$ est aussi à spectre uniformément positif borné. □

Le Lemme suivant établit que les matrices normales issues des constructions de GY et de GR induisent des approximations quadratiques majorantes pour le critère pénalisé défini par (IV.12) page 62.

Lemme IV.4.3. *Soit* $\mathcal{J}$ *le critère pénalisé défini par* (IV.12)*. Supposons que l'Hypothèse 1 (resp. Hypothèse 2) est satisfaite.*

Alors $\hat{J}_{\mathbf{A}}$ *est une approximation quadratique majorante du critère* $\mathcal{J}$ *:*

$$\hat{J}_{\mathbf{A}}(\boldsymbol{x}^{+}, \boldsymbol{x}) \geq \mathcal{J}(\boldsymbol{x}^{+}), \quad \forall \boldsymbol{x}, \boldsymbol{x}^{+} \in \mathbb{R}^{N} \tag{IV.31}$$

où $\mathbf{A} := \{\mathbf{A}_{GY}^{a}\}$ *avec* $0 < a < \hat{a}$ *(resp.* $\mathbf{A} := \{\mathbf{A}_{GR}(.)\}$*).*

Preuve. D'après [Allain *et al.*, 2006, Prop. 1], le caractère majorant de l'approximation quadratique $\hat{J}_{\mathbf{A}}$ est vérifié avec $\mathbf{A} := \{\mathbf{A}_{\mathrm{GY}}^{a}\}$ pour $0 < a < \hat{a}$ (resp. $\mathbf{A} := \{\mathbf{A}_{\mathrm{GR}}(.)\}$). Notons que dans [Allain *et al.*, 2006, Prop. 1 et Prop. 8], l'hypothèse ϕ convexe peut être relaxée en ϕ coercive. □

Ces approximations quadratiques majorantes permettent de donner une nouvelle interprétation des algorithmes SQ. Considérons la version relaxée de la méthode de Weiszfeld [Weiszfeld, 1937; Voss et Eckhardt, 1980] définie par le schéma itératif suivant

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_{k} + \theta\,(\hat{\boldsymbol{x}}_{k+1} - \boldsymbol{x}_{k}) \tag{IV.32}$$

où θ est une constante et avec

$$\begin{aligned} \hat{\boldsymbol{x}}_{k+1} &= \arg\min_{\boldsymbol{x}} \hat{J}_{\mathbf{A}}(\boldsymbol{x}, \boldsymbol{x}_{k}) \\ \hat{\boldsymbol{x}}_{k+1} &= \boldsymbol{x}_{k} - \mathbf{A}(\boldsymbol{x}_{k})^{-1} \nabla \mathcal{J}(\boldsymbol{x}_{k}) \end{aligned} \tag{IV.33}$$

car $\hat{J}_{\mathbf{A}}(., \boldsymbol{x}_{k})$ est la quadratique définie par (IV.28) page 67. Ainsi, la structure des algorithmes SQ sous la forme (IV.22)-(IV.23) page 66 est la même que celle du schéma itératif (IV.32)-(IV.33). Les algorithmes SQ peuvent donc être interprétés comme résultant de minimisations successives d'approximations quadratiques majorantes.

[C] Résultats pour les critères pénalisés généralisés

Dans [Chan et Mulet, 1999, Sec. 3], le terme de pénalisation suivant est considéré

$$\Psi(\boldsymbol{x}) = \sum_{c=1}^{C} \psi(|\mathbf{V}_{c}\boldsymbol{x}|_{\beta}) \tag{IV.34}$$

avec $|\boldsymbol{u}|_{\beta} = \sqrt{\beta + \|\boldsymbol{u}\|^{2}}$ et $\mathbf{V}_{c} \in \mathbb{R}^{2\times N}$.

En remarquant que $\psi(|\mathbf{V}_{c}\boldsymbol{x}|_{\beta}) = \phi(\|\boldsymbol{u}\|)$ où $\phi(t) = \psi(\sqrt{\beta + t^{2}})$, ce terme de pénalisation (IV.34) s'identifie au terme de pénalisation généralisé (IV.2), page 54 pour $\boldsymbol{\omega}_{c} = \mathbf{0}$ et $P = 2$.

Dans [Chan et Mulet, 1999, Sec. 4], une approximation quadratique majorante issue de la matrice normale de GR généralisée pour $P = 2$ est considérée. Dans cette section, nous proposons d'abord de généraliser ce résultat au cas $P \in \mathbb{N}^*$ et sous des hypothèses moins restrictives (sans la convexité de ϕ ni le caractère C^2). Par contre, à notre connaissance, il n'existe pas dans la littérature d'approximation quadratique majorante issue de la matrice normale de GY pour les critères pénalisés généralisés. Nous montrons également dans cette section qu'il est possible d'obtenir naturellement une telle construction.

Considérons le critère généralisé (IV.1) avec la pénalisation (IV.2). Il nous faut d'abord définir les constructions SQ généralisées associées au critère pénalisé généralisé. Soit $\mathbf{V}$ la matrice définie par (IV.3). En utilisant la matrice $\mathbf{V}$, il est possible d'obtenir des généralisations naturelles des matrices normales SQ. Pour la construction de GY on obtient la même structure matricielle (IV.26) que pour un critère pénalisé non généralisé

$$\mathbf{A}^a_{\mathrm{GY}} = 2\mathbf{H}^{\mathrm{t}}\mathbf{H} + \mathbf{V}^{\mathrm{t}}\mathbf{V}/a. \tag{IV.35}$$

La différence reposant sur le fait que la matrice $\mathbf{V}$ est de taille $CP \times N$ dans le cas d'un critère pénalisé généralisé alors qu'elle est de taille $C \times N$ dans le cas non généralisé.

Dans [Chan et Mulet, 1999], la matrice suivante est considérée :

$$\begin{aligned} \mathbf{C}(\boldsymbol{x}) &= 2\mathbf{H}^{\mathrm{t}}\mathbf{H} + \mathbf{V}^{\mathrm{t}}\mathbf{L}(\boldsymbol{x})\mathbf{V} \\ \mathbf{L}(\boldsymbol{x}) &= \mathrm{Diag}\left\{ \left\{ \frac{\psi'(|\mathbf{V}_c\boldsymbol{x}|_\beta)}{|\mathbf{V}_c\boldsymbol{x}|_\beta}\mathbf{I}_2 \right\}_c \right\} \end{aligned} \tag{IV.36}$$

En utilisant $\phi(t) = \psi(\sqrt{\beta + t^2})$, on a

$$\phi'(t) = \frac{t}{\sqrt{\beta + t^2}}\psi'(\sqrt{\beta + t^2})$$

d'où

$$\frac{\phi'(\|\mathbf{V}_c\boldsymbol{x}\|)}{\|\mathbf{V}_c\boldsymbol{x}\|} = \frac{\psi'(|\mathbf{V}_c\boldsymbol{x}|_\beta)}{|\mathbf{V}_c\boldsymbol{x}|_\beta}.$$

Ainsi, (IV.36) correspond à une généralisation de la matrice normale de GR pour $P = 2$.

Nous proposons donc la structure suivante pour la matrice normale de GR généralisée

$$\begin{aligned} \mathbf{A}_{\mathrm{GR}}(\boldsymbol{x}) &= 2\mathbf{H}^{\mathrm{t}}\mathbf{H} + \mathbf{V}^{\mathrm{t}}\mathbf{L}(\boldsymbol{x})\mathbf{V} \\ \mathbf{L}(\boldsymbol{x}) &= \mathrm{Diag}\left\{ \left\{ \frac{\phi'(\|\boldsymbol{\delta}_c\|)}{\|\boldsymbol{\delta}_c\|}\mathbf{I}_p \right\}_c \right\} \\ \boldsymbol{\delta}_c &= \mathbf{V}_c\boldsymbol{x} - \boldsymbol{\omega}_c. \end{aligned} \tag{IV.37}$$

Lorsque $P = 1$ on retrouve bien la matrice normale de GR (IV.24) page 66 non généralisée. Par la suite, lorsqu'on parlera de constructions SQ on fera référence au cas généralisé.

Il est immédiat d'établir que le Lemme IV.4.2 page 67 qui établit le caractère spectre uniformément positif borné s'applique aussi aux opérateurs SQ généralisés. A présent, on établit le caractère majorant des approximations quadratiques définies à partir des opérateurs SQ généralisés. Dans le cas de l'opérateur de GR généralisé la démonstration du résultat est une adaptation assez directe de celle de [Chan et Mulet, 1999].

Lemme IV.4.4. *Soit $\mathcal{J}$ le critère pénalisé généralisé défini par (IV.1). Supposons que l'Hypothèse 2 est satisfaite.*

Alors $\hat{J}_{\mathbf{A}_{GR}}$ est une approximation quadratique majorante du critère $\mathcal{J}$:

$$\hat{J}_{\mathbf{A}_{GR}}(\boldsymbol{x}^+, \boldsymbol{x}) \geq \mathcal{J}(\boldsymbol{x}^+), \quad \forall \boldsymbol{x}, \boldsymbol{x}^+ \in \mathbb{R}^N.$$

Preuve. Voir Annexe B.1.2, page 143. □

Dans le cas de l'opérateur de GY généralisé la démonstration du caractère majorant de l'approximation quadratique associée fait appel au caractère gradient Lipschitz du critère $\phi(\|.\|) : \mathbb{R}^N \mapsto \mathbb{R}$ et non pas seulement de la fonction ϕ. C'est l'objet du prochain Lemme.

Lemme IV.4.5. *Supposons que l'Hypothèse 1 est satisfaite.*
Alors $\phi(\|.\|) : \mathbb{R}^N \mapsto \mathbb{R}$ *est* L-$\mathcal{LC}^1$ *avec* $L = 1/\hat{a}$.

Preuve. Voir Annexe B.1.3, page 144. □

Lemme IV.4.6. *Soit* $\mathcal{J}$ *le critère pénalisé généralisé défini par* (IV.1). *Supposons que l'Hypothèse 1 est satisfaite.*
Alors $\hat{J}_{\mathbf{A}_{GY}}$ *est une approximation quadratique majorante du critère* $\mathcal{J}$ *:*

$$\hat{J}_{\mathbf{A}^a_{GY}}(\boldsymbol{x}^+, \boldsymbol{x}) \geq \mathcal{J}(\boldsymbol{x}^+), \quad \forall \boldsymbol{x}, \boldsymbol{x}^+ \in \mathbb{R}^N$$

avec $0 < a < \hat{a}$.

Preuve. La démonstration est similaire à celle de [Allain *et al.*, 2006, Appendix A] pour un critère pénalisé non généralisé. Dans le cas du critère pénalisé généralisé, d'après le Lemme IV.4.5 on peut utiliser le Lemme de descente [Bertsekas, 1999, Prop. A.24] pour $\phi(\|.\|)$. □

IV.5 Conclusion

On a vu que les algorithmes semi-quadratiques sont spécifiquement adaptés pour la minimisation des critères pénalisés. Ils tirent leur force de la prise en compte de la structure analytique du critère à minimiser, contrairement à des approches d'optimisation généralistes. Ils souffrent malheureusement d'un problème crucial. Comme nous l'exposons au chapitre suivant, ils sont généralement trop coûteux pour des problèmes de grande taille, tels que ceux couramment rencontrés en image. Ce constat nous amène à étudier deux familles d'algorithmes, contenant des ingrédients semi-quadratiques, permettant de s'affranchir des limitations de taille. Ce sera le sujet des deux prochains chapitres.

Bibliographie

[Allain, 2002] M. Allain. *Approche pénalisée en tomographie hélicoïdale. Application à la conception d'une prothèse personnalisée du genou.* Co-supervised PhD thesis, Université de Paris-Sud, Orsay, France / Ecole Polytechnique de Montréal, QC, Canada, Dec. 2002.

[Allain *et al.*, 2006] M. Allain, J. Idier et Y. Goussard. On global and local convergence of half-quadratic algorithms. *IEEE Trans. Image Processing*, 15 (5) : 1130–1142, May 2006.

[Aubert et Vese, 1997] G. Aubert et L. Vese. A variational method in image recovery. *SIAM J. Num. Anal.*, 34 (5) : 1948–1979, Oct. 1997.

[Bertsekas, 1999] D. P. Bertsekas. *Nonlinear programming.* Athena Scientific, Belmont, MA, 2nd edition, 1999.

[Blake et Zisserman, 1987] A. Blake et A. Zisserman. *Visual reconstruction.* The MIT Press, Cambridge, MA, 1987.

[Bouman et Sauer, 1996] C. A. Bouman et K. D. Sauer. A unified approach to statistical tomography using coordinate descent optimization. *IEEE Trans. Image Processing*, 5 (3) : 480–492, Mar. 1996.

[Champagnat et Idier, 2004] F. Champagnat et J. Idier. A connection between half-quadratic criteria and EM algorithms. *IEEE Signal Processing Letters*, 11 (9) : 709–712, Sep. 2004.

[Chan et Mulet, 1999] T. F. Chan et P. Mulet. On the convergence of the lagged diffusivity fixed point method in total variation image restoration. *SIAM J. Num. Anal.*, 36 (2) : 354–367, 1999.

[Charbonnier *et al.*, 1994] P. Charbonnier, L. Blanc-Féraud, G. Aubert et M. Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. In *Proc. IEEE ICIP*, volume 2, pages 168–172, Austin, TX, Nov. 1994.

[Charbonnier *et al.*, 1997] P. Charbonnier, L. Blanc-Féraud, G. Aubert et M. Barlaud. Deterministic edge-preserving regularization in computed imaging. *IEEE Trans. Image Processing*, 6 (2) : 298–311, Feb. 1997.

[Ciuciu et Idier, 2002] P. Ciuciu et J. Idier. A half-quadratic block-coordinate descent method for spectral estimation. *Signal Processing*, 82 (7) : 941–959, July 2002.

[Delaney et Bresler, 1998] A. H. Delaney et Y. Bresler. Globally convergent edge-preserving regularized reconstruction : an application to limited-angle tomography. *IEEE Trans. Image Processing*, 7 (2) : 204–221, Feb. 1998.

[Erdogan et Fessler, 1999] H. Erdogan et J. Fessler. Monotonic algorithms for transmission tomography. *IEEE Trans. Medical Imaging*, 18 (9) : 801–814, Sep. 1999.

[Geman et Reynolds, 1992] D. Geman et G. Reynolds. Constrained restoration and the recovery of discontinuities. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14 (3) : 367–383, Mar. 1992.

[Geman et Yang, 1995] D. Geman et C. Yang. Nonlinear image recovery with half-quadratic regularization. *IEEE Trans. Image Processing*, 4 (7) : 932–946, July 1995.

[Geman et Geman, 1984] S. Geman et D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, PAMI-6 (6) : 721–741, Nov. 1984.

[Hestenes et Stiefel, 1952] M. R. Hestenes et E. Stiefel. Methods of conjugate gradients for solving linear system. *J. Res. Nat. Bur. Stand.*, 49 : 409–436, 1952.

[Huber, 1981] P. J. Huber. *Robust Statistics*. John Wiley, New York, NY, 1981.

[Idier, 2001] J. Idier. Convex half-quadratic criteria and interacting auxiliary variables for image restoration. *IEEE Trans. Image Processing*, 10 (7) : 1001–1009, July 2001.

[Kaufman, 1993] L. Kaufman. Maximum likelihood, least squares, and penalized least squares for PET. *IEEE Trans. Medical Imaging*, 12 : 200–214, June 1993.

[Lange *et al.*, 2000] K. Lange, D. R. Hunter et I. Yang. Optimization transfer using surrogate objective functions (with discussion). *J. Comput. Graph. Statist.*, 9 (1) : 1–20, Mar. 2000.

[Mackens et Voss, 2000] W. Mackens et H. Voss. Computing the minimum eigenvalue of a symmetric positive definite toeplitz matrix by newton type methods. *SIAM J. Sci. Comput.*, 21 : 1650–1656, 2000.

[Nikolova, 1999] M. Nikolova. Markovian reconstruction using a GNC approach. *IEEE Trans. Image Processing*, 8 (9) : 1204–1220, Sep. 1999.

[Nikolova *et al.*, 1998] M. Nikolova, J. Idier et A. Mohammad-Djafari. Inversion of large-support ill-posed linear operators using a piecewise Gaussian MRF. *IEEE Trans. Image Processing*, 7 (4) : 571–585, Apr. 1998.

[Nikolova et Ng, 2005] M. Nikolova et M. Ng. Analysis of half-quadratic minimization methods for signal and image recovery. *SIAM J. Sci. Comput.*, 27 : 937–966, 2005.

[Nocedal et Wright, 1999] J. Nocedal et S. J. Wright. *Numerical optimization*. Springer Texts in Operations Research. Springer-Verlag, New York, NY, 1999.

[Rockafellar, 1970] R. T. Rockafellar. *Convex Analysis*. Princeton Univ. Press, 1970.

[Tseng et Bertsekas, 1987] P. Tseng et D. P. Bertsekas. Relaxation methods for problems with strictly convex separable costs and linear constraints. *Mathematical Programming*, 38 : 303–321, 1987.

[Voss et Eckhardt, 1980] H. Voss et U. Eckhardt. Linear Convergence of Generalized Weiszfeld's Method. *Computing*, 25 : 243–251, 1980.

[Weiszfeld, 1937] E. Weiszfeld. Sur le point pour lequel la somme des distances de n points donnés est minimum. *Tôhoku Mathematical Journal*, 43 : 355–386, 1937.

Chapitre V

ALGORITHMES SEMI-QUADRATIQUES APPROCHÉS

Nous avons vu dans le chapitre précédent que les algorithmes semi-quadratiques de GR et de GY sont des algorithmes séduisants à plusieurs titres pour minimiser les critères pénalisés. Ce sont des algorithmes à la fois simples et adaptés puisqu'ils mettent à profit la connaissance analytique des critères pénalisés à minimiser. En somme, ils semblent presque être parfaits pour minimiser de tels critères. Néanmoins, la simplicité structurelle des algorithmes de GR et de GY cache une complexité calculatoire importante pour les problèmes de grande taille. Ceci découle de la nécessité de résoudre à chaque itération une équation normale faisant intervenir une matrice normale semi-quadratique dont la taille est directement reliée au nombre d'inconnues. En pratique, la résolution exacte de cette équation normale est généralement une opération très coûteuse pour les problèmes de grande taille. Le temps de calcul par itération des algorithmes de GR et de GY est alors très important, ce qui finit par contre-balancer le bon taux de convergence de ces algorithmes en terme de nombres d'itérations. On ne peut donc généralement pas mettre en œuvre les algorithmes de GR et de GY tels quels pour des problèmes de grande taille.

Face à ce constat, il a été proposé d'approcher les algorithmes de GR et de GY sous une forme dont le coût d'implémentation est bien plus faible. L'inconvénient de ces formes approchées est que les résultats de convergence valables pour les formes exactes de GR et de GY ne s'appliquent plus aux formes approchées, jusqu'à présent. Il n'existe pas, à notre connaissance, de preuve de convergence des formes approchées de GR et de GY. La contribution de ce chapitre consiste précisément à établir la convergence de certaines formes d'approximation de GY et de GR, qui ont déjà été proposées et utilisées en pratique dans la communauté du traitement de l'image.

V.1 Difficultés de l'inversion des matrices semi-quadratiques

On rappelle que la mise en œuvre des algorithmes itératifs semi-quadratiques (SQ) de GR et de GY fait intervenir à chaque itération k la résolution du système linéaire suivant afin de déterminer la direction de descente courante $\boldsymbol{d}_k$:

$$\boldsymbol{d}_k = -\mathbf{A}(\boldsymbol{x}_k)^{-1}\nabla\mathcal{J}(\boldsymbol{x}_k) \tag{V.1}$$

où $\mathbf{A}(\boldsymbol{x}_k)$ est une matrice définie positive et correspond soit à la matrice normale constante de GY ($\mathbf{A}(\boldsymbol{x}_k) = \mathbf{A}^a_{\mathrm{GY}}$ définie par (IV.26) page 66), soit à la matrice normale de GR ($\mathbf{A}_k = \mathbf{A}_{\mathrm{GR}}(\boldsymbol{x}_k)$ définie par (IV.24) page 66). Notons que la complexité de la résolution du système linéaire (V.1) est toujours supérieure à $O(N)$, sauf pour certains cas triviaux.

La résolution du système linéaire (V.1) est de loin l'opération la plus coûteuse de la mise en œuvre des algorithmes de GR et de GY. En effet, la mise à jour de l'itérée $\boldsymbol{x}_k$:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \theta\boldsymbol{d}_k$$

où le pas θ constant, est de complexité $O(N)$ qui est négligeable devant celle liée à la résolution du système linéaire (V.1). Ainsi, toute la complexité des algorithmes de GR et de GY est portée par cette résolution. Il est donc crucial d'investir sur le problème de résolution du système linéaire (V.1) définissant la direction de descente courante des algorithmes de GR et de GY. Dans cette section, on se propose de faire un tour d'horizon des méthodes et des situations permettant de résoudre le système linéaire (V.1) de manière exacte. On montre que le coût de cette résolution exacte est en général trop important. Dans la section suivante, on recherche une méthode itérative permettant de résoudre (V.1) de manière approchée, en beaucoup moins de N itérations (N est trop grand dans le cas des images).

Plusieurs méthodes classiques d'analyse numérique on été envisagées pour la résolution exacte du système linéaire (V.1), telles que la méthode du pivot de Gauss et celle de la décomposition de Cholesky [Nikolova et Ng, 2005]. Le lecteur trouvera par exemple dans [Golub et Van Loan, 1996] le détail de ces méthodes classiques. La complexité de ces méthodes de résolution du système linéaire (V.1) est comparée dans [Nikolova et Ng, 2005, Sec. 4.3] pour les algorithmes de GR et de GY. La méthode du pivot de Gauss a pour complexité $O(N^3)$. Si cette méthode peut être envisagée lorsque la taille du problème est peu importante, elle devient vite d'un coût prohibitif dans le cas d'un problème de grande taille. La méthode du pivot de Gauss n'est donc pas envisageable en pratique pour résoudre les systèmes linéaires induits par les algorithmes de GR et de GY lorsqu'un problème de grande taille est à traiter.

Une distinction nette est établie dans [Nikolova et Ng, 2005] entre les algorithmes de GY et de GR. En effet, la matrice normale de GR dépend de l'itérée courante $\boldsymbol{x}_k$ contrairement à la matrice normale de GY qui est constante. Ainsi, une méthode de résolution du système linéaire (V.1) s'appuyant sur un prétraitement de la matrice normale, telle que la décomposition de Cholesky, peut être envisagée pour l'algorithme de GY. Le prétraitement de la décomposition de Cholesky, qui consiste en la factorisation en matrices triangulaires, a pour complexité $O(N^3)$ dans le cas où la matrice normale est pleine. Il s'agit de la même complexité que pour la méthode du pivot de Gauss. Par contre, la résolution du système triangulaire ainsi obtenu a pour complexité $O(N^2)$. La matrice normale de GY étant constante, ce prétraitement n'est à faire qu'une seule fois et reste valide pour toutes les itérations. Le coût global de ce prétraitement est donc à diviser par le nombre d'itérations totales. Par contre, la matrice normale de GR n'étant pas constante, un tel prétraitement n'est pas envisageable car il serait à effectuer à chaque itération.

On peut aussi tirer parti de la structure induite par certains problèmes sur la matrice de GY. Lorsque la matrice d'observation $\mathbf{H}$ présente une structure de type Toeplitz ou Toeplitz par blocs Toeplitz, par exemple dans le cadre d'un problème de déconvolution, elle peut être, sous

certaines hypothèses de bords, diagonalisée à faible coût par transformée rapide [Ng *et al.*, 1999]. Lorsque $\mathbf{V}$ est la matrice des différences finies du premier ou deuxième ordre, la structure de type Toeplitz de la matrice d'observation $\mathbf{H}$ peut être étendue à la matrice normale de GY $\mathbf{A}_{\mathrm{GY}}^{a}$. La résolution du système linéaire de la matrice normale de GY peut alors s'effectuer efficacement avec une complexité en $O(N \ln N)$ [Ng *et al.*, 1999]. Malheureusement, la matrice normale de GR ne présente pas de structure de type Toeplitz à cause de la dépendance à l'itérée courante $\boldsymbol{x}_k$. Le caractère éventuel Toeplitz ou Toeplitz par blocs Toeplitz de la matrice d'observation $\mathbf{H}$ ne peut donc pas être exploité pour l'algorithme de GR.

La distinction nette entre les algorithmes de GY et de GR, faite dans [Nikolova et Ng, 2005], réside dans le coût de résolution du système linéaire associé à l'algorithme de GR bien plus important que pour l'algorithme de GY. C'est une des raisons expliquant l'investissement porté sur l'algorithme de GY [Allain *et al.*, 2006]. Si la résolution exacte du système linéaire peut être envisagée pour l'algorithme de GY dans le cas de certains problèmes de grande taille, ce n'est pas le cas de l'algorithme de GR car cette résolution exacte est trop coûteuse. On est donc systématiquement contraint d'employer une forme approchée de l'algorithme de GR pour les problèmes de grande taille. La section suivante porte précisément sur les formes approchées des algorithmes de GY et de GR, qui ont pour intérêt d'avoir un coût par itération considérablement réduit par rapport aux formes exactes.

V.2 Inversion approchée des matrices semi-quadratiques

On a vu que la résolution exacte du système linéaire déterminant la direction de descente courante est le point faible des algorithmes de GR et de GY, car il s'agit d'une opération généralement très coûteuse pour les problèmes de grande taille. Il a donc été naturellement proposé d'employer des méthodes approchant la résolution du système linéaire, mais d'un coût bien plus faible. Ce type d'approche consiste à employer une méthode itérative tronquée, dans le sens d'un faible nombre d'itérations vis-à-vis de la dimension N du problème, afin de résoudre de manière approchée le système linéaire induit par les algorithmes de GR et de GY. Il est à noter que cette approche a été proposée dès l'introduction de l'algorithme de GR par l'utilisation de l'algorithme de Gauss-Seidel tronqué [Charbonnier *et al.*, 1997]. Comme précisé dans [Charbonnier *et al.*, 1997], il est possible d'utiliser l'algorithme du gradient conjugué pour résoudre de manière approchée le système linéaire des algorithmes de GR et de GY. Cette approche a été mise en œuvre dans [Nikolova et Ng, 2001, 2005] pour l'algorithme de GY et dans [Belge *et al.*, 2000, Sect. IV p. 601] et [Nikolova et Ng, 2005] pour l'algorithme de GR.

Il est important de noter que les résultats de convergence des algorithmes de GR et de GY exacts [Charbonnier *et al.*, 1997; Delaney et Bresler, 1998; Idier, 2001; Allain *et al.*, 2006] ne s'étendent pas tels quels aux formes approchées. L'idée d'utiliser des formes approchées des algorithmes de GR et de GY, dictée par la nécessité pratique, consiste en somme à espérer qu'une approximation suffisamment proche de la forme exacte assure la convergence. Néanmoins, la convergence de ces formes approchées n'est pas, en toute rigueur, établie si on s'appuie uniquement sur la convergence des formes exactes.

De plus, la question de la détermination du degré d'approximation des formes approchées des algorithmes de GR et de GY se pose. En effet, en ayant conscience du problème de convergence des formes approchées, il est naturel d'essayer d'approcher au mieux les formes exactes. Cela peut être mis en œuvre en pratique en déterminant le nombre d'itérations de l'algorithme résolvant de manière approchée le système linéaire des algorithmes de GR et de GY de telle sorte que l'écart entre la forme exacte et la forme approchée soit inférieur à un seuil suffisamment faible [Nikolova et Ng, 2001]. Cependant, le choix de ce seuil nous semble être une réelle difficulté dans le cadre d'une telle approche. En effet, plus ce seuil est faible et plus l'approximation est proche

de la forme exacte. En contrepartie, le nombre d'itérations à effectuer est d'autant plus grand, et donc le coût plus important, que ce seuil est faible. On voit que le réglage de ce seuil nous confronte à deux objectifs antagonistes. L'un visant à être proche de la forme exacte et l'autre visant à être peu coûteux.

Cette difficulté liée à l'existence de résultats de convergence seulement pour les formes exactes, disparaît dès lors que la convergence d'une forme approchée est assurée. Il n'y a alors plus à choisir en fonction de deux objectifs contradictoires : l'écart entre forme approchée et forme exacte peut alors être choisi uniquement en fonction de l'objectif de coût. C'est précisément une des motivations du résultat de convergence des formes approchées des algorithmes de GR et de GY par gradient conjugué tronqué présenté dans la suite du chapitre. Ce résultat de convergence est d'autant plus intéressant que cette forme approchée des algorithmes de GR et de GY est déjà employée en traitement de l'image [Nikolova et Ng, 2001, 2005].

V.2.1 Algorithme du gradient conjugué linéaire

L'algorithme du gradient conjugué linéaire [Hestenes et Stiefel, 1952] figure parmi la liste des dix meilleurs algorithmes du siècle précédent selon la revue SIAM News [Cipra, 2000]. Une telle liste est toujours discutable, mais elle laisse entendre qu'il s'agit d'un des algorithmes majeurs de l'analyse numérique. Il existe des versions du gradient conjugué non linéaire que nous étudierons dans le chapitre suivant. Ce chapitre est exclusivement consacré à l'algorithme du gradient conjugué linéaire. Par la suite nous omettrons le terme linéaire pour la désignation de l'algorithme du gradient conjugué (GC).

L'algorithme GC est une des méthodes itératives les plus utilisées pour résoudre les systèmes linéaires. Son intérêt réside dans sa simplicité et sa convergence plus rapide que l'algorithme de plus forte descente au prix d'un encombrement mémoire et d'un coût de calcul par itération très légèrement supérieur.

Considérons le système linéaire suivant

$$\mathbf{A}\boldsymbol{u} = \boldsymbol{b}$$

où $\mathbf{A} \in \mathbb{R}^{N\times N}$ est une matrice symétrique définie positive (SDP) ; $\boldsymbol{b} \in \mathbb{R}^N$ et $\boldsymbol{u} \in \mathbb{R}^N$ sont, respectivement, le vecteur connu et le vecteur à estimer. En l'absence d'erreur d'arrondi, la convergence est obtenue en au plus N itérations [Bertsekas, 1999, p. 131]. L'utilisation de la technique du préconditionnement peut améliorer de manière significative le taux de convergence de l'algorithme GC [Bertsekas, 1999, pp. 142-145].

Soit $\mathbf{M} \in \mathbb{R}^{N\times N}$ une matrice de préconditionnement SDP. Notons $\boldsymbol{u}_0 \in \mathbb{R}^N$ le vecteur initial, $\boldsymbol{r}_0 = \boldsymbol{b} - \mathbf{A}\boldsymbol{u}_0$ et $\boldsymbol{p}_0 = \mathbf{M}^{-1}\boldsymbol{r}_0$. L'algorithme GC préconditionné (GCP) peut s'écrire sous la forme presentée dans le tableau V.1 [Golub et Van Loan, 1996, Algorithm 10.3.1]. Il nous semble intéressant de souligner que la séquence des itérées $\{\boldsymbol{u}_i\}$ de l'algorithme GCP s'identifie à celle de l'algorithme BFGS [Bertsekas, 1999, Prop. 1.7.3].

La section suivante formalise la famille d'algorithmes qui utilisent l'algorithme GCP pour résoudre de manière approchée le système linéaire (V.1) déterminant la direction de descente courante des algorithmes de GR et de GY.

V.2.2 Famille d'algorithmes SQ+GCP

La famille d'algorithmes semi-quadratiques (GR ou GY) avec direction tronquée par algorithme GCP, désignée par SQ+GCP, peut s'écrire sous la forme suivante :

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \theta \boldsymbol{d}_k \tag{V.2}$$

$$\boldsymbol{d}_k = \boldsymbol{u}_{I_k}(\boldsymbol{x}_k) \tag{V.3}$$

où $\boldsymbol{u}_{I_k}(\boldsymbol{x}_k)$ est la solution courante après $I_k \in \{1, \ldots, N\}$ itérations de l'algorithme GCP défini par (V.5)-(V.9) avec $\boldsymbol{u}_0(\boldsymbol{x}_k) = \mathbf{0}$ comme initialisation, appliqué à la résolution du système linéaire suivant

$$\mathbf{A}(\boldsymbol{x}_k)\boldsymbol{u} = -\nabla\mathcal{J}(\boldsymbol{x}_k) \tag{V.4}$$

où $\mathbf{A}(\boldsymbol{x}_k)$ correspond soit à la matrice normale de GR définie par (IV.24) page 66, soit à la matrice normale de GY définie par (IV.26) page 66. Le terme *tronqué* signifie que l'algorithme GCP est arrêté avant convergence exacte, dans la mesure où cette dernière demande un nombre prohibitif d'itérations qui peut aller jusqu'à la taille N du vecteur inconnu. Notons que la suite des nombres d'itérations $\{I_k\}$ n'est pas nécessairement constante. De plus, cette famille d'algorithmes SQ+GCP peut être vue comme une généralisation des algorithmes SQ puisqu'en prenant N itérations on retrouve bien les algorithmes SQ+GCP.

A notre connaissance, il n'existe pas de preuve de convergence pour les algorithmes SQ+GCP à l'exception du cas trivial $\{I_k\} = \{N\}$ qui correspond à l'inversion exacte du système linéaire (V.4). La partie suivante présente nos résultats établissant la convergence de la famille d'algorithmes SQ+GCP sous les mêmes hypothèses que celles utilisées pour la convergence des algorithmes de GR et de GY exacts.

$\alpha_i = \dfrac{\|\boldsymbol{r}_i\|^2_{\mathbf{M}^{-1}}}{\|\boldsymbol{p}_i\|^2_{\mathbf{A}}}$	(Pas optimal)	(V.5)
$\boldsymbol{u}_{i+1} = \boldsymbol{u}_i + \alpha_i \boldsymbol{p}_i$	(Mise à jour de l'itérée)	(V.6)
$\boldsymbol{r}_{i+1} = \boldsymbol{r}_i - \alpha_i \mathbf{A}\boldsymbol{p}_i$	(Mise à jour du résidu)	(V.7)
$\beta_i = \dfrac{\|\boldsymbol{r}_{i+1}\|^2_{\mathbf{M}^{-1}}}{\|\boldsymbol{r}_i\|^2_{\mathbf{M}^{-1}}}$	(Formule de conjugaison)	(V.8)
$\boldsymbol{p}_{i+1} = \mathbf{M}^{-1}\boldsymbol{r}_{i+1} + \beta_i \boldsymbol{p}_i$	(Mise à jour de la direction de descente)	(V.9)

TAB. V.1 – L'algorithme GCP.

V.3 Convergence des algorithmes SQ+GCP

Nous avons choisi, pour des raisons de clarté, de reporter en annexe B.2 l'ensemble des résultats constituant la démonstration de convergence de la famille d'algorithmes SQ+GCP. Cette section indique les points clés utilisés dans la démonstration, précise la démarche générale et énonce le théorème de convergence avec les hypothèses associées.

Le résultat de convergence des algorithmes SQ+GCP est très proche du Théorème V.3.1 [Bertsekas, 1999, Prop. 1.2.1] énoncé ci-après. En fait, le cœur du travail de démonstration a été d'établir les hypothèses du Théorème V.3.1 que sont la *règle d'Armijo* et le caractère *gradient relié*, définis ci-dessous.

Définition V.3.1. *La séquence de pas* $\{\alpha_k\}$ *satisfait la* condition d'Armijo *pour* $\Omega \in]0,1[$ *si*

$$\mathcal{J}(\boldsymbol{x}_k) - \mathcal{J}(\boldsymbol{x}_{k+1}) + \Omega\alpha_k \boldsymbol{d}_k^{\mathrm{t}} \nabla\mathcal{J}(\boldsymbol{x}_k) \geq 0, \quad \forall k.$$

La *règle d'Armijo avec technique du rebroussement* [Bertsekas, 1999, p. 29] consiste à obtenir une séquence de pas vérifiant la condition d'Armijo tout en évitant des pas trop petits. Soit

$\Omega \in]0,1[$, $\tau \in]0,1[$ et $s > 0$. A l'itération courante k, la technique du rebroussement consiste à choisir le premier entier non négatif l_k qui permet d'obtenir un pas $\alpha_k = s\tau^{l_k}$ vérifiant

$$\mathcal{J}(\boldsymbol{x}_k) - \mathcal{J}(\boldsymbol{x}_k + s\tau^{l_k}\boldsymbol{d}_k) + \Omega s\tau^{l_k}\boldsymbol{d}_k^{\mathrm{t}}\nabla\mathcal{J}(\boldsymbol{x}_k) \geq 0.$$

Par la suite, on désignera simplement la règle d'Armijo avec technique du rebroussement par *règle d'Armijo*. Notons qu'en prenant $s = \theta$ et $l_k = 0$, il est immédiat qu'une séquence de pas constant θ vérifiant la condition d'Armijo satisfait la règle d'Armijo.

Définition V.3.2. *[Bertsekas, 1999, p. 35] La séquence de directions $\{\boldsymbol{d}_k\}$ est* gradient reliée *à $\{\boldsymbol{x}_k\}$ si pour toute sous-séquence $\{\boldsymbol{x}_k\}_{k\in\mathcal{K}}$ qui converge vers un point non stationnaire, la séquence correspondante $\{\boldsymbol{d}_k\}_{k\in\mathcal{K}}$ est bornée et vérifie*

$$\limsup_{k\to\infty, k\in\mathcal{K}} \boldsymbol{d}_k^{\mathrm{t}}\nabla\mathcal{J}(\boldsymbol{x}_k) < 0.$$

La Définition V.3.2 mérite une explication car elle peut sembler paradoxale. En effet, un algorithme itératif convergeant dont la séquence de directions est gradient reliée n'admet pas de sous-séquence qui converge vers un point non stationnaire. Le caractère gradient relié de la séquence de directions assure que, pour toute suite de points qui converge éventuellement vers un point non stationnaire, la séquence de directions est une séquence de directions de descente ($\boldsymbol{d}_k^{\mathrm{t}}\nabla\mathcal{J}(\boldsymbol{x}_k) < 0$). On peut alors faire décroître suffisamment le critère à l'aide d'une recherche de pas adaptée, ce qui contredit la convergence vers un point non stationnaire. En pratique, on utilise plutôt des conditions suffisantes plus simples à satisfaire qui assurent le caractère gradient relié de la séquence de directions, telles que celles proposées dans [Bertsekas, 1999, p. 36].

Définition V.3.3. *[Bertsekas, 1999, Def. A.2] Un vecteur $\boldsymbol{x} \in \mathbb{R}^N$ est un* point limite *d'une séquence $\{\boldsymbol{x}_k\}$ à valeur dans $\mathbb{R}^N$ s'il existe une sous-séquence de $\{\boldsymbol{x}_k\}$ qui converge vers $\boldsymbol{x}$.*

Théorème V.3.1. *[Bertsekas, 1999, Prop. 1.2.1] On suppose que le critère $\mathcal{J}$ est différentiable. Soit $\{\boldsymbol{x}_k\}$ une séquence générée par $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k\boldsymbol{d}_k$ et supposons que $\{\boldsymbol{d}_k\}$ est gradient reliée et α_k est obtenue par la règle d'Armijo. Alors tout point limite de $\{\boldsymbol{x}_k\}$ est un point stationnaire de $\mathcal{J}$.*

Afin d'établir que les propriétés gradient relié et règle d'Armijo sont vérifiées pour le pas constant θ des algorithmes SQ+GCP définis par (V.2)-(V.3) page 76, nous nous sommes appuyés principalement sur deux éléments essentiels. Le premier élément est lié à la nature même des algorithmes SQ+GCP. Il s'agit de prendre en compte la connaissance précise de la nature de l'approximation par GCP. Nous nous sommes alors appuyés sur les propriétés classiques des GCP pour établir le caractère gradient relié des algorithmes SQ+GCP (Lemme B.2.6 page 148). Le deuxième élément concerne une propriété forte des critères pénalisés généralisés définis par (IV.1)-(IV.2) page 54. Il s'agit du caractère approximation quadratique majorante induit par les matrices normales de GR et de GY sur le critère pénalisé généralisé. Cette propriété nous a permis d'établir que la condition d'Armijo est vérifiée pour le pas constant θ des algorithmes SQ+GCP (Lemme B.2.7 page 149).

Il est donc démontré que le Théorème V.3.1 s'applique pour les critères pénalisés généralisés. On peut cependant obtenir un résultat plus général que celui du Théorème V.3.1. Au lieu d'obtenir seulement que tout point limite de $\{\boldsymbol{x}_k\}$ est un point stationnaire (annulation du gradient) on montre que la séquence $\{\nabla\mathcal{J}(\boldsymbol{x}_k)\}$ tend vers le vecteur nul. Cette dernière propriété impliquant la première, elle est plus générale. Notons que nous obtenons ce résultat plus général en utilisant la propriété que les critères pénalisés généralisés sont bornés inférieurement (Théorème B.2.1 page 149).

Le Théorème V.3.2 énoncé ci-dessous établit la convergence de la famille d'algorithmes SQ+GCP. Il fait appel aux Hypothèses 1 et 2 page 62 sur la fonction de régularisation ϕ du critère pénalisé. Il est à noter que les hypothèses du Théorème V.3.2 sont les mêmes que celles de [Allain *et al.*, 2006] qui établissent la convergence des formes exactes des algorithmes de GR et de GY. Autrement dit, les algorithmes SQ+GCP peuvent être utilisés dans les mêmes situations que les algorithmes de GR et de GY exacts.

Théorème V.3.2. *Soit $\mathcal{J}$ le critère pénalisé généralisé défini par (IV.1)-(IV.2) page 54, où la condition suivante*

$$\ker(\mathbf{H}^{\mathrm{t}}\mathbf{H}) \cap \ker(\mathbf{V}^{\mathrm{t}}\mathbf{V}) = \{\mathbf{0}\}$$

est vérifiée. Soit $\boldsymbol{x}_k$ défini par (V.2)-(V.3) page 76 avec $\theta \in]0,2[$, $\mathbf{A}_k = \mathbf{A}_{GY}^a$ avec $0 < a < 1/L$ définie par (IV.35), page 69 (resp. $\mathbf{A}_k = \mathbf{A}_{GR}(\boldsymbol{x}_k)$ définie par (IV.37), page 69) et avec une matrice de préconditionnement $\mathbf{M}$ définie positive à spectre borné. Supposons que l'Hypothèse 1 (resp. Hypothèse 2) page 62 sur la fonction de régularisation ϕ est satisfaite.

Alors la séquence $\{\mathcal{J}(\boldsymbol{x}_k)\}$ est décroissante au sens large :

$$\mathcal{J}(\boldsymbol{x}_k) \geq \mathcal{J}(\boldsymbol{x}_{k+1}), \quad \forall k \tag{V.10}$$

et on a la convergence dans le sens suivant :

$$\lim_{k\to\infty} \nabla\mathcal{J}(\boldsymbol{x}_k) = \mathbf{0}. \tag{V.11}$$

Preuve. Voir Annexe B.2, page 145. L'Hypothèse 1, page 147 est vérifiée pour le critère pénalisé généralisé. L'Hypothèse 2, page 148 est vérifiée pour le critère pénalisé généralisé d'après les Lemmes IV.4.2, page 67 et IV.4.3, page 68. Ainsi, le Théorème B.2.1, page 149 s'applique pour le critère pénalisé généralisé. □

Notons que la convergence des algorithmes SQ+GCP reste valable pour un pas $\theta \in]0,2[$ constant, cette simplicité étant un des points forts des algorithmes de GR et de GY exacts. On rappelle que la convergence au sens du Théorème V.3.2 entraîne la convergence vers le minimum global si le critère pénalisé est strictement convexe. Précisons aussi qu'il est possible de considérer une séquence de matrices de préconditionnement variables et non pas une unique matrice de préconditionnement, pourvu que cette séquence de matrices soit à spectre uniformément positif borné (Téorème B.2.1, page 149).

V.4 Conclusion

L'inversion à chaque itération des matrices normales des algorithmes de GR et de GY exacts est généralement une opération très coûteuse pour les problèmes de grande taille. Il est possible de s'affranchir totalement de la difficulté de l'inversion de ces matrices normales SQ en considérant une version d'algorithme SQ coordonnée par coordonnée [Brette et Idier, 1996]. La convergence de cette version coordonnée par coordonnée est assurée dans le cas d'une pénalisation convexe d'après [Idier, 2001], mais en contrepartie le taux de convergence se trouve diminué. En somme, la version d'algorithme SQ coordonnée par coordonnée est une forme d'évitement de la difficulté de l'inversion des matrices normales de GR et de GY. Ce type d'approche perd de son intérêt dès lors que la convergence d'une forme approchée des algorithmes de GR et de GY contenant comme cas limite les formes exactes est assurée, ce qui est le cas de la famille d'algorithmes SQ+GCP.

La preuve de convergence des algorithmes SQ+GCP a été rendue possible en les considérant comme des algorithmes propres et non pas seulement comme des versions approchées. Ce changement de point de vue a des répercussions pratiques importantes. L'étude expérimentale conduite dans le chapitre VII montre clairement l'intérêt de considérer des versions des algorithmes SQ+GCP avec une approximation assez forte, c'est-à-dire avec un degré d'éloignement par rapport aux algorithmes exacts nettement plus important que ce qui a été proposé jusqu'ici [Nikolova et Ng, 2001]. Il a été fait usage dans ce chapitre de l'algorithme du gradient conjugué linéaire. Le chapitre suivant traite de l'alternative consistant à utiliser des algorithmes gradient conjugué non linéaires pour minimiser les critères pénalisés.

Bibliographie

[Allain *et al.*, 2006] M. Allain, J. Idier et Y. Goussard. On global and local convergence of half-quadratic algorithms. *IEEE Trans. Image Processing*, 15 (5) : 1130–1142, mai 2006.

[Belge *et al.*, 2000] M. Belge, M. Kilmer et E. Miller. Wavelet domain image restoration with adaptive edge-preserving regularization. *IEEE Trans. Image Processing*, 9 (4) : 597–608, avril 2000.

[Bertsekas, 1999] D. P. Bertsekas. *Nonlinear programming*. Athena Scientific, Belmont, MA, USA, 2ème edition, 1999.

[Brette et Idier, 1996] S. Brette et J. Idier. Optimized single site update algorithms for image deblurring. In *Proc. IEEE ICIP*, pages 65–68, Lausanne, Suisse, septembre 1996.

[Charbonnier *et al.*, 1997] P. Charbonnier, L. Blanc-Féraud, G. Aubert et M. Barlaud. Deterministic edge-preserving regularization in computed imaging. *IEEE Trans. Image Processing*, 6 (2) : 298–311, février 1997.

[Cipra, 2000] B. Cipra. The best of the 20th century : Editors name top 10 algorithms. *SIAM News*, 33 (4) : 1, mai 2000.

[Delaney et Bresler, 1998] A. H. Delaney et Y. Bresler. Globally convergent edge-preserving regularized reconstruction : an application to limited-angle tomography. *IEEE Trans. Image Processing*, 7 (2) : 204–221, février 1998.

[Golub et Van Loan, 1996] G. H. Golub et C. F. Van Loan. *Matrix computations*. The Johns Hopkins University Press, Baltimore, 3ème edition, 1996.

[Hestenes et Stiefel, 1952] M. R. Hestenes et E. Stiefel. Methods of conjugate gradients for solving linear system. *J. Res. Nat. Bur. Stand.*, 49 : 409–436, 1952.

[Idier, 2001] J. Idier. Convex half-quadratic criteria and interacting auxiliary variables for image restoration. *IEEE Trans. Image Processing*, 10 (7) : 1001–1009, juillet 2001.

[Ng *et al.*, 1999] M. K. Ng, R. H. Chan et W.-C. Tang. A fast algorithm for deblurring models with Neumann boundary conditions. *SIAM J. Sci. Comput.*, 21 (3) : 851–866, 1999.

[Nikolova et Ng, 2001] M. Nikolova et M. Ng. Fast image reconstruction algorithms combining half-quadratic regularization and preconditioning. In *Proc. IEEE ICIP*, pages 277–280, Thessaloniki, Grèce, octobre 2001.

[Nikolova et Ng, 2005] M. Nikolova et M. Ng. Analysis of half-quadratic minimization methods for signal and image recovery. *SIAM J. Sci. Comput.*, 27 : 937–966, 2005.

Chapitre VI

ALGORITHMES DU GRADIENT CONJUGUÉ NON LINÉAIRE

Dans le chapitre précédent, la difficulté de mise en œuvre des algorithmes de GR et de GY a été exposée. Cette difficulté est liée au coût généralement prohibitif de l'inversion des matrices normales de GR et de GY pour les problèmes de grande taille couramment rencontrés en traitement de l'image. Nous avons vu dans la section IV.4.6, page 65 que les algorithmes de GR et de GY peuvent être interprétés comme des algorithmes de type quasi-Newton. Une alternative classique aux algorithmes de type quasi-Newton est l'utilisation des algorithmes de type gradient conjugué non linéaire (GCNL). Leur intérêt réside dans leur plus faible encombrement mémoire et coût de calcul par itération inférieur.

L'utilisation des algorithmes de type GCNL pour la minimisation des critères pénalisés non quadratiques a été proposée dans la communauté du traitement de l'image, mais sans assurance de la convergence [Fessler et Booth, 1999; Rivera et Marroquin, 2003]. [Fessler et Booth, 1999] ne se préoccupe pas de la convergence et [Rivera et Marroquin, 2003] se contente d'établir le caractère non croissant de la suite des itérées. Pourtant, il existe en optimisation des résultats de convergence des algorithmes GCNL. Le lecteur trouvera en annexe A, page 139 une étude bibliographique sur les résultats de convergence des GCNL présents dans la littérature de l'optimisation. Ces résultats de convergence sont malheureusement peu connus en traitement du signal et des images.

La convergence des algorithmes de GR et de GY, ainsi que celle de leur forme approchée SQ+GCP, est assurée pour un pas constant. Par contre, un pas constant n'assure pas la convergence des algorithmes GCNL. Il est nécessaire de mettre en œuvre une recherche du pas. Les

résultats de convergence d'optimisation portent justement sur les conditions que doit vérifier la recherche du pas pour assurer la convergence des algorithmes GCNL. S'il est possible d'appliquer les résultats de convergence des algorithmes GCNL figurant dans la littérature de l'optimisation pour la minimisation des critères pénalisés, ils nous semblent insatisfaisants pour deux raisons.

D'une part, les résultats de convergence des algorithmes GCNL de la littérature de l'optimisation se contentent de proposer des conditions que doit vérifier une recherche du pas sans pour autant l'expliciter. Il existe des méthodes génériques de recherche du pas, valables pour des critères généraux. Le lecteur trouvera en annexe A, page 139 une étude bibliographique sur ces méthodes génériques de recherche du pas. Cependant, il ne semble pas y avoir une méthode standard de recherche du pas qui se démarquerait des autres. De plus, l'objectif de ces méthodes génériques de recherche du pas est radicalement opposé au nôtre. Elles visent à une certaine universalité alors que notre objectif est de prendre en compte le plus d'informations possibles sur le critère pénalisé à minimiser. En effet, en étant plus spécifique on a l'espoir d'être plus efficace. L'idée proposée de manière indépendante par [Fessler et Booth, 1999] et [Rivera et Marroquin, 2003], consiste à utiliser la forme scalaire tronquée, dans le sens d'un faible nombre d'itérations, de l'algorithme de GR pour la recherche du pas. On peut aussi envisager d'utiliser la forme scalaire tronquée de l'algorithme de GY. Ces deux recherches de pas sont appelées respectivement GR1D et GY1D. Nous désignerons l'une ou l'autre de ces recherches de pas par SQ1D et la famille d'algorithmes de minimisation du critère pénalisé qui en résulte par GCNL+SQ1D. Les algorithmes GCNL+SQ1D sont en fait intimement liés à la contribution de [Sun et Zhang, 2001] qui établit la convergence pour des conditions qui se révèlent être trop restrictives. Le fait que ces algorithmes GCNL+SQ1D ont été proposés de manière indépendante à plusieurs reprises laisse à penser qu'il s'agit d'algorithmes "profonds".

D'autre part, la mise en œuvre des résultats de convergence des algorithmes GCNL de la littérature de l'optimisation n'est pas simple. Ces résultats dépendent d'une recherche du pas générique difficile à implémenter [Nocedal et Wright, 1999, p. 61]. De plus, une recherche du pas générique doit vérifier certaines conditions qui dépendent de plusieurs paramètres. La question du réglage de ces paramètres se pose. Le résultat de convergence exposé dans ce chapitre permet de s'affranchir totalement de cette difficulté pour les critères pénalisés. En effet, on établit que la recherche de pas SQ1D assure la convergence des algorithmes GCNL sans avoir à satisfaire d'autres conditions. L'intérêt de ce résultat est donc une simplification notable par rapport à l'utilisation des résultats de la littérature de l'optimisation.

VI.1 Algorithmes du gradient conjugué non linéaire

L'algorithme du gradient conjugué (GC) défini par (V.5)-(V.9), page 77 a pour objet la résolution d'un système linéaire symétrique défini positif. C'est l'objectif historique du GC. Cependant, résoudre un système linéaire est équivalent à minimiser un critère quadratique [Bertsekas, 1999, p. 130].

Ainsi, l'algorithme du GC peut être vu comme minimisant une quadratique convexe. Ce constat a alors permis d'envisager au cours des années 1960 l'utilisation de l'algorithme GC pour des critères non quadratiques [Fletcher et Reeves, 1964]. On parle alors d'algorithmes GC non linéaires (GCNL). Cette appellation est malheureusement source d'une certaine ambiguïté. Le nom de l'algorithme GC linéaire découle d'une de ses propriétés fondamentales qui est la conjugaison de ses directions de descente pour un critère quadratique avec la formule du pas optimal [Bertsekas, 1999, p. 136]. Par contre, cette propriété de conjugaison des directions de descente est perdue dans le cas d'un critère non quadratique. L'appellation GCNL est réservée à des algorithmes qui s'identifient à l'algorithme GC dans le cas d'un critère quadratique convexe avec le pas optimal défini par (V.5), page 77. Il existe de nombreux algorithmes GCNL dont

nous rappelons les plus connus dans l'annexe A, page 139. Nous présentons dans la section suivante un des algorithmes GCNL sans doute le plus utilisé pour les critères non quadratiques.

VI.1.1 Forme de Polak-Ribiere préconditionnée

Les algorithmes GCNL sont efficaces pour minimiser des critères différentiables dans le cas de problème de grande taille [Bertsekas, 1999, p. 155]. Ce sont des algorithmes itératifs de la forme $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k$, où les directions de descente $\{\boldsymbol{d}_k\}$ sont mutuellement conjuguées dans le cas d'un problème quadratique avec pas optimal. L'algorithme GC défini par (V.5)-(V.9), page 77 est appelé forme de Fletcher-Reeves (FR) dans le cadre d'un critère non quadratique. Il est conseillé d'employer pour un critère non quadratique la forme de Polak-Ribiere plutôt que celle de Fletcher-Reeves pour des raisons de meilleure efficacité [Bertsekas, 1999, p. 140].

