Méthodes de point fixe pour la reconstruction en Cone-beam CT sur arceau interventionnel

Marion Savanier

université
PARIS-SACLAY

Iterative fixed point methods for image reconstruction in C-arm Cone-Beam CT

Méthodes de point fixe pour la reconstruction en Cone-Beam CT sur arceau interventionnel

Thèse de doctorat de l'Université Paris-Saclay

École doctorale n°580, Sciences et Technologies de l'Information et de la Communication (STIC)
Spécialité de doctorat: Sciences du traitement du signal et des images
Graduate school: Sciences de l'ingénierie et des systèmes
Référent: CentraleSupélec

Thèse préparée dans l'unité de recherche Centre de Vision Numérique (Université Paris-Saclay, CentraleSupélec), sous la direction d'Emilie CHOUZENOUX, Chargée de recherche Inria, et le co-encadrement de Cyril RIDDELL, Ingénieur de recherche à GE Healthcare.

Thèse soutenue à Paris-Saclay, le 18 novembre 2022, par

Marion SAVANIER

Composition du jury
Membres du jury avec voix délibérative

Sylvie LE HEGARAT-MASCLE Professeure, Université Paris-Saclay	Présidente
Elena LOLI PICCOLOMINI Professeure, University of Bologna	Rapporteur & Examinatrice
Dirk LORENZ Professeur, Technische Universität Braunschweig	Rapporteur & Examinateur
Johan NUYTS Professeur, Katholieke Universiteit Leuven	Examinateur

THESE DE DOCTORAT
NNT : 2022UPAST130

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre: Méthodes de point fixe pour la reconstruction en Cone-Beam CT sur arceau interventionnel
Mots clés: Tomographie, Imagerie interventionnelle, Apprentissage, Optimisation convexe, Points fixes

Résumé: L'imagerie tomographique par rayons X, appelée tomodensitométrie, est accessible en routine clinique diagnostique grâce au scanner. Le scanner permet de différencier très précisément les tissus humains mais ses performances sont obtenues par des choix techniques incompatibles avec une utilisation interventionnelle où le système d'imagerie ne doit pas empêcher l'accès au patient. Les arceaux rayons X interventionnels ont donc été dotés d'une technologie tomographique alternative, le Cone-Beam Computed Tomography (CBCT). Le CBCT est très performant en résolution spatiale mais limité pour la visualisation des faibles contrastes, surtout en présence de produit de contraste et d'objets thérapeutiques métalliques toujours sous-échantillonnés lors de l'acquisition. Des acquisitions plus économes en dose et plus rapides, obtenues en réduisant l'amplitude angulaire des mesures ou le nombre de mesures, sont souhaitables mais aggravent les problèmes liés au sous-échantillonnage.

La résolution du problème de reconstruction d'une région anatomique (ROI) à partir de projections s'effectue efficacement via la minimisation d'une fonction de coût compensant le sous-échantillonnage grâce à l'injection d'informations a priori sur la ROI. La minimisation peut être effectuée par un algorithme itératif lié à un schéma de point fixe. Quatre points essentiels sont (i) le choix du terme d'attache aux données, (ii) de la régularisation, (iii) la discrétisation des opérateurs linéaires présents dans ces deux termes et enfin, (iv) le choix de l'algorithme itératif. Souvent, ces méthodes ne sont pas utilisées dans un cadre assurant leur convergence. Dans cette thèse, nous proposons des méthodes de reconstruction itérative qui sont théoriquement convergentes et applicables à des acquisitions pour la radiologie interventionnelle.

Pour répondre à cet objectif, nous étendons les conditions de convergence d'un ensemble d'algorithmes proximaux lorsque l'adjoint du projecteur est remplacé par un autre opérateur. Une des motivations pour ce changement est l'absence de gestion des variations d'échantillonnage par le projecteur classique basé sur l'interpolation linéaire. La convergence des algorithmes est prouvée sous des conditions s'appuyant sur le caractère cocoercif de certains opérateurs linéaires.

Nous montrons ensuite qu'une modélisation des variations d'échantillonnage dans un schéma d'interpolation permet d'obtenir une discrétisation à la fois précise du projecteur et du rétroprojecteur. En partant d'un algorithme de rééchantillonnage proposé pour agrandir des images, nous concevons de nouveaux projecteurs et rétroprojecteurs adaptés à la géométrie conique avec un détecteur plan, avec différents compromis de précision et rapidité.

Par ailleurs, nous proposons une méthode itérative pour la reconstruction d'aiguilles percutanées à partir d'acquisitions ayant une amplitude angulaire limitée. Nous adoptons une stratégie de décomposition du volume pour associer différents termes de régularisation directionnelle à chaque composante et ainsi reconstruire un fond anatomique sur lequel sont superposées les aiguilles.

Enfin, nous proposons une nouvelle formulation régularisée pour la reconstruction d'une région anatomique à partir d'acquisitions ayant une faible densité angulaire. La combinaison d'un terme d'attache aux données robuste et d'une régularisation de type variation totale permet de limiter les artéfacts issus d'objets intenses présents dans les projections mais en dehors de la grille de reconstruction. Pour permettre une reconstruction précise et rapide, nous exploitons les apports des techniques d'apprentissage. Nous proposons un algorithme itératif déroulé permettant un apprentissage supervisé des paramètres du problème en un nombre d'itération restreint. Nous montrons qu'une meilleure reconstruction est obtenue en apprenant l'adjoint des opérateurs linéaires présent dans le terme de régularisation.

Title: Iterative fixed point methods for image reconstruction in C-arm Cone Beam CT
Keywords: Tomography, Interventional imaging, Deep learning, Convex Optimization, Fixed Point Theory

Abstract: X-ray tomographic imaging, known as CT scanning, is available in clinical diagnostic routines thanks to the scanner. CT scanners allow precise differentiation of human tissues. However, these performances are obtained by technical choices incompatible with interventional use where the imaging system must not prevent access to the patient. Interventional X-ray C-arms have thus been equipped with an alternative tomographic technology: Cone-Beam Computed Tomography (CBCT). CBCT offers high spatial resolution for vessel imaging but remains limited for low contrast visualization, especially in the presence of injected contrast media and metallic therapeutic objects always under-sampled during the acquisition. Simplified acquisitions that are more dose-saving and faster, obtained by reducing the angular amplitude covered by the measurements or the number of measurements, are desirable but exacerbate under-sampling issues.

The problem of reconstructing an anatomical region from projections can be efficiently solved by minimizing a cost function that compensates for under-sampling through a priori information about the region to be reconstructed. The minimization can be performed by an iterative algorithm built on a fixed point scheme. Four essential components in iterative methods are (i) the choice of the data fidelity, (ii) the regularization, and (iii) the discretization of the linear operators involved in these two terms (e.g., the projector and its adjoint, the backprojector), and (iv), the choice of the iterative algorithm. In practice, these methods are not used in a framework ensuring their convergence. In this thesis, we propose methodological contributions leading to iterative reconstruction methods that are theoretically convergent and applicable to acquisitions for interventional radiology.

To achieve this objective, we extend the convergence conditions of a set of proximal algorithms when the projector adjoint is replaced by another operator, such as the voxel-driven backprojector often used in analytical reconstruction. One of the motivations for this change is the lack of management of sampling variations by the classical projector based on linear interpolation. The convergence of the algorithms is proved under conditions based on the cocoercivity of some linear operators.

We then show that the modeling of sampling variations in an interpolation scheme allows for the derivation of an accurate discretization of the projector and the backprojector. Starting from an existing resampling algorithm for image enlargement, we design new projectors and backprojectors adapted to the cone-beam geometry with a flat detector, with different accuracy and speed trade-offs.

Next, we propose an iterative method for reconstructing percutaneous needles from acquisitions with limited angular amplitude. We adopt a volume decomposition strategy associated with different directional regularization terms for each component and thus reconstruct an anatomical background on which the needles are superimposed.

Finally, we propose a new regularized formulation for reconstructing an anatomical region from low angular density acquisitions. The combination of a robust data fidelity term with a total variation regularization limits artifacts from intense objects present in the projections but out of the reconstruction grid. To reach an accurate and fast reconstruction, we exploit recent advances in deep learning to propose an iterative unfolded algorithm allowing supervised learning of the problem parameters in a limited number of iterations. We show that reconstruction is improved by learning the adjoint of the linear operators in the regularization term.

Remerciements

Ce travail a bénéficié d'un financement CIFRE avec GE Healthcare.

Je voudrais commencer par remercier mes encadrants, Cyril Riddell (GE Healthcare), l'instigateur de cette thèse, et Émilie Chouzenoux (CVN), pour leurs précieux conseils prodigués durant ces années de thèse. Merci Cyril pour m'avoir partagé ton expertise en reconstruction tomographique et pour tes retours sur toutes les parties de mon travail, m'incitant toujours à prendre du recul. Merci Emilie et aussi merci Jean-Christophe Pesquet (CVN), qui a fait bien plus qu'apporter un soutien essentiel pendant la thèse, pour m'avoir beaucoup appris dans le domaine de l'optimisation. Les discussions que nous avons eues et votre rigueur scientifique m'ont beaucoup apporté. Je tiens aussi à remercier tous les membres de mon jury, qui ont pris le temps d'évaluer mon travail, la présidente du jury, Sylvie Le Hegarat-Mascle, ainsi que Johan Nuyts et les deux rapporteurs, Elena Loli Piccolomini et Dirk Lorenz, qui ont relu ce manuscrit.

Côté GE, je remercie Valérie Desnoux pour m'avoir accueillie dans son équipe, à Buc. Mes remerciements s'étendent ensuite à l'ensemble des ingénieurs de l'équipe qui ont contribué à un environnement professionnel stimulant : Yves, Vincent, Maxime, Thomas, Liliane et Régis. Je remercie également Pierre-Louis, mon co-bureau et également doctorant, pour avoir partagé la première année de thèse pré-covid avec moi à Buc. Je te souhaite un bel avenir professionnel et personnel.

Coté CVN, je souhaite exprimer ma gratitude à Jana Dutrey pour sa bienveillance, à Hugues Talbot pour sa bonne humeur et sa gentillesse et également à tous les doctorants du labo, que j'ai été très heureuse de (re)découvrir après de nombreux mois de télétravail. Merci aux plus anciens, Marie-Caroline, Arthur, Maissa, Sagar, Kavya, Anna et au plus récents, Ségolène, Jean-Baptiste, Mouna, Mathieu, Thomas, Gabriele, Théodore, Younes, Alexandre, Loïc, Aymen, Claire, Simona ainsi qu'à tous les autres membres du CVN que j'ai eu la chance de côtoyer. Merci à Peng Wang, ancien stagiaire à GE puis au CVN, désormais en thèse et avec qui j'ai aimé travailler. Une mention spéciale à Mario, mon acolyte de fin de thèse au CVN, mais aussi à la BNF. Je souhaite également remercier les membres de l'Academic Writing Center, notamment, Melissa Ann Thomas pour sa pédagogie et ses conseils que je garderai précieusement pour la suite de mon parcours professionnel.
Pendant ma thèse, j'ai eu la chance de collaborer avec Andrès Contreras, post-doctorant au CVN, que je tiens particulièrement à féliciter pour son poste de permanent au Chili. Je remercie également la Graduate School Computer Science de Centrale Supélec et l'ED STIC qui m'ont accordé 3 mois de financement supplémentaire à la fin de mon contrat CIFRE.

Pour finir, j'ai une pensée pour mes proches et mes amis, ceux qui sont avec moi tous les jours et ceux que je vois moins souvent car plus éloignés. Vous connaissez votre importance à mes yeux.

Résumé (French)

Cette thèse a été effectuée dans le cadre d'un partenariat CIFRE. Elle a pour but la reconstruction tomographique conique sur arceau interventionnel via des méthodes de points fixes.

L'imagerie tomographique par rayons X, aussi appelée tomodensitométrie, est utilisée en routine clinique à des fins de diagnostic via le scanner CT. Les scanners permettent d'identifier les faibles contrastes d'atténuation aux rayons X des tissus anatomiques. En fournissant des images anatomiques, ils contribuent au diagnostic d'un très grand nombre de pathologies.
La radiologie interventionnelle est une spécialité médicale regroupant des thérapies minimalement invasives faites sous contrôle radiologique rayons X, pour traiter des pathologies variées (vasculaires, tumorales, osseuses) à l'aide d'outils introduits par voie endovasculaire (cathéters, stents, coils) ou percutanée (aiguilles). Durant ce type de procédures, le radiologue interventionnel a régulièrement besoin d'acquérir des images montrant la position de son outil par rapport à l'anatomie du patient. Un scanneur CT acquiert les données tomographiques d'une façon incompatible avec une utilisation en radiologie interventionnelle où le système d'imagerie ne doit pas empêcher l'accès au patient. Ainsi, le système de guidage le plus souvent employé est l'arceau interventionnel, robot manipulant un tube à rayons X et un détecteur digital plan qui forment la chaine image. La conception des arceaux interventionnels privilégie la génération d'images projectives 2D en temps réel. Une structure mécanique permet de déplacer la chaine image suivant trois axes de rotation, de manière à l'orienter, à chaque instant de la procédure, de la manière la plus appropriée à l'anatomie d'intérêt, aussi appelée région d'intérêt (ROI). Pour permettre une visualisation 3D de l'atténuation aux rayons X de la ROI, ces arceaux sont dotés d'une technologie tomographique spécifique, le Cone-Beam Computed Tomography (CBCT), aussi appelée tomodensitométrie conique sur arceau interventionnel.
L'imagerie CBCT peut être utilisée à plusieurs étapes d'une procédure interventionnelle. Elle sert à visualiser des tissus anatomiques pour planifier la procédure, ou un outil interventionnel pour guider le geste thérapeutique, ou l'anatomie avec l'outil pour contrôler le résultat de l'intervention (Chapitre 2). Avec le CBCT, la résolution en contraste est limitée, restreignant la différentiation des tissus anatomiques. En présence de vaisseaux injectés, de métal ou de câbles électriques d'enregistrement des signaux vitaux du patient, des artéfacts en stries sont présents dans les images reconstruites. Ces artefacts proviennent des contraintes d'acquisition qui conduisent au sous-échantillonnage des données tomographiques: projections tronquées et en nombre insuffisants pour la reconstruction de la ROI. L'acquisition des données autour du patient est lente et requiert souvent de repositionner ce dernier pour permettre le geste thérapeutique. Une rotation d'amplitude angulaire réduite est donc souhaitable.
Ainsi, le sous-échantillonnage est un problème constant du CBCT sur arceau, et il y a un intérêt pratique à l'amplifier. Cela est envisageable s'il existe des informations contextuelles complémentaires aux mesures et si l'on sait les utiliser pendant la reconstruction.

Mathématiquement, la reconstruction d'une région d'intérêt (ROI) en CBCT est un problème inverse dit mal posé, et d'autant plus mal posé qu'il est sous-échantillonné. Les méthodes de reconstruction itérative ont démontré leur supériorité sur les méthodes analytiques (Chapitre 3) pour intégrer à la reconstruction de la connaissance a priori.
Ces connaissances sont modélisées sous forme de termes de contrainte et/ou de régularisation dans un problème de minimisation dont la solution donne une estimation de la ROI. Par exemple, la régularisation de type variation totale a été utilisée pour compenser les insuffisances de la trajectoire circulaire et de la faible densité angulaire d'acquisitions CBCT. Une fois le problème de minimisation formulé, sa résolution est effectuée itérativement par des méthodes variationnelles d'optimisation. Ces dernières peuvent être vues comme une forme particulière d'algorithmes de points fixes, un cadre plus général où ces difficultés théoriques peuvent être résolues. Malgré leurs bénéfices, les méthodes variationnelles sont associées às temps de calcul souvent incompatibles avec la pratique clinique. Des accélérations sont obtenues au prix de compromis théoriques, qui mettent parfois en péril la convergence de ces méthodes.

Le but de cette thèse est de proposer des méthodes de reconstruction itérative, pour le CBCT interventionnel, qui soient théoriquement convergentes et applicables à des acquisitions simplifiées.
Pour répondre à cet objectif, nous étendons les conditions de convergence d'un ensemble d'algorithmes proximaux lorsque l'adjoint du projecteur est remplacé par un autre opérateur, tel que le rétroprojecteur discret utilisé en reconstruction analytique. En effet, l'étape de discrétisation du projecteur et du rétroprojecteur en reconstruction itérative n'est pas triviale. Ces opérateurs modélisent des effets de rotation dans un repère cartésien et existe une vaste littérature sur leur discrétisation, avec différent compromis entre complexité et précision. Ainsi un projecteur rapide et précis n'induit pas forcément un rétroprojecteur adjoint ayant les mêmes qualités, d'où l'idée de découpler ces opérateurs et de s'affranchir de la propriété mathématique d'adjoint. La symmétrie du produit de l'opérateur par son adjoint est requise pour s'assurer que la méthode converge vers un minimiseur que l'on sait caractériser. Mais une image reconstruite à valeur clinique ne coincide généralement pas avec le minimiseur d'une fonction de coût. Il est donc intéressant d'étendre les résultats de convergence de plusieurs algorithmes proximaux afin de les adapter à ce cadre. Nous effectuons cette étude dans le cas où l'opérateur adjoint est modifié (Chapitres 4, 5) et où des métriques de préconditionnement différentes sont utilisées pour l'opérateur proximal et pour le pas de gradient d'une méthode de gradient proximal (Chapitre 4). Notre analyse couvre les algorithmes du gradient proximal, de Condat-Vũ - une extension de l'algorithme de Chambolle-Pock -, de Combettes-Pesquet et de Loris-Verhoeven. Tous ces algorithmes sont couramment utilisés en traitement d'images, dépassant le cadre de la reconstruction tomographique.
Ensuite, nous présentons une nouvelle perspective sur les schémas de discrétisation des opérateurs tomographiques (Chapitre 6). Nous montrons que la modélisation des variations d'échantillonnage en géométrie conique avec un détecteur plan est crucial pour obtenir une discrétisation à la fois précise du projecteur et du rétroprojecteur. Cela est possible sans d'adopter un modèle géométrique mais avec une approche par rééchantillonnage, qui est une extension d'une méthode existante pour le redimensionnement d'images. Notre approche repose sur l'optimisation de la représentation des signaux transformés par une homographie spécifique à la géométrie d'acquisition, à l'aide de B-splines ayant un support variable. Différents modèles d'interpolation peuvent être réécrits suivant cette approche, comme le modèle géométrique de l'état de l'art appelé "distance-driven".

Par ailleurs, nous proposons d'augmenter la qualité des informations a priori en adoptant une stratégie de décomposition du volume à reconstruire en composantes associées à différentes régularisations. Nous montrons l'intérêt de cette stratégie pour la reconstruction d'aiguilles métalliques à partir d'acquisitions avec une amplitude angulaire réduite (Chapitre 7). La connaissance de la direction des aiguilles est utilisée pour les reconstruire en les séparant des tissus anatomiques.
Nous montrons également que la modélisation du contenu du volume reconstruit n'est pas le seul levier, le terme dit d'attache aux données est utilisé pour modéliser le bruit statistique des mesures acquises (on parle de reconstruction statistique). Pour les problèmes de sous-échantillonnage, notamment pour une reconstruction de la ROI sur une grille réduite, des erreurs doivent être tolérées dans le terme d'attache aux données, suivant une statistique qui n'est plus celle du bruit quantique. Nous proposons une autre approche dans le Chapitre 8 i.e., une autre statistique via l'utilisation d'une fonction non-convexe issue de la théorie des M-estimateurs. L'effet joint du modèle statistique et de la régularisation permet la réduction des artefacts de sous-échantillonnage. Cette approche est combinée aux outils récents de l'apprentissage profond pour l'accélération des algorithmes proximaux et pour faciliter leur paramétrisation. Ce dernier point est la clé d'une mise en oeuvre efficace et effective de reconstructions itératives dans un contexte clinique contraint. Les réseaux de neurones profonds comportent, par nature, de nombreux degrés de liberté, ce qui rend leur analyse difficile et s'accompagne d'un questionnement concernant leur fiabilité et leur stabilité. Malgré des stratégies différentes, les approches itératives et celles d'apprentissage profond ont des avantages et des inconvénients complémentaires qui suggèrent de les combiner. Cette idée est explorée via une architecture, nommée U-RDBFB, obtenue en "déroulant" un algorithme proximal adapté à notre fonction de coût. Nous montrons l'intérêt d'apprendre l'adjoint des opérateurs de régularisation dans notre problème de reconstruction. Nous procédons à des expériences numériques pour la reconstruction de régions d'intérêt en géometrie parallèle à partir d'un faible nombre de projections sur 180°. Notre approche présente de bonnes performances par rapport à l'approche variationnelle classique et à des méthodes d'apprentissage profond, y compris d'autres architectures issues d'algorithmes déroulés.
Enfin dans le Chapitre 9, nous résumons nos principales contributions et nous proposons plusieurs pistes pour de futurs travaux.

Contents

Notation

D_N^+	diagonal matrix with positive elements in $\mathbb{R}^{N\times N}$
$L^2(\mathbb{R})$	Hilbert space of measurable, square-integrable functions from $\mathbb{R}$ to $\mathbb{R}$
M	matrices will be denoted by uppercase letters
$M^\top$, M^{-1}	transpose and inverse of M
$M^\dagger$	pseudo-inverse of M
S_N^+	symmetric positive-definite in $\mathbb{R}^{N\times N}$
$\mathbb{N}$	set of positive integers
$\mathbb{R}$, $[0,+\infty[$	sets of real, positive real scalars
$\mathbb{R}^m$	set of vectors with m entries
$\mathbb{R}^{m\times n}$	set of matrices with m rows and n columns
β_δ^n	B-spline of order n and width $\delta > 0$
β^n	B-spline of order n and width 1
ι_C	indicator function of set C
proj_C	projection onto set C
$\widehat{f}(k_t) = \mathcal{F}_1[f(t)](k_t)$	1D Fourier transform of f
$\widehat{f}(k_x,k_y) = \mathcal{F}_2[f(x,y)](k_x,k_y)$	2D Fourier transform of f
v	vectors and scalars will be denoted by lowercase letters
$\|\|\|M\|\|\|$	spectral norm of M

Acronyms

ADMM	Alternating Direction Method of Multipliers
AR	Analytical Reconstruction
ART	Algebraic Reconstruction Technique
ATV	Anisotropic Total Variation
BP	BackProjection
CAU	CAUdal
CBCT	Cone Beam Computed Tomography
CP	Combettes-Pesquet algorithm
CRA	CRAnial
CT	Computed Tomography
CV	Condat-Vũ algorithm
DBFB	Dual Block-coordinate Forward Backward algorithm
DD	Distance-Driven
DEQ	Deep EQuilibrium
DFB	Dual Forward-Backward algorithm
DTV	Directional Total Variation
DU	Deep Unfolding
FBP	Filtered BackProjection
FDK	Feldkamp-David-Kress algorithm
FISTA	Fast Iterative Shrinkage-Thresholding Algorithm
FOV	Field-Of-View
FP	Forward Projection
HU	Hounsfield Units: air is -1000 and water is 0
IR	Iterative Reconstruction
ISTA	Iterative Soft-Thresholding Algorithm
LAO	Left Anterior Oblique
LV	Loris-Verhoeven algorithm
MAR	Metal Artifacts Reduction
MIP	Maximum Intensity Projection
MRI	Magnetic Resonance Imaging
PDHG	Primal-Dual Hybrid Gradient algorithm
PGA	Proximal Gradient Algorithm
POCS	Projection Onto Convex Sets
RAO	Right Anterior Oblique
ROI	Region-of-Interest
SF	Separable Footprint
sHU	Shifted Hounsfield Units: air is 0 and water is 1000
SID	Source-to-Image Distance
SOD	Source-to-Object Distance
SPECT	Single Photon Emission Computed Tomography
TV	Total Variation

1 General introduction

1.1 Cone-Beam Computed Tomography for interventional imaging

Imaging with X-ray computed tomography, abbreviated as CT imaging [6], revolutionized the clinical diagnostic routine 50 years ago with the introduction of diagnostic CT scanners. CT scanners precisely measure small contrasts of attenuation to X-rays that differentiate soft tissues from one another, thus allowing the identification of many pathologies. The image quality achieved by diagnostic CT is possible due to technical choices, such as putting the patient into a tunnel. This is generally incompatible with therapeutic use because the imaging device must not prevent access to the patient.
In the field of interventional radiology, image-guided procedures are practiced under X-ray guidance with interventional X-ray C-arms [227]. These systems have more recently been equipped with an alternative tomographic technology called Cone-Beam Computed Tomography (CBCT). Modern practice increasingly includes CBCT to plan interventional procedures, guide the placement of therapeutic devices and control the result of the intervention. These steps correspond to distinct imaging tasks: understand the anatomical context for planning, visualize the therapeutic device for guidance, and visualize the device and/or the anatomy for assessing the success of the intervention and the absence of immediate complications. Thanks to its high spatial resolution, CBCT, combined with iodinated contrast injection, is very efficient for vessel imaging. However, it remains limited in contrast resolution when it comes to the visualization of human tissues, all the more so in the presence of injected iodine and metal. The degradation is due to various sampling limitations inherent to the design of C-arms. Other limitations of standard CBCT are that it contributes markedly to the overall dose delivered to the patient during an intervention, and acquiring the CBCT data also slows down the procedure. So a simpler acquisition with fewer measurements and, thus, a lower X-ray dose is desired. Then the loss in the number of measurements needs to be compensated for reconstruction.
In this thesis, in collaboration with GE Healthcare, we aim to increase the usability and performance of CBCT through advanced reconstruction methods to improve the interventional clinical practice in terms of confidence and safety of the procedure as well as in terms of ease of use and speed of execution.

1.2 Goal of this thesis

From a mathematical point of view, computed tomography is an inverse problem [23] for which many tomographic reconstruction methods with solid theoretical guarantees have been developed [119, 154]. A tomographic reconstruction can typically be expressed as the minimization of a convex cost function, not necessarily differentiable, with respect to a large number of variables. Proximal algorithms [59] provide efficient ways to optimize such cost functions and have become well-established in tomographic reconstruction. In

particular, they offer great flexibility in formulating the objective function, allowing elaborate models of the data acquisition and the image. Therefore, these methods offer an appropriate framework for handling CBCT sampling limitations. Note that sampling limitations are not the only cause of artifacts in C-arm CBCT; several physical effects induce nonlinear measurement errors, such as beam hardening and scatter. Although these belong to the context of this work, they will not be addressed specifically.
In this thesis, we propose novel fixed point strategies for iterative reconstruction for CBCT. Our goal is to exploit the nature of the acquisition process and provide high-quality results for the specific imaging task while offering theoretically sound approaches that are fast enough processing for clinical use.
CBCT data acquisition models involve high-dimensional linear operators such as the forward projector. Reconstruction algorithms then rely on them and their adjoints. For various reasons (e.g., computation cost, convergence rate), the discretization of each operator may differ so that they are no longer adjoint of each other, leading to an *adjoint mismatch*. Indeed, the choice of the discretization scheme is critical to minimize information loss (e.g., preserving spatial resolution, avoiding noise amplification) and computation time. However, the diagnostic information and the time budget for retrieving it are highly dependent on the clinical context: there is no one-size-fits-all discretization. Therefore, we investigate, on the one hand, how to mitigate the impact of adjoint mismatch when it is used in iterative reconstruction and, on the other hand, we propose a new discretization that provides better trade-offs with respect to information loss, computation time and symmetry. Inherent limitations or voluntary reduction of CBCT sampling can only be overcome if the image information can be modeled from *a priori* and context-dependent knowledge of the clinical task. In particular, sampling artifacts related to metallic interventional devices such as percutaneous needles can be handled to improve the image quality with standard acquisition protocols or derive context-dependent alternative protocols. We present, depending on the context, either formal or data-driven models, through deep unfolding, that increase the image quality of reconstruction from sub-sampled CBCT acquisitions.

1.3 Outline

This section provides a reading guide to this thesis with a brief description of each chapter.

Chapter 2 introduces the medical context (section 2.1) and gives an overview of C-arm systems, their 2D and 3D (CBCT) imaging capabilities and their clinical use (section 2.2). Section 2.3 introduces the mathematical basis of CBCT, with the X-ray transform as a model of the image acquisition. Artifacts in 3D imaging are covered in section 2.4 and linked to the sampling limitations imposed by the interventional context and the physical defects in image formation. Chapter 3 exposes the problem of image reconstruction as an inverse problem. Three classes of inversion are discussed. First, we recall that an analytical formulation exists through the inversion of the 2D Radon transform, which is presented in section 3.1 together with its discretization. In section 3.2, we explain how iterative reconstruction methods have been used to deal with degraded and, most notably in this work, missing data. Models of the acquisition and *a priori* knowledge over the solution space are most efficiently optimized through proximal algorithms, presented in section 3.3. section 3.4 debates why iterative reconstruction methods have not completely replaced analytical methods. Impediments to the use of iterative reconstruction, notably parameterization and execution time, are discussed as well as how practition-

ers have introduced deviations in these schemes that violate their theoretical guarantees but alleviate draw. The situation of adjoint mismatch, where the adjoint of the projector is replaced by a surrogate operator in an optimization algorithm, is presented. Iterative reconstruction is also put into perspective with the new deep learning post-processing architectures that build on analytical methods. Despite their popularity, they lack mathematical characterization and can thus produce highly variable results. Section 3.5 presents how deep learning can be used directly in the reconstruction task from the projection data in a supervised way. Hybrid methods combining iterative methods with deep learning architectures, such as deep unfolding architectures, are introduced to provide more control over the output of the learned reconstruction than post-processing deep learning approaches.

Chapter 4 is dedicated to the proximal gradient algorithm with unmatched pairs of projector and backprojector for solving a penalized least-squares optimization problem. After describing the mismatched algorithm in section 4.2 we provide our main convergence results in section 4.3 before illustrating our results to two reconstruction problems in section 4.4. In section 4.5, we show that considering an adjoint mismatch in the proximal gradient algorithm can be recast as a problem of unmatched preconditioning where the metric used in the gradient step differs from the one used in the proximity step. Application to CT reconstruction in section 4.6 concludes the chapter.

Chapter 5 extends our study to a panel of primal-dual proximal methods, namely, the Condat-Vũ algorithm, the Combettes-Pesquet algorithm, and the Loris-Verhoeven algorithm for minimizing a penalized least-squares cost function which involves two linear operators. Sections 5.2, 5.3, and 5.4 detail the properties of the three primal-dual splitting algorithms in the presence of adjoint mismatch of one or two operators. Finally, we show two applications of our results for statistical and compressed sensing reconstructions in section 5.5.

Chapter 6 presents our magnification-driven interpolation framework for discretizing cone-beam forward projection (FP) and backward projection (BP). In section 6.2, we recall how the cone-beam geometry of C-arm systems can be described with "projection matrices" (not to be confused with projectors). Such matrices provide a continuous description of the homographic transforms between any plane of the volume and the detector plane. Discretization of FP and BP reduces to that of these transforms. Section 6.3 presents an existing scheme for discretizing a 1D magnification. In section 6.4, we extend this scheme to the case of 1D and then 2D homographies. Careful analysis of the sampling rate variation allows us to derive fast implementations. Section 6.5 discusses how magnification-driven interpolation sheds light on the current advantages and limitations of standard clinical interpolations and when and how it might better answer specific usage of flat-panel based CBCT. Finally, we evaluate, in section 6.6, our interpolation framework for both analytical reconstruction and iterative reconstruction using simulated and real CBCT data.

In Chapter 7, we consider the task of reconstructing objects such as percutaneous needles from acquisitions with reduced angular amplitude. First, in section 7.2, we propose a method for reconstructing needles superimposed to an anatomical background from a limited-angle acquisition for a 2D parallel geometry. Our method makes use of a decomposition strategy coupled with a directional regularization. It is then validated on simulated data. In section 7.3, we discuss how this method can be extended to a cone-beam geometry by changing the reconstruction grid through a rotation of the projection matrices to simplify the directional regularization. Preliminary results from two actual CBCT acquisitions for biopsy conclude this chapter.

Finally, in Chapter 8, we investigate ROI reconstruction from sub-sampled measurements. We start by introducing in section 8.2 the use of the Cauchy estimator for limiting artifacts from dense objects located outside of the ROI. We describe the Dual Block Forward Backward proximal algorithm and the reweighted strategy chosen to handle the non-convex Cauchy term. We also explain how it is unrolled to yield our U-RDBFB network. Finally, in section 8.3, we evaluate U-RDBFB with respect to several state-of-the-art methods.
We conclude and draw some perspectives in Chapter 9.

1.4 Contributions

Chapter 4 studies the stability of the proximal gradient algorithm in the presence of adjoint mismatch for solving a penalized least-squares problem in an arbitrary Hilbert space. Our contributions are:

- a characterization of the fixed points of the mismatched algorithm;
- conditions of convergence with new bounds on the gradient step size and the regularization parameters;
- a characterization of the distance from the generated fixed point of the mismatched algorithm to a minimizer of the original objective function;
- an extension of our results for analyzing preconditioned iterations of the proximal gradient algorithm when different preconditioners are used for the gradient and proximity steps; and
- a validation of these results on image reconstruction scenarios.

Chapter 5 extends the analysis of Chapter 4 to a panel of primal-dual proximal algorithms, which rely on forward-backward-(forward) splitting schemes, when an adjoint mismatch occurs. Our contributions are:

- a study of the properties of three main classes of primal-dual splitting algorithms in the presence of adjoint mismatch: the Condat-Vũ algorithm and its projected version proposed by Briceño-Arias and López, the Loris-Verhoeven algorithm, and the Combettes-Pesquet algorithm;
- convergence results for all the above algorithms, with an adjoint mismatch on different operators;
- a characterization of the resulting fixed points; and
- an illustration of our theoretical findings for CT reconstruction, with two types of regularization and noise modeling.

Chapter 6 presents our magnification-driven interpolation framework for discretizing the homographic transforms that arise in C-arm CBCT projection and backprojection. Our contributions are:

- novel pairs of projector-backprojector based on least squares spline approximation;
- a unified computation pipeline for balancing spatial resolution and noise properties for both analytical and iterative reconstruction;

- an alternative and more general formulation of state-of-the-art discretizations derived from geometrical considerations, of which the distance-driven model;
- approximate implementations of our operators to reduce complexity at near-optimal performance; and
- an assessment of the performance in terms of noise propagation and spatial resolution for low-order B-splines using both simulated and real data.

Chapter 7 introduces a methodology for reconstructing objects modeled as high-intensity segments, such as metallic needles, from limited-angle acquisitions in presence of the anatomical background. Our contributions are:

- identification of acquisition conditions which are favorable, or not, for segment reconstruction;
- a 2D directional total variation regularization (DTV) to capture segments;
- a decomposition strategy to allow several *a priori* directions to be considered at once as well as exclude the anatomical background;
- a 3D extension of DTV using a more accurate estimation of the directional gradients; and
- a numerical validation of the method using simulated and clinical data.

Chapter 8 tackles ROI reconstruction of soft tissues in the presence of high-density objects. Our contributions are:

- evidence that combining M-estimators and TV regularization in our optimization problem can effectively limit under-sampling artifacts associated with dense objects out of the reconstruction grid;
- an iterative optimization algorithm combining an instance of the dual block forward-backward algorithm with an iterative reweighted scheme;
- a neural network architecture, U-RDBFB, inspired by this algorithm, allowing supervised parameter learning and fast reconstruction on the GPU; and
- numerical experiments that show that U-RDBFB compares favorably with respect to other variational and deep learning methods, including other neural networks based on deep unfolding.

1.5 Publications

For articles with the $*$ symbol, authors are listed in alphabetical order, as is customary in mathematical journals.

Published journal articles:

- M. Savanier, C. Riddell, Y. Trousset, E. Chouzenoux, J.-C. Pesquet, "Magnification-driven B-spline interpolation for cone-beam projection and backprojection", Medical Physics, vol. 48, pp. 6339-6361, 2021.

- E. Chouzenoux, J.-C. Pesquet, C. Riddell, M. Savanier, Y. Trousset, "Convergence of Proximal Gradient Algorithm in the Presence of Adjoint Mismatch", Inverse Problems, vol. 37, pp. 065009, 2021*.

- E. Chouzenoux, J.-C. Pesquet, C. Riddell, M. Savanier, "Unmatched Preconditioning of the Proximal Gradient Algorithm", IEEE Signal Processing Letters, vol. 29, pp. 1122-1126, 2022*.

Conference proceedings:

- M. Savanier, C. Riddell, Y. Trousset, E. Chouzenoux, J.-C. Pesquet, "A Matched CBCT Projector-Backprojector Based on the Convolution of B-splines", 6th International Meeting on Image Formation in X-Ray Computed Tomography (CT Meeting 2020), Aug. 2020, virtual. (https://hal.archives-ouvertes.fr/hal-03140763)

- M. Savanier, E. Chouzenoux, J.-C. Pesquet, C. Riddell, Y. Trousset, "Proximal Gradient Algorithm in the Presence of Adjoint Mismatch", 28th European Signal Processing Conference (EUSIPCO 2020), Jan. 2021, virtual.

- M. Savanier, C. Riddell, Y. Trousset, E. Chouzenoux, J.-C. Pesquet, "A Decomposition Method for Directional Total Variation With Application to Needle Reconstruction in Interventional Imaging", 7th International Meeting on Image Formation in X-Ray Computed Tomography (CT Meeting 2022), Jun. 2022, Baltimore, MD, USA. (https://hal.archives-ouvertes.fr/hal-03770330)

Submitted journal articles:

- A. Contreras Tavarez, E. Chouzenoux, J.-C. Pesquet, and M. Savanier, "Convergence Results For Primal-Dual Algorithms in the Presence of Adjoint Mismatch" accepted to SIAM Journal on Imaging Sciences.
(https://hal.archives-ouvertes.fr/hal-03654126)

- M. Savanier, E. Chouzenoux, J.-C. Pesquet, and C. Riddell, "Deep Unfolding of the DBFB Algorithm with Application to ROI CT Imaging with Limited Angular Density" submitted to IEEE Transactions on Computational Imaging.

Submitted patents:

- C. Riddell, M. Savanier, E. Chouzenoux, J.-C. Pesquet, (2020) "Systems and methods for reprojection and backprojection via homographic resampling transform" (U.S. Patent No. US 20220036605A1)

Invited Talks:

- Convergence of PGA in the presence of an adjoint mismatch; Mini-Symposium: Convex and Non-Convex First-Order Optimization for Imaging and Data Analysis, SIAM Conference on Optimization, 20-23 July 2021, virtual.

- Convergence of PGA in the presence of an adjoint mismatch, Mini-Symposium: Convex optimization with errors in the adjoint, IFIP TC7 Conference on System Modelling and Optimization, 30 Aug - 3 Sept 2021, virtual.

- A Deep Unfolding Method for Limited CT Acquisitions, Mini-Symposium: Deep Learning for tomographic image reconstruction, SIAM Conference on Imaging Science, 21-25 March 2022, virtual.

2 | Interventional imaging with C-arm systems

This Ph.D. thesis is set in the clinical context of interventional radiology. Interventional radiology is a medical discipline that involves performing minimally invasive procedures to treat various pathologies using tools introduced through percutaneous and endovascular means. During such procedures, radiologists must acquire real-time images showing the position of the tools in the patient's anatomy. For so doing, C-arm systems have been introduced and deliver two-dimensional (2D) real-time X-ray projective imaging at various orientations. Since the end of the 1990s, rotating the C-arm around the patient allows for three-dimensional (3D) CBCT imaging.

2.1 Medical context

The clinical use of C-arms is established for a wide variety of endovascular and percutaneous procedures [202]. Imaging is used at different steps of these procedures. It allows for planning the procedure, guiding the interventional tools [120, 176], and assessing the outcome of the intervention. We now review some of these use cases.

2.1.1 Endovascular interventions

Angioplasty with stenting is the most widely performed endovascular procedure. Less frequent but important procedures are the prevention of aneurysm rupture [8] and the treatment of arterio-venous malformations. Through the intra-arterial injection of an iodinated contrast agent, the lumen of the vessels becomes visible under X-ray exposure. Global or selective imaging reveals the target of the procedure (origin of the anomaly of the blood flow) as well as the path to the target (arterial tree) (see Figure 2.1).

Balloon angioplasty (Figure 2.3) is the first and foremost application of interventional radiology, in particular for coronary artery diseases (CAD). Angioplasty uses a tiny balloon catheter inserted in a blood vessel's narrowed lumen (stenosis) to widen it and restore the blood flow. It is often combined with stent placement to decrease the risk of restenosis. For CAD, stents are coated with medication (drug-eluting stent) to prevent an inflammatory reaction of the tissue against the stent that would cause restenosis.
An aneurysm is a vascular pathology in which the arterial wall is abnormally dilated under the blood flow pressure, developing a "bubble"-like shape. A widening aneurysm may compress the surrounding nerves and brain tissues when located in the brain. Most importantly, a cerebral aneurysm may rupture, resulting in a lethal hemorrhagic stroke. To avoid rupture, embolization consists of performing a vascular occlusion so that the blood flow is no longer directed toward the aneurysm. Detachable metallic coils induct occlusion. Each coil is pushed through a catheter from a puncture site (usually in the groin) up to the brain target. Several coils are packed to fill the aneurysm densely. Alternatively, a new form of a stent with dense struts, called flow diverters, is placed to

Figure 2.1: MIP image with injected vessels

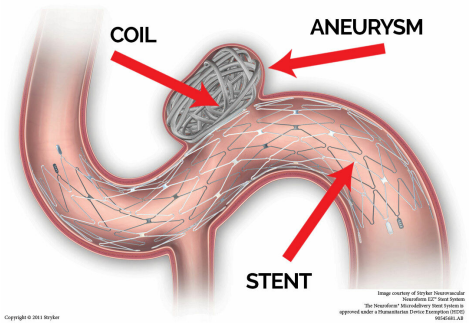


Figure 2.2: Coil embolization of an aneurysm

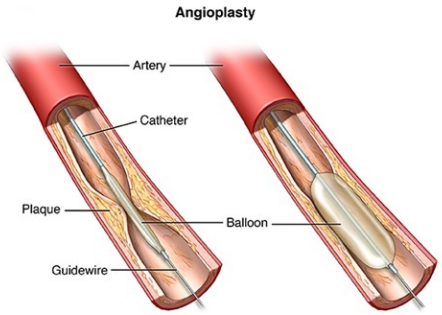
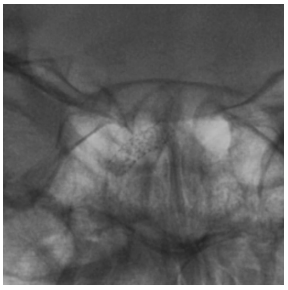


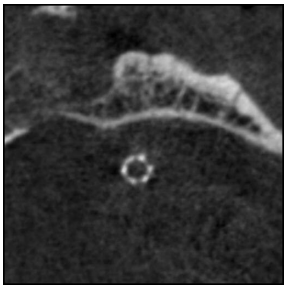
Figure 2.3: Balloon angioplasty

block the entrance of the aneurysm, and they strongly reduce the in-flow. A thrombosis then initiates within the aneurysm, further blocking any in-flux (Figure 2.2). When planning an aneurysm coil embolization in the brain, very high spatial resolution imaging is needed to assess whether there exist vessels close to the aneurysm that the device must not occlude. The endovascular treatment of aneurysms of the aorta is a more recent procedure that has gained ground as an alternative to vascular surgery. Instead of surgically replacing the diseased part of the aorta with a graft, a stent covered with tissue is introduced within the aorta to redirect the blood flow.
Liver trans-arterial embolization (TAE) [75] is used to treat malignant lesions in the liver. An embolic agent (microscopic beads or lipiodol) is injected to block the arterial supply of the lesion. Lipiodol can be mixed with chemotherapy drugs (trans-arterial chemo-embolization or TACE), and the beads can contain radioactivity (trans-arterial radio-embolization or TARE).
Guidance of catheters and intra-arterial devices is performed via real-time 2D imaging [173] called fluoroscopy. Fluoroscopy uses the minimum amount of the X-ray dose. Higher dose imaging is used for planning the gesture and assessing its success, either planar or volumetric. Planar imaging is dynamic, while volumetric imaging is static only. Figure 2.4 illustrates the superiority of 3D imaging over 2D imaging in terms of contrast for visualizing a stent within a brain vessel as well as its surroundings.

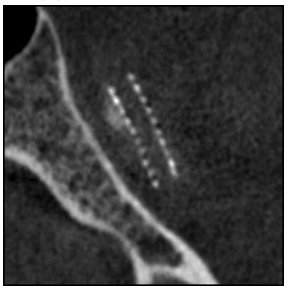
(a) 2D DSA imaging



(b) Zoom on 2D DSA



(c) Axial view from 3D imaging



(d) Coronal view from 3D imaging

Figure 2.4: Visualization of a deployed neurological stent with respect to the treatment region

2.1.2 Percutaneous interventions

The interventions are called percutaneous when the organ is reached directly rather than by its arterial supply. They include radiofrequency ablation [102], vertebroplasty [206] and biopsies [115].
Biopsies consist of sampling a suspected mass with a metallic percutaneous needle. The sample is then sent to pathology for further analysis.
Radiofrequency ablation relies on the guidance of a coaxial probe inserted into a malignant lesion to heat and destroy all tissues within a sphere of a few centimeters that must fully contain the lesion. These interventions begin with one CBCT volume to visualize the skeletal system, the third most common localization of metastases, and potential lesions in anatomical tissues. Figure 2.5 shows an axial slice of the abdomen where the liver, kidney, colon, and spinal cord are displayed.

Vertebroplasty is an effective pain treatment for compression fractures that develop in people with osteoporosis or bone tumors. To prevent further collapse of the vertebra, the physician injects bone cement into the pathological vertebral body to prevent its collapse (Figure 2.6). For intra-procedural guidance of the involved devices, the position of the patient must be optimized. A precise assessment of the device trajectory is critical to avoid puncturing the spinal cord. Radiologists usually rely on an image fusion of a pre-intervention CBCT with fluoroscopy imaging. The treatment is completed when 3D visualization assesses that the distribution of the cement fills the vertebra as planned.

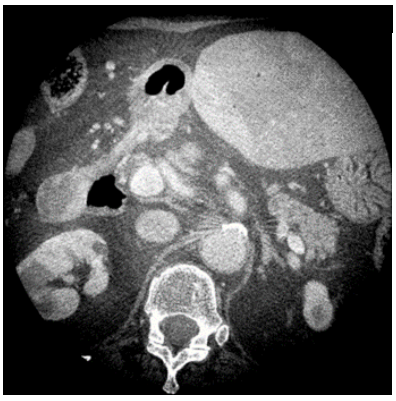


Figure 2.5: Axial slice of an abdomen

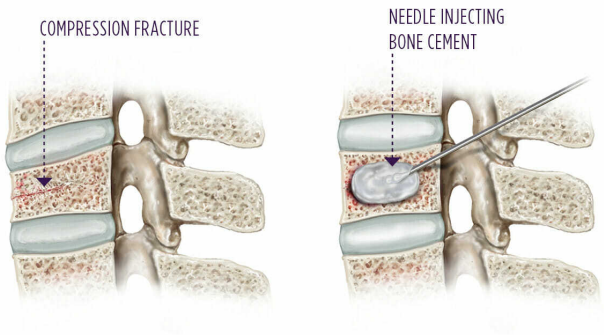


Figure 2.6: Vertebroplasty

2.2 Two-dimensional imaging

2.2.1 C-arm interventional system

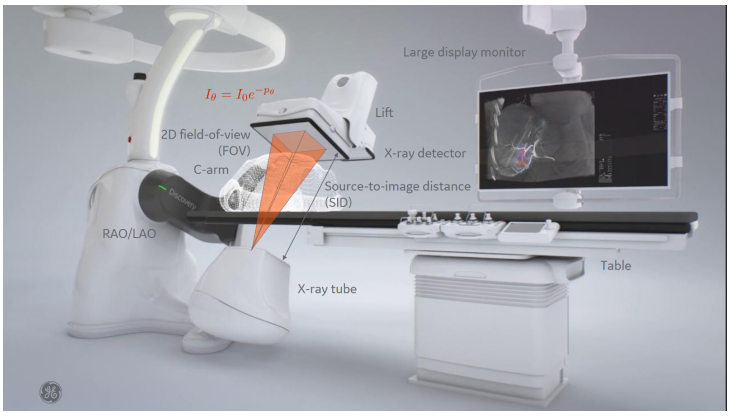


Figure 2.7: C-arm system (GE Healthcare IGS)

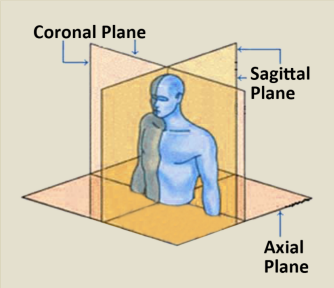


Figure 2.8: Three anatomical planes

A C-arm interventional system (Figure 2.7) is a real-time X-ray video camera composed of an open C-shaped arm holding an X-ray tube on one side and a digital flat-panel detector on the opposite side. The open design gives access to the patient, who lies on a bed table that is moved to put the anatomy of interest into the field of view (FOV) of the camera. The C-arm camera can rotate around three mechanical axes intersecting at a single point called the isocenter. Axes and directions are given with respect to the patient. One defines the axial, coronal, and sagittal planes as shown in Figure 2.8. Two anatomical angles are then considered:

- the CRAnial (CRA) or CAUdal (CAU) angle that describes rotation in the patient's sagittal plane;
- the Left Anterior Oblique (LAO) or Right Anterior Oblique (RAO) angle that describes rotation in the patient's axial plane.

