Contributions à la résolution de problèmes inverses de grande taille en traitement du signal et de l'image

Saïd Moussaoui

Habilitation à Diriger des Recherches

Université de Nantes

Ecole doctorale STIM (Sciences et Technologies de l'Information et Mathématiques)
Spécialité : Automatique, Robotique et Traitement du signal

Contributions à la résolution de problèmes inverses de grande taille en traitement du signal et de l'image

Présentée et soutenue par :

Saïd MOUSSAOUI

le 9 décembre 2014

devant le jury composé de

Président :	Christian JUTTEN	Professeur à l'Université de Grenoble, GIPSA-lab
Rapporteurs :	Olivier CAPPÉ	Directeur de Recherche au CNRS, Télécom ParisTech
	François MALGOUYRES	Professeur à l'Université Paul Sabatier, IMT, Toulouse
	Cédric RICHARD	Professeur à l'Université de Nice, Laboratoire Lagrange
Examinateurs :	David BRIE	Professeur à l'Université de Lorraine, CRAN, Nancy
	Jean-Pierre GUÉDON	Professeur à l'Université de Nantes, IRCCyN

Directeur de Recherche

Jérôme IDIER Directeur de Recherche au CNRS, IRCCyN

Laboratoire : Institut de Recherche en Communications et Cybernétique de Nantes
(UMR 6597, CNRS, Ecole centrale de Nantes, Université de Nantes, Ecole des mines de Nantes)

Préambule

Ce manuscrit, organisé en trois parties, présente une synthèse de mes activités d'enseignement et de recherche effectuées au cours des huit années passées à l'Ecole Centrale de Nantes en tant que maître de conférences, enseignant au sein du département *Automatique et Robotique* et chercheur au sein de *l'Institut de Recherche en Communications et Cybernétique de Nantes (IRCCyN).*

La première partie présente une vue d'ensemble de mes activités d'enseignement et de recherche ainsi qu'un bilan global en termes de publications scientifiques, contrats industriels, encadrement doctoral et de responsabilités collectives.

Une revue synthétique de mes travaux de recherche liés à la résolution de problèmes inverses de grande taille en traitement du signal et de l'image est fournie dans la seconde partie de ce manuscrit. Après une introduction générale visant à expliquer les motivations de ces travaux, trois chapitres seront focalisés sur des contributions méthodologiques sur les outils d'optimisation et de simulation statistique ainsi que les applications potentielles de ces développements. Plus de détails techniques sur ces méthodes et des références bibliographiques supplémentaires sont fournis dans une sélection de mes principales publications annexées à ce manuscrit

Dans la troisième partie, je présenterai mes perspectives scientifiques à court terme et je détaillerai les grandes lignes du projet de recherche que je compte mener à plus long terme.

La rédaction de ce manuscrit d'habilitation à diriger des recherches a été l'occasion de faire un point d'étape sur mon activité d'enseignant-chercheur. En plus de la projection vers l'avenir, cela a également permis de retracer un parcours scientifique et académique qui fût plein de rencontres, d'opportunités et de relations humaines très stimulantes.

Table des matières

Table des figures

Liste des tableaux

Première partie

Parcours académique et scientifique

Chapitre 1

Curriculum vitæ

1.1 Identification

Nom :	MOUSSAOUI	**Prénom :**	Saïd
Naissance :	le 31/10/1977	**Lieu :**	Makouda, Algérie
Affectation :	Ecole centrale de Nantes	**Date :**	1er septembre 2006
Corps :	Maîtres de conférences	**Section CNU :**	61ème
Grade :	4ème échelon depuis le 1er avril 2012	**Classe :**	normale
Enseignement :	Département Automatique et Robotique		
Recherche :	Institut de Recherche en Communications et Cybernétique de Nantes IRCCyN, UMR CNRS 6597		

1.2 Formation initiale et parcours académique

Depuis Sept. 2006	Maître de conférences à l'Ecole Centrale de Nantes.
Sept 2005 - Sept 2006	Attaché temporaire d'enseignement et de recherche. Université Henri Poincaré Nancy 1.
Oct. 2002 - Dec. 2005	Thèse de doctorat. Spécialité : Traitement du signal. Université Henri Poincaré, Nancy 1 Centre de Recherche en Automatique de Nancy.
Sept. 2001 - Sept. 2001	Master Contrôle Signaux et Communication Université Henri Poincaré, Nancy 1.
Sept. 1996 - Juin. 2001	Ingénieur d'état en électronique. Ecole Nationale Polytechnique, Alger

1.3 Activités d'enseignement

J'ai intégré le département *Automatique et Robotique* de l'Ecole Centrale de Nantes le 1er septembre 2006. J'assure des enseignements à tous les niveaux de la formation ingénieur : Dans le tronc commun (première et deuxième année) et en troisième année dans l'option disciplinaire ISIS (« *Ingénierie des systèmes, des images et des signaux* »). J'assure également des cours du Master 2 ARIA (*Automatique, Robotique et Informatique Appliquée*).

Pour donner une vue globale de mon activité d'enseignement, le tableau suivant fournit une liste des enseignements que j'ai réalisés durant les cinq dernières années.

Niveau	Enseignement	Volume horaire			
		CM	TD	TP	TA
1ère année	Signaux, Systèmes, Simulation	2h	16h	24h	8h
1ère année + Apprentissage	Instrumentation-Capteurs	10h	16h	16h	8h
2ème année	Commande de systèmes		16h	12h	8h
2ème année	Actionneurs électriques		8h	12h	8h
2ème année	Télédétection hyperspectrale	5h			
3ème année + Master 2	Séparation de sources	7h		8h	
3ème année + Master 2	Identification de systèmes	10h		10h	
3ème année	Implémentation	4h		8h	
3ème année	Projets en signal-image				12 h

TABLE 1.1 – Vue globale de ma charge d'enseignement annuelle durant les cinq dernières années.

Dès mon arrivée à l'ECN, des enseignements sous forme de travaux dirigés, pratiques et en autonomie dans divers cours relevant de l'EEA (Signaux-systèmes, Commande de systèmes dynamiques et Actionneurs électriques) m'ont été confiés. Au fil des années, la nature et le volume de mon activité d'enseignement ont évolué pour s'adapter aux modifications de l'offre de formation ingénieur à l'ECN et aux départs à la retraite de deux collègues enseignants-chercheurs.

Suite à une refonte du tronc commun de la formation ingénieur en 2008, de nouveaux enseignements m'ont été confiés. Tout d'abord, j'ai pris en charge le montage d'un nouveau cours intitulé « *Instrumentation–Capteurs* » en première année : à ce titre, je mets en place le contenu du cours, assure les cours magistraux (4h), organise la venue des intervenants extérieurs (6h) et gère les travaux dirigés et pratiques (12 groupes de 30 élèves-ingénieurs, en moyenne). J'ai par ailleurs, contribué à l'évolution du cours « *Signaux, systèmes et simulation* » en proposant une nouvelle partie « *Filtrage de signaux* » (2h de cours magistraux, 2h de travaux dirigés et 4h de travaux pratiques).

J'ai également mis en place deux autres enseignements en troisième année de la formation ingénieur au sein de l'option disciplinaire ISIS. Ces cours portent sur la « *Séparation de sources* », qui est également un cours destiné aux étudiants de Master 2 ARIA et « *Implémentation en traitement du signal et de l'image* ». Suite au départ à la retraite de Jean-Marie Piasco en 2009, j'ai assuré sa

succession en prenant en charge l'intégralité du cours « *Identification de systèmes dynamiques* » (destiné aux élèves-ingénieurs de l'option ISIS et du Master 2 ARIA).

Depuis la rentrée 2013-2014, j'assure la responsabilité de l'option disciplinaire ISIS (24 étudiants issus de la troisième année de la formation ingénieur ECN ou d'une mobilité internationale).

1.4 Activités de recherche

J'ai été initié au monde de la recherche scientifique à travers ma thèse de doctorat, réalisée sous la direction de David BRIE (université de Nancy 1), au *Centre de Recherche en Automatique de Nancy (CRAN, UMR CNRS 7039)*, commencée en septembre 2002 et soutenue le 7 décembre 2005. Le problème traité dans cette thèse concerne *le développement de méthodes de séparation de sources non-négatives* pour une application au traitement de signaux de spectroscopie. Ce travail de thèse m'a appris à mener un projet de recherche sur une durée déterminée en m'offrant l'opportunité de le réaliser dans un cadre pluridisciplinaire et de nouer des collaborations avec d'autres chercheurs. En effet, j'ai réalisé mon travail de thèse en étroite collaboration avec des spécialistes de la spectroscopie et notamment Cédric CARTERET et Bernard HUMBERT, du *Laboratoire de Chimie-Physique et Microbiologie pour l'environnement (LCPME, UMR CNRS 7564)* à Nancy. Par la suite, une collaboration avec Christian JUTTEN et Jocelyn CHANUSSOT (GIPSA-lab, Université Joseph Fourier, Grenoble), commencée en mars 2006 a ouvert une voie pour l'application des méthodes de séparation de sources non-négatives et d'analyse en composantes indépendantes au traitement de données hyperspectrales.

J'ai intégré l'équipe *Analyse et Décision en Traitement du Signal et de l'Image (ADTSI)* de *l'Institut de Recherche en Communications et Cybernétique de Nantes (IRCCyN, UMR CNRS 6597)* le 1er septembre 2006. Mon activité de recherche au sein l'équipe ADTSI se positionne dans un axe portant sur la résolution de problèmes inverses et ses applications en traitement de signaux et d'images, animé par Jérôme IDIER. Cette activité couvre à la fois les aspects méthodologiques (modèles statistiques et bayésiens, méthodes d'optimisation itérative et algorithmes de simulation stochastique) et les applications (télédétection par analyse d'images hyperspectrales, traitement de données de spectroscopie moléculaire). Mon arrivée a, d'une part, permis de renforcer les travaux de l'équipe ADTSI sur les méthodes de restauration ou de reconstruction d'images et, d'autre part, d'apporter de nouvelles perspectives en séparation de sources et ses applications.

En étroite collaboration avec des collègues de l'équipe *Commande* au sein de l'IRCCyN (Franck PLESTAN, notamment), je travaille sur l'intégration des techniques de traitement du signal dans la mise en œuvre de dispositifs de diagnostic ou de commande de systèmes. Ces travaux concernent l'exploitation des méthodes d'analyse spectrale et de prédiction linéaire pour la conception de systèmes de surveillance de la pression des pneumatiques d'un véhicule automobile ou encore de commande d'un dispositif de récupération de l'énergie des vagues.

La section suivante développe mes principaux thèmes de recherches en mettant l'accent sur les principaux résultats obtenus. Je signale que ces travaux sont réalisés à travers des encadrements de thèses de doctorat ou des collaborations scientifiques.

1.4.1 Modèles statistiques et méthodes bayésiennes pour la séparation de sources

La séparation de sources consiste à retrouver des signaux d'intérêt, appelés sources, à partir d'observations qui sont des mélanges de ces sources. Afin de résoudre le problème inverse de la séparation (estimation des signaux sources), il est nécessaire d'introduire des hypothèses statistiques ou des contraintes physiques sur les sources recherchées. De plus, il est parfois nécessaire d'identifier le processus de mélange et donc de lui imposer également des contraintes physiques.

Travaux de thèse. Mon activité dans cette thématique a été initiée dans le cadre de la thèse de doctorat que j'ai réalisée à l'Université Henri Poincaré, Nancy 1. Cette thèse a porté sur le développement de méthodes de séparation de sources en vue de les appliquer au traitement de signaux de spectroscopie. La problématique centrale de sujet consiste à retrouver, à partir d'une collection de spectres enregistrés dans des conditions physico-chimiques différentes, les spectres des constituants purs présente au sein des mélanges ainsi que leurs proportions. J'ai pu montrer que dans le cadre d'un tel problème de séparation de sources spectrales, il est nécessaire de rajouter des contraintes telles que la positivité des spectres recherchés, et proposé une méthode de séparation par approche basyésienne [A.2, A.3], J'ai soutenu ma thèse de doctorat, intitulée : *Séparation de sources non-négatives. Application au traitement des signaux de spectroscopie*, le 7 décembre et 2005 devant le jury suivant :

Rapporteurs :	Pierre-Olivier AMBLARD, Chargé de recherche au CNRS
	Jean-Yves TOURNERET, Professeur à l'I.N.P de Toulouse
Examinateurs :	Christian JUTTEN, Professeur à l' univ. Joseph Fourier, Grenoble
	Ali MOHAMMAD-DJAFARI, Directeur de recherche au CNRS
	Cédric CARTERET, Maître de conférences à l'univ. Henri Poincaré, Nancy 1
	David BRIE, Professeur à l'Université Henri Poincaré, Nancy 1
Invité :	Bernard HUMBERT, Professeur à l'univ. Henri Poincaré, Nancy 1

Par la suite, j'ai appliqué la méthode développée durant ma thèse au traitement de données réelles. Ainsi, j'ai pu établir des collaborations qui ont permis de résoudre des problèmes de traitement de données de spectroscopie Raman [A.9], d'analyse d'images hyperspectrales fournies par la sonde européenne Mars Express [A.8, A.17] et de décomposition du rayonnement solaire en trois sources d'activités distinctes à partir des mesures de flux EUV [A.4] ou d'images multispectrales [A.21].

Après la thèse. La deuxième phase de mon travail de recherche sur le thème de la séparation de source s'est focalisée sur le développement de méthodes de séparation bayésienne capables d'inclure la contrainte de somme à l'unité des coefficients de mélange. Cette contrainte s'avère utile pour les applications telles que le suivi de cinétiques chimiques [A.10] ou l'analyse d'imagers hyperspectrales [A.12]. Ce travail a été réalisé dans le cadre d'un projet Jeunes Chercheurs, conjointement avec Nicolas DOBIGEON [1], soutenu par le GdR ISIS.

Par ailleurs, le modèle de mélange linéaire peut ne pas être suffisant pour décrire finement le processus d'observation. J'ai donc travaillé, en collaboration avec Leonardo TOMAZELI-DUARTÉ [2], sur la mise en œuvre de méthodes de séparation de sources par approche bayésienne dans le cas de mélanges non-linéaires ou dans le cas d'un mélange linéaire-quadratique [A.16]. Un exemple concret est celui de traitement de signaux issus de capteurs d'ions chimiques à faible sélectivité [A.11]. L'inférence bayésienne dans ces problèmes requiert le recours à des modèles statistiques *a priori* adéquats pour la traduction des contraintes sur les paramètres d'intérêt et l'utilisation de techniques avancées pour la simulation des lois *a posteriori* [A.25].

Mes travaux en séparation de sources se poursuivent dans le cadre des travaux de thèse de Antoine BA [3], que je co-encadre sous la direction de Patrick LAUNEAU, sur des problématiques de télédétection littorale par fusion de données LIDAR avec des images hyperspectrales.

1.4.2 Recherche de pas pour la résolution itérative de problèmes inverses

La résolution de problèmes inverses en traitement du signal et de l'image est souvent ramenée à la minimisation d'un critère composite fusionnant adéquation aux données et satisfaction de certaines propriétés désirables de la solution recherchée. La mise en œuvre efficace, d'un point de vue théorique (garantie de convergence, taux de convergence) et pratique (coût de calcul, utilisation mémoire), d'un algorithme d'optimisation nécessite une attention particulière à la structure du critère à minimiser et ses propriétés (convexité, différentiabilité, domaine de définition, etc.), notamment dans le cadre des problèmes de grande taille. Par exemple, dans le cadre de la restauration d'images par maximum d'entropie ou en tomographie par émission de positrons, le critère présente une barrière au bord du domaine admissible. Cette barrière permet de satisfaire implicitement des contraintes d'inégalité, telles que la positivité. Cependant, comme la fonction barrière (ou sa dérivée première) tend vers l'infini lorsque l'on se rapproche du bord du domaine de définition des solutions admissibles, la minimisation du critère par un algorithme de descente itérative nécessite une méthode appropriée pour le calcul de la direction de descente ainsi que pour la recherche du pas scalaire le long de cette direction.

1. Doctorant (2004-2007) puis Enseignant-Chercheur à l'ENSEEIHT, Toulouse, France
2. Doctorant à l'INPG puis Enseignant-Chercheur à L'université d'État de Campinas (UNICAMP), Brésil
3. Doctorant Université de Nantes (2013-2016)

J'ai commencé une activité de recherche dans ce domaine dès mon intégration à l'équipe *ADTSI* le 1er septembre 2006 par le co-encadrement des thèses de master et de doctorat de Emilie CHOUZENOUX [4], sous la direction de Jérôme IDIER, et dont le sujet de recherche a porté sur la recherche de pas par des méthodes de *majoration-minimisation* (MM). Ainsi, une stratégie de recherche de pas s'appuyant sur des approximations majorantes log-quadratiques a été développée [A.18]. Cette technique a été appliquée avec succès pour proposer une algorithme de reconstruction de spectres de relaxation (T1, T2 ou T1-T2) en spectroscopie par résonance magnétique nucléaire [A.13]. Cette méthode de reconstruction est actuellement utilisée régulièrement par l'équipe IRM-Food au sein de l'IRSTEA (*Institut national de recherche en sciences et technologies pour l'environnement et l'agriculture*) à Rennes.

Par ailleurs, nous avons montré que la recherche de pas par majoration-minimisation s'étend assez naturellement aux algorithmes de descente dans un sous-espace de directions ou encore aux méthodes à base de mémoire de gradient. Cette stratégie de recherche de pas multi-dimensionnelle accélérant la convergence d'algorithmes de sous-espaces a été développée puis appliquée à des problèmes de restauration d'images [A.15].

1.4.3 Accélération algorithmique et matérielle des méthodes d'optimisation

Outre la prise en compte de contraintes sur les solutions recherchées, un deuxième aspect inhérent à la résolution de problèmes inverses est celui de la maîtrise de la complexité numérique des algorithmes. Parmi les techniques d'optimisation sous contraintes, les méthodes fondées sur les points intérieurs sont reconnues pour leur efficacité théorique mais souffrent d'un problème de coût de calcul trop élevé qui les rend inutilisables pour des problèmes de grande taille.

Le but recherché dans le cadre de la thèse de Maxime LEGENDRE [5], que je co-encadre avec Jérôme IDIER et Frédéric SCHMIDT, est de proposer des structures algorithmiques optimisées des techniques de minimisation sous-contraintes de type point-intérieurs et de les mettre en œuvre à l'aide d'outils de calcul parallèle tels que les processeurs de cartes graphiques (GPU). L'application sur laquelle j'ai focalisé ces travaux de recherche est celle du démélange linéaire d'images hyperspectrales.

Ainsi, dans le cadre d'un contrat de collaboration avec l'agence spatiale européenne, nous avons démontré dans [A.23] l'intérêt du calcul sur GPU par rapport à une approche basée sur le multi-CPU mais aussi mis en avant la nécessité de bien gérer les temps de transfert de données et des accès mémoire. Cette approche a été appliquée avec succès au problème d'estimation des cartes d'abondances en imagerie hyperspectrale dans le cas d'un mélange linéaire, d'abord par la minimisation d'un critère des moindres carrés, en utilisant une approche primale-duale des points-intérieurs, sous des contraintes de positivité, de somme à un ou encore de somme inférieure

4. Doctorante ECN (2007-2010) puis Enseignante-Chercheuse à l'Université Paris-Est Marne-la-Vallée.
5. Doctorant ECN (2013-2015)

ou égale à un [A.24], puis en ajoutant un critère régularisant permettant d'assurer une régularité de la répartition spatiale des composants de la surface imagée [A.20]. Les travaux de réduction du temps de calcul et de l'utilisation mémoire, dans le cas de la régularisation spatiale, se poursuivent par l'exploitation des techniques de majoration-minimisation pour augmenter le niveau de parallélisation des algorithmes [CI.23, CI.24]. Le point important qui ressort de ces travaux est la nécessité de concevoir des algorithmes d'optimisation ayant des propriétés théoriques bien fondées et qui présentent des étapes dont le calcul peut se faire de façon parallèle sur différents processeurs ou encore opter pour une approximation majorante séparable du critère afin d'aboutir à une structure d'algorithme se prêtant à un calcul parallélisable.

1.4.4 Algorithmes de simulation bayésienne pour des problèmes de grande taille

En inférence bayésienne par des des méthodes de Monte Carlo par chaînes de Markov, il est fréquent qu'une partie des composantes à rééchantillonner correspondent à un vecteur ou à une image de grande dimension et de loi conditionnelle gaussienne dont la covariance varie au cours des itérations (du fait de sa dépendance à des hyperparamètres eux-même rééchantillonnés, par exemple). C'est souvent le cas, par exemple, en restauration de signaux et d'images. Pour éviter le coût de calcul d'une factorisation de type Cholesky à chaque itération, une solution récente propose de passer par la résolution d'un système linéaire de grande dimension, effectuée de façon approchée par gradient conjugué tronqué. Mes travaux de recherche sur cette question consistent à (1) mettre en évidence que cette troncature empêche la convergence vers la loi cible ; (2) proposer l'incorporation d'une étape d'acceptation-rejet peu coûteuse rétablissant la convergence, (3) optimiser le choix du seuil de troncature de manière à maximiser l'efficacité statistique des échantillonneurs. Les résultats préliminaires de ce travail ont été publiés dans [CN.11, A. 26]. Ce travail de recherche, réalisé dans le cadre de la thèse de Clément GILAVERT [6] que je co-encadre avec Jérôme IDIER, se positionne dans un cadre plus général qui concerne le couplage des méthodes d'optimisation avec les méthodes de Monte Carlo pour accélérer les vitesses de convergence des algorithmes tout en assurant un coût de calcul acceptable par itération.

1.4.5 Systèmes de surveillance de la pression des pneumatiques d'un véhicule

Le pneumatique est un élément essentiel pour la tenue de route, le confort et la sécurité d'un véhicule automobile. Ses propriétés physiques, affectant son comportement mécanique ou sa dynamique, sont fortement liées à la pression. Une baisse de la pression a une incidence directe sur des caractéristiques comme l'amortissement, la raideur et la rigidité, ce qui provoque une usure rapide et une surconsommation de carburant. Il est donc important de développer des systèmes de surveillance de la pression des pneumatiques (SSPP) de façon permanente. Pour des

6. Doctorant ECN (2012-2014)

raisons d'économie et de sûreté de fonctionnement, les nouvelles générations de SSPP favorisent les méthodes, appelées indirectes, sans capteur de pression.

Mes travaux de recherche sur ce sujet sont réalisés dans le cadre d'un contrat de collaboration avec Renault et consistent à développer un système de détection d'une baisse de pression des pneumatiques d'un véhicule automobile à partir de la seule analyse des signaux de vitesse angulaire des roues. J'ai co-encadré la thèse de Charbel EL TANNOURY[7], sous la direction de Franck PLESTAN, dans laquelle un système de détection fondé sur l'analyse spectrale des signaux de vitesse angulaire des roues en employant des modèles autoregressifs a été développé. Le travail réalisé consiste en la mise en œuvre de toute la chaine d'acquisition et de traitement des données en incluant des étapes de correction d'éventuels défauts des capteurs de position angulaire des roues. Ces travaux ont fait l'objet de deux brevets déposés avec Renault [B4, B5]. Un résultat important est la mise en évidence de présence de modes de résonance dont la localisation fréquentielle dépend de la pression du pneumatique. Par ailleurs, la modélisation de la dynamique de la roue nous a amené à proposer une méthode d'analyse basée sur les observateurs du rayon dynamique et de la résistance au roulement [A.22]. Ce travail a également fait l'objet de dépôt de trois brevets [B1, B2, B3]. Ces développements se poursuivent, dans le cadre de la thèse de Joan DAVIS-VALLDAURA[8], que je co-encadre sous la direction de Franck PLESTAN, en collaboration avec Renault, pour l'optimisation des paramètres de mise en oeuvre du système de surveillance de la pression des pneumatiques d'un véhicule.

1.5 Activités d'encadrement

1.5.1 Master

[M.4] Nitish KUMAR, « Electrolocation with a bio-inspired electric sensor and electric resistance tomography ». Master EMARO. Ecole Centrale de Nantes. Septembre. 2011. Encadrants : **S. Moussaoui** (80%), A. Girin (Ingénieur de Recherche, IRCCyN, 20 %)

[M.3] Abbas ATAYA, « Electrolocation par tomographie de résistance électrique ». Master ASP. Ecole Centrale de Nantes. Septembre 2010. Encadrants : **S. Moussaoui** (80%), A. Girin (Ingénieur de Recherche, IRCCyN, 20 %)

[M.2] Xavier ARTUSI, « Reconstruction d'un spectre RMN 2D par maximum d'entropie ». Master ASP. Ecole Centrale de Nantes. Septembre 2008. Encadrants : **S. Moussaoui** (70%), J. Idier (DR CNRS, 30 %)

[M.1] Emilie CHOUZENOUX, « Reconstruction d'images 2D en tomographie par emission de positrons ». Master ASP. Ecole Centrale de Nantes. Septembre 2007. Encadrants : **S. Moussaoui** (70%), J. Idier (DR CNRS, 30 %)

7. Doctorant ECN (2009-2012)
8. Doctorant ECN (2013-2016)

1.5.2 Doctorat

[D.7] Antoine BA, « Traitement conjoint de signaux LIDAR et d'images hyperspectrales pour la surveillance des environnements côtier et urbain », Université de Nantes.

Financement	:	Programme régional des Pays de la Loire
Directeur	:	P. LAUNEAU (PR Univ. Nantes, 50 %)
Co-encadrement	:	**S. Moussaoui** (30 %), M. Robin (Univ. Nantes, 20 %)
Début	:	10/2013
Publications	:	Néant

[D.6] Joan DAVINS-VALLDAURA, « Optimisation des paramètres d'un système de surveillance de la pression des pneumatiques d'un véhicule automobile », Ecole centrale de Nantes.

Financement	:	Convention CIFRE RENAULT-ECN
Directeur	:	F. Plestan (PR ECN, 40 %)
Co-encadrement	:	**S. Moussaoui** (60 %)
Début	:	04/2013
Publications	:	Néant

[D.5] Christophe LALUC, « Prédiction de l'élévation des vagues et commande d'un houlogénérateur », Ecole centrale de Nantes.

Financement	:	Projet ANR Qualiphe
Directeur	:	F. Plestan (PR ECN, 40 %)
Co-encadrement	:	**S. Moussaoui** (40 %), A. Clément (IR CNRS, 20 %)
Début	:	04/2012. Thèse actuellement suspendue pour des raisons médicales.
Publications	:	[CN. 9]

[D.4] Maxime LEGENDRE, « Accélération algorithmique et matérielle des méthodes d'optimisation sous contraintes pour l'analyse d'images hyperspectrales », Ecole centrale de Nantes.

Financement	:	Contrat ESA et bourse de la Région des Pays de la Loire
Directeur	:	J. IDIER (DR CNRS, 40 %)
Co-encadrement	:	**S. Moussaoui** (30 %), F. Schmidt (MC Univ. Paris-Sud, 30 %)
Début	:	03/2012
Soutenance	:	fin 2014
Publications	:	[CN.8, CI.23, CI.24, A.23, A.24]

[D.3] Clément GILAVERT, « Couplage des méthodes de simulation bayésienne et d'optimisation itérative pour la résolution de problèmes inverses de grande taille », Ecole centrale de Nantes.

Financement	:	Cofinancement BDI CNRS et bourse Région des Pays de la Loire
Directeur	:	J. Idier (DR CNRS, 40 %)
Co-encadrement	:	**S. Moussaoui** (60 %)
Début	:	10/2011. Thèse actuellement suspendue pour des raisons médicales.
Publications	:	[CN.11, A.27]

[D.2] Charbel EL TANNOURY, « Surveillance de la pression des pneumatiques d'un véhicule automobile par analyse spectrale et synthèse d'observateurs », Ecole centrale de Nantes.

Financement	:	Convention CIFRE RENAULT-ECN
Directeur	:	F. Plestan (PR ECN, 50 %)
Co-encadrement	:	**S. Moussaoui** (50 %)
Début	:	12/2008
Soutenance	:	03/2012
Publications	:	[CI.16, CI.18, A.22, B.1, B.2, B.3, B.4, B.5]

[D.1] Emilie CHOUZENOUX, « Recherche de pas par majoration-minimisation pour la résolution itérative de problèmes inverses », Ecole centrale de Nantes.

Financement	:	Allocation MENRT
Directeur	:	J. Idier (DR CNRS, 50 %)
Co-encadrement	:	**S. Moussaoui** (50 %)
Début	:	10/2007
Soutenance	:	12/2010
Publications	:	[CN.6, CN.7, CI. 13, CI.15, CI.17, A.13, A.15, A.18, A.19]

1.6 Activité contractuelle, valorisation et transfert

J'ai participé à cinq contrats de recherche dont deux sont sous forme de financement institutionnels (projet jeunes chercheurs du GDR ISIS et Projet ANR) et trois contrats de collaborations industrielles, dont j'étais le responsable. Ces contrats sont résumés ci-dessous par ordre chronologique au sein de chacune de ces deux catégories.

1.6.1 Financements institutionnels

1.6.1.1 Méthodes MCMC pour l'analyse d'images hyperspectrales

- Projet Jeunes Chercheurs du GDR ISIS
- Durée : 2 ans (2007-2009),
- Responsable : **S. MOUSSAOUI**
- Autres chercheurs : Nicolas DOBIGEON, Martial COULON et Jean-Yves TOURNERET (IRIT, ENSEEIHT, Toulouse), Jérôme IDIER et Eric LE CARPENTIER (IRCCyN, ECN)

Résumé : L'analyse d'une image hyperspectrale a pour principal objectif la caractérisation qualitative et quantitative de la composition de la surface observée, en identifiant les constituants élémentaires et en déterminant leurs proportions. Ce projet vise à traiter ce problème dans un cadre bayésien en définissant des lois de probabilité *a priori* adéquates permettant de prendre en

compte explicitement toutes les contraintes auxquelles sont soumis les coefficients du mélange et les signatures spectrales des constituants. La première partie de cette étude s'est focalisée sur l'utilisation des méthodes de simulation stochastique pour réaliser l'estimation conjointe des spectres des constituants et des abondances. Il s'agit d'un problème de séparation de sources sous contraintes de non-négativité des sources et des coefficients de mélange. L'originalité de ce travail est d'intégrer, dans un premier temps, la contrainte de somme-à-un des coefficients de mélange dans le modèle de séparation bayésienne et de proposer une stratégie de simulation appropriée de la loi *a posteriori*. Dans un deuxième temps, une méthode plus adaptée au cas des images de réflectances a été développée pour exploiter une connaissance partielle sur les spectres des constituants obtenus par une méthode dite *d'extraction de pôles de mélange*.

1.6.1.2 Qualité, lissage et intégration au réseau de la production des houlogénérateurs électriques directs

- Projet ANR QUALIPHE (Référence. ANR-11-PRGE-0013)
- Durée : 3 ans (2012-2015),
- Responsable : Hamid BENAHMED (SATIE, ENS Rennes)
- Autres chercheurs : Christophe LALUC (doctorant), Alain CLÉMENT (LHEEA, ECN), Franck PLESTAN et Alain GLUMINEAU (IRCCyN, ECN)

Résumé : Le but du projet ANR QUALIPHE est d'optimiser la quantité et la qualité de l'énergie électrique issue d'un système houlogénérateur à conversion hydro-mécanique directe (c'est-à-dire sans stockage gravitaire ou mécanique tampon). Du point de vue « électrotechnique », une solution consiste à exploiter des moyens de stockage électrique locaux (embarqués) et/ou distants (mutualisés) pour lisser l'énergie électrique produite sans dégrader la productivité. Mais il faut également considérer la possibilité de profiter des effets de foisonnement des productions d'un ensemble de houlogénérateurs agencés en ferme pour poser ce problème dans toute son étendue. Le stockage électrique peut alors être envisagé comme intégré à chaque houlogénérateur et piloté de façon centralisée ou encore centralisé à l'échelle de la ferme. Les travaux proposés consistent, dans ces différentes situations, à maximiser, avec des considérations de cycle de vie, les deux critères (productivité et qualité de l'énergie) au moyen d'une gestion et d'un dimensionnement adéquats du stockage et d'une stratégie de commande optimale. Sur l'aspect « hydrodynamique », il s'agit de la problématique de la modélisation de mouvements de grande amplitude dans la tenue à la mer des flotteurs qui composent les systèmes houlomoteurs aussi bien en situation de survivabilité qu'en conditions opérationnelles. D'un point de vue « automatique », cette optimisation conjointe nécessite l'élaboration d'un modèle de la dynamique du système permettant la mise en œuvre de stratégies de commande optimale robuste. En effet, ces stratégies doivent être capables d'intégrer des critères de qualité de l'énergie déployée sur la réseau électrique et de prendre en compte la méconnaissance de l'évolution future de l'état de mer (mouvement des vagues).

1.6.2 Contrats de recherche industriels

1.6.2.1 Développement d'un système de surveillance de la pression de pneumatiques

- Contrat de collaboration ECN–RENAULT,
- Durée : 3 ans (2008-2011),
- Responsables : **S. MOUSSAOUI** et F. PLESTAN (ECN)
- Autres chercheurs : Charbel EL TANNOURY (Doctorant), Nicolas ROMANI (RENAULT).

Résumé : Ce contrat de recherche émane d'une demande exprimée par Renault et concerne la mise en œuvre d'un système de surveillance de la pression des pneumatiques (SSPP) permettant de détecter un sous-gonflage d'un ou de plusieurs pneumatiques tout en s'affranchissant des capteurs de pression dédiés. En effet, l'approche, dite « directe », qui consiste à utiliser des capteurs de pression placées dans les valves, s'avère onéreuse (installation, frais de maintenance) et peu fiable (pannes possibles des capteurs, des récepteurs). Les nouvelles générations de SSPP favorisent donc les méthodes « indirectes », c'est à dire sans capteur de pression. L'idée est d'assurer la surveillance de la pression dans les pneumatiques à partir de grandeurs physiques qui dépendent de la pression. En effet, la baisse de la pression se traduit par exemple par une augmentation de la vitesse angulaire de la roue, par un déplacement des modes vibratoires du véhicule, par une diminution du rayon effectif de la roue, par une augmentation de la résistance au roulement, etc. Pour être plus précis, le but est de développer deux types d'approches. Une approche dite « Signal » utilisant des méthodes basées sur l'analyse spectrale du signal de vitesse angulaire de chaque roue. La mise en oeuvre de toute la chaîne d'acquisition, de traitement du signal et de détection en un temps raisonnable, fût le premier objectif de cette thèse. Une approche « Système » utilisant des solutions basées sur le développement d'observateurs (capteurs logiciels) pour l'estimation de grandeurs physiques non mesurées ou, en l'occurrence, non mesurables, telles que le rayon effectif de la roue et sa résistance au roulement, dans le but de pouvoir diagnostiquer une perte de pression.

1.6.2.2 Optimisation des paramètres d'un système de surveillance de la pression de pneumatiques

- Contrat de collaboration ECN–RENAULT,
- Durée : 3 ans (2013-2016),
- Responsables : **S. MOUSSAOUI** et F. PLESTAN (ECN)
- Autres chercheurs : Joan DAVIS-VALLDAURA (Doctorant), Guillermo PITA-GIL (RENAULT).

Résumé : L'ATPMS (Advanced Tyre Pressure Monitoring System) est un système de surveillance de la pression des pneumatiques d'un véhicule sans utilisation de capteur de pression. Celui-ci détecte une baisse anormale de la pression à partir de l'analyse temporelle et fréquentielle des

signaux de vitesses angulaires des roues. La complexité du système et la grande taille de la base de données sur laquelle celui-ci doit être testé font que la mise au point manuelle du système (réglage des paramètres de mise en oeuvre) en un temps limité est quasiment impossible. L'objectif de ce contrat de recherche est de proposer une solution d'optimisation globale multi-objectif de fonctions multi-variables, coûteuses et complexes. De plus, étant donné que nous avons plusieurs sorties à optimiser (taux de fausses alarmes et non-détections), il convient de rechercher un ensemble de réglages conduisant au Front de Pareto.

La première partie du travail de recherche portera sur l'optimisation des paramètres du système ATPMS. Les paramètres à optimiser sont liés aux modules d'analyse temporelle et fréquentielle des signaux ainsi que du module de décision basé sur la fusion des deux analyses. Le problème d'optimisation est d'autant plus complexe que le coût de calcul du critère à minimiser est très élevé. La recherche se focalisera donc sur les méthodes d'optimisation globales de critères non-convexes et nécessitant un faible nombre d'évaluations du critère. La seconde partie portera sur la réalisation d'une étude prospective sur la synthèse d'un estimateur de la pression dans les pneumatiques. Pour cela, il faudra réaliser une recherche de modèles analytiques ou empiriques reliant une caractéristique dynamique de la roue à la pression du pneumatique. Un observateur sera ensuite mis en place pour fournir un bon indicateur d'une baisse anormale de la pression.

1.6.2.3 GPUs in Science Operations

- Contrat de recherche ECN–ESA,
- Durée : 18 mois (2012-2013),
- Responsable : **S. MOUSSAOUI**
- Autres chercheurs : Maxime LEGENDRE (Doctorant), Albrecht SCHMIDT (ESAC, Madrid).

Abstract : Recently, Graphics Cards have been used to offload scientific computations from traditional CPUs for greater efficiency. This project investigates the adaptation of a real-world linear system solver, which plays a central role in the data processing of the Science Ground Segment of ESA's astrometric Gaia mission. The paper quantifies the resource trade-offs between traditional CPU implementations and modern CUDA based GPU implementations. It also analyses the impact on the pipeline architecture and system development. The investigation starts from both a selected baseline algorithm with a reference implementation and a traditional linear system solver and then explores various modifications to control flow and data layout to achieve higher resource efficiency. It turns out that with the current state of the art, the modifications impact non-technical system attributes. For example, the control flow of the original modified Cholesky transform is modified so that locality of the code and verifiability deteriorate. The maintainability of the system is affected as well. On the system level, users will have to deal with more complex configuration control and testing procedures. The second application of this project concern the implementation of a spectral unmixing method to the processing of a large data set of hyperspectral images recorded during *Mars Express mission*.

1.7 Implication dans la vie collective

1.7.1 Collaborations

- D. Brie (PR, CRAN, univ. de Lorraine), C. Carteret (PR, LCPME, univ. Lorraine) : Séparation de sources, application en spectroscopie,
- E. Chouzenoux (MCF, LIGM, univ. Paris-Est) : Optimisation et problèmes inverses,
- N. Dobigeon (MCF, ENSEEIHT, Toulouse), J. Y. Tourneret (PR, ENSEEIHT, Toulouse) : Démélange non-supervisé d'images hyperspectrales
- T. Dudok de Wit (PR, Univ. Orléans), P. O. Amblard (DR CNRS, GIPSA-lab, Genoble) : Traitement de données d'observation du soleil,
- C. Jutten (PR, GIPSA-lab, Grenoble) : Séparation de sources dans le cas de mélanges non-linéaires et application en chimie.
- P. Launeau (PR, LPGN, univ. Nantes), M. Robin (PR, LETG, univ. Nantes) : fusion LIDAR et imagerie hyperspectrale pour l'analyse des végétations en environnement côtier,
- F. Mariette (DR IRSTEA, Rennes), C. Rondeau (CR HDR, IRSTEA, Rennes) : Sectroscopie RMN 2D,
- F. Schmidt (MCF HDR, IDES, univ. Paris-Sud) : Séparation de sources en planétologie,

1.7.2 Collaborations internationales

- L. Tomazeli-Duarte (Maitre Assistant, UNICAMP, Brésil) : séparation de sources dans le cas de mélanges non-linéaires,
- A. Schmidt (IR R&D, ESAC, Madrid) : Implémentation des méthodes d'imagerie hyperspectrale pour des problèmes de grande taille.

1.7.3 Animation scientifique

- Coordination, avec Nelly Pustelnik (CR CNRS, ENS Lyon), d'une action du GDR ISIS sur le thème de « l'optimisation en traitement du signal et de l'image ». Cette action consiste en l'organisation de trois journées portant sur le l'optimisation sous contraintes (22 octobre 2013) et l'optimisation non-convexe (28 mai 2014 et 16 octobre 2014),
- Participation au jury de thèse de Simon HENROT (Univ. Lorraine, 11/2014, Membre invité),
- Participation au jury de thèse de Leonardo-Tomazeli DUARTE (Univ. de Grenoble, 12/2010, Examinateur),
- Reviewer pour des revues internationales : IEEE Signal Processing Letters, IEEE Transactions on Signal Processing, IEEE Neural Networks, Signal Processing, Digital Signal Processing, IEEE Biomedical Engineering, IEEE Transactions on Geoscience and Remote sensing,
- Participation à un comité de sélection pour un recrutement de Maître de conférences (Univ. Marne-la-Vallée, avril-mai 2011).

1.7.4 Rayonnement scientifique et distinctions

- Best paper award lors du workshop IEEE WHISPERS'2012, Shanghai, Chine,
- Orateur invité, MAORI Workshop on Optimization for Image and Signal Processing, Ecole Polytechnique, Palaiseau, du 18 au 20 novembre 2013,
- Chercheur invité à l'agence spatiale européenne (ESA). Séminaire sur la séparation de sources spectrales en Astronomie. Madrid, Du 3 au 7 Août 2009,
- Titulaire de la prime d'excellence scientifique (2010-2014, évaluation du CNU : A).

1.7.5 Responsabilités

- Membre élu du conseil de laboratoire de l'IRCCyN, depuis 03/2008,
- Membre élu du conseil des études de l'ECN, depuis 10/2010,
- Coordinateur du comité de prospective de l'IRCCyN, depuis 01/2014,
- Responsable de l'option disciplinaire ISIS (ingénierie des systèmes, des images et des signaux), troisième année de la formation ingénieur à l'ECN, en 2013-2014.
- Responsable de l'option disciplinaire SIGMA (signaux-images), deuxième et troisième année de la formation ingénieur à l'ECN, depuis septembre 2014.

1.8 Production scientifique

Le tableau 1.2 donne une vue d'ensemble de mes principales publications.

Thème de recherche	Publications de rang A
1. Séparation de sources	
• cas de sources non-négatives	[A.2, A.3, A.12, O.1]
• cas de mélanges non-linéaires	[A.11, A.16]
• application à des données solaires	[A.4, A.21]
• imagerie hyperspectrale	[A.8, A.10, A.14]
2. Optimisation itérative	
• recherche de pas de descente	[A.15, A.19]
• reconstruction de distributions RMN 2D	[A.13]
3. Accélération algorithmique ou matérielle	
• intérêt de l'implementation GPU	[A.23]
• application en imagerie hyperspectrale	[A17. A.20, A.24]
4. Simulation bayésienne	
• cas de la séparation de sources	[A.5]
• cas de vecteurs gaussiens	[A.27]
5. Surveillance de la pression des pneumatiques d'un véhicule	
• estimation de la résistance au roulement	[A.22, B.1, B.2]
• traitement du signal du codeur ABS	[B.3, B.4, B.5]

TABLE 1.2 – Répartition de la production scientifique dans les principaux thèmes de recherche.

1.8.1 Chapitres dans un ouvrage scientifique

[O 1] **S. Moussaoui**, D. Brie et C. Carteret, « Bayesian approach to linear spectral mixture analysis », dans *Multivariate image processing*, Editeurs : C. Collet, J. Chanussot et K. Chehdi, pp. 143–168. John Wiley – ISTE, 2009.

1.8.2 Articles dans des revues internationales avec comité de lecture

[A.27] C. Gilavert, **S. Moussaoui** et J. Idier, « Efficient Gaussian sampling for solving large-scale inverse problems using MCMC », *IEEE Trans. on Signal Processing*, *à paraître*, 2014

[A.26] D. Brie, R. Klotz, S. Miron, **S. Moussaoui**, C. Mustin, P. Bécuwe, et S. Grandemange, « Joint analysis of flow cytometry data and fluorescence spectra as a non-negative array factorization problem », *Chemometrics and Intelligent Laboratory Systems*, vol. 137, pp. 21-32, 2014

[A.25] L.-T. Duarte, **S. Moussaoui** et C. Jutten, « Source separation in chemical analysis : recent achievements and perspectives », *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 135-146, 2014

[A.24] E. Chouzenoux, M. Legendre, **S. Moussaoui** et J. Idier, « Fast constrained least squares spectral unmixing using primal-dual interior point optimization », *IEEE Journal of Selected Togics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 1, pp. 59-69, 2014

[A.23] M. Legendre, A. Schmidt, **S. Moussaoui**, U. Lammers. « Solving Systems of Linear Equations by GPU-based Matrix Factorization in a Science Ground Segment », *Astronomy and Computing*, vol. 3-4, pp. 58-64, 2013.

[A.22] C. El Tannoury, **S. Moussaoui**, F. Plestan, N. Romani, G. Pita Gil, « Synthesis and application of nonlinear observers for the estimation of tire effective radius and rolling resistance of an automotive vehicle », *IEEE Trans. on Control Systems Technology*, vol. 21, no. 6, pp. 2408-2416, 2013.

[A.21] T. Dudok de Wit, **S. Moussaoui**, C. Guénnou, F. Auchère, G. Cessateur, M. Kretzschmar, L. E. Vieira, and F. Goryaev, « Coronal temperature maps from solar EUV images : a blind source separation approach », *Solar Physics*, vol. 283, no.1, pp. 31-47, 2013.

[A.20] E. Chouzenoux, **S. Moussaoui**, M. Legendre et J. Idier, « Algorithme primal-dual de points intérieurs pour l'estimation pénalisée des cartes d'abondances en imagerie hyperspectrale », *Traitement du Signal*, vol. 30, no.1-2, pp.35-59, 2013.

[A.19] E. Chouzenoux, **S. Moussaoui** et J. Idier, « Majorize-minimize linesearch for inversion methods involving barrier function optimization », *Inverse Problems*, vol. 28, 065011 (24 pages), 2012

[A.18] L. Chaâri, E. Chouzenoux, N. Pustelnik, C. Chaux et **S. Moussaoui**, « OPTIMED : Optimisation itérative pour la résolution de problèmes inverses de grande taille », *Traitement du Signal*, vol 28, no. 3-4, pp. 329-374, 2011

[A.17] F. Schmidt, A. Schmidt, E. Tréguier, M. Guiheneuf, **S. Moussaoui** et N. Dobigeon, « Implementation strategies for hyperspectral unmixing using Bayesian source separation », *IEEE Trans. on Geoscience and Remote Sensing*, vol. 48, no. 11, pp. 4003-4013, 2011

[A.16] L.-T. Duarte, C. Jutten et **S. Moussaoui**, « Bayesian source separation of linear and linear-quadratic mixtures using truncated priors », *Journal of Signal Processing Systems*, vol. 65, pp. 311-323, 2011

[A.15] E. Chouzenoux, J. Idier et **S. Moussaoui**, « A Majorize-Minimize strategy for subspace optimization applied to image restoration », *IEEE Trans. on Image Processing*, vol. 20, no.6, pp. 1517-1528, 2011.

[A.14] N. Dobigeon, **S. Moussaoui**, M. Coulon et J.-Y. Tourneret, « Bayesian algorithms for supervised, semi-supervised, and unsupervised unmixing of hyperspectral images », *Traitement du signal*, vol. 27, no. 1, pp. 79-108, 2010

[A.13] E. Chouzenoux, **S. Moussaoui**, J. Idier et F. Mariette, « Efficient maximum entropy reconstruction of T1-T2 spectra », *IEEE Trans. on Signal Processing*, vol. 58, no. 12, 2010

[A.12] N. Dobigeon, **S. Moussaoui**, J.-Y. Tourneret, C. Carteret, « Bayesian separation of spectral sources under non-negativity and full additivity constraints », *Signal Processing*, vol. 89, no. 12, pp. 2657-2669, 2009.

[A.11] L.-T. Duarte, C. Jutten, **S. Moussaoui**, « A Bayesian non-linear source separation method for smart ion-selective electrode arrays », *IEEE Sensors Journal*, vol. 9, no. 12, pp. 1763-1771, 2009.

[A.10] N. Dobigeon, **S. Moussaoui**, M. Coulon, J.-Y. Tourneret, A. Hero, « Joint Bayesian endmember extraction and linear unmixing for hyperspectral imagery », *IEEE Trans. on Signal Processing*, vol. 57, no. 11, pp. 4355-4368, 2009.

[A.9] C. Carteret, A. Dandeu, **S. Moussaoui**, H. Muhr, B. Humbert, E. Plasari, « Polymorphism studied by lattice phonon Raman spectroscopy and statistical mixture analysis method », *Crystal Growth and Design*, vol. 9, no. 2, pp. 807-812, 2009.

[A.8] **S. Moussaoui**, H. Hauksdottir, F. Schmidt, C. Jutten, J. Chanussot, D. Brie, S. Doute et J.A. Benediksson, « On the decomposition of Mars hyperspectral data by ICA and Bayesian positive source separation », *Neurocomputing*, vol. 71, no. 10, pp. 2194-2208, 2008.

[A.7] L. Guillemot, Y. Gaudeau, **S. Moussaoui** et J.-M. Moureaux, « Entropy-coded lattice vector quantization dedicated to the block mixture densities », *IEEE Trans. on Image Processing*, vol. 17, no. 9, pp. 1574-1586, 2008.

[A.6] J. Lilenstein, T. Dudok De Wit, M. Kretzschmar, P. O Amblard, **S. Moussaoui**, J. Aboudarham et F. Auchère, « Review on the solar spectral variability in the EUV for space weather purposes », *Annales Geophysicae*, vol. 26, no. 2, pp. 269-279, 2008.

[A.5] T. Veit, J. Idier et **S. Moussaoui**. « Rééchantillonnage de l'échelle dans les algorithmes MCMC pour les problèmes inverses bilinéaires », *Traitement du Signal*, vol. 25, no. 4. pp. 329-343, 2008.

[A.4] P. O Amblard, **S. Moussaoui**, T. Dudok De Wit, J. Aboudarham, M. Kretzschmar, J. Lilenstein et F. Auchère, « The EUV Sun as the superposition of elementary Suns », *Astronomy and Astrophysics*, vol. 487, no. 2, pp. L13-L16, 2008

[A.3] **S. Moussaoui**, D. Brie, A. Mohammad-Djafari et C. Carteret, « Separation of non-negative mixture of non-negative sources using a Bayesian approach and MCMC sampling », *IEEE Trans. on Signal Processing*, vol. 54 , no. 11, pp. 4133–4145, 2006.

[A.2] **S. Moussaoui**, C. Carteret, D. Brie et A. Mohammad-Djafari. « Bayesian analysis of spectral mixture data using Markov chain Monte Carlo Methods », *Chemometrics and Intelligent Laboratory Systems*, vol. 81, no. 2, pp. 137-148, 2006.

[A.1] **S. Moussaoui**, D. Brie et A. Richard, « Regularization aspects in continuous-time model identification », *Automatica*, vol. 41, o. 2, pp. 197-208 , 2005

1.8.3 Brevets répertoriés dans la base de données Esp@cenet de l'INPI

[B.5] C. El Tannoury, **S. Moussaoui** et G. Pita-Gil, « Procédé de détection du sens de rotation d'une roue utilisant un capteur de vitesse non signé », INPI, FR 2988850 (10-04-2014), Renault SAS, France

[B.4] C. El Tannoury, **S. Moussaoui** et G. Pita-Gil, « Perfectionnement de la mesure de vitesse de rotation d'une roue », INPI, FR 2988848 (04-10-2013), Renault SAS, France

[B.3] C. El Tannoury, **S. Moussaoui**, G. Pita-Gil et F. Plestan, « Procédé d'estimation de la résistance au roulement de roues équipant un train d'un véhicule », INPI, FR 2988645 (04-10-2013), WO 2013144469 (21-03-214), Renault SAS, France

[B.2] C. El Tannoury, **S. Moussaoui**, G. Pita-Gil, F. Plestan et N. Romani, « Procédé d'estimation de la résistance au roulement d'une roue de véhicule », INPI, FR 2980573 (29-03-2013), WO 2013041802 (01-04-2014) , Renault SAS, France

[B.1] C. El Tannoury, **S. Moussaoui**, G. Pita-Gil, F. Plestan et N. Romani, « Estimation du rayon dynamique d'une roue et de la vitesse d'un véhicule automobile », INPI, FR 2973115 (28-09-2012), WO 2012127139 (12-04-2013), Renault SAS, France

1.8.4 Conférences internationales avec actes et comité de lecture

[CI.24] M. Legendre, **S. Moussaoui**, E. Chouzenoux et J. Idier, Primal-dual interior-point optimization based on majorization-minimization for edge-preserving spectral unmixing, dans *Proc. of IEEE International Conference on Image Processing (ICIP)*, Paris, France, Octobre 2014.

[CI.23] M. Legendre, **S. Moussaoui**, F. Schmidt et J. Idier, Parallel implementation of a primal-dual interior-point optimization method for fast abundance maps estimation, dans *Proc. of IEEE International Workshop on Hyperspectral Image and Signal Processing (WHISPERS)*, Gainesville, États-Unis, Juin 2013.

[CI.22] S. Henrot, **S. Moussaoui**, C. Soussen et D. Brie, Edge-preserving nonnegative hyperspectral image restoration, dans *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, Mai 2013.

[CI.21] E. Chouzenoux, **S. Moussaoui**, J. Idier et F. Mariette. Primal-dual interior point optimization for a regularized reconstruction of NMR relaxation time distributions, dans *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, Mai 2013.

[CI.20] **S. Moussaoui**, E. Chouzenoux et J. Idier, Primal-dual interior point optimization for penalized least squares estimation of abundance maps in hyperspectral imaging, dans *Proc. of IEEE International Workshop on Hyperspectral Image and Signal Processing (WHISPERS)*, Shanghai, Chine, Juin 2012

[CI.19] Y. Gaudeau, L. Guillemot, **S. Moussaoui** et J.-M. Moureaux, Low complexity bit allocation based on a multidimensional mixture model using lattice vector quantization, dans *Proc. of Picture Coding Symposium (PCS)*, Cracovie, Pologne, Mai 2012

[CI.18] C. El Tannoury, F. Plestan, **S. Moussaoui** et G. Pita-Gil, Rolling resistance monitoring by using software sensors, dans *Proc. of 12th International Workshop on Variable Structure Systems (VSS)*, Bombay, Inde, Janvier 2012

[CI.17] E. Chouzenoux, **S. Moussaoui** et J. Idier. Efficiency of linesearch strategies in interior point methods for linearly constrained signal restoration, dans *Proc. of IEEE International Workshop on Statistical Signal Processing (SSP)*, Nice, France, Juin 2011

[CI.16] C. El Tannoury, F. Plestan, **S. Moussaoui** et N. Romani, « Tyre effective radius and vehicle velocity estimation : a variable structure observer solution », dans *Proc. of IEEE International Multi-Conference on Systems, Signals and Devices (SSD)*, Sousse, Tunisie, Mars 2011.

[CI.15] E. Chouzenoux, **S. Moussaoui**, J. Idier et F. Mariette, « Optimization of a maximum entropy criterion for 2D nuclear magnetic resonance reconstruction », dans *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, Etats-Unis, Mars 2010

[CI.14] L.-T. Duarte, C. Jutten, and **S. Moussaoui**. « Bayesian source separation of linear-quadratic and linear mixtures through a MCMC method », dans *Proc. International Workshop on Machine Learning for Signal Processing (MLSP)*, Grenoble, France, Septembre 2009

[CI.13] E. Chouzenoux, **S. Moussaoui** et J. Idier. « A majorize-minimize line search algorithm for barrier function optimization », dans *Proc. of European Signal Processing Conference (EUSIPCO)*, Glasgow, UK, Aout 2009

[CI.12] T. Dudok de Wit, **S. Moussaoui**, P.-O. Amblard, J. Aboudarham, F. Auchere, M. Kretzschmar et J. Lilensten, « Multispectral imaging the sun in the Ultraviolet », dans *Proc. of IEEE International Workshop on Hyperspectral Image and Signal Processing (WHISPERS)*, Grenoble, Aout 2009.

[CI.11] F. Schmidt, **S. Moussaoui** et N. Dobigeon, Material identification on martian hyperspectral images using Bayesian source separation, dans *Proc. of IEEE International Workshop on Hyperspectral Image and Signal Processing (WHISPERS)*, Grenoble, Aout 2009.

[CI.10] L.-T. Duarte, C. Jutten, and **S. Moussaoui**. ≪ Ion-selective electrode array based on a Bayesian nonlinear source separation method ≫, dans *Proc. International Conference on Independent Component Analysis and Signal Separation (ICA)*, Paraty, Brésil, Mars 2009.

[CI.9] N. Dobigeon, **S. Moussaoui** et J.Y. Tourneret. Blind unmixing of linear mixtures using a hierarchical Bayesian model. dans *Proc. of IEEE International Workshop on Statistical Signal Processing (SSP)*, Madison, Etats-Unis, 2007.

[CI.8] C. Jutten, **S. Moussaoui** et F. Schmidt. How to apply ICA on actual data ? Example of Mars hyperspectral image analysis. Dans *Proc. of 15th IEEE International conference on Digital Signal Processing (DSP)*, Cardiff, Royaune-Uni, 2007.

[CI.7] L. Guillemot, Y. Gaudeau, **S. Moussaoui**, J.-M. Moureaux. An analytical gamma mixture based rate-distortion model for lattice vector quantization. dans *Proc. of European Signal Processing Conference (EUSIPCO)*, Firenze, Italie, 2006.

[CI.6] M. Kretzschmar, **S. Moussaoui**, J. Lilensten, T. Dudok De Wit, F. Auchère, P.O. Amblard, Aboudarham J. ≪ Decomposition of the Solar EUV irradiance in spectral sources ≫, dans *3rd European Space Weather Week (ESWW)*, Belgique, 2006.

[CI.5] H. Hauksdottir, **S. Moussaoui**, F. Schmidt, C. Jutten, J. Chanussot et D. Brie, ≪ Mars Hyperspectral Data Processing using ICA and Bayesian Positive Source Separation ≫, dans *Proc. of International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MAXENT)*, Paris, France, Juillet 2006.

[CI.4] **S. Moussaoui**, D. Brie et C. Carteret, ≪ Non-negative source separation using the maximum likelihood approach ≫, dans *Proc. of IEEE International Workshop on Statistical Signal Processing (SSP)*, Bordeaux, France, Juillet 2005

[CI.3] **S. Moussaoui**, D. Brie et J. Idier, ≪ Non-negative source separation : range of admissible solutions and conditions for the uniqueness of the solution ≫, dans *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Philadelphia, Etats-Unis, Mars 2005

[CI.2] **S. Moussaoui**, D. Brie, C. Carteret, A. Mohammad-Djafari, ≪ Application of Bayesian non-negative source separation to mixture analysis in spectroscopy ≫, dans *Proc. of International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MAXENT)*, Garching, Allemagne, Juillet 2004

[CI.1] **S. Moussaoui**, D. Brie, O. Caspary et A. Mohammad-Djafari, ≪ A Bayesian method for positive source separation ≫, dans *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Montreal, Canada, Mai 2004

1.8.5 Conférences nationales avec actes et comité de lecture

[CN.11] C. Gilavert, **S. Moussaoui** et J. Idier, « Rééchantillonnage gaussien en grande dimension pour les problèmes inverses », dans Actes du 24ème Colloque GRETSI, Brest, France, Septembre 2013.

[CN.10] S. Henrot, C. Soussen, **S. Moussaoui** et D. Brie, « Restauration positive d'images hyperspectrales avec préservation des contours », dans Actes du 24ème Colloque GRETSI, Brest, France, Septembre 2013.

[CN.9] C. Laluc, **S. Moussaoui**, J. Idier et A. Clément, « Prédiction court-terme de la hauteur des vagues pour la commande d'un houlogénérateur », dans Actes du 24ème Colloque GRETSI, Brest, France, Septembre 2013.

[CN.8] M. Legendre, **S. Moussaoui**, F. Schmidt et J. Idier, Implémentation parallèle d'une méthode d'estimation des cartes d'abondances en imagerie hyperspectrale, dans Actes du 24ème Colloque GRETSI, Brest, France, Septembre 2013.

[CN.7] E. Chouzenoux, **S. Moussaoui** et J. Idier, « Algorithme primal-dual de points intérieurs pour l'estimation pénalisée des cartes d'abondances en imagerie hyperspectrale », dans Actes du 23ème Colloque GRETSI, Bordeaux, France, Septembre 2011.

[CN.6] E. Chouzenoux, **S. Moussaoui**, J. Idier et F. Mariette. « Reconstruction d'un spectre RMN 2D par maximum d'entropie », dans Actes du 22ème Colloque GRETSI, Dijon, France, Septembre 2009.

[CN.5] N. Dobigeon, **S. Moussaoui**, M. Coulon et J.-Y. Tourneret, « Extraction de composants purs et démélange linéaire bayésiens en imagerie hyperspectrale », dans Actes du 22ème Colloque GRETSI, Dijon, France, Septembre 2009.

[CN.4] L.-T. Duarte, C. Jutten et **S. Moussaoui**, « Séparation de sources dans le cas de mélanges linéaires quadratiques et linéaires par une approche bayésienne », dans Actes du 22ème Colloque GRETSI, Dijon, France, Septembre 2009.

[CN.3] N. Dobigeon, **S. Moussaoui** et J.Y. Tourneret, « Séparation bayéesienne de sources spectrales sous contraintes de positivité et d'additivité », dans Actes du 21ème colloque GRETSI, Troyes, Septembre 2007.

[CN.2] P. O Amblard, **S. Moussaoui**, T. Dudok De Wit, J. Lilenstein, M. Kretzschmar, J. Aboudarham et F. Auchère, « Le soleil comme superposition de soleils élémentaires », dans Actes du 21ème colloque GRETSI, Troyes, Septembre 2007.

[CN.1] **S. Moussaoui**, D. Brie et C. Carteret, « Séparation de sources non-négatives par l'approche du maximum de vraisemblance », dans Actes du 20ème Colloque GRETSI, Louvain-la-Neuve, Belgique, Septembre 2005

Deuxième partie

Synthèse des activités de recherche

Introduction générale

Dans divers domaines de l'instrumentation industrielle et scientifique, toute opération de mesure est source de signal associé à une information d'intérêt. Il est cependant assez fréquent que cette information n'apparaisse dans le signal brut que sous une forme qui dépend du principe de transduction au sein du capteur. Le traitement du signal a pour objectif de développer des techniques de transformation, de représentation, d'analyse et d'interprétation des signaux pour extraire l'information utile, ou parfois rejeter une information indésirable. On parle de *problème inverse* en traitement du signal dès lors que l'on cherche à traiter le signal fourni par un capteur afin de retrouver l'information d'intérêt.

Résolution d'un problème inverse : enjeux et difficultés

La *résolution d'un problème inverse* consiste à exploiter la connaissance des équations physiques régissant le capteur et le processus de mesure pour formuler un modèle mathématique permettant de mettre en œuvre une méthode numérique visant à restituer l'information d'intérêt. Il s'agit donc d'un champ de recherche à l'interface des mathématiques, de la physique, de l'analyse numérique et de l'instrumentation.

Considérons, à titre d'exemple, un problème inverse linéaire, au sens où le lien entre le signal d'intérêt $\boldsymbol{x}^o \in \mathbb{R}^N$ et les données observées $\boldsymbol{y} \in \mathbb{R}^M$ peut être représenté par un modèle linéaire,

$$\boldsymbol{y} = \boldsymbol{K}\boldsymbol{x}^o + \boldsymbol{e}, \tag{1.1}$$

avec $\boldsymbol{K} \in \mathbb{R}^{M \times N}$ une matrice issue de la discrétisation du processus d'observation et $\boldsymbol{e}$ un vecteur formé des M échantillons du bruit additif représentant les erreurs de mesure, de discrétisation et de modélisation. De plus, $\boldsymbol{x}$ est un vecteur contenant les N échantillons du signal d'intérêt, $\boldsymbol{y}$ un vecteur des M échantillons du signal mesuré. En réalité ce modèle linéaire peut être vu comme une approximation de premier ordre d'un modèle non-linéaire. Néanmoins, cette formulation permet de modéliser plusieurs situations réelles telles que le débruitage [Frieden 75], la déconvolution [Jansson 97], la tomographie [Gordon 71] et la séparation de sources [Jutten 91].

Une approche naïve de l'inversion consiste à se servir de la connaissance du modèle direct pour rechercher une solution $\boldsymbol{x}$ permettant de reproduire le plus fidèlement possible les observations. Cette fidélité étant évaluée par un critère dit *d'adéquation aux données* (de type moindres carrés, divergence de Kullback-Leibler ou toute autre fonction de contraste). Cependant, une telle approche se heurte à des difficultés, dues au caractère *mal-posé* du problème inverse, qui se traduisent le plus souvent par une *instabilité* ou une *non-unicité* de la solution [Tikhonov 77].

La résolution d'un problème inverse mal-posé fait appel au principe de *régularisation* [Tikhonov 63] ou encore à une *inférence statistique bayésienne* [Demoment 89, Jaynes 03, Idier 08] qui consistent en la prise en compte de contraintes traduisant certaines propriétés ou hypothèses sur la solution recherchée.

Questions liées aux outils d'optimisation

Une formulation lagrangienne des contraintes associées au problème permet d'obtenir un critère composite qui prend en compte conjointement la fidélité aux données observées $\boldsymbol{y}$ et le respect des connaissances préalables sur la solution recherchée $\boldsymbol{x}$. Il en résulte ainsi un critère composite

$$F(\boldsymbol{x}) = J(\boldsymbol{x}; \boldsymbol{y}) + \beta R(\boldsymbol{x}), \quad \beta \geqslant 0. \tag{1.2}$$

Le premier terme $J(\cdot)$ mesure l'adéquation aux données, $R(\cdot)$ est le critère de *régularisation* (appelé aussi de *pénalisation*) qui permet de quantifier le degrés de non respect des contraintes par la solution et β est un *paramètre de régularisation* dont la valeur permet d'ajuster le compromis entre les deux parties du critère. La résolution du problème inverse se réduit donc à la recherche d'un minimiseur du critère composite (1.2). Cependant, cette solution peut rarement s'exprimer sous une forme analytique et un algorithme d'optimisation doit alors être mis en œuvre pour en fournir une approximation.

Cependant, bien que cette approche pénalisée fournisse des solutions de qualité satisfaisante, sa mise en œuvre dans le cadre de problèmes de grande taille a souvent pour inconvénient de nécessiter un temps de calcul et/ou un besoin en ressources mémoire trop importants. On peut citer pour exemple les problèmes inverses où le modèle direct est de taille très grande, tels que la reconstruction tomographique 3D ou la spectroscopie RMN 2D/3D.

Le recours à des techniques d'optimisation itérative offre un cadre favorable à la résolution de problèmes de grande taille. Par contre, le coût de calcul par itération et le nombre d'itérations avant convergence sont des facteurs déterminants pour le choix de la stratégie d'optimisation. Dans le cadre de l'optimisation de critères différentiables, une grande famille d'algorithmes est basée sur une stratégie de descente itérative alternant une étape de calcul d'une *direction de descente*, définissant la direction de recherche de la solution, et une autre étape de *recherche d'un pas* quantifiant l'avancée nécessaire le long de la direction choisie.

D'un point de vue méthodologique, les questions qui se posent dans le cadre des problèmes de grande taille sont :

1. Comment choisir une direction de descente permettant une convergence rapide de l'algorithme ?
2. Comment calculer efficacement un pas de descente adapté à la structure du critère composite de manière à garantir le meilleur taux de convergence de la méthode de descente, avec un coût de calcul acceptable ?
3. Comment tirer profit des outils de calcul intensif pour accélérer le temps d'exécution des algorithmes ? Le défi majeur est de développer des méthodes présentant des structures algorithmiques qui se prêtent à une implémentation parallèle efficace.

Il est évident qu'il n'existe pas de réponses unanimes à ces questions. Néanmoins, des principes méthodologiques fondés sur des outils mathématiques éprouvés doivent être employés pour apporter des réponses adaptées aux spécificités de chaque problème à résoudre.

Questions liées aux outils de simulation bayésienne

Un problème inverse mal-posé est assez souvent résolu dans le cadre de l'inférence statistique bayésienne, ce qui sous-entend une modélisation statistique du problème d'inversion en incluant les connaissances préalables sur la solution recherchée. En effet, une formulation probabiliste des propriétés statistiques du bruit de mesure, à l'aide d'une densité de probabilité $P_{\boldsymbol{E}}(\cdot)$, et une représentation des propriétés souhaitées du signal d'intérêt à l'aide d'un modèle de densité de probabilité *a priori* $P_{\boldsymbol{X}}(\cdot)$ permet, grâce à la règle de Bayes, d'exprimer la densité de probabilité *a posteriori* de l'information d'intérêt

$$P_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{y}) = \frac{P_{\boldsymbol{E}}(\boldsymbol{y}|\boldsymbol{x})\ P_{\boldsymbol{X}}(\boldsymbol{x})}{P_{\boldsymbol{Y}}(\boldsymbol{y})}. \tag{1.3}$$

La résolution du problème inverse par inférence bayésienne se réduit ainsi au calcul d'un estimateur bayésien à partir de cette densité de probabilité *a posteriori*. Pour cela il devient nécessaire de définir un estimateur et de le calculer. Cependant, comme l'estimateur ne peut généralement pas être obtenu sous une forme analytique (loi *a posteriori* difficile à expliciter ou que le calcul analytique de l'estimateur n'est pas possible), la résolution fait appel à des méthodes de *Monte Carlo* dont une grande famille s'appuie sur la construction d'une *chaîne de Markov*.

Dans le cadre de la résolution de problèmes inverses de grande taille par inférence bayésienne, se posent alors des questions telles que :

1. Comment définir les modèles *a priori* de façon pertinente au regard de l'information que l'on souhaite représenter ?
2. Quelle est la stratégie de simulation bayésienne à mettre en œuvre dans le cas d'un problème de grande taille ?

3. Comment réduire le temps de simulation des méthodes de Monte Carlo par chaînes de Markov appliqués à la résolution de problèmes inverses ?

Je vais détailler dans la suite de ce manuscrit quelques contributions issues de travaux de thèses de doctorat, que j'ai co-encadrés, en lien avec la résolution de problèmes inverses de grande taille et visant à apporter quelques réponses aux questions précédentes. J'ai choisi de ne pas détailler dans ce manuscrit d'habilitation à diriger des recherches, mes travaux en séparation de sources, que j'ai réalisés en partie durant ma thèse puis poursuivi dans le cadre de collaborations. Celles-ci concernent la proposition de modèles bayésiens adaptés pour la prise en compte de la contrainte de somme à un des coefficients de mélange [Dobigeon 09b, Dobigeon 09a], la séparation de mélanges non-linéaires [Duarte 09, Duarte 10] ou encore d'application à des données issues de la surveillance de l'activité solaire [Amblard 08, Dudok de Wit 13].

La synthèse de mes contributions sur les outils de résolution de problèmes inverses de grande taille est organisée en trois chapitres :

1. **Chapitre 2.** Algorithmes de recherche de pas de descente par majoration-minimisation dans le cadre de la résolution itérative de problèmes inverses. Ces résultats ont été obtenus durant la thèse de Emilie CHOUZENOUX.
2. **Chapitre 3.** Accélération algorithmique et matérielle des méthodes d'optimisation itérative sous contraintes. Les développements ont été réalisés dans la thèse de Maxime LEGENDRE,
3. **Chapitre 4.** Simulation bayésienne en grande dimension appliquée à l'échantillonnage de vecteurs gaussiens. La technique d'échantillonnage proposée est issue de la thèse de Clément GILAVERT.

Une description succincte du problème traité et des outils utilisés est réalisée au début de chaque chapitre pour permettre au lecteur de mieux cerner les contributions de mes travaux.

Sur le plan des applications, ces contributions ont permis de proposer une méthode d'estimation des distributions des temps de relaxation en spectroscopie de résonance magnétique nucléaire (RMN). Il s'agit d'un problème d'inversion numérique d'une transformée de Laplace 2D par maximum d'entropie. La seconde application concerne le démélange linéaire d'images hyperspectrales pour la télédétection aéroportée. Deux publications qui détaillent ces applications sont annexées à ce manuscrit.

Chapitre 2

Recherche de pas de descente itérative par majoration-minimisation

On s'intéresse à la résolution d'un problème inverse mal-posé par la recherche du minimiseur d'un critère pénalisé $F(\cdot)$, que l'on supposera tout au long de ce manuscrit convexe et différentiable. Ce problème se résume par,

$$\begin{aligned} &\text{Trouver} \quad \hat{\boldsymbol{x}} \\ &\text{tel que} \quad \hat{\boldsymbol{x}} = \underset{\boldsymbol{x} \in \mathcal{D}_f}{\operatorname{argmin}}\, F(\boldsymbol{x}), \end{aligned} \tag{2.1}$$

où $\mathcal{D}_f = \mathbf{dom}\, F$ est le domaine de définition du critère $F(\cdot)$, supposé non vide. Il est important de signaler l'existence de travaux dans la communauté du traitement du signal et de l'image qui s'intéressent aux cas de critères non-convexes et/ou non-différentiables. Les critères non-convexes admettent souvent des minimas locaux et sont souvent traités par des méthodes d'optimisation dédiées telles que le recuit simulé [Geman 84], les algorithmes génétiques [Holland 92] ou encore en s'appuyant sur des méthodes de relaxation du critère non-convexe par des critères convexes [Chambolle 12]. Pour ce qui est de la différentiabilité, il existe également des méthodes d'optimisation spécifiques telles que les méthodes proximales [Combettes 10] ainsi que d'autres travaux qui consistent à approcher chaque terme non-différentiable du critère par une fonction différentiable paramétrique [Nesterov 05], de telle manière à ce qu'une valeur asymptotique du paramètre conduise au cas du critère non-différentiable.

Un approche très couramment utilisée pour rechercher une solution du problème d'optimisation (2.1) s'appuie sur la technique de descente itérative [Ortega 70], dont sont issus les algorithmes de gradient, de Newton et leurs différentes variantes [Kelly 99]. Le recours à une approche d'optimisation itérative dans le cadre de la résolution de problèmes inverses se justifie, d'une part, par la non existence d'une solution explicite du problème, et d'autre part par une réduction significative de la complexité arithmétique et du coût mémoire en évitant les inversions matricielles et en exploitant des implémentations moins coûteuses du problème direct.

Un problème commun à tous ces algorithmes est la recherche du pas d'avancement le long de la direction de descente à chaque itération. Après une brève description du problème de recherche de pas et un rappel de quelques techniques, ce chapitre présente les contributions liées à la proposition de méthodes s'appuyant sur la minimisation d'approximations majorantes de la fonction de recherche de pas.

2.1 Recherche de pas pour la descente itérative

Le principe des méthodes de descente itérative est très simple : partant d'un point initial $\boldsymbol{x}_0$, une suite d'itérées $\{\boldsymbol{x}_k\}$ qui converge vers une solution du problème d'optimisation (2.1) est produite. La décroissance du critère est assurée en déplaçant l'itérée courante $\boldsymbol{x}_k$ le long d'une direction $\boldsymbol{d}_k$, dite *de descente*,

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k, \tag{2.2}$$

où $\alpha_k > 0$ est le *pas de descente* et $\boldsymbol{d}_k$ une *direction de descente* satisfaisant

$$\boldsymbol{g}_k^{\mathrm{t}} \boldsymbol{d}_k < 0, \tag{2.3}$$

où $\boldsymbol{g}_k = \nabla F(\boldsymbol{x}_k)$. Un algorithme de descente itérative va donc alterner les étapes de construction de $\boldsymbol{d}_k$ et de détermination de α_k jusqu'à la satisfaction d'un test d'arrêt portant sur la petitesse du gradient $\boldsymbol{g}_k$.

2.1.1 Choix de la direction de descente

La construction de $\boldsymbol{d}_k$ est souvent réalisée en s'appuyant sur des considérations locales sur le critère $F(\cdot)$, en termes de gradient et de hessien, ou encore en ajoutant les directions de descente choisies lors des itérations précédentes (gradient conjugué, mémoire de gradient). Le tableau (2.1) donne quelques exemples de directions de descente les plus couramment utilisées. La stratégie de choix de la direction doit être élaborée afin de favoriser une convergence rapide de l'algorithme. Cependant, le coût de calcul par itération, notamment pour les problèmes de grande taille, est un facteur déterminant pour ce choix.

Une formulation plus générale de la technique de descente itérative consiste en la descente dans un sous-espace engendré par un ensemble de L directions de descente $\left\{\boldsymbol{d}_k^1, \ldots, \boldsymbol{d}_k^L\right\}$,

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{D}_k \boldsymbol{\alpha}_k \tag{2.4}$$

où $\boldsymbol{D}_k = [\boldsymbol{d}_k^1, \ldots, \boldsymbol{d}_k^L]$ et $\boldsymbol{\alpha}_k \in \mathbb{R}^L$ est un pas multi-dimensionnel. Le but de cette extension est d'augmenter la vitesse de convergence de l'algorithme de descente tout en ayant un coût maîtrisé par itération. Pour garantir cette efficacité, il faut bien évidemment prêter attention au choix du

Algorithme	Direction $\boldsymbol{d}_k$	Remarques
Gradient	$-\boldsymbol{g}_k$	$\boldsymbol{g}_k = \nabla F(\boldsymbol{x}_k)$, gradient du critère en $\boldsymbol{x}_k$
Gradient conjugué	$-\boldsymbol{g}_k + \beta_k \boldsymbol{d}_{k-1}$	β_k est le coefficient de conjugaison
Gradient conjugué préconditionné	$-\boldsymbol{M}\boldsymbol{g}_k + \beta_k \boldsymbol{d}_{k-1}$	$\boldsymbol{M}$ est une matrice de préconditionnement
Newton	$-\boldsymbol{H}_k^{-1}\boldsymbol{g}_k$	$\boldsymbol{H}_k = \nabla^2 F(\boldsymbol{x}_k)$, hessien du critère en $\boldsymbol{x}_k$
Quasi-Newton	$-\boldsymbol{B}_k^{-1}\boldsymbol{g}_k$	$\boldsymbol{B}_k$ est une approximation du hessien
Newton tronqué	$\boldsymbol{d}_k^{\ell}$	$\boldsymbol{d}_k^{\ell}$ est la ℓième itérée d'un gradient conjugué résolvant $\boldsymbol{H}_k\boldsymbol{d} = -\boldsymbol{g}_k$

TABLE 2.1 – Directions de descente utilisées dans quelques algorithmes itératifs

sous-espace de directions de descente. Le tableau (2.2) donne des exemples d'ensembles de directions utilisées dans quelques algorithmes de sous-espace.

Algorithme	Ensemble de directions
Memoire de gradient	$\{-\boldsymbol{g}_k, \boldsymbol{d}_{k-1}\}$
Super-mémoire de gradient	$\{-\boldsymbol{g}_k, \boldsymbol{d}_{k-1}, \ldots, \boldsymbol{d}_{k-m}\}$
Super-mémoire de directions	$\{-\boldsymbol{g}_k, -\boldsymbol{g}_{k-1}, \ldots, -\boldsymbol{g}_{k-m}, \}$
Sous-espace orthogonal	$\left\{ -\boldsymbol{g}_k, \boldsymbol{x}_k - \boldsymbol{x}_0, \sum_{i=1}^{k} \omega_i \boldsymbol{d}_{k-i} \right\}$; ω_i coefficients constants
Super-mémoire de gradient et sous-espace orthogonal	$\left\{ -\boldsymbol{g}_k, \boldsymbol{d}_{k-1}, \ldots, \boldsymbol{d}_{k-m}, \boldsymbol{x}_k - \boldsymbol{x}_0, \sum_{i=1}^{k} \omega_i \boldsymbol{d}_{k-i} \right\}$
Directions de Newton tronqué et super-mémoire de gradient	$\left\{ \boldsymbol{d}_k^{\ell}, \boldsymbol{H}_k \boldsymbol{d}_k^{\ell} + \boldsymbol{g}_k, \boldsymbol{d}_k^{\ell} - \boldsymbol{d}_k^{\ell-1}, \boldsymbol{d}_{k-1}, \ldots, \boldsymbol{d}_{k-m} \right\}$

TABLE 2.2 – Ensembles de directions utilisées dans quelques algorithmes de sous-espaces.

Ce schéma de descente itérative a donné lieu à plusieurs contributions [Miele 69, Cragg 69, Wolfe 76, Shi 05, Zibulevsky 10], présentées sous différentes dénominations bien qu'exploitant le même principe, qui se distinguent essentiellement par la technique de constitution du sous-espace de directions. Une synthèse bibliographique sur ces algorithmes peut être trouvée dans le papier [Chouzenoux 11a, sec. II], annexé à ce manuscrit. Cependant, ces techniques ont été peu utilisées dans la communauté du traitement du signal et des images. Des extensions et des applications récentes de ces algorithmes peuvent être trouvées dans [Chouzenoux 13a, Florescu 14].

2.1.2 Recherche de pas

La détermination de la valeur de α_k, opération dite de *recherche de pas*, consiste à réaliser une minimisation approchée de la fonction scalaire $f(\alpha)$=$F(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k)$ ou encore la minimisation de $f(\boldsymbol{\alpha}) = F(\boldsymbol{x}_k + \boldsymbol{D}_k\boldsymbol{\alpha})$ dans le cas de descente dans un sous-espace de directions. Le choix de la valeur du pas doit se faire de sorte à assurer une décroissance suffisante du critère tout en garantissant la convergence de l'algorithme de descente itérative [Bertsekas 99, Kelly 99].

Stratégie de pas	Condition sur α_k	
Règle de Cauchy	$\arg\min_{\alpha\in\mathbb{R}^+} f(\alpha)$	
Règle de Cauchy limitée	$\arg\min_{\alpha\in[0,\alpha_{\max}]} f(\alpha)$	
Règle de Curry	$\min\left\{\alpha>0;\dot{f}(\alpha)^{\mathrm{t}}\boldsymbol{d}_k=0\right\}$	
Règle d'Armijo	$f(\alpha_k)-f(0)\leqslant c_1\alpha_k\boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{d}_k$	$c_1\in]0,1[$
Règle d'Armijo non-monotone	$f(\alpha_k)-f_{\max}(0)\leqslant c_1\alpha_k\boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{d}_k$ avec $f_{\max}(0)=\max_{j\in[k-M,k]}F(\boldsymbol{x}_j)$	$c_1\in]0,1[$ $M>1$
Règle de Goldstein	$f(\alpha_k)-f(0)\leqslant c_1\alpha_k\boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{d}_k$ $f(\alpha_k)-f(0)\geqslant c_2\alpha_k\boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{d}_k$	$c_1\in]0,1[$ $c_2\in]c_1,1[$
Règle de Wolfe	$f(\alpha_k)-f(0)\leqslant c_1\alpha_k\boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{d}_k$ $\dot{f}(\alpha)^{\mathrm{t}}\boldsymbol{d}_k\geqslant c_3\boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{d}_k$	$c_1\in]0,1[$ $c_3\in]c_1,1[$
Règle de Wolfe stricte	$f(\alpha_k)-f(0)\leqslant c_1\alpha_k\boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{d}_k$ $\lvert\dot{f}(\alpha)^{\mathrm{t}}\boldsymbol{d}_k\rvert\geqslant c_3\lvert\boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{d}_k\rvert$	$c_1\in]0,1[$ $c_3\in]c_1,1[$
Condition de Zoutendijk	$f(\alpha_k)-f(0)\leqslant -c_4\lVert\boldsymbol{g}_k\rVert^2\cos^2\theta_k$ avec $\cos\theta_k=\dfrac{-\boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{d}_k}{\lVert\boldsymbol{g}_k\rVert\lVert\boldsymbol{d}_k\rVert}$	$c_4>0$

TABLE 2.3 – Règles de recherche de pas. Les trois premières sont associées à des stratégies de pas exactes alors que les quatre autres sont plutôt approchées et permettent de définir un intervalle de valeurs admissibles de α_k.

Pour la recherche d'un pas scalaire adéquat le respect de l'une des règles résumées dans le tableau (2.3) permet de garantitir la convergence. Les règles de pas de Cauchy et de Curry requièrent le calcul du minimiseur de la fonction $f(\cdot)$ et nécessitent l'utilisation d'une stratégie d'optimisation. Chacune des règles d'Armijo, de Wolfe et de Goldstein définit un ensemble de valeurs de pas admissibles. Les méthodes usuelles de recherche d'un tel pas se basent sur les stratégies de rebroussement, de dichotomie ou d'interpolation polynomiale quadratique ou cubique. Une revue très détaillée des méthodes de recherche de pas peut être trouvée dans les livres [Nocedal 99, chap. 2] et [Bertsekas 99], où il est également indiqué que la vérification de toutes les règles précédentes implique la satisfaction de la *condition de Zoutendijk* lorsque le critère est gradient Lipschitz et borné inférieurement. Cette condition, plutôt technique, est importante pour démontrer la convergence des algorithmes de descente itérative lorsqu'aucune des règles de pas n'est utilisée dans l'étape de calcul du pas.

Les différentes stratégies de recherche de pas sont souvent itératives : des valeurs de pas sont générées jusqu'à la satisfaction des conditions associées à la règle adoptée. Cependant, le nombre d'itérations nécessaires à la satisfaction de ce critère d'arrêt est inconnu et influe de façon significative sur le coût de calcul de l'algorithme, car chaque sous-itération de recherche de pas nécessite une évaluation du critère et parfois du gradient. Il existe aussi des techniques fondées sur des formules de pas analytiques [Sun 01] ou sur des récurrences simples [Labat 08] dont le nombre de sous-itérations est fixé à l'avance afin de contrôler ce coût de calcul.

Pour la recherche d'un pas multidimensionnel, le problème est plus complexe. En effet, à ma connaissance, il n'existe pas de schéma itératif permettant de déterminer un pas multi-dimensionnel respectant une règle de pas étendue au cas de descente dans un sous-espace de directions. Certes, on peut établir des liens avec les méthodes à base de *région de confiance* mais l'esprit de cette approche n'est pas tout à fait le même que celui dans la recherche d'un pas multi-dimensionnel.

Les solutions apportées aux problématiques précédentes consistent en le développement de stratégies de pas se basant sur les techniques d'optimisation par majoration-minimisation. Ces développements ont été initiés par les travaux de thèse de Marc ALLAIN [Allain 02] dans le cas des algorithmes de minimisation de critères semi-quadratiques, puis complétés successivement par les thèses de Christian LABAT [Labat 06] et Emilie CHOUZENOUX [Chouzenoux 10a].

2.2 Algorithmes de majoration-minimisation

Le schéma d'optimisation par majoration-minimisation (MM) consiste à remplacer le problème d'optimisation d'un critère, souvent difficile à minimiser, par une succession de simples minimisations de critères tangents-majorants du critère initial. Cette idée est apparue dans les travaux de [Ortega 70] et fût exploitée pour la formulation de l'algorithme EM (*Expectation-Maximization*) [Dempster 77] ou encore dans [Böhning 88] pour l'ajustement des méthodes de type quasi-Newton. Plus récemment, cette approche a servi pour le développement d'algorithmes d'optimisation de critères spécifiques en reconstruction tomographique [Lange 87, De Pierro 95, Fessler 98], en restauration d'images [Figueiredo 07] ou en factorisation de matrices non-négatives [Févotte 11]. Un tutoriel sur les méthodes MM peut être trouvé dans [Hunter 04] et une analyse de leurs propriétés théoriques est détaillée dans [Jacobson 07].

D'un point de vue méthodologique pour la résolution de problèmes de grandes tailles, les algorithmes MM offrent plusieurs perspectives intéressantes. On peut citer l'évitement d'inversion de matrices de grande taille ou mal conditionnées, l'assurance de la séparabilité des variables, la linéarisation du problème d'optimisation et la transformation de critères non-différentiables et/ou non-convexes en critères convexes et différentiables.

2.2.1 Optimisation par approximation tangente-majorante

Définition 1. *Une fonction $H(\boldsymbol{x}, \boldsymbol{y})$ est dite* approximation majorante *d'un critère $F(\boldsymbol{x})$ en $\boldsymbol{y}$ sur le domaine $\mathcal{D}_f$ si pour tout $\boldsymbol{x} \in \mathcal{D}_f$,*

$$\begin{cases} H(\boldsymbol{x}, \boldsymbol{y}) \geqslant F(\boldsymbol{x}) \\ H(\boldsymbol{y}, \boldsymbol{y}) = F(\boldsymbol{y}) \end{cases} \tag{2.5}$$

De plus, lorsque $H(\cdot, \boldsymbol{y})$ est différentiable sur $\mathcal{D}_f$, celle-ci devient tangente à $F(\cdot)$ en $\boldsymbol{y}$

$$\nabla F(\boldsymbol{y}) = \nabla_1 H(\boldsymbol{y}, \boldsymbol{y}). \tag{2.6}$$

où ∇_1 porte sur le premier argument de $H(\cdot,\cdot)$.

On se focalisera sur la majoration-minimisation quadratique [Böhning 88, Allain 06] qui se base sur des approximations tangentes majorantes quadratiques qui s'écrivent sous la forme

$$H(\boldsymbol{x},\boldsymbol{y}) = F(\boldsymbol{y}) + \nabla F(\boldsymbol{y})^{\mathrm{t}}(\boldsymbol{x}-\boldsymbol{y}) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{y})^{\mathrm{t}}\boldsymbol{B}(\boldsymbol{y})(\boldsymbol{x}-\boldsymbol{y}), \tag{2.7}$$

où $\boldsymbol{B}(\boldsymbol{y}) \in \mathbb{R}^{N\times N}$ est une matrice définie positive assurant que $H(\cdot,\cdot)$ vérifie les conditions de la définition 1. De façon générale, toute matrice $\boldsymbol{B}(\boldsymbol{y})$ telle que $\boldsymbol{B}(\boldsymbol{y}) - \nabla^2 F(\boldsymbol{y})$ soit définie positive entraîne le respect de la condition de majoration. Ces matrices sont construites en exploitant des propriétés sur la structure analytique du critère $F(\cdot)$.

2.2.1.1 Techniques de majoration

En plus des principales familles d'approches de majoration (inégalité de Jensen, définition de la convexité, courbure maximale, inégalité de Cauchy-Schwartz) présentées dans [Lange 00, Hunter 04], d'autres techniques de construction sont développées dans[De Pierro 95, Erdogan 99]. En restauration d'images par minimisation de critères de types moindres carrés pénalisés à l'aide de fonctions permettant de préserver les contours des images, (critères qualifiés de semi-quadratiques [Charbonnier 97, Idier 01]), des techniques analytiques permettant de construire la matrice hessienne de l'approximation majorante quadratique peuvent être trouvés dans[Chan 99, Allain 02].