La forme de Polak-Ribiere préconditionnée est indiquée dans le tableau VI.1 [Press *et al.*, 1992]. Nous désignerons cet algorithme par GCPPR. L'algorithme GCPPR présente deux degrés de liberté : le pas α_k et la matrice inversible de préconditionnement $\mathbf{M}_k$. Notons que dans le cas d'un critère quadratique, l'algorithme GCPPR s'identifie bien à l'algorithme classique du GC défini par (V.5)-(V.9). Dans ce cas, le pas optimal défini par

$$\hat{\alpha}_k = \arg\min_{\alpha} f(\alpha)$$

où $f(\alpha) = \mathcal{J}(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k)$, s'obtient par la formule analytique (V.5), page 77. Par contre, dans le cas d'un critère pénalisé non quadratique, il est la plupart du temps impossible d'obtenir le pas optimal, ou alors le calcul serait trop coûteux. C'est pourquoi une étape de recherche du pas, *i.e.,* une minimisation 1D de f, est requise. L'objectif d'une telle recherche du pas est de ne pas être trop coûteuse tout en assurant la convergence de l'algorithme.

$$\boldsymbol{p}_k = -(\mathbf{M}_k)^{-1} \nabla \mathcal{J}(\boldsymbol{x}_k) \qquad \text{(préconditionnement)} \qquad \text{(VI.1)}$$

$$\beta_k = \begin{cases} 0 & \text{si } k = 0 \\ \dfrac{\left(\nabla \mathcal{J}(\boldsymbol{x}_k) - \nabla \mathcal{J}(\boldsymbol{x}_{k-1})\right)^{\mathrm{t}} \boldsymbol{p}_k}{\left(\nabla \mathcal{J}(\boldsymbol{x}_{k-1})\right)^{\mathrm{t}} \boldsymbol{p}_{k-1}} & \text{si } k > 0 \end{cases} \qquad \text{(VI.2)}$$

$$\boldsymbol{d}_k = \boldsymbol{p}_k + \beta_k \boldsymbol{d}_{k-1} \qquad \text{(direction PRPCG)} \qquad \text{(VI.3)}$$

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k \qquad \text{(mise à jour)} \qquad \text{(VI.4)}$$

Tab. VI.1 – L'algorithme GCPPR.

VI.1.2 Recherche du pas par algorithmes SQ scalaires

Les méthodes génériques de recherche du pas assurant la convergence des algorithmes GCNL présentées dans l'annexe A sont difficiles à implémenter [Nocedal et Wright, 1999, p. 61]. Par contre, les algorithmes de GR et de GY ne nécessitent pas de recherche du pas car ils utilisent un pas constant, ce qui constitue une de leur principales qualités.

On peut noter, après quelques manipulations simples, que la fonction $f(\alpha) = \mathcal{J}(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k)$ présente la même structure pénalisée généralisée que le critère $\mathcal{J}$ défini par (IV.1)-(IV.2), page 54

$$f(\alpha) = a_2 \alpha^2 - 2a_1 \alpha + a_0 + \lambda \sum_{c=1}^{C} \phi(\|\alpha \boldsymbol{v}_c - \boldsymbol{\tau}_c\|), \qquad \text{(VI.5)}$$

où $\boldsymbol{v}_c = \mathbf{V}_c\boldsymbol{d}_k$, $\boldsymbol{\tau}_c = \boldsymbol{\omega}_c - \boldsymbol{V}_c\boldsymbol{x}_k$ et $a_0 = \boldsymbol{u}_2^{\mathrm{t}}\boldsymbol{u}_2$, $a_1 = \boldsymbol{u}_1^{\mathrm{t}}\boldsymbol{u}_2$, $a_2 = \boldsymbol{u}_1^{\mathrm{t}}\boldsymbol{u}_1$ avec $\boldsymbol{u}_1 = \mathbf{H}\boldsymbol{d}_k$ et $\boldsymbol{u}_2 = \boldsymbol{y} - \mathbf{H}\boldsymbol{x}_k$.

Ainsi, ce constat nous amène naturellement à utiliser comme recherche du pas de l'algorithme GCPPR une forme tronquée, dans le sens d'un faible nombre d'itérations, des algorithmes SQ scalaires (SQ1D). Cette idée n'est pas nouvelle et a déjà été proposée dans [Fessler et Booth, 1999; Rivera et Marroquin, 2003]. L'intérêt d'une telle approche est que les algorithmes SQ ont été spécifiquement développés pour minimiser l'aspect pénalisé de la fonction f, contrairement aux méthodes génériques de recherche du pas. Ils sont donc tout indiqués pour effectuer cette recherche du pas, d'autant plus que ces algorithmes SQ ne posent aucune difficulté dans le cas scalaire, contrairement à leur application aux problèmes de grande taille. En effet, la résolution du système linéaire SQ s'effectue de manière triviale dans le cas scalaire. Ainsi, les algorithmes SQ scalaires fournissent de manière simple une séquence $\alpha_k^1, \ldots, \alpha_k^i$ qui converge vers un minimum local de f (et donc vers le pas optimal lorsque ϕ est convexe).

Lemme VI.1.1. *Les algorithmes SQ scalaires (SQ1D), appliqués à la minimisation de $f(\alpha) = \mathcal{J}(\boldsymbol{x}_k + \alpha\boldsymbol{d}_k)$, peuvent s'écrire sous la forme suivante*

$$\begin{cases} \alpha_k^0 = 0 \\ \alpha_k^{i+1} = \alpha_k^i - \theta \dfrac{\boldsymbol{d}_k^{\mathrm{t}} \nabla \mathcal{J}\left(\boldsymbol{x}_k + \alpha_k^i \boldsymbol{d}_k\right)}{\boldsymbol{d}_k^{\mathrm{t}} \mathbf{Q}_k^i \boldsymbol{d}_k}, \quad 0 \leq i \leq I-1 \\ \alpha_k = \alpha_k^I \end{cases} \tag{VI.6}$$

avec $\mathbf{Q}_k^i = \mathbf{A}_{\mathrm{GY}}^a$ pour GY1D, $\mathbf{Q}_k^i = \mathbf{A}_{\mathrm{GR}}(\boldsymbol{x}_k + \alpha_k^i \boldsymbol{d}_k)$ pour GR1D et I est un paramètre indiquant le nombre de sous-itérations.

Preuve. Voir l'annexe B.3.1, page 150. □

Notons que la forme (VI.6) est une généralisation des GCNL sans recherche de pas [Sun et Zhang, 2001] car lorsque $I = 1$, la formule résultante s'identifie avec celle proposée dans [Sun et Zhang, 2001]. On pourrait être tenté d'envisager un grand nombre de sous-itérations I des algorithmes SQ1D afin d'approcher au mieux le pas optimal (au moins dans le cas convexe). Cela aurait pour conséquence un surcoût pour l'algorithme GCPPR+SQ1D d'autant plus grand que le nombre de sous-itérations I est important. De plus, on ne serait pas pour autant assuré de la convergence de l'algorithme GCPPR+SQ1D. D'ailleurs, il existe un résultat montrant que pour un certain critère non convexe, l'algorithme GCPPR ne converge pas lorsque le pas optimal est utilisé [Powell, 1984]. Ce résultat, plutôt contre-intuitif, montre bien qu'il est absolument nécessaire de se préoccuper de la convergence des algorithmes GCPPR. Ce sera l'objet de la section VI.3.

VI.2 Comparaison structurelle entre algorithmes SQ et GCNL

Dans cette section, les liens entre les algorithmes de GR et de GY et les algorithmes GCNL sont établis. Nous montrons que les algorithmes GCP+SQ1D constituent une famille d'algorithmes de minimisation contenant les algorithmes exacts de GR et de GY. En réinterprétant les algorithmes de GR et de GY comme étant des choix particuliers de préconditionnement, certaines alternatives se démarquent par leur coût d'implémentation plus faible.

VI.2.1 Le pas

La recherche du pas (VI.6) semble de prime abord bien plus compliquée que le pas constant des algorithmes semi-quadratiques. En fait, on montre que la recherche du pas (VI.6) peut être

interprétée comme une généralisation naturelle du pas constant lorsqu'on passe des algorithmes à séquence de directions gradient relié aux algorithmes GCNL.

En effet, l'algorithme GCPPR défini dans le tableau VI.1, page 83 mais sans conjugaison ($\beta_k = 0$) s'écrit sous la forme suivante :

$$\begin{aligned} \boldsymbol{d}_k &= -(\mathbf{M}_k)^{-1}\nabla\mathcal{J}(\boldsymbol{x}_k) \\ \boldsymbol{x}_{k+1} &= \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k \end{aligned}$$

Ainsi, en considérant $\mathbf{Q}_k = \mathbf{M}_k$ avec une stratégie non conjuguée, *i.e.,* $\beta_k = 0$, la formule de recherche du pas (VI.6) avec $I = 1$ se simplifie en

$$\alpha_k = -\theta \frac{\boldsymbol{d}_k^{\mathrm{t}} \nabla\mathcal{J}\left(\boldsymbol{x}_k\right)}{\boldsymbol{d}_k^{\mathrm{t}} \mathbf{Q}_k \boldsymbol{d}_k} = \theta \frac{\boldsymbol{d}_k^{\mathrm{t}} \mathbf{M}_k \boldsymbol{d}_k}{\boldsymbol{d}_k^{\mathrm{t}} \mathbf{Q}_k \boldsymbol{d}_k} = \theta. \tag{VI.7}$$

Donc, les algorithmes de GR et de GY (vus sous la forme d'algorithmes de quasi-Newton à pas fixe (IV.22)-(IV.23), page 66) correspondent à des choix particuliers de $\mathbf{M}_k$ (soit $\mathbf{A}_{\mathrm{GR}}(\boldsymbol{x}_k)$ ou $\mathbf{A}^a_{\mathrm{GY}}$) parmi cette famille. Les algorithmes GCP+SQ1D constituent donc bien une famille qui généralise les algorithmes SQ en rajoutant l'aspect conjugaison ($\beta_k \neq 0$).

VI.2.2 Le préconditionnement

Le préconditionnement est une technique particulièrement utilisée pour l'algorithme du GC car il permet d'améliorer très sensiblement le taux de convergence. L'objectif consiste alors à appliquer le GC au travers d'une transformation linéaire sur les variables du problème afin que le Hessien transformé soit le plus proche possible de la matrice identité. Il existe des résultats quantifiant le taux de convergence de l'algorithme GC en fonction de la distribution des valeurs propres du Hessien [Bertsekas, 1999, prop. 1.6.2]. Lorsque le Hessien est la matrice identité alors l'algorithme GC converge en une seule itération. La technique du préconditionnement peut aussi être envisagée pour les algorithmes GCNL. Le choix de l'inverse du Hessien comme préconditionneur entraîne pour les algorithmes GCNL un taux de convergence *superlinéaire* lorsque le pas optimal est utilisé, ce qui correspond au taux de convergence de la méthode de Newton [Al-Baali et Fletcher, 1996].

Lors de l'utilisation d'un préconditionnement il est important de distinguer deux notions :

- Le taux de convergence quantifie l'efficacité d'un algorithme en terme de nombre d'itérations.
- La vitesse de convergence expérimentale est une mesure globale d'efficacité, en temps de calcul.

La mesure d'intérêt pratique est bien la vitesse de convergence expérimentale. En effet, un algorithme nécessitant peu d'itérations mais qui sont très coûteuses, n'est pas intéressant en pratique. La vitesse de convergence expérimentale résultant d'un préconditionnement dépend du coût de sa mise en œuvre. La version préconditionnée a systématiquement un coût par itération supérieur à la version non préconditionnée. Il est alors délicat de prévoir si le coût du préconditionnement va être finalement compensé et amener à un gain pratique significatif.

En pratique, un préconditionneur efficace répond à un compromis entre un meilleur taux de convergence et un faible coût de calcul. Par exemple, pour les problèmes où la matrice d'observation $\mathbf{H}$ est proche de Toeplitz, il existe des préconditionneurs efficaces. Dans ce cas, le Hessien du critère pénalisé peut être approché par une matrice constante proche de Toeplitz sous l'hypothèse classique de régularité de l'image à restaurer. Les matrices de type Toeplitz peuvent être préconditionnées efficacement par certaines matrices [Chan et Ng, 1996]. On peut alors bénéficier de l'inversion rapide de ces dernières matrices [Ng *et al.*, 1999].

Dans la section IV.4.6 page 65, les algorithmes de GR et de GY sont interprétés comme des algorithmes de type quasi-Newton à pas fixe. Or, il est possible de voir les algorithmes de type quasi-Newton comme des algorithmes de type gradient préconditionné, sans conjugaison. Selon cette interprétation, les algorithmes de GR et de GY correspondent à une matrice de préconditionnement $\mathbf{M}_k$ imposée (soit $\mathbf{A}_{\mathrm{GR}}(\boldsymbol{x}_k)$ ou $\mathbf{A}^a_{\mathrm{GY}}$). Ce type de préconditionnement entraîne généralement de très importantes difficultés de résolution dans le cas de problèmes de grande taille, comme exposé dans la section V.1, page 74.

Par contre, l'algorithme GCPPR permet de choisir n'importe quelle matrice de préconditionnement constante, pourvu qu'elle soit définie positive. Une condition nécessaire pour un choix judicieux de préconditionnement est que la résolution ne soit pas trop coûteuse. Ainsi, les algorithmes de type de GR et de GY apparaissent comme des choix de préconditionnement généralement peu efficaces pour les problèmes de grande taille. On voit ainsi qu'il est plus judicieux d'envisager d'autres types de préconditionnement.

Pourtant, l'intérêt des algorithmes de GR et de GY réside dans la prise en compte de la spécificité des critères pénalisés. Le problème est de prendre en compte cette spécificité tout en ayant un préconditionnement efficace. Lorsqu'on considère un préconditionneur $\mathbf{M}_k$ différent de $\mathbf{Q}_k$ n'entraînant pas de difficulté de résolution du système préconditionné, le pas défini par une itération ($I = 1$) de l'algorithme SQ1D défini par (VI.6) s'écrit sous la forme

$$\alpha_k = \theta \frac{\boldsymbol{d}_k^{\mathrm{t}} \mathbf{M}_k \boldsymbol{d}_k}{\boldsymbol{d}_k^{\mathrm{t}} \mathbf{Q}_k \boldsymbol{d}_k}. \tag{VI.8}$$

Ce pas correspond à la généralisation du pas constant θ utilisé par les algorithmes de GR et de GY. En somme, le choix des algorithmes de GR et de GY consiste à privilégier la simplicité du pas (constant), au détriment de l'efficacité du préconditionnement. Pourtant, la formule de pas variable (VI.8) ne nécessite que des produits matrices vecteurs alors que le préconditionnement nécessite une inversion de matrice. Donc, le coût global du préconditionnement est bien plus important que le coût global de la formule du pas variable. Il est donc préférable de privilégier un préconditionnement efficace plutôt qu'un pas constant. On voit ainsi qu'il est souhaitable de prendre le contre-pied des algorithmes de GR et de GY en considérant, par exemple, la famille d'algorithmes GCPPR+SQ1D avec un préconditionneur efficace.

VI.3 Résultat de convergence

L'intérêt de la famille d'algorithmes GCPPR+SQ1D par rapport aux algorithmes de GR et de GY a été exposé dans la partie précédente. Cette partie s'intéresse à la convergence des algorithmes GCPPR+SQ1D pour la minimisation des critères pénalisés généralisés. Cette convergence découle de celle établie dans l'annexe C, page 155 pour des critères généraux. L'objet de cette partie consiste d'une part à établir que les critères pénalisés généralisés vérifient les hypothèses de l'annexe C. D'autre part, cette partie énonce les deux théorèmes de convergence correspondant à la présence et à l'absence de l'aspect conjugaison.

Les résultats de l'annexe C, utilisés pour établir la convergence de la famille d'algorithmes GCPPR+SQ1D nécessitent deux hypothèses sur le critère pénalisé généralisé. Ces deux hypothèses sont le caractère gradient Lipschitz et la coercivité du critère pénalisé généralisé. Notons que ces deux hypothèses n'ont pas été nécessaires pour établir le Théorème V.3.2, page 79. Cependant, ces deux hypothèses ne sont pas restrictives par rapport à celles du Théorème V.3.2. En effet, on montre dans les deux sections ci-dessous que ces deux propriétés sont vérifiées dès lors que la fonction de régularisation vérifie l'Hypothèse 1, page 62.

VI.3.1 Caractère gradient Lipschitz du critère pénalisé généralisé

Le caractère gradient Lipschitz du critère est une hypothèse classiquement utilisée en optimisation pour obtenir la convergence d'algorithmes. Ce caractère gradient Lipschitz peut simplement être vu comme une hypothèse de régularité du critère. Le lemme suivant établit le caractère gradient Lipschitz du critère pénalisé généralisé $\mathcal{J}$ défini par (IV.1)-(IV.2), page 54.

Lemme VI.3.1. *Supposons que l'Hypothèse 1 est satisfaite.*

Alors le critère pénalisé généralisé $\mathcal{J}$ défini par (IV.1)-(IV.2), page 54 est μ-$\mathcal{LC}^1$ avec la constante de Lipschitz

$$\mu = 2\|\mathbf{H}^{\mathrm{t}}\mathbf{H}\| + \lambda L \sum_{c=1}^{C} \|\mathbf{V}_c\|^2 > 0 \tag{VI.9}$$

où la fonction de régularisation ϕ est L-$\mathcal{LC}^1$.

Preuve. Voir l'annexe B.3.2, page 151. □

Le Lemme VI.3.1 montre en somme que le caractère gradient Lipschitz de la fonction de régularisation ϕ se transfert au critère pénalisé généralisé. Ainsi, la seule Hypothèse 2, page 62 sur la fonction de régularisation ϕ ne suffit pas à établir le caractère gradient Lipschitz du critère pénalisé généralisé.

Certains résultats en optimisation utilisent la valeur numérique de la constante de Lipschitz μ du critère gradient Lipschitz. Essayer d'obtenir la plus faible valeur de cette constante de Lipschitz devient alors de toute importance. Ce travail peut être difficile, d'autant plus qu'il n'est pas toujours aisé d'établir que la valeur courante de la constante de Lipschitz est la meilleure possible.

En fait, nos résultats de convergence s'appuient uniquement sur le caractère gradient Lipschitz du critère sans que la valeur numérique de la constante de Lipschitz ne soit utilisée. Autrement dit, le Lemme VI.3.1 qui n'assure pas le caractère optimal de la constante de Lipschitz nous suffit amplement. Il n'est donc pas nécessaire d'affiner la constante de Lipschitz μ définie par (VI.9).

VI.3.2 Coercivité du critère pénalisé généralisé

A l'image de la section précédente, le lemme suivant établit que le caractère coercif de la fonction de régularisation ϕ se transfert au critère pénalisé généralisé.

Lemme VI.3.2. *Supposons que l'Hypothèse 1 ou l'Hypothèse 2, page 62 est satisfaite. On suppose aussi que la condition suivante est vérifiée :*

$$\ker(\mathbf{H}^{\mathrm{t}}\mathbf{H}) \cap \ker(\mathbf{V}^{\mathrm{t}}\mathbf{V}) = \{\mathbf{0}\}\,.$$

Alors le critère pénalisé généralisé $\mathcal{J}$ défini par (IV.1)-(IV.2), page 54 est coercif.

Preuve. Voir l'annexe B.3.3, page 152. □

VI.3.3 Les hypothèses de l'annexe C sont vérifiées

L'annexe C, page 155 suppose certaines hypothèses sur le critère à minimiser. Il s'agit de l'hypothèse que les lignes de niveau définies par $L = \{\boldsymbol{x} \in \mathbb{R}^N | \mathcal{J}(\boldsymbol{x}) \leq \mathcal{J}(\boldsymbol{x}_0)\}$ sont bornées, de l'hypothèse que le critère est gradient Lipschitz (Assumption 3, annexe C page 157), et de l'hypothèse que les approximations quadratiques sont majorantes (Assumption 5, annexe C page 158).

L'hypothèse que les lignes de niveau sont bornées est immédiatement vérifiée dès lors que le critère pénalisé généralisé est coercif (Lemme VI.3.2). L'hypothèse du caractère gradient Lipschitz du critère pénalisé généralisé est vérifiée d'après le Lemme VI.3.1. Finalement, l'existence d'approximations quadratiques majorantes est fournie par les Lemmes IV.4.4, page 69 et IV.4.6, page 70.

VI.3.4 Convergence sans conjugaison

Cette partie établit la convergence de la famille d'algorithmes GCPPR+SQ1D sans conjugaison ($\beta_k = 0$) pour la minimisation des critères pénalisés généralisés. On commence par définir le caractère spectre uniformément positif borné d'une suite de matrices.

Définition VI.3.1. *Une suite de matrices symétriques définies positives $\mathcal{Q} = \{\mathbf{Q}_k\} \in \mathbb{R}^{N\times N}$ a un spectre uniformément positif borné avec une borne inférieure strictement positive s'il existe $\nu_1(\mathcal{Q})$, $\nu_2(\mathcal{Q}) \in \mathbb{R}$ tels que*

$$\nu_2(\mathcal{Q}) \geq \nu_2(\mathbf{Q}_k) \geq \nu_1(\mathbf{Q}_k) \geq \nu_1(\mathcal{Q}) > 0, \quad \forall k.$$

Dans ce cas, on dira simplement que $\mathcal{Q}$ est à spectre uniformément positif borné.

Théorème VI.3.1. *Soit $\mathcal{J}$ le critère pénalisé généralisé défini par (IV.1)-(IV.2), page 54 où la condition suivante*

$$\ker(\mathbf{H}^{\mathrm{t}}\mathbf{H}) \cap \ker(\mathbf{V}^{\mathrm{t}}\mathbf{V}) = \{\mathbf{0}\}$$

est vérifiée. Soit $\boldsymbol{x}_k$ défini par

$$\boldsymbol{d}_k = -(\mathbf{M}_k)^{-1}\nabla\mathcal{J}(\boldsymbol{x}_k) \tag{VI.10}$$

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k \tag{VI.11}$$

où la recherche de pas définie par (VI.6), page 84 est utilisée avec $\theta \in]0, 2[$, $\mathbf{Q}_k^i = \mathbf{A}_{\mathrm{GY}}^a$ avec $0 < a < 1/L$ définie par (IV.35), page 69 (resp. $\mathbf{Q}_k^i = \mathbf{A}_{\mathrm{GR}}(\boldsymbol{x}_k + \alpha_k^i \boldsymbol{d}_k)$ définie par (IV.37), page 69).

Supposons que l'Hypothèse 1, page 62 (resp. Hypothèses 1 et 2, page 62) sur la fonction de régularisation ϕ est satisfaite et que la suite de matrices de préconditionnement $\{\mathbf{M}_k\}$ est uniformément bornée.

Alors la séquence $\{\mathcal{J}(\boldsymbol{x}_k)\}$ est décroissante au sens large :

$$\mathcal{J}(\boldsymbol{x}_k) \geq \mathcal{J}(\boldsymbol{x}_{k+1}), \quad \forall k \tag{VI.12}$$

et on a la convergence dans le sens suivant :

$$\lim_{k\to\infty} \|\nabla\mathcal{J}(\boldsymbol{x}_k)\| = 0. \tag{VI.13}$$

Preuve. Voir l'annexe B.3.4, page 152. □

Le Théorème VI.3.1 permet, entre autres, de retrouver la convergence des algorithmes de GR et de GY. En effet, si on considère comme suite de matrices de préconditionnement $\mathbf{M}_k = \mathbf{A}^a_{\mathrm{GY}}$ ou $\mathbf{M}_k = \mathbf{A}_{\mathrm{GR}}(\boldsymbol{x}_k)$ avec une seule itération ($I = 1$) de la recherche de pas définie par (VI.6), page 84 alors d'après (VI.7), page 85 on retrouve bien les algorithmes de GR et de GY à pas constant. La séquence de préconditionnement définie à partir des matrices normales de GR et de GY est bien à spectre uniformément positif borné d'après le Lemme IV.4.2, page 67 qui reste valable pour le critère pénalisé généralisé.

L'intérêt du Théorème VI.3.1 est surtout de montrer que la relaxation du préconditionnement des algorithmes de GR et de GY (différent de $\mathbf{M}_k = \mathbf{A}^a_{\mathrm{GY}}$ ou $\mathbf{M}_k = \mathbf{A}_{\mathrm{GR}}(\boldsymbol{x}_k)$) fournit de nouveaux algorithmes convergents. Ce résultat présente un intérêt pratique évident en assurant la convergence d'algorithmes dont le préconditionnement peut être choisi uniquement sur la base de son efficacité. La seule condition, peu contraignante, est que la suite de matrices de préconditionnement soit à spectre uniformément positif borné. En particulier, le Théorème VI.3.1 assure la convergence de la forme de GY étendue proposée dans [Husse *et al.*, 2004], qui peut être vue comme une version avec préconditionnement efficace.

VI.3.5 Convergence avec conjugaison

Le Théorème suivant établit la convergence de la famille d'algorithmes GCPPR+SQ1D (avec conjugaison) pour la minimisation des critères pénalisés généralisés. A notre connaissannce, ce résultat de convergence n'a pas été établi auparavant. La convergence des algorithmes GCPPR+SQ1D n'est pas évoquée dans [Fessler et Booth, 1999], et la seule non croissance de la suite des itérées est établie dans [Rivera et Marroquin, 2003]. Notons que [Sun et Zhang, 2001] établit des résultats de convergence qui se révèlent être trop restrictifs.

Théorème VI.3.2. *Soit $\mathcal{J}$ le critère pénalisé généralisé défini par (IV.1)-(IV.2), page 54 où la condition suivante*

$$\ker(\mathbf{H}^{\mathrm{t}}\mathbf{H}) \cap \ker(\mathbf{V}^{\mathrm{t}}\mathbf{V}) = \{\mathbf{0}\}$$

est vérifiée. Soit $\boldsymbol{x}_k$ issu de l'algorithme GCPPR défini par (VI.1)-(VI.4), page 83 avec une suite de matrices de préconditionnement symétriques définies positives constante $\{\mathbf{M}_k\} = \{\mathbf{M}_0\}$ et où la recherche de pas définie par (VI.6), page 84 est utilisée avec $\theta \in]0, 2[$, $\mathbf{Q}^i_k = \mathbf{A}^a_{\mathrm{GY}}$ avec $0 < a < 1/L$ définie par (IV.35), page 69 (resp. $\mathbf{Q}^i_k = \mathbf{A}_{\mathrm{GR}}(\boldsymbol{x}_k + \alpha^i_k \boldsymbol{d}_k)$ définie par (IV.37), page 69).

Supposons que l'Hypothèse 1, page 62 (resp. Hypothèses 1 et 2, page 62) sur la fonction de régularisation ϕ est satisfaite et que la matrice $\mathbf{M}_0$ est symétrique définie positive.

Alors la séquence $\{\mathcal{J}(\boldsymbol{x}_k)\}$ est décroissante au sens large :

$$\mathcal{J}(\boldsymbol{x}_k) \geq \mathcal{J}(\boldsymbol{x}_{k+1}), \quad \forall k \tag{VI.14}$$

et on a la convergence dans le sens suivant : [1]

$$\liminf_{k\to\infty} \|\nabla \mathcal{J}(\boldsymbol{x}_k)\| = 0. \tag{VI.15}$$

Preuve. Voir l'annexe B.3.5, page 152. □

Il est à noter que la convergence lim inf (VI.15) du Théorème VI.3.2, plus faible que la convergence lim (VI.13) du Théorème VI.3.1, est typique des algorithmes GCNL [Hager et

[1] Soit une suite $\{u_k\}$ de scalaires réels. Soit $v_m = \inf\{u_k | k \geq m\}$. La limite de v_m est notée $\liminf_{k\to\infty} u_k$.

Zhang, 2006]. Cette convergence lim inf de la norme du gradient vers zéro implique que la proposition suivante

$$\exists \epsilon > 0,\ \forall k, \quad \|\nabla \mathcal{J}(\boldsymbol{x}_k)\| > \epsilon$$

est fausse : il n'existe pas de constante strictement positive, aussi petite soit-elle, minorant la suite des normes du gradient.

Le Théorème VI.3.2 ne porte que sur la convergence de la famille d'algorithmes GCPPR correspondant à la forme de conjugaison de Polak-Ribiere. En fait, d'après le Théorème C.4.1 de l'annexe C, page 167 on obtient aussi la convergence pour d'autres formes de conjugaison. Dans un souci de clarté, nous avons choisi de ne traiter dans cette section que la forme de Polak-Ribiere car il s'agit de la forme la plus populaire parmi les algorithmes GCNL [Nocedal et Wright, 1999, p. 121]. De plus, nos expériences numériques du chapitre VII montrent notamment que cette forme est plus efficace que les autres formes présentes dans l'annexe C.

VI.3.6 Relaxation de l'hypothèse de coercivité ?

Les Hypothèses 1 et 2, page 62 sur la fonction de régularisation ϕ supposent notamment sa coercivité. Ces Hypothèses excluent donc les fonctions de régularisation $\ell_2\ell_0$ présentées dans la section II.2.3.[B], page 37. On peut se demander s'il est possible d'obtenir des résultats de convergence en relaxant l'hypothèse de coercivité.

On peut montrer que le Théorème V.3.2, page 79 reste valide en relaxant l'hypothèse de coercivité car l'Hypothèse 1, page 147 ne nécessite que le caractère différentiable et borné inférieurement.

Par contre, ce n'est plus le cas du Théorème VI.3.2, page 89. La différence par rapport au Théorème V.3.2 réside dans l'introduction de l'hypothèse que les lignes de niveau définies par $L = \{\boldsymbol{x} \in \mathbb{R}^N | \mathcal{J}(\boldsymbol{x}) \leq \mathcal{J}(\boldsymbol{x}_0)\}$ sont bornées. Or le Lemme VI.3.3 ci-dessous montre que cette hypothèse n'est pas systématiquement vérifiée dans le cas d'une fonction de régularisation non coercive. Autrement dit, la possibilité d'obtenir les résultats de convergence du Théorème VI.3.2 pour des fonctions de régularisation non coercives est un problème ouvert.

Lemme VI.3.3. *Soit $\mathcal{J}$ un critère pénalisé généralisé défini par* (IV.1)-(IV.2)*, page 54. On suppose que la fonction de régularisation est $\ell_2\ell_0$ (asymptotiquement constante) et que* $\operatorname{Ker} \mathbf{H} \neq \{\mathbf{0}\}$.

Alors la proposition

$$\forall \boldsymbol{x}_0 \in \mathbb{R}^N, \quad \textit{les lignes de niveaux } L \textit{ sont bornées}$$

où $L = \{\boldsymbol{x} \in \mathbb{R}^N | \mathcal{J}(\boldsymbol{x}) \leq \mathcal{J}(\boldsymbol{x}_0)\}$ est fausse.

Preuve. Voir l'annexe B.3.6, page 153. □

VI.4 Conclusion

Un des intérêts des résultats de convergence de la famille d'algorithmes sans conjugaison (Théorème VI.3.1, page 88) et avec conjugaison (Théorème VI.3.2, page 89) par rapport aux résultats de convergence de la littérature de l'optimisation est l'assurance de la convergence pour n'importe quel nombre de sous-itérations de recherche du pas SQ1D. On peut donc choisir le nombre de sous-itérations avec l'efficacité, en terme de vitesse de convergence expérimentale, comme seule préoccupation. Nous étudierons expérimentalement dans le chapitre VII le choix

du nombre de sous-itérations assurant la meilleure efficacité. Nous verrons qu'il est préférable d'effectuer un très faible nombre de sous-itérations de recherche du pas SQ1D.

Le Théorème VI.3.2 se singularise par rapport au Théorème VI.3.1 en assurant la convergence pour une suite de matrices de préconditionnement constantes et non pas variables. En permettant la conjugaison on perd alors en généralité. La possibilité d'étendre la convergence de la famille d'algorithmes GCPPR+SQ1D avec conjugaison pour une suite de matrices de préconditionnement variables est à l'heure actuelle une question ouverte.

Le Lemme VI.3.3 montre que pour certaines initialisations, les lignes de niveaux ne sont pas bornées dans le cas d'une fonction de régularisation non coercive. Par contre, on peut se demander s'il est possible de trouver des conditions sur l'initialisation assurant que les lignes de niveaux sont bornées. Cependant, on aurait alors une restriction sur le domaine d'initialisation, ce qui n'est pas très satisfaisant. Il vaut peut être mieux voir s'il est possible d'obtenir une autre démonstration se passant de l'hypothèse que les lignes de niveaux sont bornées.

Bibliographie

[Al-Baali et Fletcher, 1996] M. Al-Baali et R. Fletcher. On the order of convergence of preconditioned nonlinear conjugate gradient methods. *SIAM J. Sci. Comput.*, 17 : 658–665, 1996.

[Bertsekas, 1999] D. P. Bertsekas. *Nonlinear programming.* Athena Scientific, Belmont, MA, USA, 2ème edition, 1999.

[Chan et Ng, 1996] R. H. Chan et M. K. Ng. Conjugate gradient methods for Toeplitz systems. *SIAM Rev.*, 38 (3) : 427–482, septembre 1996.

[Fessler et Booth, 1999] J. A. Fessler et S. D. Booth. Conjugate-gradient preconditionning methods for shift-variant PET image reconstruction. *IEEE Trans. Image Processing*, 8 (5) : 668–699, mai 1999.

[Fletcher et Reeves, 1964] R. Fletcher et C. M. Reeves. Function minimization by conjugate gradients. *Comp. J.*, 7 : 149–157, 1964.

[Hager et Zhang, 2006] W. W. Hager et H. Zhang. A survey of nonlinear conjugate gradient methods. *Pacific J. Optim.*, 2 (1) : 35–58, janvier 2006.

[Husse *et al.*, 2004] S. Husse, Y. Goussard et J. Idier. Extended forms of Geman and Yang algorithm : application to MRI reconstruction. In *Proc. IEEE ICASSP*, volume III, pages 513–516, Montréal, Québec, Canada, mai 2004.

[Ng *et al.*, 1999] M. K. Ng, R. H. Chan et W.-C. Tang. A fast algorithm for deblurring models with Neumann boundary conditions. *SIAM J. Sci. Comput.*, 21 (3) : 851–866, 1999.

[Nocedal et Wright, 1999] J. Nocedal et S. J. Wright. *Numerical optimization.* Springer Texts in Operations Research. Springer-Verlag, New York, NY, USA, 1999.

[Powell, 1984] M. J. D. Powell. *Nonconvex minimization calculations and the conjugate gradient method*, volume 1066 de *Lecture Notes in Mathematics.* Springer Verlag, Berlin, 1984.

[Press *et al.*, 1992] W. H. Press, S. A. Teukolsky, W. T. Vetterling et B. P. Flannery. *Numerical recipes in C, the art of scientific computing.* Cambridge Univ. Press, New York, 2ème edition, 1992.

[Rivera et Marroquin, 2003] M. Rivera et J. Marroquin. Efficient half-quadratic regularization with granularity control. *Image and Vision Comp.*, 21 (4) : 345–357, avril 2003.

[Sun et Zhang, 2001] J. Sun et J. Zhang. Global convergence of conjugate gradient methods without line search. *Annals of Operations Research*, 103 : 161–173, 2001.

Chapitre VII

COMPARAISON EXPÉRIMENTALE DE LA VITESSE DE CONVERGENCE

La convergence des familles d'algorithmes SQ+GCP et GCNL+SQ1D a été établie respectivement dans les chapitres V et VI. La question de leur efficacité se pose alors. En particulier, il s'agit d'évaluer leur efficacité par rapport aux algorithmes SQ exacts, mais également l'une par rapport à l'autre. Des études sur le taux de convergence des algorithmes de GR et de GY ont été menées dans [Allain *et al.*, 2006; Nikolova et Ng, 2005]. Elles laissent à penser que le taux de convergence de l'algorithme de GR est généralement plus grand que celui de l'algorithme de GY. Par contre, il peut arriver que l'algorithme GY présente une plus grande vitesse de convergence expérimentale (plus faible temps d'optimisation) que l'algorithme de GR [Nikolova et Ng, 2005]. Il n'est donc pas aisé d'assurer *a priori* la plus grande efficacité de l'algorithme de GR ou de GY. C'est pourquoi la vitesse de convergence expérimentale est une mesure d'efficacité pertinente.

Nous ne connaissons pas de comparaison entre les algorithmes SQ et GCNL, à l'exception de [Roullot *et al.*, 2004]. L'objet de ce chapitre est de mener une comparaison des vitesses de convergence expérimentale des algorithmes SQ+GCP et GCNL+SQ1D sur un problème de déconvolution d'image de grande taille. La question qui se pose naturellement est de voir sur la base de ces expériences si une des familles d'algorithmes est meilleure que l'autre. Le résultat de ces expériences est que ces deux familles d'algorithmes ont une vitesse de convergence expérimentale du même ordre. Plus précisément, il est toujours possible de trouver un couple d'algorithmes parmi ces deux familles dont l'un est soit meilleur que l'autre, soit moins bon. Plutôt que de privilégier une des deux familles, on pourrait envisager une forme d'hybridation entre ces deux familles algorithmes afin d'en tirer mieux parti.

VII.1 Problème de déconvolution d'image

On considère un problème de déconvolution d'image de grande taille. L'image à restaurer est une image convoluée et bruitée. Le noyau de convolution est gaussien avec un écart type de 2.24 et de taille 17×17. L'hypothèse de bord considérée est celle de Dirichlet (zéros à l'extérieur). Un bruit blanc gaussien est ajouté à l'image convoluée (RSB : 40dB). L'image d'origine est l'image `fishing boat` de taille $N = 512 \times 512$ (8 bits/pixel).

L'image est déconvoluée en utilisant une approche pénalisée. Nous utilisons le critère pénalisé $\mathcal{J}$ défini par (IV.1)-(IV.2) page 54 avec $P = 1$, ce qui correspond à un critère pénalisé non généralisé. Nous considérons la matrice des différences du premier ordre avec quatre voisins $\mathbf{V}$ pour le terme de pénalisation. Nous ne considérons pas de termes de rappels ($\omega_c = 0$). La condition

$$\ker(\mathbf{H}^{\mathrm{t}}\mathbf{H}) \cap \ker(\mathbf{V}^{\mathrm{t}}\mathbf{V}) = \{\mathbf{0}\} \tag{VII.1}$$

est alors bien vérifiée car le noyau de $\mathbf{V}^{\mathrm{t}}\mathbf{V}$ est réduit aux images constantes qui n'appartienent pas au noyau de $\mathbf{H}^{\mathrm{t}}\mathbf{H}$. On considère pour fonction de régularisation la fonction hyperbolique strictement convexe définie par $\phi_{\mathrm{hyp}}(t) = \sqrt{\delta^2 + |t|^2}$. Le critère pénalisé $\mathcal{J}$ est alors strictement convexe (voir la section IV.1.3.[A]). Il n'y a donc qu'un unique minimiseur du critère pénalisé $\mathcal{J}$.

Le critère pénalisé dépend des hyperparamètres λ et δ. Ils sont réglés afin que l'image déconvoluée soit la plus proche visuellement de l'image originale. Ainsi, leur valeur est de $\hat{\delta} = 13$ et $\hat{\lambda} = 0,2$.

VII.2 Algorithmes comparés

La fonction hyperbolique ϕ_{hyp} vérifie l'Hypothèse 2, page 62. Donc d'après (VII.1), les hypothèses du Théorème V.3.2, page 79 assurant la convergence de la famille d'algorithmes SQ+GCP sont vérifiées. De même, les hypothèses du Théorème VI.3.2, page 89 assurant la convergence de la famille d'algorithmes GCNL+SQ1D sont également vérifiées.

Les expériences ont été menées avec Matlab 7 sur un PC P4 2.8GHz RAM 1Go. L'initialisation des algorithmes est systématiquement faite avec l'image nulle ($\boldsymbol{x}_0 = \mathbf{0}$). On utilise comme critère d'arrêt pour tous les algorithmes comparés le seuil suivant sur la norme du gradient par pixel

$$\frac{\|\nabla \mathcal{J}(\boldsymbol{x}_k)\|}{\sqrt{N}} \leq 10^{-3}.$$

A présent, nous décrivons les algorithmes qui sont comparés.

VII.2.1 Algorithmes SQ+GCP

Nous considérons les algorithmes de GY et de GR approchés en utilisant l'algorithme GCP, éventuellement préconditionné, tronqué. L'algorithme GCP est utilisé avec la règle d'arrêt proposée dans [Nikolova et Ng, 2001] :

$$\|\boldsymbol{r}_i\| / \|\boldsymbol{r}_0\| < \eta, \tag{VII.2}$$

où $\boldsymbol{r}_i$ définie par (V.7), page 77 est le résidu de l'équation normale après i itérations. Il est à noter que cette règle d'arrêt correspond à une séquence $\{I_k\}$ non constante (V.3), page 76. Le paramètre η contrôle la précision de l'algorithme GCP. Plus il est faible, et plus la résolution

(a) Image convoluée et bruitée, RSB 40dB

(b) Image déconvoluée et débruitée

FIG. VII.1: Déconvolution de l'image `fishing boat` par une approche pénalisée préservant les discontinuités.

du sytème linéaire (V.4), page 77 est précise. Dans [Nikolova et Ng, 2001], le paramètre η est fixé à la faible valeur de 10^{-6}. Cette version correspond à une mise en œuvre très proche des algorithmes SQ exacts. Nous testons ici plusieurs valeurs de η dans la plage de valeurs $[10^{-6}, 1]$, qui a été choisie de manière empirique afin de permettre à chaque algorithme d'atteindre sa meilleure performance. La famille d'algorithmes qui en résulte est notée SQ+GCP(η).

Pour chaque algorithme testé, le tableau VII.1, page 98, affiche les nombres d'itérations sous la forme n_1/n_2, où n_1 est le nombre d'itérations globales, et n_2 est le nombre moyen d'itérations de l'algorithme GCP par itération globale. Le temps d'optimisation (en secondes) est aussi donné pour chaque algorithme.

VII.2.2 Algorithmes GCNL

Les algorithmes suivants sont comparés :

- GCPPR+GR1D(I) et GCPPR+GY1D(I) qui correspondent aux algorithmes GCPPR définis par (VI.1)-(VI.4), page 83 avec I itérations de la recherche du pas GR1D ou GY1D définies par (VI.6), page 84.
- GCPPR+SWOLFE(c_1, c_2) qui correspond à l'algorithme GCPPR défini par (VI.1)-(VI.4), page 83 avec pour recherche du pas celle correspondant aux algorithmes 3.2 et 3.3 de [Nocedal et Wright, 1999]. En particulier, cette recherche du pas satisfait les conditions fortes de Wolfe définies par (A.5)-(A.7), page 140. Ces conditions dépendent des paramètres c_1 et c_2 qui doivent vérifier la condition $0 < c_1 < c_2 < 1$. Cette recherche du pas est initialisée avec un pas unitaire. En suivant [Nocedal et Wright, 1999, Chap. 3] on sélectionne la valeur $c_1 = 10^{-4}$, et nous testons les valeurs $\{0.1, 0.5, 0.9\}$ pour le paramètre c_2.

Pour chaque algorithme, le tableau VII.2, page 99 affiche le nombre d'itérations nécessaire pour satisfaire la règle d'arrêt sous la forme $n_1/n_2/n_3$ où n_1 est le nombre d'itérations globales et n_2 est le nombre moyen d'évaluations du gradient par itération. Le paramètre n_3 a une signification différente selon l'algorithme considéré :

- n_3 s'identifie avec I pour les algorithmes GCPPR+GR1D(I) et GCPPR+GY1D(I),
- n_3 est le nombre moyen d'évaluations du critère par itération pour l'algorithme GCPPR+SWOLFE(c_1, c_2).

Le tableau VII.2 affiche aussi le temps mis de chaque algorithme pour satisfaire la règle d'arrêt.

VII.2.3 Préconditionnement

Il est proposé dans [Nikolova et Ng, 2005] d'utiliser l'algorithme du gradient conjugué (GC) préconditionné pour les matrices normales de GY proches d'une structure Toeplitz par blocs. Dans le problème de déconvolution d'image considéré, la matrice normale de GY est proche d'une structure Toeplitz-plus-Hankel. Ainsi, nous envisageons un préconditionnement basé sur la transformée en cosinus 2D (CT) car le noyau gaussien est symétrique [Ng *et al.*, 1999].

Bien que la matrice normale de GR ne présente pas une structure Toeplitz (à cause de sa dépendance au vecteur $\boldsymbol{x}_k$), nous proposons également d'utiliser pour l'algorithme GC un préconditionnement constant basé sur la transformée en cosinus 2D. Un tel préconditionnement pour la matrice normale de GR n'a pas été envisagé dans [Nikolova et Ng, 2005]. Cependant, lorsque l'image est composée d'un nombre suffisant de zones homogènes, ce préconditionnement peut se réveler efficace. C'est pourquoi nous avons choisi d'expérimenter ce préconditionnement pour la matrice normale de GR.

Nous utilisons également ce préconditionnement constant basé sur la transformée en cosinus 2D (CT) pour les algorithmes GCNL.

VII.3 Interprétation des résultats expérimentaux

L'algorithme de GY dépend du paramètre a_{GY}. D'après [Nikolova et Ng, 2005, remark 7], nous le fixons par défaut à la valeur $\hat{a}$ où

$$1/\hat{a} = \sup_{t\in\mathbb{R}} \phi''(t).$$

Ainsi, pour la fonction de régularisation hyperbolique $\phi_{\mathrm{hyp}}(t)$ considérée, on a

$$\hat{a} = 1/\phi''(0) = \delta. \tag{VII.3}$$

VII.3.1 Algorithmes SQ+GCP

L'influence du paramètre η des algorithmes SQ+GCP(η) contrôlant la précision de l'algorithme GCP est illustrée dans le tableau VII.1, page 98. Comme attendu, le nombre d'itérations globales n_1 diminue généralement lorsque η diminue, tandis que le nombre moyen d'itérations de l'algorithme GCP n_2 augmente. Le meilleur temps d'optimisation est obtenu lorsqu'un compromis est atteint entre le nombre d'itérations globales et le nombre d'itérations de l'algorithme GCP. Dans tous les cas, la meilleure valeur de η compte tenu du temps d'optimisation est comprise entre $0,1$ et 1, ce qui correspond à une résolution très approchée du système linéaire (V.4), page 77 par rapport à celle envisagée dans [Nikolova et Ng, 2001] ($\eta = 10^{-6}$). La charge calculatoire est alors réduite par un facteur compris entre deux et dix, ce qui correspond à un gain substantiel.

Le gain apporté par le préconditionnement CT est important pour l'algorithme GY+GCP comme attendu. Ce préconditionnement apporte aussi un gain important pour l'algorithme GR+GCP, ce qui était une question en suspens. D'ailleurs le gain en temps d'optimisation apporté par le préconditionnement CT est du même ordre pour l'algorithme GR+GCP que pour l'algorithme GY+GCP. Cette expérience montre qu'il peut être intéressant d'utiliser un préconditionnement pour l'algorithme GR+GCP.

Pour une même valeur η, l'algorithme GR+GCP(η) a systématiquement un temps d'optimisation inférieur à celui de l'algorithme GY+GCP(η). Le meilleur réglage de l'algorithme GR+GCP(η) a aussi un temps d'optimisation inférieur à celui de l'algorithme GY+GCP(η).

VII.3.2 Algorithmes GCNL

[A] Stratégie de recherche du pas

Les stratégies de recherche du pas SQ1D(I) et SWOLFE(c_1, c_2) sont comparées pour l'algorithme GCPPR dans le tableau VII.2, page 99. Les algorithmes GCPPR+GR1D(I) et GCPPR+GY1D(I) obtiennent de meilleurs résultats en terme de temps d'optimisation lorsque I est très petit ($I = 1$ ou $I = 2$). Augmenter I n'apporte pas d'amélioration dans la mesure où le nombre d'itérations globales reste presque constant. De plus, les algorithmes GCPPR+GR1D(I) et GCPPR+GY1D(I) ont des performances très proches, à l'exception du cas $I = 1$ avec préconditionnement où GCPPR+GR1D(1) nécessite moins d'itérations globales que GCPPR+GY1D(1).

L'algorithme GCPPR+SWOLFE(c_1, c_2) ne nécessite pas plus de trois évaluations du gradient et du critère par itération globale pour les trois couples de valeurs testées (c_1, c_2). Le lien entre les paramètres c_1, c_2 et le nombre d'évaluations n'est pas direct. Cependant, le nombre d'évaluation du gradient et du critère diminue lorsque c_2 augmente, car la condition forte de Wolfe devient alors moins restrictive. Dans le même temps, le nombre d'itérations augmente, parfois de manière considérable. Le meilleur choix des valeurs de (c_1, c_2) correspond à un compromis qui n'est pas aisé à trouver.

	$\theta = 1$			
	pas de precond.		CT precond.	
	Itérations	Temps (s)	Itérations	Temps (s)
GR+GCP(0.9)	37/3.1	250.5	31/1.1	132.4
GR+GCP(0.8)	26/4.1	217.5	28/1.2	126.7
GR+GCP(0.7)	21/4.8	200.0	19/1.6	101.3
GR+GCP(0.6)	18/5.6	<u>194.3</u>	15/2.1	93.8
GR+GCP(0.5)	15/7.3	200.0	14/2.3	94.3
GR+GCP(0.4)	14/8.1	205.3	12/2.6	<u>87.9</u>
GR+GCP(0.3)	13/10.2	233.8	11/3.1	93.0
GR+GCP(0.2)	12/12.0	249.3	10/3.8	99.4
GR+GCP(0.1)	11/16.3	305.5	10/5.2	127.6
GR+GCP(10^{-2})	10/32.0	523.3	9/9.9	204.6
GR+GCP(10^{-3})	9/48.2	701.1	9/14.9	297.9
GR+GCP(10^{-4})	9/65.1	945.4	9/19.7	387.9
GR+GCP(10^{-5})	9/82.4	1190.2	9/25.3	493.0
GR+GCP(10^{-6})	9/99.1	1426.3	9/30.8	596.3
	$\theta = 1,\ a_{\mathrm{GY}} = \hat{a}$			
	pas de precond.		CT precond.	
	Itérations	Temps (s)	Itérations	Temps (s)
GY+GCP(0.9)	55/2.8	300.7	59/1.0	185.3
GY+GCP(0.8)	44/3.3	275.4	38/1.2	136.8
GY+GCP(0.7)	34/3.9	244.0	30/1.6	132.4
GY+GCP(0.6)	31/4.3	<u>241.1</u>	27/1.8	<u>128.8</u>
GY+GCP(0.5)	29/4.7	246.3	28/2.0	148.6
GY+GCP(0.4)	30/6.1	323.6	29/2.4	175.5
GY+GCP(0.3)	29/7.3	364.5	28/2.7	186.7
GY+GCP(0.2)	28/8.2	390.4	32/3.0	233.5
GY+GCP(0.1)	27/11.3	510.7	27/3.7	233.7
GY+GCP(10^{-2})	26/20.7	874.4	26/6.9	396.0
GY+GCP(10^{-3})	26/31.4	1313.9	25/10.8	577.8
GY+GCP(10^{-4})	26/42.2	1755.8	25/14.6	775.3
GY+GCP(10^{-5})	25/53.4	2128.7	25/18.6	970.7
GY+GCP(10^{-6})	25/64.7	2574.7	25/22.6	1184.3

TAB. VII.1 – Comparaison des algorithmes SQ+GCP pour le problème de déconvolution d'image. Le meilleur temps (s) d'optimisation de chaque algorithme est souligné.

Finalement, les algorithmes GCPPR+SQ1D(I) présentent plusieurs caractéristiques intéressantes par rapport à l'algorithme GCPPR+SWOLFE(c_1, c_2). Leur vitesse de convergence expérimentale est au moins aussi élevée que celle de l'algorithme GCPPR+SWOLFE(c_1, c_2), mais sans avoir à régler des paramètres cruciaux. De plus, ils sont aussi plus faciles à implémenter dans la mesure où la recherche du pas se résume à l'application d'un schéma itératif SQ1D analytique.

	$\theta = 1, a_{\mathrm{GY}} = \hat{a}$			
	pas de precond.		CT precond.	
	Itérations	Temps (s)	Itérations	Temps (s)
GCPPR+GR1D(1)	76/1/1	<u>110.7</u>	24/1/1	<u>46.9</u>
GCPPR+GR1D(2)	79/1/2	120.9	25/1/2	50.2
GCPPR+GR1D(5)	80/1/5	140.4	27/1/5	60.9
GCPPR+GR1D(10)	80/1/10	169.2	27/1/10	71
GCPPR+GY1D(1)	76/1/1	<u>110.8</u>	29/1/1	56.6
GCPPR+GY1D(2)	79/1/2	120.7	24/1/2	<u>48.5</u>
GCPPR+GY1D(5)	80/1/5	139.7	25/1/5	56.4
GCPPR+GY1D(10)	80/1/10	170.1	27/1/10	72.1
GCPPR+SWOLFE($10^{-4}, 0.1$)	82/2.52/2.52	248.1	27/1.89/1.89	74.5
GCPPR+SWOLFE($10^{-4}, 0.5$)	81/1.77/1.78	<u>172.5</u>	30/1.07/1.17	54.5
GCPPR+SWOLFE($10^{-4}, 0.9$)	206/1.41/1.41	352.3	30/1/1.77	<u>53.3</u>

Tab. VII.2 – Comparaison des stratégies de recherche du pas de l'algorithme GCPPR pour le problème de déconvolution d'image. Le meilleur temps (s) d'optimisation de chaque algorithme est souligné.

[B] Autres formules de conjugaison

Le Théorème C.4.1, page 167 est valable pour d'autres formules de conjugaison que celle de Polak-Ribiere. On peut donc s'interroger sur l'influence de la formule de conjugaison pour les algorithmes GCNL. Dans le Tableau VII.3, page 99 on compare l'influence des formules de conjugaison des formes de Polak-Ribiere, Fletcher-Reeves, Hestenes-Stiefel et Liu-Storey (voir l'annexe C, page 155). On considère la même recherche du pas GR1D(1) pour toutes ces formules de conjugaison. Aucune de ces formes ne fait mieux que celle de Polak-Ribiere. Cette constatation expérimentale va dans le sens de [Nocedal et Wright, 1999] et [Bertsekas, 1999] qui insistent sur la forme de Polak-Ribiere.

	$\theta = 1$			
	pas de precond.		CT precond.	
	Itérations	Temps (s)	Itérations	Temps (s)
GCPPR+GR1D(1)	76	110.7	24	<u>47.3</u>
GCPHS+GR1D(1)	75	109.8	24	47.6
GCPFR+GR1D(1)	75	<u>109.2</u>	49	96.9
GCPLS+GR1D(1)	76	111.1	57	112.4

Tab. VII.3 – Influence de la formule de conjugaison pour les algorithmes GCNL+GR1D pour le problème de déconvolution d'image. Le meilleur temps (s) d'optimisation est souligné.

VII.3.3 Influence des paramètres θ et a_{GY}

Les algorithmes de GR et de GY dépendent respectivement du paramètre θ et du couple de paramètres θ et a_{GY} (section IV.4.6.[A], page 66). Ainsi, les algorithmes GCNL+SQ1D et SQ+GCP dépendent aussi de ces paramètres. Jusqu'ici nous avons fixé ces paramètres aux valeurs $\theta = 1$ et $a_{\mathrm{GY}} = \hat{a}$ (VII.3). Or, d'après le Théorème V.3.2, page 79 assurant la convergence de la famille d'algorithmes SQ+GCP et le Théorème VI.3.2, page 89 assurant la convergence de la famille d'algorithmes GCNL+SQ1D, la convergence est assurée pour les valeurs $\theta \in]0, 2[$ et $a_{\mathrm{GY}} \in]0, \hat{a}[$. Dans [Allain *et al.*, 2006], un domaine de convergence étendu est obtenu pour le paramètre a_{GY} de l'algorithme de GY qui correspond à $\theta \in]0, 2[$ et $a_{\mathrm{GY}} \in]0, 2\hat{a}/\theta[$. En fait, on peut montrer que les résultats de convergence portant sur les algorithmes GCNL+GY1D et GY+GCP sont également valables sur ce domaine étendu. A présent, nous testons l'influence de ces paramètres sur la vitesse de convergence expérimentale des algorithmes GCNL+SQ1D et SQ+GCP.