The C-arm is primarily used for producing a stream of 2D images acquired with the lowest X-ray dose to guide an interventional device within the patient in real time. The X-ray tube design must guarantee the availability of such X-rays during interventions that can last several hours. A medical X-ray tube is a vacuum tube that contains a cathode filament and an anode rotating disk maintaining an electrical potential difference of up to 140 kilovolts (kV) called the peak voltage (kVp). Electrons are released at the cathode by thermal excitation and hit the anode. There they release their energy primarily as heat that the anode must evacuate. Only a small fraction of them is converted into X-ray photons. The number of emitted X-ray photons is linearly proportional to the filament's current intensity (mA) and the duration of the exposure (s). The energy of the emitted photons (expressed in keV) takes all the values up to the peak voltage. The number of emitted photons per energy is called the spectrum of the X-ray tube. From now on, we call $\Omega(E)$ the X-ray spectrum and $I_0\Omega(E)$ the intensity of the X-ray beam leaving the tube. I_0 expresses the linearity of the beam intensity with respect to the exposure current (expressed in mAs = mA $\times$ second). $\Omega(E)$ depends on the voltage peak and the anode material. The beam must pass through layers of aluminum and copper, called beam filters, before exiting the tube. Beam filters shape spectrum $\Omega(E)$

because the attenuation is always stronger at low energies than at higher ones, and more so with copper and aluminum. They preventively remove the emission of rays that do not contribute to the image. The X-ray spectrum then becomes narrower with higher mean energy; it is "hardened". Notably, because of beam filtration, intensity $I_0\Omega(E)$ must be increased to output as many X-rays with filtration as without. As a result, heat increases in the tube, reducing its availability. The safety of hours-long interventions does not allow as much pre-hardening as with diagnostic systems, where the tube can be left to cool down between patients, each requiring only a few seconds of exposure. The X-ray beam is shaped into a rectangular cone-beam using collimator blades of lead, defining the rectangular exposed area over the detector (FOV).

2.2.2 Interaction of X-ray with matter

When an individual X-ray photon passes through the body, there is a range of possible interactions [69], but for image formation, it is enough to consider two types of interactions. The first is the photoelectric effect, where the X-ray photon is absorbed by ejecting a photo-electron from an inner shell. The second is Compton scattering, where the energy of the incoming photon is only partially converted into electron ejection. A secondary photon, called a scattered photon, is emitted with a new direction and an energy loss proportional to the angle of the deviation.
The probability that an X-ray photon will interact with one material is a fundamental property of the material. It is described by the Beer-Lambert law and the linear attenuation coefficient μ per unit thickness of material (cm^{-1}). A narrow beam of intensity I_0 of photons of the same energy E will be attenuated by a material of attenuation $\mu(E)$ and thickness l such that the intensity I of the transmitted beam will be

$$I(E) = I_0 \exp\left(-\mu(E)l\right). \tag{2.1}$$

The attenuation of a material increases with the material density and mean atomic number; it is inversely proportional to the photon energy (E).
Let $\mu = \mu(r; E)$ be the distribution of the linear attenuation coefficient of the patient for an energy level E, where r stands for a position vector in $\mathbb{R}^3$. Let $s_X \in \mathbb{R}^3$ denote the location of the X-ray source and $c \in \mathbb{R}^3$ the location of a detector cell, the transmitted intensity measured by the cell is

$$I(c) = I_0 \int_0^{E_{\max}} \Omega(E) \exp\left(-\int_{s_X}^{c} \mu(l, E)dl\right)dE. \tag{2.2}$$

Thanks to beam filtering, the variations due to $\Omega(E)$ can be neglected, which allows the following simplification of (2.2):

$$I(c) = I_0 \exp\left(-\int_{s_X}^{c} \mu_\Omega(l)dl\right), \tag{2.3}$$

where attenuation $\mu_\Omega(l)$ is a mathematical approximation that replaces the real attenuation functions $\mu(l, E)$ for $E \in \Omega(E)$ and $l \in]0, +\infty[$.

2.2.3 Flat-panel X-ray detector

The detector embedded in a C-arm is a square panel made of crystalline cesium iodide with a thickness of a few millimeters. It is a scintillator that transforms each incoming X-ray photon into visible light (scintillation). The intensity of the scintillation is proportional to the photon energy. The light is then turned into electricity by one photodiode per detector cell. Each cell thus integrates the energy of the incident stream of photons over a square of 200×200 square microns. The read-out electronics perform the analog-to-digital conversion of the measurements recorded by the cells.
Unfortunately, there is not a one-size-fits-all design but three different panel sizes that target different medical practices. The heart fits in a 20×20 cm^2 panel, which is the smallest size. This allows for orienting the camera towards larger values of CRA/CAU and LAO/RAO angles. Imaging the abdomen and the lower limbs does not require large rotations, so a large panel of 40×40 cm^2 that fits more anatomy is preferred. The intermediate size of 30×30 cm^2 allows all medical specialties to share the same system and is also the preferred size for neurology (Table 2.1).

Anatomy	**Panel size (cm)**	**Number of cells**
Heart	20	1000
Head	30	1500
Abdomen, extremities	40	2000

Table 2.1: Panel sizes and number of cells for different anatomies

The detector is made of 2000 lines of 2000 cells (or pixels) of size 0.2 mm for a 40 cm panel. However, the read-out electronics cannot process more than 1000 samples per image. Therefore, when the FOV is greater than 20 cm, detector cells are binned, i.e., the outputs of neighboring cells are combined in a single reading. In return, spatial resolution is reduced since binning modifies the effective bin size.

2.2.4 Imaging modes and exposure

As already mentioned, the primary imaging mode is fluoroscopy, which uses the minimum amount of the X-ray dose and is used for real-time guidance. C-arm systems provide two higher-dose imaging modes (Table 2.2).
Digital Subtracted Angiography (DSA) takes a few high-dose shots of an organ, one shot before iodine injection, and around 10 to 20 images after injection, covering the time it takes for the contrast to fill the arteries and be flushed out by the blood flow into the venous system. Subtraction of the first contrast-free image from all other images yields a very high-resolution sequence of the lumen of tiny arteries. Organ perfusion appears as an image "blush" if the noise is low with respect to the contrast of iodine. Subtraction assumes the absence of patient motion and works best for identifying vessel occlusion and abnormal brain perfusion. For liver imaging, respiratory motion often precludes background removal by subtraction.
The "record" mode is a movie of a few seconds. It has two purposes. The historical one, called "cardiac record", records at 30 frames per second the propagation of contrast within the coronaries during a few heartbeats. It is used to localize stenosis or occlusion before the intervention or to assess the restoration of the blood flow after the placement of a stent. The second purpose is to acquire images at 50 frames per second (the highest frame rate of all imaging modes) during a spin, i.e., a 200° rotation from LAO to RAO

from which a tomographic reconstruction is performed. This feature is now called Flat-panel Cone Beam Computed Tomography, shortened into CBCT. The "record" mode is the mode that is used and discussed in this work.

Protocol	**Frames per second**	**Dose per frame**
Fluoroscopy	10	Low
DSA	7.5-15	High
Record	30-50	Medium

Table 2.2: 2D imaging protocols

The dose received by the patient is proportional to the size of the FOV. The maximum FOV is as large as the detector, but the physician's focus is, most of the time, the region of intervention, which is small. To reduce unnecessary peripheral doses and only image a region-of-interest (ROI) [236], the X-ray beam is collimated in both horizontal and vertical directions. Depending on the size of the resulting FOV, binning of the detector's cells may not be necessary.

Because X-ray production and interactions with matter are random processes, the transmission measurements are random variables $\bar{I}(c)$ that follow a Poisson distribution with mean (and thus variance) equal to $I(c)$. This has two consequences: (i) the signal-to-noise ratio at a detector cell is equal to $\sqrt{I(c)}$ and (ii), to maintain the same measurement while the attenuation increases linearly, intensity I_0 must be increased exponentially. The dose received by the patient is defined as the absorption of X-ray energy per unit mass of matter. The dose is approximately linear with respect to I_0. On a C-arm, the X-ray exposure is automatic. It uses the equivalent patient thickness l_{EPT} defined such that $\int_{s_X}^{c} \mu_\Omega(l)dl = \bar{\mu} l_{\text{EPT}}$, with $\bar{\mu}$ the attenuation value of water or plexiglass measured by the C-arm (e.g., 0.15 cm^{-1}). An average value of l_{EPT} over the field of view is estimated on the last image to adjust the next image's filtration, mAs and kVp, called X-ray techniques. It sets a balance between image quality and dose. Value for l_{EPT} can go up to 45 cm in cardiac imaging. It is around 15 cm for the head.

2.3 Three-dimensional imaging

Through the acquisition of a spin with the record imaging mode, C-arm systems can generate volumetric images that offer high spatial resolution but only moderate contrast resolution. A reconstruction algorithm then computes function μ_Ω of the patient from the set of acquired measurements called projections.

2.3.1 Radon and X-ray transforms

The mathematical basis of CT is best understood in the parallel-beam geometry (see Figure 2.9). The parallel-beam acquisition geometry assumes that incoming rays are parallel to each other and orthogonal to the detector. They are parameterized by angle θ with respect to the x-axis, so that the rays are oriented along vector $\boldsymbol{\theta} = (\cos\theta, \sin\theta)^\top$. The detector axis is orthogonal to the incoming rays, hence it is oriented along vector $\boldsymbol{\theta}^\perp = (-\sin\theta, \cos\theta)^\top$. The projection view at angle θ of a two-dimensional image $f \in L_1(\mathbb{R}^2)$, the space measurable, integrable functions from $\mathbb{R}^2$ to $\mathbb{R}^2$, is a function

$p : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ defined by the Radon transform $\mathcal{R}$ as

$$p(\theta, s) = \mathcal{R}[f](\theta, s) = \mathcal{R}_\theta[f](s) = \int_{L_{\theta,s}} f(x, y)dl = \int_{-\infty}^{+\infty} f(s\boldsymbol{\theta} + t\boldsymbol{\theta}^\perp)dt$$
$$= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x, y)\delta(x\cos\theta + y\sin\theta - s)dxdy \tag{2.4}$$

where $\delta(\cdot)$ is the Dirac distribution. The Radon transform has the following properties:

- If $f = \sum_{i=1}^{I} c_i f_i$ then $\mathcal{R}[f] = \sum_{i=1}^{I} c_i \mathcal{R}[f_i]$.
- $p(\theta, s)$ is 2π-periodic in θ i.e. $p(\theta, s) = p(\theta + 2\pi, s)$.
- $p(\theta, s)$ is symmetric in θ with period π i.e. $p(\theta, s) = p(\theta \pm \pi, -s)$.

The Radon transform and the X-ray transform are equivalent descriptions of a 2D parallel tomographic setup. They are invertible. CT images are thus computed according to the inverse transform.
The cone-beam geometry induced by the C-arm flat-panel detector is 3D and samples the X-ray transform of a volume. To describe this geometry, a few parameters must be added: the X-ray source is located at point $\boldsymbol{S_m}$, which is at distance α from the center of the flat detector and at distance t_z from the center of rotation O (Figure 2.9). All integration lines cross point $\boldsymbol{S_m}$. The orthogonal projection of the source over the detector defines an angle θ with respect to the x-axis. The detector plane is orthogonal to $\boldsymbol{\theta}$. Coordinate $s = 0$ is at the focus point $\boldsymbol{S_m}$ and lies on the optical axis. A point $(x, y, z) \in \mathbb{R}^3$ projects onto the detector plane Π_m for the position $\boldsymbol{S_m}$ at coordinate $(u, v) \in \mathbb{R}^2$, according to the following relation in homogeneous coordinates $(su, sv, s) \in \mathbb{R}^3$

$$\begin{pmatrix} su \\ sv \\ s \end{pmatrix} = \underbrace{\begin{pmatrix} -\alpha\sin\theta & \alpha\cos\theta & 0 & 0 \\ 0 & 0 & \alpha & 0 \\ -\sin\theta & \cos\theta & 0 & t_z \end{pmatrix}}_{P_\theta} \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix}. \tag{2.5}$$

Matrix P_θ is called the projection matrix at view angle θ. Note that P_θ is defined up to a constant scaling. There is one projection matrix per position of the pair X-ray source/detector. Cone-beam projections can be expressed using the X-ray transform $\widehat{\mathcal{R}}$ as

$$p_\theta(u, v) = \widehat{\mathcal{R}}_\theta[f](u, v) = \int_{L(u,v,\theta)} f(x, y, z)dl = \int f(S_m + t\mathbf{l}(u, v))dt \tag{2.6}$$

where

$$\mathbf{l}(u, v) = \frac{1}{\sqrt{\alpha^2 + u^2 + v^2}}(\alpha\boldsymbol{\theta} + u\boldsymbol{\theta}^\perp + v\boldsymbol{\xi}). \tag{2.7}$$

The cone-beam projection of plane $z = 0$ falls over line $v = 0$ of the detector plane. The relation between $(x, y, z = 0)$ and $(u, v = 0)$ is called fan-beam geometry. The fan-beam geometry for $\{p_\theta \mid \theta \in [0, 2\pi]\}$ samples the 2D Radon transform exactly twice. When $(\alpha, t_z) \longrightarrow (+\infty, +\infty)$, cone-beam, fan-beam, and parallel geometries are equivalent.

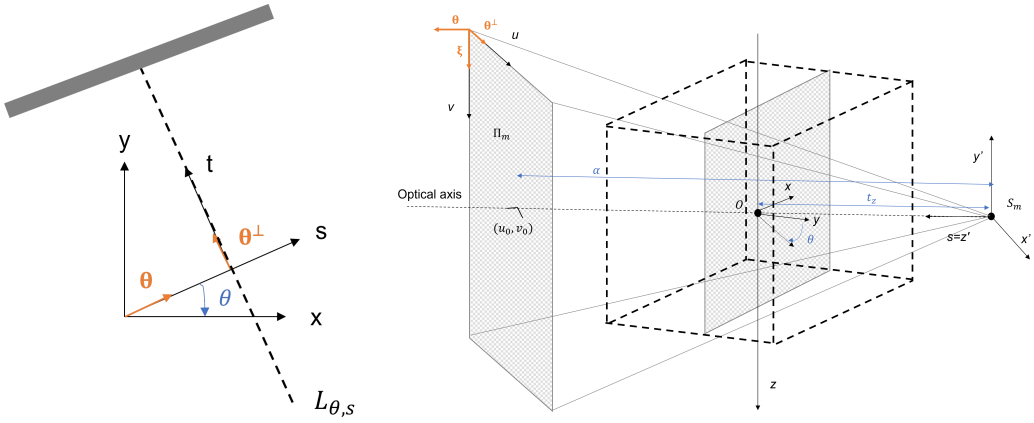


Figure 2.9: Coordinate systems in parallel geometry (left) and cone-beam geometry (right)

2.3.2 Log transform

We set the logarithm of (2.3) as

$$p_\theta(c) = \log\left(\frac{I_0}{I(c)}\right) = \log(I_0) - \log(I(c)). \tag{2.8}$$

Function p_θ is called the log-transform of the data. It converts intensities into densities which are the X-ray transforms of the attenuation map. Applying a tomographic reconstruction algorithm to p_θ will yield an estimate of attenuation μ_Ω.
Radiologists typically measure attenuation coefficients in Hounsfield units (HU). The value H_{tissue} of the attenuation coefficient of a particular tissue in HU is commonly scaled relative to the value of the attenuation coefficient μ_{water} of water, i.e.,

$$H_{\text{tissue}} = \frac{\mu_{\text{tissue}} - \mu_{\text{water}}}{\mu_{\text{water}}} \times 1000. \tag{2.9}$$

Thus, with the shifting in (2.9), the attenuation coefficient of water is 0 shifted HU (sHU), and a substance with an attenuation coefficient of 1000 sHU attenuates X-rays twice as much as water. An estimate of the Hounsfield units for materials of clinical interest, computed for $E_0 = 100$ keV, is given in Table 2.3. A large variety of intensity values can be encountered in interventional radiology imaging among metallic and anatomical objects. Note that the contrast resolution expected from diagnostic CT scanners is a few sHU to allow fine differentiation of brain white and gray matter.

Substance	HU
Air	-1000
Fat	-120 to -90
White matter	+20 to +30
Grey matter	+37 to +45
Cortical bone	+1800 to +1900
Cancelous bone	+300 to +400
Clotted blood	+50 to +75
Water	0
Lung	-700 to -600
Kidney	+20 to 45
Muscle	+35 to +55
Titanium	+6200
Silver	+17000
Steel	+20000
Gold and brass	+30000

Table 2.3: HU for different materials

2.4 Noise and artifacts in C-arm CBCT

Measuring the X-ray transform of a patient with X-rays mounted on a C-arm has limitations in terms of noise and artifacts. Noise in the measurements is propagated to the reconstructed image and decreases contrast resolution. Noise decreases as the intensity of the X-ray beam increases. We shall restrict the term "artifacts" to image defects that do not disappear, whatever the intensity of the X-ray beam.

2.4.1 Noise

As already mentioned, the measured X-ray intensity $I(c)$ is akin to a number of photons, i.e., a positive integer, realization of a Poisson process, and such that the signal-to-noise ratio is proportional to $\sqrt{I_0}$. For CBCT, values I_0 and $I(c)$ must also be high enough for (2.8) to be numerically accurate. The log transform is thus a non-linear transform and yields a potential infinite amplification of the noise from the intensity images.

2.4.2 Sampling

Lateral truncation:

Lateral truncation happens in most examinations, either voluntarily because the focus is an ROI, or because the maximal FOV of the detector does not allow for entirely exposing the patient. Truncation without adequate processing introduces strong cupping that alters the visualization of low contrasts in the reconstructed ROI.

Angular sub-sampling:

There are two main types of artifacts related to the angular sampling of the acquisition trajectory (density and amplitude).
First, ideally, CBCT imaging requires a 360° rotation as found on diagnostic CT scanners. Standard C-arms have a rotation of maximal amplitude along LAO/RAO over 200°, with

CRA/CAU = 0 and SID constant. The rotation axis is thus parallel to the patient bed. The rotation over 200° is called a *short-scan rotation*. It is the minimum coverage required by tomography. Below this value, one speaks of limited-angle tomography.
Second, the C-arm gantry is used at maximal rotation speed ($40°/s$) for acquisition with an injection of contrast to minimize the volume of injected contrast, as contrast must fill the imaged lumen for the whole duration of the acquisition. Given a detector frame rate of 50 fps, this 5-second acquisition delivers 250 views. On the contrary, the speed is kept slow for a non-angiographic exam ($16°/s$) to maximize the number of views within the scan and increase contrast resolution to differentiate human soft tissues better. Such acquisitions deliver 600 views. This number is sufficient, except in the presence of metallic devices (see Table 2.3). To reduce storage space and reconstruction time, instead of using the $1000 \times 1000 \times 0.4$ mm projections read by the detector, the projections are further binned into $500 \times 500 \times 0.8$ mm from which reconstruction of 512×512 images is performed. Angular sub-sampling of the densest objects (metal, bone) is responsible for streaks (Figure 2.10) whose intensity is higher than the contrast of soft tissues, leading to a loss in contrast detection.

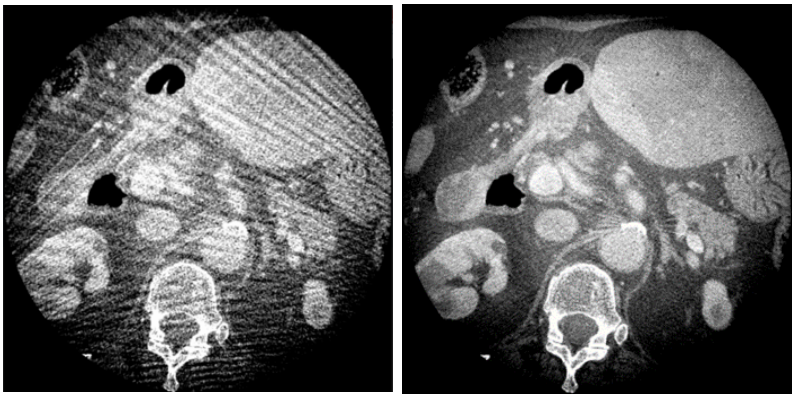


Figure 2.10: sub-sampling streaks due to a low angular density

Cone-beam artifacts:

C-arm circular source trajectory with cone-beam geometry does not allow an exact reconstruction; only the volume planes crossed by the X-ray source trajectory are fully sampled (cf. Tuy's conditions [212]). The circular trajectory fully samples the plane containing the source rotation only, i.e., the central slice of the volume. In contrast, all other slices are degraded by conic artifacts that become increasingly apparent as the slice is away from the central one. A helical trajectory combining bed translation with a continuous rotation of the imaging chain makes diagnostic systems immune from this issue. However, such continuous rotation is impossible with an open system such as a C-arm.

2.4.3 Physics

Beam Hardening:

Beam filtration is also called beam pre-hardening because as the beam penetrates the patient, the hardening effect continues, more so if the emitted spectrum is too "soft". For a single material type (e.g., water), one easily measures that the linearity between the log-transform of the measured intensity and the material thickness no longer holds and becomes closer to a square root.
The impact of beam hardening increases with the material attenuation and is specific to the tube spectrum; most tissues are considered equivalent to water, while bone, iodine, and metals will have distinct behaviors. Correction of beam hardening must thus be performed per material, with the classical case of skull beam-hardening correction introduced by Joseph [118]. After a first-pass reconstruction, the skull is segmented by thresholding. The bone contribution is estimated per cell and corrected by inverting (i.e., "squaring") the calibrated skull curve. A correction term is reconstructed and summed to the initial reconstruction.

Scatter:

As previously mentioned, scattered photons are the result of Compton interactions. If a unique detector cell is exposed, it will accumulate photons from the X-ray source. All scattered photons having other random directions will not hit the cell. If several detector cells are exposed simultaneously, then scattered photons arising from one cell measurement process will be read into another cell. Therefore, the greater the acquisition field of view, the higher the scatter cross-contamination. Image artifacts due to the scatter are low-frequency cupping, reduced contrast, and dark streaks between dense objects [200]. Scatter rejection is achieved through an anti-scatter grid placed upon the detector. It is made of lead or titanium holes that absorb all photons not coming from the X-ray source. However, the grid cannot be too strongly focused because C-arms must allow a variable geometry with a lift that moves the detector closer to or away from the patient. Reducing the field of view decreases the scattering effect. With lateral truncation, one exchanges scatter artifacts for truncation artifacts. Reduction of the number of slices acquired per rotation is very efficient but requires multiple rotations, which takes a prohibitive time on a C-arm and is thus not proposed. Scatter contamination is the most important reason why CBCT systems have a lower contrast resolution than diagnostic CT scanners.

2.4.4 Metal artifacts:

Metal artifacts are problematic because they originate from several previously described sources of artifacts. Medical X-ray tubes are not designed to image metal whose very high attenuation requires unacceptable X-ray dose levels. The detector measurements behind metal are thus small integers in (2.8), and the estimation of the log-transform becomes unreliable and amplifies these specific measurements dramatically, yielding bright streaks all over the reconstructed image. Digital filtering is insufficient because beam hardening and scatter also affect the data. Finally, the streaks due to the angular sampling are proportional to the metal density, which is itself easily 10 times the density of the anatomical background. These streaks are thus 10 times stronger than usual and also than the contrasts of the anatomical background. The tissue contrast is hidden under

the metal streaks.

For example, the evaluation of the placement of a stent in contact with a coil becomes impossible. A post-processing algorithm can reduce these artifacts and restore some clinical information, as seen on Figure 2.11. Like the skull beam hardening algorithm, the metal is segmented by thresholding after a first-pass reconstruction. The metal contribution is estimated per cell and removed from the data to reconstruct a metal-free volume that is then fused with the segmented metal.

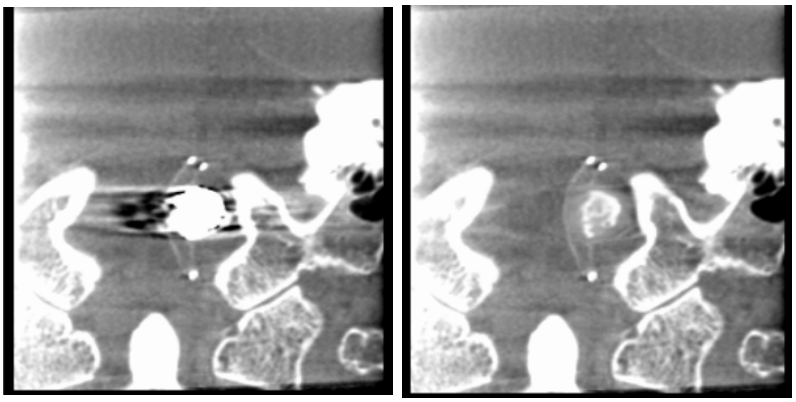


Figure 2.11: Coil in contact with a stent with (left) and without (right) artifacts.

2.5 Conclusion

In summary, flat-panel C-arm systems provide the interventional radiologist with real-time fluoroscopy, digital subtraction angiography, and recording modes such that computed tomography becomes available. The design of C-arm systems is tailored to therapeutic intervention guided by 2D imaging.
CBCT brings volumetric imaging to the interventional suite. The design constraints of C-arms provide a higher spatial resolution but a lower contrast resolution than diagnostic CT. Therefore, CBCT use is tailored to specific needs during the procedure, where lesser tissue differentiation is acceptable given that the 3D image is made available in the operating room. Examples are the planning of a gesture and the assessment of a tool position that requires high spatial resolution but limited contrast resolution. Piling up pre- and post- corrections for sub-sampling and physical degradations has increased soft-tissue contrast resolution on the newer systems to assess bleeding or ischemia in the brain or tissue perfusion in the liver.
Guidance, however, suggests the need for more than one CBCT, as in the case of percutaneous needle insertion. This means an increase in dose to the patient and disruption of the procedure as a common position must be found for the physician to work comfortably and for the system to rotate without collision over 200°.
This thesis is thus placed in the context of ROI reconstruction, for which we will focus on handling sub-sampling effects inherent to C-arm CBCT but also tackle those that arise from decreasing the rotation amplitude of the acquisition to allow better-suited acquisition protocols.

3 | CT reconstruction methods

This chapter provides an overview of the methods that have been proposed for reconstructing a function, such as a patient attenuation map from a set of its integrals over rays. Three classes of methods are detailed. First, section 3.1 presents analytical reconstruction (AR) methods that directly discretize the continuous inverse of the Radon transform. Then, section 3.2 and section 3.3 introduce the framework of iterative reconstruction (IR), which defines a discrete projection model inverted as a classical inverse problem. In contrast to analytical methods, iterative methods allow for making explicit assumptions about noise/attenuation map properties. We thus decouple the problem formulation (section 3.2) from the description of the algorithm used to solve for optimization (section 3.3). Finally, section 3.5 reviews recent deep learning methods, which can leverage large datasets and estimate the attenuation in a reduced time compared to iterative methods. We comment on the robustness of these methods to the artifacts mentioned in the previous chapter and their practical difficulties.

3.1 Analytical reconstruction

3.1.1 Inverse Radon transform

In 2D parallel geometry, the continuous inverse of the Radon transform, $\mathcal{R}$ in (2.4), relies on its adjoint denoted $\mathcal{R}^*$ and called backprojector. Let $p(\theta, s)$ be the projection data of $\mu \in L_1(\mathbb{R}^2)$. Backprojection of $p(\theta, s)$ at position (x, y) is an integration over all $\theta \in [0, \pi[$ as

$$\mathcal{R}^*[p_\theta(s)](x,y) = \int_\theta p_\theta(x\cos\theta + y\sin\theta)d\theta = \int_\theta p_\theta(r)\delta(r - x\cos\theta - y\sin\theta)d\theta. \quad (3.1)$$

Given (2.4), we get the following relationship between the backprojected image $\mathcal{R}^*[p_\theta(s)](x, y)$ and the reference object $\mu(x, y)$:

$$\mathcal{R}^*[p_\theta(s)](x,y) = \mu(x,y) * \frac{1}{\sqrt{x^2+y^2}}, \quad (3.2)$$

where the convolution product is $(f * g)(x, y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x - i, y - j)g(i, j)didj$ for functions f and g in $L^1(\mathbb{R}^2)$. (3.2) shows that the backprojection of the projection data generates a blurred version of $\mu(x, y)$.
One way to derive a closed-form expression for the inverse Radon transform is via the Fourier Slice Theorem:

$$\widehat{p}_\theta(k_s) = \mathcal{F}_1[p_\theta(s)](k_s) = \int_{-\infty}^{+\infty} p_\theta(s)e^{-j2\pi k_s s}ds = \mathcal{F}_2[\mu](k_s\cos\theta, k_s\sin\theta), \quad (3.3)$$

where k_s is the Fourier domain counterpart to the image domain spatial variables s.
The interpretation is that $(u, v) = (k_s\cos\theta, k_s\sin\theta)$ for $\theta \in [0, \pi[$ and $k_s \in \mathbb{R}$ specifies a line in the 2D Fourier space rotated by θ relative to the positive u-axis. This corresponds

to the s-axis in the image space. Thus, the 1D Fourier transform of a projection is equivalent to the corresponding slice/line through the 2D Fourier transform.
The Fourier Slice Theorem allows us to derive the inverse of the Radon transform as

$$\mu(x,y) = \mathcal{R}^*[p_\theta(s) * F(s)] = \int_0^{2\pi} p'(\theta, x\cos\theta + y\sin\theta)d\theta, \tag{3.4}$$

where $*$ now denotes a 1D convolution, $p'(\theta,s) = \int_{-\infty}^{+\infty} \widehat{p_\theta}(k_s)|k_s|e^{j\pi k_s s}dk_s$ and $F(s)$ is a filter of one variable, called ramp filter, such that $\mathcal{F}_1[F](k_s) = |k_s|$. This approach is the filter-then-backproject method (FBP) [73]. Note that (3.4) can be rewritten with a 2D Fourier transform (the 2D counterpart of the ramp filter is called the cone filter), then becoming the backproject-then-filter method.

3.1.2 Discretization

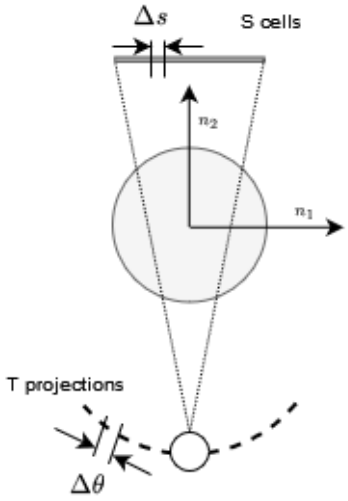


Figure 3.1: A schematic illustration of a scanning geometry with a flat panel detector and an X-ray source moving along an arc. The following parameters are shown: the number of projections T, the number of detector cells S, the angular sampling $\Delta\theta$, the detector sampling rate Δs and the spatial indexes of the image grid n_1 and n_2.

We now suppose that we have T projection views and S detector cells (see Figure 3.1). Implementation of FBP can be done either by direct discrete convolution or in the Fourier domain by 1D Fast Fourier Transform (FFT) operations. The discrete projection data, sampled with a distance Δs between the detector elements, indexed by $s \in \{1, \cdots, S\}$, represents a band-limited function with a highest frequency of $\Delta s/2 = k_s^{\max}$. Hence, in FBP, we can replace $|k_s|$ with the function

$$H(k_s) = \begin{cases} |k_s| & \text{for } k_s \leqslant k_s^{\max} \\ 0 & \text{else} \end{cases}, \tag{3.5}$$

which corresponds to the discrete impulse response (Figure 3.2)

$$(\forall l \in \mathbb{R}) \qquad h(l) = \frac{1}{\Delta_s^2} \begin{cases} 1/4 & \text{if } l = 0 \\ 0 & \text{if } l \neq 0,\ l \text{ is even} \\ -1/(\pi^2 l^2) & l \text{ is odd} \end{cases}. \tag{3.6}$$

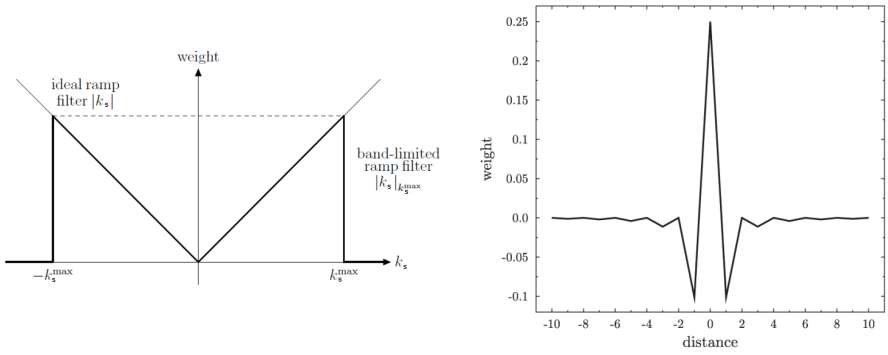


Figure 3.2: Band-limited ramp filter. Left: frequency response. Right: impulse response

As Δs tends to zero, higher frequencies are sampled, which also means that the ramp filter amplifies frequency up to $+\infty$. In practice, additional apodization of the ramp filter is used to reduce Gibbs oscillations and enable the optimized depiction of soft-tissue or high-contrast structures (e.g., cosine, Hanning, Hamming, Shepp-Logan, Butterworth). These functions provide a continuous transition between the ramp amplification and the band-limited windowing.
After filtering, FBP relies on the discretization of the continuous backprojection operation $\mathcal{R}^*$. The most straightforward choice for discretizing this operation is to directly discretize (3.1). Let $\Delta\theta$ be the rotation interval between subsequent views, indexed by $t \in \{1, \cdots, T\}$. The angle θ is replaced by a discrete set of angles $\{\theta_t = (t-1)\pi/T : 1 \leqslant t \leqslant T\}$ and the integral is replaced by a summation. Let $n_1, n_2 \in \mathbb{N}^2$ be the indexes over a reconstruction grid along the horizontal and vertical directions. Discretizing (3.1) yields the definition of the discrete backprojection B at location $(x_{n_1}, y_{n_2}) \in \mathbb{R}^2$ such that

$$B(x_{n_1}, y_{n_2}) = \sum_{t=1}^{T} (p_{\theta_t} * h)(x_{n_1} \cos\theta_t + y_{n_2} \sin\theta_t)\Delta\theta. \tag{3.7}$$

For a given angle θ_t, the discrete positions $s\Delta s$ generally do not coincide with the discrete values $x_{n_1} \cos\theta_t + y_{n_2} \sin\theta_t$. A discrete representation of the continuous projection $p(s, \theta)$ is then assumed to perform the interpolation. Here we see that backprojection at a given pixel (x_{n_1}, y_{n_2}) is independent of neighboring pixels. Depending on the representation of p, different variants of B can be obtained.

Analytical reconstruction and angular sub-sampling:

The Fourier transform of the attenuation function $\mathcal{F}_2[\mu](k_x, k_y)$ is known along radial lines with uniform sampling. The density of points in the angular direction is non-constant: it becomes smaller when the radius increases. In other words, more projections are needed for bigger images. The impulse response of the band-limited ramp filter (Figure 3.2) presents negative components around the central positive peaks. When the number of projections is infinite, the positive and negative contributions compensate each other during the backprojection step of FBP. When the number of projections decreases, non-compensation of the positive and negative contributions results in non-localized streak artifacts. The rule of thumb in clinical practice is that sampling artifacts become an issue - the acquisition being qualified as sub-sampled - when the number of projections for reconstructing an $N \times N$ image is less than N.

Short-scan acquisition:

As already mentioned, the fan-beam geometry with a 2π rotation of the source samples the parallel geometry exactly twice. On the contrary, a π rotation does not provide a sufficient sampling. The shortest scan that samples the parallel geometry at least once per sample is equal to π plus the fan angle, where the fan angle refers to the aperture covered by the detector. In this configuration, the redundancy is not uniform, with some line integrals being measured twice and others once only. In [167], Parker proposed a function to weight the short-scan fan-beam projection data before ramp-filtering that allows the same FBP-type reconstruction for the short-scan case. Parker's weights create a smooth window that sets at zero most of the redundant projection columns, does not alter most non-redundant projection columns, and smoothly applies a weight between 0 and 1 on a few projection columns. This weighting ensures that the relevant conjugate rays have weights that sum to unity and is now widely used.

ROI Imaging:

Fourier methods assume non-truncated data. In the case of full angular coverage but with all projections laterally truncated on both sides, the objective is to reconstruct the ROI that is visible (nontruncated) by all projections, despite the contamination due to external parts of the object. This problem is called the interior problem [53]. Although this problem has an infinity of solutions, the solution becomes unique with little extra *a priori* information [181]. In practice, an image of the ROI with FBP is obtained by extrapolating the projections prior to ramp filtering [132]. When the projections are not fully truncated, i.e., containing specific transverse truncations, it is no longer an interior problem, as one shows that a unique solution exists, at least in certain situations. They are computed with the alternative algorithm of differentiated backprojection with Hilbert Transform (DBP-HT) [157]. In this thesis, we will refer to ROI imaging the setting of reconstruction a ROI from projections that are at least partially truncated.

Extension to cone-beam geometry with a circular orbit:

As mentioned in Chapter 2 (section 2.4), Tuy's conditions are not satisfied for the cone-beam circular acquisition case, so no exact inversion formula is available. The Feldkamp-Davis-Kress (FDK) method [90] performs an approximate inversion of the X-ray transform by applying an FBP-type reconstruction thanks to weighting steps prior- and post-filtering. These weightings yield an exact reconstruction when the object is constant along the axis of rotation. A short formula for FDK can be found in [91].

To conclude, AR methods combine projection filtering and one (weighted) backprojection, leading to very fast implementations. However, they linearly transfer all defects and inconsistencies from the data into the reconstructed image.

3.2 Modeling in reconstruction

AR does not allow for using any *a priori* information on the volume and the nature of the acquisition. A point in space is reconstructed independently from the others, and the performance of analytical methods simply reflects the quality of the data: image quality is degraded when the sampling is not dense and regularly spaced; the ramp filer amplifies noise. Model-based reconstruction offers the flexibility to model the acquisition setup (high angular sampling but high noise, or low noise but low angular sampling) and to include *a priori* knowledge over the solution.

3.2.1 Geometric modeling

The geometry of the acquisition system can be modeled by a matrix called projector H, where each row captures one measurement that is a summing process over the volume elements. Matrix H of the ideal tomographic system is the discretization of the continuous X-ray transform appearing in AR. For real systems, H exclusively includes measurements delivered by the acquisition, thus modeling sub-sampling. Its columns model a field of view as a Cartesian grid. The grid's density and extension shape the properties of the reconstruction task. The (forward) projection operation is denoted:

$$H\overline{x} = y, \tag{3.8}$$

where $y \in \mathbb{R}^C$ ($C = T \times S$) and $\overline{x} \in \mathbb{R}^N$ ($N = N_1 \times N_2$) are obtained by stacking respectively the values of matrix $p[c_1, c_2]$ and of matrix $\mu[n_1, n_2]$ and $H \in \mathbb{R}^{N \times C}$.
A typical matrix of a tomographic system is huge. It is thus never stored but computed on-the-fly. Each matrix value contains the contribution of one point of the grid to one measurement. Because integration happens over lines, this value is generally 0, so H is sparse. Point contribution is derived from geometric and interpolation considerations, but physics constrains the result to interval $[0, 1]$, so H is stochastic. Sub-sampling makes the solution non-unique. The spectrum of eigenvalues of integration operators decreases to 0, and inversion is thus ill-posed.

Given these characteristics, reconstruction methods using standalone geometric modeling traditionally trade (3.8) for the following system of linear equations called the normal equations:

$$H^\top H\hat{x} = H^\top y, \tag{3.9}$$

where $\hat{x} \in \mathbb{R}^N$ is an estimation of the attenuation map.
ART [234] was historically proposed in the early days of CT technology during the 1970s [6] for solving (3.9). This method is an instance of the Projection Onto Convex Sets (POCS) method. ART is a fully sequential row-action method: it updates the volume estimation using one row of H at a time. The method converges to the minimum-norm solution

$$\underset{\substack{x \in \mathbb{R}^N \\ Hx = y}}{\operatorname{argmin}} \ \|x\|_2. \tag{3.10}$$

Extensions involve changes in how each row is handled in an iteration (order and number of rows) [204]. Block-row schemes (Cimmino, SART [7], DROP [40], BiCav [41]) converge to a fixed point that is not a (weighted) least-squares solution. There are also block-columns sequential methods [82, 189] such as SOR, column-Cimmino, and column-BiCav. These methods converge to a least-squares solution, though not necessarily the

one with the minimum norm. Most methods have been introduced as a faster alternative to standard gradient descent techniques. Algebraic methods are semi-convergent; the number of iterations is tuned to control the amount of noise in the solution.

In tomography, backprojector $H^\top$ is not the inverse of H while the FBP algorithm provides a linear operator that approximates it. It is therefore interesting to consider rewriting system (3.8) not with the normal equations but through approximate inversion with an adequately designed operator $H^\ddagger \in \mathbb{R}^{C\times N}$:

$$H^\ddagger H\hat{x} = H^\ddagger y. \tag{3.11}$$

The reason why $H^\ddagger H$ is not the identity is sub-sampling. Hence the conditioning of $H^\ddagger H$ is a measure of sub-sampling in the acquisition. The initial way proposed for solving (3.11) is through a successive approximation iteration

$$(\forall n \in \mathbb{N}) \quad x_{n+1} = x_n - \tau H^\ddagger (Hx_n - y), \tag{3.12}$$

that converges if $|||\operatorname{Id} -\tau H^\ddagger H||| < 1$, where $|||\cdot|||$ denotes the spectral norm. Because approximate inversion aims to satisfy $H^\ddagger H \approx \operatorname{Id}$, the eigenvalues of $H^\ddagger H$ are expected to cluster around 1, and $H^\ddagger H$ to be close to a symmetric operator. Note that in the case of a parallel geometry, in the fully-sampled continuous case, $H^\ddagger$, as provided by FBP, is exactly the (right) pseudo-inverse of H.
Using $H^\ddagger = H^\top F$ with F the ramp filter, the product $H^\ddagger H$ is indeed symmetric and (3.12) converges to the well-characterized solution of

$$\underset{x\in\mathbb{R}^N}{\operatorname{argmin}} \ \frac{1}{2}\|F^{1/2}(Hx - y)\|^2. \tag{3.13}$$

This iterative method is an improvement over an initial approximate analytical reconstruction. When $H^\ddagger$ is provided by FBP, (3.12) is referred as iterative FBP (I-FBP).
I-FBP is not only used in transmission tomography but also in emission tomography. In SPECT, the attenuated Radon transform does not have an analytical inverse when attenuation is heterogeneous. The backprojection of the filtered attenuated data still produces an image close to the true image. This leads to using the non-symmetric operator $H^\ddagger H = H^\top F H_\mathrm{a}$, where H_a is a discrete attenuated Radon transform and $H^\top$ does not model attenuation [183, 240]. If it is important to stick to the normal equation, the introduction of filtering by preconditioning (3.9) is a more generic way to reach faster convergence [54], without changing the solution.

The ability to model the acquisition does not compensate for the lack of data. If the exact contour of the object is known *a priori*, the solutions to (3.9) and (3.11) can be superior to AR. In CT, model-based reconstruction has been used to decrease the X-ray dose delivered to the patient. The dose can be reduced by decreasing the X-ray tube current, which increases the noise. However, dose reduction can also be achieved by decreasing the number of measurements, particularly the number of projection angles, while conserving a high value for the X-ray tube current. Model-based reconstruction must include more information to provide meaningful solutions to these different setups. In general, advanced model-based reconstruction methods estimate $\overline{x}$ through the solution of the following minimization problem

$$\underset{x\in\mathbb{R}^N}{\operatorname{minimize}} \ d(x) + h(x). \tag{3.14}$$

where $d : \mathbb{R}^N \mapsto]-\infty, +\infty]$ is a data fidelity term, which typically depends on vector y and matrix H, and $h : \mathbb{R}^N \mapsto]-\infty, +\infty]$ is a regularization function. We will now discuss how modeling guides the choice of these functions.

3.2.2 Modeling errors in the data

For setups where the X-ray tube current is low, the choice of the data fidelity term is often driven by the modeling of the noise properties of the acquisition. In CT, there are two main ways of modeling the noise depending on whether the modeled data are log-transformed or not.

Poisson:

In emission tomography, the reconstructed function is not an attenuation map but the number of photons emitted from each point in space, which gives the map of the radioisotope injected into the patient. The Radon transform relates the number of photons counted by the detector and the radioisotope map, with the data following Poisson statistics. In transmission tomography, the data are photon counts as in emission tomography, thus following Poisson statistics. However, in this imaging modality, the relationship with the attenuation map is the Beer-Lambert law instead of the Radon transform. Statistical reconstruction, in such a context, aims at maximizing the Poisson likelihood of the measured data with respect to the parameters (radioisotope or attenuation maps). In optimization terms, this task can be reformulated as minimizing the log-likelihood, e.g., (3.15) for transmission tomography, leading to the setting

$$(\forall x \in \mathbb{R}^N) \quad d(x) = \sum_{c=1}^{C} (I(c)[Hx]_c + I_0 \exp(-[Hx]_c)). \tag{3.15}$$

One well-established Poisson log-likelihood method is the Expectation-Maximization algorithm (EM) [194] which includes a positivity constraint on x and was first proposed for emission tomography and then extended to transmission tomography [129].

Post-log Gaussian:

For a large intensity I, the Poisson distribution of mean and variance I is well approximated by the Gaussian distribution of the same mean and variance I [95]. The two distributions differ because the probability of a "negative" intensity is 0 with the Poisson distribution, while it is only asymptotically 0 with the Gaussian distribution. This difference becomes large for low intensities. If intensity I follows a Gaussian distribution of mean and variance I, the statistics of $\log(I)$ can be described by a Gaussian distribution of mean $\log(I)$ and variance $1/I$. However, there is also empirical evidence that CT measurement of a low intensity I deviates significantly from Poisson statistics because of the noise added by the readout electronics, to the point of making negative measurements possible [158]. Even if a more complex statistical analysis can be conducted to predict the variance of CT measurements at low intensity in the presence of electronic noise [76, 147], the model is as efficiently built through variance measurement on the imaging device.
With the above analysis, statistical modeling of noise in the log-transformed data is possible using a weighted least square data fidelity term such that

$$(\forall x \in \mathbb{R}^N) \quad d(x) = \frac{1}{2}\|Hx - y\|_\Sigma^2, \tag{3.16}$$

where $\|\cdot\|_\Sigma$ denotes the Σ-weighted norm for $\Sigma \in S_C^+$, i.e., for every $z \in \mathbb{R}^C$, $\|z\|_\Sigma = \sqrt{\langle z \mid \Sigma z \rangle}$. Σ contains the estimated standard deviations of the measurement and is often chosen as a diagonal one (with values $(1/I_c)_{c=1}^C$) because CT measurements are always assumed statistically independent.
When the X-ray tube current is high and the focus is on handling sub-sampling in the acquisition, we can consider that the data is degraded by a low level of white Gaussian noise, which translates into the choice

$$(\forall x \in \mathbb{R}^N) \quad d(x) = \frac{1}{2}\|Hx - y\|^2. \tag{3.17}$$

The particularity of the Poisson model (3.15) over the Gaussian model (3.16) is that mean and variance represents a single parameter so that the signal-to-noise ratio always follows the square root of the parameter. Note that Σ in (3.16) is given *a priori*, it is not estimated with x as in the Poisson case, and there is no embedded constraint of uniform signal-to-noise ratio. The model is thus compatible with extremely noisy images. Solutions of the Poisson models have more uniform noise distributions, which forbids local high-intensity noise measurement to propagate through the entire image as happens with AR. However, this also comes at a cost as the convergence of high frequencies with EM is slow [84].