2.2.1.2 Schéma de minimisation

La minimisation du critère $F(\cdot)$ par majoration-minimisation, illustré sur la figure (2.1), consiste à résoudre le problème initial par la règle de mise à jour MM

$$\boldsymbol{x}_{k+1} = \underset{\boldsymbol{x}\in\mathcal{D}_f}{\arg\min}\, H(\boldsymbol{x},\boldsymbol{x}_k). \tag{2.8}$$

Notons que cette stratégie MM entraîne une décroissance de la suite $\{F(\boldsymbol{x}_k)\}$.

$$F(\boldsymbol{x}_{k+1}) \underset{(1)}{\leqslant} H(\boldsymbol{x}_{k+1},\boldsymbol{x}_k) \underset{(2)}{\leqslant} H(\boldsymbol{x}_k,\boldsymbol{x}_k) \underset{(3)}{=} F(\boldsymbol{x}_k),$$

où les relations (1) à (3) proviennent des opérations de

(1) majoration,
(2) minimisation,
(3) tangence.

Dans le cas d'une approximation majorante quadratique, l'étape de minimisation admet une solution analytique et le schéma de minimisation itérative qui en résulte se base sur la récurrence

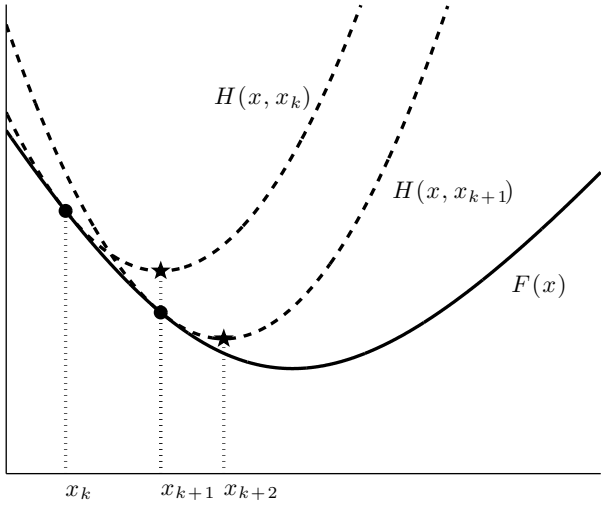


FIGURE 2.1 – Illustration du schéma de minimisation par majoration-minimisation. Dans cet exemple, la fonction $H(x, x_k)$ est une fonction quadratique de courbure fixe.

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \boldsymbol{B}(\boldsymbol{x}_k)^{-1}\boldsymbol{g}_k. \tag{2.9}$$

On peut assimiler cette récurrence à celle d'un algorithme de type quasi-Newton car la matrice $\boldsymbol{B}(\boldsymbol{x}_k)$ peut être vue comme une approximation du hessien [Böhning 88]. En outre, il été démontré dans [Chan 99, Allain 06] que les algorithmes semi-quadratiques [Geman 95, Geman 92] s'identifient à cette structure d'optimisation et ont proposé des formulations analytiques de $\boldsymbol{B}(\boldsymbol{x}_k)$.

2.3 Recherche de pas par majoration-minimisation quadratique

2.3.1 Recherche de pas scalaire

La recherche de pas consiste en la résolution approchée de

$$\begin{array}{c} \text{Trouver } \alpha_k \\ \text{tel que } \alpha_k = \underset{\alpha \geqslant 0}{\arg\min}\, f(\alpha), \end{array} \tag{2.10}$$

avec $f(\cdot)$, la restriction de $F(\cdot)$ à la droite $\{\boldsymbol{x}_k + \alpha \boldsymbol{d}_k : \alpha \in \mathbb{R}\}$, définie par $f(\alpha) = F(\boldsymbol{x}_k + \alpha\, \boldsymbol{d}_k)$.

La recherche de pas dans un cadre MM permet d'obtenir un pas α_k, défini par

$$\begin{cases} \alpha_k^0 = 0, \\ \alpha_k^{j+1} = \underset{\alpha}{\arg\min}\, h\big(\alpha, \alpha_k^j\big), \quad j = 0, \ldots, J-1, \\ \alpha_k = \alpha_k^J \end{cases} \tag{2.11}$$

où $h(\cdot, \alpha_k^j)$ est une approximation tangente majorante quadratique de $f(\cdot)$ en α_k^j,

$$h\big(\alpha, \alpha_k^j\big) = f\big(\alpha_k^j\big) + \big(\alpha - \alpha_k^j\big)\, \dot{f}\big(\alpha_k^j\big) + \frac{1}{2}\, m_k^j\, \big(\alpha - \alpha_k^j\big)^2 \tag{2.12}$$

avec

$$m_k^j = \boldsymbol{d}_k^{\mathrm{t}} \boldsymbol{B}_k^j \boldsymbol{d}_k \tag{2.13}$$

et $\boldsymbol{B}_k^j = \boldsymbol{B}(\boldsymbol{x}_k + \alpha_k^j\, \boldsymbol{d}_k)$ est la courbure de $H\big(\boldsymbol{x}, \boldsymbol{x}_k + \alpha_k^j \boldsymbol{d}_k\big)$. La récurrence MMQ 1D associée à la recherche de pas scalaire est donc définie par

$$\begin{cases} \alpha_k^0 = 0, \\ \alpha_k^{j+1} = \alpha_k^j - \theta \dot{f}(\alpha_k^j)/m_k^j \quad j = 0, \ldots, J-1, \\ \alpha_k = \alpha_k^J, \end{cases} \tag{2.14}$$

où $\theta \in]0, 2[$ est un facteur de relaxation. Cette approche proposée dans [Fessler 99, Rivera 03, Labat 08] a pour intérêt de fournir très simplement une séquence de pas qui converge vers le pas exact. Pour ce qui est de l'algorithme de descente, le résultat de convergence se résume dans le théorème suivant.

Théorème 1 ([Chouzenoux 12]). *Soit une fonction $F(\cdot)$ continûment differentiable, bornée inférieurement sur son domaine de définition $\mathcal{D}_f$ et admettant des approximations majorantes quadratiques. Si pour toute itération k, la direction de descente $\boldsymbol{d}_k$ est gradient-reliée et que le pas est calculé par la stratégie MMQ 1D avec $J > 0$ et $\theta \in]0, 2[$, alors l'algorithme de descente itérative converge au sens $\lim_{k\to\infty} \|\boldsymbol{g}_k\| = 0$.*

Discussion :

- Il faut signaler que les algorithmes classiques de descente itérative (Gradient, Newton, quasi-Newton, Newton tronqué et leurs versions préconditionnées) produisent des directions gradient-reliées, à l'exception du gradient conjugué non-linéaire (GCNL). En fait, l'interprétation d'une telle condition est d'assurer à chaque itération que le critère est susceptible de décroitre suffisamment en choisissant une direction qui ne tend pas à être orthogonale au gradient du critère.
- Un second point important qui découle de ce théorème est qu'une seule sous-itération de recherche de pas est suffisante pour garantir la convergence de l'algorithme, ce qui pourrait avoir un impact important sur la minimisation du coût de calcul par itération de descente.
- Enfin, signalons que la convergence de l'algorithme de descente par GCNL avec une recherche de pas MMQ 1D est établie dans [Labat 08] dans le cas d'un critère $F(\cdot)$ à gradient Lipschitz sur la ligne de niveau initial.

2.3.2 Recherche d'un pas multi-dimensionnel [Chouzenoux 11a]

Dans le cadre de la descente dans un sous-espace de directions, la recherche de pas correspond à la minimisation de $f(\boldsymbol{\alpha}) = F(\boldsymbol{x}_k + \boldsymbol{D}_k\boldsymbol{\alpha})$. Notons $h\big(\cdot, \boldsymbol{\alpha}_k^j\big)$ l'approximation tangente-majorante quadratique du critère $f(\cdot)$ en $\boldsymbol{\alpha}_k^j$

$$h\big(\boldsymbol{\alpha}, \boldsymbol{\alpha}_k^j\big) = f(\boldsymbol{\alpha}_k^j) + \dot{f}(\boldsymbol{\alpha}_k^j)^{\mathrm{t}}(\boldsymbol{\alpha} - \boldsymbol{\alpha}_k^j) + \frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}_k^j)^{\mathrm{t}}\boldsymbol{M}_k^j(\boldsymbol{\alpha} - \boldsymbol{\alpha}_k^j) \tag{2.15}$$

avec

$$\boldsymbol{M}_k^j = \boldsymbol{D}_k^{\mathrm{t}}\boldsymbol{B}_k^j\boldsymbol{D}_k, \tag{2.16}$$

où $\boldsymbol{B}_k^j = \boldsymbol{B}\big(\boldsymbol{x}_k + \boldsymbol{D}_k\boldsymbol{\alpha}_k^j\big)$ est la courbure de $H\big(\boldsymbol{x}, \boldsymbol{x}_k + \boldsymbol{D}_k\boldsymbol{\alpha}_k^j\big)$. Le recherche d'un pas multi-dimensionnel par approche MMQ MD est définie donc par la récurrence

$$\begin{cases} \boldsymbol{\alpha}_k^0 = 0, \\ \boldsymbol{\alpha}_k^{j+1} = \boldsymbol{\alpha}_k^j - \theta\big[\boldsymbol{M}_k^j\big]^{-1}\nabla f\big(\boldsymbol{\alpha}_k^j\big) \quad j = 0, \ldots, J-1, \\ \boldsymbol{\alpha}_k = \boldsymbol{\alpha}_k^J. \end{cases} \tag{2.17}$$

Théorème 2 ([Chouzenoux 11a]). *Soit une fonction $F(\cdot)$ continûment différentiable, bornée inférieurement sur son domaine de définition $\mathcal{D}_f$ et admettant des approximations majorantes quadratiques. Si pour toute itération k, une composante du sous-espace de directions est gradient-reliée, et que le pas est calculé par la stratégie MMQ MD avec $J > 0$ et $\theta \in]0, 2[$, alors l'algorithme de descente itérative converge au sens $\lim_{k\to\infty} \|\boldsymbol{g}_k\| = 0$.*

De même que le dans le cas scalaire, le champs d'application de ce théorème est assez large puisqu'il suffit de choisir un sous-espace de directions dont au moins une composante est gradient-reliée.

2.3.3 Recherche de pas pour un critère barrière

On s'intéresse à la minimisation d'un critère contenant un terme issu d'une fonction *barrière*, c'est-à-dire une fonction strictement convexe dont le gradient est de norme infinie aux bords de son domaine de définition $\mathcal{C}$. On peut citer comme exemples la barrière logarithmique $\psi(u) = -\log u$), la barrière entropique ($\psi(u) = u\log u$) et la barrière hyperbolique ($\psi(u) = -\sqrt{u}$).

La présence d'une telle fonction dans un critère permet d'assurer l'appartenance du minimiseur du critère à ce domaine des solutions admissibles $\mathcal{C}$. Dans le cadre de la résolution de problèmes inverses par minimisation d'un critère composite, la fonction barrière peut apparaître dans le terme d'attache aux données : par exemple, le modèle de bruit de poisson conduit à un critère d'attache au données fondé sur la divergence de Kullback-Leibler, qui correspond à une fonction barrière permettant de satisfaire la contrainte $[\boldsymbol{K}\boldsymbol{x}]_m > 0$. Dans d'autres cas, la fonction barrière apparaît

dans le critère de régularisation. Ainsi, les entropies de Shannon et Burg utilisées dans le cadre de la reconstruction d'images par maximum d'entropie [Dusaussoy 95, Johnson 03], jouent le rôle de fonctions barrières pour assurer la contrainte de positivité. Le tableau 1 de [Chouzenoux 12] (annexe A.1) donne une liste de fonctions barrières rencontrées dans le cadre de la reconstruction de signaux ou d'images.

En ce qui concerne l'optimisation itérative d'un critère contenant une fonction barrière, la singularité due à la présence de la fonction barrière implique que la restriction du critère à une droite peut présenter une ou plusieurs asymptotes verticales, ce qui rend impossible la majoration du critère (ou de sa restriction) par une fonction quadratique. De plus, les méthodes usuelles de recherche de pas peuvent être inefficaces dans ce contexte [Murray 94].

Pour palier ces limitations, il serait judicieux d'adopter le schéma de recherche de pas par majoration-minimisation en s'appuyant sur une fonction majorante log-quadratique.

2.4 Recherche de pas par majoration-minimisation log-quadratique

Ces développements sont décrits de façon plus détaillée dans [Chouzenoux 09a, Chouzenoux 12]. Tout d'abord, le critère $F(\cdot)$ est scindé en deux termes et est réécrit sous la forme

$$F(\boldsymbol{x}) = P(\boldsymbol{x}) + \mu B(\boldsymbol{x}), \quad \mu > 0, \tag{2.18}$$

où $B(\cdot)$ est une fonction barrière associée à un domaine $\mathcal{C}$ défini par M contraintes linéaires d'inégalités $C_i(\boldsymbol{x}) = \boldsymbol{a}_i^{\mathrm{t}}\boldsymbol{x} + \rho_i > \mathbf{0}$,

$$B(\boldsymbol{x}) = \sum_{i=1}^{M} \psi_i(C_i(\boldsymbol{x}_k)) \tag{2.19}$$

où $\psi_i(\cdot)$ sont des fonctions barrières scalaires et $P(\cdot)$ est un critère différentiable dont le gradient est borné sur $\mathcal{C}$.

Soit $\boldsymbol{x}_k \in \mathcal{C}$ et $\boldsymbol{d}_k$ une direction de descente de $F(\cdot)$ en $\boldsymbol{x}_k$. Il s'en suit que la fonction

$$f(\alpha) = P(\boldsymbol{x}_k + \alpha\boldsymbol{d}_k) + \mu \sum_{i=1}^{M} \psi_i(C_i(\boldsymbol{x}_k + \alpha\boldsymbol{d}_k)) \tag{2.20}$$

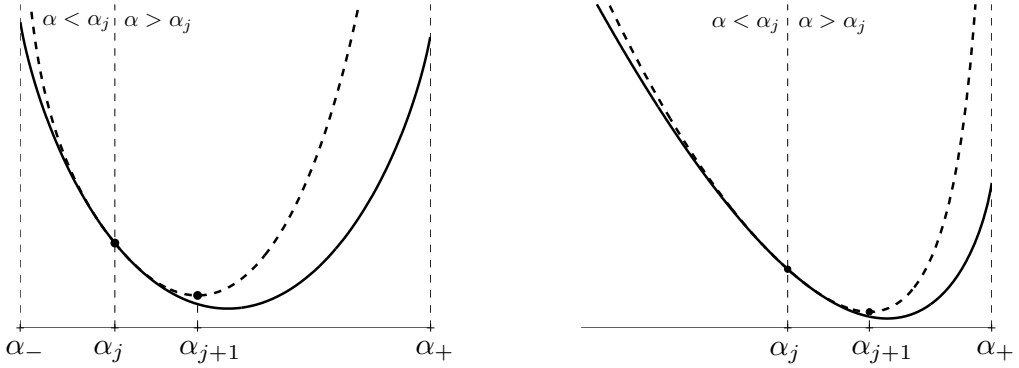
deviendrait non bornée dès lors que la valeur du pas α tend à annuler une des contraintes $\{C_i(\boldsymbol{x}_k + \alpha\boldsymbol{d}_k)\}$. Cependant, comme les contraintes sont linéaires, on peut expliciter l'intervalle $]\alpha_k^-, \alpha_k^+[$ des valeurs admissibles du pas

$$\alpha_k^- = \max_{i|\boldsymbol{a}_i^{\mathrm{t}}\boldsymbol{d}_k>0} -\frac{\boldsymbol{a}_i^{\mathrm{t}}\boldsymbol{x}_k}{\boldsymbol{a}_i^{\mathrm{t}}\boldsymbol{d}_k} \quad \text{et} \quad \alpha_k^+ = \min_{i|\boldsymbol{a}_i^{\mathrm{t}}\boldsymbol{d}_k<0} -\frac{\boldsymbol{a}_i^{\mathrm{t}}\boldsymbol{x}_k}{\boldsymbol{a}_i^{\mathrm{t}}\boldsymbol{d}_k}$$

Comme la présence de ces asymptotes verticales rend impossible la majoration de cette fonction par une parabole, un second terme composé d'une fonction logarithmique sera rajouté pour rendre possible la majoration de la fonction barrière,

$$h(\alpha,\alpha_k^j) = f(\alpha_k^j) + \dot{f}(\alpha_k^j)(\alpha-\alpha_k^j) + \frac{1}{2}m_k^j(\alpha-\alpha_k^j) + \gamma_k^j\left[(\bar{\alpha}_k^j-\alpha_k^j)\log\frac{\bar{\alpha}_k^j-\alpha_k^j}{\bar{\alpha}_k^j-\alpha} - \alpha + \alpha_k^j\right] \quad (2.21)$$

où $\alpha_k^j \in]\alpha_k^-, \alpha_k^+[$ est la valeur courante du pas et $\{m_k^j, \gamma_k^j, \bar{\alpha}_k^j\}$ sont les paramètres de construction de la fonction majorante. Cette formulation est inspirée de l'expression de la barrière adaptative décrite dans [Lange 94, Hunter 04]. Le second avantage de cette fonction est d'admettre un minimiseur exprimable analytiquement. La procédure de construction de la majorante n'est pas présentée dans ce manuscrit mais celle-ci est disponible dans [Chouzenoux 12] et d'une façon plus détaillée dans [Chouzenoux 10a]. Une illustration est fournie dans la figure (2.2).



(a) Deux barrières (b) Barrière à droite

FIGURE 2.2 – Illustration de la majoration log-quadratique

La récurrence MMLQ 1D pour la recherche de pas dans le cas de fonctions barrière se présente donc sous la forme

$$\begin{cases} \alpha_k^0 = 0, \\ \alpha_k^{j+1} = \begin{cases} \alpha_k^j - \dfrac{2q_3}{q_2+\sqrt{q_2^2-4q_1q_3}} & \text{si} \quad \dot{f}(\alpha_k^j) \leqslant 0 \\ \alpha_k^j - \dfrac{2q_3}{q_2-\sqrt{q_2^2-4q_1q_3}} & \text{si} \quad \dot{f}(\alpha_k^j) > 0 \end{cases} \quad \text{pour} \quad j = 0,\ldots,J-1, \\ \alpha_k = \alpha_k^J, \end{cases} \quad (2.22)$$

avec $q_1 = -m_k^j$, $q_2 = \gamma_k^j - \dot{f}(\alpha_k^j) + m_k^j(\bar{\alpha}_k^j - \alpha_k^j)$ et $q_3 = (\bar{\alpha}_k^j - \alpha_k^j)\dot{f}(\alpha_k^j)$

L'analyse de convergence a été réalisée dans le cas des algorithmes fondés sur des directions de descente gradient-reliées (Algorithmes de Gradient, Newton, quasi-Newton et Newton tronqué et leurs versions préconditionnées) et le résultat se résume dans le théorème suivant.

Théorème 3 ([Chouzenoux 12]). *Soit une fonction $F(\cdot)$ continûment differentiable, bornée inférieurement*

sur son domaine de définition $\mathcal{D}_f$ pouvant se réécrire comme la somme d'une fonction $P(\cdot)$ admettant une approximation majorante quadratique et une fonction barrière $B(\cdot)$. Si la fonction $F(\cdot)$ est gradient Lipschitz sur un voisinage ouvert et borné de la ligne de niveau initial et si pour toute itération k, la direction de descente $\boldsymbol{d}_k$ est gradient-reliée et que le pas est calculé par la stratégie MMLQ 1D avec $J > 0$ alors l'algorithme de descente itérative converge au sens $\lim_{k\to\infty} \|\boldsymbol{g}_k\| = 0$.

Ce résultat montre que cette stratégie de recherche de pas peut être employée au sein de plusieurs techniques usuelles de descente itératives. De plus, la convergence est également établie dans le cas de l'algorithme de gradient conjugué non-linéaire, moyennant quelques hypothèses techniques supplémentaires [Chouzenoux 10a]. Des résultats de tests numériques qui illustrant l'efficacité de cette recherche de pas MM peuvent être consultés dans [Chouzenoux 11a] pour la descente dans un sous-espace de directions, et dans le papier [Chouzenoux 12] pour la minimisation d'un critère barrière. D'un point de vue appliqué, les résultats de ce travail ont permis de proposer une méthode d'inversion numérique d'une transformée de Laplace pour la reconstruction des distributions des temps de relaxation en spectroscopie de résonance magnétique nucléaire par minimisation d'un critère régularisé par maximum d'entropie [Chouzenoux 09b, Chouzenoux 10b]. Le papier [Chouzenoux 10b] est annexé à ce manuscrit pour plus de détails. Cette méthode a fait l'objet du développement d'une application [1] qui est actuellement utilisée en routine à l'IRSTEA (Rennes).

2.5 Conclusions

La recherche de pas dans un algorithme de descente itérative est une étape qui peut affecter non seulement la convergence de l'algorithme mais aussi son coût de calcul. En effet, s'appuyer sur une règle de pas trop restrictive va induire plusieurs évaluations du critère et de son gradient, ce qui n'est pas souhaitable pour un problème de grande taille. Or, le plus important est d'assurer la convergence de l'algorithme global. L'emploi d'une stratégie de pas fondée sur une approche de majoration-minimisation est une bonne alternative d'autant plus qu'une seule sous-itération de la récurrence MM est suffisante pour garantir la convergence.

Les contributions méthodologiques fortes de mon travail sur cette question, via l'encadrement de la thèse de Emilie CHOUZENOUX, est de proposer des stratégies de calcul de pas de descente fondées sur des techniques de majoration-minimisation adaptées aux cas de descente dans un sous-espace de directions ainsi que pour la minimisation de critères barrières qui sont très rencontrés en restauration de signaux et d'images. Les perspectives à court terme de mon travail sur la recherche de pas pour l'optimisation itérative concernent essentiellement l'étude du lien entre les méthodes de descente par recherche de pas MM et l'approche fondée sur les régions de confiance.

1. Une procédure de dépôt de protection logicielle de cette application est envisagée.

Chapitre 3

Accélération algorithmique et matérielle de l'optimisation sous contraintes

Ce chapitre est dédié à la résolution d'un problème inverse par minimisation d'un critère composite, convexe et différentiable, sous des contraintes linéaires d'égalité et/ou d'inégalités

$$\begin{aligned} &\text{Trouver } \hat{\boldsymbol{x}}, \\ &\text{tel que } \hat{x} = \underset{\boldsymbol{x} \in \mathcal{D}_f}{\arg\min}\, F(\boldsymbol{x}), \\ &\text{sous contraintes } \begin{cases} \boldsymbol{C}_0\, \boldsymbol{x} - \boldsymbol{c}_0 = \boldsymbol{0}, \\ \boldsymbol{C}_1\, \boldsymbol{x} + \boldsymbol{c}_1 \geqslant \boldsymbol{0}, \end{cases} \end{aligned} \tag{3.1}$$

où $F(\cdot)$ est un critère convexe différentiable et borné inférieurement sur son domaine de définition $\mathcal{D}_f = \mathbb{R}^N$. Les matrices $\boldsymbol{C}_0 \in \mathbb{R}^{N_0 \times N}$, $\boldsymbol{C}_1 \in \mathbb{R}^{N_1 \times N}$ et les vecteurs $\boldsymbol{c}_0 \in \mathbb{R}^{N_0}$ et $\boldsymbol{c}_1 \in \mathbb{R}^{N_1}$ sont associés aux contraintes d'égalités et inégalités.

La motivation de ce travail provient de la résolution du problème d'estimation des cartes d'abondances en imagerie hyperspectrale [Chang 07]. Sur le plan méthodologique, ce problème inverse est mal-posé et est de grande taille. En effet, les données correspondent à des centaines d'images qui peuvent être de grande taille et sont parfois issues de l'observation de surfaces qui peuvent contenir plusieurs constituants. De plus, les inconnues du problème correspondent à des proportions de constituants qui doivent donc satisfaire des contraintes d'inégalité (non-négativité) et d'égalité (somme-à-un). Les questions liées à la formulation du critère de régularisation, au choix la méthode d'optimisation et son adaptation pour obtenir une structure algorithmique se prêtant à une implémentation efficace sont à l'origine des développements réalisés dans la thèse de Maxime LEGENDRE [1].

L'analyse est focalisée particulièrement sur une méthode d'optimisation sous contraintes de

1. Doctorant ECN, 2012-2015.

type points-intérieurs avec une approche primale-duale. Ces méthodes sont efficaces pour des problèmes de petite taille mais leur complexité algorithmique les rend inutilisables en grandes dimensions. Le but est de discuter l'approche qui consiste à modifier cette méthode de sorte à réduire sa complexité arithmétique et adapter sa structure algorithmique pour pouvoir l'implémenter efficacement sur des processeurs de cartes graphiques (GPUs), sans altérer les propriétés de convergence.

3.1 Méthode primale-duale des points-intérieurs

Il existe principalement trois grandes familles de méthodes d'optimisation sous contraintes [Nocedal 99] : les méthodes de *pénalité extérieure*, les méthodes de *contraintes actives* et les méthodes de *points-intérieurs*. On s'intéressera particulièrement à cette dernière famille de méthodes. Les méthodes de *points-intérieurs* (qualifiés aussi de méthodes de *pénalité intérieure*) ont la spécificité de garantir le respect des contraintes d'inégalité *strictement* (d'où le qualificatif *intérieurs*).

Le principe d'optimisation par points intérieurs est apparu dans les années cinquante, grâce à la définition de la fonction barrière logarithmique, en 1955, par Frisch [Frisch 55]. C'est dans le livre de Fiacco et McCormick [Fiacco 68] que le terme de points intérieurs a été introduit. Les travaux de Karmakar [Karmarkar 84] en 1984 fûrent à l'origine de la proposition d'un algorithme à convergence polynomiale, ce qui a ouvert la voie au développement de plusieurs techniques telles que *le suivi de chemin central*, la *barrière logarithmique* et la méthode *primale-duale* [Nocedal 99].

Dans ce qui suit, la résolution du problème d'optimisation sous contraintes se fera par un algorithme itératif de type points-intérieurs avec une approche primale-duale [Mehrotra 92, Armand 00]. D'une façon générale, pour la construction des algorithmes de points-intérieurs, il y a deux points de vue complémentaires qui conduisent au même résultat : *la pénalisation logarithmique* et *la perturbation* des conditions d'optimalité. L'approche primale-duale consiste à estimer conjointement les variables primales (variables d'intérêt) et duales (multiplicateurs de Lagrange) par la résolution d'une séquence de problèmes correspondants à des versions perturbées des conditions d'optimalité, dites de Karush-Kuhn-Tucker (KKT), pondérées par une suite de paramètres positifs $\{\mu_k\}$ convergeant vers 0. De plus, à chaque itération, la satisfaction stricte des contraintes est assurée par la minimisation d'une *fonction de mérite* présentant une barrière logarithmique à la frontière du domaine admissible des solutions [Wright 91].

3.1.1 Prise en compte de la contrainte d'égalité

Tout d'abord, à l'aide d'un changement de variable, le problème (3.1) peut être transformé en un nouveau problème faisant apparaître des contraintes d'inégalité uniquement. Comme souligné dans [Armand 00], pour tout vecteur initial $\boldsymbol{x}^{(1)}$ tel que $\boldsymbol{C}_0\boldsymbol{x}^{(1)} = \boldsymbol{c}_0$, le vecteur défini par

$\boldsymbol{a} = \boldsymbol{x}^{(1)} + \boldsymbol{Z}\boldsymbol{a}$, avec $\boldsymbol{a} \in \mathbb{R}^{N-1}$, satisfait également cette contrainte égalité si $\boldsymbol{Z} \in \mathbb{R}^{N\times(N-1)}$ est une matrice dont les colonnes forment l'espace nul de $\boldsymbol{C}_0$. La possibilité de calcul de l'espace nul de la matrice des contraintes d'égalité est une condition nécessaire pour l'emploi d'une telle approche.

Par conséquent, le problème (3.1) est réécrit sous la forme d'un problème d'optimisation sous des contraintes d'inégalités uniquement,

$$\min_{\boldsymbol{a}\in\mathbb{R}^{(N-1)}} \Phi(\boldsymbol{a}) \quad \text{s. c.} \quad \boldsymbol{T}\boldsymbol{a} + \boldsymbol{t} \geqslant 0. \tag{3.2}$$

où le critère $\Phi(\cdot)$ se déduit de $F(\cdot)$ par $\Phi(\boldsymbol{a}) = F(\boldsymbol{x}^{(1)} + \boldsymbol{Z}\boldsymbol{a})$, $\boldsymbol{T} = \boldsymbol{C}_1\boldsymbol{Z}$ et $\boldsymbol{t} = \boldsymbol{C}_1\boldsymbol{x}^{(1)} + \boldsymbol{c}_1$.

Les conditions de KKT permettant de caractériser l'optimalité de la solution $\boldsymbol{a}^*$ de (3.2) et les multiplicateurs de Lagrange associés $\boldsymbol{\lambda}^*$ sont : (1) $\nabla\Phi(\boldsymbol{a}^*) - \boldsymbol{T}^{\mathrm{t}}\boldsymbol{\lambda}^* = \boldsymbol{0}$, (2) $\mathrm{Diag}(\boldsymbol{\lambda})(\boldsymbol{T}\boldsymbol{a}^* + \boldsymbol{t}) = \boldsymbol{0}$, (3) $\boldsymbol{T}\boldsymbol{a}^* + \boldsymbol{t} \geqslant \boldsymbol{0}$ et (4) $\boldsymbol{\lambda}^* \geqslant \boldsymbol{0}$.

La perturbation de ces conditions permet de caractériser une solution intermédiaire $(\boldsymbol{a}_k, \boldsymbol{\lambda}_k)$, solution du système d'équations

$$\begin{cases} \nabla\Phi(\boldsymbol{a}) - \boldsymbol{T}^{\mathrm{t}}\boldsymbol{\lambda} = \boldsymbol{0}, \\ \boldsymbol{\Lambda}\,(\boldsymbol{T}\boldsymbol{a} + \boldsymbol{t}) = \boldsymbol{\mu}_k, \\ \boldsymbol{T}\boldsymbol{a} + \boldsymbol{t} \geqslant \boldsymbol{0}, \\ \boldsymbol{\lambda} \geqslant \boldsymbol{0}. \end{cases} \tag{3.3}$$

où $\boldsymbol{\Lambda} = \mathrm{Diag}(\boldsymbol{\lambda})$ et $\boldsymbol{\mu}_k = \mu_k \boldsymbol{1}_{(N-1)\times 1}$. Ainsi, chaque itération k de l'*algorithme primal-dual* (PDIP) pour la résolution du problème (3.2) se décompose en deux étapes. Tout d'abord, un couple $(\boldsymbol{a}_{k+1}, \boldsymbol{\lambda}_{k+1})$ est calculé en fonction de $(\boldsymbol{a}_k, \boldsymbol{\lambda}_k)$ en résolvant (3.3). Ensuite, le paramètre de perturbation μ_{k+1} est réduit selon une règle de mise à jour permettant de garantir la convergence de l'algorithme.

3.1.2 Résolution du problème perturbé

Dans le cadre des problèmes de grande taille, il n'est pas possible de résoudre (3.3) de façon exacte. En pratique, une solution approchée de (3.3) est obtenue par quelques itérations de Newton couplées avec une recherche de pas [Boyd 04, Chap.11], selon le schéma général

$$(\boldsymbol{a}_{k+1}, \boldsymbol{\lambda}_{k+1}) = (\boldsymbol{a}_k + \alpha_k \boldsymbol{d}_k^a, \boldsymbol{\lambda}_k + \alpha_k \boldsymbol{d}_k^\lambda). \tag{3.4}$$

avec des directions primale $\boldsymbol{d}_k^a$ et duale $\boldsymbol{d}_k^\lambda$ qui correspondent à un pas de Newton appliqué aux deux premières conditions du système KKT perturbé. Ces directions sont calculées en résolvant le

système linéaire

$$\begin{pmatrix} \nabla^2\Phi(\boldsymbol{a}_k) & -\boldsymbol{T}^{\mathrm{t}} \\ \boldsymbol{\Lambda}_k\boldsymbol{T} & \mathrm{Diag}(\boldsymbol{T}\boldsymbol{a}_k+\boldsymbol{t}) \end{pmatrix} \begin{pmatrix} \boldsymbol{d}_k^a \\ \boldsymbol{d}_k^\lambda \end{pmatrix} = -\boldsymbol{r}_{\mu_k}(\boldsymbol{a}_k,\boldsymbol{\lambda}_k), \tag{3.5}$$

où $\boldsymbol{r}_\mu(\boldsymbol{a},\boldsymbol{\lambda})$ est le résidu primal-dual,

$$\boldsymbol{r}_\mu(\boldsymbol{a},\boldsymbol{\lambda}) = \begin{pmatrix} \nabla\Phi(\boldsymbol{a}) - \boldsymbol{T}^{\mathrm{t}}\boldsymbol{\lambda} \\ \boldsymbol{\Lambda}(\boldsymbol{T}\boldsymbol{a}+\boldsymbol{t}) - \boldsymbol{\mu} \end{pmatrix} = \begin{pmatrix} \boldsymbol{r}_\mu^{\mathrm{prim}}(\boldsymbol{a},\boldsymbol{\lambda}) \\ \boldsymbol{r}_\mu^{\mathrm{dual}}(\boldsymbol{a},\boldsymbol{\lambda}) \end{pmatrix}. \tag{3.6}$$

3.1.2.1 Calcul des directions primale et duale

Le système (3.5) n'est pas inversé de façon directe. En effet, il est souligné dans [Wright 94, Wright 98] que ce système devient très mal conditionné, notamment à l'approche de la convergence de l'algorithme, dès lors qu'une des contraintes est active. De plus, celui-ci ne vérifie pas les propriétés de symétrie et de définie positivité, souhaitables dès lors que l'on applique une stratégie d'inversion itérative. Plusieurs stratégies de résolution de (3.5), présentées dans [Forsgren 02, Sec.5.1], permettent de pallier ces difficultés. Nous utilisons la technique de [Conn 96, Armand 00, Segalat 02], consistant à effectuer le calcul des directions en deux étapes : la direction primale $\boldsymbol{d}_k^a$ est d'abord obtenue par inversion du système réduit

$$\boldsymbol{H}_k\boldsymbol{d}_k^a = -\boldsymbol{g}_k \tag{3.7}$$

avec

$$\begin{cases} \boldsymbol{g}_k &= \nabla\Phi(\boldsymbol{a}_k) + \boldsymbol{T}^{\mathrm{t}}\mathrm{Diag}(\boldsymbol{T}\boldsymbol{a}_k+\boldsymbol{t})^{-1}\boldsymbol{\mu}_k, \\ \boldsymbol{H}_k &= \nabla^2\Phi(\boldsymbol{a}_k) + \boldsymbol{T}^{\mathrm{t}}\mathrm{Diag}(\boldsymbol{T}\boldsymbol{a}_k+\boldsymbol{t})^{-1}\boldsymbol{\Lambda}_k\boldsymbol{T}. \end{cases} \tag{3.8}$$

Rappelons que ce système réduit s'obtient par substitution de $\boldsymbol{d}_k^\lambda$ dans la première équation de (3.5) par son expression déduite de la seconde partie de ce système,

$$\boldsymbol{d}_k^\lambda = \mathrm{Diag}(\boldsymbol{T}\boldsymbol{a}_k+\boldsymbol{t})^{-1}\left[\boldsymbol{\mu}_k - \boldsymbol{\Lambda}_k\boldsymbol{T}(\boldsymbol{a}_k+\boldsymbol{d}_k^a) - \boldsymbol{\Lambda}_k\,\boldsymbol{t}\right]. \tag{3.9}$$

Finalement, après obtention de la direction primale, l'expression (3.9) est utilisée pour déterminer la direction duale $\boldsymbol{d}_k^\lambda$.

3.1.2.2 Recherche de pas

Le pas α_k est déterminé de façon à garantir la convergence de l'algorithme et à vérifier les deux contraintes d'inégalité de (3.3). La convergence de l'algorithme est garantie sous réserve que le pas entraîne une décroissance suffisante d'une fonction de mérite primale-duale $\Psi_\mu(\boldsymbol{a},\boldsymbol{\lambda})$ liée aux conditions d'optimalité du problème [Forsgren 02, Sec.5.2]. Nous employons la fonction de mérite

primale-duale [Anstreicher 94, Forsgren 98, Armand 00] définir par

$$\Psi_\mu(\boldsymbol{a},\boldsymbol{\lambda}) = \Phi(\boldsymbol{a}) - \mu\sum_{i=1}^{N}\ln([\boldsymbol{T}\boldsymbol{a}+\boldsymbol{t}]_i) + \boldsymbol{\lambda}^{\mathrm{t}}(\boldsymbol{T}\boldsymbol{a}+\boldsymbol{t}) - \mu\sum_{i=1}^{N}\ln(\lambda_i[\boldsymbol{T}\boldsymbol{a}+\boldsymbol{t}]_i). \tag{3.10}$$

On peut constater la présence des deux fonctions barrières logarithmiques pour satisfaire strictement les contraintes d'inégalités de (3.3). Une technique de rebroussement associée à la règle d'Armijo est utilisée pour la recherche de pas. Ainsi, une décroissance suffisante de la fonction de mérite se traduit par exemple par la vérification de la condition d'Armijo

$$\psi_{\mu_k}(\alpha_k) - \psi_{\mu_k}(0) \leqslant c\,\alpha_k\nabla\psi_{\mu_k}(0), \qquad c\in(0,1), \tag{3.11}$$

où $\psi_{\mu_k}(\alpha) = \Psi_{\mu_k}(\boldsymbol{a}_k+\alpha\boldsymbol{d}_k^a, \boldsymbol{\lambda}_k+\alpha\boldsymbol{d}_k^\lambda)$. Nous avons montré dans [Chouzenoux 11b] qu'une stratégie de recherche de pas plus sophistiquée, telle que par exemple l'approche MMLQ 1D, ne semble pas nécessaire dans le cadre des méthodes primales-duales des points-intérieurs.

3.1.3 Contrôle de convergence de l'algorithme primal-dual

L'arrêt de la boucle interne, liée au calcul des directions primale et duale, est régi par deux conditions [Conn 96, Johnson 00]

$$\|r_{\mu_k}^{\text{prim}}(\boldsymbol{a}_k,\boldsymbol{\lambda}_k)\|_\infty \leqslant \epsilon_k^{\text{prim}} \quad \text{et} \quad \|r_{\mu_k}^{\text{dual}}(\boldsymbol{a}_k,\boldsymbol{\lambda}_k)\|_1/N \leqslant \epsilon_k^{\text{dual}}, \tag{3.12}$$

avec $\epsilon_k^{\text{prim}} = \eta^{\text{prim}}\mu_k$, $\epsilon_k^{\text{dual}} = \eta^{\text{dual}}\mu_k$ où η^{prim} et η^{dual} sont deux paramètres positifs.

Le paramètre de perturbation μ_k est mis à jour selon la règle de μ-criticité définie dans [El-Bakry 96]

$$\mu_k = \frac{\theta}{N}(\boldsymbol{T}\boldsymbol{a}_k+\boldsymbol{t})^{\mathrm{t}}\boldsymbol{\lambda}_k, \tag{3.13}$$

où $\theta\in(0,1)$.

Enfin, les itérations de l'algorithme PDIP sont contrôlées par un test d'arrêt global [Boyd 04, Chap.11] portant sur la valeur minimale de la perturbation ou sur la norme du résidu primal-dual

$$\mu_k \leqslant \mu_{\min}, \qquad \text{et} \qquad \|\boldsymbol{r}_0(\boldsymbol{a}_k,\lambda_k)\| \leqslant \epsilon_0. \tag{3.14}$$

Les propriétés de convergence de cette méthode primale-duale des points intérieurs dans le cas de critères fortement convexes sont données dans le théorème (4).

Théorème 4 ([Armand 00]). *Supposons que la fonction $\Phi(\boldsymbol{a})$ soit fortement convexe et différentiable sur $\mathbb{R}^{N-1}$. Si les séquences $\{\mu_k\}$, $\left\{\epsilon_k^{prim}\right\}$ et $\left\{\epsilon_k^{dual}\right\}$ tendent vers 0 lorsque k tend vers l'infini, alors la suite $\{(\boldsymbol{a}_k,\boldsymbol{\lambda}_k)\}$ générée par l'algorithme PDIP est bornée et chacun de ses points d'adhérence est une solution primale-duale du problème (3.2).*

3.2 Accélération algorithmique pour des problèmes de grande taille

On considère, à présent, le cas de l'application de cet algorithme à la résolution conjointe de plusieurs problèmes donnés sous la forme

$$\boldsymbol{y}_p = \boldsymbol{K}\boldsymbol{x}_p + \boldsymbol{e}_p, \quad (\forall\, p = 1, \ldots, P.) \tag{3.15}$$

Ce modèle linéaire apparaît en imagerie hyperspectrale où $\boldsymbol{y}_p \in \mathbb{R}^M$ correspond au spectre associé au p-ème pixel de l'image dans M bandes spectrales, $\boldsymbol{K} \in \mathbb{R}^{M\times N}$ une matrice dont les colonnes contiennent des spectres caractéristiques des constituants de la zone observée et $\boldsymbol{x}_p$ le vecteur des inconnues qui sont les abondances des constituants dans la surface associée au pixel d'indice p.

Sous l'hypothèse d'un bruit de mesure additif gaussien blanc et de moyenne nulle, le critère $F(\cdot)$ s'exprime par

$$F(\boldsymbol{X}) = \frac{1}{2}\sum_{p=1}^{P} \|\boldsymbol{y}_p - \boldsymbol{K}\boldsymbol{x}_p\|_2^2 = \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{K}\boldsymbol{X}\|_F^2, \tag{3.16}$$

où $\|\cdot\|_F$ représente la norme de Frobenius et les matrices $\boldsymbol{Y} \in \mathbb{R}^{M\times P}$, $\boldsymbol{X} \in \mathbb{R}^{N\times P}$ résultent de la concaténation de tous les vecteurs $\{\boldsymbol{y}_p\}$ et $\{\boldsymbol{x}_p\}$. Il est également possible d'opter pour une résolution régularisée en ajoutant un critère de pénalisation $R(\boldsymbol{X})$, pondéré par un paramètre de régularisation β,

$$F(\boldsymbol{X}) = \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{K}\boldsymbol{X}\|_F^2 + \beta R(\boldsymbol{X}). \tag{3.17}$$

Soit l'opérateur $\boldsymbol{m} = \mathrm{vect}\big(\boldsymbol{M}\big)$ qui correspond à la transformation de la matrice $\boldsymbol{M}$ en un vecteur $\boldsymbol{m}$ dans l'ordre lexicographique ainsi qu'une solution initiale $\boldsymbol{X}^{(1)}$ satisfaisant les contraintes d'égalités. Le problème d'optimisation sous contraintes (3.2) associé au modèle (3.15) s'exprime de façon équivalente sous la forme :

$$\min_{\boldsymbol{a}\in\mathbb{R}^{(N-1)P}} \Phi(\boldsymbol{a}), \tag{3.18}$$

où le critère $\Phi(\cdot)$ se déduit de $F(\cdot)$ par $\Phi(\boldsymbol{a}) = F(\boldsymbol{X}^{(1)} + \boldsymbol{Z}\boldsymbol{A})$, $\boldsymbol{a} = \mathrm{vect}(\boldsymbol{A})$ et $\boldsymbol{t} = \mathrm{vect}(\boldsymbol{C}^{(1)})$. La matrice $\boldsymbol{T}$ est égale à $\boldsymbol{I}_N \otimes \boldsymbol{Z}$ où $\otimes$ est le produit de Kronecker et $\boldsymbol{I}_P$ est la matrice identité de taille $P \times P$.

3.2.1 Structure du système primal-dual

Le coût de calcul de l'algorithme PDIP est fortement dépendant du coût de calcul de la direction primale, qui fait appel à la résolution du système d'équations (3.7). Afin de réduire le temps de calcul dans le cas de problèmes de grande taille, il est judicieux d'alléger la complexité de cette étape en tirant profit de la structure de la matrice $\boldsymbol{H}_k$. Ainsi, des versions accélérées de l'algorithme PDIP ont été proposées en résolvant le système (3.7) de façon approchée. Un telle stratégie

d'accélération algorithmique rentre dans le cadre des méthodes de point intérieurs dites *inexactes*, dont l'analyse de convergence est fournie dans [Armand 12].

Pour cela, il faut distinguer le cas de critères sans pénalisation de ceux avec pénalisation.

1. Dans le cas d'un critère des moindres carrés non-pénalisés, la matrice $\boldsymbol{H}_k$ est une matrice bloc-diagonale contenant P blocs distincts, carrés de taille $(N-1)\times(N-1)$. L'inverse de cette matrice s'obtient simplement en calculant l'inverse de chacun des blocs,

$$\boldsymbol{H}_k = \text{Bdiag}(\boldsymbol{Z}^{\text{t}}(\boldsymbol{K}^{\text{t}}\boldsymbol{K} + \boldsymbol{D}_{p,k})\boldsymbol{Z}), \quad (\forall\, p = 1,\ldots,P), \tag{3.19}$$

où $\boldsymbol{D}_{p,k}$ est une matrice diagonale de taille $N\times N$ [Chouzenoux 14]. Cette structure a comme conséquence que la résolution du système primal se simplifie en la résolution de P systèmes indépendants de taille $(N-1)\times(N-1)$.

2. Dans le cas d'une pénalisation donnée sous une forme séparable

$$R(\boldsymbol{X}) = \sum_{p=1}^{P} \varphi(\boldsymbol{x}_p), \tag{3.20}$$

avec $\varphi(\cdot)$ est une fonction de pondération différentiable et strictement convexe, la matrice $\boldsymbol{H}_k$ sera toujours bloc-diagonale et peut être inversée simplement,

$$\boldsymbol{H}_k = \text{Bdiag}\left(\boldsymbol{Z}^{\text{t}}(\boldsymbol{K}^{\text{t}}\boldsymbol{K} + \boldsymbol{D}_{p,k} + \beta\,\text{Diag}(\ddot{\varphi}(\boldsymbol{Z}\boldsymbol{a}_{p,k}))\boldsymbol{Z}\right). \tag{3.21}$$

3. Dans le cas d'un critère de régularisation spatiale

$$R(\boldsymbol{X}) = \sum_{p=1}^{P} \varphi([\boldsymbol{\nabla}\boldsymbol{X}]_p), \tag{3.22}$$

où $\boldsymbol{\nabla} \in \mathbb{R}^{Q\times N}$ est une matrice de différentiation qui va introduire des dépendances entre pixels voisins. Il s'en suit que

$$\boldsymbol{H}_k = \text{Bdiag}(\boldsymbol{Z}^{\text{t}}(\boldsymbol{K}^{\text{t}}\boldsymbol{K} + \boldsymbol{D}_{p,k})\boldsymbol{Z}) + \beta(\boldsymbol{\nabla}\otimes\boldsymbol{Z})^{\text{t}}\text{Diag}(\ddot{\varphi}((\boldsymbol{\nabla}\otimes\boldsymbol{Z})\boldsymbol{a}_k))(\boldsymbol{\nabla}\otimes\boldsymbol{Z}), \tag{3.23}$$

où $\ddot{\varphi}(\cdot)$ est la dérivée seconde de $\varphi(\cdot)$. Dans ce cas, l'introduction de l'opérateur $\boldsymbol{\nabla}$ a comme conséquence que la matrice $\boldsymbol{H}_k$ ne sera plus bloc-diagonale. La résolution exact du système primal (3.7) va nécessiter un coût de calcul ou mémoire très important.

3.2.2 Résolution tronquée du système primal [Moussaoui 12, Chouzenoux 13b]

Une première version accélérée de l'algorithme *PDIP* consiste à tronquer la résolution du système primal (3.7). Plus précisément, la direction primale $\boldsymbol{d}_k^a$ est obtenue en appliquant J itérations

d'un algorithme itératif de résolution du système linéaire $\boldsymbol{H}_k\boldsymbol{d} = -\boldsymbol{g}_k$. L'algorithme utilisé est le gradient biconjugué [Van der Vorst 92] auquel on incorpore une stratégie de préconditionnement basée sur une décomposition LU incomplète de $\boldsymbol{H}_k$. Le nombre de sous-itérations J est contrôlé par un seuil sur la valeur de la norme du résidu

$$\|\boldsymbol{r}_J\| \leqslant \mu_k \|\boldsymbol{r}_0\|, \tag{3.24}$$

où $\boldsymbol{r}_J$ est la valeur du résidu du système (3.7) et μ_k est le paramètre de perturbation des conditions de KKT. Ce critère d'arrêt est inspiré de ceux proposés dans [Armand 12] où l'on peut trouver également d'autres critères d'arrêt. Cette approche de résolution approchée du système primal présente une similarité avec l'approche qui consiste à employer une algorithme de Newton tronqué [Dembo 83] pour la résolution du système primal-dual (3.5). L'analyse de convergence de l'algorithme primal-dual *inexact* est réalisée dans [Armand 12].

3.2.3 Résolution du système primal par majoration-minimisation [Legendre 14]

L'équation (3.7) est équivalente à la résolution du problème d'optimisation

$$\begin{array}{c}\text{Trouver } \boldsymbol{d}_k \\ \text{tel que } \boldsymbol{d}_k = \underset{\boldsymbol{d}\in\mathbb{R}^{(N-1)P}}{\arg\min}\ f(\boldsymbol{d}) = -\boldsymbol{g}_k^{\mathrm{t}}\boldsymbol{d} + \frac{1}{2}\boldsymbol{d}^{\mathrm{t}}\boldsymbol{H}_k\boldsymbol{d}.\end{array} \tag{3.25}$$

La proposition consiste à utiliser une approche MM [Hunter 04] en s'appuyant sur une approximation majorante quadratique. Ainsi, l'approximation tangente en $\boldsymbol{d}_k^j \in \mathbb{R}^{(N-1)P}$ est donnée sous la forme

$$h_j(\boldsymbol{d}, \boldsymbol{d}_k^j) = f(\boldsymbol{d}_k^j) + \frac{1}{2}(\boldsymbol{d} - \boldsymbol{d}_k^j)^{\mathrm{t}}(\boldsymbol{B}_k - \boldsymbol{H}_k)(\boldsymbol{d} - \boldsymbol{d}_k^j), \tag{3.26}$$

où $\boldsymbol{B}$ est une matrice symétrique semi-définie positive choisie de telle sorte à respecter la condition de majoration de la fonction $f(\cdot)$ par $h_j(\cdot)$. Une valeur possible de $\boldsymbol{B}_k$, obtenue par la technique de majoration par maximum de courbure, est fournie dans [Legendre 14]. La minimisation de $h_j(\cdot)$ conduit à la récurrence MM suivante,

$$\begin{cases} \boldsymbol{d}_k^0 = \boldsymbol{0}, \\ \boldsymbol{d}_k^{j+1} = \boldsymbol{d}_k^j - \boldsymbol{B}_k^{-1}(\boldsymbol{g}_k + \boldsymbol{H}_k\boldsymbol{d}_k^j), \ \text{ pour } \ j = 1, \ldots, J, \\ \boldsymbol{d}_k^a = \boldsymbol{d}_k^J, \end{cases} \tag{3.27}$$

dont le nombre d'itérations J est contrôlé en utilisant le test d'arrêt (3.24).

3.2.4 Avantages de la résolution approchée

En plus de l'évitement de l'inversion d'une matrice de grande taille, la résolution tronquée du système primal présente deux autres avantages

1. La résolution du système primal par gradient biconjugué ou par approche MM fait appel au calcul de produits matrice-vecteur qui peuvent se faire avec une complexité arithmétique réduite. En fait, les matrices $\boldsymbol{H}_k$ et $\boldsymbol{B}_k$ sont déduites de la matrice $\boldsymbol{K}$ qui correspond à des opérateurs qui peuvent se calculer à l'aide d'algorithmes rapides. De plus la coût mémoire sera réduit car cette approche ne nécessite pas d'expliciter la matrice $\boldsymbol{K}$.
2. Un gain supplémentaire est obtenu dans le cadre de l'approche MM en optant pour des matrices $\boldsymbol{B}_k$ bloc-diagonales (*i.e.*, des approximations quadratiques majorantes séparables). Par conséquent, la résolution du système linéaire se réduit à la résolution de P systèmes linéaires indépendants de taille $(N-1) \times (N-1)$.

3.3 Accélération matérielle pour des problèmes de grande taille

Compte tenu de l'évolution croissante des outils de calcul intensif, l'augmentation de la puissance des calculateurs est également exploitée pour la mise en place de stratégies d'accélération des approches de résolution itérative. L'idée serait de proposer des méthodes ayant des structures algorithmiques qui se prêtent à une implémentation parallèle ou à un calcul distribué. Pour cela, on explorera dans ce qui suit le potentiel des processeurs de cartes graphiques (GPUs) et les contraintes d'implémentation que leur utilisation impose.

3.3.1 Calcul scientifique sur des processeurs de cartes graphiques

Un GPU est un processeur contenant plusieurs centaines d'unités de calcul pouvant effectuer de façon parallèle les mêmes opérations sur des données mémoire différentes. Les GPU sont de plus en plus utilisés dans le milieu scientifique car ils offrent une puissance de calcul importante pour un prix raisonnable [Owens 08]. Comme illustré par la figure 3.1, le GPU se distingue du CPU par le grand nombre d'Unités Arithmétiques et Logiques (UAL) qu'il comporte, souvent plusieurs centaines contre 2 ou 4 pour un CPU. Ces unités effectuent nécessairement les mêmes instructions sur des données différentes, suivant le modèle SIMD (*Single Instruction Multiple Data*).

L'implémentation réalisée utilise la technologie CUDA qui est développée par Nvidia [Nvidia 12] et se présente comme une extension du langage C. Elle permet d'exécuter des portions de code appelés *noyaux* de façon parallèle sur un GPU. Un noyau est exécuté à travers un grand nombre de *threads*. Selon ce modèle, les threads sont autant de processus indépendants réalisant la suite d'instructions définie dans le noyau sur des données différentes. Un thread est donc exécuté sur une

FIGURE 3.1 – Illustration de la différence entre GPU et CPU. Extrait de [Nvidia 12].

UAL. Les threads étant souvent bien plus nombreux que les UAL disponibles, le modèle CUDA dispose de règles de contrôle d'exécution permettant d'optimiser l'exécution des threads en fonction du nombre de d'UAL disponibles. Une présentation plus détaillée sur la programmation GPU est disponible dans [Wilt 13].

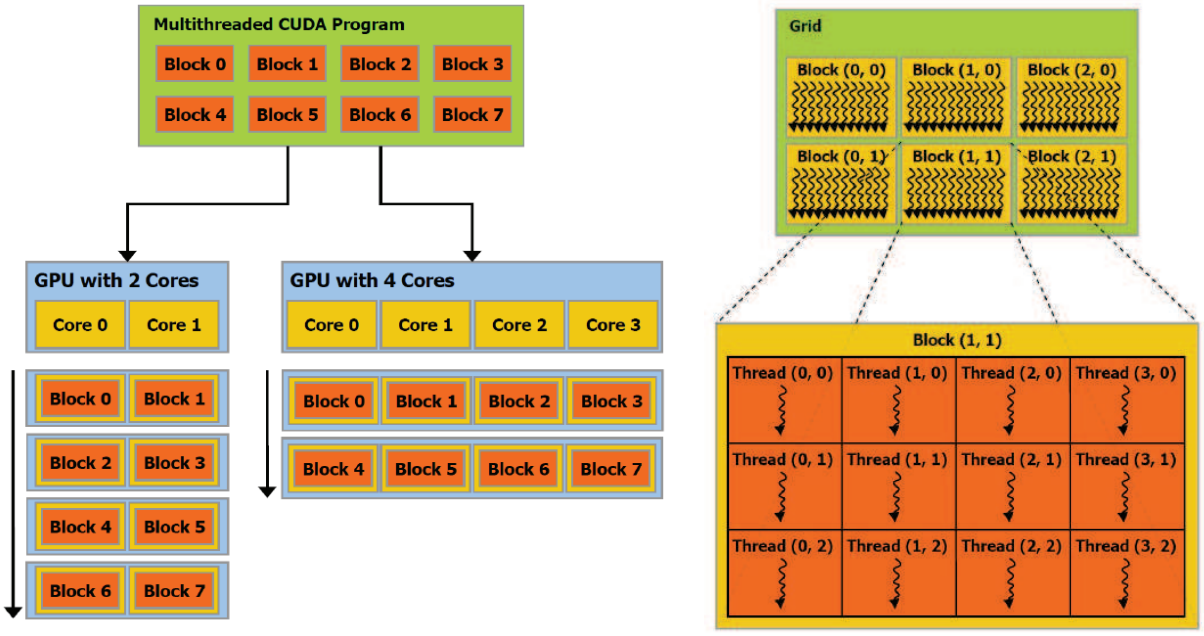


FIGURE 3.2 – Illustration du principe du multithreading selon le nombre de multiprocesseurs GPU ou d'UAL disponibles et du mode d'organisation des threads en blocs et grille.

Des travaux précurseurs sur l'utilisation des GPU pour la résolution de problèmes d'optimisation en traitement du signal et de l'image peuvent être trouvés dans [Ruggiero 10, Sánchez 11].

3.3.2 Implémentation GPU de l'algorithme primal-dual des points intérieurs [Legendre 13b, Legendre 13a]

Il est naturel de penser que c'est en rendant le problème totalement parallélisable que le GPU sera le mieux exploité. Cela est possible en remarquant que la minimisation du critère des moindres carrés défini par l'équation (3.16), est assurée en résolvant des minimisations indépendantes des critères

$$F_p(\boldsymbol{x}) = \|\boldsymbol{y}_p - \boldsymbol{K}\boldsymbol{x}_p\|_2^2, \qquad (\forall p = 1, \ldots, P).$$

Ainsi, P threads peuvent être lancés sur le GPU, chacun exécutant entièrement l'algorithme PDIP pour chaque sous-problème. Néanmoins, cette méthode présente un inconvénient : Les structures conditionnelles de type « if, then, else » au sein d'un noyau sont à éviter. Dans le modèle défini par CUDA, les threads sont organisés par groupes de 32, appelés *warps*, se comportant comme autant d'unités SIMD. En cas de structure conditionnelle, si les deux conditions sont vérifiées par des threads différents d'un même warp (groupe de 32 threads), alors les instructions liées aux deux conditions sont exécutées par tous les threads de ce warp. Des opérations inutiles sont alors effectuées bien que leurs résultats soient ignorés, ce qui a pour effet d'augmenter le temps d'exécution. Par conséquent, lors de l'exécution de l'algorithme PDIP le temps total est fixé par le sous-problème dont le traitement nécessite le plus d'itérations dans chaque sous-groupe.

L'alternative est de revenir au problème initial et de minimiser le critère global en tirant parti au mieux des caractéristiques différentes du CPU et du GPU. Le CPU est utilisé pour implémenter la structure de l'algorithme, contenant les conditions d'arrêt. Le GPU est utilisé au sein de chaque étape afin d'en accélérer l'exécution. Cette version présente l'avantage de ne pas introduire de calculs inutiles, cependant certaines étapes de l'algorithme nécessitent des transferts de données entre la mémoire du CPU et celle du GPU, ce qui ralentit leur exécution.

D'après la figure (3.3), on peut distinguer deux types d'étapes : celles pour lesquelles une parallélisation totale est possible car elles contiennent des calculs indépendants sur chaque sous-problème, et celles dont le résultat est une variable unique pour le critère global. C'est le cas pour le calcul du paramètre barrière, du pas de Newton, ainsi que des deux tests d'arrêt. Ce type d'étape est appelée *réduction* et n'est effectuée que partiellement sur le GPU. En effet il est plus avantageux pour ce type d'étape d'effectuer une réduction partielle sur le GPU, puis de transférer ce résultat intermédiaire dans la mémoire du CPU pour y terminer le calcul. Une organisation optimale de la mémoire est indispensable pour réduire les temps d'accès aux données et donc éviter les temps d'attente [Legendre 13a]. Il est à noter que cette stratégie d'implémentation est indispensable lorsqu'un critère de régularisation spatiale $R(\boldsymbol{X})$ est introduit dans le critère à minimiser.

Les résultats de l'application de cette approche au traitement d'images hyperspectrales sont décrits dans [Chouzenoux 14, Legendre 14], en annexe de ce manuscrit. Un travail en cours concerne l'implémentation sur GPU de la méthode intégrant une pénalisation spatiale sur le cartes de distribution spatiale des abondances.

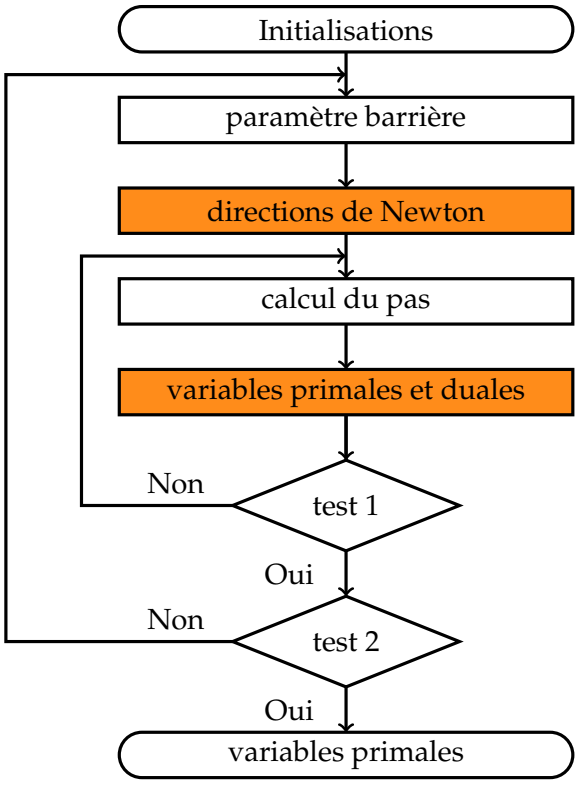


FIGURE 3.3 – Organigramme de la méthode primale-duale des points-intérieurs. Les étapes grisées sont réalisées sur le GPU alors que les autres sont réalisées partiellement sur le GPU, puis finalisées sur le CPU.

3.4 Conclusions

J'ai présenté dans ce chapitre deux stratégies d'accélération des méthodes d'optimisation sous contraintes pour la résolution de problèmes inverses de grande taille. La première proposition concerne l'utilisation de versions tronquées des algorithmes de résolution itératives et d'adapter les stratégie de barrière pour assurer la convergence de l'algorithme primal-dual des points intérieurs. La seconde concerne l'utilisation des algorithmes de majoration-minimisation pour la résolution approchée de manière à assurer la séparabilité du critère majorant. On exploite ainsi l'un des points forts des méthodes MM pour que l'étape de minimisation pour la mise à jour des variables primales et duales puisse être implémentée sur le GPU. Une perspective à court-terme de ce travail concerne l'implémentation GPU de l'algorithme primal-dual incluant une stratégie de résolution du système primal par un algorithme MM ainsi que l'évaluation du gain apporté par cette implémentation.

Ce qu'il faut retenir du travail présenté dans ce chapitre est le succès apporté par l'approche consistant à agir en amont de l'implémentation matérielle en modifiant la méthode de traitement de telle manière à ce que l'algorithme résultant présente une structure fortement parallélisable tout en préservant des propriétés théoriques saines.

Chapitre 4

Simulation bayésienne pour l'inférence statistique en grandes dimensions

Ce chapitre est consacré aux travaux que j'ai réalisés récemment sur les méthodes de Monte Carlo par chaînes de Markov (MCMC) en grande dimension. Le recours aux méthodes MCMC pour l'inférence bayésienne permet de résoudre des problèmes inverses tout en incluant naturellement une étape d'estimation de tous les paramètres d'un modèle bayésien hiérarchique [Robert 01]. Cependant, l'utilisation de ces techniques pour la résolution de problèmes de grande taille est parfois inenvisageable à cause d'un coût de calcul trop élevé qui résulte, soit d'un besoin mémoire excessif ou d'une convergence trop lente de l'algorithme.

Ces travaux sont motivés par des problématiques constatées lors de la mise en œuvre des méthodes de séparation de sources par approche bayésienne ainsi que d'un questionnement sur la possibilité d'exploiter des outils issus de l'optimisation au sein des techniques d'échantillonnage. La finalité étant d'augmenter l'efficacité des algorithmes MCMC dans ce contexte. Bien que cette interrogation est connue de longue date dans la communauté des mathématiques appliquées, celle-ci suscite ces dernières années un très fort engouement au sein de la communauté du traitement du signal et des images. Citons pour exemple les travaux précurseurs sur des méthodes de Langevin-Hastings exploitant les informations sur les gradient de la distribution cible [Rossky 78, Roberts 96] ou encore, plus récemment celles, utilisant le hessien [Vacar 11, Zhang 11, Martin 12]. On retrouve cette stratégie également dans les approches variationnelles [Fraysse 11].

Le cas d'étude qui sera discuté dans ce chapitre est celui où une des variables à rééchantillonner est un vecteur gaussien de grande dimension dont la covariance varie au cours des itérations. Les développements présentés dans ce chapitre sont issus des travaux de thèse de doctorat de Clément GILAVERT [1]. Ces développements ont fait l'objet d'un papier récemment accepté pour publication dans une revue internationale [Gilavert 14] et d'une communication nationale [Gilavert 13].

1. Doctorant ECN, 2012-2014. Cette thèse actuellement suspendue pour des raisons médicales.

4.1 Echantillonnage gaussien en grande dimension

On s'intéresse à l'échantillonnage gaussien dans le cadre de la résolution d'un problème inverse linéaire, au sens que les mesures $\boldsymbol{y} \in \mathbb{R}^M$ sont exprimées par

$$\boldsymbol{y} = \boldsymbol{K}\boldsymbol{x} + \boldsymbol{e}, \tag{4.1}$$

avec $\boldsymbol{K} \in \mathbb{R}^{M \times N}$ est la matrice associée au modèle d'observation et $\boldsymbol{e}$ un terme de bruit additif, correspondant aux erreurs de mesure et de modélisation, indépendant de la variable d'intérêt $\boldsymbol{x}$. L'estimation de $\boldsymbol{x}$ par inférence bayésienne [Robert 01, Idier 08] requière tout d'abord la formulation de la densité de probabilité *a posteriori* $P(\boldsymbol{x}, \Theta|\boldsymbol{y})$, où Θ est l'ensemble des hyperparamètres. Une suite de réalisations $\{\boldsymbol{x}_k\}$ est alors simulée à partir de cette densité et utilisée pour l'approximation des statistiques d'intérêt, telles que la moyenne, le mode ou la covariance de la variable aléatoire associée au vecteur $\boldsymbol{x}$. Idéalement, les réalisations $\{\boldsymbol{x}_k\}$ doivent être indépendantes mais, la génération d'une telle suite est rarement simple en pratique. L'alternative consiste à construire une chaîne de Markov ayant comme distribution asymptotique la densité *a posteriori* [Gilks 99]. Une méthode classique pour la construction d'une telle chaîne est l'échantillonneur de Gibbs [Geman 84] dont le principe est de simuler itérativement des réalisations $(\Theta_k, \boldsymbol{x}_k)$ à partir des densités conditionnelles $P(\Theta|\boldsymbol{x}_{k-1}, \boldsymbol{y})$ et $P(\boldsymbol{x}|\Theta_k, \boldsymbol{y})$.

4.1.1 Formulation du problème

Dans le cas où des modèles gaussions $\mathcal{N}\left(\boldsymbol{\mu}_y, \boldsymbol{R}_y\right)$ et $\mathcal{N}\left(\boldsymbol{\mu}_x, \boldsymbol{R}_x\right)$ sont affectés aux statistiques du bruit $\boldsymbol{e}$ et au vecteur des inconnues $\boldsymbol{x}$, l'ensemble des hyperparamètres Θ correspond aux moyennes et les matrices de covariances de ces deux distributions. Notons que cette loi *a priori* couvre une famille plus large de modèles bayésiens hiérarchiques tels que le modèle de mélange continu de gaussiennes [Andrews 74, Champagnat 04] et les modèles à base de champs de Markov gaussiens [Geman 84, Papandreou 10].

En s'appuyant sur un tel modèle, la densité conditionnelle $P(\boldsymbol{x}|\Theta, \boldsymbol{y})$ est une distribution gaussienne, $\mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{Q}^{-1}\right)$, dont la matrice de précision $\boldsymbol{Q}$ (*i.e.*, l'inverse de la matrice de covariance) est donnée par

$$\boldsymbol{Q} = \boldsymbol{H}^{\mathrm{t}}\boldsymbol{R}_y^{-1}\boldsymbol{H} + \boldsymbol{R}_x^{-1}, \tag{4.2}$$

et sa moyenne $\boldsymbol{\mu}$ est telle que

$$\boldsymbol{Q}\boldsymbol{\mu} = \boldsymbol{H}^{\mathrm{t}}\boldsymbol{R}_y^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y) + \boldsymbol{R}_x^{-1}\boldsymbol{\mu}_x. \tag{4.3}$$

On peut constater que la matrice de précision $\boldsymbol{Q}$ dépend des hyperparamètres Θ à travers $\boldsymbol{R}_y$ et $\boldsymbol{R}_x$, ce qui induit une variation de cette matrice tout au long des itérations de l'algorithme de Gibbs. De plus, la moyenne $\boldsymbol{\mu}$ est solution d'un système d'équations dépendant de $\boldsymbol{Q}$.

4.1.2 Echantillonnage gaussien

L'approche classique pour la génération de vecteurs aléatoires gaussiens [Wold 48, Scheuer 62] consiste à (1) calculer la factorisation de Cholesky de la matrice de covariance $\boldsymbol{R} = \boldsymbol{L}_r\boldsymbol{L}_r^{\mathrm{t}}$, (2) générer un échantillon $\boldsymbol{x} = \boldsymbol{R}(\boldsymbol{Q}\boldsymbol{\mu}) + \boldsymbol{L}_r\boldsymbol{\omega}$, avec $\boldsymbol{\omega} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Or, l'équation (4.2) ne donne pas la matrice de covariance. Pour éviter l'inversion matricielle, la solution proposée par [Rue 01] est de réaliser la factorisation de Cholesky de la matrice de précision $\boldsymbol{Q} = \boldsymbol{L}_q\boldsymbol{L}_q^{\mathrm{t}}$ et de calculer $\boldsymbol{x} = \boldsymbol{L}_q^{-\mathrm{t}}\left(\boldsymbol{L}_q^{-1}\boldsymbol{Q}\boldsymbol{\mu} + \boldsymbol{\omega}\right)$. Cette dernière étape est moins couteuse car elle consiste en la résolution séquentielle de deux systèmes d'équations triangulaires. Mais ces approches fondées sur la factorisation de Cholesky nécessitent $\mathcal{O}(N^3)$ opérations si la matrice n'a pas de structure spécifique, ce qui les rend infaisables en grande dimension. Il est vraie que si la matrice $\boldsymbol{Q}$ (ou $\boldsymbol{R}$) possède une structure particulière, la factorisation peut se faire avec un coût réduit : $\mathcal{O}(N^2)$ lorsque $\boldsymbol{Q}$ est Toeplitz [Trench 64] ou encore $\mathcal{O}(N \log N)$ lorsque $\boldsymbol{Q}$ est circulante [Geman 95]. Les matrices sparses se factorisent aussi avec un coût réduit. C'est d'ailleurs ce qui a motivé la proposition de Rue [Rue 01]. Cependant, même dans le cas favorable, une telle factorisation reste toujours une opération lourde à réaliser à chaque itération de l'échantillonneur de Gibbs.

Une autre approche, appelée *Perturbation-Optimization* [Orieux 12] (appelée aussi *Independent Factor Perturbation* dans [Papandreou 10]), consiste à simuler un échantillon $\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{Q}\boldsymbol{\mu}, \boldsymbol{Q})$, puis à retenir la solution du système $\boldsymbol{Q}\boldsymbol{x} = \boldsymbol{\eta}$ comme réalisation du vecteur gaussien. On peut vérifier aisément que le nouvel échantillon est distribué selon $\mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{Q}^{-1}\right)$. On retrouve la même stratégie dans les travaux de [Lalanne 01, Tan 10, Bardsley 12].

Bien que la simulation de $\boldsymbol{\eta}$ soit très facile dans le cadre de problèmes inverses linéaires, le problème lié au coût de calcul élevé reste toujours d'actualité car la complexité de la résolution exacte du système est identique à celle de la factorisation de Cholesky. C'est pourquoi ces mêmes travaux préconisent de retenir une solution tronquée de ce système $\boldsymbol{Q}\boldsymbol{x} = \boldsymbol{\eta}$ en employant un algorithme de gradient conjugué linéaire avec un nombre réduit d'itérations ($J_k << N$). Cependant, l'effet de cette troncature en termes de convergence vers la loi cible n'a pas été étudié.

La première contribution de ce travail consiste à mettre en évidence que cette troncature empêche la convergence vers la loi cible et de proposer l'incorporation d'une étape d'acceptation-rejet peu coûteuse rétablissant la convergence. Pour cela, l'échantillonnage gaussien est formulé dans le cadre des méthode fondées sur les méthodes MCMC à sauts réversibles [Green 95, Waagepetersen 01] à dimension invariante. La seconde contribution est d'analyser l'efficacité statistique de cet algorithme afin de développer une stratégie de réglage adaptatif du niveau de troncature de la résolution du système linéaire de sorte à optimiser le coût de calcul de l'échantillonneur.

4.2 Méthode MCMC à sauts réversibles

La méthode d'échantillonnage consiste à construire une chaîne de Markov dont la distribution converge asymptotiquement vers la loi cible $P_{\boldsymbol{X}}(\cdot)$. Pour cela, plaçons nous dans le cadre des méthodes MCMC à saut réversibles (RJ-MCMC) [Green 95].

Dans ce schéma, une variable auxiliaire $\boldsymbol{z} \in \mathbb{R}^L$, est d'abord simulée à partir d'une distribution $P_{\boldsymbol{Z}}(\boldsymbol{z}|\underline{\boldsymbol{x}})$ qui dépend de la réalisation courante $\underline{\boldsymbol{x}} \in \mathbb{R}^N$. Ensuite, un mouvement deterministe est réalisé selon une transformation $\phi(\cdot)$,

$$\begin{aligned} \boldsymbol{\phi} : \left(\mathbb{R}^N \times \mathbb{R}^L\right) &\mapsto \left(\mathbb{R}^N \times \mathbb{R}^L\right) \\ (\underline{\boldsymbol{x}}, \boldsymbol{z}) &\mapsto (\boldsymbol{x}, \boldsymbol{s}) \end{aligned}$$

qui doit être réversible, c'est à dire $\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{s}) = (\underline{\boldsymbol{x}}, \boldsymbol{z})$. Le nouvel échantillon $\bar{\boldsymbol{x}}$ est ensuite obtenu en soumettant $\boldsymbol{x}$, résultant de la transformation, à un test d'acceptation-rejet avec une probabilité d'acceptation

$$\alpha(\underline{\boldsymbol{x}}, \boldsymbol{x}) = \min\left(1, \frac{P_{\boldsymbol{X}}(\boldsymbol{x}) P_{\boldsymbol{Z}}(\boldsymbol{s}|\boldsymbol{x})}{P_{\boldsymbol{X}}(\underline{\boldsymbol{x}}) P_{\boldsymbol{Z}}(\boldsymbol{z}|\underline{\boldsymbol{x}})} |J_{\boldsymbol{\phi}}(\underline{\boldsymbol{x}}, \boldsymbol{z})|\right),$$

où $J_{\boldsymbol{\phi}}(\underline{\boldsymbol{x}}, \boldsymbol{z})$ est le determinant du jacobien de la transformation $\boldsymbol{\phi}(\cdot)$ en $(\underline{\boldsymbol{x}}, \boldsymbol{z})$. En pratique, le choix de la loi conditionnelle $P_{\boldsymbol{Z}}(\cdot)$ ansi que de la transformation $\boldsymbol{\phi}(\cdot)$ doit être adapté à la distribution cible $P_{\boldsymbol{X}}(\cdot)$.

4.2.1 Cas de l'échantillonnage de vecteurs gaussiens

Pour échantillonner une distribution gaussienne $\boldsymbol{x} \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{Q}^{-1}\right)$, une généralisation du schéma adopté dans [De Forcrand 99] consiste à prendre $L = N$, et à définir une variable auxiliaire $\boldsymbol{z} \in \mathbb{R}^N$ distribuée selon

$$P_{\boldsymbol{Z}}(\boldsymbol{z}|\underline{\boldsymbol{x}}) = \mathcal{N}\left(\boldsymbol{A}\underline{\boldsymbol{x}} + \boldsymbol{b}, \boldsymbol{B}\right), \tag{4.4}$$

où $\boldsymbol{A} \in \mathbb{R}^{N \times N}$, $\boldsymbol{B} \in \mathbb{R}^{N \times N}$ et $\boldsymbol{b} \in \mathbb{R}^N$ sont des paramètres dont le choix sera discuté plus loin. De plus, la transformation déterministe $\phi(\cdot)$ sera choisie telle que

$$\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{s} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\phi}_1(\underline{\boldsymbol{x}}, \boldsymbol{z}) \\ \boldsymbol{\phi}_2(\underline{\boldsymbol{x}}, \boldsymbol{z}) \end{pmatrix} = \begin{pmatrix} -\underline{\boldsymbol{x}} + \boldsymbol{f}(\boldsymbol{z}) \\ \boldsymbol{z} \end{pmatrix}, \tag{4.5}$$

avec des fonctions $\left(\boldsymbol{f} : \mathbb{R}^N \mapsto \mathbb{R}^N\right)$, $\left(\boldsymbol{\phi}_1 : \left(\mathbb{R}^N \times \mathbb{R}^N\right) \mapsto \mathbb{R}^N\right)$ et $\left(\boldsymbol{\phi}_2 : \left(\mathbb{R}^N \times \mathbb{R}^N\right) \mapsto \mathbb{R}^N\right)$.

Proposition 1 ([Gilavert 14]). *Soit une variable auxiliaire $\boldsymbol{z}$ obtenue selon (4.4) et une proposition d'échantillon $\boldsymbol{x}$ résultant de (4.5). Le probabilité d'acceptation de cet échantillon vaut*

$$\alpha(\underline{\boldsymbol{x}}, \boldsymbol{x}|\boldsymbol{z}) = \min\left(1, e^{-\boldsymbol{r}(\boldsymbol{z})^{\mathrm{t}}(\underline{\boldsymbol{x}}-\boldsymbol{x})}\right), \tag{4.6}$$

avec

$$\boldsymbol{r}(\boldsymbol{z}) = \boldsymbol{Q}\boldsymbol{\mu} + \boldsymbol{A}^{\mathrm{t}}\boldsymbol{B}^{-1}(\boldsymbol{z}-\boldsymbol{b}) - \frac{1}{2}\left(\boldsymbol{Q} + \boldsymbol{A}^{\mathrm{t}}\boldsymbol{B}^{-1}\boldsymbol{A}\right)\boldsymbol{f}(\boldsymbol{z}). \tag{4.7}$$

En particulier, cette probabilité d'acceptation vaut 1 lorsque $\boldsymbol{f}(\boldsymbol{z})$ set définie comme la solution exacte de

$$\frac{1}{2}\left(\boldsymbol{Q} + \boldsymbol{A}^{\mathrm{t}}\boldsymbol{B}^{-1}\boldsymbol{A}\right)\boldsymbol{f}(\boldsymbol{z}) = \boldsymbol{Q}\boldsymbol{\mu} + \boldsymbol{A}^{\mathrm{t}}\boldsymbol{B}^{-1}(\boldsymbol{z}-\boldsymbol{b}). \tag{4.8}$$

De plus, dans le cas d'une résolution exacte de ce système, la corrélation entre deux échantillons consécutifs vaut zéro seulement et seulement si les matrices $\boldsymbol{A}$ et $\boldsymbol{B}$ sont choisies tel que

$$\boldsymbol{A}^{\mathrm{t}}\boldsymbol{B}^{-1}\boldsymbol{A} = \boldsymbol{Q}. \tag{4.9}$$

4.2.2 Algorithme RJPO

Considérons la variable auxiliaire $\boldsymbol{z}$ dont la distribution est définie par (4.4) avec

$$\boldsymbol{A} = \boldsymbol{B} = \boldsymbol{Q} \quad \text{and} \quad \boldsymbol{b} = \boldsymbol{Q}\boldsymbol{\mu}. \tag{4.10}$$

Ce choix permet, d'une part, de respecter la condition (4.9) de la proposition 1 et, d'autre part, de réduire l'équation (4.8) à la résolution d'un système linéaire de la forme $\boldsymbol{Q}\boldsymbol{f}(\boldsymbol{z}) = \boldsymbol{z}$, ce qui permet d'établir le lien avec la méthode PO. En effet, d'après une telle paramétrisation, la variable auxiliaire $\boldsymbol{z}$ aura comme distribution $\mathcal{N}(\boldsymbol{Q}\underline{\boldsymbol{x}} + \boldsymbol{Q}\boldsymbol{\mu}, \boldsymbol{Q})$ et peut ainsi être exprimée sous la forme $\boldsymbol{z} = \boldsymbol{Q}\underline{\boldsymbol{x}} + \boldsymbol{\eta}$, avec $\boldsymbol{\eta} \sim (\boldsymbol{Q}\boldsymbol{\mu}, \boldsymbol{Q})$. Par conséquent, la simulation de la variable $\boldsymbol{z}$ se réduit à la simulation de $\boldsymbol{\eta}$, ce qui correspond à l'étape de *perturbation* dans l'algorithme PO.

Génération de la variable auxiliaire. Dans [Papandreou 10, Orieux 12], une méthode simple de simulation de $\boldsymbol{\eta}$ est proposée. Celle-ci consiste à exploiter l'expression (4.3) et à perturber chaque terme séparément

1. Simuler $\boldsymbol{\eta}_y \sim \mathcal{N}\left(\boldsymbol{y} - \boldsymbol{\mu}_y, \boldsymbol{R}_y\right)$,
2. Simuler $\boldsymbol{\eta}_x \sim \mathcal{N}\left(\boldsymbol{\mu}_x, \boldsymbol{R}_x\right)$,
3. Définir $\boldsymbol{\eta} = \boldsymbol{H}^{\mathrm{t}}\boldsymbol{R}_y^{-1}\boldsymbol{\eta}_y + \boldsymbol{R}_x^{-1}\boldsymbol{\eta}_x$, un échantillon de $\mathcal{N}(\boldsymbol{Q}\boldsymbol{\mu}, \boldsymbol{Q})$.

Il est important de noter qu'une telle astuce est très intéressante car les matrices $\boldsymbol{R}_y$ et $\boldsymbol{R}_x$ possèdent des structures très simples et sont parfois même diagonales.

En fait, l'étape de perturbation peut être appliquée pour la simulation de toute distribution gaussienne dont la matrice de précision $\boldsymbol{Q}$ est disponible sous une forme factorisée $\boldsymbol{Q} = \boldsymbol{F}^{\mathrm{t}}\boldsymbol{F}$, avec une matrice $\boldsymbol{F} \in \mathbb{R}^{N' \times N}$. Dans ce cas, $\boldsymbol{\eta} = \boldsymbol{Q}\boldsymbol{\mu} + \boldsymbol{F}^{\mathrm{t}}\boldsymbol{w}$, où $\boldsymbol{\omega} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{N'})$.

Génération du nouvel échantillon. Dans le cas (4.10), l'équation (4.7) se simplifie en

$$\boldsymbol{r}(\boldsymbol{z}) = \boldsymbol{z} - \boldsymbol{Q}\boldsymbol{f}(\boldsymbol{z}). \tag{4.11}$$

Par conséquent, une première version de l'algorithme RJPO est la suivante

1. Simuler $\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{Q}\boldsymbol{\mu}, \boldsymbol{Q})$,
2. Prendre $\boldsymbol{z} = \boldsymbol{Q}\underline{\boldsymbol{x}} + \boldsymbol{\eta}$. Résoudre le système linéaire $\boldsymbol{Q}\boldsymbol{u} = \boldsymbol{z}$, de façon approchée. Soit $\widehat{\boldsymbol{u}}$ la solution retenue avec un résidu $\boldsymbol{r}(\boldsymbol{z}) = \boldsymbol{z} - \boldsymbol{Q}\widehat{\boldsymbol{u}}$ et noter $\widehat{\boldsymbol{x}} = -\underline{\boldsymbol{x}} + \widehat{\boldsymbol{u}}$, l'échantillon proposé.
3. Avec une probabilité $\min\left(1, e^{-\boldsymbol{r}(\boldsymbol{z})^{\mathrm{t}}(\underline{\boldsymbol{x}} - \widehat{\boldsymbol{x}})}\right)$ prendre $\bar{\boldsymbol{x}} = \widehat{\boldsymbol{x}}$, ou retenir $\bar{\boldsymbol{x}} = \underline{\boldsymbol{x}}$.

Remarques :

- Dans le cas d'une résolution tronquée du système dans l'étape 2, la solution retenue peut dépendre du point initial $\boldsymbol{u}_0$. Or, $\boldsymbol{f}(\boldsymbol{z})$ ne doit pas dépendre de $\underline{\boldsymbol{x}}$, pour que le mouvement (4.5) soit toujours réversible. Donc le point initial $\boldsymbol{u}_0$ ne doit pas dépendre de $\underline{\boldsymbol{x}}$. Un choix, par défaut, est $\boldsymbol{u}_0 = \boldsymbol{0}$.
- Une version plus compacte de l'échantillonneur peut être obtenue en substituant $\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{z}) - \underline{\boldsymbol{x}}$ dans l'équation (4.11). Celle-ci se réduit à la résolution du système $\boldsymbol{Q}\boldsymbol{x} = \boldsymbol{\eta}$. L'étape 2 de l'algorithme RJPO se simplifie donc en :

  2. Résoudre le système linéaire $\boldsymbol{Q}\boldsymbol{x} = \boldsymbol{\eta}$ de façon approchée. Soit $\widehat{\boldsymbol{x}}$ la solution retenue et $\boldsymbol{r}(\boldsymbol{z}) = \boldsymbol{\eta} - \boldsymbol{Q}\widehat{\boldsymbol{x}}$.

4.2.3 Lien avec l'algorithme PO

D'après la proposition 1, la résolution exacte du système (4.8) conduit à une probabilité d'acceptation qui vaut 1. La procédure d'échantillonnage résultante est comme suit

1. Simuler $\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{Q}\boldsymbol{\mu}, \boldsymbol{Q})$,
2. Calculer $\boldsymbol{z} = \boldsymbol{Q}\underline{\boldsymbol{x}} + \boldsymbol{\eta}$,
3. Prendre $\bar{\boldsymbol{x}} = -\underline{\boldsymbol{x}} + \boldsymbol{Q}^{-1}\boldsymbol{z}$.

Notons que $\bar{\boldsymbol{x}} = -\underline{\boldsymbol{x}} + \boldsymbol{Q}^{-1}(\boldsymbol{Q}\underline{\boldsymbol{x}} + \boldsymbol{\eta}) = \boldsymbol{Q}^{-1}\boldsymbol{\eta}$. Par conséquent, la variable auxilliarie $\boldsymbol{z}$ n'est plus nécessaire car les étapes 2 et 3 de l'algorithme peuvent être fusionnées en une seule

2. Prendre $\bar{\boldsymbol{x}} = \boldsymbol{Q}^{-1}\boldsymbol{\eta}$.

Discussion :

- Dans le cas de la résolution exacte, l'algorithme de simulation par RJMCMC coïncide avec l'algorithme PO proposé dans [Orieux 12].
- Pour les mêmes raisons que précédemment, le point initial $\boldsymbol{x}_0$ de résolution du système linéaire doit être choisi de sorte à ce que $\boldsymbol{u}_0 = \boldsymbol{x}_0 + \underline{\boldsymbol{x}}$ soit indépendant de $\underline{\boldsymbol{x}}$. Par conséquent, des choix tels quel $\boldsymbol{x}_0 = \boldsymbol{0}$ ou $\boldsymbol{x}_0 = \underline{\boldsymbol{x}}$ ne sont pas autorisés, alors que $\boldsymbol{x}_0 = -\underline{\boldsymbol{x}}$ est le choix par défaut correspondant à $\boldsymbol{u}_0 = \boldsymbol{0}$.