Dans l'annexe D, page 171 on teste pour le paramètre de sur ou sous-relaxation θ les valeurs $\{0.5, 0.75, 0.9, 1.1, 1.25, 1.5\}$.

Le Tableau D.1, page 171 montre que pour l'algorithme GCPPR+GR1D(1) on ne fait pas mieux qu'avec $\theta = 1$. D'ailleurs, plus on s'écarte de la valeur $\theta = 1$ et plus c'est contre-productif. Par contre, pour l'algorithme GCPPR+GY1D(1) avec $a_{\mathrm{GY}} = \hat{a}$, la valeur $\theta = 1.1$ permet d'obtenir un léger gain de la vitesse de convergence expérimentale par rapport à la valeur $\theta = 1$. La plage de valeur proche de $\theta = 1$ apporte donc pour ce problème de déconvolution d'image la meilleure vitesse de convergence expérimentale pour les algorithmes GCPPR+SQ1D(1).

Le Tableau D.2, page 172 montre que pour l'algorithme GR+GCP(η) la plage de valeurs $\theta \in \{1.1, 1.25, 1.5\}$ correspondant à une sur-relaxation de θ permet d'obtenir un très léger gain de la vitesse de convergence expérimentale par rapport à $\theta = 1$. Par contre, la sous-relaxation de θ est contre-productive.

Le réglage de l'algorithme GY+GCP(η) est plus délicat que celui de l'algorithme GR+GCP(η) car il dépend de deux paramètres θ et a_{GY}. Nous avons choisi de fixer le paramètre θ à $\theta = 1$ et de faire varier le paramètre a_{GY} dans la plage de valeurs $a_{\mathrm{GY}} \in \{0.9, 1, 1.1, 1.25, 1.5, 1.75\}\ \hat{a}$. Le Tableau D.3, page 173 montre un léger gain de la vitesse de convergence pour la plage de valeurs $a_{\mathrm{GY}} \in \{1.1, 1.25, 1.5\}\ \hat{a}$ par rapport à la valeur $a_{\mathrm{GY}} = \hat{a}$. Par contre, les valeurs $0.9\hat{a}$ et $1.75\hat{a}$ sont contre-productives.

VII.3.4 Comparaison entre algorithmes SQ+GCP et GCNL

Dans [Roullot *et al.*, 2004] l'algorithme de GY est expérimentalement comparé à l'algorithme GCPFR (forme de Fletcher-Reeves) sans préconditionnement. La stratégie de recherche du pas utilisée n'est pas évoquée. Les résultats expérimentaux qui y sont présentés indiquent que l'algorithme de GY est plus efficace que l'algorithme GCPFR sans préconditionnement. Pour notre part, d'après le Tableau VII.1 et le Tableau VII.3, l'algorithme GCPFR+GR1D(1) sans préconditionnement a une vitesse de convergence expérimentale similaire à l'algorithme GY+GCP avec préconditionnement. Cependant, nous ne sommes pas dans la même situation que [Roullot *et al.*, 2004]. En effet, nous considérons ici un problème de déconvolution avec hypoyhèse de bords de Dirichlet qui ne permet pas d'inverser la matrice normale de GY par transformées de Fourier rapides comme c'est le cas dans [Roullot *et al.*, 2004].

Dans le Tableau VII.4, page 101 nous récapitulons les performances des algorithmes SQ+GCP et GCNL avec leur meilleur réglage. Les algorithmes GCNL+SQ1D avec leur meilleur réglage sont plus performants que les algorithmes SQ+GCP avec leur meilleur réglage. Néanmoins, l'écart du temps d'optimisation est inférieur à un facteur deux. Ces algorithmes présentent donc une performance de même ordre de grandeur. D'ailleurs, d'après le Tableau VII.5, ce constat

est préservé lorsqu'on considère d'autres critères pénalisés en faisant varier les hyperparamètres δ et λ autour des valeurs $\hat{\delta}$ et $\hat{\lambda}$.

Ce qui diffère véritablement entre les algorithmes SQ+GCP et GCNL c'est la simplicité de réglage. Bien qu'ils dépendent tous les deux d'un paramètre, soit respectivement la précision de la résolution du sytème linéaire et le nombre d'itérations de recherche du pas SQ1D, leur facilité de réglage n'est pas la même. En effet, il est apparu expérimentalement que le meilleur réglage pour I se trouve à une des extrémités de sa plage de valeur ($I = 1$), ce qui n'est pas le cas pour le meilleur réglage du paramètre η (autour de $\eta = 0.5$).

D'après le Tableau VII.4, l'algorithme GR+GCP préconditionné présente un nombre d'itérations plus faibles que l'algorithme GCPPR+GR1D préconditionné. On pourrait alors envisager une hybridation entre ces deux derniers algorithmes. Il s'agirait en somme de considérer un algorithme GCPPR+GR1D avec un préconditionnement variable de type GR+GCP.

	$\theta = 1,\ a_{\mathrm{GY}} = \hat{a}$			
	pas de precond.		CT precond.	
	Itérations	Temps (s)	Itérations	Temps (s)
GR+GCP(0.4)	14/8.1	205.3	12/2.6	87.9
GY+GCP(0.6)	31/4.3	241.1	27/1.8	128.8
GCPPR+GR1D(1)	76/1	<u>110.7</u>	24/1	<u>46.9</u>
GCPPR+GY1D(1)	76/1	110.8	29/1	56.6

Tab. VII.4 – Récapitulatif des algorithmes SQ+GCP et GCPPR+SQ1D avec leur meilleur réglage pour le problème de déconvolution d'image. Le meilleur temps (s) d'optimisation est souligné.

	pas de precond.		CT precond.	
	Itérations	Temps (s)	Itérations	Temps (s)
	$\delta = \hat{\delta}$ et $\lambda = \hat{\lambda}$			
GR+GCP(0.4)	14/8.1	205.3	12/2.6	87.9
GCPPR+GR1D(1)	76/1	<u>110.7</u>	24/1	<u>46.9</u>
	$\delta = 0.5\hat{\delta}$ et $\lambda = 0.5\hat{\lambda}$			
GR+GCP(0.4)	19/9.8	330.1	18/2.6	134.4
GCPPR+GR1D(1)	112/1	<u>165.5</u>	31/1	<u>61.8</u>
	$\delta = 0.5\hat{\delta}$ et $\lambda = 2\hat{\lambda}$			
GR+GCP(0.4)	28/5.7	304.8	30/3.4	268.8
GCPPR+GR1D(1)	73/1	<u>108.1</u>	53/1	<u>105.5</u>
	$\delta = 2\hat{\delta}$ et $\lambda = 0.5\hat{\lambda}$			
GR+GCP(0.4)	13/7.6	183	13/2.8	101.2
GCPPR+GR1D(1)	82/1	<u>120.6</u>	33/1	<u>66</u>
	$\delta = 2\hat{\delta}$ et $\lambda = 2\hat{\lambda}$			
GR+GCP(0.4)	12/4.9	117.2	12/2	74
GCPPR+GR1D(1)	52/1	<u>76.4</u>	20/1	<u>39.8</u>

Tab. VII.5 – Influence de la valeur des hyperparamètres. Comparaison des algorithmes GR+GCP et GCPPR+GR1D avec $\theta = 1$ pour le problème de déconvolution d'image. Le meilleur temps (s) d'optimisation est souligné.

VII.4 Conclusion

Nous avons expérimentalement montré qu'il est largement préférable pour les familles d'algorithmes SQ+GCP et GCNL+SQ1D de ne pas effectuer un nombre trop grand de sous-itérations. De plus, les algorithmes SQ+GCP ne devraient pas être vus comme des versions approchées des algorithmes SQ, mais bien comme des algorithmes propres, avec leur paramètre de réglage.

Si les familles d'algorithmes SQ+GCP et GCNL+SQ1D ont une vitesse de convergence expérimentale de même ordre, elles se distinguent néanmoins par leur simplicité de réglage. En effet, nous avons été capable de trouver facilement un réglage efficace pour les algorithmes GCNL+SQ1D ($I = 1$). Par contre, nous n'avons pas trouvé de réglage simple et efficace pour les algorithmes SQ+GCP. Dans nos expériences, il est apparu que les versions utilisant la forme de GR ne sont pas moins efficaces que celles utilisant la forme de GY. L'intérêt majeur des premières formes est leur plus grande simplicité de mise en œuvre. En effet, elles ne dépendent que d'un seul paramètre de réglage (θ) alors que les algorithmes utilisant la forme de GY dépendent de deux paramètres de réglage (θ et a_{GY}).

A l'heure actuelle, l'algorithme qui se démarque le plus des autres par son efficacité et sa simplicité est l'algorithme GCPPR+GR1D(1) avec pour paramètre $\theta = 1$. Cet algorithme nous semble être une référence à considérer systématiquement avant tout autre algorithme ou réglage. Cependant, on pourrait envisager comme perspective une hybridation entre les algorithmes SQ+GCP et GCNL+SQ1D. L'intérêt résiderait dans la possibilité que cet algorithme hybride présente des performances meilleures que ses deux parents.

Bibliographie

[Allain *et al.*, 2006] M. Allain, J. Idier et Y. Goussard. On global and local convergence of half-quadratic algorithms. *IEEE Trans. Image Processing*, 15 (5) : 1130–1142, mai 2006.

[Bertsekas, 1999] D. P. Bertsekas. *Nonlinear programming.* Athena Scientific, Belmont, MA, USA, 2ème edition, 1999.

[Ng *et al.*, 1999] M. K. Ng, R. H. Chan et W.-C. Tang. A fast algorithm for deblurring models with Neumann boundary conditions. *SIAM J. Sci. Comput.*, 21 (3) : 851–866, 1999.

[Nikolova et Ng, 2001] M. Nikolova et M. Ng. Fast image reconstruction algorithms combining half-quadratic regularization and preconditioning. In *Proc. IEEE ICIP*, pages 277–280, Thessaloniki, Grèce, octobre 2001.

[Nikolova et Ng, 2005] M. Nikolova et M. Ng. Analysis of half-quadratic minimization methods for signal and image recovery. *SIAM J. Sci. Comput.*, 27 : 937–966, 2005.

[Nocedal et Wright, 1999] J. Nocedal et S. J. Wright. *Numerical optimization.* Springer Texts in Operations Research. Springer-Verlag, New York, NY, USA, 1999.

[Roullot *et al.*, 2004] E. Roullot, A. Herment, I. Bloch, A. de Cesare, M. Nikolova et E. Mousseaux. Modeling anisotropic undersampling of magnetic resonance angiographies and reconstruction of a high-resolution isotropic volume using half-quadratic regularization techniques. *Signal Processing*, 84 : 743–762, avril 2004.

Chapitre VIII

DÉCONVOLUTION 2D POUR LE CONTRÔLE NON DESTRUCTIF PAR ULTRASONS

Le travail présenté dans ce chapitre a été initié dans le cadre d'un contrat de recherche entre l'IRCCyN et le GSDTI d'EDF (voir l'introduction I.2.[B], page 24). La méthode de restauration proposée dans ce chapitre pour le contrôle non destructif par ultrasons a répondu aux attentes exprimées par le GSDTI d'EDF. Il s'agissait d'une part de proposer une extension 2D de la méthode de déconvolution 1D de [Gautier *et al.*, 2001]. Cependant, cette extension n'est pas directe et il a fallu mettre en œuvre une procédure de recalage pour en assurer la robustesse. D'autre part, il s'agissait d'investir sur le problème d'optimisation résultant de la méthode de restauration. La méthode d'optimisation considérée par EDF avant cette collaboration consistait en un algorithme du gradient conjugué non linéaire non préconditionné. Nous en avons proposé une version préconditionnée qui permet d'obtenir un gain considérable de la vitesse de convergence. La convergence de cet algorithme est assurée par les résultats du chapitre VI.

La méthode considérée dans ce chapitre n'a pas été proposée dans la littérature. De plus, nous abordons l'estimation de la hauteur de défauts dans le cadre du contrôle non destructif par ultrasons qui n'est pas, de manière surprenante, traité dans la littérature. Nous nous basons pour cela sur une généralisation de la méthode proposée dans le brevet [Gautier *et al.*, 2002].

La méthode proposée, illustrée sur des données réelles, se révèle simple et robuste. Elle permet d'une part d'augmenter la résolution temporelle et latérale des images ultrasonores et d'en faciliter ainsi l'interprétation par un expert. D'autre part, elle répond de manière très satisfaisante au problème crucial de l'estimation de la hauteur de défauts de petite taille.

VIII.1 Introduction

VIII.1.1 Contrôle non destructif par ultrasons

Sous l'effet de contraintes mécaniques et du vieillissement affectant les pièces métalliques, des défauts surviennent. Le contrôle non destructif (CND) est un ensemble de méthodes qui permet de caractériser l'état d'intégrité des structures industrielles, sans les dégrader. L'inspection ultrasonore est une des principales techniques de CND pour de nombreuses applications industrielles. Dans l'industrie nucléaire, par exemple, le CND par ultrasons est une méthode particulièrement adaptée pour vérifier l'intégrité de larges conduits en acier de grande dimension, sous pression. L'inspection de ces conduits s'effectue pendant leur utilisation, ce qui présente un intérêt économique évident. Un transducteur ultrasonore fonctionnant en émetteur et récepteur est déplacé le long de la surface extérieure du conduit. Le transducteur émet une *ondelette* (une brève émission) qui se propage dans le milieu constitutif du conduit. Cette ondelette est partiellement réfléchie dès qu'elle rencontre une inhomogénéité. Le signal réfléchi est reçu par le même transducteur comme une fonction du temps appelée un *A-scan*.

Ici, on s'intéresse à l'inspection de blocs en acier soudés comme illustré dans la figure VIII.1. Dans le cas de la présence d'une entaille au fond de la soudure, deux types d'échos sont mesurés [Sallard et Paradis, 1998; Gautier *et al.*, 2001; Chalmond *et al.*, 2003]. Un écho de *coin* se propage jusqu'au transducteur après deux réflexions au pied de l'entaille. Un autre écho est émis par *diffraction* au sommet de l'entaille. La figure VIII.2(a) illustre un A-scan et la figure VIII.2(b) illustre un B-scan (*i.e.,* un ensemble d'A-scans obtenus pour des positions successives du transducteur, affichées en colonnes dont l'intensité est codée en niveau de gris). Les lignes inclinées indiquent la présence d'une entaille. Les lignes du dessus correspondent à l'écho de diffraction. En ne tenant pas compte des échos parasites, les autres lignes sont dûes à l'écho de coin. Ceci illustre deux types de difficultés liées à l'interprétation visuelle des B-scans bruts :

1. Une faible résolution temporelle : les échos de diffraction et de coin peuvent se superposer selon l'axe temporel lorsque la hauteur de l'entaille est faible devant la longueur d'onde de l'ondelette propagée
2. Une faible résolution latérale : les deux échos se retrouvent parmi plusieurs A-scans à cause de la largeur du faisceau ultrasonore.

La présence du bruit est une autre source de difficultés. Il est souvent dû à la chaîne de mesure, mais aussi à la structure du matériau inspecté. Par conséquent, l'estimation de l'amplitude et du temps d'arrivée des échos réfléchis est loin d'être évidente.

La géométrie du matériau inspecté peut être entièrement caractérisée par une réflectivité composée de pics localisés indiquant les présences d'inhomogénéités. L'intérêt étant par exemple d'estimer la hauteur d'un défaut éventuel. Tout le problème consiste alors à estimer la réflectivité à partir des mesures qui en sont les versions bruitées et filtrées. Nous sommes alors confrontés à un problème de déconvolution.

VIII.1.2 Etat de l'art

[A] Déconvolution

Les méthodes de déconvolution ont été l'objet d'une attention soutenue pour les données ultrasonores [Fatemi et Kak, 1980; Jensen, 1992]. Plusieurs auteurs ont publié des résultats pour la déconvolution 1D latérale permettant d'améliorer la résolution latérale à l'aide d'un noyau de convolution 1D représentant la largeur du faisceau le long de l'axe latéral [Hundt et Trautenberg, 1980; Vollmann, 1982; Schomberg *et al.*, 1983]. D'un autre côté, la déconvolution 1D temporelle

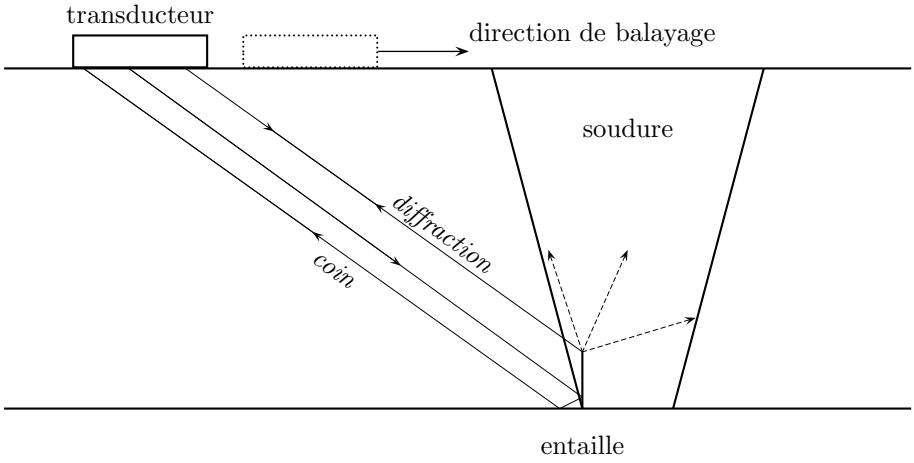


FIG. VIII.1: Procédure d'inspection d'un bloc d'acier soudé.

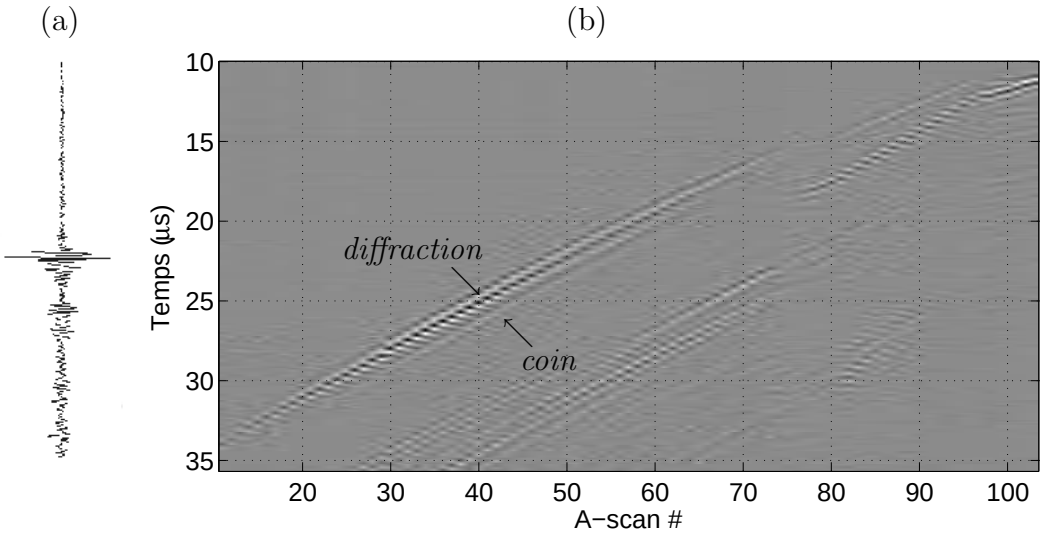


FIG. VIII.2: (a) Exemple d'un A-scan. (b) Exemple d'un B-scan avec un écho de diffraction et de coin. Les autres échos sont des artefacts.

permettant d'améliorer la résolution temporelle a aussi été étudiée. Elle est basée sur un noyau de convolution 1D qui modélise l'ondelette ultrasonore se propageant dans le milieu [Demoment *et al.*, 1984; Hayward et Lewis, 1989; Sallard et Paradis, 1998; Kaaresen et Bølviken, 1999; O'Brien *et al.*, 1994; Honarvar *et al.*, 2004]. Dans [Champagnat et Idier, 1993], un déphasage est observé entre les échos de diffraction et de coin, ce qui ne peut pas être expliqué par les méthodes de déconvolution temporelle classiques. La raison d'un tel déphasage pourrait être dûe à la nature de l'interaction entre le faisceau ultrasonore et le défaut. Une nouvelle méthode de déconvolution temporelle a alors été introduite dans [Champagnat et Idier, 1993], qui se base sur la transformée de Hilbert de l'ondelette pour pouvoir en décrire les versions déphasées. Le nouveau problème de déconvolution fait intervenir une réflectivité complexe. Un tel modèle a pour avantage de permettre la conservation de la linéarité du modèle de convolution. Par la suite, de nombreux auteurs ont adopté le même modèle de convolution avec réflectivité complexe

pour rendre compte des déphasages de l'ondelette [Faur *et al.*, 1998; Gautier *et al.*, 2001; Demirli et Saniie, 2001a,b; Mu *et al.*, 2002; Chalmond *et al.*, 2003].

L'augmentation de résolution peut être obtenue temporellement et latéralement en utilisant des méthodes de déconvolution 2D [Vollmann, 1982; Sciacca et Evans, 1992; Abeyratne *et al.*, 1995; Husby *et al.*, 2001; Taxt et Jirik, 2004]. Dans ces dernières contributions, le faisceau ultrasonore est supposé être orienté perpendiculairement au plan de balayage. L'ondelette 2D utilisée dans les méthodes de déconvolution est alors exprimée comme la convolution d'un noyau 1D et de l'ondelette précédemment définie. A notre connaissance, aucun modèle de convolution 2D rendant compte des déphasages de l'ondelette n'a été proposé.

[B] Déconvolution par approche pénalisée

Les méthodes de déconvolution ont été l'objet d'une attention soutenue pour les données ultrasonores [Fatemi et Kak, 1980; Jensen, 1992]. Les problèmes de déconvolution sont des problèmes mal posés et une régularisation s'impose afin d'obtenir une solution acceptable [Tikhonov et Arsenin, 1977]. La plupart des efforts pour régulariser les problèmes mal posés se sont concentrés sur l'utilisation d'une connaissance *a priori* de la solution [Idier, 1999]. Dans le cadre du CND par ultrasons, la réflectivité recherchée est classiquement supposée être composée de seulement quelques pics, dont les positions et les amplitudes sont inconnues. Ainsi, plusieurs contributions se basent sur un modèle de séquence d'impulsions parcimonieuses pour décrire la réflectivité [Champagnat et Idier, 1993; Sallard et Paradis, 1998; Kaaresen, 1998]. La plupart des contributions plus anciennes sont basées sur le filtrage de Wiener [Fatemi et Kak, 1980; Jeurens *et al.*, 1987]. Une variante basée sur le filtrage de Kalman se trouve aussi dans [Demoment *et al.*, 1984]. Cette simple approche linéaire est restée populaire jusqu'à présent [Abeyratne *et al.*, 1997; Taxt et Jirik, 2004; Honarvar *et al.*, 2004]. Il est à noter que la solution du filtre de Wiener peut être vue comme le minimiseur d'un critère des moindres carrés pénalisés, où la fonction de régularisation est le carré de la norme l_2 [Hunt, 1977]. Une fonction de régularisation utilisant la norme l_1 fournit de bien meilleurs résultats, mais au prix d'un coût de calcul supérieur [Hayward et Lewis, 1989; O'Brien *et al.*, 1994]. Un autre inconvénient du cadre l_1, est que l'unicité de la solution n'est pas assurée car la norme l_1 n'est pas strictement convexe. C'est pourquoi nous avons choisi une fonction de régularisation l_2l_1, qui assure le caractère strictement convexe et différentiable (contrairement à la norme l_1). Il s'agit de la fonction hyperbolique $\phi_{\text{hyp}}(t) = \sqrt{\delta^2 + t^2}$ qui est utilisée dans le cadre de la déconvolution ultrasonore 1D [Gautier *et al.*, 2001].

L'utilisation de fonctions de régularisation l_2l_1 a été appliquée avec succès en déconvolution d'image préservant les discontinuités [Charbonnier *et al.*, 1997]. En considérant un B-scan comme une image, la préservation des discontinuités est similaire à la restitution du caractère marqué des pics de la réflectivité. Dans [Champagnat et Idier, 1993], un déphasage est observé entre les échos de diffraction et de coin, qui ne peut pas être expliqué par les méthodes de déconvolution classiques. Par contre, l'introduction d'une réflectivité complexe tenant compte des déphasages de l'ondelette permet de conserver la linéarité du modèle de convolution [Champagnat et Idier, 1993]. Par la suite, de nombreux auteurs ont adopté le même modèle de convolution avec réflectivité complexe pour rendre compte des déphasages de l'ondelette [Faur *et al.*, 1998; Gautier *et al.*, 2001; Demirli et Saniie, 2001a,b; Mu *et al.*, 2002; Chalmond *et al.*, 2003].

La méthode proposée dans ce document pour l'imagerie ultrasonore (B-scans) est une extension 2D de la méthode de déconvolution 1D de [Gautier *et al.*, 2001] tenant compte des déphasages de l'ondelette avec une fonction de régularisation l_2l_1.

VIII.1.3 Contributions

La méthode proposée ici consiste à effectuer la déconvolution 2D en estimant la réflectivité complexe rendant compte des déphasages de l'ondelette via la minimisation d'un critère pénalisé avec une fonction convexe l_2l_1 comme fonction de régularisation. L'ondelette 2D du modèle de convolution est identifiée à partir d'un écho isolé d'un bloc test. Cette identification se base sur la possibilité de négliger la dépendance par rapport à la profondeur de l'ondelette pour les petits objets [Fatemi et Kak, 1980].

La méthode proposée nécessite l'utilisation d'un algorithme de minimisation. Bien qu'une fonction de régularisation l_2l_1 permette une plus grande augmentation de la résolution, elle induit un coût de calcul plus grand que le filtrage de Wiener. Il est donc primordial d'utiliser un algorithme de minimisation convergent et efficace. Pour les critères convexes, les méthodes du gradient conjugué (GC) sont intéressantes pour leur simplicité et leur bon compromis entre taux de convergence et vitesse de convergence expérimentale [Bertsekas, 1999]. Nous considérons aussi un préconditionnement circulaire qui permet une amélioration notable du taux de convergence des méthodes GC. Les préconditionneurs circulaires sont particulièrement séduisants dans la mesure où on peut utiliser la transformée de Fourier rapide pour une implémentation efficace [Chan et Ng, 1996]. De plus, on ne peut pas utiliser la transformée en cosinus car l'ondelette n'est pas symétrique, comme illustré sur la figure VIII.7(b), page 119.

La méthode proposée ne se contente pas d'effectuer une déconvolution 2D, mais effectue un recalage préalable des A-scans. En effet, nous avons observé un certain nombre de décalages entre A-scans. Afin de les traiter, nous introduisons une procédure empirique de recalage réversible où les A-scans consécutifs sont recalés selon une approche basée sur le maximum de corrélation. Cette procédure met les échos à l'horizontale ce qui permet alors aussi d'utiliser un modèle 2D séparable pour l'ondelette sous la forme de la convolution de l'ondelette 1D et du profil latéral 1D du faisceau. Ceci est rendu possible grâce à cette procédure de recalage alors même que le faisceau ultrasonore n'est pas orienté perpendiculairement au plan de balayage. Ce modèle séparable pour l'ondelette a pour intérêt d'accroître la qualité de l'estimation de la réflectivité.

Les résultats expérimentaux dans le cadre de l'inspection ultrasonore de blocs d'acier soudés montrent la pertinence de la méthode proposée pour résoudre le problème de la superposition des échos de coin et de diffraction. La méthode 2D proposée fournit une plus grande augmentation de la résolution que l'approche 1D traitée dans [Gautier *et al.*, 2001]. La possibilité d'estimer la hauteur de défaut à partir des résultats de la méthode proposée est aussi illustrée expérimentalement. Nous n'avons pas trouvé de référence dans la littérature sur l'estimation de la hauteur de défaut dans le cadre CND par ultrasons. L'estimation de la hauteur du défaut utilisée ici se base sur une généralisation de la méthode proposée dans [Gautier *et al.*, 2002].

VIII.2 Méthode proposée

VIII.2.1 Modèle direct

[A] A-scan

Dans cette section, le modèle direct des A-scans est présenté. Considérons tout d'abord un modèle continu. Les mesures dépendant du temps $z(t)$ sont supposées être issues d'une convolution 1D entre la réflectivité dépendant de la hauteur dans le milieu $r(y)$ et l'ondelette dépendant du temps $h(t)$ qui se propage avec une célérité constante C. Ainsi, l'interaction entre le milieu du matériau et les ultrasons se modélise simplement par la convolution bruitée suivante

[Fatemi et Kak, 1980]

$$z(t) = \int h(t')r(C(t-t'))dt' + n(t) \tag{VIII.1}$$

où $n(t)$ est un bruit blanc gaussien de variance σ^2 qui représente le bruit de mesure et les erreurs de modèle. Dans un cadre entièrement discret, le modèle (VIII.1) devient

$$z[k] = \sum_{i=1}^{I} h[i]r[k-i] + n[k], \quad k = 1, \ldots, M \tag{VIII.2}$$

où les A-scan discrets $z[k]$ sont de taille M. L'équation (VIII.2) peut s'écrire sous la forme suivante

$$\boldsymbol{z} = \boldsymbol{h} \star_1 \boldsymbol{r} + \boldsymbol{n} \tag{VIII.3}$$

où le signe '$\star_1$' indique une convolution 1D discrète.

Cependant, ce modèle de convolution simple n'est pas capable de rendre suffisamment bien compte des données. Dans le cas, par exemple, de l'inspection de blocs en acier soudés il ne permet pas d'estimer correctement simultanément les échos de diffraction et de coin. En effet, ces deux échos ne sont pas de nature comparable et un déphasage différent de l'ondelette est observé. Le modèle simple (VIII.3) ne permet pas de tenir compte de ce déphasage. Par conséquent, un modèle plus élaboré est nécessaire afin de tenir compte des déphasages de l'ondelette. Un tel déphasage a été considéré dans [Chalmond *et al.*, 2003; Faur *et al.*, 1998; Mu *et al.*, 2002; Demirli et Saniie, 2001a]. En détection ultrasonore pulsée, l'ondelette $h(t)$ se modélise sous la forme d'une amplitude et d'une phase

$$h(t) = A(t)\cos(2\pi f_c t + \Phi(t)).$$

Cette amplitude et cette phase dépendent à la fois des caractéristiques du transducteur et du milieu de propagation. A partir de cette ondelette $h(t)$, la famille d'ondelettes déphasées $\{h_\varphi(t)\}$ peut se définir par

$$h_\varphi(t) = A(t)\cos(2\pi f_c t + \Phi(t) + \varphi). \tag{VIII.4}$$

Ainsi, la généralisation du modèle (VIII.2) permettant de tenir compte des déphasages inconnus de l'ondelette peut s'écrire sous la forme suivante

$$z[k] = \sum_{i=1}^{I} h_{\varphi_{k-i}}[i]r[k-i] + n[k], k = 1, \ldots, M. \tag{VIII.5}$$

où φ_k sont les déphasages inconnus de l'ondelette $h_{\varphi_k}[k]$ par rapport à l'ondelette $h[k]$.

L'équation (VIII.5) ne correspond clairement pas à un modèle convolutif bruité. Néanmoins, ce modèle peut être linéarisé par l'intermédiaire d'une transformée appropriée. En utilisant la transformée de Hilbert, $h_\varphi(t)$ s'obtient alors comme une combinaison linéaire de $h(t)$ et de $g(t)$

$$h_\varphi[k] = \cos\varphi\, h[k] + \sin\varphi\, g[k] \tag{VIII.6}$$

où $g(t) = \mathcal{H}(\boldsymbol{h}(t))$ est la transformée de Hilbert de $h(t)$, qui permet de tenir compte d'un déphasage inconnu [Champagnat et Idier, 1993]. Il est à noter que $g(t)$ s'obtient directement

à partir de $h(t)$ et donc $g(t)$ est déterminé aussitôt que $h(t)$ est identifiée. Dès lors, d'après (VIII.6), (VIII.5) se linéarise sous la forme suivante

$$z[k] = \sum_{i=1}^{I} (h[i]a[k-i] + g[i]b[k-i]) + n[k] \tag{VIII.7}$$

où $a[k] = r[k]\cos\varphi_k$ et $b[k] = r[k]\sin\varphi_k$.

Finalement, l'équation (VIII.7) s'écrit sous la forme suivante

$$\boldsymbol{z} = \boldsymbol{h} \star_1 \boldsymbol{a} + \boldsymbol{g} \star_1 \boldsymbol{b} + \boldsymbol{n}. \tag{VIII.8}$$

La différence entre les modèles (VIII.3) et (VIII.8) peut être interprétée comme le passage d'une réflectivité réelle $\boldsymbol{r}$ à une réflectivité complexe $\boldsymbol{a} + i\boldsymbol{b}$ [Faur *et al.*, 1998; Mu *et al.*, 2002]. Ces deux réflectivités s'identifient en l'absence de déphasage de l'ondelette $\boldsymbol{h}$. Ainsi, on parlera par la suite de réflectivité complexe pour la réflectivité associée au modèle de convolution avec déphasage.

[B] B-scan

Dans la section précédente, un unique A-scan est considéré. Pourtant, les données ultrasonores sont habituellement visualisées sous la forme d'un B-scan composé de plusieurs A-scans issus de positions successives du transducteur sur la surface balayée. Un B-scan typique est illustré figure VIII.2(b). Un B-scan $\mathbf{Z}$ n'est pas véritablement une image mais plutôt une série de A-scans : $\mathbf{Z} = \left[\, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_N \,\right] \in \mathbb{R}^{M \times N}$, où chaque $\boldsymbol{z}_k$ est un A-scan. Les échos observés dans le B-scan de la figure VIII.2(b) présentent une importante composante latérale. Cette composante latérale est due au profil du faisceau ultrasonore qui induit une corrélation forte entre A-scans voisins. Le modèle de convolution 1D précédent n'est pas capable de rendre compte d'une composante latérale des échos d'un B-scan. Ainsi, en utilisant le modèle de convolution 1D ces composantes latérales des échos subsistent dans les réflectivités restaurées, comme illustré figure VIII.8(c).

A présent, afin de tenir compte de la composante latérale des échos, on considère un modèle de convolution latérale et temporelle alors que le modèle de convolution 1D précédent (VIII.8) ne tient compte que de l'aspect temporel. Un modèle de convolution 2D avec réflectivité complexe n'a pas, à notre connaissance, été considéré jusqu'à présent. Un tel modèle correspond naturellement à l'extension 2D du modèle de convolution tenant compte des déphasages de l'ondelette et peut s'écrire sous la forme suivante

$$\mathbf{Z} = \mathbf{H} \star_2 \mathbf{A} + \mathbf{G} \star_2 \mathbf{B} + \mathbf{N} \tag{VIII.9}$$

où le signe '$\star_2$' indique une convolution 2D discrète, la matrice $\mathbf{H}$ est l'ondelette 2D dont la composante latérale est due au profil du faisceau ultrasonore, la matrice $\mathbf{G}$ contient les transformées de Hilbert des colonnes de $\mathbf{H}$ et les matrices $\mathbf{A}$ et $\mathbf{B}$ sont respectivement les parties réelles et imaginaires de la réflectivité complexe 2D. $\mathbf{N}$ est la matrice du bruit supposé blanc gaussien de matrice de covariance $\sigma^2\mathbf{I}$, avec $\mathbf{I}$ la matrice identité. L'identification de l'ondelette 2D $\mathbf{H}$ est traitée dans la partie VIII.3.2.

[C] Condition de bords

Nous avons testé deux conditions de bords pour le modèle de convolution 2D : la condition de Dirichlet (zéros à l'extérieur) et la condition périodique [Ng *et al.*, 1999]. Il apparaît que, dans nos résultats expérimentaux, les deux réflectivités 2D sont très similaires. Dans la mesure

où la convolution et la déconvolution peuvent être calculées un peu plus rapidement avec la condition périodique, nous n'avons retenu ici que la condition de bords périodique. Ainsi, les signes '$\star_1$' et '$\star_2$' indiquent par la suite une convolution discrète périodique.

VIII.2.2 Réflectivité estimée par utilisation d'*a priori*

[A] Information *a priori*

Par souci de simplicité nous considérons les notations du cas 1D. En utilisant le modèle (VIII.8), le problème qui consiste à estimer $\boldsymbol{a}$ et $\boldsymbol{b}$ lorsque les données $\boldsymbol{z}$ et les ondelettes $\boldsymbol{h}$ et $\boldsymbol{g}$ sont connues est un problème de déconvolution qu'il est nécessaire de régulariser. La plupart des efforts pour régulariser les problèmes mal posés se sont concentrés sur l'utilisation d'une connaissance *a priori* sur la solution [Idier, 1999]. Dans le cadre du CND par ultrasons, la réflectivité recherchée est classiquement supposée être composée de seulement quelques pics, dont les positions et les amplitudes sont inconnues. Il s'agit d'une hypothèse de parcimonie. Dans le cas du modèle de convolution incorporant les déphasages de l'ondelette définie par la section VIII.2.1.[A], l'hypothèse de parcimonie s'applique naturellement au module de la réflectivité complexe $|\boldsymbol{x}| = |\boldsymbol{a} + i\boldsymbol{b}|$. L'estimée de la réflectivité complexe est alors définie comme le minimiseur du critère pénalisé suivant

$$\mathcal{J}(\boldsymbol{x}) = \|\boldsymbol{e}(\boldsymbol{x})\|^2 + \lambda\Phi(|\boldsymbol{x}|) \tag{VIII.10}$$

où $\boldsymbol{e}(\boldsymbol{x}) = \boldsymbol{z}_n - \boldsymbol{h}\star_1\boldsymbol{a} - \boldsymbol{g}\star_1\boldsymbol{b}$.

Le premier terme $\|\boldsymbol{e}(\boldsymbol{x})\|^2$ est un terme d'adéquation aux données tandis que le terme $\Phi(|\boldsymbol{x}|)$ pénalise les valeurs non nulles de $\boldsymbol{x}$ et λ est un paramètre réel positif. Le paramètre λ est choisi de telle sorte qu'un compromis soit trouvé entre l'adéquation aux données et la conformité à l'*a priori*. Ici, chaque composante de la réflectivité complexe est pénalisée de la même manière, c'est-à-dire que la pénalisation $\Phi(|\boldsymbol{x}|)$ s'écrit sous la forme

$$\Phi(|\boldsymbol{x}|) = \sum_{m=1}^{M} \phi(|x[m]|)$$

où $\boldsymbol{x}$ est de taille M et ϕ est une fonction l_2l_1 convexe. Un choix possible de fonction l_2l_1 est la fonction hyperbolique convexe $\phi_{\text{hyp}}(t) = \sqrt{\delta^2 + t^2}$. Ainsi, le critère $\mathcal{J}$ défini par (VIII.10) est convexe.

[B] Estimation 1D de la réflectivité

Dans [Gautier *et al.*, 2001], une solution basée sur un modèle de convolution 1D avec prise en compte des déphasages (VIII.8) et avec une hypothèse de parcimonie permet de traiter la superposition d'échos proches le long de l'axe temporel. A partir de chaque A-scan $\boldsymbol{z}_n$, une réflectivité complexe $\widehat{\boldsymbol{a}}_n + i\,\widehat{\boldsymbol{b}}_n$ est calculée comme le minimiseur du critère pénalisé (VIII.10)

$$(\widehat{\boldsymbol{a}}_n, \widehat{\boldsymbol{b}}_n) = \arg\min_{\boldsymbol{a},\boldsymbol{b}} \mathcal{J}_1^n(\boldsymbol{a}, \boldsymbol{b}), \tag{VIII.11}$$

$$\mathcal{J}_1^n(\boldsymbol{a}, \boldsymbol{b}) = \|\boldsymbol{z}_n - \boldsymbol{h}\star_1\boldsymbol{a} - \boldsymbol{g}\star_1\boldsymbol{b}\|^2 + \lambda \sum_{m=1}^{M} \phi(|a[m] + i\,b[m]|)$$

où $\boldsymbol{a}$, $\boldsymbol{b}$ et les ondelettes $\boldsymbol{h}$, $\boldsymbol{g}$ sont de taille M.

Cette méthode de déconvolution 1D permet de discriminer les échos de diffraction et de coin selon l'axe temporel même dans des situations difficiles telles que celle illustrée figure VIII.8(c). Cependant, elle ne fournit pas la position latérale des échos avec précision. C'est la raison pour laquelle une extension 2D est proposée.

[C] Estimation 2D de la réflectivité

La méthode de déconvolution 2D avec prise en compte des déphasages (VIII.9) et avec une hypothèse de parcimonie consiste à définir à partir du B-scan $\mathbf{Z}$ une réflectivité complexe $\mathbf{A}+i\mathbf{B}$ calculée comme le minimiseur du critère pénalisé (VIII.10)

$$(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}) = \arg\min_{\mathbf{A},\mathbf{B}} \mathcal{J}_2(\mathbf{A}, \mathbf{B}), \tag{VIII.12}$$

$$\mathcal{J}_2(\mathbf{A}, \mathbf{B}) = \|\mathbf{Z} - \mathbf{H}\star_2\mathbf{A} - \mathbf{G}\star_2\mathbf{B}\|^2 + \lambda \sum_{m,n} \phi(|a[m,n] + i\, b[m,n]|)$$

où $\mathbf{A}$, $\mathbf{B}$ et les ondelettes $\mathbf{H}$, $\mathbf{G}$ sont de taille $M \times N$.

VIII.2.3 Minimisation du critère pénalisé 2D

Les estimées des réflectivités complexes 1D et 2D $\widehat{\boldsymbol{a}}_n + i\widehat{\boldsymbol{b}}_n$ et $\widehat{\mathbf{A}} + i\widehat{\mathbf{B}}$ sont définies respectivement par (VIII.11) et par (VIII.12). Dans le cas d'une fonction de régularisation l_2l_1 (contrairement au cas l_2) le minimiseur des critères pénalisés ne peut pas s'exprimer sous la forme d'une formule analytique. Ainsi, la mise en œuvre d'un algorithme de minimisation est nécessaire pour estimer les minimiseurs de (VIII.11) et (VIII.12).

Afin d'estimer les réflectivités 1D, nous proposons la mise en œuvre d'une déconvolution 2D virtuelle, dans le sens où les N réflectivités 1D sont estimées simultanément en utilisant une ondelette 2D dont l'unique colonne non nulle est l'ondelette 1D $\boldsymbol{h}$. Il en résulte que les estimées des réflectivités complexes 1D et 2D sont toutes deux issues d'un critère pénalisé 2D qui ne diffère que par l'ondelette 2D employée. L'objet de cette partie est de présenter l'algorithme de minimisation utilisé pour le critère pénalisé 2D. Si l'utilisation d'une fonction de régularisation l_2l_1 permet un meilleur résultat que l'utilisation de la norme l_2, elle est en contrepartie plus coûteuse. C'est pourquoi il est important d'employer un algorithme de minimisation convergent et efficace pour les fonctions de régularisation l_2l_1.

[A] Critère pénalisé 2D

La réflectivité complexe à estimer est définie par la matrice $\widehat{\mathbf{A}} + i\widehat{\mathbf{B}}$ (VIII.12). Afin de pouvoir manipuler le gradient du critère $\mathcal{J}_2$, on réindexe la réflectivité matricielle sous la forme vectorielle. Ainsi les vecteurs $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{z}$ de taille MN correspondent respectivement aux matrices $\mathbf{A}, \mathbf{B}, \mathbf{Z}$ de taille $M \times N$. La réflectivité complexe est alors équivalente au vecteur $\boldsymbol{x} = \left[\boldsymbol{a}^{\mathrm{t}}\ \boldsymbol{b}^{\mathrm{t}}\right]^{\mathrm{t}} \in \mathbb{R}^{2MN}$. On utilise aussi la notation complexe compacte $\tilde{x}[m] = a[m] + ib[m]$.

On introduit les matrices circulante-bloc-circulante (CBC) $\mathbb{H}$ et $\mathbb{G}$ de taille $MN \times MN$ correspondant respectivement à la convolution périodique 2D par les ondelettes 2D $\mathbf{H}$ et $\mathbf{G}$ [Ng *et al.*, 1999]. On considère alors la matrice $\boldsymbol{\Pi} = [\mathbb{H}\ \mathbb{G}] \in \mathbb{R}^{MN\times 2MN}$. Ainsi, le problème de minimisation (VIII.12) s'écrit sous la forme

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \mathcal{J}_2(\boldsymbol{x}), \tag{VIII.13}$$

$$\mathcal{J}_2(\boldsymbol{x}) = \|\boldsymbol{z} - \boldsymbol{\Pi}\boldsymbol{x}\|^2 + \lambda \sum_{m=1}^{MN} \phi(|\tilde{x}[m]|) \tag{VIII.14}$$

Il est à noter qu'une fois la minimisation achevée, le vecteur $\widehat{\boldsymbol{x}}$ peut être facilement réindexé pour former l'estimée de la réflectivité complexe 2D $\widehat{\mathbf{A}} + i\,\widehat{\mathbf{B}}$.

[B] Utilisation de l'algorithme GCPPR+GR1D préconditionné

Les algorithmes GCNL sont efficaces pour minimiser des critères différentiables dans le cas de problème de grande taille [Bertsekas, 1999, p. 155]. On considère pour minimiser le critère pénalisé $\mathcal{J}_2$ l'algorithme GCPPR+GR1D défini dans le chapitre VI. La recherche du pas GR1D est considérée avec une seule itération ($I = 1$). La convergence de cet algorithme est assurée par le Théorème VI.3.2.

Le gradient $\nabla\mathcal{J}_2(\boldsymbol{x})$ du critère $\mathcal{J}_2$ s'obtient sous la forme

$$\nabla\mathcal{J}_2(\boldsymbol{x}) = 2\boldsymbol{\Pi}^{\mathrm{t}}(\boldsymbol{\Pi}\boldsymbol{x} - \boldsymbol{z}) + \lambda \begin{bmatrix} \mathrm{Vect}\left[\frac{\partial\phi}{\partial a_m}(|\tilde{x}[m]|)\right] \\ \mathrm{Vect}\left[\frac{\partial\phi}{\partial b_m}(|\tilde{x}[m]|)\right] \end{bmatrix}. \tag{VIII.15}$$

La partie suivante décrit le préconditionneur considéré pour l'algorithme GCPPR+GR1D.

[C] Préconditionnement

Le taux de convergence de l'algorithme GCPPR dépend crucialement du choix de la matrice de préconditionnement $\mathbf{M}$ (VI.1). L'utilisation de l'inverse du Hessien comme préconditionneur avec pas optimal permet d'obtenir un taux de convergence *superlinéaire* comme la méthode de Newton [Al-Baali et Fletcher, 1996]. Cependant, l'utilisation d'un tel préconditionnement est en général trop coûteuse. En pratique, un compromis doit être trouvé entre le taux de convergence et le coût d'implémentation.

L'hypothèse de parcimonie de la réflectivité définie dans la section VIII.2.2.[A] permet de définir un préconditionneur qui est à la fois proche de l'inverse du Hessien et qui se calcule de manière efficace. Le Hessien $\nabla^2\mathcal{J}_2(\boldsymbol{x})$ du critère $\mathcal{J}_2$ peut s'écrire sous la forme

$$\nabla^2\mathcal{J}_2(\boldsymbol{x}) = 2\boldsymbol{\Pi}^{\mathrm{t}}\,\boldsymbol{\Pi} + \lambda \begin{bmatrix} \mathrm{Diag}\left[\frac{\partial^2\phi}{\partial a_m^2}(|\tilde{x}[m]|)\right] & \mathrm{Diag}\left[\frac{\partial^2\phi}{\partial a_m \partial b_m}(|\tilde{x}[m]|)\right] \\ \mathrm{Diag}\left[\frac{\partial^2\phi}{\partial a_m \partial b_m}(|\tilde{x}[m]|)\right] & \mathrm{Diag}\left[\frac{\partial^2\phi}{\partial b_m^2}(|\tilde{x}[m]|)\right] \end{bmatrix}.$$

En supposant que la majorité des composantes de la réflectivité sont très petites, typiquement toutes à l'exception de celles correspondantes aux échos de diffraction et coin, on peut alors considérer l'approximation circulante suivante du Hessien

$$\nabla^2\mathcal{J}_2(\boldsymbol{x}) \simeq \nabla^2\mathcal{J}_2(\mathbf{0})$$

où

$$\nabla^2\mathcal{J}_2(\mathbf{0}) = 2\boldsymbol{\Pi}^{\mathrm{t}}\,\boldsymbol{\Pi} + \mu\,\mathbf{I}_{2MN} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{12}^{\mathrm{t}} & \mathbf{C}_{22} \end{bmatrix},$$

avec $\mu := \lambda/\delta$, $\mathbf{C}_{11} = 2\mathbb{H}^{\mathrm{t}}\mathbb{H} + \mu\mathbf{I}_{MN}$, $\mathbf{C}_{22} = 2\mathbb{G}^{\mathrm{t}}\mathbb{G} + \mu\mathbf{I}_{MN}$ et $\mathbf{C}_{12} = 2\mathbb{H}^{\mathrm{t}}\mathbb{G}$. Les matrices $\mathbf{C}_{ij}$ sont CBC de taille $MN \times MN$. De plus, la matrice $\mathbf{C}_{12}$ est symétrique. Notons que la matrice $\nabla^2\mathcal{J}_2(\mathbf{0})$ correspond à la matrice normale de Geman et Yang définie par (IV.26), page 66 avec $\hat{a} = 1/\phi''(0) = \delta$ [Nikolova et Ng, 2005, p. 955, ligne 5].

Le préconditionneur proposé $\mathbf{M}$ est alors défini par

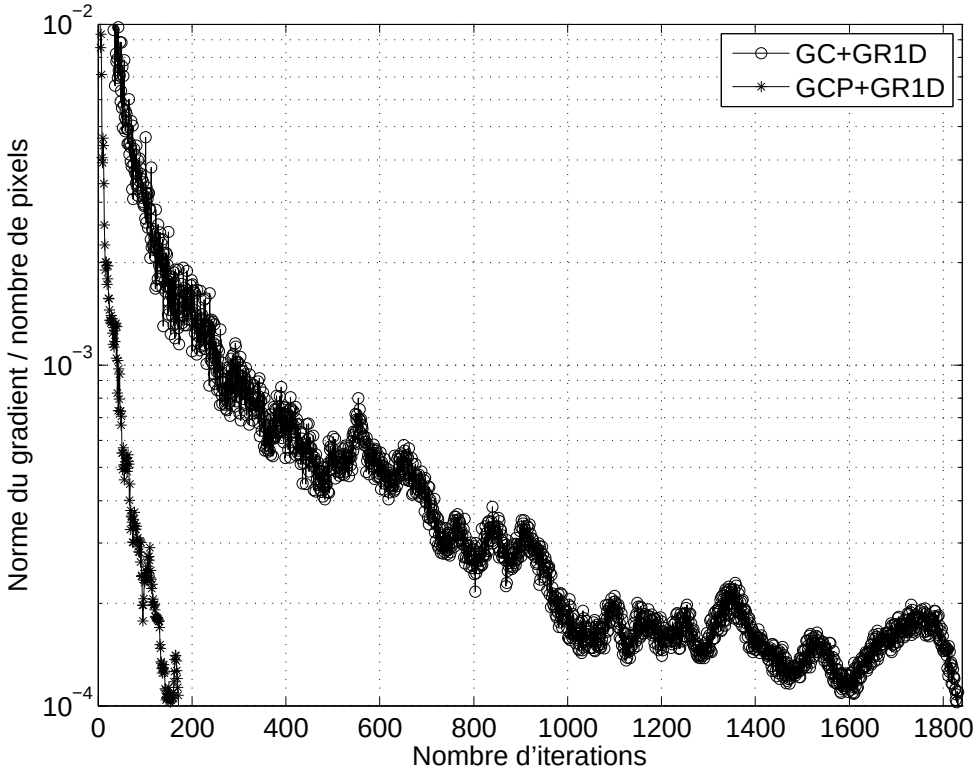
$$\mathbf{M} = \nabla^2\mathcal{J}_2(\mathbf{0})^{-1}, \tag{VIII.16}$$

afin d'exploiter le fait que les matrices CBC $\mathbf{C}_{ij}$ peuvent être diagonalisées par la transformée de Fourier 2D $\mathbf{W}_2$ [Ng *et al.*, 1999]

$$\mathbf{C}_{ij} = \mathbf{W}_2^*\boldsymbol{\Lambda}_{ij}\mathbf{W}_2, \quad i,j \in \{1,2\} \tag{VIII.17}$$

où les valeurs propres de $\mathbf{C}_{ij}$ sont déterminées par transformée de Fourier 2D.

(a)



(b)

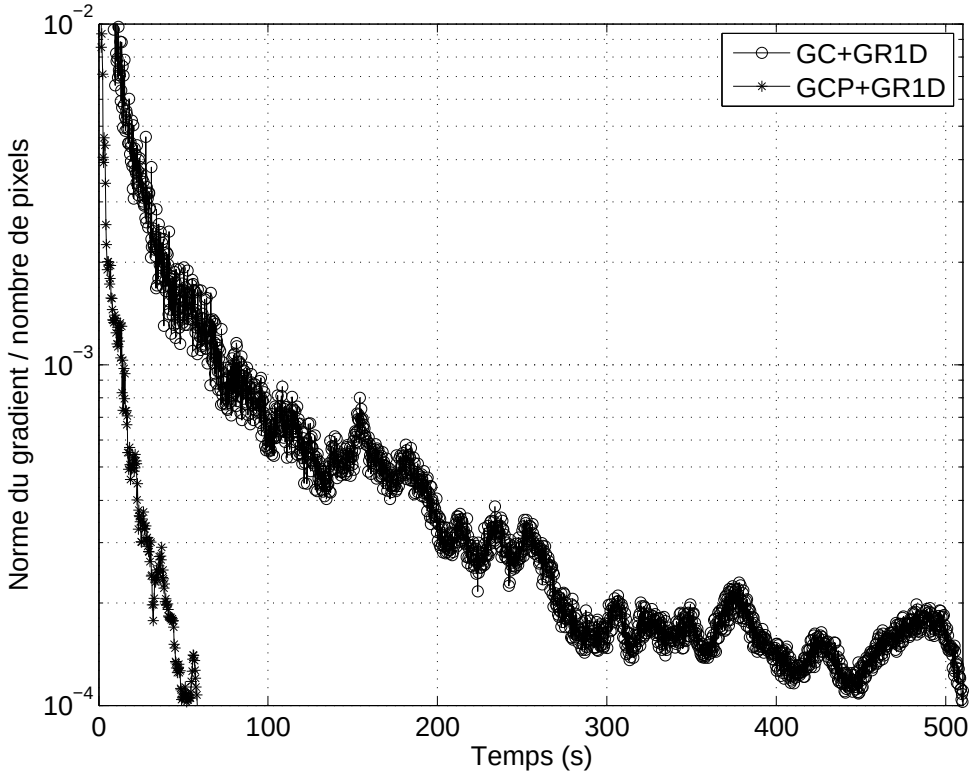


FIG. VIII.3: Influence du préconditionnement sur le taux de convergence pour la minimisation du critère $\mathcal{J}_2$. (a) $\|\nabla\mathcal{J}(\boldsymbol{x}_k)\|/MN$ en fonction du nombre d'itérations, (b) $\|\nabla\mathcal{J}(\boldsymbol{x}_k)\|/MN$ en fonction du temps.