3.2.3 Modeling properties of the image

Regularization function(s) usually favor prior knowledge or expectations of the image's characteristics. Regularization is also useful to mitigate the problem of ill-posedness of the reconstruction task.
In statistical terms, this is done through providing an *a priori* known distribution of the possible solution maps, which, in a Bayesian framework, is the Maximum A Posteriori (MAP) estimator, the penalty function h being the negative logarithm of the prior. Common prior assume spatial correlation of the pixels of the image [241]. Successful examples for diagnostic CT are Markov random field model-based priors (GMRF, q-GMRF) [28, 207]. Additional box or positivity constraints can easily be embedded in h. The interaction between the regularization and data fidelity terms is critical for statistical reconstruction to ensure proper noise filtering. It gives a complex *a priori* parameterization of the reconstruction problem that locally sets the balance between the noise level and the spatial resolution to enable the diagnostic.
For sub-sampled acquisitions, h is often a combination of functions that sparsely represent x. Examples of sparsity-promoting functions are the total variation (TV) [187], and its various improvements [38, 67], and frame-based regularization [45]. Sparse priors have seen a revival in popularity with the framework of Compressed Sensing [39, 78], which offers means to recover sparse images from fewer projection angles. The goal is to generate missing content in the nullspace of operator H according to the prior. Methods based on this framework have yielded impressive sampling compression factors for recovering images that perfectly fit sparse priors. Their applicability is long observed in MRI [104]. Dynamic MRI monitors dynamic processes such as brain hemodynamics and cardiac motion. Dynamic MR images are highly compressible. For example, the quasi-periodicity of heart images has a sparse temporal Fourier transform. Compressed sensing then relies on this sparsity to reconstruct a time-varying volume of the patient's heart. However, in the spatial domain, sparse priors do not perfectly fit clinical images. Only a biased estimation of these images can thus be obtained. For example, clinical images are only approximately piecewise constant. So TV regularization yields an unwanted

patchy look and an overall decrease in contrast resolution. Anatomical images are mixed with geometric devices (needles) in our context. Exploiting the sparsity of the devices is done by building the regularization on several priors applied to separate components of the volume (device, anatomical background). In CT, most of the works is focused on TV regularization [25, 117, 199]. TV methods built on Compressed Sensing consider a constrained optimization problem that reads

$$\underset{\substack{x\in\mathbb{R}^N \\ \|Hx-y\|^2\leqslant\epsilon}}{\text{minimize}} \ \|\nabla x\|_{1,2}, \tag{3.18}$$

where $\nabla \in \mathbb{R}^{N\times 2N}$ is the discrete 2D gradient operator and $(\forall z = (z_1, z_2) \in \mathbb{R}^{2N})$, $\|z\|_{1,2} = \sum_{n=1}^{N} \sqrt{(z_1)_n^2 + (z_2)_n^2}$ and $\epsilon = 0$ in the noise-free case. Using the regularizing term $h = \lambda\|\nabla \cdot\|_{1,2}$ and d as in (3.14), could sacrifice data fidelity for image regularity compared to the constrained formulation (3.18). One should still note that there is some equivalence between the regularized form (3.14) and a constrained version of it, in the spirit of (3.18). For any ϵ strictly greater than zero, there is a small value λ such that the constrained optimization problem is equivalent to the unconstrained problem.

In CBCT, the conic geometry and the detector's limited frame rate and size lead to an inherent lack of data. Even in the noise-free, densely sampled case, intense objects are never well sampled in the Fourier domain, so sub-sampling is always part of the problem. The setup of low noise data but a limited number of projections is thus commonly encountered in C-arm CBCT.

3.3 Optimization with fixed-point proximal algorithms

A closed-form solution to (3.14) is rarely available, so the solution is estimated iteratively. Proximal algorithms provide an iterative reconstruction (IR) of the volume as the solution of (3.14). They rely on proximal mappings to compute solutions to non-smooth optimization problems. At the core of the convergence analysis of most proximal splitting methods is the interpretation of the system of first-order optimality conditions associated with the optimization problem as an instance of a fixed-point problem and, more precisely, of a monotone inclusion problem.
We now introduce the mathematical tools required for the proximal algorithms used in this work. We present these methods in the general setting of optimization problems over real Hilbert spaces. We denote by $\mathcal{H}$, $\mathcal{G}$, $\mathcal{L}$ some real Hilbert spaces, and $\mathcal{B}(\mathcal{H}, \mathcal{G})$ the set of bounded and linear operators from $\mathcal{H}$ to $\mathcal{G}$. Most of our notations follow from the reference book [16]. From this point forward, our focus will be on the following regularized least-squares optimization problem

$$\underset{x\in\mathbb{R}^N}{\text{minimize}} \ \frac{1}{2}\|Hx - y\|_{\mathcal{G}}^2 + h(x), \tag{3.19}$$

where we now assume that $y \in \mathcal{G}$, $H \in \mathcal{B}(\mathcal{H}, \mathcal{G})$, $h : \mathcal{H} \mapsto]-\infty, +\infty]$ is convex but non-differentiable. Let us define functions $f : \mathcal{H} \mapsto]-\infty, +\infty]$ and $g : \mathcal{L} \mapsto]-\infty, +\infty]$ and operator $D \in \mathcal{B}(\mathcal{H}, \mathcal{L})$ such that

$$h(x) = f(x) + g(Dx), \tag{3.20}$$

so that (3.19) can be rewritten as

$$\underset{x\in\mathbb{R}^N}{\text{minimize}} \ \frac{1}{2}\|Hx - y\|_{\mathcal{G}}^2 + g(Dx) + f(x). \tag{3.21}$$

3.3.1 Mathematical analysis tools

The following definitions set up our framework for convex analysis and subdifferential calculus.

Notions of convex optimization:

Definition 3.3.1.1 Let $f : \mathcal{H} \mapsto]-\infty, +\infty]$.

(i) The *domain* of f is the set defined by $\operatorname{dom} f = \{x \in \mathcal{H} \mid f(x) < +\infty\}$.

(ii) The function f is *proper* if $\operatorname{dom} f$ is not empty.

(iii) The function f is *convex* if for every $\alpha \in]0, 1[$ the following holds,

$$(\forall (x, y) \in \operatorname{dom} f) \qquad f(\alpha x + (1-\alpha)y) \leqslant \alpha f(x) + (1-\alpha) f(y). \tag{3.22}$$

In addition, f is *ρ-strongly convex* function if $f - \frac{\rho}{2}\|x\|_{\mathcal{H}}^2$ is convex.

(iv) The function f is *coercive* if

$$\lim_{\|x\|_{\mathcal{H}} \to +\infty} f(x) = +\infty. \tag{3.23}$$

(v) The function f is *lower semicontinuous* (l.s.c) if, for every $x_0 \in \mathcal{H}$,

$$\lim_{x \mapsto x_0} \inf f(x) \geqslant f(x_0). \tag{3.24}$$

(vi) The *conjugate* of f is the function $f^* : \mathcal{H} \mapsto [-\infty, +\infty]$ defined by

$$(\forall x \in \mathcal{H}) \qquad f^*(x) = \sup_{y \in \mathcal{H}} \left(\langle x \mid y \rangle - f(y) \right). \tag{3.25}$$

An important remark is that f is ρ-strongly convex if and only if its conjugate f^* is continuously differentiable with $1/\rho$-Lipschitz gradient. Hence duality allows trading strong convexity with smoothness.

(vii) The *subdifferential* of f at the point x is the set-valued operator

$$\partial f : \mathcal{H} \mapsto 2^{\mathcal{H}} : x \mapsto \{v \in \mathcal{H} : f(z) \geqslant f(x) + v^\top (z - x) \qquad \forall z, x \in \operatorname{dom}(z)\}, \tag{3.26}$$

where $2^{\mathcal{H}}$ denotes the power set of $\mathcal{H}$.
Any such element is called a *subgradient*. If the function is differentiable, the subdifferential is a singleton set comprising the ordinary gradient.

(viii) The function f is *coercive* if

$$\lim_{\|x\|_{\mathcal{H}} \to +\infty} f(x) = +\infty \tag{3.27}$$

and *supercoercive* if

$$\lim_{\|x\|_{\mathcal{H}} \to +\infty} \frac{f(x)}{\|x\|_{\mathcal{H}}} = +\infty. \tag{3.28}$$

(ix) The *Moreau envelope* of f is the function $x \mapsto f_\lambda(x) = \min_{u \in \mathcal{H}} f(u) + \frac{1}{2\lambda}\|x - u\|^2$, where $\lambda > 0$. It can be used to smooth a non-smooth convex function without altering its minimizer. For example, the Moreau envelope of $f(x) = |x|$ is the Huber function from robust statistics.

The class of functions which are proper, convex, lower-semicontinuous on $\mathcal{H}$ and take values in $\mathbb{R} \cup \{+\infty\}$ is denoted by $\Gamma_0(\mathcal{H})$. The following optimality property is the basis of many optimization algorithms. For $f \in \Gamma_0(\mathcal{H})$,

$$x^* = \underset{x \in \mathcal{H}}{\operatorname{argmin}} \ f(x) \Leftrightarrow 0 \in \partial f(x^*). \tag{3.29}$$

A solution to (3.21) is thus a solution to the following variational inclusion problem

$$0 \in \partial f(x) + D^* \partial g(Dx) + H^*(Hx - y). \tag{3.30}$$

Notions of operator theory:

Definition 3.3.1.2 Let $\mathcal{M} : \mathcal{H} \to 2^{\mathcal{H}}$ be a set-valued operator.

(i) $\operatorname{dom} \mathcal{M} = \{x \in \mathcal{H} \mid \mathcal{M}x \neq \varnothing\}$ is the *domain* of $\mathcal{M}$, $\operatorname{ran} \mathcal{M}$ and $\operatorname{gra} \mathcal{M} = \{(x, y) \in \mathcal{H}^2 \mid y \in \mathcal{M}x\}$, its *range* and *graph*, respectively.

(ii) $\mathcal{M}^{-1} : \mathcal{H} \to 2^{\mathcal{H}}$ denotes the inverse operator of $\mathcal{M}$, with domain $\operatorname{ran} \mathcal{M}$ and range $\operatorname{dom} \mathcal{M}$, and $\mathcal{M}^{-1}(0) = \{x \in \mathcal{H} \mid 0 \in \mathcal{M}x\}$ the set of zeros of $\mathcal{M}$.

(iii) $\mathcal{M}$ is *monotone* if

$$(\forall (x, y) \in \mathcal{H}^2)(\forall u \in \mathcal{M}x)(\forall v \in \mathcal{M}y) \qquad \langle u - v \mid x - y \rangle_{\mathcal{H}} \geqslant 0. \tag{3.31}$$

$\mathcal{M}$ is *maximal monotone* if, in addition, its graph is not properly contained in the graph of any other monotone operator.

(iv) $\mathcal{M}$ is *strictly monotone* if

$$(\forall (x, y) \in \mathcal{H}^2)(\forall u \in \mathcal{M}x)(\forall v \in \mathcal{M}y) \quad x \neq y \Rightarrow \langle u - v \mid x - y \rangle_{\mathcal{H}} > 0. \tag{3.32}$$

(v) $\mathcal{M}$ is *strongly monotone* if there exists $\eta \in]0, +\infty[$ such that $\mathcal{M} - \eta \operatorname{Id}$ is monotone.

(vi) The *resolvent* of a maximally monotone operator $\mathcal{M}$ scaled with a parameter $\gamma > 0$ is the mapping $J_{\gamma\mathcal{M}} : \mathcal{H} \to \mathcal{H} : x \mapsto J_{\gamma\mathcal{M}}(x) = (\operatorname{Id}_{\mathcal{H}} + \gamma\mathcal{M})^{-1} x$, where $\operatorname{Id}_{\mathcal{H}}$ refers to the identity operator in $\mathcal{H}$.

(vii) The *Yosida approximation* of $\mathcal{M}$ of index γ is

$$^{\gamma}\mathcal{M} = \frac{1}{\gamma}(\operatorname{Id} - J_{\gamma\mathcal{M}}). \tag{3.33}$$

A prominent example of a monotone operator is the subdifferential operator. The resolvent of $(\gamma \partial f)$ is a proximity operator $\operatorname{prox}_{\gamma f}$ and the Yosida approximation of ∂f is the Fréchet derivative of the Moreau envelope; more precisely $^{\gamma}(\partial f) = \nabla(^{\gamma} f)$.

Definition 3.3.1.3 Let $T \colon \mathcal{H} \to \mathcal{H}$ be a single-valued operator

(i) A *fixed point* of T is any $x^* \in \mathcal{H}$ satisfying $x^* = Tx^*$.

(ii) T is η-*cocoercive* with $\eta \in [0, +\infty[$ if

$$(\forall x \in \mathcal{H})(\forall y \in \mathcal{H}) \qquad \eta \|Tx - Ty\|_{\mathcal{H}}^2 \leqslant \langle x - y \mid Tx - Ty \rangle_{\mathcal{H}}. \tag{3.34}$$

For example, let $f : \mathcal{H} \mapsto \mathbb{R}$ be a differentiable convex function, whose gradient is β-Lipschitz continuous, for some $\beta > 0$ then ∇f is $\frac{1}{\beta}$-cocoercice.

(iii) T is *nonexpansive* if it is Lipschitz continuous with constant 1, i.e.,

$$(\forall (x,y) \in \mathcal{H}^2) \qquad \|Tx - Ty\|_{\mathcal{H}} \leqslant \|x - y\|_{\mathcal{H}}. \tag{3.35}$$

T is α-*averaged* with $\alpha \in]0,1]$ if there exists a nonexpansive operator Q such that $T = (1-\alpha)\,\mathrm{Id} + \alpha Q$. *Firmly nonexpansive* means (1/2)-averaged.

Proximal algorithms aim at constructing α-averaged operators T, for some $\alpha \in]0,1[$, whose fixed points are solutions to optimization problems. They combine gradients and proximity operators of functions and iterate through

$$\begin{aligned} z^{n+\frac{1}{2}} &= Tz^n \\ z^{n+1} &= z^n + \theta_n(z^{n+\frac{1}{2}} - z^n), \end{aligned} \tag{3.36}$$

where $z_0 \in \mathcal{H}$ and $(\theta_n)_{n\in\mathbb{N}}$ are nonnegative relaxation parameters.

Proximity operators:

Proximal algorithms handle non-smooth functions through their proximity operators. The use of proximity operators can be viewed as a regularized and implicit way of dealing with their set-valued subdifferentials.

Definition 3.3.1.4 If $x \in \mathcal{H}$, the *proximity operator* of f at x is defined as [152]

$$\mathrm{prox}_f(x) = \underset{z\in\mathcal{H}}{\mathrm{argmin}} \quad \left(f(z) + \frac{1}{2}\|x - z\|_{\mathcal{H}}^2 \right). \tag{3.37}$$

It is the point that attains the minimum of the Moreau envelope.

This definition can be extended to an Q-weighted space, where $Q \in S_N^+$ is self-adjoint and positive-definite, as $\mathrm{prox}_f^Q(x) = \underset{z\in\mathcal{H}}{\mathrm{argmin}} \;\; f(z) + \frac{1}{2}\|x - z\|_Q^2$.

Proposition 3.3.1.5

- $\mathrm{prox}_f(x)$ is related to ∂f through the following inclusion

$$p = \mathrm{prox}_f(x) \Leftrightarrow x - p \in \partial f(p). \tag{3.38}$$

- $\mathrm{prox}_f(x)$ is firmly non-expansive

$$\forall x, u \in \mathbb{R}^n \qquad \|\mathrm{prox}_f(x) - \mathrm{prox}_f(u)\|_{\mathcal{H}} - \|(x - \mathrm{prox}_f(x)) + (u - \mathrm{prox}_f(u))\|_{\mathcal{H}} \leqslant \|x - u\|_{\mathcal{H}}. \tag{3.39}$$

- Moreau identity: $x = \mathrm{prox}_{\lambda f}(x) + \lambda \mathrm{prox}_{\frac{1}{\lambda} f^*}(\frac{x}{\lambda})$ $(\lambda > 0)$. It shows that the proximity operator of a function f is as easy to compute as the proximity operator of its convex conjugate.

- Evaluating the proximal operator can be viewed as a gradient-descent step for the Moreau envelope, with λ as a step size parameter i.e.

$$\nabla f_\lambda(\cdot) = \frac{1}{\lambda}(\mathrm{Id} - \mathrm{prox}_{\lambda f}). \tag{3.40}$$

- When $f(x) = \iota_C(x)$ is the set indicator function of some convex set C, $\operatorname{prox}_f(x) = \underset{z \in C}{\operatorname{argmin}} \ \|x - z\|_2^2 = \operatorname{proj}_C(x)$ is the ordinary Euclidean projection of x onto C. This suggests that, for other functions, the proximal operator can be thought of as a generalized projection. A constrained optimization problem $\underset{x \in C}{\text{minimize}} \ f(x)$ has an equivalent solution as an unconstrained proximal operator problem.

Although the definition of the proximity operator is implicit, it has a closed form for many functions of practical interest. For instance, for the absolute value (and by extension for the ℓ_1 norm by element-wise application), this is soft-thresholding: we have, for any $\lambda > 0$,

$$(\forall t \in \mathbb{R}) \qquad \operatorname{prox}_{\lambda|\cdot|}(t) = \operatorname{sign}(t) \max(|t| - \lambda, 0). \tag{3.41}$$

There are closed-form expressions for the proximity operators of a large class of functions http://proximity-operator.net.

3.3.2 First order splitting schemes

Now that we have seen that finding a minimizer of a function was equivalent to finding a zero of a monotone operator, we will focus on different operators $T : \mathcal{Z} \mapsto \mathcal{Z}$ in (3.36) whose fixed points are designed to be the minimizer of (3.21). Considering (3.19) and following the previous section, the resulting minimizer, denoted $\hat{x}$, verifies the following monotone inclusion

$$0 \in (\mathscr{A} + B)\hat{x}, \tag{3.42}$$

where $\mathscr{A} = H^*(H \cdot -y)$ and $B = \partial h$.
We now present two algorithms whose fixed point verify (3.42).

Forward-backward splitting:

The forward-backward algorithm assumes that $\mathscr{A}$ is a maximally monotone operator and B is a cocoercive operator. This method provides a sequence obtained from the fixed point iteration (3.36) of the nonexpansive operator

$$T = J_{\gamma B} \circ (\mathrm{Id} \ -\gamma \mathscr{A}), \tag{3.43}$$

where $\gamma \in \left]0, +\infty\right[$ is the algorithm step size.
Iterating (3.36) with (3.43) can be equivalently written under the implicit form

$$0 \in \mathcal{A} z_{n+\frac{1}{2}} + B z_n + \frac{1}{\gamma}(z_{n+\frac{1}{2}} - z_n), \tag{3.44}$$

where $z_0 \in \mathcal{Z}$.
For problem (3.19) over Hilbert space $\mathcal{H}$, it becomes

$$(\forall n \in \mathbb{N}) \quad x_{n+1} = x_n + \theta_n\big(\operatorname{prox}_{\gamma h}(x_n - \gamma H^*(H x_n - y)) - x_n\big), \tag{3.45}$$

where $x_0 \in \mathcal{H}$. This method is also called the proximal gradient algorithm (PGA). Several special cases for regularization are of interest:

- when $h \equiv 0$, (3.45) becomes the gradient method.
- When $h = \iota_C$, (3.45) becomes the projected gradient method.

- When $h = \|\cdot\|_1$, (3.45) reduces to the well-known ISTA that was developed for the purpose of wavelet-based signal restoration [21, 70].
- When $H = 0$, (3.45) becomes the proximal point method. In addition, if $h = \sum_{i=1}^{I} \iota_{C_i}$, (3.45) is the Projection Onto Convex Sets (POCS) for finding a solution in the intersection of convex sets [32, 234].

If $\theta_n \in [\epsilon, 1]$ with $\epsilon \in]0, 1[$ and $\gamma \in]0, 2/||H||^2_{\mathcal{H},\mathcal{G}}[$, the sequence $(x_n)_{n\in\mathbb{N}}$ generated by algorithm (3.45) converges weakly to a solution to Problem (3.19) when such a solution exists [64]. Strong convergence is even achieved in some contexts [31, 63, 70]. Recent results on overrelaxed versions of (3.45) can be found in [68] for special cases of gradient operators. The flexibility introduced by an iteration-dependent step size can be used to improve the algorithm convergence pattern.
In our general template, h is a composite function (3.20); $\text{prox}_{\gamma h}(x)$ often has to be computed iteratively by solving

$$\underset{z\in\mathcal{H}}{\text{minimize}} \ \frac{1}{2}\|z - x\|^2_{\mathcal{H}} + \gamma(f(z) + g(Dz)). \tag{3.46}$$

The forward-backward splitting can also be applied to solve sub-problem (3.46) by considering its dual problem

$$\underset{u\in\mathcal{G}}{\text{minimize}} \ \ \phi(-D^*u) + g^*(u), \tag{3.47}$$

where $\phi = f^* \square \frac{1}{2}\|\cdot\|^2_{\mathcal{H}}$ is the Moreau envelope of parameter 1 of f^*.
This yields the dual forward-backward algorithm (DFB)

$$z_{n+1} = \text{prox}_{\sigma g^*}(z_n - \sigma\nabla(\phi \circ (-D^*\cdot))(z_n)), \tag{3.48}$$

where $\sigma \in]0, 2/\|D\|^2_{\mathcal{H},\mathcal{L}}[$. Note that

$$\nabla(\phi \circ (-D^*\cdot)) = -D\text{prox}_{\gamma f}(-D^*\cdot) \tag{3.49}$$

Thus (3.48) becomes

$$\begin{aligned} z_n &= \text{prox}_{\gamma f}(x - D^*u_n) \\ u_{n+1} &= \text{prox}_{\sigma g^*}(u_n + \sigma D z_n). \end{aligned} \tag{3.50}$$

It can be shown that the sequences $(z_n)_{n\in\mathbb{N}}$ and $(u_n)_{n\in\mathbb{N}}$ converge to the solutions to the primal and dual problems $\hat{z}$ and $\hat{u}$, respectively and

$$\hat{z} = \text{prox}_{\gamma f}(D^*\hat{u}). \tag{3.51}$$

Acceleration of the forward-backward:

Inertial extensions improve the theoretical convergence rate of the forward-backward algorithm leading to the FISTA algorithm [42].
In an attempt to reach greater accelerations, several authors have explored preconditioning strategies [48, 156] to improve the conditioning of the gradient step [19, 27, 52, 62]. Let P be a self-adjoint strongly positive bounded linear operator on $\mathcal{Z}$, an equivalent problem is to multiply (3.42) by P^{-1}. The preconditioned forward-backward iteration to solve (3.42)-(3.19) is:

$$0 \in \mathcal{A}z_{n+\frac{1}{2}} + Bz_n + P(z_{n+\frac{1}{2}} - z_n), \tag{3.52}$$

leading to

$$x_{n+1} = x_n + \theta_n\big(\mathrm{prox}^P_{\gamma h}(x_n - \gamma P^{-1}H^*(Hx_n - y)) - x_n\big). \tag{3.53}$$

Other authors [130, 155, 208, 209] have proposed improving the conditioning of the data fidelity term using approximate inversion (see subsection 3.2.1), thus leading to a different solution contrary to the case of preconditioning. When the reconstruction focuses on subsampling, the precise statistical knowledge of the noise in the data is often disregarded, and the choice of the unweighted least-squares data fidelity term in (3.21) is convenient from a computational point of view. In this context, [130] proposed to replace $\frac{1}{2}\|Hx-y\|^2_{\mathcal{G}}$ by $\frac{1}{2}\|F^{1/2}(Hx-y)\|^2_{\mathcal{G}}$ in (3.19), where F is the ramp filter of AR. Note that even for applications other than tomography where a fast analytical inverse is unavailable, authors have used the pseudo-inverse $H^\dagger$ instead of F [208, 209]. [208] showed empirically that better reconstruction results could be achieved with approximate inversion.
Approximate homotopy strategies have also been proposed to solve ℓ_1-regularized least-squares problems (3.19) ($h = \lambda\|\cdot\|_1$) [79, 159, 230]. The idea is to first solve (3.19) with a large regularization parameter λ and then gradually decrease λ until the target regularization is reached. For each fixed λ, the forward-backward algorithm is used to solve the minimization problem up to an adequate precision. Then, the approximate solution serves as the initial point for the next value of λ. This strategy was also used for TV-regularized problems in [130], which reported superior empirical performance.

Tseng's splitting:

Tseng's Forward-Backward-Forward splitting relies on less restrictive assumptions on operator B that is only assumed to be monotone and ϑ-lipschitz continuous for some $\vartheta > 0$ but not necessarily cocoercive. The method relies on the fixed point iteration (3.36) of the operator

$$T = (\mathrm{Id} - \gamma B) \circ J_{\gamma\mathscr{A}} \circ (\mathrm{Id} - \gamma B) + \gamma B, \tag{3.54}$$

with $\theta_n \equiv 1$. For problem (3.19), it becomes

$$(\forall n \in \mathbb{N}) \quad \begin{aligned} v_n &= x_n - \gamma\nabla f(x_n) \\ p_n &= \mathrm{prox}_{\gamma h}(v_n) \\ q_n &= p_n - \gamma\nabla f(p_n) \\ x_{n+1} &= x_n - v_n + q_n. \end{aligned} \tag{3.55}$$

According to [34, Theorem 2.5 (ii)], if $\mathrm{zer}(\mathscr{A}+B) \neq \varnothing$ and $\gamma \in]0, 1/\vartheta[$, sequences $(x_n)_{n\in\mathbb{N}}$ and $(p_n)_{n\in\mathbb{N}}$ converge weakly to some $\widehat{x} \in \mathrm{zer}(\mathscr{A} + B)$. Therefore, sequences $(x_n)_{n\in\mathbb{N}}$ generated by (3.55) converge weakly to $\widehat{x}$, which is a solution to (3.19).

Note that other splittings, which do not exploit any Lipschitz assumptions for B, exist, such as the Douglas-Rachford splitting [29, 96], but were not used in this thesis.

3.3.3 Primal-dual methods defined on a product space

It is possible to avoid computing the proximity operator of h. This section shows how the two aforementioned splitting techniques can be applied to an equivalent problem to handle f and g separately and decouple D from g, yielding simpler iterations.
In general, a solution to (3.30) is a solution to (3.21), but the converse may not be true. From now on, the solution set of (3.30) is supposed to be nonempty, and then so is the solution set of (3.21). Under mild qualification constraints, the solutions to (3.21) and (3.30) are the same. One of such qualification conditions is

$$0 \in \operatorname{sri}\left(D(\operatorname{dom} f) - \operatorname{dom} g\right). \tag{3.56}$$

where $\operatorname{sri} C$ denotes the strong relative interior of $C \subset \mathcal{H}$. We refer the reader to [16, Proposition 27.5, Corollary 27.6] for other examples of qualification constraints.
A way to simplify the problem consists in introducing an auxiliary variable $u \in \partial g(Dx)$, called the dual variable so that the problem now consists in finding x and u such that

$$\begin{cases} u \in \partial g(Dx) \\ 0 = D^*u + \nabla h(x) \end{cases} \tag{3.57}$$

So equivalently, the problem is to find a pair of objects $z = (x, u) \in \mathcal{Z} = \mathcal{H} \times \mathcal{L}$ which satisfies the following system of decoupled monotone inclusions

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} \in \begin{pmatrix} \partial f(x) + D^*u + H^*Hx - H^*y \\ -Dx + \partial g^*(u) \end{pmatrix}. \tag{3.58}$$

This is also equivalent to

$$\begin{cases} Dx \in (\partial g)^{-1}(u) \\ x \in (\nabla q)^{-1}(-D^*u) \end{cases} \tag{3.59}$$

where $q\colon \mathcal{H} \to \mathbb{R}\colon x \mapsto \frac{1}{2}\|y - Hx\|^2_{\mathcal{G}}$, which implies that

$$0 \in (\partial g)^{-1}u - D(\nabla q)^{-1}(-D^*u). \tag{3.60}$$

This is the first-order characterization of the dual problem of (3.21)

$$\underset{u \in \mathcal{G}}{\text{minimize}} \quad (f + q)^*(-D^*u) + g^*(u), \tag{3.61}$$

According to [16, Theorem 19.1], x is a solution to (3.21) and u is a solution to (3.61) if $z = (x, u) \in \mathcal{Z} = \mathcal{H} \times \mathcal{L}$ is a solution to (3.58). We now show several ways of splitting (3.58) which lead to different proximal primal-dual algorithms.

Condat-Vũ algorithm:

The Condat-Vũ (CV) algorithm [66, 221] can be obtained from the preconditioned forward-backward splitting (3.52) where

$$(\forall z = (x, u) \in \mathcal{Z}) \quad \mathcal{A}z = \begin{pmatrix} \partial f(x) + D^*u \\ -Dx + \partial g^*(u) \end{pmatrix} \tag{3.62}$$

$$Bz = \begin{pmatrix} H^*Hx - H^*y \\ 0 \end{pmatrix} \tag{3.63}$$

$$Pz = \begin{pmatrix} \frac{1}{\tau}x - D^*u \\ -Dx + \frac{1}{\sigma}u \end{pmatrix}, \tag{3.64}$$

with $(\tau, \sigma) \in]0, +\infty[^2$. Operator $\mathcal{A}$ is maximally monotone [16, Proposition 26.32 (iii)] and operator B is θ-cocoercive, with $\theta = 1/\|H\|^2_{\mathcal{H},\mathcal{G}}$. Let $\{\Theta_n\}_{n\in\mathbb{N}}$ be a sequence of relaxation parameters. Let $z_n = (x_n, u_n) \in \mathcal{Z}$, and $z_{n+\frac{1}{2}} = (x_{n+\frac{1}{2}}, u_{n+\frac{1}{2}}) \in \mathcal{Z}$. Plugging P, $\mathcal{A}$ and B in (3.52) yields

CV iterations for solving (3.21):

$$\text{for } n = 0, 1, \dots \quad \left\lfloor \begin{array}{l} x_{n+\frac{1}{2}} = \operatorname{prox}_{\tau f}\left(x_n - \tau(H^*(Hx_n - y) + D^* u_n)\right) \\ u_{n+\frac{1}{2}} = \operatorname{prox}_{\sigma g^*}\left(u_n + \sigma D(2x_{n+\frac{1}{2}} - x_n)\right) \\ x_{n+1} = x_n + \Theta_n(x_{n+\frac{1}{2}} - x_n) \\ u_{n+1} = u_n + \Theta_n(u_{n+\frac{1}{2}} - u_n), \end{array} \right. \tag{3.65}$$

with initialization $x_0 \in \mathcal{H}$ and $u_0 \in \mathcal{L}$.
The convergence of (3.65) to a solution pair (x, u) to (3.58) is guaranteed by [66, Theorem 3.1] for step sizes $\tau > 0$, $\sigma > 0$ such that $\tau\left(\sigma\|D\|^2_{\mathcal{H},\mathcal{L}} + \dfrac{1}{2\theta}\right) < 1$, and relaxation parameters $\{\Theta_n\}_{n\in\mathbb{N}} \subset [0, \delta]$, with $\delta = 2 - \dfrac{1}{2\theta}\left(\dfrac{1}{\tau} - \sigma\|D\|^2_{\mathcal{H},\mathcal{L}}\right)$ and θ the cocoercivity constant of operator B such that $\sum_{n\in\mathbb{N}} \Theta_n\,(\delta - \Theta_n) = +\infty$.

Remark 3.3.3.1 The condition on the step sizes (τ, σ) implies that $\tau\sigma\|D\|^2_{\mathcal{H},\mathcal{L}} < 1$, which allows us to conclude that operator P in (3.52) is strongly positive (bounded from below) and therefore invertible.

Remark 3.3.3.2 We can observe that if we set the smooth term in (3.21) to 0 i.e., we remove the quadratic term $\frac{1}{2}\|Hx - y\|^2_{\mathcal{G}}$, the Condat–Vũ iteration reverts to the Chambolle–Pock iteration [43] (PDHG). CV can be viewed as a generalization of PDHG, taking into account the gradient of Lipschitz differentiable terms in the cost function.

Loris-Verhoeven algorithm:

We now present another forward-backward primal-dual algorithm proposed by Loris and Verhoeven (LV) [143] that can be derived for a special case of f such that $f = \frac{\kappa}{2}\|\cdot\|^2_{\mathcal{H}}$ with $\kappa \in]0, +\infty[$. This algorithm also appears under the name of Primal-Dual Fixed-Point algorithm based on the Proximity Operator (PDFP2O) [49] and Proximal Alternating Predictor-Corrector (PAPC) algorithm [80].
The LV iterations can still be described by means of the implicit inclusion (3.52) where $\mathcal{A}$, B, and P are now given by

$$(\forall z = (x, u) \in \mathcal{Z}) \quad \mathcal{A}z = \begin{pmatrix} D^* u \\ -Dx + \partial g^*(u) \end{pmatrix} \tag{3.66}$$

$$Bz = \begin{pmatrix} (H^*H + \kappa\,\mathrm{Id}_{\mathcal{H}})x - H^* y \\ 0 \end{pmatrix} \tag{3.67}$$

$$Pz = \begin{pmatrix} \frac{1}{\tau} x \\ (\frac{1}{\sigma}\,\mathrm{Id}_{\mathcal{L}} - \tau D^* D)u \end{pmatrix} \tag{3.68}$$

with $(\tau, \sigma) \in]0, +\infty[^2$. The iterations of LV algorithm are then given by

LV iterations:

$$\text{for } n = 0, 1, \ldots \quad \left\lfloor \begin{array}{l} t_n = H^*\left(Hx_n - y\right) + \kappa x_n \\ u_{n+\frac{1}{2}} = \operatorname{prox}_{\sigma g^*}\Big(u_n + \sigma D\Big(x_n - \tau(t_n + D^* u_n)\Big)\Big) \\ x_{n+1} = x_n - \Theta_n \tau\Big(t_n + D^* u_{n+\frac{1}{2}}\Big) \\ u_{n+1} = u_n + \Theta_n\Big(u_{n+\frac{1}{2}} - u_n\Big) \end{array} \right. \tag{3.69}$$

where $\{\Theta_n\}_{n\in\mathbb{N}}$ is a sequence of relaxation parameters and $(x_0, u_0) \in \mathcal{Z}$.
The Condat–Vũ algorithm and the Loris–Verhoeven algorithms are both primal-dual forward-backward algorithms, but they are different. When $f = 0$, larger values of τ and σ are allowed in the latter than in the former; this may benefit the convergence speed in practice.

Combettes-Pesquet algorithm:

We now present the Combettes - Pesquet algorithm (CP), which relies on Tseng's splitting (3.54) to solve (3.21). This algorithm was introduced in [58, Theorem 4.2] and generalizes the one in [34]. In (3.42), we can set

$$(\forall z = (x, u) \in \mathcal{Z}) \quad \mathcal{A}z = \begin{pmatrix} \partial f(x) \\ \partial g^*(u) \end{pmatrix} \tag{3.70}$$

$$Bz = \begin{pmatrix} H^*Hx - H^*y + D^*u \\ -Dx \end{pmatrix}, \tag{3.71}$$

where $\mathscr{A}$ is a maximally monotone operator (see [16, Theorem 26.32 (iii)]) and B is a monotone operator which is Lipschitz continuous with constant $\vartheta \leqslant \|H\|^2_{\mathcal{H}} + \|D\|_{\mathcal{H},\mathcal{L}}$ (see [58]).
It reads

CP-iterations for (3.21):

$$\text{for } n = 0, 1, \ldots \quad \left\lfloor \begin{array}{l} v_{1,n} = x_n - \gamma\left(H^*(Hx_n - y) + D^* u_n\right) \\ p_{1,n} = \operatorname{prox}_{\gamma f}(v_{1,n}) \\ v_{2,n} = u_n + \gamma D x_n \\ p_{2,n} = \operatorname{prox}_{\gamma g^*}(v_{2,n}) \\ q_{2,n} = p_{2,n} + \gamma D p_{1,n} \\ q_{1,n} = p_{1,n} - \gamma\left(H^*(Hp_{1,n} - y) + D^* p_{2,n}\right) \\ x_{n+1} = x_n - v_{1,n} + q_{1,n} \\ u_{n+1} = u_n - v_{2,n} + q_{2,n}, \end{array} \right. \tag{3.72}$$

where $\gamma > 0$ and $(x_0, u_0) \in \mathcal{Z}$. If $\operatorname{zer}(\mathscr{M} + Q) \neq \varnothing$ and $\gamma \in]0, 1/\vartheta[$, sequences $(z_n)_{n\in\mathbb{N}} = \big((x_n, u_n)\big)_{n\in\mathbb{N}}$ and $(p_n)_{n\in\mathbb{N}}$ converge weakly to some $\widehat{z} = (\widehat{x}, \widehat{u}) \in \operatorname{zer}(\mathscr{M} + Q)$. Therefore, sequences $(x_n)_{n\in\mathbb{N}}$ (resp. $(u_n)_{n\in\mathbb{N}}$) generated by (3.72) converge weakly to $\widehat{x}$ (resp. $\widehat{u}$), which is a solution to (3.21) (resp. (3.61)).
These primal and primal-dual proximal schemes will appear throughout this thesis in our numerical experiments and mathematical analysis. Other primal-dual algorithms exist and can be deduced from the Douglas-Rachford iteration, such as the well-known Alternating Direction Method of Multipliers (ADMM) algorithm.
We now discuss the practical limitations of model-based iterative reconstruction methods.

3.4 Why do commercial CT scanners still employ traditional, filtered back-projection over iterative reconstruction? [164]

The authors of this 2009 paper stated that "the title poses a question meant to provoke applied mathematicians and image-reconstruction experts to consider closer collaboration with engineers who design tomographic systems and vice versa.". Since 2009, iterative reconstruction has been introduced into all commercial scanners for dose reduction through noise modeling and filtering only. No commercial C-arm systems offer an iterative reconstruction solution. The question is thus still open given that, in the meantime, fixed point proximal algorithms have reached image-reconstruction experts [197]. We now explore contingencies that still block the introduction of iterative reconstruction into clinical systems. They form the rationale for the rest of this manuscript.

3.4.1 Computation time

Model-based reconstruction offers a lot of flexibility to take into account different sources of data degradation simultaneously. However, in the clinical practice of interventional imaging, if an analytical reconstruction takes T seconds, performing n iterations of proximal algorithms will take at least nT, which quickly becomes prohibitive. C-arm commercial algorithms thus add simpler pre- and post-processing methods to tackle a few sources of artifacts separately. For instance, dedicated beam hardening and scatter corrections have been proposed [118, 211]. Metal artifacts reduction (MAR) algorithms are used to handle data degradation in the presence of metal [101]. Contrary to proximal algorithms for model-based reconstruction, MAR algorithms are not implicitly defined; they do not minimize any global criteria or solve a fixed point problem. A common point with the homotopy strategy of [130] with TV-regularization is to segment the metal over a previous reconstruction. However, with MAR, iterations are avoided by interpolating the data.

3.4.2 Theoretical issues

Reconstruction experts have allowed changes in the system models that break the theoretical guarantees of proximal algorithms to reduce the computation time.

Mismatched projection pairs

In the continuous domain, backward projection or backprojection is the adjoint operation of forward projection or projection. In the discrete domain, the expression projection/backprojection pair (FP/BP) refers to the choices made for the object basis and the system model. If H is a projection matrix, $H^\top$ is a backprojection matrix. However, one may find, for instance, that discretizing the continuous backprojector $\mathcal{R}^\top$ directly as $B \neq H^\top$ is better suited to one's needs. In that case, H and B are said to be unmatched. There are two main categories of discretization strategies for deriving such operators; the ones that use resampling transforms and those that adopt a geometric viewpoint. Geometric models for FP are ray-summation methods [137, 195] which model the volume as a set of cubic voxels. They consider rays or strips connecting the X-ray source and the detector bins according to the acquisition geometry and compute the intersection length between each voxel and each ray/strip. Alternative resampling methods consider the center of each voxel only. They project the center of each detector bin onto the volume

and perform separable interpolations with the volume samples. For example, the Joseph method is such a method using linear interpolation. Resampling methods for FP also virtually involve summing values on a ray drawn from the detector bin center to the X-ray source. As such, these methods are often referred to as ray-driven methods. The sampling of the values is constant for a given ray but varies between rays.
Discretization strategies designed for BP usually use resampling transforms, which are now voxel-driven [88]: each voxel is visited in a loop and projected on the detector across lines. A voxel's contribution is split between two neighboring detector elements using typically linear interpolation.
Ray-driven FP and voxel-driven BP are destination-driven, one endpoint of the rays or lines being either a detector bin or a voxel center. An efficient FP does not necessarily imply that the adjoint BP is also efficient and vice-versa. A precise and fast projector uses the Joseph method with an increased number of samples in the volume. However, the adjoint backprojector introduces interpolation redundancies in the volume domain that must then be suppressed by the iterative reconstruction method. The presence of redundancies has led to the development of a geometric approach equally suited for the discretization of FP and BP [72, 138], which is, however, not easily adapted to GPUs. Due to the difficulty to use adjoint FP and BP which are both fast, accurate, and easily parallelizable, it has become very common in the CT community to work with unmatched FP/BP pairs.

Non-symmetrical approximate inversion:

In subsection 3.3.2, we pointed out that the ramp filter could be used to improve the conditioning of the data fidelity term, resulting in an acceleration of the forward-backward algorithm. For parallel geometry, operator $H^\ddagger H$ is symmetrical, but for cone-beam geometry, the voxel-wise weighting within FDK backprojection makes $H^\ddagger H$ paradoxically closer to the identity, inducing faster convergence, but formally not symmetrical as it is the case for product BH. When $H^\ddagger$ is provided by an analytical algorithm, in addition to faster convergence in early iterations [240], the reconstruction criterion does not differ much from the proximal operator of the regularization. The behavior and parameterization of the regularization become independent from model H and easily predictable [182].

Inadequate minimizer

In IR, the prior remains simple enough to design tractable proximal algorithms. Simple priors lead to an "unnatural look" of the images because they do not perfectly capture real anatomies. For instance, the minimum of a TV criterion will be a piece-wise constant image that clinicians do not trust. In [130], the regularization strength goes to zero in the last step of the homotopy method. The final image is thus not the minimizer of the criterion, but it is much better than FDK while keeping its preferred "look".

3.4.3 Hyper-parameters

Model-based reconstruction involves one or several hyper-parameters that appear in the regularization. These parameters are critical to image quality, yet they are difficult to tune because they depend on both the patient and the clinical task. Specializing the prior accordingly increases the number of parameters. There is no universal automatic metric to set these parameters, so the reconstruction is empirically tuned per context. Bayesian methods tune the parameters during the reconstruction by assigning them probability distribution functions [223]. This is attractive but makes the objective function more complicated, which entails more computational burden.

3.4.4 Deep learning pre-/post-processing

Since the publication of [164], deep learning methods combined with AR have shown tremendous potential, resulting in rich literature surrounding this subject and making the question raised in this section even more relevant.
Deep learning methods are fast and do not require manually defining prior or tuning hyper-parameters at inference time. Multi-resolution convolutional neural networks (CNN) now strongly challenge statistical iterative reconstruction methods for diagnostic CT [22, 201, 228].
Recently, researchers also started investigating the benefits of deep learning for subsampled acquisitions [4, 222] especially to accelerate the reconstruction [86]. Most of the proposed deep learning methods either restore, interpolate, or extrapolate the projections before the application of $H^\ddagger$ [18,97,112,134], or post-process the result of $H^\ddagger y$ [245] while evidence that post-processing approaches achieve better performance than pre-processing approaches were provided in [135]. The most popular multi-scale CNN used in the literature for image reconstruction is the U-Net [111, 116], first proposed for biomedical image segmentation [186]. Unlike in IR, it is difficult to incorporate prior knowledge about the reconstructed images into a neural network, which in most cases is viewed as a black box. The reliability of such methods can thus be questioned as theoretical foundations have still to be developed. Notably, in [198], the authors suggest being cautious when using popular post-processing CNNs as they demonstrate that such networks may not compute the solution of the CT inverse problem even when it exists. They can introduce structures not belonging to the scanned objects in the reconstructions. Their performance is highly dependent on the quality of the training data to properly learn both the artifacts to remove and the anatomical details to preserve. This dependency is a challenge, especially for medical applications, where precise training data is often lacking. In contrast, with IR, a general inverse problem framework for sparsity exploiting image reconstruction underpins the algorithm so that inverse problem results are testable and repeatable. The observations of [198] thus suggest that IR is still a competitive alternative to deep learning post-processing methods.

3.5 Deep Learning reconstruction

We now review methods that integrate deep learning into the image reconstruction framework as an entirely data-driven mapping from measured projections to images. Our focus is on supervised learning approaches which rely on pairs of projections $(y_i)_{i=1}^I$ and artifact-free images $(\overline{x}_i)_{i=1}^I$[1]. One first attempt in using supervised learning for reconstruction is to directly estimate the mapping between the projection data and the reconstructed image by minimizing the ℓ_2 loss $\theta \mapsto \frac{1}{2}\sum_{i=1}^I \|x_i - f_\theta(y_i)\|_2^2$ where f_θ is the function representing the parameterized network. Learning such a mapping using a neural network, though not impossible, was shown to require a very large training set, to be computationally expensive, and to rely heavily on good initialization of the model parameters (e.g., AUTOMAP [246]). More successful approaches have been developed by building on the previous AR or IR methods.

3.5.1 Learning in analytical reconstruction

The first approaches to using deep learning within reconstruction are built on AR methods [93, 114, 128, 168, 229]. Most of them focus on learning the reconstruction filter. In [229], the authors note that AR can be encoded as a CNN in one-to-one correspondence; $H^\top H$ being of convolutional type, the ramp filter also acts as a convolution. This representation allows the authors to learn joint correction steps in the volume and projection domains and improve Parker's weighting for cone-beam limited-angle problems with an angular range of 180°. In [168], the authors learn a non-linear combination of FBP reconstruction operators. These non-linear combinations of the reconstruction filters in the FBP reconstruction operators are learned by training against the outcome of the reconstruction given by the TV regularization. The authors then extend this approach for cone-beam geometry [128].

3.5.2 Extension of post-processing methods

A direct way to ensure data consistency is to combine a post-processing network such as a U-net with data consistency layers. The resulting network is not necessarily a feed-forward architecture [83, 203].

3.5.3 Deep Unfolding

Deep unfolding (DU) translates a fixed number N of iterations of an iterative algorithm to a neural network with an architecture adapted to that algorithm. For example, unrolling the scheme defining T in (3.36) and stopping the iterates after N steps allow us to express the N-th iterate as

$$T_\theta^N = (T_{\omega_N} \circ \cdots \circ T_{\omega_1}). \tag{3.73}$$

T_θ^N can be seen as a feed-forward neural network where each layer in the network evaluates T_{ω_n} and the parameters of the network are $\theta = (\omega_1, \cdots, \omega_N)$. Moreover, if the parameters are shared across layers, i.e. $\omega_1 = \cdots = \omega_N = \omega$ for some ω, T_θ^N can in fact be interpreted as a recurrent neural network. The non-linear operations that may appear in T_θ^N play the role of an activation function (soft thresholding is very close to the well-known ReLU activation function).

[1] Note that there are unsupervised approaches which either rely on unpaired data (adversarial regularizers [145]) or can be trained only from images $(\overline{x}_i)_{i=1}^I$ (Compressed-Sensing generative models [113]).