4.2.4 Illustration du comportement pathologique du PO tronqué

Considérons une distribution gaussienne multivariée en dimension $N = 20$, de moyenne $\boldsymbol{\mu}$ et de matrice de covariance $\boldsymbol{R}$ définis par

$$R_{ij} = \sigma^2 \rho^{|i-j|}, \ (\forall i = 1, \ldots, N; \forall j = 1, \ldots, N), \tag{4.12}$$

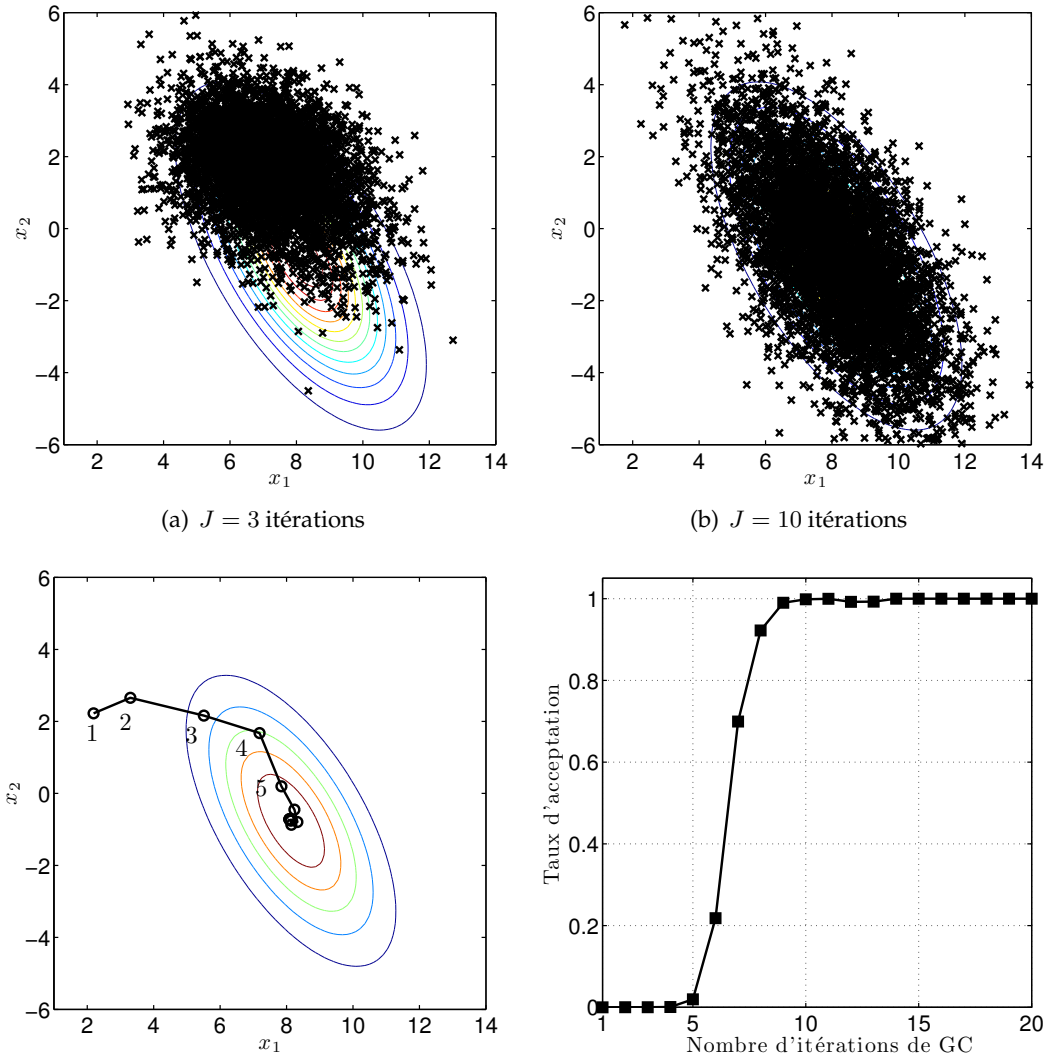
$$\mu_i \sim \mathcal{U}[0, 10], \quad (\forall i = 1, \ldots, N), \tag{4.13}$$

avec $\sigma^2 = 1$ et $\rho = 0.8$. Après calcul de la matrice de précision $\boldsymbol{Q}$ et du produit $\boldsymbol{Q}\boldsymbol{\mu}$, l'algorithme PO tronqué est appliqué pour générer 5000 échantillons avec plusieurs niveaux de troncature (nombre de sous-itérations de gradient conjugué, noté J).

On peut constater sur la figure 4.1 qu'une troncature prématurée, avec $J < 5$, conduirait à une distribution complètement différente de la loi cible. Cependant, grâce à la formulation de l'échantillonnage dans le cadre des RJMCMC, il est possible de déduire à partir de la figure 4.1(d) qu'il faut augmenter le nombre de sous-itérations de gradient conjugué pour obtenir une taux d'acceptation suffisant. On peut observer aussi qu'une résolution exacte n'est pas nécessaire car le taux d'acceptation est pratiquement égal à un, dès lors que $J > 9$.

Discussion.

- Ce résultat permet de conclure que l'idée de tronquer la résolution est judicieuse, mais une étape d'acceptation-rejet est nécessaire pour assurer un comportement sain de l'échantillonneur.
- Le choix du seuil de troncature doit permettre d'optimiser le coût de calcul de l'échantillonneur en réalisant le meilleur compromis entre nombre d'itérations par échantillon et convergence rapide de la chaîne. En effet, le seuil de troncature va dépendre à la fois de la dimension du problème ainsi que du conditionnement de la matrice $\boldsymbol{Q}$.
- Une analyse plus détaillée de l'influence du seuil de troncature sur le comportement de l'échantillonneur est réalisée dans le papier [Gilavert 14], fourni dans l'annexe A.5 de ce manuscrit .



(a) $J = 3$ itérations

(b) $J = 10$ itérations

(c) Moyenne empirique

(d) Taux d'acceptation

FIGURE 4.1 – Illustration du comportement pathologique du PO tronqué sur un problème de petite taille. On peut constater que l'effet néfaste de la troncature est perceptible pour un faible nombre d'itérations de gradient conjugué.

4.3 Optimisation du coût de calcul des échantillonneurs

Afin de proposer une stratégie de réglage automatique du seuil de troncature, une analyse de l'efficacité statistique de l'échantillonneur est d'abord réalisée.

4.3.1 Efficacité statistique

Les performances de l'algorithme RJPO sont analysées en analysant les propriétés de la chaîne de Markov en termes de nombre effectif d'échantillons (*effective sample size*) [Liu 08, p. 125]. Ce

critère statistique, proposé par Goodman et Sokal dans [Goodman 89], donne le nombre d'échantillons indépendants, n_{eff}, qui conduiraient à la même variance d'approximation empirique de l'estimateur bayésien, calculé avec $n_{\max}$ échantillons de la chaîne simulée. Ce critère est relié à la fonction d'autocorrélation de la chaîne par

$$n_{\text{eff}} = \frac{n_{\max}}{1 + 2\sum_{k=1}^{\infty} \rho_k}, \tag{4.14}$$

où ρ_k est le coefficient d'autocorrelation d'ordre k. Sous l'hypothèse d'une chaîne modélisée par un processus autoregressif d'ordre 1, $\rho_k = \rho^k$, l'équation (4.14) conduit à un ESS relatif (ESSR pour ESS *ratio*)

$$\text{ESSR} = \frac{n_{\text{eff}}}{n_{\max}} = \frac{1-\rho}{1+\rho}. \tag{4.15}$$

On peut constater que ESSR=1 lorsque les échantillons sont indépendants $(\rho = 0)$ et diminue lorsque la corrélation de la chaîne augmente. En pratique, il serait judicieux d'exprimer le coût de calcul par échantillon effectif (CCES pour *Computation Cost per Effective Sample*), que l'on pourrait définir dans le cadre de l'algorithme RJPO, par

$$\text{CCES} = \frac{J_{\text{tot}}}{n_{\text{eff}}} = \frac{J}{\text{ESSR}} \tag{4.16}$$

où $J = J_{\text{tot}}/n_{\max}$ est le nombre d'itérations de gradient conjugué par échantillon de la chaîne. En pratique, cette relation permet de définir le nombre d'itérations requises, $n_{\max}$, pour chaque niveau de troncature, pour obtenir des estimateurs équivalents (c'est à dire, ayant le même ESS).

4.3.2 Optimisation du coût de calcul

Tout niveau de troncature J (ou valeur de résidu relatif correspondant, ϵ) va induire un niveau de corrélation de la chaîne qui va permettre de déduire l'ESSR et le CCES d'après (4.16). L'objectif est donc de trouver une stratégie de réglage du seuil de résolution en se basant sur le résidu relatif de telle manière à minimiser le CCES. L'ESSR exprimé par (4.15) dépend de la corrélation de la chaîne ρ, qui est une fonction implicite du taux d'acceptation α. Pour $\alpha = 1$, $\rho = 0$ d'après la proposition (1). Pour $\alpha = 0$, $\rho = 1$ puisqu'aucun nouvel échantillon ne sera accepté. Pour des valeurs intermédiaires de α, la corrélation décroit de 1 à 0. Celle-ci peut être décomposée en deux termes

- Avec une probabilité $(1 - \alpha)$, la procédure d'acceptation-rejet produit des échantillons identiques en cas de rejection,
- Dans le cas de l'acceptation, l'échantillon proposé sera corrélé avec l'échantillon précédent à cause de la résolution tronquée du système linéaire

Alors qu'on arrive assez facilement à exprimer la corrélation induite par réjection, il n'est pas aisé de trouver une formulation explicite de la corrélation en cas d'acceptation. Cependant, des tests empiriques ont montré que cette corrélation reste négligeable comparée à celle induite par la

réjection. En négligeant cette corrélation, il en résulte que $\rho = 1 - \alpha$ et ESSR $= \dfrac{\alpha}{2-\alpha}$.

Par conséquent, le réglage optimal J_{opt} du seuil de troncature, permettant de minimiser le CCES, coût de calcul par échantillon effectif, est solution de

$$J\,\frac{d\alpha}{dJ} - \alpha + \frac{\alpha^2}{2} = 0. \tag{4.17}$$

La figure 4.2 illustre, dans le cas d'une gaussienne de dimension $N = 128$, de paramètres définis par (4.12), l'existence d'un réglage optimal permettant de réaliser un compromis entre résolution exacte du système (qui va nécessiter beaucoup de sous-itérations de gradient conjugué et un taux d'acceptation égal à 1) et une résolution grossière qui aura comme conséquence une chaîne fortement corrélée et beaucoup d'itérations n_{max} avant convergence.

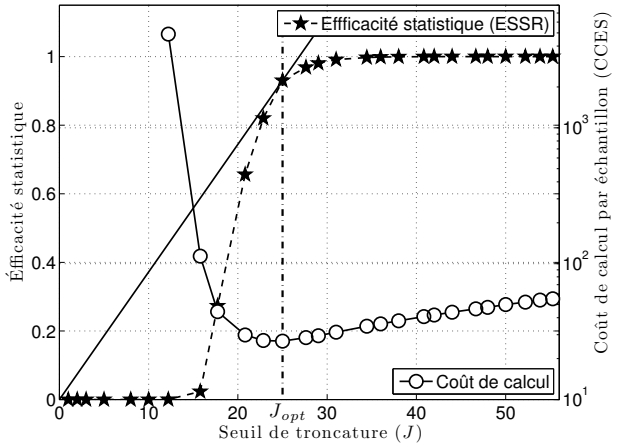


FIGURE 4.2 – Evolution du coût de calcul par échantillon effectif (CCES) et du taux d'acceptation α en fonction du seuil de troncature J , pour l'échantillonnage d'une distribution gaussiene en dimension $N = 128$. Le meilleur réglage du seuil de troncature est $J_{\text{opt}} = 26$.

4.3.3 Réglage auto-adaptatif du niveau de troncature

La courbe $\alpha(J)$ n'étant pas connue au préalable, la recherche du réglage du seuil de troncature permettant de satisfaire (4.17) est réalisé récursivement en utilisant l'algorithme de Robbins-Monro [Robbins 51, Bercu 12]. Il s'agit d'un algorithme adaptatif stochastique [Benveniste 12] permettant de résoudre une équation non-linéaire $g(\theta) = 0$ en utilisant une récurrence

$$\theta_{n+1} = \theta_n + \gamma_n \left[g(\theta_n) + \nu_n \right] \tag{4.18}$$

avec ν is une variable aléatoire traduisant l'incertitude associée à l'évaluation de $g(\cdot)$ et $\{\gamma_n\}$ est une séquence de pas permettant d'assurer la convergence de l'algorithme [Andrieu 01, Andrieu 06]. Une telle procédure a été déjà employée pour le réglage optimal des algorithmes MCMC adap-

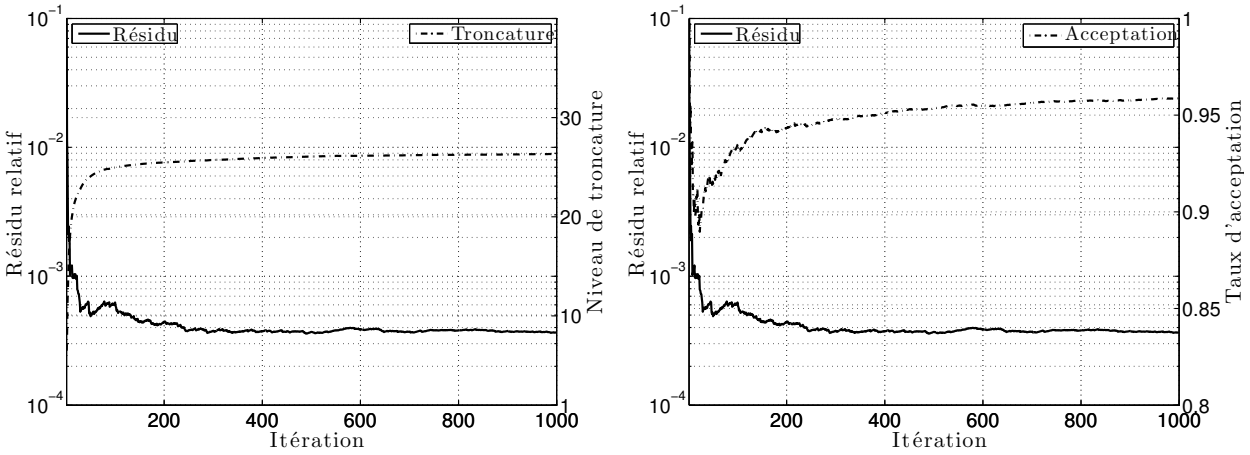
tatifs [Andrieu 08], tels que l'algorithme RWMH (*Random Walk Metropolis-Hastings*) [Haario 01] et MALA (*Metropolis-Adjusted Langevin Algorithm*) [Atchadé 05]. Le réglage adaptatif a été mis en oeuvre dans [Haario 01, Atchadé 05, Achadé 06] de façon à définir les paramètres des algorithmes RWMH et MALA permettant d'atteindre le taux d'acceptation optimal suggéré dans [Roberts 97, Gelman 96] pour RWMH et dans [Roberts 98] pour MALA.

Afin d'assurer la positivité de résidu relatif ϵ, la mise à jour est réalisée sur son logarithme. A chaque itération n de l'échantillonneur la valeur du résidu relatif est ajusté selon

$$\log \epsilon_{n+1} = \log \epsilon_n + \gamma_n \left(J_n \frac{d\alpha_n}{dJ} - \alpha_n + \frac{\alpha_n^2}{2} \right), \tag{4.19}$$

où le gradient $\frac{d\alpha_n}{dJ}$ est évalué numériquement et le pas de mise à jour sont issus d'une séquence décroissante vers 0 pour assurer la convergence vers la loi cible. Comme suggéré dans [Andrieu 08], une solution simple consiste à prendre $\gamma_n = \gamma_0/n^\beta$, avec $\kappa \in]0, 1]$.

La figure 4.3 illustre l'évolution du niveau de troncature lors de l'incorporation du réglage adaptatif de l'algorithme RJPO. On peut constater une convergence rapide du résidu relatif vers une valeur permettant d'obtenir une nombre de sous-itérations de gradient conjugué proche de la valeur optimale J_{opt}, constatée sur la figure 4.2. De plus, le taux d'acceptation est $\alpha = 0.96$.



(a) Niveau de troncature

(b) Taux d'acceptation

FIGURE 4.3 – Evolution de la valeur de la norme du résidu relatif au cours des itérations de l'échantillonneur RJPO et niveau de troncature moyen pour l'échantillonnage d'une gaussien en dimension $N = 128$. La valeur moyenne de J est proche de la valeur optimale $J_{\text{opt}} = 26$ constatée dans la figure 4.2.

4.4 Conclusions

J'ai présenté dans ce chapitre les premiers résultats de travaux sur les méthodes de Monte Carlo en grande dimension. La contribution majeure consiste, d'une part, à corriger l'erreur commise dans des algorithmes d'échantillonnage gaussien faisant appel à une résolution tronquée d'un système linéaire et, d'autre part, à proposer une stratégie de réglage automatique du seuil de troncature de cette résolution de telle sorte à optimiser le temps de calcul. L'algorithme résultant rentre dans la famille de algorithmes MCMC adaptatifs qui ne nécessite aucun paramètre de réglage, contrairement aux versions adaptatives des algorithme RWMH et MALA, car le taux d'acceptation optimal s'ajuste naturellement en fonction du seuil de troncature permettant d'optimiser le temps de calcul par échantillon effectif. Les illustrations ont été réalisées sur des problèmes de faible dimension mais la méthode proposée a été effectivement appliquée avec succès à un problème de taille réaliste de restauration d'images [Gilavert 14]. Un travail en cours concerne l'exploitation de cette technique pour la restauration non-supervisée d'images de spectroscopie Raman.

Troisième partie

Perspectives et projet scientifique

Chapitre 5

Perspectives et projet scientifique

5.1 Sur le volet enseignement

Mon activité d'enseignement à l'école centrale de Nantes est réalisée au sein d'un département couvrant le domaine de l'EEA (robotique, commande de systèmes, informatique embarquée et traitement du signal). Paradoxalement, bien que le numérique occupe une place très importante dans l'industrie et dans la vie de tous les jours, le recrutement d'étudiants dans cette discipline devient de plus en plus difficile. En tant que porteur d'une option disciplinaire « signaux-images » à l'école centrale de Nantes, je souhaite mener un travail de fond pour d'abord comprendre les attentes des élèves-ingénieurs, en termes de projet professionnel individuel, avant de réorienter cette offre de formation pour concilier ces attentes avec les besoins du monde industriel.

Une autre piste de travail à mener consiste à faire comprendre que toute application issue de l'informatique graphique ou tactile s'appuie sur une base de données qui doit nécessairement être alimentée régulièrement. Or, la fréquence de rafraîchissement de cette base et la fiabilité des informations injectées sont fortement liés aux performances des méthodes de traitement de signal.

Je pense, par ailleurs, qu'il faut orienter la formation ingénieur vers une mode de fonctionnement proche de celui de la formation par apprentissage : alterner les phases d'acquisition de connaissances théoriques fondamentales et de maîtrise de techniques appliquées à travers des projets réalistes. L'implication des industriels dans la formation est aussi un point important pour ce type de fonctionnement dans une offre de formation ingénieur.

Un autre chantier important sur lequel un travail de taille est a réaliser concerne l'amélioration de la visibilité du parcours *signal-image* au sein de la formation du Master. En effet, cette formation est non seulement un point d'entrée pour le recrutement de nouveaux chercheurs mais aussi un bon moyen de diffusion des résultats de la recherche.

5.2 Sur le volet recherche

Sur le plan de la recherche fondamentale, l'apport essentiel de mes travaux se situe dans les outils d'optimisation et de simulation de Monte Carlo et dans leur mise en œuvre pour la résolution de problèmes inverses de grande taille. Ce type de problèmes apparaissent dans plusieurs domaines de l'instrumentation industrielle et scientifique. On peut citer à titre d'exemples, les réseaux de capteurs, les instruments d'imagerie à haute résolution spatiale et temporelle, les techniques d'imagerie multimodale ou hyperspectrale. Les catégories de problèmes inverses sur lesquels j'ai travaillé jusqu'à présent ont concerné l'identification de systèmes, la séparation de sources, la décomposition de signaux et la restauration d'images. Sur la plan de la recherche appliquée, j'ai eu l'opportunité de m'impliquer dans des collaborations industrielles ou académiques qui ont permis de remplir pleinement l'une des missions de mon métier d'enseignant-chercheur qui consiste en la veille technologique et le transfert de connaissances.

Mon projet de recherche s'inspire des développements réalisés et des résultats obtenus durant ces travaux. Ses grandes lignes ont pour objectif commun de concilier mes compétences acquises dans le domaine de l'optimisation, de la simulation statistique et la résolution de problèmes inverses. Il y'aura évidemment une continuité à court et moyen termes sur certains aspects et, comme je l'ai fait jusqu'à présent, des ouvertures thématiques ne seront pas exclues à plus long terme, du moment que celles-ci seraient liées au traitement statistique du signal et de l'image. Le fil conducteur de ce projet de recherche part du principe que l'évolution des techniques de mesure doit être accompagnée par des développements méthodologiques en traitement du signal capables de répondre, d'une part, aux contraintes pratiques liées aux temps de calcul et, d'autre part, à la nécessité d'améliorer les résultats de traitement en s'appuyant sur des modèles et des méthodes mathématiques adaptés.

Je vais détailler dans les sections suivantes ont les différentes facettes de ce projet focalisé sur le traitement de données massives.

5.2.1 Autour de la résolution par minimisation d'un critère composite

Un grand effort méthodologique a été consenti pour le développement et la mise en œuvre de méthodes d'optimisation adaptées aux propriétés du critères à minimiser. J'ai travaillé sur la technique de descente itérative pour laquelle une stratégie de recherche de pas fondée sur les méthodes de majoration-minimisation a été proposée.

A court terme, deux points restent à réaliser pour compléter cette étude. Le premier concerne le cas d'une descente itérative dans un sous-espace de directions en présence d'un critère de barrière. La technique de descente dans un sous-espace permet d'accélérer la convergence des algorithmes. Néanmoins, l'approche actuelle est dédiée aux critères à gradient Lipschitz. La méthode de sous-espaces a été appliquée dans [Skilling 84] pour la minimisation d'un critère de maximum d'en-

tropie, sans garantie théorique de convergence. Je pense que la formulation du problème dans un cadre MM avec une technique de majoration telle que celle proposée par De Pierro [De Pierro 95] serait une bonne approche pour proposer un algorithme efficace. Le deuxième point est lié à l'établissement du lien entre l'approche MM et les techniques à base de régions de confiance. Ces techniques se fondent sur une approximation du critère dans un périmètre défini par le rayon de la région de confiance. Or, on peut songer à une stratégie similaire en terme de majoration locale.

A long terme, un sujet qui a été peu traité dans la communauté du traitement du signal et de l'image est celui de l'estimation du paramètre de régularisation ainsi que les différents paramètres du critère de pénalisation. Il y'a, certes, les méthodes basées sur la courbe en L [Hansen 93], la courbe en S [Butler 81] ou la validation croisée [Lukas 93] mais je pense qu'il faut traiter le problème de façon globale. En fait, la recherche du paramètre de régularisation peut être vu comme un problème qui fait appel à plusieurs résolutions du problème de minimisation du critère composite pour des valeurs fixes du paramètre de régularisation. L'idée que je souhaite développer consiste à exploiter le potentiel des méthodes d'optimisation globale, utilisées actuellement dans la thèse de Joan DAVIS-VALLDAURA, pour la recherche du meilleure réglage avec un minimum de résolutions de problèmes intermédiaires, qui sont très coûteux dans le cas de problèmes de grande taille. Une telle question peut faire l'objet de travail de thèse de doctorat.

5.2.2 Autour des méthodes de simulation statistique

Les résultats de la thèse de Clément Gilavert, ont permis de développer une approche originale pour l'échantillonnage de vecteurs gaussiens ainsi que l'optimisation de sa mise en œuvre.

A court terme, un point qui mérite une attention particulière est la formulation de l'échantillonnage gaussien dans le cadre des méthodes de RWMH ou MALA et établir le lien avec la méthode RJPO. Je pense qu'il est possible d'établir un choix des paramètres du RJPO pour aboutir à des structures algorithmiques identiques. Cette analyse aura comme objectif la recherche des paramètres du RJPO (paramètres de la densité conditionnelle de la variable auxiliaire z) de telle manière à minimiser le coût de calcul de l'algorithme. Le lien avec des travaux récents sur l'utilisation des sous-espaces de Krylov pour l'échantillonnage gaussien [Parker 12] peut aussi être établi.

A long terme, je pense que la technique proposée pour l'optimisation du coût de calcul du RJPO peut être étendue à l'optimisation des algorithmes de simulation adaptative. Par exemple, les études empiriques sur le taux d'acceptation associé à un réglage optimal de l'algorithme MALA [Roberts 98] et de l'algorithme de RWMH [Roberts 97] ne tient pas compte du coût de calcul par itération. Une telle approche aura certainement des retombées importantes pour des problèmes de simulation en grande dimension car l'objectif est de maximiser l'efficacité statistique au prix d'un coût de calcul minimal. Cette perspective peut faire l'objet d'un travail d'une thèse de Master.

5.2.3 Coopération optimisation-simulation statistique-calcul intensif

La proposition de l'algorithme de Langevin dans [Rossky 78] fait apparaître explicitement le gradient de la distribution cible alors que dans les méthodes de Monte Carlo hybrides [Duane 87], une succession de minimisations d'une fonction d'énergie sont réalisées. On voit donc un apport significatif de l'optimisation pour améliorer les performances d'un algorithme d'échantillonnage. Cela s'est traduit par des travaux récents qui ont porté sur l'exploitation du hessien dans la mise en œuvre des méthodes de Langevin [Vacar 11, Zhang 11, Martin 12].

A court terme, il serait intéressant de coupler les méthodes de Monte Carlo hybrides avec la technique de descente dans un sous-espace de directions. On peut penser qu'une approximation majorante du critère pourrait être vue comme une loi instrumentale pour la proposition dans un formalisme de type Metropolis-Hastings. L'intérêt d'une telle approche est de pouvoir maîtriser le coût de calcul de l'algorithme. Une extension de ces approches est liée à la prise en compte de contraintes sur les variables à échantillonner telles que la positivité.

A long terme, comme il a été montré dans le chapitre 3, le temps de résolution d'un problème inverse peut être réduit significativement en exploitant les ressources offertes par les outils de calcul parallèle tels que les GPU. L'idée est d'exploiter les compétences acquises sur les trois outils pour proposer des méthodes de simulation incluant des ingrédients issus de l'optimisation avec des structures algorithmiques adaptées pour un calcul parallèle. Une telle approche serait intéressante pour des problèmes inverses de grande taille. Je pense par exemple à mes collaborations avec l'agence spatiale européenne qui souhaite explorer des masses de données importantes à l'aide des algorithmes de traitement non-supervisés.

5.2.4 A propos de l'application en spéctroscopie RMN

L'application à la spectroscopie RMN 2D (mode T1-T2) est un véritable cas de test pour les méthodes que j'ai développées jusqu'à présent. Il s'agit d'un problème inverse de grande taille, avec des matrices de problème direct fortement mal-conditionnées et des images à estimer qui doivent respecter des contraintes de non-négativité et de parcimonie. J'ai pu exploiter le fait que le noyau 2D est séparable pour proposer une stratégie d'implémentation adaptée.

A court terme, un premier travail consiste en l'application de la même approche au cas de la spectroscopie de diffusion (mode dit D-T2). Bien que ce problème ne présente pas de difficulté particulière, il permet de préparer le terrain pour un travail à **long terme** consistant en la reconstruction RMN 3D (D-T1-T2). Toute la difficulté de ce problème est la non séparabilité du modèle direct dans le plan (D,T1). Un sujet de thèse pour ce développement est en perspective avec Corinne Rondeau (IRSTEA, Rennes).

5.2.5 A propos de la fusion de données en séparation de sources

Les techniques d'instrumentation font appel de plus en plus à une diversification des modalité de mesure. En fait, selon les conditions expérimentales et environnementales, une information peut apparaître plus clairement selon un certain mode d'observation. Le traitement conjoint des données doit donc prendre en compte cette masse de données. Les problèmes à résoudre rentre dans le cadre de la fusion de données.

On sait que dans le cadre de la séparation de sources, l'introduction de nouvelles données va engendrer une restriction de l'intervalle des solutions admissibles mais le traitement des données doit s'accompagner de méthode adaptées.

A court terme, un problème à traiter concerne la séparation de sources de spectroscopie Raman dans le cas mélanges convolutif avec des opérateurs de convolution qui dépendent de l'interaction entre l'objet et l'instrument de mesure. On se place ainsi dans le cadre d'un problème de séparation convolutive paramétrique.

A long terme, un couplage avec une modalité de mesure permettant d'avoir une information sur la tomographie de l'échantillon va permettre de mieux caractériser cette interaction. Un montage de projet à ce sujet est en discussion avec Bernard HUMBERT de l'IMN (Institut des Matériaux à Nantes).

Bibliographie

[Achadé 06] Y. F. Achadé. *An Adaptive Version for the Metropolis Adjusted Langevin Algorithm with a Truncated Drift*. Methodology and Computing in Applied Probability, vol. 8, no. 2, pages 235–254, 2006.

[Allain 02] M. Allain. *Approche pénalisée en tomographie hélicoïdale en vue de l'application à la conception d'une prothèse personnalisée du genou*. Thèse de Doctorat, Ecole Polytechnique Montréal, Canada, http://tel.archives-ouvertes.fr/tel-00003756, déc. 2002.

[Allain 06] M. Allain, J. Idier & Y. Goussard. *On global and local convergence of half-quadratic algorithms*. IEEE Transactions on Image Processing, vol. 15, no. 5, pages 1130–1142, 2006.

[Amblard 08] P.O Amblard, S. Moussaoui, T. Dudok De Wit, J. Aboudarham, M. Kretzschmar, J. Lilensten & F. Auchère. *The EUV Sun as the superposition of elementary Suns*. Astronomy and Astrophysics, vol. 487, pages L13–L16, 2008.

[Andrews 74] D.F. Andrews & C.L. Mallows. *Scale mixtures of normal distributions*. Journal of the Royal Statistical Society B, pages 99–102, 1974.

[Andrieu 01] C. Andrieu & C.P Robert. Controlled MCMC for optimal sampling. Tech. Report No. 0125, Cahiers de Mathématiques du Ceremade, Université Paris-Dauphine, 2001.

[Andrieu 06] Christophe Andrieu, Eric Moulines & Pierre Priouret. *Stability of stochastic approximation under verifiable conditions*. SIAM Journal on Control, vol. 44, no. 1, pages 283–312, 2006.

[Andrieu 08] C. Andrieu & J. Thoms. *A tutorial on adaptive MCMC*. Statistics and Computing, vol. 18, no. 4, pages 343–373, 2008.

[Anstreicher 94] K. M. Anstreicher & J.-P. Vial. *On the convergence of an infeasible primal-dual interior-point method for convex programming*. Optimization Methods and Software, vol. 3, no. 4, pages 273–283, 1994.

[Armand 00] P. Armand, J. C. Gilbert & S. Jan-Jégou. *A feasible BFGS interior point algorithm for solving strongly convex minimization problems*. SIAM Journal on Optimization, vol. 11, pages 199–222, 2000.

[Armand 12] P. Armand, J. Benoist & J.-P. Dussault. *Local path-following property of inexact interior methods in nonlinear programming*. Computational Optimization and Applications, vol. 52, no. 1, pages 209–238, 2012.

[Atchadé 05] Y. F. Atchadé & J. S. Rosenthal. *On adaptive Markov chain Monte Carlo algorithms*. Bernouilli, vol. 11, no. 5, pages 815–828, 2005.

[Bardsley 12] J. M. Bardsley & C. Fox. *An MCMC method for uncertainty quantification in nonnegativity constrained inverse problems*. Inverse Problems in Science and Engineering, vol. 20, no. 4, pages 477–498, 2012.

[Benveniste 12] A. Benveniste, M. Métivier & P. Priouret. Adaptive algorithms and stochastic approximations. Springer Publishing Company, Incorporated, 2012.

[Bercu 12] B. Bercu & P. Fraysse. *A Robbins–Monro procedure for estimation in semiparametric regression models*. Annals of Statistics, vol. 40, no. 2, pages 666–693, 2012.

[Bertsekas 99] D. P. Bertsekas. Nonlinear programming. Athena Scientific, Belmont, MA, USA, 2ème edition, 1999.

[Böhning 88] D. Böhning & B. G. Lindsay. *Monotonicity of quadratic-approximation algorithms*. Annals of the Institute of Statistical Mathematics, vol. 40, no. 4, pages 641–663, 1988.

[Boyd 04] S. Boyd & L. Vandenberghe. Convex optimization. Cambridge University Press, New York, 1ère edition, 2004.

[Butler 81] J. P. Butler, J. A. Reeds & S. V. Dawson. *Estimating Solutions of First Kind Integral Equations with Nonnegative Constraints and Optimal Smoothing*. SIAM Journal on Numerical Analysis, vol. 18, no. 3, pages 381–397, juin 1981.

[Chambolle 12] A. Chambolle, D. Cremers & T. Pock. *A Convex Approach to Minimal Partitions*. SIAM Journal on Imaging Sciences, vol. 5, no. 4, pages 1113–1158, 2012.

[Champagnat 04] F. Champagnat & J. Idier. *A connection between half-quadratic criteria and EM algorithms*. IEEE Signal Processing Letters, vol. 11, no. 9, pages 709–712, 2004.

[Chan 99] T. F. Chan & P. Mulet. *On the convergence of the lagged diffusivity fixed point method in total variation image restoration*. SIAM Journal on Numerical Analysis, vol. 36, no. 2, pages 354–367, 1999.

[Chang 07] C.-I. Chang. Hyperspectral data exploitation. Wiley Interscience, 2007.

[Charbonnier 97] P. Charbonnier, L. Blanc-Féraud, G. Aubert & M. Barlaud. *Deterministic edge-preserving regularization in computed imaging*. IEEE Transactions on Image Processing, vol. 6, pages 298–311, 1997.

[Chouzenoux 09a] E. Chouzenoux, S. Moussaoui & J. Idier. *A Majorize-Minimize line search algorithm for barrier function optimization*. In Proc. EURASIP European Signal and Image Processing Conference (EUSIPCO), pages 1379–1383, Glasgow, UK, Aug. 2009.

[Chouzenoux 09b] E. Chouzenoux, S. Moussaoui, J. Idier & F. Mariette. *Reconstruction d'un spectre RMN 2D par maximum d'entropie*. In Actes du XXIIième Colloque GRETSI sur le Traitement du Signal et des Images, Dijon, France, Sep. 2009.

[Chouzenoux 10a] E. Chouzenoux. *Recherche de pas par Majoration-Minoration. Application à la résolution de problèmes inverses*. Thèse de Doctorat, Ecole Centrale de Nantes, France, http://tel.archives-ouvertes.fr/tel-00555643, décembre 2010.

[Chouzenoux 10b] E. Chouzenoux, S. Moussaoui, J. Idier & F. Mariette. *Efficient maximum entropy reconstruction of nuclear magnetic resonance T1-T2 spectra*. IEEE Transactions on Signal Processing, vol. 58, no. 12, pages 6040–6051, 2010.

[Chouzenoux 11a] E. Chouzenoux, J. Idier & S. Moussaoui. *A majorize-minimize strategy for subspace optimization applied to image restoration*. IEEE Transactions on Image Processing, vol. 20, no. 6, pages 1517–1528, 2011.

[Chouzenoux 11b] E. Chouzenoux, S. Moussaoui & J. Idier. *Efficiency of linesearch strategies in interior point methods for linearly constrained signal restoration*. In Proc. of IEEE Workshop on Statistical Signal Processing (SSP), Nice, France, juin 2011.

[Chouzenoux 12] E. Chouzenoux, S. Moussaoui & J. Idier. *Majorize-minimize linesearch for inversion methods involving barrier function optimization*. Inverse Problems, vol. 28, no. 6, page 065011, mai 2012.

[Chouzenoux 13a] E. Chouzenoux, A. Jezierska, J.-C. Pesquet & H. Talbot. *A Majorize-Minimize Subspace Approach for $\ell 2$-$\ell 0$ Image Regularization*. SIAM Journal on Imaging Sciences, vol. 6, no. 1, pages 563–591, 2013.

[Chouzenoux 13b] E. Chouzenoux, S. Moussaoui, M. Legendre & J. Idier. *Algorithme primal-dual de points intérieurs pour l'estimation pénalisée des cartes d'abondances en imagerie hyperspectrale*. Traitement du Signal, vol. 30, no. 1-2, pages 35–59, 2013.

[Chouzenoux 14] E. Chouzenoux, M. Legendre, S. Moussaoui & J. Idier. *Fast Constrained Least Squares Spectral Unmixing Using Primal-Dual Interior-Point Optimization*. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 7, no. 1, pages 59–69, 2014.

[Combettes 10] P. L. Combettes & J.-C. Pesquet. *Proximal splitting methods in signal processing*. In Fixed-Point Algorithms for Inverse Problems in Science and Engineering. Springer-Verlag, NewYork, NY, USA, 2010.

[Conn 96] A. Conn, N. Gould & P. L. Toint. *A primal-dual algorithm for minimizing a nonconvex function subject to bounds and nonlinear constraints*. In G. Di Pillo & F. Giannessi, editeurs, Nonlinear Optimization and Applications. Kluwer Academic Publishers, 2 edition, 1996.

[Cragg 69] E. E. Cragg & A. V. Levy. *Study on a supermemory gradient method for the minimization of functions*. Journal of Optimization Theory and Applications, vol. 4, no. 3, pages 191–205, 1969.

[De Forcrand 99] P. De Forcrand. *Monte Carlo Quasi-heatbath by approximate inversion*. Physical Review D, vol. 59, no. 3, pages 3698–3701, 1999.

[De Pierro 95] A.R. De Pierro. *A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography*. IEEE Transactions on Medical Imaging, vol. 14, no. 1, pages 132–137, mars 1995.

[Dembo 83] R. S. Dembo & T. Steihaug. *Truncated-Newton methods algorithms for large scale unconstrained optimization*. Mathematical Programming, vol. 26, pages 190–212, 1983.

[Demoment 89] G. Demoment. *Image Reconstruction and Restoration : Overview of Common Estimation Structure and Problems*. IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 37, no. 12, pages 2024–2036, Décembre 1989.

[Dempster 77] A. P. Dempster, N. M. Laird & D. B. Rubin. *Maximum likelihood from incomplete data via the EM algorithm*. Journal of the Royal Statistical Society B, vol. 39, pages 1–38, 1977.

[Dobigeon 09a] Nicolas Dobigeon, Saïd Moussaoui, Martical Coulon, Jean-Yves Tourneret & Alfred Hero. *Joint Bayesian Endmember Extraction and Linear Unmixing for Hyperspectral Imagery*. IEEE Transactions on Signal Processing, vol. 57, no. 11, pages 4355–4368, November 2009.

[Dobigeon 09b] Nicolas Dobigeon, Saïd Moussaoui, Jean-Yves Tourneret & Cédric Carteret. *Bayesian separation of spectral sources under non-negativity and full additivity constraints*. Signal Processing, vol. 89, no. 12, pages 2657 – 2669, December 2009.

[Duane 87] S. Duane, A. Kennedy, B Pendleton & D Roweth. *Hybrid Monte Carlo*. Physics Letters B, vol. 195, no. 2, pages 216–222, 1987.

[Duarte 09] L. T. Duarte, C. Jutten & S. Moussaoui. *A Bayesian Nonlinear Source Separation Method for Smart Ion-Selective Electrode Arrays*. IEEE Sensors Journal, vol. 9, no. 12, pages 1763–1771, December 2009.

[Duarte 10] L. T. Duarte, C. Jutten & S. Moussaoui. *Bayesian Source Separation of Linear and Linear-quadratic Mixtures Using Truncated Priors*. Journal of Signal Processing Systems, vol. 65, no. 3, pages 311–323, mai 2010.

[Dudok de Wit 13] T. Dudok de Wit, S. Moussaoui, C. Guénnou, F. Auchère, G. Cessateur, M Kretzschmar, L.E. Vieira & F. Goryaev. *Coronal Temperature Maps from Solar EUV Images : A Blind Source Separation Approach*. Solar Physics, vol. 283, no. 1, pages 31–47, 2013.

[Dusaussoy 95] N. J. Dusaussoy & I. E. Abdou. *The extended MENT algorithm : a maximum entropy type algorithm using prior knowledge for computerized tomography*. IEEE Transactions on Signal Processing, vol. 39, no. 5, pages 1164–1180, 1995.

[El-Bakry 96] A. S. El-Bakry, R. A. Tapia, T. Tsuchiya & Y. Zhang. *On the formulation and theory of the Newton interior-point method for nonlinear programming*. Journal of Optimization Theory and Applications, vol. 89, no. 3, pages 507–541, 1996.

[Erdogan 99] H. Erdogan & J.A. Fessler. *Monotonic Algorithms for Transmission Tomography*. IEEE Transactions on Medical Imaging, vol. 18, no. 9, pages 801–814, sep. 1999.

[Fessler 98] J.A. Fessler & H. Erdogan. *A paraboloidal surrogates algorithm for convergent penalized-likelihood emission image reconstruction*. IEEE Nuclear Science Symposium, vol. 2, pages 1132–1135, 1998.

[Fessler 99] J. A. Fessler & S. D. Booth. *Conjugate-Gradient Preconditiong Methods for Shift-Variant PET Image Reconstruction*. IEEE Transactions on Image Processing, vol. 8, no. 5, pages 688–699, mai 1999.

[Févotte 11] C. Févotte & J. Idier. *Algorithms for nonnegative matrix factorization with the beta-divergence*. Neural Computation, vol. 23, no. 9, pages 2421–2456, sep. 2011.

[Fiacco 68] A. V. Fiacco & G. P. McCormick. Nonlinear programming : Sequential unconstrained minimization techniques. John Wiley and Sons, Inc., 1968.

[Figueiredo 07] M.A.T. Figueiredo, J.M. Bioucas-Dias & R.D. Nowak. *Majorization-Minimization Algorithms for Wavelet-Based Image Restoration*. IEEE Transactions on Image Processing, vol. 16, no. 12, pages 2980–2991, 2007.

[Florescu 14] A. Florescu, E. Chouzenoux, J.-C. Pesquet, P. Ciuciu & S. Ciochina. *A Majorize-Minimize Memory Gradient Method for Complex-Valued Inverse Problems*. Signal Processing, vol. 103, pages 285–295, 2014.

[Forsgren 98] A. Forsgren & P. E. Gill. *Primal-Dual Interior Methods for Nonconvex Nonlinear Programming*. SIAM Journal on Optimization, vol. 8, pages 1132–1152, avr. 1998.

[Forsgren 02] A. Forsgren, P.E. Gill & M.H. Wright. *Interior Methods for Nonlinear Optimization*. SIAM Review, vol. 44, no. 4, pages 525–597, 2002.

[Fraysse 11] A. Fraysse & T. Rodet. *A gradient-like variational Bayesian algorithm*. In Proc. of IEEE Workshop on Statistical Signal Processing (SSP), pages pp. 605–608, 2011.

[Frieden 75] B. Frieden. *Image enhancement and restoration*. In Picture Processing and Digital Filtering, volume 6 of *Topics in Applied Physics*, pages 177–248. Springer-Verlag, New York, NY, USA, 1975.

[Frisch 55] K. R. Frisch. *The logarithmic potential method of convex programming*. Memorandum, University Institute of Economics, Oslo, Norway, may 1955.

[Gelman 96] A. Gelman, G. O. Roberts & W. R. Gilks. *Efficient Metropolis jumping rules*. In Bayesian statistics 5, pages 599–607. Oxford University Press, 1996.

[Geman 84] S. Geman & D. Geman. *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-6, no. 6, pages 721–741, nov. 1984.

[Geman 92] D. Geman & G. Reynolds. *Constrained restoration and the recovery of discontinuities*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 14, no. 3, pages 367–383, mars 1992.

[Geman 95] D. Geman & C. Yang. *Nonlinear image recovery with half-quadratic regularization*. IEEE Transactions on Image Processing, vol. 4, no. 7, pages 932–946, juil. 1995.

[Gilavert 13] C. Gilavert, S. Moussaoui & J. Idier. *Rééchantillonnage gaussien en grande dimension pour les problèmes inverses*. In Actes 24e coll. GRETSI, Brest, France, 2013.

[Gilavert 14] C. Gilavert, S. Moussaoui & J. Idier. Efficient gaussian sampling for solving large-scale inverse problems using MCMC methods. En révision dans IEEE Transaction on Signal Processing, 2014.

[Gilks 99] W.R. Gilks, S. Richardson & D.J. Spiegehalter. Markov Chain Monte Carlo in practice. Chapman & Hall, London, UK, 1999.

[Goodman 89] J. Goodman & A. D. Sokal. *Multigrid Monte Carlo method. Conceptual foundations*. Physical Review D, vol. 40, no. 6, 1989.

[Gordon 71] R. Gordon & G. T. Herman. *Reconstruction of pictures from their projections*. Communications of the ACM, vol. 14, no. 12, pages 759–768, 1971.

[Green 95] P. J. Green. *Reversible jump Markov chain Monte Carlo computation and Bayesian model determination*. Biometrika, vol. 82, pages 711–732, 1995.

[Haario 01] H. Haario, E. Saksman & J. Tamminen. *An adaptive Metropolis algorithm*. Bernouilli, vol. 7, no. 2, pages 223–242, 2001.

[Hansen 93] P. C. Hansen & D. P. O'Leary. *The use of the L-curve in the regularization of discrete ill-posed problems*. SIAM Journal on Scientific Computing, vol. 14, no. 6, pages 1497–1503, nov. 1993.

[Holland 92] J. H. Holland. Adaptation in natural and artificial systems. MIT Press, Cambridge, MA, USA, 1992.

[Hunter 04] D. R. Hunter & Kenneth L. *A Tutorial on MM Algorithms*. The American Statistician, vol. 58, no. 1, pages 30–37, fév. 2004.

[Idier 01] J. Idier. *Convex Half-Quadratic Criteria and Interacting Auxiliary Variables for Image Restoration*. IEEE Transactions on Image Processing, vol. 10, no. 7, pages 1001–1009, juil. 2001.

[Idier 08] J. Idier. Bayesian approach to inverse problems. ISTE Ltd and John Wiley & Sons Inc, 2008.

[Jacobson 07] M.W. Jacobson & J.A. Fessler. *An Expanded Theoretical Treatment of Iteration-Dependent Majorize-Minimize Algorithms*. IEEE Transactions on Image Processing, vol. 16, no. 10, pages 2411–2422, oct. 2007.

[Jansson 97] P. Jansson. Deconvolution of images and spectra. Academic Press, San Diego, California, USA, 2ème edition, 1997.

[Jaynes 03] E. T. Jaynes. Probability theory : The logic of science. Cambridge, 2003.

[Johnson 00] C. A. Johnson, J. Seidel & A. Sofer. *Interior-Point Methodology for 3-D PET Reconstruction*. IEEE Transactions on Medical Imaging, vol. 19, no. 4, avr. 2000.

[Johnson 03] C. A. Johnson & D. McGarry. *Maximum Entropy Reconstruction Methods in Electron Paramagnetic Resonance Imaging*. Annals of Operations Research, vol. 119, pages 101–118, 2003.

[Jutten 91] C. Jutten & J. Hérault. *Blind separation of sources, part I : An adaptive algorithm based on neuromimetic architecture*. Signal Processing, vol. 24, no. 1, pages 1–10, 1991.

[Karmarkar 84] N. K. Karmarkar. *A new polynomial–time algorithm for linear programming*. Combinatorica, vol. 4, pages 373–395, 1984.

[Kelly 99] C. T. Kelly. Iterative methods for optimization. Society for Industrial and Applied Mathematics, 1999.

[Labat 06] C. Labat. *Algorithmes d'optimisation de critères pénalisés pour la restauration d'images. Application à la déconvolution de trains d'impulsions en imagerie ultrasonore*. Thèse de Doctorat, Ecole Centrale de Nantes, France, http://tel.archives-ouvertes.fr/tel-00132861, Décembre 2006.

[Labat 08] C. Labat & J. Idier. *Convergence of conjugate gradient methods with a closed-form stepsize formula*. Journal of Optimization Theory and Applications, vol. 136, no. 1, pages 43–60, jan. 2008.

[Lalanne 01] P. Lalanne, D. Prévost & P. Chavel. *Stochastic artificial retinas : algorithm, optoelectronic circuits, and implementation*. Applied Optics, vol. 40, no. 23, pages 3861–3876, 2001.

[Lange 87] K. Lange, M. Bahn & R. Little. *A theoretical study of some maximum likelihood algorithms for emission and transmission tomography*. IEEE Transactions on Image Processing, vol. 6, no. 2, pages 106–114, juin 1987.

[Lange 94] K. Lange. *An adaptive barrier method for convex programming*. Methods and Applications of Analysis, vol. 1, no. 4, pages 392–402, 1994.

[Lange 00] K. Lange, R. R. Hunter & I Yang. *Optimization Transfer Using Surrogate Objective Functions*. Journal of Computational and Graphical Statistics, vol. 9, no. 1, pages 1–20, 2000.

[Legendre 13a] M. Legendre, S. Moussaoui, F. Schmidt & J. Idier. *Parallel implementation of a primal-dual interior-point optimization method for fast abundance maps estimation*. In Proc. of IEEE Workshop on Hyperspectral Image and Signal Processing : Evolution in Remote Sensing (WHISPERS), Gainesville, Etats-Unis, 2013.

[Legendre 13b] M. Legendre, A. Schmidt, S. Moussaoui & U. Lammers. *Solving Systems of Linear Equations by GPU-based Matrix Factorization in a Science Ground Segment*. Astronomy and Computing, vol. 3-4, pages 58–64, 2013.

[Legendre 14] M. Legendre, S. Moussaoui, E. Chouzenoux & J. Idier. *Primal-dual interior-point optimization based on majorization-minimization for edge-preserving spectral unmixing*. In Proc. of IEEE International Conference on Image Processing, Paris, France, octobre 2014.

[Liu 08] J. S. Liu. Monte carlo strategies in scientific computing. Springer Series in Statistics. Springer, 2nd edition, 2008.

[Lukas 93] M. A. Lukas. *Asymptotic optimality of generalized cross-validation for choosing the regularization parameter*. Numerische Mathematik, vol. 66, no. 1, pages 41–66, 1993.

[Martin 12] J. Martin, L.C. Wilcox, C. Burstedde & O. Ghattas. *A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion*. SIAM Journal on Scientific Computing, vol. 34, no. 3, pages A1460–A11487, 2012.

[Mehrotra 92] S. Mehrotra. *On the Implementation of a Primal-Dual Interior Point Method*. SIAM Journal on Optimization, vol. 4, pages 575–601, 1992.

[Miele 69] A. Miele & J. W. Cantrell. *Study on a memory gradient method for the minimization of functions*. Journal of Optimization Theory and Applications, vol. 3, no. 6, pages 459–470, 1969.

[Moussaoui 12] S. Moussaoui, E. Chouzenoux & J. Idier. *Primal-dual interior point optimization for penalized least squares estimation of abundance maps in hyperspectral imaging*. In Proc. of IEEE Workshop on Hyperspectral Image and Signal Processing : Evolution in Remote Sensing (WHISPERS), Shanghai, Chine, juin 2012.

[Murray 94] W. Murray & M. H. Wright. *Line Search Procedures for the Logarithmic Barrier Function*. SIAM Journal on Optimization, vol. 4, no. 2, pages 229–246, 1994.

[Nesterov 05] Y. E. Nesterov. *Smooth minimization of non-smooth functions*. Mathematical Programming, vol. 103, no. 1, pages 127–152, 2005.

[Nocedal 99] J. Nocedal & S. J. Wright. Numerical optimization. Springer-Verlag, New York, NY, USA, 1999.

[Nvidia 12] C. Nvidia. CUDA C programming guide version 4.2. NVIDIA, 2012.

[Orieux 12] F. Orieux, O. Féron & JF Giovannelli. *Sampling High-Dimensional Gaussian Distributions for General Linear Inverse Problems*. IEEE Signal Processing Letters, vol. 19, no. 5, page 251, 2012.

[Ortega 70] J. M. Ortega & W. C. Rheinboldt. Iterative solution of nonlinear equations in several variables. Academic Press, New York, USA, 1970.

[Owens 08] J.D. Owens, M. Houston, D. Luebke, S. Green, J.E. Stone & J.C. Phillips. *GPU computing*. Proceedings of the IEEE, vol. 96, no. 5, pages 879–899, 2008.

[Papandreou 10] G. Papandreou & A. Yuille. *Gaussian sampling by local perturbations*. In Proc. of Neural Information Processing Systems (NIPS), 2010.

[Parker 12] A. Parker & C. Fox. *Sampling Gaussian distributions in Krylov spaces with conjugate gradients*. SIAM Journal on Scientific Computing, vol. 34, no. 3, pages B312–B334, 2012.

[Rivera 03] M. Rivera & J. Marroquin. *Efficient half-quadratic regularization with granularity control*. Image and Vision Computing, vol. 21, no. 4, pages 345–357, avr. 2003.

[Robbins 51] H. Robbins & S. Monro. *A stochastic approximation method*. Annales of Mathematical Statistics, vol. 22, pages 400–407, 1951.

[Robert 01] C.P. Robert. The bayesian choice. Springer-Verlag, 2nd edition, 2001.

[Roberts 96] G. O. Roberts & R. L. Tweedie. *Exponential convergence of Langevin distributions and their discrete approximations*. Bernoulli, vol. 2, no. 4, pages 341–363, 1996.

[Roberts 97] G. O. Roberts, A. Gelman & W. R. Gilks. *Weak convergence and optimal scaling of random walk Metropolis algorithms*. The Annals of Applied Probability, vol. 7, no. 1, pages 110–120, 1997.

[Roberts 98] G. O. Roberts & J. S. Rosenthal. *Optimal scaling of discrete approximations to Langevin's diffusions*. Journal of the Royal Statistical Society B, vol. 60, no. 1, pages 255–268, 1998.

[Rossky 78] P.J. Rossky, J.D. Doll & H.L. Friedman. *Brownian dynamics as smart Monte Carlo simulation*. Journal of Chemical Physics, vol. 69, pages 4628–4633, 1978.

[Rue 01] H. Rue. *Fast sampling of Gaussian Markov random fields*. Journal of the Royal Statistical Society B, vol. 63, no. 2, pages 325–338, 2001.

[Ruggiero 10] V. Ruggiero, T. Serafini, R. Zanella & L. Zanni. *Iterative regularization algorithms for constrained image deblurring on graphics processors*. Journal of Global Optimization, vol. 48, pages 145–157, 2010.

[Sánchez 11] S. Sánchez, A. Paz, G. Martín & A. Plaza. *Parallel unmixing of remotely sensed hyperspectral images on commodity graphics processing units*. Concurrency and Computation : Practice and Experience, vol. 23, no. 13, pages 1538–1557, 2011.

[Scheuer 62] E. M. Scheuer & D.S. Stoller. *On the Generation of Normal Random Vectors*. Technometrics, vol. 4, no. 2, pages 278–281, 1962.

[Segalat 02] P. Segalat. *Méthodes de points intérieurs et de Quasi-Newton*. Thèse de doctorat, Université de Limoges, 2002.

[Shi 05] Z.-J. Shi & J. Shen. *A new super-memory gradient method with curve search rule*. Applied Mathematics and Computations, vol. 170, pages 1–16, 2005.

[Skilling 84] J. Skilling & R. K. Bryan. *Maximum entropy image reconstruction : General algorithm*. Monthly Notices of the Royal Astronomical Society, vol. 211, pages 111–124, 1984.

[Sun 01] J. Sun & J. Zhang. *Global Convergence of Conjugate Gradient Methods without Line Search*. Annals of Operations Research, vol. 103, pages 161–173, mars 2001.

[Tan 10] X. Tan, J. Li & P. Stoica. *Efficient sparse Bayesian learning via Gibbs sampling*. In Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 3634–3637, 2010.

[Tikhonov 63] A. Tikhonov. *Regularization of incorrectly posed problems*. Soviet. Math. Dokl., vol. 4, pages 1624–1627, 1963.

[Tikhonov 77] A. Tikhonov & V. Arsenin. Solutions of ill-posed problems. Winston, Washington, DC, USA, 1977.

[Trench 64] W. F. Trench. *An algorithm for the inversion of finite Toeplitz matrices*. Journal of the Society for Industrial Application of Mathematics, vol. 12, no. 3, pages 515–522, 1964.

[Vacar 11] C. Vacar, J.-F. Giovannelli & Y. Berthoumieu. *Langevin and Hessian with Fisher approximation stochastic sampling for parameter estimation of structured covariance*. In Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2011.

[Van der Vorst 92] H. A. Van der Vorst. *BI-CGSTAB : a fast and smoothly converging variant of BI-CG for the solution of nonsymmetric linear systems*. SIAM Journal on Scientific Computing, vol. 13, pages 631–644, mars 1992.

[Waagepetersen 01] R. Waagepetersen & D. Sorensen. *A Tutorial on Reversible Jump MCMC with a View toward Applications in QTL-mapping*. International Statistical Review, vol. 69, no. 1, pages 49–61, 2001.

[Wilt 13] N. Wilt. The CUDA handbook : A comprehensive guide to GPU programming. Addison-Wesley, 2013.

[Wold 48] H. Wold. Random normal deviates, volume 25 of *Tracts for Computers*. Cambridge University Press, 1948.

[Wolfe 76] M.A. Wolfe & C. Viazminsky. *Supermemory Descent Methods for Unconstrained Minimization*. Journal of Optimization Theory and Applications, vol. 18, no. 4, pages 455–468, 1976.

[Wright 91] M. H. Wright. *Interior methods for constrained optimization*. In Acta Numerica 1992, pages 341–407. Cambridge University Press, 1991.

[Wright 94] M. H. Wright. *Some properties of the Hessian of the logarithmic barrier function*. Mathematical Programming, vol. 67, no. 2, pages 265–295, 1994.

[Wright 98] M. H. Wright. *Ill-conditioning and computational error in interior methods for nonlinear programming*. SIAM Journal on Optimization, vol. 9, no. 1, pages 84–111, 1998.

[Zhang 11] Y. Zhang & C. Sutton. *Quasi-Newton methods for Markov chain Monte Carlo*. In Proc. of Neural Information Processing Systems (NIPS), 2011.

[Zibulevsky 10] M. Zibulevsky & M. Elad. $\ell_2 - \ell_1$ *optimization in signal and image processing*. IEEE Signal Processing Magazine, vol. 27, no. 3, pages 76–88, mai 2010.

Quatrième partie

Annexes

Annexe A

Sélection de publications

Cette annexe contient une sélection de mes publications en lien avec les travaux de recherche décrits dans la deuxième partie de ce manuscrit. Un article présentant le travail de recherche réalisé dans le cadre du contrat industriel avec RENAULT sur l'estimation du rayon effectif et de la résistance au roulement d'une roue d'un véhicule automobile est également fourni.

A.1 Majorize-Minimize strategy for subspace optimization applied to image restoration

E. Chouzenoux, J. Idier et **S. Moussaoui**, *IEEE Trans. on Image Processing*, vol. 20, no.6, pp. 1517-1528, 2011.

Cet article est consacré à la présentation de la stratégie de recherche pas de descente dans un sous-espace de directions en utilisant une technique de majoration-minimisation quadratique décrite dans le chapitre 2 de ce manuscrit.

A Majorize–Minimize Strategy for Subspace Optimization Applied to Image Restoration

Emilie Chouzenoux, Jérôme Idier, *Member, IEEE*, and Saïd Moussaoui

***Abstract*—This paper proposes accelerated subspace optimization methods in the context of image restoration. Subspace optimization methods belong to the class of iterative descent algorithms for unconstrained optimization. At each iteration of such methods, a stepsize vector allowing the best combination of several search directions is computed through a multidimensional search. It is usually obtained by an inner iterative second-order method ruled by a stopping criterion that guarantees the convergence of the outer algorithm. As an alternative, we propose an original multidimensional search strategy based on the majorize–minimize principle. It leads to a closed-form stepsize formula that ensures the convergence of the subspace algorithm whatever the number of inner iterations. The practical efficiency of the proposed scheme is illustrated in the context of edge-preserving image restoration.**

***Index Terms*—Conjugate gradient, image restoration, memory gradient, quadratic majorization, stepsize strategy, subspace optimization.**

I. Introduction

THIS work addresses a wide class of problems where an input image $\boldsymbol{x}^o \in \mathbb{R}^N$ is estimated from degraded data $\boldsymbol{y} \in \mathbb{R}^T$. A typical model of image degradation is

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x}^o + \boldsymbol{\epsilon}$$

where $\boldsymbol{H}$ is a linear operator, described as a $T \times N$ matrix, that models the image degradation process, and $\boldsymbol{\epsilon}$ is an additive noise vector. This simple formalism covers many real situations such as deblurring, denoising, inverse-Radon transform in tomography, and signal interpolation.

Two main strategies emerge in the literature for the restoration of $\boldsymbol{x}^o$ [1]. The first one uses an *analysis-based* approach, solving the following problem [2], [3]:

$$\min_{\boldsymbol{x}\in\mathbb{R}^N} \left(F(\boldsymbol{x}) = \|\boldsymbol{H}\boldsymbol{x} - \boldsymbol{y}\|^2 + \lambda\Psi(\boldsymbol{x}) \right). \tag{1}$$

In Section V, we will consider an image deconvolution problem that calls for the minimization of this criterion form.

The second one employs a *synthesis-based* approach, looking for a decomposition $\boldsymbol{z}$ of the image in some dictionary $\boldsymbol{K} \in \mathbb{R}^{T\times R}$ [4], [5]:

$$\min_{\boldsymbol{z}\in\mathbb{R}^R} \left(F(\boldsymbol{z}) = \|\boldsymbol{H}\boldsymbol{K}\boldsymbol{z} - \boldsymbol{y}\|^2 + \lambda\Psi(\boldsymbol{z}) \right). \tag{2}$$

Manuscript received September 06, 2010; revised December 17, 2010; accepted December 20, 2010. Date of publication December 30, 2010; date of current version May 18, 2011. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Brendt Wohlberg.

The authors are with IRCCyN (CNRS UMR 6597), Ecole Centrale Nantes, 44321 Nantes Cedex 03, France.

Digital Object Identifier 10.1109/TIP.2010.2103083

This method is applied to a set of image reconstruction problems [6] in Section IV.

In both cases, the penalization term Ψ, whose weight is set through the regularization parameter λ, aims at guaranteeing the robustness of the solution to the observation noise and at favorizing its fidelity to *a priori* assumptions [7].

From the mathematical point a view, problems (1) and (2) share a common structure. In this paper, we will focus on the resolution of the first problem (1), but we will also provide numerical results regarding the second one. On the other hand, we restrict ourselves to regularization terms of the form

$$\Psi(\boldsymbol{x}) = \sum_{c=1}^{C} \psi\left(\|\boldsymbol{V}_c\boldsymbol{x} - \boldsymbol{\omega}_c\|\right)$$

where $\boldsymbol{V}_c \in \mathbb{R}^{P\times N}$, $\boldsymbol{\omega}_c \in \mathbb{R}^P$ for $c = 1, \ldots, C$ and $\|.\|$ stands for the Euclidian norm. In the analysis-based approach, $\boldsymbol{V}_c$ is typically a linear operator yielding either the differences between neighboring pixels (e.g., in the Markovian regularization approach), or the local spatial gradient vector (e.g., in the total variation framework), or wavelet decomposition coefficients in some recent works such as [1]. In the synthesis-based approach, $\boldsymbol{V}_c$ usually identifies with the identity matrix.

The strategy used for solving the penalized least squares (PLS) optimization problem (1) strongly depends on the objective function properties (i.e., differentiability and convexity). Moreover, these mathematical properties contribute to the quality of the reconstructed image. In that respect, we particularly focus on differentiable, coercive, edge-preserving functions ψ, e.g., ℓ_p norm with $1 < p < 2$, Huber, hyperbolic, or Geman and McClure functions [8]–[10], since they give rise to locally smooth images [11]–[13]. In contrast, some restoration methods rely on nondifferentiable regularizing functions to introduce priors such as sparsity of the decomposition coefficients [5] and piecewise constant patterns in the images [14]. As emphasized in [6], the nondifferentiable penalization term can be replaced by a smoothed version without altering the reconstruction quality. Moreover, the use of a smoother penalty can reduce the staircase effect that appears in the case of total variation regularization [15].

In the case of large-scale nonlinear optimization problems as encountered in image restoration, direct resolution is impossible. Instead, iterative optimization algorithms are used to solve (1). Starting from an initial guess $\boldsymbol{x}_0$, they generate a sequence of updated estimates $(\boldsymbol{x}_k)$ until sufficient accuracy is obtained. A fundamental update strategy is to produce a decrease of the objective function at each iteration: from the current value $\boldsymbol{x}_k$, $\boldsymbol{x}_{k+1}$ is obtained according to

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k \tag{3}$$

where $\alpha_k > 0$ is the *stepsize* and $\boldsymbol{d}_k$ is a *descent direction* i.e., a vector such that $\boldsymbol{g}_k^T \boldsymbol{d}_k < 0$, where $\boldsymbol{g}_k = \nabla F(\boldsymbol{x}_k)$ denotes the gradient of F at $\boldsymbol{x}_k$. The determination of α_k is called the *line search*. It is usually obtained by partially minimizing the scalar function $f^{(k)}(\alpha) = F(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k)$ until the fulfillment of some sufficient conditions related to the overall algorithm convergence [16].

In the context of the minimization of PLS criteria, the determination of the descent direction $\boldsymbol{d}_k$ is customarily addressed using a half-quadratic (HQ) approach that exploits the PLS structure [11], [12], [17], [18]. A constant stepsize is then used while $\boldsymbol{d}_k$ results from the minimization of a quadratic majorizing approximation of the criterion [13], either resulting from Geman and Reynolds (GR) or from Geman and Yang (GY) constructions [2], [3].

Another effective approach for solving (1) is to consider subspace acceleration [6], [19]. As emphasized in [20], some descent algorithms (3) have a specific subspace feature: they produce search directions spanned in a low-dimension subspace, with examples given here.

- The nonlinear conjugate gradient (NLCG) method [21] uses a search direction in a 2-D space spanned by the opposite gradient and the previous direction.
- The L-BFGS quasi-Newton method [22] generates updates in a subspace of size $2m + 1$, where m is the limited memory parameter.

Subspace acceleration consists in relying on iterations more explicitly aimed at solving the optimization problem within such low dimension subspaces [23]–[27]. The acceleration is obtained by defining $\boldsymbol{x}_{k+1}$ as the approximate minimizer of the criterion over the subspace spanned by a set of M directions

$$\boldsymbol{D}_k = \left[\boldsymbol{d}_k^1, \ldots, \boldsymbol{d}_k^M\right]$$

with $1 \leq M \ll N$. More precisely, the iterates are given by

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{D}_k \boldsymbol{s}_k \tag{4}$$

where $\boldsymbol{s}_k$ is a multidimensional stepsize that aims at partially minimizing

$$f^{(k)}(\boldsymbol{s}) = F(\boldsymbol{x}_k + \boldsymbol{D}_k \boldsymbol{s}). \tag{5}$$

The prototype scheme (4) defines an *iterative subspace optimization* algorithm that can be viewed as an extension of (3) to a search subspace of dimension larger than one. The subspace algorithm has been shown to outperforms standard descent algorithms, such as NLCG and L-BFGS, in terms of computational cost and iteration number before convergence, over a set of PLS minimization problems [6], [19].

The implementation of subspace algorithms requires a strategy to determine the stepsize $\boldsymbol{s}_k$ that guarantees the convergence of the recurrence (4). However, it is difficult to design a practical multidimensional stepsize search algorithm gathering suitable convergence properties and low computational time [26], [28]. Recently, GY and GR HQ approximations have led to an efficient majorization–minimization (MM) line search strategy for the computation of α_k when $\boldsymbol{d}_k$ is the NLCG direction [29] (see also [30] for a general reference on MM algorithms). In this paper, we generalize this strategy to define the multidimensional stepsize $\boldsymbol{s}_k$ in (4). We prove the mathematical convergence of the resulting subspace algorithm under mild conditions on $\boldsymbol{D}_k$. We illustrate its efficiency on four image restoration problems.

The remainder of this paper is organized as follows. Section II gives an overview of existing subspace constructions and multidimensional search procedures. In Section III, we introduce the proposed HQ/MM strategy for the stepsize calculation and we establish general convergence properties for the overall subspace algorithm. Finally, Sections IV and V give some illustrations and a discussion of the algorithm performances by means of a set of experiments in image restoration.

II. Subspace Optimization Methods

The first subspace optimization algorithm is the memory gradient method, proposed in the late 1960s by Miele and Cantrell [23]. It corresponds to

$$\boldsymbol{D}_k = [-\boldsymbol{g}_k, \boldsymbol{d}_{k-1}]$$

and the stepsize $\boldsymbol{s}_k$ results from the exact minimization of $f^{(k)}(\boldsymbol{s})$. When F is quadratic, it is equivalent to the nonlinear conjugate gradient algorithm [31].

More recently, several other subspace algorithms have been proposed. Some of them are briefly reviewed here. We first focus on the subspace construction, and then we describe several existing stepsize strategies.

A. Subspace Construction

Choosing subspaces $\boldsymbol{D}_k$ of dimensions larger than one may allow faster convergence in terms of iteration number. However, it requires a multidimensional stepsize strategy, which can be substantially more complex (and computationaly costly) than the usual line search. Therefore, the choice of the subspace must achieve a tradeoff between the iteration number to reach convergence and the cost per iteration. Let us review some existing iterative subspace optimization algorithms and their associated set of directions. For the sake of compactness, their main features are summarized in Table I. Two families of algorithms are distinguished.

1) Memory Gradient Algorithms: In the first seven algorithms, $\boldsymbol{D}_k$ mainly gathers successive gradient and direction vectors.

The third one, introduced in [32] as supermemory descent (SMD) method, generalizes SMG by replacing the steepest descent direction by any direction $\boldsymbol{p}_k$ nonorthogonal to $\boldsymbol{g}_k$ i.e., $\boldsymbol{g}_k^T \boldsymbol{p}_k \neq 0$. PCD-SESOP and SSF-SESOP algorithms from [6], [19] identify with SMD algorithm, when $\boldsymbol{p}_k$ equals respectively the parallel coordinate descent (PCD) direction and the separable surrogate functional (SSF) direction, both described in [19].

Although the fourth algorithm was introduced in [33]–[35] as a supermemory gradient method, we rather refer to it as a *gradient subspace* (GS) algorithm in order to make the distinction with the supermemory gradient (SMG) algorithm introduced in [24].

The orthogonal subspace (ORTH) algorithm was introduced in [36] with the aim to obtain a first order algorithm with an optimal worst case convergence rate. The ORTH subspace

TABLE I
SET OF DIRECTIONS CORRESPONDING TO THE MAIN EXISTING ITERATIVE SUBSPACE ALGORITHMS. THE WEIGHTS w_i AND THE VECTORS $\boldsymbol{\delta}_i$ ARE DEFINED BY (6) AND (7), RESPECTIVELY. $\boldsymbol{G}_k$ IS DEFINED BY (8), AND $\boldsymbol{d}_k^\ell$ IS THE ℓTH OUTPUT OF A CG ALGORITHM TO SOLVE $\boldsymbol{G}_k(\boldsymbol{d}) = \boldsymbol{0}$

Acronym	Algorithm	Set of directions $\boldsymbol{D}_k$	Subspace size
MG	Memory gradient [23, 31]	$\left[-\boldsymbol{g}_k, \boldsymbol{d}_{k-1}\right]$	2
SMG	Supermemory gradient [24]	$\left[-\boldsymbol{g}_k, \boldsymbol{d}_{k-1}, \ldots, \boldsymbol{d}_{k-m}\right]$	$m+1$
SMD	Supermemory descent [32]	$\left[\boldsymbol{p}_k, \boldsymbol{d}_{k-1}, \ldots, \boldsymbol{d}_{k-m}\right]$	$m+1$
GS	Gradient subspace [33, 34, 37]	$\left[-\boldsymbol{g}_k, -\boldsymbol{g}_{k-1}, \ldots, -\boldsymbol{g}_{k-m}\right]$	$m+1$
ORTH	Orthogonal subspace [36]	$\left[-\boldsymbol{g}_k, \boldsymbol{x}_k - \boldsymbol{x}_0, \sum_{i=0}^k w_i \boldsymbol{g}_i\right]$	3
SESOP	Sequential Subspace Optimization [26]	$\left[-\boldsymbol{g}_k, \boldsymbol{x}_k - \boldsymbol{x}_0, \sum_{i=0}^k w_i \boldsymbol{g}_i, \boldsymbol{d}_{k-1}, \ldots, \boldsymbol{d}_{k-m}\right]$	$m+3$
QNS	Quasi-Newton subspace [20, 25, 38]	$\left[-\boldsymbol{g}_k, \boldsymbol{\delta}_{k-1}, \ldots, \boldsymbol{\delta}_{k-m}, \boldsymbol{d}_{k-1}, \ldots, \boldsymbol{d}_{k-m}\right]$	$2m+1$
SESOP-TN	Truncated Newton subspace [27]	$\left[\boldsymbol{d}_k^\ell, \boldsymbol{G}_k(\boldsymbol{d}_k^\ell), \boldsymbol{d}_k^\ell - \boldsymbol{d}_k^{\ell-1}, \boldsymbol{d}_{k-1}, \ldots, \boldsymbol{d}_{k-m}\right]$	$m+3$

corresponds to the opposite gradient augmented with the two so-called Nemirovski directions, $\boldsymbol{x}_k - \boldsymbol{x}_0$ and $\sum_{i=0}^k w_i \boldsymbol{g}_i$, where w_i are prespecified, recursively defined weights

$$w_i = \begin{cases} 1, & \text{if } i = 0 \\ \frac{1}{2} + \sqrt{\frac{1}{4} + w_{i-1}^2}, & \text{otherwise.} \end{cases} \tag{6}$$

In [26], the Nemirovski subspace is augmented with previous directions, leading to the SESOP algorithm whose efficiency over ORTH is illustrated on a set of image reconstruction problems. Moreover, experimental tests showed that the use of Nemirovski directions in SESOP does not improve practical convergence speed. Therefore, in their recent paper [6], Zibulevsky *et al.* do not use these additionnal vectors so that their modified SESOP algorithm actually reduces to the SMG algorithm from [24].

2) Newton-Type Subspace Algorithms: The last two algorithms introduce additional directions of the Newton type.

In the quasi-Newton subspace (QNS) algorithm proposed in [25], $\boldsymbol{D}_k$ is augmented with

$$\boldsymbol{\delta}_{k-i} = \boldsymbol{g}_{k-i+1} - \boldsymbol{g}_{k-i}, \qquad i = 1, \ldots, m. \tag{7}$$

This proposal is reminiscent from the L-BFGS algorithm [22], since the latter produces directions in the space spanned by the resulting set $\boldsymbol{D}_k$.

SESOP-TN has been proposed in [27] to solve the problem of sensitivity to an early break of conjugate gradient (CG) iterations in the truncated Newton (TN) algorithm. Let $\boldsymbol{d}_k^\ell$ denote the current value of $\boldsymbol{d}$ after ℓ iterations of CG to solve the Gauss-Newton system $\boldsymbol{G}_k(\boldsymbol{d}) = \boldsymbol{0}$, where

$$\boldsymbol{G}_k(\boldsymbol{d}) = \nabla^2 F(\boldsymbol{x}_k)\boldsymbol{d} + \boldsymbol{g}_k. \tag{8}$$

In the standard TN algorithm, $\boldsymbol{d}_k^\ell$ defines the search direction [39]. In SESOP-TN, it is only the first component of $\boldsymbol{D}_k$, while the second and third components of $\boldsymbol{D}_k$ also result from the CG iterations.

Finally, to accelerate optimization algorithms, a common practice is to use a preconditioning matrix. The principle is to introduce a linear transform on the original variables, so that the new variables have a Hessian matrix with more clustered eigenvalues. Preconditioned versions of subspace algorithms are easily defined by using $\boldsymbol{P}_k \boldsymbol{g}_k$ instead of $\boldsymbol{g}_k$ in the previous direction sets [26].

B. Stepsize Strategies

The aim of the multidimensional stepsize search is to determine $\boldsymbol{s}_k$ that ensures a sufficient decrease of function $f^{(k)}$ defined by (5) in order to guarantee the convergence of recurrence (4). In the scalar case, typical line search procedures generate a series of stepsize values until the fulfillment of sufficient convergence conditions such as Armijo *et al.* [40]. An extension of these conditions to the multidimensional case can easily be obtained (e.g., the multidimensional Goldstein rule in [28]). However, it is difficult to design practical multidimensional stepsize search algorithms allowing to check these conditions [28].

Instead, in several subspace algorithms, the stepsize results from an iterative descent algorithm applied to function $f^{(k)}$, stopped before convergence. In SESOP and SESOP-TN, the minimization is performed by a Newton method. However, unless the minimizer is found exactly, the resulting subspace algorithms are not proved to converge. In the QNS and GS algorithms, the stepsize results from a trust region recurrence on $f^{(k)}$. It is shown to ensure the convergence of the iterates under mild conditions on $\boldsymbol{D}_k$ [25], [34], [35]. However, except when the quadratic approximation of the criterion in the trust region is separable [34], the trust region search requires to solve a nontrivial constrained quadratic programming problem at each inner iteration.

In the particular case of modern SMG algorithms [41]–[44], $\boldsymbol{s}_k$ is computed in two steps. First, a descent direction is constructed by combining the vectors $\boldsymbol{d}_k^i$ with some predefined weights. Then, a scalar stepsize is calculated through an iterative line search. This strategy leads to the recurrence

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \left(-\beta_k^0 \boldsymbol{g}_k + \sum_{i=1}^m \beta_k^i \boldsymbol{d}_{k-i} \right).$$

Different expressions for the weights β_k^i have been proposed. To our knowledge, their extension to the preconditioned version of SMG or to other subspaces is an open issue. Moreover, since the computation of (α_k, β_k^i) does not aim at minimizing $f^{(k)}$ in

the SMG subspace, the resulting schemes are not true subspace algorithms.

In Section III, we propose an original strategy to define the multidimensional stepsize $\boldsymbol{s}_k$ in (4). The proposed stepsize search is proved to ensure the convergence of the whole algorithm, under low assumptions on the subspace, and to require low computationnal cost.

III. Proposed Multidimensional Stepsize Strategy

A. GR and GY Majorizing Approximations

Let us first introduce Geman and Yang [3] and Geman and Reynolds [2] matrices $\boldsymbol{A}_{\mathrm{GY}}$ and $\boldsymbol{A}_{\mathrm{GR}}$, which play a central role in the multidimensional stepsize strategy proposed in this paper:

$$\boldsymbol{A}^a_{\mathrm{GY}} = 2\boldsymbol{H}^T\boldsymbol{H} + \frac{\lambda}{a}\boldsymbol{V}^T\boldsymbol{V} \tag{9}$$

$$\boldsymbol{A}_{\mathrm{GR}}(\boldsymbol{x}) = 2\boldsymbol{H}^T\boldsymbol{H} + \lambda\boldsymbol{V}^T\mathrm{Diag}\left\{\boldsymbol{b}(\boldsymbol{x})\right\}\boldsymbol{V} \tag{10}$$

where $\boldsymbol{V}^T = [\boldsymbol{V}_1^T|\ldots|\boldsymbol{V}_C^T]$, $a > 0$ is a free parameter, and $\boldsymbol{b}(\boldsymbol{x})$ is a $CP \times 1$ vector with entries

$$b_{cp}(\boldsymbol{x}) = \frac{\dot{\psi}\left(\|\boldsymbol{V}_c\boldsymbol{x} - \boldsymbol{\omega}_c\|\right)}{\|\boldsymbol{V}_c\boldsymbol{x} - \boldsymbol{\omega}_c\|}.$$

Both GY and GR matrices allow the construction of majorizing approximation for F. More precisely, let us introduce the following second order approximation of F in the neighborhood of $\boldsymbol{x}_k$

$$Q(\boldsymbol{x},\boldsymbol{x}_k) = F(\boldsymbol{x}_k) + \nabla F(\boldsymbol{x}_k)^T(\boldsymbol{x}-\boldsymbol{x}_k) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_k)^T\boldsymbol{A}(\boldsymbol{x}_k)(\boldsymbol{x}-\boldsymbol{x}_k). \tag{11}$$

Let us also introduce the following assumptions on the function ψ:

(H1) ψ is $\mathcal{C}^1$ and coercive.
$\dot{\psi}$ is L-Lipschitz.
(H2) ψ is $\mathcal{C}^1$, even and coercive.
$\psi(\sqrt{.})$ is concave on $\mathbb{R}^+$.
$0 < \dot{\psi}(t)/t < \infty, \forall t \in \mathbb{R}$.

Then, the following lemma holds.

Lemma 1 [13]: Let F defined by (1) and $\boldsymbol{x}_k \in \mathbb{R}^N$. If Assumption H1 holds and $\boldsymbol{A} = \boldsymbol{A}^a_{\mathrm{GY}}$ with $a \in (0, 1/L)$ (resp. Assumption H2 holds and $\boldsymbol{A} = \boldsymbol{A}_{\mathrm{GR}}$), then, for all $\boldsymbol{x}$, (11) is a *tangent majorant* for F at $\boldsymbol{x}_k$ i.e., for all $\boldsymbol{x} \in \mathbb{R}^n$,

$$\begin{cases} Q(\boldsymbol{x},\boldsymbol{x}_k) \geq F(\boldsymbol{x}), \\ Q(\boldsymbol{x}_k,\boldsymbol{x}_k) = F(\boldsymbol{x}_k). \end{cases} \tag{12}$$

The majorizing property (12) ensures that the MM recurrence

$$\boldsymbol{x}_{k+1} = \arg\min_{\boldsymbol{x}} Q(\boldsymbol{x},\boldsymbol{x}_k) \tag{13}$$

produces a nonincreasing sequence $\left(F(\boldsymbol{x}_k)\right)$ that converges to a stationary point of F [30], [45]. Half-quadratic algorithms [2], [3] are based on the relaxed form

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \theta(\hat{\boldsymbol{x}}_{k+1} - \boldsymbol{x}_k) \tag{14}$$

where $\hat{\boldsymbol{x}}_{k+1}$ is obtained by (13). The convergence properties of recurrence (14) are analyzed in [12], [13], [46].

B. Majorize–Minimize Line Search

In [29], $\boldsymbol{x}_{k+1}$ is defined as (3) where $\boldsymbol{d}_k$ is the NLCG direction and the stepsize value α_k results from $J \geq 1$ successive minimizations of quadratic tangent majorant functions for the scalar function $f^{(k)}(\alpha) = F(\boldsymbol{x}_k + \alpha\boldsymbol{d}_k)$, expressed as

$$q^{(k)}\left(\alpha,\alpha_k^j\right) = f^{(k)}\left(\alpha_k^j\right) + \left(\alpha-\alpha_k^j\right)\dot{f}^{(k)}\left(\alpha_k^j\right) + \frac{1}{2}b_k^j\left(\alpha-\alpha_k^j\right)^2$$

at α_k^j. The scalar parameter b_k^j is defined as

$$b_k^j = \boldsymbol{d}_k^T\boldsymbol{A}\left(\boldsymbol{x}_k + \alpha_k^j\boldsymbol{d}_k\right)\boldsymbol{d}_k$$

where $\boldsymbol{A}(.)$ is either the GY or the GR matrix, respectively defined by (9) and (10). The stepsize values are produced by the relaxed MM recurrence

$$\begin{cases} \alpha_k^0 = 0 \\ \alpha_k^{j+1} = \alpha_k^j - \theta\dot{f}\left(\alpha_k^j\right)/b_k^j, \quad j = 0,\ldots,J-1 \end{cases} \tag{15}$$

and the stepsize α_k corresponds to the last value α_k^J. The distinctive feature of the MM line search is to yield the convergence of standard descent algorithms without any stopping condition whatever the number of MM subiterations J and relaxation parameter $\theta \in (0,2)$ [29]. Here, we propose to extend this strategy to the determination of the multidimensional stepsize $\boldsymbol{s}_k$, and we prove the convergence of the resulting family of subspace algorithms.

C. MM Multidimensional Search

Let us define the $M \times M$ symmetric positive definite (SPD) matrix

$$B_k^j = \boldsymbol{D}_k^T\boldsymbol{A}_k^j\boldsymbol{D}_k$$

with $\boldsymbol{A}_k^j \triangleq \boldsymbol{A}(\boldsymbol{x}_k + \boldsymbol{D}_k\boldsymbol{s}_k^j)$ and $\boldsymbol{A}$ is either the GY matrix or the GR matrix. According to Lemma 1,

$$q^{(k)}\left(\boldsymbol{s},\boldsymbol{s}_k^j\right) = f^{(k)}\left(\boldsymbol{s}_k^j\right) + \nabla f^{(k)}\left(\boldsymbol{s}_k^j\right)^T\left(\boldsymbol{s}-\boldsymbol{s}_k^j\right) + \frac{1}{2}\left(\boldsymbol{s}-\boldsymbol{s}_k^j\right)^T\boldsymbol{B}_k^j\left(\boldsymbol{s}-\boldsymbol{s}_k^j\right) \tag{16}$$

is quadratic tangent majorant for $f^{(k)}(\boldsymbol{s})$ at $\boldsymbol{s}_k^j$. Then, let us define the MM multidimensional stepsize by $\boldsymbol{s}_k = \boldsymbol{s}_k^J$, with

$$\begin{cases} \boldsymbol{s}_k^0 = \boldsymbol{0}, \\ \hat{\boldsymbol{s}}_k^{j+1} = \arg\min_{\boldsymbol{s}} q^{(k)}\left(\boldsymbol{s},\boldsymbol{s}_k^j\right), \quad j = 0,\ldots,J-1. \\ \boldsymbol{s}_k^{j+1} = \boldsymbol{s}_k^j + \theta\left(\hat{\boldsymbol{s}}_k^{j+1} - \boldsymbol{s}_k^j\right) \end{cases} \tag{17}$$

Given (16), we obtain an explicit stepsize formula

$$\boldsymbol{s}_k^{j+1} = \boldsymbol{s}_k^j - \theta\left(\boldsymbol{B}_k^j\right)^{-1}\nabla f^{(k)}\left(\boldsymbol{s}_k^j\right).$$

Moreover, according to [13], the update rule (17) produces monotonically decreasing values $(f^{(k)}(\boldsymbol{s}_k^j))$ if $\theta \in (0,2)$. Let us emphasize that this stepsize procedure identifies with the HQ/MM iteration (14) when $\mathrm{span}(\boldsymbol{D}_k) = \mathbb{R}^N$, and to the HQ/MM line search (15) when $\boldsymbol{D}_k = \boldsymbol{d}_k$.

D. Convergence Analysis

Here, we establish the convergence of the iterative subspace algorithm (4) when $\boldsymbol{s}_k$ is chosen according to the MM strategy (17).

We introduce the following assumption, which is a necessary condition to ensure that the penalization term $\Psi(\boldsymbol{x})$ regularizes the problem of estimating $\boldsymbol{x}$ from $\boldsymbol{y}$ in a proper way

(H3) $\boldsymbol{H}$ and $\boldsymbol{V}$ are such that

$$\ker(\boldsymbol{H}^T\boldsymbol{H}) \cap \ker(\boldsymbol{V}^T\boldsymbol{V}) = \{\boldsymbol{0}\}.$$

Lemma 2 [13]: Let F be defined by (1), where $\boldsymbol{H}$ and $\boldsymbol{V}$ satisfy Assumption H3. If Assumption H1 or H2 holds, F is continuously differentiable and bounded below. Moreover, if for all k, j, $\boldsymbol{A} = \boldsymbol{A}_{\mathrm{GY}}^a$ with $0 < a < 1/L$ (resp., $\boldsymbol{A} = \boldsymbol{A}_{\mathrm{GR}}$), then $(\boldsymbol{A}_k^j)$ has a *positive bounded spectrum*, i.e., there exists $\nu_1 \in \mathbb{R}$ such that

$$0 < \boldsymbol{v}^T\boldsymbol{A}_k^j\boldsymbol{v} \leq \nu_1\|\boldsymbol{v}\|^2, \ \forall k, j \in \mathbb{N}, \forall \boldsymbol{v} \in \mathbb{R}^N.$$

Let us also assume that the set of directions $\boldsymbol{D}_k$ fulfills the following condition.

(H4) For all $k \geq 0$, the matrix of directions $\boldsymbol{D}_k$ is of size $N \times M$ with $1 \leq M \leq N$ and the first subspace direction $\boldsymbol{d}_k^1$ fulfills

$$\boldsymbol{g}_k^T\boldsymbol{d}_k^1 \leq -\gamma_0\|\boldsymbol{g}_k\|^2 \tag{18}$$

$$\left\|\boldsymbol{d}_k^1\right\| \leq \gamma_1\|\boldsymbol{g}_k\| \tag{19}$$

with $\gamma_0, \gamma_1 > 0$.

Then, the convergence of the MM subspace scheme holds according to the following theorem.

Theorem 1: Let F defined by (1), where $\boldsymbol{H}$ and $\boldsymbol{V}$ satisfy Assumption H3. Let $\boldsymbol{x}_k$ defined by (4)–(17) where $\boldsymbol{D}_k$ satisfies Assumption H4, $J \geq 1$, $\theta \in (0,2)$ and $\boldsymbol{B}_k^j = \boldsymbol{D}_k^T\boldsymbol{A}_{\mathrm{GY}}^a\boldsymbol{D}_k$ with $0 < a < 1/L$ (resp., $\boldsymbol{B}_k^j = \boldsymbol{D}_k^T\boldsymbol{A}_{\mathrm{GR}}(\boldsymbol{x}_k + \boldsymbol{D}_k\boldsymbol{s}_k^j)\boldsymbol{D}_k$). If Assumption H1 (resp., Assumption H2) holds, then

$$F(\boldsymbol{x}_{k+1}) \leq F(\boldsymbol{x}_k). \tag{20}$$

Moreover, we have convergence in the following sense:

$$\lim_{k\to\infty} \|\boldsymbol{g}_k\| = 0.$$

Proof: See Appendix A. ∎

Remark 1: Assumption H4 is fulfilled by a large family of descent directions. In particular, the following results hold.

- Let $(\boldsymbol{P}_k)$ be a series of SPD matrices with eigenvalues that are bounded below and above, respectively by γ_1 and $\gamma_0 > 0$. Then, according to [16Sec. 1.2], Assumption H4 holds if $\boldsymbol{d}_k^1 = -\boldsymbol{P}_k\boldsymbol{g}_k$.
- According to [47], Assumption H4 also holds if $\boldsymbol{d}_k^1$ results from any fixed positive number of CG iterations on the linear system $\boldsymbol{M}_k\boldsymbol{d} = -\boldsymbol{g}_k$, provided that $(\boldsymbol{M}_k)$ is a matrix series with a positive bounded spectrum.
- Finally, Lemma 3 in Appendix B ensures that Assumption H4 holds if $\boldsymbol{d}_k^1$ is the PCD direction, provided that F is strongly convex and has a Lipschitz gradient.

Remark 2: For a preconditioned NLCG algorithm with a variable preconditioner $\boldsymbol{P}_k$, the generated iterates belong to the subspace spanned by $-\boldsymbol{P}_k\boldsymbol{g}_k$ and $\boldsymbol{d}_{k-1}$. Whereas the convergence of the PNLCG scheme with a variable preconditioner is still an open problem [21], [48], the preconditioned MG algorithm using $\boldsymbol{D}_k = [-\boldsymbol{P}_k\boldsymbol{g}_k, \boldsymbol{d}_{k-1}]$ and the proposed MM stepsize is guaranteed to converge for bounded SPD matrices $\boldsymbol{P}_k$, according to Theorem 1.

E. Implementation Issues

In the proposed MM multidimensional search, the main computational burden originates from the need to multiply the spanning directions with linear operators $\boldsymbol{H}$ and $\boldsymbol{V}$, in order to compute $\nabla f^{(k)}(\boldsymbol{s}_k^j)$ and $\boldsymbol{B}_k^j$. When the problem is large scale, these products become expensive and may counterbalance the efficiency obtained when using a subset of larger dimension. In this section, we give a strategy to reduce the computational cost of the product $\boldsymbol{M}_k \triangleq \boldsymbol{\Delta}\boldsymbol{D}_k$ when $\boldsymbol{\Delta} = \boldsymbol{H}$ or $\boldsymbol{V}$. This generalizes the strategy proposed in [26, Sec. 3] for the computation of $\nabla f^{(k)}(\boldsymbol{s})$ and $\nabla^2 f^{(k)}(\boldsymbol{s})$ during the Newton search of the SESOP algorithm.