D'après le Lemme des matrices partitionnées [Golub et Van Loan, 1996], (VIII.16) fournit

$$\mathbf{M} = \begin{bmatrix} \mathbf{C}_{11}^{-1} + \mathbf{C}_{11}^{-1}\mathbf{C}_{12}\Delta_{\mathbf{C}}^{-1}\mathbf{C}_{12}\mathbf{C}_{11}^{-1} & -\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\Delta_{\mathbf{C}}^{-1} \\ -\Delta_{\mathbf{C}}^{-1}\mathbf{C}_{12}\mathbf{C}_{11}^{-1} & \Delta_{\mathbf{C}}^{-1} \end{bmatrix}$$

où $\Delta_{\mathbf{C}} = \mathbf{C}_{22} - \mathbf{C}_{12}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}$.

Finalement, d'après (VIII.17) et $\mathbf{W}_2^{-1} = \mathbf{W}_2^*$, on obtient

$$\mathbf{M} = \begin{bmatrix} \mathbf{W}_2^* & 0 \\ 0 & \mathbf{W}_2^* \end{bmatrix} \begin{bmatrix} \mathbf{D}_{11} & \mathbf{D}_{12} \\ \mathbf{D}_{21} & \mathbf{D}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{W}_2 & 0 \\ 0 & \mathbf{W}_2 \end{bmatrix} \tag{VIII.18}$$

où $\mathbf{D}_{11} = \mathbf{\Lambda}_{11}^{-1} + \mathbf{\Lambda}_{11}^{-2}\mathbf{\Lambda}_{12}^2\Delta^{-1}$, $\mathbf{D}_{12} = -\mathbf{\Lambda}_{11}^{-1}\mathbf{\Lambda}_{12}\Delta^{-1}$, $\mathbf{D}_{21} = -\Delta^{-1}\mathbf{\Lambda}_{12}\mathbf{\Lambda}_{11}^{-1}$, $\mathbf{D}_{22} = \Delta^{-1}$ avec $\Delta = \mathbf{\Lambda}_{22} - \mathbf{\Lambda}_{12}^2\mathbf{\Lambda}_{11}^{-1}$.

La direction de descente préconditionnée $\boldsymbol{p}_k = -\mathbf{M}\,\mathcal{J}_2(\boldsymbol{x}_k)$ (VI.1), page 83 se déduit efficacement du gradient courant $\nabla\mathcal{J}_2(\boldsymbol{x}_k)$ avec une complexité en $O(MN\log(MN))$ par l'utilisation des transformées rapides de Fourier 2D [Cooley et Tukey, 1965]. L'efficacité du préconditionnement est illustrée figure VIII.3 en comparant les versions avec et sans préconditionnement. La figure figure VIII.3(a) donnant le taux de convergence en fonction du nombre d'itérations est très proche de la figure figure VIII.3(b) donnant le taux de convergence en fonction du temps. Ceci indique que le surcoût du préconditionnement est faible devant le coût global de l'algorithme GCPPR+GR1D. D'autre part, le gain apporté par le préconditionnement est considérable, ce qui montre clairement l'intérêt du préconditionnement proposé.

[D] Règle d'arrêt

L'algorithme GCPPR+GR1D nécessite l'utilisation d'une règle d'arrêt. Une règle d'arrêt classique utilise la norme du gradient du critère à minimiser. L'algorithme de minimisation est arrêté lorsque la norme du gradient passe sous un seuil. Ici nous utilisons la règle d'arrêt suivante pour l'algorithme GCPPR+GR1D défini dans le chapitre VI, page 81.

$$\|\nabla\mathcal{J}_2(\boldsymbol{x}_{k+1})\|/\sqrt{MN} < 10^{-4}$$

où MN est la taille du vecteur $\boldsymbol{x}_k$. Cette règle d'arrêt fait intervenir le gradient normalisé afin qu'elle soit indépendante de la taille de la réflectivité.

VIII.3 Résultats de la méthode

VIII.3.1 Procédure de recalage

Dans les résultats de l'application de la méthode proposée, $\mathbf{Z}$ n'est pas le B-scan brut tel que celui de la figure figure VIII.2(b). En fait, les B-scans bruts ne sont pas bien décrits par le modèle de convolution $\mathbf{H}\star_2\mathbf{A} + \mathbf{G}\star_2\mathbf{B} + \textit{bruit}$ (VIII.9). D'après notre expérience, ceci est principalement dû à la présence de certains décalages entre A-scans. Par conséquent, nous proposons une procédure de recalage empirique, où chaque A-scan est recalé selon une approche de type maximum de corrélation. L'objectif principal de cette procédure de recalage est de traiter les décalages entre A-scans. On pourrait se contenter de cet objectif en conservant l'inclinaison originale des échos. Nous proposons plutôt de mettre les échos à l'horizontale. Cette procédure est intéressante à plusieurs titres :

- elle est robuste par rapport à l'état de surface qui peut ne pas être parfaitement plate,
- elle est robuste par rapport à la variation possible de l'angle du faisceau,

– elle permet naturellement d'utiliser un modèle séparable pour l'ondelette 2D $\mathbf{H}$ (*cf.* section VIII.3.2).
– elle transforme en rectangle la zone contenant les échos qui est initialement un parallélogramme.

Nous insistons sur la réversibilité d'une telle procédure de recalage dans la mesure où le déplacement de chaque A-scan est mémorisé. La figure VIII.4(a) montre qu'un simple recalage constant entre A-scan (figure VIII.2(b)) n'est pas suffisant. Nous proposons plutôt d'utiliser une approche par maximum de corrélation pour estimer le meilleur recalage pour chaque A-scan :

– un A-scan de référence $\boldsymbol{r}$ est choisi comme celui contenant les échos de diffraction et de coin avec les plus fortes énergies.
– les autres A-scans $\boldsymbol{z}_n$ sont recalés en fonction de l'A-scan de référence $\boldsymbol{r}$.

L'approche par maximum de corrélation utilisée consiste à trouver le meilleur recalage τ_n et la meilleure amplitude μ_n pour l'A-scan $\boldsymbol{z}_n$ par rapport à l'A-scan de référence $\boldsymbol{r}$, qui sont tous deux de taille K. Ceci est réalisé par l'intermédiaire de la minimisation du critère des moindres carrés suivant

$$(\mu_n, \tau_n) = \arg\min_{(\mu,\tau)} \|\mu \boldsymbol{r}(\tau) - \boldsymbol{z}_n\|^2$$

où $\boldsymbol{r}(\tau) = [\{r(\tau)\}_k]$ de taille K est la version décalée de $\boldsymbol{r}$

$$\{r(\tau)\}_k = \begin{cases} \{r\}_{k-\tau} & \text{si } k-\tau \in \{1,\dots,K\} \\ 0 & \text{sinon.} \end{cases}$$

Afin d'assurer que la version recalée $\boldsymbol{r}(\tau)$ contient exactement la même information que l'original $\boldsymbol{r}$, il est possible de rajouter des zéros aux bords à $\boldsymbol{r}$ avant d'effectuer le recalage.

Les valeurs μ_n et τ_n s'obtiennent simplement en utilisant la corrélation normalisée entre l'A-scan courant $\boldsymbol{z}$ et l'A-scan de référence $\boldsymbol{r}$ recalé de τ pixels. Cette dernière s'exprime par

$$\mathrm{NCorr}_{\boldsymbol{z},\boldsymbol{r}}(\tau) = \frac{\boldsymbol{z}^{\mathrm{t}}\boldsymbol{r}(\tau)}{\|\boldsymbol{z}\|\|\boldsymbol{r}(\tau)\|}. \tag{VIII.19}$$

En utilisant (VIII.19), les valeurs μ_n et τ_n peuvent être déterminées séparément par

$$\begin{aligned} \tau_n &= \arg\max_{\tau} \mathrm{NCorr}_{\boldsymbol{z}_n,\boldsymbol{r}}(\tau), \\ \mu_n &= \frac{\|\boldsymbol{z}_n\|}{\|\boldsymbol{r}(\tau_n)\|}\mathrm{NCorr}_{\boldsymbol{z}_n,\boldsymbol{r}}(\tau_n). \end{aligned} \tag{VIII.20}$$

L'utilisation d'une approche par maximum de corrélation permet d'estimer la valeur du recalage τ_n (figure VIII.5(a)) et donc d'aligner les A-scans horizontalement (figure VIII.4(b)), par rapport à l'A-scan de référence. Pour les A-scans situés aux extrémités gauche et droite des deux échos, le rapport signal à bruit (RSB) est faible, donc la procédure de recalage n'est pas fiable. Il est donc préférable d'arrêter la procédure de recalage dans ce cas.

Cette procédure de recalage est systématiquement utilisée pour obtenir $\mathbf{Z}$ (figure VIII.4(b)) et $\mathbf{H}$ (figure VIII.6(b)) à partir respectivement du B-scan brut et d'un écho isolé.

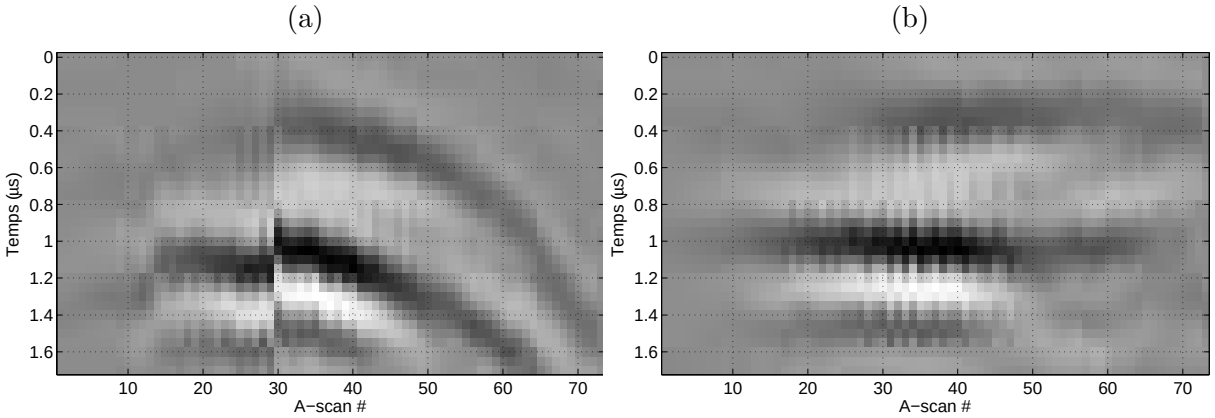


FIG. VIII.4: (a) Zoom de la figure VIII.2(b) après un recalage vertical constant de six pixels entre chaque A-scan. (b) Zoom de la figure VIII.2(b) après recalage par l'approche maximum de corrélation.

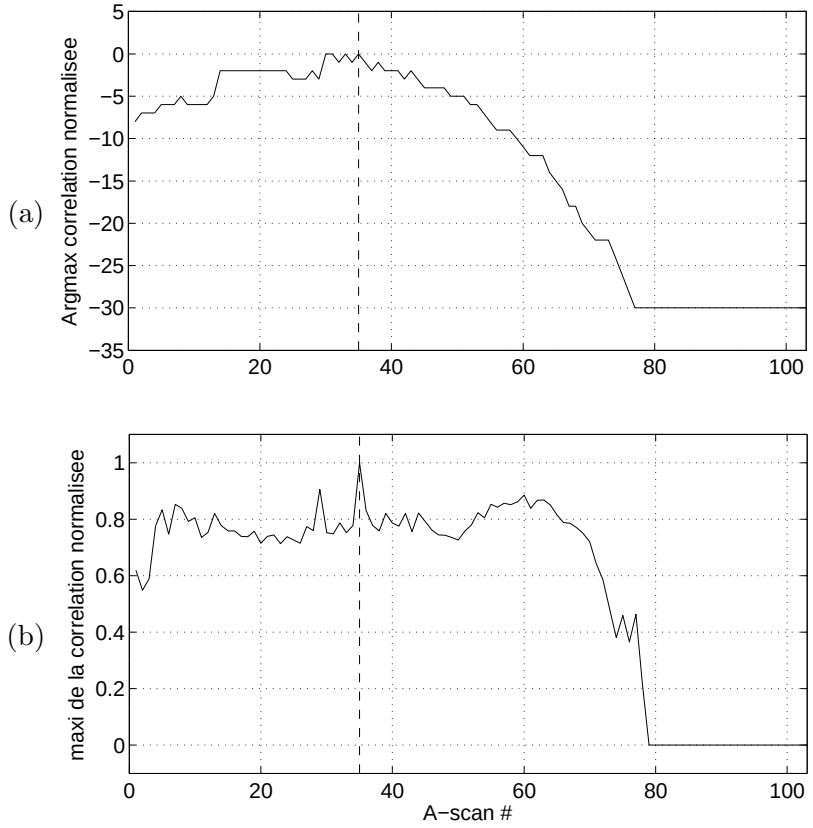


FIG. VIII.5: A partir de la figure VIII.4(a). Corrélation normalisée entre l'A-scan # et l'A-scan de référence #35 : (a) valeur du recalage τ_n (VIII.20), (b) $\mathrm{NCorr}_{\boldsymbol{z}_n,\boldsymbol{r}}(\tau_n)$ (VIII.19). La procédure de recalage est arrêtée lorsque la valeur de $\mathrm{NCorr}_{\boldsymbol{z}_n,\boldsymbol{r}}(\tau_n)$ passe sous le seuil de $0,3$ ce qui se produit ici pour l'A-scan #79.

VIII.3.2 Identification de l'ondelette 2D

Cette partie traite de l'identification de l'ondelette 2D $\mathbf{H}$ utilisée pour la déconvolution de tous les B-scans. Cette étape d'identification est cruciale. La précision de la réflectivité estimée dépend fortement de la qualité de l'identification de l'ondelette 2D. Il est donc important de prendre un soin tout particulier pour cette étape.

Il y a plusieurs possibilités pour identifier l'ondelette 2D. Une approche classique en CND consiste à utiliser un bloc test qui a les mêmes caractéristiques de propagation ultrasonore que le matériau testé, mais présentant des caractéristiques géométriques telles qu'un écho isolé se retrouve dans les mesures. L'identification de l'ondelette 2D est alors basée sur cet écho isolé.

L'objectif n'est pas seulement d'obtenir un écho isolé, mais aussi un écho impulsionnel. En effet, si l'ondelette 2D candidate est plus large que certains échos mesurés, la déconvolution risque de ne pas pouvoir bien restituer ces échos. La caractérisation de la géométrie du matériau risque alors d'être compromise. Par contre, une ondelette 2D plus étroite que les échos n'entraîne pas ces inconvénients.

Nous avons conduit des essais d'identification de l'ondelette 2D en considérant deux types de blocs test :

- avec un trou au fond
- avec une haute entaille de telle sorte que les échos de diffraction et de coin soient bien séparés

Les échos issus du trou et l'écho de coin sont des échos de réflexion. Ils ne sont pas de nature comparable à l'écho de diffraction. L'écho de diffraction est créé uniquement par le haut de l'entaille alors que l'écho de coin est créé par l'entaille dans son ensemble. Ainsi, l'écho de coin est plus large (axe latéral) que l'écho de diffraction, comme illustré sur les figure VIII.8(c) et (d). Par conséquent, un candidat naturel pour l'écho impulsionnel est l'écho de diffraction. Plus précisément, la version réalignée $\mathbf{H}^{\text{align}}$ de l'écho de diffraction avec la procédure de recalage de la section VIII.3.1 est le candidat naturel pour estimer l'ondelette 2D.

Malheureusement, la déconvolution 2D basée sur $\mathbf{H}^{\text{align}}$ fournit des résultats de mauvaise qualité, car la réflectivité ainsi estimée ne se compose pas de deux échos ponctuels comme attendu mais contient plutôt de nombreux échos sous la forme de hachures. Ce problème est dû à la présence de composantes hautes fréquences dans $\mathbf{H}^{\text{align}}$ (figure VIII.6(a)). Ceci peut s'expliquer simplement par le fait que le recalage par maximum de corrélation repose sur un modèle simple, qui ne fait intervenir que des décalages entiers.

Les hautes fréquences de $\mathbf{H}^{\text{align}}$ pourraient être annulées par filtrage. Nous introduisons plutôt une forme *séparable* de l'ondelette 2D : $\mathbf{H} = \boldsymbol{h}\,\boldsymbol{f}^{\text{t}}$, où $\boldsymbol{f}$ représente le profil 1D du faisceau et $\boldsymbol{h}$ est l'ondelette 1D [Jensen, 1992; Sciacca et Evans, 1992; Abeyratne *et al.*, 1995]. Cela fournit naturellement une version lissée de $\mathbf{H}^{\text{align}}$ (figure VIII.6(b)). La matrice $\mathbf{G}$ peut aussi être mise sous une forme 2D séparable : $\mathbf{G} = \boldsymbol{g}\,\boldsymbol{f}^{\text{t}}$ où $\boldsymbol{g} = \mathcal{H}(\boldsymbol{h})$. De plus, l'identification du modèle 2D séparable est effectuée à partir de J échos réalignés, $\mathbf{H}_1^{\text{align}}, \ldots, \mathbf{H}_J^{\text{align}}$, au lieu d'un seul pour une meilleure qualité d'estimation.

Pour estimer $\boldsymbol{f}$ et $\boldsymbol{h}$, on utilise l'approche moindre carrés :

$$(\widehat{\boldsymbol{f}}, \widehat{\boldsymbol{h}}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\tau}}) = \underset{(\boldsymbol{f},\boldsymbol{h},\boldsymbol{\mu},\boldsymbol{\tau})}{\arg\min}\, \mathcal{L}(\boldsymbol{f}, \boldsymbol{h}, \boldsymbol{\mu}, \boldsymbol{\tau}) \tag{VIII.21}$$

avec

$$\mathcal{L}(\boldsymbol{f}, \boldsymbol{h}, \boldsymbol{\mu}, \boldsymbol{\tau}) = \|\mathbf{H}_1^{\text{align}} - \boldsymbol{h}\,\boldsymbol{f}^{\text{t}}\|^2 + \sum_{j=2}^{J} \|\mu_j \mathbf{H}_j^{\text{align}}(\tau_i) - \boldsymbol{h}\,\boldsymbol{f}^{\text{t}}\|^2 \tag{VIII.22}$$

où $\|\cdot\|$ est la norme de Frobenius, $\boldsymbol{\mu} = [\mu_2, \ldots, \mu_J]$ est le vecteur des amplitudes relatives et $\boldsymbol{\tau} = [\tau_2, \ldots, \tau_J]$ avec $\tau_j = (\tau_j^x, \tau_j^y)$ est le vecteur des recalages 2D. Les variables τ_j et μ_j

permettent de normaliser et de centrer les $\mathbf{H}_j^{\text{align}}$. $\mathbf{H}_j^{\text{align}}(\tau_j) = \left[\{H_j^{\text{align}}(\tau_j)\}_{p,q}\right]$ de taille $P \times Q$ est la version recalée de $\mathbf{H}_j^{\text{align}}$ avec un décalage 2D de valeur τ_j :

$$\{H_j^{\text{align}}(\tau_j)\}_{p,q} = \begin{cases} \{H_j^{\text{align}}\}_{p-\tau_j^x, q-\tau_j^y} & \text{si } \begin{cases} p-\tau_j^x \in \{1,\dots,P\} \\ q-\tau_j^y \in \{1,\dots,Q\} \end{cases} \\ 0 & \text{autrement.} \end{cases} \tag{VIII.23}$$

Notons que le critère $\mathcal{L}$ défini par (VIII.22) possède plusieurs minima locaux. Nous proposons d'effectuer une recherche locale initialisée suffisamment proche du minimum global qui se révèle satisfaisante en pratique. Afin de minimiser le critère $\mathcal{L}$, nous proposons un schéma de minimisation alternée de type méthode coordonnée par coordonnée de la forme suivante

$$(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\tau}})_{k+1} = \underset{(\boldsymbol{\mu},\boldsymbol{\tau})}{\arg\min}\, \mathcal{L}(\boldsymbol{f}_k, \boldsymbol{h}_k, \boldsymbol{\mu}, \boldsymbol{\tau}), \tag{VIII.24}$$

$$\widehat{\boldsymbol{f}}_{k+1} = \underset{\boldsymbol{f}}{\arg\min}\, \mathcal{L}(\boldsymbol{f}, \boldsymbol{h}_k, \boldsymbol{\mu}_{k+1}, \boldsymbol{\tau}_{k+1}), \tag{VIII.25}$$

$$\widehat{\boldsymbol{h}}_{k+1} = \underset{\boldsymbol{h}}{\arg\min}\, \mathcal{L}(\boldsymbol{f}_{k+1}, \boldsymbol{h}, \boldsymbol{\mu}_{k+1}, \boldsymbol{\tau}_{k+1}) \tag{VIII.26}$$

où $\widehat{\boldsymbol{h}}_0$ est la moyenne des colonnes de $\mathbf{H}_1^{\text{align}}$ et

$$\widehat{\boldsymbol{f}}_0 = \underset{\boldsymbol{f}}{\arg\min}\, \|\mathbf{H}_1^{\text{align}} - \widehat{\boldsymbol{h}}_0\, \boldsymbol{f}^{\text{t}}\|^2 = \frac{(\mathbf{H}_1^{\text{align}})^{\text{t}} \widehat{\boldsymbol{h}}_0}{\|\widehat{\boldsymbol{h}}_0\|^2}.$$

D'après l'indépendance des $(\mu_j, \tau_j)_{j\in[2,\dots,J]}$ on obtient

$$(\widehat{\mu}_j, \widehat{\tau}_j)_{k+1} = \underset{(\mu,\tau)}{\arg\min}\, \|\mu \mathbf{H}_j^{\text{align}}(\tau) - \boldsymbol{h}_k\, \boldsymbol{f}_k^{\text{t}}\|^2 \tag{VIII.27}$$

qui est déterminé en utilisant une approche par maximum de corrélation avec pour référent $\boldsymbol{h}_k \boldsymbol{f}_k^{\text{t}}$ (VIII.3.1).

D'après la bilinéarité de $\mathcal{L}(\boldsymbol{f}, \boldsymbol{h}, \boldsymbol{\mu}_{k+1}, \boldsymbol{\tau}_{k+1})$ en fonction de $\boldsymbol{f}$ et $\boldsymbol{h}$, $\widehat{\boldsymbol{f}}_{k+1}$ et $\widehat{\boldsymbol{h}}_{k+1}$ s'obtiennent sous la forme des formules analytiques suivantes

$$\widehat{\boldsymbol{f}}_{k+1} = \frac{1}{J\|\widehat{\boldsymbol{h}}_k\|^2}(\mathbf{H}_1^{\text{align}} + \sum_{j=2}^{J} \mu_{j,k+1} \mathbf{H}_j^{\text{align}}(\tau_{j,k+1}))^{\text{t}} \widehat{\boldsymbol{h}}_k \tag{VIII.28}$$

$$\widehat{\boldsymbol{h}}_{k+1} = \frac{1}{J\|\widehat{\boldsymbol{f}}_{k+1}\|^2}(\mathbf{H}_1^{\text{align}} + \sum_{j=2}^{J} \mu_{j,k+1} \mathbf{H}_j^{\text{align}}(\tau_{j,k+1}))\widehat{\boldsymbol{f}}_{k+1}. \tag{VIII.29}$$

Considérons à titre pédagogique le cas plus simple de l'estimation de l'ondelette 2D séparable à partir d'un unique écho recalé ($J = 1$). Le schéma précédent de minimisation alternée est alors simplifié de telle sorte que (VIII.24) et (VIII.27) ne sont plus nécessaires, tandis que les autres équations se simplifient en utilisant la convention que la somme $\sum_{j=2}^{1}$ vaut zéro. La convergence de ce schéma de minimisation alternée vers un minimum local du critère $\mathcal{L}$ est assurée. En pratique, la rapidité de convergence de ce schéma se révèle très élevée. Seulement deux itérations permettent expérimentalement d'atteindre la convergence.

Malheureusement, l'ondelette 2D séparable estimée à partir des échos de diffraction ne permet pas d'estimer correctement l'écho de coin de certains B-scans. Ceci est dû au faible RSB de l'écho de diffraction ce qui entraîne une estimation de l'ondelette 1D de qualité assez moyenne. Dans la mesure où les échos de réflexion ont un meilleure RSB, nous les avons considérés pour

l'identification de l'ondelette 2D séparable. En fait, comme il a déjà été précisé, les échos issus de trous ne sont pas directement de bons candidats car ils sont trop larges latéralement. Par contre, en réduisant leur largeur ils permettent une estimation de l'ondelette 2D de bonne qualité.

Pour réduire la largeur des échos issus des trous dans les blocs tests, nous utilisons un modèle gaussien du profil 1D [Vollmann, 1982; Demirli et Saniie, 2001a]. Une fois que l'ondelette 2D séparable avec un modèle gaussien pour le profil 1D est identifiée à partir des échos de trous, il devient évident d'en réduire la largeur en réglant la variance du modèle gaussien. Notons que cette variance peut être déterminée en considérant que l'écho de diffraction doit apparaître sous la forme d'un seul pixel dans la réflectivité estimée. L'identification de l'ondelette 2D séparable avec modèle gaussien pour le profil 1D est résumée tableau VIII.1. L'ondelette identifiée est de bonne qualité comme illustrée figure VIII.6 avec l'utilisation de six échos ($J = 6$). Le profil 1D et l'ondelette 1D sont représentés dans la figure VIII.7.

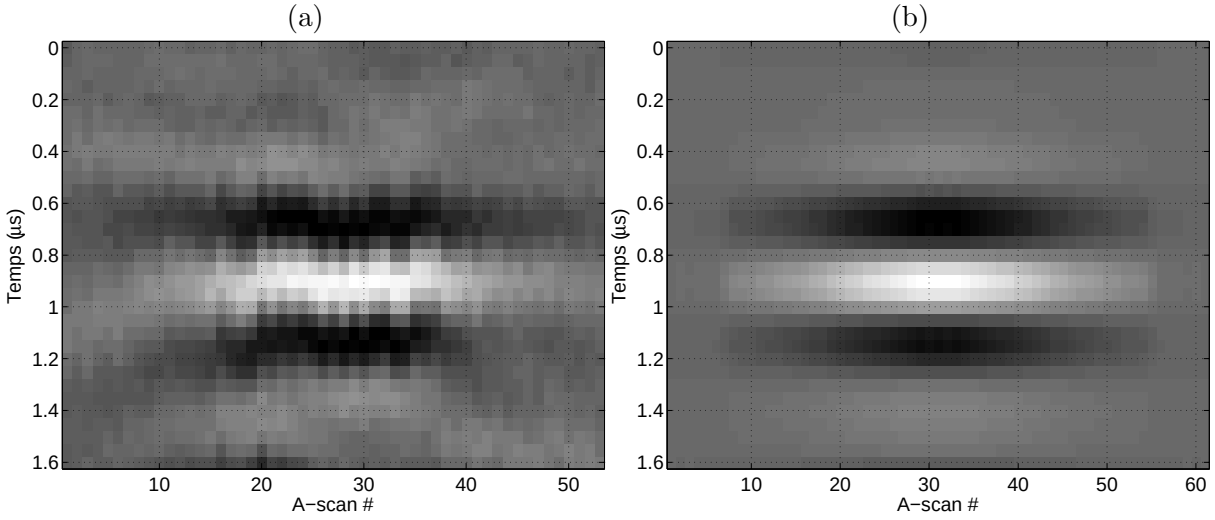


FIG. VIII.6: (a) Version recalée $\mathbf{H}^{\text{align}}$ à partir d'un écho de trou ; (b) Ondelette 2D séparable $\mathbf{H} = \boldsymbol{h}\,\boldsymbol{f}^{\text{t}}$ identifiée à partir de six $\mathbf{H}^{\text{align}}$ selon l'algorithme tableau VIII.1 ($K = 2$ et $J = 6$).

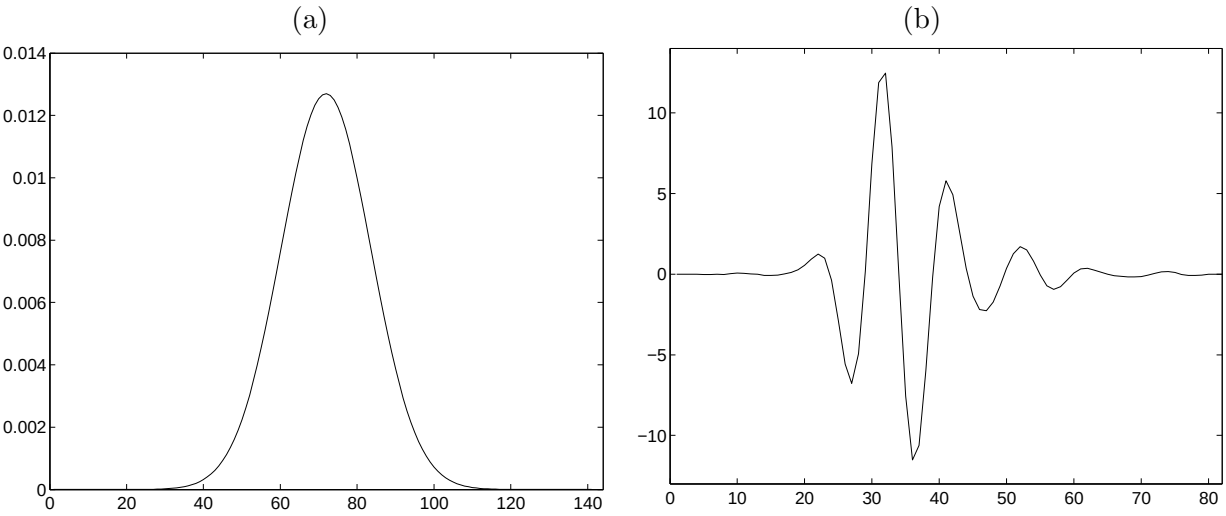


FIG. VIII.7: Profil 1D et ondelette 1D de l'ondelette 2D séparable $\mathbf{H} = \boldsymbol{h}\,\boldsymbol{f}^{\text{t}}$ (figure VIII.6(b)). (a) profil 1D $\boldsymbol{f}$; (b) ondelette 1D $\boldsymbol{h}$.

$\widehat{\boldsymbol{h}}_0$: moyenne des colonnes de $\mathbf{H}_1^{\text{align}}$

$\widehat{\boldsymbol{f}}_0 = \dfrac{(\mathbf{H}_1^{\text{align}})^{\text{t}}\widehat{\boldsymbol{h}}_0}{\|\widehat{\boldsymbol{h}}_0\|^2}$

Pour $k = 0 : K-1$

  $\mathbf{H}_k = \boldsymbol{h}_k \boldsymbol{f}_k^{\text{t}}$

  Pour $j = 2 : J$

    $\tau_{j,k+1} = \arg\max_{\tau} \text{NCorr}_{\mathbf{H}_k, \mathbf{H}_j^{\text{align}}}(\tau)$

    $\mu_{j,k+1} = \dfrac{\|\mathbf{H}_k\|}{\|\mathbf{H}_j^{\text{align}}(\tau_{j,k+1}))\|} \text{NCorr}_{\mathbf{H}_k, \mathbf{H}_j^{\text{align}}}(\tau_{j,k+1})$

  Fin

  $\widehat{\boldsymbol{f}}_{k+1}$ est déterminé par (VIII.28)

  $\widehat{\boldsymbol{h}}_{k+1}$ est déterminé par (VIII.29)

Fin

$\mathbf{H}_K = \boldsymbol{h}_K \boldsymbol{f}_K^{\text{t}}$

$\boldsymbol{f}_G$: est identifié avec un modèle gaussien à partir de $\boldsymbol{f}_K$

$\mathbf{H} = \boldsymbol{h}_K \boldsymbol{f}_G^{\text{t}}$

TAB. VIII.1 – L'algorithme d'identification de l'ondelette 2D séparable $\mathbf{H}$ où $\text{NCorr}_{\mathbf{A},\mathbf{B}}(\tau) = \sum_{m,n}\{A\}_{m,:}\{B(\tau)\}_{:,n}/(\|\mathbf{A}\|\|\mathbf{B}(\tau)\|)$.

VIII.3.3 Résultats expérimentaux

Un des objectifs principaux est de caractériser la géométrie du défaut par la méthode proposée ici. Pour illustrer, on considère l'estimation de la hauteur d'une entaille. Les résultats expérimentaux peuvent être décomposés en trois étapes. Dans un premier temps, nous indiquons d'où proviennent les données brutes. Puis nous indiquons comment ont été estimées les réflectivités à partir des données brutes. Enfin nous détaillons la méthode employée pour estimer la hauteur des entailles à partir des réflectivités estimées.

[A] Obtention des données brutes

La méthode proposée est illustrée sur des données réelles obtenues à partir de blocs d'acier austénitique présentant une entaille électro-érodée. La vitesse de propagation ultrasonore est de 5862 m/s. Trois hauteurs d'entailles ont été réalisées : 3, 5 et 8 mm. L'onde ultrasonore émise par le transducteur dans le bloc d'acier est de type longitudinal, de fréquence 2 MHz et a pour angle réfracté $\theta = 59.7^{\circ}$ (voir figure VIII.1). Pour ce qui est de l'échantillonnage, le pas temporel est de $\Delta_t = 5\ 10^{-8}$ s et le pas spatial (entre 2 A-scans) est de $\Delta_x = 0,5$ mm.

La cale à trou utilisée pour l'identification de l'ondelette 2D (section VIII.3.2) est de même caractéristique que les blocs d'acier contenant les entailles. On a donc 4 blocs : 3 pour les données et un pour identifier l'ondelette 2D. Chaque bloc fournit 6 B-scans car un balayage selon la profondeur z est aussi effectué.

[B] Obtention des réflectivités estimées

Les réflectivités sont estimées à partir des B-scans recalés selon la procédure décrite dans section VIII.3.1. L'obtention de l'ondelette 2D utilisée pour la déconvolution est décrite dans section VIII.3.2. L'ondelette 1D considérée ici est tout simplement la partie temporelle de l'ondelette 2D. La taille de la zone d'interêt des B-scans recalés contenant les échos de coin et de diffraction est d'environ 100×100.

L'algorithme GCPPR+GR1D défini dans le chapitre VI est employé pour estimer les réflectivités. Les réflectivités estimées dépendent du paramètre λ gérant l'équilibre entre information *a priori* et adéquation aux données ainsi que du choix de la fonction ϕ. Ici nous considérons la fonction hyperbolique convexe $\phi_{\text{hyp}}(t) = \sqrt{\delta^2 + t^2}$ où δ est un paramètre de seuil entre le régime quadratique et le régime linéaire [Gautier *et al.*, 2001]. Les noyaux 1D et 2D sont normalisés de telle sorte que leur énergie soit unitaire ($\|\text{noyau}\|_2$=1) afin d'utiliser le même réglage des hyperparamètres λ et δ pour tous les B-scans ainsi que pour les noyaux 1D et 2D. Nous avons choisi de régler les hyperparamètres aux valeurs $\lambda = 40$ et $\delta = 0,5$ de manière expérimentale. Ils ont été réglés afin d'obtenir dans les réflectivités estimées un écho de diffraction le plus ponctuel possible. On peut noter que la résolution latérale de la réflectivité estimée est grandement améliorée par le cadre 2D (figure VIII.8(d)) comparé au cadre 1D (figure VIII.8(b)). De plus, la méthode de déconvolution 2D induit un filtrage des échos secondaires plus efficace que la méthode 1D, comme observé dans figure VIII.8. Ceci permet une détection du défaut plus aisée. Du point de vue qualitatif les résultats sont donc très satisfaisants. Il reste à évaluer les résultats du point de vue quantitatif. En particulier, on cherche à évaluer la qualité de l'estimation de la hauteur des entailles.

[C] Obtention des hauteurs d'entailles estimées

Nous n'avons pas trouvé d'articles dans la littérature estimant la hauteur de défauts dans un cadre CND par ultrasons. Une méthode pour estimer la hauteur de défaut à partir des réflectivités est proposée dans le brevet [Gautier *et al.*, 2002]. Il s'agit d'une méthode s'appuyant sur la déconvolution 1D (VIII.11). Nous proposons de généraliser cette estimation de hauteur du défaut au cas de la déconvolution 2D. La méthode décrite dans le brevet [Gautier *et al.*, 2002] utilise un A-scan contenant les échos de coin et de diffraction. Notons h la distance entre les échos de diffraction et de coin. Cette distance h correspond à la hauteur de l'entaille car on la suppose orthogonale. D'après le brevet [Gautier *et al.*, 2002], il est possible de relier h à la différence du temps de vol δ_t entre les échos au sein d'un A-scan et aux caractéristiques de l'expérience (angle réfracté du faisceau θ et vitesse de propagation ultrasonore V) :

$$h = \frac{V \delta_t}{2 \cos(\theta)}. \tag{VIII.30}$$

Le coefficient 2 est dû à l'aller-retour du faisceau ultrasonore. On pourrait prendre en compte, comme cela est fait dans [Gautier *et al.*, 2002], l'angle d'inclinaison du défaut. Ce n'est pas le cas ici : on suppose le défaut orthogonal.

On utilise la notation suivante pour tenir compte de la nature discrétisée des B-scans. Soit u la valeur d'intérêt :

$$\delta_u = \#_u \, \Delta_u, \tag{VIII.31}$$

où δ_u est la vraie mesure, $\#_u$ est le nombre de pixels correspondants et Δ_u le pas d'échantillonnage.

Ainsi, dans le cas 1D la formule (VIII.30) devient

$$h^{1\mathrm{D}} = \frac{V\Delta_t \left|\#_t^{1\mathrm{D}}\right|}{2\cos(\theta)}, \tag{VIII.32}$$

où $\#_t^{1\mathrm{D}} = \#_{t,\mathrm{coin}}^{1\mathrm{D}} - \#_{t,\mathrm{diffr.}}^{1\mathrm{D}}$. Les maximums d'énergie des échos de coin et de diffraction se trouvent dans le A-scan considérée de la réflectivité estimée (figure VIII.8(c)-(d)) respectivement aux instants $\#_{t,\mathrm{coin}}^{1\mathrm{D}}$ et $\#_{t,\mathrm{diffr.}}^{1\mathrm{D}}$.

Comme on peut le voir sur la figure VIII.8(h), les échos de coin et de diffraction estimés à partir de la déconvolution 2D ne se trouvent pas sur un même A-scan. Cet écart est parfaitement normal car il est dû au fait que l'écho de coin est vu avant l'écho de diffraction. Cependant, on ne peut pas appliquer directement la formule (VIII.32) estimant la hauteur du défaut car elle suppose que ces deux échos se trouvent sur un même A-scan. Il faut donc rajouter une étape estimant le temps de vol entre les échos de coin et de diffraction à partir de la réflectivité issue de la déconvolution 2D.

Le temps de vol entre les échos de coin et de diffraction $\#_t^{2\mathrm{D}}$ peut être estimé à partir de la réflectivité issue de la déconvolution 2D par

$$\#_t^{2\mathrm{D}} = \left|\nabla\#_t^{2\mathrm{D}}\right| - \tan\omega \left|\nabla\#_x^{2\mathrm{D}}\right|, \tag{VIII.33}$$

où $\nabla\#_\bullet^{2\mathrm{D}} = \#_{\bullet,\mathrm{coin}}^{2\mathrm{D}} - \#_{\bullet,\mathrm{diffr.}}^{2\mathrm{D}}$. ω est l'angle d'inclinaison des échos dans le plan B-scan (figure VIII.8(a)). Les maximums d'énergie des échos de coin et de diffraction se trouvent dans la réflectivité estimée (figure VIII.8(g)) respectivement aux positions $(\#_{x,\mathrm{coin}}^{2\mathrm{D}}, \#_{t,\mathrm{coin}}^{2\mathrm{D}})$ et $(\#_{x,\mathrm{diffr.}}^{2\mathrm{D}}, \#_{t,\mathrm{diffr.}}^{2\mathrm{D}})$.

La formule (VIII.30) estimant la hauteur du défaut suppose qu'on se trouve dans le plan B-scan où les échos sont inclinés. Or, des problèmes de décalage entre A-scans peuvent se manifester (section VIII.3.1). Ceci n'affecte pas l'utilisation de (VIII.30) dans le cadre d'une déconvolution 1D car on se base sur un unique A-scan. Par contre, l'utilisation directe de (VIII.33) à partir de la réflectivité 2D avec échos inclinés (figure VIII.8(g)) pour estimer le temps de vol entre échos de coin et de diffraction souffre alors d'un manque de robustesse.

On propose donc d'utiliser un plan B-scan virtuel avec échos inclinés robuste aux problèmes de décalage entre A-scans. Cette fois l'application de la formule (VIII.33) dans ce plan B-scan virtuel devient robuste par rapport au problème de décalage entre A-scans. A partir du plan B-scan* avec échos horizontaux (figure VIII.8(h)), le plan B-scan virtuel se détermine en conservant les $\#_x$, mais en effectuant l'opération suivante sur les $\#_t$:

$$\left|\nabla\#_t^{2\mathrm{D}}\right| = \left|\nabla\#_t^{2\mathrm{D}^*}\right| + \tan\omega^\dagger \left|\nabla\#_x^{2\mathrm{D}}\right|, \tag{VIII.34}$$

où $\omega^\dagger$ est l'angle d'inclinaison des échos dans le plan B-scan virtuel. Comme l'angle d'inclinaison $\omega^\dagger$ est très important (environ 80°), il n'est que très peu modifié par des décalages entre A-scans de quelques pixels. Ainsi, $\omega^\dagger$ est très proche de ω.

En combinant (VIII.33) et (VIII.34) et en approchant $\omega^\dagger$ par ω on obtient alors

$$\#_t^{2\mathrm{D}} = \left|\nabla\#_t^{2\mathrm{D}^*}\right|$$

d'où d'après (VIII.30) et (VIII.31) l'estimation de h dans le cadre de la déconvolution 2D :

$$h^{2\mathrm{D}} = \frac{V\Delta_t \left|\nabla\#_t^{2\mathrm{D}^*}\right|}{2\cos(\theta)}. \tag{VIII.35}$$

Il est intéressant de noter que cette formule est similaire à celle du cas 1D (VIII.32), mais que les espaces d'où sont tirées les mesures $\nabla\#_t^{2\mathrm{D}^*}$ et $\nabla\#_t^{1\mathrm{D}}$ ne sont pas les mêmes. Dans le premier cas

on utilise le plan B-scan* avec échos horizontaux (figure VIII.8(h)) tandis que dans le deuxième cas on utilise le plan B-scan avec échos inclinés (figure VIII.8(c)). De plus, la formule (VIII.35) ne fait pas intervenir l'angle ω d'inclinaison des échos dans le plan B-scan, ce qui en facilite l'utilisation. Une erreur d'un pixel induit une incertitude associée aux estimations de h qui est la même dans le cas 1D (VIII.32) et dans le cas 2D (VIII.35) et qui vaut

$$\frac{V\Delta_t}{2\cos(\theta)} = 0,29 \text{ mm}. \tag{VIII.36}$$

[D] Analyse quantitative des résultats

La méthode d'estimation de la hauteur du défaut proposée dans la partie précédente est appliquée sur les réflectivités estimées par déconvolution 1D et 2D. L'analyse des résultats des estimations de la hauteur des défauts (tableau VIII.2, page 125 et tableau VIII.3, page 126) montre tout d'abord que la répétabilité est bonne d'un B-scan à l'autre.

La répétabilité de la caractérisation du maximum des échos est bonne pour les angles. Par contre, la dispersion du module du maximum des échos est assez forte. Ceci s'explique très certainement par le fait que le même jeu de paramètres δ et λ est ici utilisé pour l'ensemble des données. Un réglage adapté à chaque échantillon de ces deux paramètres diminuerait cette dispersion du module du maximum des échos. Notons que la valeur du module du maximum des échos n'intervient pas dans la détermination de la hauteur du défaut, seule sa position dans le B-scan est utilisée.

Les résultats d'estimation de la hauteur des défauts sont tout à fait corrects comparés aux vraies valeurs de la hauteur des entailles. D'après la moyenne des six échantillons d'une même entaille (tableau VIII.2 et tableau VIII.3), la méthode 2D estime légèrement mieux la hauteur des défauts que la méthode 1D. Ce résultat est à relativiser car ce gain est inférieur à l'incertitude associée aux estimations de la hauteur des entailles donnée par (VIII.36). Malgré tout, l'estimation de la hauteur de l'entaille se révèle plus simple avec la méthode 2D qu'avec la méthode 1D car il n'y a plus à s'interroger sur le choix des A-scans à utiliser.

VIII.4 Conclusion

Dans ce chapitre, nous avons illustré la méthode proposée sur des données issues d'entailles. Les résultats de la méthode proposée se sont révélés très satisfaisants. Cependant, ce type de données est assez éloigné des véritables défauts rencontrés sur les sites d'inspection. Nous avons alors également testé la méthode proposée sur des données issues de fissures (fournies par le GSDTI d'EDF). Les résultats de la méthode proposée se sont aussi révélés très satisfaisants dans ce cadre. Par contre, nous avons observé une sensibilité plus marquée de la méthode proposée pour ce qui est du réglage des hyperparamètres. Si un même réglage d'hyperparamètres fournit de bons résultats pour les données issues d'une même entaille, ce n'est plus le cas pour les données issue d'une même fissure. Il est alors nécessaire de régler les hyperparamètres de la méthode proposée pour chaque B-scan issu d'une fissure. Ceci s'explique sans doute par la nature plus irrégulière d'une fissure que d'une entaille.

Une perspective possible serait l'estimation automatique de ces hyperparamètres. Cela permettrait d'accroître la simplicité de la méthode proposée, notamment dans le cas de données issues de fissures. On pourrait alors envisager une extension 2D de la méthode de déconvolution aveugle présentée dans l'annexe E, page 175. Une autre perspective est d'envisager une extension 3D de la méthode proposée en considérant plusieurs B-scans consécutifs plutôt qu'un seul, afin de tenir également compte de la composante transversale du faisceau ultrasonore.

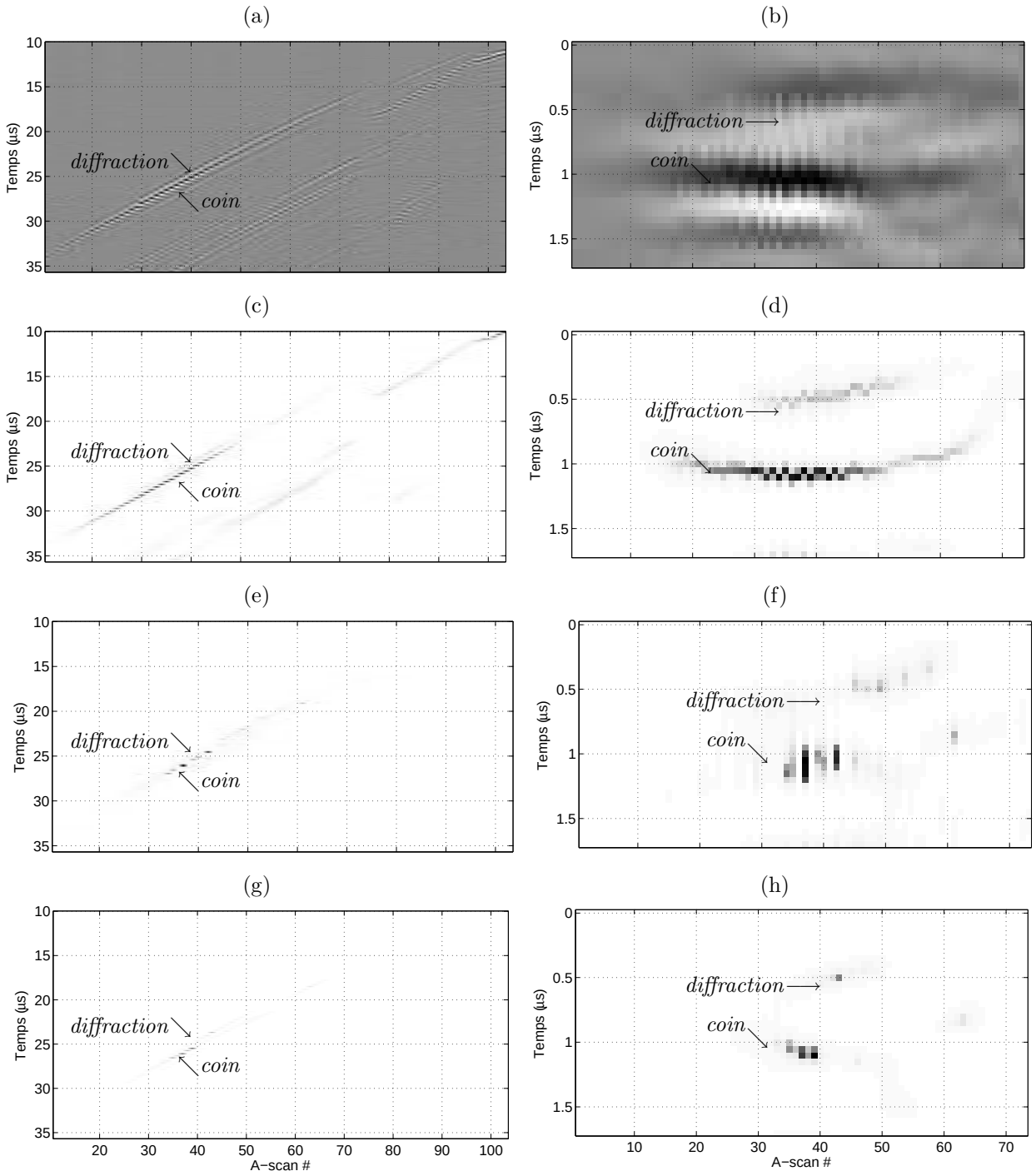


FIG. VIII.8: (a) Exemple d'un B-scan. (b) Version zoomée et recalée de (a). (c) Résultat après déconvolution 1D (module de $\widehat{\boldsymbol{a}}_m + i\,\widehat{\boldsymbol{b}}_m$, $m = 1, \ldots, M$) avec $\phi = \phi_{\text{hyp}}$, selon [Gautier *et al.*, 2001]. (d) Version zoomée et recalée de (c). (e) Résultat après déconvolution 2D (module de $\widehat{\mathbf{A}} + i\,\widehat{\mathbf{B}}$ avec $\mathbf{H}^{\text{align}}$ donnée par la figure VIII.6(a), avec $\phi = \phi_{\text{hyp}}$). (f) Version zoomée et recalée de (e). (g) Résultat après déconvolution 2D (module de $\widehat{\mathbf{A}} + i\,\widehat{\mathbf{B}}$ avec ondelette 2D séparable de la figure VIII.6(b), avec $\phi = \phi_{\text{hyp}}$). (h) Version zoomée et recalée de (g).

3mm	3a	3b	3c	3d	3e	3f	moyenne
angle de l'écho de coin	-176.9	-166.7	-167	-166.6	158.8	167.7	-178.5
module de l'écho de coin	100.2	63	122	91.8	137.7	83.5	99.7
angle de l'écho de diffraction	116.5	120.5	117.7	120.3	123.8	106.7	117.6
module de l'écho de diffraction	26.8	22.6	52.9	59.6	64.9	55.1	47
$\nabla\#_{t,\text{horiz}}$	12	12	12	12	13	12	12.17
h	3.49	3.49	3.49	3.49	3.78	3.49	**3.54**
5mm	5a	5b	5c	5d	5e	5f	moyenne
angle de l'écho de coin	-156	-148.4	-126.9	-126.3	-124.9	-122.1	-134.1
module de l'écho de coin	150.6	76.7	78.5	54.7	69	39.7	78.2
angle de l'écho de diffraction	-67.5	-55.8	-58.7	-49.6	-47.2	-74.4	-58.5
module de l'écho de diffraction	42.9	18.3	31.9	15.1	23.1	22.4	25.6
$\nabla\#_{t,\text{horiz}}$	15	15	14	14	14	14	14.33
h	4.36	4.36	4.07	4.07	4.07	4.07	**4.17**
8mm	8a	8b	8c	8d	8e	8f	moyenne
angle de l'écho de coin	-125.9	-134.7	-130.5	-130.4	-91.3	-118.8	-121.9
module de l'écho de coin	190.6	205.2	352.1	180.8	189.8	152.8	211.9
angle de l'écho de diffraction	58.7	83.9	81.0	81.3	69.2	87.1	76.9
module de l'écho de diffraction	76.3	68.1	84.2	69.5	62.6	40.1	66.8
$\nabla\#_{t,\text{horiz}}$	27	28	28	28	27	28	27.67
h	7.84	8.13	8.13	8.13	7.84	8.13	**8.03**

TAB. VIII.2 – Estimation de la hauteur des entailles avec l'ondelette 2D séparable (section VIII.3.2) d'après la formule (VIII.35). Les angles sont en degrés.

3mm	3a	3b	3c	3d	3e	3f	moyenne
nombre d'A-scans	12	16	13	16	17	19	12.83
minimum de $\nabla\#_t$	12	11	12	12	13	12	12
maximum de $\nabla\#_t$	14	14	14	14	15	14	14.17
moyenne de $\nabla\#_t$	13	12.81	13.23	13.44	13.59	13.05	13.19
minimum de h	3.49	3.2	3.49	3.49	3.78	3.49	3.49
maximum de h	4.07	4.07	4.07	4.07	4.36	4.07	4.12
moyenne de h	3.78	3.72	3.85	3.9	3.95	3.79	**3.83**
5mm	5a	5b	5c	5d	5e	5f	moyenne
nombre d'A-scans	16	16	14	15	14	14	14.83
minimum de $\nabla\#_t$	12	12	12	12	12	13	12.17
maximum de $\nabla\#_t$	17	16	16	16	16	16	16.17
moyenne de $\nabla\#_t$	14.63	14.44	14.14	14.2	13.86	14.36	14.27
minimum de h	3.49	3.49	3.49	3.49	3.49	3.78	3.54
maximum de h	4.94	4.65	4.65	4.65	4.65	4.65	4.7
moyenne de h	4.25	4.19	4.11	4.12	4.03	4.17	**4.15**
8mm	8a	8b	8c	8d	8e	8f	moyenne
nombre d'A-scans	11	10	11	9	7	8	9.33
minimum de $\nabla\#_t$	26	26	26	27	27	26	26.33
maximum de $\nabla\#_t$	27	27	28	28	28	28	27.67
moyenne de $\nabla\#_t$	26.55	26.9	27.27	27.56	27.29	26.75	27.05
minimum de h	7.55	7.55	7.55	7.84	7.84	7.55	7.65
maximum de h	7.84	7.84	8.13	8.13	8.13	8.13	8.03
moyenne de h	7.71	7.81	7.92	8	7.93	7.77	**7.86**

TAB. VIII.3 – Estimation de la hauteur des entailles avec l'ondelette 1D (section VIII.3.2) d'après la formule (VIII.32). Les A-scans considérés sont ceux dont les deux points correspondant aux échos de coin et de diffraction sont d'énergie suffisante (au moins 30% de l'énergie des points de plus forte énergie des deux échos).

Bibliographie

[Abeyratne *et al.*, 1997] U. R. Abeyratne, A. Petropulu, T. Golas, J. Reid, E. Conant et F. Forsberg. Higher-order vs. second-order statistics in ultrasound image deconvolution. *IEEE Trans. Ultrasonics Ferroelectrics Frequency Control*, 44 (6) : 1409–1416, novembre 1997.