The corresponding architecture T_θ^N is trained from end-to-end such that

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \ \frac{1}{2} \sum_{i=1}^{I} loss(\overline{x}_i, T_\theta^N(x_i)). \tag{3.74}$$

Prior works have unrolled the gradient method [3, 47], PGA [11], ADMM [233], PDHG [2, 224] and ISTA [106]. No clear consensus has emerged about which methods perform best in general or even for specific problems. The only formal requirement is that each iteration of the unrolled algorithm is almost everywhere differentiable.
The set of parameters θ in an unrolled iterative algorithm can be the algorithm's hyper-parameters, such as step sizes and regularization parameters. Learning only the algorithm's hyper-parameters [24] yields a network with very few parameters that remains close to the original solver. One can also learn linear operators such as convolution kernels that can be tied [106] or untied [177]. [36] also learns the backprojector in the context of limited-angle CT. In [243], the authors replaced all linear operators in the regularization in ISTA with non-linear operators moving further from the original algorithm. Another choice is to learn proximity operators as it is done in ADMM-Net [233], and Primal-Dual Networks (PD-Net) [2] where a neural network replaces each proximity operator in the sub-problems of PDHG. [108] replaced the projector in a projected gradient descent method with a CNN. Another example of using unrolling is the work of [99], which constructs a deep neural network by unrolling a truncated Neumann series for the inverse of a linear operator. The closest architectures to IR benefit from their interpretability. Other works focus on designing learned optimization solvers with a better convergence rate than those achieved with IR [15]. When increasing the capacity of the unrolled scheme, we also increase the expressivity of the network and allow for further acceleration, potentially at the price of convergence guarantees.

3.5.4 Deep Equilibrium

Deep equilibrium models (DEQ) is a recent extension of DU to an arbitrary number of iterations [13]. DEQ is an implicit network; it can be implemented by replacing $T_\theta^N(x_i)$ in (3.74) by a fixed-point x_i^* of a given operator T_θ, and using implicit differentiation for updating the weights θ. Therefore, contrary to the approaches above, DEQ models do not have prescribed explicit computation graphs. The benefit of DEQ over DU is that it does not require the storage of the intermediate variables for solving (3.74); the algorithm that drives the model to fulfill this equilibrium criterion is not prescribed. Therefore, DEQ models can leverage black-box solvers in their forward passes and enjoy analytical backward passes independent of the forward pass trajectories, thus reducing the memory complexity of training. However, the computation of the fixed-point $\hat{x}_i$ can increase the computational complexity for training. Rigorous optimization of DEQ models can be challenging because they involve bi-level optimization. DEQ has been used for MRI reconstruction [100] and multi-scale learning problems [14].

DU and DEQ are at the crossroads of classical deep learning architectures and fixed-point algorithms and aim at enjoying the benefits of both approaches [10, 231]. When seen as a modification of a fixed-point algorithm, they provide a framework to possibly make current fixed-point algorithms evolve towards better adapting to clinical data through a learned prior. When seen as a modification of deep learning architectures, they offer a way to embed prior information about the data or the reconstructed image (the most simple form of information being data consistency). DU and DEQ architectures

typically have much fewer trainable parameters than black-box deep neural networks and are more suitable for learning on relatively small data sets.

3.6 Conclusion

In the following chapters, we will propose ways to maintain the theoretical guarantees of proximal methods when the "tricks" mentioned in subsection 3.4.2 are used in practice, addressing all aspects of iterative reconstruction. Firstly, we will investigate the convergence properties of some of the aforementioned proximal algorithms when unmatched FP/BP are used. Secondly, we will propose alternative discretizations for H with new competitive matched pairs; thirdly, a new regularization based on a sum of directional TV terms will be shown to capture better the content of images in the context of percutaneous interventions. We will conclude with an alternative data fidelity term and explore the benefits of DU for a task of ROI reconstruction with an expected fast convergence thanks to an untied parameterization of the learned parameters across layers.

4 | Convergence of the proximal gradient algorithm with an adjoint mismatch

4.1 Introduction

As outlined in Chapter 3 (section 3.4), unmatched FP/BP pairs are frequently used in CT reconstruction. The motivation is often to obtain an estimate of the patient's attenuation to X-rays more rapidly (by improving the conditioning [240], reducing the computational complexity per iteration [88]) or to improve the quality of the reconstruction [238]. However, unmatched pairs lead to an adjoint mismatch on the forward operator H. With an adjoint mismatch, the convergence guarantees of classical minimization algorithms no longer hold; errors might accumulate over iterations [9, 238]. Hence, existing results and schemes must be adapted to these situations.

The effect of an adjoint mismatch has first been studied for CT in the context of row-action algebraic algorithms based on POCS. The convergence of these schemes was analyzed with tools from linear algebra. Among them, [77,81,139,240] gave convergence conditions and focused on fixing the convergence of these schemes [77]. In contrast, [161,196] proposed to use more general optimization schemes that can directly deal with non-symmetric normal fixed-point equations. Studying the impact of adjoint mismatch also finds application in deep learning. Very recently, Bubba et al. have proposed a CNN-based reconstruction algorithm ΦDONet in [36] where the BP operator is replaced with a partially learned operator in an unrolled ISTA architecture to improve backprojection for limited-angle acquisitions. They considered an ℓ_1 penalization applied to the wavelets coefficients of the object. By exploiting some conditions on H, ΦDONet was presented as a perturbed version of ISTA. Adopting a probabilistic approach, the authors establish the convergence in mean of the output of their optimally trained network with respect to the ground truth in finite dimensions. Though targeted to a specific application, their approach could be extended to any convolutional forward operator H, which is a pseudo-differential or a Fourier integral operator.

As mentionned in Chapter 3, PGA is an algorithm that involves non-smooth operators for solving the penalized least-squares problems arising in CT reconstruction. It is very popular because of its simplicity and ability to handle general non-differentiable convex priors. In this chapter, we extend the stability analysis proposed in [77] to PGA in the presence of adjoint mismatch in an arbitrary Hilbert space.

This chapter is organized as follows: section 4.2 introduces the mismatched PGA iteration. Section 4.3 gives necessary conditions to preserve the convergence of this iteration and gives a bound on the discrepancy induced by the mismatch between the resulting fixed point and the minimizer of the original objective function. Examples of linear inverse problems with sparsity constraints arising from computed tomography are discussed in section 4.4. Then, we show in section 4.5 how our theoretical results can be

exploited to study an unmatched preconditioning approach where two different metrics are used in the gradient step and the proximity step of PGA. Associated examples are provided in section 4.6.

4.2 Mismatched PGA

We denote by $\mathcal{H}$, $\mathcal{G}$ some real Hilbert spaces. We suppose that our FP, H, belongs to $\mathcal{B}(\mathcal{H},\mathcal{G})$ and that the data y and the attenuation map $\overline{x}$ evolve respectively in $\mathcal{G}$ and $\mathcal{H}$. In this chapter, we consider the following penalized least squares criterion for finding an estimate of $\overline{x}$:

$$\underset{x\in\mathcal{H}}{\text{minimize}} \ \frac{1}{2}\|y-Hx\|_{\mathcal{G}}^2 + g(x) + \frac{\kappa}{2}\|x\|_{\mathcal{H}}^2, \tag{4.1}$$

where $g \in \Gamma_0(\mathcal{H})$ is a suitable possibly non-smooth regularization function and $\kappa \in [0,+\infty[$. Note that, when $\kappa > 0$, an elastic net-like penalization is introduced and the objective function in (4.1) is thus strongly convex [248].
PGA applied on (4.1) reads, for every $n \in \mathbb{N}$,

$$x_{n+1} = x_n + \theta_n\big(\text{prox}_{\gamma g}((1-\gamma\kappa)x_n - \gamma H^*(Hx_n - y)) - x_n\big), \tag{4.2}$$

where $x_0 \in \mathcal{H}$ is the initial estimate, $(\theta_n)_{n\in\mathbb{N}}$ are nonnegative relaxation parameters, and $\gamma \in]0,+\infty[$ is the algorithm step size.
If $\theta_n \in [\epsilon, 1]$ with $\epsilon \in]0,1[$ and $\gamma \in]0, 2/(||H||_{\mathcal{H},\mathcal{G}}^2 + \kappa)[$, the sequence $(x_n)_{n\in\mathbb{N}}$ generated by Algorithm (4.2) converges weakly to a solution to Problem (4.1) when such a solution exists [63, 64, 70]. Without loss of generality, the step size is hereinafter assumed to be constant [1].
In the context of an adjoint mismatch, operator H^* is purposefully replaced by surrogate operators $(K_n)_{n\in\mathbb{N}}$, iteration (4.2) thus becoming:
For every $n \in \mathbb{N}$,

$$x_{n+1} = x_n + \theta_n\big(\text{prox}_{\gamma g}((1-\gamma\kappa)x_n - \gamma K_n(Hx_n - y)) - x_n\big). \tag{4.3}$$

We now list our assumptions to analyze iteration (4.3).

Assumption 4.2.0.1

(i) $g \in \Gamma_0(\mathcal{H})$

(ii) *For every* $n \in \mathbb{N}$, $K_n \in \mathcal{B}(\mathcal{G},\mathcal{H})$

(iii) *There exist* $\overline{K} \in \mathcal{B}(\mathcal{G},\mathcal{H})$ *and* $\{\omega_n\}_{n\in\mathbb{N}} \subset]0,+\infty[$ *with* $\sum_{n\in\mathbb{N}}\omega_n < +\infty$ *such that*

$$\overline{K}H \neq 0 \tag{4.4}$$

$$(\forall n \in \mathbb{N}) \qquad \|K_n - \overline{K}\|_{\mathcal{G},\mathcal{H}} \leqslant \omega_n. \tag{4.5}$$

The last assumption covers two scenarios of particular interest:

- When $\overline{K} = H^*$, we get a sequence of operators $(K_n)_{n\in\mathbb{N}}$ providing asymptotically the adjoint of H.
- When, for every $n \in \mathbb{N}$, $\omega_n = 0$, a constant difference $\overline{K} - H^*$ is introduced on the adjoint.

[1] Extending our analysis to varying step sizes is straightforward.

In the context of convergence analysis of fixed point iterations [61] of the modified PGA Algorithm (4.3), the following notation is central.

Notation 4.2.0.2 Let $\gamma \in]0, +\infty[$. We define

$$L = \overline{K}H + \kappa \,\mathrm{Id} \tag{4.6}$$

$$\begin{aligned} T_\gamma \colon \quad & \mathcal{H} \to \mathcal{H} \\ & x \mapsto \mathrm{prox}_{\gamma g}(x - \gamma L x + \gamma \overline{K} y) \end{aligned} \tag{4.7}$$

$$\lambda_{\min} = \inf_{\substack{x \in \mathcal{H} \\ \|x\|_{\mathcal{H}} = 1}} \langle x \mid Lx \rangle_{\mathcal{H}} \tag{4.8}$$

$$\lambda^+_{\min} = \inf_{\substack{x \in (\mathrm{Ker}\, L)^\perp \\ \|x\|_{\mathcal{H}} = 1}} \langle x \mid Lx \rangle_{\mathcal{H}} \tag{4.9}$$

$$\lambda_{\max} = \sup_{\substack{x \in \mathcal{H} \\ \|x\|_{\mathcal{H}} = 1}} \langle x \mid Lx \rangle_{\mathcal{H}} \tag{4.10}$$

$$\beta = \frac{1}{2}\|L - L^*\|_{\mathcal{H},\mathcal{H}}. \tag{4.11}$$

Note that $\lambda_{\min}$ (resp. $\lambda_{\max}$) is the minimum (resp. maximum) spectral value of $(L+L^*)/2$ and that $\lambda^+_{\min} \geqslant \lambda_{\min}$.
We now show that the convergence of Algorithm (4.3) is guaranteed under cocoercivity conditions on operator L.

4.3 Convergence analysis

4.3.1 Regularity of the surrogate gradient operator

When $\overline{K} \neq H^*$, the gradient of the smooth part of our objective function is replaced by operator $\kappa \,\mathrm{Id} + \overline{K}(H \cdot - y)$, which is not guaranteed to be a cocoercive operator. We now propose conditions preserving this property. First, we prove certain properties induced by the cocoercivity of operator L.

Lemma 4.3.1.1 *Let $\eta \in]0, +\infty[$. If L is η-cocoercive, then the following hold:*

(i) $\lambda_{\min} \geqslant 0$

(ii) $\mathrm{Ker}\,(L + L^*) = \mathrm{Ker}\, L = \mathrm{Ker}\, L^*$

(iii) $L + L^* \neq 0$.

Proof: L is η-cocoercive if and only if, for every $x \in \mathcal{H}$,

$$\langle x \mid Lx \rangle_{\mathcal{H}} \geqslant \eta \|Lx\|^2_{\mathcal{H}}. \tag{4.12}$$

(i): The fact that $\lambda_{\min} \geqslant 0$ directly follows from (4.12).
(ii): If $x \in \mathrm{Ker}\, L$, then

$$\begin{aligned} & \langle x \mid Lx \rangle_{\mathcal{H}} = 0 \\ \Leftrightarrow \quad & \langle x \mid (L + L^*)x \rangle_{\mathcal{H}} = 0 \end{aligned} \tag{4.13}$$

According to (i), $L+L^*$ is self-adjoint positive. It thus admits a self adjoint square root $(L+L^*)^{1/2}$ and (4.13) is equivalent to

$$\|(L+L^*)^{1/2}x\|_{\mathcal{H}}^2 = 0 \quad \Leftrightarrow \quad (L+L^*)^{1/2}x = 0, \tag{4.14}$$

which yields $(L+L^*)x = 0$. We have thus proved that $\operatorname{Ker} L \subset \operatorname{Ker}(L+L^*)$. By reexpressing (4.12),

$$(\forall x \in \mathcal{H}) \quad \frac{1}{2}\langle x \mid (L+L^*)x\rangle_{\mathcal{H}} \geqslant \eta\|Lx\|_{\mathcal{H}}^2. \tag{4.15}$$

Consequently, if $x \in \operatorname{Ker}(L+L^*)$, then $x \in \operatorname{Ker} L$. In summary, $\operatorname{Ker}(L+L^*) = \operatorname{Ker} L$. By symmetry, $\operatorname{Ker}(L+L^*) = \operatorname{Ker} L^*$

(iii): $L+L^* = 0$ if and only if $\operatorname{Ker}(L+L^*) = \mathcal{H}$ which, according to (ii), would imply that $\operatorname{Ker} L = \mathcal{H}$, that is $L = 0$. This contradicts our assumption in (4.4). □

Whenever cocoercivity is present, we show that the behavior of iterative scheme (4.3) remains stable. Conditions for cocoercivity are summarized below.

Proposition 4.3.1.2

(i) Assume that $\lambda_{\min} \geqslant 0$.
If $\lambda_{\min}^+ \in]0,+\infty[$ and $\operatorname{Ker}(L+L^*) = \operatorname{Ker} L$, then L is $\underline{\eta}$-cocoercive with

$$\underline{\eta} = 1/\Big(\sqrt{\lambda_{\max}} + \frac{\beta}{\sqrt{\lambda_{\min}^+}}\Big)^2. \tag{4.16}$$

If $\beta = 0$, then L is $(1/\lambda_{\max})$-cocoercive.

(ii) Suppose that $\operatorname{ran}(L+L^*)$ is closed. L is η-cocoercive with $\eta \in]0,+\infty[$ if and only if $\lambda_{\min} \geqslant 0$, $\operatorname{Ker}(L+L^*) = \operatorname{Ker} L$, and

$$\eta \leqslant \overline{\eta} = \frac{2}{\|(\mathrm{Id} + (L-L^*)(L+L^*)^{\dagger})(L+L^*)^{1/2}\|_{\mathcal{H},\mathcal{H}}^2}. \tag{4.17}$$

Proof: (i): Let A and B be the self-adjoint and skewed parts of L, respectively given by

$$A = \frac{L+L^*}{2} \tag{4.18}$$

$$B = \frac{L-L^*}{2}. \tag{4.19}$$

Assume first that $V = \operatorname{Ker} A = \operatorname{Ker} L = \operatorname{Ker} L^*$. Let $x \in \mathcal{H}$ and let $x_{V^\perp}$ denote its projection onto the orthogonal complement of V. We have

$$\begin{aligned} \|Lx\|_{\mathcal{H}}^2 &= \|Lx_{V^\perp}\|_{\mathcal{H}}^2 \\ &\leqslant (\|Ax_{V^\perp}\|_{\mathcal{H}} + \|Bx_{V^\perp}\|_{\mathcal{H}})^2. \end{aligned} \tag{4.20}$$

Since $\lambda_{\min} \geqslant 0$, A is a positive operator and we have then

$$\|Ax_{V^\perp}\|_{\mathcal{H}}^2 \leqslant \|A\|_{\mathcal{H},\mathcal{H}}\langle x_{V^\perp} \mid Ax_{V^\perp}\rangle_{\mathcal{H}} = \lambda_{\max}\langle x_{V^\perp} \mid Ax_{V^\perp}\rangle_{\mathcal{H}}. \tag{4.21}$$

In turn,

$$\|Bx_{V^\perp}\|_{\mathcal{H}} \leqslant \beta\|x_{V^\perp}\|_{\mathcal{H}}. \tag{4.22}$$

and

$$\langle x_{V^\perp} \mid Ax_{V^\perp}\rangle_{\mathcal{H}} \geqslant \lambda^+_{\min}\|x_{V^\perp}\|^2_{\mathcal{H}}. \tag{4.23}$$

Then, if $\lambda^+_{\min} > 0$,

$$\|Bx_{V^\perp}\|^2_{\mathcal{H}} \leqslant \frac{\beta^2}{\lambda^+_{\min}}\langle x_{V^\perp} \mid Ax_{V^\perp}\rangle_{\mathcal{H}} \tag{4.24}$$

Altogether (4.20), (4.21) and (4.24) yield

$$\|Lx\|^2_{\mathcal{H}} \leqslant \frac{1}{\underline{\eta}}\langle x_{V^\perp} \mid Ax_{V^\perp}\rangle_{\mathcal{H}} = \frac{1}{\underline{\eta}}\langle x \mid Ax\rangle_{\mathcal{H}} = \frac{1}{\underline{\eta}}\langle x \mid Lx\rangle_{\mathcal{H}}, \tag{4.25}$$

where $\underline{\eta}$ is given by (4.16). This shows that L is $\underline{\eta}$-cocoercive.

If $\beta = 0$, then $L = A$ and the result follows from the inequality

$$\|Ax\|^2_{\mathcal{H}} \leqslant \lambda_{\max}\langle x \mid Ax\rangle_{\mathcal{H}}. \tag{4.26}$$

(ii): According to Lemma 4.3.1.1, if L is cocoercive then $\lambda_{\min} \geqslant 0$ and $\operatorname{Ker}(L+L^*) = \operatorname{Ker} L$. To establish the result, we thus assume that these two conditions are satisfied and prove that L is cocoercive if and only if (4.17) holds.

Let us use the same notation as in the proof of (i). Since $V = \operatorname{Ker} A = \operatorname{Ker} L$, L is η-cocoercive with $\eta \in]0, +\infty[$ if and only if

$$(\forall x \in V^\perp) \quad \eta\|Lx\|^2_{\mathcal{H}} \leqslant \langle x \mid Ax\rangle_{\mathcal{H}}. \tag{4.27}$$

Let $x \in V^\perp$ and let $y = Ax$. Since $\operatorname{ran} A$ is closed, this is equivalent to $x = A^\dagger y$. We have then

$$\begin{aligned}
\|Lx\|^2_{\mathcal{H}} &= \|Ax + Bx\|^2_{\mathcal{H}} \\
&= \|Ax + BA^\dagger y\|^2_{\mathcal{H}} \\
&= \|(\mathrm{Id} + BA^\dagger)Ax\|^2_{\mathcal{H}} \\
&\leqslant \|(\mathrm{Id} + BA^\dagger)A^{1/2}\|^2_{\mathcal{H},\mathcal{H}}\|A^{1/2}x\|^2_{\mathcal{H}} \\
&= \|(\mathrm{Id} + BA^\dagger)A^{1/2}\|^2_{\mathcal{H},\mathcal{H}}\,\langle x \mid Ax\rangle_{\mathcal{H}}.
\end{aligned} \tag{4.28}$$

Note that $\|(\mathrm{Id} + BA^\dagger)A^{1/2}\|_{\mathcal{H},\mathcal{H}} \neq 0$ (since L is nonzero). We have thus shown that L is cocoercive with constant $1/\|(\mathrm{Id} + BA^\dagger)A^{1/2}\|^2_{\mathcal{H},\mathcal{H}} = \overline{\eta}$, hence for any constant $\eta > 0$ satisfying (4.17).

In addition, the maximum cocoercivity constant $\eta_{\max}$ is such that

$$\frac{1}{\eta_{\max}} = \sup_{v \in V^\perp\setminus\{0\}} \frac{\|Lx\|^2_{\mathcal{H}}}{\langle x \mid Ax\rangle_{\mathcal{H}}} = \sup_{v \in V^\perp\setminus\{0\}} \frac{\|(\mathrm{Id} + BA^\dagger)Ax\|^2_{\mathcal{H}}}{\|A^{1/2}x\|^2_{\mathcal{H}}}. \tag{4.29}$$

On the other hand,

$$\|(\mathrm{Id} + BA^\dagger)A^{1/2}\|^2_{\mathcal{H},\mathcal{H}} = \sup_{z \in \mathcal{H}\setminus\{0\}} \frac{\|(\mathrm{Id} + BA^\dagger)A^{1/2}z\|^2_{\mathcal{H}}}{\|z\|^2_{\mathcal{H}}}. \tag{4.30}$$

Every $z \in \mathcal{H}$ can be decomposed as $z_V + z_{V^\perp}$, where $(z_V, z_{V^\perp}) \in V \times V^\perp$. Since A is self-adjoint positive, $V = \operatorname{Ker} A^{1/2} = \operatorname{Ker} A$. We can thus reexpress (4.30) as

$$\begin{aligned}
\|(\mathrm{Id} + BA^\dagger)A^{1/2}\|_{\mathcal{H},\mathcal{H}} &= \sup_{z \in \mathcal{H}\setminus\{0\}} \frac{\|(\mathrm{Id} + BA^\dagger)A^{1/2}z_{V^\perp}\|^2_{\mathcal{H}}}{\|z_V\|^2_{\mathcal{H}} + \|z_{V^\perp}\|^2_{\mathcal{H}}} \\
&= \sup_{z_{V^\perp} \in V^\perp\setminus\{0\}} \frac{\|(\mathrm{Id} + BA^\dagger)A^{1/2}z_{V^\perp}\|^2_{\mathcal{H}}}{\|z_{V^\perp}\|^2_{\mathcal{H}}}.
\end{aligned} \tag{4.31}$$

We know however that $V^\perp = (\operatorname{Ker} A^{1/2})^\perp = \overline{\operatorname{ran}\left((A^{1/2})^*\right)} = \overline{\operatorname{ran}(A^{1/2})}$. The expressions in (4.29) and (4.31) are thus equal. □

Remark 4.3.1.3 When $\beta \neq 0$, (4.16) suggests that η is higher when the nonzero spectral values of $(L + L^*)/2$ are clustered together.

The following special cases are worth being mentioned.

Corollary 4.3.1.4

(i) If $\lambda_{\min} > 0$, then L is cocoercive with constant

$$\begin{aligned}\overline{\eta} &= \frac{2}{\|(\mathrm{Id} + (L - L^*)(L + L^*)^{-1})(L + L^*)^{1/2}\|^2_{\mathcal{H},\mathcal{H}}} \\ &\geqslant 1/\Big(\sqrt{\lambda_{\max}} + \frac{\beta}{\sqrt{\lambda_{\min}}}\Big)^2. \end{aligned} \tag{4.32}$$

(ii) Assume that $\mathcal{H}$ is finite dimensional and $\lambda_{\min} \geqslant 0$. If the dimensions of $\operatorname{Ker} L$ *and* $\operatorname{Ker}(L + L^*)$ *are equal, then L is cocoercive with constant $\overline{\eta} \geqslant \underline{\eta}$ where $\underline{\eta}$ and $\overline{\eta}$ are given by* (4.16) *and* (4.17)*, respectively.*

Proof:

(i) If $\lambda_{\min} > 0$, then $L + L^*$ is strongly positive. It is thus invertible, $\operatorname{ran}(L + L^*) = \mathcal{H}$ is closed, and $\operatorname{Ker}(L + L^*) = \{0\}$. In the proof of Proposition 4.3.1.2(ii), we have seen that $\operatorname{Ker} L \subset \operatorname{Ker}(L + L^*)$. Therefore, $\operatorname{Ker}(L + L^*)$ and $\operatorname{Ker} L$ reduce to the null space and, according to Proposition 4.3.1.2(ii), L is $\overline{\eta}$-cocoercive.
In addition, in this case, $\lambda^+_{\min} = \lambda_{\min}$, it then follows from Proposition 4.3.1.2(i) and the fact that $\overline{\eta}$ is the maximum cocoercivity constant of L that the lower bound in (4.32) holds.

(ii) We have seen that $\lambda_{\min} \geqslant 0$ implies that $\operatorname{Ker} L \subset \operatorname{Ker}(L + L^*)$. Therefore, $\operatorname{Ker} L$ is equal to $\operatorname{Ker}(L + L^*)$ if and only if the dimensions of $\operatorname{Ker} L$ and $\operatorname{Ker}(L + L^*)$ are equal. In addition, $\operatorname{ran}(L + L^*)$ is closed and $\lambda^+_{\min}$ is necessarily positive when $\mathcal{H}$ is finite-dimensional. The result then follows from Proposition 4.3.1.2.

□

Remark 4.3.1.5 Let $\widetilde{\lambda}_{\min}$ be the minimum spectral value of $(\overline{K}H + H^*\overline{K}^*)/2$. We have $\lambda_{\min} = \widetilde{\lambda}_{\min} + \kappa$. A practical choice for κ to ensure that $\lambda_{\min}$ is positive is thus to set $\kappa > -\widetilde{\lambda}_{\min}$.

To ensure the convergence of (4.3), it is straightforward to rely upon conditions based on the nonexpansiveness of $\mathrm{Id} - \gamma L$. We next show that such conditions are directly related to the cocoercivity of L.

Proposition 4.3.1.6 If L is η-cocoercive with $\eta \in]0, +\infty[$ and $\gamma \leqslant 2\eta$, then

$$\|\mathrm{Id} - \gamma L\|^2_{\mathcal{H},\mathcal{H}} \leqslant 1 + \gamma\Big(\frac{\gamma}{\eta} - 2\Big)\lambda_{\min} \leqslant 1. \tag{4.33}$$

Conversely, if $\|\mathrm{Id} - \gamma L\|_{\mathcal{H},\mathcal{H}} \leqslant 1$ for some $\gamma \in]0, +\infty[$, then L is η-cocoercive for every $\eta \in]0, \gamma/2]$.

Proof: For every $x \in \mathcal{H}$,

$$\|(\mathrm{Id} - \gamma L)x\|_{\mathcal{H}}^2 = \|x\|_{\mathcal{H}}^2 - 2\gamma\langle x \mid Lx\rangle_{\mathcal{H}} + \gamma^2\|Lx\|_{\mathcal{H}}^2. \tag{4.34}$$

Because of (4.12)

$$\|(\mathrm{Id} - \gamma L)x\|_{\mathcal{H}}^2 \leqslant \|x\|_{\mathcal{H}}^2 - 2\gamma\langle x \mid Lx\rangle_{\mathcal{H}} + \frac{\gamma^2}{\eta}\langle x \mid Lx\rangle_{\mathcal{H}}. \tag{4.35}$$

Therefore, since $\gamma \leqslant 2\eta$,

$$\|\,\mathrm{Id} - \gamma L\|_{\mathcal{H},\mathcal{H}}^2 \leqslant 1 + \gamma\Big(\frac{\gamma}{\eta} - 2\Big)\lambda_{\min}. \tag{4.36}$$

According to Proposition 4.3.1.2(i), $\lambda_{\min} \geqslant 0$ and the obtained upper bound is thus less than or equal to 1.
Conversely, if $\mathrm{Id} - \gamma L$ is nonexpansive, then

$$(\forall x \in \mathcal{H}) \quad \|x - \gamma Lx\|_{\mathcal{H}}^2 \leqslant \|x\|_{\mathcal{H}}^2 \tag{4.37}$$

The cocoercivity of L thus straightforwardly follows from (4.34). □

4.3.2 Characterization of the fixed points of the mismatched iteration

We now characterize the existence and uniqueness of the fixed point set of operator T_γ. Such a fixed point generally no longer coincides with the global solution to (4.1).

Proposition 4.3.2.1

(i) Let $\gamma \in]0, +\infty[$ and let $\widetilde{x} \in \mathcal{H}$. We have $\widetilde{x} \in \operatorname{Fix} T_\gamma$ if and only if $\widetilde{x}$ belongs to

$$\mathcal{F} = \big\{x \in \mathcal{H} \;\big|\; 0 \in Lx - \overline{K}y + \partial g(x)\big\}. \tag{4.38}$$

In addition, $\mathcal{F}$ is non empty if $L + \partial g$ is surjective.

(ii) If $\lambda_{\min} \geqslant 0$, then $\mathcal{F}$ is a closed and convex set.

(iii) $\mathcal{F}$ has at most one element if one of the following conditions holds:

(a) $L + \partial g$ is strictly monotone;
(b) $L + L^*$ is positive definite;
(c) $\lambda_{\min} \geqslant 0$ and g is strictly convex.

In addition, $\mathcal{F}$ is a singleton if $\lambda_{\min} \geqslant 0$ and one of the following conditions holds:

(d) $L + \partial g$ is strongly monotone;
(e) $\lambda_{\min} \neq 0$;
(f) g is strongly convex.

(iv) Assume that L is cocoercive. $\mathcal{F}$ is nonempty if one of the following conditions holds:

(a) $\operatorname{dom} \partial g = \mathcal{H}$ and $(L + L^*)/2 + \partial g$ is surjective;

(b) $\operatorname{dom} \partial g = \mathcal{H}$ and

$$x \mapsto \frac{1}{2}\langle x \mid Lx\rangle_{\mathcal{H}} + g(x) \tag{4.39}$$

is coercive;

(c) g is supercoercive;

(d) $\operatorname{dom} g$ is bounded.

Proof: (i): We have

$$\begin{aligned} \widetilde{x} \in \operatorname{Fix} T_\gamma &\Leftrightarrow \widetilde{x} = \operatorname{prox}_{\gamma g}((1-\gamma\kappa)\widetilde{x} - \gamma\overline{K}(H\widetilde{x} - y)) \\ &\Leftrightarrow (1-\gamma\kappa)\widetilde{x} - \gamma\overline{K}(H\widetilde{x} - y) \in (\operatorname{Id} + \gamma\partial g)(\widetilde{x}) \\ &\Leftrightarrow \widetilde{x} \in \mathcal{F}. \end{aligned} \tag{4.40}$$

Under the considered surjectivity condition, there straightforwardly exists $\widetilde{x} \in \mathcal{H}$ for which (4.38) holds.

(ii): If $\lambda_{\min} \geqslant 0$, then L is monotone. Since it is continuous, it is maximally monotone, and $x \mapsto Lx - \overline{K}y$ is also maximally monotone. As the domain of this operator is $\mathcal{H}$ and ∂g is maximally monotone, $x \mapsto Lx - \overline{K}y + \partial g(x)$ is maximally monotone. It then follows from [16, Proposition 23.39] that $\mathcal{F}$ is closed and convex.

(iii)(a): This follows from [16, Proposition 23.35].

(iii)(b): If $L + L^*$ is positive definite then, for every $x \in \mathcal{H} \setminus \{0\}$,

$$\langle x \mid Lx\rangle_{\mathcal{H}} = \frac{1}{2}\langle x \mid (L + L^*)x\rangle_{\mathcal{H}} > 0, \tag{4.41}$$

which shows that L is strictly monotone. Since ∂g is monotone, we deduce that $L + \partial g$ is strictly monotone, and (iii)(a) allows us to conclude that T_γ has at most one fixed point.

(iii)(c): According to [16, Example 22.4(ii)], if g is strictly convex, then ∂g is strictly monotone. $\lambda_{\min} \geqslant 0$ if and only if L is monotone. $L + \partial g$ is then strictly monotone. Thus the result still follows from (iii)(a).

(iii)(d): If $\lambda_{\min} \geqslant 0$, because of the monotonicity and the continuity of L, $L + \partial g$ is maximally monotone. The result then follows from [16, Proposition 23.37].

(iii)(e): For every $x \in \mathcal{H}$,

$$\langle x \mid Lx\rangle_{\mathcal{H}} \geqslant \lambda_{\min}\|x\|_{\mathcal{H}}^2, \tag{4.42}$$

which shows that L is strongly monotone. We deduce that $L + \partial g$ is strongly monotone, and (iii)(d) allows us to conclude that $\mathcal{F}$ is a singleton.

(iii)(f): According to [16, Example 22.4(iv)], if g is strongly convex, then ∂g is strongly monotone. Since L is monotone, $L + \partial g$ is strongly monotone and the result follows from (iii)(d).

(iv)(a): Let A and B be defined by (4.18) and (4.19), respectively. We have thus

$$L + \partial g = A + \partial g + B. \tag{4.43}$$

According to Proposition 4.3.1.2,

$$\lambda_{\min} = \inf_{\substack{x \in \mathcal{H} \\ \|x\|_{\mathcal{H}} = 1}} \langle x \mid Lx\rangle_{\mathcal{H}} = \inf_{\substack{x \in \mathcal{H} \\ \|x\|_{\mathcal{H}} = 1}} \langle x \mid Ax\rangle_{\mathcal{H}} \geqslant 0, \tag{4.44}$$

which implies that A is maximally monotone. As ∂g is maximally monotone and $\operatorname{dom} A = \mathcal{H}$, $A + \partial g$ is maximally monotone. Since B is a skewed continuous linear operator, it

is also maximally monotone and $A + \partial g + B$ is maximally monotone. According to Lemma 4.3.1.1(iii), $A \neq 0$ and, since it is self-adjoint, it is $1/\|A\|_{\mathcal{H},\mathcal{H}}$-cocoercive. It then follows from [16, Proposition 25.16] that A is 3^* monotone. According to [16, Example 2.13], ∂g is 3^* monotone. By invoking [16, Proposition 25.22], $A+\partial g$ is thus 3^* monotone. Since $\operatorname{dom} B = \mathcal{H} = \operatorname{dom}(A + \partial g)$, it can be deduced from the Brézis-Haraux theorem (see [16, Corollary 25.27(ii)]) that $A + \partial g + B$ is surjective.

(iv)(b): The function defined by (4.39) also reads

$$h\colon x \mapsto \frac{1}{2}\langle x \mid Ax\rangle_{\mathcal{H}} + g(x). \tag{4.45}$$

We have seen that A is self-adjoint and monotone (i.e. positive semi-definite and self-adjoint). Let $z \in \mathcal{H}$. Minimizing

$$x \mapsto h(x) - \langle x \mid z\rangle_{\mathcal{H}} \tag{4.46}$$

is thus a convex optimization problem. A classical necessary condition for this problem to admit a solution is that h is coercive. In turn, if x is a solution to the optimization problem (4.45), it follows from Fermat's rule that

$$z \in \partial h(x) = Ax + \partial g(x). \tag{4.47}$$

Since z can be chosen arbitrarily, this shows that $A + \partial g$ is surjective. The fact that $\mathcal{F} \neq \varnothing$ then follows from (iv)(a).

(iv)(c)-(iv)(d): Let $\gamma \in]0, 2\eta]$ where η is the cocoercivity constant of L and let $W = \operatorname{Id} - \gamma L$. According to Proposition 4.3.1.6, $\|W\|_{\mathcal{H},\mathcal{H}} \leqslant 1$ and $\operatorname{Id} - W = \gamma L$ is monotone. In addition, $\operatorname{dom} g^* = \mathcal{H}$ if and only if g is supercoercive [16, Proposition 14.15]. The results then follow from [60, Proposition 4.3(vi)(d)]. □

By design, (4.38) shows that any fixed point of T_γ is a solution to an equilibrium instead of being defined from some optimality condition. In the context of Remark 4.3.1.5, the existence of a unique fixed point $\widetilde{x}$ for T_γ follows from the above result.

4.3.3 Convergence conditions and error bound

The fixed point of T_γ can be viewed as an approximation to the minimizer of Problem (4.1) whose error is bounded in the following theorem.

Theorem 4.3.3.1 *Assume that the following hold.*

(i) L *is cocoercive.*

(ii) Let $\nu \in [0, +\infty[$ *be the strong convexity modulus of* g*. Either* $\nu > 0$ *or* $\lambda_{\min} \neq 0$*.*

(iii) $\widehat{x}$ *is a solution to the minimization Problem* (4.1)*.*

Then there exists a unique solution $\widetilde{x}$ *to* (4.38) *and the following upper bound on the error incurred by the mismatch holds:*

$$\|\widetilde{x} - \widehat{x}\|_{\mathcal{H}} \leqslant \chi \,\|(H^* - \overline{K})(H\widehat{x} - y)\|_{\mathcal{H}}, \tag{4.48}$$

where

$$\chi = \inf_{\gamma \in]0, 2\eta[} \frac{\gamma}{1 + \gamma\nu - \|\operatorname{Id} - \gamma L\|_{\mathcal{H},\mathcal{H}}} \leqslant \frac{1}{\nu + 2\lambda_{\min}}. \tag{4.49}$$

Proof: According to Proposition 4.3.1.2(i), $\lambda_{\min} \geqslant 0$.
If $\lambda_{\min} > 0$, according to Proposition 4.3.2.1(iii)(e), (4.38) has a unique solution $\widetilde{x}$.
If $\lambda_{\min} = 0$, then $\nu > 0$, which means that g is ν-strongly convex. It then follows from Proposition 4.3.2.1(iii)(f) that (4.38) has a unique solution $\widetilde{x}$.
Let $\gamma \in]0, +\infty[$. According to Proposition 4.3.2.1(i), $\widetilde{x} \in \operatorname{Fix} T_\gamma$, that is

$$\widetilde{x} = \operatorname{prox}_{\gamma g}((1 - \gamma\kappa)\widetilde{x} - \gamma\overline{K}(H\widetilde{x} - y)), \tag{4.50}$$

and we also know that

$$\widehat{x} = \operatorname{prox}_{\gamma g}((1 - \gamma\kappa)\widehat{x} - \gamma H^*(H\widehat{x} - y)). \tag{4.51}$$

We can write $g = h + \nu/2\|\cdot\|_{\mathcal{H}}^2$ where $h \in \Gamma_0(\mathcal{H})$, which implies that

$$(\forall x \in \mathcal{H}) \quad \operatorname{prox}_{\gamma g}(x) = \operatorname{prox}_{\frac{\gamma}{1+\gamma\nu}h}\Big(\frac{x}{1+\gamma\nu}\Big). \tag{4.52}$$

As $\operatorname{prox}_{\frac{\gamma}{1+\gamma\nu}h}$ is nonexpansive, we deduce that

$$\begin{aligned}
\|\widetilde{x} - \widehat{x}\|_{\mathcal{H}} &\leqslant \frac{1}{1+\gamma\nu}\|(1 - \gamma\kappa)\widetilde{x} - \gamma\overline{K}(H\widetilde{x} - y) - (1 - \gamma\kappa)\widehat{x} + \gamma H^*(H\widehat{x} - y)\|_{\mathcal{H}} \\
&= \frac{1}{1+\gamma\nu}\|(\operatorname{Id} - \gamma L)(\widetilde{x} - \widehat{x}) + \gamma(H^* - \overline{K})(H\widehat{x} - y)\|_{\mathcal{H}} \\
&\leqslant \tau_\gamma\|\widetilde{x} - \widehat{x}\|_{\mathcal{H}} + \frac{\gamma}{1+\gamma\nu}\|(H^* - \overline{K})(H\widehat{x} - y)\|_{\mathcal{H}}
\end{aligned} \tag{4.53}$$

with

$$\tau_\gamma = \frac{\|\operatorname{Id} - \gamma L\|_{\mathcal{H},\mathcal{H}}}{1+\gamma\nu}. \tag{4.54}$$

In addition, according to Proposition 4.3.1.6, when $\gamma \leqslant 2\eta$,

$$\|\operatorname{Id} - \gamma L\|_{\mathcal{H},\mathcal{H}} \leqslant 1. \tag{4.55}$$

This ensures that $\tau_\gamma < 1$, when $\nu > 0$.
Assume now that $\lambda_{\min} > 0$. If $\gamma < 2\eta$, (4.33) yields

$$\|\operatorname{Id} - \gamma L\|_{\mathcal{H},\mathcal{H}} < 1, \tag{4.56}$$

which also guarantees that $\tau_\gamma < 1$.
In summary, if $\gamma < 2\eta$, it can be deduced from (4.53) that

$$\|\widetilde{x} - \widehat{x}\|_{\mathcal{H}} \leqslant \frac{\gamma}{(1 - \tau_\gamma)(1+\gamma\nu)}\|(H^* - \overline{K})(H\widehat{x} - y)\|_{\mathcal{H}}, \tag{4.57}$$

which leads to (4.48). In addition, according to (4.54) and (4.33),

$$(1 - \tau_\gamma)(1+\gamma\nu) \geqslant \gamma\left(\nu + \Big(2 - \frac{\gamma}{\eta}\Big)\lambda_{\min}\right) > 0. \tag{4.58}$$

By noticing that

$$\sup_{\gamma\in]0,2\eta[} \nu + \Big(2 - \frac{\gamma}{\eta}\Big)\lambda_{\min} = \nu + 2\lambda_{\min}, \tag{4.59}$$

the upper bound on χ is obtained. □

Remark 4.3.3.2

(i) Under the assumptions of the above proposition, we deduce from (4.48) that

$$\|\widetilde{x} - \widehat{x}\|_{\mathcal{H}} \leqslant \chi \, \|H^* - \overline{K}\|_{\mathcal{G},\mathcal{H}} \|H\widehat{x} - y\|_{\mathcal{H}}. \tag{4.60}$$

This upper bound indicates that the error depends on the data error (accounting for noise and modeling errors) and the norm of the mismatch on the adjoint.

(ii) In addition, the parameter χ depends on the strong convexity modulus ν and on the quadratic regularization parameter κ. Indeed, the larger κ, the larger $\lambda_{\min}$. The upper bound in (4.49) shows that increasing ν or κ allows us to decrease the distance to the true minimizer $\hat{x}$. At the same time, these parameters control the regularization term in (4.1) so that large values of them can introduce a bias in the recovery of the true signal. One should therefore seek values of these parameters balancing these two effects.

Remark 4.3.3.3 It follows from [77, Theorem 3.3] that, when $g = 0$ and $H^*H + \kappa \operatorname{Id}$ is invertible,

$$\|\widetilde{x} - \widehat{x}\|_{\mathcal{H}} \leqslant \frac{1}{\kappa} \, \|(H^* - \overline{K})(H\widehat{x} - y)\|_{\mathcal{H}} + o(\|H^* - \overline{K}\|_{\mathcal{G},\mathcal{H}}). \tag{4.61}$$

This bound is less tight than the one in (4.48)-(4.49) if $2\lambda_{\min} > \kappa \Leftrightarrow \kappa > -2\widetilde{\lambda}_{\min}$, where $\widetilde{\lambda}_{\min}$ is the minimum spectral value of $(\overline{K}H + H^*\overline{K}^*)/2$.

We present a first result concerning the averageness properties of T_γ with $\gamma \in]0, +\infty[$.

Lemma 4.3.3.4 *Let $\eta \in]0, +\infty[$, let $\gamma \in]0, 2\eta[$, let*

$$\overline{\alpha} = \frac{1}{2 - \gamma/(2\eta)} \in \Big]\frac{1}{2}, 1\Big[, \tag{4.62}$$

and let $\overline{W} = \operatorname{Id} - \gamma L$. The following properties hold.

$$L \text{ is } \eta\text{-cocoercive} \tag{4.63}$$

$$\Leftrightarrow \overline{W} \text{ is } \gamma/(2\eta)\text{-averaged} \tag{4.64}$$

$$\Rightarrow (\forall x \in \mathcal{H}) \quad \|\overline{W}x - 2(1-\overline{\alpha})x\|_{\mathcal{H}} + \|\overline{W}x\|_{\mathcal{H}} \leqslant 2\overline{\alpha}\|x\|_{\mathcal{H}} \tag{4.65}$$

$$\Rightarrow T_\gamma \text{ is } \overline{\alpha}\text{-averaged.} \tag{4.66}$$

Proof: If $\gamma < 2\eta$ and L is η-cocoercive, then the first equivalence holds [16, Proposition 4.39].
Let us now show that (4.64) implies (4.65). Set $\alpha = \gamma/(2\eta)$. Since $\overline{W}$ is α-averaged, there exists a nonexpansive operator $Q\colon \mathcal{H} \to \mathcal{H}$ such that $\overline{W} = (1-\alpha)\operatorname{Id} + \alpha Q$. We have then, for every $x \in \mathcal{H}$,

$$\begin{aligned}
&\|\overline{W}x - 2(1-\overline{\alpha})x\|_{\mathcal{H}} + \|\overline{W}x\|_{\mathcal{H}} \\
&= \|(1-\alpha)x + \alpha Qx - 2(1-(2-\alpha)^{-1})x\|_{\mathcal{H}} + \|(1-\alpha)x + \alpha Qx\|_{\mathcal{H}} \\
&= \alpha\|Qx - (1-\alpha)(2-\alpha)^{-1}x\|_{\mathcal{H}} + \|(1-\alpha)x + \alpha Qx\|_{\mathcal{H}} \\
&= \alpha\sqrt{\|Qx\|_{\mathcal{H}}^2 + \Big(\frac{1-\alpha}{2-\alpha}\Big)^2\|x\|_{\mathcal{H}}^2 - 2\frac{1-\alpha}{2-\alpha}\langle x \mid Qx\rangle_{\mathcal{H}}} \\
&\quad + \sqrt{\alpha^2\|Qx\|_{\mathcal{H}}^2 + (1-\alpha)^2\|x\|_{\mathcal{H}}^2 + 2\alpha(1-\alpha)\langle x \mid Qx\rangle_{\mathcal{H}}} \\
&\leqslant \varphi(\theta)\|x\|_{\mathcal{H}},
\end{aligned} \tag{4.67}$$

where

$$\varphi(\theta) = \alpha\sqrt{1 + \Big(\frac{1-\alpha}{2-\alpha}\Big)^2 - 2\frac{1-\alpha}{2-\alpha}\theta} + \sqrt{\alpha^2 + (1-\alpha)^2 + 2\alpha(1-\alpha)\theta}.$$

In the last inequality, we have set $\langle x \mid Qx\rangle_{\mathcal{H}} = \theta\|x\|_{\mathcal{H}}\|Qx\|_{\mathcal{H}}$ and used the nonexpansiveness of Q. Let us now study function φ on $[-1,1]$. The derivative φ' of this function is such that

$$\alpha^{-1}(1-\alpha)^{-1}\varphi'(\theta) = \frac{1}{\sqrt{\alpha^2 + (1-\alpha)^2 + 2\alpha(1-\alpha)\theta}} - \frac{1}{\sqrt{(2-\alpha)^2 + (1-\alpha)^2 - 2(1-\alpha)(2-\alpha)\theta}}.$$

Therefore, $\varphi'(\theta) \geqslant 0$ since

$$\begin{aligned}\alpha^2 + (1-\alpha)^2 + 2\alpha(1-\alpha)\theta &\leqslant (2-\alpha)^2 + (1-\alpha)^2 - 2(1-\alpha)(2-\alpha)\theta\\ &\Leftrightarrow \theta \leqslant 1,\end{aligned} \tag{4.68}$$

where we used the fact that $\alpha \in]0,1[$. This shows that φ is increasing on $[-1,1]$ and we deduce from Lemma 4.67 that

$$\|\overline{W}x - 2(1-\overline{\alpha})x\|_{\mathcal{H}} + \|\overline{W}x\|_{\mathcal{H}} \leqslant \varphi(1)\|x\|_{\mathcal{H}} = \frac{2}{2-\alpha}\|x\|_{\mathcal{H}} = 2\overline{\alpha}\|x\|_{\mathcal{H}}. \tag{4.69}$$

Since $\operatorname{prox}_{\gamma g}$ is firmly nonexpansive, it follows from [60, Theorem 3.8], that when (4.65) holds, T_γ is $\overline{\alpha}$-averaged. □

Given the above properties, the convergence of the mismatched PGA is guaranteed by the following result.

Proposition 4.3.3.5 Assume that L is η-cocoercive with $\eta \in]0,+\infty[$. Let $\gamma \in]0, 2\eta[$ and $\delta = 2 - \gamma/(2\eta)$. Let $(\theta_n)_{n\in\mathbb{N}}$ be a sequence in $[0,\delta]$ such that $\sum_{n\in\mathbb{N}} \theta_n(\delta - \theta_n) = +\infty$. Suppose that $\mathcal{F} \neq \varnothing$. Then the sequence $(x_n)_{n\in\mathbb{N}}$ generated by Algorithm (4.3) converges weakly to a point $\widetilde{x} \in \mathcal{F}$. In addition, if $\lambda_{\min} \neq 0$ and, for every $n \in \mathbb{N}$, $\omega_n = 0$ and $\theta_n \in [\underline{\theta}, 1]$ with $\underline{\theta} \in]0,+\infty[$, then $(x_n)_{n\in\mathbb{N}}$ converges linearly.