For all subspace algorithms, the set $\boldsymbol{D}_k$ can be expressed as the sum of a new matrix and a weighted version of the previous set

$$\boldsymbol{D}_k = [\boldsymbol{N}_k|\boldsymbol{0}] + [\boldsymbol{0}|\boldsymbol{D}_{k-1}\boldsymbol{W}_k]. \tag{21}$$

The obtained expressions for $\boldsymbol{N}_k$ and $\boldsymbol{W}_k$ are given in Table II. According to (21), $\boldsymbol{M}_k$ can be obtained by the recurrence

$$\boldsymbol{M}_k = [\boldsymbol{\Delta}\boldsymbol{N}_k|\boldsymbol{0}] + [\boldsymbol{0}|\boldsymbol{M}_{k-1}\boldsymbol{W}_k].$$

Assuming that $\boldsymbol{M}_k$ is stored at each iteration, the computationnal burden reduces to the product $\boldsymbol{\Delta}\boldsymbol{N}_k$. This strategy is efficient as far as $\boldsymbol{N}_k$ has a small number of columns. Moreover, the cost of the latter product does not depend on the subspace dimension, by contrast with the direct computation of $\boldsymbol{M}_k$.

IV. Application to the Set of Image Processing Problems From [6]

Here, we consider three image processing problems, namely image deblurring, tomography, and compressive sensing, generated with Zibulevsky's code.[1] For all problems, the synthesis-based approach is used for the reconstruction. The image is assumed to be well described as $\boldsymbol{x}^o = \boldsymbol{K}\boldsymbol{z}^o$ with a known dictionary $\boldsymbol{K}$ and a sparse vector $\boldsymbol{z}^o$. The restored image is then defined as $\boldsymbol{x}^* = \boldsymbol{K}\boldsymbol{z}^*$ where $\boldsymbol{z}^*$ minimizes the PLS criterion

$$F(\boldsymbol{z}) = \|\boldsymbol{H}\boldsymbol{K}\boldsymbol{z} - \boldsymbol{y}\|^2 + \lambda\sum_{i=1}^{N}\psi(z_i)$$

[1][Online]. Available: http://iew3.technion.ac.il/mcib

TABLE II
RECURSIVE MEMORY FEATURE AND DECOMPOSITION (21) OF SEVERAL ITERATIVE SUBSPACE ALGORITHMS. HERE, $\boldsymbol{D}(i:j)$ DENOTES THE SUBMATRIX OF $\boldsymbol{D}$ MADE OF COLUMNS i TO j, AND $\boldsymbol{I}_{i:j}$ DENOTES THE MATRIX SUCH THAT $\boldsymbol{D}, \boldsymbol{I}_{i:j} = \boldsymbol{D}(i:j)$

Acronym	Recursive form of $\boldsymbol{D}_k$	$\boldsymbol{N}_k$	$\boldsymbol{W}_k$
MG	$[-\boldsymbol{g}_k, \boldsymbol{D}_{k-1}\boldsymbol{s}_{k-1}]$	$-\boldsymbol{g}_k$	$\boldsymbol{s}_{k-1}$
SMG	$[-\boldsymbol{g}_k, \boldsymbol{D}_{k-1}\boldsymbol{s}_{k-1}, \boldsymbol{D}_{k-1}(2:m)]$	$-\boldsymbol{g}_k$	$[\boldsymbol{s}_{k-1}, \boldsymbol{I}_{2:m}]$
GS	$[-\boldsymbol{g}_k, \boldsymbol{D}_{k-1}(1:m)]$	$-\boldsymbol{g}_k$	$\boldsymbol{I}_{1:m}$
ORTH	$[-\boldsymbol{g}_k, \boldsymbol{x}_k - \boldsymbol{x}_0, \omega_k \boldsymbol{g}_k + \boldsymbol{D}_{k-1}(3)]$	$[-\boldsymbol{g}_k, \boldsymbol{x}_k - \boldsymbol{x}_0, \omega_k \boldsymbol{g}_k]$	$\boldsymbol{I}_3$
QNS	$[-\boldsymbol{g}_k, \boldsymbol{g}_k + \boldsymbol{D}_{k-1}(1), \boldsymbol{D}_{k-1}(2:m), \boldsymbol{D}_{k-1}\boldsymbol{s}_{k-1}, \boldsymbol{D}_{k-1}(m+2:2m)]$	$[-\boldsymbol{g}_k, \boldsymbol{g}_k]$	$[\boldsymbol{I}_1, \boldsymbol{I}_{2:m}, \boldsymbol{s}_{k-1}, \boldsymbol{I}_{m+2:2m}]$
SESOP-TN	$[\boldsymbol{d}_k^\ell, \boldsymbol{G}_k(\boldsymbol{d}_k^\ell), \boldsymbol{d}_k^\ell - \boldsymbol{d}_k^{\ell-1}, \boldsymbol{D}_{k-1}(4:m+2)]$	$[\boldsymbol{d}_k^\ell, \boldsymbol{G}_k(\boldsymbol{d}_k^\ell), \boldsymbol{d}_k^\ell - \boldsymbol{d}_k^{\ell-1}]$	$\boldsymbol{I}_{4:m+2}$

where ψ is the logarithmic smooth version of the ℓ_1 norm

$$\psi(u) = |u| - \delta \log(1 + |u|/\delta)$$

that aims at sparsifying the solution.

In [6], several subspace algorithms are compared in order to minimize F. In all cases, the multidimensional stepsize results from a fixed number of Newton iterations. The aim of this section is to test the convergence speed of the algorithms when the Newton procedure is replaced by the proposed MM stepsize strategy.

A. Subspace Algorithm Settings

SESOP [26] and PCD-SESOP [19] direction sets are considered here. The latter uses SMD vectors with $\boldsymbol{p}_k$ defined as the PCD direction

$$p_{i,k} = \arg\min_{\alpha} F(\boldsymbol{x}_k + \alpha \boldsymbol{e}_i), \qquad i = 1, \ldots, N \tag{22}$$

where $\boldsymbol{e}_i$ stands for the ith elementary unit vector. Following [6], the memory parameter is tuned to $m = 7$ (i.e., $M = 8$). Moreover, the Nemirovski directions are discarded, so that SESOP identifies with the SMG subspace.

Let us define SESOP-MM and PCD-SESOP-MM algorithms by associating SESOP and PCD-SESOP subspaces with the multidimensional MM stepsize strategy (17). The latter is fully specified by the curvature matrix $\boldsymbol{A}_k^j$, the number of MM sub-iterations J and the relaxation parameter θ. For all k, j, we define $\boldsymbol{A}_k^j = \boldsymbol{A}_{\mathrm{GR}}(\boldsymbol{x}_k + \boldsymbol{D}_k \boldsymbol{s}_k^j)$ where $\boldsymbol{A}_{\mathrm{GR}}(.)$ is given by (10), and $J = \theta = 1$. Function ψ is strictly convex and fulfills both Assumptions H1 and H2. Therefore, Lemma 1 applies. Matrix $\boldsymbol{V}$ identifies with the identity matrix, so Assumption H3 holds and Lemma 2 applies. Moreover, according to Lemma 3, Assumption H4 holds and Theorem 1 ensures the convergence of SESOP-MM and PCD-SESOP-MM schemes.

MM versions of SESOP and PCD-SESOP are compared to the original algorithms from [6], where the inner minimization uses Newton iterations with backtracking line search, until the tight stopping criterion

$$\left\| \nabla f^{(k)}(\boldsymbol{s}) \right\| < 10^{-10}$$

is met, or seven Newton updates are achieved.

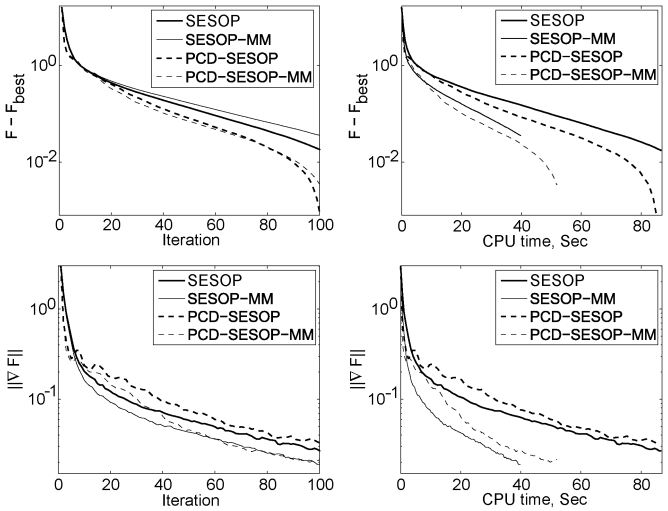


Fig. 1. Deblurring problem taken from [6] (128 × 128 pixels). The objective function and the gradient norm value as a function of iteration number (left) and CPU time in seconds (right) for the four tested algorithms.

For each test problem, the results were plotted as functions of either iteration numbers, or of computational times in seconds, on an Intel Pentium 4 PC (3.2-GHz CPU and 3-GB RAM).

B. Results and Discussion

1) Choice Between Subspace Strategies: According to Figs. 1–3, the PCD-SESOP subspace leads to the best results in terms of objective function decrease per iteration, while the SESOP subspace leads to the largest decrease of the gradient norm, independently from the stepsize strategy. Moreover, when considering the computational time, it appears that SESOP and PCD-SESOP algorithms have quite similar performances.

2) Choice Between Stepsize Strategies: The impact of the stepsize strategy is the central issue in this paper. According to a visual comparison between thin and thick plots in Figs. 1–3, the MM stepsize strategy always leads to significantly faster algorithms compared with the original versions based on Newton search, mainly because of a reduced computational time per iteration.

Moreover, let us emphasize that the theoretical convergence of SESOP-MM and PCD-SESOP-MM is ensured according to Theorem 1. In contrast, unless the Newton search reaches

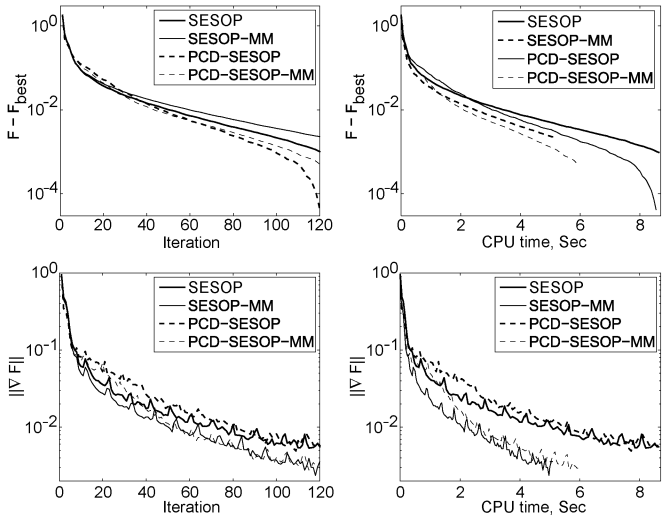


Fig. 2. Tomography problem taken from [6] (32 × 32 pixels). The objective function and the gradient norm value as a function of iteration number (left) and CPU time in seconds (right) for the four tested algorithms.

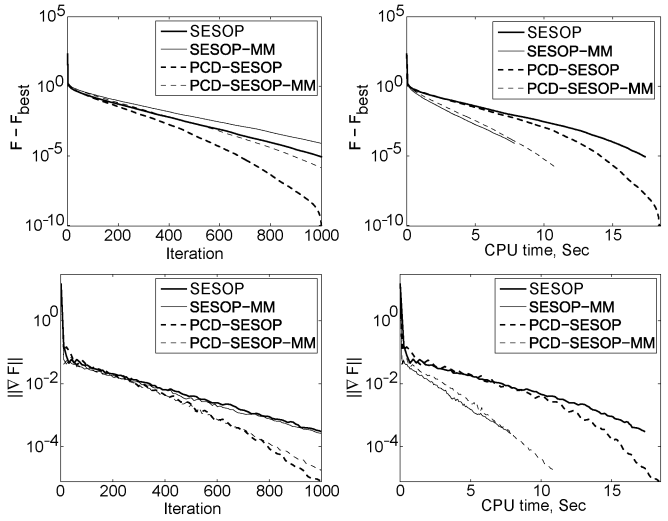


Fig. 3. Compressed sensing problem taken from [6] (64 × 64 pixels). The objective function and the gradient norm value as a function of iteration number (left) and CPU time in seconds (right) for the four tested algorithms.

the exact minimizer of $f^{(k)}(\boldsymbol{s})$, the convergence of SESOP and PCD-SESOP is not guaranteed theoretically.

V. Application to Edge-Preserving Image Restoration

The problem considered here is the restoration of the well-known images `boat`, `lena`, and `peppers` of size $N = 512 \times 512$. These images are firstly convolved with a Gaussian point spread function of standard deviation 2.24 and of size 17×17. Second, a white Gaussian noise is added with a variance adjusted to get a signal-to-noise ratio (SNR) of 40 dB. The following analysis-based PLS criterion is considered:

$$F(\boldsymbol{x}) = \|\boldsymbol{H}\boldsymbol{x} - \boldsymbol{y}\|^2 + \lambda \sum_c \sqrt{\delta^2 + [\boldsymbol{V}\boldsymbol{x}]_c^2}$$

TABLE III
Values of Hyperparameters λ, δ and Reconstruction Quality in Terms of PSNR and RMSE

	`boat`	`lena`	`peppers`
λ	0.2	0.2	0.2
δ	13	13	8
PSNR	28.4	30.8	31.6
RMSE	$5 \cdot 10^{-3}$	$3.3 \cdot 10^{-3}$	$2 \cdot 10^{-3}$

where $\boldsymbol{V}$ is the first-order difference matrix. This criterion depends on the parameters λ and δ. They are assessed to maximize the peak signal to noise ratio (PSNR) between each image $\boldsymbol{x}^o$ and its reconstruction version $\boldsymbol{x}$. Table III gives the resulting values of PSNR and relative mean square error (RMSE), defined by

$$\mathrm{PSNR}(\boldsymbol{x}, \boldsymbol{x}^o) = 20 \log_{10} \left(\frac{\max_i(x_i)}{\sqrt{1/N \sum_i (x_i - x_i^o)^2}} \right)$$

and

$$\mathrm{RMSE}(\boldsymbol{x}, \boldsymbol{x}^o) = \frac{\|\boldsymbol{x} - \boldsymbol{x}^o\|^2}{\|\boldsymbol{x}\|^2}.$$

The purpose of this section is to test the convergence speed of the multidimensional MM stepsize strategy (17) for different subspace constructions. Furthermore, these performances are compared with standard iterative descent algorithms associated with the MM line search described in Section III-B.

A. Subspace Algorithm Settings

The MM stepsize search is used with the Geman and Reynolds HQ matrix and $\theta = 1$. Since the hyperbolic function ψ is a strictly convex function that fulfills both Assumptions H1 and H2, Lemma 1 applies. Furthermore, Assumption H3 holds [29] so Lemma 2 applies.

Our study deals with the preconditioned form of the following direction sets: SMG, GS, QNS, and SESOP-TN. The preconditioner $\bar{\boldsymbol{P}}$ is a SPD matrix based on the 2-D Cosine Transform. Thus, Assumption H4 holds and Theorem 1 ensures the convergence of the proposed scheme whatever the number of MM subiterations $J \geq 1$. Moreover, the implementation strategy described in Section III-E will be used.

For each subspace, we first consider the reconstruction of `peppers`, illustrated in Fig. 4, allowing us to discuss the tuning of the memory parameter m, related to the size of the subspace M as described in Table I, and the performances of the MM search. The latter is again compared with the Newton search from [6].

Then, we compare the subspace algorithms with iterative descent methods in association with the MM scalar line search.

The global stopping rule $\|\boldsymbol{g}_k\|/\sqrt{N} < 10^{-4}$ is considered. For this setting, no significant differences between algorithms have been observed in terms of reconstruction quality. For each tested scheme, the performance results are displayed under the

(a) (b)

Fig. 4. (a) Noisy, blurred peppers image, 40 dB. (b) Restored image.

TABLE IV
RECONSTRUCTION OF peppers: ITERATION NUMBER K/TIME T (s) BEFORE CONVERGENCE FOR MM AND NEWTON STRATEGIES FOR THE MULTIDIMENSIONAL SEARCH IN SMG ALGORITHM

SMG(m)		1	2	5	10
Newton		76/578	75/630	76/701	74/886
MM (J)	1	**67/119**	68/125	67/140	67/163
	2	66/141	66/147	67/172	67/206
	5	74/211	72/225	71/255	72/323
	10	76/297	74/319	73/394	74/508

TABLE V
RECONSTRUCTION OF peppers: ITERATION NUMBER K/TIME T (s) BEFORE CONVERGENCE FOR THE MULTIDIMENSIONAL SEARCH IN GS ALGORITHM

GS(m)		1	5	10	15
Newton		458/3110	150/1304	96/1050	81/1044
MM (J)	1	315/534	128/258	76/180	**67/175**
	2	316/656	134/342	86/257	70/232
	5	317/856	137/481	91/400	78/386
	10	317/1200	137/709	92/619	78/598

form K/T where K is the number of global iterations and T is the global minimization time in seconds.

B. Gradient and Memory Gradient Subspaces

The aim of this section is to analyze the performances of SMG and GS algorithms.

1) Influence of Tuning Parameters: According to Tables IV, V, the algorithms perform better when the stepsize is obtained with the MM search. Furthermore, it appears that $J = 1$ leads to the best results in terms of computation time which indicates that the best strategy corresponds to a rough minimization of $f^{(k)}(\boldsymbol{s})$. Such a conclusion meets that of [29]. In contrast, the MM strategy with high values of J leads to poor performances in term of iteration number K, comparable with those obtained when using Newton search.

The effect of the memory size m differs according to the subspace construction. For the SMG algorithm, an increase of the size of the memory m does not accelerate the convergence. On the contrary, it appears that the number of iterations for GS decreases when more gradients are saved and the best tradeoff is obtained with $m = 15$.

TABLE VI
RECONSTRUCTION OF peppers: ITERATION NUMBER K/TIME T (s) BEFORE CONVERGENCE FOR MG AND NLCG FOR DIFFERENT CONJUGACY STRATEGIES. IN ALL CASES, THE STEPSIZE RESULTS FROM J ITERATIONS OF THE MM RECURRENCE

J	1	2	5	10
NLCG-FR	145/270	137/279	143/379	143/515
NLCG-DY	234/447	159/338	144/387	143/516
NLCG-PRP	77/137	69/139	75/202	77/273
NLCG-HS	68/122	67/134	75/191	77/289
NLCG-LS	82/149	67/135	74/190	76/266
MG	**67/119**	66/141	74/211	76/297

TABLE VII
ITERATION NUMBER K/TIME T (s) BEFORE CONVERGENCE FOR MG AND NLCG ALGORITHMS. IN ALL CASES, THE NUMBER OF MM SUBITERATIONS IS SET TO $J = 1$

	boat	lena	peppers
NLCG-FR	77/141	98/179	145/270
NLCG-DY	86/161	127/240	234/447
NLCG-PRP	40/74	55/99	77/137
NLCG-HS	39/71	50/93	68/122
NLCG-LS	42/81	57/103	82/149
MG	**37/67**	**47/85**	**67/119**

2) Comparison With Conjugate Gradient Algorithms: Let us compare the MG algorithm (i.e., SMG with $m = 1$) with the NLCG algorithm making use of the MM line search strategy proposed in [29]. The latter is based on the following descent recurrence:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k(-\boldsymbol{g}_k + \beta_k \boldsymbol{d}_{k-1})$$

where β_k is the conjugacy parameter. Table VI summarizes the performances of NLCG for five different conjugacy strategies described in [21]. The stepsize α_k in NLCG results from J iterations of (15) with $\boldsymbol{A} = \boldsymbol{A}_{\mathrm{GR}}$ and $\theta = 1$. According to Table VI, the convergence speed of the conjugate gradient method is very sensitive to the conjugacy strategy. The last line of Table VI reproduces the first column of Table IV. The five tested NLCG methods are outperformed by the MG subspace algorithm with $J = 1$, both in terms of iteration number K and computational time T.

The two other cases lena and boat lead to the same conclusion, as reported in Table VII. Finally, Table VIII reports the results obtained with $\mathrm{SNR} = 20$ dB. While the iteration number K and computational time T before convergence globally increased due to the higher noise level, the best results were still observed with MG algorithm.

C. Quasi-Newton Subspace

Dealing with the QNS algorithm, the best results were observed with $J = 1$ iteration of the MM stepsize strategy and

TABLE VIII
ITERATION NUMBER K/TIME T (s) BEFORE CONVERGENCE FOR MG AND NLCG ALGORITHMS FOR $\mathrm{SNR} = 20$ dB. IN ALL CASES, THE NUMBER OF MM SUBITERATIONS IS SET TO $J = 1$

	boat	lena	peppers
NLCG-FR	120/220	171/318	383/713
NLCG-DY	136/255	227/430	532/1016
NLCG-PRP	72/133	100/177	191/339
NLCG-HS	71/129	94/171	177/318
NLCG-LS	73/141	106/192	199/361
MG	**69/125**	**91/162**	**174/309**

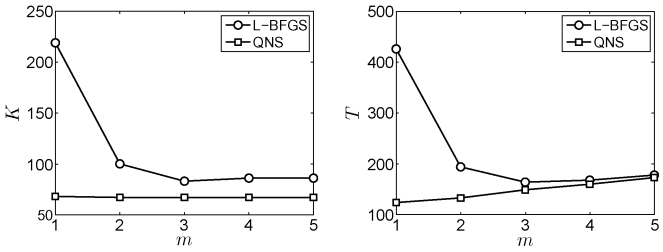


Fig. 5. Reconstruction of peppers: Influence of memory m for algorithms L-BFGS and QNS in terms of iteration number K and computation time T in seconds. In all cases, the number of MM subiterations is set to $J = 1$.

TABLE IX
ITERATION NUMBER K/TIME T (s) BEFORE CONVERGENCE FOR QNS AND L-BFGS ALGORITHMS FOR $J = 1$

	boat	lena	peppers
L-BFGS ($m = 3$)	45/94	62/119	83/164
QNS ($m = 1$)	**38/83**	**48/107**	**68/124**

the memory parameter $m = 1$. For this setting, the peppers image is restored after 68 iterations, which takes 124 s. As a comparison, when the Newton search is used and $m = 1$, the QNS algorithm requires 75 iterations that take more than 1000 s.

Let us now compare the QNS algorithm with the standard L-BFGS algorithm from [22]. Both algorithms require the tuning of the memory size m. Fig. 5 illustrates the performances of the two algorithms. In both cases, the stepsize results from one iteration of MM recurrence. Contrary to L-BFGS, QNS is not sensitive to the size of the memory m. Moreover, according to Table IX, the QNS algorithm outperforms the standard L-BFGS algorithm with its best memory setting for the three restoration problems.

D. Truncated Newton Subspace

Now, let us focus on the second order subspace method SESOP-TN. The first component of $\boldsymbol{D}_k^\ell$, $\boldsymbol{d}_k^\ell$, is computed by applying ℓ iterations of the preconditioned CG method to the Newton equations. Akin to the standard TN algorithm, ℓ is chosen according to the following convergence test:

$$\left\| \boldsymbol{g}_k + \boldsymbol{H}_k \boldsymbol{d}_k^\ell \right\| / \|\boldsymbol{g}_k\| < \eta$$

where $\eta > 0$ is a threshold parameter. Here, the setting $\eta = 0.5$ has been adopted since it leads to lowest computation time for the standard TN algorithm.

TABLE X
RECONSTRUCTION OF peppers: ITERATION NUMBER K/TIME T (s) BEFORE CONVERGENCE FOR MM AND NEWTON STEPSIZE STRATEGIES IN SESOP-TN ALGORITHM

SESOP-TN(m)		0	1	2	5
Newton		159/436	155/427	128/382	151/423
MM (J)	1	415/870	410/864	482/979	387/840
	2	253/532	232/506	239/525	345/731
	5	158/380	132/316	143/359	139/351
	10	122/322	134/323	**119/301**	128/334
	15	114/320	134/365	117/337	127/389

TABLE XI
ITERATION NUMBER K/TIME T (s) BEFORE CONVERGENCE FOR SESOP-TN AND TN ALGORITHMS FOR $\eta = 0.5$ AND $J = 10$

	boat	lena	peppers
TN	65/192	**74/199**	137/322
SESOP-TN(2)	**55/180**	76/218	**119/301**

In Tables X and XI, the results are reported in the form K/T where K denotes the total number of CG steps.

According to Table X, SESOP-TN-MM behaves differently from the previous algorithms. A quite large value of J is necessary to obtain the fastest version. In this example, the MM search is still more efficient than the Newton search, provided that we choose $J \geq 5$. Concerning the memory parameter, the best results are obtained for $m = 2$.

Finally, Table XI summarizes the results for the three test images, in comparison with the standard TN (not fully standard, though, since the MM line search has been used). Our conclusion is that the subspace version of TN does not seem to bring a significant acceleration compared to the standard version. Again, this contrasts with the results obtained for the other tested subspace methods.

VI. CONCLUSION

This paper explored the minimization of penalized least squares criteria in the context of image restoration, using the subspace algorithm approach. We pointed out that the existing strategies for computing the multidimensional stepsize suffer either from a lack of convergence results (e.g., Newton search) or from a high computational cost (e.g., trust region method). As an alternative, we proposed an original stepsize strategy based on a MM recurrence. The stepsize results from the minimization of a half-quadratic approximation over the subspace. Our method benefits from mathematical convergence results, whatever the number of MM iterations. Moreover, it can be implemented efficiently by taking advantage of the recursive structure of the subspace.

On practical restoration problems, the proposed search is significantly faster than the Newton minimization used in [6], [26], [27], in terms of computational time before convergence. Quite remarkably, the best performances have almost always been obtained when only one MM iteration was performed ($J = 1$),

and when the size of the memory was reduced to one stored iterate ($m = 1$), which means that simplicity and efficiency meet in our context. In particular, the resulting algorithmic structure contains no nested iterations.

Finally, among all of the tested variants of subspace methods, the best results were obtained with the memory gradient subspace (i.e., where the only stored vector is the previous direction), using a single MM iteration for the stepsize. The resulting algorithm can be viewed as a new form of preconditioned, nonlinear conjugate gradient algorithm, where the conjugacy parameter and the stepsize are jointly given by a closed-form formula that amounts to solve a 2×2 linear system.

APPENDIX

A. Proof of Theorem 1

Let us introduce the scalar function

$$h^{(k)}(\alpha) \triangleq q^{(k)}\left([\alpha, 0, \ldots, 0]^T, \mathbf{0}\right), \quad \forall \alpha \in \mathbb{R}. \tag{23}$$

According to the expression of $q(., \mathbf{0})$, h reads

$$h^{(k)}(\alpha) = f^{(k)}(\mathbf{0}) + \alpha \boldsymbol{g}_k^T \boldsymbol{d}_k^1 + \frac{1}{2}\alpha^2 \boldsymbol{d}_k^{1T} \boldsymbol{A}_k^0 \boldsymbol{d}_k^1. \tag{24}$$

Its minimizer $\hat{\alpha}_k$ is given by

$$\hat{\alpha}_k = -\frac{\boldsymbol{g}_k^T \boldsymbol{d}_k^1}{\boldsymbol{d}_k^{1T} \boldsymbol{A}_k^0 \boldsymbol{d}_k^1}. \tag{25}$$

Therefore

$$h^{(k)}(\hat{\alpha}_k) = f^{(k)}(\mathbf{0}) + \frac{1}{2}\hat{\alpha}_k \boldsymbol{g}_k^T \boldsymbol{d}_k^1. \tag{26}$$

Moreover, according to the expression of $\hat{\boldsymbol{s}}_k^1$, we have

$$q^{(k)}\left(\hat{\boldsymbol{s}}_k^1, \mathbf{0}\right) = f^{(k)}(\mathbf{0}) + \frac{1}{2}\nabla f^{(k)}(\mathbf{0})^T \hat{\boldsymbol{s}}_k^1. \tag{27}$$

$\hat{\boldsymbol{s}}_k^1$ minimizes $q^{(k)}(\boldsymbol{s}, \mathbf{0})$, hence $q^{(k)}(\hat{\boldsymbol{s}}_k^1, \mathbf{0}) \leq h^{(k)}(\hat{\alpha}_k)$. Thus, using (26) and (27), we have

$$\hat{\alpha}_k \boldsymbol{g}_k^T \boldsymbol{d}_k^1 \geq \nabla f^{(k)}(\mathbf{0})^T \hat{\boldsymbol{s}}_k^1. \tag{28}$$

According to (24) and (25), the relaxed stepsize $\alpha_k = \theta\hat{\alpha}_k$ fulfills

$$h^{(k)}(\alpha_k) = f^{(k)}(\mathbf{0}) + \delta\hat{\alpha}_k \boldsymbol{g}_k^T \boldsymbol{d}_k^1 \tag{29}$$

where $\delta = \theta(1 - \theta/2)$. Moreover,

$$q^{(k)}\left(\boldsymbol{s}_k^1, \mathbf{0}\right) = f^{(k)}(\mathbf{0}) + \delta\nabla f^{(k)}(\mathbf{0})^T \hat{\boldsymbol{s}}_k^1. \tag{30}$$

Thus, using (28)–(30), we obtain $q^{(k)}(\boldsymbol{s}_k^1, \mathbf{0}) \leq h^{(k)}(\alpha_k)$ and

$$f^{(k)}(\mathbf{0}) - q^{(k)}\left(\boldsymbol{s}_k^1, \mathbf{0}\right) \geq -\delta\hat{\alpha}_k \boldsymbol{g}_k^T \boldsymbol{d}_k^1. \tag{31}$$

Furthermore, $q^{(k)}(\boldsymbol{s}_k^1, \mathbf{0}) \geq f^{(k)}(\boldsymbol{s}_k^1) \geq f^{(k)}(\boldsymbol{s}_k)$ according to Lemma 1 and [13, Prop. 5]. Thus,

$$f^{(k)}(\mathbf{0}) - f^{(k)}(\boldsymbol{s}_k) \geq -\delta\hat{\alpha}_k \boldsymbol{g}_k^T \boldsymbol{d}_k^1. \tag{32}$$

According to Lemma 2

$$\hat{\alpha}_k \geq -\frac{\boldsymbol{g}_k^T \boldsymbol{d}_k^1}{\nu_1 \left\|\boldsymbol{d}_k^1\right\|^2}. \tag{33}$$

Hence, according to (32), (33), and Assumption H4,

$$f^{(k)}(\mathbf{0}) - f^{(k)}(\boldsymbol{s}_k) \geq \frac{\delta\gamma_0^2}{\nu_1\gamma_1^2}\|\boldsymbol{g}_k\|^2 \tag{34}$$

which also reads

$$F(\boldsymbol{x}_k) - F(\boldsymbol{x}_{k+1}) \geq \frac{\delta\gamma_0^2}{\nu_1\gamma_1^2}\|\boldsymbol{g}_k\|^2. \tag{35}$$

Thus, (20) holds. Moreover, F is bounded below according to Lemma 2. Therefore, $\lim_{k\to\infty} F(\boldsymbol{x}_k)$ is finite. Thus

$$\infty > \left(\frac{\delta\gamma_0^2}{\nu_1\gamma_1^2}\right)^{-1}\left(F(\boldsymbol{x}_0) - \lim_{k\to\infty} F(\boldsymbol{x}_k)\right) \geq \sum_k \|\boldsymbol{g}_k\|^2$$

and finally

$$\lim_{k\to\infty} \|\boldsymbol{g}_k\| = 0.$$

B. Relations Between the PCD and the Gradient Directions

Lemma 3: Let the PCD direction be defined by $\boldsymbol{p} = (p_i)$, with

$$p_i = \arg\min_{\alpha} F(\boldsymbol{x} + \alpha\boldsymbol{e}_i), \qquad i = 1, \ldots, N,$$

where $\boldsymbol{e}_i$ stands for the ith elementary unit vector. If F is gradient Lipschitz and strongly convex on $\mathbb{R}^N$, then there exist $\gamma_0, \gamma_1 > 0$ such that $\boldsymbol{p}$ fulfills

$$\boldsymbol{g}^T\boldsymbol{p} \leq -\gamma_0\|\boldsymbol{g}\|^2 \tag{36}$$

$$\|\boldsymbol{p}\| \leq \gamma_1\|\boldsymbol{g}\| \tag{37}$$

for all $\boldsymbol{x} \in \mathbb{R}^N$.

Proof: Let us introduce the scalar functions $f_i(\alpha) \triangleq F(\boldsymbol{x} + \alpha\boldsymbol{e}_i)$, so that

$$p_i = \arg\min_{\alpha} f_i(\alpha). \tag{38}$$

F is gradient Lipschitz, so there exists $L > 0$ such that for all i

$$\left|\dot{f}_i(a) - \dot{f}_i(b)\right| \leqslant L|a - b|, \quad \forall a, b \in \mathbb{R}.$$

In particular, for $a = 0$ and $b = p_i$, we obtain

$$|p_i| \geqslant \left|\dot{f}_i(0)\right|/L$$

given that $\dot{f}_i(p_i) = 0$ according to (38). According to the expression of f_i,

$$\boldsymbol{g}^T\boldsymbol{p} = \sum_{i=1}^{N} \dot{f}_i(0)p_i.$$

Moreover, p_i minimizes the convex function f_i on $\mathbb{R}$ so

$$p_i \dot{f}_i(0) \leqslant 0, \qquad i = 1, \ldots, N. \tag{39}$$

Therefore

$$\boldsymbol{g}^T \boldsymbol{p} = -\sum_{i=1}^{N} \left| \dot{f}_i(0) \right| |p_i| \leqslant -\frac{1}{L} ||\boldsymbol{g}||^2. \tag{40}$$

F is strongly convex, so there exists $\nu > 0$ such that, for all i,

$$\left(\dot{f}_i(a) - \dot{f}_i(b) \right) (a - b) \geqslant \nu (a - b)^2, \quad \forall a, b \in \mathbb{R}.$$

In particular, $a = 0$ and $b = p_i$ give

$$-\dot{f}_i(0) p_i \geqslant \nu p_i^2, \qquad i = 1, \ldots, N. \tag{41}$$

Using (39), we obtain

$$p_i^2 \leqslant \nu \left| \dot{f}_i(0) \right|^2 / \nu^2, \qquad i = 1, \ldots, N. \tag{42}$$

Therefore

$$||\boldsymbol{p}||^2 = \sum_{i=1}^{N} p_i^2 \leqslant \frac{1}{\nu^2} ||\boldsymbol{g}||^2. \tag{43}$$

Thus, (36) and (37) hold for $\gamma_0 = 1/L$ and $\gamma_1 = 1/\nu$. ■

REFERENCES

[1] M. Elad, P. Milanfar, and R. Rubinstein, "Analysis versus synthesis in signal priors," *Inverse Prob.*, vol. 23, no. 3, pp. 947–968, 2007.

[2] S. Geman and G. Reynolds, "Constrained restoration and the recovery of discontinuities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 3, pp. 367–383, Mar. 1992.

[3] D. Geman and C. Yang, "Nonlinear image recovery with half-quadratic regularization," *IEEE Trans. Image Process.*, vol. 4, no. 7, pp. 932–946, Jul. 1995.

[4] A. Chambolle, R. A. De Vore, L. Nam-Yong, and B. Lucier, "Nonlinear wavelet image processing: Variational problems, compression, and noise removal through wavelet shrinkage," *IEEE Trans. Image Process.*, vol. 7, no. 3, pp. 319–335, Mar. 1998.

[5] M. Figueiredo, J. Bioucas-Dias, and R. Nowak, "Majorization-minimization algorithms for wavelet-based image restoration," *IEEE Trans. Image Process.*, vol. 16, no. 12, pp. 2980–2991, 2007.

[6] M. Zibulevsky and M. Elad, "$\ell_2 - \ell_1$ optimization in signal and image processing," *IEEE Signal. Process. Mag.*, vol. 27, no. 3, pp. 76–88, May 2010.

[7] G. Demoment, "Image reconstruction and restoration: Overview of common estimation structure and problems," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 37, no. 12, pp. 2024–2036, Dec. 1989.

[8] P.-J. Huber, *Robust Statistics*. New York: Wiley, 1981.

[9] S. Geman and D. McClure, "Statistical methods for tomographic image reconstruction," in *Proc. 46th Session ICI, Bull. ICI*, 1987, vol. 52, pp. 5–21.

[10] C. Bouman and K. D. Sauer, "A generalized Gaussian image model for edge-preserving MAP estimation," *IEEE Trans. Image Process.*, vol. 2, no. 3, pp. 296–310, Jul. 1993.

[11] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud, "Deterministic edge-preserving regularization in computed imaging," *IEEE Trans. Image Process.*, vol. 6, no. 2, pp. 298–311, Feb. 1997.

[12] M. Nikolova and M. K. Ng, "Analysis of half-quadratic minimization methods for signal and image recovery," *SIAM J. Sci. Comput.*, vol. 27, no. 3, pp. 937–966, 2005.

[13] A. Allain, J. Idier, and Y. Goussard, "On global and local convergence of half-quadratic algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 5, pp. 1130–1142, May 2006.

[14] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Phys. Rev. D*, vol. 60, pp. 259–268, 1992.

[15] M. Nikolova, "Weakly constrained minimization: Application to the estimation of images and signals involving constant regions," *J. Math. Imag. Vis.*, vol. 21, pp. 155–175, 2004.

[16] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed. Belmont, MA: Athena Scientific, 1999.

[17] P. Ciuciu and J. Idier, "A half-quadratic block-coordinate descent method for spectral estimation," *Signal Process.*, vol. 82, no. 7, pp. 941–959, Jul. 2002.

[18] M. Nikolova and M. K. Ng, "Fast image reconstruction algorithms combining half-quadratic regularization and preconditioning," in *Proc. Int. Conf. Image Process.*, 2001, pp. 277–280.

[19] M. Elad, B. Matalon, and M. Zibulevsky, "Coordinate and subspace optimization methods for linear least squares with non-quadratic regularization," *Appl. Comput. Harmon. Anal.*, vol. 23, pp. 346–367, 2006.

[20] Y. Yuan, , R. Jeltsh, T.-T. Li, and H. I. Sloan, Eds., "Subspace techniques for nonlinear optimization," in *Some Topics in Industrial and Applied Mathematics*, ser. Contemporary Applied Mathematics. Beijing, China: Higher Education, 2007, vol. CAM 8, pp. 206–218.

[21] W. W. Hager and H. Zhang, "A survey of nonlinear conjugate gradient methods," *Pacific J. Optim.*, vol. 2, no. 1, pp. 35–58, Jan. 2006.

[22] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Math. Program.*, ser. B, vol. 45, no. 3, pp. 503–528, 1989, (Ser. B).

[23] A. Miele and J. W. Cantrell, "Study on a memory gradient method for the minimization of functions," *J. Optim. Theor. Appl.*, vol. 3, no. 6, pp. 459–470, 1969.

[24] E. E. Cragg and A. V. Levy, "Study on a supermemory gradient method for the minimization of functions," *J. Optim. Theor. Appl.*, vol. 4, no. 3, pp. 191–205, 1969.

[25] Z. Wang, Z. Wen, and Y. Yuan, "A subspace trust region method for large scale unconstrained optimization," in *Numerical Linear Algebra and Optimization*. Marrickville, Australia: Science, 2004, pp. 264–274.

[26] G. Narkiss and M. Zibulevsky, "Sequential subspace optimization method for large-scale unconstrained problems," Israel Inst. Technol., Tech. Rep. 559, Oct. 2005 [Online]. Available: http://iew3.technion.ac.il/mcib/sesop_report_version301005.pdf

[27] M. Zibulevsky, "SESOP-TN: Combining sequential subspace optimization with truncated Newton method," Israel Inst. Technol., Sep. 2008 [Online]. Available: http://www.optimization-online.org/DB_FILE/2008/09/2098.pdf

[28] A. R. Conn, N. Gould, A. Sartenaer, and P. L. Toint, "On iterated-subspace minimization methods for nonlinear optimization," Rutherford Appleton Lab., Oxfordshire, U.K., Tech. Rep. 94-069, May 1994 [Online]. Available: ftp://130.246.8.32/pub/reports/cgstRAL94069.ps.Z

[29] C. Labat and J. Idier, "Convergence of conjugate gradient methods with a closed-form stepsize formula," *J. Optim. Theor. Appl.*, vol. 136, no. 1, pp. 43–60, Jan. 2008.

[30] D. R. Hunter and K. L. , "A tutorial on MM algorithms," *Am. Stat.*, vol. 58, no. 1, pp. 30–37, Feb. 1, 2004.

[31] J. Cantrell, "Relation between the memory gradient method and the fletcher-reeves method," *J. Optim. Theor. Appl.*, vol. 4, no. 1, pp. 67–71, 1969.

[32] M. Wolfe and C. Viazminsky, "Supermemory descent methods for unconstrained minimization," *J. Optim. Theor. Appl.*, vol. 18, no. 4, pp. 455–468, 1976.

[33] Z.-J. Shi and J. Shen, "A new class of supermemory gradient methods," *Appl. Math. Comp.*, vol. 183, pp. 748–760, 2006.

[34] Z.-J. Shi and J. Shen, "Convergence of supermemory gradient method," *Appl. Math. Comp.*, vol. 24, no. 1–2, pp. 367–376, 2007.

[35] Z.-J. Shi and Z. Xu, "The convergence of subspace trust region methods," *J. Comput. Appl. Math.*, vol. 231, no. 1, pp. 365–377, 2009.

[36] A. Nemirovski, "Orth-method for smooth convex optimization," *Izvestia SSSR,Transl.: Eng. Cybern. Soviet J. Comput. Syst. Sci.*, vol. 2, 1982.

[37] Z.-J. Shi and J. Shen, "A new super-memory gradient method with curve search rule," *Appl. Math. Comp.*, vol. 170, pp. 1–16, 2005.

[38] Z. Wang and Y. Yuan, "A subspace implementation of quasi-Newton trust region methods for unconstrained optimization," *Numer. Math.*, vol. 104, pp. 241–269, 2006.

[39] S. G. Nash, "A survey of truncated-Newton methods," *J. Comput. Appl. Math.*, vol. 124, pp. 45–59, 2000.

[40] J. Nocedal and S. J. Wright, *Numerical Optimization*. New York: Springer-Verlag, 1999.

[41] Z.-J. Shi, "Convergence of line search methods for unconstrained optimization," *Appl. Math. Comp.*, vol. 157, pp. 393–405, 2004.

[42] Y. Narushima and Y. Hiroshi, "Global convergence of a memory gradient method for unconstrained optimization," *Comput. Optim. Applic.*, vol. 35, no. 3, pp. 325–346, 2006.
[43] Z. Yu, "Global convergence of a memory gradient method without line search," *J. Appl. Math. Comput.*, vol. 26, no. 1–2, pp. 545–553, Feb. 2008.
[44] J. Liu, H. Liu, and Y. Zheng, S. Berlin, Ed., "A new supermemory gradient method without line search for unconstrained optimization," in *Proc. 6th Int. Symp. Neural Netw.*, 2009, vol. 56, pp. 641–647.
[45] M. W. Jacobson and J. A. Fessler, "An expanded theoretical treatment of iteration-dependent majorize-minimize algorithms," *IEEE Trans. Image Process.*, vol. 16, no. 10, pp. 2411–2422, Oct. 2007.
[46] J. Idier, "Convex half-quadratic criteria and interacting auxiliary variables for image restoration," *IEEE Trans. Image Process.*, vol. 10, no. 7, pp. 1001–1009, Jul. 2001.
[47] R. S. Dembo and T. Steihaug, "Truncated-Newton methods algorithms for large scale unconstrained optimization," *Math. Prog.*, vol. 26, pp. 190–212, 1983.
[48] M. Al-Baali and R. Fletcher, "On the order of convergence of preconditioned nonlinear conjugate gradient methods," *SIAM J. Sci. Comput.*, vol. 17, no. 3, pp. 658–665, 1996.

Emilie Chouzenoux received the engineering degree from École Centrale, Nantes, France, in 2007, and the Ph. D. degree in signal processing from the Institut de Recherche en Communications et Cybernétique, Nantes, France, in 2010.

She is currently a Graduate Teaching Assistant with the University of Paris-Est, Champs-sur-Marne, France (LIGM, UMR CNRS 8049). Her research interests are in optimization algorithms for large scale problems of image and signal reconstruction.

Jérôme Idier (M'09) was born in France in 1966. He received the Diploma degree in electrical engineering from École Supérieure d'Électricité, Gif-sur-Yvette, France, in 1988, and the Ph.D. degree in physics from University of Paris-Sud, Orsay, France, in 1991.

In 1991, he joined the Centre National de la Recherche Scientifique. He is currently a Senior Researcher with the Institut de Recherche en Communications et Cybernétique, Nantes, France. His major scientific interests are in probabilistic approaches to inverse problems for signal and image processing. He is currently serving as an associate editor for the *Journal of Electronic Imaging*.

Dr. Idier is currently serving as an associate editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING.

Saïd Moussaoui received the State engineering degree from Ecole Nationale Polytechnique, Algiers, Algeria, in 2001, and the Ph.D. degree in automatic control and signal processing from Université Henri Poincaré, Nancy, France, in 2005.

He is currently Assistant Professor with École Centrale de Nantes, Nantes, France. Since September 2006, he has been with the Institut de Recherche en Communications et Cybernétique, Nantes (IRCCYN, UMR CNRS 6597). His research interests are in statistical signal and image processing including source separation, Bayesian estimation, and their applications.

A.2 A Majorize-minimize linesearch for inversion methods involving barrier function optimization

E. Chouzenoux, **S. Moussaoui** et J. Idier, *Inverse Problems*, vol. 28, 065011 (24 pages), 2012.

Cet article est dédié à la stratégie de recherche pas de descente dans le cas d'un critère barrière en utilisant une technique de majoration-minimisation log-quadratique décrite dans le chapitre 2 de ce manuscrit.

Inverse Problems **28** (2012) 065011 (24pp) doi:10.1088/0266-5611/28/6/065011

Majorize–minimize linesearch for inversion methods involving barrier function optimization

E Chouzenoux[1], **S Moussaoui and J Idier**

L'UNAM Université, Ecole Centrale Nantes, CNRS, IRCCyN UMR 6597, 1 rue de la Noë, BP 92101, F-44321 Nantes Cedex 3, France

E-mail: emilie.chouzenoux@univ-mlv.fr

Received 17 February 2011, in final form 28 February 2012
Published 15 May 2012
Online at stacks.iop.org/IP/28/065011

Abstract

This paper focuses on the issue of stepsize determination (linesearch) in iterative descent algorithms applied to the minimization of a criterion containing a barrier function associated with linear constraints. Such an issue arises in inversion methods involving the minimization of a penalized criterion where the barrier function comes either from the data fidelity term or from the regularizing functional. In order to circumvent the inefficiency of general-purpose linesearch strategies in the case of barrier functions, we propose to adopt a majorization–minimization scheme by deriving a new form of a majorant function well suited to approximate a criterion containing barrier terms. We also establish the convergence of classical descent algorithms when this linesearch strategy is employed. Its efficiency is illustrated by means of numerical examples of signal and image restoration.

(Some figures may appear in colour only in the online journal)

1. Introduction

A common inverse problem arising in many application domains is to estimate an object from a set of observations depending on this object through a measurement process. In this paper, we consider the frequent situation where the dependence of the observations $\boldsymbol{y} \in \mathbb{R}^M$ on the unknown discretized object $\boldsymbol{x}^o \in \mathbb{R}^N$ is represented by a linear model:

$$\boldsymbol{y} = \boldsymbol{K}\boldsymbol{x}^o + \boldsymbol{\epsilon}, \tag{1}$$

with $\boldsymbol{K}$ being a known ill-conditioned matrix and $\boldsymbol{\epsilon}$ an additive noise term representing measurement errors and model uncertainties. This simple formalism covers many real situations such as deblurring, denoising and inverse-Radon transform in tomography [1]. It can also be used as a first-order approximation of a nonlinear observation model [2]. To

[1] Present address: LIGM, CNRS-UMR 8049, Université Paris-Est, France.

handle the ill-posedness of such problems, several efficient inversion methods are based on the minimization of a composite criterion (see for instance [3, 4] and references therein)

$$F(\boldsymbol{x}) = S(\boldsymbol{x}) + \lambda R(\boldsymbol{x}), \quad \lambda \geqslant 0. \tag{2}$$

The first term $S(\boldsymbol{x})$ aims at enforcing some fidelity of the solution to the data. It typically corresponds to a neg-log-likelihood, which is derived from the statistics of the noise $\boldsymbol{\epsilon}$. The second term $R(\boldsymbol{x})$, whose weight is set by the parameter λ, is a regularization term that allows us to account for additional information not carried out by the data alone. Its design is linked to some *a priori* assumptions one can have concerning the sought object. Both terms will be assumed differentiable in the following.

The effective resolution of the inverse problem is then expressed as that of finding the minimizer of the composite criterion (2). However, in several cases, the solution cannot be given explicitly or cannot be computed directly since it requires the inversion of large-scale matrices. Instead, iterative descent algorithms are employed. Starting from an initial guess $\boldsymbol{x}_0$, these algorithms generate a sequence of iterates $\{\boldsymbol{x}_k\}$ until the fulfillment of a stopping condition [6]. In practice, from the current value $\boldsymbol{x}_k$, the update $\boldsymbol{x}_{k+1}$ is obtained according to

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k, \tag{3}$$

where $\alpha_k > 0$ is the *stepsize* and $\boldsymbol{d}_k$ is a *descent direction,* i.e. a vector such that $\boldsymbol{g}_k^T \boldsymbol{d}_k < 0$, where $\boldsymbol{g}_k = \nabla F(\boldsymbol{x}_k)$ denotes the gradient of F at $\boldsymbol{x}_k$. The determination of α_k is called the *linesearch*. Linesearch strategies perform an inexact minimization of $f(\alpha) = F(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k)$ to find a stepsize value that ensures the convergence of the whole descent algorithm [5, 6].

The strategies used for computing the direction and the stepsize strongly depend on the mathematical properties of the criterion. In this paper, we focus on penalized criteria that contain a *barrier* function associated with some constraints $\boldsymbol{x} \in \mathcal{C}$. A fundamental property of barrier functions is to ensure that any minimizer of F belongs to the interior of a feasible domain $\mathcal{C}$ by making the gradient of F unbounded at the boundary of $\mathcal{C}$. This property is used by interior-point algorithms [7] to solve inequality-constrained optimization problems, a barrier function being artificially introduced to the objective function. Interior-point methods have been applied for instance to sparse signal reconstruction [8] and to image reconstruction under positivity constraints [9].

Table 1 reports several examples of barrier criteria that can be encountered in the context of signal or image reconstruction. In the first two examples, the barrier results from the presence of singular terms in the data fidelity term. For example, when a Poisson noise distribution is assumed, $S(\boldsymbol{x})$ corresponds to the Kullback–Leibler divergence of $\boldsymbol{Kx}$ from $\boldsymbol{y}$, which exhibits a barrier function associated with the constraints $[\boldsymbol{Kx}]_m > 0$, $m = 1, \ldots, M$. In section 4.1, we will consider a positron emission tomography (PET) problem [10] which involves this form of likelihood. In the other examples, the barrier function is part of the regularization term. For instance, Shannon and Burg entropic penalty terms, used in the maximum entropy strategy for image reconstruction [11], act as barrier functions for the positivity orthant. The maximum entropy approach will be applied in section 4.2 to the reconstruction of one-dimensional nuclear magnetic resonance (NMR) spectra.

As discussed in [23], general-purpose linesearch techniques tend to be inefficient in the case of criteria containing a barrier function. In this paper, we propose a majorization–minimization (MM) approach by constructing a tangent majorant function suitable for a wide set of barriers. As will be shown hereafter, the main advantage of this approach is to yield

Table 1. Examples of barrier functions encountered in penalized signal or image reconstruction. The first two functions are data fidelity functions $S(\boldsymbol{x})$ while the others are penalty functions $R(\boldsymbol{x})$. We emphasize that Gamma log-likelihood and the two roughness penalties do not fall within the scope of this study.

Name	Function	Constraints
Log-likelihoods		
Poisson [12]	$\sum_{m=1}^{M} [\boldsymbol{K}\boldsymbol{x}]_m - y_m + y_m \log \frac{y_m}{[\boldsymbol{K}\boldsymbol{x}]_m}$	$[\boldsymbol{K}\boldsymbol{x}]_m > 0$
Gamma [13, 14]	$\sum_{m=1}^{M} -\log \frac{y_m}{[\boldsymbol{K}\boldsymbol{x}]_m} + \frac{y_m}{[\boldsymbol{K}\boldsymbol{x}]_m}$	$[\boldsymbol{K}\boldsymbol{x}]_m > 0$
Prior log-densities		
Gamma [15, 16]	$\sum_{n=1}^{N} (1-\alpha_n)\log x_n + \frac{\alpha_n}{\beta_n} x_n$	$x_n > 0$
Beta [16]	$\sum_{n=1}^{N} (1-\alpha_n)\log(x_n - a_n) + (1-\beta_n)\log(b_n - x_n)$	$x_n \in (a_n, b_n)$
Rayleigh [17]	$\sum_{n=1}^{N} -\log(x_n) + \alpha_n x_n^2$	$x_n > 0$
Entropies		
Shannon [18]	$\sum_{n=1}^{N} x_n \log x_n$	$x_n > 0$
Burg [11]	$-\sum_{n=1}^{N} \log x_n$	$x_n > 0$
Hyperbolic [19]	$-\sum_{n=1}^{N} \sqrt{x_n}$	$x_n > 0$
Cross entropy [20]	$\sum_{n=1}^{N} x_n \log \frac{x_n}{r_n} + x_n - r_n$	$x_n > 0$
Generalized Fermi-Dirac [21]	$\sum_{n=1}^{N} (x_n - a_n)\log(x_n - a_n) + (b_n - x_n)\log(b_n - x_n)$	$x_n \in (a_n, b_n)$
Roughness penalties		
Kullback–Leibler [22]	$\sum_{n=1}^{N} x_n \log \frac{x_n}{x_{n-1}} + x_n - x_{n-1}$	$x_n > 0$
Itakura–Saito [22]	$\sum_{n=1}^{N} -\log \frac{x_n}{x_{n-1}} + \frac{x_n}{x_{n-1}}$	$x_n > 0$

a simple scheme for stepsize determination that ensures the convergence of many descent algorithms whatever the number of linesearch iterations.

The rest of this paper is organized as follows. In section 2, we recall the main properties of the barrier functions and discuss why specific linesearch strategies are called for when dealing with the optimization of a criterion containing a barrier function. The proposed linesearch procedure is introduced and its properties are studied in section 3. Section 4 illustrates the efficiency of the proposed approach through numerical examples in the field of signal and image processing.

Table 2. Examples of scalar barrier functions associated with $u > 0$. The first two are strict barriers since they grow to infinity as $u \to 0$. Note that property (7) does not hold for the inverse barrier function.

Barrier name	Logarithmic	Inverse	Entropic	Hyperbolic
Function $\psi(u)$	$-\log u$	u^{-1}	$u \log u$	$-u^r$, $r \in (0, 1)$

2. Linesearch strategies for barrier functions

2.1. Formulation of the criterion involving barrier functions

In this paper, we focus on the cases when the composite criterion (2) can be rewritten as

$$F(\boldsymbol{x}) = P(\boldsymbol{x}) + B(\boldsymbol{x}), \tag{4}$$

where B is a barrier function associated with $\boldsymbol{x} \in \mathcal{C}$, with $\mathcal{C}$ being defined by linear inequalities

$$\mathcal{C} = \{\boldsymbol{x} \in \mathbb{R}^N | C_i(\boldsymbol{x}) = \boldsymbol{c}_i^T \boldsymbol{x} + \rho_i > 0, \ \forall i \in \{1, \ldots, I\}\}, \tag{5}$$

and P is a differentiable criterion over $\mathcal{C}$. A barrier function associated with $\boldsymbol{x} \in \mathcal{C}$ is built as

$$B(\boldsymbol{x}) = \sum_{i=1}^{I} \psi_i(C_i(\boldsymbol{x})), \tag{6}$$

where for all $i \in \{1, \ldots, I\}$, $\psi_i(u)$ are scalar barrier functions associated with $u > 0$, i.e.

$$\begin{aligned} &\psi_i \text{ is continuous and strictly convex on } (0, +\infty[\\ &\psi_i(u) \text{ is differentiable on } (0, +\infty[\\ &\lim_{u \to 0} \dot{\psi}_i(u) = +\infty. \end{aligned}$$

When $\lim_{u \to 0} \psi_i(u) = +\infty$, the scalar barrier ψ_i is said *strict*. In the particular case where ψ_i is a strict scalar barrier function for all $i \in \{1, \ldots, I\}$, $B(\boldsymbol{x})$ is called a strict barrier function. We restrict ourselves to barrier functions (6) formed of scalar barriers ψ_i that fulfil

$$-\frac{2}{u}\ddot{\psi}_i(u) \leqslant \dddot{\psi}_i(u) \leqslant 0, \quad \forall u > 0, \quad \forall i \in \{1, \ldots, I\}. \tag{7}$$

This assumption allows us to consider, for the ψ_i in (6), logarithmic, entropic and hyperbolic scalar barrier functions presented in table 2. Therefore, all barrier functions from table 1 fall within the scope of (6)–(7) except the Gamma log-likelihood and the two roughness penalties.

2.2. Determination of the stepsize

Let $\boldsymbol{x}_k \in \mathcal{C}$ and $\boldsymbol{d}_k$ a descent direction for F at $\boldsymbol{x}_k$. In order to compute the new iterate $\boldsymbol{x}_{k+1}$, one has to perform a linesearch that identifies a step length α_k achieving sufficient decrease in $f(\alpha) = F(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k)$ [6, chapter 3]. The presence of scalar barrier functions ψ_i implies that the derivative of the scalar function

$$f(\alpha) = P(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k) + \sum_{i=1}^{I} \psi_i(C_i(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k)) \tag{8}$$

is unbounded when α is such that $C_i(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k) = 0$ for some i. Since functions C_i are assumed to be linear, this limits the stepsize value α_k to an interval (α_-, α_+) where

$$\begin{cases} \alpha_- = \max_{i \in \mathcal{I}_-} -\dfrac{\theta_i}{\delta_i} \\ \alpha_+ = \min_{i \in \mathcal{I}_+} -\dfrac{\theta_i}{\delta_i} \end{cases} \quad \text{with} \quad \begin{cases} \mathcal{I}_- = \{i \in \{1, \ldots, I\} | \, \delta_i > 0\} \\ \mathcal{I}_+ = \{i \in \{1, \ldots, I\} | \, \delta_i < 0\} \end{cases}, \tag{9}$$

where for all $i \in \{1, \ldots, I\}$, $\theta_i = \boldsymbol{c}_i^T \boldsymbol{x}_k + \rho_i$, $\delta_i = \boldsymbol{c}_i^T \boldsymbol{d}_k$, and it is understood that $\alpha_- = -\infty$ (respectively, $\alpha_+ = +\infty$) if $\mathcal{I}_-$ (resp., $\mathcal{I}_+$) is empty. Moreover, the stepsize should fulfil some sufficient convergence conditions. The most popular are the strong Wolfe conditions that state that a stepsize series $\{\alpha_k\}$ is acceptable if there exist $\sigma_1, \sigma_2 \in (0, 1)$ such that for all k and for all $\boldsymbol{x}_k$,

$$F(\boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k) \leqslant F(\boldsymbol{x}_k) + \sigma_1 \alpha_k \boldsymbol{g}_k^T \boldsymbol{d}_k, \tag{10}$$

$$|\nabla F(\boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k)^T \boldsymbol{d}_k| \leqslant \sigma_2 |\boldsymbol{g}_k^T \boldsymbol{d}_k|. \tag{11}$$

The stepsize is then determined with an iterative procedure that generates candidate values, until (10)–(11) are satisfied. One iteration usually consists in a bracketing phase that finds an interval containing acceptable stepsizes, followed by a polynomial cubic interpolation phase that computes a particular stepsize within this interval [5, 6]. However, cubic interpolation is not suited to interpolate function (8) since its derivative $\dot{f}(\alpha)$ tends to $-\infty$ when α tends to α_- or α_+. Therefore, new interpolating functions have been proposed in [23, 24] to account for the barrier singularity.

2.3. Interpolation based linesearch for barrier functions optimization

The particular case $\psi_i(u) = -\mu \log(u)$, $\mu > 0$, for all $i \in \{1, \ldots, I\}$, is considered in [23–25]. In order to account for the logarithmic barrier term, Murray *et al* proposed a log-quadratic interpolating function of the form

$$f_0 + f_1 \alpha + f_2 \alpha^2 - \mu \log(f_3 - \alpha), \tag{12}$$

where the coefficients f_i are chosen to fit f and its derivative at two or three trial points. More precisely, four interpolating schemes are considered where in each case, f_0, f_1 and f_2 have an analytical expression, while the computation of f_3 requires to solve a scalar equation. In order to guarantee the uniqueness of f_3, some inequality has to be fulfilled. If this is the case, f_3 is computed from an iterative Newton procedure. Otherwise, f_3 is undefined and a cubic interpolation is rather used. The linesearch strategy consists in repeating this interpolation process over intervals $[a, b]$ until the fulfilment of Wolfe conditions [24] or Armijo condition [25]. Let us remark that the resulting algorithms are not often used in practice, possibly because the interpolating function is difficult to compute.

3. Majorize–minimize linesearch

Recently, a linesearch procedure based on the MM principle has been introduced [26]. In this strategy, the stepsize α_k results from successive minimizations of quadratic tangent *majorant functions* for $f(.)$. The function $h(., \alpha')$ is said tangent majorant for $f(.)$ at α' if for all α,

$$\begin{cases} h(\alpha, \alpha') \geqslant f(\alpha), \\ h(\alpha', \alpha') = f(\alpha'). \end{cases} \tag{13}$$

The convergence of conjugate-gradient [27, 28] and truncated Newton (TN) algorithms [29] associated with quadratic MM linesearch strategy has been established. A major advantage of quadratic majorization is that it gives an analytical formulation of the stepsize value. However, its application is not possible in the case of a strict barrier function since there exists no quadratic function that majorizes f in the set (α_-, α_+). For ensuring convergence properties, it would be sufficient to find a function h majorizing f in the interval defined by $\{\alpha \in \mathbb{R} | f(\alpha) \leqslant f(0)\}$. However, such an interval can be difficult to compute or even impossible to approximate.

In the case of nonstrict barriers, f is bounded at the boundary of the set (α_-, α_+). However, the curvature of f is unbounded and one can expect suboptimal results by majorizing the scalar function with a parabola. In particular, very high curvature will be obtained for stepsize values close to the singularity. In this section, we propose a new form of a tangent majorant function that is well suited to approximate a criterion containing a barrier function.

3.1. A new tangent majorant for MM linesearch

Let $\alpha' \in (\alpha_-, \alpha_+)$ be a current stepsize value. Instead of a quadratic, we propose the following form of the tangent majorant function of f at α':

$$h(\alpha, \alpha') = h_0 + h_1\alpha + h_2\alpha^2 - h_3 \log(h_4 - \alpha). \tag{14}$$

The majorant function (14) takes a similar form to (12) but, here, parameters h_i are chosen to ensure the majorization properties (13) for all α and α' in (α_-, α_+). According to the MM principle, the stepsize is defined by $\alpha_k = \alpha^J$, with

$$\begin{aligned} &\alpha^0 = 0 \\ &\alpha^{j+1} = \operatorname*{argmin}_{\alpha} h(\alpha, \alpha^j), \quad j = 0, \ldots, J-1, \end{aligned} \tag{15}$$

where $h(\cdot, \alpha^j)$ is the tangent majorant function

$$\begin{aligned} h(\alpha, \alpha^j) = f(\alpha^j) + (\alpha - \alpha^j)\dot{f}(\alpha^j) + \frac{1}{2}m^j(\alpha - \alpha^j)^2 \\ + \gamma^j\left[(\bar{\alpha}^j - \alpha^j)\log\frac{\bar{\alpha}^j - \alpha^j}{\bar{\alpha}^j - \alpha} - \alpha + \alpha^j\right], \end{aligned} \tag{16}$$

which depends on three parameters m_j, γ_j and $\bar{\alpha}^j$. It is easy to check that $h(\alpha, \alpha) = f(\alpha)$ for all α. There remains to find values of m^j, γ^j, $\bar{\alpha}^j$ such that $h(\alpha, \alpha^j) \geqslant f(\alpha)$ for all $\alpha \in (\alpha_-, \alpha_+)$.

3.2. Construction of the majorant function

Let us introduce the following assumption on P.

Assumption 1. *For all $\boldsymbol{x}' \in \mathbb{R}^N$, there exists a symmetric matrix $\boldsymbol{A}(\boldsymbol{x}')$ such that*

$$Q(\boldsymbol{x}, \boldsymbol{x}') = P(\boldsymbol{x}') + (\boldsymbol{x} - \boldsymbol{x}')^T\nabla P(\boldsymbol{x}') + \tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{x}')^T\boldsymbol{A}(\boldsymbol{x}')(\boldsymbol{x} - \boldsymbol{x}') \geqslant P(\boldsymbol{x}) \tag{17}$$

for all $\boldsymbol{x}$. Moreover, for any bounded set $\mathcal{V}$ included in the definition domain of $P(.)$, the set $\{\boldsymbol{A}(\boldsymbol{x}) | \boldsymbol{x} \in \mathcal{V}\}$ has a positive bounded spectrum with bounds $(\nu^{\mathcal{A}}_{\min}, \nu^{\mathcal{A}}_{\max})$, i.e. for all $\boldsymbol{x} \in \mathcal{V}$,

$$0 < \nu^{\mathcal{A}}_{\min} \leqslant \frac{\boldsymbol{v}^T\boldsymbol{A}(\boldsymbol{x})\boldsymbol{v}}{\|\boldsymbol{v}\|^2} \leqslant \nu^{\mathcal{A}}_{\max}, \ \forall \boldsymbol{v} \in \mathbb{R}^N \setminus \{\boldsymbol{0}\}. \tag{18}$$

As emphasized in [28, lemma 2.1], assumption 1 holds if $P(.)$ is gradient Lipschitz with constant L_p by setting $\boldsymbol{A}(\boldsymbol{x}) = L_p\boldsymbol{I}_N$ for all $\boldsymbol{x}$, where $\boldsymbol{I}_N$ states for the identity matrix with size $N \times N$. Useful methods for constructing $\boldsymbol{A}(\boldsymbol{x})$ without requiring the knowledge of L_p are developed in [30, 31].

Under assumption 1, the majorization of (8) is given by the following theorem.

Theorem 1. *Let $F = P + B$, where P fulfils assumption 1 and B takes the form (6) where ψ_i fulfils (7) for all $i \in \{1, \ldots, I\}$. For all $\boldsymbol{x}_k \in \mathcal{C}$ and $\boldsymbol{d}_k \in \mathbb{R}^N$, the log-quadratic function (16) is tangent majorant for (8) at α^j for the following parameters*

$$\begin{cases} \bar{\alpha}^j = \alpha_- \\ m^j = m_p^j + Z_2(\alpha^j) \\ \gamma^j = (\alpha_- - \alpha^j)Z_1(\alpha^j) \end{cases} \quad \forall \alpha \in (\alpha_-, \alpha^j], \tag{19}$$

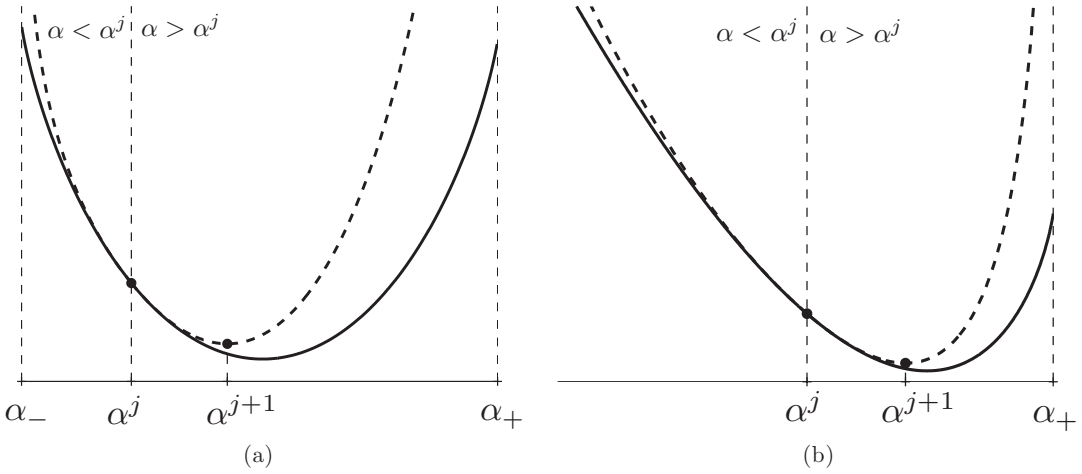


Figure 1. Schematic principle of the MM linesearch procedure. The tangent majorant function $h(\cdot, \alpha^j)$ (dashed line) for $f(\cdot)$ (solid line) at α^j is piecewise defined on the sets $(\alpha_-, \alpha^j]$ and $[\alpha^j, \alpha_+)$. The new iterate α^{j+1} is taken as the minimizer of $h(., \alpha^j)$. Two cases are illustrated. The third and last case, where α_- is finite and $\alpha_+ = +\infty$ is the mirror image of case (b). (a) Case α_- and α_+ finite. (b) Case $\alpha_- = -\infty$ and α_+ finite.

$$\begin{cases} \bar{\alpha}^j = \alpha_+ \\ m^j = m_p^j + Z_1(\alpha^j) \\ \gamma^j = (\alpha_+ - \alpha^j) Z_2(\alpha^j) \end{cases} \qquad \forall \alpha \in [\alpha^j, \alpha_+) \tag{20}$$

where

$$\begin{cases} m_p^j = \boldsymbol{d}_k^T \boldsymbol{A}(\boldsymbol{x}_k + \alpha^j \boldsymbol{d}_k) \boldsymbol{d}_k, \\ Z_1(\alpha^j) = \sum\limits_{i \in \mathcal{I}_-} \zeta_i(\alpha^j), \\ Z_2(\alpha^j) = \sum\limits_{i \in \mathcal{I}_+} \zeta_i(\alpha^j), \end{cases}$$

with $\zeta_i(\alpha) = \delta_i^2 \ddot{\psi}_i(\theta_i + \alpha\delta_i)$ *for all* $i = 1, \ldots, I$.

Proof. See appendix A. □

Remark 1. If the set $\mathcal{I}_-$ is empty (i.e. $\alpha_- = -\infty$), it is understood that $Z_1(\alpha^j)$ equals zero. Thus, for all $\alpha \in (-\infty, \alpha^j]$, $\gamma^j = 0$ and the tangent majorant function has the following expression:

$$h(\alpha, \alpha^j) = f(\alpha^j) + (\alpha - \alpha^j)\dot{f}(\alpha^j) + \tfrac{1}{2}\big(m_p^j + Z_2(\alpha^j)\big)(\alpha - \alpha^j)^2. \tag{21}$$

Correspondingly, if $\mathcal{I}_+$ is empty (i.e. $\alpha_+ = +\infty$), $Z_2(\alpha^j) = 0$ so that for all $\alpha \in [\alpha^j, +\infty)$,

$$h(\alpha, \alpha^j) = f(\alpha^j) + (\alpha - \alpha^j)\dot{f}(\alpha^j) + \tfrac{1}{2}\big(m_p^j + Z_1(\alpha^j)\big)(\alpha - \alpha^j)^2. \tag{22}$$

Although theorem 1 separately defines $h(\alpha, \alpha^j)$ whether α is in $(\alpha_-, \alpha^j]$ or $[\alpha^j, \alpha+)$ (see figure 1 for an illustration), the resulting function is twice differentiable and convex according to the following lemma.

Lemma 1. *Under assumption 1,* $h(., \alpha^j)$ *is* C^2 *and strictly convex in* (α_-, α_+).

Proof. See appendix B. □

3.3. Minimization of the tangent majorant

According to the MM theory, the stepsize α_k is defined by (15) where $h(\cdot, \alpha^j)$ is the tangent majorant function (16). The MM recurrence (15) involves the computation of the minimizer of $h(\cdot, \alpha^j)$ for $j \in \{0, \ldots, J-1\}$. Thanks to lemma 1, the tangent majorant $h(., \alpha^j)$ has a unique minimizer, which can be expressed as an explicit function of $\dot{f}(\alpha^j)$ as follows:

$$\alpha^{j+1} = \begin{cases} \alpha^j - \dfrac{2q_3}{q_2 + \sqrt{q_2^2 - 4q_1q_3}} & \text{if} \quad |\bar{\alpha}^j| < \infty \quad \text{and} \quad \dot{f}(\alpha^j) \leqslant 0 \\ \alpha^j - \dfrac{2q_3}{q_2 - \sqrt{q_2^2 - 4q_1q_3}} & \text{if} \quad |\bar{\alpha}^j| < \infty \quad \text{and} \quad \dot{f}(\alpha^j) > 0 \\ \alpha^j - \dfrac{\dot{f}(\alpha^j)}{m^j} & \text{if} \quad |\bar{\alpha}^j| = \infty \end{cases} \tag{23}$$

with

$$\begin{cases} q_1 = -m^j \\ q_2 = \gamma^j - \dot{f}(\alpha^j) + m^j(\bar{\alpha}^j - \alpha^j) \\ q_3 = (\bar{\alpha}^j - \alpha^j)\dot{f}(\alpha^j). \end{cases} \tag{24}$$

Finally, (15) produces monotonically decreasing values $\{f(\alpha^j)\}$ and the series $\{\alpha^j\}$ converges to a stationary point of $f(\alpha)$ [32].

3.4. Convergence analysis

This section studies the convergence of the iterative descent algorithm

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k, \quad k \geqslant 0, \tag{25}$$

when $\boldsymbol{d}_k$ satisfies $\boldsymbol{g}_k^T \boldsymbol{d}_k < 0$ and the stepsize value α_k results from (15). The proposed analysis requires the following assumption on F.

Assumption 2. *For some $\boldsymbol{x}_0 \in \mathcal{C}$, there exists a neighborhood $\mathcal{V}_0$ of the level set $\mathcal{L}_0 = \{\boldsymbol{x} \in \mathbb{R}^N | F(\boldsymbol{x}) \leqslant F(\boldsymbol{x}_0)\}$ such that*

- *$\mathcal{V}_0$ is bounded;*
- *F is differentiable on $\mathcal{V}_0$ and $\nabla F(\boldsymbol{x})$ is Lipschitz continuous on $\mathcal{V}_0$ with the Lipschitz constant $L > 0$, i.e.*

$$\|\nabla F(\boldsymbol{x}) - \nabla F(\boldsymbol{y})\| \leqslant L\|\boldsymbol{x} - \boldsymbol{y}\|, \ \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{V}_0. \tag{26}$$

Let us emphasize that assumption 1 holds for the particular case $\mathcal{V} = \mathcal{V}_0$. Moreover, the boundedness assumption on $\mathcal{V}_0$ holds if F is coercive, that is,

$$\lim_{\|\boldsymbol{x}\| \to +\infty} F(\boldsymbol{x}) = +\infty. \tag{27}$$

3.4.1. Properties of the stepsize series. First, let us recall some essential properties of the MM recurrence.

Lemma 2 ([31, 32])**.** *Let $\boldsymbol{x}_k \in \mathcal{V}_0$ and $\boldsymbol{d}_k$ such that $\boldsymbol{g}_k^T \boldsymbol{d}_k < 0$. For all $j \geqslant 1$, the series $\{\alpha^j\}$ defined by (15) fulfils*

- *$f(\alpha^j) \leqslant f(\alpha^{j-1})$;*
- *$\text{sign}(\alpha^j - \alpha^{j-1}) = -\text{sign}(\dot{f}(\alpha^{j-1}))$;*
- *$\alpha^j > 0$*

and converges to a stationary point of f.

The first item of lemma 2 implies that for all k,

$$F(\boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k) \leqslant F(\boldsymbol{x}_k),$$

which means that the iterates $\{\boldsymbol{x}_k\}$ remain in $\mathcal{V}_0$. However, this simple monotonicity condition does not imply that the descent algorithm (3) converges [6, chapter 3].

3.4.2. Sufficient decrease condition. In order to ensure that the descent algorithm makes reasonable progress, the stepsize value must yield a *sufficient decrease* in the objective function F, as measured by the first Wolfe condition (10).

Property 1. *Let $\boldsymbol{x}_k \in \mathcal{V}_0$ and $\boldsymbol{d}_k$ satisfy $\boldsymbol{g}_k^T \boldsymbol{d}_k < 0$. Under assumptions 1 and 2, the iterates of (15) fulfil*

$$F(\boldsymbol{x}_k + \alpha^j \boldsymbol{d}_k) \leqslant F(\boldsymbol{x}_k) + \sigma_1^j \alpha^j \nabla F(\boldsymbol{x}_k)^T \boldsymbol{d}_k, \tag{28}$$

for all $j \geqslant 1$, with $\sigma_1^j = (2\sigma_{\max}^j)^{-1} \in (0, 1)$ for some $\sigma_{\max}^j > 0$.

Proof. See appendix C. □

Property 1 is a strong result since it means that the MM linesearch produces a sufficient decrease of the criterion, whatever the number of linesearch iterates J.

3.4.3. Stepsize minoration. The first Wolfe condition alone is not sufficient to ensure the convergence since it does not prevent arbitrarily small steps. A second condition is required, such as the second Wolfe condition (11). Here, the proposed convergence study rather relies on a direct minoration of the stepsize values.

Property 2. *Let $\boldsymbol{x}_k \in \mathcal{V}_0$ and $\boldsymbol{d}_k$ satisfy $\boldsymbol{g}_k^T \boldsymbol{d}_k < 0$. Under assumptions 1 and 2, for all $j \geqslant 1$, the iterates of (15) fulfil*

$$\alpha^j \geqslant \sigma_{\min} \alpha^1 \tag{29}$$

and

$$\alpha^j \geqslant \sigma_{\min} \frac{-\boldsymbol{g}_k^T \boldsymbol{d}_k}{\nu \|\boldsymbol{d}_k\|^2}, \tag{30}$$

for some $\sigma_{\min},\ \nu > 0$.

Proof. See appendix D. □

3.4.4. Zoutendijk condition. Obviously, the global convergence of a descent direction method is not only ensured by a good stepsize strategy, but also by well-chosen search directions $\boldsymbol{d}_k$. Convergence proofs often rely on the fulfilment of the Zoutendijk condition:

$$\sum_{k=0}^{\infty} \|\boldsymbol{g}_k\|^2 \cos^2 \theta_k < \infty, \tag{31}$$

where θ_k is the angle between $\boldsymbol{d}_k$ and the steepest descent direction $-\boldsymbol{g}_k$:

$$\cos \theta_k = \frac{-\boldsymbol{g}_k^T \boldsymbol{d}_k}{\|\boldsymbol{g}_k\| \, \|\boldsymbol{d}_k\|}. \tag{32}$$

In the case of the proposed linesearch, properties 1 and 2 lead to the following result.

Property 3. *Let α_k be defined for all k by (15) with $J \geqslant 1$. Under assumptions 1 and 2, $(\boldsymbol{g}_k, \boldsymbol{d}_k)_{k \geqslant 0}$ fulfils the Zoutendijk condition (31).*

Proof. See appendix E. □

3.4.5. Gradient-related directions. Finally, a general convergence result can be established from property 3 by using the concept of *gradient-related* direction [33].

Definition 1. *A direction sequence* $\{\boldsymbol{d}_k\}$ *is said gradient related to* $\{\boldsymbol{x}_k\}$*, if the following property holds: 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*

$$\limsup_{k\to\infty, k\in\mathcal{K}} \boldsymbol{g}_k^T \boldsymbol{d}_k < 0. \tag{33}$$

Property 4 [33]**.** *If there exist* $\nu_1 > 0$, $\nu_2 > 0$ *such that for all k,* $\boldsymbol{d}_k$ *fulfils*

$$\nu_1 \|\boldsymbol{g}_k\|^2 \leqslant -\boldsymbol{g}_k^J \boldsymbol{d}_k, \qquad \|\boldsymbol{d}_k\|^2 \leqslant \nu_2 \|\boldsymbol{g}_k\|^2, \tag{34}$$

then $\{\boldsymbol{d}_k\}$ *is gradient related to* $\{\boldsymbol{x}_k\}$.

Theorem 2. *Let* $\{\boldsymbol{x}_k\}$ *be a sequence generated by a descent method* $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k$. *Assume that the sequence* $\{\boldsymbol{d}_k\}$ *fulfils (34) and* α_k *is defined by (15). Under assumptions 1 and 2, the descent algorithm (25) converges in the sense* $\lim_{k\to\infty} \|\boldsymbol{g}_k\| = 0$.

Proof. According to property 3, the Zoutendijk condition (31) holds. theorem 2 results from [34, theorem 5.1]. □

As emphasized in [35, section 6.2], theorem 2 yields convergence of several classical descent optimization schemes such as the steepest descent method, TN method and the projected gradient method for constrained optimization. Let us remark that it does not cover nonlinear conjugate-gradient algorithms (NLCG) defined by the following recurrence

$$\begin{aligned} \boldsymbol{x}_{k+1} &= \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k, \\ \boldsymbol{d}_{k+1} &= -\boldsymbol{g}_{k+1} + \beta_{k+1} \boldsymbol{d}_k, \end{aligned} \tag{35}$$

where β_k is the conjugacy parameter. However, a deeper analysis of the MM stepsize properties shows that specific convergence results hold for standard NLCG methods such as Fletcher–Reeves, Polak–Ribière–Polyak and the modified Polak–Ribière–Polyak [35, section 6.3].

4. Numerical results

In this section, two application examples are considered to illustrate the practical efficiency of the proposed linesearch procedure. In both cases, a reference descent optimization algorithm is considered, and the MM linesearch is tested against two Wolfe linesearch strategies taken from [5] and [23] based on polynomial and log-quadratic interpolation.

4.1. Image reconstruction under Poisson noise

A simulated PET ([10]) reconstruction problem is first considered. The measurements in PET are modeled as Poisson random variables:

$$\boldsymbol{y} \sim \text{Poisson}(\boldsymbol{K}\boldsymbol{x} + \boldsymbol{r}), \tag{36}$$

where the nth entry of $\boldsymbol{x} \in \mathbb{R}^N$ represents the radioisotope amount in pixel n and $\boldsymbol{K} \in \mathbb{R}^{M\times N}$ is the projection matrix whose elements K_{mn} model the contribution of the nth pixel to the mth datapoint. The components of $\boldsymbol{y} \in \mathbb{R}^M$ are the counts measured by the detector pairs and $\boldsymbol{r} \in \mathbb{R}^M$ models the background events (scattered events and accidental coincidences). The aim is to reconstruct the image $\boldsymbol{x} \geqslant 0$ from the noisy measurements $\boldsymbol{y}$.

4.1.1. Objective function. According to the noise statistics, the neg-log-likelihood of the emission data is

$$S(\boldsymbol{x}) = \sum_{m=1}^{M} \left([\boldsymbol{K}\boldsymbol{x}]_m + r_m - y_m \log([\boldsymbol{K}\boldsymbol{x}]_m + r_m)\right). \tag{37}$$

A useful penalization promoting smoothness of the estimated image is given by

$$R(\boldsymbol{x}) = \sum_{\ell} \omega_\ell \phi([\boldsymbol{D}\boldsymbol{x}]_\ell),$$

where ϕ is the edge-preserving potential function $\phi(u) = \sqrt{\delta^2 + u^2} - \delta$ and $\boldsymbol{D}\boldsymbol{x}$ is the vector of difference between neighboring pixel intensities [36]. The weights depend on the relative position of the neighbors: $\omega_\ell = 1$ for vertical and horizontal neighbors and $\omega_\ell = 2^{-\frac{1}{2}}$ for diagonal neighbors. The estimated image is the minimizer of the following objective function

$$F(\boldsymbol{x}) = S(\boldsymbol{x}) + \lambda R(\boldsymbol{x}), \tag{38}$$

over the positive orthant $\{\boldsymbol{x} \geqslant \boldsymbol{0}\}$.

An efficient approach for solving this constrained optimization problem is to use the split-gradient method (SGM) from [37] associated with a convergent linesearch strategy ([12, 38]). The first part of the criterion implies the presence of a logarithmic barrier in $S(.)$ associated with the domain $\boldsymbol{K}\boldsymbol{x} + \boldsymbol{r} > \boldsymbol{0}$. We propose to analyze the performance of the SGM algorithm when the stepsize is computed using the proposed MM linesearch.

4.1.2. Optimization strategy. The SGM is a descent algorithm aimed at minimizing a criterion under non-negativity constraints. Assuming that the gradient can be split into positive and negative parts $\nabla F(\boldsymbol{x}) = V(\boldsymbol{x}) - U(\boldsymbol{x})$, $U(\boldsymbol{x}), V(\boldsymbol{x}) \geqslant 0$, for all $\boldsymbol{x} \geqslant \boldsymbol{0}$, the SGM iteration is defined as

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + s_k \boldsymbol{d}_k, \quad \boldsymbol{d}_k = -\mathrm{Diag}\left(\frac{\boldsymbol{x}_k}{V(\boldsymbol{x}_k)}\right) \boldsymbol{g}_k, \quad \boldsymbol{g}_k = V(\boldsymbol{x}_k) - U(\boldsymbol{x}_k), \tag{39}$$

where s_k is a stepsize ensuring the positivity of the iterates. When $s_k = 1$ for all k, iteration (39) takes a very simple multiplicative form,

$$\boldsymbol{x}_{k+1} = \mathrm{Diag}\left(\frac{\boldsymbol{x}_k}{V(\boldsymbol{x}_k)}\right) U(\boldsymbol{x}_k). \tag{40}$$

However, the unit stepsize does not guarantee the convergence of the iterates and a linesearch along $\boldsymbol{d}_k$ has to be performed. More precisely, according to [39], the convergence is ensured if

$$s_k = \min(\tau s_{\max}, \alpha_k), \quad s_{\max} = \max\{s | \boldsymbol{x}_k + s\boldsymbol{d}_k \geqslant 0\}, \quad \tau \in (0, 1), \tag{41}$$

as soon as α_k results from a linesearch along $\boldsymbol{d}_k$ that satisfies both (10) and (31) conditions.

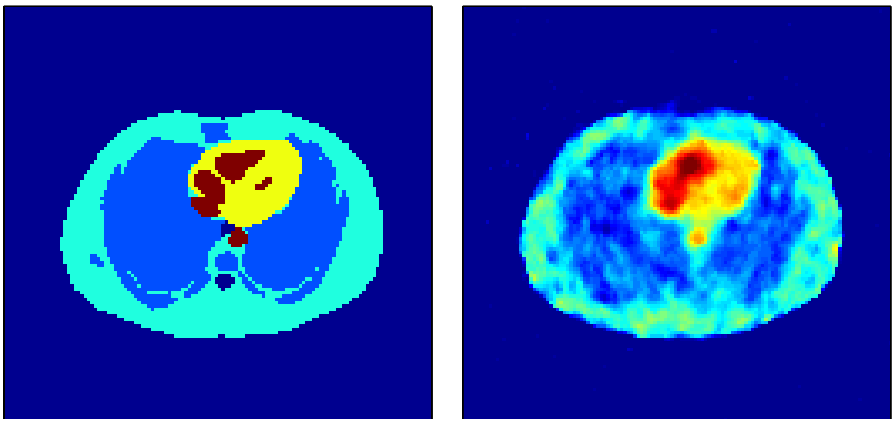
Let $F = P + B$ with

$$B(\boldsymbol{x}) = \sum_{m=1}^{M} -y_m \log([\boldsymbol{K}\boldsymbol{x}]_m + r_m),$$

$$P(\boldsymbol{x}) = \sum_{m=1}^{M} [\boldsymbol{K}\boldsymbol{x}]_m + r_m + \lambda \sum_{\ell} \omega_\ell \phi([\boldsymbol{D}\boldsymbol{x}]_\ell).$$

The linear operators $\boldsymbol{K}$ and $\boldsymbol{D}$ are such that $\ker(\boldsymbol{K}^T\boldsymbol{K}) \cap \ker(\boldsymbol{D}^T\boldsymbol{D}) = \{\boldsymbol{0}\}$. Thus, it is straightforward that assumption 2 holds for all $\boldsymbol{x}_0 > \boldsymbol{0}$. Moreover, according to [30], assumption 1 holds for

$$\boldsymbol{A}(\boldsymbol{x}) = \boldsymbol{D}^T \mathrm{Diag}(\omega(\boldsymbol{D}\boldsymbol{x}))\boldsymbol{D}, \quad \omega(u) = \frac{1}{u}\dot{\phi}(u).$$



(a) (b)

Figure 2. Simulated PET reconstruction with the split-gradient method. (a) Simulated PET phantom. (b) Reconstruction with similarity error 6%.

Therefore, according to properties 1 and 3, the proposed MM linesearch ensures the convergence of the SGM algorithm. We propose to compare its performance with the Moré and Thuente linesearch [5] (MT) based on the fulfilment of the strong Wolfe conditions (10)–(11). Two interpolation schemes will be considered for the MT linesearch, namely the cubic interpolation procedure (MTcubic) and the Murray and Wright log-quadratic interpolation procedure (MTlog) [23].