[Abeyratne *et al.*, 1995] U. R. Abeyratne, A. Petropulu et J. Reid. Higher-order spectra based deconvolution of ultrasound images. *IEEE Trans. Ultrasonics Ferroelectrics Frequency Control*, 42 (6) : 1064–1075, novembre 1995.

[Al-Baali et Fletcher, 1996] M. Al-Baali et R. Fletcher. On the order of convergence of preconditioned nonlinear conjugate gradient methods. *SIAM J. Sci. Comput.*, 17 : 658–665, 1996.

[Bertsekas, 1999] D. P. Bertsekas. *Nonlinear programming.* Athena Scientific, Belmont, MA, USA, 2ème edition, 1999.

[Chalmond *et al.*, 2003] B. Chalmond, F. Coldefy, E. Goubet et B. Lavayssière. Coherent 3-D echo detection for utrasonic imaging. *IEEE Trans. Signal Processing*, 51 (3) : 592–601, 2003.

[Champagnat et Idier, 1993] F. Champagnat et J. Idier. Deconvolution of sparse spike trains accounting for wavelet phase shifts and colored noise. In *Proc. IEEE ICASSP*, pages 452–455, Minneapolis, MN, USA, 1993.

[Chan et Ng, 1996] R. H. Chan et M. K. Ng. Conjugate gradient methods for Toeplitz systems. *SIAM Rev.*, 38 (3) : 427–482, septembre 1996.

[Charbonnier *et al.*, 1997] P. Charbonnier, L. Blanc-Féraud, G. Aubert et M. Barlaud. Deterministic edge-preserving regularization in computed imaging. *IEEE Trans. Image Processing*, 6 (2) : 298–311, février 1997.

[Cooley et Tukey, 1965] J. W. Cooley et J. W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Math. Comp.*, 19 (90) : 297–301, 1965.

[Demirli et Saniie, 2001a] R. Demirli et J. Saniie. Model-based estimation of ultrasonic echoes. Part I : Analysis and algorithms. *IEEE Trans. Ultrasonics Ferroelectrics Frequency Control*, 48 (3) : 787–802, mai 2001.

[Demirli et Saniie, 2001b] R. Demirli et J. Saniie. Model-based estimation of ultrasonic echoes. Part II : Nondestructive evaluation applications. *IEEE Trans. Ultrasonics Ferroelectrics Frequency Control*, 48 (3) : 803–811, mai 2001.

[Demoment *et al.*, 1984] G. Demoment, R. Reynaud et A. Herment. Range resolution improvement by a fast deconvolution method. *Ultrasonic Imaging*, 6 : 435–451, 1984.

[Fatemi et Kak, 1980] M. Fatemi et A. C. Kak. Ultrasonic B-scan imaging : Theory of image formation and a technique for restoration. *Ultrasonic Imaging*, 2 : 1–47, 1980.

[Faur *et al.*, 1998] M. Faur, L. Paradis, J. Oksman et P. Morisseau. A two-step inverse procedure for outer surface defects characterization from ultrasonic bscan images. In *Review of Progress in Quantitative Nondestructive Evaluation*, volume 17, pages 815–822, 1998.

[Gautier *et al.*, 2001] S. Gautier, J. Idier, F. Champagnat et D. Villard. Restoring separate discontinuities from ultrasonic data. In *Review of Progress in Quantitative Nondestructive Evaluation, AIP Conf. Proc. Vol 615(1)*, pages 686–690, Brunswick, ME, USA, juillet 2001.

[Gautier *et al.*, 2002] S. Gautier, J. Idier, F. Champagnat et D. Villard. Procédé de mesure par ondes sonores et ultrasonores - Méthode de déconvolution. Brevet WO02086485, EDF/CNRS, France, 2002.

[Golub et Van Loan, 1996] G. H. Golub et C. F. Van Loan. *Matrix computations*. The Johns Hopkins University Press, Baltimore, 3ème edition, 1996.

[Hayward et Lewis, 1989] G. Hayward et J. E. L. Lewis. Comparison of some non-adaptive deconvolution techniques for resolution enhancement of ultrasonic data. *Ultrasonics*, 27 (3) : 155–164, mai 1989.

[Honarvar *et al.*, 2004] F. Honarvar, H. Sheikhzadeh, M. Moles et A. Sinclair. Improving the time-resolution and signal-to-noise ratio of ultrasonic NDE signals. *Ultrasonics*, 41 : 755–63, mars 2004.

[Hundt et Trautenberg, 1980] E. Hundt et E. Trautenberg. Digital processing of ultrasonic data by deconvolution. *IEEE Trans. Ultrasonics Ferroelectrics Frequency Control*, 27 (5) : 249–252, septembre 1980.

[Hunt, 1977] B. R. Hunt. Bayesian methods in nonlinear digital image restoration. *IEEE Trans. Communications*, C-26 : 219–229, mars 1977.

[Husby *et al.*, 2001] O. Husby, T. Lie, T. Lango, J. Hokland et H. Rue. Bayesian 2-D deconvolution : a model for diffuse ultrasound scattering. *IEEE Trans. Ultrasonics Ferroelectrics Frequency Control*, 48 (1) : 121–130, janvier 2001.

[Idier, 1999] J. Idier. Regularization tools and models for image and signal reconstruction. In *3rd Int. Conf. Inverse Problems in Engng.*, pages 23–29, Port Ludlow, WA, USA, juin 1999.

[Jensen, 1992] J. A. Jensen. Deconvolution of ultrasound images. *Ultrasonic Imaging*, 14 (1) : 1–15, janvier 1992.

[Jeurens *et al.*, 1987] T. J. M. Jeurens, J. C. Somer, F. A. M. Smeets et A. P. G. Hoeks. The practical significance of two-dimensional deconvolution in echography. *Ultrasonic Imaging*, 9 (2) : 106–116, avril 1987.

[Kaaresen, 1998] K. F. Kaaresen. Evaluation and applications of the iterated window maximization method for sparse deconvolution. *IEEE Trans. Signal Processing*, 46 (3) : 609–624, mars 1998.

[Kaaresen et Bølviken, 1999] K. F. Kaaresen et E. Bølviken. Blind deconvolution of ultrasonic traces accounting for pulse variance. *IEEE Trans. Ultrasonics Ferroelectrics Frequency Control*, 46 (3) : 564–573, mai 1999.

[Mu *et al.*, 2002] Z. Mu, R. Plemmons et P. Santago. Estimation of complex ultrasonic medium responses by deconvolution. In *Proc. IEEE Conf. on Medical Imaging*, 2002.

[Ng *et al.*, 1999] M. K. Ng, R. H. Chan et W.-C. Tang. A fast algorithm for deblurring models with Neumann boundary conditions. *SIAM J. Sci. Comput.*, 21 (3) : 851–866, 1999.

[Nikolova et Ng, 2005] M. Nikolova et M. Ng. Analysis of half-quadratic minimization methods for signal and image recovery. *SIAM J. Sci. Comput.*, 27 : 937–966, 2005.

[O'Brien *et al.*, 1994] M. S. O'Brien, A. N. Sinclair et S. M. Kramer. Recovery of a sparse spike time series by l_1 norm deconvolution. *IEEE Trans. Signal Processing*, 42 (12) : 3353–3365, décembre 1994.

[Sallard et Paradis, 1998] J. Sallard et L. Paradis. Use of a priori information for the deconvolution of ultrasonic signals. In *Review of Progress in Quantitative Nondestructive Evaluation*, volume 17, pages 735–742, 1998.

[Schomberg *et al.*, 1983] H. Schomberg, W. Vollmann et G. Mahnke. Lateral inverse filtering of ultrasonic B-scan images. *Ultrasonic Imaging*, 5 (1) : 38–54, janvier 1983.

[Sciacca et Evans, 1992] L. J. Sciacca et R. J. Evans. Signal processing applied to ultrasonic imaging. In *IEEE Trans. Acoust. Speech, Signal Processing*, pages 225–228, octobre 1992.

[Taxt et Jirik, 2004] T. Taxt et R. Jirik. Superresolution of ultrasound images using the first and second harmonic signal. *IEEE Trans. Ultrasonics Ferroelectrics Frequency Control*, 51 (2) : 163–175, février 2004.

[Tikhonov et Arsenin, 1977] A. Tikhonov et V. Arsenin. *Solutions of Ill-Posed Problems*. Winston, Washington, DC, USA, 1977.

[Vollmann, 1982] W. Vollmann. Resolution enhancement of ultrasonic B-scan images by deconvolution. *IEEE Trans. Sonics Ultrasonics*, 29 (2) : 78–83, mars 1982.

Chapitre IX

CONCLUSION ET PERSPECTIVES

Dans le cadre de ce travail de thèse, nous nous sommes intéressés à des questions algorithmiques liées à la mise en œuvre de l'approche pénalisée préservant les discontinuités pour des problèmes de restauration et de reconstruction de grande taille. Ce travail de thèse nous a amené à nous intéresser de près au domaine de l'optimisation de critères différentiables. Notre point de départ a été de chercher des alternatives aux algorithmes semi-quadratiques (SQ) qui sont généralement trop coûteux pour les problèmes de grande taille. Nous avons alors considéré des algorithmes contenant des ingrédients semi-quadratiques qui permettent de s'affranchir des limitations de taille. Il s'agit des algorithmes SQ approchés (SQ+GCP) et du gradient conjugué non linéaire avec pour recherche du pas la forme scalaire des algorithmes SQ (GCNL+SQ1D). Une part importante de nos efforts a porté sur l'établissement des preuves de convergence de ces algorithmes. Nous avons illustré expérimentalement sur un problème de déconvolution d'image la pertinence de ces algorithmes par rapport aux algorithmes SQ.

Nous indiquons maintenant quelques perspectives qui nous semblent intéressantes à court et moyen terme. Une des questions qui se posent naturellement à l'issue de ce travail est de savoir s'il est possible de trouver d'autres algorithmes convergents plus efficaces que les algorithmes SQ+GCP et GCNL+SQ1D. Dans le cadre de ce travail nous avons établi des liens forts entre les algorithmes SQ, SQ+GCP et GCNL+SQ1D. Il nous semble également intéressant de poursuivre l'étude des liens de ces algorithmes avec d'autres algorithmes figurant dans la littérature de l'optimisation.

IX.1 Liens entre algorithmes SQ+GCP et Newton tronqué

Des algorithmes de type Newton tronqué, dans le sens où la direction de Newton est approchée, ont été proposés en optimisation [Nash, 2000]. Or, la nature des algorithmes SQ+GCP est similaire à celle de certains algorithmes de type Newton tronqué. Une perspective serait d'utiliser un algorithme de type Newton tronqué, lorsque le critère pénalisé est strictement convexe. En particulier, il serait intéressant de voir si l'utilisation du Hessien à la place des matrices SQ apporte un gain de performance.

IX.2 Lien entre algorithmes GCNL et méthodes à mémoire de gradient

Les algorithmes GCNL font intervenir une direction de descente définie par

$$\boldsymbol{d}_k = -\nabla \mathcal{J}(\boldsymbol{x}_k) + \beta_k \boldsymbol{d}_{k-1} \tag{IX.1}$$

qui dépend de la direction précédente et du gradient courant. La nouvelle itération s'exprime alors sous la forme suivante

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k$$

et une recherche du pas α_k scalaire doit alors être envisagée. Cette recherche du pas constitue en fait la mise en œuvre d'un algorithme de minimisation monodimensionnel. Nous avons proposé dans ce travail l'utilisation des algorithmes SQ scalaires (SQ1D) pour les critères pénalisés. Une des difficultés associées aux algorithmes GCNL est le choix de la formule de conjugaison β_k (IX.1). En effet, il existe de nombreuses formules de conjugaison [Hager et Zhang, 2006]. Or, le choix de la formule de conjugaison a des répercussions importantes sur les performances de l'algorithme GCNL. On peut envisager de rechercher simultanément les coefficients α_k et β_k en ayant recours à un algorithme de minimisation bidimensionnel. L'intérêt de cette nouvelle approche est qu'on se débarrasse alors de manière séduisante de la difficulté du choix de la formule de conjugaison. On peut montrer que dans le cas d'un critère quadratique cette approche d'identifie bien à l'algorithme du gradient conjugué. On pourrait alors étendre l'approche de recherche du pas SQ1D étudiée dans le cadre de ce travail à une approche bidimensionnelle SQ2D.

Les algorithmes GCNL font en fait partie d'une classe d'algorithmes plus générale qui est celle des algorithmes de type mémoire de gradient [Shi et Shen, 2004, 2005]. Les algorithmes de type mémoire de gradient consistent à définir une direction de descente qui ne dépend plus seulement de la direction précédente, mais des M directions précédentes :

$$\boldsymbol{d}_k = -\nabla \mathcal{J}(\boldsymbol{x}_k) + \sum_{m=1}^{M} \beta_{k+1-m} \boldsymbol{d}_{k-m} \tag{IX.2}$$

où M est l'ordre de mémoire considéré. Lorsque $M = 1$, on retrouve une direction de descente de même structure que celle des algorithmes GCNL. Ces algorithmes de type mémoire de gradient restent dans le même esprit que les algorithmes GCNL dans le sens où les coefficients β_{k+1-m} sont également définis par des formules analytiques. Un exemple de ce type d'algorithme est l'algorithme de type BFGS à mémoire limitée (l-BFGS) [Nocedal et Wright, 1999, Sec. 9.1]. L'intérêt de cet algorithme est qu'il nécessite un volume de calcul et de stockage à chaque itération plus restreint que l'algorithme BFGS (section IV.3.2.[D], page 60).

Cependant, on peut envisager une extension multidimensionnelle de la recherche du pas SQ1D. En effet, d'après (IX.2), la nouvelle itération peut s'exprimer sous la forme suivante

$$\begin{aligned} \boldsymbol{x}_{k+1} &= \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k \\ &= \boldsymbol{x}_k - \alpha_k \nabla \mathcal{J}(\boldsymbol{x}_k) + \sum_{m=1}^{M} \alpha_k^{(m)} \boldsymbol{d}_{k-m} \end{aligned}$$

où $\alpha_k^{(m)} = \alpha_k \beta_{k+1-m}$. On peut alors considérer un pas de dimension $M+1$ formé des valeurs $\{\alpha_k, \alpha_k^{(1)}, \ldots, \alpha_k^{(M)}\}$. On pourrait alors mettre en œuvre une recherche du pas semi-quadratique multidimensionnelle SQMD. A notre connaissance, cette approche est originale pour les critères pénalisés. Il serait intéressant de voir si en prenant un ordre de mémoire M plus grand que 1 (mais pas trop grand), cette approche est plus performante que les algorithmes GCNL+SQ1D.

IX.3 Préconditionnement variable pour les algorithmes GCNL

Nous n'avons établi la convergence des algorithmes GCNL+SQ1D dans le Théorème VI.3.2, page 89 que dans le cas d'un préconditionnement constant. La possibilité d'étendre la convergence des algorithmes GCNL+SQ1D pour une suite de matrices de préconditionnement variables

est à l'heure actuelle un problème ouvert. Une telle extension pourrait être envisagée simplement sous la forme d'une hybridation entre les algorithmes SQ+GCP et GCNL+SQ1D. En effet, les algorithmes SQ+GCP sont des algorithmes sans conjugaison permettant un préconditionnement variable tandis que les algorithmes GCNL+SQ1D sont des algorithmes conjugués permettant pour l'instant uniquement un préconditionnement constant. L'intérêt résiderait dans la possibilité que cet algorithme hybride présente des performances meilleures que ses deux parents.

IX.4 La question de la simplicité

Les algorithmes d'optimisation considérés présentent tous des paramètres de réglage. Ils se distinguent par le nombre de paramètres de réglage mais aussi par leur sensibilité de réglage en terme de performance. Ces deux aspects constituent selon nous une notion de simplicité d'utilisation. Cette notion nous semble d'un intérêt pratique majeur. En effet, il est naturel de choisir l'algorithme le plus simple entre deux algorithmes de performance similaire. On pourrait même envisager de mettre en avant la notion de simplicité d'utilisation. Ainsi, on pourrait être amené à sélectionner un algorithme qui ne serait pas systématiquement le plus performant, mais qui aurait le mérite d'être le plus simple à utiliser. Cette question de la simplicité d'utilisation nous semble être un enjeu important pour la recherche de nouveaux algorithmes de minimisation des critères pénalisés.

En résumé, les algorithmes proposés et étudiés dans ce travail de thèse, pour la minimisation des critères pénalisés préservant les discontinuités dans le cas de problèmes de grande taille, sont convergents et établissent un bon compromis entre efficacité et simplicité. Cependant, on peut espérer trouver d'autres algorithmes convergents qui établissent un compromis aussi bon, voire meilleur que ces algorithmes. Pour finir, indiquons qu'une partie du travail prévu dans un projet accepté en 2005 par l'Agence Nationale de la Recherche suite à l'appel à projet *Masse de Données* (ARA - MDSA) va consister à développer certains résultats présentés dans cette thèse dans le cadre d'une application spécifique à l'imagerie biomédicale, notamment avec l'hypothèse de bruit poissonien plutôt que gaussien.

Bibliographie

[Hager et Zhang, 2006] W. W. Hager et H. Zhang. A survey of nonlinear conjugate gradient methods. *Pacific J. Optim.*, 2 (1) : 35–58, janvier 2006.

[Nash, 2000] S. G. Nash. A survey of truncated-Newton methods. *J. Comput. Appl. Math.*, 124 : 45–59, 2000.

[Nocedal et Wright, 1999] J. Nocedal et S. J. Wright. *Numerical optimization*. Springer Texts in Operations Research. Springer-Verlag, New York, NY, USA, 1999.

[Shi et Shen, 2004] Z.-J. Shi et J. Shen. A gradient-related algorithm with inexact line searches. *J. Comput. Appl. Math.*, 170 (2) : 349–370, 2004.

[Shi et Shen, 2005] Z.-J. Shi et J. Shen. Convergence property and modifications of a memory gradient method. *Asia-Pac. J. Oper. Res.*, 22 (4) : 463–477, 2005.

Annexes

Annexes

Annexe A

BIBLIOGRAPHIE COMMENTÉE SUR LES MÉTHODES GCNL ET LA RECHERCHE DU PAS

Les méthodes du gradient conjugué non linéaire (GCNL) sont des algorithmes d'optimisation pour les critères différentiables qui se caractérisent par un faible encombrement mémoire et par l'existence de résultats de convergence généraux. Cette annexe a pour objectif d'établir une brève bibliographie sur les algorithmes GCNL et sur leurs résultats de convergence. Ces résultats de convergence sont malheureusement peu connus en traitement du signal et des images. Dans ce domaine, les algorithmes de type GCNL sont souvent utilisés sans assurance de la convergence. D'ailleurs, on ne trouve pas d'implémentation des algorithmes GCNL dans le logiciel de calcul scientifique `Matlab` ni dans la `Toolbox Optimization`. Une bibliographie détaillée de l'algorithme GC pour la période 1948-1976 se trouve dans [Golub et O'Leary, 1989]. On peut trouver une bibliographie plus récente sur les algorithmes GCNL dans [Hager et Zhang, 2006].

A.1 Algorithmes GCNL

L'histoire des algorithmes GCNL débute avec l'algorithme du gradient conjugué (GC) défini par (V.5)-(V.9), page 77. L'objectif de l'algorithme GC est la résolution de systèmes linéaires symétriques définis positifs [Hestenes et Stiefel, 1952]. Cependant, résoudre un système linéaire est équivalent à minimiser un critère quadratique [Bertsekas, 1999, p. 130]. Ainsi, l'algorithme du GC peut être vu comme minimisant une quadratique convexe. Ce constat a alors permis d'envisager au cours des années 1960 d'utiliser l'algorithme GC pour des critères non quadratiques [Fletcher et Reeves, 1964]. On parle alors d'algorithmes GC non linéaires (GCNL). Les algorithmes GCNL sont appliqués au problème d'optimisation non linéaire et non contraint suivant

$$\min \mathcal{J}(\boldsymbol{x}) : \boldsymbol{x} \in \mathbb{R}^N$$

où $\mathcal{J} : \mathbb{R}^N \mapsto \mathbb{R}$ est continûment différentiable. Un algorithme GCNL génère une suite d'itérées $\{\boldsymbol{x}_k\}$ d'après la formule de récurrence

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \, \boldsymbol{d}_k$$

où le pas positif α_k est obtenu par une recherche de pas et les directions $\{\boldsymbol{d}_k\}$ sont générées par la formule suivante

$$\boldsymbol{d}_{k+1} = -\boldsymbol{g}_{k+1} + \beta_k \boldsymbol{d}_k$$

avec $\boldsymbol{d}_0 = -\boldsymbol{g}_0$ et en notant $\boldsymbol{g}_k = \nabla \mathcal{J}(\boldsymbol{x}_k)$.

Les algorithmes GCNL se distinguent selon le paramètre scalaire β_k utilisé. Nous indiquons ci-dessous les formes les plus connues

$$\beta_k^{\mathrm{HS}} = \boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{y}_{k-1}/\boldsymbol{d}_{k-1}^{\mathrm{t}}\boldsymbol{y}_{k-1} \quad \text{[Hestenes et Stiefel, 1952]} \tag{A.1}$$
$$\beta_k^{\mathrm{FR}} = \|\boldsymbol{g}_k\|^2/\|\boldsymbol{g}_{k-1}\|^2 \quad \text{[Fletcher et Reeves, 1964]} \tag{A.2}$$
$$\beta_k^{\mathrm{PRP}} = \boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{y}_{k-1}/\|\boldsymbol{g}_{k-1}\|^2 \quad \text{[Polak et Ribière, 1969; Polyak, 1969]} \tag{A.3}$$
$$\beta_k^{\mathrm{LS}} = -\boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{y}_{k-1}/\boldsymbol{d}_{k-1}^{\mathrm{t}}\boldsymbol{g}_{k-1} \quad \text{[Liu et Storey, 1991]} \tag{A.4}$$

où $\boldsymbol{y}_{k-1} = \boldsymbol{g}_k - \boldsymbol{g}_{k-1}$.

Tous les algorithmes GCNL ont pour point commun de s'identifier à l'algorithme GC dans le cas d'un critère quadratique convexe lorsque le pas optimal défini par (V.5), page 77 est utilisé.

A.2 Recherche du pas

A.2.1 Nécessité d'une recherche du pas

Lors de chaque itération k d'un algorithme GCNL, le pas α_k est déterminé par une recherche du pas. La recherche du pas a pour objectif de trouver un pas qui fait décroître suffisament le critère tout en assurant la convergence. Le pas qui fait décroître le plus le critère est le pas optimal défini par

$$\hat{\alpha} = \arg\min_{\alpha} f(\alpha)$$

où $f(\alpha) = \mathcal{J}(\boldsymbol{x}_k + \alpha\boldsymbol{d}_k)$. Dans le cas d'un critère pénalisé non quadratique, il est en général impossible d'obtenir le pas optimal, ou alors un tel résultat serait trop coûteux. Une étape de recherche du pas, *i.e.,* une minimisation 1D de f, est requise. L'objectif d'une telle recherche du pas est de ne pas être trop coûteuse tout en assurant la convergence de l'algorithme.

La nécessité d'avoir recours à une recherche du pas n'est pas propre aux algorithmes GCNL. Il existe en optimisation une littérature qui s'intéresse à la recherche de pas de manière générale [Moré et Thuente, 1994]. Notons qu'il n'y a pas véritablement de recherche du pas universelle. Par contre, les résultats de convergence des algorithmes GCNL s'appuient très souvent sur une variante des conditions de Wolfe présentées ci-dessous.

A.2.2 Conditions de Wolfe

Les conditions standard de Wolfe pour la recherche du pas α_k sont les suivantes [Wolfe, 1971]

$$\mathcal{J}(\boldsymbol{x}_k + \alpha_k\boldsymbol{d}_k) - \mathcal{J}(\boldsymbol{x}_k) \leq c_1\alpha_k\nabla\boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{d}_k \tag{A.5}$$
$$\nabla\mathcal{J}(\boldsymbol{x}_k + \alpha_k\boldsymbol{d}_k)^{\mathrm{t}}\boldsymbol{d}_k \geq c_2\boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{d}_k \tag{A.6}$$

avec $0 < c_1 < c_2 < 1$. L'équation (A.5) est appelée *condition d'Armijo* et l'équation (A.6) est appelée *condition de courbure*.

Ces conditions de Wolfe méritent une explication. La condition d'Armijo (A.5) est une condition de descente suffisante du critère $\mathcal{J}$. Cette condition n'est pas suffisante à elle seule pour assurer la convergence de l'algorithme. En effet, elle est vérifiée pour tout pas suffisamment petit. On court alors le risque de converger vers un point qui n'est pas stationnaire. C'est pourquoi la condition de courbure (A.6) est introduite afin d'éviter les pas trop petits.

La partie gauche de la condition de courbure (A.6) n'est autre que la dérivée $f'(\alpha_k)$. Ainsi, la condition de courbure assure simplement que la pente de la fonction f en α_k est plus grande

que c_2 fois la pente en 0. Lorsque la pente $f'(0)$ est fortement négative, cela laisse à penser qu'on peut diminuer le critère $\mathcal{J}$ de manière significative en se déplaçant davantage selon la direction courante $\boldsymbol{d}_k$. Par contre, si la pente $f'(0)$ est faiblement négative ou positive, on s'attend à ne pas pouvoir diminuer le critère $\mathcal{J}$ de manière significative selon la direction courante $\boldsymbol{d}_k$.

On peut aussi ne pas autoriser les pas α_k induisant une pente $f'(\alpha_k)$ trop positive. Les conditions standard de Wolfe peuvent être renforcées sous la forme des conditions fortes de Wolfe [Nocedal et Wright, 1999, p. 39] en substituant

$$\left|\nabla\mathcal{J}(\boldsymbol{x}_k+\alpha_k\boldsymbol{d}_k)^{\mathrm{t}}\boldsymbol{d}_k\right| \leq c_2\left|\boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{d}_k\right| \tag{A.7}$$

à la condition de courbure (A.6).

Il est à noter que pour un critère continûment différentiable et borné inférieurement les conditions standard de Wolfe (A.5)-(A.6) et les conditions fortes de Wolfe (A.5)-(A.7) sont toujours vérifiables [Nocedal et Wright, 1999, Lemma 3.1].

La fonction de minimisation scalaire de Matlab `fminbnd` n'implémente pas les conditions de Wolfe. Par contre, le lecteur trouvera dans [Nocedal et Wright, 1999, Algorithms 3.2 et 3.3] une implémentation de recherche du pas qui garantit de trouver un pas satisfaisant les conditions fortes de Wolfe.

A.3 Résultats de convergence

Il nous semble important de souligner que, de manière contre-intuitive, l'utilisation du pas optimal n'assure pas toujours la convergence des algorithmes GCNL. En effet, il est établi dans [Powell, 1984] que l'utilisation de la méthode de Polak-Ribiere avec pas optimal peut, pour un critère non convexe, cycler infiniment sans atteindre un point stationnaire. Ce résultat montre que se préoccuper des conditions que doit vérifier la recherche du pas fournissant le pas α_k est de toute importance pour assurer la convergence des algorithmes GCNL.

Le théorème suivant établit la convergence de la forme de Fletcher-Reeves en supposant que la recherche du pas vérifie les conditions fortes de Wolfe.

Théorème A.3.1. *[Nocedal et Wright, 1999, Theorem 5.8] On suppose que les lignes de niveau* $L = \{\boldsymbol{x} \in \mathbb{R}^n | \mathcal{J}(\boldsymbol{x}) \leq \mathcal{J}(\boldsymbol{x}_0)\}$ *sont bornées et que le critère* $\mathcal{J}$ *est gradient Lipschitz sur un voisinage de* L.

Alors l'algorithme GCNL avec la forme de Fletcher-Reeves (A.2) *implémenté avec une recherche du pas satisfaisant les conditions fortes de Wolfe* (A.5)-(A.7) *avec* $c_1 < c_2 < 1/2$ *est convergent dans le sens suivant*

$$\liminf_{k\to\infty} \|\nabla\mathcal{J}(\boldsymbol{x}_k)\| = 0.$$

Il existe des résultats de convergence pour les autres formes de GCNL [Hager et Zhang, 2006]. En général, les hypothèses du Théorème A.3.1 ne suffisent pas à assurer la convergence des autres formes de GCNL. Pour ne traiter que de la forme de Polak-Ribiere, une condition portant sur la recherche du pas ne suffit pas. Il faut rajouter notamment des hypothèses sur la suite des directions $\{\boldsymbol{d}_k\}$ [Gilbert et Nocedal, 1992, Theorem 4.3]. Il s'agit en particulier de la *condition de descente suffisante* définie ci-dessous

$$\boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{d}_k < -c\|\boldsymbol{g}_k\|^2$$

avec $c > 0$. Ce type de condition fournit, à notre sens, un théorème peu exploitable en pratique. En effet, on n'est pas assuré de pouvoir vérifier la condition de descente suffisante à chaque itération. De plus, la question du choix du paramètre c se pose. On voit que l'implémentation

de la forme de Polak-Ribiere dans le cadre du Théorème [Gilbert et Nocedal, 1992, Theorem 4.3] est plus lourde que l'implémentation de la forme Fletcher-Reeves car elle demande de vérifier à la fois des conditions sur la recherche du pas et sur les directions générées. Néanmoins, selon [Nocedal et Wright, 1999, p. 130], la méthode de Polak-Ribiere est plus performante que celle de Fletcher-Reeves.

Bibliographie

[Bertsekas, 1999] D. P. Bertsekas. *Nonlinear programming*. Athena Scientific, Belmont, MA, USA, 2ème edition, 1999.

[Fletcher et Reeves, 1964] R. Fletcher et C. M. Reeves. Function minimization by conjugate gradients. *Comp. J.*, 7 : 149–157, 1964.

[Gilbert et Nocedal, 1992] J. C. Gilbert et J. Nocedal. Global convergence properties of conjugate gradient methods for optimization. *SIAM J. Optimization*, 2 (1) : 21–42, 1992.

[Golub et O'Leary, 1989] G. Golub et D. O'Leary. Some history of the conjugate gradient and Lanczos methods : 1948-1976. *SIAM Rev.*, 31 (1) : 50–102, mars 1989.

[Hager et Zhang, 2006] W. W. Hager et H. Zhang. A survey of nonlinear conjugate gradient methods. *Pacific J. Optim.*, 2 (1) : 35–58, janvier 2006.

[Hestenes et Stiefel, 1952] M. R. Hestenes et E. Stiefel. Methods of conjugate gradients for solving linear system. *J. Res. Nat. Bur. Stand.*, 49 : 409–436, 1952.

[Liu et Storey, 1991] Y. Liu et C. Storey. Efficient generalized conjugate gradient algorithms, part 1 : Theory. *Journal of Optimisation Theory and Applications*, 69 : 129–137, 1991.

[Moré et Thuente, 1994] J. J. Moré et D. J. Thuente. Line search algorithms with guaranteed sufficient decrease. *ACM trans. on Mathematical Software*, 20 (3) : 286–307, septembre 1994.

[Nocedal et Wright, 1999] J. Nocedal et S. J. Wright. *Numerical optimization*. Springer Texts in Operations Research. Springer-Verlag, New York, NY, USA, 1999.

[Polak et Ribière, 1969] E. Polak et G. Ribière. Note sur la convergence de méthodes des directions conjuguées. *Rev. Française d'Informatique et de Recherche Opérationelle*, 16 : 35–43, 1969.

[Polyak, 1969] B. T. Polyak. The conjugate gradient method in extreme problems. *USSR Comp. Math. and Math. Phys.*, 9 : 94–112, 1969.

[Powell, 1984] M. J. D. Powell. *Nonconvex minimization calculations and the conjugate gradient method*, volume 1066 de *Lecture Notes in Mathematics*. Springer Verlag, Berlin, 1984.

[Wolfe, 1971] P. Wolfe. Convergence conditions for ascent methods. II : Some corrections. *SIAM Rev.*, 13 : 185–188, 1971.

Annexe B

PREUVES DES RÉSULTATS

B.1 Preuves du chapitre IV

B.1.1 Preuve du Lemme IV.4.1

Preuve. Soit $g(u) = u^2/2 - a\phi(u)$ avec $0 < a < \hat{a}$. On a $g'(u) = u - a\phi'(u)$. Soit $u \leq v$. On a

$$\begin{aligned} g'(u) - g'(v) &= a(\phi'(v) - \phi'(u)) - (v - u) \\ &= a(\phi'(v) - \phi'(u)) - |v - u| \\ &\leq \hat{a}\left|\phi'(v) - \phi'(u))\right| - |v - u| \end{aligned}$$

Or d'après l'hypothèse (IV.10b) on a $|\phi'(v) - \phi'(u))| \leq \frac{1}{\hat{a}}|v - u|$ d'où

$$g'(u) - g'(v) \leq 0$$

c'est-à-dire g' est croissante donc g est convexe.

□

B.1.2 Preuve du Lemme IV.4.4

Preuve. On s'inspire de la démonstration de [Chan et Mulet, 1999]. Soit $\boldsymbol{x}, \boldsymbol{x}^+ \in \mathbb{R}^N$. En notant que la première partie du critère $\mathcal{J}$ est quadratique, d'après (IV.28) page 67 on a

$$\begin{aligned} \widehat{\mathcal{J}}_{\boldsymbol{Q}_{GR}}(\boldsymbol{x}, \boldsymbol{x}^+) - \mathcal{J}(\boldsymbol{x}^+) = \\ \Phi(\boldsymbol{x}) - \Phi(\boldsymbol{x}^+) + (\boldsymbol{x}^+ - \boldsymbol{x})^{\mathrm{t}} \nabla\Phi(\boldsymbol{x}) + (\boldsymbol{x}^+ - \boldsymbol{x})^{\mathrm{t}} \mathbf{V}^{\mathrm{t}} \mathbf{L}(\boldsymbol{x}) \mathbf{V} (\boldsymbol{x}^+ - \boldsymbol{x})/2. \end{aligned} \quad \text{(B.1)}$$

Soit $\boldsymbol{u}_c = \mathbf{V}_c \boldsymbol{x} - \boldsymbol{\omega}_c$, $\boldsymbol{u}_c^+ = \mathbf{V}_c \boldsymbol{x}^+ - \boldsymbol{\omega}_c$ et $\boldsymbol{\Delta}_c = \boldsymbol{u}_c^+ - \boldsymbol{u}_c$. D'après la structure du gradient de la pénalisation $\nabla\Phi(\boldsymbol{x})$ (IV.4) page 55, on peut écrire (B.1) sous la forme

$$\begin{aligned} &\sum_{c=1}^{C} \phi(\|\boldsymbol{u}_c\|) - \phi(\|\boldsymbol{u}_c^+\|) + \boldsymbol{\Delta}_c^{\mathrm{t}} \boldsymbol{u}_c \frac{\phi'(\|\boldsymbol{u}_c\|)}{\|\boldsymbol{u}_c\|} + \|\boldsymbol{\Delta}_c\|^2 \frac{\phi'(\|\boldsymbol{u}_c\|)}{2\|\boldsymbol{u}_c\|} \\ =&\sum_{c=1}^{C} \phi(\|\boldsymbol{u}_c\|) - \phi(\|\boldsymbol{u}_c^+\|) + (\|\boldsymbol{u}_c^+\|^2 - \|\boldsymbol{u}_c\|^2) \frac{\phi'(\|\boldsymbol{u}_c\|)}{2\|\boldsymbol{u}_c\|}. \end{aligned} \quad \text{(B.2)}$$

En posant $\delta_c = \|\boldsymbol{u}_c\|$ et $\delta_c^+ = \|\boldsymbol{u}_c^+\|$, (B.2) s'écrit

$$\sum_{c=1}^{C} \phi(\delta_c) - \phi(\delta_c^+) + ((\delta_c^+)^2 - \delta_c^2) \frac{\phi'(\delta_c)}{2\delta_c}. \quad \text{(B.3)}$$

La somme (B.3) est la même que celle dans [Chan et Mulet, 1999]. La suite de la démonstration est alors identique. Chaque terme de la somme (B.3) est de la forme

$$\phi(a) - \phi(b) + (a^2 - b^2)\frac{\phi'(a)}{2a}. \tag{B.4}$$

Soit $\psi(t) = \phi(\sqrt{t})$. L'équation (B.4) s'écrit sous la forme

$$\psi(a^2) - \psi(b^2) + \psi'(a^2)(a^2 - b^2)$$

qui est positive car ψ est concave d'après l'hypothèse (IV.11b) page 62. □

B.1.3 Preuve du Lemme IV.4.5

Preuve. On a $\nabla\phi(\|\boldsymbol{x}\|) = \frac{\phi'(\|\boldsymbol{x}\|)}{\|\boldsymbol{x}\|}\boldsymbol{x}$, qui est bien défini en effectuant un prolongement par continuité en zéro d'après $\phi'(0) = 0$.

On a

$$\begin{aligned}
&\|\nabla\phi(\|\boldsymbol{y}\|) - \nabla\phi(\|\boldsymbol{x}\|)\|^2 - L^2\|\boldsymbol{y} - \boldsymbol{x}\|^2 \\
&= \left\| \frac{\phi'(\|\boldsymbol{y}\|)}{\|\boldsymbol{y}\|}\boldsymbol{y} - \frac{\phi'(\|\boldsymbol{x}\|)}{\|\boldsymbol{x}\|}\boldsymbol{x} \right\|^2 - L^2\|\boldsymbol{y} - \boldsymbol{x}\|^2 \\
&= \phi'(\|\boldsymbol{y}\|)^2 + \phi'(\|\boldsymbol{x}\|)^2 - L^2\|\boldsymbol{y}\|^2 - L^2\|\boldsymbol{x}\|^2 - 2\boldsymbol{x}^{\mathrm{t}}\boldsymbol{y}\left(\frac{\phi'(\|\boldsymbol{y}\|)\phi'(\|\boldsymbol{x}\|)}{\|\boldsymbol{y}\|\|\boldsymbol{x}\|} - L^2\right).
\end{aligned} \tag{B.5}$$

D'après le caractère gradient Lipschitz de ϕ on a

$$\begin{aligned}
&(\phi'(\|\boldsymbol{y}\| - \phi'(\|\boldsymbol{x}\|)^2 \leq L^2(\|\boldsymbol{y}\| - \|\boldsymbol{x}\|)^2 \\
&\phi'(\|\boldsymbol{y}\|)^2 + \phi'(\|\boldsymbol{x}\|)^2 - L^2\|\boldsymbol{y}\|^2 - L^2\|\boldsymbol{x}\|^2 \leq 2\left(\phi'(\|\boldsymbol{y}\|)\phi'(\|\boldsymbol{x}\|) - L^2\|\boldsymbol{y}\|\|\boldsymbol{x}\|\right).
\end{aligned} \tag{B.6}$$

D'après (B.5) et (B.6) on a

$$\begin{aligned}
&\|\nabla\phi(\|\boldsymbol{y}\|) - \nabla\phi(\|\boldsymbol{x}\|)\|^2 - L^2\|\boldsymbol{y} - \boldsymbol{x}\|^2 \\
&\leq 2\left(\phi'(\|\boldsymbol{y}\|)\phi'(\|\boldsymbol{x}\|) - L^2\|\boldsymbol{y}\|\|\boldsymbol{x}\|\right) - 2\boldsymbol{x}^{\mathrm{t}}\boldsymbol{y}\left(\frac{\phi'(\|\boldsymbol{y}\|)\phi'(\|\boldsymbol{x}\|)}{\|\boldsymbol{y}\|\|\boldsymbol{x}\|} - L^2\right) \\
&= \frac{2}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|}\left(\phi'(\|\boldsymbol{y}\|)\phi'(\|\boldsymbol{x}\|) - L^2\|\boldsymbol{y}\|\|\boldsymbol{x}\|\right)\left(\|\boldsymbol{y}\|\|\boldsymbol{x}\| - \boldsymbol{x}^{\mathrm{t}}\boldsymbol{y}\right).
\end{aligned} \tag{B.7}$$

D'après le caractère gradient Lipschitz de ϕ et $\phi'(0) = 0$ on a

$$\left|\phi'(\|\boldsymbol{u}\|)\right| \leq L\|\boldsymbol{u}\|$$

d'où

$$\phi'(\|\boldsymbol{y}\|)\phi'(\|\boldsymbol{x}\|) - L^2\|\boldsymbol{y}\|\|\boldsymbol{x}\| \leq 0. \tag{B.8}$$

D'après (B.7), (B.8) et l'inégalité de Cauchy-Schwartz on a

$$\|\nabla\phi(\|\boldsymbol{y}\|) - \nabla\phi(\|\boldsymbol{x}\|)\|^2 - L^2\|\boldsymbol{y} - \boldsymbol{x}\|^2 \leq 0$$

d'où

$$\|\nabla\phi(\|\boldsymbol{y}\|) - \nabla\phi(\|\boldsymbol{x}\|)\| \leq L\|\boldsymbol{y} - \boldsymbol{x}\|.$$

□

B.2 Preuves du chapitre V

B.2.1 Résultats pour l'algorithme GCP

Let us first restate the following orthogonal property of the GCP algorithm [Bertsekas, 1999, p. 137].

Lemma B.2.1. *For any initial guess $\boldsymbol{u}_0 \in \mathbb{R}^N$, the GCP algorithm (V.5)-(V.9) page 77 ensures that the residual $\boldsymbol{r}_i$ is orthogonal to the previous descent directions $\boldsymbol{p}_0, \ldots, \boldsymbol{p}_{i-1}$*

$$\boldsymbol{r}_i^{\mathrm{t}} \boldsymbol{p}_j = 0, \quad \forall i, \forall j < i. \tag{B.9}$$

In the sequel, a zero initial guess $\boldsymbol{u}_0 = \mathbf{0}$ is assumed. We did not find in the literature the following lemmas B.2.2 and B.2.3 on the GCP algorithm.

Lemma B.2.2. *Let $\boldsymbol{u}_0 = \mathbf{0}$. The GCP algorithm (V.5)-(V.9) page 77 ensures that the residual $\boldsymbol{r}_i$ is orthogonal to the iterate $\boldsymbol{u}_i$*

$$\boldsymbol{r}_i^{\mathrm{t}} \boldsymbol{u}_i = \mathbf{0}, \quad \forall i, \tag{B.10}$$

and

$$\boldsymbol{b}^{\mathrm{t}} \boldsymbol{u}_i = \boldsymbol{u}_i^{\mathrm{t}} \mathbf{A} \boldsymbol{u}_i, \quad \forall i. \tag{B.11}$$

Proof. According to (V.6) page 77, we have

$$\boldsymbol{u}_j = \boldsymbol{u}_0 + \sum_{\ell=0}^{j-1} \alpha_\ell \boldsymbol{p}_\ell, \quad \forall j.$$

According to (B.9), we deduce that

$$\boldsymbol{r}_i^{\mathrm{t}} \boldsymbol{u}_j = \boldsymbol{r}_i^{\mathrm{t}} \boldsymbol{u}_0, \quad \forall i, \forall j \leq i.$$

Hence we obtain (B.10) according to $\boldsymbol{u}_0 = \mathbf{0}$. On the other hand, (V.6) and (V.7) imply $\boldsymbol{r}_{i+1} = \boldsymbol{r}_i - \mathbf{A}(\boldsymbol{u}_{i+1} - \boldsymbol{u}_i)$. Given $\boldsymbol{u}_0 = \mathbf{0}$, we deduce

$$\boldsymbol{r}_i = \boldsymbol{r}_0 - \mathbf{A}(\boldsymbol{u}_i - \boldsymbol{u}_0) = \boldsymbol{b} - \mathbf{A}\boldsymbol{u}_i, \quad \forall i \tag{B.12}$$

by immediate recursion. The latter identity yields

$$\boldsymbol{r}_i^{\mathrm{t}} \boldsymbol{u}_i = \boldsymbol{b}^{\mathrm{t}} \boldsymbol{u}_i - \boldsymbol{u}_i^{\mathrm{t}} \mathbf{A} \boldsymbol{u}_i, \quad \forall i$$

and according to (B.10) we obtain (B.11). □

Lemma B.2.3. *Let $\boldsymbol{u}_0 = \mathbf{0}$. The GCP algorithm (V.5)-(V.9) page 77 ensures that*

$$\|\boldsymbol{u}_{i+1}\|_{\mathbf{M}}^2 \geq \|\boldsymbol{u}_i\|_{\mathbf{M}}^2, \quad \forall i. \tag{B.13}$$

Proof. According to (V.6) page 77 we have

$$\|\boldsymbol{u}_{i+1}\|_{\mathbf{M}}^2 = \|\boldsymbol{u}_i\|_{\mathbf{M}}^2 + \alpha_i^2 \|\boldsymbol{p}_i\|_{\mathbf{M}}^2 + 2\alpha_i \boldsymbol{u}_i^{\mathrm{t}} \mathbf{M} \boldsymbol{p}_i \geq \|\boldsymbol{u}_i\|_{\mathbf{M}}^2 + 2\alpha_i \boldsymbol{u}_i^{\mathrm{t}} \mathbf{M} \boldsymbol{p}_i.$$

Since $\alpha_i \geq 0$, we deduce that (B.13) holds if

$$\boldsymbol{u}_i^{\mathrm{t}} \mathbf{M} \boldsymbol{p}_i \geq 0 \tag{B.14}$$

is true. Let us show the latter inequality by recursion on i. Since $\boldsymbol{u}_0 = 0$, we have $\boldsymbol{u}_0^{\mathrm{t}}\mathbf{M}\boldsymbol{p}_0 = 0$.

Let us assume now that (B.14) holds, and let us show that $\boldsymbol{u}_{i+1}^{\mathrm{t}}\mathbf{M}\boldsymbol{p}_{i+1} \geq 0$. According to (V.9) page 77, we have

$$\boldsymbol{u}_{i+1}^{\mathrm{t}}\mathbf{M}\boldsymbol{p}_{i+1} = \boldsymbol{u}_{i+1}^{\mathrm{t}}\boldsymbol{r}_{i+1} + \beta_i \boldsymbol{u}_{i+1}^{\mathrm{t}}\mathbf{M}\boldsymbol{p}_i.$$

According to (B.10) we deduce

$$\boldsymbol{u}_{i+1}^{\mathrm{t}}\mathbf{M}\boldsymbol{p}_{i+1} = \beta_i \boldsymbol{u}_{i+1}^{\mathrm{t}}\mathbf{M}\boldsymbol{p}_i.$$

Given (V.6) page 77 we get

$$\boldsymbol{u}_{i+1}^{\mathrm{t}}\mathbf{M}\boldsymbol{p}_{i+1} = \beta_i(\boldsymbol{u}_i^{\mathrm{t}}\mathbf{M}\boldsymbol{p}_i + \alpha_i \left\|\boldsymbol{p}_i\right\|_{\mathbf{M}}^2),$$

which is nonnegative since $\beta_i \geq 0$, $\alpha_i \geq 0$ and according to the recursion assumption. □

Consider a SPD matrix $\mathbf{Q} \in \mathbb{R}^{N\times N}$. Let $\nu_1(\mathbf{Q}) > 0$ and $\nu_2(\mathbf{Q}) > 0$ denote the smallest and largest eigenvalues of $\mathbf{Q}$, respectively, so that we have

$$\nu_1(\mathbf{Q}) \left\|\boldsymbol{v}\right\|^2 \leq \left\|\boldsymbol{v}\right\|_{\mathbf{Q}}^2 \leq \nu_2(\mathbf{Q}) \left\|\boldsymbol{v}\right\|^2, \quad \forall \boldsymbol{v} \in \mathbb{R}^N. \tag{B.15}$$

Lemma B.2.4. *Let $\boldsymbol{u}_0 = \mathbf{0}$. The GCP algorithm (V.5)-(V.9) page 77 ensures that*

$$\boldsymbol{b}^{\mathrm{t}}\boldsymbol{u}_i \geq \frac{\tau^2}{\nu_1(\mathbf{A})} \left\|\boldsymbol{b}\right\|^2, \quad \forall i, \tag{B.16}$$

where

$$\tau = \frac{\nu_1(\mathbf{M})\nu_1(\mathbf{A})}{\nu_2(\mathbf{M})\nu_2(\mathbf{A})} \in (0,1). \tag{B.17}$$

Proof. Given (B.15) we have

$$\mu_1 \left\|\boldsymbol{v}\right\|_{\mathbf{A}}^2 \leq \left\|\boldsymbol{v}\right\|_{\mathbf{M}}^2 \leq \mu_2 \left\|\boldsymbol{v}\right\|_{\mathbf{A}}^2, \quad \forall \boldsymbol{v} \in \mathbb{R}^N, \tag{B.18}$$

where $\mu_1 = \nu_1(\mathbf{M})/\nu_2(\mathbf{A}) > 0$ and $\mu_2 = \nu_2(\mathbf{M})/\nu_1(\mathbf{A}) > 0$.

From (B.13), by immediate recursion we get $\left\|\boldsymbol{u}_1\right\|_{\mathbf{M}}^2 \leq \left\|\boldsymbol{u}_i\right\|_{\mathbf{M}}^2$. According to (B.18) we deduce

$$\mu_1 \left\|\boldsymbol{u}_1\right\|_{\mathbf{A}}^2 \leq \left\|\boldsymbol{u}_1\right\|_{\mathbf{M}}^2 \leq \left\|\boldsymbol{u}_i\right\|_{\mathbf{M}}^2 \leq \mu_2 \left\|\boldsymbol{u}_i\right\|_{\mathbf{A}}^2 \tag{B.19}$$

On the other hand, (B.12) yields

$$\left\|\boldsymbol{u}_i\right\|_{\mathbf{A}}^2 = \boldsymbol{u}_i^{\mathrm{t}}\mathbf{A}\boldsymbol{u}_i = \boldsymbol{b}^{\mathrm{t}}\boldsymbol{u}_i - \boldsymbol{r}_i^{\mathrm{t}}\boldsymbol{u}_i.$$

According to (B.10) we deduce

$$\left\|\boldsymbol{u}_i\right\|_{\mathbf{A}}^2 = \boldsymbol{b}^{\mathrm{t}}\boldsymbol{u}_i, \tag{B.20}$$

$$\left\|\boldsymbol{u}_1\right\|_{\mathbf{A}}^2 = \boldsymbol{b}^{\mathrm{t}}\boldsymbol{u}_1. \tag{B.21}$$

Given $\boldsymbol{u}_1 = \boldsymbol{u}_0 + \alpha_0\boldsymbol{p}_0 = \alpha_0\boldsymbol{p}_0 = \alpha_0\mathbf{M}^{-1}\boldsymbol{b}$ and

$$\alpha_0 = \frac{\left\|\boldsymbol{r}_0\right\|_{\mathbf{M}^{-1}}^2}{\left\|\boldsymbol{p}_0\right\|_{\mathbf{A}}^2} = \frac{\left\|\boldsymbol{b}\right\|_{\mathbf{M}^{-1}}^2}{\left\|\mathbf{M}^{-1}\boldsymbol{b}\right\|_{\mathbf{A}}^2},$$

we have

$$\boldsymbol{b}^{\mathrm{t}}\boldsymbol{u}_1 = \frac{\|\boldsymbol{b}\|^4_{\mathbf{M}^{-1}}}{\|\mathbf{M}^{-1}\boldsymbol{b}\|^2_{\mathbf{A}}} = \frac{\left\|\mathbf{M}^{-1}\boldsymbol{b}\right\|^4_{\mathbf{M}}}{\|\mathbf{M}^{-1}\boldsymbol{b}\|^2_{\mathbf{A}}}.$$

According to (B.18) we deduce

$$\boldsymbol{b}^{\mathrm{t}}\boldsymbol{u}_1 \geq \mu_1 \left\|\mathbf{M}^{-1}\boldsymbol{b}\right\|^2_{\mathbf{M}} = \mu_1 \|\boldsymbol{b}\|^2_{\mathbf{M}^{-1}}.$$

According to (B.18) and (B.15) we obtain

$$\boldsymbol{b}^{\mathrm{t}}\boldsymbol{u}_1 \geq \mu_1\nu_1(\mathbf{M}^{-1}) \|\boldsymbol{b}\|^2 = \frac{\mu_1}{\nu_2(\mathbf{M})} \|\boldsymbol{b}\|^2 = \frac{\mu_1}{\mu_2\nu_1(\mathbf{A})} \|\boldsymbol{b}\|^2. \tag{B.22}$$

Finally, according to (B.19), (B.20), (B.21) and (B.22) we deduce

$$\boldsymbol{b}^{\mathrm{t}}\boldsymbol{u}_i \geq \tau\boldsymbol{b}^{\mathrm{t}}\boldsymbol{u}_1 \geq \frac{\tau^2}{\nu_1(\mathbf{A})} \|\boldsymbol{b}\|^2,$$

since $\tau = \mu_1/\mu_2$. □

Lemma B.2.5. *Let $\boldsymbol{u}_0 = \mathbf{0}$. The GCP algorithm (V.5)-(V.9) page 77 ensures that*

$$\|\boldsymbol{u}_i\| \leq \frac{1}{\tau^{1/2}\nu_1(\mathbf{A})}\|\boldsymbol{b}\|, \quad \forall i, \tag{B.23}$$

where τ is defined by (B.17).

Proof. From (B.13), by immediate recursion we get

$$\|\boldsymbol{u}_i\|^2_{\mathbf{M}} \leq \|\boldsymbol{u}_N\|^2_{\mathbf{M}}.$$

According to (B.18) we obtain

$$\|\boldsymbol{u}_i\|^2_{\mathbf{A}} \leq \frac{1}{\tau} \|\boldsymbol{u}_N\|^2_{\mathbf{A}}. \tag{B.24}$$

According to (B.15) we have

$$\|\boldsymbol{u}_i\|^2 \leq \frac{1}{\nu_1(\mathbf{A})} \|\boldsymbol{u}_i\|^2_{\mathbf{A}}. \tag{B.25}$$

The GCP algorithm (V.5)-(V.9) ensures that $\boldsymbol{u}_N = \mathbf{A}^{-1}\boldsymbol{b}$ after N iterations. Hence, we have

$$\|\boldsymbol{u}_N\|^2_{\mathbf{A}} = \left\|\mathbf{A}^{-1}\boldsymbol{b}\right\|^2_{\mathbf{A}} = \|\boldsymbol{b}\|^2_{\mathbf{A}^{-1}} \leq \nu_2(\mathbf{A}^{-1})\|\boldsymbol{b}\|^2 = \frac{1}{\nu_1(\mathbf{A})}\|\boldsymbol{b}\|^2, \tag{B.26}$$

according to (B.15). Finally, (B.23) is easily deduced from (B.25), (B.24) and (B.26). □

B.2.2 Convergence pour un critère général

For sake of generality, the convergence of algorithm (V.2)-(V.3) page 76 is established considering a general criterion $\mathcal{J}$. Let $\mathcal{N}$ be a neighborhood of the level set $\mathcal{L} = \{\boldsymbol{x} \in \mathbb{R}^N | \mathcal{J}(\boldsymbol{x}) \leq \mathcal{J}(\boldsymbol{x}_0)\}$. The following assumptions are adopted.