Proof: For every $n \in \mathbb{N}$, let $W_n = (1-\gamma\kappa)\operatorname{Id} - \gamma K_n H$, let $b_n = \gamma K_n y$, let $\overline{W} = \operatorname{Id} - \gamma L$, and let $\overline{b} = \gamma\overline{K}y$. Then (4.3) reads, for every $n \in \mathbb{N}$,

$$x_{n+1} = x_n + \theta_n\big(\operatorname{prox}_{\gamma g}(W_n x_n + b_n) - x_n\big). \tag{4.70}$$

The algorithm can thus be interpreted as an instance of the recurrent neural network investigated in [60] with $m = 1$ layer. It follows from Lemma 4.3.3.4 that [60, Condition 3.1] holds. In addition, as a consequence of Assumption 4.2.0.1, [60, Assumption 5.1] is satisfied since

$$\sum_{n\in\mathbb{N}} \|W_n - \overline{W}\|_{\mathcal{H},\mathcal{H}} \leqslant \gamma\|H\|_{\mathcal{H},\mathcal{H}} \sum_{n\in\mathbb{N}} \omega_n < +\infty \tag{4.71}$$

$$\sum_{n\in\mathbb{N}} \|b_n - \overline{b}\|_{\mathcal{H}} \leqslant \gamma\|y\|_{\mathcal{G}} \sum_{n\in\mathbb{N}} \omega_n < +\infty. \tag{4.72}$$

The convergence of $(x_n)_{n\in\mathbb{N}}$ to $\widetilde{x} \in \mathcal{F}$ can then be deduced from [60, Theorem 5.4].

Assume now that $\lambda_{\min} \neq 0$, $(\forall n \in \mathbb{N})$ $\omega_n \equiv 0$ and $\theta_n \in [\underline{\theta}, 1]$. It follows from (4.3) and (4.50) that, for every $n \in \mathbb{N}$,

$$x_{n+1} - \widetilde{x} = (1-\theta_n)(x_n - \widetilde{x}) + \theta_n\big(\operatorname{prox}_{\gamma g}\big((\operatorname{Id} - \gamma L)x_n + \gamma\overline{K}y\big) - \operatorname{prox}_{\gamma g}\big((\operatorname{Id} - \gamma L)\widetilde{x} + \gamma\overline{K}y\big)\big). \tag{4.73}$$

Using the nonexpansivity of the proximity operator and the triangle inequality yield

$$\begin{aligned}\|x_{n+1}-\widetilde{x}\|_{\mathcal{H}} &\leqslant (1-\theta_n)\|x_n-\widetilde{x}\|_{\mathcal{H}}+\theta_n\|(\mathrm{Id}\,-\gamma L)(x_n-\widetilde{x})\|_{\mathcal{H}}\\ &\leqslant (1-\theta_n+\theta_n\|\,\mathrm{Id}\,-\gamma L\|_{\mathcal{H},\mathcal{H}})\|x_n-\widetilde{x}\|_{\mathcal{H}}. \end{aligned} \tag{4.74}$$

By using now Proposition 4.3.1.6, we deduce that

$$\begin{aligned}\|x_{n+1}-\widetilde{x}\|_{\mathcal{H}} &\leqslant \left(1-\theta_n+\theta_n\sqrt{1-\gamma\Big(2-\frac{\gamma}{\eta}\Big)\lambda_{\min}}\right)\|x_n-\widetilde{x}\|_{\mathcal{H}}\\ &\leqslant \rho\|x_n-\widetilde{x}\|_{\mathcal{H}}, \end{aligned} \tag{4.75}$$

where

$$\rho = 1-\left(1-\sqrt{1-\gamma\Big(2-\frac{\gamma}{\eta}\Big)\lambda_{\min}}\right)\underline{\theta}\in]0,1[. \tag{4.76}$$

We deduce that, for every $n\in\mathbb{N}$, $\|x_n-\widetilde{x}\|_{\mathcal{H}}\leqslant\rho^n\|x_0-\widetilde{x}\|_{\mathcal{H}}$, which shows the linear convergence of $(x_n)_{n\in\mathbb{N}}$. □

We now see that the cocoercivity constant of L is useful to obtain an upper bound on the gradient descent parameter.

Remark 4.3.3.6 If L is self-adjoint positive (i.e. $\beta=0$ and $\lambda_{\min}\geqslant 0$), then it follows from Proposition 4.3.1.2 that L is η-cocoercive with $1/\eta=\lambda_{\max}=\|L\|_{\mathcal{H},\mathcal{H}}$. Proposition 4.3.3.5 thus leads to $2/\|L\|_{\mathcal{H},\mathcal{H}}$ as a strict upper bound on step size γ to guarantee the convergence of the algorithm. This allows us to recover the classical upper bound on the step size for Algorithm (4.3) in the special case when $\overline{K}=H^*$.

Remark 4.3.3.7 When $g=0$, $\theta_n\equiv 1$, and $\mathcal{H}=\mathbb{R}^N$, (4.3) becomes a linear recursive equation and tools from matrix analysis can be employed to derive the following necessary and sufficient convergence conditions [77, Theorem 3.1]:

$$(\forall j\in\mathbb{J})\quad \gamma<2\frac{\mathrm{Re}\,\zeta_j}{|\zeta_j|^2} \tag{4.77}$$

$$\mathrm{Re}\,\zeta_j>0, \tag{4.78}$$

where $(\zeta_j)_{i\in\mathbb{J}}$ are the nonzero eigenvalues of L. It is easy to show that, for every $j\in\mathbb{J}$, $\lambda_{\min}\leqslant\mathrm{Re}\,\zeta_j$. Therefore, if $\lambda_{\min}>0$, (4.78) is satisfied. Then, it follows from Propositions 4.3.1.2(ii) and Corollary 4.3.1.4(i) that a sufficient and necessary condition for L to be η-cocoercive is $\eta\leqslant\overline{\eta}$ where $\overline{\eta}$ is given by (4.32). Since Proposition 4.3.3.5 guarantees the convergence of (4.3) when $\gamma\in]0,2\overline{\eta}[$, we deduce that

$$(\forall j\in\mathbb{J})\quad \overline{\eta}\leqslant\frac{\mathrm{Re}\,\zeta_j}{|\zeta_j|^2}. \tag{4.79}$$

This emphasizes that, in the presence of adjoint mismatch, the cocoercivity of L only provides a sufficient condition for the convergence of PGA.

4.4 Application

We now show the applicability of the proposed approach through two examples of reconstruction of piecewise constant numerical phantoms where instability arises from the presence of truncation. The FP is ray-driven, while the BP is pixel-driven. In our first experiment, a geometric abdomen phantom is reconstructed using a wavelet-based regularization. In the second experiment, we perform a joint reconstruction and segmentation of a metallic device (e.g., needles) present in a region of interest of another geometric phantom from sub-sampled projections. Here we relied on a geometric decomposition of the phantom into two components (a needle and a background).

4.4.1 Reconstruction of a geometric abdomen from undersampled projections

4.4.1.1 Problem statement:

We simulated a scan of an abdomen of size 45 cm, made of a vertebra set to 3000, metallic inserts ranging from 4000 sHU to 4500 sHU, and a liver area set to 1840 sHU. The source-to-object and source-to-image distances were respectively set to 800 mm and 1200 mm leading to a magnification factor of 1.5, as can be found on clinical scanners. The associated sinogram was computed in fan-beam geometry over 180° using 50 regularly spaced angular steps. The projection and backprojection operators were rescaled by $\pi/50$ to make the parametrization independent from the number of projections. The detector has 62 bins of size 6.4 mm, so that $M = 62 \times 50$, and the image is reconstructed on a discrete grid of $N = 128 \times 128$ pixels, with size $1.5 \times 6.4 = 4.26$ mm. The image reconstruction problem is undetermined due to the small detector FOV and the limited angular coverage. The noise standard deviation is set to $\sigma = 0.69$, so that $\|b\|_{\mathcal{G}}/\|H\overline{x}\|_{\mathcal{G}} \approx 6.3 \times 10^{-5}$. Figure 4.1 shows the phantom $\overline{x}$ and the data y.

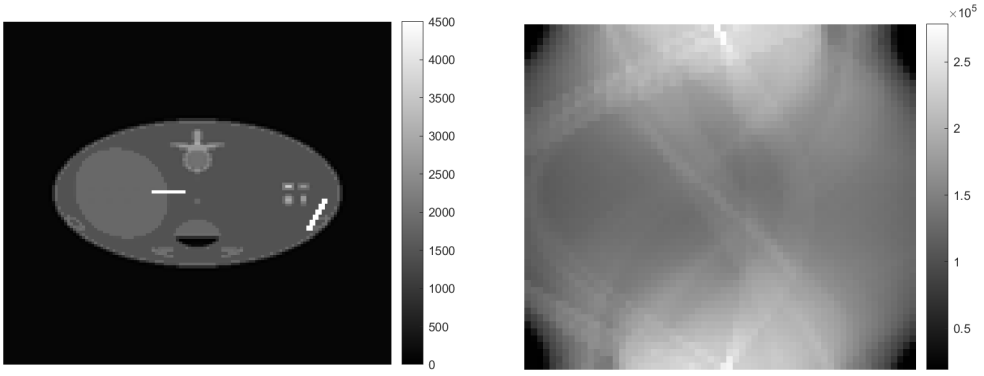


Figure 4.1: Phantom $\overline{x}$ (left) and sinogram y (right)

With those settings, H^* contains 1.08% nonzero elements, whereas this proportion decreases to 0.89% for $\overline{K}$. The coupling ratio is $\xi = 1.151$, and the asymmetry metric μ equals 0.159. Figure 4.2 shows the backprojection of constant measurements at a single angle using either $\overline{K}$ or H^*. A high-frequency Moire pattern is visible when using H^* (right image) due to the redundancy introduced by oversampling the projection. In contrast, the backprojected view remains uniform with $\overline{K}$ (left image).
An estimate of $\overline{x}$ is obtained by adopting a sparse inducing compressed-sensing formulation. We solve the penalized least squares Problem (4.1) with $g = \rho\|W \cdot\|_1$, $W \in \mathbb{R}^{N\times N}$

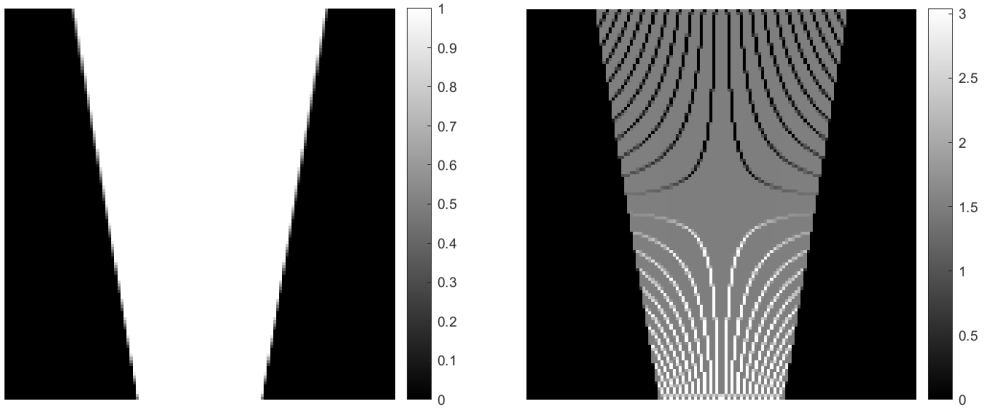


Figure 4.2: Backprojection of a uniform view with $\overline{K}$ (left) and H^* (right)

being the orthogonal Symlet 2 wavelet transform on 2 resolution levels, and $\rho > 0$ the associated regularization parameter.
We ran Algorithms (4.2) and (4.3), for two settings κ_1 and κ_2 of parameter κ such that L is not cocoercive with κ_1, but becomes cocoercive with κ_2. In such case, the condition given in Proposition 4.3.2.1(iii)(e) holds, which proves the existence of a unique fixed point of scheme (4.3) and its convergence is ensured according to Proposition 4.3.3.5. We set $\kappa_1 = 10^{-2}$. Moreover, following Remark 4.3.1.5, κ_2 is set as $-\widetilde{\lambda}_{\min} + 10^{-2}$. The eigenvalue $\widetilde{\lambda}_{\min} = -1.61$ is computed using the Matlab function eigs. Note that although matrices H and $\overline{K}$ were stored in these experiments, matrix-free iterative methods can be used to compute the dominant and the smallest eigenvalues of operator $(L + L^*)/2$, thus complying with practical implementations of the FP-BP pair for higher dimensional problems. Moreover, to bypass the need for the exact adjoints of H and $\overline{K}^*$ while computing minimum eigenvalues, we refer to the strategy in [77]. We additionally set the regularization hyperparameter ρ to 600 and the relaxation parameter $\theta_n \equiv 1$.
For the coupled settings (H^*, κ_1), $(\overline{K}, \kappa_1)$ and (H^*, κ_2), step size γ was set to $1.9/(\|H\|^2_{\mathcal{H},\mathcal{H}} + \kappa) = 2.9 \times 10^{-3}$. For $(\overline{K}, \kappa_2)$, γ is chosen equal to 1.82×10^{-5} in accordance with Corollary 4.3.1.4 and Proposition 4.3.3.5. The algorithms are run until a stopping precision on the relative distance between two consecutive iterates is below 10^{-7} or a maximum number of iterations of 10^4 is reached.

4.4.1.2 Results:

Figure 4.3 displays the normalized mean square error (NMSE) defined as $(\|\overline{x} - x_n\|_{\mathcal{H}} / \|\overline{x}\|_{\mathcal{H}})_n$, computed along the iterations when applying Algorithms (4.2) and (4.3). The plots confirm that with value κ_1, PGA converges when the exact adjoint H^* is used but diverges when H^* is replaced by $\overline{K}$, as was expected from our theoretical analysis. In the latter case, Algorithm (4.3) shows an initial convergence trend that reaches a minimum discrepancy point close to the minimizer obtained with H^* before diverging. For value κ_2, both Algorithms (4.2) and (4.3) converge to fixed points that are close to each other, again confirming our theoretical analysis. The corresponding NMSE values are 0.4432 and 0.4572, respectively. PGA without mismatch requires fewer iterations to reach convergence than its perturbed version using $\overline{K}$. Note that, in a real context, practitioners often use early stopping to avoid the potential negative effects of the adjoint mismatch. Nevertheless, it is difficult for the user to know when the iterations should be stopped to reach this good intermediary solution; hence the result is often suboptimal. Our analysis

shows that one can still use an inaccurate adjoint without using an empirical rule.

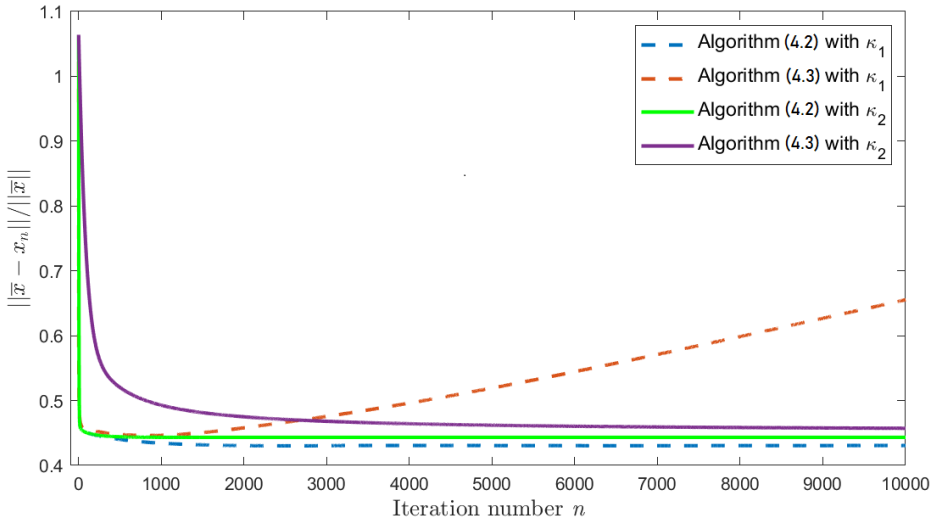


Figure 4.3: Decay of the error along iterations for Algorithms (4.2) and (4.3) and two choices of κ parameter.

Reconstruction results are displayed in Figure 4.4 and Figure 4.5 using the same windowing. Let us remark that, due to a short detector, the projections suffer from axial truncation. The set of pixels of the image whose projections belong to the detector FOV then defines the image FOV. We added a comparison with two reconstructions obtained from the standard FBP approach in Figure 4.6. The image is obtained from FBP by zero-padding the sinogram on the left. On the right, the borders of the sinogram have been replicated before FBP. Only the image FOV is depicted since the FBP reconstruction outside this zone is irrelevant. We also indicated the NMSE and the maximum absolute error (MAE), defined as $\max_{i\in\{1,\dots,N\}} |\overline{x}_i - x_i|$, for all the reconstructed images when compared with the ground truth. Both FBP reconstructions suffer from various artifacts (peripheral bright-band artifacts, cupping, over-estimation of the values as shown in Figure 4.7) [163,181], in contrast with the solutions provided by our regularized iterative approach. Furthermore, when parameter κ_2 is used, the reconstructed image obtained by PGA with the mismatched adjoint $\overline{K}$ is very similar to the image obtained without mismatch. In contrast, combining the setting κ_1 with the mismatched adjoint in PGA yields a reconstruction that is deteriorated by artifacts propagating from the exterior of the FOV. It leads to a higher NMSE compared to the solution obtained when using the exact H^* as shown on the reconstructed image in Figure 4.4 (bottom left) and the FOV error map in Figure 4.8 (top right). As soon as the convergence of PGA is ensured, an unmatched FP/BP pair gives a similar reconstruction quality to the matched pair but may lead to a slower convergence. Let us emphasize that, in practice, the decrease in the convergence rate in terms of iterations could be compensated by a reduced computation cost for operator $\overline{K}$. Finally, note that computing the infimum in (4.49) in Theorem 4.3.3.1 with a grid search gives an upper bound of 3.25×10^5. The actual error is 1.0934×10^4, which is indeed lower than this upper bound, as expected from our theoretical analysis.

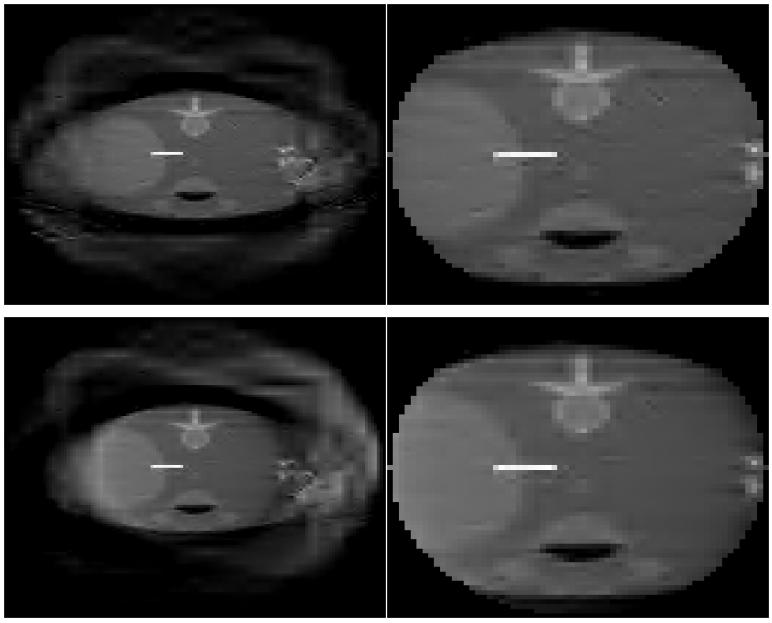


Figure 4.4: Reconstructions (left) and zoomed versions within the FOV (right) obtained using κ_1 and either Algorithm (4.2), NMSE = 0.1207, MAE = 2330 (top) or Algorithm (4.3), NMSE = 0.1610, MAE = 3141 (bottom).

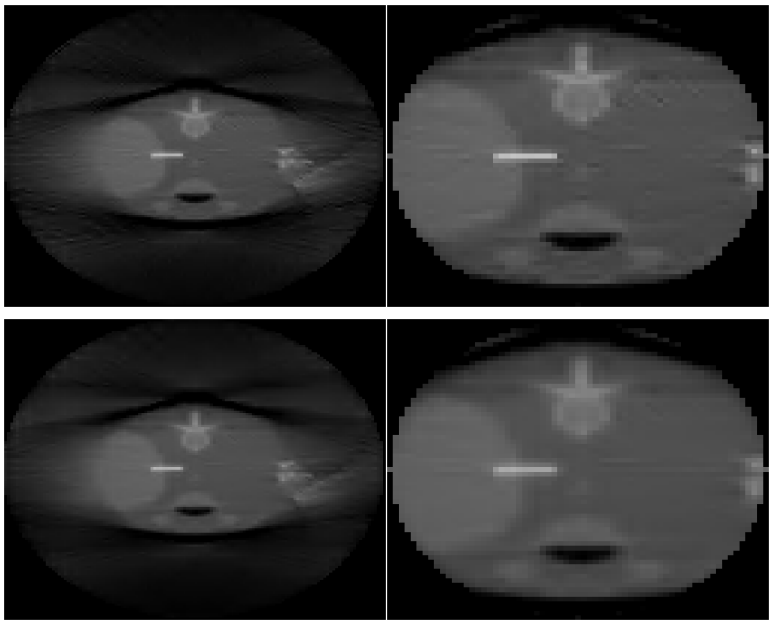


Figure 4.5: Reconstructions (left) and zoomed versions within the FOV (right) obtained using κ_2 and either Algorithm (4.2), NMSE = 0.16, MAE = 2205 (top) or Algorithm (4.3), NMSE = 0.1534, MAE = 2399 (bottom).

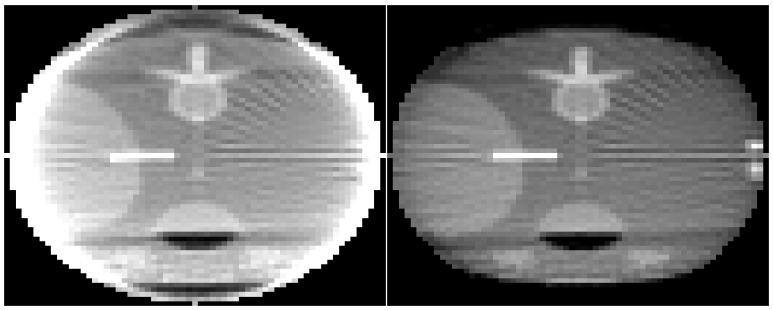


Figure 4.6: FBP reconstructions, in the FOV, with zero-padded FBP, NMSE = 1.776, MAE = 8534 (left) and extrapolated FBP by replicating the borders of the sinogram, NMSE = 0.366, MAE = 1871 (right).

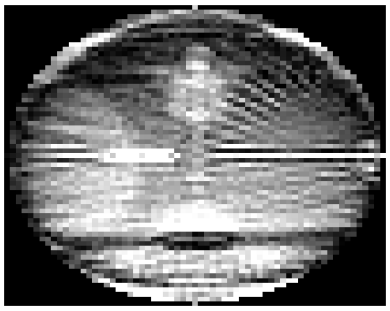


Figure 4.7: Absolute difference between the reconstructed image from FBP using replicated sinogram borders and the ground truth within the FOV.

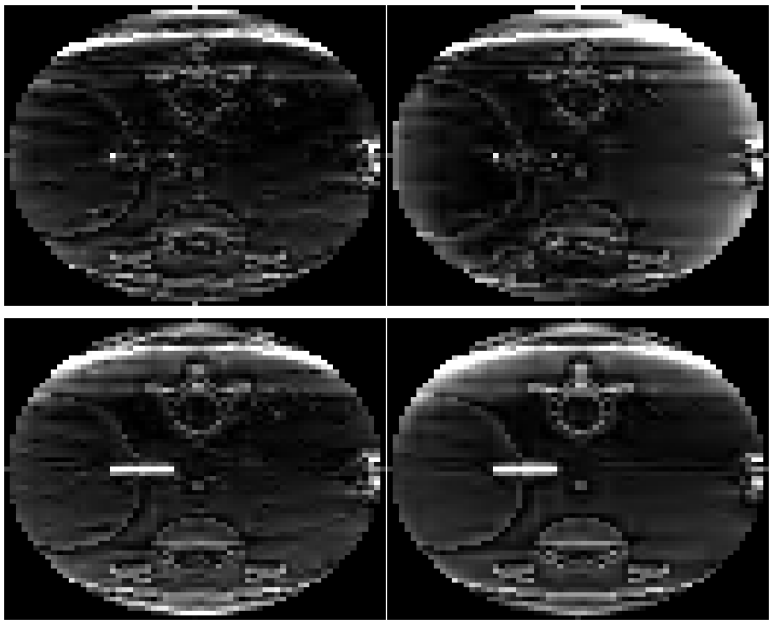


Figure 4.8: Absolute difference between the reconstructed image and the ground truth, within the FOV, using κ_1 (top) or κ_2 (bottom), and either Algorithm (4.2) (left) or Algorithm (4.3) (right).

4.4.2 Joint object-background decomposition and reconstruction

4.4.2.1 Problem statement:

This example focuses on a joint reconstruction and decomposition task. Flat-panel detectors commonly sample projections with small pixels but at a slow frame rate, so the angular sampling is comparatively poor. Then reconstructing the entire object on a fine grid is time-consuming and produces large volumes that are also difficult to manipulate. We thus look at reconstructing a relevant ROI only, as is the case when the clinical goal is to assess the precise position of metallic needles within a soft-tissue background. A priori knowledge about the device (e.g., sparsity, high contrast, and direction [125]) can be used, given that the object is separated from the background.
The phantom grid is of 256×256 pixels of size 0.53 mm, of which the ROI, denoted by $\overline{x}_{\mathrm{r}}$, is a patch of size 88×88. The simulated phantom $\overline{x}$ and the ROI are displayed in Figure 4.9. The phantom projection is computed for a detector of 500 bins of 0.4 mm. The detector bins are sampled on a twice thinner grid than the pixels. The number of uniformly spaced angular positions is set to 100 only over the interval $[0°, 180°]$ leading to $M = 100 \times 500$. The source-to-object and source-to-image distances were set as in our first experiment. The operators are rescaled by $\pi/100$. The noise standard deviation σ is chosen equal to 0.35, so that $\|b\|_{\mathcal{G}}/\|H\overline{x}\|_{\mathcal{G}} \approx 3.32 \times 10^{-5}$.
The acquired projections contain information regarding pixels outside the ROI. In order to reduce reconstruction artifacts, we define a larger reconstruction grid, with size $N = 140 \times 140$ containing the ROI. Let us introduce the sampling operator $S \in \mathbb{R}^{88^2 \times 140^2}$, which selects the ROI within this extended image. We then aim at decomposing the spacial contents $\overline{x}_{\mathrm{r}}$ within this ROI into two maps $S\overline{x}_{\mathrm{m}} \in \mathbb{R}^{88 \times 88}$ and $S\overline{x}_{\mathrm{b}} \in \mathbb{R}^{88 \times 88}$ which describe respectively the metal component of the ROI (needles) and the tissues of the ROI as shown in Figure 4.10, so that $\overline{x}_{\mathrm{r}} = S\overline{x}_{\mathrm{m}} + S\overline{x}_{\mathrm{b}}$.

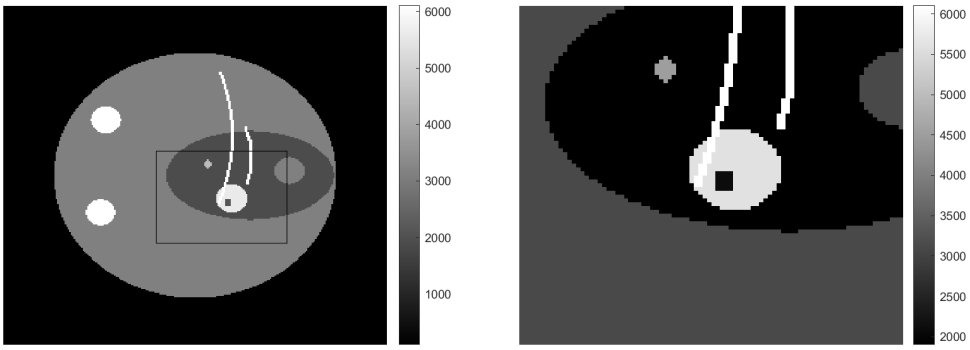


Figure 4.9: Phantom $\overline{x}$ (left) and ROI $\overline{x}_r$ (right).

Estimates of the two maps $(\overline{x}_{\mathrm{m}}, \overline{x}_{\mathrm{b}}) \in \mathbb{R}^{2N}$ on the extended grid of size N, are obtained by solving the following penalized least-squares problem:

$$\underset{(x_{\mathrm{m}}, x_{\mathrm{b}}) \in \mathbb{R}^{2N}}{\text{minimize}} \quad \frac{1}{2}\|y - H_r(x_{\mathrm{m}} + x_{\mathrm{b}})\|_2^2 + g(x_{\mathrm{m}}, x_{\mathrm{b}}) + \frac{\kappa}{2}(\|x_{\mathrm{m}}\|_2^2 + \|x_{\mathrm{b}}\|_2^2). \tag{4.80}$$

Furthermore, we define $g(x_{\mathrm{m}}, x_{\mathrm{b}}) = \rho \mathrm{DTV}_{\Omega}(x_{\mathrm{m}}) + \beta \mathrm{TV}(x_{\mathrm{b}}) + \alpha\|x_{\mathrm{m}}\|_1 + \iota_{[0,+\infty[^N}(x_{\mathrm{m}}) + \iota_{[0,+\infty[^N}(x_{\mathrm{b}})$ where ι_C denotes the indicator function of set C and $(\rho, \beta, \alpha, \kappa) \in [0, +\infty[^4$

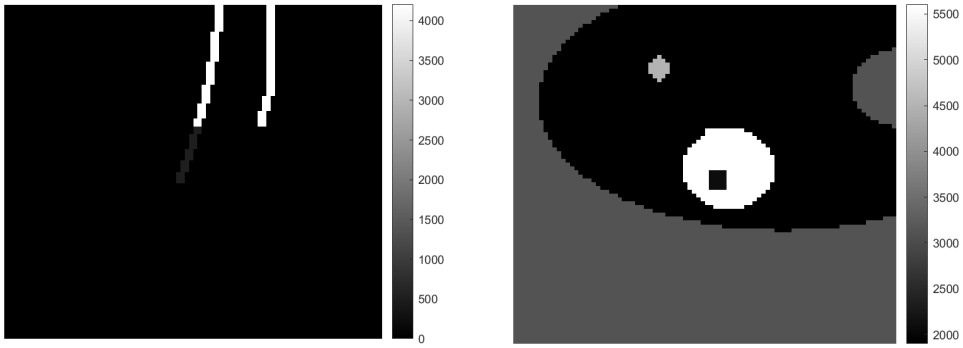


Figure 4.10: $S\overline{x}_m$ (left) and $S\overline{x}_b$ (right).

and the TV term, acting on the background image, is defined as

$$(\forall u \in \mathbb{R}^N) \quad \mathrm{TV}(u) = \sum_{i=1}^{N} \sqrt{(\Delta_i^{\mathrm{h}} u)^2 + (\Delta_i^{\mathrm{v}} u)^2} \tag{4.81}$$

with $\Delta_i^{\mathrm{h}} \in \mathbb{R}^N$, $\Delta_i^{\mathrm{v}} \in \mathbb{R}^N$, the horizontal and vertical discrete gradient operators at location i (assuming zero-padding), respectively. Furthermore, given that $\overline{x}_{\mathrm{m}}$ is sparse, and contains needles following about the same direction, we use both an ℓ_1 penalty and the directional total variation introduced in [17], defined, for every $u \in \mathbb{R}^N$, as $\mathrm{DTV}_\Omega(u) = \sum_{i=1}^{N} \|(\nabla_\Omega u)_i\|$ where $\nabla_\Omega \in \mathbb{R}^{2\times N}$ computes the two directional derivatives at the pixel i, parameterized by an angular direction $\theta \in [0^\circ, 180^\circ[$, and a scaling factor $s > 0$ ($\Omega = \{\theta, s\}$), i.e.

$$(\nabla_\Omega u)_i = \begin{pmatrix} 1 & 0 \\ 0 & s \end{pmatrix} \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \Delta_i^{\mathrm{h}} u \\ \Delta_i^{\mathrm{v}} u \end{pmatrix}. \tag{4.82}$$

Let $H = (H_r, H_r)$ in $\mathbb{R}^{M\times 2N}$. Equation (4.80) can be rewritten as

$$\underset{z=(x_{\mathrm{m}}^\top, x_{\mathrm{b}}^\top)^\top \in \mathbb{R}^{2N}}{\text{minimize}} \quad \frac{1}{2}\|y - Hz\|_2^2 + h(z) + \frac{\kappa}{2}\|z\|_2^2 + i_{[0,+\infty[^{2N}}(z), \tag{4.83}$$

with $h\colon z = (x_{\mathrm{m}}^\top, x_{\mathrm{b}}^\top)^\top \mapsto \rho\mathrm{DTV}_\Omega(x_{\mathrm{m}}) + \alpha\|x_{\mathrm{m}}\|_1 + \beta\mathrm{TV}(x_{\mathrm{b}})$.
The coupling ratio between H and its associated adjoint approximation $\overline{K} = (\overline{K}_{\mathrm{r}}, \overline{K}_{\mathrm{r}})$ is $\xi = 0.75$ and the asymmetry metric is $\mu = 0.0418$. The proximity operator of h does not have a closed form; hence it is approximated by using inner iterations of the dual forward-backward algorithm [55, 56] with a stopping precision of 10^{-8}. We set $\theta_n \equiv 1$, $\rho = 5500$, $\beta = 2950$, $\alpha = 500$, $s = 0.2$, and $\theta = 10^\circ$. Initial estimates for both maps are zero-valued. As in our first experiment, two values of κ are tested, namely $\kappa_1 = 10^{-2}$ and $\kappa_2 = 0.2438$. L is guaranteed to be cocoercive for $\kappa = \kappa_2$, but not for $\kappa = \kappa_1$. Here again, the existence and uniqueness of the fixed point of scheme (4.3) are guaranteed for $\kappa = \kappa_2$ because the condition in Proposition 4.3.2.1(iii)(e) is fulfilled. Furthermore, for the settings (H^*, κ_1), $(\overline{K}, \kappa_1)$ and (H^*, κ_2), the step size γ is set respectively to 2×10^{-3} while for $(\overline{K}, \kappa_2)$, γ is set to 1.5×10^{-3}. The stopping precision on the relative distance between two consecutive iterates is 10^{-7}, and the maximum number of iterations is 2×10^4.

4.4.2.2 Results:

In Figure 4.11 and Figure 4.12, we plot the relative errors between the ground truth metal map $S\overline{x}_{\mathrm{m}}$ and tissues map $S\overline{x}_{\mathrm{b}}$, cropped to the ROI, and their estimates along the iterations. In Figure 4.12, one sees that the iterates obtained from (4.3) with κ_1 are unstable. Oscillations hamper the convergence of scheme (4.3). The stopping convergence criterion is never met, and the maps cannot be reconstructed at the end of the 2×10^4 iterations. Figure 4.11 shows that for the three other cases, the algorithm stops in a phase where the errors associated with both maps are simultaneously decreasing. These plots confirm that, with setting κ_1, only Algorithm (4.2) (i.e., PGA without adjoint mismatch) converges. For κ_2, Algorithms (4.2) and (4.3) converge to two fixed points that are quite close to each other and the exact solution. Figure 4.13 shows the reconstructed maps within the ROI, obtained with Algorithm (4.2) and κ_1, Algorithm (4.2) and κ_2, Algorithm (4.3) and κ_2. Upon visual inspection, the two restored components $S\hat{x}_{\mathrm{b}}$ and $S\hat{x}_{\mathrm{m}}$ are efficiently separated and well reconstructed in all three cases. Furthermore, no visible deterioration arises on the images reconstructed with κ_2.

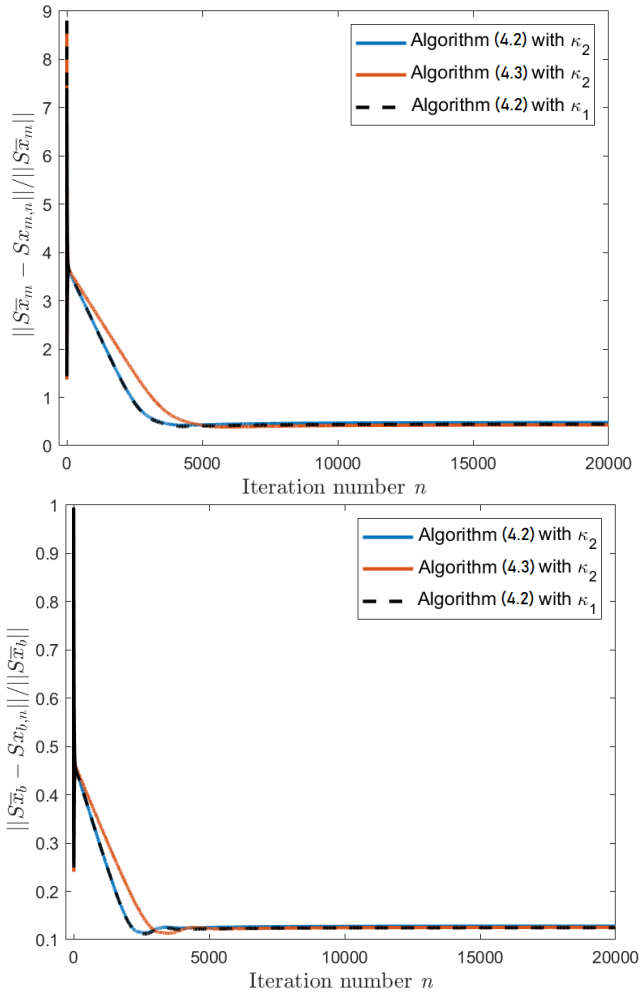


Figure 4.11: Evolution of the error, inside the ROI, of the metal and tissue maps $(Sx_{\mathrm{m,n}})_n$ and $(Sx_{\mathrm{b,n}})_n$ estimated along iterations by Algorithms (4.2) and two choices of κ parameter and Algorithm (4.3) with κ_2.

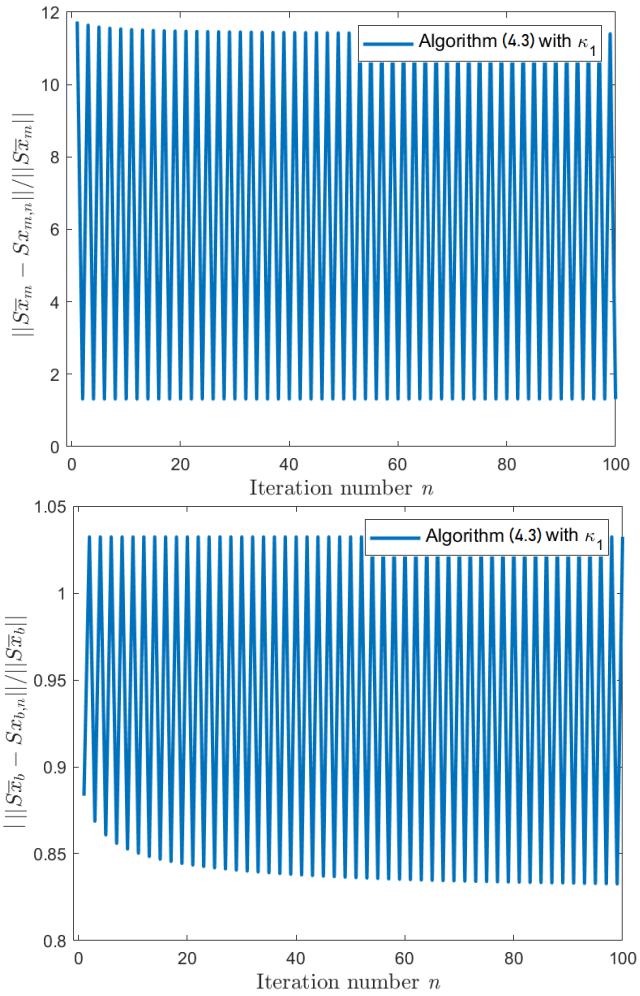


Figure 4.12: Evolution of the error, inside the ROI, of the metal and tissue maps $(Sx_{\mathrm{m,n}})_n$ and $(Sx_{\mathrm{b,n}})_n$ estimated along iterations by Algorithm (4.3) with κ_1.

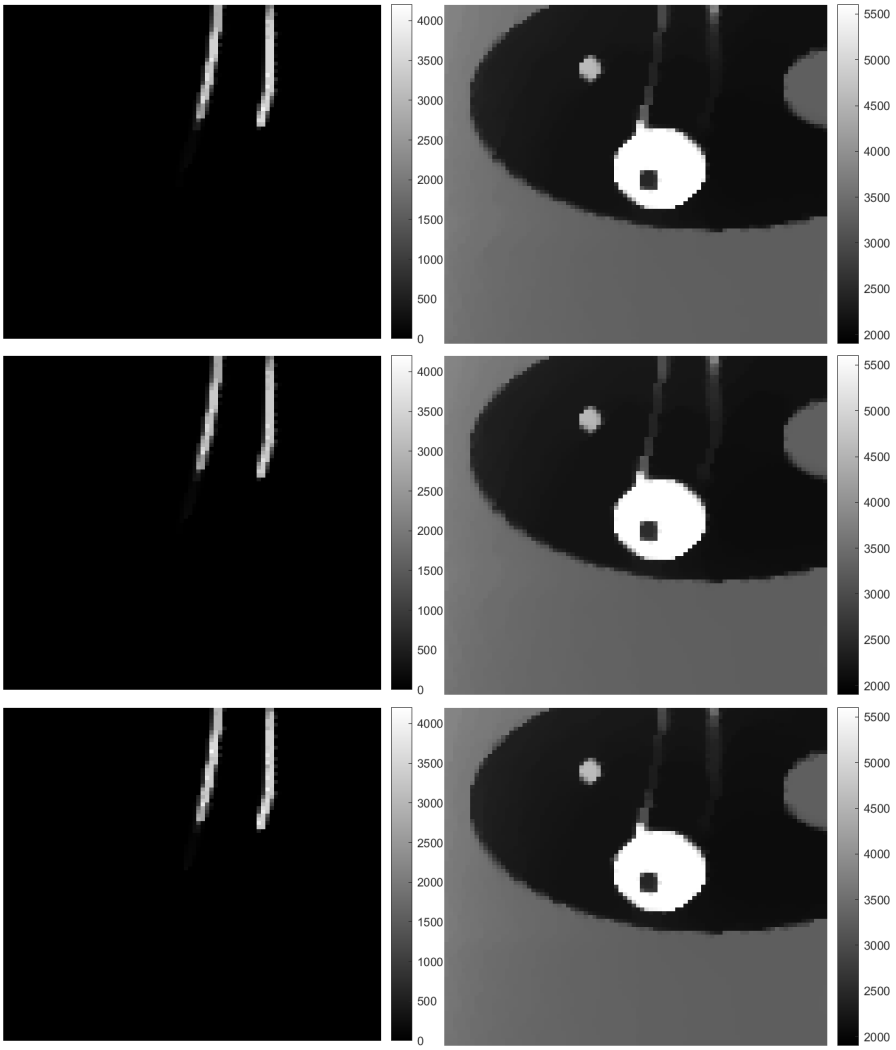


Figure 4.13: Reconstructed maps within the ROI $S\hat{x}_m$ (left) and $S\hat{x}_b$ (right) using κ_1 (first row) or κ_2 (last two rows), and either Algorithm (4.2) (first two rows) and Algorithm (4.3) (last row).

4.5 Unmatched preconditioning of the proximal gradient algorithm

The problem of adjoint mismatch is closely related to the use of unmatched preconditioning metrics in PGA.
As a first-order optimization method, the convergence of PGA is slow [46], bringing into question the relevance of the algorithm for high-dimensional applications. Preconditioning strategies (Chapter 3-subsection 3.3.2) offer a way to accelerate PGA. They consist of performing a change of metric to reduce the condition number of the linear operator (typically the Hessian) involved in updating the smooth part of the cost function. Preconditioning does not affect the fixed points of PGA. However, one main issue is that the metric used in the gradient term must theoretically also be included, via its inverse, in the proximity operator. Such an inversion can be computationally costly. Computationally cheap metrics resulting from rough approximations to the Hessian operator can be conveniently inverted (e.g., diagonal matrices), but they may fail to achieve significant acceleration. Moreover, the proximity operator might no longer have a closed form in the chosen metric. In this case, the proximity operator must be computed through inner iterations so that the extra computations outweigh the benefit in terms of convergence rate. Designing both effective and efficient preconditioners is then especially challenging for proximal methods [20].
This section investigates the use of unmatched preconditioners, which relax the constraints associated with preconditioning for penalized least-squares problems.

4.5.1 Preconditioning for CT reconstruction

We now consider a slightly more general problem than (4.1) with a weighted least-squares data fidelity term, and we focus on the case where $\mathcal{H} = \mathbb{R}^N$ and $\mathcal{G} = \mathbb{R}^M$:

$$\underset{x\in\mathbb{R}^N}{\text{minimize}} \ \frac{1}{2}\|y - Hx\|_W^2 + g(x) + \frac{\kappa}{2}\|x\|^2. \tag{4.84}$$

For ease of notation, we define $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ as the standard scalar product and the norm associated with $\mathbb{R}^N$.
Problem (4.84) can be solved by PGA preconditioned by $Q \in \mathcal{S}_N^+$ [27] and the resulting iteration reads, for every $n \in \mathbb{N}$,

$$x_{n+1} = (1-\theta_n)x_n + \theta_n \text{prox}_{\gamma g}^Q((x_n - \gamma Q^{-1}(Mx_n - H^\top W y))), \tag{4.85}$$

where M is the Hessian of the smooth part of the cost function:

$$M = H^\top W H + \kappa \,\text{Id}\ . \tag{4.86}$$

The convergence conditions of PGA, given in the previous section, can be adapted to the preconditioned scheme: if $\theta_n \in [\epsilon, 1]$ with $\epsilon \in]0, 1[$ and $\gamma \in]0, 2/\alpha[$, where α is the Lipschitz constant of the preconditioned gradient operator $Q^{-1}(M\cdot - H^\top W y)$, i.e. $\alpha = |||Q^{-1/2}MQ^{-1/2}|||$, then the sequence $(x_n)_{n\in\mathbb{N}}$ generated by Algorithm (4.85) converges to a solution to Problem (4.84).
We see that the convergence of PGA depends on the norm of the Hessian of the quadratic part of the objective function and, more particularly, on its maximum eigenvalue. A large value results in small gradient steps leading to slow convergence.
The criteria for identifying an effective preconditioner Q are twofold: (i) the action of

Q and Q^{-1} on an element $x \in \mathbb{R}^N$ should be easily computed; (ii) the conditioning number of $Q^{-1}M$ should be small. These criteria often limit the use of preconditioning to a simple form of regularizer g, e.g., a positivity constraint [26], an ℓ_1, or a quadratic penalty.
For CT, most preconditioners were proposed in the case when no proximity operator is involved [92,94,107,179,210]. Diagonal preconditioners [105] provide limited acceleration so Fourier preconditioners are preferred [54, 87, 92]. Preconditioning PGA when the precision matrix $W = \mathrm{Id}$ in Problem (4.84) derives from the Fourier-slice theorem (see Chapter 3), which states that the continuous version of the normal operator $M = H^\top H$ is a convolutional circulant operator. This theorem provides two inversion filters. The first is the 2D cone filter, which is applied after $H^\top$, yielding a potential choice of preconditioner Q^{-1} in Algorithm (4.85). However, this 2D filter requires infinite support and is computationally expensive. Many variants have been proposed to simplify [172] or improve it by considering the underlying sampling in $H^\top H$. Improved filters were obtained by approximating the SIRT algorithm [170], or in a learning framework [169, 205]. The second choice is the 1D ramp filter, which is simpler to compute and does not require infinite support. However, since it is applied before $H^\top$, it does not fit the form of Algorithm (4.85): the product $PM = PH^\top H$ is replaced by the more general structure $H^\ddagger H$. Both filters are particularly efficient because they provide a close approximate inversion of H. Regularized I-FBP/I-FDK methods, presented in Chapter 3, speed up convergence by using this approximate inversion. This method applies $H^\ddagger H$ rather than $H^\top H$ in the gradient step, followed by an unweighted proximity step. These methods thus relate to a version of Algorithm (4.85) where Q^{-1} is kept while Q is replaced by Id.
We now present the generic unmatched preconditioned version of Algorithm (4.85) under study.