The SGM iteration (39) is employed with the same splitting functionals $U(.)$ and $V(.)$ as in [38, equations (20)–(22)]. The algorithm is initialized with a uniform positive object and the convergence is checked using the following stopping rule ([40]):

$$\|\nabla_{\mathcal{P}} F(\boldsymbol{x}_k)\|_\infty < 10^{-3} \|\nabla_{\mathcal{P}} F(\boldsymbol{x}_0)\|_\infty , \tag{42}$$

where $\nabla_{\mathcal{P}} F(\boldsymbol{x})$ denotes the projected gradient of $F(.)$ at $\boldsymbol{x}$,

$$\nabla_{\mathcal{P}} F(\boldsymbol{x}) = \max(\boldsymbol{x} - \nabla F(\boldsymbol{x}), \boldsymbol{0}) - \boldsymbol{x}. \tag{43}$$

4.1.3. Results and discussion. We present a simulated example using data generated with Fessler's code available at http://www.eecs.umich.edu/∼fessler. For this simulation, we consider an image $\boldsymbol{x}^o$ of size $N = 128 \times 128$ pixels and $M = 30720$ pairs of detectors. 10^5 counts are considered in the Poisson degradation model (36). The regularization parameters (λ, δ) are set to $\lambda = 10^{-1}$, $\delta = 50$ to obtain the best result in terms of similarity between the simulated and the estimated images (in the sense of the quadratic error). The two images are illustrated in figure 2.

Table 3 summarizes the performance results in terms of iteration number and computation time in seconds on an Intel Core 2 CPU 6700, 3 GHz, 3 GB RAM. The same strategy as in [9, section 4.1.2] has been used for the implementation of the three linesearch methods, reducing the gradient computation counts to the descent algorithm outer iteration number. The design parameters are the number of sub-iterations J for the MM procedure, and the Wolfe condition constants (σ_1, σ_2) for the MTcubic and MTlog methods. For the two latter methods, we give the mean number J of sub-iterations that are necessary to fulfil the two Wolfe conditions.

Table 3. Comparison between MM, MTcubic and MTlog linesearch strategies for a PET reconstruction problem solved using the split-gradient algorithm, in terms of iteration number and time (in seconds) before convergence, considered in the sense of (42). As a comparison, the multiplicative split-gradient (i.e. $s_k = 1, \forall k \geqslant 0$) requires 788 iterations and 58 s to fulfil the stopping criterion.

SGM–MTcubic					SGM–MTlog					SGM–MM		
σ_1	σ_2	J	Iter.	Time	σ_1	σ_2	J	Iter.	Time	J	Iter.	Time
10^{-4}	0.5	5.9	353	66	10^{-4}	0.5	5.4	349	60	1	353	**44**
10^{-4}	0.9	4.3	356	56	10^{-4}	0.9	4.9	350	58	2	349	48
10^{-4}	0.99	2.5	389	54	10^{-4}	0.99	3.3	350	54	3	350	54
10^{-3}	0.99	2.5	389	57	10^{-3}	0.99	3.3	350	54	4	350	60
10^{-2}	0.99	2.5	389	57	10^{-2}	0.99	3.3	350	53	5	349	66
10^{-1}	0.99	2.5	389	56	10^{-1}	0.99	3.3	350	53	10	349	94

It can be noted that the split-gradient algorithm with MM linesearch (SGM-MM) requires about the same number of iterations than the MTcubic or MTlog approaches (SGM-MTcubic and SGM-MTlog), provided that the parameters (σ_1, σ_2) are appropriately chosen.

The effect of the Wolfe parameters (σ_1, σ_2) differs according to the interpolation strategy. For the cubic linesearch, a decrease of the first Wolfe parameter σ_1 accelerates the convergence rate, but at a price of a larger cost per iteration. In contrast, it appears that the number of iterations for SGM-MTlog remains stable toward the tuning (σ_1, σ_2), which shows that the use of the Murray and Wright log-quadratic interpolation enhances the performances of the MT linesearch.

In terms of time before convergence, the SGM algorithm performs better when the stepsize is obtained with the proposed MM search, because of smaller computation cost per sub-iteration. The MM strategy admits a unique tuning parameter, namely the sub-iteration number J, and it appears that the simplest choice $J = 1$ leads to the best results. This indicates that the best strategy corresponds to a rough minimization of $f(\alpha)$. Such a conclusion meets that of [28] in the context of quadratic MM linesearch.

4.2. NMR reconstruction

We consider a mono-dimensional NMR reconstruction problem [18]. The NMR decay $y(t)$ associated with a continuous distribution of relaxation constants $x(T)$ is described in terms of a Fredholm integral of the first kind:

$$y(t) = \int_{T_{\min}}^{T_{\max}} x(T)\, k(t, T)\, \mathrm{d}T, \tag{44}$$

with $k(t, T) = \exp\left\{-\frac{t}{T}\right\}$. In practice, the measured signal $\boldsymbol{y}$ is a set of discrete experimental noisy data points $y_m = y(t_m)$ modeled as

$$\boldsymbol{y} = \boldsymbol{K}\boldsymbol{x} + \boldsymbol{\epsilon}, \tag{45}$$

where $\boldsymbol{K}$ and $\boldsymbol{x}$ are discretized versions of $k(t, T)$ and $x(T)$ with dimensions $M \times N$ and $N \times 1$, and $\boldsymbol{\epsilon}$ is an additive noise assumed centered white Gaussian. Given $\boldsymbol{y}$, the aim is to determine $\boldsymbol{x} \geqslant 0$. This problem is equivalent to a numerical inversion of the Fredholm integral (44) and is known as very ill conditioned ([41]).

4.2.1. Objective function. In order to obtain a stabilized solution, an often used method minimizes the expression

$$F(\boldsymbol{x}) = S(\boldsymbol{x}) + \lambda R(\boldsymbol{x}), \tag{46}$$

under positivity constraints, where $S(.)$ is a fidelity to the data term

$$S(\boldsymbol{x}) = \tfrac{1}{2}\|\boldsymbol{K}\boldsymbol{x} - \boldsymbol{y}\|_2^2, \tag{47}$$

and $R(.)$ is an entropic regularization term, e.g., the Shannon entropy measure

$$R(\boldsymbol{x}) = \sum_{n=1}^{N} x_n \log x_n \tag{48}$$

Moreover, the positivity constraint is implicitly handled because of the barrier property of the entropy function.

4.2.2. Optimization strategy. The TN algorithm is employed for minimizing (46). The direction $\boldsymbol{d}_k$ is computed by approximately solving the Newton system $\nabla^2 F(\boldsymbol{x}_k)\boldsymbol{d} = -\boldsymbol{g}_k$ using preconditioned conjugate-gradient (PCG) iterations. We propose a preconditioning matrix $\boldsymbol{M}_k$ built as an approximation of the inverse Hessian of $F(.)$ at $\boldsymbol{x}_k$

$$\boldsymbol{M}_k = (\boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^T + \lambda \mathrm{Diag}(\boldsymbol{x}_k)^{-1})^{-1}, \tag{49}$$

where $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ is the truncated singular value decomposition of $\boldsymbol{K}$ with rank equal to 5, and $\boldsymbol{D} = \boldsymbol{\Sigma}^T\boldsymbol{\Sigma}$. The direction $\boldsymbol{d}_k$ being gradient related, the convergence of the TN algorithm with the proposed linesearch is established in theorem 2 under assumptions 1 and 2, by defining $L \equiv P$ and $R \equiv B$. The verification of assumption 1 is straightforward for $\boldsymbol{A}(\boldsymbol{x}) = \boldsymbol{K}^T\boldsymbol{K}$. The fulfilment of assumption 2 is more difficult to check since the level set $\mathcal{L}_0$ may contain an element $\boldsymbol{x}$ with zero components, contradicting the gradient Lipschitz assumption. In practice, we initialized the algorithm with $\boldsymbol{x}_0 > 0$ and we never noticed convergence issues in our practical tests. The extension of the convergence results under a weakened version of assumption 2 remains an open issue in our convergence analysis.

The algorithm is initialized with a uniform positive object and, following [6], the convergence is checked using

$$\|\nabla F(\boldsymbol{x}_k)\|_\infty \leqslant 10^{-9}(1 + |F(\boldsymbol{x}_k)|). \tag{50}$$

Following [42], the PCG iterations are stopped when

$$\|\nabla F(\boldsymbol{x}_k) + \nabla^2 F(\boldsymbol{x}_k)\boldsymbol{d}_k\| \leqslant 10^{-5}\|F(\boldsymbol{x}_k)\|. \tag{51}$$

We propose to compare the performances of the MM linesearch with both MTcubic and MTlog strategies.

4.2.3. Results and discussion. Let $\boldsymbol{x}(T)$ a distribution to estimate. We consider the resolution of (45) when data $\boldsymbol{y}$ are simulated from $\boldsymbol{x}(T)$ via the NMR model (45) over $M = 10\,000$ sampled times t_m, with a SNR of 25 dB (figure 3). The regularization parameter λ is set to $\lambda = 7.2 \times 10^{-4}$ to obtain the best result in terms of similarity between the simulated and the estimated spectra (in the sense of quadratic error). Table 4 summarizes the performance results in terms of the iteration number and computation time in seconds.

According to table 4, the TN algorithm with the MM linesearch performs better than with Wolfe-based strategies with their best settings for σ_1 and σ_2. Concerning the choice of the sub-iteration number, it appears that $J = 1$ leads again to the best results in terms of the computation time.

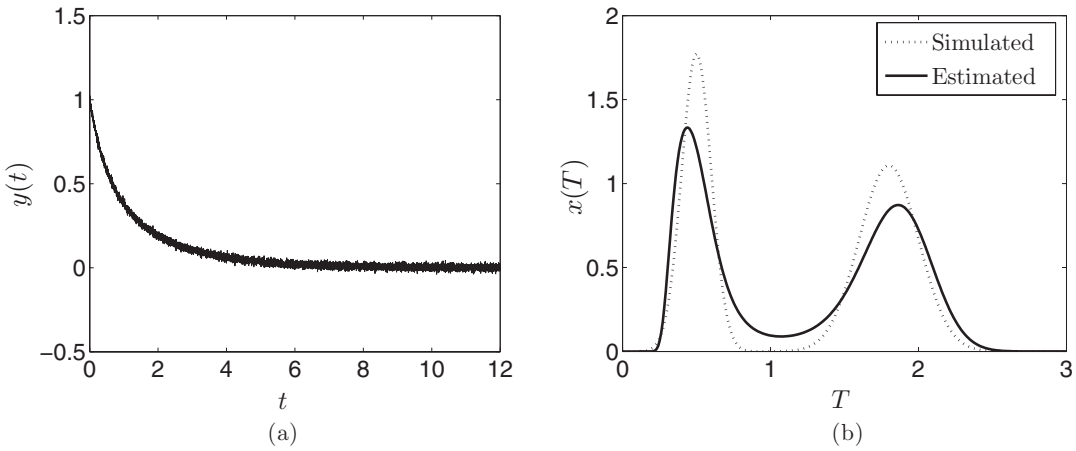


Figure 3. Simulated NMR reconstruction with the maximum entropy method. (a) Simulated NMR measurement (SNR: 25 dB). (b) NMR reconstruction (similarity error: 8.5%).

Table 4. Comparison between MM, MTcubic and MTlog linesearch strategies for a maximum entropy NMR reconstruction problem solved with TN algorithm, in terms of iteration number and time (in seconds) before convergence. Convergence is considered in the sense of (50).

TN–MTcubic					TN–MTlog					TN–MM		
σ_1	σ_2	J	Iter.	Time	σ_1	σ_2	J	Iter.	Time	J	Iter.	Time
10^{-4}	0.5	1.9	35	8	10^{-4}	0.5	2.6	35	10	1	36	**6**
10^{-4}	0.9	1.4	41	10	10^{-4}	0.9	2	35	9	2	40	7
10^{-4}	0.99	1	70	15	10^{-4}	0.99	2	35	9	3	40	7
10^{-3}	0.99	1	70	15	10^{-3}	0.99	2	35	9	4	40	7
10^{-2}	0.99	1	70	15	10^{-2}	0.99	2	35	9	5	40	7
10^{-1}	0.99	1	70	15	10^{-1}	0.99	2	35	9	10	40	8

5. Conclusion

This paper extends the linesearch strategy of [28] to the case of criteria containing barrier functions, by proposing a non-quadratic majorant approximation of the criterion in the linesearch direction. The proposed majorant has the same form as the interpolating function proposed in [23]. However, in the majorization approach, the construction of the approximation is easier and its minimization leads to a closed-form stepsize formula, which guarantees the convergence of several descent algorithms. Numerical experiments indicate that this linesearch strategy outperforms interpolating-based linesearch methods.

Two extensions of this work are envisaged. On the one hand, the analysis could be extended to additional forms of barrier functions, such as barriers for nonlinear constraints ([43]), roughness penalties ([22]) or inverse function ([44]). For the latter, the main difficulty will come from the fact that the inverse barrier grows faster than a logarithmic barrier near zero. Therefore, the proposed log-quadratic majorization will not be suited, and another form of majorant function should probably be envisaged.

On the other hand, the applicability of the proposed procedure for non-negative matrix factorization (NMF) could be studied. NMF is usually based on the minimization of a Bregman divergence between two unknown matrices [45]. Particular Bregman divergences such as Kullback–Leibler and Itakuro–Saito contain barrier terms that fall within the scope of the

present study. It would be interesting to analyze the performances of NMF iterative algorithms when the proposed linesearch is incorporated into the update schemes.

Appendix A. Proof of theorem 1

A.1. Majorizing property

Let $\boldsymbol{x}_k \in \mathcal{C}$, $\boldsymbol{d}_k \in \mathbb{R}^N$ a search direction and $\alpha^j \in (\alpha_-, \alpha_+)$. Let us show that $h(\alpha, \alpha^j)$ whose parameters $(m^j, \gamma^j, \bar{\alpha}^j)$ are given by (19) and (20) is a tangent majorant for $F(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k) = f(\alpha)$ at α^j, over (α_-, α_+).

First, according to assumption 1, $q(\alpha, \alpha^j) = p(\alpha^j) + (\alpha - \alpha^j)\dot{p}(\alpha^j) + \frac{1}{2} m_p^j (\alpha - \alpha^j)^2$ is a tangent majorant for $p(\alpha) = P(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k)$ at α^j for all $\alpha \in \mathbb{R}$. There remains to show that

$$\phi(\alpha, \alpha^j) = b(\alpha^j) + (\alpha - \alpha^j)\dot{b}(\alpha^j) + \frac{1}{2} m_b^j (\alpha - \alpha^j)^2 + \gamma^j \left[(\bar{\alpha}^j - \alpha^j) \log \frac{\bar{\alpha}^j - \alpha^j}{\bar{\alpha}^j - \alpha} + \alpha^j - \alpha \right], \tag{A.1}$$

with $m_b^j = m^j - m_p^j$ being a tangent majorant for $b(\alpha) = B(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k)$ at α^j. Let us define

$$b_1(\alpha) = \sum_{i \in \mathcal{I}_-} \psi_i(\theta_i + \alpha \delta_i), \qquad b_2(\alpha) = \sum_{i \in \mathcal{I}_+} \psi_i(\theta_i + \alpha \delta_i), \tag{A.2}$$

so that $b(\alpha) = b_1(\alpha) + b_2(\alpha) + b_0$, where b_0 is constant with respect to α. First, we will prove that

$$\begin{cases} \phi_1(\alpha, \alpha^j) = b_1(\alpha^j) + (\alpha - \alpha^j)\dot{b}_1(\alpha^j) + \frac{1}{2} m_b^j (\alpha - \alpha^j)^2 \\ \phi_2(\alpha, \alpha^j) = b_2(\alpha^j) + (\alpha - \alpha^j)\dot{b}_2(\alpha^j) + \gamma^j \left[(\alpha_+ - \alpha^j) \log \frac{\alpha_+ - \alpha^j}{\alpha_+ - \alpha} + \alpha^j - \alpha \right], \end{cases}$$

respectively, majorize b_1 and b_2 for all $\alpha \in [\alpha^j, \alpha_+)$.

Let us assume that $\mathcal{I}_-$ is not empty. Then, according to the expression of b_1, $Z_1(\alpha) = \ddot{b}_1(\alpha)$ so $m_b^j = \ddot{b}_1(\alpha^j)$. The strict convexity of functions ψ_i, for all $i \in \{1, \ldots, I\}$ implies that b_1 is strictly convex and $\dot{b}_1$ is strictly concave. Then, for all $\alpha \in [\alpha_j, \alpha^+)$, $\ddot{b}_1(\alpha) \leqslant \ddot{b}_1(\alpha^j) = m_b^j$. Hence, $\phi_1(., \alpha^j)$ majorizes b_1 on $[\alpha_j, \alpha^+)$. If $\mathcal{I}_-$ is empty, both $b_1(.)$ and $\phi_1(., \alpha^j)$ equal zero so the latter majorizing property still holds.

Let us assume that $\mathcal{I}_+$ is not empty. The expression of b_2 leads to $Z_2(\alpha) = \ddot{b}_2(\alpha)$ so $\gamma^j = (\alpha_+ - \alpha^j)\ddot{b}_2(\alpha^j)$. Let us define $T(\alpha) = \dot{b}_2(\alpha)(\alpha_+ - \alpha)$ and $l(\alpha) = \dot{b}_2(\alpha^j)(\alpha_+ - \alpha) + \gamma^j(\alpha - \alpha^j)$. Given $\gamma^j = (\alpha_+ - \alpha^j)\ddot{b}_2(\alpha^j)$, the linear function l also reads

$$l(\alpha) = \dot{\phi}_2(\alpha, \alpha^j)(\alpha_+ - \alpha). \tag{A.3}$$

Thus, we have $l(\alpha^j) = T(\alpha^j)$ and $\dot{l}(\alpha^j) = \dot{T}(\alpha^j)$. Moreover,

$$\ddot{T}(\alpha) = \dddot{b}_2(\alpha)(\alpha_+ - \alpha) - 2\ddot{b}_2(\alpha) = \sum_{i \in \mathcal{I}_+} \delta_i^3 \dddot{\psi}_i(\theta_i + \alpha \delta_i)(\alpha_+ - \alpha) - 2\delta_i^2 \ddot{\psi}_i(\theta_i + \alpha \delta_i). \tag{A.4}$$

According to the definition of α_+,

$$\alpha_+ - \alpha < -\frac{\theta_i + \alpha \delta_i}{\delta_i}, \quad \forall i \in \mathcal{I}_+. \tag{A.5}$$

According to (7), the third derivative of ψ_i is negative, so

$$\ddot{T}(\alpha) < \sum_{i \in \mathcal{I}_+} \delta_i^2 [-\dddot{\psi}_i(\theta_i + \alpha \delta_i)(\theta_i + \alpha \delta_i) - 2\ddot{\psi}_i(\theta_i + \alpha \delta_i)] < 0, \tag{A.6}$$

where the last inequality is a consequence of (7). Thus, T is concave. Since l is a linear function tangent to T, we have

$$l(\alpha) \geqslant T(\alpha), \quad \forall \alpha \in [\alpha_j, \alpha^+). \tag{A.7}$$

Given $\alpha_+ > \alpha$, (A.7) also reads

$$\dot{\phi}_2(\alpha, \alpha^j) \geqslant \dot{b}_2(\alpha), \quad \forall \alpha \in [\alpha_j, \alpha^+), \tag{A.8}$$

so $\phi_2(., \alpha^j)$ majorizes b_2 over $[\alpha_j, \alpha^+)$. This property still holds if $\mathcal{I}_+$ is empty, since both $b_2(.)$ and $\phi_2(., \alpha^j)$ equal zero in that case. Finally, $\phi(., \alpha^j) = \phi_1(., \alpha^j) + \phi_2(., \alpha^j)$ majorizes b for $\alpha \geqslant \alpha_j$. The same elements of proof apply to the case $\alpha \leqslant \alpha^j$. We can thus conclude that $h(\alpha, \alpha^j) = q(\alpha, \alpha^j) + \phi(\alpha, \alpha^j)$ is a tangent majorant for f at α^j.

Appendix B. Proof of lemma 1

First, $h(., \alpha^j)$ is C^∞ over (α_-, α^j) and (α^j, α_+). Moreover, it is easy to check that h and its first two derivatives are continuous at α^j according to (19)–(20). Then, $h(., \alpha^j)$ is C^2 over (α_-, α_+). On the other hand, (19)–(20) imply, for all $\alpha \in (\alpha_-, \alpha^j]$,

$$\ddot{h}(\alpha, \alpha^j) = \begin{cases} m_p^j + Z_2(\alpha^j) + Z_1(\alpha^j) \dfrac{(\alpha_- - \alpha^j)^2}{(\alpha_- - \alpha)^2} & \text{if} \quad \mathcal{I}_- \neq \emptyset \\ m_p^j + Z_2(\alpha^j) & \text{otherwise,} \end{cases} \tag{B.1}$$

and for all $\alpha \in [\alpha^j, \alpha_+)$,

$$\ddot{h}(\alpha, \alpha^j) = \begin{cases} m_p^j + Z_1(\alpha^j) + Z_2(\alpha^j) \dfrac{(\alpha_+ - \alpha^j)^2}{(\alpha_+ - \alpha)^2} & \text{if} \quad \mathcal{I}_+ \neq \emptyset \\ m_p^j + Z_1(\alpha^j) & \text{otherwise,} \end{cases} \tag{B.2}$$

$Z_1(\cdot)$ and $Z_2(\cdot)$ are positive since ψ_i is strictly convex for all $i \in \{1, \ldots, I\}$. Moreover, $m_p^j > 0$ according to assumption 1. Thus, $h(., \alpha^j)$ is strictly convex.

Appendix C. Proof of property 1

Let us consider $\boldsymbol{x} \in \mathcal{V}_0$ and $\boldsymbol{d}$ a descent direction. First, we establish some preliminary results arising from the expression of the majorant function. Then, some lower and upper bounds for the stepsize values are derived. Finally, property 1 is proved.

C.1. Preliminary results

Lemma 3. *Let $j \in \{0, \ldots, J-1\}$. If $\dot{f}(\alpha^j) \leqslant 0$ and $|\bar{\alpha}^j| < \infty$, then α^{j+1} fulfils*

$$-\frac{q_3}{q_2} \leqslant \alpha^{j+1} - \alpha^j \leqslant -\frac{2q_3}{q_2}, \tag{C.1}$$

where q_1, q_2 and q_3 are given by (24).

Proof. If $\dot{f}(\alpha^j) \leqslant 0$ and $|\bar{\alpha}^j| < \infty$, then α^{j+1} reads

$$\alpha^{j+1} = \alpha^j - \frac{2q_3}{q_2 + \sqrt{q_2^2 - 4q_1 q_3}}, \tag{C.2}$$

with

$$\begin{cases} q_1 = -m^j, \\ q_2 = \gamma^j - \dot{f}(\alpha^j) + m^j(\bar{\alpha}^j - \alpha^j), \\ q_3 = (\bar{\alpha}^j - \alpha^j)\dot{f}(\alpha^j), \end{cases} \tag{C.3}$$

where parameters $(m^j, \gamma^j, \bar{\alpha}^j)$ are given by

$$\begin{cases} \bar{\alpha}^j = \alpha_+, \\ m^j = m_p^j + \sum\limits_{i\in\mathcal{I}_-} \phi_i(\alpha^j), \\ \gamma^j = (\alpha_+ - \alpha^j) \sum\limits_{i\in\mathcal{I}_+} \phi_i(\alpha^j), \end{cases} \tag{C.4}$$

with $\phi_i(\alpha) = \delta_i^2 \ddot{\psi}_i(\theta_i + \alpha\delta_i)$ being a positive function. Therefore, we have $q_1 < 0$, $q_3 < 0$ and $q_2 > 0$, which yield (C.1). □

Lemma 4. *Let $j \in \{0, \ldots, J-1\}$. If $\dot{f}(\alpha^j) \leqslant 0$, then*

$$f(\alpha^j) - f(\alpha^{j+1}) + \tfrac{1}{2}(\alpha^{j+1} - \alpha^j)\dot{f}(\alpha^j) \geqslant 0. \tag{C.5}$$

Proof. The property is trivial if $\dot{f}(\alpha^j) = 0$. Let us assume that $\dot{f}(\alpha^j) < 0$ so that $\alpha_+ > \alpha^{j+1} > \alpha^j$. Let us define

$$\tau(\alpha) = h(\alpha, \alpha^j) - (f(\alpha^j) + (\alpha - \alpha^j)\dot{f}(\alpha^j)). \tag{C.6}$$

If $\mathcal{I}_+$ is not empty, $\tau(\alpha) = Q(\alpha) + \gamma^j(\alpha_+ - \alpha^j)\varphi(\alpha)$ with $Q(\alpha) = \frac{1}{2}m^j(\alpha - \alpha^j)^2$ and $\varphi(\alpha) = \xi\left(\frac{\alpha-\alpha^j}{\alpha_+-\alpha^j}\right)$, where $\xi(u) = -\log(1-u) - u$ for all $u \in (0,1)$. A straightforward analysis shows that

$$\frac{\xi(u)}{u\dot{\xi}(u)} \leqslant \frac{1}{2}, \quad \forall u \in (0,1). \tag{C.7}$$

Taking $u = \frac{\alpha-\alpha^j}{\alpha_+-\alpha^j}$ in (C.7) leads to

$$\frac{\varphi(\alpha)}{(\alpha-\alpha^j)\dot{\varphi}(\alpha)} \leqslant \frac{1}{2}, \quad \forall \alpha \in (\alpha^j, \alpha_+). \tag{C.8}$$

Furthermore, according to the expression of $Q(\alpha)$, we have

$$Q(\alpha) = \tfrac{1}{2}(\alpha - \alpha^j)\dot{Q}(\alpha). \tag{C.9}$$

Thus, using (C.8) and (C.9),

$$\frac{\tau(\alpha)}{(\alpha-\alpha^j)\dot{\tau}(\alpha)} \leqslant \frac{1}{2}, \quad \forall \alpha \in (\alpha^j, \alpha_+). \tag{C.10}$$

If $\mathcal{I}_+$ is empty, $\tau(\alpha) = Q(\alpha)$ so (C.10) still holds, according to (C.9). $h(., \alpha^j)$ is a tangent majorant for f so

$$h(\alpha, \alpha^j) - f(\alpha) = f(\alpha^j) - f(\alpha) + (\alpha - \alpha^j)\dot{f}(\alpha^j) + \tau(\alpha) \geqslant 0. \tag{C.11}$$

Taking $\alpha = \alpha^{j+1} > \alpha^j$ in (C.10) and (C.11), we obtain

$$f(\alpha^j) - f(\alpha^{j+1}) + (\alpha^{j+1} - \alpha^j)\dot{f}(\alpha^j) + \tfrac{1}{2}(\alpha^{j+1} - \alpha^j)\dot{\tau}(\alpha^{j+1}) \geqslant 0. \tag{C.12}$$

Finally, the result holds since $\dot{\tau}(\alpha^{j+1}) = \dot{h}(\alpha^{j+1}, \alpha^j) - \dot{f}(\alpha^j) = -\dot{f}(\alpha^j)$. □

Lemma 5. *Let $j \in \{0, \ldots, J-1\}$. Under assumptions 1 and 2, there exist $\nu_{\min}$, $\nu_{\max}$, $0 < \nu_{\min} \leqslant \nu_{\max}$, such that for all $\boldsymbol{x} \in \mathcal{V}_0$ and for all descent direction $\boldsymbol{d}$ at $\boldsymbol{x}$:*

$$\nu_{\min}\|\boldsymbol{d}\|^2 \leqslant \ddot{h}(\alpha^j, \alpha^j) \leqslant \nu_{\max}\|\boldsymbol{d}\|^2, \quad \forall j \geqslant 0. \tag{C.13}$$

Proof. Let us first remark that, according to assumption 2, there exists $\eta > 0$ such that

$$\|\nabla F(\boldsymbol{x})\| \leqslant \eta, \quad \forall \boldsymbol{x} \in \mathcal{V}_0. \tag{C.14}$$

Moreover, because the gradient of B is unbounded at the boundary of $\mathcal{C}$, (C.14) leads to the existence of $C_{\min} > 0$ such that

$$C_i(\boldsymbol{x}) \geqslant C_{\min}, \quad \forall \boldsymbol{x} \in \mathcal{V}_0, \ \forall i \in \{1, \ldots, I\}, \tag{C.15}$$

and the boundedness assumption on $\mathcal{V}_0$ implies that there exists $C_{\max} > 0$ such that

$$C_i(\boldsymbol{x}) \leqslant C_{\max}, \quad \forall \boldsymbol{x} \in \mathcal{V}_0, \ \forall i \in \{1, \ldots, I\}. \tag{C.16}$$

According to lemma 1,

$$\ddot{h}(\alpha^j, \alpha^j) = m_p^j + \ddot{b}(\alpha^j), \tag{C.17}$$

where $b(\alpha) = B(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k)$, so that

$$\ddot{b}(\alpha^j) = \boldsymbol{d}^T \nabla^2 B(\boldsymbol{x} + \alpha^j \boldsymbol{d}) \boldsymbol{d}. \tag{C.18}$$

On the other hand,

$$m_p^j = \boldsymbol{d}^T \boldsymbol{A}(\boldsymbol{x} + \alpha^j \boldsymbol{d}) \boldsymbol{d}. \tag{C.19}$$

Since $\boldsymbol{x} + \alpha^j \boldsymbol{d} \in \mathcal{V}_0$, it is sufficient to show that the set $\{\boldsymbol{A}(\boldsymbol{x}) + \nabla^2 B(\boldsymbol{x}) | \boldsymbol{x} \in \mathcal{V}_0\}$ has a positive bounded spectrum. According to (6),

$$\nabla^2 B(\boldsymbol{x}) = \boldsymbol{C}^T \mathrm{Diag}(\ddot{\psi}_i(C_i(\boldsymbol{x}))) \boldsymbol{C}, \tag{C.20}$$

with $\boldsymbol{C} = [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_I]^T$. According to (7), for all $i \in \{1, \ldots, I\}$, $\ddot{\psi}_i$ is decreasing on $\mathbb{R}^+$. Therefore, (C.15) and (C.16) yield

$$\boldsymbol{d}^T \boldsymbol{H}(C_{\max}) \boldsymbol{d} \leqslant \boldsymbol{d}^T \nabla^2 B(\boldsymbol{x}) \boldsymbol{d} \leqslant \boldsymbol{d}^T \boldsymbol{H}(C_{\min}) \boldsymbol{d}, \quad \forall \boldsymbol{x} \in \mathcal{V}_0, \tag{C.21}$$

with $\boldsymbol{H}(c) = \boldsymbol{C}^T \mathrm{Diag}(\ddot{\psi}_i(c)) \boldsymbol{C}$. Since $\ddot{\psi}_i$ is strictly convex, matrix $\boldsymbol{H}(c)$ is symmetric and has a non-negative bounded spectrum with bounds $(\nu_{\min}^{\mathcal{H}}(c), \nu_{\max}^{\mathcal{H}}(c))$. Moreover, according to assumption 1, $\boldsymbol{A}(\boldsymbol{x})$ has a positive bounded spectrum with bounds $(\nu_{\min}^{\mathcal{A}}, \nu_{\max}^{\mathcal{A}})$ on $\mathcal{V}_0$. Thus, lemma 5 holds with $\nu_{\min} = \nu_{\min}^{\mathcal{A}} + \nu_{\min}^{\mathcal{H}}(C_{\max}) > 0$ and $\nu_{\max} = \nu_{\max}^{\mathcal{A}} + \nu_{\max}^{\mathcal{H}}(C_{\min})$. □

C.2. Upper and lower bounds for the stepsize series

Lemma 6. *Under assumptions 1 and 2, there exist $\nu, \nu' > 0$ such that for all $\boldsymbol{x} \in \mathcal{V}_0$ and for all descent direction $\boldsymbol{d}$ at $\boldsymbol{x}$,*

$$\frac{-\boldsymbol{g}^T \boldsymbol{d}}{\nu \|\boldsymbol{d}\|^2} \leqslant \alpha^1 \leqslant \frac{-\boldsymbol{g}^T \boldsymbol{d}}{\nu' \|\boldsymbol{d}\|^2}, \tag{C.22}$$

where $\boldsymbol{g}$ denotes the gradient of $F(.)$ at $\boldsymbol{x}$.

Proof. $\boldsymbol{d}$ is a descent direction, so $\dot{f}(0) < 0$ and $h(., 0)$ has a barrier at $\bar{\alpha}^0 = \alpha_+$.

If $\alpha_+ = +\infty$ then $h(., 0)$ is a quadratic function with curvature m^0. This majorant is minimized at $\alpha^1 = \frac{-\dot{f}(0)}{m^0}$ and according to lemma 5, we have

$$\frac{-\boldsymbol{g}^T \boldsymbol{d}}{\nu_{\max} \|\boldsymbol{d}\|^2} \leqslant \alpha^1 \leqslant \frac{-\boldsymbol{g}^T \boldsymbol{d}}{\nu_{\min} \|\boldsymbol{d}\|^2}. \tag{C.23}$$

If $\alpha_+ < +\infty$, according to lemma 3

$$\frac{-\boldsymbol{g}^T \boldsymbol{d}}{\frac{\gamma^0}{\alpha_+} - \frac{\boldsymbol{g}^T \boldsymbol{d}}{\alpha_+} + m^0} \leqslant \alpha^1 \leqslant \frac{-2\boldsymbol{g}^T \boldsymbol{d}}{\frac{\gamma^0}{\alpha_+} - \frac{\boldsymbol{g}^T \boldsymbol{d}}{\alpha_+} + m^0}. \tag{C.24}$$

Using lemma 5 and the positivity of $-\boldsymbol{g}^T\boldsymbol{d}$, we obtain

$$\nu_{\min}\|\boldsymbol{d}\|^2 \leqslant \frac{\gamma^0}{\alpha_+} - \frac{\boldsymbol{g}^T\boldsymbol{d}}{\alpha_+} + m^0. \tag{C.25}$$

On the other hand, taking $\iota = \operatorname{argmax}_{i\in\{1,\ldots,I\}} -\boldsymbol{c}_i^T\boldsymbol{d}$, we deduce from (C.15) that

$$\alpha^+ \geqslant \frac{C_{\min}}{|\boldsymbol{c}_\iota^T\boldsymbol{d}|}. \tag{C.26}$$

Thus, using Cauchy–Schwartz inequality and (C.14),

$$\frac{-\boldsymbol{g}^T\boldsymbol{d}}{\alpha_+} = \frac{|\boldsymbol{g}^T\boldsymbol{d}|}{\alpha_+} \leqslant |\boldsymbol{g}^T\boldsymbol{d}||\boldsymbol{c}_\iota^T\boldsymbol{d}|\frac{1}{C_{\min}} \leqslant \|\boldsymbol{g}\|\|\boldsymbol{c}_\iota\|\|\boldsymbol{d}\|^2\frac{1}{C_{\min}} \leqslant \frac{\eta\bar{C}}{C_{\min}}\|\boldsymbol{d}\|^2, \tag{C.27}$$

with $\bar{C} = \max_{i\in\{1,\ldots,I\}}\|\boldsymbol{c}_i\| > 0$. Moreover, according to lemma 5, there exists $\nu_{\max}$ such that

$$m^0 + \frac{\gamma^0}{\alpha_+} \leqslant \nu_{\max}\|\boldsymbol{d}\|^2. \tag{C.28}$$

Therefore, (C.25)–(C.28) allow us to check that lemma 6 holds for $\nu = \nu_{\max} + \frac{\eta\bar{C}}{C_{\min}}$ and $\nu' = \frac{\nu_{\min}}{2}$. □

Property 5. *Under assumptions 1 and 2, for all* $j \in \{1,\ldots,J\}$,

$$\alpha^j \leqslant \sigma_{\max}^j\alpha^1, \tag{C.29}$$

where

$$\sigma_{\max}^j = \left(1 + \frac{2\nu_{\max}L}{\nu_{\min}^2}\right)^{j-1}\left(1 + \frac{\nu}{L}\right) - \frac{\nu}{L} \geqslant 1. \tag{C.30}$$

Proof. It is easy to check (C.29) for $j = 1$, with $\sigma_{\max}^1 = 1$. Let us prove that (C.29) holds for $j > 1$. Assume that $\dot{f}(\alpha^j) < 0$. Then, $\bar{\alpha}^j = \alpha_+$. Let us denote

$$\hat{m}^j = \begin{cases} m^j + \dfrac{\gamma^j}{\alpha_+ - \alpha^j} & \text{if} \quad \mathcal{I}_+ \neq \emptyset, \\ m^j & \text{otherwise.} \end{cases} \tag{C.31}$$

If $\mathcal{I}_+$ is not empty, i.e. $\alpha_+ < +\infty$, we deduce from lemma 3 that

$$\alpha^{j+1} - \alpha^j \leqslant \frac{-2\dot{f}(\alpha^j)}{\frac{\gamma^j - \dot{f}(\alpha^j)}{\alpha_+ - \alpha^j} + m^j}. \tag{C.32}$$

Since $\dot{f}(\alpha^j)$ is negative,

$$\alpha^{j+1} - \alpha^j \leqslant \frac{-2\dot{f}(\alpha^j)}{\hat{m}^j}. \tag{C.33}$$

If $\mathcal{I}_+$ is empty, then $\alpha^{j+1} - \alpha^j = \frac{-\dot{f}(\alpha^j)}{m^j}$, so (C.33) also holds. According to lemma 5,

$$\|\boldsymbol{d}\|^2 \geqslant \frac{\hat{m}^0}{\nu_{\max}}, \tag{C.34}$$

and

$$\hat{m}^j \geqslant \nu_{\min}\|\boldsymbol{d}\|^2, \tag{C.35}$$

thus, we have

$$\hat{m}^j \geqslant (\hat{m}^0)\frac{\nu_{\min}}{\nu_{\max}} > 0. \tag{C.36}$$

Then, from (C.33)

$$\alpha^{j+1} \leqslant \alpha^j + \frac{2|\dot{f}(\alpha^j)|}{\hat{m}^0} \frac{\nu_{\max}}{\nu_{\min}}. \tag{C.37}$$

If $\dot{f}(\alpha^j) \geqslant 0$, α^{j+1} is lower than α^j so (C.37) still holds. According to assumption 2, ∇F is Lipschitz, so that $|\dot{f}(\alpha^j) - \dot{f}(0)| \leqslant L\|\boldsymbol{d}\|^2\alpha^j$. Using the fact that $|\dot{f}(\alpha^j)| \leqslant |\dot{f}(\alpha^j) - \dot{f}(0)| + |\dot{f}(0)|$, and $\dot{f}(0) < 0$, we obtain

$$|\dot{f}(\alpha^j)| \leqslant L\alpha^j\|\boldsymbol{d}\|^2 - \dot{f}(0). \tag{C.38}$$

Using lemma 6 and (C.34),

$$-\dot{f}(0) \leqslant \alpha^1 \nu \|\boldsymbol{d}\|^2 \leqslant \alpha^1 \frac{\nu}{\nu_{\min}} \hat{m}^0. \tag{C.39}$$

Given (C.34)–(C.39), we obtain

$$\alpha^{j+1} \leqslant \alpha^j + \frac{2\nu_{\max}}{\nu_{\min}} \frac{1}{\hat{m}^0} \left[L\alpha^j \left(\frac{\hat{m}^0}{\nu_{\min}} \right) + \alpha^1 \frac{\nu}{\nu_{\min}} \hat{m}^0 \right]. \tag{C.40}$$

Hence,

$$\alpha^{j+1} \leqslant \alpha^j \left(1 + \frac{2\nu_{\max} L}{\nu_{\min}^2} \right) + 2\alpha^1 \frac{\nu_{\max}\nu}{\nu_{\min}^2}. \tag{C.41}$$

This corresponds to a recursive definition of the series $(\sigma_{\max}^j)$ with

$$\sigma_{\max}^{j+1} = \sigma_{\max}^j \left(1 + 2\frac{\nu_{\max} L}{\nu_{\min}^2} \right) + 2\frac{\nu\nu_{\max}}{\nu_{\min}^2}. \tag{C.42}$$

Given $\sigma_{\max}^1 = 1$, (C.30) is the general term of the series. □

C.3. First Wolfe condition

First, for $j = 1$, the first Wolfe condition (28) holds according to lemma 4, since it identifies with (C.5) when $j = 0$, given $\sigma_{\max}^1 = 1$. For all $j > 1$, (28) holds by immediate recurrence, given property 5, hence the result.

Appendix D. Proof of property 2

First, let us show that (29) holds for all $j \geqslant 1$ with

$$\sigma_{\min} = \frac{\sqrt{1 + 2\frac{L}{\nu_{\min}}} - 1}{2\frac{L}{\nu_{\min}}} \in \left(0, \frac{1}{2}\right). \tag{D.1}$$

Let ϕ be the concave quadratic function $\phi(\alpha) = f(0) + \alpha\dot{f}(0) + m\frac{\alpha^2}{2}$, with $m = -\frac{L}{\nu_{\min}}\hat{m}^0$, where $\hat{m}^0$ is defined in (C.31). We have $\phi(0) = f(0)$ and $\dot{\phi}(0) = \dot{f}(0) < 0$, so ϕ is decreasing on $\mathbb{R}^+$. Let us consider $\alpha \in [0, \alpha^j]$, so that $\boldsymbol{x} + \alpha\boldsymbol{d} \in \mathcal{V}_0$. According to assumption 2, we have $|\dot{f}(\alpha) - \dot{f}(0)| \leqslant \|\boldsymbol{d}\|^2 L|\alpha|$, and according to lemma 5,

$$|\dot{f}(\alpha) - \dot{f}(0)| \leqslant \frac{L\alpha}{\nu_{\min}} \hat{m}^0. \tag{D.2}$$

Then we obtain

$$|\dot{f}(\alpha)| \leqslant \frac{L\alpha}{\nu_{\min}} \hat{m}^0 - \dot{f}(0). \tag{D.3}$$

Hence,

$$\dot{\phi}(\alpha) \leqslant \dot{f}(\alpha), \quad \forall \alpha \in [0, \alpha^j]. \tag{D.4}$$

Integrating (D.4) between 0 and α^j yields

$$\phi(\alpha^j) \leqslant f(\alpha^j). \tag{D.5}$$

On the other hand, the expression of ϕ at $\alpha_{\min} = \sigma_{\min}\alpha^1$ reads $\phi(\alpha_{\min}) = f(0) + S\alpha^1\dot{f}(0)$, where

$$S = \sigma_{\min} - \sigma_{\min}^2 L\alpha^1 \frac{\hat{m}^0}{2\dot{f}(0)\nu_{\min}}. \tag{D.6}$$

According to (C.33),

$$\alpha^1 \leqslant \frac{-2\dot{f}(0)}{\hat{m}^0}, \tag{D.7}$$

so that

$$S \leqslant \sigma_{\min} + \sigma_{\min}^2 \frac{L}{\nu_{\min}} = \frac{1}{2}, \tag{D.8}$$

where the latter equality directly stems from the expression of $\sigma_{\min}$. Since ϕ is decreasing on $\mathbb{R}^+$, we obtain

$$\phi(\alpha_{\min}) \geqslant f(0) + \tfrac{1}{2}\alpha^1\dot{f}(0) \geqslant f(\alpha^1), \tag{D.9}$$

where the last inequality is the first Wolfe condition (28) for $j = 1$.

Finally, $\alpha^j > 0$ for all $j \geqslant 1$. Assume that there exists j such that $\alpha^j < \alpha_{\min}$. According to (D.5) and given that ϕ is decreasing on $\mathbb{R}^+$, we obtain

$$f(\alpha^j) \geqslant \phi(\alpha^j) > \phi(\alpha_{\min}) \geqslant f(\alpha^1), \tag{D.10}$$

which contradicts the fact that $f(\alpha^j)$ is nonincreasing. Thus, (29) holds. So does (30), according to lemma 6.

Appendix E. Proof of property 3

Let us first remark that for all k, $\boldsymbol{d}_k \neq \mathbf{0}$, since $\boldsymbol{g}_k^T\boldsymbol{d}_k < 0$. According to property 1, the first Wolfe condition holds for $\sigma_1 = \sigma_1^J$:

$$F(\boldsymbol{x}_k) - F(\boldsymbol{x}_{k+1}) \geqslant -\sigma_1^J\alpha_k\boldsymbol{g}_k^T\boldsymbol{d}_k. \tag{E.1}$$

According to property 2,

$$\alpha_k \geqslant \sigma_{\min}\frac{-\boldsymbol{g}_k^T\boldsymbol{d}_k}{\nu\|\boldsymbol{d}_k\|^2}, \tag{E.2}$$

so

$$0 \leqslant \sigma_0\frac{(\boldsymbol{g}_k^T\boldsymbol{d}_k)^2}{\|\boldsymbol{d}_k\|^2} \leqslant F(\boldsymbol{x}_k) - F(\boldsymbol{x}_{k+1}), \tag{E.3}$$

with $\sigma_0 = \frac{1}{\nu}\sigma_{\min}\sigma_1^J > 0$. According to assumption 2, the level set $\mathcal{L}_0$ is bounded, so $\lim_{k\to\infty} F(\boldsymbol{x}_k)$ is finite. Therefore,

$$\sum_{k=0}^{\infty}\frac{(\boldsymbol{g}_k^T\boldsymbol{d}_k)^2}{\|\boldsymbol{d}_k\|^2} \leqslant \frac{1}{\sigma_0}\left[F(\boldsymbol{x}_0) - \lim_{k\to\infty} F(\boldsymbol{x}_k)\right] < \infty. \tag{E.4}$$

References

[1] Bertero M and Boccacci P 1998 *Introduction to Inverse Problems in Imaging* (Bristol: Institute of Physics Publishing)

[2] Boerner W M, Brand H, Cram L A, Jordan A K, Keydel W D T G, Schwierz G and Vogel M 1985 *Inverse Methods in Electromagnetic Imaging: part I* (*NATO ASI Series* vol 143) (Reidel: Dordrecht)

[3] Demoment G 1989 Image reconstruction and restoration: overview of common estimation structure and problems *IEEE Trans. Acoust. Speech Signal Process.* **37** 2024–36

[4] Tikhonov A N and Arsenin V Y 1977 *Solutions of Ill-Posed Problems* (New York: Wiley)

[5] Moré J J and Thuente D J 1994 Line search algorithms with guaranteed sufficient decrease *ACM Trans. Math. Softw.* **20** 286–307

[6] Nocedal J and Wright S J 1999 *Numerical Optimization* (New York: Springer)

[7] Wright M H 1992 Interior methods for constrained optimization *Acta Numerica* (Cambridge: Cambridge University Press) pp 341–407

[8] Kim S-J, Koh K, Lustig M, Boyd S and Gorinevsky D 2007 An interior-point method for large-scale ℓ_1-regularized least-squares *IEEE J. Sel. Top. Signal Process.* **1** 606–17

[9] Johnson C A and Sofer A 2000 A primal-dual method for large-scale image reconstruction in emission tomography *SIAM J. Optim.* **11** 691–715

[10] Ollinger J M and Fessler J A 1997 Positron-emission tomography *IEEE Signal Process. Mag.* **14** 43–55

[11] Johnson C A and McGarry D 2003 Maximum entropy reconstruction methods in electron paramagnetic resonance imaging *Ann. Oper. Res.* **119** 101–18

[12] Bertero M, Boccacci P, Desidera G and Vicidomini G 2009 Image deblurring with Poisson data: from cells to galaxies *Inverse Problems* **25** 123006

[13] Cao Y, Eggermont P P B and Terebey S 1999 Cross Burg entropy maximization and its application to ringing suppression in image reconstruction *IEEE Trans. Image Process.* **8** 286–92

[14] Aubert G and Aujol J-F 2008 A variational approach to removing multiplicative noise *SIAM J. Appl. Math.* **68** 925–46

[15] Hsiao I T, Rangarajan A and Gindi G 2002 Joint-MAP Bayesian tomographic reconstruction with a gamma-mixture prior *IEEE Trans. Image Process.* **11** 1466–77

[16] Le Besnerais G, Bercher J-F and Demoment G 1999 A new look at entropy for solving linear inverse problems *IEEE Trans. Inform. Theory* **45** 1565–78

[17] Mohammad-Djafari A 1994 Maximum d'entropie et problèmes inverses en imagerie *Trait. Signal* **2** 87–116

[18] Gull S F and Skilling J 1984 Maximum entropy method in image processing *IEEE Proc.* F **131** 646–59

[19] Nityananda R and Narayan R 1982 Maximum entropy image reconstruction—a practical non-information-theoretic approach *J. Astrophys. Astron.* **3** 419–50

[20] Byrne C L 1993 Iterative image reconstruction algorithms based on cross-entropy minimization *IEEE Trans. Image Process.* **2** 96–103

[21] Narayanan M V, Byrne C L and King M A 2001 An interior point iterative maximum-likelihood reconstruction algorithm incorporating upper and lower bounds with application to SPECT transmission imaging *IEEE Trans. Med. Imaging* **20** 342–53

[22] O'Sullivan J A 1995 Roughness penalties on finite domains *IEEE Trans. Image Process.* **4** 1258–68

[23] Murray W and Wright M H 1994 Line search procedures for the logarithmic barrier function *SIAM J. Optim.* **4** 229–46

[24] Lin C H 2002 A null-space primal-dual algorithm for nonlinear network optimization *PhD Thesis* University of Stanford

[25] Nash S G and Sofer A 1993 A barrier method for large-scale constrained optimization *ORSA J. Comput.* **5** 40–53

[26] Fessler J A and Booth S D 1999 Conjugate-gradient preconditioning methods for shift-variant PET image reconstruction *IEEE Trans. Image Process.* **8** 688–99

[27] Sun J and Zhang J 2001 Global convergence of conjugate gradient methods without line search *Ann. Oper. Res.* **103** 161–73

[28] Labat C and Idier J 2008 Convergence of conjugate gradient methods with a closed-form stepsize formula *J. Optim. Theory Appl.* **136** 43–60

[29] Labat C and Idier J 2007 Convergence of truncated half-quadratic and Newton algorithms, with application to image restoration *Technical Report* IRCCyN www.irccyn.ec-nantes.fr/∼idier/pub/labat07b.pdf

[30] Allain M, Idier J and Goussard Y 2006 On global and local convergence of half-quadratic algorithms *IEEE Trans. Image Process.* **15** 1130–42

[31] Hunter D R and Lange K 2004 A tutorial on MM algorithms *Am. Stat.* **58** 30–7

[32] Jacobson M W and Fessler J A 2007 An expanded theoretical treatment of iteration-dependent majorize–minimize algorithms *IEEE Trans. Image Process.* **16** 2411–22

[33] Bertsekas D P 1999 *Nonlinear Programming* 2nd edn (Belmont, MA: Athena Scientific)
[34] Shi Z-J 2004 Convergence of line search methods for unconstrained optimization *Appl. Math. Comput.* **157** 393–405
[35] Chouzenoux E, Moussaoui S and Idier J 2009 A majorize–minimize line search algorithm for barrier functions. *Technical Report* IRCCyN http://hal.archives-ouvertes.fr/hal-00362304
[36] Charbonnier P, Blanc-Féraud L, Aubert G and Barlaud M 1997 Deterministic edge-preserving regularization in computed imaging *IEEE Trans. Image Process.* **6** 298–311
[37] Lanteri H, Roche M and Aime C 2002 Penalized maximum likelihood image restoration with positivity constraints: multiplicative algorithms *Inverse Problems* **18** 1397–419
[38] Vicidomini G, Boccacci P, Diaspro A and Bertero M 2009 Application of the split-gradient method to 3D image deconvolution in fluorescence microscopy *J. Microsc.* **234** 47–61
[39] Zhang Y 2004 Interior-point gradient methods with diagonal-scalings for simple-bound constrained optimization *Technical Report* TR04-06 (Department of Computational and Applied Mathematics, Rice University, Houston, TX)
[40] Hager W W and Zhang H 2006 Recent advances in bound constrained optimization *System Modeling and Optimization* (*IFIP International Federation for Information Processing* vol 199) (Berlin: Springer) pp 67–82
[41] Butler J P, Reeds J A and Dawson S V 1981 Estimating solutions of first kind integral equations with nonnegative constraints and optimal smoothing *SIAM J. Numer. Anal.* **18** 381–97
[42] Nash S G 2000 A survey of truncated-Newton methods *J. Comput. Appl. Math.* **124** 45–59
[43] Nesterov Y and Nemirovskii A 1994 *Interior-Point Polynomial Alogorithms in Convex Programming (Studies in Applied Mathematics)* (Philadelphia: Society for Industrial and Applied Mathematics)
[44] Den Hertog D, Roos C and Terlaky T 1994 Inverse barrier methods for linear programming *Rev. Fr. Autom. Inform. Rech. Oper.* **28** 135–63
[45] Dhillon I S and Sra S 2005 Generalized nonnegative matrix approximations with Bregman divergences *Adv. Neural Inform. Process. Syst.* **19** 283–90

A.3 Efficient maximum entropy reconstruction of T1-T2 spectra

E. Chouzenoux, **S. Moussaoui**, J. Idier et F. Mariette, *IEEE Trans. on Signal Processing*, vol. 58, no. 12, 2010.

Dans cet article, une méthode d'optimisation itérative de type Newton tronqué est appliquée avec succès à la reconstruction de distributions des temps de relation T1-T2 en RMN bi-dimensionnelle. Le critère composite est de type moindres carrés pénalisés par maximum d'entropie. L'efficacité de la méthode est garantie en réalisant la recherche de pas par une approche MMLQ adaptée à un critère barrière et en proposant une mise œuvre exploitant la séparabilité du modèle direct.

Efficient Maximum Entropy Reconstruction of Nuclear Magnetic Resonance T1-T2 Spectra

Émilie Chouzenoux, Saïd Moussaoui, Jérôme Idier, *Member, IEEE*, and François Mariette

***Abstract*—This paper deals with the reconstruction of T1-T2 correlation spectra in nuclear magnetic resonance relaxometry. The ill-posed character and the large size of this inverse problem are the main difficulties to tackle. While maximum entropy is retained as an adequate regularization approach, the choice of an efficient optimization algorithm remains a challenging task. Our proposal is to apply a truncated Newton algorithm with two original features. First, a theoretically sound line search strategy suitable for the entropy function is applied to ensure the convergence of the algorithm. Second, an appropriate preconditioning structure based on a singular value decomposition of the forward model matrix is used to speed up the algorithm convergence. Furthermore, we exploit the specific structures of the observation model and the Hessian of the criterion to reduce the computation cost of the algorithm. The performances of the proposed strategy are illustrated by means of synthetic and real data processing.**

***Index Terms*—Laplace inversion, line search, maximum entropy, nuclear magnetic resonance, SVD preconditioning, T1-T2 spectrum, truncated Newton.**

I. Introduction

NUCLEAR magnetic resonance (NMR) relaxometry is a measurement technique used to analyze the properties of matter in order to determine its molecular structure and dynamics. After the immersion of the matter in a strong magnetic field, all the nuclear spins align to an equilibrium state along the field orientation. The application of a short magnetic pulse in resonance with the spin motion perturbates the spin orientation with a predefined angle Φ, called *flip angle* or *pulse angle*. The NMR experiment aims at analyzing the relaxation process which corresponds to the re-establishment of the spin into its equilibrium state.

This movement is decomposed into longitudinal and transverse dynamics, characterized by relaxation times T_1 and T_2, respectively. In practice, the longitudinal magnetization after τ_1 seconds of relaxation is measured by applying a 90° impulsion in the transverse plane. The transverse magnetization after $\tau_1+\tau_2$ seconds of relaxation is obtained by a series of dephasing impulsions in the transverse plane [1, Ch.4], [2, Ch.4], [3]. Classical NMR experiments are conducted to analyze the samples independently, either in terms of longitudinal or transverse relaxation, leading to one-dimensional (1-D) distributions [4], [5]. On the contrary, joint measurements with respect to the two relaxation parameters allow to build two-dimensional (2-D) T_1-T_2 spectra. Such spectra reveal couplings between T_1 and T_2 relaxations that are very useful for structure determination [6]–[8].

The physical model behind NMR relaxometry states that the measured NMR data $X(\tau_1,\tau_2)$ are related to the T_1-T_2 spectrum $S(\mathrm{T}_1,\mathrm{T}_2)$, according to a 2-D Fredholm integral of the first kind

$$X(\tau_1,\tau_2)=\iint k_1(\tau_1,\mathrm{T}_1)S(\mathrm{T}_1,\mathrm{T}_2)k_2(\tau_2,\mathrm{T}_2)d\mathrm{T}_1 d\mathrm{T}_2 \quad (1)$$

where k_1 and k_2 are kernels modeling the longitudinal and transverse relaxations

$$\begin{aligned} k_1(\tau_1,\mathrm{T}_1)&=1-\gamma e^{-\tau_1/\mathrm{T}_1}\\ k_2(\tau_2,\mathrm{T}_2)&=e^{-\tau_2/\mathrm{T}_2}\end{aligned} \quad (2)$$

with $\gamma=1-\cos\Phi$. In practice, an uncertainty in this observation model can occur if the pulse angle Φ is not set exactly to its desired value.

The associated inverse problem involving the recovery of the continuous distribution $S(\mathrm{T}_1,\mathrm{T}_2)$ is known to be an ill-posed problem [9].

Experimental data are collected at $m_1\times m_2$ discrete values in the $\tau_1-\tau_2$ domain. Thus, the data function $X(\tau_1,\tau_2)$ is replaced by a data matrix $\boldsymbol{X}\in\mathbb{R}^{m_1\times m_2}$. Similarly, the kernels k_1 and k_2 are discretized as matrices $\boldsymbol{K}_1\in\mathbb{R}^{m_1\times N_1}$ and $\boldsymbol{K}_2\in\mathbb{R}^{m_2\times N_2}$. Equation (1) takes a discrete form $\boldsymbol{X}=\boldsymbol{K}_1\boldsymbol{S}\boldsymbol{K}_2^{\mathrm{t}}$, where the spectrum $\boldsymbol{S}$ is a real-valued matrix of size $N_1\times N_2$. In practice, measurements are modeled by

$$\boldsymbol{Y}=\boldsymbol{K}_1\boldsymbol{S}\boldsymbol{K}_2^{\mathrm{t}}+\boldsymbol{E} \quad (3)$$

with $\boldsymbol{E}$ a noise term assumed white Gaussian. 2-D NMR reconstruction amounts to estimating $\boldsymbol{S}$ given $\boldsymbol{Y}$ subject to $\boldsymbol{S}\succeq 0$ (in the sense $S_{ij}\geqslant 0\ \forall i,j$). Attention must be paid to the size of the 2-D NMR problem. Indeed, when converted to a standard one-dimensional representation, (3) reads

$$\boldsymbol{y}=\boldsymbol{K}\boldsymbol{s}+\boldsymbol{e} \quad (4)$$

with $\boldsymbol{y}=\mathrm{vect}[\boldsymbol{Y}]$, $\boldsymbol{s}=\mathrm{vect}[\boldsymbol{S}]$, $\boldsymbol{e}=\mathrm{vect}[\boldsymbol{E}]$, $\mathrm{vect}[\cdot]$ denoting a column vector obtained by stacking all the elements of a matrix in lexicographic order and

$$\boldsymbol{K}=\boldsymbol{K}_1\otimes\boldsymbol{K}_2 \quad (5)$$

is the Kronecker product between matrices $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$. Matrix $\boldsymbol{K}$ is thus of size $m_1m_2\times N_1N_2$. Typical values are $m_1=50$,

Manuscript received March 05, 2010; accepted August 03, 2010. Date of publication September 02, 2010; date of current version November 17, 2010. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Isao Yamada.

É. Chouzenoux, S. Moussaoui, and J. Idier are with IRCCyN (CNRS UMR 6597), École Centrale Nantes, 44321 Nantes Cedex 03, France (e-mail: emilie.chouzenoux@irccyn.ec-nantes.fr; said.moussaoui@irccyn.ec-nantes.fr; jerome.idier@irccyn.ec-nantes.fr).

F. Mariette is with Cemagref, UR TERE, F-35044 Rennes, France, and Université Européenne de Bretagne, France (e-mail: francois.mariette@cemagref.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2010.2071870

$m_2 = 10^4$, $N_1 \times N_2 = 200 \times 200$, so $\boldsymbol{K}$ is a huge matrix whose explicit handling is almost impossible. It is one of the two main contributions of this paper to make use of the factored form (3) to solve this issue without any approximation.

Adopting the well-known least-squares approach would lead to define a spectrum estimate as the minimizer of

$$C(\boldsymbol{S}) = \frac{1}{2} \left\| \boldsymbol{Y} - \boldsymbol{K}_1 \boldsymbol{S} \boldsymbol{K}_2^{\mathrm{t}} \right\|_F^2 \tag{6}$$

where $\| \cdot \|_F$ denotes the Frobenius norm, under the positivity constraint $\boldsymbol{S} \succeq 0$. However, $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$ are rank-deficient and very badly conditioned matrices [10]. Therefore, such a solution is numerically unstable and regularized solutions must be sought instead. Given that the maximum entropy approach provides acknowledged methods for conventional (i.e., one-dimensional) NMR [4], [11], this paper explores T_1-T_2 spectrum estimation based on maximum entropy regularization and proposes a specific descent algorithm. According to our experience, the barrier shape of the entropy function makes the minimization problem quite specific. In particular, general-purpose nonlinear programming algorithms can be extremely inefficient in terms of convergence speed. More surprisingly, the more specific scheme adapted from [12] also turns out to be very slow to converge. This motivated us to devise an alternative optimization strategy that is provably convergent and shows a good tradeoff between simplicity and efficiency. The proposed algorithm belongs to the truncated Newton algorithm family but possesses original features regarding the line search and the preconditioning strategy.

The rest of the paper is organized as follows. Section II gives an overview of different regularization strategies that can be applied to solve this problem. Section III proposes an efficient reconstruction method for maximum entropy regularization, based on a truncated Newton algorithm associated with an original line search strategy well suited to the form of the criterion. The computation cost of the algorithm is reduced by working directly with the factored form (6) to calculate quantities such as gradient and Hessian-vector products. In Section IV, the efficiency of the proposed scheme is illustrated by means of synthetic and real data examples.

II. Problem Statement and Existing Solutions

The mathematical methods developed to solve (1) can be classified into two groups. The first approach is to fit the decay curves with a minimal number of discrete exponentials terms. The parametric minimization is usually handled with the Levenberg-Marquardt algorithm [13]. In this paper, we rather focus on the second approach which analyzes the data in terms of a continuous distribution of relaxation components $S(\mathrm{T}_1, \mathrm{T}_2)$. This model gives rise to the linear equation (3). In this section, we give an overview of different inversion strategies for this problem.

A. Direct Resolution: TSVD and Tikhonov Methods

NMR reconstruction is a linear ill-posed problem. To tackle it, truncated singular value decomposition (TSVD) and Tikhonov penalization (TIK) are commonly used methods [9]. Each of them calls for its own regularization principle to compensate the ill-conditioned character of the observation matrix.

1) TSVD: The TSVD approach consists in replacing the inverse (or the generalized inverse) of $\boldsymbol{K}$ by a matrix of reduced rank, in order to avoid the amplification of noise due to the inversion of small nonzero singular values [14]. In practice, computing the TSVD requires the explicit decomposition of $\boldsymbol{K}$ in terms of singular elements, which can be numerically burdensome.

2) Tikhonov Penalization: While TSVD tackles the ill-posed character by control of dimensionality, Tikhonov method follows a penalization approach by which a tradeoff is sought between fidelity-to-data and regularity. It leads to the minimization of a mixed objective function

$$L(\boldsymbol{S}) = C(\boldsymbol{S}) + \lambda R(\boldsymbol{S}) \tag{7}$$

where the regularization parameter $\lambda > 0$ controls the respective weight of the two terms, C is a least-square term

$$C(\boldsymbol{S}) = \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{s}\|^2 = \frac{1}{2} \left\| \boldsymbol{Y} - \boldsymbol{K}_1 \boldsymbol{S} \boldsymbol{K}_2^{\mathrm{t}} \right\|_F^2 \tag{8}$$

and the additional term R is also a quadratic term. In the context of NMR reconstruction, the regularization functional R is usually chosen as the squared ℓ_2-norm of the spectrum [5], [10], [15], [16]

$$R(\boldsymbol{S}) = \frac{1}{2} \|\boldsymbol{s}\|^2 = \frac{1}{2} \|\boldsymbol{S}\|_F^2. \tag{9}$$

Tikhonov solution is then obtained by solving the linear system $(\boldsymbol{K}^{\mathrm{t}}\boldsymbol{K} + \lambda \boldsymbol{I})\boldsymbol{s} = \boldsymbol{K}^{\mathrm{t}}\boldsymbol{y}$.

B. Iterative Minimization

Both TSVD and TIK solutions provide results of limited resolution. Moreover, they tend to exhibit oscillatory excursions, especially in the peripheral regions of the recovered peaks, which usually violate the positivity of the spectrum components [17]. Enforcing the positivity of the spectrum is obviously desirable from the viewpoint of physical interpretation, but it has also a favorable effect on the resolution of the estimated spectrum.

1) Tikhonov Under Positivity Constraint (TIK^+): The positivity constraint $\boldsymbol{S} \succeq 0$ is naturally incorporated into Tikhonov approach by constraining the minimization of L to the positive orthant. However, there is no closed-form expression for the minimizer anymore, so the solution must be computed iteratively using a fixed-point algorithm.

Butler–Reeds–Dawson algorithm (BRD) [10] is a rather simple and efficient technique based on the resolution of the Karush–Kuhn–Tucker conditions [18]. Although commonly used in materials science, it is scarcely referenced in the quadratic programming literature. For the sake of clarification, Appendix A proposes a very simple interpretation of the BRD scheme as iteratively minimizing a dual function of the criterion in the sense of Legendre–Fenchel duality [19].

However, the BRD scheme requires the inversion of a system of size $m \times m$ at each iteration, where m is the number of measurements. In the case of 2-D NMR problems, $m = m_1 m_2$, and usual values of m_1 and m_2 lead to a prohibitive computation cost. To solve this issue, a data compression step is proposed in [15], prior to the application of BRD. It relies on strongly truncated singular value decompositions of $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$ $\boldsymbol{K}_i \approx$

$\boldsymbol{U}_i\boldsymbol{\Sigma}_i\boldsymbol{V}_i^{\mathrm{t}}$, $i = 1, 2$, with $\tilde{m}_i = \mathrm{rank}(\boldsymbol{K}_i) \ll m_i$. The fidelity to data term is then approximated by

$$\tilde{C}(\boldsymbol{S}) = \frac{1}{2}\left\|\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{K}}_1\boldsymbol{S}\tilde{\boldsymbol{K}}_2^{\mathrm{t}}\right\|_F^2 \tag{10}$$

where $\tilde{\boldsymbol{K}}_1 = \boldsymbol{\Sigma}_1\boldsymbol{V}_1^{\mathrm{t}}$, $\tilde{\boldsymbol{K}}_2 = \boldsymbol{\Sigma}_2\boldsymbol{V}_2^{\mathrm{t}}$ and $\tilde{\boldsymbol{Y}} = \boldsymbol{U}_1^{\mathrm{t}}\boldsymbol{Y}\boldsymbol{U}_2$ are of size $\tilde{m}_1 \times N_1$, $\tilde{m}_2 \times N_2$ and $\tilde{m}_1 \times \tilde{m}_2$, respectively.

2) Maximum Entropy: A different regularization approach will be considered here, based on Shannon entropy penalization $\phi(s) = -s\log s$. Maximum entropy (ME) [12], [20] is an acknowledged approach in the context of 1-D NMR relaxometry [4], [11]. An interesting feature of entropy penalization is that it implicitly handles the positivity constraint since the norm of the gradient of the entropy term is unbounded at the boundary of the positive orthant. Thus, the minimizer of the resulting penalized least-square criterion cancels its gradient, and computing it is essentially similar to solving an unconstrained optimization problem.

Formally, the extension to the 2-D case is easily obtained by minimization of

$$L(\boldsymbol{S}) = \frac{1}{2}\left\|\boldsymbol{Y} - \boldsymbol{K}_1\boldsymbol{S}\boldsymbol{K}_2^{\mathrm{t}}\right\|_F^2 + \lambda\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} S_{ij}\log S_{ij}. \tag{11}$$

However, the practical computation of the solution is clearly more difficult in the 2-D case because the optimization problem is much larger-scale. The choice of a specific minimization scheme suited to maximum entropy 2-D NMR reconstruction is a challenging task.

In the context of maximum entropy, [21] proposed the fixed-point multiplicative algebraic reconstruction technique (MART) that maximizes the entropy term subject to $\boldsymbol{K}\boldsymbol{s} = \boldsymbol{y}$. The simplicity of MART is attractive. However, as emphasized in [22], the presence of inherent noise in projection data makes this method less effective than an approach based on the minimization of the penalized criterion (11). In [12], an iterative minimization algorithm based on a quadratic approximation of the criterion over a low-dimension subspace is developed. However, according to [23, p. 1022], the convergence of this algorithm is not established. We have tested its behavior in the 2-D NMR context. Our conclusions are that this algorithm does not ensure a monotonic decrease of the criterion, and that its convergence is very slow [24]. Finally, in a preliminary version of the present work, we have proposed to make use of a preconditioned nonlinear conjugate gradient algorithm [25]. Although the latter shows a good practical behavior, its theoretical convergence is not ensured, since the preconditioner is a variable matrix.

The goal of the next section is to derive an optimization algorithm that would benefit from stronger theoretical properties and sufficiently low computational cost to avoid any data compression step.

III. Proposed Truncated Newton Algorithm

A. Minimization Strategy

The truncated Newton (TN) algorithm [26], [27] is based on iteratively decreasing the objective function $L(\boldsymbol{s})$ by moving the current solution $\boldsymbol{s}_k$ along a descent direction $\boldsymbol{d}_k$

$$\boldsymbol{s}_{k+1} = \boldsymbol{s}_k + \alpha_k\boldsymbol{d}_k \tag{12}$$

where $\alpha_k > 0$ is the stepsize and $\boldsymbol{d}_k$ is a search direction computed by solving approximately the Newton equations

$$\boldsymbol{H}_k\boldsymbol{d}_k = -\boldsymbol{g}_k \tag{13}$$

with $\boldsymbol{H}_k \triangleq \nabla^2 L(\boldsymbol{s}_k)$ and $\boldsymbol{g}_k \triangleq \nabla L(\boldsymbol{s}_k)$. The TN algorithm has been widely used in the context of interior point algorithms with logarithmic [28], [29] and entropic [22] barrier functions.

In practice, the TN method consists in alternating the construction of $\boldsymbol{d}_k$ and the computation of the stepsize α_k by a line search procedure. The direction $\boldsymbol{d}_k$ results from preconditioned conjugate gradient (PCG) iterations on (13) stopped before convergence. The stepsize α_k is obtained by iteratively minimizing the scalar function $\ell(\alpha) = L(\boldsymbol{s}_k + \alpha\boldsymbol{d}_k)$ until some convergence conditions are met [18, Ch. 3]. Typically, the strong Wolfe conditions are considered

$$\ell(\alpha_k) \leqslant \ell(0) + c_1\alpha_k\dot{\ell}(0) \tag{14}$$

$$\left|\dot{\ell}(\alpha_k)\right| \leqslant c_2\left|\dot{\ell}(0)\right| \tag{15}$$

where $(c_1, c_2) \in (0, 1)$ are tuning parameters that do not depend on k. There exist several procedures to find an acceptable stepsize: exact minimization of $\ell(.)$, backtracking, approximation of $\ell(.)$ using cubic interpolations [18], [30] or quadratic majorizations [31], [32]. However, the entropic penalty term implies that the derivative of $\ell(\alpha)$ takes the value $-\infty$ as soon as any of the components of the vector $\boldsymbol{s}_k + \alpha\boldsymbol{d}_k$ vanishes, hence when α is equal to one of the two limit values

$$\alpha_- = \max_{i,d_{k,i}>0}\left(\frac{-s_i}{d_{k,i}}\right), \quad \alpha_+ = \min_{i,d_{k,i}<0}\left(\frac{-s_i}{d_{k,i}}\right). \tag{16}$$

The function ℓ is undefined outside (α_-, α_+), therefore, we must ensure that during the line search, the stepsize values remain in the interval (α_-, α_+). Moreover, because of the vertical asymptotes at α_- and α_+, standard methods using cubic interpolations or quadratic majorizations are not well suited. Our proposal is to adopt the specific majorization-based line search proposed in [33] and [34] for barrier function optimization. Using an adequate form of majorization, we now derive an analytical stepsize formula preserving strong convergence properties.

B. Line Search Strategy

The minimization of $\ell(\cdot)$ using the Majorization–Minimization (MM) principle [35] is performed by successive minimizations of majorant functions for $\ell(.)$. Function $h(\alpha, \alpha')$ is said to be majorant for $\ell(\alpha)$ at α' if for all α,

$$\begin{cases} h(\alpha, \alpha') \geqslant \ell(\alpha) \\ h(\alpha', \alpha') = \ell(\alpha'). \end{cases} \tag{17}$$

As illustrated in Fig. 1, the initial minimization of $\ell(\alpha)$ is then replaced by a sequence of easier subproblems, corresponding to the MM update rule

$$\begin{cases} \alpha_k^0 = 0, \\ \alpha_k^j = \arg\min_\alpha h^j\left(\alpha, \alpha_k^{j-1}\right), \quad j = 1, \ldots, J_k, \\ \alpha_k = \alpha_k^{J_k}. \end{cases} \tag{18}$$

Following [34], we propose a majorant function $h^j(., \alpha_k^j)$ that incorporates barriers to account for the entropy term. It is piece-

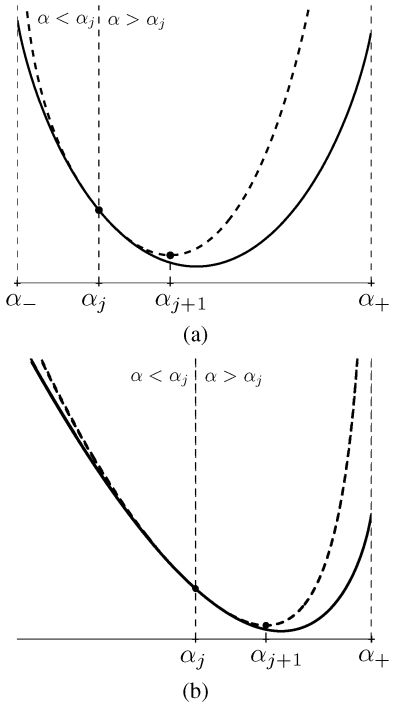


Fig. 1. Schematic principle of the MM line search procedure. The tangent majorant function $h^j(\alpha, \alpha^j)$ (dashed line) for $\ell(\alpha)$ (solid line) at α_j is piecewise defined on the sets $(\alpha_-, \alpha_j]$ and $[\alpha_j, \alpha_+)$. The new iterate α_{j+1} is taken as the minimizer of $h_j(., \alpha_j)$. Two cases are illustrated. The third and last case where α_- is finite and $\alpha_+ = +\infty$ is the mirror image of case (b). (a) Case α_- and α_+ finite; (b) case $\alpha_- = -\infty$ and α_+ finite.

wise defined under the following form (whenever unambiguous, the iteration index k will be dropped for the sake of simplicity)

$$h^j(\alpha, \alpha^j) = \begin{cases} p_0^- + p_1^- \alpha + p_2^- \alpha^2 \\ \quad -p_3^- \log(\alpha - \alpha_-) & \text{for all } \alpha \in (\alpha_-; \alpha^j] \\ p_0^+ + p_1^+ \alpha + p_2^+ \alpha^2 \\ \quad -p_3^+ \log(\alpha_+ - \alpha) & \text{for all } \alpha \in [\alpha^j; \alpha_+). \end{cases} \tag{19}$$

The parameters $p_n^{\pm}$, $n = 0, \ldots, 3$ must be defined to ensure that $h^j(., \alpha^j)$ is actually a majorant of $\ell(\cdot)$ at α^j (see Fig. 1 for an illustration). A direct application of [34, Prop. 2] allows to establish expressions for these parameters. The resulting form of $h^j(., \alpha^j)$ is rather simple, though lengthy to express, so it is reported in Appendix B. According to [34, Lemma 2], it corresponds to a strictly convex, twice differentiable function in the set (α_-, α_+). Moreover, its unique minimizer takes an explicit form, the latter being also found in Appendix B.

Finally, (18) produces monotonically decreasing values $\{\ell(\alpha^j)\}$ and the series $\{\alpha^j\}$ converges to a stationary point of $\ell(\alpha)$ [36].

C. Convergence Result

Let us focus on the convergence of the truncated Newton algorithm when α_k is chosen according to the proposed MM strategy. A detailed analysis can be found in [34] in a more general framework. According to [34], the proposed line search procedure ensures that

$$\sum_k \frac{\left(\boldsymbol{g}_k^{\mathrm{t}} \boldsymbol{d}_k\right)^2}{\|\boldsymbol{d}_k\|^2} < \infty \tag{20}$$

and that the directions generated by the TN algorithm are *gradient related* in the sense of [37]. According to [38], inequality (20), known as *Zoutendijk condition*, is sufficient to prove the convergence of the algorithm in the sense $\lim_{k\to\infty} \|\boldsymbol{g}_k\| = 0$. Finally, the objective function being strictly convex, the proposed algorithm converges to its unique minimizer.

D. Preconditioning

As emphasized in [39], the Hessian of the Shannon entropy regularization term is very ill-conditioned for points that are close to the boundary of the positive orthant since some of its eigenvalues tend to infinity. Furthermore, the exponential decays in kernels k_1 and k_2 imply that $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$ are also very ill-conditioned. Preconditioning is a well-known technique to obtain more clustered eigenvalues of the Hessian of the criterion and to accelerate the convergence of descent algorithms. The principle is to transform the space of original variables into a space in which the Hessian has more clustered eigenvalues by using a preconditioning matrix $\boldsymbol{P}_k$ that approximates the inverse $\boldsymbol{H}_k^{-1}$ of the Hessian. A good preconditioner achieves a tradeoff between the approximation quality and the computation cost. General-purpose preconditioning strategies have been proposed in the literature including symmetric successive overrelaxation and incomplete LU or Cholesky factorizations [40, Ch. 10], [41]. In the context of ME optimization, [22] takes $\boldsymbol{P}_k$ as a diagonal matrix defined using the Hessian diagonal elements

$$\boldsymbol{P}_k = \left[\mathrm{diag}\left(\mathrm{diag}(\boldsymbol{K}^{\mathrm{t}}\boldsymbol{K})\right) + \lambda\, \mathrm{diag}(\boldsymbol{s}_k)^{-1}\right]^{-1}. \tag{21}$$

We rather propose a more specific preconditioner. It is based on the fact that, as a consequence of (5), the singular value decomposition of $\boldsymbol{K}$ is given by $\boldsymbol{K} = \boldsymbol{U}^{\mathrm{t}}\boldsymbol{\Sigma}\boldsymbol{V}$, with $\boldsymbol{U} = \boldsymbol{U}_1 \otimes \boldsymbol{U}_2$, $\boldsymbol{V} = \boldsymbol{V}_1 \otimes \boldsymbol{V}_2$, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1 \otimes \boldsymbol{\Sigma}_2$, $\boldsymbol{U}_i^{\mathrm{t}}\boldsymbol{\Sigma}_i\boldsymbol{V}_i$ being the singular value decomposition of $\boldsymbol{K}_i$, $i = 1, 2$. Then, let us define

$$\boldsymbol{P}_k = \left[\tilde{\boldsymbol{V}}\tilde{\boldsymbol{\Sigma}}^2\tilde{\boldsymbol{V}}^{\mathrm{t}} + \lambda\, \mathrm{diag}(\boldsymbol{s}_k)^{-1}\right]^{-1} \tag{22}$$

where $\tilde{\boldsymbol{V}}$ and $\tilde{\boldsymbol{\Sigma}}$ correspond to truncated versions of $\boldsymbol{V}$ and $\boldsymbol{\Sigma}$. In the nontruncated case, $\tilde{\boldsymbol{V}} = \boldsymbol{V}$ and $\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}$, and $\boldsymbol{P}_k$ then is equal to the Hessian of $\boldsymbol{L}$ at $\boldsymbol{s}_k$. It remains to define the way we truncate the singular value decomposition of $\boldsymbol{K}$. Akin to [15], [42], we separately truncate the decompositions of $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$ at ranks v_1, v_2 and we define $\tilde{\boldsymbol{V}}$ and $\tilde{\boldsymbol{\Sigma}}$ according to

$$\tilde{\boldsymbol{V}} = \tilde{\boldsymbol{V}}_1 \otimes \tilde{\boldsymbol{V}}_2 \tag{23}$$

$$\tilde{\boldsymbol{\Sigma}} = \tilde{\boldsymbol{\Sigma}}_1 \otimes \tilde{\boldsymbol{\Sigma}}_2. \tag{24}$$

Let us remark that the resulting approximation of $\boldsymbol{K}$ may slightly differ from the TSVD of $\boldsymbol{K}$. The reason is simple: although $\tilde{\boldsymbol{\Sigma}}_1$ and $\tilde{\boldsymbol{\Sigma}}_2$ separately gather the largest singular values of $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, $\tilde{\boldsymbol{\Sigma}}$ does not necessarily gather the largest singular values of $\boldsymbol{\Sigma}$. As a consequence, our approximation may be suboptimal compared to the TSVD, the latter being optimal in the least-square sense [43], but the fact that we maintain factored expressions for matrices $\tilde{\boldsymbol{V}}$ and $\tilde{\boldsymbol{\Sigma}}$ is essential in terms of computation cost.

E. Memory Storage and Computation Cost Reduction

The computation cost can be reduced by exploiting the factored form of the observation model. Three main operations are involved in the iterative optimization algorithm: the computation of the gradient vector $\boldsymbol{g}_k = \nabla L(\boldsymbol{s}_k)$, and the products of

$\bar{\boldsymbol{P}}_k$ and $\bar{\boldsymbol{H}}_k$ with a vector. The three resulting quantities can be calculated using low cost operations, as described below.

1) Gradient: The gradient of the criterion can be computed without explicitly handling matrix $\boldsymbol{K}$, according to

$$\boldsymbol{g}_k = -\text{vect}\left[\boldsymbol{K}_1^{\text{t}}\left(\boldsymbol{Y} - \boldsymbol{K}_1\boldsymbol{S}_k\boldsymbol{K}_2^{\text{t}}\right)\boldsymbol{K}_2\right] + \lambda(1+\log \boldsymbol{s}_k). \quad (25)$$

2) Hessian: In the same manner, products between the Hessian matrix and any vector $\boldsymbol{w} = \text{vect}[\boldsymbol{W}]$ can be computed as follows:

$$\boldsymbol{H}_k\boldsymbol{w} = \text{vect}\left[\boldsymbol{K}_1^{\text{t}}\boldsymbol{K}_1\boldsymbol{W}\boldsymbol{K}_2^{\text{t}}\boldsymbol{K}_2\right] + \lambda(\boldsymbol{w}./\boldsymbol{s}_k) \quad (26)$$

where "./" denotes componentwise division.

3) Preconditioner: In order to compute products involving $\boldsymbol{P}_k$, the matrix inversion lemma is applied to (22). Thus,

$$\boldsymbol{P}_k = \boldsymbol{A}_k - \boldsymbol{A}_k\tilde{\boldsymbol{V}}(\tilde{\boldsymbol{\Sigma}}^{-2} + \tilde{\boldsymbol{V}}^{\text{t}}\boldsymbol{A}_k\tilde{\boldsymbol{V}})^{-1}\tilde{\boldsymbol{V}}^{\text{t}}\boldsymbol{A}_k, \quad (27)$$

with $\boldsymbol{A}_k = \lambda^{-1}\text{diag}(\boldsymbol{s}_k)$. Moreover, the following factored expression can be deduced from (23) for the entries of matrix $\boldsymbol{M} = \tilde{\boldsymbol{V}}^{\text{t}}\boldsymbol{A}_k\tilde{\boldsymbol{V}} \in \mathbb{R}^{v_1v_2\times v_1v_2}$

$$M_{ij} = \frac{1}{\lambda}\sum_{m=1}^{N_1}\sum_{n=1}^{N_2}(S_k)_{mn}(\tilde{V}_1)_{ma}(\tilde{V}_2)_{nb}(\tilde{V}_1)_{mc}(\tilde{V}_2)_{nd}$$

where (a,b) and (c,d) are row and column subscripts that correspond to the linear indexes i and j, respectively. Thus, the product $\boldsymbol{P}_k\boldsymbol{w}$ can be efficiently computed according to

$$\begin{aligned}\boldsymbol{P}_k\boldsymbol{w} &= \boldsymbol{b}_k - \boldsymbol{A}_k\tilde{\boldsymbol{V}}(\tilde{\boldsymbol{\Sigma}}^{-2} + \boldsymbol{M})^{-1}\tilde{\boldsymbol{V}}^{\text{t}}\boldsymbol{b}_k \\ &= \boldsymbol{b}_k - \boldsymbol{A}_k\text{vect}\left[\tilde{\boldsymbol{V}}_1\boldsymbol{Q}_k\tilde{\boldsymbol{V}}_2^{\text{t}}\right] \quad (28)\end{aligned}$$

where $\boldsymbol{b}_k = \boldsymbol{A}_k\boldsymbol{w}$, $\boldsymbol{q}_k = (\tilde{\boldsymbol{\Sigma}}^{-2} + \boldsymbol{M})^{-1}\text{vect}[\tilde{\boldsymbol{V}}_1^{\text{t}}\boldsymbol{B}_k\tilde{\boldsymbol{V}}_2]$ and $\boldsymbol{Q}_k$, $\boldsymbol{B}_k$ denote the equivalent square matrix representations of $\boldsymbol{q}_k$ and $\boldsymbol{b}_k$ respectively.

F. Resulting Algorithm

The resulting TN algorithm is given in Algorithm 1. The algorithm convergence is checked using the following stopping rule [18]:

$$\|\boldsymbol{g}_k\|_\infty < \epsilon\left(1 + |L(\boldsymbol{s}_k)|\right), \quad (29)$$

and the PCG iterations in Algorithm 2 are stopped when [27]

$$\|\boldsymbol{g}_k + \boldsymbol{H}_k\boldsymbol{d}_k\| \leqslant \eta\,\|L(\boldsymbol{s}_k)\|. \quad (30)$$

Typical values of (ϵ, η) are $(10^{-8}, 10^{-4})$.

Algorithm 1: TN Algorithm for ME Optimization

Require: Initial value $\boldsymbol{s}_0 \succeq 0$, parameters v_1, v_2, λ, J and accuracies ϵ, η.

Ensure: Resolution of (11)

Compute the TSVD of $\bar{\boldsymbol{K}}_1$ and $\bar{\boldsymbol{K}}_2$ at ranks v_1, v_2.

while (29) does not hold **do**

Compute $\boldsymbol{g}_k$, $\boldsymbol{P}_k$ and $\boldsymbol{H}_k$ using (25), (26) and (27).

Compute $\boldsymbol{d}_k$ using PCG algorithm (Table II).

Set α_k after J iterations of (18).

Update $\boldsymbol{s}_k$ according to (12).

end while

Algorithm 2: PCG Algorithm

Require: $\boldsymbol{g}_k$, $\bar{\boldsymbol{H}}_k$, $\bar{\boldsymbol{P}}_k$, η

Ensure: Approximate solution $\boldsymbol{d}_k$ of (13)

$\boldsymbol{u}_0 \leftarrow \boldsymbol{0}$

$\boldsymbol{r}_0 \leftarrow -\boldsymbol{g}_k - \bar{\boldsymbol{H}}_k\boldsymbol{u}_0$

$\boldsymbol{p}_0 \leftarrow \bar{\boldsymbol{P}}_k\boldsymbol{r}_0$

while (30) does not hold **do**

$\theta_i \leftarrow \left(\boldsymbol{r}_i^{\text{t}}\bar{\boldsymbol{P}}_k\boldsymbol{r}_i\right) / \left(\boldsymbol{p}_i^{\text{t}}\bar{\boldsymbol{H}}_k\boldsymbol{p}_i\right)$

$\boldsymbol{u}_{i+1} \leftarrow \boldsymbol{u}_i + \theta_i\boldsymbol{p}_i$

$\boldsymbol{r}_{i+1} \leftarrow \boldsymbol{r}_i - \theta_i\bar{\boldsymbol{H}}_k\boldsymbol{p}_i$

$\beta_i \leftarrow \left(\boldsymbol{r}_{i+1}^{\text{t}}\bar{\boldsymbol{P}}_k\boldsymbol{r}_{i+1}\right) / \left(\boldsymbol{r}_i^{\text{t}}\bar{\boldsymbol{P}}_k\boldsymbol{r}_i\right)$

$\boldsymbol{p}_{i+1} \leftarrow \bar{\boldsymbol{P}}_k\boldsymbol{r}_{i+1} + \beta_i\boldsymbol{p}_i$

$\boldsymbol{d}_k \leftarrow \boldsymbol{u}_{i+1}$

end while

IV. Experimental Results

This section discusses the performances of the proposed method and illustrates its applicability. First, we consider synthetic data in order to discuss the influence of the tuning parameters on the algorithm behavior. Then, the proposed method applicability is illustrated through the processing of real NMR data.

In NMR experiments, the pulse angle Φ may not be set exactly to its desired value. Therefore, we analyze the effect of a potential error in the value of γ in the observation model and propose an original strategy allowing to estimated this parameter.

The different results are obtained with Matlab 7.5 running on an Intel Pentium IV 3.2-GHz, 3-GB RAM.

A. Synthetic Data

We consider two spectra A and B (Fig. 2) and the corresponding decays (Fig. 3) according to the observation model (4) with a signal-to-noise ratio (SNR) of 10 dB, $m_1 = 100$, $m_2 = 1000$, and $\gamma = 1$ (i.e., $\Phi = 90°$). The synthetic spectrum A has a symmetric Gaussian shape located at $[\text{T}_1, \text{T}_2] = [0.5\text{ s}, 1\text{ s}]$ while spectrum B is the sum of two Gaussian patterns. The first one is symmetric and located at $[\text{T}_1, \text{T}_2] = [0.5\text{ s}, 0.5\text{ s}]$. The second pattern is located at $[\text{T}_1, \text{T}_2] = [1.5\text{ s}, 1.5\text{ s}]$ and simulates a positive T_1-T_2 correlation. The reconstruction is performed for $N_1 = N_2 = 100$ and the algorithm is initialized with a uniform positive 2-D spectrum. The regularization parameter λ is set to minimize the normalized quadratic error

$$Q = 100\,\|\boldsymbol{s}(\lambda) - \boldsymbol{s}^o\|_2^2 / \|\boldsymbol{s}^o\|_2^2 \quad (31)$$

and the preconditioner truncation parameters v_1, v_2 are set to the same value v.