Assumption 1. *Let us assume that $\mathcal{J} : \mathbb{R}^N \mapsto \mathbb{R}$ is differentiable and bounded below on $\mathcal{N}$.*

Definition B.2.1. *A sequence of SPD matrices $\mathcal{Q} = \{\mathbf{Q}_k\} \in \mathbb{R}^{N\times N}$ has a uniformly bounded spectrum with a strictly positive lower bound if there exist $\nu_1(\mathcal{Q})$, $\nu_2(\mathcal{Q}) \in \mathbb{R}$ such that*

$$\nu_2(\mathcal{Q}) \geq \nu_2(\mathbf{Q}_k) \geq \nu_1(\mathbf{Q}_k) \geq \nu_1(\mathcal{Q}) > 0, \quad \forall k.$$

For the sake of brevity, $\mathcal{Q}$ will be said uniformly bounded.

Assumption 2. *Let us introduce local quadratic models of $\mathcal{J}$ in the neighborhood of $\boldsymbol{x}$ under the following form :*

$$\widehat{\mathcal{J}}_k(\boldsymbol{x}^+, \boldsymbol{x}) = \mathcal{J}(\boldsymbol{x}) + (\boldsymbol{x}^+ - \boldsymbol{x})^{\mathrm{t}} \nabla\mathcal{J}(\boldsymbol{x}) + (\boldsymbol{x}^+ - \boldsymbol{x})^{\mathrm{t}} \mathbf{A}_k (\boldsymbol{x}^+ - \boldsymbol{x})/2. \tag{B.27}$$

It is assumed here that there exists a uniformly bounded matrix sequence $\mathcal{A} = \{\mathbf{A}_k\}$ such that $\mathcal{J}$ is upper bounded by $\widehat{\mathcal{J}}_k(\cdot, \boldsymbol{x})$, i.e.,

$$\widehat{\mathcal{J}}_k(\boldsymbol{x}^+, \boldsymbol{x}) \geq \mathcal{J}(\boldsymbol{x}^+), \quad \forall \boldsymbol{x}, \boldsymbol{x}^+ \in \mathcal{N}, \forall k. \tag{B.28}$$

Definition B.2.2. *[Bertsekas, 1999, p. 35] The direction sequence $\{\boldsymbol{d}_k\}$ is gradient related to $\{\boldsymbol{x}_k\}$ if for any subsequence $\{\boldsymbol{x}_k\}_{k\in\mathcal{K}}$ that converges to a nonstationary point, the corresponding subsequence $\{\boldsymbol{d}_k\}_{k\in\mathcal{K}}$ is bounded and satisfies*

$$\limsup_{k\to\infty, k\in\mathcal{K}} \boldsymbol{d}_k^{\mathrm{t}} \nabla\mathcal{J}(\boldsymbol{x}_k) < 0.$$

Lemma B.2.6. *Let $\boldsymbol{x}_k$ be defined by (V.2)-(V.3) page 76 with $\theta \in (0, 2)$, let Assumptions 1 and 2 hold, and let the preconditioning sequence $\mathcal{M}$ be uniformly bounded. Then we have*

$$\frac{\mathcal{T}^2}{\nu_1(\mathcal{A})} \|\nabla\mathcal{J}(\boldsymbol{x}_k)\|^2 \leq -\boldsymbol{d}_k^{\mathrm{t}} \nabla\mathcal{J}(\boldsymbol{x}_k) \tag{B.29}$$

and

$$\|\boldsymbol{d}_k\| \leq \frac{1}{\mathcal{T}^{1/2}\nu_1(\mathcal{A})} \|\nabla\mathcal{J}(\boldsymbol{x}_k)\| \tag{B.30}$$

where

$$\mathcal{T} = \frac{\nu_1(\mathcal{M})\nu_1(\mathcal{A})}{\nu_2(\mathcal{M})\nu_2(\mathcal{A})} \in (0, 1). \tag{B.31}$$

Thus, the direction sequence $\{\boldsymbol{d}_k\}$ is gradient related to $\{\boldsymbol{x}_k\}$.

Proof. For any value of k, let $\boldsymbol{b} := -\nabla\mathcal{J}(\boldsymbol{x}_k)$, $\mathbf{A} := \mathbf{A}_k$, $\mathbf{M} := \mathbf{M}_k$ and $\boldsymbol{d}_k := \boldsymbol{u}_{I_k}(\boldsymbol{x}_k)$. Since both sequences $\mathcal{A}$ and $\mathcal{M}$ are uniformly bounded, it is easy to see that (B.17) and (B.31) imply $\mathcal{T} \leq \tau$, as well as

$$\frac{\mathcal{T}^2}{\nu_1(\mathcal{A})} \leq \frac{\tau^2}{\nu_1(\mathbf{A})}$$

and

$$\frac{1}{\tau^{1/2}\nu_1(\mathbf{A})} \leq \frac{1}{\mathcal{T}^{1/2}\nu_1(\mathcal{A})}.$$

Then (B.29) and (B.30) are obvious consequences of (B.16) and (B.23), respectively.

According to [Bertsekas, 1999, p. 36], if for some scalars $c_1 > 0$, $c_2 > 0$, $p_1 \geq 0$, $p_2 \geq 0$ and all k

$$c_1\|\nabla\mathcal{J}(\boldsymbol{x}_k)\|^{p_1} \leq -\boldsymbol{d}_k^{\mathrm{t}} \nabla\mathcal{J}(\boldsymbol{x}_k), \quad \|\boldsymbol{d}_k\| \leq c_2\|\nabla\mathcal{J}(\boldsymbol{x}_k)\|^{p_2}$$

then $\{\boldsymbol{d}_k\}$ is gradient related to $\{\boldsymbol{x}_k\}$.

Thus, with $c_1 = \mathcal{T}^2/\nu_1(\mathcal{A})$, $c_2 = 1/\mathcal{T}^{1/2}\nu_1(\mathcal{A})$, $p_1 = 2$ and $p_2 = 1$ inequalities (B.29) and (B.30) are sufficient conditions to ensure that $\{\boldsymbol{d}_k\}$ is gradient related to $\{\boldsymbol{x}_k\}$. □

Definition B.2.3. *The stepsize sequence* $\{\alpha_k\}$ *satisfies the Armijo condition with* $\Omega \in (0,1)$ *if*

$$\mathcal{J}(\boldsymbol{x}_k) - \mathcal{J}(\boldsymbol{x}_{k+1}) + \Omega\alpha_k \boldsymbol{d}_k^{\mathrm{t}} \nabla\mathcal{J}(\boldsymbol{x}_k) \geq 0, \quad \forall k. \tag{B.32}$$

Lemma B.2.7. *Let* $\boldsymbol{x}_k$ *be defined by* (V.2)-(V.3) *page 76 with* $\theta \in (0,2)$*, and let Assumption 1 and Assumption 2 hold, and the preconditioning sequence* $\mathcal{M}$ *be uniformly bounded. Then the constant stepsize sequence* $\{\theta\}$ *satisfies the Armijo condition with* $\Omega = 1-\theta/2 \in (0,1)$*. Moreover, the sequence* $\{\mathcal{J}(\boldsymbol{x}_k)\}$ *is nonincreasing :*

$$\mathcal{J}(\boldsymbol{x}_k) \geq \mathcal{J}(\boldsymbol{x}_{k+1}), \quad \forall k. \tag{B.33}$$

Thus, the sequence $\{\boldsymbol{x}_k\}$ *belongs to* $\mathcal{N}$*.*

Proof. According to (B.28) we have

$$\widehat{\mathcal{J}}_k(\boldsymbol{x}_{k+1}, \boldsymbol{x}_k) \geq \mathcal{J}(\boldsymbol{x}_{k+1}).$$

Then, given (V.2) page 76 and (B.27) we get

$$\mathcal{J}(\boldsymbol{x}_k) - \mathcal{J}(\boldsymbol{x}_{k+1}) + \theta \boldsymbol{d}_k^{\mathrm{t}} \nabla\mathcal{J}(\boldsymbol{x}_k) + \theta^2 \boldsymbol{d}_k^{\mathrm{t}} \mathbf{A}_k \boldsymbol{d}_k/2 \geq 0. \tag{B.34}$$

Let $\boldsymbol{b} := -\nabla\mathcal{J}(\boldsymbol{x}_k)$, $\mathbf{A} := \mathbf{A}_k$ and $\boldsymbol{d}_k := \boldsymbol{u}_{I_k}(\boldsymbol{x}_k)$. According to (B.11) we have

$$-\boldsymbol{d}_k^{\mathrm{t}} \nabla\mathcal{J}(\boldsymbol{x}_k) = \boldsymbol{d}_k^{\mathrm{t}} \mathbf{A}_k \boldsymbol{d}_k. \tag{B.35}$$

According to (B.34), (B.35) and the positiveness of $\mathbf{A}_k$ we have

$$\mathcal{J}(\boldsymbol{x}_k) - \mathcal{J}(\boldsymbol{x}_{k+1}) \geq -\theta\Omega \boldsymbol{d}_k^{\mathrm{t}} \nabla\mathcal{J}(\boldsymbol{x}_k) \geq 0, \tag{B.36}$$

where $\Omega = 1 - \theta/2 \in (0,1)$. □

Theorem B.2.1. *Let* $\boldsymbol{x}_k$ *be defined by* (V.2)-(V.3) *page 76 with* $\theta \in (0,2)$*, and let Assumption 1 and Assumption 2 hold, and the preconditioning sequence* $\mathcal{M}$ *be uniformly bounded. Then we have convergence in the sense*

$$\lim_{k\to\infty} \nabla\mathcal{J}(\boldsymbol{x}_k) = \mathbf{0}.$$

Proof. Let $\Omega_1 = \theta\Omega\mathcal{T}^2/\nu_1(\mathcal{A}) > 0$. Inequalities (B.29) and (B.36) yield

$$\mathcal{J}(\boldsymbol{x}_k) - \mathcal{J}(\boldsymbol{x}_{k+1}) \geq \Omega_1 \|\nabla\mathcal{J}(\boldsymbol{x}_k)\|^2 \geq 0. \tag{B.37}$$

On the other hand, given Assumption 1 and (B.33), we have

$$-\infty < \inf_{\boldsymbol{x}\in\mathcal{N}} \mathcal{J}(\boldsymbol{x}) \leq \mathcal{J}(\boldsymbol{x}_\ell), \quad \forall \ell.$$

Then (B.37) allows to deduce

$$\infty > \mathcal{J}(\boldsymbol{x}_0) - \inf_{\boldsymbol{x}\in\mathcal{N}} \mathcal{J}(\boldsymbol{x}) \geq \mathcal{J}(\boldsymbol{x}_0) - \mathcal{J}(\boldsymbol{x}_\ell) \geq \Omega_1 \sum_{k=0}^{\ell-1} \|\nabla\mathcal{J}(\boldsymbol{x}_k)\|^2, \quad \forall \ell. \tag{B.38}$$

Hence, $\lim_{k\to\infty} \nabla\mathcal{J}(\boldsymbol{x}_k) = \mathbf{0}$. □

B.3 Preuves du chapitre VI

B.3.1 Preuve du Lemme VI.1.1

Preuve. La fonction scalaire $f(\alpha) = \mathcal{J}(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k)$ présente la même structure pénalisée généralisée que le critère $\mathcal{J}$ défini par (IV.1)-(IV.2), page 54 :

$$f(\alpha) = a_2\alpha^2 - 2a_1\alpha + a_0 + \lambda \sum_{c=1}^{C} \phi(\|\alpha \boldsymbol{v}_c - \boldsymbol{\tau}_c\|),$$

où $\boldsymbol{v}_c = \mathbf{V}_c\boldsymbol{d}_k$, $\boldsymbol{\tau}_c = \boldsymbol{\omega}_c - \boldsymbol{V}_c\boldsymbol{x}_k$ et $a_0 = \boldsymbol{u}_2^{\mathrm{t}}\boldsymbol{u}_2$, $a_1 = \boldsymbol{u}_1^{\mathrm{t}}\boldsymbol{u}_2$, $a_2 = \boldsymbol{u}_1^{\mathrm{t}}\boldsymbol{u}_1$ avec $\boldsymbol{u}_1 = \mathbf{H}\boldsymbol{d}_k$ et $\boldsymbol{u}_2 = \boldsymbol{y} - \mathbf{H}\boldsymbol{x}_k$.

D'après (IV.22)-(IV.23), page 66 les algorithmes SQ1D appliqués à la minimisation de la fonction scalaire f peuvent s'écrire sous la forme

$$\alpha_k^{i+1} = \alpha_k^i - \theta \frac{f'(\alpha_k^i)}{a(\alpha_k^i)} \tag{B.39}$$

avec les scalaires $a(\alpha) := a_{\mathrm{GY}}$ pour GY1D et $a(\alpha) := a_{\mathrm{GR}}(\alpha)$ pour GR1D.

Soit $\boldsymbol{v}$ le vecteur de taille CP défini par $\boldsymbol{v}^{\mathrm{t}} = \left[\boldsymbol{v}_1^{\mathrm{t}}|\ldots|\boldsymbol{v}_C^{\mathrm{t}}\right]$. D'après la définition de la matrice normale de GY (IV.26), page 66, la forme scalaire de l'opérateur de GY s'écrit sous la forme

$$a_{\mathrm{GY}} = 2a_2 + \boldsymbol{v}^{\mathrm{t}}\boldsymbol{v}/a$$

d'où d'après $a_2 = \boldsymbol{d}_k^{\mathrm{t}}\mathbf{H}^{\mathrm{t}}\mathbf{H}\boldsymbol{d}_k$, $\boldsymbol{v}_c = \mathbf{V}_c\boldsymbol{d}_k$ et la définition de la matrice $\mathbf{V}$ (IV.3), page 54 on obtient

$$\begin{aligned} a_{\mathrm{GY}} &= \boldsymbol{d}_k^{\mathrm{t}}\left(2\mathbf{H}^{\mathrm{t}}\mathbf{H} + \mathbf{V}^{\mathrm{t}}\mathbf{V}/a\right)\boldsymbol{d}_k \\ &= \boldsymbol{d}_k^{\mathrm{t}}\mathbf{A}_{\mathrm{GY}}^a\boldsymbol{d}_k. \end{aligned} \tag{B.40}$$

D'après la définition de la matrice normale de GR (IV.37), page 69, la forme scalaire de l'opérateur de GR s'écrit sous la forme

$$\begin{aligned} a_{\mathrm{GR}}(\alpha) &= 2a_2 + \boldsymbol{v}^{\mathrm{t}}\mathbf{L}(\alpha)\boldsymbol{v} \\ \mathbf{L}(\alpha) &= \mathrm{Diag}\{\dot{\phi}(\|\boldsymbol{\delta}_c\|)/\|\boldsymbol{\delta}_c\|\} \\ \boldsymbol{\delta}_c &= \boldsymbol{v}_c\alpha - \boldsymbol{\tau}_c \end{aligned}$$

d'où d'après $a_2 = \boldsymbol{d}_k^{\mathrm{t}}\mathbf{H}^{\mathrm{t}}\mathbf{H}\boldsymbol{d}_k$ et $\boldsymbol{v}_c = \mathbf{V}_c\boldsymbol{d}_k$ et la définition de la matrice $\mathbf{V}$ (IV.3), page 54 on obtient

$$a_{\mathrm{GR}}(\alpha) = \boldsymbol{d}_k^{\mathrm{t}}\left(2\mathbf{H}^{\mathrm{t}}\mathbf{H} + \mathbf{V}^{\mathrm{t}}\mathbf{L}(\alpha)\mathbf{V}\right)\boldsymbol{d}_k.$$

D'après $\boldsymbol{v}_c = \mathbf{V}_c\boldsymbol{d}_k$ et $\boldsymbol{\tau}_c = \boldsymbol{\omega}_c - \boldsymbol{V}_c\boldsymbol{x}_k$ on a $\boldsymbol{\delta}_c = \boldsymbol{v}_c\alpha - \boldsymbol{\tau}_c = \mathbf{V}_c(\boldsymbol{x}_k + \alpha\boldsymbol{d}_k) - \boldsymbol{\omega}_c$ d'où

$$a_{\mathrm{GR}}(\alpha) = \boldsymbol{d}_k^{\mathrm{t}}\mathbf{A}_{\mathrm{GR}}(\boldsymbol{x}_k + \alpha\boldsymbol{d}_k)\boldsymbol{d}_k \tag{B.41}$$

où $\mathbf{A}_{\mathrm{GR}}(\boldsymbol{x})$ est définie par (IV.37), page 69.

Finalement, d'après (B.40), (B.41) et $f'(\alpha) = \boldsymbol{d}_k^{\mathrm{t}}\nabla\mathcal{J}\left(\boldsymbol{x}_k + \alpha\boldsymbol{d}_k\right)$, (B.39) se met sous la forme

$$\alpha_k^{i+1} = \alpha_k^i - \theta \frac{\boldsymbol{d}_k^{\mathrm{t}}\nabla\mathcal{J}\left(\boldsymbol{x}_k + \alpha_k^i\boldsymbol{d}_k\right)}{\boldsymbol{d}_k{}^{\mathrm{t}}\mathbf{Q}_k^i\boldsymbol{d}_k}$$

avec $\mathbf{Q}_k^i = \mathbf{A}_{\mathrm{GY}}^a$ pour GY1D et $\mathbf{Q}_k^i = \mathbf{A}_{\mathrm{GR}}(\boldsymbol{x}_k + \alpha_k^i\boldsymbol{d}_k)$ pour GR1D. □

B.3.2 Preuve du Lemme VI.3.1

Preuve. Le gradient du terme de pénalisation Φ défini par (IV.2), page 54 peut s'écrire sous la forme

$$\nabla\Phi(\boldsymbol{x}) = \sum_{c=1}^{C} \mathbf{V}_c^{\mathrm{t}} \nabla\phi(\|\mathbf{V}_c\boldsymbol{x} - \boldsymbol{\omega}_c\|)$$

d'où

$$\|\nabla\Phi(\boldsymbol{y}) - \nabla\Phi(\boldsymbol{x})\| = \left\| \sum_{c=1}^{C} \mathbf{V}_c^{\mathrm{t}} \left(\nabla\phi(\|\mathbf{V}_c\boldsymbol{y} - \boldsymbol{\omega}_c\|) - \nabla\phi(\|\mathbf{V}_c\boldsymbol{x} - \boldsymbol{\omega}_c\|)\right) \right\|.$$

D'après l'inégalité triangulaire on a

$$\|\nabla\Phi(\boldsymbol{y}) - \nabla\Phi(\boldsymbol{x})\| \leq \sum_{c=1}^{C} \|\mathbf{V}_c^{\mathrm{t}} \left(\nabla\phi(\|\mathbf{V}_c\boldsymbol{y} - \boldsymbol{\omega}_c\|) - \nabla\phi(\|\mathbf{V}_c\boldsymbol{x} - \boldsymbol{\omega}_c\|)\right) \|.$$

D'après la propriété de sous-multiplicativité de la 2-norme matricielle [Golub et Van Loan, 1996, p. 55] on a

$$\|\nabla\Phi(\boldsymbol{y}) - \nabla\Phi(\boldsymbol{x})\| \leq \sum_{c=1}^{C} \|\mathbf{V}_c^{\mathrm{t}}\| \|\nabla\phi(\|\mathbf{V}_c\boldsymbol{y} - \boldsymbol{\omega}_c\|) - \nabla\phi(\|\mathbf{V}_c\boldsymbol{x} - \boldsymbol{\omega}_c\|)\|. \tag{B.42}$$

D'après le Lemme IV.4.5, page 70 on a

$$\|\nabla\phi(\|\boldsymbol{u}\|) - \nabla\phi(\|\boldsymbol{v}\|)\| \leq L\|\boldsymbol{u} - \boldsymbol{v}\|, \quad \forall \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^N$$

d'où d'après (B.42) on obtient

$$\|\nabla\Phi(\boldsymbol{y}) - \nabla\Phi(\boldsymbol{x})\| \leq L \sum_{c=1}^{C} \|\mathbf{V}_c^{\mathrm{t}}\| \|\mathbf{V}_c(\boldsymbol{y} - \boldsymbol{x})\|.$$

De nouveau d'après la propriété de sous-multiplicativité de la 2-norme matricielle on obtient

$$\|\nabla\Phi(\boldsymbol{y}) - \nabla\Phi(\boldsymbol{x})\| \leq \mu_\Phi \|\boldsymbol{y} - \boldsymbol{x}\| \tag{B.43}$$

avec

$$\mu_\Phi = L \sum_{c=1}^{C} \|\mathbf{V}_c^{\mathrm{t}}\| \|\mathbf{V}_c\| = L \sum_{c=1}^{C} \|\mathbf{V}_c\|^2.$$

Soit $\mathcal{Q}(\boldsymbol{x}) = \|\boldsymbol{y} - \mathbf{H}\boldsymbol{x}\|^2$. On a $\nabla\mathcal{Q}(\boldsymbol{x}) = 2\mathbf{H}^{\mathrm{t}}(\mathbf{H}\boldsymbol{x} - \boldsymbol{y})$ d'où d'après la propriété de sous-multiplicativité de la 2-norme matricielle on obtient

$$\|\nabla\mathcal{Q}(\boldsymbol{y}) - \nabla\mathcal{Q}(\boldsymbol{x})\| = \|2\mathbf{H}^{\mathrm{t}}\mathbf{H}(\boldsymbol{y} - \boldsymbol{x})\| \leq 2\|\mathbf{H}^{\mathrm{t}}\mathbf{H}\| \|\boldsymbol{y} - \boldsymbol{x}\| \tag{B.44}$$

Finalement, d'après $\nabla\mathcal{J}(\boldsymbol{x}) = \nabla\mathcal{Q}(\boldsymbol{x}) + \lambda\nabla\Phi(\boldsymbol{x})$, (B.43) et (B.44) on obtient

$$\|\nabla\mathcal{J}(\boldsymbol{y}) - \nabla\mathcal{J}(\boldsymbol{x})\| \leq \|\nabla\mathcal{Q}(\boldsymbol{y}) - \nabla\mathcal{Q}(\boldsymbol{x})\| + \lambda\|\nabla\Phi(\boldsymbol{y}) - \nabla\Phi(\boldsymbol{x})\| \leq \mu\|\boldsymbol{y} - \boldsymbol{x}\|$$

avec

$$\mu = 2\|\mathbf{H}^{\mathrm{t}}\mathbf{H}\| + \lambda\mu_\Phi.$$

□

B.3.3 Preuve du Lemme VI.3.2

Preuve. On a

$$\begin{aligned}\|\mathbf{V}\boldsymbol{x}\|^2 + \|\mathbf{H}\boldsymbol{x}\|^2 &= \boldsymbol{x}^{\mathrm{t}}(\mathbf{V}^{\mathrm{t}}\mathbf{V} + \mathbf{H}^{\mathrm{t}}\mathbf{H})\boldsymbol{x} \\ &= \|\boldsymbol{x}\|^2_{\mathbf{V}^{\mathrm{t}}\mathbf{V}+\mathbf{H}^{\mathrm{t}}\mathbf{H}}\,. \end{aligned} \tag{B.45}$$

Soit $\{\boldsymbol{x}_k\}$ tel que $\lim_{k\to\infty}\|\boldsymbol{x}_k\| = \infty$. On a alors $\lim_{k\to\infty}\|\boldsymbol{x}_k\|_{\mathbf{V}^{\mathrm{t}}\mathbf{V}+\mathbf{H}^{\mathrm{t}}\mathbf{H}} = \infty$.

D'après (B.45) on a

$$\lim_{k\to\infty}\|\mathbf{V}\boldsymbol{x}_k\|^2 + \|\mathbf{H}\boldsymbol{x}_k\|^2 = \infty$$

ce qui assure la coercivité du critère pénalisé généralisé $\mathcal{J}$. □

B.3.4 Preuve du Théorème VI.3.1

Preuve. D'après le Lemme C.3.1, page 159 on a (VI.12), page 88.

D'après le Lemme C.4.1, page 164 qui reste valide sans conjugaison ($\beta_k = 0$) on a

$$\sum_{k,\,\boldsymbol{d}_k\neq 0} (\nabla\mathcal{J}(\boldsymbol{x}_k)^{\mathrm{t}}\boldsymbol{d}_k)^2/\|\boldsymbol{d}_k\|^2 < \infty. \tag{B.46}$$

D'après (VI.10), page 88 on a $\nabla\mathcal{J}(\boldsymbol{x}_k) = -\mathbf{M}_k\boldsymbol{d}_k$, d'où d'après (B.46) on obtient

$$\sum_{k,\,\boldsymbol{d}_k\neq 0} (\boldsymbol{d}_k^{\mathrm{t}}\mathbf{M}_k\boldsymbol{d}_k)^2/\|\boldsymbol{d}_k\|^2 < \infty. \tag{B.47}$$

Or $\{\mathbf{M}_k\}$ est uniformément bornée d'où

$$(\boldsymbol{d}_k^{\mathrm{t}}\mathbf{M}_k\boldsymbol{d}_k) \geq \nu_1(\mathcal{M})\|\boldsymbol{d}_k\|^2 \tag{B.48}$$

avec $\nu_1(\mathcal{M}) > 0$. D'après (B.47) et (B.48) on a

$$\sum_{k,\,\boldsymbol{d}_k\neq 0} \nu_1(\mathcal{M})^2\|\boldsymbol{d}_k\|^2 < \infty$$

d'où $\lim_{k\to\infty}\|\boldsymbol{d}_k\| = 0$ et donc d'après $\nabla\mathcal{J}(\boldsymbol{x}_k) = -\mathbf{M}_k\boldsymbol{d}_k$ on obtient la convergence définie par (VI.13), page 88. □

B.3.5 Preuve du Théorème VI.3.2

Preuve. D'après le Lemme C.3.1, page 159 on a (VI.14), page 89.

D'après le Théorème C.4.1, page 167 on a la convergence définie par (VI.15), page 89 en l'absence de préconditionnement ($\{\mathbf{M}_k\} = \{\mathbf{I}\}$). Il nous reste à établir la convergence pour un préconditionnement constant $\{\mathbf{M}_k\} = \{\mathbf{M}_0\}$. Effectuons le changement de variables $\boldsymbol{x} = \mathbf{S}^{\mathrm{t}}\boldsymbol{y}$, où $\mathbf{S}$ est la matrice inversible telle que $\mathbf{M}_0^{-1} = \mathbf{S}^{\mathrm{t}}\mathbf{S}$. Soit $\mathcal{H}(\boldsymbol{y}) = \mathcal{J}(\mathbf{S}^{\mathrm{t}}\boldsymbol{y})$.

D'après [Bertsekas, 1999, p 138], l'algorithme GCPPR défini par (VI.1)-(VI.4), page 83 avec préconditionnement constant ($\{\mathbf{M}_k\} = \{\mathbf{M}_0\}$) appliqué à la minimisation du critère $\mathcal{J}$ s'identifie avec l'algorithme GCPPR sans préconditionnement ($\{\mathbf{M}_k\} = \{\mathbf{I}\}$) appliqué à la minimisation du critère $\mathcal{H}$.

Le Théorème C.4.1, page 167 appliqué à la minimisation du critère $\mathcal{H}$ sans préconditionnement implique que

$$\liminf_{k\to\infty}\|\nabla\mathcal{H}(\boldsymbol{y}_k)\| = 0. \tag{B.49}$$

On a

$$\nabla\mathcal{H}(\boldsymbol{y}) = \mathbf{S}\nabla\mathcal{J}(\mathbf{S}^{\mathrm{t}}\boldsymbol{y}) = \mathbf{S}\nabla\mathcal{J}(\boldsymbol{x}). \tag{B.50}$$

D'après (B.50) et l'inversibilité de la matrice $\mathbf{S}$ on a que (B.49) est équivalent à (VI.15), page 89. □

B.3.6 Preuve du Lemme VI.3.3

Preuve. Afin de simplifier les notations, le paramètre de régularisation λ est intégré à la fonction de régularisation ϕ.

La positivité de la fonction de régularisation ϕ entraîne que

$$\mathcal{J}(\boldsymbol{x}) \geq \|\mathbf{H}\boldsymbol{x} - \boldsymbol{y}\|^2, \quad \forall \boldsymbol{x} \in \mathbb{R}^N.$$

Il existe donc un $\boldsymbol{x}_0 \notin \operatorname{Ker}\mathbf{H}$ tel que

$$\mathcal{J}(\boldsymbol{x}_0) > KC + \|\boldsymbol{y}\|^2. \tag{B.51}$$

Comme $\operatorname{Ker}\mathbf{H} \neq \{\mathbf{0}\}$, il existe une suite $\{\boldsymbol{x}_k\} \in \operatorname{Ker}\mathbf{H}$ telle que $\lim_{k\to\infty}\|\boldsymbol{x}_k\| = \infty$. On a alors

$$\mathcal{J}(\boldsymbol{x}_k) = \|\boldsymbol{y}\|^2 + \sum_{c=1}^{C} \phi(\|\mathbf{V}_c\boldsymbol{x}_k - \boldsymbol{\omega}_c\|)$$

et d'après $\phi(t) < K, \forall t \in \mathbb{R}$ on obtient

$$\mathcal{J}(\boldsymbol{x}_k) < \|\boldsymbol{y}\|^2 + KC.$$

D'où d'après (B.51) on a

$$\mathcal{J}(\boldsymbol{x}_k) < \mathcal{J}(\boldsymbol{x}_0).$$

Comme $\lim_{k\to\infty}\|\boldsymbol{x}_k\| = \infty$, les lignes de niveau définies par $L = \{\boldsymbol{x} \in \mathbb{R}^N | \mathcal{J}(\boldsymbol{x}) \leq \mathcal{J}(\boldsymbol{x}_0)\}$ ne sont donc pas bornées. □

Bibliographie

[Bertsekas, 1999] D. P. Bertsekas. *Nonlinear programming*. Athena Scientific, Belmont, MA, USA, 2ème edition, 1999.

[Chan et Mulet, 1999] T. F. Chan et P. Mulet. On the convergence of the lagged diffusivity fixed point method in total variation image restoration. *SIAM J. Num. Anal.*, 36 (2) : 354–367, 1999.

[Golub et Van Loan, 1996] G. H. Golub et C. F. Van Loan. *Matrix computations*. The Johns Hopkins University Press, Baltimore, 3ème edition, 1996.

Annexe C

CONVERGENCE OF CONJUGATE GRADIENT METHODS WITH A CLOSED-FORM STEPSIZE FORMULA

Abstract. [1] Conjugate gradient methods are efficient to minimize differentiable objective functions in large dimension spaces. However, converging line search strategies are usually not easy to choose, nor to implement. In [Sun et Zhang, 2001; Chen et Sun, 2002], Sun and colleagues introduced a simple stepsize formula. However, the associated convergence domain happens to be overrestrictive, since it precludes the optimal stepsize in the convex quadratic case. Here, we identify this stepsize formula with one iteration of Weiszfeld's algorithm in the scalar case. More generally, we propose to make use of a finite number of iterates of such an algorithm to compute the stepsize. In this framework, we establish a new convergence domain, that incorporates the optimal stepsize in the convex quadratic case.
Key Words. Conjugate gradient methods, convergence, stepsize formula, Weiszfeld's method.

C.1 Introduction

Let us consider the following unconstrained minimization problem :

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \mathcal{J}(\boldsymbol{x}) \tag{C.1}$$

where J is a differentiable objective function. In the implementation of a conjugate gradient (CG) method, the stepsize strategy often incorporates a stopping criterion such as to satisfy the Wolfe conditions [Wolfe, 1971]. For instance, the Armijo condition is used as stopping criterion in [Dixon, 1973]. Most recently, a simple stepsize formula was proposed by Sun and Zhang [Sun et Zhang, 2001] and by Chen and Sun [Chen et Sun, 2002] for several CG methods. Its distinctive feature it to yield convergence results without any stopping condition. Here, we pursue in the same direction, by proposing a generalized stepsize formula. We also reexamine the convergence conditions, which leads us to a broadened convergence domain for several types of conjugacy.

[1] Rapport technique IRCCyN [Labat et Idier, 2005] accepté pour publication [Labat et Idier, 2007].

In this paper, we restrict ourselves to the following family of CG algorithm :

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k \tag{C.2}$$
$$\boldsymbol{c}_k = -\boldsymbol{g}_k + \beta_k \boldsymbol{d}_{k-1} \tag{C.3}$$
$$\boldsymbol{d}_k = \begin{cases} \boldsymbol{c}_k & \text{if } \boldsymbol{g}_k^{\mathrm{t}} \boldsymbol{c}_k \leq 0 \\ -\boldsymbol{c}_k & \text{otherwise} \end{cases} \tag{C.4}$$
$$\beta_k = \begin{cases} 0, & \text{for } k = 0 \\ \beta_k^{\mu_k,\omega_k}, & \text{for } k \geq 1 \end{cases}$$

where $k \in \mathbb{N}$, $\boldsymbol{g}_k = \nabla \mathcal{J}(\boldsymbol{x}_k)$ and with the following conjugacy formulas :

$$D_k = (1 - \mu_k - \omega_k)\|\boldsymbol{g}_{k-1}\|^2 + \mu_k \boldsymbol{d}_{k-1}^{\mathrm{t}} \boldsymbol{y}_{k-1} - \omega_k \boldsymbol{d}_{k-1}^{\mathrm{t}} \boldsymbol{g}_{k-1} \tag{C.5}$$
$$\beta_k^{\mu_k,\omega_k} = \boldsymbol{g}_k^{\mathrm{t}} \boldsymbol{y}_{k-1} / D_k \tag{C.6}$$

where $\|\,.\,\|$ is the Euclidean norm, "t" stands for the transpose, $\boldsymbol{y}_{k-1} = \boldsymbol{g}_k - \boldsymbol{g}_{k-1}$, and D_k depends on parameters $\mu_k \in [0, 1]$ and $\omega_k \in [0, 1 - \mu_k]$. Let us remark that the descent direction $\boldsymbol{d}_k$ is defined such that $\boldsymbol{g}_k^{\mathrm{t}} \boldsymbol{d}_k \leq 0$.

The parametrized expression (C.6) is taken from [Chen et Sun, 2002]. It only covers a subset of a larger family introduced by Dai and Yan in [Dai et Yuan, 2001]. Three classical versions of nonlinear CG are particular cases of formula (C.6) :

$$\beta_k^{1,0} = \beta_k^{\mathrm{HS}} = \boldsymbol{g}_k^{\mathrm{t}} \boldsymbol{y}_{k-1} / \boldsymbol{d}_{k-1}^{\mathrm{t}} \boldsymbol{y}_{k-1} \qquad \text{[Hestenes et Stiefel, 1952]}$$
$$\beta_k^{0,0} = \beta_k^{\mathrm{PRP}} = \boldsymbol{g}_k^{\mathrm{t}} \boldsymbol{y}_{k-1} / \|\boldsymbol{g}_{k-1}\|^2 \qquad \text{[Polak et Ribière, 1969; Polyak, 1969]}$$
$$\beta_k^{0,1} = \beta_k^{\mathrm{LS}} = -\boldsymbol{g}_k^{\mathrm{t}} \boldsymbol{y}_{k-1} / \boldsymbol{d}_{k-1}^{\mathrm{t}} \boldsymbol{g}_{k-1} \qquad \text{[Liu et Storey, 1991]}$$

Other important cases are not covered by the present study, such as the Fletcher-Reeves method [Fletcher et Reeves, 1964], the Conjugate Descent method [Fletcher, 1987] and the Dai-Yuan method [Dai et Yuan, 1999].

On the other hand, we focus on the following stepsize strategy :

$$\alpha_k = \alpha_k^1 = 0 \quad \text{if } \boldsymbol{d}_k = \boldsymbol{0}; \tag{C.7}$$

otherwise,

$$\begin{cases} \alpha_k^0 = 0 & \text{(C.8a)} \\ \alpha_k^{i+1} = \alpha_k^i - \theta \boldsymbol{d}_k^{\mathrm{t}} \nabla \mathcal{J}(\boldsymbol{x}_k + \alpha_k^i \boldsymbol{d}_k) / \boldsymbol{d}_k^{\mathrm{t}} \boldsymbol{Q}_k^i \boldsymbol{d}_k, \quad i \in \{0, \ldots, I-1\} & \text{(C.8b)} \\ \alpha_k = \alpha_k^I & \text{(C.8c)} \end{cases}$$

where $I \in \mathbb{N} - \{0\}$, $\theta \in \mathbb{R}$ is a parameter, $\{\boldsymbol{Q}_k^i\} \in \mathbb{R}^{n \times n}$ is a series of symmetric, positive definite matrices with a uniformly bounded spectrum and a strictly positive lower bound, *i.e.,* there exist $\nu_1, \nu_2 \in \mathbb{R}$ with $\nu_2 \geq \nu_1 > 0$ such that

$$\nu_1 \|\boldsymbol{v}\|^2 \leq \boldsymbol{v}^{\mathrm{t}} \boldsymbol{Q}_k^i \boldsymbol{v} \leq \nu_2 \|\boldsymbol{v}\|^2, \quad \forall k \in \mathbb{N}, \forall i \in \{0, \ldots, I-1\}, \forall \boldsymbol{v} \in \mathbb{R}^n. \tag{C.9}$$

The fixed number of iterations of (C.8b) yields a family of CG methods with a *closed-form stepsize formula* (CFSF). In the case of a single application of (C.8b) ($I = 1$), our stepsize formula boils down to

$$\alpha_k = \alpha_k^1 = -\theta \boldsymbol{g}_k^{\mathrm{t}} \boldsymbol{d}_k / \boldsymbol{d}_k^{\mathrm{t}} \boldsymbol{Q}_k^0 \boldsymbol{d}_k, \tag{C.10}$$

which is exactly the formula introduced in [Sun et Zhang, 2001; Chen et Sun, 2002]. According to (C.4), note that the latter expression for α_k is nonnegative provided that $\theta > 0$.

To ensure convergence, the condition $\theta \in]0, \nu_1/\mu[$ is introduced in [Sun et Zhang, 2001; Chen et Sun, 2002], where μ is a Lipschitz constant (see Assumption 3 below). In Section C.5, we show that this condition is overrestrictive, so that the stepsize formula proposed in [Sun et Zhang, 2001; Chen et Sun, 2002] produces too small steps. This becomes obvious in the convex quadratic case, since the optimal stepsize $\theta = 1$ does not belong to the interval $]0, \nu_1/\mu[\subset]0, 1[$.

In this paper, we propose relaxed convergence conditions. In particular, the optimal stepsize becomes admissible in the convex quadratic case. The key ingredient we incorporate consists in approximating J by a convex quadratic function from above, which is the basic principle of the Weiszfeld's method [Weiszfeld, 1937; Voss et Eckhardt, 1980]. First of all, we put forward that the stepsize formula proposed in [Sun et Zhang, 2001; Chen et Sun, 2002] identifies with one iteration of Weiszfeld's algorithm in the scalar case. More generally, our iterated version (C.8b) corresponds to a fixed number of the same scalar algorithm. The majorizing convex quadratic approximation framework provides altered convergence conditions compared to the conditions found in [Sun et Zhang, 2001; Chen et Sun, 2002] : in particular, $\theta \in]0, \nu_1/\mu[$ is replaced by $\theta \in]0, 2[$ for any finite value of I.

The paper is organized as follows. Some preliminary results on the family of CG methods with the closed-form stepsize formula (C.7)-(C.8) are given in Section C.2. We also introduce the additional assumption of a majorizing convex quadratic function that allow us to make the connection between the closed-form stepsize formula and the scalar Weiszfeld's method. Section C.3 gathers some properties concerning the stepsize series generated by (C.8) useful for the next section. Section C.4 includes the main convergence properties of the two-parameter family of CG methods defined by (C.2)-(C.8). Finally, discussions on the convex quadratic case, the general case and the case of edge preserving image restoration are given in Section C.5.

C.2 Preliminaries

Let N be a neighborhood of the level set $L = \{\boldsymbol{x} \in \mathbb{R}^n | \mathcal{J}(\boldsymbol{x}) \leq \mathcal{J}(\boldsymbol{x}_0)\}$, which is assumed to be bounded in the sequel. The following assumption is also adopted.

Assumption 3. *Let us assume that $\mathcal{J} : \mathbb{R}^n \mapsto \mathbb{R}$ is differentiable on N, and that $\nabla\mathcal{J}$ is Lipschitz continuous on N with the Lipschitz constant $\mu > 0$:*

$$\|\nabla\mathcal{J}(\boldsymbol{x}) - \nabla\mathcal{J}(\boldsymbol{x}')\| \leq \mu\|\boldsymbol{x} - \boldsymbol{x}'\|, \quad \forall \boldsymbol{x}, \boldsymbol{x}' \in N.$$

In short, it will be said that $\mathcal{J}$ is μ-$\mathcal{LC}^1$.

In the sequel, Assumption 3 will appear to be sufficient for the global convergence of the CG method when $\mu_k = 0$ and $\omega_k \in [0, 1]$, which encompasses the PRP and the LS cases, but not the HS case. Thus, we consider the following stronger assumption for the more general case $\mu_k \in [0, 1]$, $\omega_k \in [0, 1 - \mu_k]$.

Assumption 4. *Let Assumption 3 hold, and let $\mathcal{J}$ be strongly convex on N : there exists $\lambda > 0$ such that*

$$\left[\nabla\mathcal{J}(\boldsymbol{x}) - \nabla\mathcal{J}(\boldsymbol{x}')\right]^{\mathrm{t}} (\boldsymbol{x} - \boldsymbol{x}') \geq \lambda\|\boldsymbol{x} - \boldsymbol{x}'\|^2, \quad \forall \boldsymbol{x}, \boldsymbol{x}' \in N.$$

Note that Assumption 4 implies that L bounded since a strongly convex function has bounded level sets.

Finally, let us introduce convex quadratic majorizing functions through the following assumption.

Assumption 5. *Let*

$$\widehat{\mathcal{J}}_k^i(\boldsymbol{x}', \boldsymbol{x}) = \mathcal{J}(\boldsymbol{x}) + (\boldsymbol{x}' - \boldsymbol{x})^{\mathrm{t}} \nabla \mathcal{J}(\boldsymbol{x}) + (\boldsymbol{x}' - \boldsymbol{x})^{\mathrm{t}} \boldsymbol{Q}_k^i (\boldsymbol{x}' - \boldsymbol{x})/2 \tag{C.11}$$

where $\{\boldsymbol{Q}_k^i\}$ *is a series of positive definite matrices, such that*

$$\widehat{\mathcal{J}}_k^i(\boldsymbol{x}', \boldsymbol{x}) \geq \mathcal{J}(\boldsymbol{x}'), \quad \forall \boldsymbol{x}, \boldsymbol{x}' \in N, \tag{C.12}$$

for all $k \in \mathbb{N}, i \in \{0, \ldots, I-1\}$.

For sake of notational simplicity, let $f(\alpha) = \mathcal{J}(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k)$. Moreover, the current iteration index k will remain implicit whenever unambiguous : typically, the stepsize α_k^i will be abridged into α^i. Using such compact notations, the stepsize update (C.8) also reads

$$\begin{cases} \alpha^0 = 0 \\ \alpha^{i+1} = \alpha^i - \theta \dot{f}(\alpha^i)/a_i, \quad i \in \{0, \ldots, I-1\} \\ \alpha_k = \alpha^I \end{cases} \tag{C.13}$$

with $\dot{f}(\alpha^i) = \boldsymbol{d}_k^{\mathrm{t}} \nabla \mathcal{J}\left(\boldsymbol{x}_k + \alpha^i \boldsymbol{d}_k\right)$ and $a_i = \boldsymbol{d}_k^{\mathrm{t}} \boldsymbol{Q}_k^i \boldsymbol{d}_k$.

According to (C.9) we have

$$0 < \nu_1 \|\boldsymbol{d}_k\|^2 \leq a_i \leq \nu_2 \|\boldsymbol{d}_k\|^2. \tag{C.14}$$

According to (C.11), let

$$q_i(\alpha', \alpha) = \widehat{\mathcal{J}}_k^i(\boldsymbol{x}_k + \alpha' \boldsymbol{d}_k, \boldsymbol{x}_k + \alpha \boldsymbol{d}_k) = f(\alpha) + (\alpha' - \alpha)\dot{f}(\alpha) + (\alpha' - \alpha)^2 a_i/2, \tag{C.15}$$

which is a convex parabola as a function of α'.

Let us rely on a fixed number I of iterations of Weiszfeld's method for the determination of the stepsize. The (scalar) function to minimize is f and, according to Assumption 5, $q_i(\alpha', \alpha)$ is an upper convex quadratic approximation of $f(\alpha')$. Then, the successive iterations of Weiszfeld's method are defined by

$$\alpha^{i+1} = \arg\min_{\alpha'} q_i(\alpha', \alpha^i). \tag{C.16}$$

According to the parabolic nature of $q_i(\alpha', \alpha)$ as a function of α', (C.16) takes the following expression :

$$\alpha^{i+1} = \alpha^i - \dot{f}(\alpha^i)/a_i.$$

Hence, (C.13) identifies with a relaxed version of Weiszfeld's method to minimize f. Note that the convergence results in Section C.4 hold regardless of the value of I. This is in contrast with usual line search procedures, where appropriate stopping conditions (*e.g.,* Wolfe conditions) must be checked to ensure convergence.

As already mentioned, the stepsize formula (C.10) proposed in [Sun et Zhang, 2001; Chen et Sun, 2002] formally corresponds to one iteration of the same relaxed Weiszfeld's method. We are now led to a deeper result : the condition $\theta \in]0, \nu_1/\mu[$ stated in [Sun et Zhang, 2001; Chen et Sun, 2002] for the convergence of their CG method implies that our Assumption 5 holds. First, let us give an equivalent formulation for (C.10).

Let $\tilde{\boldsymbol{Q}}_k^0 = \boldsymbol{Q}_k^0/\theta$, so that (C.10) also reads $\alpha_k = -\boldsymbol{g}_k^{\mathrm{t}} \boldsymbol{d}_k / \boldsymbol{d}_k^{\mathrm{t}} \tilde{\boldsymbol{Q}}_k^0 \boldsymbol{d}_k$. From (C.9) and $\theta \in]0, \nu_1/\mu[$, we deduce that

$$\boldsymbol{v}^{\mathrm{t}} \tilde{\boldsymbol{Q}}_k^0 \boldsymbol{v} \geq \mu \|\boldsymbol{v}\|^2, \quad \forall k \in \mathbb{N}, \forall \boldsymbol{v} \in \mathbb{R}^n, \tag{C.17}$$

i.e., the spectrum of matrices $\tilde{\boldsymbol{Q}}_k^0$ is bounded from below by μ. Such a simple change of notations stresses that the constraint $\theta \in]0, \nu_1/\mu[$ stated in [Sun et Zhang, 2001; Chen et Sun, 2002] can be translated as a constraint on the matrices $\boldsymbol{Q}_k^i$. The following lemma shows that matrices $\tilde{\boldsymbol{Q}}_k^0$ yield convex quadratic majorizing approximations in the sense of Assumption 5 (provided that N is a convex set).

Lemma C.2.1. *Suppose that Assumption 3 holds, and also that the lower bound ν_1 is not smaller than the Lipschitz constant μ. Let us restrict ourselves to the case where N is a convex set. Then Assumption 5 holds, i.e., the function $\widehat{\mathcal{J}}_k^i(\boldsymbol{x}', \boldsymbol{x})$ defined by (C.11) fulfills (C.12) over N.*

Proof. According to the Descent Lemma [Bertsekas, 1999, Prop. A.24], we have

$$\mathcal{J}(\boldsymbol{x}') - \mathcal{J}(\boldsymbol{x}) - (\boldsymbol{x}' - \boldsymbol{x})^{\mathrm{t}} \nabla \mathcal{J}(\boldsymbol{x}) \leq \mu \|\boldsymbol{x}' - \boldsymbol{x}\|^2/2 \tag{C.18}$$

for any $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^n$ if $\mathcal{J}$ is μ-$\mathcal{LC}^1$on $\mathbb{R}^n$. Actually, it is easy to check that (C.18) still holds for any $\boldsymbol{x}, \boldsymbol{x}' \in N$ if $\mathcal{J}$ is μ-$\mathcal{LC}^1$on N, provided that N is convex.

Since the spectrum of $\{\boldsymbol{Q}_k^i\}$ is bounded from below by $\nu_1 \geq \mu$, we have

$$\mu \|\boldsymbol{x}' - \boldsymbol{x}\|^2 \leq \nu_1 \|\boldsymbol{x}' - \boldsymbol{x}\|^2 \leq \left(\boldsymbol{x}' - \boldsymbol{x}\right)^{\mathrm{t}} \boldsymbol{Q}_k^i \left(\boldsymbol{x}' - \boldsymbol{x}\right)$$

Jointly with (C.18), the latter yields

$$\mathcal{J}(\boldsymbol{x}') - \mathcal{J}(\boldsymbol{x}) + (\boldsymbol{x} - \boldsymbol{x}')^{\mathrm{t}} \nabla \mathcal{J}(\boldsymbol{x}) \leq \left(\boldsymbol{x}' - \boldsymbol{x}\right)^{\mathrm{t}} \boldsymbol{Q}_k^i \left(\boldsymbol{x}' - \boldsymbol{x}\right)/2,$$

i.e., $\widehat{\mathcal{J}}_k^i(\boldsymbol{x}', \boldsymbol{x}) \geq \mathcal{J}(\boldsymbol{x}')$. □

Lemma C.2.1 indicates that Assumption 5 is not a restrictive condition compared to the hypotheses found in [Sun et Zhang, 2001; Chen et Sun, 2002]. On the contrary, it is a weaker assumption (let alone the fact that Lemma C.2.1 only applies when N is a convex set), so that a convergence proof based on Assumption 5 would be of broader applicability. This is the goal reached in Section C.4, where $\mathcal{J}$ is not necessarily assumed ν_1-$\mathcal{LC}^1$(and N is not necessarily convex).

C.3 Properties of the stepsize series

The present section gathers technical results concerning the stepsize series $\alpha^i = \alpha_k^i$ generated by (C.8), which will be useful to derive the global convergence properties of the next section. Remark that Assumption 4 is never used in this section.

Let us introduce the notation $\Gamma(a, b) = [\min(a, b), \max(a, b)]$ to handle with intervals with unordered endpoints.

Lemma C.3.1. *Suppose that Assumption 3 and Assumption 5 hold and that $\theta \in]0, 2[$. Then*

$$\mathcal{J}(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k) \leq \mathcal{J}(\boldsymbol{x}_k^i), \quad \forall \alpha \in \Gamma(\alpha_k^i, \alpha_k^{i+1}) \tag{C.19}$$

for all $k \geq 0$, $i \in \{0, \ldots, I-1\}$, where $\boldsymbol{x}_k^i = \boldsymbol{x}_k + \alpha_k^i \boldsymbol{d}_k$.

Proof. Let us first assume $\boldsymbol{x}_k^0 \in N$, and then let us show that (C.19) holds, recursively on i.

$\dot{f}(\alpha_k^i)$ exists since $\mathcal{J}$ is differentiable on N. We have $\alpha_k^0 = 0$ and $\dot{f}(0) = \boldsymbol{g}_k^{\mathrm{t}} \boldsymbol{d}_k \leq 0$, but the sign of $\dot{f}(\alpha_k^i) = \boldsymbol{d}_k^{\mathrm{t}} \nabla \mathcal{J}\left(\boldsymbol{x}_k + \alpha_k^i \boldsymbol{d}_k\right)$ is indeterminate for $i > 0$. Let us study each case separately (the index k is omitted in the rest of the proof).

– Suppose $\dot{f}(\alpha^i) = 0$. According to (C.13), $\alpha^{i+1} = \alpha^i$ so (C.19) is true.

– Suppose $\dot{f}(\alpha^i) < 0$. According to (C.13) and $a_i > 0$ we have $\alpha^{i+1} > \alpha^i$. Let us prove (C.19) by contradiction : suppose, on the contrary, that there exists $\alpha' \in (\alpha^i, \alpha^{i+1}]$ such that

$$f(\alpha') > f(\alpha^i). \tag{C.20}$$

Let $\ell^i = \{\alpha \in \mathbb{R} | f(\alpha) \leq f(\alpha^i)\}$. Since f is continuous on ℓ^i, according to (C.20) and $\dot{f}(\alpha^i) < 0$, there exists $\alpha'' \in \left]\alpha^i, \alpha'\right[$ such that $f(\alpha'') < f(\alpha^i)$. There also exists $\alpha''' \in]\alpha'', \alpha'[$ such that

$$f(\alpha''') = f(\alpha^i); \tag{C.21}$$

in the contrary case, since f is continuous on ℓ^i, the inequality $f(\alpha) < f(\alpha^i)$ would hold for all $\alpha \in]\alpha'', \alpha'[$. In particular, we would get

$$\lim_{\substack{\alpha \to \alpha' \\ \alpha < \alpha'}} f(\alpha) < f(\alpha^i), \tag{C.22}$$

thus $\alpha \in \ell^i$ for all values of $\alpha \in (\alpha'', \alpha')$ arbitrary close to α'. (C.22) would be incompatible with (C.20) given the continuity of f on ℓ^i.
Let $q(\alpha) = q_i(\alpha, \alpha^i)$, where q_i is defined by (C.15). Since $\dot{q}(\alpha^{i+1}) = \dot{f}(\alpha^i)(1-\theta)$, $\dot{q}(\alpha^i) = \dot{f}(\alpha^i) < 0$ and $\theta \in]0, 2[$, we have $\dot{q}(\alpha^{i+1}) \in \left]\dot{q}(\alpha^i), -\dot{q}(\alpha^i)\right[$. Because q is a convex parabola and $\alpha''' \in]\alpha'', \alpha'[\subset \left]\alpha^i, \alpha^{i+1}\right[$, we can conclude that $q(\alpha''') < q(\alpha^i) = f(\alpha^i)$. Hence, according to (C.21), we get $q(\alpha''') < f(\alpha''')$, which contradicts the majorizing character (C.12) of $\widehat{\mathcal{J}}_k^i$ w.r.t. $\mathcal{J}$ at $\boldsymbol{x}_k + \alpha''' \boldsymbol{d}_k \in N$.

– Suppose $\dot{f}(\alpha^i) > 0$. According to (C.13) and $a_i > 0$, we have $\alpha^{i+1} < \alpha^i$. We are led back to the previous case if we replace $f(\alpha)$ by $f(-\alpha)$.

Our intermediate conclusion is that (C.19) holds for all $i \in \{0, \ldots, I-1\}$: in particular $\mathcal{J}(\boldsymbol{x}_k^{i+1}) \leq \mathcal{J}(\boldsymbol{x}_k^i)$.

Since $\boldsymbol{x}_0 \in N$ and $\boldsymbol{x}_{k+1}^0 = \boldsymbol{x}_k^I = \boldsymbol{x}_k$, we get

$$\mathcal{J}(\boldsymbol{x}_k^0) = \mathcal{J}(\boldsymbol{x}_{k-1}^I) \leq \cdots \leq \mathcal{J}(\boldsymbol{x}_{k-1}^0) \leq \cdots \leq \mathcal{J}(\boldsymbol{x}_1^0) = \mathcal{J}(\boldsymbol{x}_0), \quad \forall k$$

by immediate recursion. Hence $\boldsymbol{x}_k^0 \in N$, which proves that (C.19) holds for all $k \geq 0$, $i \in \{0, \ldots, I-1\}$. □

An immediate consequence of Lemma C.3.1 is

$$\boldsymbol{x}_k + \alpha \boldsymbol{d}_k \in N, \quad \forall \alpha \in [0, \alpha_k^i], \tag{C.23}$$

for all $k \geq 0$, $i \in \{0, \ldots, I-1\}$ since $\boldsymbol{x}_0 \in N$. Thus, according to (C.12),

$$q_i(\alpha^j, \alpha^i) \geq f(\alpha^j), \quad \forall i, j \in \{0, \ldots, I-1\}. \tag{C.24}$$

The following three lemmas are specific to the case when $\dot{f}(0) = \boldsymbol{g}_k^{\mathrm{t}} \boldsymbol{d}_k$ does not vanish for the current iteration k, *i.e.*, $\boldsymbol{g}_k^{\mathrm{t}} \boldsymbol{d}_k < 0$. Then $\boldsymbol{d}_k \neq 0$, and the sequence $\left\{\alpha_k^i\right\}$ is well defined according to (C.8b).