4.5.2 Preconditioning with unmatched metrics

The standard form of preconditioned PGA with matrix Q converges in a reduced number of iterations for a suitably chosen Q and a limited panel of regularization functions. To include a more generic prior, we introduce a second matrix $P \in \mathbb{R}^{N\times N}$ such that $PM \neq 0$, that leads to a relaxed version of iteration (4.85) which reads

$$x_{n+1} = (1-\theta_n)x_n + \theta_n \mathrm{prox}^Q_{\gamma g}((x_n - \gamma P(Mx_n - H^\top W y))). \tag{4.87}$$

We say that P and Q are matched when $PQ = \mathrm{Id}$, hence we recover (4.85). Note that iteration (4.87) can be viewed as a deviation of (4.85) where the error involved in the argument of the proximity operator is $e(x_n)$ with

$$(\forall x \in \mathbb{R}^N) \quad e(x) = \gamma(Q^{-1} - P)(Mx - H^\top W y). \tag{4.88}$$

However, the assumption of summability of the error often adopted in the literature [64,192] is generally not satisfied in our context.

4.5.3 Adaptation of previous results

Similar to section 4.3, we introduce the nonlinear operator

$$T_\gamma \colon \mathbb{R}^N \to \mathbb{R}^N \colon x \mapsto \operatorname{prox}_{\gamma g}^{Q}((\mathrm{Id}\, -\gamma L)x + \gamma P H^\top W y). \tag{4.89}$$

Iteration (4.87) can be expressed more concisely as

$$x_{n+1} = (1-\theta_n)x_n + \theta_n T_\gamma x_n,$$

$$\text{with} \qquad L = P(H^\top W H + \kappa\, \mathrm{Id}\,) = PM \in \mathbb{R}^{N\times N}. \tag{4.90}$$

We also define $L_Q = Q^{1/2} L Q^{-1/2}$. It results from simple algebra that

$$|||L|||_Q = \sup_{x\in\mathbb{R}^N} \frac{\|Lx\|_Q}{\|x\|_Q} = |||L_Q||| \tag{4.91}$$

and the adjoint of L in metric $\|\cdot\|_Q$ is

$$L^* = Q^{-1} L^\top Q = Q^{-1/2} L_Q^\top Q^{1/2}. \tag{4.92}$$

We redefine the gap β, the spectral values $\lambda_{\min}$, $\lambda_{\min}^+$, $\lambda_{\max}$ as

$$\lambda_{\min} = \inf_{\substack{x\in\mathbb{R}^N \\ \|x\|_Q=1}} \langle x \mid Lx\rangle_Q = \inf_{\substack{x\in\mathbb{R}^N \\ \|x\|=1}} \langle x \mid L_Q x\rangle, \tag{4.93}$$

$$\lambda_{\min}^+ = \inf_{\substack{x\in(\operatorname{Ker} L_Q)^\perp \\ \|x\|=1}} \langle x \mid L_Q x\rangle,\ \lambda_{\max} = \sup_{\substack{x\in\mathbb{R}^N \\ \|x\|=1}} \langle x \mid L_Q x\rangle, \tag{4.94}$$

$$\beta = \frac{1}{2}|||L - L^*|||_Q = \frac{1}{2}|||L_Q - L_Q^\top|||. \tag{4.95}$$

The notion of cocoercivity can also be redefined in a weighted space.

Definition 4.5.3.1 Let $\eta \in]0,+\infty[$. L is η-cocoercive in $(\mathbb{R}^N, \|\cdot\|_Q)$ if

$$(\forall x \in \mathbb{R}^N) \qquad \langle x \mid Lx\rangle_Q \geqslant \eta \|Lx\|_Q^2. \tag{4.96}$$

We now provide conditions for this property to be satisfied.

Proposition 4.5.3.2 L is cocoercive in $(\mathbb{R}^N, \|\cdot\|_Q)$ with $\eta \in]0,+\infty[$ if and only if

$$\begin{cases} \lambda_{\min} \geqslant 0 & (4.97) \\ \operatorname{Ker}(L_Q + L_Q^\top) = \operatorname{Ker} L_Q. & (4.98) \end{cases}$$

Then, the maximum cocoercivity constant of L is

$$\overline{\eta} = \frac{2}{|||(\mathrm{Id}\, + (L_Q - L_Q^\top)(L_Q + L_Q^\top)^\dagger)(L_Q + L_Q^\top)^{1/2}|||^2}. \tag{4.99}$$

In particular, L is cocoercive in $(\mathbb{R}^N, \|\cdot\|_Q)$ with constant $\underline{\eta} = 1/\left(\sqrt{\lambda_{\max}} + \beta/\sqrt{\lambda_{\min}^+}\right)^2 \leqslant \overline{\eta}$.

Proof: For $z \in \mathbb{R}^N$, by setting $x = Q^{-1/2}z$, it follows that condition (4.96) is equivalent to

$$(\forall z \in \mathbb{R}^N) \qquad \langle z \mid L_Q z\rangle \geqslant \eta \|L_Q z\|^2. \tag{4.100}$$

In other words, L is η-cocoercive in $(\mathbb{R}^N, \|\cdot\|_Q)$ if and only if L_Q is η-cocoercive in $\mathbb{R}^N$. The result then follows from Proposition 4.3.1.2, which provides cocoercivity conditions for L_Q. □

From (4.90) and (4.92), some special cases are worth being considered:

(i) If $P = Q^{-1}$ (matched preconditioning) then $L_Q = Q^{-1/2}MQ^{-1/2}$ and $\beta = 0$, which leads to $\overline{\eta} = \underline{\eta} = 1/|||L_Q|||$.

(ii) If M is invertible and $P = M^{-1}$ (Newton preconditioning) then $L = \mathrm{Id}$, $L_Q = Q^{1/2}Q^{-1/2} = \mathrm{Id}$, and $\beta = 0$, which leads to $\overline{\eta} = \underline{\eta} = 1$.

Let $\widehat{x}$ be a solution to Problem (4.84). Then, $\widehat{x}$ satisfies the following first-order optimality condition:

$$0 \in Q^{-1}\partial g(\widehat{x}) + Q^{-1}M\widehat{x} - Q^{-1}H^\top W y, \tag{4.101}$$

where $\partial g(\widehat{x})$ denotes the subdifferential of g at $\widehat{x}$. Similarly, for every $\gamma \in]0, +\infty[$, the fixed point set of operator T_γ is

$$\mathcal{F} = \big\{\widetilde{x} \in \mathbb{R}^N \;\big|\; 0 \in Q^{-1}\partial g(\widetilde{x}) + L\widetilde{x} - PH^\top W y\big\}. \tag{4.102}$$

One can notice that $\mathcal{F}$ is no longer the set of minimizers of the objective function in (4.84). We first characterize the existence and uniqueness of a fixed point of T_γ.

Proposition 4.5.3.3 Assume that conditions (4.97)-(4.98) hold.

(i) $\mathcal{F}$ is nonempty if $\operatorname{dom}\partial g = \mathbb{R}^N$ and the function defined as

$$(\forall x \in \mathbb{R}^N) \quad h(x) = \frac{1}{2}\langle x \mid Lx\rangle_Q + g(x) \tag{4.103}$$

is coercive, i.e., $\lim_{\|x\|\to+\infty} h(x) = +\infty$.

(ii) $\mathcal{F}$ is a singleton if g is strongly convex or $\lambda_{\min} > 0$.

Proof.

(i) According to (4.90) and (4.102), $\widetilde{x} \in \mathcal{F}$ if and only if $\widetilde{x} = Q^{-1/2}\widetilde{z}$ and

$$Q^{1/2}PH^\top W y \in Q^{-1/2}\partial g(Q^{-1/2}\widetilde{z}) + L_Q\widetilde{z}. \tag{4.104}$$

To prove the existence of a solution to this inclusion, let us define the auxiliary function $\widetilde{h}$ as

$$\widetilde{h}\colon x \mapsto h(Q^{-1/2}x) = \frac{1}{2}\langle x \mid L_Q x\rangle + g(Q^{-1/2}x). \tag{4.105}$$

If $\operatorname{dom}\partial g = \mathbb{R}^N$ and h (hence $\widetilde{h}$, since $Q^{-1/2}$ is invertible) is coercive, the cocoercivity of L_Q allows us to apply Proposition 4.3.2.1(iv)(a)-(b) that establishes the surjectivity of $\partial(g\circ Q^{-1/2}) + L_Q$. Given that $Q^{-1/2}\circ\partial g\circ Q^{-1/2} + L_Q = \partial(g\circ Q^{-1/2}) + L_Q$, this shows the existence of $\tilde{z}$ in (4.104).

(ii) According to (4.104), $\widetilde{x} \in \mathcal{F}$ is uniquely defined if and only if $\widetilde{z}$ is. From the assumption of convexity made on g, and the cocoercivity of L_Q, it follows from [16, Corollary 23.37] that $\widetilde{x}$ is unique if either $\partial(g \circ Q^{-1/2})$ or L_Q is strongly monotone [16]. This holds in particular if g is strongly convex or $\lambda_{\min} > 0$.

The unmatched preconditioned scheme (4.87) benefits from convergence results similar to those existing for the standard preconditioned PGA. However, there usually exists a discrepancy between the limit point $\widetilde{x}$ of (4.87) and any minimizer $\widehat{x}$ of the objective function in (4.84).

Proposition 4.5.3.4 Suppose that $\mathcal{F} \neq \varnothing$. Assume that conditions (4.97) and (4.98) hold. Let $\overline{\eta}$ be defined by (4.99), let $\gamma \in]0, 2\overline{\eta}[$, let $\delta = 2 - \gamma/(2\overline{\eta})$, and let $(\theta_n)_{n\in\mathbb{N}}$ be a sequence in $[0, \delta]$ such that $\sum_{n\in\mathbb{N}} \theta_n(\delta - \theta_n) = +\infty$. Then any sequence $(x_n)_{n\in\mathbb{N}}$ generated by iteration (4.87) converges to a point $\widetilde{x} \in \mathcal{F}$.
In addition, let $\nu \in [0, +\infty[$ be such that $g = h + \nu/2\|\cdot\|_Q^2$ where $h \in \Gamma_0(\mathbb{R}^N)$.[2] If $\nu > 0$ or $\lambda_{\min} > 0$, then

$$\|\widetilde{x} - \widehat{x}\|_Q \leqslant \inf_{\gamma\in]0,2\overline{\eta}[} \frac{\|e(\widehat{x})\|_Q}{1 + \gamma\nu - |||\operatorname{Id} - \gamma L_Q|||}. \tag{4.106}$$

Proof: In the renormed space $(\mathbb{R}^N, \|\cdot\|_Q)$, Algorithm (4.87) with L defined by (4.90) takes the same form as the mismatched PGA studied in the previous section (4.3). The convergence thus follows from Proposition 4.3.3.5.
The existence of a unique point $\widetilde{x} \in \mathcal{F}$ is a direct consequence of Proposition 4.3.2.1(iv). Let $\widehat{x}$ be a solution to Problem (4.84) and let $\gamma \in]0, +\infty[$. We have

$$\begin{cases} \widetilde{x} = \operatorname{prox}^Q_{\gamma g}(\widetilde{x} - \gamma P(M\widetilde{x} - H^\top W y)), & (4.107) \\ \widehat{x} = \operatorname{prox}^Q_{\gamma g}(\widehat{x} - \gamma Q^{-1}(M\widehat{x} - H^\top W y)). & (4.108) \end{cases}$$

From properties of the proximity operator [59], since $\nu \geqslant 0$, for every $x \in \mathbb{R}^N$, $\operatorname{prox}^Q_{\gamma g}(x) = \operatorname{prox}^Q_{\frac{\gamma}{1+\gamma\nu}h}(\frac{x}{1+\gamma\nu})$. Since $\operatorname{prox}^Q_{\frac{\gamma}{1+\gamma\nu}h}$ is nonexpansive in $(\mathbb{R}^N, \|\cdot\|_Q)$, we deduce from (4.107) and (4.108) that

$$(1 - \tau_\gamma)\|\widetilde{x} - \widehat{x}\|_Q \leqslant \frac{\gamma}{1 + \gamma\nu}\|(Q^{-1} - P)(M\widehat{x} - H^\top W y)\|_Q$$

with $\tau_\gamma = \frac{|||\operatorname{Id} - \gamma L|||_Q}{1+\gamma\nu} = \frac{|||\operatorname{Id} - \gamma L_Q|||}{1+\gamma\nu}$. In addition, according to Proposition 3.9, for $\lambda_{\min} > 0$ or $\nu > 0$, if $\gamma \in]0, 2\overline{\eta}[$, then $\tau_\gamma < 1$.
In summary, if $\gamma < 2\overline{\eta}$ and either $\lambda_{\min} > 0$ or $\nu > 0$,

$$\|\widetilde{x} - \widehat{x}\|_Q \leqslant \frac{\|e(\widehat{x})\|_Q}{(1 - \tau_\gamma)(1 + \gamma\nu)} = \frac{\|e(\widehat{x})\|_Q}{1 + \gamma\nu - |||\operatorname{Id} - \gamma L_Q|||},$$

which leads to (4.48). □
It is worth noting that the symmetry and positiveness of P are not mandatory to ensure the convergence of (4.87).
Usually, the regularization function g depends on some parameter vector ω. The previous analysis assumes that ω is set to the same value for the exact and unmatched preconditioned PGA. A better strategy may be to adjust ω in the unmatched case to reduce the discrepancy between $\widetilde{x}$ and $\widehat{x}$.

[2] ν is the strong convexity modulus of g in $(\mathbb{R}^N, \|\cdot\|_Q)$.

4.6 Application

From now on, H is a FP in parallel-beam geometry, y represents limited tomographic measurements, and $\overline{x}$ is the imaged phantom.
FP and BP are often the most computation-intensive operations in IR for CT. To limit the number of PGA iterations and thus multiplications with these operators, we use an effective preconditioner P for the gradient step. The numerical experiments were conducted in MATLAB with the ASTRA Toolbox [219].

Tomographic geometry:

Reference image $\overline{x}$ is a slice of size $N = 256 \times 256$ extracted from a computerized tomographic scan of an abdomen with values belonging to $[1000, 2100]$ sHU. The image contains simulated small structures of comparatively high intensity (3000 sHU). We simulate 60 projections at uniformly spaced angular positions within the interval $[0°, 180°]$. The detector has $363 \simeq \sqrt{2} \times 256$ bins of the same size as the pixels so that the data is not truncated and $K = 60 \times 363$. Operator H is based on the line-length model, which corresponds to ASTRA GPU implementation [162]. For the measurement y, we simulate a noise-free sinogram and add some low level of noise b drawn from the Gaussian distribution $\mathcal{N}(0, 100 \times \mathrm{Id}\,)$, so that $W = 10^{-2} \times \mathrm{Id}\,$.

Regularization:

For such an underdetermined problem, we adopt an Ivanov variational formulation [141] combined with a TV bound [57]: an estimate $\rho > 0$ of the value range for the TV bound of our target image is supposed to be known. Thus, in (4.1), $g = \iota_{B_{1,2}(0,\rho)}(\nabla \cdot)$, where ∇ is the 2D discrete spatial gradient operator and $\iota_{B_{1,2}}$ is the indicator function of the ball of radius ρ associated with the $\ell_{1,2}$ norm. More precisely, $B_{1,2}(0, \rho) = \{u \in \mathbb{R}^{2N} \mid \|u\|_{1,2} \leqslant \rho\}$. Even without any metric, the proximity operator of g does not have a closed form and is computed by DFB with a maximum of 500 iterations and warm restart.

Parameter selection:

Except otherwise stated, we choose ρ equal to $\overline{\rho} = \|\nabla \overline{x}\|_{1,2}$. Parameter κ in (4.86) is set to 10^{-5}, which implies that M is invertible and we set $P = M^{-1}$. To compute products between P and any vector, we use the conjugate gradient algorithm (CG) with a tolerance of 10^{-4} initialized with the previous iterate (warm restart). Preconditioning the gradient term with P might require extra projections and backprojections per iteration due to the CG inner loop. However, we demonstrate in the sequel that this inversion of the Hessian effectively accelerates the method, especially when combined with a simpler preconditioner for the proximity operator. Following Proposition 4.3.1.2, the cocoercivity constant of L equals 1. In addition, $\lambda_{\min} = 1$ and the conditions in Proposition 4.3.2.1(iv) are satisfied; $\mathcal{F}$ is thus a singleton. In PGA, the step size γ is set to 0.9, which, according to Proposition 4.3.3.5, guarantees the convergence of Algorithm (4.87). In the DFB iterations, the metric Q weighting the proximity operator is inverted. Setting $Q = P^{-1} = M$ would require using CG again in the DFB step, which would be computationally heavy. Several other choices of proximity metrics are considered instead, as described hereinafter.

Matched versus unmatched preconditioning:

We first set metric Q to the diagonal majorant matrix proposed in [148]: $Q_1 = W\,\mathrm{diag}(S^\top 1_N)+\kappa\,\mathrm{Id}$, where $1_N = [1,\dots,1]^\top \in \mathbb{R}^N$ and $S = \left(|H_{i,k}|\sum_{k'=1}^{K}|H_{i,k'}|\right)_{1\leqslant i\leqslant N, 1\leqslant k\leqslant K}$. We also compare PGA schemes (4.85) and (4.87) to PDHG, which involves neither sub-iterations nor preconditioning. We assess the performance of the three methods in Figure 4.14 showing the overall normalized reconstruction error NRMSE $= \|x_n - \overline{x}\|/\|\overline{x}\|$ as a function of the execution time in seconds. We notice that Algorithm (4.87) converges faster than both Algorithm (4.85) and PDHG. The convergence curve of PDHG oscillates in the first iterations. After only 10 seconds, the reconstruction error associated with Algorithm (4.87) is lower than the error corresponding to the estimate delivered by Algorithm (4.85) at convergence. Interestingly, the resulting fixed point was observed to be closer to the ground truth $\overline{x}$ than the minimizer of Problem (4.1). The reconstructed solutions are displayed in Figure 4.15. Residual deconvolution artifacts (undershooting) and sub-sampling streaks are present in the estimate produced by Algorithm (4.85) with Q_1 but not in the one yielded by our unmatched scheme.

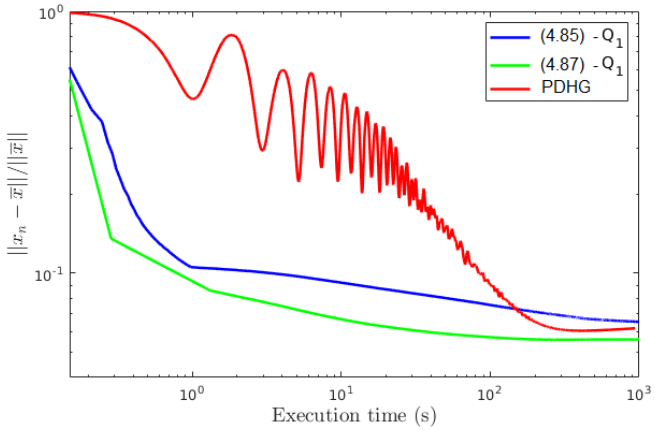


Figure 4.14: Evolution of the NRMSE along iterations for Algorithms (4.85) and (4.87), $\rho = \overline{\rho}$, and $Q = Q_1$.

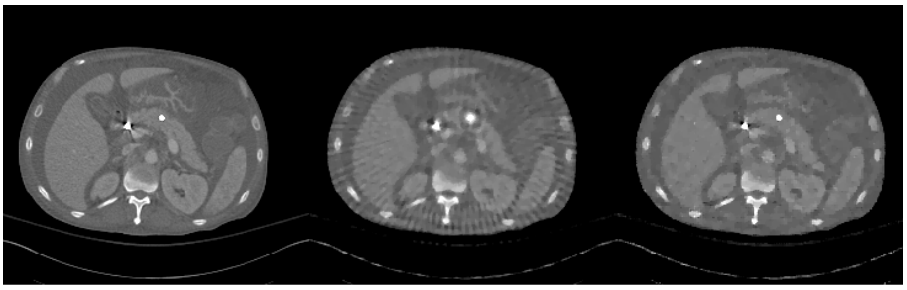


Figure 4.15: From left to right: $\overline{x}$, reconstructed images for $\rho = \overline{\rho}$ with Alg. (4.85) using Q_1, and with Alg. (4.87) using Q_1.

Alternative choices for Q:

Four additional approximations to M in $\mathcal{S}_N^+$ have been tested:

- $Q_2 = \mathrm{Id}$,

- Q_3 is the inverse of 2D Laplacian filter[3],
- $Q_4 = \mathrm{Diag}\left((M_{i,i})_{1\leqslant i\leqslant N}\right)$ (Jacobi preconditioner),
- $Q_5 = \mathrm{argmin}_{Q\in\mathcal{D}_N^+} \|Q^{1/2} - Q^{-1/2}M\|_{\mathrm{F}}^2$,
- Q_6 is a tridiagonal approximation to M i.e., the symmetric matrix whose elements on the main diagonal are $(M_{i,i})_{1\leqslant i\leqslant N}$ and those on the next upper / lower diagonals are $\frac{1}{2}(M_{i+1,i} + M_{i,i+1})_{1\leqslant i\leqslant N-1}$.

Note that the entries of H correspond to line integrals between a ray and a pixel, which are positive. Since $\lambda_{\min} > 0$, the entries of M are strictly positive. Table 4.1 (first row) contains the NRMSE values after 1000 iterations obtained with Algorithm (4.87) for the different metrics, and $\rho = \overline{\rho}$. First, we see that all choices are competitive compared to the baseline Algorithm (4.85), and their NRMSE values are close. Metric Q_1 leads to the best quantitative results. Metric Q_3 provides the poorest reconstruction with the highest NRMSE and a patchy look (see Figure 4.16).

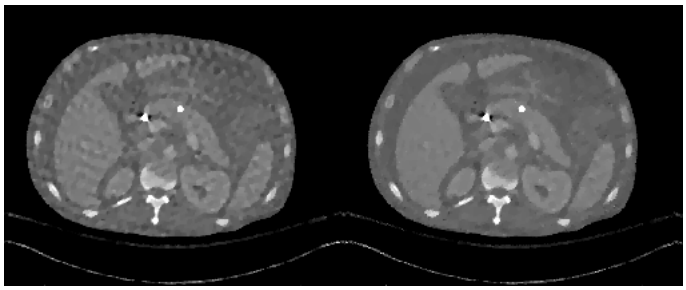


Figure 4.16: Reconstructed images for $\rho = \overline{\rho}$ with Alg. (4.87) using Q_3 (left) and Q_1 (right).

Sensitivity to ρ:

Table 4.1 also shows the sensitivity of the reconstruction to the TV bound when performing sets of trials reconstructions for $\rho \in \{\rho_-, \rho_+\}$, with $\rho_+ = 1.1\times\overline{\rho}$, $\rho_- = 0.9\times\overline{\rho}$. We observe that ρ_- generally leads to lower reconstruction errors. Q_1 is consistently associated with the lowest NRMSE. The choice of $Q = Q_2$ was noticeably shown to rank second-best for the lowest bound ρ_-.

	Q_1	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6
$\overline{\rho}$	0.0621	0.0543	0.0562	0.0622	0.0558	0.0559	0.0560
ρ_-	0.0559	0.0484	0.0489	0.0544	0.0494	0.0494	0.0490
ρ_+	0.0664	0.0614	0.0622	0.0708	0.0621	0.0628	0.0626

Table 4.1: NRMSE after 1000 iterations for Algorithms (4.85) (first column) and (4.87) (all other columns), for various choices of Q and ρ.

[3]The cone filter is decomposed into a local Laplacian operator coupled with a nonlocal logarithmic kernel filtering [230].

4.7 Conclusion

In this chapter, we established some necessary conditions to ensure the convergence of PGA when the adjoint of the forward operator involved in the quadratic part of the objective function was changed. Using cocoercivity and monotone operators, we derived conditions on the step size and the gradient of the smooth part of the objective function under which convergence of the algorithm to a fixed point is guaranteed. We also derived bounds on the error between this point and the solution to the original minimization problem. In addition to generalizing the original PGA method, our results give foundations for unmatched preconditioned PGA schemes where the metric used in the gradient step of PGA differs from the one used in the proximity step. Our simulations demonstrated that an unmatched preconditioning strategy offers an effective solution for X-ray tomographic imaging. Note that the considered minimization problem captures a broader class of image recovery problems than CT reconstruction.

5 Convergence of primal-dual algorithms with an adjoint mismatch

5.1 Introduction

One main limitation of PGA is that when the cost function involves a non-smooth term composed with a linear operator, the proximal step may not have a closed form, thus requiring inner iterations. For instance, this is the case when TV regularization is used as in Chapter 4 (section 4.6). A way to avoid these sub-iterations is to rely on primal-dual proximal splitting algorithms [5, 30, 34, 58, 85, 124]. These algorithmic schemes are grounded on splitting strategies such as the forward-backward, the Douglas-Rachford, or Tseng's forward-backward-forward algorithms, presented in Chapter 3 (subsection 3.3.2). This chapter extends the analysis conducted in Chapter 4 to these primal-dual algorithms, namely the CV, the LV, and the CP algorithms.

To our knowledge, the first proposal to analyze a primal-dual proximal splitting method under an adjoint mismatch is [140], which studied a mismatched form of PDHG with fixed step sizes. The analysis was conducted under strong convexity assumptions. The authors gave convergence conditions on the strong convexity modulus of the involved functions and derived update rules for the step sizes to recover a similar convergence rate to the matched scheme. However, they did not investigate the existence and uniqueness of the fixed points of the mismatched algorithm. Similarly, in the context of microscopy imaging, where the forward operator satisfies a specific orthogonality condition, [180] gave conditions for the Douglas-Rachford/ADMM iterations to converge in the Multi-Agent Consensus Equilibrium framework [37]. The authors highlighted that using an adjoint mismatch on the forward operator in a quadratic term was equivalent to using the proximity operator (or agent) of the quadratic term with a different prior model for the image that depends on the mismatched adjoint.

The chapter is organized as follows. Section 5.2 focuses on a mismatched CV algorithm. We consider the case when an adjoint mismatch appears on the linear operator in the quadratic term. Note that this complements the work of [140], where the mismatch appears on the second linear operator. This analysis is then extended to the projected form of the CV algorithm proposed by Briceño-Arias and López [35]. In section 5.3, we perform an analysis of the LV algorithm in the case of an adjoint mismatch on the linear operator involved in the quadratic term. In section 5.4, we analyze the CP algorithm when the adjoints of both linear operators are changed. Finally, in section 5.5, we illustrate our theoretical findings on examples involving two different inverse problems in CT reconstruction, with two types of regularization and noise modeling.

5.2 The mismatched Condat-Vũ Algorithm

5.2.1 Algorithm

Let $H \in \mathcal{B}(\mathcal{H}, \mathcal{G})$, $D \in \mathcal{B}(\mathcal{H}, \mathcal{L})$, $f \in \Gamma_0(\mathcal{H})$, and $g \in \Gamma_0(\mathcal{L})$. Given $y \in \mathcal{G}$, we are now interested in solving the following problem

$$\underset{x \in \mathcal{H}}{\text{minimize}} \ \frac{1}{2}\|y - Hx\|_{\mathcal{G}}^2 + f(x) + g(Dx). \tag{5.1}$$

As we have seen in Chapter 3 (subsection 3.3.3), the CV algorithm (5.6) can solve (5.1). It is derived from a preconditioned form of the forward-backward algorithm of the form, for every $n \in \mathbb{N}$,

$$0 \in \mathcal{A}z_{n+\frac{1}{2}} + Bz_n + P(z_{n+\frac{1}{2}} - z_n), \tag{5.2}$$

where $z_n = (x_n, u_n) \in \mathcal{Z} = \mathcal{H} \times \mathcal{L}$, $z_{n+\frac{1}{2}} = (x_{n+\frac{1}{2}}, u_{n+\frac{1}{2}}) \in \mathcal{Z}$ and

$$(\forall z = (x, u) \in \mathcal{Z}) \quad \mathcal{A}z = \begin{pmatrix} \partial f(x) + D^* u \\ -Dx + \partial g^*(u) \end{pmatrix} \tag{5.3}$$

$$Bz = \begin{pmatrix} H^*Hx - H^*y \\ 0 \end{pmatrix}. \tag{5.4}$$

The preconditioning metric P is defined as

$$(\forall z = (x, u) \in \mathcal{Z}) \quad Pz = \begin{pmatrix} \frac{1}{\tau}x - D^*u \\ -Dx + \frac{1}{\sigma}u \end{pmatrix}, \tag{5.5}$$

where τ and σ are two positive real parameters.
We recall that a CV iteration is obtained by rewriting (5.2) as

CV iterations:

$$\text{for } n = 0, 1, \ldots \quad \left\lfloor \begin{aligned} x_{n+\frac{1}{2}} &= \operatorname{prox}_{\tau f}\left(x_n - \tau(H^*(Hx_n - y) + D^*u_n)\right) \\ u_{n+\frac{1}{2}} &= \operatorname{prox}_{\sigma g^*}\left(u_n + \sigma D(2x_{n+\frac{1}{2}} - x_n)\right) \\ x_{n+1} &= x_n + \Theta_n(x_{n+\frac{1}{2}} - x_n) \\ u_{n+1} &= u_n + \Theta_n(u_{n+\frac{1}{2}} - u_n). \end{aligned} \right. \tag{5.6}$$

with initialization $x_0 \in \mathcal{H}$ and $u_0 \in \mathcal{L}$.
In this section, we study the impact of replacing the operator H^* in the n-th iteration of the CV algorithm by a surrogate operator $K_n \in \mathcal{B}(\mathcal{G}, \mathcal{H})$.
This yields:

Mismatched CV iterations:

$$\text{for } n = 0, 1, \ldots \quad \left\lfloor \begin{aligned} x_{n+\frac{1}{2}} &= \operatorname{prox}_{\tau f}\left(x_n - \tau(K_n(Hx_n - y) + D^*u_n)\right) \\ u_{n+\frac{1}{2}} &= \operatorname{prox}_{\sigma g^*}\left(u_n + \sigma D(2x_{n+\frac{1}{2}} - x_n)\right) \\ x_{n+1} &= x_n + \Theta_n(x_{n+\frac{1}{2}} - x_n) \\ u_{n+1} &= u_n + \Theta_n(u_{n+\frac{1}{2}} - u_n), \end{aligned} \right. \tag{5.7}$$

where $\{\Theta_n\}_{n\in\mathbb{N}}$ is a sequence of relaxation parameters.
As in Chapter 4, we suppose that the sequence of surrogates $(K_n)_{n\in\mathbb{N}}$ is related to a constant linear operator $\overline{K} \in \mathcal{B}(\mathcal{G}, \mathcal{H})$ through Assumption 4.2.0.1(iii).

5.2.2 Adaptation of previous results

To perform the convergence analysis of (5.7), we introduce notation involved in characterizing the spectra of the linear operators of the cost function.

Notation 5.2.2.1 For every $n \in \mathbb{N}$,

(i) $L = \overline{K}H$

(ii) $\lambda_{\min} = \inf\{\langle x, Lx\rangle_{\mathcal{H}} \mid x \in \mathcal{H}, \|x\|_{\mathcal{H}} = 1\}$

(iii) $\widetilde{L} = \Pi L$, $\widetilde{L}_n = \Pi K_n H$, $\widetilde{K} = \Pi K$, and $\widetilde{K}_n = \Pi K_n$, with $\Pi\colon \mathcal{Z} \to \mathcal{Z}\colon (x,u) \mapsto (x,0)$.

(iv) $T_n\colon \mathcal{Z} \to \mathcal{Z}\colon z \mapsto J_{P^{-1}\mathcal{A}}\left(z - P^{-1}\left(\widetilde{L}_n z - \widetilde{K}_n y\right)\right)$ and
$\overline{T}\colon \mathcal{Z} \to \mathcal{Z}\colon z \mapsto J_{P^{-1}\mathcal{A}}\left(z - P^{-1}\left(\widetilde{L} z - \widetilde{K} y\right)\right)$.

Under this notation, Algorithm (5.7) can be rewritten more concisely with a simple update rule on the pair $z_n = (x_n, u_n)$:

$$(\forall n \in \mathbb{N}) \quad z_{n+1} = z_n + \Theta_n\left(T_n\left(z_n\right) - z_n\right). \tag{5.8}$$

We introduce the Hilbert space $\mathcal{Z}_P$ obtained by equipping $\mathcal{Z}$ with the inner product $\langle\cdot,\cdot\rangle_P\colon (z,z') \mapsto \langle z, z'\rangle_P = \langle z, Pz'\rangle_{\mathcal{Z}}$. In the case of CV, operators $P^{-1}\mathcal{A}$ and $P^{-1}B$ are maximally monotone [16, Proposition 20.24] and cocoercive [66, Theorem 3.1] in $\mathcal{Z}_P$, respectively. (5.8) can be viewed as a mismatched form of the forward-backward algorithm for finding a zero of a sum of maximally monotone operators in $\mathcal{Z}_P$. Similar to our analysis in Chapter 4, we rely on the cocoercivity properties of operator $\widetilde{L}$ to study the convergence of (5.8).

Proposition 5.2.2.2 Assume that $(\tau,\sigma) \in]0,+\infty[^2$ are such that $\tau\sigma\|D\|^2_{\mathcal{H},\mathcal{L}} < 1$. We have the following properties.

(i) $P^{-1}\widetilde{L}$ is cocoercive in $\mathcal{Z}_P$ if and only if L is cocoercive.

(ii) Suppose that $\mathrm{Ran}(L + L^*)$ is closed. Then $P^{-1}\widetilde{L}$ is cocoercive in $\mathcal{Z}_P$ with constant $\widetilde{\eta} > 0$ if and only if $\lambda_{\min} \geqslant 0$, $\mathrm{Ker}(L + L^*) = \mathrm{Ker}\, L$, and

$$\widetilde{\eta} \leqslant \widetilde{\eta}_{\max} = \frac{2}{\|P^{-1/2}\Pi M\|^2_{\mathcal{H},\mathcal{Z}}}, \tag{5.9}$$

where

$$M = (\mathrm{Id}_{\mathcal{H}} + (L - L^*)(L + L^*)^{\dagger})(L + L^*)^{1/2}. \tag{5.10}$$

In addition,

$$\widetilde{\eta}_{\max} \geqslant \tau^{-1}(1 - \tau\sigma\|D\|^2_{\mathcal{H},\mathcal{L}})\eta_{\max} \tag{5.11}$$

where $\eta_{\max} = 2/\|M\|^2_{\mathcal{H},\mathcal{H}}$ is the largest cocoercivity constant of L.

Proof:

(i) $P^{-1}\widetilde{L}$ is cocoercive in $\mathcal{Z}_P$ if and only if there exists $\widetilde{\eta} > 0$ such that, for every $z \in \mathcal{Z}$,

$$\begin{aligned} &(\forall z \in \mathcal{Z}) \quad \langle z, P^{-1}\widetilde{L}z\rangle_{\mathcal{Z}_P} \geqslant \widetilde{\eta}\|P^{-1}\widetilde{L}z\|^2_{\mathcal{Z}_P} \\ \Leftrightarrow \quad &(\forall z' \in \mathcal{Z}) \quad \langle z', P^{-1/2}\widetilde{L}P^{-1/2}z'\rangle_{\mathcal{Z}} \geqslant \widetilde{\eta}\|P^{-1/2}\widetilde{L}P^{-1/2}z'\|^2_{\mathcal{Z}}. \end{aligned} \tag{5.12}$$

Therefore $P^{-1}\widetilde{L}$ is cocoercive in $\mathcal{Z}_P$ if and only if $P^{-1/2}\widetilde{L}P^{-1/2}$ is cocoercive in $\mathcal{Z}$. In turn, it follows from [16, Proposition 4.12] that, if $\widetilde{L}$ is cocoercive, then $P^{-1/2}\widetilde{L}P^{-1/2}$ is cocoercive. Conversely, if $P^{-1/2}\widetilde{L}P^{-1/2}$ is cocoercive, then $P^{1/2}(P^{-1/2}\widetilde{L}P^{-1/2})P^{1/2} = \widetilde{L}$ is cocoercive, that is

$$\begin{aligned} & (\forall z \in \mathcal{Z}) \quad \langle z, \widetilde{L}z\rangle_{\mathcal{Z}} \geqslant \eta\|\widetilde{L}z\|_{\mathcal{Z}}^2 \\ \Leftrightarrow \quad & (\forall x \in \mathcal{H}) \quad \langle x, Lx\rangle_{\mathcal{H}} \geqslant \eta\|Lx\|_{\mathcal{H}}^2, \end{aligned} \tag{5.13}$$

for some $\eta > 0$.

(ii) Let $\widetilde{L}_P^*$ denote the adjoint of $P^{-1}\widetilde{L}$ in $\mathcal{Z}_P$ and let

$$\widetilde{\lambda}_{\min} = \inf\left\{\langle z, P^{-1}\widetilde{L}z\rangle_{\mathcal{Z}_P} \mid z \in \mathcal{Z}_P, \|z\|_{\mathcal{Z}_P} = 1\right\}. \tag{5.14}$$

According to Proposition 4.3.1.2(ii), provided that $\operatorname{Ran}(P^{-1}\widetilde{L}+\widetilde{L}_P^*)$ is closed, $P^{-1}\widetilde{L}$ is cocoercive in $\mathcal{Z}_P$ with constant $\widetilde{\eta}$ if and only if $\widetilde{\lambda}_{\min} \geqslant 0$, $\operatorname{Ker}(P^{-1}\widetilde{L} + \widetilde{L}_P^*) = \operatorname{Ker}(P^{-1}\widetilde{L})$, and

$$\widetilde{\eta} \leqslant \frac{2}{\|(\mathrm{Id}_{\mathcal{Z}} + (P^{-1}\widetilde{L} - \widetilde{L}_P^*)(P^{-1}\widetilde{L} + \widetilde{L}_P^*)^\dagger)(P^{-1}\widetilde{L} + \widetilde{L}_P^*)^{1/2}\|_{\mathcal{Z}_P,\mathcal{Z}_P}^2}. \tag{5.15}$$

Since $\widetilde{L}_P^* = P^{-1}\widetilde{L}^*$, $\operatorname{Ker}(P^{-1}\widetilde{L} + \widetilde{L}_P^*) = \operatorname{Ker}(\widetilde{L} + \widetilde{L}^*) = \operatorname{Ker}(L + L^*) \times \mathcal{L}$ and, similarly, $\operatorname{Ker}(P^{-1}\widetilde{L}) = \operatorname{Ker} L \times \mathcal{L}$. Theqrefore $\operatorname{Ker}(P^{-1}\widetilde{L} + \widetilde{L}_P^*) = \operatorname{Ker}(P^{-1}\widetilde{L})$ if and only if $\operatorname{Ker}(L+L^*) = \operatorname{Ker} L$. Similarly, $\operatorname{Ran}(P^{-1}\widetilde{L}+\widetilde{L}_P^*) = P^{-1}\operatorname{Ran}(\widetilde{L}+\widetilde{L}^*) = P^{-1}(\operatorname{Ran}(L+L^*)\times\{0\})$. Thus, $\operatorname{Ran}(P^{-1}\widetilde{L}+\widetilde{L}_P^*)$ is closed if and only if $\operatorname{Ran}(L+L^*)$ is closed.

For every $z = (x, u) \in \mathcal{Z}$,

$$\begin{aligned} & \langle z, P^{-1}\widetilde{L}z\rangle_{\mathcal{Z}_P} \geqslant \widetilde{\lambda}_{\min}\|z\|_{\mathcal{Z}_P}^2 \\ \Leftrightarrow \quad & \langle z, \widetilde{L}z\rangle_{\mathcal{Z}} \geqslant \widetilde{\lambda}_{\min}\|z\|_{\mathcal{Z}_P}^2 \\ \Leftrightarrow \quad & \langle x, Lx\rangle_{\mathcal{H}} \geqslant \widetilde{\lambda}_{\min}\|(x,u)\|_{\mathcal{Z}_P}^2. \end{aligned} \tag{5.16}$$

So, $\widetilde{\lambda}_{\min} \geqslant 0$ if and only if, for every $x \in \mathcal{H}$, $\langle x, Lx\rangle_{\mathcal{H}} \geqslant 0$, that is $\lambda_{\min} \geqslant 0$. In addition, when condition $\lambda_{\min} \geqslant 0$ is met, $\widetilde{\lambda}_{\min} = 0$.

Since $\|\cdot\|_{\mathcal{Z}_P,\mathcal{Z}_P} = \|P^{1/2}\cdot P^{-1/2}\|_{\mathcal{Z},\mathcal{Z}}$, we have

$$\begin{aligned} & \|(\mathrm{Id}_{\mathcal{Z}} + (P^{-1}\widetilde{L} - \widetilde{L}_P^*)(P^{-1}\widetilde{L} + \widetilde{L}_P^*)^\dagger)(P^{-1}\widetilde{L} + \widetilde{L}_P^*)^{1/2}\|_{\mathcal{Z}_P,\mathcal{Z}_P}^2 \\ &= \|(\mathrm{Id}_{\mathcal{Z}} + (P^{-1}\widetilde{L} - \widetilde{L}_P^*)(P^{-1}\widetilde{L} + \widetilde{L}_P^*)^\dagger)(P^{-1}\widetilde{L} + \widetilde{L}_P^*)(\mathrm{Id}_{\mathcal{Z}} + \\ &\qquad (P^{-1}\widetilde{L} + \widetilde{L}_P^*)^\dagger(\widetilde{L}_P^* - P^{-1}\widetilde{L})\|_{\mathcal{Z}_P,\mathcal{Z}_P} \\ &= \|P^{1/2}(\mathrm{Id}_{\mathcal{Z}} + P^{-1}(\widetilde{L} - \widetilde{L}^*)(\widetilde{L} + \widetilde{L}^*)^\dagger P)P^{-1}(\widetilde{L} + \widetilde{L}^*)(\mathrm{Id}_{\mathcal{Z}} + \\ &\qquad (\widetilde{L} + \widetilde{L}^*)^\dagger P P^{-1}(\widetilde{L}^* - \widetilde{L}))P^{-1/2}\|_{\mathcal{Z},\mathcal{Z}} \\ &= \|P^{-1/2}(\mathrm{Id}_{\mathcal{Z}} + (\widetilde{L} - \widetilde{L}^*)(\widetilde{L} + \widetilde{L}^*)^\dagger)(\widetilde{L} + \widetilde{L}^*)(\mathrm{Id}_{\mathcal{Z}} + (\widetilde{L} + \widetilde{L}^*)^\dagger(\widetilde{L}^* - \widetilde{L}))P^{-1/2}\|_{\mathcal{Z},\mathcal{Z}} \\ &= \|P^{-1/2}(\mathrm{Id}_{\mathcal{Z}} + (\widetilde{L} - \widetilde{L}^*)(\widetilde{L} + \widetilde{L}^*)^\dagger)(\widetilde{L} + \widetilde{L}^*)^{1/2}\|_{\mathcal{Z},\mathcal{Z}}^2. \end{aligned}$$

By using the specific form of $\widetilde{L}$, we deduce that

$$\|(\mathrm{Id}_{\mathcal{Z}} + (P^{-1}\widetilde{L} - \widetilde{L}_P^*)(P^{-1}\widetilde{L} + \widetilde{L}_P^*)^\dagger)(P^{-1}\widetilde{L} + \widetilde{L}_P^*)^{1/2}\|_{\mathcal{Z}_P,\mathcal{Z}_P}^2 = \|P^{-1/2}\Pi M\|_{\mathcal{H},\mathcal{Z}}^2. \tag{5.17}$$

Altogether with (5.15), this yields (5.9).

In addition,

$$\|P^{-1/2}\Pi M\|_{\mathcal{H},\mathcal{Z}} \leqslant \|P^{-1/2}\Pi\|_{\mathcal{Z},\mathcal{Z}}\|M\|_{\mathcal{H},\mathcal{H}} = \|P^{-1/2}\Pi\|_{\mathcal{Z},\mathcal{Z}}\sqrt{\frac{2}{\eta_{\max}}}, \tag{5.18}$$

where

$$\begin{aligned}\|P^{-1/2}\Pi\|_{\mathcal{Z},\mathcal{Z}} = \|\Pi^* P^{-1/2}\|_{\mathcal{Z},\mathcal{Z}} &= \|\Pi P^{-1/2}\|_{\mathcal{Z},\mathcal{Z}} \\ &= \sup_{z'\in\mathcal{Z}\setminus\{0\}} \frac{\|\Pi P^{-1/2} z'\|_{\mathcal{Z}}}{\|z'\|_{\mathcal{Z}}} \\ &= \sup_{z=(x,u)\in\mathcal{Z}\setminus\{0\}} \frac{\|x\|_{\mathcal{H}}}{\sqrt{\langle z, Pz\rangle_{\mathcal{Z}}}}.\end{aligned} \tag{5.19}$$

For every $z = (x, u) \in \mathcal{Z}$,

$$\begin{aligned}\langle z, Pz\rangle_{\mathcal{Z}} &= \frac{1}{\tau}\|x\|_{\mathcal{H}}^2 - 2\langle Dx, u\rangle_{\mathcal{L}} + \frac{1}{\sigma}\|u\|_{\mathcal{L}}^2 \\ &= \frac{1}{\tau}\|x\|_{\mathcal{H}}^2 + \frac{1}{\sigma}\|u - \sigma Dx\|_{\mathcal{L}}^2 - \sigma\|Dx\|_{\mathcal{L}}^2 \\ &\geqslant \frac{1}{\tau}(1 - \tau\sigma\|D\|_{\mathcal{H},\mathcal{L}}^2)\|x\|_{\mathcal{H}}^2.\end{aligned} \tag{5.20}$$

We deduce from (5.19) and (5.20) that

$$\|P^{-1/2}\Pi\|_{\mathcal{Z},\mathcal{Z}}^2 \leqslant \frac{\tau}{1 - \tau\sigma\|D\|_{\mathcal{H},\mathcal{L}}^2}, \tag{5.21}$$

which, combined with (5.18), yields (5.11).

□

The following results provide a characterization of the fixed point set of $\overline{T}$.

Proposition 5.2.2.3 Let $(\widetilde{x}, \widetilde{u}) \in \mathcal{Z}$. Then $(\widetilde{x}, \widetilde{u}) \in \operatorname{Fix}\left(\overline{T}\right)$ if and only if $(\widetilde{x}, \widetilde{u})$ belongs to

$$\overline{\mathcal{F}} = \left\{(x, u) \in \mathcal{Z} \mid \overline{K}y \in \partial f(x) + Lx + D^* u,\ u \in \partial g(Dx)\right\}, \tag{5.22}$$

which is nonempty if $L + \partial f + D^* \circ \partial g \circ D$ is surjective.

(i) If $\lambda_{\min} \geqslant 0$, then $\overline{\mathcal{F}}$ is closed and convex.

(ii) Let $\overline{\mathcal{F}}_1 = \{x \in \mathcal{H} \mid (\exists u \in \mathcal{L})\ (x, u) \in \overline{\mathcal{F}}\}$.
$\overline{\mathcal{F}}_1$ has at most one element if one of the following conditions holds:

(a) $L + \partial f + D^* \circ \partial g \circ D$ is strictly monotone.

(b) $\lambda_{\min} \geqslant 0$ and $g \circ D + f$ is strictly convex.

$\overline{\mathcal{F}}_1$ is a singleton if

$$0 \in \operatorname{sri}\left(D(\operatorname{dom} f) - \operatorname{dom} g\right). \tag{5.23}$$

is satisfied and one of the following conditions holds:

(c) $\lambda_{\min} \geqslant 0$ and $L + \partial f + D^* \circ \partial g \circ D$ is strongly monotone.

(d) $\lambda_{\min} > 0$.

(e) $\lambda_{\min} \geqslant 0$, and f is strongly convex or [g is strongly convex and D^*D is strongly positive].