1) PCG Subiterations: The parameter η controls the accuracy of the PCG minimization. The smaller it is, the more accurate the solving of (13). Here, several values are tested within the range $[10^{-7}, 10^{-1}]$. Let I_k denotes the number of PCG subiterations (inner loop) at iteration k. As expected, the average value of I_k generally increases with η [Fig. 4(a)] while the number of TN iterations K (outer loop) decreases [Fig. 4(b)]. The number of PCG subiterations depends also on the truncation rank v of

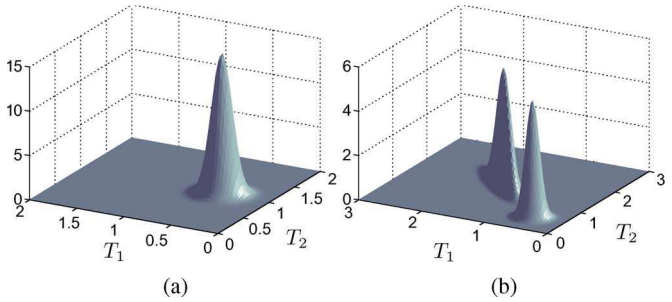


Fig. 2. Simulated 2-D spectra: (a) Dataset A and (b) dataset B.

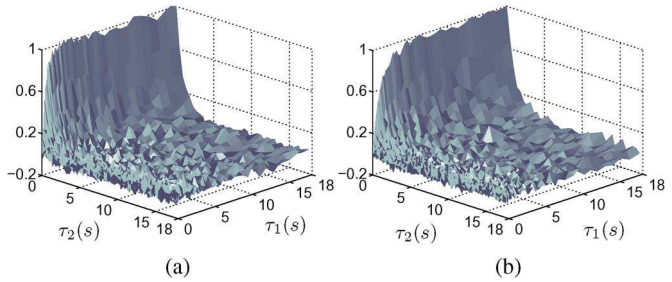


Fig. 3. NMR decays: (a) Dataset A and (b) dataset B.

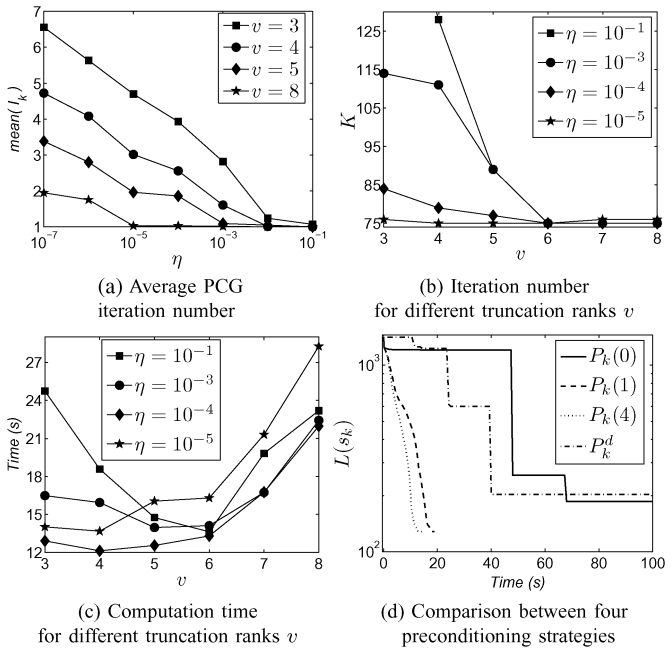


Fig. 4. Dataset A: Analysis of the TN algorithm performances for different PCG strategies. The SVD preconditioner with truncature parameter v was used for (a)–(c) while the truncature parameter η is set to 10^{-4} for (d). Moreover, in all cases, the stepsize results from $J = 1$ subiteration of MM line search. (a) Average PCG iteration number; (b) iteration number for different truncation ranks v; (c) computation time for different truncation ranks v; and (d) comparison between four preconditioning strategies.

the SVD preconditioner, it can be noted that I_k decreases as this rank increases, corresponding to a more accurate approximation of the inverse Hessian matrix. The smallest overall minimization time is achieved when a tradeoff is reached between the number of outer iterations and the number of inner iterations [Fig. 4(c)].

TABLE I
DATASET A: COMPARISON BETWEEN MM AND MT LINE SEARCH STRATEGIES IN TERMS OF ITERATION NUMBER AND TIME BEFORE CONVERGENCE FOR THE TN ALGORITHM

	c_1	c_2	K	T (s)
MT	10^{-1}	0.5	93	19.84
	10^{-1}	0.9	90	15.64
	10^{-1}	0.99	170	25.72
	10^{-3}	0.5	93	16.98
	10^{-3}	0.9	90	**15.36**
	10^{-3}	0.99	170	25.14
MM	J		K	T (s)
	1		79	**13.56**
	2		85	15.09
	3		84	15.06
	4		84	15.11
	5		85	15.31

In this example, the best compromise is $(v, \eta) = (4, 10^{-4})$. This setting will be retained in the sequel.

2) Preconditioning: Fig. 4(d) illustrates the criterion evolution for different preconditioners: the proposed approximation $\boldsymbol{P}_k(v)$ given by (22) with $v_1 = v_2 = v = 0$, 1, 4 and the diagonal preconditioner $\boldsymbol{P}_k^d$ resulting from (21). The stopping criterion (29) is fulfilled after 93 and 80 iterations for $\boldsymbol{P}_k(1)$ and $\boldsymbol{P}_k(4)$, whereas it is not fulfilled after 1000 iterations neither for $\boldsymbol{P}_k(0)$ nor for $\boldsymbol{P}_k^d$. Moreover, according to Fig. 4(a), the TN iteration number decreases as the SVD truncation rank v increases. However, the choice of v involves a compromise between an acceleration of the algorithm and an increase of the computational cost [Fig. 4(b) and (c)].

3) Line Search: Let us compare the performances of the algorithm when the stepsize is obtained either by the proposed MM line search or by Moré and Thuente's cubic interpolation procedure (MT) [30]. The latter performs an iterative minimization of $\ell(.)$ based on cubic interpolation until identifying α_k that fulfills the strong Wolfe conditions (14) and (15).

According to Table I, the TN algorithm with the MM line search performs better than with the MT line search with the best settings for c_1 and c_2. Concerning the choice of the sub-iteration number, it appears that $J = 1$ leads to the best results in terms of computation time which shows that an exact minimization of the scalar function $\ell(\alpha)$ during line search is not necessary.

4) Regularization Term: As explained in the introduction, the application of BRD algorithm to 2-D NMR reconstruction requires data compression. This preprocessing step calls for the tuning of two additional parameters, $\tilde{m}_1$ and $\tilde{m}_2$. Table II illustrates the reconstruction quality and algorithmic properties of BRD method for different values of $\tilde{m}_i$. As expected, the computation cost decreases with $\tilde{m}_i$. However, according to Fig. 5, below a certain compression value $\tilde{m}_1^{\min}$, the reconstruction error quickly grows. We observe that $\tilde{m}_1^{\min} = 3$ for dataset A and $\tilde{m}_1^{\min} = 5$ for dataset B. The same behavior was observed when varying $\tilde{m}_2$. This shows that the compression tuning not only depends on spectral properties of matrices $\boldsymbol{K}_i$ [15], but also on the spectra shape. Therefore, the setting of these parameters may be problematic when processing real data.

In order to compare the ME and TIK^+ regularizations, we apply the same compression level $\tilde{m}_1 = \tilde{m}_2 = 5$. We have

TABLE II
RECONSTRUCTION QUALITY Q, ITERATION NUMBER K AND TIME BEFORE CONVERGENCE T FOR TIK^+-BRD RECONSTRUCTION WITH DIFFERENT LEVELS OF DATA COMPRESSION

	$\tilde{m}_1$	10	10	10	5	2	1
	$\tilde{m}_2$	100	50	10	5	5	1
A	Q	4.53	4.66	4.79	3.92	81.6	97.9
A	K	31	30	20	22	13	12
A	T (s)	43	29	4	2	< 1	< 1
B	Q	12.8	12.8	12.6	10.7	84.4	94.3
B	K	19	18	18	20	11	2
B	T (s)	50	27	3	< 1	< 1	< 1

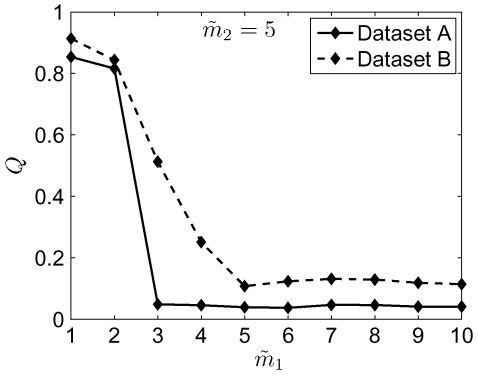


Fig. 5. TIK^+-BRD reconstruction quality of dataset A and B with different level of data compression. In both cases, the compression parameter $\bar{m}_2$ is equal to 5 while $\bar{m}_1$ is varying. ($\mathrm{SNR} = 10$ dB).

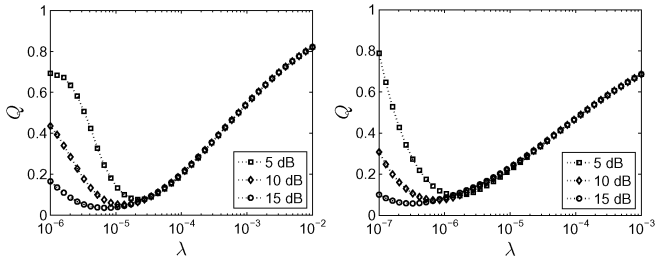


Fig. 6. Dataset A: Similarity error for ME (left) and TIK^+ (right) reconstructions. Average of Monte Carlo simulations with 100 random realizations for $\mathrm{SNR} = 5$, 10, and 15 dB ($\bar{m}_i = 5$).

tested different noise realizations with $\mathrm{SNR} = 5$, 10, and 15 dB. According to Fig. 6, the minimum value of $Q(\lambda)$ decreases with the noise level, for both ME and TIK^+ regularizations, as expected. Moreover, the two strategies lead to similar reconstruction minimum errors for the three noise levels. Furthermore, their sensitivity to λ is similar. However, as illustrated in Figs. 7 and 8, the entropy penalization leads to spectra whose shape is closer to the simulated one. More precisely, the ME spectra are smoother. This regularity is evaluated in Table III, which compares the reconstructions in terms of the Euclidian norm of the first-order difference $\|\Delta \boldsymbol{s}\|$.

5) Hyperparameter Estimation: In the previous experiments, the regularization parameter λ is tuned by minimizing a quadratic error whose evaluation requires the knowledge of the reference spectrum. This strategy is impractical in an experimental context but it can be replaced by different procedures proposed in the literature. In NMR reconstruction [4], [10],

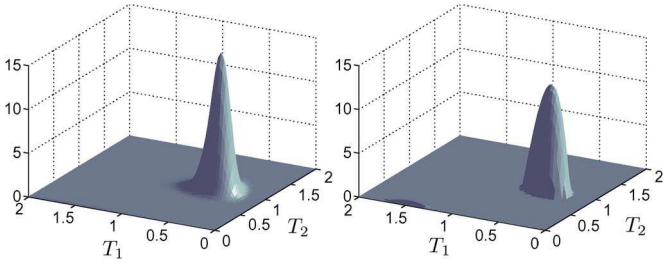


Fig. 7. Dataset A: Reconstructed spectra with optimal setting of λ for ME (left) and TIK^+ (right) regularization ($\mathrm{SNR} = 10$ dB and $\bar{m}_i = 5$).

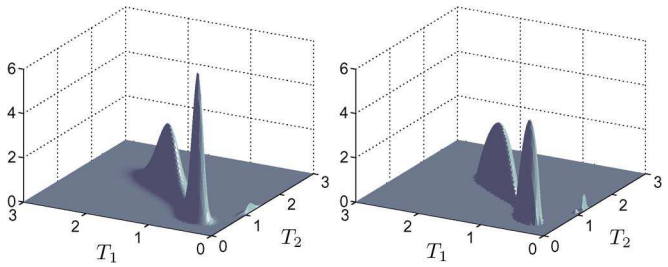


Fig. 8. Dataset B: Reconstructed spectra with optimal setting of λ for ME (left) and TIK (right) regularization ($\mathrm{SNR} = 10$ dB and $\bar{m}_i = 5$).

TABLE III
REGULARITY OF RECONSTRUCTED SPECTRA FOR ME AND TIK^+ RECONSTRUCTIONS ($\mathrm{SNR} = 10$ dB AND $\bar{m}_i = 5$). Δ IS THE FIRST-ORDER DIFFERENCE MATRIX

	Dataset A		Dataset B	
	ME	TIK^+	ME	TIK^+
$\|\Delta \boldsymbol{s}\|$	51.9	57.8	23.6	26.1
$\|\Delta \boldsymbol{s}\| / \|\boldsymbol{s}\|$	0.5484	0.5891	0.5256	0.5324

[15] and ME optimization [12], [44], a frequently used strategy is the Chi-square approach.

Given measurements $\boldsymbol{Y}$ and an estimate of the noise standard deviation $\hat{\sigma}$, statistical considerations state that the error

$$\chi^2(\boldsymbol{S}) = \left\| \boldsymbol{K}_1 \boldsymbol{S} \boldsymbol{K}_2^{\mathrm{t}} - \boldsymbol{Y} \right\|_F^2 / \hat{\sigma}^2 \tag{32}$$

follows a Chi-square distribution [45], [46]. In the limit of a large number of independent measurements $m_1 m_2$, the latter tends to a standard normal distribution with expected value $m_1 m_2$ and variance $2 m_1 m_2$.

Thus, a classical method for setting the regularization parameter and avoiding over-smoothed reconstructions [44], [46] is to find the value of λ allowing to reach

$$\chi^2_{\mathrm{aim}} = m_1 m_2 - \sqrt{2 m_1 m_2}. \tag{33}$$

However, when the noise level is high or when the estimation of σ is too rough, one can have $\chi^2(\lambda) > \chi^2_{\mathrm{aim}}$ for all values of the regularization parameter so that the Chi-square test cannot be achieved.

An alternative approach, based on the S-curve [47], consists in choosing λ such that its reduction does not lead to a significant decrease in $\chi^2(\lambda)$

$$\frac{\partial \log_{10} \chi^2(\lambda)}{\partial \log_{10} \lambda} \ll 1. \tag{34}$$

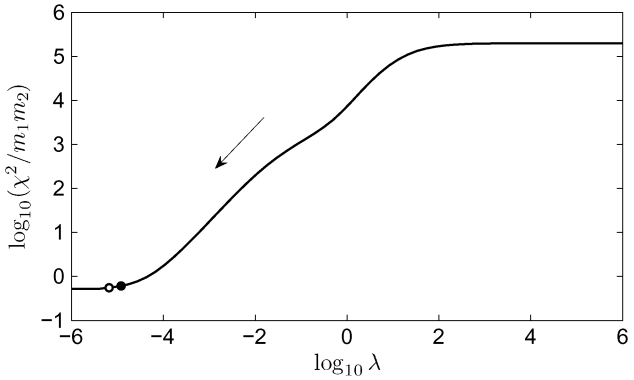


Fig. 9. Dataset A ($\mathrm{SNR} = 10$ dB and $\bar{m}_i = 5$): Estimation of the regularization parameter for ME reconstruction. The fulfillment of the Chi-square test (33) and the S-curve test (34) are illustrated by black and white dots respectively. According to Algorithm 3, the result of the Chi-square test is retained.

TABLE IV
REGULARIZATION PARAMETER ESTIMATES (λ_Q, λ_S) OBTAINED RESPECTIVELY BY MINIMIZING Q AND BY APPLYING THE ALGORITHM SUMMARIZED IN ALGORITHM 3. ($\mathrm{SNR} = 10$ dB AND $\bar{m}_i = 5$). (a) DATASET A; (b) DATASET B

	$-\log_{10}\lambda_Q$	$-\log_{10}\lambda_S$	$Q(\lambda_Q)$	$Q(\lambda_S)$
ME	4.92	5.05	2.05	2.43
TIK^+	6.19	5.91	3.92	4.67

(a)

	$-\log_{10}\lambda_Q$	$-\log_{10}\lambda_S$	$Q(\lambda_Q)$	$Q(\lambda_S)$
ME	5.32	5.59	13.8	22.9
TIK^+	5.92	5.92	10.7	10.7

(b)

Here, we suggest to combine the two latter strategies for the determination of λ, as detailed in Algorithm 3 and Fig. 9. We emphasize that the minimizations (35) can be performed at very low cost by initializing the TN algorithm of Algorithm 1 with the solution at previous λ. Table IV illustrates the efficiency of the proposed scheme for finding λ.

Algorithm 3: Chi-Square Method for Regularization Parameter Estimation

Require: Initial values $\boldsymbol{s}_0 \succeq 0$, λ_0, parameter $\theta \in (0,1)$ and accuracy η

Ensure: ME resolution with Chi-square tuned λ

while (33) **and** (34) do not hold **do**

Using Table I, compute

$$\hat{\boldsymbol{S}} = \arg\min L(\boldsymbol{S}) + \lambda_n R(\boldsymbol{S}). \tag{35}$$

Compute $\chi^2(\hat{\boldsymbol{S}})$ using (32).

$\lambda_{n+1} \leftarrow \theta\lambda_n$

end while

6) Pulse Angle Effect: In NMR experiments, the pulse angle Φ may not be set exactly to its desired value. This uncertainty introduces a potential error in the value of γ in the observation model. Let us first discuss the effect of an inexact value of this parameter on the reconstruction results. Several reconstructions using an observation model with $\gamma \neq 1$ have been performed. Fig. 10 shows the optimal value of the regularization parameter λ and the reconstruction error Q for different values of γ, for ME and TIK^+ algorithms. As expected, an error on the value of γ leads to a larger reconstruction error. Moreover, a larger value of λ has to be chosen to compensate the increase of the modelization error. We can conclude that the pulse angle parameter has an influence on the reconstruction results whatever the employed inversion algorithm.

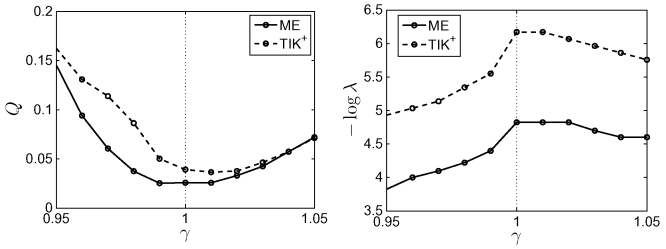


Fig. 10. Dataset A ($\mathrm{SNR} = 10$ dB, $\gamma = 1$, $\bar{m}_i = 5$): Sensitivity to a wrong estimation of γ in terms of reconstruction error Q (left) and optimal regularization parameter λ (right).

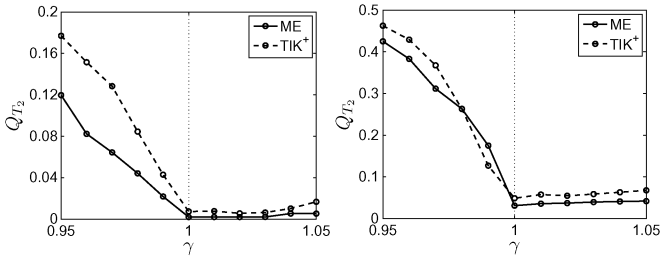


Fig. 11. Sensitivity to a wrong estimation of γ in terms of error Q_{T_2} between the T_2 marginalized spectra and the reference T_2 spectrum. (a) Dataset A; (b) dataset B.

7) Pulse Angle Estimation: In [47], some data preprocessing strategies are proposed to handle systematic errors, including pulse angle inaccuracy, in NMR experiments. An alternative strategy allowing to assess the pulse angle value is proposed here. The basic idea is to use the reconstructed T_2 spectrum, obtained from T_2 relaxation data, as a reference spectrum. Since these data are obtained for high values of τ_1, the underlying spectrum is not affected by the value of γ. After performing several 2-D reconstructions with different values of γ, we retain the pulse angle value maximizing the similarity between the marginalized T_2 spectrum and the reference T_2 spectrum.

Fig. 11 illustrates the relative Euclidian distance Q_{T_2} between the 1-D recovered T_2 spectrum and the marginalized T_2 spectra for several values of γ. The best matching is reached when γ equals its actual value, i.e., $\gamma = 1$.

B. Application to Experimental Data

Measurements have been performed on a plant matter sample (apple) to test the applicability of the proposed algorithm on experimental data. In the experiment, $m_1 = 50$ values of τ_1, nonuniformly spaced between 30 ms and 12 s were retained. In all cases, $m_2 = 10\,000$ echoes with a uniform time spacing of 800 μs between 600 μs and 8 s were acquired.

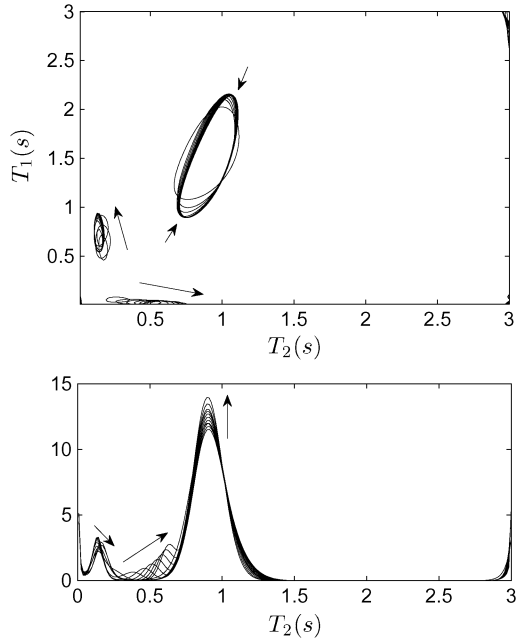


Fig. 12. 2-D ME spectra (top) from experimental data and 1-D distributions resulting from T_2 marginalization (bottom), for different values of the pulse angle parameter in the interval [0.9, 1]. The effect of increasing γ onto peak positions and amplitudes is indicated by arrows.

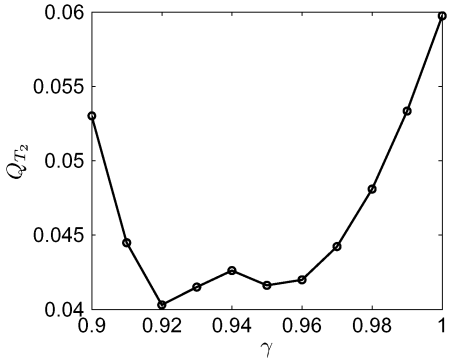


Fig. 13. Error between 1-D ME T_2 reconstruction and 2-D ME marginalized spectrum. The minimum is reached for $\gamma = 0.92$ which corresponds to $\Phi = 85°24'$.

The proposed algorithm was applied to reconstruct a spectrum with $N_1 = N_2 = 200$ values of T_1 and T_2 relaxation times, equally spaced between 25 ms and 3 s.

1) Reconstruction Algorithm Tuning: The lowest computation time was reached when using only one sub-iteration of MM line search and computing the preconditioner with TSVDs at rank $v = 7$. The proposed strategy in Algorithm 3 was used to set the regularization parameter.

2) Pulse Angle Parameter Setting: Fig. 12 summarizes the reconstruction results for different values of γ between 0.9 and 1. It can be noted that the positions and the amplitudes of some peaks are highly affected by the pulse angle value. Therefore, the reconstruction of a reliable spectrum requires the use of an accurate value of this parameter. The same strategy as that proposed in Section IV-A-7 is used to set the pulse angle value. According to Fig. 13, the retained value corresponds to $\Phi \approx 85°$ (i.e., $\gamma = 0.92$).

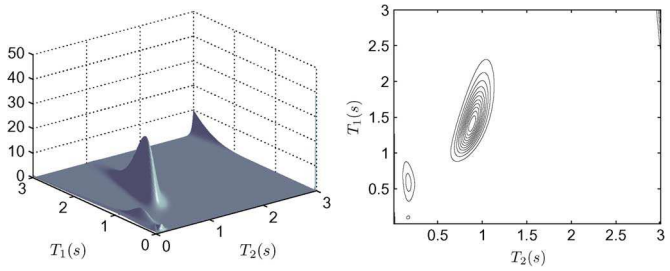
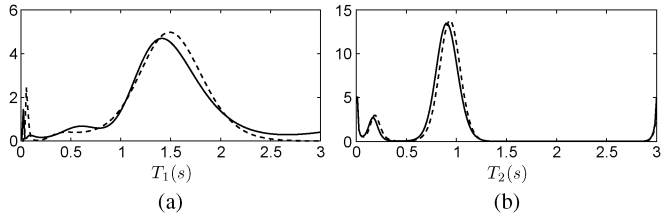


Fig. 14. Reconstructed spectrum from 2-D NMR experimental data with ME method.



(a) (b)

Fig. 15. 1-D distributions resulting from marginalization of the 2-D ME spectrum (solid line) or 1-D ME reconstruction (dashed line). (a) T_1 spectrum; (b) T_2 spectrum.

3) Comparison of Algorithms: Fig. 14 shows the reconstructed 2-D spectrum for $\gamma = 0.92$. It can be noted that this spectrum allows to analyze the correlation between T_1 and T_2 relaxation times. This correlation appears, for example, in the peak located around $[T_1 = 1.4\,\mathrm{s}, T_2 = 0.9\,\mathrm{s}]$. Such information is very useful to obtain the T_1/T_2 ratio which gives insights related to the molecular structure of the analyzed sample [7]. Concerning the reconstruction algorithm performances, the computation time was 59 s for 67 iterations and the final value of λ was $1.3 \cdot 10^{-4}$.

Since there is no ground truth regarding the T_1-T_2 correlation spectrum of the apple, we compare the 1-D distributions (T_1 and T_2) obtained by 1-D inversion with the 1-D distributions deduced by marginalization of the reconstructed 2-D distribution. It can be noted from Fig. 15 the similarity between the 1-D spectra which shows the relevance of the 2-D spectrum.

We also compare these results with the ones obtained by the TIK^+ algorithm of [47]. This algorithm was tuned with a compression rank $\tilde{m}_i = 10$ and the same strategy as in [47] was used to determine the regularization parameter. The algorithm requires a computation time of 11 s for 14 iterations and the final value of $\lambda = 2 \cdot 10^{-5}$. The reconstructed 2-D spectrum and the corresponding 1-D distributions are shown in Figs. 16 and 17. Even if the two reconstruction methods led to similar measurement data fit (98%), a visual comparison reveals significant differences between the two spectra shapes in terms of regularity and amplitude.

V. Conclusion

The reconstruction of a T_1-T_2 spectrum in NMR requires a numerical inversion of a 2-D Laplace transform. This is known to be an ill posed inverse problem. In this paper, we presented an efficient inversion method based on maximum entropy regularization and truncated Newton optimization. A second difficulty is related to the large scale of the 2-D model. To handle this problem, rather than compressing the data matrix, we rely on an

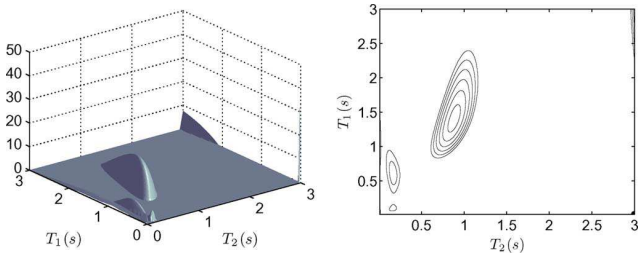


Fig. 16. Reconstructed spectrum from 2-D NMR experimental data with TIK^+ method.

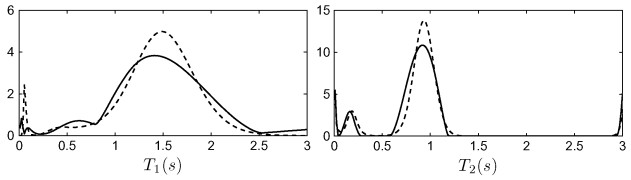


Fig. 17. 1-D distributions resulting from marginalization of the 2-D TIK^+ spectrum (solid line) or 1-D ME reconstruction (dashed line). (a) T_1 spectrum; (b) T_2 spectrum.

exact data model thanks to an iterative algorithm exploiting the separability of the convolution kernel. All required quantities such as gradient, Hessian-vector product are computed with reduced memory storage and computation time. Moreover, since the entropy criterion introduces a barrier in the criterion to minimize, an appropriate line search strategy is used. This procedure is fast and ensures the theoretical convergence of the truncated Newton algorithm. Finally, the convergence speed of the algorithm is increased by applying an adequate preconditioner using TSVDs of the convolution kernels. The applicability of the proposed method has been demonstrated through the processing of simulated and real data and a comparison with the constrained Tikhonov approach of [15]. Our conclusion is that the two methods produce reconstructions of similar quality. The constrained Tikhonov approach is noticeably faster, at the price of resorting to a data compression step that needs the tuning of two parameters. In contrast, our approach remains fast without data compression.

The processing of real data measurements allowed us to point out the difficulty of setting the pulse angle parameter appearing in the observation model. We have shown that an inaccurate value of this parameter tends to produce a significant error in peak positions and amplitudes. Up to our knowledge, this point is only partially addressed in NMR literature where data preprocessing strategies are suggested. Therefore, we proposed an original strategy allowing to estimate this parameter. Although this strategy seems to give satisfying results in our tests, further investigations and experiments would be needed to validate this approach. Another perspective would be to build a criterion allowing to reduce the number of peaks in the reconstructed spectrum or to propose a strategy based on a parametric 2-D reconstruction where the number of peaks will be imposed.

From the methodological point of view, we restricted our analysis to the case of separable convolution kernels. However, in some NMR measurement models [7], the separability is no longer valid. It would be interesting to generalize our approach by considering the case where the observation model can be expressed as a linear superposition of several separable kernels.

APPENDIX

A. Interpretation of BRD Algorithm Using Legendre–Fenchel Duality

Let us consider the constrained minimization problem

$$\min_{\boldsymbol{s}\geqslant 0}\left\{L(\boldsymbol{s})=\frac{1}{2}\|\boldsymbol{K}\boldsymbol{s}-\boldsymbol{y}\|^2+\frac{\lambda}{2}\|\boldsymbol{s}\|^2\right\}. \tag{36}$$

The BRD algorithm [10] is based on the equivalence between the KKT conditions of problem (36) and the following unconstrained problem

$$\min_{\boldsymbol{c}\in\mathbb{R}^m}\left\{\chi(\boldsymbol{c})=\frac{1}{2}\boldsymbol{c}^{\mathrm{t}}\left(G(\boldsymbol{c})+\lambda\boldsymbol{I}\right)\boldsymbol{c}-\boldsymbol{c}^{\mathrm{t}}\boldsymbol{y}\right\} \tag{37}$$

with the reparametrization $\boldsymbol{s}=\max(\boldsymbol{0},\boldsymbol{K}^{\mathrm{t}}\boldsymbol{c})$ and

$$G(\boldsymbol{c})=\boldsymbol{K}^{\mathrm{t}}\mathrm{Diag}\left(\mathbb{H}(\boldsymbol{K}^{\mathrm{t}}\boldsymbol{c})\right)\boldsymbol{K} \tag{38}$$

where $\mathbb{H}$ denotes a componentwise unit step function that takes the value zero for negative or zero arguments and one for positive arguments. Let us show that this equivalence can also be obtained from the Legendre–Fenchel conjugacy theory (see [19] for a reminder on Legendre–Fenchel theory).

First, let us introduce the Legendre–Fenchel conjugate f^* of the quadratic $f(\boldsymbol{u})=(1/2)\|\boldsymbol{u}-\boldsymbol{y}\|^2$, i.e.,

$$f^*(\boldsymbol{u})=\sup_{\boldsymbol{v}}\left(\boldsymbol{v}^{\mathrm{t}}\boldsymbol{u}-\frac{1}{2}\|\boldsymbol{v}-\boldsymbol{y}\|^2\right)=\frac{1}{2}\|\boldsymbol{u}\|^2+\boldsymbol{y}^{\mathrm{t}}\boldsymbol{u}. \tag{39}$$

According to the conjugacy theorem [19, Prop. 7.1.1],

$$L(\boldsymbol{s})=\sup_{\boldsymbol{u}\in\mathbb{R}^m}\left(\boldsymbol{s}^{\mathrm{t}}\boldsymbol{K}^{\mathrm{t}}\boldsymbol{u}-f^*(\boldsymbol{u})\right)+\frac{\lambda}{2}\|\boldsymbol{s}\|^2. \tag{40}$$

Moreover, according to the minimax theorem [19, Prop. 2.6.2], (40) implies

$$\begin{aligned}\min_{\boldsymbol{s}\geqslant 0}L(\boldsymbol{s})&=\max_{\boldsymbol{u}\in\mathbb{R}^m}\min_{\boldsymbol{s}\geqslant 0}\left(\boldsymbol{s}^{\mathrm{t}}\boldsymbol{K}^{\mathrm{t}}\boldsymbol{u}-f^*(\boldsymbol{u})+\frac{\lambda}{2}\|\boldsymbol{s}\|^2\right)\\&=\max_{\boldsymbol{u}\in\mathbb{R}^m}\left(\varphi(\boldsymbol{u})-f^*(\boldsymbol{u})\right)\end{aligned} \tag{41}$$

where

$$\varphi(\boldsymbol{u})=\min_{\boldsymbol{s}\geqslant 0}\left(\boldsymbol{s}^{\mathrm{t}}\boldsymbol{K}^{\mathrm{t}}\boldsymbol{u}+\frac{\lambda}{2}\|\boldsymbol{s}\|^2\right). \tag{42}$$

The minimization problem (42) is convex, separable and the following expression of the minimizer is easy to derive

$$\boldsymbol{s}^*(\boldsymbol{u})=\frac{1}{\lambda}\max(\boldsymbol{0},-\boldsymbol{K}^{\mathrm{t}}\boldsymbol{u}) \tag{43}$$

where max is to be considered component-wise. Moreover, we have

$$\varphi(\boldsymbol{u})=\left(\boldsymbol{s}^*(\boldsymbol{u})\right)^{\mathrm{t}}\boldsymbol{K}^{\mathrm{t}}\boldsymbol{u}+\frac{\lambda}{2}\left\|\boldsymbol{s}^*(\boldsymbol{u})\right\|^2=\frac{1}{2}\left(\boldsymbol{s}^*(\boldsymbol{u})\right)^{\mathrm{t}}\boldsymbol{K}^{\mathrm{t}}\boldsymbol{u} \tag{44}$$

the latter expression being a consequence of $(\max(0,x))^2=x\max(0,x)$ for all $x\in\mathbb{R}$. Finally, given (39), (43), and (44), (41) also reads

$$\begin{aligned}&\min_{\boldsymbol{s}\geqslant 0}L(\boldsymbol{s})\\&\quad=\max_{\boldsymbol{u}\in\mathbb{R}^m}\left(-\frac{1}{2\lambda}\left(\max(\boldsymbol{0},-\boldsymbol{K}^{\mathrm{t}}\boldsymbol{u})\right)^{\mathrm{t}}\boldsymbol{K}^{\mathrm{t}}\boldsymbol{u}+\frac{1}{2}\|\boldsymbol{u}\|^2+\boldsymbol{y}^{\mathrm{t}}\boldsymbol{u}\right)\\&\quad=-\lambda\min_{\boldsymbol{c}\in\mathbb{R}^m}\chi(\boldsymbol{c})\end{aligned}$$

where the last identity is obtained using the change of variable $\boldsymbol{c} = -\boldsymbol{u}/\lambda$. Thus, (36) and (37) are equivalent through Legendre–Fenchel duality, and $\boldsymbol{c}^*$ minimizes $\chi(\boldsymbol{c})$ in $\mathbb{R}^m$ if and only if $\boldsymbol{s}^* = \max(\boldsymbol{0}, \boldsymbol{K}^{\mathrm{t}}\boldsymbol{c}^*)$ minimizes $\boldsymbol{L}(\boldsymbol{s})$ in $\mathbb{R}^m_+$.

B. Expression of the Majorant Function $h^j(\cdot, \alpha^j)$ and of Its Minimizer

The majorant function $h^j(\cdot, \alpha^j)$ is piecewise defined, whether $\alpha \in (\alpha_-; \alpha^j]$ or $\alpha \in [\alpha^j; \alpha_+)$. In both cases, it takes the following form:

$$h^j(\alpha, \alpha^j) = \ell(\alpha^j) + (\alpha - \alpha^j)\dot{\ell}(\alpha^j) + \frac{1}{2}m^j(\alpha - \alpha^j)^2 + \gamma^j\left[(\bar{\alpha}^j - \alpha^j)\log\frac{\bar{\alpha}^j - \alpha^j}{\bar{\alpha}^j - \alpha} - \alpha + \alpha^j\right] \quad (45)$$

while the expressions of parameters $\bar{\alpha}^j$, m^j, and γ^j are specific to each case. The notation $\dot{\ell}$ refers to the derivative of ℓ, also defined as $\dot{\ell}(\alpha) = \boldsymbol{d}_k^T \nabla L(\boldsymbol{s}_k + \alpha \boldsymbol{d}_k)$.

1) Case $\alpha \in (\alpha_-; \alpha^j]$

$$\begin{cases} \bar{\alpha}^j = \alpha_- \\ m^j = \boldsymbol{d}_k^{\mathrm{t}}\boldsymbol{K}^{\mathrm{t}}\boldsymbol{K}\boldsymbol{d}_k + \lambda\sum_{i|d_{k,i}<0}\phi_i(\alpha^j) \\ \gamma^j = \lambda(\alpha_- - \alpha^j)\sum_{i|d_{k,i}>0}\phi_i(\alpha^j) \end{cases} \quad (46)$$

2) Case $\alpha \in [\alpha^j; \alpha_+)$

$$\begin{cases} \bar{\alpha}^j = \alpha_+ \\ m^j = \boldsymbol{d}_k^{\mathrm{t}}\boldsymbol{K}^{\mathrm{t}}\boldsymbol{K}\boldsymbol{d}_k + \lambda\sum_{i|d_{k,i}>0}\phi_i(\alpha^j) \\ \gamma^j = \lambda(\alpha_+ - \alpha^j)\sum_{i|d_{k,i}<0}\phi_i(\alpha^j) \end{cases} \quad (47)$$

where $\phi_i(\alpha) = d_{k,i}^2/(s_i + \alpha d_{k,i})$ in both cases.

The minimizer of $h^j(\cdot, \alpha^j)$ can be expressed as follows:

$$\alpha^j + \operatorname{sign}\left(\dot{\ell}(\alpha^j)\right)\frac{2|A_3|}{|A_2| + \sqrt{A_2^2 - 4A_1A_3}} \quad (48)$$

with

$$\begin{cases} A_1 = -m^j \\ A_2 = \gamma^j - \dot{\ell}(\alpha^j) + m^j(\bar{\alpha}^j - \alpha^j) \\ A_3 = (\bar{\alpha}^j - \alpha^j)\dot{\ell}(\alpha^j). \end{cases} \quad (49)$$

REFERENCES

[1] D. Canet, J.-C. Boubel, and E. Canet-Soulas, *La RMN. Concepts, Méthodes et Applications*, 2nd ed. Paris, France: Dunod, 2002.

[2] R. R. Ernst, G. Bodenhausen, and A. Wokaun, *Principles of Nuclear Magnetic Resonance in One and Two Dimensions*, ser. International Series of Monographs on Chemistry, 2nd ed. Oxford, NY: Oxford Univ. Press, 1997.

[3] F. J. M. Van de Ven, *Multidimensional NMR in Liquids. Basic Principles and Experimental Methods*. New York: Wiley-VCH, 1995.

[4] F. Mariette, J. P. Guillement, C. Tellier, and P. Marchal, "Continuous relaxation time distribution decomposition by MEM," *Signal Treat. Signal Anal. NMR*, pp. 218–234, 1996.

[5] R. Lamanna, "On the inversion of multicomponent NMR relaxation and diffusion decays in heterogeneous systems," *Concepts Magn. Reson. Part A*, vol. 26, no. 2, pp. 78–90, 2005.

[6] A. E. English, K. P. Whittall, M. L. G. Joy, and R. M. Henkelman, "Quantitative two-dimensional time correlation relaxometry," *Magn. Reson. Med.*, vol. 22, pp. 425–434, 1991.

[7] M. D. Hürlimann and L. Venkataramanan, "Quantitative measurement of two-dimensional distribution functions of diffusion and relaxation in grossly inhomogeneous fields," *J. Magn. Reson.*, vol. 157, pp. 31–42, 2002.

[8] M. Fleury and J. Soualem, "Quantitative analysis of diffusional pore coupling from T_2-store- T_2 NMR experiments," *J. Colloid Interface Sci.*, vol. 336, pp. 250–259, 2009.

[9] E. Sternin, "Use of inverse theory algorithms in the analysis of biomembrane NMR data," in *Methods in Membrane Lipids*, ser. Methods in Molecular Biology, A. M. Dopico, Ed. Totowa, NJ: Humana Press, 2008, vol. 400, pp. 103–125.

[10] J. P. Butler, J. A. Reeds, and S. V. Dawson, "Estimating solutions of first kind integral equations with nonnegative constraints and optimal smoothing," *SIAM J. Numer. Anal.*, vol. 18, no. 3, pp. 381–397, Jun. 1981.

[11] E. D. Laue, J. Skilling, J. Staunton, S. Sibisi, and R. G. Brereton, "Maximum entropy method in nuclear magnetic resonance spectroscopy," *J. Magn. Reson.*, vol. 62, no. 3, pp. 437–452, 1985.

[12] J. Skilling and R. K. Bryan, "Maximum entropy image reconstruction: General algorithm," *Month. Not. Roy. Astron. Soc.*, vol. 211, pp. 111–124, 1984.

[13] D. W. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *SIAM J. Appl. Math.*, vol. 11, pp. 431–441, 1963.

[14] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. Baltimore, MD: The Johns Hopkins Univ. Press, 1996.

[15] L. Venkataramanan, Y. Q. Song, and M. D. Hürlimann, "Solving Fredholm integrals of the first kind with tensor product structure in 2 and 2.5 dimensions," *IEEE Trans. Signal Process.*, vol. 50, no. 5, pp. 1017–1026, May 2002.

[16] G. Bruckner and J. Cheng, "Tikhonov regularization for an integral equation of the first kind with logarithmic kernel," Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany, Tech. rep., 1998 [Online]. Available: http://www.wias-berlin.de/publications/preprints/523/wias_preprints_523.pdf

[17] Y.-W. Chiang, P. P. Borbat, and J. H. Freed, "Maximum entropy: A complement to Tikhonov regularization for determination of pair distance distributions by pulsed ESR," *J. Magn. Reson.*, vol. 177, no. 2, pp. 184–196, Dec. 2005.

[18] J. Nocedal and S. J. Wright, *Numerical Optimization*. New York: Springer-Verlag, 1999.

[19] D. P. Bertsekas, *Convex Analysis and Optimization*, 1st ed. Belmont, MA: Athena Scientific, 2003.

[20] P. P. B. Eggermont, "Maximum entropy regularization for Fredholm integral equations of the first kind," *SIAM J. Math. Anal.*, vol. 24, no. 6, pp. 1557–1576, 1993.

[21] R. Gordon, R. Bender, and G. T. Herman, "Algebric reconstruction techniques (ART) for three-dimensional electron microscopy and X-ray photography," *J. Theor. Biol.*, vol. 29, pp. 471–481, 1970.

[22] C. A. Johnson and D. McGarry, "Maximum entropy reconstruction methods in electron paramagnetic resonance imaging," *Ann. Oper. Res.*, vol. 119, pp. 101–118, 2003.

[23] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes: The Art of Scientific Computing*, 3rd ed. New York: Cambridge Univ. Press, 1992.

[24] E. Chouzenoux, S. Moussaoui, J. Idier, and F. Mariette, "Reconstruction d'un spectre RMN 2-D par maximum d'entropie," presented at the GRETSI, Dijon, France, Sep. 2009.

[25] E. Chouzenoux, S. Moussaoui, J. Idier, and F. Mariette, "Optimization of a maximum entropy criterion for 2-D nuclear magnetic resonance reconstruction," presented at the IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP), Dallas, TX, Mar. 2010.

[26] R. Dembo, S. C. Eisenstat, and S. Steihaug, "Inexact Newton methods," *SIAM J. Numer. Anal.*, vol. 19, no. 2, pp. 400–408, Apr. 1982.

[27] S. G. Nash, "A survey of truncated-Newton methods," *J. Comput. Appl. Math.*, vol. 124, pp. 45–59, 2000.

[28] S. G. Nash and A. Sofer, "On the complexity of a practical interior-point method," *SIAM J. Optim.*, vol. 8, no. 3, pp. 833–849, 1998.

[29] S. Bellavia, "Inexact interior-point method," *J. Optim. Theory Appl.*, vol. 96, pp. 109–121, 1998.

[30] J. J. Moré and D. J. Thuente, "Line search algorithms with guaranteed sufficient decrease," *ACM Trans. Math. Softw.*, vol. 20, no. 3, pp. 286–307, 1994.

[31] J. Sun and J. Zhang, "Global convergence of conjugate gradient methods without line search," *Ann. Oper. Res.*, vol. 103, pp. 161–173, Mar. 2001.

[32] C. Labat and J. Idier, "Convergence of conjugate gradient methods with a closed-form stepsize formula," *J. Optim. Theory Appl.*, vol. 136, no. 1, pp. 43–60, Jan. 2008.

[33] E. Chouzenoux, S. Moussaoui, and J. Idier, "A majorize-minimize line search algorithm for barrier function optimization," presented at the EUSIPCO, Glasgow, U.K., Aug. 2009.

[34] E. Chouzenoux, S. Moussaoui, and J. Idier, "A new line search method for barrier functions with strong convergence properties," IRCCyN, Nantes, France, Tech. Rep., 2009 [Online]. Available: http://hal.archives-ouvertes.fr/IRCCYN-ADTSI

[35] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *Amer. Stat.*, vol. 58, no. 1, pp. 30–37, Feb. 2004.
[36] M. Jacobson and J. Fessler, "An expanded theoretical treatment of iteration-dependent majorize-minimize algorithms," *IEEE Trans. Image Process.*, vol. 16, no. 10, pp. 2411–2422, Oct. 2007.
[37] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed. Belmont, MA: Athena Scientific, 1999.
[38] Z.-J. Shi, "Convergence of line search methods for unconstrained optimization," *Appl. Math. Comput.*, vol. 157, pp. 393–405, 2004.
[39] P. C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*. Philadelphia, PA: SIAM, 1998.
[40] Y. Saad, *Iterative Methods for Sparse Linear Systems*. Philadelphia, PA: SIAM, 2003.
[41] K. Chen, *Matrix Preconditioning Techniques and Applications*, 1st ed. Cambridge, U.K.: Cambridge Univ. Press, 2005.
[42] R. Piché, "Regularization operators for multidimensional inverse problems with Kronecker product structure," presented at the ECCOMAS, Jyväskylä, Finland, Jul. 2004.
[43] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, no. 3, pp. 211–218, 1936.
[44] C. Pichon and E. Thiébaut, "Non-parametric reconstruction of distribution functions from observed galactic discs," *Month. Not. Roy. Astron. Soc.*, vol. 301, no. 2, pp. 419–434, 1998.
[45] H. Trussell, "Convergence criteria for iterative restoration methods," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 31, no. 1, pp. 129–136, 1983.
[46] N. Galatsanos and A. Katsaggelos, "Methods for choosing the regularization parameter and estimating the noise variance in image restoration and their relation," *IEEE Trans. Image Process.*, vol. 1, no. 3, pp. 322–336, Jul. 1992.
[47] Y. Q. Song, L. Venkataramanan, M. D. Hürlimann, M. Flaum, P. Frulla, and C. Straley, "T1-T2 correlation spectra obtained using a fast two-dimensional Laplace inversion," *J. Magn. Reson.*, vol. 154, pp. 261–268, 2002.

Émilie Chouzenoux received the M.Sc. degree in signal processing and the Engineering degree from École Centrale, Nantes, France, in 2007. She is currently working towards the Ph.D. degree in signal processing at the Institut de Recherche en Communications et Cybernétique, Nantes, France (IRCCYN, UMR CNRS 6597).

Her research interests are in optimization algorithms for large-scale problems of image and signal reconstruction.

Saïd Moussaoui received the State Engineering degree from Ecole Nationale Polytechnique, Algiers, Algeria, in 2001, and the Ph.D. degree in automatic control and signal processing from the Université Henri Poincaré, Nancy, France, in 2005.

He is currently Assistant Professor at the Ecole Centrale de Nantes. Since September 2006, he has been with the Institut de Recherche en Communications et Cybernétique, Nantes, France (IRCCYN, UMR CNRS 6597). His research interests are in statistical signal and image processing, including source separation, Bayesian estimation, and their applications.

Jérôme Idier (M'09) was born in France in 1966. He received the Diploma degree in electrical engineering from École Supérieure d'Électricité, Gif-sur-Yvette, France, in 1988 and the Ph.D. degree in physics from University of Paris-Sud, Orsay, France, in 1991.

Since 1991, he joined the Centre National de la Recherche Scientifique. He is currently a Senior Researcher at the Institut de Recherche en Communications et Cybernétique, Nantes, France. His major scientific interests are in probabilistic approaches to inverse problems for signal and image processing.

Dr. Idier is serving as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING and for the *Journal of Electronic Imaging*, co-published by SPIE and IS&T.

François Mariette received the Ph.D. degree in physical chemistry from the University of Nantes, France, in 1992.

From 1993 to 1999, he was a Researcher in the Food Process Engineering research unit at the Cemagref research institute. In 2001, he was a Visiting Scientist in the Physical Chemistry 1 Department at the Lund University, Sweden. Since 2004, he has been Research Director and Leader of the NMR/MRI research team in the Food Process Engineering research unit at Cemagref. Since 2006, he has been the coordinator of PRISM (Structural and metabolic imaging and spectroscopy research platform), Rennes, France, a national research platform under the responsibility of the University of Rennes 1, INRA, and Cemagref. He is permanent member of the scientific committee of a French association of dairy companies. He is author of 83 papers in scientific journals and chapters in textbook, and of 66 presentations to international conference in the field of NMR and MRI applied to food science and food processing.

A.4 Fast constrained least squares spectral unmixing using primal-dual interior point optimization

E. Chouzenoux, M. Legendre, **S. Moussaoui** et J. Idier, *IEEE Journal of Selected Togics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 1, pp. 59-69, 2014 *Ce papier est consacré au démélange supervisé d'images hyperspectrales par minimisation d'un critère de type moindres carrés sous des contraintes linéaires d'égalité et ou d'inégalités. La particularité est d'adopter un algorithme primal-dual des points intérieurs et de discuter les différentes stratégies d'implémentation CPU ou GPU permettant de réduire le temps de calcul et l'occupation mémoire.*

Fast Constrained Least Squares Spectral Unmixing Using Primal-Dual Interior-Point Optimization

Emilie Chouzenoux, *Member, IEEE*, Maxime Legendre, Saïd Moussaoui, *Member, IEEE*, and Jérôme Idier, *Member, IEEE*

***Abstract*—Hyperspectral data unmixing aims at identifying the components (endmembers) of an observed surface and at determining their fractional abundances inside each pixel area. Assuming that the spectral signatures of the surface components have been previously determined by an endmember extraction algorithm, or to be part of an available spectral library, the main problem is reduced to the estimation of the fractional abundances. For large hyperspectral image data sets, the estimation of the abundance maps requires the resolution of a large-scale optimization problem subject to linear constraints such as non-negativity and sum less or equal to one. This paper proposes a primal-dual interior-point optimization algorithm allowing a constrained least squares estimation approach. In comparison with existing methods, the proposed algorithm is more flexible since it can handle any linear equality and/or inequality constraint and has the advantage of a reduced computational cost. It also presents an algorithmic structure suitable for a parallel implementation on modern intensive computing devices such as Graphics Processing Units (GPU). The implementation issues are discussed and the applicability of the proposed approach is illustrated with the help of examples on synthetic and real hyperspectral data.**

***Index Terms*—Spectral unmixing, constrained least squares, interior-point optimization, primal-dual algorithm, GPU computing.**

I. Introduction

HYPERSPECTRAL imaging corresponds to the measurement of the incident light reflection at the ground surface of an observed scene in several contiguous spectral bands. Despite of the high spatial resolution that can be attained by recent imaging devices, the surface area covered by any pixel of the image may contain different components. Therefore, the measured reflectance spectrum in each pixel can be explained as a mixture of the individual component reflectance spectra weighted by the proportion (abundance) of each component in this pixel area.

Unmixing hyperspectral data aims at the identification of the observed surface components (endmembers) and the determination of their fractional abundances inside each pixel area [1], [2]. Fast hyperspectral data unmixing approaches are supervised, by assuming that the endmember spectra are part of an available spectral library or can be provided by an endmember extraction algorithm [3], [4]. Then, the remaining step of the unmixing is the estimation of the fractional abundances. Actually, there is an increasing interest to joint estimation methods based either on non-negative source separation [5], [6] or constrained non-negative matrix factorization [7], [8]. However, the purpose of this paper is to focus on the second step of the supervised approach with the aim to present a fast computation method adapted to the case of large data sets.

Usual algorithms for solving the spectral unmixing problem consist in minimizing a data fitting measure (generally a least squares criterion) under the physical constraints of non-negativity and sum-to-one. For instance, the former constraint leads to the *non-negative least squares* algorithm (NNLS) [9], [10], and the latter is handled by the *sum-to-one constrained least squares* (SCLS) method [11]. Both constraints are accounted for by the *fully constrained least squares* (FCLS) algorithm [12]. In [13], a Bayesian inference algorithm incorporating jointly these constraints is proposed. It is based on Monte Carlo Markov chain methods and offers the advantage of estimating the number of components. However, all these mentioned methods suffer from a significant increase of the computation time in the case of large data sets (in terms of image size, number of components or number of spectral bands). In order to reduce the computation time, many recent contributions have investigated the use of parallel computing tools [14] such as graphics processing units (GPUs) [15] and FPGA based-design [16]. A geometrical formulation of the abundance estimation step has been recently proposed in [17], the computation cost being reduced by retrieving some quantities computed during the endmember extraction step or by using simplex projection methods [18]. However, the geometrical formulation is restricted to the case of full-additivity and is not suitable for general linear constraints such as partial additivity (sum less than or equal to one) or bound constraints on the abundances. In [19], a modern convex optimization approach based on the alternating method of multipliers [20] was adopted for solving the constrained optimization problem arising in spectral unmixing.

In this paper, we propose a new flexible spectral unmixing algorithm based on constrained least squares estimation and interior-point optimization [21], [22]. The main originality of our approach is to exploit the potential of primal-dual interior-point techniques, which have shown their efficiency for solving large-scale constrained signal and image processing problems [23], [24]. The proposed optimization method allows to minimize any convex objective function under equality (e.g., sum-to-one) and inequality (e.g., non-negativity or

Manuscript received December 21, 2012; revised March 26, 2013, and May 17, 2013; accepted May 29, 2013. Date of publication July 02, 2013; date of current version December 18, 2013. This work was supported by the région Pays de la Loire (France).

E. Chouzenoux is with LIGM (CNRS UMR 8049), University Paris Est Marne-La-Vallée, Paris, France (corresponding author, e-mail: emilie.chouzenoux@univ-mlv.fr).

M. Legendre, S. Moussaoui, and J. Idier are with IRCCyN (CNRS UMR 6597), Ecole Centrale Nantes, France.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTARS.2013.2266732

sum-less-than-one) constraints. From the numerical optimization point of view, the choice of a primal-dual interior-point optimization scheme leads to an algorithm that can be implemented efficiently using modern parallel computing tools such as GPUs.

The rest of this paper is organized as follows. Section II formulates the constrained optimization problem arising in spectral unmixing. Section III presents the adopted interior-point optimization scheme for the estimation of the abundance maps. Its implementation issues accounting for memory storage and computing time are discussed in Section IV. Finally, Section V illustrates the performances of the proposed approach in terms of computation time and unmixing accuracy, through applications to both synthetic and real data.

II. Problem Statement

Let us consider N pixels of a hyperspectral image acquired in L spectral bands and assume a linear mixing model. This linear model is widely accepted in many practical situations since it offers a first-order approximation of the radiative transfer model [25]. According to this model, the observed spectrum $\boldsymbol{y}_n \in \mathbb{R}^L$ in the n-th pixel is explained as a linear combination of P endmember spectra and corrupted by an additive noise $\boldsymbol{\epsilon}_n$,

$$\boldsymbol{y}_n = \boldsymbol{S}\,\boldsymbol{a}_n + \boldsymbol{\epsilon}_n \tag{1}$$

where $\boldsymbol{S} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_P] \in \mathbb{R}^{L\times P}$ contains the P endmember spectra and $\boldsymbol{a}_n = [a_{n,1}, \ldots, a_{n,P}]^{\mathrm{t}} \in \mathbb{R}^P$ is the vector of endmember abundances in the n-th pixel. Using matrix notations, the mixing model is rewritten as,

$$\boldsymbol{Y} = \boldsymbol{S}\,\boldsymbol{A} + \boldsymbol{E} \tag{2}$$

where $\boldsymbol{Y} \in \mathbb{R}^{L\times N}$ is the observation data matrix, $\boldsymbol{A} \in \mathbb{R}^{P\times N}$ the fractional abundance matrix and $\boldsymbol{E} \in \mathbb{R}^{L\times N}$ the measurement noise.

The abundance matrix should satisfy the non-negativity constraint

$$(\forall\, n \in \{1, \ldots, N\})\,(\forall\, p \in \{1, \ldots, P\}) \quad A_{pn} \geqslant 0. \tag{3}$$

This constraint will be denoted as $\boldsymbol{A} \geqslant \boldsymbol{0}$. Moreover, under the assumption that all the endmembers comprising the pixel spectrum in $\boldsymbol{Y}$ are present in the columns of $\boldsymbol{S}$, the abundances coefficients should satisfy the full additivity constraint,

$$(\forall\, n \in \{1, \ldots, N\}) \quad \sum_{p=1}^{P} A_{pn} = 1 \tag{4}$$

which can be summarized by $\mathbf{1}_P^{\mathrm{t}}\boldsymbol{A} = \mathbf{1}_N^{\mathrm{t}}$, where $\mathbf{1}_N^{\mathrm{t}}$ denotes a vector of $\mathbb{R}^N$ with all entries equal to one.

When the set of endmembers is incomplete, or when the pixel area are subject to illumination variability or attenuation, only partial additivity should be required, i.e.:

$$(\forall\, n \in \{1, \ldots, N\}) \quad \sum_{p=1}^{P} A_{pn} \leqslant 1 \tag{5}$$

which can be noted shortly by $\mathbf{1}_P^{\mathrm{t}}\boldsymbol{A} \leqslant \mathbf{1}_N^{\mathrm{t}}$.

The estimation of $\boldsymbol{A}$ given $\boldsymbol{S}$ and $\boldsymbol{Y}$ is firstly formulated as the minimization of a convex criterion $F(\,\cdot\,)$, under linear inequality constraints such as non-negativity and partial additivity. Then, the case of the sum-to-one constraint is addressed. Finally, an interior-point algorithm based on a primal-dual approach is proposed for the resolution of the constrained optimization problem.

A. Criterion Formulation

The criterion $F(\,\cdot\,)$ to minimize results from the statistical modeling of the observation process and the sought abundances properties. Adopting the well-known least squares approach leads to define $F(\,\cdot\,)$ as the quadratic function:

$$F(\boldsymbol{A}) = \frac{1}{2}\sum_{\ell=1}^{L}\sum_{n=1}^{N}((\boldsymbol{S}\boldsymbol{A})_{\ell n} - Y_{\ell n})^2. \tag{6}$$

In a statistical estimation framework, (6) corresponds to the neg-log-likelihood associated to a spatially and spectrally uncorrelated Gaussian noise $\boldsymbol{E}$.

Note that the proposed approach can be adapted to a wider class of convex criteria which can be expressed as:

$$F(\boldsymbol{A}) = \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{S}\boldsymbol{a}_n - \boldsymbol{y}_n)^{\mathrm{t}}\boldsymbol{\Sigma}_n^{-1}(\boldsymbol{S}\boldsymbol{a}_n - \boldsymbol{y}_n) + \sum_{n=1}^{N}\varphi(\boldsymbol{a}_n), \tag{7}$$

where $\boldsymbol{\Sigma}_n \in \mathbb{R}^{L\times L}$ is a positive spectral covariance matrix and φ is a convex regularization function. However, in the sequel, the presentation will be focused on the case of the least squares criterion (6) since it is widely used in hyperspectral imaging with reflectance spectroscopy.

B. Constraint Formulation

We focus on the following general formulation of the constrained optimization problem for the estimation of the abundances maps:

$$\min_{\boldsymbol{A}\in\mathbb{R}^{P\times N}} F(\boldsymbol{A}) \quad \text{s.t.} \quad \boldsymbol{T}_1\boldsymbol{A} + \boldsymbol{T}_0 \geqslant \boldsymbol{0} \tag{8}$$

where $\boldsymbol{T}_1 \in \mathbb{R}^{Q\times P}$ and $\boldsymbol{T}_0 \in \mathbb{R}^{Q\times N}$. This formulation allows to take into account

- constraint (3) when $Q = P$, $\boldsymbol{T}_1 = \boldsymbol{I}_P$ and $\boldsymbol{T}_0 = \boldsymbol{0}_P$,
- constraint (5) when $Q = 1$, $\boldsymbol{T}_1 = -\mathbf{1}_P^{\mathrm{t}}$ and $\boldsymbol{T}_0 = \mathbf{1}_N^{\mathrm{t}}$,
- constraints (3) and (5) jointly, by setting $Q = P + 1$, $\boldsymbol{T}_1 = [\boldsymbol{I}_P \mid -\mathbf{1}_P]^{\mathrm{t}}$ and $\boldsymbol{T}_0 = [\boldsymbol{0}^{\mathrm{t}} \mid \mathbf{1}_N]^{\mathrm{t}}$,

where $\boldsymbol{I}_N$ denotes the identity matrix of $\mathbb{R}^N$. The equality constraint (4) can be implicitly handled by introducing a reparametrization so that the optimization problem is reduced to an inequality constrained minimization [26].

Property 1: For each matrix $\boldsymbol{A}^{(0)} \in \mathbb{R}^{P\times N}$ satisfying the equality constraint (4), the transformed vector $\boldsymbol{A} = \boldsymbol{A}^{(0)} + \boldsymbol{Z}\boldsymbol{U}$ also satisfies (4) as soon as the columns of matrix $\boldsymbol{Z} \in \mathbb{R}^{P\times P-1}$ are formed with vectors of the null space of $\mathbf{1}_P^{\mathrm{t}}$.

For the sum-to-one constraint, a null space matrix can be defined by

$$Z_{ij} = \begin{cases} 1 & \text{if} \quad i = j, \\ -1 & \text{if} \quad i = j+1, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

According to this reparametrization, the constrained optimization problem when constraints (3) and (4) are imposed becomes equivalent to

$$\min_{\boldsymbol{U} \in \mathbb{R}^{(P-1)\times N}} F(\boldsymbol{A}^{(0)} + \boldsymbol{Z}\boldsymbol{U}) \quad \text{s.t.} \quad \boldsymbol{T}_1^u \boldsymbol{U} + \boldsymbol{T}_0^u \geqslant \boldsymbol{0} \tag{10}$$

which takes the general form (8) with $\boldsymbol{T}_1^u = \boldsymbol{T}_1 \boldsymbol{Z}$ and $\boldsymbol{T}_0^u = \boldsymbol{T}_1 \boldsymbol{A}^{(0)} + \boldsymbol{T}_0$.

III. Primal-Dual Optimization for Abundance Maps Estimation

The main feature of interior-point optimization is to keep the solution inside the strictly feasible domain [21], [22]. At each iteration, the constraint fulfillment is ensured by adding a logarithmic barrier function making the criterion unbounded at the boundary of the feasible solution domain. Let us present our *Interior-Point Least Squares* (IPLS) algorithm to solve problem (8). By introducing the operator $\boldsymbol{m} = \mathrm{vec}(\boldsymbol{M})$ which corresponds to the transformation of a matrix $\boldsymbol{M}$ to a vector $\boldsymbol{m}$ in the lexicographic order, problem (8) is equivalent to the standard form inequality constrained optimization problem:

$$\min_{\boldsymbol{a} \in \mathbb{R}^{PN}} \Phi(\boldsymbol{a}) \quad \text{s.t.} \quad \boldsymbol{T}\boldsymbol{a} + \boldsymbol{t} \geqslant \boldsymbol{0} \tag{11}$$

with $\boldsymbol{a} = \mathrm{vec}(\boldsymbol{A})$, $\boldsymbol{T} = \boldsymbol{I}_N \otimes \boldsymbol{T}_1$ and $\boldsymbol{t} = \mathrm{vec}(\boldsymbol{T}_0)$, where $\otimes$ is the Kronecker product.

The IPLS algorithm is based on a primal-dual interior-point approach which consists in jointly estimating $\boldsymbol{a} \in \mathbb{R}^{PN}$, and their associated Lagrange multipliers $\boldsymbol{\lambda} \in \mathbb{R}^{QN}$ through the resolution of a sequence of optimization problems obtained from perturbed versions of the Karush-Kuhn-Tucker (KKT) optimality conditions for problem (11):

$$\begin{cases} \nabla\Phi(\boldsymbol{a}) - \boldsymbol{T}^{\mathrm{t}}\boldsymbol{\lambda} = \boldsymbol{0}, \\ \boldsymbol{\Lambda}(\boldsymbol{T}\boldsymbol{a} + \boldsymbol{t}) = \boldsymbol{\mu}_k, \\ \boldsymbol{T}\boldsymbol{a} + \boldsymbol{t} \geqslant \boldsymbol{0}, \\ \boldsymbol{\lambda} \geqslant \boldsymbol{0}, \end{cases} \tag{12}$$

where $\boldsymbol{\Lambda} = \mathrm{Diag}(\boldsymbol{\lambda})$ and $\boldsymbol{\mu}_k = \mu_k \boldsymbol{1}_{QN}$ results from a sequence of perturbation parameters $\{\mu_k\}_{k\in\mathbb{N}}$ converging to 0 as k is growing.

At each iteration k of the algorithm, $\boldsymbol{a}_{k+1}$ and $\boldsymbol{\lambda}_{k+1}$ are firstly calculated from the perturbed KKT conditions. The perturbation parameter μ_{k+1} is then updated in order to ensure the algorithm convergence. More precisely, an approximate solution of (12) is retained from a Newton algorithm step on the equality conditions, in association with a linesearch strategy allowing to ensure the inequality conditions [22, ch. 11]. The update strategy is then given by

$$(\boldsymbol{a}_{k+1}, \boldsymbol{\lambda}_{k+1}) = \left(\boldsymbol{a}_k + \alpha_k \boldsymbol{d}_k^a, \boldsymbol{\lambda}_k + \alpha_k \boldsymbol{d}_k^\lambda\right) \tag{13}$$

where α_k is the step size and $(\boldsymbol{d}_k^a, \boldsymbol{d}_k^\lambda)$ are the primal and dual Newton directions.

Based on the iterative scheme (13), several primal-dual interior-point methods have been proposed in the literature, each of them calling for its own strategy for the computation of the primal-dual directions, the derivation of a suitable step size, and the update of the perturbation parameter (See [27], [28] for a review). The proposed IPLS algorithm for spectral unmixing relies on the iterative scheme of [26] into which additional tools, that are described in the following, have been included to accelerate the practical convergence, as well as to reduce the computational cost per iteration.

1) Primal-Dual Directions: The Newton directions $(\boldsymbol{d}_k^a, \boldsymbol{d}_k^\lambda)$ are obtained by solving the linear system,

$$\begin{bmatrix} \nabla^2\Phi(\boldsymbol{a}_k) & -\boldsymbol{T}^{\mathrm{t}} \\ \boldsymbol{\Lambda}_k \boldsymbol{T} & \mathrm{Diag}(\boldsymbol{T}\boldsymbol{a}_k + \boldsymbol{t}) \end{bmatrix} \begin{bmatrix} \boldsymbol{d}_k^a \\ \boldsymbol{d}_k^\lambda \end{bmatrix} = -\boldsymbol{r}_{\mu_k}(\boldsymbol{a}_k, \boldsymbol{\lambda}_k), \tag{14}$$

where $\nabla\Phi(\,\cdot\,)$ and $\nabla^2\Phi(\,\cdot\,)$ are, respectively, the gradient and the Hessian of criterion $\Phi(\,\cdot\,)$, and $\boldsymbol{r}_\mu(\boldsymbol{a}, \boldsymbol{\lambda})$ is the primal-dual residual defined by,

$$\boldsymbol{r}_\mu(\boldsymbol{a}, \boldsymbol{\lambda}) = \begin{pmatrix} \nabla\Phi(\boldsymbol{a}) - \boldsymbol{T}^{\mathrm{t}}\boldsymbol{\lambda} \\ \boldsymbol{\Lambda}(\boldsymbol{T}\boldsymbol{a} + \boldsymbol{t}) - \boldsymbol{\mu} \end{pmatrix} = \begin{pmatrix} \boldsymbol{r}_\mu^{\mathrm{prim}}(\boldsymbol{a}, \boldsymbol{\lambda}) \\ \boldsymbol{r}_\mu^{\mathrm{dual}}(\boldsymbol{a}, \boldsymbol{\lambda}) \end{pmatrix}. \tag{15}$$

As pointed out in [29], [30], the primal-dual matrix in the left side of (14) suffers from ill-conditioning as soon as $(\boldsymbol{T}\boldsymbol{a}_k + \boldsymbol{t})_i \ll 1$ or $\lambda_i \ll 1$. Moreover, this matrix is not guarenteed to be symmetric or definite positive [28], so that the linear system (14) is difficult to solve. Therefore, rather than solving directly (14), [26], [31] propose to proceed by variable substitution. From the second equation of (14) one deduces

$$\boldsymbol{d}_k^\lambda = \mathrm{Diag}(\boldsymbol{T}\boldsymbol{a}_k + \boldsymbol{t})^{-1} \left[\boldsymbol{\mu}_k - \boldsymbol{\Lambda}_k(\boldsymbol{T}\boldsymbol{a}_k + \boldsymbol{t}) - \boldsymbol{\Lambda}_k \boldsymbol{T} \boldsymbol{d}_k^a\right]. \tag{16}$$

Then, the primal direction $\boldsymbol{d}_k^a$ is obtained by solving the reduced linear system

$$\begin{aligned} [\nabla^2\Phi(\boldsymbol{a}_k) + \boldsymbol{T}^{\mathrm{t}}\mathrm{Diag}(\boldsymbol{T}\boldsymbol{a}_k + \boldsymbol{t})^{-1}\boldsymbol{\Lambda}_k \boldsymbol{T}]\boldsymbol{d}_k^a = \\ -\nabla\Phi(\boldsymbol{a}_k) + \boldsymbol{T}^{\mathrm{t}}\mathrm{Diag}(\boldsymbol{T}\boldsymbol{a}_k + \boldsymbol{t})^{-1}\boldsymbol{\mu}_k. \end{aligned} \tag{17}$$

Finally, the dual direction $\boldsymbol{d}_k^\lambda$ is calculated according to (16). Note that our computation of the primal direction differs from [26]. Indeed, instead of a low rank approximation of $\nabla^2\Phi(\boldsymbol{a}_k)$, we keep the true Hessian matrix in (17), with the will to accelerate the convergence of our algorithm (see Remark 1 at the end of this section).

2) Linesearch: The step size value α_k should be chosen so as to ensure the convergence of the IPLS algorithm and the fulfillment of the inequalities of the pertubed KKT system (12). The convergence study of the primal-dual algorithm presented in [26] requests that α_k ensures a sufficient decrease of the primal-dual merit function $\Psi_{\mu_k}(\boldsymbol{a}, \boldsymbol{\lambda})$,

$$\begin{aligned} \Psi_{\mu_k}(\boldsymbol{a}, \boldsymbol{\lambda}) = \Phi(\boldsymbol{a}) - \mu \sum_{i=1}^{QN} \ln([\boldsymbol{T}\boldsymbol{a} + \boldsymbol{t}]_i) \\ + \boldsymbol{\lambda}^{\mathrm{t}}(\boldsymbol{T}\boldsymbol{a} + \boldsymbol{t}) - \mu \sum_{i=1}^{QN} \ln(\lambda_i [\boldsymbol{T}\boldsymbol{a} + \boldsymbol{t}]_i). \end{aligned} \tag{18}$$

One can note that (18) contains two logarithmic barrier functions enforcing the fulfillment of the KKT inequalities. The sufficient decrease is assessed using the Armijo condition,

$$\psi_{\mu_k}(\alpha_k) - \psi_{\mu_k}(0) \leqslant \sigma\, \alpha_k \nabla\psi_{\mu_k}(0) \quad \text{with} \quad \sigma \in (0, 1/2), \tag{19}$$

where $\psi_{\mu_k}(\alpha) \triangleq \Psi_{\mu_k}(\boldsymbol{a}_k + \alpha \boldsymbol{d}_k^a, \boldsymbol{\lambda}_k + \alpha \boldsymbol{d}_k^\lambda)$. A step size α_k satisfying (19) is obtained by a backtracking algorithm [22]: starting from an initial step size α_k^0, and if the latter does not satisfy (19), smaller values are tested, $\alpha_k^0\tau$, $\alpha_k^0\tau^2, \ldots, \tau \in (0, 1)$, until (19) holds.

In order to ensure that $\psi_{\mu_k}(\,\cdot\,)$ remains finite valued, the backtracking strategy is initialized as follows,

$$\begin{cases} \alpha_k^0 = 1 & \text{if} \quad \alpha_k^+ = +\infty, \\ \alpha_k^0 = \min(1, 0.99\,\alpha_k^+) & \text{elsewhere}, \end{cases} \tag{20}$$

where α_k^+ is the largest positive value such that,

$$\boldsymbol{\lambda}_k + \alpha \boldsymbol{d}_k^\lambda > \boldsymbol{0}, \quad \boldsymbol{T}(\boldsymbol{a}_k + \alpha \boldsymbol{d}_k^a) + \boldsymbol{t} > \boldsymbol{0}. \tag{21}$$

3) Perturbation Parameter Update: According to [26], the convergence is ensured as soon as the sequence $\{\mu_k\}_{k\in\mathbb{N}}$ tends to 0 when k tends to infinity. We propose to update the parameter μ_k by using the μ-criticity rule defined in [32] by:

$$\mu_k = \theta \frac{\delta_k}{QN} \tag{22}$$

where $\delta_k = (\boldsymbol{T}\boldsymbol{a}_k + \boldsymbol{t})^{\mathrm{t}}\boldsymbol{\lambda}_k$ is the duality gap and $\theta \in (0, 1)$.

4) Stopping Criteria: The main steps of the proposed optimization method are summarized in Algorithm 1. Following [23], [31], the accuracy of the primal and dual directions (inner loop) is controlled by:

$$\left\|\boldsymbol{r}_{\mu_k}^{\mathrm{prim}}\right\|_\infty \leqslant \epsilon_k^{\mathrm{prim}} \text{ and } \delta_k/QN \leqslant \epsilon_k^{\mathrm{dual}} \tag{23}$$

where $\boldsymbol{r}_{\mu_k}^{\mathrm{prim}}$ is the primal residual at $\boldsymbol{a}_k$, $\epsilon_k^{\mathrm{prim}} = \eta^{\mathrm{prim}}\mu_k$, $\epsilon_k^{\mathrm{dual}} = \eta^{\mathrm{dual}}\mu_k$ with $\eta^{\mathrm{prim}} > 0$ and $\eta^{\mathrm{dual}} \in (1, \theta^{-1})$. The outer iterations of Algorithm 1 are run until the fulfillment of the stopping condition proposed in [22, ch. 11]:

$$\mu_k \leqslant \mu_{\min} \quad \text{or} \quad \left(\left\|\boldsymbol{r}_0^{\mathrm{prim}}\right\| + \left\|\boldsymbol{r}_0^{\mathrm{dual}}\right\|\right) \leqslant \epsilon_0. \tag{24}$$

5) Convergence Result: The convergence of Algorithm 1 is guaranteed by the following result.

Theorem III.1: Let $\Phi(\,\cdot\,)$ a twice differentiable convex function on $\mathbb{R}^{PN}$. Assume that the set $\mathcal{S} = \{\boldsymbol{a} \in \mathbb{R}^{PN} \mid \boldsymbol{T}\boldsymbol{a} + \boldsymbol{t} > 0\}$ is nonempty and bounded, and that either $\Phi(\,\cdot\,)$ is strictly convex or $\boldsymbol{T}^{\mathrm{t}}\boldsymbol{T}$ is inversible. Then, for every fixed $\mu > 0$, there exists k_μ such that the sequence $\{(\boldsymbol{a}_k, \boldsymbol{\lambda}_k)\}_{k \geqslant k_\mu}$ generated by (13) converges q-superlinearly to the unique minimizer of Ψ_μ. Moreover, the outer loop of Algorithm 1 generates a bounded sequence $\{(\boldsymbol{a}_k, \boldsymbol{\lambda}_k)\}$ whose accumulation points are primal-dual solutions of problem (11). Finally, if $\Phi(\,\cdot\,)$ is strictly convex, the outer iterates $\{\boldsymbol{a}_k\}$ converge to the unique solution of (11).

Proof: See Appendix A. ∎

We now comment the differences between Theorem III.1 and the convergence result in [26].

Require: Initial values $\boldsymbol{\lambda}_0 > \boldsymbol{0}$ and $\boldsymbol{a}_0$ such that $\boldsymbol{T}\boldsymbol{a}_0 + \boldsymbol{t} > \boldsymbol{0}$
Ensure: Resolution of (11)
While (condition (24) is not satisfied) **do**
  While (condition (23) is not satisfied) **do**
    Calculate $\boldsymbol{d}_k^a$ by solving (17)
    Deduce $\boldsymbol{d}_k^\lambda$ from (16)
    Find $\alpha_k > 0$ satisfying (19)
    Update $(\boldsymbol{a}_{k+1}, \boldsymbol{\lambda}_{k+1})$ according to (13)
  done
  Define μ_{k+1} according to (22).
done

Algorithm 1. Interior-Point Least Squares algorithm.

Remark 1:

i) The convergence result of [26] is established under the assumption that the criterion and the constraints are convex, and at least one of them is strongly convex. In our study, the convexity of Φ is sufficient, under the additional assumption that the constraints are linearly independant (i.e., $\boldsymbol{T}^{\mathrm{t}}\boldsymbol{T}$ inversible) and that the set $\mathcal{S}$ is nonempty and bounded. Note that these assumptions hold in particular for the constraints (3)–(4) or (3)–(5).

ii) The q-superlinear convergence rate of the inner loop of our algorithm is mainly due to the use of the exact Hessian matrix in the primal system (17). A weaker result in terms of convergence speed is obtained in [26], since a quasi-Newton approximation of the Hessian matrix is considered.

IV. Memory Requirement and Computation Time Reduction for Large Scale Spectral Unmixing

According to the expression of the least squares criterion (6), the constrained optimization problem (8) is separable with respect to the image pixels. A first implementation strategy, denoted hereafter by *pixel-based* strategy, is to solve problem (2) by applying Algorithm 1 for unmixing each n-th pixel individually. A second approach is to adopt an *image-based* strategy, that is, solving the whole problem (8) with the primal-dual algorithm. A discussion on the numerical efficiency of both strategies will be given in Section V-A.

When the *image-based* strategy is adopted, the numerical complexity of Algorithm 1 is highly dominated by the primal direction calculation through the resolution of the linear system (17). This section presents an analysis of the structure of this system with the aim to reduce the computational cost and the memory requirement of Algorithm 1.

A. Primal System Structure

The linear system (17) can be expressed as

$$\boldsymbol{H}_k \boldsymbol{d}_k^a = -\boldsymbol{g}_k \tag{25}$$

where

$$\boldsymbol{H}_k = \nabla^2\Phi(\boldsymbol{a}_k) + \boldsymbol{T}^{\mathrm{t}}\mathrm{Diag}(\boldsymbol{T}\boldsymbol{a}_k + \boldsymbol{t})^{-1}\boldsymbol{\Lambda}_k\boldsymbol{T}, \tag{26}$$

$$\boldsymbol{g}_k = \nabla\Phi(\boldsymbol{a}_k) - \boldsymbol{T}^{\mathrm{t}}\mathrm{Diag}(\boldsymbol{T}\boldsymbol{a}_k + \boldsymbol{t})^{-1}\boldsymbol{\mu}_k. \tag{27}$$

An analysis of the structure of matrix $\boldsymbol{H}_k$ is necessary in order to find an appropriate implementation strategy.

Firstly, we recall that $\boldsymbol{T} = \boldsymbol{I}_N \otimes \boldsymbol{T}_1$. Thus, $\boldsymbol{T}$ is block-diagonal composed by N identical blocks equal to $\boldsymbol{T}_1$. The notation $\boldsymbol{T} = \mathrm{Bdiag}_N(\boldsymbol{T}_1)$ is used in the sequel. For every $n \in \{1, \ldots, N\}$, let $\boldsymbol{a}_{n,k} \in \mathbb{R}^P$ (resp. $\boldsymbol{\lambda}_{n,k} \in \mathbb{R}^Q$) be the n-th column of $\boldsymbol{A}_k = \mathrm{mat}(\boldsymbol{a}_k) \in \mathbb{R}^{P \times N}$ (resp. of $\Lambda_k = \mathrm{mat}(\boldsymbol{\lambda}_k) \in \mathbb{R}^{Q \times N}$) where $\mathrm{mat}(\cdot)$ is the reciprocal operator of $\mathrm{vec}(\cdot)$. Moreover, let $\boldsymbol{t}_{0,n} \in \mathbb{R}^Q$ be the n-th column of $\boldsymbol{T}_0$. It follows that,

$$\mathrm{Diag}(\boldsymbol{T}\boldsymbol{a}_k + \boldsymbol{t})^{-1}\boldsymbol{\Lambda}_k = \mathrm{Bdiag}_N(\boldsymbol{D}_{n,k}) \tag{28}$$

where $\boldsymbol{D}_{n,k}$ is a diagonal matrix of size $P \times P$ whose diagonal elements are $\mathrm{Diag}(\boldsymbol{T}_1\boldsymbol{a}_{n,k} + \boldsymbol{t}_{0,n})^{-1}\boldsymbol{\lambda}_{n,k}$. Therefore,

$$\begin{aligned}\boldsymbol{T}^{\mathrm{t}}&\mathrm{Diag}(\boldsymbol{T}\boldsymbol{a}_k + \boldsymbol{t})^{-1}\boldsymbol{\Lambda}_k\boldsymbol{T} \\ &= \mathrm{Bdiag}_N(\boldsymbol{T}_1)^{\mathrm{t}}\mathrm{Bdiag}_N(\boldsymbol{D}_{n,k})\mathrm{Bdiag}_N(\boldsymbol{T}_1), \\ &= \mathrm{Bdiag}_N(\boldsymbol{T}_1^{\mathrm{t}}\boldsymbol{D}_{n,k}\boldsymbol{T}_1).\end{aligned} \tag{29}$$

Secondly, the Hessian $\nabla^2\Phi(\boldsymbol{a}_k)$ of the least squares criterion reads

$$\begin{aligned}\nabla^2\Phi(\boldsymbol{a}_k) &= (\boldsymbol{I}_N \otimes \boldsymbol{S})^{\mathrm{t}}(\boldsymbol{I}_N \otimes \boldsymbol{S}), \\ &= \mathrm{Bdiag}_N(\boldsymbol{S}^{\mathrm{t}}\boldsymbol{S}),\end{aligned} \tag{30}$$

where (30) is a consequence of Kronecker product properties [33]:

$$\begin{cases}(\boldsymbol{A} \otimes \boldsymbol{B})^{\mathrm{t}} = \boldsymbol{A}^{\mathrm{t}} \otimes \boldsymbol{B}^{\mathrm{t}}, \\ (\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{C} \otimes \boldsymbol{D}) = \boldsymbol{A}\boldsymbol{C} \otimes \boldsymbol{B}\boldsymbol{D}.\end{cases} \tag{31}$$

Finally, (29) and (30) yield,

$$\boldsymbol{H}_k = \mathrm{Bdiag}_N(\boldsymbol{S}^{\mathrm{t}}\boldsymbol{S} + \boldsymbol{T}_1^{\mathrm{t}}\boldsymbol{D}_{n,k}\boldsymbol{T}_1). \tag{32}$$

Consequently, $\boldsymbol{H}_k$ is a block-diagonal matrix formed by N blocks of size $P \times P$. Note that, in the case of problem (10), a similar analysis leads to

$$\boldsymbol{H}_k = \mathrm{Bdiag}_N\left(\boldsymbol{Z}^{\mathrm{t}}(\boldsymbol{S}^{\mathrm{t}}\boldsymbol{S} + \boldsymbol{T}_1^{\mathrm{t}}\boldsymbol{D}_{n,k}\boldsymbol{T}_1)\boldsymbol{Z}\right) \tag{33}$$

that is, $\boldsymbol{H}_k$ block-diagonal with N blocks of size $(P-1)\times(P-1)$.

B. Memory Issues

When applying the *image-based* strategy to large scale problems, the memory space required to store matrix $\boldsymbol{H}_k$ can exceed the available memory, even when using a sparse coding. A less memory demanding calculation of the primal direction can be achieved by solving separately, for each iteration, the N lower-size linear systems

$$(\forall\, n \in \{1, \ldots, N\}) \quad \boldsymbol{H}_{n,k}\boldsymbol{d}^a_{n,k} = -\boldsymbol{g}_{n,k} \tag{34}$$

where $\boldsymbol{d}^a_{n,k}$ (resp. $\boldsymbol{g}_{n,k}$) is the n-th column of the matrix $\mathrm{mat}(\boldsymbol{d}^a_k)$ (resp. $\mathrm{mat}(\boldsymbol{g}_k)$). This implementation will now be referred to as the *image-based pixel-wise* implementation, as opposed to the *image-based full-wise* implementation where $\boldsymbol{H}_k$ is entirely built.

The pixel-wise strategy being based on the resolution at each iteration of N independent linear systems, it is straightforward to implement in parallel. An intermediate implementation dividing system (32) (or (33)) into $1 \leqslant K \leqslant N$ blocks can also be considered, with the advantage to adapt the normal equations size K to the available memory. This latter approach will be referred to as *image-based block-wise* implementation. The performance of each implementation strategy will be discussed in Section V-A.

V. Experimental Results

This section discusses the performances and illustrates the applicability of Algorithm 1. The latter is referred to as IPLS, the type of constraints being indicated in prefix, namely NN for non-negativity constraint (3), STO for sum-to-one and non-negativity constraints (3)–(4) and SLO for sum-lower-than-one and non-negativity constraints (3)–(5).

We first consider synthetic data in order to discuss the choice of implementation strategy, to perform a comparative analysis with existing unmixing methods, and to illustrate the relevance of the partial additivity constraints. Then, the parallel implementation of IPLS using GPUs is addressed. Finally, its applicability is emphasized through the processing of real hyperspectral data.

The computation of the proposed primal-dual algorithm requires specifying the parameters $(\eta^{\mathrm{prim}}, \eta^{\mathrm{dual}}, \theta)$ and $(\mu_{\min}, \epsilon_0)$, controlling the precision of the inner and outer loops, respectively. Following [23], [31], we set:

$$\eta^{\mathrm{prim}} = 100, \quad \eta^{\mathrm{dual}} = 1.9, \quad \theta = 0.5. \tag{35}$$

Moreover, the values $\mu_{\min} = 10^{-9}$ and $\epsilon_0 = 10^{-7}$ are retained for the stopping condition (24).

A. Synthetic Data

In order to simulate realistic synthetic hyperspectral data, reflectance spectra from the USGS (U.S. Geological Survey) spectral library [34] are retained[1]. These reflectance spectra contain $L = 224$ spectral bands from 383 nm to 2508 nm. A subset of P spectra is then randomly picked up to create synthetic mixtures with abundances simulated from a Dirichlet distribution. Only realizations with maximum abundance value lower than a specified level $A_{\max}$ are retained. Finally, a random Gaussian noise is added to each resulting mixture spectrum, in order to get a signal to noise ratio (SNR) of 30 dB. The unmixing algorithms are implemented on Matlab 2012b and the calculations are performed using a HP Compaq Elite desktop having an Intel Core i7 3.4 GHz processor and 8 GB of RAM.

The first step of the experiment consists in choosing the best implementation strategy adapted to this hardware and software configuration. Then, some comparisons are performed between our method and NNLS, FCLS and ADMM algorithms, in terms of computation time and estimation accuracy. Finally, we illustrate through two examples, the relevance of the partial addi-

[1][Online.] Available: http://pubs.usgs.gov/of/2003/ofr-03-395/datatable.html

TABLE I
AVERAGE TIME PER PIXEL, NUMBER OF ITERATIONS, AND RESIDUAL NORM OVER 100 MONTE-CARLO SIMULATIONS, FOR DIFFERENT NUMBER OF ENDMEMBERS USING ACTUAL (LIB) OR ESTIMATED ENDMEMBERS (EEA): COMPARISON BETWEEN PIXEL-BASED (PXL) AND IMAGE-BASED (IMG) IMPLEMENTATIONS OF IPLS

Endmembers	P	Time (μs)		Iterations		r ($\times 10^{-4}$)	
		PXL	IMG	PXL	IMG	PXL	IMG
LIB	3	2150	15	25.7	20.3	3.14	3.14
	6	2198	49	25.2	20.4	3.38	3.38
	10	2307	118	24.6	20.9	3.65	3.65
	15	2649	266	24.0	21.1	3.75	3.76
EEA	3	2138	15	25.7	20.2	3.15	3.14
	6	2192	48	25.2	20.2	3.38	3.39
	10	2302	117	24.5	20.9	3.69	3.70
	15	2643	277	23.9	21.2	3.86	3.86

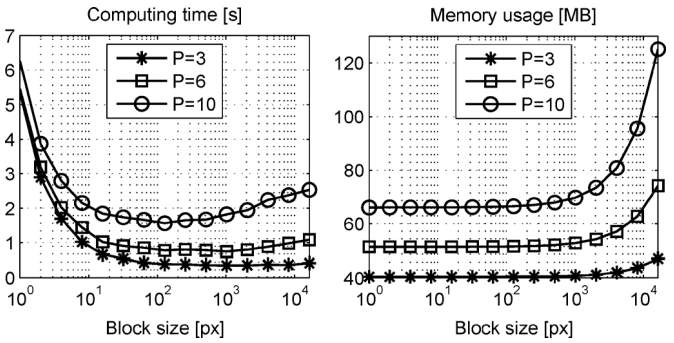


Fig. 1. Unmixing computation time (in seconds) and memory usage (MB) for different block sizes of the block-wise implementation.

tivity constraint over the full additivity and the non-negativity constraints.