Lemma C.3.2. *Suppose that Assumption 3 and Assumption 5 hold. Assume also that $\dot{f}(0) < 0$ and $\theta \in]0, 2[$. Then the whole sequence $\{\alpha^i\}$ is strictly positive :*

$$\alpha^i > 0, \quad \forall i \in \{0, \ldots, I-1\}. \tag{C.25}$$

Proof. According to (C.24), we have

$$q_i(0,\alpha^i) = f(\alpha^i) - \alpha^i \dot{f}(\alpha^i) + (\alpha^i)^2 a_i/2 \geq f(0).$$

Since $\{f(\alpha^i)\}$ is a nonincreasing sequence according to Lemma C.3.1, we deduce that

$$-\alpha^i \dot{f}(\alpha^i) + (\alpha^i)^2 a_i/2 \geq 0$$

so that, according to (C.13) and $a_i > 0$,

$$\alpha^i(\alpha^{i+1} - 2\delta \dot{f}(\alpha^i)/a_i) \geq 0 \tag{C.26}$$

with

$$\delta = 1 - \theta/2 \in]0,1[\,. \tag{C.27}$$

Now, let us show (C.25) by a recurrence on i. We have $\alpha^1 > 0$ according to (C.13) and $\dot{f}(0) < 0$. Let us assume now that $\alpha^i > 0$ for some i.

- If $\dot{f}(\alpha^i) \leq 0$, then $\alpha^{i+1} > 0$ according to (C.13).
- If $\dot{f}(\alpha^i) > 0$, then given $\alpha^i > 0$, inequality (C.26) yields $\alpha^{i+1} > 0$.

□

Lemma C.3.3. *Suppose that Assumption 3 and Assumption 5 hold. Assume also that* $\dot{f}(0) < 0$ *and* $\theta \in]0,2[$. *Then, for all* $i \in \mathbb{N} - \{0\}$,

$$f(\alpha^i) \leq q_0(\alpha^1, 0), \tag{C.28}$$

$$c^{\min}\alpha^1 \leq \alpha^i, \tag{C.29}$$

where

$$c^{\min} = \left(\sqrt{1 + 2\mu\theta\delta/\nu_1} - 1\right)\nu_1/\theta\mu \in]0,1[\,. \tag{C.30}$$

Proof. The proof of (C.28) is straightforward : according to (C.24), we have $q_0(\alpha^1, 0) \geq f(\alpha^1)$. Then (C.28) holds, because $\{f(\alpha^i)\}$ is a decreasing sequence according to Lemma C.3.1.

The derivation of (C.29) is not so direct. Let g the concave parabola defined by

$$g(\alpha) = f(0) + \alpha \dot{f}(0) - \mu a_0 \alpha^2/2\nu_1. \tag{C.31}$$

Remark that $g(0) = f(0)$ and that g is decreasing on $\mathbb{R}^+$ since $\dot{g}(0) = \dot{f}(0) < 0$.

Let us first show that

$$g(\alpha^i) \leq f(\alpha^i). \tag{C.32}$$

Let us consider $\alpha \in [0, \alpha^i]$: $\boldsymbol{x}_k + \alpha \boldsymbol{d}_k \in N$ according to (C.23). Since $f(\alpha) = \mathcal{J}(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k)$ and Assumption 3 holds, we have

$$|\dot{f}(\alpha) - \dot{f}(0)| = \left|\boldsymbol{d}_k^{\mathrm{t}} \left(\nabla \mathcal{J}(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k) - \nabla \mathcal{J}(\boldsymbol{x}_k)\right)\right| \leq \|\boldsymbol{d}_k\|^2 \mu |\alpha|$$

and according to (C.14), we get

$$|\dot{f}(\alpha) - \dot{f}(0)| \leq a_0 \mu \alpha/\nu_1.$$

Given $|\dot{f}(\alpha)| \leq |\dot{f}(\alpha) - \dot{f}(0)| + |\dot{f}(0)|$ and $\dot{f}(0) < 0$, we obtain

$$|\dot{f}(\alpha)| \leq a_0 \mu \alpha/\nu_1 - \dot{f}(0). \tag{C.33}$$

Thus,

$$\dot{f}(0) - a_0\mu\alpha/\nu_1 \leq \dot{f}(\alpha)$$

or equivalently

$$\dot{g}(\alpha) \leq \dot{f}(\alpha), \quad \forall \alpha \in [0, \alpha^i] \tag{C.34}$$

according to (C.31). Since $g(0) = f(0)$, (C.32) is obtained by integrating (C.34) between 0 and α^i.

According to (C.10), (C.13), (C.15) and (C.27), we have

$$\begin{aligned} q_0(\alpha^1, 0) &= f(0) + \alpha^1\dot{f}(0) + (\alpha^1)^2 a_0/2 \\ &= f(0) + \delta\alpha^1\dot{f}(0). \end{aligned} \tag{C.35}$$

Then let us show that $q_0(\alpha^1, 0) = g(\alpha^{\min})$, where $\alpha^{\min} = c^{\min}\alpha^1$. From (C.30) we have

$$\begin{aligned} (c^{\min})^2 &= \left(2 + 2\mu\theta\delta/\nu_1 - 2\sqrt{1 + 2\mu\theta\delta/\nu_1}\right)\nu_1^2/(\theta\mu)^2 \\ &= \left(\delta - c^{\min}\right) 2\nu_1/\theta\mu. \end{aligned} \tag{C.36}$$

According to (C.31), we have also

$$g(\alpha^{\min}) = f(0) + c^{\min}\alpha^1\dot{f}(0) - \mu a_0(c^{\min}\alpha^1)^2/2\nu_1$$

which also reads

$$g(\alpha^{\min}) = f(0) + \alpha^1\dot{f}(0)\left(c^{\min} + (c^{\min})^2\theta\mu/2\nu_1\right)$$

according to $\alpha^1 = -\theta\dot{f}(0)/a_0$. Jointly with (C.36), the latter identity yields

$$g(\alpha^{\min}) = f(0) + \delta\alpha^1\dot{f}(0),$$

so that $g(\alpha^{\min})$ identifies with $q_0(\alpha^1, 0)$ according to (C.35).

On the other hand, $\{\alpha^i\}$ is positive according to Lemma C.3.2. We are now in position to show (C.29) by contradiction : assume that there exists $i > 0$ such that $0 \leq \alpha^i < \alpha^{\min}$. According to (C.32) and given that g is decreasing on $\mathbb{R}^+$, we get $f(\alpha^i) \geq g(\alpha^i) > g(\alpha^{\min}) = q_0(\alpha^1, 0)$, which contradicts (C.28).

Finally, it is obvious that $c^{\min} > 0$. Let us consider the alternate expression

$$c^{\min} = 2\delta/\left(\sqrt{1 + 2\mu\theta\delta/\nu_1} + 1\right),$$

so it becomes also apparent that $c^{\min} < \delta < 1$. □

Lemma C.3.4. *Suppose that Assumption 3 and Assumption 5 hold. Assume also that $\dot{f}(0) < 0$ and $\theta \in]0, 2[$. Then*

$$\alpha^i \leq c_i^{\max}\alpha^1 \tag{C.37}$$

$\forall i \in \mathbb{N} - \{0\}$, with

$$c_i^{\max} = \left(1 + \nu_2\theta\mu/\nu_1^2\right)^{i-1}\left(1 + \nu_1/\theta\mu\right) - \nu_1/\theta\mu \geq 1. \tag{C.38}$$

Proof. It is easy to check that $c_i^{\max}$ is not smaller than 1 for all $i > 0$. Let us show the inequality (C.37) recursively on i. It is valid for $i = 1$, since $c_1^{\max} = 1$. Now let us suppose that $\alpha^i \leq c_i^{\max}\alpha^1$, and let us prove that $\alpha^{i+1} \leq c_{i+1}^{\max}\alpha^1$.

According to (C.13), we have $\alpha^{i+1} \leq \alpha^i + |\dot{f}(\alpha^i)|\,\theta/a_i$ and according to (C.14), we have also $a_i \geq a_0\nu_1/\nu_2$. Thus,

$$\alpha^{i+1} \leq \alpha^i + |\dot{f}(\alpha^i)|\,\theta\nu_2/\nu_1 a_0. \tag{C.39}$$

On the other hand, (C.33) implies

$$|\dot{f}(\alpha^i)| \leq a_0\mu\alpha^i/\nu_1 - \dot{f}(0).$$

In combination with the latter inequality and with $\alpha^1 = -\theta\dot{f}(0)/a_0$, (C.39) yields

$$\alpha^{i+1} \leq \alpha^i(1 + \nu_2\theta\mu/\nu_1^2) + \nu_2\alpha^1/\nu_1,$$

which corresponds to a recursive definition of the series $(c_i^{\max})$ according to

$$c_{i+1}^{\max} = c_i^{\max}(1 + \nu_2\theta\mu/\nu_1^2) + \nu_2/\nu_1.$$

Given $c_1^{\max} = 1$, it can be checked that (C.38) is the general term of the series. □

Definition C.3.1. *The stepsize sequence $\{\alpha_k\}$ satisfies the Armijo condition with $\Omega \in]0,1[$ if*

$$\mathcal{J}(\boldsymbol{x}_k) - \mathcal{J}(\boldsymbol{x}_{k+1}) + \Omega\alpha_k\boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{d}_k \geq 0, \quad \forall k. \tag{C.40}$$

Lemma C.3.5. *Suppose that Assumption 3 and Assumption 5 hold. Assume also that $\theta \in]0,2[$. Then the stepsize sequence defined by (C.8) satisfies the Armijo condition with*

$$\Omega = \Omega_I = \delta/c_I^{\max} \in]0,1[\,, \tag{C.41}$$

where δ and $c_I^{\max}$ are defined by (C.27) and (C.38), respectively.

Proof. We have $\dot{f}(0) = \boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{d}_k \leq 0$. Let us first examine the particular case $\dot{f}(0) = 0$: according to (C.8), α_k vanishes, so that (C.40) holds trivially.

Suppose now $\dot{f}(0) < 0$. According to (C.35), (C.28) also reads

$$f(0) - f(\alpha^I) + \delta\dot{f}(0)\alpha^1 \geq 0. \tag{C.42}$$

Finally, since $\dot{f}(0) < 0$ and $\alpha^1 \geq \alpha^I/c_I^{\max} > 0$ according to (C.37), (C.42) implies that

$$f(0) - f(\alpha^I) + \delta\dot{f}(0)\alpha^I/c_I^{\max} \geq 0,$$

which identifies with (C.40) with $\Omega = \Omega_I$. □

Remark C.3.1. *In Lemma C.3.5, $\Omega_I = \delta/c_I^{\max}$ does not depend on k, which is an essential point for the fulfillment of the Armijo condition.*

The following theorem sums up the main results that will be useful in the next section.

Theorem C.3.1. *Let $\boldsymbol{x}_k$ be defined by (C.2)-(C.8) with $\theta \in]0,2[$, and let Assumption 3 and Assumption 5 hold. Then the Armijo condition (C.40) is satisfied by the stepsize sequence $\{\alpha_k\}$ with $\Omega = \Omega_I = \delta/c_I^{\max}$, where δ and $c_I^{\max}$ are defined by (C.27) and (C.38), respectively. Moreover, we have*

$$0 \leq c^{\min}\alpha_k^1 \leq \alpha_k \leq c_I^{\max}\alpha_k^1, \quad \forall k, \tag{C.43}$$

where $c^{\min}$ is defined by (C.30).

Proof. Lemma C.3.5 corresponds to the fulfillment of the Armijo condition.

On the other hand, we have $\dot{f}(0) \leq 0$. If $\dot{f}(0) = 0$, then $\alpha_k = 0$, so (C.43) trivially holds. Otherwise, we have $\dot{f}(0) < 0$, so (C.43) is a joint consequence of Lemmas C.3.3 and C.3.4. □

C.4 Global convergence

The two following lemmas establish results for the whole two-parameter family of conjugacy coefficient $\beta_k = \beta_k^{\mu_k,\omega_k}$. Then, we will draw conclusions for specific CG methods.

Lemma C.4.1. *Under the conditions of Theorem C.3.1, we have*

$$\sum_{k,\, \boldsymbol{d}_k \neq 0} (\boldsymbol{g}_k^{\mathrm{t}} \boldsymbol{d}_k)^2 / \|\boldsymbol{d}_k\|^2 < \infty. \tag{C.44}$$

Proof. According to Theorem C.3.1, the Armijo condition (C.40) is satisfied with $\Omega = \Omega_I$. Given (C.43) and $\boldsymbol{g}_k^{\mathrm{t}} \boldsymbol{d}_k \leq 0$, we deduce that

$$\mathcal{J}(\boldsymbol{x}_k) - \mathcal{J}(\boldsymbol{x}_{k+1}) \geq -\Omega_I c^{\min} \alpha_k^1 \boldsymbol{g}_k^{\mathrm{t}} \boldsymbol{d}_k. \tag{C.45}$$

If $\boldsymbol{d}_k \neq \boldsymbol{0}$, we have

$$\alpha_k^1 = -\theta \boldsymbol{g}_k^{\mathrm{t}} \boldsymbol{d}_k / \boldsymbol{d}_k^{\mathrm{t}} \boldsymbol{Q}_k^0 \boldsymbol{d}_k \geq -\theta \boldsymbol{g}_k^{\mathrm{t}} \boldsymbol{d}_k / \nu_2 \|\boldsymbol{d}_k\|^2 \tag{C.46}$$

according to (C.10) and (C.9), so that

$$\mathcal{J}(\boldsymbol{x}_k) - \mathcal{J}(\boldsymbol{x}_{k+1}) \geq c_0 (\boldsymbol{g}_k^{\mathrm{t}} \boldsymbol{d}_k)^2 / \|\boldsymbol{d}_k\|^2 \geq 0 \tag{C.47}$$

with $c_0 = \Omega_I c^{\min} \theta / \nu_2 > 0$.

Given Assumption 3 and that L is assumed bounded, (C.47) implies $\lim_{k\to\infty} \mathcal{J}(\boldsymbol{x}_k)$ is finite. Finally, we obtain

$$\infty > (\mathcal{J}(\boldsymbol{x}_0) - \lim_{k\to\infty} \mathcal{J}(\boldsymbol{x}_k))/c_0 \geq \sum_{k,\, \boldsymbol{d}_k \neq 0} (\boldsymbol{g}_k^{\mathrm{t}} \boldsymbol{d}_k)^2 / \|\boldsymbol{d}_k\|^2.$$

□

Lemma C.4.2. *Let $k \in \mathbb{N}$. Under the conditions of Theorem C.3.1, we have*

$$|\boldsymbol{g}_{k+1}^{\mathrm{t}} \boldsymbol{d}_k| \leq -\boldsymbol{g}_k^{\mathrm{t}} \boldsymbol{d}_k \left(1 + c_I^{\max} \theta \mu / \nu_1\right). \tag{C.48}$$

Moreover, if Assumption 4 holds, then

$$-\boldsymbol{g}_{k+1}^{\mathrm{t}} \boldsymbol{d}_k \leq -\boldsymbol{g}_k^{\mathrm{t}} \boldsymbol{d}_k (1 - c^{\min} \theta \lambda / \nu_2). \tag{C.49}$$

Proof. (C.48) and (C.49) are trivial assertions if $\boldsymbol{d}_k = \boldsymbol{0}$. Otherwise, following [Sun et Zhang, 2001; Chen et Sun, 2002], let us define

$$\phi_k = \begin{cases} (\boldsymbol{g}_{k+1} - \boldsymbol{g}_k)^{\mathrm{t}} (\boldsymbol{x}_{k+1} - \boldsymbol{x}_k) / \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2 = \boldsymbol{y}_k^{\mathrm{t}} \boldsymbol{d}_k / \alpha_k \|\boldsymbol{d}_k\|^2 & \text{for } \alpha_k \neq 0 \\ 0 & \text{for } \alpha_k = 0. \end{cases} \tag{C.50}$$

Note that according to (C.23), $\boldsymbol{x}_k \in N$. If Assumption 3 holds, then $|\phi_k| \leq \mu$ according to Cauchy-Schwartz inequality. If Assumption 4 holds, then $\phi_k \geq \lambda > 0$.

According to (C.10) and (C.50), we have

$$\boldsymbol{g}_{k+1}^{\mathrm{t}} \boldsymbol{d}_k = \boldsymbol{g}_k^{\mathrm{t}} \boldsymbol{d}_k + \boldsymbol{y}_k^{\mathrm{t}} \boldsymbol{d}_k = \boldsymbol{g}_k^{\mathrm{t}} \boldsymbol{d}_k + \alpha_k \phi_k \|\boldsymbol{d}_k\|^2. \tag{C.51}$$

According to (C.43), (C.51), $\mu \geq |\phi_k|$, and $\boldsymbol{g}_k^{\mathrm{t}} \boldsymbol{d}_k \leq 0$, we deduce that

$$|\boldsymbol{g}_{k+1}^{\mathrm{t}} \boldsymbol{d}_k| \leq -\boldsymbol{g}_k^{\mathrm{t}} \boldsymbol{d}_k + \mu c_I^{\max} \alpha_k^1 \|\boldsymbol{d}_k\|^2.$$

According to (C.10), we have also

$$|\boldsymbol{g}_{k+1}^{\mathrm{t}}\boldsymbol{d}_k| \leq -\boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{d}_k - \boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{d}_k \mu c_I^{\max}\theta\|\boldsymbol{d}_k\|^2/\boldsymbol{d}_k^{\mathrm{t}}\boldsymbol{Q}_k^0\boldsymbol{d}_k.$$

Finally, since $\nu_1 > 0$ is a lower bound for the spectrum of $\boldsymbol{Q}_k^0$, and $\boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{d}_k \leq 0$, we obtain (C.48).

Let us suppose now that Assumption 4 holds. Given (C.43) and $\phi_k \geq \lambda > 0$, (C.51) implies

$$\boldsymbol{g}_{k+1}^{\mathrm{t}}\boldsymbol{d}_k \geq \boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{d}_k + \lambda c^{\min}\alpha_k^1\|\boldsymbol{d}_k\|^2$$

and according to (C.46), we obtain (C.49). □

Lemma C.4.3. *Suppose that Assumption 4 holds, as well as the conditions of Theorem C.3.1. Then*

$$D_k \geq (1-\mu_k-\omega_k)\|\boldsymbol{g}_{k-1}\|^2 - \boldsymbol{d}_{k-1}^{\mathrm{t}}\boldsymbol{g}_{k-1}(\omega_k + \mu_k c^{\min}\theta\lambda/\nu_2) \geq 0, \quad \forall k \in \mathbb{N} - \{0\}. \tag{C.52}$$

Proof. Since $\boldsymbol{y}_{k-1} = \boldsymbol{g}_k - \boldsymbol{g}_{k-1}$, (C.49) also reads

$$\boldsymbol{d}_{k-1}^{\mathrm{t}}\boldsymbol{y}_{k-1} \geq -\boldsymbol{d}_{k-1}^{\mathrm{t}}\boldsymbol{g}_{k-1}c^{\min}\theta\lambda/\nu_2, \quad \forall k \in \mathbb{N} - \{0\}.$$

Then, given the expression (C.5) of D_k and $\boldsymbol{d}_{k-1}^{\mathrm{t}}\boldsymbol{g}_{k-1} \leq 0$, the conclusion is immediate. □

Remark C.4.1. *Let us examine the case where the denominator D_k of $\beta_k^{\mu_k,\omega_k}$ vanishes. Here, we assume that the conditions of Theorem C.3.1 hold.*

Let us suppose first that Assumption 4 is valid. If D_k vanishes, then (C.52) implies

$$(1-\mu_k-\omega_k)\|\boldsymbol{g}_{k-1}\|^2 - (\omega_k + \mu_k c^{\min}\theta\lambda/\nu_2)\boldsymbol{d}_{k-1}^{\mathrm{t}}\boldsymbol{g}_{k-1} = 0.$$

Since the left-hand side is the sum of two nonnegative terms, we obtain

$$\begin{cases} (1-\mu_k-\omega_k)\|\boldsymbol{g}_{k-1}\|^2 = 0, & \text{(C.53a)} \\ (\omega_k + \mu_k c^{\min}\theta\lambda/\nu_2)\boldsymbol{d}_{k-1}^{\mathrm{t}}\boldsymbol{g}_{k-1} = 0. & \text{(C.53b)} \end{cases}$$

- *Case 1 : If $\mu_k + \omega_k < 1$, (C.53a) boils down to $\|\boldsymbol{g}_{k-1}\|^2 = 0$, which means that convergence is reached at iteration $k-1$. This case includes the PRP method.*
- *Case 2 : If $\mu_k + \omega_k = 1$, (C.53b) implies $\boldsymbol{d}_{k-1}^{\mathrm{t}}\boldsymbol{g}_{k-1} = 0$, so that $\alpha_{k-1} = 0$. Thus, $\boldsymbol{x}_k = \boldsymbol{x}_{k-1}$, $\boldsymbol{y}_{k-1} = 0$, and the numerator of $\beta_k^{\mu_k,\omega_k}$ vanishes. In this case, we let $\beta_k^{\mu_k,\omega_k} = 0$, conventionally. This case includes the HS and the LS method.*

In the situation where Assumption 4 is not necessarily valid, our study only covers the case $\mu_k = 0$: then D_k is the sum of two nonnegative terms, so $D_k = 0$ implies that both vanish :

$$\begin{cases} (1-\omega_k)\|\boldsymbol{g}_{k-1}\|^2 = 0, \\ \omega_k\boldsymbol{d}_{k-1}^{\mathrm{t}}\boldsymbol{g}_{k-1} = 0. \end{cases}$$

- *If $\omega_k < 1$, the conclusion is the same as in Case 1. This case includes the PRP method.*
- *If $\omega_k = 1$, the conclusion is the same as in Case 2. This case includes the LS method.*

Lemma C.4.4. *Under the conditions of Theorem C.3.1, we have*

$$\liminf_{k\to\infty} \|\boldsymbol{g}_k\| > 0 \implies \lim_{k\to\infty} \beta_k^{0,\omega_k} = 0.$$

Moreover, if Assumption 4 is valid, then

$$\liminf_{k\to\infty} \|\boldsymbol{g}_k\| > 0 \implies \lim_{k\to\infty} \beta_k^{\mu_k,\omega_k} = 0.$$

Proof. According to (C.2) and (C.43), we have

$$\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\|^2=\alpha_k^2\|\boldsymbol{d}_k\|^2\leq(c_I^{\max})^2(\alpha_k^1)^2\|\boldsymbol{d}_k\|^2.$$

Given that (C.10) holds unless $\boldsymbol{d}_k=\boldsymbol{0}$, we deduce that

$$\begin{aligned}\sum_k\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\|^2&\leq(c_I^{\max}\theta)^2\sum_{k,\,\boldsymbol{d}_k\neq 0}\left(\boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{d}_k\right)^2\|\boldsymbol{d}_k\|^2/\left(\boldsymbol{d}_k^{\mathrm{t}}\boldsymbol{Q}_k^0\boldsymbol{d}_k\right)^2,\\&\leq(c_I^{\max}\theta/\nu_1)^2\sum_{k,\,\boldsymbol{d}_k\neq 0}\left(\boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{d}_k\right)^2/\|\boldsymbol{d}_k\|^2\end{aligned}$$

according to (C.9). Given (C.44), we conclude that $\lim_{k\to\infty}\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\|^2=0$. Because $\mathcal{J}$ is continuously differentiable and $\|\boldsymbol{g}_k\|$ is bounded according to Assumption 3 and the boundedness of L, we have also $\lim_{k\to\infty}\boldsymbol{y}_{k-1}=0$ and

$$\lim_{k\to\infty}\boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{y}_{k-1}=0. \tag{C.55}$$

If $\liminf_{k\to\infty}\|\boldsymbol{g}_k\|>0$, then there exists $\gamma>0$ such that

$$\|\boldsymbol{g}_k\|\geq\gamma>0\quad\forall k. \tag{C.56}$$

According to (C.6), we have

$$|\boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{y}_{k-1}|=|\beta_k^{\mu_k,\omega_k}||D_k|. \tag{C.57}$$

On the one hand, suppose that Assumption 4 is valid.

Firstly, let us consider the iteration indices k such that $\mu_k+\omega_k\in[0,1/2]$. According to (C.52) and $\boldsymbol{d}_{k-1}^{\mathrm{t}}\boldsymbol{g}_{k-1}\leq 0$, (C.57) implies that

$$|\boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{y}_{k-1}|\geq|\beta_k^{\mu_k,\omega_k}|(1-\mu_k-\omega_k)\|\boldsymbol{g}_{k-1}\|^2,$$

which leads to

$$|\boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{y}_{k-1}|\geq|\beta_k^{\mu_k,\omega_k}|\gamma^2/2, \tag{C.58}$$

given (C.56).

Let us establish a similar result in the more complex case $\mu_k+\omega_k\in(1/2,1]$. As a preliminary step, let us show that

$$\boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{d}_k\leq-\gamma^2/2 \tag{C.59}$$

for all sufficiently large values of k.

According to Remark C.4.1, in the case $\boldsymbol{g}_{k-1}^{\mathrm{t}}\boldsymbol{d}_{k-1}=0$, we have $\beta_k^{\mu_k,\omega_k}=0$, so $\boldsymbol{d}_k=-\boldsymbol{g}_k$ and (C.59) is valid according to (C.56).

Now let us consider the case where $\boldsymbol{g}_{k-1}^{\mathrm{t}}\boldsymbol{d}_{k-1}<0$. Given (C.3) and (C.6), we have

$$\boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{c}_k=\boldsymbol{g}_k^{\mathrm{t}}(-\boldsymbol{g}_k+\beta_k^{\mu_k,\omega_k}\boldsymbol{d}_{k-1})=-\|\boldsymbol{g}_k\|^2+(\boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{y}_{k-1})(\boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{d}_{k-1})/D_k.$$

Given $\mu_k+\omega_k\in(1/2,1]$, we have $\omega_k+\mu_k c^{\min}\theta\lambda/\nu_2\geq m$, where $m=\min\{1/2,c^{\min}\theta\lambda/\nu_2\}$. As a consequence, (C.52) implies that $D_k\geq-m\,\boldsymbol{d}_{k-1}^{\mathrm{t}}\boldsymbol{g}_{k-1}$. Jointly with (C.48) and (C.56), the latter inequality yields

$$\boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{c}_k\leq-\gamma^2+|\boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{y}_{k-1}|\left(1+c_I^{\max}\theta\mu/\nu_1\right)/m.$$

Given (C.55), we deduce that $\boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{c}_k\leq-\gamma^2/2$ for all sufficiently large k. Because of (C.4), we can conclude that (C.59) holds.

Given (C.52) and (C.59), (C.57) implies

$$\begin{aligned}|\boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{y}_{k-1}| &\geq |\beta_k^{\mu_k,\omega_k}|\left((1-\mu_k-\omega_k)\gamma^2+(\omega_k+\mu_k c^{\min}\theta\lambda/\nu_2)\gamma^2/2\right)\\ &= |\beta_k^{\mu_k,\omega_k}|\left(1-\omega_k/2-(1-c^{\min}\theta\lambda/\nu_2)\mu_k\right)\gamma^2\end{aligned}$$

for all sufficiently large values of k. Given $\mu_k+\omega_k\in(1/2,1]$, the latter inequality implies

$$|\boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{y}_{k-1}| \geq |\beta_k^{\mu_k,\omega_k}|m\gamma^2. \tag{C.60}$$

Since $m=\min\{1/2, c^{\min}\theta\lambda/\nu_2\}\leq 1/2$, (C.60) is implied by (C.58), so that (C.60) holds in the whole domain $\mu_k\in[0,1]$, $\omega_k\in[0,1-\mu_k]$. Finally, (C.55) and (C.60) jointly imply $\lim_{k\to\infty}|\beta_k^{\mu_k,\omega_k}|=0$.

On the other hand, consider the case where Assumption 4 is not necessarily valid. If $\mu_k=0$, then we have

$$|\boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{y}_{k-1}| \geq |\beta_k^{0,\omega_k}|\gamma^2/2.$$

The proof is similar to that of (C.60), where the two cases to examine are $\omega_k\in[0,1/2]$ and $\omega_k\in(1/2,1]$. Finally, according to (C.55), we have $\lim_{k\to\infty}|\beta_k^{0,\omega_k}|=0$. □

Remark C.4.2. *The proof of Lemma C.4.4 is inspired from that of [Chen et Sun, 2002, Lemma 3.2], but we deal with the more general case of the iterated formula (C.8). Moreover, μ_k and ω_k are possibly varying, while they are constant parameters in [Chen et Sun, 2002].*

Theorem C.4.1. *Let $\boldsymbol{x}_k$ be defined by (C.2)-(C.8) with $\theta\in\,]0,2[$, and let Assumption 3 and Assumption 5 hold.*

Then we have convergence in the sense $\liminf_{k\to\infty}\boldsymbol{g}_k=\boldsymbol{0}$ for the PRP and LS methods, and more generally for $\mu_k=0$ and $\omega_k\in[0,1]$.

Moreover, if Assumption 4 holds, then we have also $\liminf_{k\to\infty}\boldsymbol{g}_k=\boldsymbol{0}$ in all cases.

Proof. Assume on the contrary that $\|\boldsymbol{g}_k\|\geq\gamma>0$ for all k. Since L is bounded, both $\{\boldsymbol{x}_k\}$ and $\{\boldsymbol{g}_k\}$ are bounded.

Let us first suppose that Assumption 4 holds. Since $\liminf_{k\to\infty}\|\boldsymbol{g}_k\|>0$, by Lemma C.4.4 we have $\lim_{k\to\infty}\beta_k^{\mu_k,\omega_k}=0$.

Since

$$\|\boldsymbol{d}_k\| = \|\boldsymbol{c}_k\| \leq \|\boldsymbol{g}_k\| + |\beta_k^{\mu_k,\omega_k}|\|\boldsymbol{d}_{k-1}\|,$$

we conclude that $\{\|\boldsymbol{d}_k\|\}$ is uniformly bounded for sufficient large k. Thus we have

$$\begin{aligned}|\boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{d}_k| &= |\boldsymbol{g}_k^{\mathrm{t}}(-\boldsymbol{g}_k+\beta_k^{\mu_k,\omega_k}\boldsymbol{d}_{k-1})|\\ &\geq \|\boldsymbol{g}_k\|^2 - |\beta_k^{\mu_k,\omega_k}|\|\boldsymbol{g}_k\|\|\boldsymbol{d}_{k-1}\|\\ &\geq \|\boldsymbol{g}_k\|^2/2\end{aligned}$$

for sufficient large k. Then there exists $\epsilon>0$ so that

$$\boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{d}_k/\|\boldsymbol{d}_k\|\|\boldsymbol{g}_k\| \geq \|\boldsymbol{g}_k\|/2\|\boldsymbol{d}_k\| \geq \epsilon$$

for sufficient large k. Finally, we conclude that

$$\sum_{k,\,\boldsymbol{d}_k\neq 0}\|\boldsymbol{g}_k\|^2\left(\boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{d}_k/\|\boldsymbol{d}_k\|\|\boldsymbol{g}_k\|\right)^2=\infty.$$

This is a contradiction to Lemma C.4.1.

The same proof applies to the case where $\mu_k=0$, and Assumption 4 is not necessarily valid. □

Remark C.4.3. *The proof of Theorem C.4.1 is partly inspired from that of [Chen et Sun, 2002, Theorem 3.3]. However, our result deals with variable parameters μ_k, ω_k. Moreover, Assumption 4 is not necessary in the case $\mu_k=0$, which contains the PRP and LS methods.*

C.5 Discussion

C.5.1 The convex quadratic case

Let us show that the convergence condition $\theta \in]0, \nu_1/\mu[$ is too restrictive for the stepsize formula proposed in [Sun et Zhang, 2001; Chen et Sun, 2002], in the case of a convex quadratic objective function. Let

$$\mathcal{Q}(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{t}}\boldsymbol{Q}\boldsymbol{x}/2 - \boldsymbol{b}^{\mathrm{t}}\boldsymbol{x}, \quad \boldsymbol{x} \in \mathbb{R}^n$$

where $\boldsymbol{Q}$ is a symmetric, positive definite matrix. Let ν_1 and ν_2 respectively denote the smallest and largest eigenvalue of $\boldsymbol{Q}$, so that (C.9) holds.

Now, consider the stepsize formula (C.10) with $\boldsymbol{Q}_k^0 = \boldsymbol{Q}$. When $\theta = 1$, it yields the optimal stepsize $\alpha_k = \arg\min_\alpha \mathcal{J}(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k)$.

In the convex quadratic case, Theorem C.4.1 ensures the convergence for $\theta \in]0, 2[$ and for any fixed $I > 0$. Remark that Assumption 5 is easily checked, since

$$\hat{\mathcal{Q}}(\boldsymbol{x}', \boldsymbol{x}) = \mathcal{Q}(\boldsymbol{x}) + (\boldsymbol{x}' - \boldsymbol{x})^{\mathrm{t}} \nabla\mathcal{Q}(\boldsymbol{x}) + (\boldsymbol{x}' - \boldsymbol{x})^{\mathrm{t}} \boldsymbol{Q} (\boldsymbol{x}' - \boldsymbol{x})/2 = \mathcal{Q}(\boldsymbol{x}').$$

In the case of the optimal stepsize, *i.e.,* $\theta = 1$, the classical linear CG algorithm is covered, since all conjugacy formulas reduce to the FR method.

On the contrary, the convergence domain deduced from [Sun et Zhang, 2001; Chen et Sun, 2002] is $\theta \in]0, \nu_1/\nu_2[\subset]0, 1[$, since $\nabla\mathcal{Q}$ is μ-$\mathcal{LC}^1$with μ not larger than ν_2. Hence, the convergence of the classical linear CG algorithm is not recovered. In particular, the condition $\theta \in]0, \nu_1/\nu_2[$ will produce excessively small and inefficient stepsizes when the Hessian matrix $\boldsymbol{Q}$ is ill-conditioned.

C.5.2 The general case

In the general case, the nontrivial computation of ν_1 and μ is a prerequisite to check the convergence condition $\theta \in]0, \nu_1/\mu[$. In [Chen et Sun, 2002], it is rather proposed to ensure the convergence empirically by choosing an arbitrarily small value of θ. Unfortunately, the resulting algorithm will be hardly competitive, compared to CG methods with a usual line search procedure.

Our convergence results do not share the same drawback, provided that, as a preliminary step, a convex quadratic function has been found to approximate the objective function from above. According to Lemma C.2.1, finding such a convex quadratic majorizing function is always possible when N is a convex set.

In practice, case-by-case considerations may provide tighter convex quadratic approximations, that will result in larger stepsizes. This issue is actually not new, since finding a good convex quadratic majorizing function is already a crucial step in the use of Weiszfeld's method [Voss et Eckhardt, 1980]. The latter reference provides examples in the field of optimal location (which was Weiszfeld's original concern), and in structural mechanics. Robust regression is another classical area where Weiszfeld's method is widely applied, under the name of *Iterative Reweighting* [Huber, 1981]. More recently, the latter has also become a standard approach for edge preserving image restoration, under the name of *Half-Quadratic Scheme* [Charbonnier *et al.*, 1997; Allain *et al.*, 2006].

Bibliographie

[Allain *et al.*, 2006] M. Allain, J. Idier et Y. Goussard. On global and local convergence of half-quadratic algorithms. *IEEE Trans. Image Processing*, 15 (5) : 1130–1142, mai 2006.

[Bertsekas, 1999] D. P. Bertsekas. *Nonlinear programming*. Athena Scientific, Belmont, MA, USA, 2ème edition, 1999.

[Charbonnier *et al.*, 1997] P. Charbonnier, L. Blanc-Féraud, G. Aubert et M. Barlaud. Deterministic edge-preserving regularization in computed imaging. *IEEE Trans. Image Processing*, 6 (2) : 298–311, février 1997.

[Chen et Sun, 2002] X. Chen et J. Sun. Global convergence of two-parameter family of conjugate gradient methods without line search. *Journal of Computational and Applied Mathematics*, 146 : 37–45, 2002.

[Dai et Yuan, 1999] Y. Dai et Y. Yuan. A nonlinear conjugate gradient method with a strong global convergence property. *SIAM J. Optimization*, 10 (1) : 177–182, 1999.

[Dai et Yuan, 2001] Y. Dai et Y. Yuan. A three-parameter family of nonlinear conjugate gradient methods. *Math. Comp.*, 70 : 1155–1167, 2001.

[Dixon, 1973] L. C. W. Dixon. Conjugate directions without linear searches. *IMA Journal of Applied Mathematics*, 11 : 317–328, 1973.

[Fletcher, 1987] R. Fletcher. *Practical Methods of Optimization*. Wiley&Sons, New York, USA, 2ème edition, 1987.

[Fletcher et Reeves, 1964] R. Fletcher et C. M. Reeves. Function minimization by conjugate gradients. *Comp. J.*, 7 : 149–157, 1964.

[Hestenes et Stiefel, 1952] M. R. Hestenes et E. Stiefel. Methods of conjugate gradients for solving linear system. *J. Res. Nat. Bur. Stand.*, 49 : 409–436, 1952.

[Huber, 1981] P. J. Huber. *Robust Statistics*. John Wiley, New York, NY, USA, 1981.

[Labat et Idier, 2005] C. Labat et J. Idier. Convergence of conjugate gradient methods with a closed-form stepsize formula. Tech. rep., IRCCyN RI2005_11, dec. 2005.

[Labat et Idier, 2007] C. Labat et J. Idier. Convergence of conjugate gradient methods with a closed-form stepsize formula. à paraître, *Journal of Optimisation Theory and Applications*, 2007.

[Liu et Storey, 1991] Y. Liu et C. Storey. Efficient generalized conjugate gradient algorithms, part 1 : Theory. *Journal of Optimisation Theory and Applications*, 69 : 129–137, 1991.

[Polak et Ribière, 1969] E. Polak et G. Ribière. Note sur la convergence de méthodes des directions conjuguées. *Rev. Française d'Informatique et de Recherche Opérationelle*, 16 : 35–43, 1969.

[Polyak, 1969] B. T. Polyak. The conjugate gradient method in extreme problems. *USSR Comp. Math. and Math. Phys.*, 9 : 94–112, 1969.

[Sun et Zhang, 2001] J. Sun et J. Zhang. Global convergence of conjugate gradient methods without line search. *Annals of Operations Research*, 103 : 161–173, 2001.

[Voss et Eckhardt, 1980] H. Voss et U. Eckhardt. Linear Convergence of Generalized Weiszfeld's Method. *Computing*, 25 : 243–251, 1980.

[Weiszfeld, 1937] E. Weiszfeld. Sur le point pour lequel la somme des distances de n points donnés est minimum. *Tôhoku Mathematical Journal*, 43 : 355–386, 1937.

[Wolfe, 1971] P. Wolfe. Convergence conditions for ascent methods. II : Some corrections. *SIAM Rev.*, 13 : 185–188, 1971.

Annexe D

RÉSULTATS EXPÉRIMENTAUX SUR L'INFLUENCE DES PARAMÈTRES θ ET $A_{\mathbf{GY}}$

	$\theta = 0.5$		$\theta = 0.75$		$\theta = 0.9$	
	Itérations	Temps (s)	Itérations	Temps (s)	Itérations	Temps (s)
GCPPR+GR1D(1)	42	81.7	28	54.6	25	<u>48.4</u>
GCPPR+GY1D(1)	56	108.8	40	78.1	33	64.2
	$\theta = 1.1$		$\theta = 1.25$		$\theta = 1.5$	
	Itérations	Temps (s)	Itérations	Temps (s)	Itérations	Temps (s)
GCPPR+GR1D(1)	25	48.9	31	60.7	75	146.1
GCPPR+GY1D(1)	27	<u>52.6</u>	27	52.7	63	123

Tab. D.1 – Influence du paramètre θ sur les algorithmes GCPPR+GR1D(1) et GCPPR+GY1D(1) avec $a_{\mathrm{GY}} = \hat{a}$ pour le problème de déconvolution d'image. Un préconditionnement CT est utilisé. Le meilleur temps (s) d'optimisation de chaque algorithme est souligné.

	$\theta = 0.5$				$\theta = 0.75$			
	pas de precond.		CT precond.		pas de precond.		CT precond.	
	Iter.	Temps (s)	Iter.	Temps (s)	Iter.	Temps (s)	Iter.	Temps (s)
GR+GCP(0.9)	33/3.6	246.8	39/1.2	171.3	25/4.1	208.4	30/1.3	137.1
GR+GCP(0.8)	32/3.7	242.7	28/1.5	140.1	23/4.5	207.4	27/1.4	130.9
GR+GCP(0.7)	32/3.7	244.7	26/1.6	136.4	22/4.8	206.7	20/1.8	111.4
GR+GCP(0.6)	31/3.9	249.7	25/1.6	134.6	21/5.5	220.8	17/2.0	103.9
GR+GCP(0.5)	31/5.6	331.1	25/1.6	133.4	20/7.0	260.0	16/2.2	104.5
GR+GCP(0.4)	30/7.0	386.3	24/1.9	142.8	19/8.1	277.0	15/2.4	104.4
GR+GCP(0.3)	29/8.1	424.8	23/3.0	188.3	18/9.4	301.5	15/2.8	117.2
GR+GCP(0.2)	28/9.2	456.7	23/3.5	213.7	17/10.9	324.0	14/3.8	138.1
GR+GCP(0.1)	27/12.3	578.2	22/4.3	241.1	16/13.9	381.8	13/4.8	157.6
	$\theta = 0.9$				$\theta = 1.1$			
	pas de precond.		CT precond.		pas de precond.		CT precond.	
	Iter.	Temps (s)	Iter.	Temps (s)	Iter.	Temps (s)	Iter.	Temps (s)
GR+GCP(0.9)	26/3.8	207.6	32/1.1	136.4	?/1	∞	60/1.0	240.0
GR+GCP(0.8)	23/4.3	201.0	26/1.4	127.6	38/3.3	265.4	30/1.2	132.4
GR+GCP(0.7)	20/5.0	195.4	20/1.8	111.9	23/4.3	198.4	22/1.5	109.6
GR+GCP(0.6)	17/6.2	197.6	17/1.9	102.0	20/5.2	202.8	17/1.8	96.2
GR+GCP(0.5)	17/7.1	221.1	14/2.4	96.2	16/6.5	194.7	15/2.2	98.5
GR+GCP(0.4)	15/8.3	224.0	13/2.6	95.5	15/7.4	204.1	13/2.3	88.7
GR+GCP(0.3)	15/9.9	262.3	12/3.2	104.4	13/9.0	210.3	11/2.8	86.2
GR+GCP(0.2)	14/11.9	288.8	11/4.0	113.7	12/11.0	230.9	10/3.2	86.5
GR+GCP(0.1)	13/15.6	343.4	11/5.2	141.1	11/14.4	270.3	9/4.6	103.6
	$\theta = 1.25$				$\theta = 1.5$			
	pas de precond.		CT precond.		pas de precond.		CT precond.	
	Iter.	Temps (s)	Iter.	Temps (s)	Iter.	Temps (s)	Iter.	Temps (s)
GR+GCP(0.9)	?/1	∞	61/1.0	242.3	?/1	∞	61/1.0	243.6
GR+GCP(0.8)	682/1.0	2339.3	46/1.0	187.1	682/1.0	2327	46/1.0	186.9
GR+GCP(0.7)	38/3.2	258.5	26/1.3	123.2	38/3.2	258.7	26/1.3	122.3
GR+GCP(0.6)	25/4.6	225.2	21/1.6	111.6	25/4.6	226.5	21/1.6	110.9
GR+GCP(0.5)	19/5.2	191.9	17/2.0	104.1	19/5.2	190.5	17/2.0	104.1
GR+GCP(0.4)	17/6.0	192.7	15/2.1	94.1	17/6.0	191.5	15/2.1	93.6
GR+GCP(0.3)	15/7.2	197.5	12/2.6	88.6	15/7.2	198.6	12/2.6	88.4
GR+GCP(0.2)	13/8.6	201.1	11/2.8	86.8	13/8.6	199.9	11/2.8	86.8
GR+GCP(0.1)	12/11.1	232.4	10/3.5	93.6	12/11.1	231.4	10/3.5	92.9

TAB. D.2 – Influence du paramètre θ sur les algorithmes GR+GCP(η) pour le problème de déconvolution d'image. Le meilleur temps (s) d'optimisation de chaque algorithme est souligné. Le symbole ∞ indique que nous avons délibérément arrêté l'algorithme car il était trop lent (temps supérieur à plusieurs heures).

	$\theta = 1$							
	$a_{\mathrm{GY}} = 0.9\hat{a}$				$a_{\mathrm{GY}} = \hat{a}$			
	pas de precond.		CT precond.		pas de precond.		CT precond.	
	Iter.	Temps (s)	Iter.	Temps (s)	Iter.	Temps (s)	Iter.	Temps (s)
GY+GCP(0.9)	60/2.8	322.1	59/1.0	183.9	55/2.8	300.7	59/1.0	185.3
GY+GCP(0.8)	53/3.5	344.9	39/1.3	143.7	44/3.3	275.4	38/1.2	136.8
GY+GCP(0.7)	40/3.6	272.9	33/1.6	145.0	34/3.9	244.0	30/1.6	132.4
GY+GCP(0.6)	33/4.2	253.4	31/1.7	146.1	31/4.3	241.1	27/1.8	128.8
GY+GCP(0.5)	32/4.5	263.2	31/2.1	166.2	29/4.7	246.3	28/2.0	148.6
GY+GCP(0.4)	32/5.7	322.1	32/2.4	196.0	30/6.1	323.6	29/2.4	175.5
GY+GCP(0.3)	31/6.9	373.9	31/2.7	205.7	29/7.3	364.5	28/2.7	186.7
GY+GCP(0.2)	31/7.8	413.8	35/3.0	257.3	28/8.2	390.4	32/3.0	233.5
GY+GCP(0.1)	30/10.8	548.0	29/3.7	252.5	27/11.3	510.7	27/3.7	233.7
	$a_{\mathrm{GY}} = 1.1\hat{a}$				$a_{\mathrm{GY}} = 1.25\hat{a}$			
	pas de precond.		CT precond.		pas de precond.		CT precond.	
	Iter.	Temps (s)	Iter.	Temps (s)	Iter.	Temps (s)	Iter.	Temps (s)
GY+GCP(0.9)	49/2.7	256.4	59/1.0	184.9	49/2.9	271.8	60/1.0	188.0
GY+GCP(0.8)	43/3.4	276.1	39/1.2	137.6	40/3.6	268.5	38/1.2	135.9
GY+GCP(0.7)	36/4.0	264.1	26/1.7	118.0	33/4.1	248.6	25/1.7	115.0
GY+GCP(0.6)	31/4.3	243.6	25/1.8	120.5	29/4.5	236.0	24/1.8	117.6
GY+GCP(0.5)	27/4.8	232.6	25/2.0	131.6	25/5.1	226.0	23/2.0	118.8
GY+GCP(0.4)	27/6.4	303.7	27/2.4	164.5	25/6.7	291.5	24/2.4	145.2
GY+GCP(0.3)	26/7.6	340.4	27/2.6	176.4	24/8.0	328.4	24/2.6	157.7
GY+GCP(0.2)	26/8.7	383.6	29/2.9	209.1	23/9.1	358.4	26/3.0	187.4
GY+GCP(0.1)	25/11.8	491.4	24/3.7	207.2	22/12.5	456.9	21/3.7	185.1
	$a_{\mathrm{GY}} = 1.5\hat{a}$				$a_{\mathrm{GY}} = 1.75\hat{a}$			
	pas de precond.		CT precond.		pas de precond.		CT precond.	
	Iter.	Temps (s)	Iter.	Temps (s)	Iter.	Temps (s)	Iter.	Temps (s)
GY+GCP(0.9)	47/2.7	246.6	61/1.0	190.5	43/2.8	231.1	62/1.0	194.7
GY+GCP(0.8)	41/3.7	278.8	41/1.1	140.7	44/3.6	299.2	45/1.1	153.7
GY+GCP(0.7)	36/4.1	273.4	27/1.6	120.3	40/4.0	293.1	33/1.5	135.0
GY+GCP(0.6)	32/4.6	264.9	26/1.8	128.3	38/4.4	305.7	31/1.8	147.6
GY+GCP(0.5)	23/5.7	233.6	22/2.0	115.7	34/4.9	300.1	27/2.0	139.0
GY+GCP(0.4)	22/7.3	276.7	21/2.4	126.9	21/7.5	269.2	24/2.2	138.4
GY+GCP(0.3)	21/8.5	306.7	20/2.9	140.0	19/8.9	287.7	22/2.6	143.1
GY+GCP(0.2)	20/9.8	329.8	22/3.0	159.7	18/10.4	318.1	21/3.0	153.5
GY+GCP(0.1)	19/13.4	422.8	18/3.9	164.7	17/14.3	401.2	17/3.9	157.8

TAB. D.3 – Influence du paramètre a_{GY} sur les algorithmes GY+GCP(η) avec $\theta = 1$ pour le problème de déconvolution d'image. Le meilleur temps (s) d'optimisation de chaque algorithme est souligné.

Annexe E

DÉCONVOLUTION AVEUGLE DE TRAINS D'IMPULSIONS ROBUSTE À L'AMBIGUÏTÉ DE DÉCALAGE TEMPOREL

[Labat et Idier, 2006] C. Labat et J. Idier, Sparse blind deconvolution accounting for time-shift ambiguity, in Proc. IEEE ICASSP, Toulouse, mai 2006, vol. III, pp. 616-619[1].

SPARSE BLIND DECONVOLUTION ACCOUNTING FOR TIME-SHIFT AMBIGUITY

Christian Labat, Jérôme Idier

IRCCyN (CNRS UMR 6597)
ECN, 1 rue de la Noë, BP 92101, F44321 Nantes Cedex 3, France
firstname.lastname@irccyn.ec-nantes.fr

ABSTRACT

Our contribution deals with blind deconvolution of sparse spike trains. More precisely, we examine the problem in the Markov chain Monte-Carlo (MCMC) framework, where the unknown spike train is modeled as a Bernoulli-Gaussian process. In this context, we point out that time-shift and scale ambiguities jeopardize the robustness of basic MCMC methods, in quite a similar manner to the label switching effect studied by Stephens (2000) in mixture model identification. Finally, we propose proper modifications of the MCMC approach, in the same spirit as Stephens' contribution.

1. INTRODUCTION

The problem of the restoration of sparse spike trains distorted by a linear system and additive noise arises in many fields such as seismic exploration [1–3], and non-destructive evaluation [4]. It is classically dealt with a discrete-time noisy convolution model for the observations:

$$\boldsymbol{z} = \boldsymbol{h} \star \boldsymbol{x} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{h}$ is the impulse response (IR) of the system (assumed finite here), $\boldsymbol{x}$ is the sparse spike train to be restored and $\boldsymbol{\epsilon}$ is a stationary white Gaussian noise.

The deconvolution problem is said *blind* when the IR $\boldsymbol{h}$ is unknown. In the present study, we will assume that parameters such as the noise variance are also unknown. It is clear however that some statistical information must be available, at least to distinguish the input signal from the IR. Here, we adopt a Bernoulli-Gaussian process (BG) for $\boldsymbol{x}$, following [1] and many posterior contributions such as [2–4]. Moreover, we adopt a Markov chain Monte-Carlo (MCMC) approach, akin to that of [2].

Blind deconvolution is a fundamentally information-deficient issue: since $\boldsymbol{h} \star \boldsymbol{x}$ is equal to $(\boldsymbol{f} \star \boldsymbol{h}) \star (\boldsymbol{f}^{-1} \star \boldsymbol{x})$ for any invertible filter $\boldsymbol{f}$, the solution of a blind deconvolution problem is not unique. Here, the BG prior helps to raise the main ambiguities, but there remain the following ones:

- Scale ambiguity: $\boldsymbol{h} \star \boldsymbol{x} = (a\boldsymbol{h}) \star (\boldsymbol{x}/a)$, $\forall a \neq 0$.
- Time-shift ambiguity: $\boldsymbol{h} \star \boldsymbol{x} = (d_\tau \star \boldsymbol{h}) \star (d_{-\tau} \star \boldsymbol{x})$, $\forall \tau \in \mathbb{Z}$, where d_τ is the time delay filter of τ samples.

Such ambiguities must be taken into account. Otherwise, classical estimators for $\boldsymbol{h}$ and $\boldsymbol{x}$ such as posterior expectations (as typically approximated by MCMC computations) become meaningless, as averaged quantities over the variations of a and τ.

Whereas the scale ambiguity is rather easily raised by an arbitrary scaling, the time-shift ambiguity is more difficult to handle within the MCMC framework. In [2], it is simply circumvented by constraining the maximum of $\boldsymbol{h}$ to a prescribed position. However, as discussed in Section 5, such a solution is not always satisfying (see also Section 7).

The potential effect of time-shifts can be compared to the *label-switching* effect dealt by Stephens in [5] in the case of mixture model identification. It is our goal here to analyze the time-shift effect and to compensate for it, in the same spirit as Stephens' contribution.

The formulation of the blind deconvolution problem is presented in Section 2. Section 3 introduces the fully Bayesian framework. A Gibbs sampling scheme quite similar to that of [2] is proposed in Section 4. Sections 5 and 6 contain our main contributions: Section 5 examinates the time-shift problem, and Section 6 proposes an hybrid Gibbs sampler to compensate for it. The proposed method is compared to the method of [2] in Section 7, and conclusive remarks are made in Section 8.

2. PROBLEM STATEMENT

Let $\boldsymbol{z} = \{z_1, z_2, \ldots, z_N\}$, $\boldsymbol{h} = \{h_0, h_1 \ldots, h_P\}$, and $\boldsymbol{x} = \{x_1, x_2 \ldots, x_M\}$. Here, we adopt a "zero boundary" condition: the input coefficients x_m are assumed to vanish for all $m < 1$ and $m > M$), so that $N = M + P$. The length P of the unknown wavelet is assumed to be available in the sequel.

3. BAYESIAN APPROACH

3.1. Prior laws

The unknown input signal $\boldsymbol{x}$ is modeled as a BG sequence:

$$q_m \sim \mathrm{Bi}(\lambda), \quad (x_m \,|\, q_m) \sim \mathcal{N}(0, q_m \sigma_1^2), \tag{2}$$

where $\mathrm{Bi}(\lambda)$ is the Bernoulli law of parameter λ, so that $p(q_m = 1) = \lambda$. For the sake of simple notations, "p" will indifferently denote probabilities, probability densities, and products of the two. Moreover, random variables will not be distinguished from their realizations. With such simplified notations, we have

$$\begin{aligned} p(\boldsymbol{q}, \boldsymbol{x} \,|\, \lambda, \sigma_1^2) &= p(\boldsymbol{q} \,|\, \lambda)\, p(\boldsymbol{x} \,|\, \boldsymbol{q}, \sigma_1^2) \\ &= \lambda^L (1-\lambda)^{M-L} \prod_{m, q_m = 1} g(x_m; \sigma_1^2) \prod_{m, q_m = 0} \delta(x_m), \end{aligned}$$

where $L = \sum_{m=1}^{M} q_m$ and $g(\cdot\,;\,\mathbf{A})$ stands for the zero-mean Gaussian density of covariance $\mathbf{A}$.