(f) L is cocoercive, and f is strongly convex or [g is strongly convex and D^*D is strongly positive].

Proof:

$$\begin{aligned}\widetilde{z} = (\widetilde{x}, \widetilde{u}) \in \mathrm{Fix}\overline{T} &\Leftrightarrow \widetilde{z} = \overline{T}\widetilde{z} \\ &\Leftrightarrow 0 \in P^{-1}\left(\widetilde{L}\widetilde{z} - \widetilde{K}y\right) + P^{-1}\mathcal{A}\widetilde{z} \\ &\Leftrightarrow \overline{K}y \in \partial f(\widetilde{x}) + L\widetilde{x} + D^*\widetilde{u},\ D\widetilde{x} \in \partial g^*(\widetilde{u}),\end{aligned}$$

that is $\widetilde{z} \in \overline{\mathcal{F}}$. In addition,

$$(\exists \widetilde{u} \in \mathcal{L}) \quad \begin{cases} \overline{K}y \in \partial f(\widetilde{x}) + L\widetilde{x} + D^*\widetilde{u} \\ D\widetilde{x} \in \partial g^*(\widetilde{u}) \end{cases} \quad \Leftrightarrow \quad \overline{K}y \in \partial f(\widetilde{x}) + L\widetilde{x} + D^*\partial g(D\widetilde{x}), \quad (5.24)$$

and the latter condition is satisfied if $\partial f + L + D^* \circ \partial g \circ D$ is surjective.

(i) If $\lambda_{\min} \geqslant 0$, then, for every $z = (x, u) \in \mathcal{Z}$, $\langle \widetilde{L}z, z\rangle_{\mathcal{Z}} = \langle Lx, x\rangle_{\mathcal{H}} \geqslant 0$ and, since $\widetilde{L}$ is continuous, $\widetilde{L}$ is maximally monotone on $\mathcal{Z}$. Operator $\mathcal{A}$ being maximally monotone, by [16, Proposition 23.39], we conclude that $\overline{\mathcal{F}} = \mathrm{zer}(\widetilde{L} - \widetilde{K}y + \mathcal{A})$ is closed and convex.

(ii) According to (5.24), $\overline{\mathcal{F}}_1 = \mathrm{zer}(L - \overline{K}y + \partial f + D^* \circ \partial g \circ D)$.

(a) Follows from [16, Proposition 23.35].

(b) From standard subdifferential calculus rules, $\overline{\mathcal{F}}_1 \subset \mathrm{zer}(L - Ky + \partial(f + g \circ D))$. $\lambda_{\min} \geqslant 0 \Leftrightarrow L + L^*$ positive $\Leftrightarrow L$ is monotone. In addition, $f + g \circ D$ is strictly convex if and only if $\partial(f + g \circ D)$ is strictly monotone. Thus, under the stated condition, $L - Ky + \partial(f + g \circ D)$ is strictly monotone, and $\overline{\mathcal{F}}_1$ has at most one element.

(c) Since (5.23) holds, it follows from [16, Theorem 16.47] that $\partial f + D^* \circ \partial g \circ D = \partial(f + g \circ D)$ is maximally monotone. Since L is maximally monotone and has a full domain, $L + \partial f + D^* \circ \partial g \circ D$ is thus maximally monotone. Then we deduce the result from [16, Proposition 23.37].

(d) If $\lambda_{\min} > 0$, L is strongly monotone and (c) is satisfied.

(e) If g is ρ_{G}-strongly convex with $\rho_{\mathrm{G}} > 0$, there exists $h \in \Gamma_0(\mathcal{L})$ such that $g = h + \rho_{\mathrm{G}}\|\cdot\|^2_{\mathcal{L}}/2$ and $D^* \circ \partial g \circ D = D^* \circ \partial h \circ D + \rho_{\mathrm{G}} D^*D$ is strongly monotone as D^*D is strongly positive. If f is strongly convex, then ∂f is strongly monotone. The result then follows from (c).

(f) Since L is cocoercive, from Lemma 4.3.1.1(i) $\lambda_{\min} \geqslant 0$ and the result follows from (e).

□

We now focus on the averagedness properties of operator $\overline{T}$.

Lemma 5.2.2.4 *Let $\widetilde{\eta} \in]\frac{1}{2}, +\infty[$, $\overline{\alpha} = \dfrac{1}{2 - \frac{1}{2\widetilde{\eta}}}$, and $\overline{W} = \mathrm{Id}_{\mathcal{Z}_P} - P^{-1}\widetilde{L}$. Then, if $P^{-1}\widetilde{L}$ is $\widetilde{\eta}$-cocoercive in $\mathcal{Z}_P$, then $\overline{T}$ is $\overline{\alpha}$-averaged in $\mathcal{Z}_P$.*

Proof: If $P^{-1}\widetilde{L}$ is $\widetilde{\eta}$-cocoercive in $\mathcal{Z}_P$ and $\widetilde{\eta} > 1/2$, according to [16, Proposition 4.39], $\overline{W}$ is $\frac{1}{2\widetilde{\eta}}$-averaged in $\mathcal{Z}_P$. Similarly to the proof of Lemma 4.3.3.4, we deduce that

$$(\forall z \in \mathcal{Z}_P) \quad \|\overline{W}z - 2(1-\overline{\alpha})z\|_{\mathcal{Z}_P} + \|\overline{W}z\|_{\mathcal{Z}_P} \leqslant 2\overline{\alpha}\|z\|_{\mathcal{Z}_P}. \tag{5.25}$$

Since $P^{-1}\mathcal{A}$ is maximally monotone on $\mathcal{Z}_P$, then $J_{P^{-1}\mathcal{A}}$ is firmly nonexpansive [16, Corollary 23.9]. Finally, it follows from (5.25) and [60, Theorem 3.8], that $\overline{T} = J_{P^{-1}\mathcal{A}}(\overline{W}\cdot + P^{-1}\widetilde{K}y)$ is $\overline{\alpha}$-averaged. □

The following theorem provides conditions under which iteration (5.7) converges to a fixed point of $\overline{T}$.

Theorem 5.2.2.5 *Let $(\tau,\sigma) \in]0,+\infty[^2$ be such that $\tau\sigma\|D\|^2_{\mathcal{H},\mathcal{L}} < 1$. Assume that $\widetilde{\eta} \in]1/2,+\infty[$ is a cocoercivity constant of $P^{-1}\widetilde{L}$ in $\mathcal{Z}_P$. For $\delta = 2 - 1/(2\widetilde{\eta})$, let $\{\Theta_n\}_{n\in\mathbb{N}} \subset [0,\delta]$ be a sequence such that $\sum_{n\in\mathbb{N}} \Theta_n(\delta - \Theta_n) = +\infty$, and suppose that $\overline{\mathcal{F}} \neq \varnothing$. Then the sequence $((x_n,u_n))_{n\in\mathbb{N}}$ given by (5.7) converges weakly to some point in $\overline{\mathcal{F}}$.*

Proof: Let $z_0 = r_0 \in \mathcal{Z}$, let $(z_n)_{n\geqslant 1}$ be given by (5.8), and let $(r_n)_{n\geqslant 1}$ be defined as

$$(\forall n \in \mathbb{N}) \quad r_{n+1} = r_n + \Theta_n\left(\overline{T}(r_n) - r_n\right). \tag{5.26}$$

By applying Lemma 5.2.2.4, operator $\overline{T}$ is $\overline{\alpha}$-averaged in $\mathcal{Z}_P$ with $\overline{\alpha} = 1/(2 - \frac{1}{2\widetilde{\eta}})$. Thus, the sequence $(r_n)_{n\in\mathbb{N}}$ converges weakly to some $\overline{r} \in \operatorname{Fix}\left(\overline{T}\right)$ [16, Proposition 5.16], which implies that $\varrho = \sup_{n\in\mathbb{N}} \|r_n\|_{\mathcal{Z}_P} < +\infty$.
For every $n \in \mathbb{N}$, let us bound $z_{n+1} - r_{n+1}$ as follows

$$\begin{aligned}\|z_{n+1} - r_{n+1}\|_{\mathcal{Z}_P} &= \left\|z_n - r_n + \Theta_n(r_n - z_n) + \Theta_n\left(T_n(z_n) - \overline{T}(r_n)\right)\right\|_{\mathcal{Z}_P} \\ &\leqslant \left\|z_n - r_n + \Theta_n\left(\overline{T}(z_n) - \overline{T}(r_n) - z_n + r_n\right)\right\|_{\mathcal{Z}_P} + \Theta_n\left\|T_n(z_n) - \overline{T}(z_n)\right\|_{\mathcal{Z}_P}.\end{aligned} \tag{5.27}$$

$\overline{T}$ being $\overline{\alpha}$-averaged in $\mathcal{Z}_P$, there exist a nonexpansive operator $W\colon \mathcal{Z}_P \to \mathcal{Z}_P$ such that $\overline{T} = (1-\overline{\alpha})\mathrm{Id} + \overline{\alpha}W$. Since $\{\Theta_n\}_{n\in\mathbb{N}} \subset [0, 1/\overline{\alpha}]$, we have

$$\begin{aligned}\|z_n - r_n + \Theta_n(\overline{T}(z_n) - \overline{T}(r_n) - z_n + r_n))\|_{\mathcal{Z}_P} &= \|\,(1-\overline{\alpha}\Theta_n)(z_n - r_n) + \overline{\alpha}\Theta_n(W(z_n) - W(r_n))\,\|_{\mathcal{Z}_P} \\ &\leqslant (1-\overline{\alpha}\Theta_n)\|z_n - r_n\|_{\mathcal{Z}_P} + \overline{\alpha}\Theta_n\|W(z_n) - W(r_n)\|_{\mathcal{Z}_P} \\ &\leqslant \|z_n - r_n\|_{\mathcal{Z}_P}.\end{aligned} \tag{5.28}$$

Since $J_{P^{-1}\mathcal{A}}$ is firmly nonexpansive in $\mathcal{Z}_P$, for every $z \in \mathcal{Z}$,

$$\begin{aligned}\left\|T_n(z) - \overline{T}(z)\right\|_{\mathcal{Z}_P} &\leqslant \|P^{-1}(\widetilde{L}_n z - \widetilde{K}_n y - \widetilde{L}z + \widetilde{K}y)\|_{\mathcal{Z}_P} \\ &\leqslant \|P^{-1}(\widetilde{L}_n - \widetilde{L})z\|_{\mathcal{Z}_P} + \|P^{-1}(\widetilde{K}_n - \widetilde{K})y\|_{\mathcal{Z}_P} \\ &\leqslant \|P^{-1}(\widetilde{L}_n - \widetilde{L})\|_{\mathcal{Z}_P,\mathcal{Z}_P}\|z\|_{\mathcal{Z}_P} + \|P^{-1}(\widetilde{K}_n - \widetilde{K})\|_{\mathcal{Z}_P,\mathcal{Z}_P}\|y\|_{\mathcal{Z}_P} \\ &= \|P^{-1/2}(\widetilde{L}_n - \widetilde{L})P^{-1/2}\|_{\mathcal{Z},\mathcal{Z}}\|z\|_{\mathcal{Z}_P} + \|P^{-1/2}(\widetilde{K}_n - \widetilde{K})P^{-1/2}\|_{\mathcal{Z},\mathcal{Z}}\|y\|_{\mathcal{Z}_P} \\ &\leqslant \|P^{-1}\|_{\mathcal{Z},\mathcal{Z}}(\|\widetilde{L}_n - \widetilde{L}\|_{\mathcal{Z},\mathcal{Z}}\|z\|_{\mathcal{Z}_P} + \|\widetilde{K}_n - \widetilde{K}\|_{\mathcal{Z},\mathcal{Z}}\|y\|_{\mathcal{Z}_P}) \\ &= \|P^{-1}\|_{\mathcal{Z},\mathcal{Z}}(\|L_n - L\|_{\mathcal{H},\mathcal{H}}\|z\|_{\mathcal{Z}_P} + \|K_n - K\|_{\mathcal{G},\mathcal{H}}\|y\|_{\mathcal{Z}_P}) \\ &= \|P^{-1}\|_{\mathcal{Z},\mathcal{Z}}(\|(K_n - K)H\|_{\mathcal{H},\mathcal{H}}\|z\|_{\mathcal{Z}_P} + \|K_n - K\|_{\mathcal{G},\mathcal{H}}\|y\|_{\mathcal{Z}_P}) \\ &\leqslant \|P^{-1}\|_{\mathcal{Z},\mathcal{Z}}(\|H\|_{\mathcal{H},\mathcal{G}}\|z\|_{\mathcal{Z}_P} + \|y\|_{\mathcal{Z}_P})\|K_n - K\|_{\mathcal{G},\mathcal{H}} \\ &\leqslant \|P^{-1}\|_{\mathcal{Z},\mathcal{Z}}(\|H\|_{\mathcal{H},\mathcal{G}}\|z\|_{\mathcal{Z}_P} + \|y\|_{\mathcal{Z}_P})\,\omega_n,\end{aligned} \tag{5.29}$$

where the last inequality follows from Assumption 4.2.0.1(iii). Altogether (5.27), (5.28), and (5.29) yield, for every $n \in \mathbb{N}$,

$$\begin{aligned}\|z_{n+1} - r_{n+1}\|_{\mathcal{Z}_P} &\leqslant \|z_n - r_n\|_{\mathcal{Z}_P} + \Theta_n \|P^{-1}\|_{\mathcal{Z},\mathcal{Z}} (\|H\|_{\mathcal{H},\mathcal{G}} \|z_n\|_{\mathcal{Z}_P} + \|y\|_{\mathcal{Z}_P}) \omega_n \\ &\leqslant \|z_n - r_n\|_{\mathcal{Z}_P} + \Theta_n \|P^{-1}\|_{\mathcal{Z},\mathcal{Z}} \big(\|H\|_{\mathcal{H},\mathcal{G}} (\|z_n - r_n\|_{\mathcal{Z}_P} + \|r_n\|_{\mathcal{Z}_P}\big) + \|y\|_{\mathcal{Z}_P}) \omega_n \\ &\leqslant (1 + \mu_n) \|z_n - r_n\|_{\mathcal{Z}_P} + \nu_n \end{aligned} \tag{5.30}$$

with

$$\mu_n = \delta \|P^{-1}\|_{\mathcal{Z},\mathcal{Z}} \|H\|_{\mathcal{H},\mathcal{G}}\, \omega_n \tag{5.31}$$

$$\nu_n = \delta \|P^{-1}\|_{\mathcal{Z},\mathcal{Z}} (\varrho \|H\|_{\mathcal{H},\mathcal{G}} + \|y\|_{\mathcal{Z}_P})\, \omega_n. \tag{5.32}$$

Since $(\mu_n) \in \ell^1_+$ and $(\nu_n) \in \ell^1_+$, according to [16, Lemma 5.31], $\|z_n - r_n\|_{\mathcal{Z}_P} < +\infty$. Consequently, $\delta' = \sup_{n\in\mathbb{N}} \|z_n\|_{\mathcal{Z}_P} < \varrho + \sup_{n\in\mathbb{N}} \|z_n - r_n\|_{\mathcal{Z}_P} < +\infty$.
Let us define

$$(\forall n \in \mathbb{N}) \quad e_n = \frac{T_n(z_n) - \overline{T}(z_n)}{\overline{\alpha}}, \tag{5.33}$$

and it follows from (5.29) that

$$\sum_{n\in\mathbb{N}} \|e_n\|_{\mathcal{Z}_P} \leqslant \frac{\|P^{-1}\|_{\mathcal{Z},\mathcal{Z}}}{\overline{\alpha}} (\|H\|_{\mathcal{H},\mathcal{G}} \delta' + \|y\|_{\mathcal{Z}_P}) \sum_{n\in\mathbb{N}} \omega_n < +\infty. \tag{5.34}$$

Algorithm (5.8) can be re-expressed as

$$(\forall n \in \mathbb{N}) \quad z_{n+1} = z_n + \Theta'_n (Q z_n + e_n - z_n) \quad \text{with} \quad \Theta'_n = \overline{\alpha}\Theta_n \in]0,1[. \tag{5.35}$$

Therefore, the weak convergence of $(z_n)_{n\in\mathbb{N}}$ to some $\widetilde{z} \in \operatorname{Fix}(Q) = \operatorname{Fix}\big(\overline{T}\big) = \overline{\mathcal{F}}$ follows from [16, Theorem 5.5]. □
We deduce the following more restrictive convergence result which is an extension of standard convergence results for CV.

Corollary 5.2.2.6 *Assume that* $\operatorname{Ran}(L + L^*)$ *is closed,* $\lambda_{\min} \geqslant 0$*, and* $\operatorname{Ker}(L + L^*) = \operatorname{Ker} L$*. Let* $(\tau, \sigma) \in]0, +\infty[^2$ *be such that* $\tau^{-1} - \sigma\|D\|^2_{\mathcal{H},\mathcal{L}} > \|M\|^2_{\mathcal{H},\mathcal{H}}/4$*, where* M *is given by* (5.10)*. For*

$$\delta = 2 - \frac{1}{4}\Big(\frac{1}{\tau} - \sigma\|D\|^2_{\mathcal{H},\mathcal{L}}\Big)^{-1} \|M\|^2_{\mathcal{H},\mathcal{H}}, \tag{5.36}$$

let $\{\Theta_n\}_{n\in\mathbb{N}} \subset [0, \delta]$ *be a sequence such that* $\sum_{n\in\mathbb{N}} \Theta_n (\delta - \Theta_n) = +\infty$*, and suppose that* $\overline{\mathcal{F}} \neq \varnothing$*. Then the sequence* $((x_n, u_n))_{n\in\mathbb{N}}$ *given by* (5.7) *converges weakly to some point in* $\overline{\mathcal{F}}$*.*

Proof: Note that, if $\tau^{-1} - \sigma\|D\|^2_{\mathcal{H},\mathcal{L}} > \|M\|^2_{\mathcal{H},\mathcal{H}}/4$, then $\tau\sigma\|D\|^2_{\mathcal{H},\mathcal{L}} < 1$. In addition, it follows from Proposition 5.2.2.2 that, when $\operatorname{Ran}(L + L^*)$ is closed, $\lambda_{\min} \geqslant 0$, and $\operatorname{Ker}(L + L^*) = \operatorname{Ker} L$, $P^{-1}\widetilde{L}$ is $\widetilde{\eta}$ -cocoercive in $\mathcal{Z}_P$ with

$$\widetilde{\eta} = \tau^{-1}(1 - \tau\sigma\|D\|^2_{\mathcal{H},\mathcal{L}})\eta_{\max} \tag{5.37}$$

and $\eta_{\max} = 2/\|M\|^2_{\mathcal{H},\mathcal{H}}$. The result then follows by applying Theorem 5.2.2.5. □
We now provide an estimate of the distance between a Kuhn-Tucker pair $(\widehat{x}, \widehat{u})$ of the original optimization problem and a fixed point $(\widetilde{x}, \widetilde{u})$ of $\overline{T}$.

Proposition 5.2.2.7 Assume that (5.23) holds and L is a cocoercive operator.
If $f + g \circ D$ is strongly convex with modulus $\rho > 0$, then there exists a unique vector $\widetilde{x}$ in $\overline{\mathcal{F}}_1$ and a unique solution $\widehat{x}$ to the primal problem (5.1), and we have

$$\|\widetilde{x} - \widehat{x}\|_{\mathcal{H}} \leqslant \frac{1}{\rho} \left\| \left(\overline{K} - H^*\right) (H\widehat{x} - y) \right\|_{\mathcal{H}}. \tag{5.38}$$

In addition, if g is β-Lipschitz differentiable with $\beta \in [0, +\infty[$, there exists a unique $(\widetilde{x}, \widetilde{u}) \in \overline{\mathcal{F}}$ and a unique solution $\widehat{u}$ to the dual problem, and we have

$$\|\widetilde{u} - \widehat{u}\|_{\mathcal{L}} \leqslant \frac{\beta}{\rho} \|D\|_{\mathcal{H},\mathcal{L}} \left\| \left(\overline{K} - H^*\right) (H\widehat{x} - y) \right\|_{\mathcal{H}}. \tag{5.39}$$

Proof: Since (5.23) is satisfied, $\partial f + D^* \circ \partial g \circ D = \partial(f + g \circ D)$ [16, Theorem 16.47 (i)]. As we have assumed that $f + g \circ D$ is ρ-strongly convex and L is cocoercive, $L + \partial f + D^* \circ \partial g \circ D$ is strongly monotone. It follows from Proposition 5.2.2.3(c) that there exists a single element $\widetilde{x}$ in $\overline{\mathcal{F}}_1$ which is such that

$$\overline{K}y \in L\widetilde{x} + \partial f(\widetilde{x}) + D^* \partial g(D\widetilde{x}). \tag{5.40}$$

For any $\gamma > 0$, (5.40) is equivalent to

$$\widetilde{x} = \mathrm{prox}_{\gamma(f+g\circ D)} \left(\widetilde{x} - \gamma\overline{K}(H\widetilde{x} - y)\right). \tag{5.41}$$

For similar reasons, there exists a unique solution $\widehat{x}$ to the primal problem, which satisfies the fixed point equation

$$\widehat{x} = \mathrm{prox}_{\gamma(f+g\circ D)} \left(\widehat{x} - \gamma H^*(H\widehat{x} - y)\right). \tag{5.42}$$

Because of the ρ-strong convexity of $f + g \circ D$, $\mathrm{prox}_{\gamma(f+g\circ D)}$ is $(1 + \gamma\rho)^{-1}$-Lipschitzian. The error bound in (5.38) is thus derived by the same arguments as in the proof of Theorem 4.3.3.1.
In addition, if g is Gâteaux differentiable, there exists a unique $\widetilde{u} \in \mathcal{L}$ such that $(\widetilde{x}, \widetilde{u}) \in \overline{\mathcal{F}}$, which is given by

$$\widetilde{u} = \nabla g(D\widetilde{x}), \tag{5.43}$$

where ∇g is the gradient of g. Similarly, there exists a unique solution $\widehat{u}$ to the dual problem, given by

$$\widehat{u} = \nabla g(D\widehat{x}). \tag{5.44}$$

By using the fact that ∇g is β-Lipschitzian, we deduce that

$$\begin{aligned} \|\widetilde{u} - \widehat{u}\|_{\mathcal{L}} &\leqslant \beta \|D(\widetilde{x} - \widehat{x})\|_{\mathcal{L}} \\ &\leqslant \beta \|D\|_{\mathcal{H},\mathcal{L}} \|\widetilde{x} - \widehat{x}\|_{\mathcal{H}}. \end{aligned} \tag{5.45}$$

The upper error bound in (5.39) then follows from (5.38). □

5.2.3 Remarks on the mismatched projected Condat-Vũ algorithm

When a constraint is added to primal problem (5.1), the latter becomes

$$\text{Find } \widehat{x} \in C \cap \operatorname*{argmin}_{x\in\mathcal{H}} \ \frac{1}{2}\|Hx-y\|_{\mathcal{G}}^2 + f(x) + g(Dx), \tag{5.46}$$

where C is a closed and convex nonempty subset of $\mathcal{H}$. The dual problem reads

$$\text{Find } \widehat{u} \in E \cap \operatorname*{argmin}_{u\in\mathcal{L}} \ (f+h)^*(-D^*u) + g^*(u), \tag{5.47}$$

where E is a closed vector subspace of $\mathcal{L}$ such that $\operatorname{ran} D \subset E$. Such a problem can be solved by the projected form of CV proposed by Briceño-Arias and López in [35]. We are interested in a mismatched form of this algorithm with a fixed operator $\overline{K}$: Given $(x_0, u_0) \in \mathcal{Z}$ and $(\tau, \sigma) \in]0, +\infty[^2$,

Projected mismatched primal-dual algorithm:

$$\text{for } n = 0, 1, \ldots \quad \left\lfloor \begin{array}{l} p_n = \operatorname{prox}_{\tau f}\left(x_n - \tau\left(\overline{K}\left(Hx_n - y\right) + D^*u_n\right)\right) \\ x_{n+1} = \operatorname{proj}_C(p_n) \\ \overline{x}_n = x_{n+1} + p_n - x_n \\ q_n = \operatorname{prox}_{\sigma g^*}\left(u_n + \sigma D\overline{x}_n\right) \\ u_{n+1} = \operatorname{proj}_E\left(q_n\right) \end{array} \right. . \tag{5.48}$$

In particular, if $\overline{K} = H^*$, $C = \mathcal{H}$, and $E = \mathcal{L}$, we recover the CV Iteration (5.6) in the case when, for every $n \in \mathbb{N}$, $\Theta_n = 1$.
Then we obtain the following convergence result for the same L and $\overline{\mathcal{F}}$.

Proposition 5.2.3.1 Assume that $\operatorname{Ran}(L + L^*)$ is closed, $\lambda_{\min} \geqslant 0$, and $\operatorname{Ker}(L + L^*) = \operatorname{Ker} L$. Let $(\tau, \sigma) \in]0, +\infty[^2$ be such that $\tau^{-1} - \sigma\|D\|_{\mathcal{H},\mathcal{L}}^2 > \|M\|_{\mathcal{H},\mathcal{H}}^2/4$, where M is given by (5.10). Suppose that $\overline{\mathcal{F}} \cap (C \times E) \neq \varnothing$. Then the sequence $((x_n, u_n))_{n\in\mathbb{N}}$ generated by (5.48) converges weakly to some point in $\overline{\mathcal{F}} \cap (C \times E)$. In addition $(p_n - x_n)_{n\in\mathbb{N}}$ and $(q_n - u_n)_{n\in\mathbb{N}}$ converge strongly to 0.

Proof: On the one hand, under the assumptions made on L, we have seen that $x \mapsto \overline{K}(Hx - y)$ is $\eta_{\max}$-cocoercive with $\eta_{\max} = 2/\|M\|_{\mathcal{H},\mathcal{H}}^2$. On the other hand, [35, Theorem 3.2 (ii)] guarantees the weak convergence of $((x_n, u_n))_{n\in\mathbb{N}}$ under the condition

$$\|D\|_{\mathcal{H},\mathcal{L}}^2 < \frac{1}{\sigma}\left(\frac{1}{\tau} - \frac{1}{2\eta_{\max}}\right). \tag{5.49}$$

The strong convergence properties of $(p_n - x_n)_{n\in\mathbb{N}}$ and $(q_n - u_n)_{n\in\mathbb{N}}$ are also stated in the proof of [35, Theorem 3.2]. □
Note that the error related to the mismatch can still be quantified by Proposition 5.2.2.7.

5.3 The mismatched Loris-Verhoeven algorithm

In this section, we consider a specific instance of our original template (5.1) where $f = \frac{\kappa}{2}\|\cdot\|^2_{\mathcal{H}}$ with $\kappa \in]0,+\infty[$. This problem can be solved by the primal-dual algorithm proposed by Loris and Verhoeven [143] shown in Chapter 3 (section 3.3).
Like the CV iterations, the LV iterations can be described by means of the implicit inclusion (5.2) where $\mathcal{A}$, B, and P are now given by

$$(\forall z = (x,u) \in \mathcal{Z}) \quad \mathcal{A}z = \begin{pmatrix} D^*u \\ -Dx + \partial g^*(u) \end{pmatrix} \tag{5.50}$$

$$Bz = \begin{pmatrix} (H^*H + \kappa\,\mathrm{Id}_{\mathcal{H}})x - H^*y \\ 0 \end{pmatrix} \tag{5.51}$$

$$Pz = \begin{pmatrix} \frac{1}{\tau}x \\ (\frac{1}{\sigma}\,\mathrm{Id}_{\mathcal{L}} - \tau DD^*)u \end{pmatrix} \tag{5.52}$$

with $(\tau,\sigma) \in]0,+\infty[^2$. In detail, the LV iterations take the form

LV iterations:

$$\text{for } n = 0,1,\ldots \quad \left\lfloor \begin{array}{l} t_n = H^*\left(Hx_n - y\right) + \kappa x_n \\ u_{n+\frac{1}{2}} = \mathrm{prox}_{\sigma g^*}\Big(u_n + \sigma D\Big(x_n - \tau(t_n + D^*u_n)\Big)\Big) \\ x_{n+1} = x_n - \Theta_n\tau\Big(t_n + D^*u_{n+\frac{1}{2}}\Big) \\ u_{n+1} = u_n + \Theta_n\Big(u_{n+\frac{1}{2}} - u_n\Big) \end{array} \right. \tag{5.53}$$

where $\{\Theta_n\}_{n\in\mathbb{N}}$ is a sequence of relaxation parameters and $(x_0,u_0) \in \mathcal{Z}$.
When, at iteration $n \in \mathbb{N}$, H^* is replaced by an operator $K_n \in \mathcal{B}(\mathcal{H},\mathcal{G})$ satisfying Assumption 4.2.0.1(iii), the mismatched form of LV algorithm reads

Mismatched LV iterations:

$$\text{for } n = 0,1,\ldots \quad \left\lfloor \begin{array}{l} t_n = K_n\left(Hx_n - y\right) + \kappa x_n \\ u_{n+\frac{1}{2}} = \mathrm{prox}_{\sigma g^*}\Big(u_n + \sigma D\Big(x_n - \tau(t_n + D^*u_n)\Big)\Big) \\ x_{n+1} = x_n - \Theta_n\tau\Big(t_n + D^*u_{n+\frac{1}{2}}\Big) \\ u_{n+1} = u_n + \Theta_n\Big(u_{n+\frac{1}{2}} - u_n\Big) \end{array} \right. \tag{5.54}$$

This iteration can be reexpressed as (5.8) where Notation 5.2.2.1(ii)-(iv) holds, but L is now defined as

$$L = \overline{K}H + \kappa\mathrm{Id}_{\mathcal{H}}. \tag{5.55}$$

It follows that all the results in subsection 5.2.2 can be extended to the mismatched LV algorithm.
Proposition 5.3.0.1 is a straightforward adaptation of Proposition 5.2.2.3 to characterize the fixed point set of the nonlinear mapping $\overline{T}$ (see Notation 5.2.2.1(iv)).

Proposition 5.3.0.1 Let $(\widetilde{x},\widetilde{u}) \in \mathcal{Z}$. Then $(\widetilde{x},\widetilde{u}) \in \mathrm{Fix}\left(\overline{T}\right)$ if and only if $(\widetilde{x},\widetilde{u})$ belongs to

$$\overline{\mathcal{F}} = \left\{(x,u) \in \mathcal{Z} \mid \overline{K}y \in Lx + D^*u,\ u \in \partial g(Dx)\right\}, \tag{5.56}$$

which is nonempty if $L + D^* \circ \partial g \circ D$ is surjective.

(i) If $\lambda_{\min} \geqslant 0$, then $\overline{\mathcal{F}}$ is closed and convex.

(ii) Let $\overline{\mathcal{F}}_1 = \{x \in \mathcal{H} \mid (\exists u \in \mathcal{L})\ (x,u) \in \overline{\mathcal{F}}\}$.
$\overline{\mathcal{F}}_1$ has at most one element if one of the following conditions holds:

(a) $L + D^* \circ \partial g \circ D$ is strictly monotone.

(b) $\lambda_{\min} \geqslant 0$ and $g \circ D$ is strictly convex.

$\overline{\mathcal{F}}_1$ is a singleton if (5.23) is satisfied and one of the following conditions holds:

(c) $\lambda_{\min} \geqslant 0$ and $L + D^* \circ \partial g \circ D$ is strongly monotone.

(d) $\lambda_{\min} > 0$.

(e) $\lambda_{\min} \geqslant 0$, g is strongly convex, and D^*D is strongly positive.

(f) L is cocoercive, g is strongly convex, and D^*D is strongly positive.

Similarly, we derive an equivalent of Corollary 5.2.2.6 concerning the convergence of the mismatched LV.

Proposition 5.3.0.2 Assume that $\mathrm{Ran}(L+L^*)$ is closed, $\lambda_{\min} \geqslant 0$, and $\mathrm{Ker}(L+L^*) = \mathrm{Ker}\, L$. Let $(\tau,\sigma) \in]0,+\infty[^2$ be such that $\tau < 4/\|M\|^2_{\mathcal{H},\mathcal{H}}$ and $\tau\sigma\|D\|^2_{\mathcal{H},\mathcal{L}} < 1$, where M is given by (5.10). For

$$\delta = 2 - \frac{\tau}{4}\|M\|^2_{\mathcal{H},\mathcal{H}}, \tag{5.57}$$

let $\{\Theta_n\}_{n\in\mathbb{N}} \subset [0,\delta]$ be a sequence such that $\sum_{n\in\mathbb{N}} \Theta_n\,(\delta - \Theta_n) = +\infty$, and suppose that $\overline{\mathcal{F}} \neq \varnothing$. Then the sequence $((x_n,u_n))_{n\in\mathbb{N}}$ given by (5.54) converges weakly to some point in $\overline{\mathcal{F}}$.

Proof: The result from Proposition 5.2.2.2 stating that, when $\mathrm{Ran}(L+L^*)$ is closed, $\lambda_{\min} \geqslant 0$, and $\mathrm{Ker}(L+L^*) = \mathrm{Ker}\, L$, $P^{-1}\widetilde{L}$ is cocoercive in $\mathcal{Z}_P$ with constant $\widetilde{\eta}_{\max} = 2/\|P^{-1/2}\Pi M\|^2_{\mathcal{H},\mathcal{Z}}$ still holds for LV algorithm. We also have

$$\widetilde{\eta}_{\max} \geqslant \frac{2}{\|P^{-1/2}\Pi\|^2_{\mathcal{H},\mathcal{Z}}\|M\|^2_{\mathcal{H},\mathcal{H}}} = \frac{2}{\tau\|M\|^2_{\mathcal{H},\mathcal{H}}} = \widetilde{\eta}, \tag{5.58}$$

which shows that $P^{-1}\widetilde{L}$ is $\widetilde{\eta}$-cocoercive in $\mathcal{Z}_P$. The result then follows from Theorem 5.2.2.5, which remains valid in this context. □

Remark 5.3.0.3 When there is no mismatch, $M = \sqrt{2}(H^*H + \kappa\,\mathrm{Id}_{\mathcal{H}})^{1/2}$ and we recover the conditions derived in [143, Theorem 3.1], [68, Theorem 3.1] for the convergence of sequences $(x_n)_{n\in\mathbb{N}}$ and $(u_n)_{n\in\mathbb{N}}$ generated by (5.53).

Similarly to Proposition 5.2.2.7, we provide an estimate of the distance between a Kuhn-Tucker pair $(\widehat{x},\widehat{u})$ of the original problem and a fixed point $(\widetilde{x},\widetilde{u})$ of $\overline{T}$.

Proposition 5.3.0.4 Assume that $0 \in \mathrm{sri}\,(\mathrm{Ran}(D) - \mathrm{dom}(g))$ and L, given in (5.55), is a cocoercive operator.
Let $\rho \in]0,+\infty[$. If $\kappa \geqslant \rho$ or $g \circ D$ is strongly convex with modulus ρ, then there exists a unique vector $\widetilde{x}$ in $\overline{\mathcal{F}}_1$, defined in Proposition 5.3.0.1(ii), and a unique solution $\widehat{x}$ to the primal problem (5.1). Moreover, inequality (5.38) holds.
In addition, if g is β-Lipschitz differentiable with $\beta \in [0,+\infty[$, there exists a unique $(\widetilde{x},\widetilde{u}) \in \overline{\mathcal{F}}$, defined in (5.56), and a unique solution $\widehat{u}$ to the dual problem. Finally, inequality (5.39) is satisfied.

5.4 The mismatched Combettes-Pesquet algorithm

5.4.1 Algorithm

The last algorithm explored in this chapter is the Combettes - Pesquet algorithm, which relies on Tseng's splitting to solve (5.1). We recall that CP iterations read

CP-iterations for (5.1):

$$\text{for } n = 0, 1, \dots \quad \left\lfloor \begin{array}{l} v_{1,n} = x_n - \gamma \left(H^*(Hx_n - y) + D^* u_n \right) \\ p_{1,n} = \operatorname{prox}_{\gamma f}(v_{1,n}) \\ v_{2,n} = u_n + \gamma D x_n \\ p_{2,n} = \operatorname{prox}_{\gamma g^*}(v_{2,n}) \\ q_{2,n} = p_{2,n} + \gamma D p_{1,n} \\ q_{1,n} = p_{1,n} - \gamma \left(H^*(Hp_{1,n} - y) + D^* p_{2,n} \right) \\ x_{n+1} = x_n - v_{1,n} + q_{1,n} \\ u_{n+1} = u_n - v_{2,n} + q_{2,n}, \end{array} \right. \tag{5.59}$$

where $\gamma > 0$ and $(x_0, u_0) \in \mathcal{Z}$. By setting, for every $n \in \mathbb{N}$,

$$z_n = (x_n, u_n)\,,\ v_n = (v_{1,n}, v_{2,n})\,,\ p_n = (p_{1,n}, p_{2,n})\,,\ \text{ and } \ q_n = (q_{1,n}, q_{2,n})\,, \tag{5.60}$$

(5.59) can be rewritten as

Tseng iterations:

$$\text{for } n = 0, 1, \dots \quad \left\lfloor \begin{array}{l} v_n = z_n - \gamma Q(z_n) \\ p_n = J_{\gamma \mathscr{M}}(v_n) \\ q_n = p_n - \gamma Q(p_n) \\ z_{n+1} = z_n - v_n + q_n, \end{array} \right. \tag{5.61}$$

with

$$\mathscr{M} \colon \mathcal{Z} \to 2^{\mathcal{Z}} \colon (x, u) \mapsto (\partial f(x), \partial g^*(u)) \tag{5.62}$$

$$Q \colon \mathcal{Z} \to \mathcal{Z} \colon (x, u) \mapsto (H^*Hx + D^*u - H^*y, -Dx)\,. \tag{5.63}$$

For Algorithm (5.59), we consider a mismatched form obtained by substituting a fixed operator $\overline{K} \in \mathcal{B}(\mathcal{G}, \mathcal{H})$ for H^* as well as an operator $V^* \in \mathcal{B}(\mathcal{L}, \mathcal{H})$ for D^*. This leads to:

Mismatched CP-iterations:

$$\text{for } n = 0, 1, \dots \quad \left\lfloor \begin{array}{l} v_{1,n} = x_n - \gamma \left(\overline{K}(Hx_n - y) + V^* u_n \right) \\ p_{1,n} = \operatorname{prox}_{\gamma f}(v_{1,n}) \\ v_{2,n} = u_n + \gamma(Dx_n + \varepsilon u_n) \\ p_{2,n} = \operatorname{prox}_{\gamma g^*}(v_{2,n}) \\ q_{2,n} = p_{2,n} + \gamma(Dp_{1,n} + \varepsilon p_{2,n}) \\ q_{1,n} = p_{1,n} - \gamma \left(\overline{K}(Hp_{1,n} - y) + V^* p_{2,n} \right) \\ x_{n+1} = x_n - v_{1,n} + q_{1,n} \\ u_{n+1} = u_n - v_{2,n} + q_{2,n}, \end{array} \right. \tag{5.64}$$

where $\varepsilon > 0$ is an extra parameter. Note that there are two algorithmic modifications other than the presence of mismatched adjoints in the update rules of variables $v_{2,n}$ and $q_{2,n}$.

By making the change of variables (5.60), Algorithm (5.64) can be rewritten in the product space $\mathcal{Z}$, as

$$\text{Tseng form of (5.64):}$$
$$\text{for } n = 0, 1, \ldots \quad \left| \begin{array}{l} v_n = z_n - \gamma \widetilde{Q}(z_n) \\ p_n = J_{\gamma \mathscr{M}}(v_n) \\ q_n = p_n - \gamma \widetilde{Q}(p_n) \\ z_{n+1} = z_n - v_n + q_n, \end{array} \right. \tag{5.65}$$

where

$$\widetilde{Q} \colon \mathcal{Z} \to \mathcal{Z} \colon (x, u) \mapsto \big(Lx + V^* u - \overline{K} y, -Dx + \varepsilon u\big) \tag{5.66}$$

and

$$L = \overline{K} H, \tag{5.67}$$

as in section 5.2.

5.4.2 Convergence analysis

5.4.2.1 Regularity of the surrogate gradient operator

We first provide some preliminary results on operator $\widetilde{Q}$.

Proposition 5.4.2.1 Let $\lambda_{\min}$ be defined in Notation 5.2.2.1 and

$$^{\varepsilon}\lambda_{\min} = \lambda_{\min} - \frac{1}{4\varepsilon} \|V - D\|^2_{\mathcal{H},\mathcal{L}}, \tag{5.68}$$
$$^{\varepsilon}\vartheta_1 = \max\{\|L\|_{\mathcal{H},\mathcal{H}}, \varepsilon\} + \max\{\|D\|_{\mathcal{H},\mathcal{L}}, \|V\|_{\mathcal{H},\mathcal{L}}\}, \tag{5.69}$$
$$^{\varepsilon}\vartheta_2 = \sqrt{\|L\|^2_{\mathcal{H},\mathcal{H}} + \|V\|^2_{\mathcal{H},\mathcal{L}} + \|D\|^2_{\mathcal{H},\mathcal{L}} + \varepsilon^2}, \tag{5.70}$$

and $^{\varepsilon}\vartheta = \min\{^{\varepsilon}\vartheta_1, {}^{\varepsilon}\vartheta_2\}$. We have the following properties:

(i) $\widetilde{Q}$ is Lipschitz continuous with constant $^{\varepsilon}\vartheta$.

(ii) If $^{\varepsilon}\lambda_{\min} \geqslant 0$, then $\widetilde{Q}$ is monotone.

(iii) If $^{\varepsilon}\lambda_{\min} > 0$, then $\widetilde{Q}$ is strongly monotone and cocoercive.

Proof: Since $\widetilde{Q}$ is an affine operator, its monotonicity, Lipschitz continuity, and cocoercivity properties are the same as those of the linear operator

$$\overline{Q} = \widetilde{Q} + (\overline{K} y, 0). \tag{5.71}$$

(i) On the one hand

$$\begin{aligned} \|\overline{Q}\|_{\mathcal{Z},\mathcal{Z}} &\leqslant \left\| \begin{bmatrix} L & 0 \\ 0 & \varepsilon \operatorname{Id}_{\mathcal{L}} \end{bmatrix} \right\|_{\mathcal{Z},\mathcal{Z}} + \left\| \begin{bmatrix} 0 & V^* \\ -D & 0 \end{bmatrix} \right\|_{\mathcal{Z},\mathcal{Z}} \\ &\leqslant \max\{\|L\|_{\mathcal{H},\mathcal{H}}, \varepsilon\} + \max\{\|D\|_{\mathcal{H},\mathcal{L}}, \|V\|_{\mathcal{H},\mathcal{L}}\} = {}^{\varepsilon}\vartheta_1. \end{aligned} \tag{5.72}$$

On the other hand, for every $z = (x, u) \in \mathcal{Z}$,

$$\begin{aligned} \|\overline{Q} z\|^2_{\mathcal{Z}} &= \|Lx + V^* u\|^2_{\mathcal{H}} + \| - Dx + \varepsilon u\|^2_{\mathcal{L}} \\ &\leqslant \|LL^* + V^* V\|_{\mathcal{H},\mathcal{H}} \|z\|^2_{\mathcal{Z}} + \|DD^* + \varepsilon^2 \operatorname{Id}_{\mathcal{L}}\|_{\mathcal{L},\mathcal{L}} \|z\|^2_{\mathcal{Z}} \\ &\leqslant (\|L\|^2_{\mathcal{H},\mathcal{H}} + \|V\|^2_{\mathcal{H},\mathcal{L}} + \|D\|^2_{\mathcal{H},\mathcal{L}} + \varepsilon^2) \|z\|^2_{\mathcal{Z}}, \end{aligned} \tag{5.73}$$

which implies that

$$\|\overline{Q}\|_{\mathcal{Z},\mathcal{Z}} \leqslant {}^{\varepsilon}\vartheta_2. \tag{5.74}$$

In summary, $\overline{Q}$, and thus $\widetilde{Q}$, are Lipschitz continuous with constant $\|\overline{Q}\|_{\mathcal{Z},\mathcal{Z}} \leqslant {}^{\varepsilon}\vartheta$.

(ii) By using Cauchy-Schwarz inequality, for every $z = (x, u) \in \mathcal{Z}$,

$$\begin{aligned}\left\langle \overline{Q}z, z\right\rangle_{\mathcal{Z}} &= \langle Lx, x\rangle_{\mathcal{H}} + \varepsilon\|u\|_{\mathcal{L}}^2 + \langle u, (V-D)x\rangle_{\mathcal{L}} \\ &\geqslant \lambda_{\min}\|x\|_{\mathcal{H}}^2 + \varepsilon\|u\|_{\mathcal{L}}^2 - \|u\|_{\mathcal{L}}\|(V-D)x\|_{\mathcal{L}} \\ &\geqslant \lambda_{\min}\|x\|_{\mathcal{H}}^2 + \varepsilon\|u\|_{\mathcal{L}}^2 - \|V-D\|_{\mathcal{H},\mathcal{L}}\|u\|_{\mathcal{L}}\|x\|_{\mathcal{H}} \\ &= \begin{bmatrix}\|x\|_{\mathcal{H}} & \|u\|_{\mathcal{L}}\end{bmatrix} C \begin{bmatrix}\|x\|_{\mathcal{H}} \\ \|u\|_{\mathcal{L}}\end{bmatrix},\end{aligned} \tag{5.75}$$

where

$$C = \begin{bmatrix}\lambda_{\min} & -\frac{1}{2}\|V-D\|_{\mathcal{H},\mathcal{L}} \\ -\frac{1}{2}\|V-D\|_{\mathcal{H},\mathcal{L}} & \varepsilon\end{bmatrix}. \tag{5.76}$$

C is positive semidefinite if and only if

$$\begin{cases}\operatorname{tr}(C) = \lambda_{\min} + \varepsilon \geqslant 0 \\ \det(C) = \lambda_{\min}\varepsilon - \frac{1}{4}\|V-D\|_{\mathcal{H},\mathcal{L}}^2 \geqslant 0,\end{cases} \tag{5.77}$$

that is ${}^{\varepsilon}\lambda_{\min} \geqslant 0$. we deduce from (5.75) that, subject to this condition, $\overline{Q}$ and thus $\widetilde{Q}$ are monotone.

(iii) Assume now that ${}^{\varepsilon}\lambda_{\min} > 0$. Then C is positive definite and its smallest eigenvalue is

$${}^{\varepsilon}\upsilon = \frac{\lambda_{\min} + \varepsilon - \sqrt{(\lambda_{\min}+\varepsilon)^2 - 4\,{}^{\varepsilon}\lambda_{\min}\varepsilon}}{2} > 0. \tag{5.78}$$

It follows from (5.75) that $\overline{Q}$ is strongly monotone with constant ${}^{\varepsilon}\upsilon$.
We have then, for every $z \in \mathcal{Z}$,

$$\left\langle \overline{Q}z, z\right\rangle_{\mathcal{Z}} \geqslant {}^{\varepsilon}\upsilon\,\|z\|_{\mathcal{Z}}^2 \geqslant {}^{\varepsilon}\upsilon\,\frac{\|\overline{Q}z\|_{\mathcal{Z}}^2}{\|\overline{Q}\|_{\mathcal{Z},\mathcal{Z}}^2} \geqslant \frac{{}^{\varepsilon}\upsilon}{({}^{\varepsilon}\vartheta)^2}\,\|\overline{Q}z\|_{\mathcal{Z}}^2.$$

This shows that $\overline{Q}$ (and thus $\widetilde{Q}$) is cocoercive with constant

$${}^{\varepsilon}\eta = \frac{{}^{\varepsilon}\upsilon}{({}^{\varepsilon}\vartheta)^2}. \tag{5.79}$$

□

5.4.2.2 Characterization of the fixed points of the mismatched iterations

We now characterize the set of limit points.

Proposition 5.4.2.2 Let ${}^{\varepsilon}g$ be the Moreau envelope of g of parameter $\varepsilon > 0$ defined as

$$(\forall v \in \mathcal{L}) \quad {}^{\varepsilon}g(v) = \inf_{w\in\mathcal{L}} g(w) + \frac{1}{2\varepsilon}\|w - v\|_{\mathcal{L}}^2. \tag{5.80}$$

Let $(\widetilde{x}, \widetilde{u}) \in \mathcal{Z}$. Then $(\widetilde{x}, \widetilde{u}) \in \operatorname{zer}(\mathscr{M} + \widetilde{Q})$ if and only if $(\widetilde{x}, \widetilde{u})$ belongs to

$${}^{\varepsilon}\overline{\mathcal{F}} = \left\{(x,u) \in \mathcal{Z} \mid \overline{K}y \in \partial f(x) + Lx + V^*u,\ u = \nabla\,{}^{\varepsilon}g(Dx) = \frac{Dx - \operatorname{prox}_{\varepsilon g}(Dx)}{\varepsilon}\right\}, \tag{5.81}$$

which is nonempty if $L + \partial f + V^* \circ \nabla\,{}^{\varepsilon}g \circ D$ is surjective.