1) Pixel-Based or Image-Based Unmixing: In order to compare the performances of pixel-based and image-based full-wise implementations, we consider the unmixing of synthetic images of size $N = 64^2$, with different number of endmembers, and $A_{\max} = 1$. The unmixing is performed under constraints (3)–(4), using either the exact endmembers or those extracted from the image using the N-FINDR method [35]. For each test realized, Table I reports the computation time per pixel, the average number of iterations (outer iterations of Algorithm 1), and the average residual norm:

$$r = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{L}\|\boldsymbol{y}_n - \boldsymbol{S}\boldsymbol{a}_n\|_2. \tag{36}$$

It can be noted that the image-based implementation is significantly faster than the pixel-based. This is explained by the better management of the vectorized calculations compared to sequential ones in Matlab. Moreover, less iterations are required to reach the stopping criterion in the case of the image-based implementation. The average residual norm of the two strategies certifies that the quality of the reconstruction is equivalent in both cases. Finally, let us emphasize that the performances of IPLS are not degraded when replacing the exact endmembers by their estimation with N-FINDR.

2) Image-Based Unmixing Alternatives: We now analyse the influence of the block-size parameter K on the performance of the image-based block-wise implementation. In that respect, we consider a hyperspectral image of size $N = 128^2$ built using a subset of P endmembers from the USGS library, and $A_{\max} = 1$. The computation time required by Algorithm 1 to unmix this image under full additivity constraints (3)–(4), is presented in Fig. 1 for several numbers of endmembers and different values of K. Note that a block size $K = 1$ corresponds to the pixel-wise implementation and the full-size strategy is obtained by setting $K = N$. The other configurations correspond to intermediate block-wise alternatives. The memory space required for the unmixing function in Matlab is also reported.

Ideally, the best implementation should correspond to both a low computing time and a low memory usage. According to our results, the computation time decreases as the block size rises, reaching a minimum for a block size above 100, whatever the value of P. On the other hand, as expected, the memory usage grows with the block size. Consequently, the block size should be set to an intermediate value in order to achieve the best computing time. The block-wise implementation with a block size $K = 256$ allowing to get the best compromise between computing time and memory requirement is retained for the remaining experiments presented in the paper.

3) Comparison With State-of-the-Art Unmixing Algorithms: IPLS is now compared with FCLS when both non-negativity constraint (3) and full additivity constraint (4) are imposed, and with NNLS when only non-negativity is considered. We also compare our algorithm with the alternating method of multipliers (ADMM) from [19], using the Matlab code available at `http://www.lx.it.pt/~bioucas`. Synthetic hyperspectral images of size $N = 64^2$ are generated, using different number P of endmembers and $A_{\max} = 1$. For each image, the set of spectra employed to perform the spectral unmixing is either the one used to create the image or is estimated using the N-FINDR endmember extraction algorithm.

The three methods have led to the same unmixing quality in terms of residual value r. The results in terms of average computation time per pixel over 100 Monte-Carlo simulations are reported in Table II. For all the tests realized, both STO-IPLS and STO-ADMM appear to be faster than FCLS. The ratio between STO-IPLS or STO-ADMM and FCLS computation times seems to be independent from the number of endmember used. NN-IPLS and NN-ADMM are also faster than NNLS under the conditions tested. This superiority tends to decrease as the number of endmembers increases. Finally, the ranking between IPLS and ADMM methods depends on the experimental conditions. According to our tests, STO-IPLS seems slightly faster than STO-ADMM, while NN-IPLS and NN-ADMM perform similarly in terms of computation time.

4) Relevance of the Partial Additivity Constraint: When dealing with real data, the abundance estimation performances depend on the used endmember spectra and on the constraints that are imposed on the abundance values. The aim of this section is to show that it may be suitable, in some situations, to relax the sum-to-one constraint (4) and, eventually, to replace it by the partial additivity constraint (5). A hyperspectral image of size $N = 128^2$ built using a subset of $P = 6$ endmembers from the USGS library is considered. The accuracy of the

TABLE II
AVERAGE TIME PER PIXEL FOR DIFFERENT NUMBER OF ENDMEMBERS USING ACTUAL (LIB) OR ESTIMATED ENDMEMBERS (EEA): COMPARISON BETWEEN FCLS, ADMM AND IPLS FULLY CONSTRAINED (STO) AND NNLS, ADMM AND IPLS WITH NON-NEGATIVITY CONSTRAINT ONLY (POS)

Endmembers	P	Time (μs)		
		FCLS	STO-ADMM	STO-IPLS
LIB	3	46	22	**18**
	6	84	65	**45**
	10	210	124	**90**
	15	479	198	**177**
EEA	3	45	**18**	**18**
	6	84	71	**42**
	10	144	132	**89**
	15	314	197	**179**
Endmembers	P	Time (μs)		
		NNLS	NN-ADMM	NN-IPLS
LIB	3	66	**18**	20
	6	117	53	**46**
	10	177	109	**94**
	15	246	**183**	190
EEA	3	67	**15**	20
	6	118	54	**45**
	10	175	121	**93**
	15	237	189	**187**

abundance maps estimation is assessed using the normalized mean square error

$$\mathrm{NMSE}(\%) = \frac{100}{P}\sum_{p=1}^{P}\left(\|\boldsymbol{m}_p - \hat{\boldsymbol{m}}_p\|^2/\|\boldsymbol{m}_p\|^2\right) \qquad (37)$$

which measures the relative mean difference between the actual abundances maps $\boldsymbol{m}_p = [A_{p1}, \ldots, A_{pN}]^{\mathrm{t}}$ and the estimated ones $\hat{\boldsymbol{m}}_p$.

a) Effect of Illumination Variability: We first analyse the relevance of the full additivity constraint when the image pixels are subject to illumination variability. In that respect, each pixel spectrum generated from the linear mixing model is multiplied by a scale factor η modeling the illumination variability due to surface topography or atmospheric attenuation [36]. As in [37], this scale factor is simulated from a Beta distribution with a specified mean value ν_η in the interval [0.9, 1]. The hyperspectral image is then unmixed using the IPLS algorithm on the exact endmembers, with either constraint (4) or (5) in addition to the non-negativity constraint (3).

Table III summarizes our results for different maximum abundance values and attenuation levels. The number of pixels was set to $N = 50^2$ and 100 Monte-Carlo simulations have been considered. It can be noted that the full additivity constraint leads to the best estimation results in the absence of illumination variability ($\nu_\eta = 1$) and endmember spectra taken either from the library or extracted from the image using an endmember extraction algorithm (VCA in this experiment). However, the performances decrease when the value of ν_η equals 0.95 or 0.9. It can be, for instance, noted that the partial additivity is relevant when the endmembers are taken from the library and that the non-negativity constraint alone leads to the best results when endmembers are extracted from the image.

b) Effect of an Incomplete Set of Endmembers: We now consider the case when the set of endmembers used to unmix the image spectra is incomplete. Such situation arises, for instance, when the number of components is underestimated. An image of size $N = 50^2$ containing P endmembers is simulated using the same strategy as in the previous experiment, using $A_{\max} = 1$ and $\nu_\eta = 1$. The unmixing is performed using a subset of $\hat{P} \leqslant P$ endmembers arbitrarily taken from the actual set of endmembers or estimated using the VCA algorithm. The IPLS algorithm is applied with either the full additivity (3)–(4), partial additivity (3)–(5) or non-negativity (3) constraints. From Table IV, one can note that imposing the partial additivity constraint is very useful when the number of unmixed endmembers is lower than the number of actual endmembers.

TABLE III
ACCURACY (NMSE) OF ABUNDANCE ESTIMATION FROM EITHER ACTUAL (LIB) OR EXTRACTED ENDMEMBERS (EEA), USING IPLS UNDER FULL ADDITIVITY (STO), PARTIAL ADDITIVITY (SLO) AND NON-NEGATIVITY (NN) CONSTRAINTS

Endmembers	$A_{\max}$	ν_η	Constraints		
			STO	SLO	NN
LIB	(1, 0.95, 0.9)	1.00	**0.04**	0.10	0.19
		0.95	3.83	**0.64**	0.69
		0.90	12.36	**2.03**	2.06
EEA	1	1.00	**0.47**	0.58	0.72
		0.95	4.57	3.39	**1.44**
		0.90	16.45	12.71	**5.08**
	0.95	1.00	**0.63**	0.74	**0.90**
		0.95	5.65	4.47	**1.94**
		0.90	17.18	13.28	**5.02**
	0.9	1.00	**0.99**	**1.12**	1.24
		0.95	4.67	3.45	**1.42**
		0.90	17.25	13.25	**5.72**

TABLE IV
ACCURACY (NMSE IN (%)) OF ABUNDANCE ESTIMATION USING IPLS UNDER FULL ADDITIVITY (STO), PARTIAL ADDITIVITY (SLO) AND NON-NEGATIVITY (NN) CONSTRAINTS: EFFECT OF AN INCOMPLETE SET OF ACTUAL (LIB) OR EXTRACTED ENDMEMBERS (EEA)

P	$\hat{P}$	LIB			EEA		
		STO	SLO	NN	STO	SLO	NN
6	6	**0.06**	0.11	0.17	**0.47**	0.51	0.60
	5	18.22	**15.39**	19.34	18.28	**15.47**	19.66
	4	48.31	**36.00**	39.15	48.17	**35.99**	39.19
10	10	**0.14**	0.21	0.31	**1.68**	1.84	2.15
	9	10.13	**7.41**	8.69	11.64	**9.05**	10.17
	8	22.47	**16.40**	22.41	23.66	**17.98**	23.50
15	15	**0.52**	0.73	1.57	**6.43**	6.64	9.23
	14	6.17	**5.65**	16.19	15.79	**12.31**	19.20
	13	12.32	**10.97**	26.75	21.56	**16.77**	27.96

B. Parallel Implementation

A parallel implementation of Algorithm 1, for both image-based and pixel-based strategies, has been realized using CUDA (for *Compute Unified Device Architecture*), a programming model created by Nvidia based on a language designed as an extension of the C language.

In the pixel-based implementation, the entire algorithm is run independently for each pixel. One thread per pixel is used, each thread containing the whole IPLS algorithm. The image-based implementation does not present such a degree of parallelization since some steps, namely the linesearch, the perturbation parameter update, and the convergence check, require the computation of one variable for the entire image. These global steps,

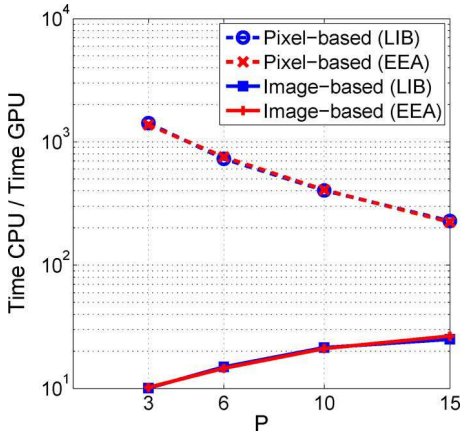


Fig. 2. Comparison in terms of computation time per pixel, between GPU and CPU implementation for pixel-based and image-based implementations of IPLS algorithm. Average results over 100 Monte-Carlo simulations, for different number of endmembers using actual (LIB) or estimated endmembers (EEA).

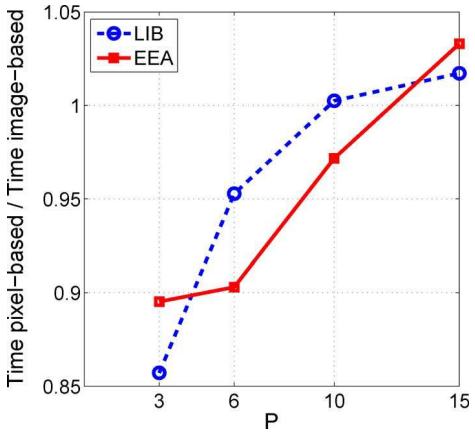


Fig. 3. Comparison in terms of computation time per pixel, between pixel-based and image-based GPU implementations of IPLS algorithm. Average results over 100 Monte-Carlo simulations, for different number of endmembers using actual (LIB) or estimated endmembers (EEA).

called *reductions*, are optimized using the combination of the GPU and the CPU as it is described in [38, ch. 6].

For the same experiments than those conducted in Section V-A1, we present in Fig. 2 the speed up in terms of average computation time per pixel obtained when using parallel programming. The IPLS algorithm was run on a Dell Precision T7400 having an Intel Xeon X5472 3 GHz processor and 16 GB of RAM. It embeds the Nvidia Tesla C1060 GPU (Graphics Processing Unit) allowing to do parallel computation on its 240 processor cores running at 1.3 GHz. Note that the iterations number and the residual norms resulting from these tests were the same than those presented in Table I, which shows the validity of our GPU program. The ratio between GPU and CPU computation time follows different behaviour for pixel-based and image-based implementations. When the first approach is retained, the gain of GPU computing tends to decrease, as the number of endmembers grows. On the opposite, the image-based GPU implementation tends to be more efficient when P increases. Up to our knowledge, this difference could be due to the use of the strategy by [38, ch. 6] in the GPU programs of the image-based implementation. Indeed, it implies that the computing time necessary for performing data transfers remains constant, whatever the value of P. For small size unmixing problems, this transfer time becomes preponderant over other operations, thus limiting the GPU speed-up.

Fig. 3 illustrates the ratio between the average computation time per pixel for pixel-based and image-based GPU implementations. Although the pixel-based approach would seem better suited for parallelization, it presents similar computation time than the image-based approach, when GPU programming is employed. This can be explained by the fact that, according to the CUDA model [39], the threads are actually processed by groups of 32 called *warps*. Each warp works as an SIMD (Single Instruction Multiple Data) unit. At a given instant, all the threads of one warp are necessarily executing the same instruction. In the case of a conditional structure such as *if, then, else*, if two conditions are satisfied by different threads within a warp, then two series of instructions corresponding to these conditions are executed by all the threads of this warp, although some of the results are ignored. Therefore, during the execution of the pixel-based IPLS, the computing time depends on the pixels having the slowest convergence rate in each group of 32 consecutive pixels. The gain when using the pixel-based approach can thus be small if the convergence rates highly differ from one pixel to another. Another reason is that, as emphasized in Section V-A1, the IPLS algorithm requires more iterations to reach convergence in its pixel-based version.

C. Real Data Processing

We consider in this section the unmixing of the well known AVIRIS Cuprite dataset available online[2]. This image originally contains 250×191 pixels and 224 spectral bands between 0.4 and 2.5 μm. Only 188 bands are preserved after removing the corrupted ones.

1) Number of Endmembers Estimation: The number of endmembers is estimated with the SGDE method proposed in [40] based on Gerschgorin disks' radii, leading to the reasonable number of 14 endmembers. This estimated number is retained during the rest of the experiment. Other methods such as Virtual Dimentionality estimation [41] or ELM [42] could have been used, possibly leading to a different number.

2) Endmembers Extraction: The endmembers are extracted from the scene using the N-FINDR algorithm. For each endmember, Table V gives the two closest components of the USGS library according to the Spectral Information Divergence (SID) [43]. Other endmember extraction algorithms and spectral distance measurements could have been used, possibly leading to different substances. A survey on EEA algorithms is conducted in [44].

3) Abundance Estimation: Computing times for different constraints are reported in Table VI, for an image-based block-wise implementation of the IPLS algorithm, run on Matlab R2011b, using the same architecture as in Section V-B.

[2][Online.] Available: http://aviris.jpl.nasa.gov/html/aviris.freedata.html.

TABLE V
SPECTRAL INFORMATION DIVERGENCE (SID) BETWEEN EXTRACTED ENDMEMBERS AND LABORATORY REFLECTANCES

Index	Mineral	SID ($\times 10^{-3}$)
1	Pyrope WS474	6.52
	Sphene HS189.3B	8.36
2	Buddingtonite GDS85 D-206	6.51
	Kaolin/Smect KLF511 12%K	7.39
3	Nontronite SWa-1.a	13.31
	Kaolin/Smect H89-FR-5 30K	13.73
4	Nontronite NG-1.a	6.33
	Montmorillonite+Illi CM37	8.63
5	Nontronite SWa-1.a	15.39
	Kaolin/Smect KLF508 85%K	17.51
6	Rectorite ISR202 (RAr-1)	8.46
	Montmorillonite+Illi CM42	8.49
7	Montmorillonite+Illi CM42	9.52
	Kaolin/Smect H89-FR-5 30K	10.36
8	Kaolin/Smect KLF511 12%K	3.28
	Rectorite ISR202 (RAr-1)	4.08
9	Montmorillonite CM20	5.38
	Alunite GDS82 Na82	6.91
10	Thenardite GDS146	3.60
	Kaolin/Smect H89-FR-2 50K	3.35
11	Cookeite CAr-1.c ¡30um	2.54
	Thenardite GDS146	2.65
12	Montmorillonite+Illi CM42	8.22
	Rectorite ISR202 (RAr-1)	9.37
13	Kaolin/Smect KLF511 12%K	1.82
	Montmorillonite+Illi CM37	3.21
14	Barite HS79.3B	5.36
	Richterite HS336.3B	5.51

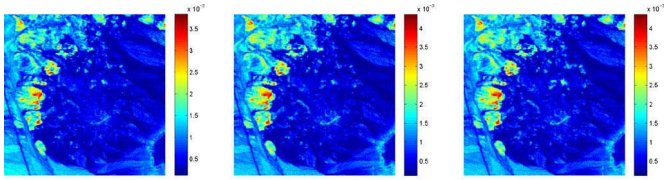


Fig. 4. Cuprite residual norm per pixel after unmixing with IPLS subject to different constraints.

TABLE VI
COMPUTING TIME AND RESIDUAL NORM AFTER UNMIXING CUPRITE SCENE WITH IPLS SUBJECT TO DIFFERENT CONSTRAINTS

Constraint	NN	SLO	STO
Time (s)	13.2	17.5	16.4
r ($\times 10^{-4}$)	7.37	8.34	8.52

We can note on both Table VI and Fig. 4 that the residual error is not strongly affected by the constraint choice.

Fig. 5 illustrates the effect of the constraint choice on the distribution of the abundance sum per pixel. With positivity only, a sum higher than one is observed in a significant part of the image, which has no physical meaning. Adding the partial additivity constraint provides a sum close to one in most of the pixels. Those who have an abundance sum far from one may reveal a lack of luminosity, an underestimated number of endmembers, or a non-linear phenomenon that cannot be handled with the proposed mixing model.

Using the GPU implementation described in Section V-B, the computational time for unmixing the real data under constraints (3) and (4) becomes 0.50 s for the image-based version, and

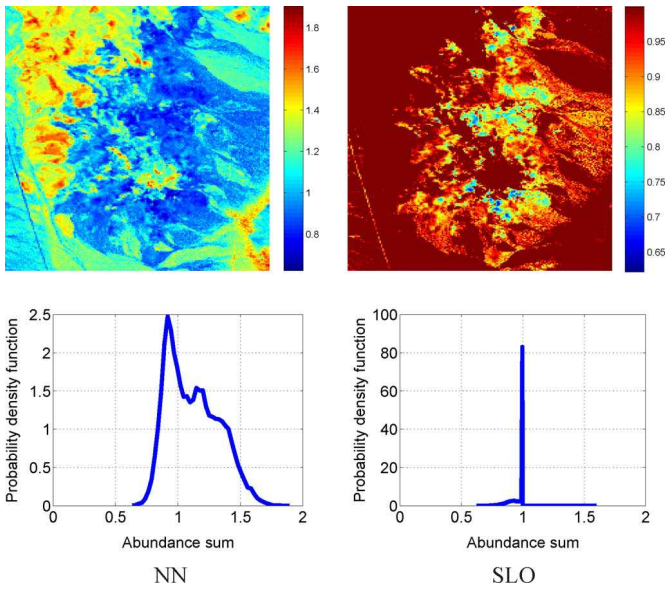


Fig. 5. Abundance sum per pixel after unmixing Cuprite scene subject to different constraints: maps and probability density functions.

0.59 s for the pixel-based version. The speed up of 33 compared to the CPU version exhibits the suitability of our method for parallel programming.

VI. CONCLUSION

We have proposed in this paper a spectral unmixing algorithm allowing to estimate the abundance maps using a primal-dual interior-point optimization method. The main feature of the proposed approach is to handle various linear constraints such as full additivity, partial additivity and non-negativity. The second advantage of this approach is its suitability for an efficient parallel implementation using GPUs. These features have been illustrated by processing a real hyperspectral data using two GPU variants of the proposed method (pixel-based or image-based). Implementing the pixel-based version of the method can be extended to the case of sparse unmixing of a single pixel spectrum using either a sparse recovery approach on a large library or to the case of a large number of spectral bands. On the other hand, the image-based implementation opens the way to fast processing methods including spatial penalization on the abundance maps.

In that respect, the proposed method can be naturally extended to the case of abundance estimation using penalized or weighted least squares estimation, when the regularization function preserves the block diagonal structure of the Hessian matrix. This is for instance the case with Tikhonov regularization or sparse regularization approaches. Future works will be directed to addressing the case of penalization functions that incorporate a spatial regularization of the abundance maps, such as total variation [45] and roughness penalties [46]. Our preliminary results, presented in [47], have shown that the spatial regularization enhances the estimation quality at the price of a significant increase of the numerical complexity. Additional mathematical development are required in order to adapt the primal-dual approach to solve the minimization problem in this context with a reduced memory requirement and computation cost.

APPENDIX

A. Proof of Theorem III.1

1) Convergence of the Inner Loop: Let us firstly prove the convergence of the inner loop of Algorithm 1 for a fixed perturbation parameter $\mu > 0$. According to the definition (18), we have, for all $k \in \mathbb{N}$,

$$\nabla \Psi_\mu(\boldsymbol{a}_k, \boldsymbol{\lambda}_k) = \begin{bmatrix} \nabla \Phi(\boldsymbol{a}_k) - 2\boldsymbol{T}^{\mathrm{t}} \mathrm{Diag}(\boldsymbol{T}\boldsymbol{a}_k + \boldsymbol{t})^{-1}\boldsymbol{\mu} + \boldsymbol{T}^{\mathrm{t}}\boldsymbol{\lambda}_k \\ \boldsymbol{T}\boldsymbol{a}_k + \boldsymbol{t} - \boldsymbol{\Lambda}_k^{-1}\boldsymbol{\mu} \end{bmatrix} \quad (38)$$

and

$$\nabla^2 \Psi_\mu(\boldsymbol{a}_k, \boldsymbol{\lambda}_k) = \begin{bmatrix} \nabla^2 \Phi(\boldsymbol{a}_k) + 2\boldsymbol{T}^{\mathrm{t}} \mathrm{Diag}(\boldsymbol{T}\boldsymbol{a}_k + \boldsymbol{t})^{-2}\boldsymbol{T}\boldsymbol{\mu} & \boldsymbol{T}^{\mathrm{t}} \\ \boldsymbol{T} & \boldsymbol{\Lambda}_k^{-2}\boldsymbol{\mu} \end{bmatrix} \quad (39)$$

For all $k \in \mathbb{N}$, the linesearch ensures that $\boldsymbol{T}\boldsymbol{a}_k + \boldsymbol{t} > \boldsymbol{0}$ and $\boldsymbol{\lambda}_k > \boldsymbol{0}$. Since either $\Phi(\,\cdot\,)$ is strictly convex or $\boldsymbol{T}^{\mathrm{t}}\boldsymbol{T}$ is inversible, $\nabla^2 \Psi_\mu(\boldsymbol{a}_k, \boldsymbol{\lambda}_k)$ is positive definite so that Ψ_μ has a unique minimizer $(\hat{\boldsymbol{a}}_\mu, \hat{\boldsymbol{\lambda}}_\mu)$. Then, the same analysis as in [26, sec. 3] allows us to deduce that the sequence $\{(\boldsymbol{a}_k, \boldsymbol{\lambda}_k)\}_{k\in\mathbb{N}}$ resulting from the update equation (13) converges to $(\hat{\boldsymbol{a}}_\mu, \hat{\boldsymbol{\lambda}}_\mu)$.

2) Convergence Rate of the Inner Loop: Lengthy but straightforward calculations show that (16), (17), (38) and (39) lead to

$$\nabla^2 \Psi_\mu(\boldsymbol{a}_k, \boldsymbol{\lambda}_k)\boldsymbol{d}_k + \nabla \Psi_\mu(\boldsymbol{a}_k, \boldsymbol{\lambda}_k) = \begin{bmatrix} \mu \boldsymbol{T}^{\mathrm{t}} \mathrm{Diag}(\boldsymbol{T}\boldsymbol{a}_k + \boldsymbol{t})^{-1} \left(\mathrm{Diag}\left(\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\mu} - (\boldsymbol{T}\boldsymbol{a}_k + \boldsymbol{t})\right)\right)^{-1} \boldsymbol{T}\boldsymbol{d}_k^a \\ \boldsymbol{\Lambda}_k^{-1} \left(\mathrm{Diag}\left(\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\mu} - (\boldsymbol{T}\boldsymbol{a}_k + \boldsymbol{t})\right)\right) \boldsymbol{d}_k^\lambda \end{bmatrix}$$

with $\boldsymbol{d}_k = [(\boldsymbol{d}_k^a)^{\mathrm{t}} \quad (\boldsymbol{d}_k^\lambda)^{\mathrm{t}}]^{\mathrm{t}}$. Let us study the behaviour of $\|\nabla^2 \Psi_{\mu_k}(\boldsymbol{a}_k, \boldsymbol{\lambda}_k)\boldsymbol{d}_k + \nabla \Psi_{\mu_k}(\boldsymbol{a}_k, \boldsymbol{\lambda}_k)\|$ for large values of k. According to ([26], Lemma 3.1), $\{\boldsymbol{\lambda}_k\}_{k\in\mathbb{N}}$ and $\{\boldsymbol{T}\boldsymbol{a}_k + \boldsymbol{t}\}_{k\in\mathbb{N}}$ are bounded, and bounded away from zero. Moreover, the gradient of Ψ_μ tends to $\boldsymbol{0}$ as k goes to infinity, so that (38) yields $\lim_{k\to\infty} \boldsymbol{\Lambda}_k^{-1}\boldsymbol{\mu} - (\boldsymbol{T}\boldsymbol{a}_k + \boldsymbol{t}) = \boldsymbol{0}$. Therefore,

$$\|\nabla^2 \Psi_{\mu_k}(\boldsymbol{a}_k, \boldsymbol{\lambda}_k)\boldsymbol{d}_k + \nabla \Psi_{\mu_k}(\boldsymbol{a}_k, \boldsymbol{\lambda}_k)\| \underset{k\to\infty}{=} o(\|\boldsymbol{d}_k\|) \quad (40)$$

Hence, applying ([48], Theorem 3.5), there exists k_μ such that the stepsize $\alpha_k = 1$ is admissible for all $k \geqslant k_\mu$ and the sequence $\{(\boldsymbol{a}_k, \boldsymbol{\lambda}_k)\}_{k\geqslant k_\mu}$ resulting from (13) converges q-superlinearly to $(\hat{\boldsymbol{a}}_\mu, \hat{\boldsymbol{\lambda}}_\mu)$.

3) Convergence of the Outer Loop: For all k such that (23) holds:

$$\|\boldsymbol{r}_{\mu_k}^{\mathrm{prim}}\| \leqslant \sqrt{PN}\|\boldsymbol{r}_{\mu_k}^{\mathrm{prim}}\|_\infty \leqslant \sqrt{PN}\eta^{\mathrm{prim}}\mu_k, \quad (41)$$

$$\|\boldsymbol{r}_{\mu_k}^{\mathrm{dual}}\| \leqslant \delta_k + \mu_k \leqslant (QN\eta^{\mathrm{dual}} + 1)\mu_k. \quad (42)$$

Furthermore,

$$\mu_{k+1} \leqslant \theta\eta^{\mathrm{dual}}\mu_k \quad (43)$$

with $\theta\eta^{\mathrm{dual}} \in (0, 1)$ and $\mu_0 > 0$ so that $\{\mu_k\}_{k\in\mathbb{N}}$ converges to 0 as k tends to infinity. Thus, according to (41), (42) and ([26], Theorem 5.1), the outer loop of Algorithm 1 generates a bounded sequence $\{(\boldsymbol{a}_k, \boldsymbol{\lambda}_k)\}$ whose accumulation points are primal-dual solutions of problem (11). Finally, if $\Phi(\,\cdot\,)$ is strictly convex, the solution $\hat{\boldsymbol{a}}$ of (11) is unique, and the outer iterates $\{\boldsymbol{a}_k\}$ converge to $\hat{\boldsymbol{a}}$.

REFERENCES

[1] J. R. Scott, *Remote Sensing: The Image Chain Approach*. New York, NY, USA: Oxford Univ. Press, 1997.

[2] C.-I Chang, *Hyperspectral Data Exploitation*. New York, NY, USA: Wiley Interscience, 2007.

[3] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, D. Qian, P. Gader, and J. Chanussot, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 2, pp. 354–379, Apr. 2012.

[4] N. Keshava and J. F. Mustard, "Spectral unmixing," *IEEE Signal Process. Mag.*, vol. 19, no. 1, pp. 44–57, Jan. 2002.

[5] S. Moussaoui, H. Hauksdóttir, F. Schmidt, C. Jutten, J. Chanussot, D. Brie, S. Douté, and J. Benediktsson, "On the decomposition of Mars hyperspectral data by ICA and Bayesian positive source separation," *Neurocomputing*, vol. 71, no. 10–12, pp. 2194–2208, Jun. 2008.

[6] N. Dobigeon, S. Moussaoui, J.-Y. Tourneret, and C. Carteret, "Bayesian separation of spectral sources under non-negativity and full additivity constraints," *Signal Process.*, vol. 89, no. 12, pp. 2657–2669, Dec. 2009.

[7] S. Jia and Y. Qian, "Constrained nonnegative matrix factorization for hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 1, pp. 161–173, 2009.

[8] A. Huck, M. Guillaume, and J. Blanc-Talon, "Minimum dispersion constrained nonnegative matrix factorization to unmix hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 6, pp. 2590–2602, 2010.

[9] C. L. Lawson and R. J. Hanson, *Solving Least-Squares Problems*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1974.

[10] R. Bro and S. De Jong, "A fast non-negativity constrained least squares algorithm," *J. Chemometr.*, vol. 11, pp. 393–401, 1997.

[11] J. J. Settle and N. A. Drake, "Linear mixing and the estimation of ground cover proportions," *Int. J. Remote Sens.*, vol. 14, no. 6, pp. 1159–1177, 1993.

[12] D. C. Heinz and C.-I Chang, "Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 3, pp. 529–545, 2001.

[13] N. Dobigeon, J.-Y. Tourneret, and C.-I Chang, "Semi-supervised linear spectral unmixing using a hierarchical Bayesian model for hyperspectral imagery," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 2684–2695, 2008.

[14] A. Plaza, J. Plaza, A. Paz, and S. Sanchez, "Parallel hyperspectral image and signal processing," *IEEE Signal Process. Mag.*, vol. 28, pp. 119–126, May 2011.

[15] S. Sanchez, A. Paz, G. Martin, and A. Plaza, "Parallel unmixing of remotely sensed hyperspectral images on commodity graphics processing units," *Concurrency and Computation: Practice and Experience*, vol. 23, pp. 1538–1557, 2011.

[16] C. González, J. Resano, A. Plaza, and D. Mozos, "FPGA implementation of abundance estimation for spectral unmixing of hyperspectral data using the image space reconstruction algorithm," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 1, pp. 248–261, 2012.

[17] P. Honeine and C. Richard, "Geometric unmixing of large hyperspectral images: A barycentric coordinate approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 6, pp. 2185–2195, Jun. 2012.

[18] R. Heylen, D. Burazerovic, and P. Scheunders, "Fully constrained least squares spectral unmixing by simplex projection," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 11, pp. 4112–4122, Nov. 2011.

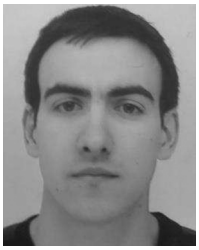
[19] J. M. Bioucas-Dias and M. A. T. Figueiredo, "Alternating direction algorithms for constrained sparse regression: Application to hyperspectral unmixing," in *Proc. 2nd IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS'10)*, Reykjavik, Iceland, 2010.

[20] M. V. Afonso, J. M. Bioucas-Dias, and M. A. T. Figueiredo, "An augmented Lagrangian aproach to the constrained optimization formulation of imaging inverse problems," *IEEE Trans. Image Process.*, vol. 20, no. 3, pp. 681–695, 2011.

[21] M. H. Wright, "Interior methods for constrained optimization," in *Acta Numerica 1992*. New York, NY, USA: Cambridge Univ. Press, 1991, pp. 341–407.
[22] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge Univ. Press, 2004.
[23] C. A. Johnson, J. Seidel, and A. Sofer, "Interior-point methodology for 3-D PET reconstruction," *IEEE Trans. Med. Imag.*, vol. 19, no. 4, Apr. 2000.
[24] S. Bonettini and T. Serafini, "Non-negatively constrained image deblurring with an inexact interior point method," *J. Comput. Appl. Math.*, vol. 231, no. 1, pp. 236–248, 2009.
[25] N. Keshava, "A survey of spectral unmixing," *Lincoln Lab. J.*, vol. 14, no. 1, pp. 55–78, 2003.
[26] P. Armand, J. C. Gilbert, and S. Jan-Jégou, "A feasible BFGS interior point algorithm for solving strongly convex minimization problems," *SIAM J. Optimization*, vol. 11, pp. 199–222, 2000.
[27] S. J. Wright, *Primal-Dual Interior-Point Methods*. Philadelphia, PA, USA: SIAM, 1997.
[28] A. Forsgren, P. E. Gill, and M. H. Wright, "Interior methods for nonlinear optimization," *SIAM Rev.*, vol. 44, no. 4, pp. 525–597, 2002.
[29] M. H. Wright, "Some properties of the Hessian of the logarithmic barrier function," *Math. Programm.*, vol. 67, no. 2, pp. 265–295, 1994.
[30] M. H. Wright, "Ill-conditioning and computational error in interior methods for nonlinear programming," *SIAM J. Optimization*, vol. 9, no. 1, pp. 84–111, 1998.
[31] A. Conn, N. Gould, and P. L. Toint, , G. Di Pillo and F. Giannessi, Eds., "A primal-dual algorithm for minimizing a nonconvex function subject to bounds and nonlinear constraints," in *Nonlinear Optimization and Applications*, 2nd ed. Boston, MA, USA: Kluwer Academic, 1996.
[32] A. S. El-Bakry, R. A. Tapia, T. Tsuchiya, and Y. Zhang, "On the formulation and theory of the Newton interior-point method for nonlinear programming," *J. Optimization Theory and Applications*, vol. 89, pp. 507–541, June 1996.
[33] C. F. Van Loan, "The ubiquitous Kronecker product," *J. Computational and Applied Mathematics*, vol. 123, no. 1–2, pp. 85–100, 2000.
[34] R. N. Clark, G. A. Swayze, A. Gallagher, T. V. King, and W. M. Calvin, "The U.S. Geological Survey Digital Spectral Library: Version 1: 0.2 to 3.0 μm," U.S. Geological Survey, Denver, CO, USA, Open File Rep. 93-592, 1993.
[35] A. Plaza and C.-I Chang, "An improved N-FINDR algorithm in implementation," in *Proc. SPIE Vol. 5806*, 2005, p. 299.
[36] L. R. Gaddis, L. A. Soderblom, H. H. Kieffer, K. J. Becker, J. Torsen, and K. F. Mullins, "Decomposition of AVIRIS spectra: Extraction of spectral reflectance, atmospheric, and instrumental components," *IEEE Trans. Geosci. Remote Sens.*, vol. 34, no. 1, pp. 163–178, Jan. 1996.
[37] J. M. P. Nascimento and J. M. Bioucas-Dias, "Vertex component analysis: A fast algorithm to unmix hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 4, pp. 898–910, Apr. 2005.
[38] D. B. Kirk and W. H. Wen-Mei, *Programming Massively Parallel Processors: A Hands-On Approach*. Norwood, MA, USA: Morgan Kaufmann, 2010.
[39] NVIDIA CUDA C Programming Guide Version 4.2, 2012.
[40] O. Caspary, P. Nus, and T. Cecchin, "The source number estimation based on Gerschgorin radii," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'98)*, 1998, vol. 4, pp. 1993–1996.
[41] C.-I Chang and D. Qian, "Estimation of number of spectrally distinct signal sources in hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 3, pp. 608–619, 2004.
[42] B. Luo, J. Chanussot, D. Sylvain, and L. Zhang, "Empirical automatic estimation of the number of endmembers in hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 1, pp. 24–28, Jan. 2013.
[43] C.-I Chang, "Spectral information divergence for hyperspectral image analysis," in *Proc. IEEE Int. Geoscience and Remote Sensing Symp. (IGARSS'99)*, 1999, vol. 1, pp. 509–511.
[44] A. Plaza, P. Martinez, R. Perez, and J. Plaza, "A quantitative and comparative analysis of endmember extraction algorithms from hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 3, pp. 650–663, Mar. 2004.
[45] M. D. Iordache, J. M. Bioucas-Dias, and A. Plaza, "Total variation spatial regularization for sparse hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 11, pp. 4484–4502, 2012.
[46] J. Idier, "Convex half-quadratic criteria and interacting auxiliary variables for image restoration," *IEEE Trans. Image Process.*, vol. 10, no. 7, pp. 1001–1009, 2001.
[47] S. Moussaoui, E. Chouzenoux, and J. Idier, "Primal-dual interior point optimization for penalized least squares estimation of abundance maps in hyperspectral imaging," in *Proc. IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS'12)*, Shangai, China, Jun. 2012.
[48] J. Nocedal and S. J. Wright, *Numerical Optimization*. New York, NY, USA: Springer-Verlag, 1999.

Émilie Chouzenoux received the engineering degree from École Centrale, Nantes, France, in 2007, and the Ph.D. degree in signal processing from the Institut de Recherche en Communications et Cybernétique, Nantes, France, in 2010.

She is currently an Assistant Professor with the University of Paris-Est, Champs-sur-Marne, France (LIGM, UMR CNRS 8049). Her research interests are in convex and nonconvex optimization algorithms for large scale problems of image and signal reconstruction.



Maxime Legendre received the engineering degree from École Centrale, Nantes, France, in 2011. He is currently a Ph.D. student with the Institut de Recherche en Communications et Cybernétique, Nantes (IRCCYN, UMR CNRS 6597).

His research interests are in constrained optimization algorithms for inverse problems of image processing, and their applications to spatial imaging.

Saïd Moussaoui received the State engineering degree from Ecole Nationale Polytechnique, Algiers, Algeria, in 2001, and the Ph.D. degree in automatic control and signal processing from Université Henri Poincaré, Nancy, France, in 2005.

He is currently Assistant Professor with École Centrale de Nantes, Nantes, France. Since September 2006, he has been with the Institut de Recherche en Communications et Cybernétique, Nantes (IRCCYN, UMR CNRS 6597). His research interests are in statistical signal and image processing including source separation, Bayesian estimation, and their applications to spectrometry and hyperspectral imaging.

Jérôme Idier was born in France in 1966. He received the Diploma degree in electrical engineering from École Supérieure d'Électricité, Gif-sur-Yvette, France, in 1988, and the Ph.D. degree in physics from University of Paris-Sud, Orsay, France, in 1991.

In 1991, he joined the Centre National de la Recherche Scientifique. He is currently a Senior Researcher with the Institut de Recherche en Communications et Cybernétique, Nantes, France. His major scientific interests are in probabilistic approaches to inverse problems for signal and image processing.

Dr. Idier is currently elected member of the French National Committee for Scientific Research.

A.5 Efficient Gaussian sampling for solving large-scale inverse problems using MCMC

C. Gilavert, **S. Moussaoui** et J. Idier, *IEEE Trans. Signal Processing*, accepté, octobre 2014.

Ce papier est consacré à la proposition d'une méthode de simulation de vecteurs gaussiens pour la résolution de problèmes inverses. L'apport de ce papier est de corriger un comportement erroné de certaines stratégies employées par certains auteurs et de proposer une technique de réglage adaptatif de l'algorithme de telle manière à optimiser son temps de calcul.

Efficient Gaussian Sampling for Solving Large-Scale Inverse Problems using MCMC

Clément Gilavert, Saïd Moussaoui, *Member, IEEE,* and Jérôme Idier, *Member, IEEE*

***Abstract*—The resolution of many large-scale inverse problems using MCMC methods requires a step of drawing samples from a high dimensional Gaussian distribution. While direct Gaussian sampling techniques, such as those based on Cholesky factorization, induce an excessive numerical complexity and memory requirement, sequential coordinate sampling methods present a low rate of convergence. Based on the reversible jump Markov chain framework, this paper proposes an efficient Gaussian sampling algorithm having a reduced computation cost and memory usage, while maintaining the theoretical convergence of the sampler. The main feature of the algorithm is to perform an approximate resolution of a linear system with a truncation level adjusted using a self-tuning adaptive scheme allowing to achieve the minimal computation cost per effective sample. The connection between this algorithm and some existing strategies is given and its performance is illustrated on a linear inverse problem of image resolution enhancement.**

***Index Terms*—Multivariate Gaussian sampling, Gibbs algorithm, reversible jump Monte Carlo, adaptive MCMC, conjugate gradient.**

I. Introduction

A common inverse problem arising in many signal and image processing applications is to estimate a hidden object $\boldsymbol{x} \in \mathbb{R}^N$ (*e.g.,* an image or a signal) from a set of measurements $\boldsymbol{y} \in \mathbb{R}^M$ given an observation model [1, 2]. The most frequent case is that of a linear model between $\boldsymbol{x}$ and $\boldsymbol{y}$ according to

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n}, \tag{1}$$

with $\boldsymbol{H} \in \mathbb{R}^{M\times N}$ the known observation matrix and $\boldsymbol{n}$ an additive noise term representing measurement errors and model uncertainties. Such a linear model covers many real problems such as, for instance, denoising [3], deblurring [4], and reconstruction from projections [5, 6].

The statistical estimation of $\boldsymbol{x}$ in a Bayesian simulation framework [7, 8] firstly requires the formulation of the posterior distribution $P(\boldsymbol{x}, \Theta|\boldsymbol{y})$, with Θ a set of unknown hyper-parameters. Pseudo-random samples of $\boldsymbol{x}$ are then drawn from this posterior distribution. Finally, a Bayesian estimator (posterior mean, maximum *a posteriori*) is computed from the set of generated samples. Other quantities of interest, such as posterior variances, can be estimated likewise. Within the *standard Monte Carlo* framework, independent realizations of the posterior law must be generated, which is rarely possible in realistic cases of inverse problems. One rather resorts to *Markov Chain Monte Carlo* (MCMC) schemes, where Markovian dependencies between successive samples are allowed. A very usual sampling scheme is then to iteratively draw realizations from the conditional posterior densities $P(\Theta|\boldsymbol{x}, \boldsymbol{y})$ and $P(\boldsymbol{x}|\Theta, \boldsymbol{y})$, according to a *Gibbs sampler* [11].

In such a context, when independent Gaussian models $\mathcal{N}\left(\boldsymbol{\mu}_y, \boldsymbol{R}_y\right)$ and $\mathcal{N}\left(\boldsymbol{\mu}_x, \boldsymbol{R}_x\right)$ are assigned to the noise statistics and to the unknown object distribution, respectively, the set of hyper-parameters Θ determines the mean and the covariance of the latter two distributions. This framework also covers the case of priors based on hierarchical or latent Gaussian models such as Gaussian scale mixtures [9, 10] and Gaussian Markov fields [11, 12]. The additional parameters of such models are then included in Θ. According to such modeling, the conditional posterior distribution $P(\boldsymbol{x}|\Theta, \boldsymbol{y})$ is also Gaussian, $\mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{Q}^{-1}\right)$, with a precision matrix $\boldsymbol{Q}$ (*i.e.*, the inverse of the covariance matrix) given by:

$$\boldsymbol{Q} = \boldsymbol{H}^{\mathrm{t}}\boldsymbol{R}_y^{-1}\boldsymbol{H} + \boldsymbol{R}_x^{-1}, \tag{2}$$

and a mean vector $\boldsymbol{\mu}$ such that:

$$\boldsymbol{Q}\boldsymbol{\mu} = \boldsymbol{H}^{\mathrm{t}}\boldsymbol{R}_y^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y) + \boldsymbol{R}_x^{-1}\boldsymbol{\mu}_x. \tag{3}$$

Let us remark that the precision matrix $\boldsymbol{Q}$ generally depends on the hyper-parameter set Θ through $\boldsymbol{R}_y$ and $\boldsymbol{R}_x$, so that $\boldsymbol{Q}$ is a varying matrix along the Gibbs sampler iterations. Moreover, the mean vector $\boldsymbol{\mu}$ is expressed as the solution of a linear system where $\boldsymbol{Q}$ is the normal matrix.

In order to draw samples from the conditional posterior distribution $P(\boldsymbol{x}|\Theta, \boldsymbol{y})$, a usual way is to firstly perform the Cholesky factorization of the covariance matrix [13, 14]. Since equation (2) yields the precision matrix $\boldsymbol{Q}$ rather than the covariance matrix, Rue [15] proposed to compute the Cholesky decomposition of $\boldsymbol{Q}$, *i.e.*, $\boldsymbol{Q} = \boldsymbol{C}_q\boldsymbol{C}_q^{\mathrm{t}}$, and to solve the triangular system $\boldsymbol{C}_q^{\mathrm{t}}\boldsymbol{x} = \boldsymbol{\omega}$, where $\boldsymbol{\omega}$ is a vector of independent Gaussian variables of zero mean and unit variance. Moreover, the Cholesky factorization is exploited to calculate the mean $\boldsymbol{\mu}$ from (3) by solving two triangular systems sequentially.

The Cholesky factorization of $\boldsymbol{Q}$ generally requires $\mathcal{O}(N^3)$ operations. Spending such a numerical cost at each iteration of the sampling scheme rapidly becomes prohibitive for large values of N. In specific cases where $\boldsymbol{Q}$ belongs to certain families of structured matrices, the factorization can be obtained with a reduced numerical complexity, *e.g.*, $\mathcal{O}(N^2)$ when $\boldsymbol{Q}$ is Toeplitz [16] or even $\mathcal{O}(N \log N)$ when $\boldsymbol{Q}$ is circulant [17]. Sparse matrices can be also factorized at a reduced cost [15, 18]. Alternative approaches to the Cholesky factorization are

C. Gilavert, S. Moussaoui, and J. Idier are with IRCCyN, CNRS UMR 6597, Ecole Centrale Nantes, 1 rue de la Noë, 44321, Nantes Cedex 3, France.

Corresponding author: said.moussaoui@@ec-nantes.fr.

This work was supported by the CNRS and the French Région des Pays de la Loire (France).

based on using an iterative method for the calculation of the inverse square root matrix of $\boldsymbol{Q}$ using Krylov subspace methods [19–21]. In practice, even in such favorable cases, the factorization often remains a burdensome operation to be performed at each iteration of the Gibbs sampler.

The numerical bottleneck represented by the factorization techniques can be removed by using alternate schemes that bypass the step of exactly sampling $P(\boldsymbol{x}|\Theta,\boldsymbol{y})$. For instance, a simple alternative solution is to sequentially sample each entry of $\boldsymbol{x}$ given the other variables according to a scalar Gibbs scheme. However, such a scalar approach reveals extremely inefficient when $P(\boldsymbol{x}|\Theta,\boldsymbol{y})$ is strongly correlated, since each conditional sampling step will produce a move of very small variance. As a consequence, a huge number of iterations will be required to reach convergence [22]. A better trade-off between the numerical cost of each iteration and the overall convergence speed of the sampler must be found.

In this paper, we focus on a two-step approach named *Independent Factor Perturbation* in [12] and *Perturbation-Optimization* in [23] (see also [18, 24]). It consists in

- drawing a sample $\boldsymbol{\eta}$ from $\mathcal{N}(\boldsymbol{Q}\boldsymbol{\mu},\boldsymbol{Q})$,
- solving the linear system $\boldsymbol{Q}\boldsymbol{x}=\boldsymbol{\eta}$.

It can be easily checked that, when the linear system is solved exactly, the new sample $\boldsymbol{x}$ is distributed according to $\mathcal{N}\left(\boldsymbol{\mu},\boldsymbol{Q}^{-1}\right)$. Hereafter, we refer to this method as *Exact Perturbation Optimization* (E-PO). However, the numerical cost of E-PO is typically as high as the Cholesky factorization of $\boldsymbol{Q}$. Therefore, an essential element of the Perturbation Optimization approach is to truncate the linear system solving [12, 23, 24], by running a limited number of iterations of the conjugate gradient method (CG). For the sake of clarity, let us call the resulting version *Truncated Perturbation Optimization* (T-PO).

Skipping from E-PO to T-PO allows to strongly reduce the numerical cost of each iteration. However, let us stress that no convergence analysis of T-PO exists, to our best knowledge. It is only argued that a well-chosen truncation level will induce a significant reduction of the numerical cost and a small estimation error. The way the latter error alters the convergence towards the target distribution remains a fully open issue, that has not been discussed in existing contributions. Moreover, how the resolution accuracy should be chosen in practice is also an open question.

A first contribution of the present paper is to bring practical evidence that the T-PO does not necessarily converge towards the target distribution (see Section IV). In practice, the implicit trade-off within T-PO is between the computational cost and the error induced on the target distribution, depending on the adopted truncation level. Our second contribution is to propose a new scheme similar to T-PO, but with a guarantee of convergence to the target distribution, whatever the truncation level. We call the resulting scheme *Reversible Jump Perturbation Optimization* (RJPO), since it incorporates an accept-reject step derived within the Reversible Jump MCMC (RJ-MCMC) framework [25, 26]. The numerical cost of the proposed test is marginal, so that RJPO has nearly the same cost per iteration as T-PO. Finally, we propose an unsupervised tuning of the truncation level allowing to automatically achieve a pre-specified overall acceptance rate or even to minimize the computation cost per effective sample. Consequently, the resulting algorithm can be viewed as an adaptive MCMC sampler [27–29].

The rest of the paper is organized as follows: Section II introduces the global framework of RJ-MCMC and presents a general scheme to sample Gaussian vectors. Section III considers a specific application of the previous results, which finally boils down to the proposed RJPO sampler. Section IV analyses the performance of RJPO compared to T-PO on simple toy problems and presents the adaptive RJPO which incorporates an automatic tuning of the truncation level. Finally, in section V, an example of linear inverse problem, the unsupervised image resolution enhancement is presented to illustrate the applicability of the method. These results confirm the superiority of the RJPO algorithm over the usual sampling approach, based on Cholesky factorization, in terms of computational cost and memory usage.

II. The reversible jump MCMC framework

The sampling procedure consists on iteratively constructing a Markov chain whose distribution asymptotically converges to the target distribution $P_{\boldsymbol{X}}$. Let $\underline{\boldsymbol{x}} \in \mathbb{R}^N$ be the current sample of the Markov chain and $\overline{\boldsymbol{x}}$ the new sample obtained according to a transition kernel derived in the reversible jump framework [25, 26] .

A. General framework

In the constant dimension case, the Reversible Jump MCMC strategy introduces an auxiliary variable $\boldsymbol{z} \in \mathbb{R}^L$, obtained from a distribution $P_{\boldsymbol{Z}}(\boldsymbol{z}|\underline{\boldsymbol{x}})$ and a deterministic move according to a differentiable transformation

$$\begin{aligned}\boldsymbol{\phi} : \left(\mathbb{R}^N \times \mathbb{R}^L\right) &\mapsto \left(\mathbb{R}^N \times \mathbb{R}^L\right)\\ (\underline{\boldsymbol{x}},\boldsymbol{z}) &\mapsto (\boldsymbol{x},\boldsymbol{s}).\end{aligned}$$

This transformation must be reversible, that is $\boldsymbol{\phi}(\boldsymbol{x},\boldsymbol{s}) = (\underline{\boldsymbol{x}},\boldsymbol{z})$. The new sample $\overline{\boldsymbol{x}}$ is thereby obtained by submitting $\boldsymbol{x}$ (resulting from the deterministic move) to an accept-reject step with an acceptance probability given by

$$\alpha(\underline{\boldsymbol{x}},\boldsymbol{x}|\boldsymbol{z}) = \min\left(1, \frac{P_{\boldsymbol{X}}(\boldsymbol{x})P_{\boldsymbol{Z}}(\boldsymbol{s}|\boldsymbol{x})}{P_{\boldsymbol{X}}(\underline{\boldsymbol{x}})P_{\boldsymbol{Z}}(\boldsymbol{z}|\underline{\boldsymbol{x}})}|J_{\boldsymbol{\phi}}(\underline{\boldsymbol{x}},\boldsymbol{z})|\right),$$

with $J_{\boldsymbol{\phi}}(\underline{\boldsymbol{x}},\boldsymbol{z})$ the Jacobian determinant of the transformation $\boldsymbol{\phi}$ at $(\underline{\boldsymbol{x}},\boldsymbol{z})$. In fact, the choice of the conditional distribution $P_{\boldsymbol{Z}}(\cdot)$ and the transformation $\boldsymbol{\phi}(\cdot)$ must be adapted to the target distribution $P_{\boldsymbol{X}}(\cdot)$ and affects the resulting Markov chain properties in terms of correlation and convergence rate.

B. Gaussian Case

To draw samples from a Gaussian distribution $\mathcal{N}\left(\boldsymbol{\mu},\boldsymbol{Q}^{-1}\right)$, we generalize the scheme adopted in [30]. We consider $L = N$, and take an auxiliary variable $\boldsymbol{z} \in \mathbb{R}^N$ sampled from

$$P_{\boldsymbol{Z}}(\boldsymbol{z}|\underline{\boldsymbol{x}}) = \mathcal{N}\left(\boldsymbol{A}\underline{\boldsymbol{x}} + \boldsymbol{b}, \boldsymbol{B}\right), \tag{4}$$

where $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{b}$ denote a $N \times N$ real matrix, a $N \times N$ real positive definite matrix and a $N \times 1$ real vector, respectively. The choice of the latter three quantities will be discussed later.

The proposed deterministic move is performed using the transformation ϕ such that

$$\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{s} \end{pmatrix} = \begin{pmatrix} \phi_1(\underline{\boldsymbol{x}}, \boldsymbol{z}) \\ \phi_2(\underline{\boldsymbol{x}}, \boldsymbol{z}) \end{pmatrix} = \begin{pmatrix} -\underline{\boldsymbol{x}} + \boldsymbol{f}(\boldsymbol{z}) \\ \boldsymbol{z} \end{pmatrix}, \quad (5)$$

with functions $\left(\boldsymbol{f} : \mathbb{R}^N \mapsto \mathbb{R}^N\right)$, $\left(\phi_1 : (\mathbb{R}^N \times \mathbb{R}^N) \mapsto \mathbb{R}^N\right)$ and $\left(\phi_2 : (\mathbb{R}^N \times \mathbb{R}^N) \mapsto \mathbb{R}^N\right)$.

Proposition 1. *Let an auxiliary variable $\boldsymbol{z}$ be obtained according to* (4) *and a proposed sample $\boldsymbol{x}$ resulting from* (5)*. Then the acceptance probability is*

$$\alpha(\underline{\boldsymbol{x}}, \boldsymbol{x}|\boldsymbol{z}) = \min\left(1, e^{-\boldsymbol{r}(\boldsymbol{z})^{\mathrm{t}}(\underline{\boldsymbol{x}}-\boldsymbol{x})}\right), \quad (6)$$

with

$$\boldsymbol{r}(\boldsymbol{z}) = \boldsymbol{Q}\boldsymbol{\mu} + \boldsymbol{A}^{\mathrm{t}}\boldsymbol{B}^{-1}(\boldsymbol{z} - \boldsymbol{b}) - \frac{1}{2}\left(\boldsymbol{Q} + \boldsymbol{A}^{\mathrm{t}}\boldsymbol{B}^{-1}\boldsymbol{A}\right)\boldsymbol{f}(\boldsymbol{z}). \quad (7)$$

In particular, the acceptance probability equals one when $\boldsymbol{f}(\boldsymbol{z})$ is defined as the exact solution of the linear system

$$\frac{1}{2}\left(\boldsymbol{Q} + \boldsymbol{A}^{\mathrm{t}}\boldsymbol{B}^{-1}\boldsymbol{A}\right)\boldsymbol{f}(\boldsymbol{z}) = \boldsymbol{Q}\boldsymbol{\mu} + \boldsymbol{A}^{\mathrm{t}}\boldsymbol{B}^{-1}(\boldsymbol{z} - \boldsymbol{b}). \quad (8)$$

Proof: See appendix A. ∎

Let us remark that $\boldsymbol{b}$ is a dummy parameter, since the residual $\boldsymbol{r}(\boldsymbol{z})$ (and thus $\alpha(\underline{\boldsymbol{x}}, \boldsymbol{x}|\boldsymbol{z})$) depends on $\boldsymbol{b}$ through $\boldsymbol{z}-\boldsymbol{b}$ only. However, choosing a specific expression of $\boldsymbol{b}$ jointly with $\boldsymbol{A}$ and $\boldsymbol{B}$ will lead to a simplified expression of $\boldsymbol{r}(\boldsymbol{z})$ in the next section.

Proposition 1 plays a central role in our proposal. When the exact resolution of (8) is numerically costly, it allows to derive a procedure where the resolution is performed only approximately, at the expense of a lowered acceptance probability. The conjugate gradient algorithm stopped before convergence, is a typical example of an efficient algorithm to approximately solve (8).

Proposition 2. *Let an auxiliary variable $\boldsymbol{z}$ be obtained according to* (4)*, a proposed sample $\boldsymbol{x}$ resulting from* (5) *and $\boldsymbol{f}(\boldsymbol{z})$ be the exact solution of* (8)*. The correlation between two successive samples is zero if and only if matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are chosen such that*

$$\boldsymbol{A}^{\mathrm{t}}\boldsymbol{B}^{-1}\boldsymbol{A} = \boldsymbol{Q}. \quad (9)$$

Proof: See Appendix B. ∎

Many couples $(\boldsymbol{A}, \boldsymbol{B})$ fulfill condition (9):

- Consider the Cholesky factorization $\boldsymbol{Q} = \boldsymbol{C}_q\boldsymbol{C}_q^{\mathrm{t}}$ and take $\boldsymbol{A} = \boldsymbol{C}_q^{\mathrm{t}}$, $\boldsymbol{B} = \boldsymbol{I}$. It leads to $\boldsymbol{z} = \boldsymbol{C}_q^{\mathrm{t}}\underline{\boldsymbol{x}} + \boldsymbol{b} + \boldsymbol{\omega}$ with $\boldsymbol{\omega} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_N)$. According to (8), the next sample $\overline{\boldsymbol{x}} = -\underline{\boldsymbol{x}} + \boldsymbol{f}(\boldsymbol{z})$, will be obtained as

$$\begin{aligned}\overline{\boldsymbol{x}} &= -\underline{\boldsymbol{x}} + \left(\boldsymbol{C}_q\boldsymbol{C}_q^{\mathrm{t}}\right)^{-1}\left(\boldsymbol{Q}\boldsymbol{\mu} + \boldsymbol{C}_q(\boldsymbol{z} - \boldsymbol{b})\right), \\ &= \left(\boldsymbol{C}_q^{\mathrm{t}}\right)^{-1}\left(\boldsymbol{C}_q^{-1}\boldsymbol{Q}\boldsymbol{\mu} + \boldsymbol{\omega}\right).\end{aligned}$$

  Such an update scheme is exactly the same as the one proposed by Rue in [15].
- The particular configuration

$$\boldsymbol{A} = \boldsymbol{B} = \boldsymbol{Q} \quad \text{and} \quad \boldsymbol{b} = \boldsymbol{Q}\boldsymbol{\mu}. \quad (10)$$

is retained in the sequel, since:

i) $\boldsymbol{A}^{\mathrm{t}}\boldsymbol{B}^{-1}\boldsymbol{A} = \boldsymbol{Q}$ is a condition of Proposition 2,

ii) $\boldsymbol{b} = \boldsymbol{Q}\boldsymbol{\mu}$ simplifies equation (8) to a linear system $\boldsymbol{Q}\boldsymbol{f}(\boldsymbol{z}) = \boldsymbol{z}$.

In particular, it allows to make a clear connection between our RJ-MCMC approach and the E-PO algorithm in the case of an exact resolution of the linear system. It also allows to simplify the accept-reject step that must be considered when an approximate resolution is retained.

III. RJ-MCMC algorithms for Sampling Gaussian Distributions

A. Sampling the Auxiliary Variable

According to the configuration (10), the auxiliary variable $\boldsymbol{z}$ is distributed according to $\mathcal{N}(\boldsymbol{Q}\underline{\boldsymbol{x}} + \boldsymbol{Q}\boldsymbol{\mu}, \boldsymbol{Q})$. It can then be expressed as $\boldsymbol{z} = \boldsymbol{Q}\underline{\boldsymbol{x}} + \boldsymbol{\eta}$, $\boldsymbol{\eta}$ being distributed according to $\mathcal{N}(\boldsymbol{Q}\boldsymbol{\mu}, \boldsymbol{Q})$. Consequently, the auxiliary variable sampling step is reduced to the simulation of $\boldsymbol{\eta}$, which is the perturbation step in the PO algorithm. In [12, 23], a subtle way of sampling $\boldsymbol{\eta}$ is proposed. It consists in exploiting equation (3) and *perturbing* each factor separately:

1) Sample $\boldsymbol{\eta}_y \sim \mathcal{N}\left(\boldsymbol{y} - \boldsymbol{\mu}_y, \boldsymbol{R}_y\right)$,
2) Sample $\boldsymbol{\eta}_x \sim \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{R}_x)$,
3) Set $\boldsymbol{\eta} = \boldsymbol{H}^{\mathrm{t}}\boldsymbol{R}_y^{-1}\boldsymbol{\eta}_y + \boldsymbol{R}_x^{-1}\boldsymbol{\eta}_x$, a sample of $\mathcal{N}(\boldsymbol{Q}\boldsymbol{\mu}, \boldsymbol{Q})$.

It is important to notice that such a tricky method is interesting since matrices $\boldsymbol{R}_y$ and $\boldsymbol{R}_x$ have often a simple structure if not diagonal.

We emphasize that this perturbation step can be applied more generally for the sampling of any Gaussian distribution, for which a factored expression of the precision matrix $\boldsymbol{Q}$ is available under the form $\boldsymbol{Q} = \boldsymbol{F}^{\mathrm{t}}\boldsymbol{F}$, with matrix $\boldsymbol{F} \in \mathbb{R}^{N' \times N}$. In such a case, $\boldsymbol{\eta} = \boldsymbol{Q}\boldsymbol{\mu} + \boldsymbol{F}^{\mathrm{t}}\boldsymbol{w}$, where $\boldsymbol{\omega} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{N'})$.

B. Exact resolution case: the E-PO algorithm

As stated by proposition 1, the exact resolution of system (8) implies an acceptance probability equal to one. The resulting sampling procedure is thus based on the following steps:

1) Sample $\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{Q}\boldsymbol{\mu}, \boldsymbol{Q})$,
2) Set $\boldsymbol{z} = \boldsymbol{Q}\underline{\boldsymbol{x}} + \boldsymbol{\eta}$,
3) Take $\overline{\boldsymbol{x}} = -\underline{\boldsymbol{x}} + \boldsymbol{Q}^{-1}\boldsymbol{z}$.

Let us remark that $\overline{\boldsymbol{x}} = -\underline{\boldsymbol{x}} + \boldsymbol{Q}^{-1}(\boldsymbol{Q}\underline{\boldsymbol{x}} + \boldsymbol{\eta}) = \boldsymbol{Q}^{-1}\boldsymbol{\eta}$, so the handling of variable $\boldsymbol{z}$ can be skipped and Steps 2 and 3 can be merged to an equivalent but more direct step:

2) Set $\overline{\boldsymbol{x}} = \boldsymbol{Q}^{-1}\boldsymbol{\eta}$.

In the exact resolution case, the obtained algorithm is thus identical to the E-PO algorithm [23].

According to Proposition 2, E-PO enjoys the property that each sample is totally independent from the previous ones. However, a drawback is that the exact resolution of the linear system $\boldsymbol{Q}\boldsymbol{x} = \boldsymbol{\eta}$ often leads to an excessive numerical complexity and memory usage in high dimensions [12]. In practice, early stopping of an iterative solver such as the linear conjugate gradient algorithm is used, yielding the Truncated Perturbation Optimization (T-PO) version. The main point is that, up to our knowledge, there is no theoretical analysis

of the efficiency of T-PO and of its convergence to the target distribution. Actually, the simulation tests provided in Section IV indicate that convergence to the target distribution is not guaranteed. However, as shown in the next subsection, two slight but decisive modifications of T-PO lead us to the RJPO version, which is a provably convergent algorithm.

C. Approximate resolution case: the RJPO algorithm

In the case of configuration (10), equation (7) reduces to

$$\boldsymbol{r}(\boldsymbol{z}) = \boldsymbol{z} - \boldsymbol{Q}\boldsymbol{f}(\boldsymbol{z}). \tag{11}$$

Therefore, a first version of the RJPO algorithm is as follows:

1) Sample $\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{Q}\boldsymbol{\mu}, \boldsymbol{Q})$.
2) Set $\boldsymbol{z} = \boldsymbol{Q}\underline{\boldsymbol{x}} + \boldsymbol{\eta}$. Solve the linear system $\boldsymbol{Q}\boldsymbol{u} = \boldsymbol{z}$, in an approximate way. Let $\widehat{\boldsymbol{u}}$ denote the obtained solution, $\boldsymbol{r}(\boldsymbol{z}) = \boldsymbol{z} - \boldsymbol{Q}\widehat{\boldsymbol{u}}$ and propose $\widehat{\boldsymbol{x}} = -\underline{\boldsymbol{x}} + \widehat{\boldsymbol{u}}$.
3) With probability $\min\left(1, e^{-\boldsymbol{r}(\boldsymbol{z})^{\mathrm{t}}(\underline{\boldsymbol{x}} - \widehat{\boldsymbol{x}})}\right)$, set $\overline{\boldsymbol{x}} = \widehat{\boldsymbol{x}}$, otherwise let $\overline{\boldsymbol{x}} = \underline{\boldsymbol{x}}$.

An important point concerns the initialization of the linear solver in Step 2: in the case of an early stopping, the computed approximate solution may depend on the initial point $\boldsymbol{u}_0$. On the other hand, $\boldsymbol{f}(\boldsymbol{z})$ must not depend on $\underline{\boldsymbol{x}}$, otherwise the reversibility of the deterministic move (5) would not be ensured. Hence, the initial point $\boldsymbol{u}_0$ must not depend on $\underline{\boldsymbol{x}}$ either. In the rest of the paper, $\boldsymbol{u}_0 = \boldsymbol{0}$ is the default choice.

A more compact and direct version of the sampler can be obtained by substituting $\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{z}) - \underline{\boldsymbol{x}}$ in equation (11). The latter reduces to the solving of the system $\boldsymbol{Q}\boldsymbol{x} = \boldsymbol{\eta}$. Step 2 of the RJPO algorithm is then simplified to:

2) Solve the linear system $\boldsymbol{Q}\boldsymbol{x} = \boldsymbol{\eta}$ in an approximate way. Let $\widehat{\boldsymbol{x}}$ denote the obtained solution and $\boldsymbol{r}(\boldsymbol{z}) = \boldsymbol{\eta} - \boldsymbol{Q}\widehat{\boldsymbol{x}}$.

For the reason just discussed above, the initial point $\boldsymbol{x}_0$ of the linear solver must be such that $\boldsymbol{u}_0 = \boldsymbol{x}_0 + \underline{\boldsymbol{x}}$ does not depend on $\underline{\boldsymbol{x}}$. Hence, as counterintuitive as it may be, choices such as $\boldsymbol{x}_0 = \boldsymbol{0}$ or $\boldsymbol{x}_0 = \underline{\boldsymbol{x}}$ are not allowed, while $\boldsymbol{x}_0 = -\underline{\boldsymbol{x}}$ is the default choice corresponding to $\boldsymbol{u}_0 = \boldsymbol{0}$.

It is remarkable that both T-PO and the proposed RJPO algorithm rely on the approximate resolution of the same linear system $\boldsymbol{Q}\boldsymbol{x} = \boldsymbol{\eta}$. However, RJPO algorithm incorporates two additional ingredients that make the difference in terms of mathematical validity:

- RJPO relies on an accept-reject strategy to ensure the sampler convergence in the case of an approximate system solving.
- There is a constraint on the initial point $\boldsymbol{x}_0$ of the linear solver: $\boldsymbol{x}_0 + \underline{\boldsymbol{x}}$ must not depend on $\underline{\boldsymbol{x}}$.

D. Implementation issues

Let us stress that there is no constraint on the choice of the linear solver, nor on its initialization and the early stopping rule, except that they must not depend on the value of $\underline{\boldsymbol{x}}$. Indeed, any linear system solver, or any quadratic programming method could be employed. In the sequel, we have adopted the linear conjugate gradient algorithm for two reasons:

- Early stopping (*i.e.*, *truncating*) the conjugate gradient iterations is a very usual procedure to approximately solve a linear system, with well-known convergence properties towards the exact solution [31]. Moreover, a preconditioned conjugate gradient could well be used to accelerate the convergence speed.
- It lends itself to a matrix-free implementation with reduced memory requirements, as far as matrix-vector products involving matrix $\boldsymbol{Q}$ can be performed without explictly manipulating such a matrix.

On the other hand, we have selected a usual stopping rule based on a threshold on the relative residual norm:

$$\epsilon = \frac{\|\boldsymbol{\eta} - \boldsymbol{Q}\boldsymbol{x}\|_2}{\|\boldsymbol{\eta}\|_2}. \tag{12}$$

IV. Practical Performance Analysis and Optimization

The aim of this section is to analyze the performance of the RJPO algorithm and to discuss the influence of the relative residual norm (and hence, the truncation level of the conjugate gradient algorithm) on the practical efficiency of both RJPO and T-PO schemes.

A. Performance Analysis

A Gaussian distribution with randomly generated positive definite precision matrix $\boldsymbol{Q}$ and mean vector $\boldsymbol{\mu}$ is considered. First, we focus on a small size problem ($N = 16$) to discuss the influence of the truncation level on the numerical performance in terms of acceptance rate and estimation error. For the retained Gaussian sampling schemes, both RJPO and T-PO are run for a number of CG iterations allowing to reach a predefined value of relative residual norm (12). We also discuss the influence of the problem dimension on the best value of the truncation level allowing to minimize the total number of CG iterations before convergence.

1) Acceptance rate: Figure 1 shows the average acceptance probability (acceptance rate) obtained over $n_{\max} = 10^5$ iterations of the sampler for different relative residual norm values.

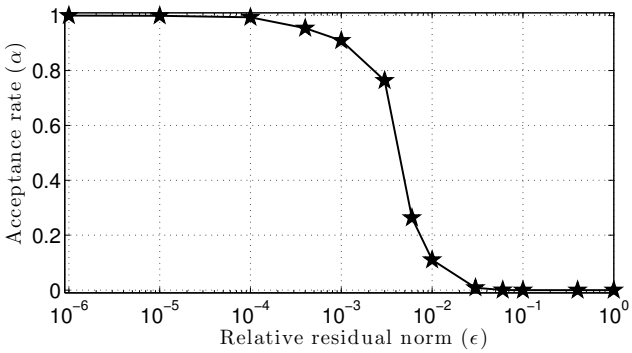


Fig. 1. Acceptance rate α of the RJPO algorithm for different values of the relative residual norm in the case of a small size problem ($N = 16$).

It can be noted that the acceptance rate is almost zero when the relative residual norm is larger than 10^{-2} and monotonically increases for higher resolution accuracies. Moreover, a

relative residual norm lower than 10^{-5} leads to an acceptance probability almost equal to one. Such a curve indicates that the stopping criterion of the CG must be chosen carefully in order to run the RJPO algorithm efficiently and to get non-zero acceptance probabilities. Finally, we emphasize that this curve mainly depends on the condition number of the precision matrix $\boldsymbol{Q}$. Even if the shape of the acceptance curve stays the same for different problems, it happens to be difficult to determine the value of the relative residual norm that corresponds to a given acceptance rate.

2) Estimation error: The estimation error is assessed as the relative mean square error (RMSE) on the estimated mean vector and covariance matrix

$$\begin{cases} \text{RMSE}(\boldsymbol{\mu}) = \dfrac{\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|_2}{\|\boldsymbol{\mu}\|_2}, \\ \text{RMSE}(\boldsymbol{R}) = \dfrac{\|\boldsymbol{R} - \hat{\boldsymbol{R}}\|_F}{\|\boldsymbol{R}\|_F}, \end{cases} \tag{13}$$

where $\|\cdot\|_F$ and $\|\cdot\|_2$ represent the Frobenius and the ℓ_2 norms, respectively. $\boldsymbol{\mu}$, $\boldsymbol{R}$, $\hat{\boldsymbol{\mu}}$, and $\hat{\boldsymbol{R}}$ are respectively the mean and the covariance matrix of the Gaussian vector, and their empirical estimates using the generated Markov chain samples:

$$\hat{\boldsymbol{\mu}} = \frac{1}{n_{\max} - n_{\min}} \sum_{n=n_{\min}+1}^{n_{\max}} \boldsymbol{x}_n,$$

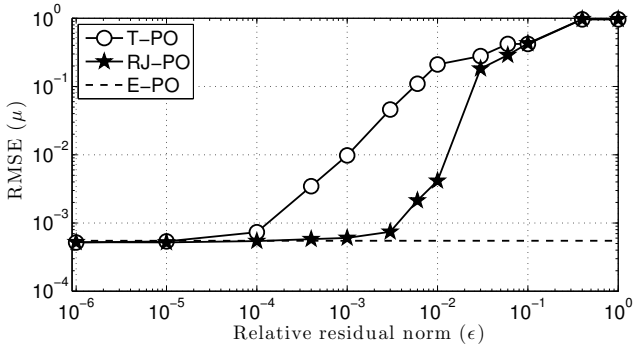
and

$$\hat{\boldsymbol{R}} = \frac{1}{n_{\max} - n_{\min} - 1} \sum_{n=n_{\min}+1}^{n_{\max}} (\boldsymbol{x}_n - \hat{\boldsymbol{\mu}})(\boldsymbol{x}_n - \hat{\boldsymbol{\mu}})^{\mathrm{t}},$$

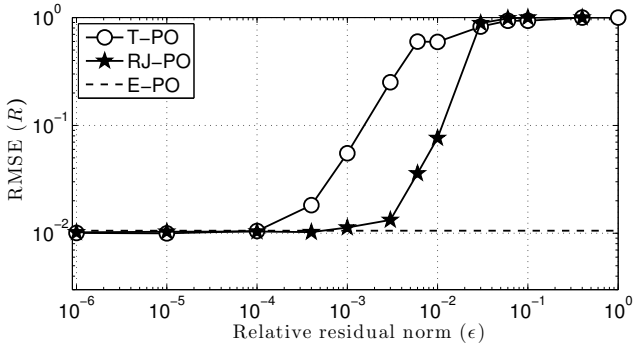
with $n_{\min}$ iterations of burn-in and $n_{\max}$ total iterations.

As expected, Figure 2 indicates that the estimation error is very high if the acceptance rate is zero (when the relative residual norm is lower than 10^{-2}), even for RJPO after $n_{\max} = 10^5$ iterations. This is due to the very low acceptance rate which slows down the chain convergence. However, as soon as new samples are accepted, RJPO leads to the same performance as when the system is solved exactly (E-PO algorithm). On the other hand, T-PO keeps a significant error for small and moderate resolution accuracies. Naturally, both methods present similar performance when the relative residual norm is sufficiently low since these methods tend to provide almost the same samples with an acceptance probability equal to one. This experimental result clearly highlights the deficiency of T-PO: the system must be solved with a relatively high accuracy to avoid an important estimation error. On the other hand, in the RJPO algorithm the acceptance rate is a good indicator whether the value of the relative residual norm threshold is appropriate to ensure a sufficient mixing of the chain.

3) Computation cost: Since the CG iterations correspond to the only burdensome task, the numerical complexity of the sampler can expressed in terms of the total number J_{tot} of CG iterations to be performed before convergence and the number of required samples to get efficient empirical approximation of the estimators. The Markov chain convergence is firstly assessed using the Gelman-Rubin criterion, which requires the running of several chains [32]. It consists in computing a scale reduction factor based on the between and within-chain variances. In this experiment, 100 parallel chains are considered. The results are summarized in Figure 3. It can be noted that a lower acceptance rate induces a higher number of iterations since the Markov chain converges more slowly towards its stationary distribution.



(a) Mean vector



(b) Covariance matrix

Fig. 2. Estimation error for different values of the truncation level after $n_{\max} = 10^5$ iterations of E-PO, T-PO and RJPO algorithms: (a) mean vector, (b) covariance matrix.

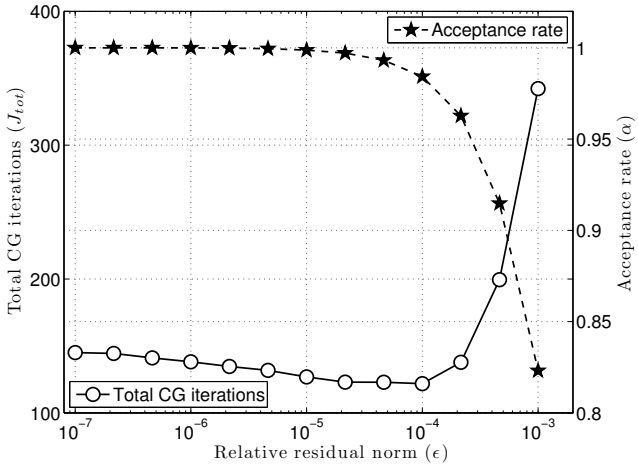


Fig. 3. Number of CG iterations before convergence and acceptance probability of the RJPO algorithm for different values of relative residual norm for a small size problem ($N = 16$).

One can also see that a minimal cost can be reached and, according to Figure 1, it corresponds to an acceptance rate of almost one. As the acceptance rate decreases, even a little, the computational cost rises very quickly. Conversely, if the relative residual is too small, the computation effort per

sample will decrease but additional sampling iterations will be needed before convergence, which naturally increases the overall computation cost. The latter result points out the need to appropriately choose the truncation level to jointly avoid a low acceptance probability and a high resolution accuracy of the linear system since both induce unnecessary additional computations.

4) Statistical efficiency: The performance of the RJPO sampler can also be analyzed using the effective sample size (ESS) [33, p. 125]. This indicator gives the number of independent samples, n_{eff}, that would yield the same statistical efficiency in approximating the Bayesian estimator as $n_{\max}$ successive samples of the simulated chain [34]. It is related to the chain autocorrelation function according to

$$n_{\text{eff}} = \frac{n_{\max}}{1 + 2\sum_{k=1}^{\infty} \rho_k} \tag{14}$$

where ρ_k the autocorrelation coefficient at lag k. In the Gaussian sampling context, such a relation allows to define how many iterations $n_{\max}$ are needed for each resolution accuracy in order to get estimators with comparable statistical performance (*e.g.*, using chains having the same effective sample size). Under the hypothesis of a first-order autoregressive chain, $\rho_k = \rho^k$, (14) leads to the ESS ratio

$$\text{ESSR} = \frac{n_{\text{eff}}}{n_{\max}} = \frac{1-\rho}{1+\rho}. \tag{15}$$

It can be noted that the ESSR is equal to one when the samples are independent $(\rho = 0)$ and decreases as the correlation between successive samples grows. In the RJPO case, we propose to define the *computing cost per effective sample* (CCES) as

$$\text{CCES} = \frac{J_{\text{tot}}}{n_{\text{eff}}} = \frac{J}{\text{ESSR}} \tag{16}$$

where $J = J_{\text{tot}}/n_{\max}$ is the average number of CG iterations per sample. Figure 4 shows the ESSR and the CCES in the case of a Gaussian vector of dimension $N = 16$. It can be seen that an early stopped CG algorithm induces a very small ESSR, due to a large sample correlation value, and thus a high effective cost to produce accurate estimates. On the contrary, a very precise resolution of the linear system induces a larger number of CG iterations per sample but a shorter Markov chain since the ESSR is almost equal to 1. The best trade-off is produced by intermediate values of the relative residual norm threshold around $\epsilon = 2 \cdot 10^{-4}$.

To conclude, the Gelman-Rubin convergence diagnostic and the ESS approach both confirm that the computation cost of the RJPO can be reduced by appropriately truncating the CG iterations. Although the Gelman-Rubin convergence test is probably more accurate, since it is based on several independent sequences, the CCES based test is far simpler and provides nearly the same trade-off in the tested example. Such encouraging results have motivated us to develop a self-tuning strategy to automatically adjust the threshold parameter ϵ by tracking the minimizer of the CCES. Our proposed strategy is presented in Subsection IV-B.

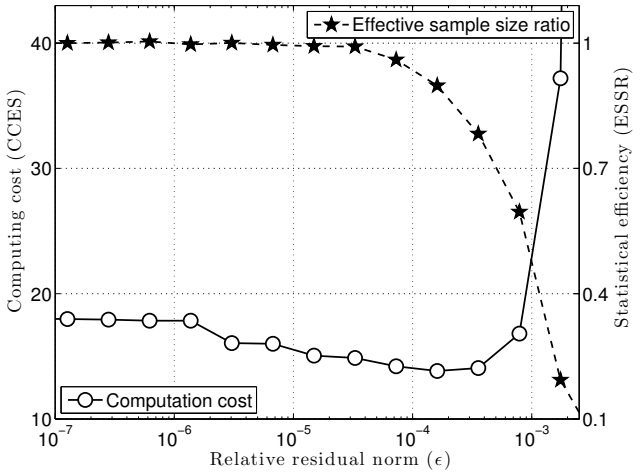


Fig. 4. Computing cost per effective sample of the RJPO algorithm for different relative residual norm values on a small size problem $(N = 16)$ estimated from $n_{\max} = 10^4$ samples.

5) Influence of the dimension: Figure 5 summarizes the optimal values of the truncation level ϵ that allows to minimize the CCES for different values of N. The best trade-off is reached for decreasing values of ϵ as N grows. More generally, the same observation can be made as the problem conditioning deteriorates. In practice, predicting the appropriate truncation level for a given problem is difficult. Fortunately, Figure 5 also indicates that the optimal setting is obtained for an acceptance probability that remains almost constant. The best trade-off is clearly obtained for an acceptance rate α lower than one ($\alpha = 1$ corresponds to $\epsilon = 0$, *i.e.*, to the exact solving of $\boldsymbol{Qx} = \boldsymbol{\eta}$). In the tested example, the optimal truncation level ϵ rather corresponds to an acceptance rate around 0.99. However, finding an explicit mathematical correspondence between ϵ and α is not a simple task. In the next subsection, we propose an unsupervised tuning strategy of the relative residual norm allowing either to achieve a predefined target acceptance rate, or even to directly optimize the computing cost per effective sample.

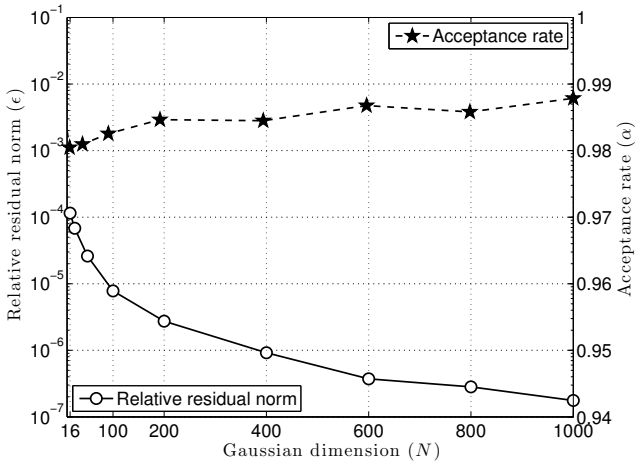


Fig. 5. Influence of the problem dimension on the optimal values of the relative residual norm and the acceptance rate.

B. Adaptive Tuning of the Resolution Accuracy

The suited value of the relative residual norm ϵ to achieve a desired acceptance rate α_t can be adjusted recursively using a Robbins-Monro type algorithm [35, 36]. Such an adaptive scheme is formulated in the stochastic approximation framework [37] in order to solve a non-linear equation of the form $g(\theta) = 0$ using an update

$$\theta_{n+1} = \theta_n + K_n \left[g(\theta_n) + \nu_n \right] \tag{17}$$

where ν is a random variable traducing the uncertainty on each evaluation of function $g(\cdot)$ and $\{K_n\}$ is a sequence of step-sizes ensuring stability and convergence [38]. Such a procedure has been widely used for the optimal scaling of adaptive MCMC algorithms [27, 28] since it does not alter the chain convergence towards the target distribution. For instance, it is used in [39, 40] to set adaptively the scale parameters of a Random walk Metropolis-Hastings (RWMH) algorithm in order to reach the optimal acceptance rate suggested by theoretical or empirical analysis [41, 42]. The same procedure was also used by [43] for the adaptive tuning of a Metropolis-adjusted Langevin algorithm (MALA) to reach the optimal acceptance rate proposed by [44].

1) Tuning to achieve a target acceptance rate: In order to ensure the positivity of the relative residual norm ϵ, the update is performed on its logarithm. At each iteration n of the sampler, the relative residual norm is adjusted according to

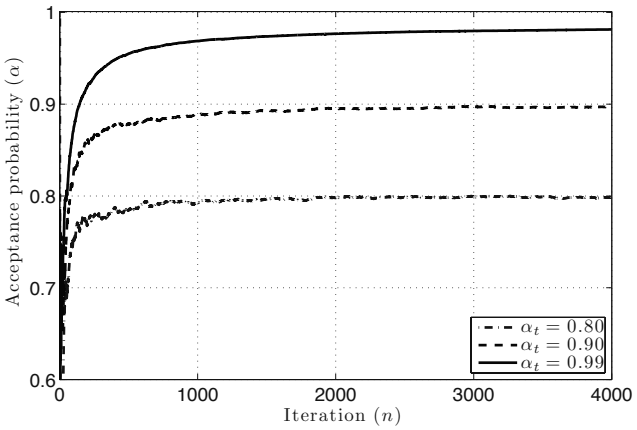
$$\log \epsilon_{n+1} = \log \epsilon_n + K_n \left[\alpha(\underline{\boldsymbol{x}}_n, \boldsymbol{x}_n) - \alpha_t \right]. \tag{18}$$

where α_t is a given target acceptance probability and $\{K_n\}$ is a sequence of step-sizes decaying to 0 as n grows in order to ensure the convergence of the Markov chain to the target distribution. As suggested in [28], the step-sizes are chosen according to $K_n = K_0/n^\kappa$, with $\kappa \in]0, 1]$. We emphasize that more sophisticated methods, such as those proposed in [36] could be used to approximate the acceptance rate curve and to derive a more efficient adaptive strategy for choosing this parameter.

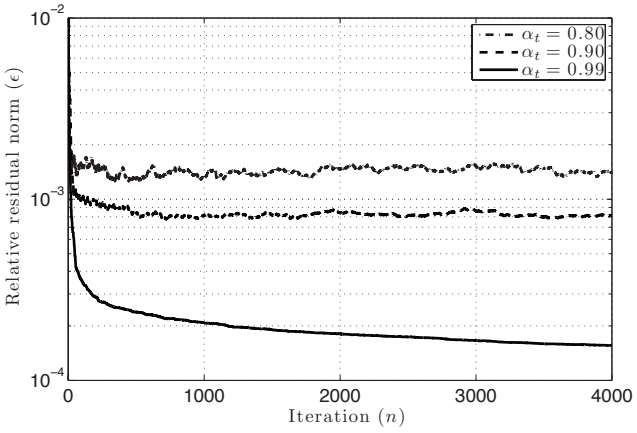
The adaptive RJPO is applied to the sampling of the previously described Gaussian distribution using the adopted step-size with parameters $K_0 = 1$ and $\kappa = 0.5$. Figure 6 presents the evolution of the average acceptance probability and the obtained relative residual norm for three different values of the target acceptance rate α_t. One can note that the average acceptance rate converges to the desired value. Moreover, the relative residual norm also converges to the expected value according to Figure 1 (for example, the necessary relative residual norm to get an acceptance probability $\alpha_t = 0.8$ is equal to $1.5 \cdot 10^{-3}$).

In practice, it remains difficult to *a priori* determine which acceptance rate should be targeted to achieve the faster convergence. The next subsection proposes to modify the target of the adaptive strategy to directly minimize the CCES.

2) Tuning to optimize the numerical efficiency: A given threshold ϵ on the relative residual norm induces an average truncation level J and an ESSR value, from which the CCES can be deduced according to (16). Our goal is to adaptively adjust the threshold value ϵ in order to minimize the CCES.



(a) Acceptance probability



(b) Relative residual norm

Fig. 6. Behavior of the adaptive RJPO for 1000 iterations and three values of the target acceptance probability: (a) Evolution of the average acceptance probability and (b) Evolution of the computed relative residual norm.

Let J_{opt} be the average number of CG iterations per sample corresponding to the optimal threshold value. In the plane (J, ESSR), it is easy to see that J_{opt} is the abscissa of the point at which the tangent of the ESSR curve intercepts the origin (see Figure 7).

The ESSR is expressed by (15) as a function of the chain correlation ρ, the latter being an implicit function of the acceptance rate α. For $\alpha = 1$, $\rho = 0$ according to Proposition 2. For $\alpha = 0$, $\rho = 1$ since no new sample can be accepted. For intermediate values of α, the correlation lies between 0 and 1, and it is typically decreasing. It can be decomposed on two terms:

- With a probability $1 - \alpha$, the accept-reject procedure produces identical (*i.e.*, maximally correlated) samples in case of rejection.
- In case of acceptance, the new sample is slightly correlated with the previous one, because of the early stopping of the CG algorithm.