We also assume that $\boldsymbol{\epsilon}$ is a Gaussian white noise with unknown variance σ^2:

$$\boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0},\, \sigma^2 \mathbf{I}_N\right), \tag{3}$$

where $\mathbf{I}_N$ is the identity matrix of size N.

Within our fully Bayesian framework, prior laws are also needed for $\boldsymbol{h}$ and for the hyperparameters. Our choices are much similar to those found in [2]:

- $\boldsymbol{h}$ has a Gaussian prior of known variance:

$$\boldsymbol{h} \sim \mathcal{N}\left(\mathbf{0},\, \sigma_h^2 \mathbf{I}_{P+1}\right). \tag{4}$$

- The noise variance σ^2 follows an inverse gamma prior $\mathrm{IG}(\nu, \eta)$ with known parameters ν, $\eta > 0$.
- λ is assigned a beta law $\mathrm{Be}(a,\, b)$ with known parameters a, $b > 0$.

Finally, we assume that σ_1^2 is a known constant. For instance, we can take $\sigma_1^2 = 1$. This assumption is not restrictive because of the scale ambiguity

$$\boldsymbol{h} \star \boldsymbol{x} = (\pm\sigma_1^{-1}\boldsymbol{h}) \star (\pm\sigma_1 \boldsymbol{x}), \quad \forall \sigma_1 > 0.$$

3.2. Joint posterior law

Let $\boldsymbol{\theta} = (\boldsymbol{q}, \boldsymbol{x}, \boldsymbol{h}, \lambda, \sigma^2)$. Given our previous assumptions, the joint *posterior* probability reads [2]:

$$p(\boldsymbol{\theta} \,|\, \boldsymbol{z}) = g(\boldsymbol{z} - \boldsymbol{h} \star \boldsymbol{x}\,;\, \sigma^2 \mathbf{I}_N)\, p(\boldsymbol{\theta}) / p(\boldsymbol{z}), \tag{5}$$

with $p(\boldsymbol{\theta}) = p(\boldsymbol{q} \,|\, \lambda)\, p(\boldsymbol{x} \,|\, \boldsymbol{q})\, p(\lambda)\, p(\boldsymbol{h})\, p(\sigma^2)$.

4. GIBBS SAMPLING

Using conditional posterior laws, the Gibbs sampler generates a Markov chain of random samples $\boldsymbol{\theta}^{(k)}$ whose equilibrium law coincides with the joint posterior law [6]. For the sake of simplicity, the iteration number k is omitted. The following conditional posterior laws can be deduced from (5) (see also [2]):

① Let $\boldsymbol{r} = \{\boldsymbol{z}, \boldsymbol{\theta}\} \setminus \{q_m, x_m\}$. Then, $(x_m \,|\, \boldsymbol{r})$ follows a BG law with a nonzero mean:

$$(q_m \,|\, \boldsymbol{r}) \sim \mathrm{Bi}(\lambda_{1,m}), \quad (x_m \,|\, q_m, \boldsymbol{r}) \sim \mathcal{N}(\mu_{1,m}, q_m \sigma_{1,m}^2)$$

with

$$\lambda_{1,m} = \frac{\widetilde{\lambda}_{1,m}}{\widetilde{\lambda}_{1,m} + 1 - \lambda}, \quad \widetilde{\lambda}_{1,m} = \lambda \frac{\sigma_{1,m}}{\sigma_1} \exp\left(\frac{\mu_{1,m}^2}{2\sigma_{1,m}^2}\right),$$

$$\sigma_{1,m}^2 = \frac{\sigma^2 \sigma_1^2}{\sigma^2 + \sigma_1^2 \|\boldsymbol{h}\|^2}, \quad \mu_{1,m} = \frac{\sigma_{1,m}^2}{\sigma^2} \sum_{i=0}^{P} h_i e_{m+i},$$

where $e_n = (\boldsymbol{z} - \boldsymbol{h} \star \boldsymbol{x})_n + h_{n-m} x_m$.

② Let $\boldsymbol{r} = \{\boldsymbol{z}, \boldsymbol{\theta}\} \setminus \{\boldsymbol{h}\}$. Then $(\boldsymbol{h} \,|\, \boldsymbol{r}) \sim \mathcal{N}(\boldsymbol{m}, \mathbf{R})$ with

$$\mathbf{R} = \left(\sigma^{-2} \mathbf{X}^{\mathrm{t}} \mathbf{X} + \sigma_h^{-2} \mathbf{I}_{P+1}\right)^{-1}, \quad \boldsymbol{m} = \sigma^{-2} \mathbf{R} \mathbf{X}^{\mathrm{t}} \boldsymbol{z}, \tag{6}$$

where $\mathbf{X}$ is the Toeplitz matrix of size $N \times (P+1)$ with first row $[x_1\ \mathbf{0}_P]$ and first column $\left[\boldsymbol{x}^{\mathrm{t}}\ \mathbf{0}_P\right]^{\mathrm{t}}$.

③ Conditionally to all other variables, the law of σ^2 is $\mathrm{IG}(N/2 + \nu,\, \|\boldsymbol{z} - \boldsymbol{h} \star \boldsymbol{x}\|^2/2 + \eta)$.

④ Conditionally to all other variables, the law of λ is $\mathrm{Be}(a + L,\, b + M - L)$, with $L = \sum_{m=1}^{M} q_m$.

The Gibbs sampler iterates steps ①-④, all of which corresponding to classical sampling operations.

5. DEALING WITH TIME-SHIFT AND SCALE AMBIGUITIES

5.1. Principle

In [2], it is proposed to raise the time-shift ambiguity by constraining the maximizer of $\boldsymbol{h}$ to a given position i^*:

$$\text{Time constraint [2]:} \qquad |h_{i^*}| = \max_i |h_i| \tag{7}$$

In practice, the sampling procedure is only slightly modified: Step ② is repeated until constraint (7) is met. The following constraint is also introduced to deal with scale ambiguity:

$$\text{Scale constraint [2]:} \qquad h_0 = 1. \tag{8}$$

The resulting posterior mean estimates can then be approximated by averages over the last $K - D$ samples:

$$\hat{\boldsymbol{\theta}} = \frac{1}{K - D} \sum_{k=D+1}^{K} \boldsymbol{\theta}^{(k)}$$

Enforcing conditions (7)-(8) is typical of the *identifiability constraint* approach, whose limitations are underlined in [5] in the context of label-switching.

In particular, the scale constraint $h_0 = 1$ is not always appropriate: if the true value of h_0 vanishes (or nearly so), such a constraint will be artificial and unsuited to impose a common dynamic to the samples $(\boldsymbol{h}^{(k)})$.

Similarly, condition (7) will be ineffective when the maximum magnitude $\max_i |h_i|$ of the true response is reached at several positions. There is another more specific drawback in imposing (7): a slight error in positioning i^* may yield severely degraded estimation results, as checked in Section 7. The reason is the following: in contrast with the perfect label-switching effect, where all label permutations are equally likely, different values of time-shifts do not yield perfectly equivalent solutions, because both $\boldsymbol{h}$ and $\boldsymbol{x}$ are defined as *finite length vectors*. In particular, the shape of the IR will no longer fit the time window $\{0, 1, \ldots, P\}$ after an arbitrary time-shift.

Following [5], we propose to get rid of constraints on the sampler. We rather cope with time ambiguities by shifting the samples $(\boldsymbol{h}^{(k)}, \boldsymbol{x}^{(k)})$ w.r.t. the time index, prior to computing averages. Scale ambiguities are dealt simultaneously, in the same spirit.

5.2. Scaling-shifting algorithm

Once a series of random samples $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(K)}$ is available, the question is to compute appropriate averages to estimate the unknown quantities. Usual averages are suited for scalar parameters λ and σ^2. Only the series $(\boldsymbol{h}^{(k)})$ and $(\boldsymbol{x}^{(k)})$ may be affected by time-shifts and scale fluctuations.

To compensate for label switching in mixture model identification, Stephens proposes a relabelling algorithm. The principle is to find the best permutation for each sample, in the sense of a well-chosen loss function to be minimized [5].

In the same spirit, we propose a scaling-shifting algorithm to remove the time and scale ambiguities in the series $(\boldsymbol{x}^{(k)})$ and $(\boldsymbol{h}^{(k)})$. Let the following loss function:

$$\mathcal{R}(\boldsymbol{h},\, \boldsymbol{a},\, \boldsymbol{\tau}) = \frac{1}{K-D} \sum_{k=D+1}^{K} \|\boldsymbol{h} - a_k\, d_{\tau_k} \star \boldsymbol{h}^{(k)}\|^2,$$

to be minimized w.r.t. $\boldsymbol{h}$, $\boldsymbol{a}$, $\boldsymbol{\tau}$. We are led to a combinatorial problem, for which we propose a suboptimal solution inspired from [5, Algorithm 4.1]. Each iteration of the following two-step scheme reduces $\mathcal{R}(\boldsymbol{h},\, \boldsymbol{a},\, \boldsymbol{\tau})$ until a fixed point is reached:

1. For all $k = D+1, \ldots, K$, choose $(\tilde{a}_k,\, \tilde{\tau}_k)$ as the minimizer of $\|\tilde{\boldsymbol{h}} - a_k\, d_{\tau_k} \star \boldsymbol{h}^{(k)}\|^2$ when $\tilde{\boldsymbol{h}}$ is held constant.
2. Choose $\tilde{\boldsymbol{h}}$ as the minimizer of $\mathcal{R}(\boldsymbol{h},\, \tilde{\boldsymbol{a}},\, \tilde{\boldsymbol{\tau}})$, *i.e.,*

$$\tilde{\boldsymbol{h}} = \frac{1}{K-D} \sum_{k=D+1}^{K} \tilde{a}_k\, d_{\tilde{\tau}_k} \star \boldsymbol{h}^{(k)}.$$

Step 1 yields $(\tilde{a}_k, \tilde{\tau}_k)$ as a matched filtering solution with reference to $\tilde{\boldsymbol{h}}$, while Step 2 updates $\tilde{\boldsymbol{h}}$ as the average of the scaled, shifted versions of $\boldsymbol{h}^{(k)}$. We propose to initialize the procedure by $\tilde{\boldsymbol{h}} = \boldsymbol{h}^{(K)}$. Convergence is observed after a few iterations.

The procedure applies to $(\boldsymbol{h}^{(k)})$, since our loss function is defined as a function of $\boldsymbol{h}$ only, but it also provides corrections for $(\boldsymbol{x}^{(k)})$, from which an estimated input signal is computed:

$$\tilde{\boldsymbol{x}} = \frac{1}{K-D} \sum_{k=D+1}^{K} \frac{1}{\tilde{a}_k}\, d_{-\tilde{\tau}_k} \star \boldsymbol{x}^{(k)}. \tag{9}$$

6. HYBRID GIBBS SAMPLING

6.1. Metropolis-Hastings within Gibbs sampling

In practice, removing constraints (7)-(8) in favor of our scaling-shifting scheme is not sufficient to provide a fully satisfying procedure. There is still an issue to deal with: the Gibbs sampling procedure of Section 4 scarcely produces any time-shifts. As a consequence, the estimates produced after the scaling-shifting procedure will be strongly influenced by the initialization of the Gibbs procedure: typically, the position of the maximum value of $\boldsymbol{h}^{(0)}$ nearly plays the role of i^* in constraint (7).

It is actually quite easy to cope with such a deficiency by slightly modifying the Gibbs sampler of Section 4 in order to stimulate time-shifts. Our proposition only affects Step ②. Instead of merely resampling $\boldsymbol{h}$ according to its posterior law, we envisage a Metropolis-Hastings (MH) procedure involving both $\boldsymbol{h}$ and $(\boldsymbol{q}, \boldsymbol{x})$. To simplify notations in the whole section, we drop dependence of probability terms on current parameter values $\lambda^{(k)}$, $\sigma^{(k)}$, and we use $\boldsymbol{\theta}$ as a shorthand for $(\boldsymbol{q}, \boldsymbol{x}, \boldsymbol{h})$. Let us define the proposal kernel of the MH step as $\pi(\boldsymbol{\theta}' \,|\, \boldsymbol{z}, \boldsymbol{\theta}) = p(\boldsymbol{h}' \,|\, \boldsymbol{q}', \boldsymbol{x}', \boldsymbol{z})\, \pi(\boldsymbol{q}', \boldsymbol{x}' \,|\, \boldsymbol{q}, \boldsymbol{x})$, where

$$\pi(\boldsymbol{q}', \boldsymbol{x}' \,|\, \boldsymbol{q}, \boldsymbol{x}) = \begin{cases} 1-2\alpha & \text{if } (\boldsymbol{q}', \boldsymbol{x}') = (\boldsymbol{q}, \boldsymbol{x}), \\ \alpha & \text{if } (\boldsymbol{q}', \boldsymbol{x}') = (\mathbf{C}\boldsymbol{q}, \mathbf{C}^{-1}\boldsymbol{x}), \\ \alpha & \text{if } (\boldsymbol{q}', \boldsymbol{x}') = (\mathbf{C}^{-1}\boldsymbol{q}, \mathbf{C}\boldsymbol{x}), \end{cases}$$

with $\mathbf{C}$ defined as the right circular shift operator, and $\alpha \in (0, 1/2)$. With probability $1-2\alpha$, the MH procedure boils down to Step ②. Otherwise, a time-shift of ± 1 is proposed. Circular shifting provides a good trade-off between easy implementation and a fair acceptation probability. According to our "MH within Gibbs" sampler, Step ② is replaced by:

2 Propose $\boldsymbol{\theta}'$ according to $\pi(\boldsymbol{\theta}' \,|\, \boldsymbol{z}, \boldsymbol{\theta}^{(k)})$. Accept $\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}'$ with probability $\min\{1, \rho(\boldsymbol{z}, \boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}')\}$, where

$$\rho(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\theta}') = \frac{\pi(\boldsymbol{\theta} \,|\, \boldsymbol{z}, \boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}' \,|\, \boldsymbol{z}, \boldsymbol{\theta})} \frac{p(\boldsymbol{\theta}' \,|\, \boldsymbol{z})}{p(\boldsymbol{\theta} \,|\, \boldsymbol{z})}. \tag{10}$$

Otherwise, let $\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)}$.

Let us establish a simple expression for $\rho(\boldsymbol{z}, \boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}')$.

6.2. Acceptation probability

According to $\pi(\boldsymbol{q}', \boldsymbol{x}' \,|\, \boldsymbol{q}, \boldsymbol{x}) = \pi(\boldsymbol{q}, \boldsymbol{x} \,|\, \boldsymbol{q}', \boldsymbol{x}')$, (10) reads

$$\begin{aligned} \rho(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\theta}') &= \frac{p(\boldsymbol{h} \,|\, \boldsymbol{q}, \boldsymbol{x}, \boldsymbol{z})}{p(\boldsymbol{h}' \,|\, \boldsymbol{q}', \boldsymbol{x}', \boldsymbol{z})} \frac{p(\boldsymbol{h}' \,|\, \boldsymbol{q}', \boldsymbol{x}', \boldsymbol{z})}{p(\boldsymbol{h} \,|\, \boldsymbol{q}, \boldsymbol{x}, \boldsymbol{z})} \frac{p(\boldsymbol{q}', \boldsymbol{x}' \,|\, \boldsymbol{z})}{p(\boldsymbol{q}, \boldsymbol{x} \,|\, \boldsymbol{z})} \\ &= \frac{p(\boldsymbol{q}', \boldsymbol{x}' \,|\, \boldsymbol{z})}{p(\boldsymbol{q}, \boldsymbol{x} \,|\, \boldsymbol{z})}. \end{aligned}$$

Then, according to $p(\boldsymbol{q}', \boldsymbol{x}') = p(\boldsymbol{q}, \boldsymbol{x})$, (10) also reads

$$\rho(\boldsymbol{\theta}, \boldsymbol{\theta}', \boldsymbol{z}) = \frac{p(\boldsymbol{z} \,|\, \boldsymbol{q}', \boldsymbol{x}')}{p(\boldsymbol{z} \,|\, \boldsymbol{q}, \boldsymbol{x})} = \frac{p(\boldsymbol{z} \,|\, \boldsymbol{x}')}{p(\boldsymbol{z} \,|\, \boldsymbol{x})}, \tag{11}$$

which shows that $\rho(\boldsymbol{\theta}, \boldsymbol{\theta}', \boldsymbol{z})$ depends neither on $\boldsymbol{h}$ nor on $\boldsymbol{h}'$. Moreover, $\boldsymbol{x}' = \boldsymbol{x}$ implies $\rho(\boldsymbol{\theta}, \boldsymbol{\theta}', \boldsymbol{z}) = 1$, as expected.

In all cases, it is easy to establish that $\rho(\boldsymbol{\theta}, \boldsymbol{\theta}', \boldsymbol{z})$ is the ratio of two Gaussian densities. More precisely, it can be deduced from (1), (4) and (3) that $(\boldsymbol{z} \,|\, \boldsymbol{x}) \sim \mathcal{N}(\mathbf{0},\, \mathbf{P}^{-1})$, with

$$\mathbf{P} = (\sigma^2 \mathbf{I}_N + \sigma_h^2 \mathbf{X}\mathbf{X}^{\mathrm{t}})^{-1} = \sigma^{-2}\mathbf{I}_N - \sigma^{-4}\mathbf{X}\mathbf{R}\mathbf{X}^{\mathrm{t}}.$$

given the matrix inversion lemma. Hence,

$$\begin{aligned} 2\ln \rho(\boldsymbol{\theta}, \boldsymbol{\theta}', \boldsymbol{z}) &= \boldsymbol{z}^{\mathrm{t}}(\mathbf{P} - \mathbf{P}')\boldsymbol{z} + \ln|\mathbf{P}^{-1}\mathbf{P}'| \\ &= (\boldsymbol{m}')^{\mathrm{t}}(\mathbf{R}')^{-1}\boldsymbol{m}' - \boldsymbol{m}^{\mathrm{t}}\mathbf{R}^{-1}\boldsymbol{m} + \ln|\mathbf{R}^{-1}\mathbf{R}'|, \end{aligned}$$

where we make use of the matrix inversion lemma again. The latter expression can be evaluated in $O(P^2)$ operations given that matrices $\mathbf{R}$, $\mathbf{R}'$ are Toeplitz.

7. SIMULATION RESULTS

A simulation is proposed to compare the original MCMC method of [2] with our modified approach incorporating the MH step and the scaling-shifting procedure. It is based on an example found in [2]. An IR of order $P = 40$ is defined by $\boldsymbol{h} = \boldsymbol{h}^*/h_0^*$, with $h_i^* = \sin(\pi(i+1)/6.4)\exp(-0.12\,|i-9|)$. The input signal $\boldsymbol{x}$ is generated from a BG law (2) with $\lambda = 0.1$ and $\sigma_1^2 = 4$. The observed signal is obtained from (1) for $N = 2000$ and $\sigma^2 = 1$, which corresponds to a signal-to-noise ratio of 18.6dB. The parameter values for the priors are taken from [2]:

$$\boldsymbol{h} \sim \mathcal{N}\left(\mathbf{0},\, 100\,\mathbf{I}_{P+1}\right), \quad \lambda \sim \mathrm{Be}(10, 50), \quad \sigma^2 \sim \mathrm{IG}(1,\, 0.3),$$

as well as the initial values: $\sigma^{(0)} = 1$, $\boldsymbol{x}^{(0)} = \mathbf{0}$, $\lambda^{(0)} = 0.1$. Both sampling schemes were carried out for 4000 iterations,

and only the last 1000 samples were considered to compute estimates.

The robustness of the method presented in [2] is tested by assuming a slight error in the position of the maximizer $i^* = \arg\max_i |h_i|$: $i^* = 6$ is considered instead of $i^* = 9$. Accordingly, $\boldsymbol{h}^{(0)}$ is chosen as a Kronecker function at $i = 6$, instead of $i = 9$. Figure 1(a) shows that the left part of the resulting estimate $\hat{\boldsymbol{h}}$ is slightly altered, but the consequence on the estimated input $\hat{\boldsymbol{x}}$ is a more severe degradation as shown in Figure 1(b). The corresponding estimated value of λ is $\hat{\lambda} = 0.18$, which significantly differs from the true $\lambda = 0.1$, while the noise variance is rather well estimated $\hat{\sigma}^2 = 1.13$.

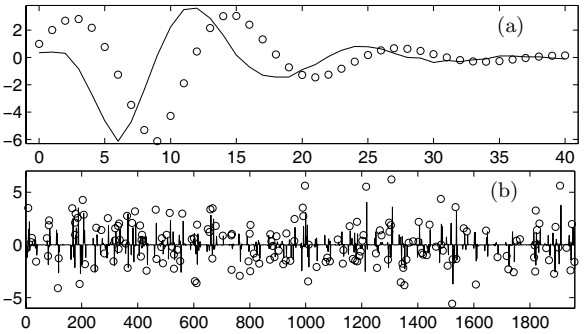


Fig. 1. Application of the standard MCMC method found in [2], tested with an error of -3 time samples on the position of $i^* = \arg\max_i |h_i|$. The circles indicate true values. (a) Estimated IR $\hat{\boldsymbol{h}}$ after proper scaling. (b) Estimated BG $\hat{\boldsymbol{x}}$ signal after proper scaling.

When tested in the same conditions, our modified MCMC method clearly shows a better robustness according to Figure 2. The estimated value of λ is now $\tilde{\lambda} = 0.099$ while the noise variance is still well estimated $\tilde{\sigma}^2 = 1.07$. Table 1 provides additional information about the frequency of accepted shifts within the MH step of Subsection 6.1. During the first 200 iterations, nearly 20% of right shifts are accepted, because the algorithm compensates for the initialization of the IR as a Kronecker function at the left of the time domain. Then the proportion of accepted left and right shifts gets balanced, at about 7%. After the burn-in, the proportion is below 4%, which does not mean that our MH procedure is not efficient, but rather that the correct position of the IR is significantly more probable than the others.

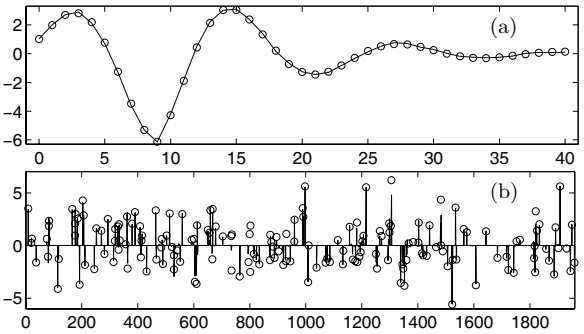


Fig. 2. Proposed modified MCMC method in the same conditions as in Figure 1, with $\alpha = 0.1$. The circles indicate true values. (a) Estimated IR $\tilde{\boldsymbol{h}}$ after proper scaling. (b) Estimated BG $\tilde{\boldsymbol{x}}$ signal after proper scaling.

Sample index $k \in$	$[1, 200]$	$[201, 3000]$	$[3001, 4000]$
# of proposed right shifts	22	267	76
# of accepted right shifts	4	20	3
# of proposed left shifts	9	277	98
# of accepted left shifts	1	19	3

Table 1. Number of proposed and accepted shifts at different stages of the MH algorithm.

8. DISCUSSION

In this paper, we have pointed out that time and scale ambiguities jeopardize the robustness of basic MCMC methods applied to sparse blind deconvolution problem. We have established a formal link between this issue and the label switching effect studied by Stephens in [5]. We have proposed to introduce some modifications in the light of Stephens' contribution. The proposed method is based on an "MH within Gibbs" sampler, and estimation of the unknowns are only obtained after an operation of scaling-shifting on the generated samples. The additional cost is negligible compared to a more standard application of the MCMC approach, as found in [2], while it is no more necessary to assume that the position of the maximum value of the IR is known in advance.

Our main perspective is to consider an IR $\boldsymbol{h}$ of unknown length. In our opinion, the present contribution is a prerequisite step towards estimating the length of $\boldsymbol{h}$, since it provides a robust way of letting the IR make best use of the alloted time window. Different window lengths could then be explored, for instance using a reversible jump procedure.

9. REFERENCES

[1] J. M. Mendel, *Optimal Seismic Deconvolution*, Academic Press, New York, NY, 1983.

[2] Q. Cheng, R. Chen, and T.-H. Li, "Simultaneous wavelet estimation and deconvolution of reflection seismic signals", *IEEE Trans. Geosci. Remote Sensing*, vol. 34, pp. 377–384, Mar. 1996.

[3] O. Rosec, J. Boucher, B. Nziri, and T. Chonavel, "Blind marine seismic deconvolution using statistical mcmc methods", *IEEE oceanic engineering*, vol. 28, pp. 502–512, July 2003.

[4] F. Champagnat and J. Idier, "Deconvolution of sparse spike trains accounting for wavelet phase shifts and colored noise", in *Proc. IEEE ICASSP*, Minneapolis, MN, 1993, pp. 452–455.

[5] M. Stephens, "Dealing with label-switching in mixture models", *J. R. Statist. Soc. B*, vol. 62, pp. 795–809, 2000.

[6] C. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer Texts in Statistics. Springer Verlag, New York, NY, 2nd edition, 2004.

RÉFÉRENCES BIBLIOGRAPHIQUES

[Abeyratne *et al.*, 1997] U. R. Abeyratne, A. Petropulu, T. Golas, J. Reid, E. Conant et F. Forsberg. Higher-order vs. second-order statistics in ultrasound image deconvolution. *IEEE Trans. Ultrasonics Ferroelectrics Frequency Control*, 44 (6) : 1409–1416, novembre 1997.

[Abeyratne *et al.*, 1995] U. R. Abeyratne, A. Petropulu et J. Reid. Higher-order spectra based deconvolution of ultrasound images. *IEEE Trans. Ultrasonics Ferroelectrics Frequency Control*, 42 (6) : 1064–1075, novembre 1995.

[Al-Baali et Fletcher, 1996] M. Al-Baali et R. Fletcher. On the order of convergence of preconditioned nonlinear conjugate gradient methods. *SIAM J. Sci. Comput.*, 17 : 658–665, 1996.

[Allain, 2002] M. Allain. *Approche pénalisée en tomographie hélicoïdale. Application à la conception d'une prothèse personnalisée du genou.* thèse de doctorat en cotutelle, Université de Paris-Sud, Orsay, France / Ecole Polytechnique de Montréal, Québec, Canada, décembre 2002.

[Allain *et al.*, 2006] M. Allain, J. Idier et Y. Goussard. On global and local convergence of half-quadratic algorithms. *IEEE Trans. Image Processing*, 15 (5) : 1130–1142, mai 2006.

[Antoniadis et Fan, 2001] A. Antoniadis et J. Fan. Regularization of wavelet approximations. *J. Amer. Statist. Assoc.*, 96 (455) : 939–967, 2001.

[Antonini *et al.*, 1992] M. Antonini, M. Barlaud, P. Mathieu et I. Daubechies. Image coding using wavelet transform. *IEEE Trans. Image Processing*, 1 (2) : 205–220, 1992.

[Aubert et Kornprobst, 2002] G. Aubert et P. Kornprobst. *Mathematical problems in images processing.* Springer-Verlag, Berlin, 2002.

[Aubert et Vese, 1997] G. Aubert et L. Vese. A variational method in image recovery. *SIAM J. Num. Anal.*, 34 (5) : 1948–1979, octobre 1997.

[Belge *et al.*, 2000] M. Belge, M. Kilmer et E. Miller. Wavelet domain image restoration with adaptive edge-preserving regularization. *IEEE Trans. Image Processing*, 9 (4) : 597–608, avril 2000.

[Bertsekas, 1999] D. P. Bertsekas. *Nonlinear programming.* Athena Scientific, Belmont, MA, USA, 2ème edition, 1999.

[Besag *et al.*, 1995] J. Besag, P. Green, D. Higdon et K. Mengersen. Bayesian computation and stochastic systems. *Statistical Science*, 10 : 3–66, 1995.

[Black et Rangarajan, 1996] M. J. Black et A. Rangarajan. On the unification of line processes, outlier rejection, and robust statistics with applications in early vision. *Int. J. Computer Vision*, 19 (1) : 57–91, 1996.

[Blake, 1989] A. Blake. Comparison of the efficiency of deterministic and stochastic algorithms for visual reconstruction. *IEEE Trans. Pattern Anal. Mach. Intell.*, PAMI-11 (1) : 2–12, janvier 1989.

[Blake et Zisserman, 1987] A. Blake et A. Zisserman. *Visual reconstruction*. The MIT Press, Cambridge, MA, USA, 1987.

[Blanc-Féraud *et al.*, 1995] L. Blanc-Féraud, P. Charbonnier, G. Aubert et M. Barlaud. Nonlinear image processing : Modelling and fast algorithm for regularisation with edge detection. In *Proc. IEEE ICIP*, volume 2, pages 474–477, 1995.

[Bouman et Sauer, 1993] C. A. Bouman et K. D. Sauer. A generalized Gaussian image model for edge-preserving MAP estimation. *IEEE Trans. Image Processing*, 2 (3) : 296–310, juillet 1993.

[Bouman et Sauer, 1996] C. A. Bouman et K. D. Sauer. A unified approach to statistical tomography using coordinate descent optimization. *IEEE Trans. Image Processing*, 5 (3) : 480–492, mars 1996.

[Brette et Idier, 1996] S. Brette et J. Idier. Optimized single site update algorithms for image deblurring. In *Proc. IEEE ICIP*, pages 65–68, Lausanne, Suisse, septembre 1996.

[Brémaud, 1998] P. Brémaud. *Markov chains. Gibbs fields and Monte Carlo*. Cours ENSTA, Paris, 1998.

[Brémaud, 1999] P. Brémaud. *Markov Chains. Gibbs fields, Monte Carlo Simulation, and Queues*. Texts in Applied Mathematics 31. Springer, New York, NY, USA, 1999.

[Chalmond *et al.*, 2003] B. Chalmond, F. Coldefy, E. Goubet et B. Lavayssière. Coherent 3-D echo detection for utrasonic imaging. *IEEE Trans. Signal Processing*, 51 (3) : 592–601, 2003.

[Champagnat et Idier, 1993] F. Champagnat et J. Idier. Deconvolution of sparse spike trains accounting for wavelet phase shifts and colored noise. In *Proc. IEEE ICASSP*, pages 452–455, Minneapolis, MN, USA, 1993.

[Champagnat et Idier, 2004] F. Champagnat et J. Idier. A connection between half-quadratic criteria and EM algorithms. *IEEE Signal Processing Letters*, 11 (9) : 709–712, septembre 2004.

[Chan et Ng, 1996] R. H. Chan et M. K. Ng. Conjugate gradient methods for Toeplitz systems. *SIAM Rev.*, 38 (3) : 427–482, septembre 1996.

[Chan et Mulet, 1999] T. F. Chan et P. Mulet. On the convergence of the lagged diffusivity fixed point method in total variation image restoration. *SIAM J. Num. Anal.*, 36 (2) : 354–367, 1999.

[Charbonnier, 1994] P. Charbonnier. *Reconstruction d'image : régularisation avec prise en compte des discontinuités*. thèse de doctorat, Université de Nice-Sophia Antipolis, Nice, septembre 1994.

[Charbonnier *et al.*, 1994] P. Charbonnier, L. Blanc-Féraud, G. Aubert et M. Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. In *Proc. IEEE ICIP*, volume 2, pages 168–172, Austin, TX, USA, novembre 1994.

[Charbonnier *et al.*, 1997] P. Charbonnier, L. Blanc-Féraud, G. Aubert et M. Barlaud. Deterministic edge-preserving regularization in computed imaging. *IEEE Trans. Image Processing*, 6 (2) : 298–311, février 1997.

[Chen et Sun, 2002] X. Chen et J. Sun. Global convergence of two-parameter family of conjugate gradient methods without line search. *Journal of Computational and Applied Mathematics*, 146 : 37–45, 2002.

[Cipra, 2000] B. Cipra. The best of the 20th century : Editors name top 10 algorithms. *SIAM News*, 33 (4) : 1, mai 2000.

[Ciuciu et Idier, 2002] P. Ciuciu et J. Idier. A half-quadratic block-coordinate descent method for spectral estimation. *Signal Processing*, 82 (7) : 941–959, juillet 2002.

[Cooley et Tukey, 1965] J. W. Cooley et J. W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Math. Comp.*, 19 (90) : 297–301, 1965.

[Dai et Yuan, 1999] Y. Dai et Y. Yuan. A nonlinear conjugate gradient method with a strong global convergence property. *SIAM J. Optimization*, 10 (1) : 177–182, 1999.

[Dai et Yuan, 2001] Y. Dai et Y. Yuan. A three-parameter family of nonlinear conjugate gradient methods. *Math. Comp.*, 70 : 1155–1167, 2001.

[Delaney et Bresler, 1998] A. H. Delaney et Y. Bresler. Globally convergent edge-preserving regularized reconstruction : an application to limited-angle tomography. *IEEE Trans. Image Processing*, 7 (2) : 204–221, février 1998.

[Demirli et Saniie, 2001a] R. Demirli et J. Saniie. Model-based estimation of ultrasonic echoes. Part I : Analysis and algorithms. *IEEE Trans. Ultrasonics Ferroelectrics Frequency Control*, 48 (3) : 787–802, mai 2001.

[Demirli et Saniie, 2001b] R. Demirli et J. Saniie. Model-based estimation of ultrasonic echoes. Part II : Nondestructive evaluation applications. *IEEE Trans. Ultrasonics Ferroelectrics Frequency Control*, 48 (3) : 803–811, mai 2001.

[Demoment *et al.*, 1984] G. Demoment, R. Reynaud et A. Herment. Range resolution improvement by a fast deconvolution method. *Ultrasonic Imaging*, 6 : 435–451, 1984.

[Descombes *et al.*, 1999] X. Descombes, M. Sigelle et F. Preteux. Estimating Gaussian Markov random field parameters in a nonstationnary framework : Application to remote sensing imaging. *IEEE Trans. Image Processing*, 8 : 490–503, avril 1999.

[Dixon, 1973] L. C. W. Dixon. Conjugate directions without linear searches. *IMA Journal of Applied Mathematics*, 11 : 317–328, 1973.

[Erdogan et Fessler, 1999] H. Erdogan et J. Fessler. Monotonic algorithms for transmission tomography. *IEEE Trans. Medical Imaging*, 18 (9) : 801–814, septembre 1999.

[Fatemi et Kak, 1980] M. Fatemi et A. C. Kak. Ultrasonic B-scan imaging : Theory of image formation and a technique for restoration. *Ultrasonic Imaging*, 2 : 1–47, 1980.

[Faur *et al.*, 1998] M. Faur, L. Paradis, J. Oksman et P. Morisseau. A two-step inverse procedure for outer surface defects characterization from ultrasonic bscan images. In *Review of Progress in Quantitative Nondestructive Evaluation*, volume 17, pages 815–822, 1998.

[Fessler et Booth, 1999] J. A. Fessler et S. D. Booth. Conjugate-gradient preconditionning methods for shift-variant PET image reconstruction. *IEEE Trans. Image Processing*, 8 (5) : 668–699, mai 1999.

[Fletcher, 1987] R. Fletcher. *Practical Methods of Optimization*. Wiley&Sons, New York, USA, 2ème edition, 1987.

[Fletcher et Reeves, 1964] R. Fletcher et C. M. Reeves. Function minimization by conjugate gradients. *Comp. J.*, 7 : 149–157, 1964.

[Fourgeaud et Fuchs, 1972] C. Fourgeaud et A. Fuchs. *Statistique*. Dunod, Paris, 2ème edition, 1972.

[Gautier *et al.*, 2001] S. Gautier, J. Idier, F. Champagnat et D. Villard. Restoring separate discontinuities from ultrasonic data. In *Review of Progress in Quantitative Nondestructive Evaluation, AIP Conf. Proc. Vol 615(1)*, pages 686–690, Brunswick, ME, USA, juillet 2001.

[Gautier *et al.*, 2002] S. Gautier, J. Idier, F. Champagnat et D. Villard. Procédé de mesure par ondes sonores et ultrasonores - Méthode de déconvolution. Brevet WO02086485, EDF/CNRS, France, 2002.

[Geman et Reynolds, 1992] D. Geman et G. Reynolds. Constrained restoration and the recovery of discontinuities. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14 (3) : 367–383, mars 1992.

[Geman et Yang, 1995] D. Geman et C. Yang. Nonlinear image recovery with half-quadratic regularization. *IEEE Trans. Image Processing*, 4 (7) : 932–946, juillet 1995.

[Geman et Geman, 1984] S. Geman et D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, PAMI-6 (6) : 721–741, novembre 1984.

[Geman et McClure, 1987] S. Geman et D. McClure. Statistical methods for tomographic image reconstruction. In *Proceedings of the 46th Session of the ICI, Bulletin of the ICI*, volume 52, pages 5–21, 1987.

[Gilbert et Nocedal, 1992] J. C. Gilbert et J. Nocedal. Global convergence properties of conjugate gradient methods for optimization. *SIAM J. Optimization*, 2 (1) : 21–42, 1992.

[Golub et O'Leary, 1989] G. Golub et D. O'Leary. Some history of the conjugate gradient and Lanczos methods : 1948-1976. *SIAM Rev.*, 31 (1) : 50–102, mars 1989.

[Golub et Van Loan, 1996] G. H. Golub et C. F. Van Loan. *Matrix computations*. The Johns Hopkins University Press, Baltimore, 3ème edition, 1996.

[Hadamard, 1901] J. Hadamard. Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton Univ. Bull.*, 13, 1901.

[Hager et Zhang, 2006] W. W. Hager et H. Zhang. A survey of nonlinear conjugate gradient methods. *Pacific J. Optim.*, 2 (1) : 35–58, janvier 2006.

[Hayward et Lewis, 1989] G. Hayward et J. E. L. Lewis. Comparison of some non-adaptive deconvolution techniques for resolution enhancement of ultrasonic data. *Ultrasonics*, 27 (3) : 155–164, mai 1989.

[Hebert et Leahy, 1989] T. Hebert et R. Leahy. A generalized EM algorithm for 3-D Bayesian reconstruction from Poisson data using Gibbs priors. *IEEE Trans. Medical Imaging*, 8 (2) : 194–202, juin 1989.

[Hestenes et Stiefel, 1952] M. R. Hestenes et E. Stiefel. Methods of conjugate gradients for solving linear system. *J. Res. Nat. Bur. Stand.*, 49 : 409–436, 1952.

[Honarvar *et al.*, 2004] F. Honarvar, H. Sheikhzadeh, M. Moles et A. Sinclair. Improving the time-resolution and signal-to-noise ratio of ultrasonic NDE signals. *Ultrasonics*, 41 : 755–63, mars 2004.

[Huber, 1981] P. J. Huber. *Robust Statistics*. John Wiley, New York, NY, USA, 1981.

[Hundt et Trautenberg, 1980] E. Hundt et E. Trautenberg. Digital processing of ultrasonic data by deconvolution. *IEEE Trans. Ultrasonics Ferroelectrics Frequency Control*, 27 (5) : 249–252, septembre 1980.

[Hunt, 1977] B. R. Hunt. Bayesian methods in nonlinear digital image restoration. *IEEE Trans. Communications*, C-26 : 219–229, mars 1977.

[Husby *et al.*, 2001] O. Husby, T. Lie, T. Lango, J. Hokland et H. Rue. Bayesian 2-D deconvolution : a model for diffuse ultrasound scattering. *IEEE Trans. Ultrasonics Ferroelectrics Frequency Control*, 48 (1) : 121–130, janvier 2001.

[Husse *et al.*, 2004] S. Husse, Y. Goussard et J. Idier. Extended forms of Geman and Yang algorithm : application to MRI reconstruction. In *Proc. IEEE ICASSP*, volume III, pages 513–516, Montréal, Québec, Canada, mai 2004.

[Idier, 1999] J. Idier. Regularization tools and models for image and signal reconstruction. In *3rd Int. Conf. Inverse Problems in Engng.*, pages 23–29, Port Ludlow, WA, USA, juin 1999.

[Idier, 2001a] J. Idier, éditeur. *Approche bayésienne pour les problèmes inverses.* Traité IC2, Série traitement du signal et de l'image, Hermès, Paris, novembre 2001.

[Idier, 2001b] J. Idier. Convex half-quadratic criteria and interacting auxiliary variables for image restoration. *IEEE Trans. Image Processing*, 10 (7) : 1001–1009, juillet 2001.

[Idier et Blanc-Féraud, 2001] J. Idier et L. Blanc-Féraud. *Déconvolution en imagerie*, chapitre 6, pages 139–165. In , Idier [2001a], novembre 2001.

[Jalobeanu *et al.*, 2004] A. Jalobeanu, L. Blanc-Feraud et J. Zerubia. An adaptive gaussian model for satellite image deblurring. *IEEE Trans. Image Processing*, 13 (4) : 613–621, avril 2004.

[Jensen, 1992] J. A. Jensen. Deconvolution of ultrasound images. *Ultrasonic Imaging*, 14 (1) : 1–15, janvier 1992.

[Jeurens *et al.*, 1987] T. J. M. Jeurens, J. C. Somer, F. A. M. Smeets et A. P. G. Hoeks. The practical significance of two-dimensional deconvolution in echography. *Ultrasonic Imaging*, 9 (2) : 106–116, avril 1987.

[Kaaresen, 1998] K. F. Kaaresen. Evaluation and applications of the iterated window maximization method for sparse deconvolution. *IEEE Trans. Signal Processing*, 46 (3) : 609–624, mars 1998.

[Kaaresen et Bølviken, 1999] K. F. Kaaresen et E. Bølviken. Blind deconvolution of ultrasonic traces accounting for pulse variance. *IEEE Trans. Ultrasonics Ferroelectrics Frequency Control*, 46 (3) : 564–573, mai 1999.

[Kaufman, 1993] L. Kaufman. Maximum likelihood, least squares, and penalized least squares for PET. *IEEE Trans. Medical Imaging*, 12 : 200–214, juin 1993.

[Koulibaly *et al.*, 1996] P. Koulibaly, P. Charbonnier, L. Blanc Feraud, I. Laurette, J. Darcourt et M. Barlaud. Poisson statistic and half-quadratic regularization for emission tomography reconstruction algorithm. In *Proc. IEEE ICIP*, pages 729–732, 1996.

[Labat et Idier, 2005] C. Labat et J. Idier. Convergence of conjugate gradient methods with a closed-form stepsize formula. Tech. rep., IRCCyN RI2005_11, dec. 2005.

[Labat et Idier, 2007] C. Labat et J. Idier. Convergence of conjugate gradient methods with a closed-form stepsize formula. à paraître, *Journal of Optimisation Theory and Applications*, 2007.

[Labat *et al.*, 2005] C. Labat, J. Idier et Y. Goussard. Comparison between half-quadratic and preconditioned conjugate gradient algorithms for mri reconstruction. In *PSIP'2005 : Physics in signal and Image processing*, Toulouse, janvier 2005.

[Lange *et al.*, 2000] K. Lange, D. R. Hunter et I. Yang. Optimization transfer using surrogate objective functions (with discussion). *J. Comput. Graph. Statist.*, 9 (1) : 1–20, mars 2000.

[Li, 1995] S. Z. Li. On discontinuity-adaptive smoothness priors in computer vision. *IEEE Trans. Pattern Anal. Mach. Intell.*, PAMI-17 (6) : 576–586, juin 1995.

[Liu et Storey, 1991] Y. Liu et C. Storey. Efficient generalized conjugate gradient algorithms, part 1 : Theory. *Journal of Optimisation Theory and Applications*, 69 : 129–137, 1991.

[Mackens et Voss, 2000] W. Mackens et H. Voss. Computing the minimum eigenvalue of a symmetric positive definite toeplitz matrix by newton type methods. *SIAM J. Sci. Comput.*, 21 : 1650–1656, 2000.

[Mallat, 1989] S. G. Mallat. A theory for multiresolution signal decomposition : The wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 11 (7) : 674–693, juillet 1989.

[Mazet *et al.*, 2005] V. Mazet, C. Carteret, D. Brie, J. Idier et B. Humbert. Background removal from spectra by designing and minimising a non-quadratic cost function. *Chemometrics and Intelligent Laboratory Systems*, 76 : 121–133, 2005.

[Moré et Thuente, 1994] J. J. Moré et D. J. Thuente. Line search algorithms with guaranteed sufficient decrease. *ACM trans. on Mathematical Software*, 20 (3) : 286–307, septembre 1994.

[Moulin et Liu, 1999] P. Moulin et J. Liu. Analysis of multiresolution image denoising schemes using generalized - gaussian and complexity priors. *IEEE Trans. Inf. Theory*, 45 (3) : 909–919, avril 1999.

[Mu *et al.*, 2002] Z. Mu, R. Plemmons et P. Santago. Estimation of complex ultrasonic medium responses by deconvolution. In *Proc. IEEE Conf. on Medical Imaging*, 2002.

[Nash, 2000] S. G. Nash. A survey of truncated-Newton methods. *J. Comput. Appl. Math.*, 124 : 45–59, 2000.

[Nashed, 1981] M. Z. Nashed. Operator-theoretic and computational approaches to ill-posed problems with applications to antenna theory. *IEEE Trans. Ant. Propag.*, 29 : 220–231, 1981.

[Ng *et al.*, 1999] M. K. Ng, R. H. Chan et W.-C. Tang. A fast algorithm for deblurring models with Neumann boundary conditions. *SIAM J. Sci. Comput.*, 21 (3) : 851–866, 1999.

[Nikolova, 1999] M. Nikolova. Markovian reconstruction using a GNC approach. *IEEE Trans. Image Processing*, 8 (9) : 1204–1220, septembre 1999.

[Nikolova, 2002] M. Nikolova. Minimizers of cost-functions involving non-smooth data-fidelity terms. Application to the processing of outliers. *SIAM J. Num. Anal.*, 40 (3) : 965–994, 2002.

[Nikolova *et al.*, 1998] M. Nikolova, J. Idier et A. Mohammad-Djafari. Inversion of large-support ill-posed linear operators using a piecewise Gaussian MRF. *IEEE Trans. Image Processing*, 7 (4) : 571–585, avril 1998.

[Nikolova et Ng, 2001] M. Nikolova et M. Ng. Fast image reconstruction algorithms combining half-quadratic regularization and preconditioning. In *Proc. IEEE ICIP*, pages 277–280, Thessaloniki, Grèce, octobre 2001.

[Nikolova et Ng, 2005] M. Nikolova et M. Ng. Analysis of half-quadratic minimization methods for signal and image recovery. *SIAM J. Sci. Comput.*, 27 : 937–966, 2005.

[Nocedal et Wright, 1999] J. Nocedal et S. J. Wright. *Numerical optimization*. Springer Texts in Operations Research. Springer-Verlag, New York, NY, USA, 1999.

[O'Brien *et al.*, 1994] M. S. O'Brien, A. N. Sinclair et S. M. Kramer. Recovery of a sparse spike time series by l_1 norm deconvolution. *IEEE Trans. Signal Processing*, 42 (12) : 3353–3365, décembre 1994.

[Polak et Ribière, 1969] E. Polak et G. Ribière. Note sur la convergence de méthodes des directions conjuguées. *Rev. Française d'Informatique et de Recherche Opérationelle*, 16 : 35–43, 1969.

[Polyak, 1969] B. T. Polyak. The conjugate gradient method in extreme problems. *USSR Comp. Math. and Math. Phys.*, 9 : 94–112, 1969.

[Powell, 1984] M. J. D. Powell. *Nonconvex minimization calculations and the conjugate gradient method*, volume 1066 de *Lecture Notes in Mathematics*. Springer Verlag, Berlin, 1984.

[Press *et al.*, 1992] W. H. Press, S. A. Teukolsky, W. T. Vetterling et B. P. Flannery. *Numerical recipes in C, the art of scientific computing*. Cambridge Univ. Press, New York, 2ème edition, 1992.

[Rivera et Marroquin, 2003] M. Rivera et J. Marroquin. Efficient half-quadratic regularization with granularity control. *Image and Vision Comp.*, 21 (4) : 345–357, avril 2003.

[Rockafellar, 1970] R. T. Rockafellar. *Convex Analysis*. Princeton Univ. Press, 1970.

[Roullot *et al.*, 2004] E. Roullot, A. Herment, I. Bloch, A. de Cesare, M. Nikolova et E. Mousseaux. Modeling anisotropic undersampling of magnetic resonance angiographies and reconstruction of a high-resolution isotropic volume using half-quadratic regularization techniques. *Signal Processing*, 84 : 743–762, avril 2004.

[Sallard et Paradis, 1998] J. Sallard et L. Paradis. Use of a priori information for the deconvolution of ultrasonic signals. In *Review of Progress in Quantitative Nondestructive Evaluation*, volume 17, pages 735–742, 1998.

[Schomberg *et al.*, 1983] H. Schomberg, W. Vollmann et G. Mahnke. Lateral inverse filtering of ultrasonic B-scan images. *Ultrasonic Imaging*, 5 (1) : 38–54, janvier 1983.

[Sciacca et Evans, 1992] L. J. Sciacca et R. J. Evans. Signal processing applied to ultrasonic imaging. In *IEEE Trans. Acoust. Speech, Signal Processing*, pages 225–228, octobre 1992.

[Shi et Shen, 2004] Z.-J. Shi et J. Shen. A gradient-related algorithm with inexact line searches. *J. Comput. Appl. Math.*, 170 (2) : 349–370, 2004.

[Shi et Shen, 2005] Z.-J. Shi et J. Shen. Convergence property and modifications of a memory gradient method. *Asia-Pac. J. Oper. Res.*, 22 (4) : 463–477, 2005.

[Sun et Zhang, 2001] J. Sun et J. Zhang. Global convergence of conjugate gradient methods without line search. *Annals of Operations Research*, 103 : 161–173, 2001.

[Taxt et Jirik, 2004] T. Taxt et R. Jirik. Superresolution of ultrasound images using the first and second harmonic signal. *IEEE Trans. Ultrasonics Ferroelectrics Frequency Control*, 51 (2) : 163–175, février 2004.

[Tikhonov, 1963] A. Tikhonov. Regularization of incorrectly posed problems. *Soviet. Math. Dokl.*, 4 : 1624–1627, 1963.

[Tikhonov et Arsenin, 1977] A. Tikhonov et V. Arsenin. *Solutions of Ill-Posed Problems*. Winston, Washington, DC, USA, 1977.

[Tikhonov et Arsénine, 1976] A. Tikhonov et V. Arsénine. *Méthodes de résolution de problèmes mal posés*. Éditions MIR, Moscou, Russie, 1976.

[Tseng et Bertsekas, 1987] P. Tseng et D. P. Bertsekas. Relaxation methods for problems with strictly convex separable costs and linear constraints. *Mathematical Programming*, 38 : 303–321, 1987.

[Vogel et Oman, 1996] R. V. Vogel et M. E. Oman. Iterative methods for total variation denoising. *SIAM J. Sci. Comput.*, 17 (1) : 227–238, janvier 1996.

[Vollmann, 1982] W. Vollmann. Resolution enhancement of ultrasonic B-scan images by deconvolution. *IEEE Trans. Sonics Ultrasonics*, 29 (2) : 78–83, mars 1982.

[Voss et Eckhardt, 1980] H. Voss et U. Eckhardt. Linear Convergence of Generalized Weiszfeld's Method. *Computing*, 25 : 243–251, 1980.

[Wang *et al.*, 1995] G. Wang, J. Zhang et G.-W. Pan. Solution of inverse problems in image processing by wavelet expansion. *IEEE Trans. Image Processing*, 4 (5) : 579–593, mai 1995.

[Weiszfeld, 1937] E. Weiszfeld. Sur le point pour lequel la somme des distances de n points donnés est minimum. *Tôhoku Mathematical Journal*, 43 : 355–386, 1937.

[Winkler, 1995] G. Winkler. *Image Analysis, Random Fields and Dynamic Monte Carlo Methods*. Springer Verlag, Berlin, Allemagne, 1995.

[Wolfe, 1971] P. Wolfe. Convergence conditions for ascent methods. II : Some corrections. *SIAM Rev.*, 13 : 185–188, 1971.

Titre : *Algorithmes d'optimisation de critères pénalisés pour la restauration d'images. Application à la déconvolution de trains d'impulsions en imagerie ultrasonore.*

Résumé : La solution de nombreux problèmes de restauration et de reconstruction d'images se ramène à celle de la minimisation d'un critère pénalisé qui prend en compte conjointement les observations et les informations préalables. Ce travail de thèse s'intéresse à la minimisation des critères pénalisés préservant les discontinuités des images. Nous discutons des aspects algorithmiques dans le cas de problèmes de grande taille. Il est possible de tirer parti de la structure des critères pénalisés pour la mise en œuvre algorithmique du problème de minimisation. Ainsi, des algorithmes d'optimisation semi-quadratiques (SQ) convergents exploitant la forme analytique des critères pénalisés ont été utilisés. Cependant, ces algorithmes SQ sont généralement lourds à manipuler pour les problèmes de grande taille. L'utilisation de versions approchées des algorithmes SQ a alors été proposée. On peut également envisager d'employer des algorithmes du gradient conjugué non linéaire GCNL+SQ1D utilisant une approche SQ scalaire pour la recherche du pas. En revanche, plusieurs questions liées à la convergence de ces différentes structures algorithmiques sont restées sans réponses jusqu'à présent. Nos contributions consistent à :

- Démontrer la convergence des algorithmes SQ approchés et GCNL+SQ1D.
- Etablir des liens forts entre les algorithmes SQ approchés et GCNL+SQ1D.
- Illustrer expérimentalement en déconvolution d'images le fait que les algorithmes SQ approchés et GCNL+SQ1D sont préférables aux algorithmes SQ exacts.
- Appliquer l'approche pénalisée à un problème de déconvolution d'images en contrôle non destructif par ultrasons.

Mots-clés : Problèmes inverses, restauration et reconstruction d'images, déconvolution, approche bayésienne, critères pénalisés, optimisation, convergence, algorithmes semi-quadratiques, algorithmes du gradient conjugué non linéaires, contrôle non destructif par ultrasons.

Title : *Optimization of penalized criteria for image restoration. Application to sparse spike train deconvolution in ultrasonic imaging.*

Abstract : The solution to many image restoration and reconstruction problems is often defined as the minimizer of a penalized criterion that accounts simultaneously for the data and the prior. This thesis deals more specifically with the minimization of edge-preserving penalized criteria. We focus on algorithms for large-scale problems. The minimization of penalized criteria can be addressed using a half-quadratic approach (HQ). Converging HQ algorithms have been proposed. However, their numerical cost is generally too high for large-scale problems. An alternative is to implement inexact HQ algorithms. Nonlinear conjugate gradient algorithms can also be considered using scalar HQ algorithms for the line search (NLCG+HQ1D). Some issues on the convergence of the aforementioned algorithms remained open until now. In this thesis we :

- Prove the convergence of inexact HQ algorithms and NLCG+HQ1D.
- Point out strong links between HQ algorithms and NLCG+HQ1D.
- Experimentally show that inexact HQ algorithms and NLCG+HQ1D perform better than exact HQ algorithms, for an image deconvolution test problem.
- Apply the penalized approach to a deconvolution problem in the field of ultrasonic imaging for nondestructive testing.

Keywords : Inverse problems, image restoration and reconstruction, deconvolution, Bayesian framework, penalized criterion, optimization, convergence, half-quadratic algorithms, nonlinear conjugate gradient methods, ultrasonic nondestructive testing.