While it is easy to express the correlation induced by rejection, it is difficult to find an explicit expression for the correlation between accepted samples. However, we have checked that the latter source of correlation is negligible compared to the

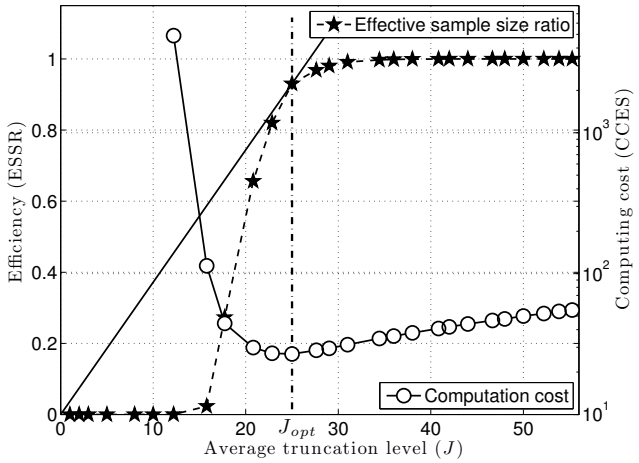


Fig. 7. Influence of the CG truncation level on the overall computation cost and the statistical efficiency of the RJPO for sampling a Gaussian of dimension $N = 128$.

correlation induced by rejection. If we approximately assume that accepted samples are independent, we get $\rho = 1 - \alpha$. Consequently, the best tuning of the relative residual norm leading to the lowest computation cost per effective sample is the minimizer of $\dfrac{(2-\alpha)}{\alpha} J$. By necessary condition, we get

$$J \frac{d\alpha}{dJ} - \alpha + \frac{\alpha^2}{2} = 0.$$

Finally, a similar procedure as in the previous section is applied to adaptively adjust the optimal value of the relative residual norm according to

$$\log \epsilon_{n+1} = \log \epsilon_n + K_n \left(J_n \frac{d\alpha_n}{dJ} - \alpha_n + \frac{\alpha_n^2}{2} \right), \qquad (19)$$

where $\dfrac{d\alpha_n}{dJ}$ is evaluated numerically. Figure 8 illustrates that the proposed adaptive scheme efficiently adjusts ϵ to minimize the CCES where J_{opt} is around 26 , which is in agreement with Figure 7.

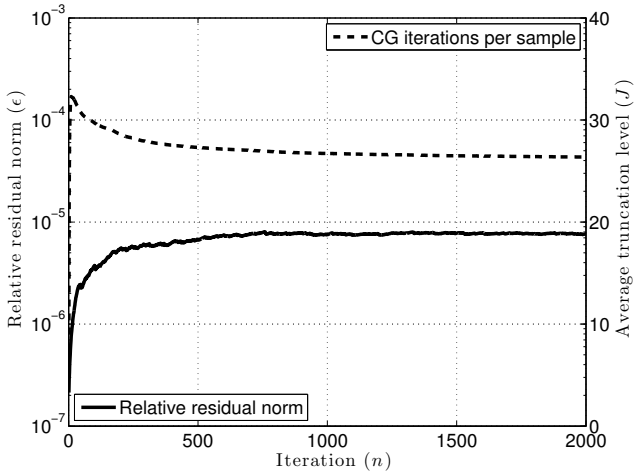


Fig. 8. Evolution of the relative residual norm and the acceptance rate for a Gaussian sampling problem of size $N = 128$. The adaptive algorithm leads to a relative residual norm $\epsilon_{\text{opt}} = 7.79 \cdot 10^{-6}$ leading to $\alpha_{\text{opt}} = 0.977$.

V. Application to Unsupervised Super-Resolution

In the linear inverse problem of unsupervised image super-resolution, several images are observed with a low spatial resolution. In addition, the measurement process presents a point spread function (PSF) that introduces a blur on the images. The purpose is then to reconstruct the original image with a higher resolution using an unsupervised method. Such an approach allows to also estimate the model hyper-parameters and the PSF [23, 45, 46]. In order to discuss the relevance of the previously presented Gaussian sampling algorithms we apply a Bayesian approach and MCMC methods for solving this inverse problem.

A. Problem Statement

The observation model is given by $\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n}$, where $\boldsymbol{H} = \boldsymbol{P}\boldsymbol{F}$, with $\boldsymbol{y} \in \mathbb{R}^M$ the vector containing the pixels of the observed images in a lexicographic order, $\boldsymbol{x} \in \mathbb{R}^N$ the sought high resolution image, $\boldsymbol{F}$ the $N \times N$ circulant convolution matrix associated with the blur, $\boldsymbol{P}$ the $M \times N$ decimation matrix and $\boldsymbol{n}$ the additive noise. This linear model also includes classical image deconvolution problems [1, 2]. The noise is assumed to follow a zero-mean Gaussian distribution with an unknown precision matrix $\boldsymbol{Q}_y = \gamma \boldsymbol{I}$. We also assume a zero-mean Gaussian distribution for the prior of the sought variable $\boldsymbol{x}$, with a precision matrix $\boldsymbol{Q}_x = \delta \boldsymbol{D}^{\mathrm{t}}\boldsymbol{D}$, where $\boldsymbol{D}$ is the circulant convolution matrix associated to a Laplacian filter. Non-informative Jeffrey's priors [47] are also assigned to the two hyper-parameters γ and δ.

According to Bayes' theorem, the posterior distribution is given by

$$P(\boldsymbol{x}, \gamma, \delta | \boldsymbol{y}) \propto \delta^{(N-1)/2-1} \gamma^{M/2-1} \times e^{-\frac{\gamma}{2}(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x})^{\mathrm{t}}(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x}) - \frac{\delta}{2}\boldsymbol{x}^{\mathrm{t}}\boldsymbol{D}^{\mathrm{t}}\boldsymbol{D}\boldsymbol{x}}$$

To explore this posterior distribution, a Gibbs sampler iteratively draws

1) γ_n from $P(\gamma | \boldsymbol{x}_{n-1}, \boldsymbol{y})$ given as

$$\mathcal{G}\left(1 + \frac{M}{2}, 2||\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}_{n-1}||^{-2}\right),$$

2) δ_n from $P(\delta | \boldsymbol{x}_{n-1})$ given as

$$\mathcal{G}\left(1 + \frac{N-1}{2}, 2||\boldsymbol{D}\boldsymbol{x}_{n-1}||^{-2}\right)$$

3) $\boldsymbol{x}_n$ from $P(\boldsymbol{x} | \delta_n, \gamma_n, \boldsymbol{y})$ which is

$$\mathcal{N}\left(\boldsymbol{\mu}_n, [\boldsymbol{Q}_n]^{-1}\right)$$

with

$$\boldsymbol{Q}_n = \gamma_n \boldsymbol{H}^{\mathrm{t}}\boldsymbol{H} + \delta_n \boldsymbol{D}^{\mathrm{t}}\boldsymbol{D}$$
$$\boldsymbol{Q}_n \boldsymbol{\mu}_n = \gamma_n \boldsymbol{H}^{\mathrm{t}}\boldsymbol{y}$$

The third step of the sampler requires an efficient sampling of a multivariate Gaussian distribution whose parameters change along the sampling iterations. In the sequel, direct sampling with Cholesky factorization [15] is firstly employed as a reference method. It yields the same results as the E-PO

algorithm. For the inexact resolution case, the T-PO algorithm using a CG controlled by the relative residual norm, and the adaptive RJPO directly tuned with the acceptance probability are performed. For these two methods, the product matrix-vector used in the CG algorithm is done by exploiting the structure of the precision matrix $\boldsymbol{Q}$ and thus only implies circulant convolutions, performed by FFT, and decimations.

B. MCMC Results

We consider the observation of five images of dimension 128×128 pixels ($M = 81920$) and we reconstruct the original one of dimension 256×256 ($N = 65536$). The convolution part $\boldsymbol{F}$ has a Laplace shape with of full width at half maximum (FWHM) of 4 pixels. A white Gaussian noise is added to get a signal-to-noise ratio (SNR) equal to 20dB. The original image and one of the observations are shown in Figure 9.

Fig. 9. Unsupervised super-resolution - image reconstruction using the adaptive RJPO algorithm with $\alpha_t = 0.99$.

The Gibbs sampler is run for 1000 iterations and a burn-in period of 100 iterations is considered after a visual inspection of the chains. The performances are evaluated in terms of the mean and standard deviation of both hyper-parameters γ, δ and one randomly chosen pixel x_i of the reconstructed image. Table I presents the mean and standard deviation of the variable of interest. As we can see, the T-PO algorithm is totally inappropriate even with a precision of 10^{-8}. Conversely, the estimation from the samples given by the adaptive RJPO and Cholesky method are very similar, which demonstrates the correct behavior of the proposed algorithm.

		γ	$\delta \times 10^{-4}$	x_i
Cholesky		102.1 (0.56)	6.1 (0.07)	104.6 (9.06)
T-PO	$\epsilon = 10^{-4}$	0.3 (0.06)	45 (0.87)	102.2 (3.30)
T-PO	$\epsilon = 10^{-8}$	71.7 (0.68)	21 (0.29)	102.7 (2.51)
A-RJPO,	$\alpha_t = 0.99$	101.2 (0.55)	6.1 (0.07)	101.9 (8.89)

TABLE I
COMPARISON BETWEEN DIFFERENT GAUSSIAN SAMPLING STRATEGIES: THE CHOLESKY FACTORIZATION BASED APPROACH, THE T-PO CONTROLLED BY THE RELATIVE RESIDUAL NORM AND THE ADAPTIVE RJPO TUNED BY THE ACCEPTANCE RATE. THE PERFORMANCES ARE EXPRESSED IN TERMS OF EMPIRICAL MEAN AND STANDARD DEVIATION OF HYPER-PARAMETERS AND ONE RANDOMLY CHOSEN PIXEL.

Figure 10 shows the evolution of the acceptance rate with respect to the number of CG iterations. We can notice that at least 400 iterations are required to have a nonzero acceptance probability. Moreover, more than 800 iterations seems unnecessary. For this specific problem, the E-PO algorithm needs theoretically $N = 65536$ iterations to have a new sample while the adaptive RJPO only requires around 700. Concerning the computation time, on a Intel Core i7-3770 with 8GB of RAM and a 64bit system, it took about 20.3s on average and about 6GB of RAM for the Cholesky sampler to generate one sample and only 15.1s and less than 200MB for the RJPO. This last result is due to the use of a conjugate gradient on which each matrix-vector product is performed without explicitly writing the matrix $\boldsymbol{Q}$. Finally, note that if we consider images of higher resolution, for instance $N = 1024 \times 1024$, the Cholesky factorization would require around 1TB of RAM and the adaptive RJPO only about 3GB (when using double precision floating-point format).

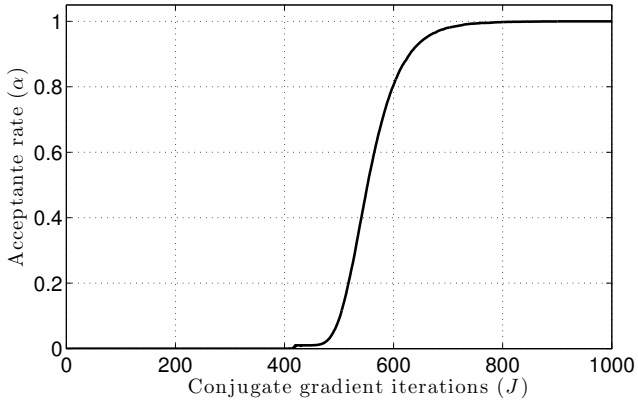


Fig. 10. Evolution of the acceptance rate with respect to average conjugate gradient iterations for sampling a Gaussian of dimension $N = 65536$.

VI. CONCLUSION

The sampling of high dimensional Gaussian distributions appears in the resolution of many linear inverse problems using MCMC methods. Alternative solutions to the Cholesky factorization are needed to reduce the computation time and to limit the memory usage. Based on the theory of reversible jump MCMC, we derived a sampling method allowing to introduce an approximate solution of a linear system during the sample generation step. The approximate resolution of a linear system was already adopted in methods like IFP and PO to reduce the numerical complexity, but without any guarantee of convergence to the target distribution. The proposed algorithm RJPO is based on an accept-reject step that is absent from the existing PO algorithms. Indeed, the difference between RJPO and existing PO algorithms is much comparable to the difference between the Metropolis-adjusted Langevin algorithm (MALA) [48] and a plainly discretized Langevin diffusion [49].

Our results pointed out that the required resolution accuracy in these methods must be carefully tuned to prevent a significant error. It was also shown that the proposed RJ-MCMC framework allows to ensure the convergence through the accept-reject step whatever the truncation level. In addition, thanks to the simplicity of the acceptance probability, the resolution accuracy can be adjusted automatically using an adaptive scheme allowing to achieve a pre-defined acceptation rate. We have also proposed a significant improvement of the same adaptive tuning approach, where the target is directly

formulated in terms of minimal computing cost per effective sample.

Finally, the linear system resolution using the conjugate gradient algorithm offers the possibility to implement the matrix-vector products with a limited memory usage by exploiting the structure of the forward model operators. The adaptive RJPO has thus proven to be less consuming in both computational cost and memory usage than any approach based on Cholesky factorization.

This work opens some perspectives in several directions. Firstly, preconditioned conjugate gradient or alternative methods can be envisaged for the linear system resolution with the aim to reduce the computation time per iteration. Such an approach will highly depend on the linear operator and the ability to compute a preconditioning matrix. A second direction concerns the connection between the RJ-MCMC framework and other sampling methods such as those based on Krylov subspace [19, 20], particularly with appropriate choices of the parameters $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{b}$ and $\boldsymbol{f}(\cdot)$ defined in section II. Another perspective of this work is to analyze more complex situations involving non-gaussian distributions with the aim to be able to formulate the perturbation step and to perform an approximate optimization allowing to reduce the computation cost. Finally, the proposed adaptive tuning scheme allowing to optimize the computation cost per effective sample could be generalized to other Metropolis adjusted sampling strategies.

APPENDIX

A. Expression of the acceptance probability

According to the RJ-MCMC theory, the acceptance probability is given by

$$\alpha(\underline{\boldsymbol{x}}, \boldsymbol{x}|\boldsymbol{z}) = \min\left(1, \frac{P_{\boldsymbol{X}}(\boldsymbol{x})P_{\boldsymbol{Z}}(\boldsymbol{s}|\boldsymbol{x})}{P_{\boldsymbol{X}}(\underline{\boldsymbol{x}})P_{\boldsymbol{Z}}(\boldsymbol{z}|\underline{\boldsymbol{x}})}|J_{\boldsymbol{\phi}}(\underline{\boldsymbol{x}}, \boldsymbol{z})|\right),$$

with $\boldsymbol{s} = \boldsymbol{z}$ and $\boldsymbol{x} = -\underline{\boldsymbol{x}} + \boldsymbol{f}(\boldsymbol{z})$. The Jacobian determinant of the deterministic move is $|J_{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{z})| = 1$. Since

$$P_{\boldsymbol{X}}(\boldsymbol{x}) \propto e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{t}}\boldsymbol{Q}(\boldsymbol{x}-\boldsymbol{\mu})},$$

and

$$P_{\boldsymbol{Z}}(\boldsymbol{z}|\underline{\boldsymbol{x}}) \propto e^{-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{A}\underline{\boldsymbol{x}}-\boldsymbol{b})^{\mathrm{t}}\boldsymbol{B}^{-1}(\boldsymbol{z}-\boldsymbol{A}\underline{\boldsymbol{x}}-\boldsymbol{b})},$$

the acceptance probability can be written as

$$\alpha(\underline{\boldsymbol{x}}, \boldsymbol{x}|\boldsymbol{z}) = \min\left(1, e^{-\frac{1}{2}\Delta S}\right)$$

with $\Delta S = \Delta S_1 + \Delta S_2$ and

$$\begin{aligned}
\Delta S_1 &= \boldsymbol{x}^{\mathrm{t}}\boldsymbol{Q}\boldsymbol{x} - 2\boldsymbol{x}^{\mathrm{t}}\boldsymbol{Q}\boldsymbol{\mu} - \underline{\boldsymbol{x}}^{\mathrm{t}}\boldsymbol{Q}\underline{\boldsymbol{x}} + 2\underline{\boldsymbol{x}}^{\mathrm{t}}\boldsymbol{Q}\boldsymbol{\mu},\\
\Delta S_2 &= -(\boldsymbol{z} - \boldsymbol{A}\underline{\boldsymbol{x}} - \boldsymbol{b})^{\mathrm{t}}\boldsymbol{B}^{-1}(\boldsymbol{z} - \boldsymbol{A}\underline{\boldsymbol{x}} - \boldsymbol{b}),\\
&= \boldsymbol{x}^{\mathrm{t}}\boldsymbol{A}^{\mathrm{t}}\boldsymbol{B}^{-1}\boldsymbol{A}\boldsymbol{x} - 2\boldsymbol{x}^{\mathrm{t}}\boldsymbol{A}^{\mathrm{t}}\boldsymbol{B}^{-1}(\boldsymbol{z} - \boldsymbol{b})\\
&\quad - \underline{\boldsymbol{x}}^{\mathrm{t}}\boldsymbol{A}^{\mathrm{t}}\boldsymbol{B}^{-1}\boldsymbol{A}\underline{\boldsymbol{x}} + 2\underline{\boldsymbol{x}}^{\mathrm{t}}\boldsymbol{A}^{\mathrm{t}}\boldsymbol{B}^{-1}(\boldsymbol{z} - \boldsymbol{b}).
\end{aligned}$$

Since $\boldsymbol{x} = -\underline{\boldsymbol{x}} + \boldsymbol{f}(\boldsymbol{z})$, we get

$$\begin{aligned}
\Delta S_1 &= (\boldsymbol{x} - \underline{\boldsymbol{x}})^{\mathrm{t}}\boldsymbol{Q}\left(\boldsymbol{f}(\boldsymbol{z}) - 2\boldsymbol{\mu}\right)\\
\Delta S_2 &= (\boldsymbol{x} - \underline{\boldsymbol{x}})^{\mathrm{t}}\left(\boldsymbol{A}^{\mathrm{t}}\boldsymbol{B}^{-1}\boldsymbol{A}\,\boldsymbol{f}(\boldsymbol{z}) - 2\boldsymbol{A}^{\mathrm{t}}\boldsymbol{B}^{-1}(\boldsymbol{z} - \boldsymbol{b})\right)
\end{aligned}$$

Finally

$$\begin{aligned}
\Delta S &= (\boldsymbol{x} - \underline{\boldsymbol{x}})^{\mathrm{t}}\\
&\quad \left[\left(\boldsymbol{Q} + \boldsymbol{A}^{\mathrm{t}}\boldsymbol{B}^{-1}\boldsymbol{A}\right)\boldsymbol{f}(\boldsymbol{z}) - 2\left(\boldsymbol{Q}\boldsymbol{\mu} + \boldsymbol{A}^{\mathrm{t}}\boldsymbol{B}^{-1}\left(\boldsymbol{z} - \boldsymbol{b}\right)\right)\right]\\
&= 2\,\left(\underline{\boldsymbol{x}} - \boldsymbol{x}\right)^{\mathrm{t}}\boldsymbol{r}(\boldsymbol{z}).
\end{aligned}$$

Finally, when the system is solved exactly, $\Delta S = 0$ and thus $\alpha(\underline{\boldsymbol{x}}, \boldsymbol{x}|\boldsymbol{z}) = 1$.

B. Correlation between two successive samples

Since

$$\overline{\boldsymbol{x}} = -\underline{\boldsymbol{x}} + 2\left(\boldsymbol{Q} + \boldsymbol{A}^{\mathrm{t}}\boldsymbol{B}^{-1}\boldsymbol{A}\right)^{-1}\left(\boldsymbol{Q}\boldsymbol{\mu} + \boldsymbol{A}^{\mathrm{t}}\boldsymbol{B}^{-1}(\boldsymbol{z} - \boldsymbol{b})\right)$$

and $\boldsymbol{z}$ is sampled from $\mathcal{N}\left(\boldsymbol{A}\underline{\boldsymbol{x}} + \boldsymbol{b}, \boldsymbol{B}\right)$, we have

$$\begin{aligned}
\overline{\boldsymbol{x}} = -\underline{\boldsymbol{x}} &+ 2\left(\boldsymbol{Q} + \boldsymbol{A}^{\mathrm{t}}\boldsymbol{B}^{-1}\boldsymbol{A}\right)^{-1}\\
&\left(\boldsymbol{Q}\boldsymbol{\mu} + \boldsymbol{A}^{\mathrm{t}}\boldsymbol{B}^{-1}\boldsymbol{A}\underline{\boldsymbol{x}} + \boldsymbol{A}^{\mathrm{t}}\boldsymbol{B}^{-1}\boldsymbol{\omega}_{\boldsymbol{B}}\right)
\end{aligned}$$

with $\boldsymbol{\omega}_{\boldsymbol{B}}$ totally independent of $\underline{\boldsymbol{x}}$. One can firstly check that $\mathbb{E}\left[\underline{\boldsymbol{x}}\right] = \mathbb{E}\left[\overline{\boldsymbol{x}}\right] = \boldsymbol{\mu}$. Consequently, the correlation between two successive samples is given by

$$\begin{aligned}
&\mathbb{E}\left[(\overline{\boldsymbol{x}} - \boldsymbol{\mu})(\underline{\boldsymbol{x}} - \boldsymbol{\mu})^{\mathrm{t}}\right] =\\
&\qquad \left(2\left(\boldsymbol{Q} + \boldsymbol{A}^{\mathrm{t}}\boldsymbol{B}^{-1}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^{\mathrm{t}}\boldsymbol{B}^{-1}\boldsymbol{A} - \boldsymbol{I}\right)\boldsymbol{Q}^{-1}
\end{aligned}$$

which is zero if and only if $\boldsymbol{A}^{\mathrm{t}}\boldsymbol{B}^{-1}\boldsymbol{A} = \boldsymbol{Q}$.

REFERENCES

[1] M. Bertero and P. Boccacci, *Introduction to Inverse Problems in Imaging*. London, UK: IOP Publishing Ltd., 1998.

[2] J. Idier, *Bayesian approach to Inverse problems*. ISTE Ltd and John Wiley & Sons Inc, 2008.

[3] B. Frieden, "Image enhancement and restoration," in *Picture Processing and Digital Filtering*, ser. Topics in Applied Physics. New York, NY, USA: Springer-Verlag, 1975, vol. 6, pp. 177–248.

[4] G. Demoment, "Image reconstruction and restoration: Overview of common estimation structure and problems," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 37, no. 12, pp. 2024–2036, dec 1989.

[5] G. T. Gordon, R.and Herman, "Reconstruction of pictures from their projections," *Communications of the ACM*, vol. 14, no. 12, pp. 759–768, 1971.

[6] R. M. Lewitt and S. Matej, "Overview of methods for image reconstruction from projections in emission computed tomography," *Proceedings of the IEE*, vol. 91, no. 1, pp. 1588–1611, 2003.

[7] W. Gilks, S. Richardson, and D. Spiegehalter, *Markov Chain Monte Carlo in Practice*. London, UK: Chapman & Hall, 1999.

[8] C. Robert, *The Bayesian Choice*, 2nd ed. Springer-Verlag, 2001.

[9] D. Andrews and C. Mallows, "Scale mixtures of normal distributions," *J. R. Statist. Soc. B*, pp. 99–102, 1974.

[10] F. Champagnat and J. Idier, "A connection between half-quadratic criteria and EM algorithms," *IEEE Signal Processing Lett.*, vol. 11, no. 9, pp. 709–712, 2004.

[11] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-6, no. 6, pp. 721–741, Nov. 1984.

[12] G. Papandreou and A. Yuille, "Gaussian sampling by local perturbations," in *Proceedings of NIPS*, 2010.

[13] E. M. Scheuer and D. S. Stoller, "On the generation of normal random vectors," *Technometrics*, vol. 4, no. 2, pp. 278–281, 1962.

[14] D. R. Barr and N. L. Slezak, "A comparison of multivariate normal generators," *Commun. ACM*, vol. 15, no. 12, pp. 1048–1049, Dec. 1972.

[15] H. Rue, "Fast sampling of Gaussian Markov random fields," *J. R. Statist. Soc. B*, vol. 63, no. 2, pp. 325–338, 2001.

[16] W. F. Trench, "An algorithm for the inversion of finite Toeplitz matrices," *J. Soc. Indust. Appl. Math.*, vol. 12, no. 3, pp. 515–522, 1964.

[17] D. Geman and C. Yang, "Nonlinear image recovery with half-quadratic regularization," *IEEE Trans. Image Processing*, vol. 4, no. 7, pp. 932–946, 1995.

[18] P. Lalanne, D. Prévost, and P. Chavel, "Stochastic artificial retinas: algorithm, optoelectronic circuits, and implementation," *Applied Optics*, vol. 40, no. 23, pp. 3861–3876, 2001.
[19] A. Parker and C. Fox, "Sampling Gaussian distributions in Krylov spaces with conjugate gradients," *SIAM J. Sci. Comput.*, vol. 34, no. 3, pp. B312–B334, 2012.
[20] E. Aune, J. Eidsvik, and Y. Pokern, "Iterative numerical methods for sampling from high dimensional Gaussian distributions," *Statist. and Comp.*, pp. 1–21, 2013.
[21] E. Chow and Y. Saad, "Preconditioned Krylov subspace methods for sampling multivariate Gaussian distributions," *SIAM*, vol. 36, no. 2, pp. A588–A608, 2014.
[22] Y. Amit and U. Grenander, "Comparing sweep strategies for stochastic relaxation," *Journal of Multivariate Analysis*, vol. 37, no. 2, pp. 197 – 222, 1991.
[23] F. Orieux, O. Féron, and J. Giovannelli, "Sampling high-dimensional Gaussian distributions for general linear inverse problems," *IEEE Signal Processing Lett.*, vol. 19, no. 5, p. 251, 2012.
[24] X. Tan, J. Li, and P. Stoica, "Efficient sparse Bayesian learning via Gibbs sampling," in *Proceedings of ICASSP*, 2010, pp. 3634–3637.
[25] P. J. Green, "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, vol. 82, pp. 711–732, 1995.
[26] R. Waagepetersen and D. Sorensen, "A tutorial on reversible jump MCMC with a view toward applications in QTL-mapping," *Int. Statist. Rev.*, vol. 69, no. 1, pp. 49–61, 2001.
[27] C. Andrieu and C. Robert, "Controlled MCMC for optimal sampling," 2001, tech. Report No. 0125, Cahiers de Mathématiques du Ceremade, Université Paris-Dauphine.
[28] C. Andrieu and J. Thoms, "A tutorial on adaptive MCMC," *Statist. and Comp.*, vol. 18, no. 4, pp. 343–373, 2008.
[29] Y. Atchade, G. Fort, E. Moulines, and P. Priouret, "Adaptive Markov chain Monte Carlo: theory and methods," in *Bayesian Time Series Models*. Cambridge Univ. Press., 2011, pp. 33–53.
[30] P. De Forcrand, "Monte Carlo quasi-heatbath by approximate inversion," *Phys. Rev. D*, vol. 59, no. 3, pp. 3698–3701, 1999.
[31] R. Barrett, M. Berry, T. F. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, and H. van der Vorst, *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*. SIAM, 1994.
[32] A. Gelman and D. B. Rubin, "Inference from iterative simulation using multiple sequences," *Statist. and Comp.*, pp. 457–472, 1992.
[33] J. S. Liu, *Monte Carlo Strategies in Scientific Computing*, 2nd ed., ser. Springer Series in Statistics. Springer, 2008.
[34] J. Goodman and A. D. Sokal, "Multigrid Monte Carlo method. Conceptual foundations," *Physical Review D*, vol. 40, no. 6, 1989.
[35] H. Robbins and S. Monro, "A stochastic approximation method," *Annals of Mathematical Statistics*, vol. 22, pp. 400–407, 1951.
[36] B. Bercu and P. Fraysse, "A Robbins–Monro procedure for estimation in semiparametric regression models," *Annals Statist.*, vol. 40, no. 2, pp. 666–693, 2012.
[37] A. Benveniste, M. Métivier, and P. Priouret, *Adaptive algorithms and stochastic approximations*. Springer, 2012.
[38] C. Andrieu, E. Moulines, and P. Priouret, "Stability of stochastic approximation under verifiable conditions," *SIAM J. Control Optimization*, vol. 44, no. 1, pp. 283–312, 2006.
[39] H. Haario, E. Saksman, and J. Tamminen, "An adaptive Metropolis algorithm," *Bernouilli*, vol. 7, no. 2, pp. 223–242, 2001.
[40] Y. F. Atchadé and J. S. Rosenthal, "On adaptive Markov chain Monte Carlo algorithms," *Bernouilli*, vol. 11, no. 5, pp. 815–828, 2005.
[41] G. O. Roberts, A. Gelman, and W. R. Gilks, "Weak convergence and optimal scaling of random walk Metropolis algorithms," *The Annals of Applied Probability*, vol. 7, no. 1, pp. 110–120, 1997.
[42] A. Gelman, G. O. Roberts, and W. R. Gilks, "Efficient Metropolis jumping rules," in *Bayesian statistics 5*. Oxford University Press, 1996, pp. 599–607.
[43] Y. F. Achadé, "An adaptive version for the metropolis adjusted Langevin algorithm with a truncated drift," *Methodology and Computing in Applied Probability*, vol. 8, no. 2, pp. 235–254, 2006.
[44] G. O. Roberts and J. S. Rosenthal, "Optimal scaling of discrete approximations to Langevin's diffusions," *J. R. Statist. Soc. B*, vol. 60, no. 1, pp. 255–268, 1998.
[45] S. C. Park, M. K. Park, and M. G. Kang, "Super-resolution image reconstruction: a technical overview," *IEEE Trans. Signal Processing*, vol. 20, no. 3, pp. 21–36, 2003.
[46] G. Rochefort, F. Champagnat, G. Le Besnerais, and J.-F. Giovannelli, "An improved observation model for super-resolution under affine motion," *IEEE Trans. Image Processing*, vol. 15, no. 11, pp. 3325–3337, 2006.
[47] H. Jeffreys, "An invariant form for the prior probability in estimation problems," *Proc. of the Royal Society of London. Series A. Mathematical and Physical Sciences*, vol. 186, no. 1007, pp. 453–461, 1946.
[48] G. O. Roberts and R. L. Tweedie, "Exponential convergence of Langevin distributions and their discrete approximations," *Bernoulli*, vol. 2, no. 4, pp. 341–363, 1996.
[49] G. Parisi, "Correlation functions and computer simulations," *Nuclear Physics B*, vol. 180, pp. 378–384, 1981.

Clément Gilavert received the engineering degree from École Centrale, Nantes, France, in 2011.
From october 2011 to march 2014, he was a Ph.D. student with the Institut de Recherche en Communications et Cybernétique, Nantes (IRCCYN, UMR CNRS 6597).
His Ph.D. thesis work is related to statistical inference methods for large-scale inverse problems in signal and image processing.

Saïd Moussaoui received the State engineering degree from Ecole Nationale Polytechnique, Algiers, Algeria, in 2001, and the Ph.D. degree in automatic control and signal processing from Université Henri Poincaré, Nancy, France, in 2005.
He is currently Assistant Professor with École Centrale de Nantes, Nantes, France. Since September 2006, he has been with the Institut de Recherche en Communications et Cybernétique, Nantes (IRCCYN, UMR CNRS 6597).
His research interests are in lstatistical signal and image processing methods including source separation, image restoration, and their applications to spectrometry and hyperspectral imaging.

Jérôme Idier was born in France in 1966. He received the Diploma degree in electrical engineering from École Supérieure d'Électricité, Gif-sur-Yvette, France, in 1988, and the Ph.D. degree in physics from University of Paris-Sud, Orsay, France, in 1991. In 1991, he joined the Centre National de la Recherche Scientifique.
He is currently a Senior Researcher with the Institut de Recherche en Communications et Cybernétique, Nantes, France.
His major scientific interests are in probabilistic approaches to inverse problems for signal and image processing. Dr. Idier is currently elected member of the French National Committee for Scientific Research.

A.6 Synthesis and application of nonlinear observers for the estimation of tire effective radius and rolling resistance of an automotive vehicle

C. El Tannoury, **S. Moussaoui**, F. Plestan, N. Romani et G. Pita Gil, *IEEE Trans. on Control Systems Technology*, vol. 21, no. 6, pp. 2408-2416, 2013.

Ce papier est dédié à la présentation d'un méthode d'estimation récursive de la résistance au roulement d'un véhicule automobile. Le but étant de se servir de cette estimation pour détecter une éventuelle baisse de la pression du pneumatique.

Synthesis and Application of Nonlinear Observers for the Estimation of Tire Effective Radius and Rolling Resistance of an Automotive Vehicle

Charbel El Tannoury, Saïd Moussaoui, Franck Plestan, Nicolas Romani, and Guillermo Pita-Gil

***Abstract*— The rolling resistance and the effective radius of a vehicle's tires are two important characteristics that affect its dynamics, performance, and comfort. Because of their dependence on tire inflation pressure, online estimation of such parameters could be used to monitor tire pressures using an indirect approach. By considering rotational and longitudinal dynamics, the aim of this paper proposes to apply observers for this online estimation using measurements of the wheels' angular velocities and the engine torque. Because these signals are available on major vehicle controller area networks, the proposed solutions do not require additional sensors. These nonlinear observers are based, first, on a high-gain approach, and then on a high-order sliding-mode approach, allowing robustness and finite time convergence. The originality of the presented results consists in providing a joint estimation of both variables, i.e., wheel effective radius and rolling resistance force. The observers offer the very first solution for dynamical estimation of rolling resistance in standard driving conditions, but the rolling resistance is very difficult to estimate by an online procedure. Simulations and experimental results allow the discussion of the effect of tire pressure on these parameters and illustrate the applicability of the proposed approach.**

***Index Terms*— Effective radius, high-gain observer, high-order sliding-mode observer, rolling resistance, tire pressure.**

NOMENCLATURE

SYMBOL	DEFINITION
Ω	Wheel angular velocity.
v_x	Vehicle's longitudinal speed.
d_c	Vertical displacement of the car body.
d_r	Vertical displacement of the wheel.
d_{pro}	Road profile.
Γ	Traction torque.
F_x	Tractive force.
F_d	Aerodynamic drag force.
F_r	Rolling resistance force.
J	Wheel inertia.
C_f	Wheel viscous friction coefficient.
R	Wheel effective radius.
R_0	Wheel nominal radius.
λ	Slip ratio.
μ	Tire–road friction coefficient.

Manuscript received December 21, 2011; revised November 6, 2012; accepted November 23, 2012. Manuscript received in final form December 4, 2012. Date of publication January 17, 2013; date of current version October 15, 2013. Recommended by Associate Editor A. Behal.

Ch. El Tannoury is with IRCCyN (CNRS UMR 6597), LUNAM Université, Ecole Centrale Nantes, 44321 Nantes Cedex 03, France. He is also with Renault Corporation, Guyancourt 78288, France (e-mail: Charbel.El-Tannoury@irccyn.ec-nantes.fr).

S. Moussaoui and F. Plestan are with IRCCyN (CNRS UMR 6597), LUNAM Université, Ecole Centrale Nantes, 44321 Nantes Cedex 03, France (e-mail: Said.Moussaoui@irccyn.ec-nantes.fr; Franck.Plestan@irccyn.ec-nantes.fr).

N. Romani and G. Pita-Gil are with Renault Corporation, Guyancourt 78288, France (e-mail: Nicolas.Romani@renault.com; guillermo.pita-gil@renault.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCST.2012.2232669

I. INTRODUCTION

TIRE pressure monitoring constitutes one of the most important challenges in improving modern vehicle safety [1], [2]. Indeed, a low tire pressure greatly influences vehicle behavior and fuel consumption [3], [4] and increases the risk of accident [5]. A first solution to tire pressure monitoring is based on direct measurement of pressure using dedicated sensors (see [6] for a bibliographical review). However, the use of such sensors (located in the tire valve) presents some drawbacks, such as possible failures, the need for specialized tires, and additional costs which car manufacturers wish to avoid. There is therefore a real interest in removing these sensors and finding indirect solutions of tire deflation detection [7].

Tire rolling resistance and effective radius[1] are two characteristics depending on several physical parameters, among which is the tire inflation pressure [8, Ch. 3]. Experimental studies [9], [10] have shown that the lack of adequate pressure will simultaneously cause a decrease of the effective radius and an increase of the rolling resistance. Thus, these parameters are possible indicators for tire pressure monitoring.

However, to our knowledge, there is no method that allows the estimation of the rolling resistance during the vehicle's motion. The aim of this brief paper is therefore to propose online estimation methods of the rolling resistance and of the effective radius of the wheels; the online estimation of such variables is a novelty, as are the application conditions. In fact, the latter are standard, i.e., the vehicle on which the observers are evaluated has no specific measurement equipment, and the driving conditions are such that there are time-varying velocity, change of load mass, etc. The estimation results could then be used by a tire pressure monitoring system (TPMS). However, the development of such a TPMS is not addressed in this brief paper.

An estimate of the effective radius can be obtained from the vehicle speed, the wheel angular velocity, and the slip ratio [11], [12]. However, in the absence of a vehicle speed sensor such as a GPS (global positioning system), this direct calculation will be inaccurate. In practice, the vehicle's longitudinal velocity is derived from the angular velocity of the nondriven wheels and their *a priori* known nominal radii, by supposing a negligible slip ratio on these wheels.

Several authors have proposed online estimation schemes of effective radius, in addition to parameters such as longitudinal

[1]The effective radius of the wheel is obtained by dividing the vehicle velocity by the angular velocity of the wheel.

stiffness, tire–road friction, and wheel slip using nonlinear observation methods [13]–[15], and have proposed the use of this estimation for inflation pressure diagnosis [16]–[18]. The underlying models consider that the rolling resistance can be measured *a priori* under normal driving conditions and then linked to the vehicle velocity and load using empirical models. However, according to [8], [19], the rolling resistance force is highly dependent on tire parameters such as inflation pressure, temperature, road surface type, and vehicle speed. Consequently, such approaches cannot be used in real driving conditions.

In [17], a strategy using second-order sliding-mode observers is proposed for the joint estimation of the effective radius and inflation pressure of a tire. However, even though this method does not require the rolling resistance force, it assumes an empirical relation between the tire's vertical deformation and the inflation pressure. Concerning the rolling resistance force, to our knowledge, prior works on its estimation were based only on experimental tests performed under particular driving conditions [20]–[23] and coast-down tests [24].

By considering rotational and longitudinal dynamics models, nonlinear (high-gain and variable-structure) observers are designed for the online estimation of the tire's effective radius and rolling resistance by using wheel angular velocities and engine torque. These latter signals being available on the controller area networks (CANs), the proposed solution does not require additional sensors other than standard equipment on a vehicle. The observers developed in the sequel are based on the following.

1) A high-gain approach [25]. The advantage of this well-known nonlinear approach is its implementation simplicity. Furthermore, it has been already used in many applications.
2) The high-order sliding mode (HOSM) approach [26]–[28]. HOSM observers are applicable to a very large class of observable systems, allow robustness, and ensure finite time convergence. Moreover, such observers have already shown their efficiency in automotive applications [29].

The rest of this brief paper is organized as follows. In Section II, reference physical models of vertical and longitudinal dynamics of a wheel are used to discuss the influence of inadequate tire inflation pressure on three parameters: 1) effective radius, 2) rolling resistance force, and 3) vertical stiffness coefficient. The results of this section motivate rolling resistance estimation, since this parameter is the most sensitive to tire inflation pressure. The observer design strategies in the case of a single wheel monitoring are detailed in Section III. The application of these observers to simulated data and to real measurements are detailed in Section IV. The extension to the front axle is also proposed (Section V). Finally, conclusions and future perspectives are given in Section VI.

II. Inflation Pressure Influence on Tire Dynamics

A complete model accounting for vertical and longitudinal tire dynamics is first derived. This model will serve as a reference simulator allowing us to discuss the effect of tire inflation pressure variations on three parameters, namely, rolling resistance, effective radius, and vertical stiffness. It will be also used in order to validate by simulation the observers and to derive a simplified model for the observer design.

A. Quarter-Car Model

The suspended part of the vehicle's body can be represented as a mass m_c, whereas the wheel and the unsuspended mechanical part are modeled by a mass m_r. The vehicle's suspension is modeled by a spring K_s and a damper C_s. A spring K_v and a damper C_v are also used to model the vertical behavior of the tire. The additional parameters used for the modeling are given in Nomenclature.

Consider a vehicle moving on a plane road during an acceleration phase, and assume negligible lateral motion. The wheel dynamics can be represented by combining all the vertical and longitudinal forces [30]. The main forces that are acting on each wheel are the following.

1) Normal Force F_z: This force depends mainly on the mass $M = m_c + m_r$ of the quarter-car, but also on the vertical displacement of the tire-road contact point

$$F_z = Mg - K_v(d_r - d_{\text{pro}}) - C_v(\dot{d}_r - \dot{d}_{\text{pro}}) \tag{1}$$

with g the gravitational acceleration.

2) Traction Force F_x: This force results from tire–road interaction due to the applied wheel torque. This force is characterized by the friction coefficient μ and the slip ratio λ. The slip ratio is defined as

$$\lambda = \frac{R\,\Omega - v_x}{R\Omega} = 1 - \frac{v_x}{R\Omega}. \tag{2}$$

The friction coefficient μ is theoretically given by semiempirical formulas [11]. An acceptable approximation [30], [31] is expressed as a function of λ by

$$\mu(\lambda) = 2\,\mu_0\,\frac{\lambda_0\lambda}{\lambda_0^2 + \lambda^2} \tag{3}$$

where λ_0 is the optimal slip ratio, leading to the maximum friction value μ_0. The tractive force is then given by

$$F_x = \mu(\lambda)\,F_z. \tag{4}$$

3) Aerodynamic Drag Force F_d: This force is proportional to the square of the vehicle's velocity [19]

$$F_d = \frac{1}{2}\,\rho\,A_d\,C_d\,v_x^2 \tag{5}$$

with ρ the air density, A_d the quarter-car frontal area, and C_d the aerodynamic drag coefficient.

4) Rolling Resistance Force F_r: This force results from tire–road contact and deformation [8]. The rolling resistance force is applied from the wheel–road contact point and is directed toward the center of the wheel. Therefore, its moment is zero [32, pp. 41–43]. However, its longitudinal component only appears in the longitudinal dynamics and is proportional to the normal force F_z [32, pp. 41–43]

$$F_r = C_r\,F_z \tag{6}$$

with C_r the rolling resistance coefficient. This coefficient depends mainly on the tire inflation pressure, temperature, velocity, and road surface type [19, p. 110].

According to Newton's laws, the resulting model reads

$$\begin{aligned}
\dot{x}_1 &= \frac{1}{J}\left[\Gamma - RF_x - C_f x_1\right] \\
\dot{x}_2 &= \frac{1}{m_c + m_r}\left[F_x - F_d - F_r\right] \\
\dot{x}_3 &= x_4 \\
\dot{x}_4 &= \frac{1}{m_c}\left[-K_s(x_3 - x_5) - C_s(x_4 - x_6)\right] \\
\dot{x}_5 &= x_6, \\
\dot{x}_6 &= \frac{1}{m_r}\big[K_s(x_3 - x_5) + C_s(x_4 - x_6) \\
&\qquad - K_v(x_5 - d_{\text{pro}}) - C_v(x_6 - \dot{d}_{\text{pro}})\big] \qquad (7)
\end{aligned}$$

with $[x_1\ x_2\ x_2\ x_4\ x_5\ x_6]^T = [\Omega\ v_x\ d_c\ \dot{d}_c\ d_r\ \dot{d}_r]^T$. The first two equations of this model are related to the longitudinal and rotational dynamics of the wheel, whereas the last four are concerned with its vertical dynamics [8], [11], [19].

B. Analysis of Inflation Pressure Effect on Tire Parameters

This section discusses how the rolling resistance force, the effective radius, and the tire vertical stiffness are affected by pressure variation. The idea consists in finding the parameters that are more sensitive to inflation pressure. Once these sensitive variables are identified, it will be necessary to check the feasibility of their estimation by using a solution based on observer design. Finally, it will be possible to conclude on the variables that will be retained for inflation pressure diagnosis.

1) Complete Model Simulation: The complete model (7) is used for the simulation where lookup tables, provided by Renault co., are used to set the values of the quarter-car model parameters which depend on inflation pressure (rolling resistance, vertical stiffness, and effective radius). The chosen values for the pressure-independent parameters are summarized in Table I. The road profile is assumed a band-limited random noise with a maximum amplitude of 1 mm with zero average. A controller based on input–output linearization and linear state feedback [33] is used to adjust the wheel torque in order to simulate the vehicle going ahead at a constant longitudinal velocity.

2) Sensitivity Analysis: Table II summarizes the effect of the pressure deflation on the relative variation of R, F_r, and K_v. It appears clearly that the most sensitive parameter is the rolling resistance force. Therefore, the detection of a pressure deflation can be achieved through the estimation of this force.

TABLE I
WHEEL AND CAR MODEL PARAMETERS USED [19], [30]

Parameter	Symbol	Value	Unit
Wheel inertia	J	1.6	$\text{kg} \cdot \text{m}^2$
Nominal radius	R_0	0.32	m
Quarter-car mass	M	440	kg
Frontal area	A_d	0.325	m^2
Air density	ρ	1.205	$\text{kg} \cdot \text{m}^{-3}$
Gravitational constant	g	9.807	$\text{m} \cdot \text{s}^{-2}$
Viscous coefficient	C_f	0.08	$\text{kg} \cdot \text{m}^2 \cdot \text{s}^{-1}$
Drag coefficient	C_d	0.25	No unit
Suspension damping	C_s		$\text{kg} \cdot \text{s}^{-1}$
Suspension stiffness	K_s		$\text{kg} \cdot \text{s}^{-2}$
Peak friction	μ_0	0.9	No unit
Optimal slip	λ_0	0.25	No unit

TABLE II
RELATIVE VARIATIONS OF THE EFFECTIVE RADIUS R, THE ROLLING RESISTANCE F_r, AND THE VERTICAL STIFFNESS K_v FOR TIRE PRESSURE FALL EQUAL TO 20%, AND FOR SEVERAL VEHICLE VELOCITIES v_x

v_x (kmh)	50	70	90	110
R (%)	−0.26	−0.26	−0.3	−0.32
F_r (%)	+21	+24.2	+26.7	+28.6
K_v (%)	−5.3	−5.2	−5.6	−6

III. ROLLING RESISTANCE AND EFFECTIVE RADIUS ESTIMATION USING NONLINEAR OBSERVERS

In addition to the estimation of the rolling resistance F_r, the estimation of the effective radius R is also included since it does not strongly increase the observation task when the observation model is based only on longitudinal dynamics. Furthermore, the estimation of R is interesting for the localization of the deflated wheel when several wheels are considered (see Section V). Since the variations of R and F_r are unknown, their dynamics read as

$$\dot{R} = \eta_1(t), \quad \dot{F}_r = \eta_2(t) \qquad (8)$$

with η_1 and η_2 bounded unknown functions.

A. Reduced Model

The main assumption leading to the complete model simplification is due to the expression of the normal force F_z. One can assume

$$K_v(x_5 - d_{\text{pro}}) + C_v(x_6 - \dot{d}_{\text{pro}}) \ll Mg$$

which gives $F_z = Mg$. In this case, vertical dynamics are not used in the simplified model, which is

$$\begin{aligned}
J\,\dot{x}_1 &= \Gamma - RF_x(v_x, R, \Omega) - C_f x_1 \\
M\,\dot{x}_2 &= F_x(v_x, R, \Omega) - F_d(v_x) - F_r. \qquad (9)
\end{aligned}$$

B. Observation Model

By denoting the state vector $x = [x_1\ x_2\ x_7\ x_8]^T = [\Omega\ v_x\ R\ F_r]^T$, $u = \Gamma$ the control input, and the measured output $y = [y_1\ y_2]^T = [\Omega\ v_x]^T = [x_1\ x_2]^T$, the following state system is obtained from (9):

$$\dot{x} = f(x) + \Delta f + \chi(y, u) \qquad (10)$$

with

$$f(x) = \begin{bmatrix} -\frac{1}{J}x_7 F_x(x) \\ \frac{1}{M}\Big(F_x(x) - F_d(x) - x_8\Big) \\ 0 \\ 0 \end{bmatrix}, \Delta f = \begin{bmatrix} 0, \\ 0, \\ \eta_1, \\ \eta_2 \end{bmatrix}$$

and

$$\chi(y,u) = \left[-\frac{C_f}{J}x_1 + \frac{1}{J}u,\ 0,\ 0,\ 0\right]^T. \tag{11}$$

H1: The uncertainty term Δf does not change the observability of (10). ∎

Consider now the nonlinear system (10) without any uncertainty ($\Delta f = 0$)

$$\dot{x} = f(x) + \chi(y,u). \tag{12}$$

The term $\chi(y,u)$ only depends on well-known (measured) variables. As shown in [34] and [35], this term is not required for the observer design, and can be removed thanks to an input–output injection function $-\chi(y,u)$.

H2: The input–output injection function $-\chi(y,u)$ does not change the observability feature. ∎

It yields that the system (12) is transformed into (with abuse of notation)

$$\dot{x} = f(x), \quad y = [x_1\ x_2]^T. \tag{13}$$

C. Observability Analysis

Consider the integer pair $(k_1, k_2) = (2,2)$, and define the following function:[2]

$$\Psi(x) = \begin{bmatrix} y_1 \\ y_1^{(k_1-1)} \\ y_2 \\ y_2^{(k_2-1)} \end{bmatrix} = \begin{bmatrix} y_1 \\ \dot{y}_1 \\ y_2 \\ \dot{y}_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ \dot{x}_1 \\ x_2 \\ \dot{x}_2. \end{bmatrix}. \tag{14}$$

Definition 1 [36]: Denote $\mathcal{M}_x \subset \mathbb{R}^n$ the operating physical domain in which x is evolving. The system (13) is locally observable (i.e., observable $\forall x \in \mathcal{M}_x$) if $\Psi(x)$ is a state coordinates transformation, i.e., $\zeta = \Psi(x)$ is invertible $\forall x \in \mathcal{M}_x$. In this case, the integers (k_1, k_2) are called observability indices. ∎

Given the complexity of Ψ, it is difficult (even with formal computation software) to analytically establish its invertibility. Thus, this latter will be numerically evaluated. If its Jacobian never equals 0 on the operating trajectories, it yields that the transformation Ψ is invertible; then, system (13) is locally observable. Furthermore, under assumptions $H1$ and $H2$, one can conclude that the system (10) with both measured variables x_1 and x_2 is observable for the operating conditions.

[2]Let $y_i^{(k_i-1)}$ ($i \in \{1,2\}$) denote the (k_i-1)th time derivative of the function $y_i(x)$.

D. Observer Design

Given that $\Psi(x)$ is invertible under the proposed operating conditions, it defines a state coordinates transformation $\zeta = \Psi(x)$. Then, it is trivial to show that the nonlinear system (13) is locally equivalent to

$$\dot{\zeta} = \underbrace{\begin{bmatrix} 0\ 1\ 0\ 0 \\ 0\ 0\ 0\ 0 \\ 0\ 0\ 0\ 1 \\ 0\ 0\ 0\ 0 \end{bmatrix}}_{A} \zeta + \underbrace{\begin{bmatrix} 0 \\ \Phi_1(\zeta) \\ 0 \\ \Phi_2(\zeta) \end{bmatrix}}_{\Phi(\zeta)}. \tag{15}$$

Given the presence of unknown dynamics as η_1 and η_2, the vector $\Phi(x)$ can be decomposed into a nominal part Φ_n (composed of known parameters and dynamics), and an uncertain part $\Delta\Phi$, which gives

$$\dot{\zeta} = A\zeta + \Phi_n(\zeta) + \Delta\Phi. \tag{16}$$

Following these notations, one has $\Phi_1(\zeta) = \Phi_{1n} + \Delta\Phi_1$ and $\Phi_2(\zeta) = \Phi_{2n} + \Delta\Phi_2$.

Proposition 1: An observer for system (16) reads as

$$\dot{\hat{\zeta}} = A\hat{\zeta} + \Phi_n(\hat{\zeta}) + \kappa(y,\hat{\zeta}) \tag{17}$$

with $\hat{\zeta}$ the estimated state of ζ and the function $\kappa(y,\hat{\zeta})$ called "correction term" and forcing $\hat{\zeta} \to \zeta$. ∎

It is obvious that the correction term $\kappa(y,\hat{\zeta})$ is not unique and can be obtained by several different methods depending on the desired features (robustness, finite time convergence, etc.,). Given that estimation error dynamics reads as (with $e = \hat{\zeta} - \zeta$)

$$\dot{e} = Ae + \Phi_n(\hat{\zeta}) - \Phi_n(\zeta) - \Delta\Phi + \kappa(y,\hat{\zeta}) \tag{18}$$

$\kappa(y,\hat{\zeta})$ has to make the observer converging (exponentially or in a finite time) to the real system in spite of the initial error $e(0)$ and the uncertain term $\Delta\Phi$. From $\hat{\zeta} = \Psi(\hat{x})$, one gets

$$\dot{\hat{\zeta}} = \frac{\partial\Psi}{\partial\hat{x}}\dot{\hat{x}} \to \dot{\hat{x}} = \left[\frac{\partial\Psi}{\partial\hat{x}}\right]^{-1}\dot{\hat{\zeta}}. \tag{19}$$

Then, in a similar way as in [37], an observer for system (13) reads as

$$\dot{\hat{x}} = f(\hat{x}) + \left[\frac{\partial\Psi}{\partial x}\right]^{-1}\kappa(y,\hat{x}). \tag{20}$$

The application of the inverse input–output injection transformation $\chi(y,u)$ allows us to get an observer for system (12)

$$\dot{\hat{x}} = f(\hat{x}) + \chi(y,u) + \left[\frac{\partial\Psi}{\partial\hat{x}}\right]^{-1}\kappa(y,\hat{x}). \tag{21}$$

The function $\kappa(y,\hat{x})$ has to be designed such that the state vector $\hat{x}$ of the previous system is reaching the vicinity of state vector of system (10) in finite time in spite of the uncertain dynamics Δf of R and F_r.

High-Gain Observer [25]: The observer (21) for the system (10) admits a correction term $\kappa(y,\hat{x})$ defined as

$$\kappa(y,\hat{x}) = \Lambda^{-1}K(y - C\hat{x}) \tag{22}$$

with

$$K = \begin{bmatrix} K_1 & 0 \\ K_2 & 0 \\ 0 & K_1 \\ 0 & K_2 \end{bmatrix}$$

so that $A - KC$ is Hurwitz

$$\Lambda(T) = \begin{bmatrix} \tau_1 & 0 & 0 & 0 \\ 0 & \tau_1^2 & 0 & 0 \\ 0 & 0 & \tau_2 & 0 \\ 0 & 0 & 0 & \tau_2^2 \end{bmatrix}$$

with τ_1 and τ_2 strictly positive. ∎

Second-Order Sliding-Mode Observer [27]: A solution for $\kappa(y, \hat{x})$ is proposed in order to obtain an accurate and robust estimation of $\hat{x}$; it is based on high-order sliding-mode differentiation. Consider the system (17) and suppose that

H3: For $\zeta \in \mathcal{M}_\zeta$ ($\mathcal{M}_\zeta$ being the operating domain in ζ-state space)

$$|\Phi_{1n}(\zeta)| \leq L_{\Phi_1}, \quad |\Phi_{2n}(\zeta)| \leq L_{\Phi_2}$$

with L_{Φ_1} and L_{Φ_2} being known Lipschitz positive constants. Furthermore

$$|\Delta\Phi_1| \leq L_{\Delta\Phi_1}, \quad |\Delta\Phi_2| \leq L_{\Delta\Phi_2}$$

with $0 < L_{\Delta\Phi_1} < \infty$ and $0 < L_{\Delta\Phi_2} < \infty$. ∎

An observer based on high-order sliding mode [27], [38] for system (16) reads as (with $\zeta = [\zeta_1, \ \zeta_2, \ \zeta_3, \ \zeta_4]^T$)

$$\begin{aligned} \dot{\hat{\zeta}}_1 &= \hat{\zeta}_2 + \underbrace{a_1 \ L_{\Phi_1}^{\frac{1}{2}} | \ \zeta_1 - \hat{\zeta}_1|^{\frac{1}{2}} \text{sign}(\zeta_1 - \hat{\zeta}_1)}_{\gamma_1} \\ \dot{\hat{\zeta}}_2 &= \Phi_{n1}(\hat{\zeta}) + a_2 \ L_{\Phi_1} \text{sign}(\gamma_1) \\ \dot{\hat{\zeta}}_3 &= \hat{\zeta}_4 + \underbrace{a_3 \ L_{\Phi_2}^{\frac{1}{2}} \ | \ \zeta_3 - \hat{\zeta}_3|^{\frac{1}{2}} \text{sign}(\zeta_3 - \hat{\zeta}_3)}_{\gamma_3} \\ \dot{\hat{\zeta}}_4 &= \Phi_{n2}(\hat{\zeta}) + a_4 \ L_{\Phi_2} \text{sign}(\gamma_3). \end{aligned} \tag{23}$$

Coefficients a_1, a_2, a_3, a_4 must be fixed as proposed in [27]

$$a_1 = a_3 = 1.5, \quad a_2 = a_4 = 1.1.$$

The finite time convergence of the estimation $\hat{\zeta} - \zeta$ to a vicinity of 0 (see Theorem 6 in [27]) can be proved by rewriting (23) using the differential inclusion understood in the Filippov sense [39]. Then, the observer (21) for the system (10) admits a correction term $\kappa(y, \hat{x})$ defined as (by replacing in γ_1 and γ_3, ζ_1 by x_1, $\hat{\zeta}_1$ by $\hat{x}_1$, ζ_3 by x_2, and $\hat{\zeta}_3$ by $\hat{x}_2$)

$$\kappa(y, \hat{x}) = \begin{bmatrix} \gamma_1 & a_2 L_{\Phi_1} \text{sign}(\gamma_1) & \gamma_3 & a_4 L_{\Phi_2} \text{sign}(\gamma_3) \end{bmatrix}^T.$$

IV. Simulation and Experimental Results

A. Description of Experimental Setup and Experimentation Scenario

All experiments were carried out on a Renault Laguna II vehicle in order to validate the algorithms. An AutoBox was also used: it communicates with vehicle CAN bus and, by wired connections, with the sensors. This equipped vehicle allows recordings for offline analysis, as well as developing, implementing, and testing many designed observation/control strategies. To summarize, the equipment consists of the following:

Hardware Tools:

1) a dSpace AutoBox with PHS bus and PCMCIA host interface *DS*815;
2) a dSpace *DS*4302 CAN interface board;
3) a dSpace *DS*4002 timing and digital I/O board for the acquisition of the ABS signals;
4) a PC for development and supervision.

Software Tools:

1) MATLAB *R*2006*b*/simulink;
2) dSpace real time interface;
3) control desk standard developer v6.0.

The vehicle CAN operates at a rate of 500 kb/s and the ABS signal acquisition period is 1 ms. The vehicle is equipped with Dunlop tires of 195 mm width and 127 mm height. The inner diameter equals 38.1 cm. The experimental data were collected during tests made on roads in Spain and on an experimental driving circuit at the Renault Technical Center at Aubevoye (France). Simulations were performed in order to check the observability feature and to evaluate observers on the complete model (7). However, its application to real vehicle data requires the knowledge of the wheel torque, the wheel angular velocity, and the vehicle speed. In that respect as follows.

1) The wheel angular velocities are given by the ABS sensors.
2) The vehicle velocity is derived from the average velocity of the rear wheels by assuming a negligible slip ratio on these wheels.[3] In fact, in most situations, there appears a smaller slip ratio in the rear wheels as compared to the front wheels.
3) The engine torque being accessible on the CAN bus, the wheel torque is deduced by knowing the gear ratio and transmission factor. The torque is assumed to be the same on both left and right wheels.

Note that all these assumptions introduce some uncertainties. In order to counter the effect of uncertainties/modeling approximations, the observer (especially the sliding-mode one) has to be sufficiently robust and tuned in order to get sufficient accuracy.[4]

Two observation strategies are evaluated in the experiments:

1) *Single-Wheel Observer:* The proposed observer is applied separately to the left and right front wheels. It will lead to a degraded solution because it does not take into account the coupling between the left and the right wheel dynamics. However, its advantage is to estimate the rolling resistance of each wheel separately.
2) *Axle Observer:* This observer (Section V) is used to jointly estimate the front axle wheels' radii and the rolling resistance force of the whole axle. It therefore takes into account the coupling between the both wheels, but only gives a global rolling resistance (not for each wheel).

[3] A more accurate solution would be to use a dedicated velocity sensor; however, the objective of this brief paper is to use standard equipment on automotive vehicles.

[4] It is clear that it would be necessary to increase the number of scenarios in order to find optimal tunings (or a "map" of gains) for the observers. It will be done in future works, the purpose of this brief paper being to show the feasibility of the approach.

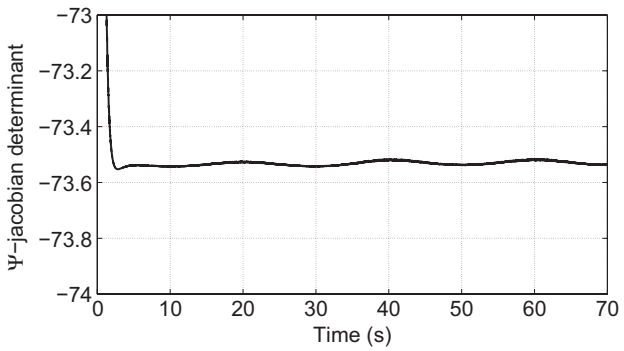


Fig. 1. Determinant of the Jacobian of matrix Ψ versus time (s).

B. Application to a Simulated Quarter-Car model

The complete model (7) is used for the simulation according to the procedure described in Section II-B. The simulation results are displayed in Fig. 2. The proposed observers (21) are initialized by

$$\left[\hat{\Omega}(0)\ \hat{v}_x(0)\ \hat{R}(0)\ \hat{F}_r(0)\right] = [15/0.31\ 15\ 0.305\ 74] \quad (24)$$

whereas the initial conditions of the complete model (7) are such that $\Omega(0) = 15/0.3$ rad/s, $v_x(0) = 15$ m/s, $d_c(0) = d_r(0) = 0$ m, and $\dot{d}_c(0) = \dot{d}_r(0) = 0$ m/s. The vehicle is going ahead at a slow time-varying velocity

$$v_x^d = v^d(1 + \delta_v \sin(\omega t))$$

with $v^d = 40$ kmh, omega $= 0.314$ rad/s, and $\delta_v = 0.01$. A fall of pneumatic pressure of 20% to its nominal value 2.5 bar between the instants $t_1 = 30$ s and $t_2 = 40$ s is simulated. A zero mean additive random noise has been added to the measurements (vehicle velocity, wheel speed, and wheel torque) in order to simulate sensor noises, their variances being fixed at 0.01 for the torque and for the velocities. The first simulations, named "nominal," were carried out with the parameter values given in Table I.

Note that, at the initial time, there is an error of 1.6 rad/s between the actual and estimated angular velocities of the wheel and an error of 5 mm between the actual and estimated radii. The results remain the same for any reasonably different initial value of this state vector. The observer gains have been stated as follows (fixed by simulation).

1) *High-Gain Observer (22):* $\tau_1 = \tau_2 = 80$.

$$K = \begin{bmatrix} 2\alpha & \alpha^2 & 0 & 0 \\ 0 & 0 & 2\alpha & \alpha^2 \end{bmatrix}^T \text{ with } \alpha = 100.$$

2) *Sliding-Mode Observer (23):* $L_{\Phi_1} = L_{\Phi_2} = 0.1$.

The first step of observer design consists in checking the observability of the system (independent of the observer strategy) through the analysis of the invertibility of (14). Then, the determinant of the Jacobian of the transformation Ψ is given by Fig. 1. As this determinant never equals 0, one concludes that Ψ establishes a state transformation.

Fig. 2 displays the effective radius estimation and the rolling resistance force. It appears that the proposed observers provide correct estimations of the two variables R and F_r. Note that the

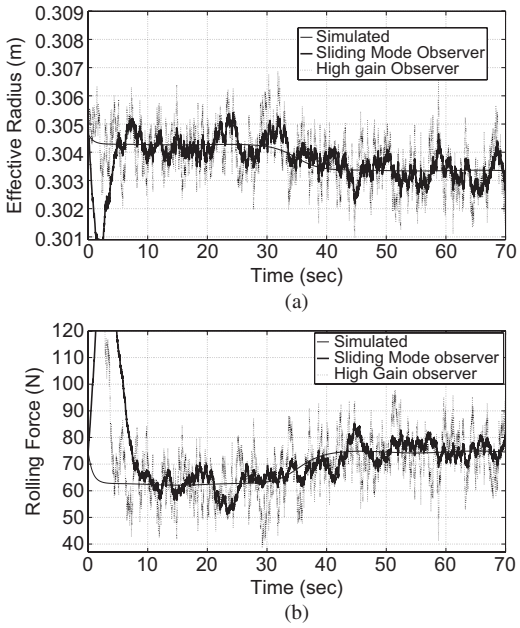


Fig. 2. Simulation results. (a) Current (dotted line) effective radius R (m) and its estimated value [solid line (m)] versus time (s). (b) Current (dotted line) resistance rolling force F_r (N) and its estimated value [solid line (N)] versus time (s).

both observers are evaluated by having no information of the dynamics of R and F_r which induces a very large uncertainty. Furthermore, there is also the simplifying assumption of F_z (see Section III-A on reduced model). In order to evaluate the quality of the estimation, the following two mean estimation errors have been computed for each observer:

$$\bar{e}_R = \frac{1}{N}\sum_{i=1}^{N} |R(t_i) - \hat{R}(t_i)|$$
$$\bar{e}_F = \frac{1}{N}\sum_{i=1}^{N} |F_r(t_i) - \hat{F}_r(t_i)| \quad (25)$$

for $t_i \in [15\text{ s}, 70\text{ s}]$. Table III displays the values of these errors for the both observers, in several cases of parametric variations (PV) as follows:

1) PV1: nominal case;
2) PV2: noise variance = 0.1 (0.01 in the nominal case);
3) PV3: mass = 352 kg (440 kg in the nominal case);
4) PV4: road profile = 2 cm (1 mm in the nominal case);
5) PV5: peak friction $\mu_0 = 0.78$ (0.9 in the nominal case);
6) PV6: optimal slip $\lambda_0 = 0.2$ (0.25 in the nominal case).

From the previous robustness tests, it appears that the sliding-mode observer is more robust than the high-gain one, especially in case of noise and variation of road–tire contact. For this reason, and for the sake of brevity, only the sliding-mode observer is applied on the experimental data.

C. Single-Wheel Monitoring on Experimental Data

The experimental data were obtained on a prototype car going ahead at a speed of approximatively 40 km/h. In order to get different values of effective radius and rolling resistance

TABLE III
MEAN ESTIMATION ERRORS FOR SLIDING-MODE (SM) AND HIGH-GAIN (HG) OBSERVERS. THE MEAN ESTIMATION ERRORS ARE FOR $t \in [15\text{ s}, 70\text{ s}]$

PV	$\bar{e}_R$ (m) (SM)	$\bar{e}_R$ (m) (HG)	$\bar{e}_F$ (N) (SM)	$\bar{e}_F$ (N) (HG)
PV1	$1.8\ 10^{-4}$	$2.2\ 10^{-4}$	2.77	2.4
PV2	$3.5\ 10^{-4}$	$6.9\ 10^{-4}$	3.23	7.02
PV3	$1.8\ 10^{-4}$	$2.4\ 10^{-4}$	2.71	2.35
PV4	$1.7\ 10^{-4}$	$2.4\ 10^{-4}$	2.32	2.48
PV5	$2.1\ 10^{-4}$	$2.7\ 10^{-4}$	2.66	2.53
PV6	$2.8\ 10^{-4}$	$2.6\ 10^{-4}$	2.88	6.53

force, two inflation pressure values were considered ($P_1 = 2.3$ bar, which is the nominal pressure; and $P_2 = 1.9$ bar) for the left front wheel, whereas the pressures are maintained at their nominal values for the right front wheel (2.3 bar) and the rear wheels (2.2 bar).

For the sake of brevity, only the observer application results to the left front wheel are presented. The estimated wheel radius and rolling resistance force are displayed in Fig. 3. This figure shows that, when the tire inflation pressure decreases, the observer gives a smaller value of the radius and an increased value for the rolling resistance force.

The observer allows the detection of the tire inflation variation and it also presents a convergence time (about 10 s) which is compatible with the objectives of the car manufacturers. In fact, it is smaller than the imposed time responses by standard norms on wheel monitoring systems. In order to show more clearly the influence of the pressure (and to attempt to propose a diagnosis tool for pressure fall), Fig. 4 displays the histograms of the both estimated variables. These graphical representations show the distribution of R and F_r and confirm the direct influence of the pressure fall on the both variables. Thus, the joint estimation of the rolling resistance and the effective radius appears as an adequate tool for pressure monitoring.

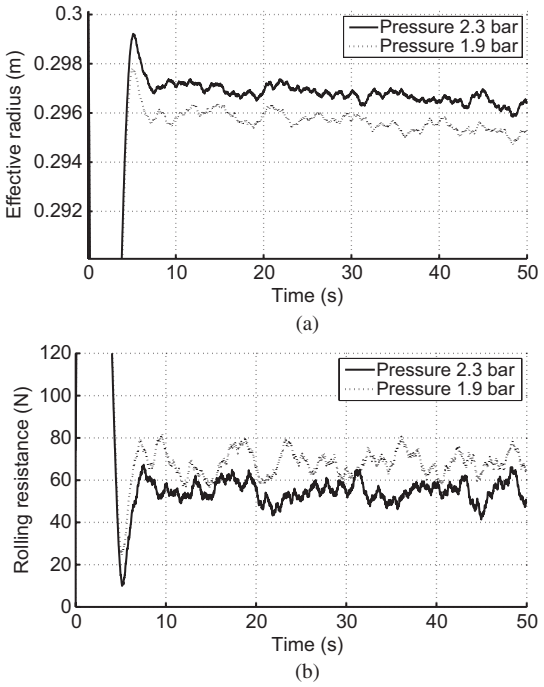


Fig. 3. Experimental results. (a) Estimated wheel effective radii (m) versus time (s) for two different tire pressures (solid line: 2.3 bar and dotted line: 1.9 bar). (b) Estimated rolling resistance forces (N) versus time (s) for two different tire pressures (solid line: 2.3 bar and dotted line: 1.9 bar).

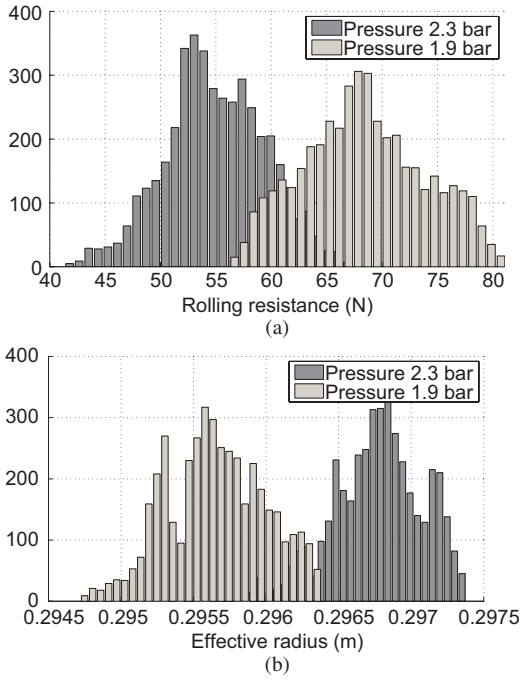


Fig. 4. Experimental results. (a) Statistical distribution of estimated rolling resistance force (N) for two different tire pressures (2.3 and 1.9 bar). (b) Statistical distribution of estimated wheel effective radius (m) for two different tire pressures (2.3 and 1.9 bar).

V. EXTENSION TO THE VEHICLE FRONT AXLE

The goal here consists in monitoring the tire pressures of two wheels located in the front axle of the vehicle. The problem is then to extend the previous methodology for the observer design by estimating the two wheel radii and the front axle resistance rolling force. The proposed observer takes into account the coupling between the both wheels, which leads to a more accurate estimation of the wheel radii. Starting from the increase of the rolling resistance due to tire pressure fall, the estimation of the radius of each wheel is an important point in order to establish whether there is a pressure fall and to locate it. Moreover, the single-wheel observer can also be used to estimate the rolling resistance of each wheel.

Consider the rotational dynamics of two wheels axle with $M_{1/2}$, the mass of the half-car, and $F_{d1/2}(x)$, its aerodynamic drag force. In the sequel, index l (resp. r) refers to the front left (resp. right) wheel. Denote by $F_{rlr} = F_{rl} + F_{rr}$ the global rolling resistance force on the axle. The effective radii R_l and R_r of the front wheels and the global rolling resistance force F_{rlr} on the front axle are unknown, as their dynamics; these

TABLE IV
WHEEL AND CAR MODEL PARAMETERS FOR THE AXLE MODEL

Parameter	Value	Unit
J_l	1.672	$\text{kg} \cdot \text{m}^2$
J_r	1.672	$\text{kg} \cdot \text{m}^2$
$M_{1/2}$	607.5	kg
$A_{d1/2}$	0.815	m^2
C_f	0	$\text{kg} \cdot \text{m}^2 \cdot \text{s}^{-1}$
$C_{d1/2}$	0.3125	
λ_0	0.15	

latter ones read as

$$\dot{R}_l = \eta_l(t), \quad \dot{R}_r = \eta_r(t), \quad \dot{F}_{rlr} = \eta_F(t) \tag{26}$$

with $\eta_l(t)$, $\eta_r(t)$, and $\eta_F(t)$ unknown and bounded. The torques applied to each wheel are, respectively, denoted as Γ_r and Γ_l. By denoting $x = [x_1\ x_2\ x_3\ x_4\ x_5\ x_6]^T = [\Omega_l\ \Omega_r\ v_x\ F_{rlr}\ R_l\ R_r]^T$ with $u = [u_1\ u_2]^T = [\Gamma_l\ \Gamma_r]^T$ the control input, the dynamic behavior of the whole axle is given by

$$\dot{x} = f_M(x) + \chi_M(y,u) + \Delta f_M \tag{27}$$

where

$$f_M(x) = \begin{bmatrix} -\frac{1}{J_l}x_5 F_{xl}(x) \\ -\frac{1}{J_r}x_6 F_{xr}(x) \\ \frac{1}{M_{1/2}}\Big(F_{xl}(x) + F_{xr}(x) - F_{d1/2}(x) - x_4\Big) \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\chi_M(y,u) = \begin{bmatrix} -\frac{1}{J_l}C_f x_1 + \frac{1}{J_l}u_1 \\ -\frac{1}{J_r}C_f x_2 + \frac{1}{J_r}u_2 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \qquad \Delta f_M = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \eta_l(t) \\ \eta_r(t) \\ \eta_F(t) \end{bmatrix}$$

with $F_{d1/2}$ derived from (5), and F_{xl}, F_{xr} derived from (4). The measured variables are the wheel velocities and the vehicle's longitudinal speed, $y=[x_1\ x_2\ x_3]^T$. Note that the structure of (27) is similar to that of (10). As before, it can be shown that system (27) is observable and that each output variable admits an observability index equal to 2. Then, an observer for system (27) reads as

$$\dot{\hat{x}} = f_M(\hat{x}) + \chi(y,u) + \left[\frac{\partial \Psi_M}{\partial \hat{x}}\right]^{-1} \kappa_M(y,\hat{x}) \tag{28}$$

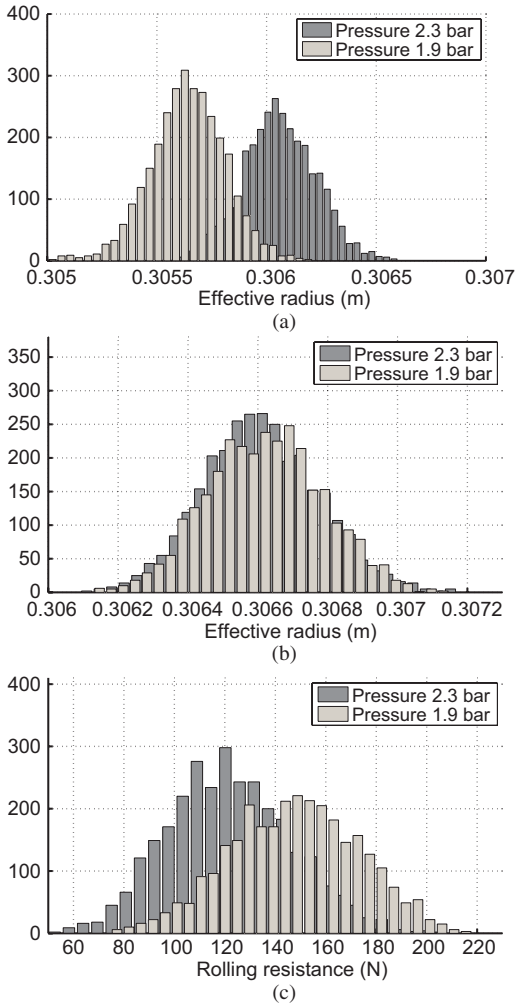


Fig. 5. Experimental results. (a) Statistical distribution of estimated left wheel radius R_l (m) for two different tire pressures (2.3 and 1.9 bar). (b) Statistical distribution of estimated right wheel radius R_r (m) for two different tire pressures (2.3 and 1.9 bar). (c) Statistical distribution of the rolling resistance force F_{rlr} (N) of the axle for two different tire pressures (2.3 and 1.9 bar).

with $\Psi_M = [x_1\ \dot{x}_1\ x_2\ \dot{x}_2\ x_3\ \dot{x}_3]^T$ and

$$\kappa_M(y,\hat{x}) = \begin{bmatrix} \underbrace{a_1\ L_{\Phi_1}^{\frac{1}{2}}|x_1-\hat{x}_1|^{\frac{1}{2}}\text{sign}(x_1-\hat{x}_1)}_{\gamma_1} \\ a_2\ L_{\Phi_1}\text{sign}(\gamma_1) \\ \underbrace{a_3\ L_{\Phi_2}^{\frac{1}{2}}\ |x_2-\hat{x}_2|^{\frac{1}{2}}\text{sign}(x_2-\hat{x}_2)}_{\gamma_3} \\ a_4\ L_{\Phi_2}\text{sign}(\gamma_3) \\ \underbrace{a_5\ L_{\Phi_3}^{\frac{1}{2}}\ |x_3-\hat{x}_3|^{\frac{1}{2}}\text{sign}(x_3-\hat{x}_3)}_{\gamma_5} \\ a_6\ L_{\Phi_3}\text{sign}(\gamma_5) \end{bmatrix}.$$

The parameters used for the axle model are summarized in Table IV. The observer coefficients are defined as $a_1 = a_3 = a_5 = 1.5$; $a_2 = a_4 = a_6 = 1.1$; and $L_{\Phi_1} = L_{\Phi_2} = L_{\Phi_3} = 1$.

The experimental evaluation of the observer performances was made on the same vehicle and under similar conditions as before. The initial values of the observer state variables read as $\hat{x}(0) = [16/0.31\ 16/0.31\ 16\ 102.5\ 0.305\ 0.305]^T$.

Fig. 5 displays the histograms of the left/right wheels' radii, and that of the axle rolling resistance force. It appears that the following.

1) The rolling resistance force increases when the tire pressure decreases; however, this information alone does not allow us to establish which tire presents a pressure fall.
2) The radii of the both wheels are directly connected to their pressure. The histograms allow us to establish that the pressure fall comes from the left wheel.

VI. Conclusion

Nonlinear (high-gain and high-order sliding mode) observers were applied in order to get a robust and online estimation of tire effective radius and rolling resistance force. A future objective is to integrate this observation solution in tire pressure monitoring. This brief paper presented the very first experimental results, which appeared promising given that, from vehicle signals, the observers were able to provide information allowing us to conclude (or not) to a pressure fall. Future research will be conducted in several directions. The first one is methodological in order to propose an integrated strategy for observer design using adaptive gain solutions which will strongly reduce the time for parameter tuning. Another research direction concerns the embedded integration of the approaches by taking into account all the constraints as computation time. Finally, as claimed previously, there is also the necessity to integrate the observation solutions in a tire pressure monitoring system.

References

[1] L. Li and F. Wang, *Advanced Motion Control and Sensing for Intelligent Vehicles*. Berlin, Germany, Springer Verlag, 2007.

[2] R. Matsuzaki and A. Todoroki, "Wireless monitoring of automobile tires for intelligent tires," *Sensors*, vol. 8, no. 12, pp. 8123–8138, 2008.

[3] D. J. Schuring, "Effects of tire rolling loss on vehicle fuel consumption," *Tire Sci. Technol.*, vol. 22, no. 3, pp. 149–161, 1994.

[4] W. H. Waddell, "Inflation pressure retention effects on tire rolling resistance and vehicle fuel economy," ExxonMobil Chem. Co., Sacramento, CA, Tech. Rep. PYBA31, 2008.

[5] D. Stein, "Tires and passenger vehicle fuel economy: Informing consumers and improving performances," Trans. Res. Board, Washington, DC., TRB Special Rep. 286-78, Aug. 2006.

[6] L. Li, F.-Y. Wang, and Q. Zhou, "A watch in developments of intelligent tire inspection and monitoring," in *Proc. IEEE Int. Conf. Veh. Electron. Safety*, 2005, pp. 333–338.

[7] S. Velupillai and L. Guveng, "Tire pressure monitoring," *IEEE Control Syst. Mag.*, vol. 27, no. 6, pp. 22–25, Jun. 2007.

[8] G. Jazar, *Vehicle Dynamics: Theory and Applications*, 1st ed. New York: Springer Verlag, 2008.

[9] D. J. Schuring, "The rolling loss of pneumatic tires," *Rubber Chem. Technol.*, vol. 53, no. 3, pp. 600–727, 1980.

[10] H. Mayer, "Comparative diagnosis of tyre pressures," in *Proc. IEEE Conf. Control Appl.*, Aug. 1994, pp. 627–632.

[11] H. B. Pacejka, *Tyre and Vehicles Dynamics*. Amsterdam, The Netherlands: Elsevier, 2005.

[12] R. Rajamani, *Vehicle Dynamics and Control*. New York: Springer-Verlag, 2006.

[13] N. Sirdi, A. Rabhi, L. Fridman, J. Davila, and Y. Delanne, "Second order sliding mode observer for estimation of velocities, wheel slip, radius and stiffness," in *Proc. Amer. Control Conf.*, Jun. 2006, pp. 1–5.

[14] C. Carlson and J. Gerdes, "Consistent nonlinear estimation of longitudinal tire stiffness and effective radius," *IEEE Trans. Control Syst. Technol.*, vol. 13, no. 6, pp. 1010–1020, Nov. 2005.

[15] N. Sirdi, A. Rabhi, L. Fridman, J. Davila, and Y. Delanne, "Second-order sliding-mode observer for estimation of vehicle dynamic parameters," *Int. J. Veh. Design*, vol. 48, nos. 3–4, pp. 190–207, 2008.

[16] C. Carlson and J. Gerdes, "Identifying tire pressure variation by nonlinear estimation of longitudinal stiffness and effective radius," in *Proc. Int. Symp. Adv. Veh. Control*, 2002, pp. 1–8.

[17] H. Shraim, B. Ananou, M. Ouladsine, and L. Fridman, "A new diagnosis strategy based on the online estimation of the tire pressure," in *Proc. Eur. Control Conf.*, Jul. 2007, pp. 1–7.

[18] D. Lee and Y. Park, "Sliding-mode-based parameter identification with application to tire pressure and tire-road friction," *Int. J. Autom. Technol.*, vol. 12, no. 4, pp. 571–577, 2011.

[19] T. Gillespie, *Fundamentals of Vehicle Dynamics*. SAE Int., Warrendale, PA, 1992.

[20] J. C. Beebe and B. D. Cargould, "Tire rolling resistance measurement system," U.S. Patent 4 489 598 A, Mar. 1984.

[21] R. S. Petrovich and T. V. Nikolaevich, "Definition method of coefficient of rolling resistance of wheel with pneumatic tire and device for its fulfilment," U.S. Patent 2 327 968 A, Jan. 14, 2006.

[22] G. R. Potts, "Methods and systems for measurement of tire rolling resistance," U.S. Patent 0 115 563 A1, Sep. 22, 2008.

[23] M. Atsushi, "Method for evaluating rolling resistance of tire, system for evaluating tire using the same, and program for evaluating rolling resistance of tire," U.S. Patent 0 249 527 A, Jul. 30, 2010.

[24] V. A. Petrushov, "Improvement in vehicle aerodynamic drag and rolling resistance determination from coast-down tests," in *Proc. Institut. Mech. Eng. Part D, J. Automob. Eng.*, vol. 212, no. 5, pp. 369–380, 1998.

[25] J. P. Gauthier, H. Hammouri, and S. Othman, "A simple observer for nonlinear systems, application to bioreactors," *IEEE Trans. Autom. Control*, vol. 37, no. 6, pp. 875–880, Jun. 1992.

[26] A. Levant, "Sliding order and sliding accuracy in sliding mode control," *Int. J. Control*, vol. 58, no. 6, pp. 1247–1263, 1993.

[27] A. Levant, "High-order sliding modes: Differentiation and output-feedback control," *Int. J. Control*, vol. 76, nos. 9–10, pp. 924–941, 2003.

[28] L. Fridman, J. Moreno, and R. Iriarte, *Sliding Modes After the 1st Decade of the 21st Century: State of the Art*, ser. Lecture Notes in Control and Inf. Sci., New York: Springer-Verlag, 2011.

[29] H. Imine, L. Fridman, H. Shraim, and M. Djemai, *Sliding Mode Based Analysis and Identification of Vehicle Dynamics*, ser. Lecture Notes in Control and Inf. Sci., New York: Springer-Verlag, 2011, vol. 414.

[30] J.-S. Lin and W.-E. Ting, "Nonlinear control design of anti-lock braking systems with assistance of active suspension," *IET Control Theory Appl.*, vol. 1, no. 1, pp. 343–348, 2007.

[31] C. Unsal and P. Kachroo, "Sliding mode measurement feedback control for antilock braking systems," *IEEE Trans. Control Syst. Technol.*, vol. 7, no. 2, pp. 271–281, Mar. 1999.

[32] G. Genta, *Motor Vehicle Dynamics: Modeling and Simulation*. World Scientific Publishing Co., World Scientific, Singapore, 2011.

[33] C. El Tannoury, F. Plestan, S. Moussaoui, and N. Romani, "Tyre effective radius and vehicle velocity estimation: A variable structure observer solution," in *Proc. IEEE Int. Multi-Conf. Syst., Signals Dev.*, Mar. 2011, pp. 1–6.

[34] C. Moog, F. Plestan, G. Conte, and A. Perdon, "On canonical forms of nonlinear systems," in *Proc. Eur. Control Conf.*, 1993, pp. 1514–1517.

[35] F. Plestan and A. Glumineau, "Linearization by generalized input-output injection," *Syst. Control Lett.*, vol. 31, no. 31, pp. 115–128, 1997.

[36] A. J. Krener and W. Respondek, "Nonlinear observers with linearizable error dynamics," *SIAM J. Control Optim.*, vol. 23, no. 2, pp. 197–216, 1985.

[37] V. Lebastard, Y. Aoustin, and F. Plestan, "Estimation of absolute orientation for a bipedal robot: Experimental results," *IEEE Trans. Robot.*, vol. 27, no. 1, pp. 170–174, Feb. 2011.

[38] A. Levant, "Finite-time stability and high relative degrees in sliding-mode control," in *Sliding Modes after the First Decade of the 21st Century* (Lecture Notes in Control and Information Sciences), vol. 412. New-York: Springer-Verlag, 2012, pp. 59–92.

[39] A. Filippov and F. Arscott, *Differential Equations with Discontinuous Right Hand Sides*, ser. Mathematics and its Applications Series., Kluwer Academic Publishers, 1988.

Résumé. Les travaux présentés dans ce mémoire d'habilitation à diriger des recherches s'articulent autour de la résolution de problèmes inverses en traitement du signal et d'image. La particularité de ces travaux est liée à la résolution de problèmes de grande taille par une approche bayésienne et d'outils issus de l'optimisation convexe, de la simulation stochastique et du calcul numérique. La première partie de ce manuscrit est un bilan synthétique de mes activités d'enseignement, au sein du département Automatique et Robotique de l'Ecole Centrale de Nantes, et de recherche à l'Institut de Recherche en Communications et Cybernétique de Nantes. Dans la seconde partie, et après une introduction générale visant à expliquer les questions majeures liés à la résolution de problèmes inverses de grande taille, seront présentées quelques contributions méthodologiques issues de mes travaux de recherche. Ces contributions consistent en la conception de méthodes de majoration-minimisation pour l'accélération des algorithmes d'optimisation itérative ainsi que le recours à des outils de calcul intensif sur des processeurs de cartes graphiques pour réduire le temps de calcul. Le fil conducteur de ces travaux réside dans le développement de méthodes ayant des propriétés théoriques éprouvées et possédant une structure algorithmique adaptée pour une implémentation parallélisable. Par ailleurs, dans le cadre de la résolution de problèmes inverses linéaires par des méthodes de Monte Carlo, une contribution récente, couplant optimisation itérative et simulation stochastique, concerne la proposition d'une approche originale d'échantillonnage de vecteurs gaussiens en grande dimension. La dernière partie de ce manuscrit portera sur les perspectives scientifiques à court terme et sur les grandes lignes de mon projet de recherche à plus long terme lié notamment à la coopération entre optimisation, simulation stochastique et calcul intensif pour la résolution de problèmes inverses.

Abstract. This manuscript of accreditation to supervise research concerns the resolution of inverse problems in signal and image processing. The particularity of this research is the focus on the context of large-scale inverse problems using convex optimization, stochastic simulation and scientific computing. The first part of this report gives a review of my teaching activities in the department Automatic and Robotic of Ecole Centrale de Nantes and research at the Research Institute on Communications and Cybernetics (IRCCyN). In the second part, after a short introduction explaining the main questions related to large-scale inverse problem resolution, some methodological contributions rising from my research activities are given. These are related to the conception of majorization-minimization methods for the algorithmic acceleration of iterative optimization methods and the use of intensive computing methods to reduce the processing time. A common point in these works is the development of theoretically sound methods and having an algorithmic structure adapted to a parallel implementation. In the case of a linear inverse problem resolution using Monte Carlo methods, a recent contribution that uses stochastic simulation and numerical optimization methods is related to an original approach for the simulation of Gaussian vectors in high dimensions. The last part of this report, gives some perspectives related to the resolution of large-scale inverse problems and its applications. The main direction that will be privileged concerns the cooperation between numerical optimization, stochastic simulation and high parallel computing tools for the resolution of inverse problems in the big data context.