(i) If ${}^{\varepsilon}\lambda_{\min}$ defined by (5.68) is nonnegative, then ${}^{\varepsilon}\overline{\mathcal{F}}$ is closed and convex.

(ii) Let ${}^{\varepsilon}\overline{\mathcal{F}}_1 = \{x \in \mathcal{H} \mid (x, \nabla {}^{\varepsilon}g(Dx)) \in {}^{\varepsilon}\overline{\mathcal{F}}\}$ and let

$$ {}^{\varepsilon}\lambda_{1,\min} = \lambda_{\min} - \frac{1}{\varepsilon}\|V - D\|_{\mathcal{H},\mathcal{L}}\|D\|_{\mathcal{H},\mathcal{L}} \geqslant 0. \tag{5.82} $$

${}^{\varepsilon}\overline{\mathcal{F}}_1$ has at most one element if one of the following conditions holds:

(a) $L + \partial f + V^* \circ \nabla {}^{\varepsilon}g \circ D$ is strictly monotone.

(b) ${}^{\varepsilon}\lambda_{1,\min} \geqslant 0$ and ${}^{\varepsilon}g \circ D + f$ is strictly convex.

$\overline{\mathcal{F}}_1$ is a singleton if one of the following conditions holds:

(c) $L + \partial f + V^* \circ \nabla {}^{\varepsilon}g \circ D$ is strongly monotone.

(d) ${}^{\varepsilon}\lambda_{1,\min} > 0$

(e) ${}^{\varepsilon}\lambda_{1,\min} \geqslant 0$, and f is strongly convex or [g^* is Lipschitz-differentiable and D^*D is strongly positive].

Proof: The proof follows the same lines as in the proof of Proposition 5.2.2.3. In the following, we point out the main differences.
By using (5.10) and (5.66), we have

$$ \begin{aligned} &(\widetilde{x}, \widetilde{u}) \in \operatorname{zer}(\mathscr{M} + \widetilde{Q}) \\ \Leftrightarrow &\begin{cases} 0 \in \partial f(\widetilde{x}) + L\widetilde{x} + V^*\widetilde{u} - \overline{K}y \\ 0 \in \partial g^*(\widetilde{u}) - D\widetilde{x} + \varepsilon\widetilde{u}. \end{cases} \end{aligned} \tag{5.83} $$

We know that $({}^{\varepsilon}g)^* = g^* + \frac{\varepsilon}{2}\|\cdot\|^2 \Rightarrow \partial({}^{\varepsilon}g)^*(\widetilde{u}) = \partial g^*(\widetilde{u}) + \varepsilon\widetilde{u}$ [16, Proposition 14.1] and ${}^{\varepsilon}g$ is differentiable with gradient $\nabla {}^{\varepsilon}g = \varepsilon^{-1}(\mathrm{Id}_{\mathcal{L}} - \operatorname{prox}_{\varepsilon g})$ [16, Proposition 12.30]. We thus deduce that

$$ \begin{aligned} &(\widetilde{x}, \widetilde{u}) \in \operatorname{zer}(\mathscr{M} + \widetilde{Q}) \\ \Leftrightarrow &\begin{cases} \overline{K}y \in \partial f(\widetilde{x}) + L\widetilde{x} + V^*\widetilde{u} \\ D\widetilde{x} \in \partial({}^{\varepsilon}g)^*(\widetilde{u}) \end{cases} \\ \Leftrightarrow &\begin{cases} \overline{K}y \in \partial f(\widetilde{x}) + L\widetilde{x} + V^*\widetilde{u} \\ \widetilde{u} = \nabla {}^{\varepsilon}g(D\widetilde{x}). \end{cases} \end{aligned} \tag{5.84} $$

This shows that $(\widetilde{x}, \widetilde{u}) \in {}^{\varepsilon}\overline{\mathcal{F}}$.

(i) According to Proposition 5.4.2.1(i)-(ii), if ${}^{\varepsilon}\lambda_{\min} \geqslant 0$, $\widetilde{Q}$ is monotone and continuous. Since $\mathscr{M}$ is maximally monotone, $\mathscr{M} + \widetilde{Q}$ is also maximally monotone and $\operatorname{zer}(\mathscr{M} + \widetilde{Q})$ is closed and convex.

(ii) We can perform the decomposition

$$ L + \partial f + V^* \circ \nabla {}^{\varepsilon}g \circ D = \partial f + D^* \circ \nabla {}^{\varepsilon}g \circ D + L + (V - D)^* \circ \nabla {}^{\varepsilon}g \circ D. \tag{5.85} $$

Subdifferential $\partial f + D^* \circ \nabla {}^{\varepsilon} g \circ D = \partial(f + {}^{\varepsilon} g \circ D)$ is maximally monotone. Using the Cauchy-Schwarz inequality, for every $(x, x') \in \mathcal{H}^2$,

$$\begin{aligned}
&\langle L(x - x'), x - x' \rangle_{\mathcal{H}} + \langle (V - D)^* \nabla {}^{\varepsilon} g(Dx) - (V - D)^* \nabla {}^{\varepsilon} g(Dx'), x - x' \rangle_{\mathcal{H}} \\
&\geqslant \lambda_{\min} \|x - x'\|_{\mathcal{H}}^2 - \left\| (V - D)^* \left(\nabla {}^{\varepsilon} g(Dx) - \nabla {}^{\varepsilon} g(Dx') \right) \right\|_{\mathcal{H}} \|x - x'\|_{\mathcal{H}} \\
&\geqslant \lambda_{\min} \|x - x'\|_{\mathcal{H}}^2 - \|V - D\|_{\mathcal{H},\mathcal{L}} \|\nabla {}^{\varepsilon} g(Dx) - \nabla {}^{\varepsilon} g(Dx')\|_{\mathcal{L}} \|x - x'\|_{\mathcal{H}} \\
&\geqslant \lambda_{\min} \|x - x'\|_{\mathcal{H}}^2 - \frac{1}{\varepsilon} \|V - D\|_{\mathcal{H},\mathcal{L}} \|D(x - x')\|_{\mathcal{L}} \|x - x'\|_{\mathcal{H}} \\
&\geqslant {}^{\varepsilon}\lambda_{1,\min} \|x - x'\|_{\mathcal{H}}^2, \qquad (5.86)
\end{aligned}$$

where we have used the ε^{-1}-Lipschitz continuity of $\nabla {}^{\varepsilon} g$. This shows that $L + (V - D)^* \circ \nabla {}^{\varepsilon} g \circ D$ is monotone, if ${}^{\varepsilon}\lambda_{1,\min} \geqslant 0$. In addition, it is strongly monotone (hence, strictly monotone) if ${}^{\varepsilon}\lambda_{1,\min} > 0$. Since this operator is also continuous, it is maximally monotone in both cases. The rest of the proof is similar to that of Proposition 5.2.2.3, by noticing that ${}^{\varepsilon} g$ is strongly convex $\Leftrightarrow$ $({}^{\varepsilon} g)^*$ is Lipschitz-differentiable $\Leftrightarrow$ g^* is Lipschitz-differentiable.

□

5.4.2.3 Convergence conditions and error bound

Conditions of convergence for our mismatched CP algorithm are deduced from this result.

Proposition 5.4.2.3 Let ${}^{\varepsilon}\lambda_{\min}$ and ${}^{\varepsilon}\vartheta$ be defined as in Proposition 5.4.2.1. Let $\gamma \in]0, ({}^{\varepsilon}\vartheta)^{-1}[$. Assume that $\operatorname{zer}(\mathscr{M} + \widetilde{Q}) \neq \varnothing$ and ${}^{\varepsilon}\lambda_{\min} \geqslant 0$. Then the sequences $((x_n, u_n))_{n \in \mathbb{N}}$ and $((p_{1,n}, p_{2,n}))_{n \in \mathbb{N}}$ generated by Algorithm (5.64) converge weakly to $(\widetilde{x}, \widetilde{u}) \in \operatorname{zer}(\mathscr{M} + \widetilde{Q})$. In addition, if ${}^{\varepsilon}\lambda_{\min} > 0$, then $((x_n, u_n))_{n \in \mathbb{N}}$ and $((p_{1,n}, p_{2,n}))_{n \in \mathbb{N}}$ converge strongly to the unique zero of $\mathscr{M} + \widetilde{Q}$.

Proof: Under the considered assumptions, $\widetilde{Q}$ is monotone and ${}^{\varepsilon}\vartheta$-Lipschitzian. Thus, the result follows from standard conditions for the convergence of Tseng's algorithm applied to (5.65). Then, if ${}^{\varepsilon}\lambda_{\min} > 0$, $\widetilde{Q}$ is strongly monotone and the strong convergence property follows from [16, Theorem 26.17 (iii)]. □

We now provide a bound on the distance between an optimal pair of solutions $(\widehat{x}, \widehat{u})$ to Problem (5.1) and $(\widetilde{x}, \widetilde{u}) \in {}^{\varepsilon}\overline{\mathcal{F}}$.

Proposition 5.4.2.4 Let ${}^{\varepsilon}\lambda_{\min}$ be defined by (5.68). Assume that ${}^{\varepsilon}\lambda_{\min} > 0$, f is strongly convex with modulus $\rho > 0$, and g is Lipschitz-differentiable with constant $\beta > 0$. Then, there exists a unique pair $\widetilde{z} = (\widetilde{x}, \widetilde{u}) \in {}^{\varepsilon}\overline{\mathcal{F}}$ and a unique solution $(\widehat{x}, \widehat{u})$ to the primal-dual problem. In addition,

$$\sqrt{\|\widetilde{x} - \widehat{x}\|_{\mathcal{H}}^2 + \|\widetilde{u} - \widehat{u}\|_{\mathcal{L}}^2} \leqslant \frac{1}{\mu} \left(\|(\overline{K} - H^*)(H\widehat{x} - y)\|_{\mathcal{H}} + \sqrt{\|V - D\|_{\mathcal{H},\mathcal{L}}^2 + \varepsilon^2}\, \|\widehat{u}\|_{\mathcal{L}} \right), \qquad (5.87)$$

where $\mu = \min\{\rho, 1/\beta\}$.

Proof: f is ρ-strongly convex and g is β-Lispchitz differentiable (i.e., g^* is β^{-1}-strongly convex), ∂f and ∂g^* are strongly monotone with constants ρ and $1/\beta$, respectively. $\mathscr{M}$ is thus strongly monotone with constant μ. Since Q is continuous and monotone, $\mathscr{M} + Q$ is maximally monotone and strongly monotone. The existence of a unique zero $\widehat{z}$ to

$\mathscr{M}+Q$ is thus guaranteed by [16, Corollary 23.37]. Similarly, it follows from Proposition 5.4.2.1(i) and Proposition 5.4.2.1(iii) that $\widetilde{Q}$ is continuous and strongly monotone. Hence $\mathscr{M}+\widetilde{Q}$ is maximally monotone and strongly monotone and $\operatorname{zer}(\mathscr{M}+\widetilde{Q})$ is a singleton $\{\widetilde{z}\}$.
For every $\gamma>0$,

$$\widehat{z}\in\operatorname{zer}(\mathscr{M}+Q) \quad\Leftrightarrow\quad \widehat{z}=J_{\gamma.\mathscr{M}}(\widehat{z}-\gamma Q\widehat{z}) \tag{5.88}$$

$$\widetilde{z}\in\operatorname{zer}(\mathscr{M}+\widetilde{Q}) \quad\Leftrightarrow\quad \widetilde{z}=J_{\gamma.\mathscr{M}}(\widetilde{z}-\gamma\widetilde{Q}\widetilde{z}) \tag{5.89}$$

Since $\mathscr{M}$ is strongly monotone with constant μ, $J_{\gamma.\mathscr{M}}$ is Lipschitz continuous with constant $1/(1+\gamma\mu)$ [16, Proposition 23.13]. From (5.88) and (5.89), we deduce that

$$\begin{aligned}\|\widetilde{z}-\widehat{z}\|_{\mathcal{Z}} &\leqslant \frac{1}{1+\gamma\mu}\|\widetilde{z}-\widehat{z}-\gamma(\widetilde{Q}\widetilde{z}-Q\widehat{z})\|_{\mathcal{Z}}\\ &\leqslant \frac{1}{1+\gamma\mu}\|(\mathrm{Id}_{\mathcal{Z}}-\gamma\widetilde{Q})(\widetilde{z}-\widehat{z})-\gamma(\widetilde{Q}-Q)\widehat{z}\|_{\mathcal{Z}}\\ &\leqslant \frac{1}{1+\gamma\mu}\left(\|\mathrm{Id}_{\mathcal{Z}}-\gamma\widetilde{Q}\|_{\mathcal{Z},\mathcal{Z}}\|\widetilde{z}-\widehat{z}\|_{\mathcal{Z}}+\gamma\|(\widetilde{Q}-Q)\widehat{z}\|_{\mathcal{Z}}\right).\end{aligned} \tag{5.90}$$

According to Proposition 5.4.2.1(iii), $\widetilde{Q}$ is cocoercive with constant ${}^{\varepsilon}\eta$ given by (5.79). Therefore, by assuming that $\gamma\in]0,{}^{\varepsilon}\eta]$, we have $\|\mathrm{Id}_{\mathcal{Z}}-\gamma\widetilde{Q}\|_{\mathcal{Z},\mathcal{Z}}\leqslant 1$. We deduce from (5.90) that

$$\begin{aligned}\|\widetilde{z}-\widehat{z}\|_{\mathcal{Z}} &\leqslant \frac{1}{\mu}\|(\widetilde{Q}-Q)\widehat{z}\|_{\mathcal{Z}}\\ &= \frac{1}{\mu}\left\|\left((\overline{K}-H^*)(H\widehat{x}-y)+(V-D)^*\widehat{u},\varepsilon\widehat{u}\right)\right\|_{\mathcal{Z}}\\ &\leqslant \frac{1}{\mu}\left(\left\|\left((\overline{K}-H^*)(H\widehat{x}-y),0\right)\right\|_{\mathcal{Z}}+\left\|\left((V-D)^*\widehat{u},\varepsilon\widehat{u}\right)\right\|_{\mathcal{Z}}\right)\\ &\leqslant \frac{1}{\mu}\left(\|(\overline{K}-H^*)(H\widehat{x}-y)\|_{\mathcal{H}}+\sqrt{\|V-D\|_{\mathcal{H},\mathcal{L}}^2+\varepsilon^2}\,\|\widehat{u}\|_{\mathcal{L}}\right).\end{aligned} \tag{5.91}$$

□

Remark 5.4.2.5

(i) In the absence of mismatch on D^* (i.e., $V^*=D^*$), one can choose $\varepsilon=0$ in (5.64) and (5.66). $\operatorname{zer}(\mathscr{M}+\widetilde{Q})$ is then equal to the set $\overline{\mathcal{F}}$ characterized in Proposition 5.2.2.3. If $\overline{\mathcal{F}}\neq\varnothing$, $\lambda_{\min}\geqslant 0$, and $\gamma\in]0,({}^{0}\vartheta)^{-1}[$, then the sequences $((x_n,u_n))_{n\in\mathbb{N}}$ and $((p_{1,n},p_{2,n}))_{n\in\mathbb{N}}$ generated by Algorithm (5.64) converge weakly to $(\widetilde{x},\widetilde{u})\in\overline{\mathcal{F}}$. Proposition 5.2.2.7 applies to evaluate the mismatch error.

(ii) When $H=\overline{K}=0$, it follows from [140, Theorem 1.2] that, if f is strongly convex with constant $\rho>0$, $(\widetilde{x},\widetilde{u})\in{}^{\varepsilon}\overline{\mathcal{F}}$ and $\widehat{x}$ is the solution to the primal problem, then

$$\|\widetilde{x}-\widehat{x}\|_{\mathcal{H}}\leqslant\frac{1}{\rho}\|V-D\|_{\mathcal{H},\mathcal{L}}\|\widetilde{u}\|_{\mathcal{L}}. \tag{5.92}$$

5.5 Application

This section illustrates our theoretical results applied to the resolution of 2D image reconstruction problems. All the simulations presented in this section are performed using the ASTRA Toolbox [218, 219] in Matlab.

5.5.1 Example 1: reconstruction from few CT views

We aim at recovering an image $\overline{x}$ with N pixels, reshaped as a vector in the Euclidean space $\mathcal{H} = \mathbb{R}^N$. A set of noisy tomographic projections $p \in \mathcal{G} = \mathbb{R}^S$ of the original image $\overline{x}$ is available, according to the following observation model:

$$p = R\overline{x} + b \tag{5.93}$$

where $R \in \mathbb{R}^{S\times N}$ is the FP, and b is an additive i.i.d. zero-mean Gaussian noise. R is chosen as the line-length ray-driven projector [239]. A surrogate adjoint of R, denoted by $B \in \mathbb{R}^{N\times S}$, is the pixel-driven backprojector which is particularly suited for a GPU implementation compared to the adjoint of R [88]. For (u, v) i.i.d. uniformly sampled on $([0, 1]^N)^2$, the average over 20 realizations of the ratio $\langle Ru \mid v\rangle / \langle u \mid Bv\rangle$ is 1.005.
We aim to retrieve an estimate of $\overline{x}$ given p, the projector R, and its surrogate adjoint B.

5.5.1.1 Data

In model (5.93), $\overline{x}$ is an axial slice of an abdomen of size 41 cm whose values range from 1000 sHU to 3000 sHU, containing intense inserts with pixel intensity ranges in $[3500, 4200]$ sHU (see Figure 5.2). The source-to-object distance is 800 mm, and the source-to-image distance is 1200 mm. R describes a fan-beam geometry over 180° using 50 regularly spaced angular steps. The detector has a length of 40 cm. The bin grid is twice up-sampled with respect to the pixel grid: the detector has 250 bins of size 1.6 mm so that $S = 250 \times 250$. The image is reconstructed on a grid of $N = 160 \times 160$ pixels, with size $2 \times 1.6/1.5 = 2.13$ mm. Data p is obtained after adding 1% relative Gaussian noise on $R\overline{x}$. The inverse problem (5.93) is highly ill-posed because of the small detector FOV and the limited angular density. This problem corresponds to a setup where the detector is not large enough to measure the projections of large body parts such as the abdomen. The set of pixels in the image whose projections belong to the detector FOV defines an image FOV. Because of truncation, we estimate the exterior of the FOV so that the FOV can be accurately reconstructed. The size of the reconstruction grid is thus slightly larger than the support of the FOV.

5.5.1.2 Regularization

We provide an estimate of $\overline{x}$ by solving the following penalized least squares problem:

$$\underset{x\in\mathbb{R}^N}{\text{minimize}} \ \frac{1}{2}\|p - Rx\|^2 + f(x) + g(Dx) + \frac{\kappa}{2}\|x\|^2 \tag{5.94}$$

with $\kappa \in [0, +\infty[$. We promote sparsity of the image vertical and horizontal gradients [39]. We additionally constrain the nonnegativity of the reconstructed pixel intensities. This leads us to set $f = \iota_{[0,+\infty[^N}$ where ι_C denotes the indicator function of a set C. Moreover, we set $g = \xi\|\cdot\|_{1,2}$ with $D = \nabla = \begin{bmatrix}\nabla^{\mathrm{h}} \\ \nabla^{\mathrm{v}}\end{bmatrix}$, where $\nabla^{\mathrm{h}} \in \mathbb{R}^{N\times N}$, $\nabla^{\mathrm{v}} \in \mathbb{R}^{N\times N}$

are, respectively, the horizontal and vertical discrete gradient operators (assuming zero-padding) and $\|\cdot\|_{1,2}$ is the $\ell_{1,2}$-norm of $\mathbb{R}^N$, so that $g \circ D$ is the discrete total variation penalty weighted by $\xi \in [0, +\infty[$ [187]. We set the regularization hyperparameter ξ to 800 through a grid search, minimizing the error on the image FOV.

5.5.1.3 Condat-Vũ algorithm

Problem (5.94) can be rewritten as (5.1) with

$$H = \begin{bmatrix} R \\ \sqrt{\kappa}\mathrm{Id}_{\mathbb{R}^N} \end{bmatrix},$$
$$y = \begin{bmatrix} p \\ 0 \end{bmatrix},$$

where $\kappa \in [0, +\infty[$. The surrogate adjoint of H is

$$\overline{K} = \begin{bmatrix} B & \sqrt{\kappa}\mathrm{Id}_{\mathbb{R}^N} \end{bmatrix}.$$

For such a problem, we can apply the CV approach presented in section 5.2.
We run Algorithms (5.6) and (5.7) (i.e., CV algorithm without/with an adjoint mismatch, respectively) for $\kappa \in \{\kappa_1, \kappa_2\}$ where $L = \overline{K}H = BR + \kappa\mathrm{Id}_{\mathbb{R}^N}$ is only cocoercive for $\kappa = \kappa_2$. In the latter case, the condition given in Proposition 5.2.2.3(d) holds, which proves the existence of a unique fixed point $\widetilde{x}$ of scheme (5.7) and its convergence is ensured according to Corollary 5.2.2.6. In contrast, nothing can be said about the convergence of the scheme in the case involving κ_1. We set $\kappa_1 = 1$ and κ_2 to $-\widetilde{\lambda}_{\min} + 0.01$ where $\widetilde{\lambda}_{\min} = -28.24$ is the minimum spectral value of $(BR + R^*B^*)/2$ estimated from the power iterative method. The cocoercivity constant η is computed using Proposition 5.2.2.2(ii).
The convergence parameter σ is set to 10^{-3}. For Algorithm (5.6), the step size τ is set to $0.99/(8\sigma + 0.5/\theta)$ with $\theta = 1/\|H\|^2_{\mathcal{H},\mathcal{G}}$. To illustrate the instabilities incurred by the use of the mismatched adjoint $\overline{K}$ when using κ_1, the same step size value τ is used as in the matched case. With κ_2, the convergence of Algorithm (5.7) is ensured, as stated by Corollary 5.2.2.6 and Proposition 5.2.2.2, by setting $\tau = 0.99/(8\sigma + \|M\|^2_{\mathcal{H},\mathcal{H}}/4)$, where M is defined in (5.10). Both algorithms are run until a maximum number 3×10^4 of iterations is reached. Initial iterates x_0 and u_0 are set to zero.

5.5.1.4 Loris-Verhoeven algorithm

Problem (5.94) can also be solved with the LV approach presented in section 5.3. The cost function can be rewritten as in (5.1) by setting

$$H = R,$$
$$y = p,$$
$$D = \begin{bmatrix} \nabla \\ \mathrm{Id}_{\mathbb{R}^N} \end{bmatrix},$$
$$f = \frac{\kappa}{2}\|\cdot\|^2,$$

and

$$\left(\forall (z_1, z_2) \in (\mathbb{R}^N)^2\right) \quad g\left(\begin{bmatrix} z_1 \\ z_2 \end{bmatrix}\right) = \|z_1\|_{1,2} + \iota_{[0,+\infty[^N}(z_2). \tag{5.95}$$

Similarly to the CV case, $L = BR + \kappa\mathrm{Id}_{\mathbb{R}^N}$. Therefore, for the mismatched LV algorithm (5.54), the existence and uniqueness of a fixed point $\widetilde{x}$ in $\overline{\mathcal{F}}_1$, defined in Proposition

5.3.0.1(ii)(c), is only guaranteed when $\kappa = \kappa_2$, but not when $\kappa = \kappa_1$. The convergence parameter σ is set as $1.99/(9\tau)$. For Algorithm (5.53) with both values of κ and Algorithm (5.54) with κ_1, step size τ is set to $1.99/(\|H\|^2_{\mathcal{H},\mathcal{G}} + \kappa)$. The convergence of Algorithm (5.54) with κ_2 is ensured by setting τ to $3.99/\|M\|^2_{\mathcal{H},\mathcal{H}}$, where M is the same as in the CV case, in accordance with Proposition 5.3.0.2.

5.5.1.5 Results

Figure 5.1 displays the normalized root mean square error (NMSE) defined as $(\|\overline{x} - x_n\|/\|\overline{x}\|)_n$, computed along the iterations when applying CV Algorithms (5.6)-(5.7) and LV Algorithms (5.53)-(5.54). We recall that Algorithms (5.6) and (5.53) require the use of the exact adjoint of H. The plots confirm that with κ_1, both CV and LV algorithms converge when this exact adjoint H^* is used, but diverge when H^* is replaced by $\overline{K}$, as was expected from our results.
In the latter case, CV and LV algorithms show an initial convergence trend before diverging. We notice that on this example, the mismatched CV (5.7) diverges faster than the mismatched LV (5.54). When reaching the maximal number of iterations, the NMSE associated with (5.7) is 0.93 whereas the NMSE associated with (5.54) is 0.57.
When using κ_2, all algorithms converge to close fixed points, as expected by our theory. The corresponding NMSE values are 0.251 for (5.6), 0.242 for (5.7), 0.253 for (5.53) and 0.252 for (5.54). Remarkably, our mismatched algorithms (5.7)-(5.54) with κ_2 lead to lower reconstruction error in the first iterations of the algorithms.
Reconstructed images and their FOVs are displayed in Figure 5.2 using the same windowing. Note that the reconstructions obtained using (5.53) look the same as those obtained with (5.6). Likewise, the same reconstruction is obtained with (5.7) and (5.54), when κ_2 is used. For all the reconstructed images, we also provide the NMSE and the maximum absolute error (MAE) computed in the FOV image, defined as $\max_{i\in\{1,\dots,N\}} |[\mathrm{mask}_{\mathrm{FOV}}(\overline{x} - x)]_i|$ where $[\mathrm{mask}_{\mathrm{FOV}}(x_n)]_i = [x_n]_i$ if pixel i of x_n is in the FOV, and $[\mathrm{mask}_{\mathrm{FOV}}(x_n)]_i = 0$ otherwise. When parameter κ_2 is used, the reconstructed image obtained by CV/LV with the mismatched adjoint $\overline{K}$ (MAE=461, NMSE=0.040) is very similar to the image obtained without mismatch (MAE=495, NMSE=0.041). In contrast, combining the setting κ_1 with the mismatched adjoint yields reconstructions that are highly deteriorated by high-frequency patterns leading to a higher error (MAE=2735, NMSE=0.637 for CV and MAE=1249, NMSE=0.249 for LV) compared to the solution obtained when using the exact adjoint (MAE=215, NMSE=0.026 for both CV and LV).

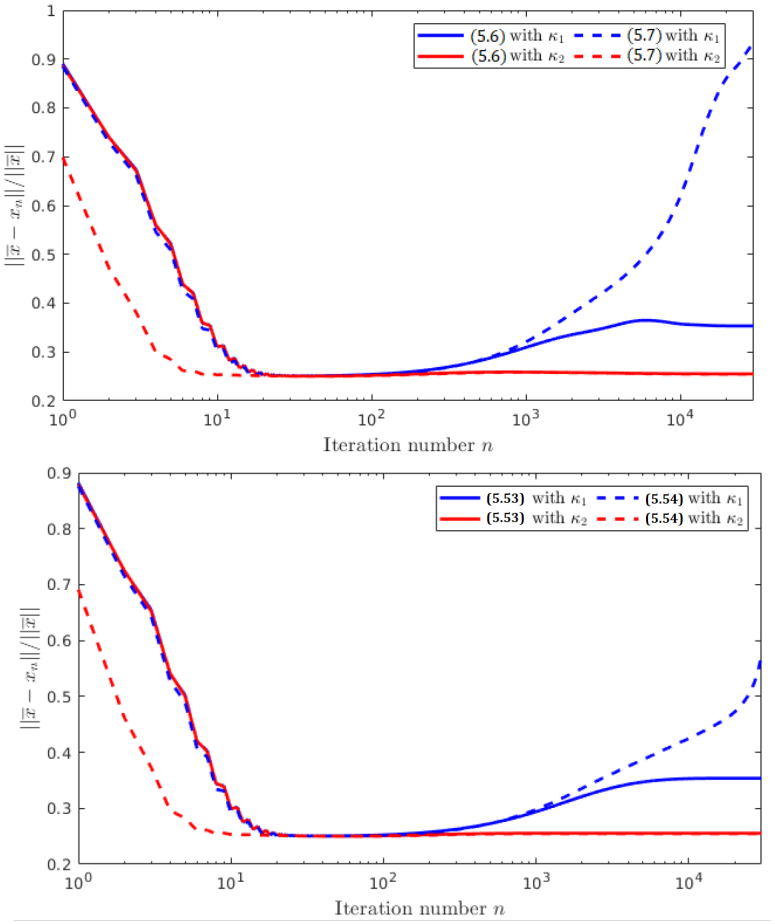


Figure 5.1: Evolution of the error $(\|\overline{x} - x_n\|/\|\overline{x}\|)_n$ along iterations for Algorithms (5.6)-(5.7) (top) and Algorithms (5.53)-(5.54) (bottom), for two settings of parameter κ.

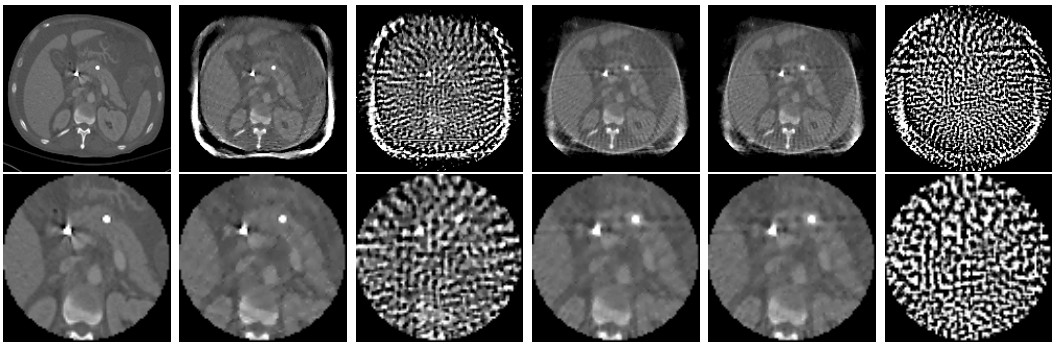


Figure 5.2: Reconstructed images (top) and zoomed FOVs (bottom). From left to right: $\overline{x}$, reconstructions obtained using (5.6) with κ_1, (5.7) with κ_1, (5.6) with κ_2, (5.7) with κ_2, (5.54) with κ_1.

5.5.2 Example 2: reconstruction from Poisson data

In this second example, we focus on another acquisition scenario: the projection data p contain photon count views that follow a Poisson distribution:

$$p = \mathcal{P}(R\overline{x}), \tag{5.96}$$

that each component p_s of $p \in \mathbb{R}^S$, $s \in \{1, \dots, S\}$, is drawn independently from a Poisson distribution with mean $[R\overline{x}]_s$. The goal is again to restore an estimate of $\overline{x}$ given p, R, and its mismatched adjoint B.

5.5.2.1 Data

This second test problem uses a fan-beam geometry with 200 views. In model (5.96), $\overline{x}$ is an axial slice of an abdomen. Contrary to Example 1, it is now made only of an anatomical background (see Figure 5.3). We set the source-to-object distance, the source-to-image distance, the detector length, the number of bins, and the bin size as in Example 1. Hence, $S = 250 \times 200$. Projections p are then simulated by using model (5.96). The image is reconstructed on a discrete grid of $N = 220 \times 220$ pixels, with size 2.13 mm. B is derived from the same discretization scheme as in Example 1.

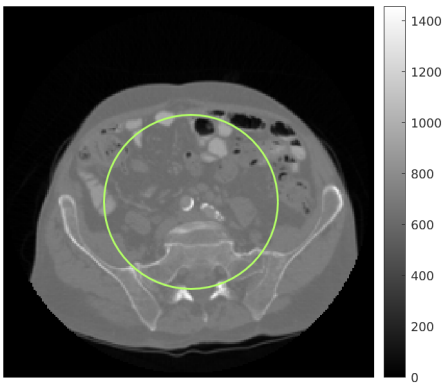


Figure 5.3: Phantom $\overline{x}$ with highlighted FOV

5.5.2.2 Regularization

Due to Poisson noise, the data discrepancy term in the cost function differs from the one in Example 1. Namely, we introduce the negative log-likelihood of the image given the data [127] to define $\ell \circ R$ with

$$(\forall z = (z_s)_{1 \leqslant s \leqslant S} \in \mathbb{R}^S) \qquad \ell(z) = \sum_{s=1}^{S} \mathcal{KL}(z_s, p_s) \tag{5.97}$$

and

$$(\forall (u, v) \in \mathbb{R}^2) \qquad \mathcal{KL}(u, v) = \begin{cases} -v \log u + u & \text{if } u > 0, v > 0 \\ u & \text{if } u \geqslant 0, v = 0 \\ +\infty & \text{otherwise.} \end{cases} \tag{5.98}$$

Furthermore, we introduce a nonnegativity constraint on the components of the solution and a Tikhonov-based penalty. Altogether, the resulting minimization problem reads

$$\underset{x \in \mathbb{R}^N}{\text{minimize}} \;\; \chi\, \ell(Rx) + \frac{1}{2}\|Hx\|^2 + \iota_{[0,+\infty[^N}(x) \tag{5.99}$$

where $\chi \in [0, +\infty[$ weights the Poisson data fidelity term and we define the linear operator

$$H = \begin{bmatrix} \Delta \\ \kappa \mathrm{Id}_{\mathbb{R}^N} \end{bmatrix}.$$

Moreover, $\Delta \in \mathbb{R}^{N \times N}$ refers to the 2D discrete Laplacian operator (here, implemented in the 2D Fourier domain), and $\kappa \in [0, +\infty[$.
Problem (5.99) can be rewritten as (5.1) with

$$\begin{aligned} y &= 0_{\mathbb{R}^{2N}}, \\ D &= R, \\ g &= \chi\, \ell, \\ f &= \iota_{[0,+\infty[^N}. \end{aligned}$$

5.5.2.3 Combettes-Pesquet algorithm

Problem (5.99) is solved with CP algorithm (5.59)-(5.64). In our configuration, an adjoint mismatch only arises on D (i.e., the projector R), and again denoting by B the mismatched adjoint. We hence have $V = B^*$ and $\overline{K} = H^*$ in Algorithm (5.64).
To guarantee the convergence of Algorithm (5.64) to a unique fixed point, we must choose (κ, ε) to satisfy the conditions of Proposition 5.4.2.1 and Proposition 5.4.2.2. We first set κ and then choose $4\varepsilon = \|V - D\|^2_{\mathcal{H},\mathcal{L}}/(\kappa^2 - 0.02)$ so that ${}^{\varepsilon}\lambda_{\min} = 0.02 > 0$ in (5.68). In particular, we consider the setting $(\kappa_2, \varepsilon_2) = (3.3, 6.6)$ for which convergence is guaranteed and the setting $(\kappa_1, \varepsilon_1) = (0.6, 0)$ for which it is not guaranteed.
In Algorithm (5.59), parameter γ is set to $0.99/(4 + \kappa^2 + \|D\|_{\mathcal{H},\mathcal{L}})$. When Algorithm (5.64) is run with κ_1, γ is set as in the matched case. When Algorithm (5.64) is run with κ_2 and ε_2, we ensure its convergence using Proposition 5.4.2.3 by setting $\gamma = 0.99/{}^{\varepsilon}\vartheta$ with ${}^{\varepsilon}\vartheta = \min({}^{\varepsilon}\vartheta_1, {}^{\varepsilon}\vartheta_2) = {}^{\varepsilon}\vartheta_1$ where ${}^{\varepsilon}\vartheta_1$ and ${}^{\varepsilon}\vartheta_2$ are defined respectively by (5.69) and (5.70). We set the data fidelity parameter χ to 5000, and we perform 4000 iterations of CP algorithm. Similarly to Example 1, the initial iterates x_0 and u_0 are set to zero.

5.5.2.4 Results

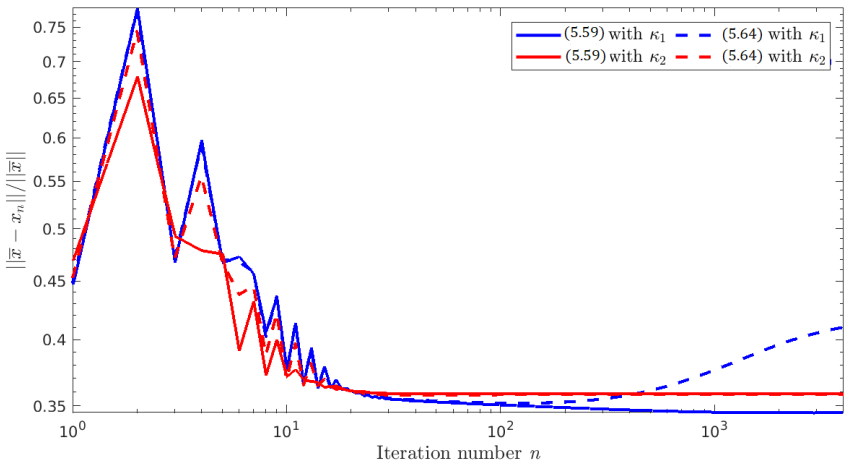


Figure 5.4: Evolution of the error $\|\overline{x} - x_n\|/\|\overline{x}\|$ along iterations for CP Algorithms (5.59) and (5.64), for two settings of parameter κ.

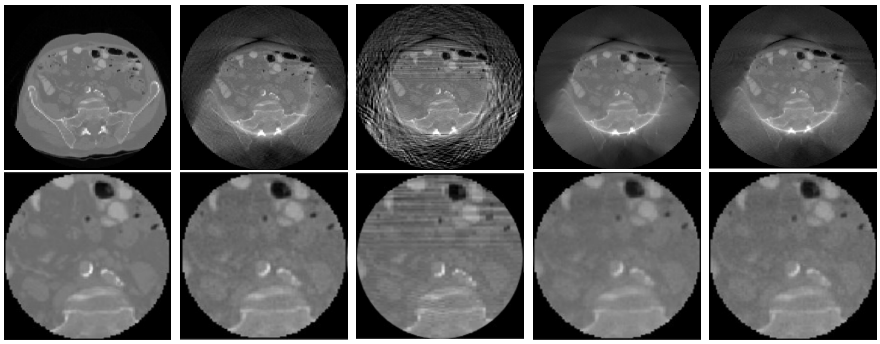


Figure 5.5: Reconstructed images (top) and zoomed FOVs (bottom). From left to right: $\overline{x}$, reconstructed images using (5.59) with κ_1 (NMSE = 0.052, MAE = 201), (5.64) with κ_1 (NMSE = 0.079, MAE = 401), (5.59) with κ_2 (NMSE = 0.056, MAE = 232), (5.64) with κ_2 (NMSE = 0.055, MAE = 206).

In Figure 5.4, we plot the relative error between the ground truth $\overline{x}$ and the estimate along the iterations. We observe the same behavior as in our previous example. Algorithm (5.64) with $(\kappa_1, \varepsilon_1)$ diverges quickly, while it converges to a fixed point with $(\kappa_2, \varepsilon_2)$. This fixed point is indistinguishable from the minimizer of (5.99) with $\kappa = \kappa_2$, both in terms of NMSE/MAE and visual inspection (see Figure 5.5).

5.6 Conclusion

In this chapter, we analyzed the stability of a set of primal-dual proximal splitting algorithms when the adjoints of the involved linear operators have been replaced by surrogates. By relying on the results of Chapter 4, we established necessary conditions to ensure the convergence of these modified algorithms when applied to non-smooth convex penalized least-squares problems. We illustrated our results through two numerical examples of image reconstruction where an adjoint mismatch occurs on FP. A quadratic and a more sophisticated Poisson fidelity term have been considered in our experiments. In both cases, we showed that convergence can still be guaranteed for an unmatched FP/BP pair.

6 | Magnification-driven cone-beam tomographic operators

6.1 Introduction

In Chapter 3 (section 3.4), we outlined one motivation for using unmatched FP/BP pairs, i.e., avoiding the redundancies that may be produced by the adjoint of the FP in each IR step. Standard FP/BP use resampling transforms. They are ray-driven and voxel-driven; they fit a continuous function on given known points and then perform (linear) interpolation on new points. Ray-driven FP and voxel-driven BP are not each other transposes. Other matched FP/BP pairs have been proposed for IR, namely the Distance-Driven (DD) [71, 72] and its generalization, the Separable Footprint (SF) [138] pairs. Since X-ray detector bins are small surfaces over which the X-ray energy is integrated, and the volume is reconstructed on a Cartesian grid, DD and SF pairs share a geometrical perspective: given a cubic shape for the voxel of the volume, the projective anisotropic footprint of the shape over the detector, and its relation to the detector bin surface, are modeled with respect to the rotation of the system. One essential aspect of such a footprint approach is making explicit assumptions regarding the shape, thus size and sampling, of both the volume voxels and the detector bins. These assumptions apply equally to FP and BP, yielding symmetry. This property contrasts with ray-driven and voxel-driven models, which specify the sampling on either the volume or the detector but cannot handle both.
Since the FP/BP pair is used during each iteration of IR at least once, an optimal practical implementation should be fast, accurate, and memory-saving. Model separability is an essential driver for selecting a fast FP. The DD model, in particular, offers one of the best compromises between computation cost and image quality for diagnostic CT. However, implementing the DD and SF pairs on GPUs is not straightforward [44, 136].
This chapter will show that it is unnecessary to use the viewpoint of a footprint approach to model varying sampling levels. A seminal image resizing scheme based on the convolution of polynomial B-splines displays this key feature and provides an optimal approximation [215] in terms of L_2-norm. Hereinafter, we propose a resampling scheme based on families of B-splines of varying widths to account for the magnifications introduced by the homographies found in flat-panel cone-beam projection. This defines a new *magnification-driven* interpolation framework for discretizing the projection and backprojection operations in CBCT.
The chapter is organized as follows. Section 6.2 recalls cone-beam projection on a plane. Then section 6.3 describes the B-spline-based scheme for image resizing, and section 6.4 extends it to derive a resampling scheme for tomographic homographies. New discrete projection and backprojection operators based on the convolution of B-splines are proposed for either AR or IR. Different approximations are given in subsection 6.4.4 for a fast

implementation of the FP/BP schemes. Furthermore, we highlight the relation between state-of-the-art discretizations and our approach in section 6.5, which concludes with the connection between these pairs and data pre-processing in clinical practice. Experiments on simulated data are presented in section 6.6 and discussed for analytical and iterative reconstruction. Finally, an illustration of FDK reconstruction with several interpolation options on a real data case is provided.

6.2 Flat-panel cone-beam geometry

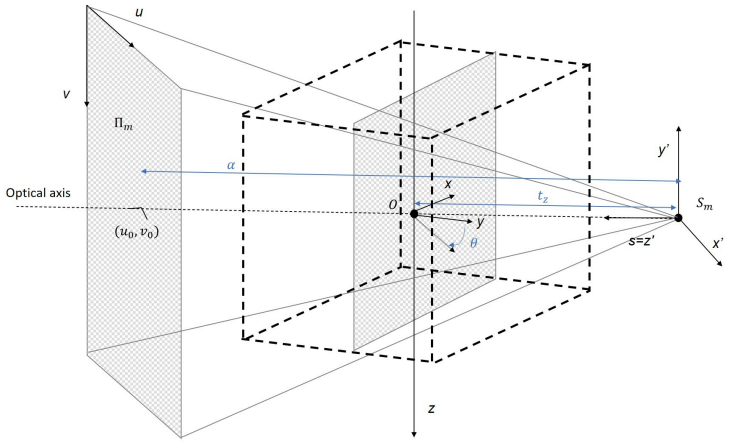


Figure 6.1: Cone-beam geometry. (O, x, y, z) is the volume coordinate system; (S_m, x', y', z') is the source coordinate system ; (u, v) is the detector plane. Ideal acquisition: z, v and y' are aligned.

The projective geometry defines the relationship between voxel coordinates $(x, y, z) \in \mathbb{R}^3$ and the coordinates of the projected pixels $(u, v) \in \mathbb{R}^2$. In X-ray cone-beam computed tomography with a flat-panel detector, the data acquisition is characterized by a set of $M \in \mathbb{N}^*$ projection matrices $(\mathbf{P}_m)_{1 \leqslant m \leqslant M}$ of size 3×4, that is one projection matrix per position of the pair X-ray source/detector. For a given projection matrix $\mathbf{P}_m$, coordinates (u, v) of the projection of point (x, y, z) onto the detection plane Π_m for the position S_m of the source can be written with homogeneous coordinates (su, sv, s) [89] as

$$\big(su, sv, s\big)^\top = \mathbf{P}_m \, \big(x, y, z, 1\big)^\top . \tag{6.1}$$

Coordinate $s = 0$ is at the focus point S_m and lies on the optical axis, which is orthogonal to the detector and crosses S_m. Point O is the center of the volume and the center of rotation. Projection matrices can be measured very accurately. They provide a precise, compact, and powerful way of capturing cone-beam geometry in a continuous space. Axes orientations are shown on Figure 6.1 as well as an additional intermediate 3D coordinate system (S_m, x', y', z') attached to S_m.
Forward projection with one matrix is independent of the other matrices. In contrast, backprojection requires the M projected images as it is the sum over $m \in \{1, \dots, M\}$ of the backprojection of each single projected image obtained using matrix $\mathbf{P}_m$. For every m, the only common condition we set, pertaining to tomography, is that the projection matrices operate on the same coordinate system (O, x, y, z) such that z is an axis of rotation always aligned with the axis v of the detector. We can thus focus the discussion on a single matrix and drop index m.

Any projection matrix $\mathbf{P} = (p_{i,j})_{1 \leqslant i \leqslant 3, 1 \leqslant j \leqslant 4}$ can be decomposed into the product of a matrix $\mathbf{P}_\mathrm{i}$ of intrinsic parameters relating (x', y', z') to (u, v) and matrix $\mathbf{P}_\mathrm{e}$ of extrinsic parameters relating (x, y, z) to (x', y', z') [146]. Matrix $\mathbf{P}_\mathrm{i}$ is defined by

$$\mathbf{P}_\mathrm{i} = \begin{pmatrix} \alpha & 0 & u_0 \\ 0 & -\alpha & v_0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{6.2}$$

where α is the source-to-detector distance in the unit of pixel size and (u_0, v_0) are the coordinates of the orthogonal projection of point S over the detector, also called the piercing point where the optical axis crosses the detector plane. Here again, the unit of length is the pixel size, which is given with the data at backprojection, while it is a parameter for projection. Matrix $\mathbf{P}_\mathrm{e}$ is a 3D rotation and translation operator that, given our specified tomographic conditions, is given by

$$\mathbf{P}_\mathrm{e} = \begin{pmatrix} \cos\theta & \sin\theta & 0 & t_x \\ 0 & 0 & -1 & t_y \\ -\sin\theta & \cos\theta & 0 & t_z \end{pmatrix}, \tag{6.3}$$

where θ is the rotation angle within plane (x, y), (t_x, t_y) are translations that, when not equal to 0, capture a centering shift of the detector, and t_z is the distance from source S to origin O which is also set as the center of rotation. The unit of length is the voxel size, which is a parameter at backprojection, while it is given with the volume at projection. The optical axis is positioned at angle $\theta + \pi$ in this configuration. It follows that $\mathbf{P}$ has two null coefficients:

$$\mathbf{P} = \mathbf{P}_\mathrm{i}\mathbf{P}_\mathrm{e} = \begin{pmatrix} p_{1,1} & p_{1,2} & 0 & p_{1,4} \\ p_{2,1} & p_{2,2} & p_{2,3} & p_{2,4} \\ p_{3,1} & p_{3,2} & 0 & p_{3,4} \end{pmatrix}. \tag{6.4}$$

Matrix $\mathbf{P}$ provides direct access to key parameters since $p_{3,1} = -\sin\theta$, $p_{3,2} = \cos\theta$, $p_{3,4} = t_z$, and $p_{2,3} = \alpha$. However, it does not give access to the voxel and pixel units of length but only to their ratio given by α/t_z, i.e., the magnification factor at the origin O. This ratio is used as a reference. Projection operations are performed at equivalent sampling for a ratio of 1 (called isosampling). A ratio greater than 1 oversamples the detector side or undersamples the volume side, and inversely for a ratio lower than 1. Discretizing tomographic operators over the Cartesian grid can be seen as decomposing the cone-beam projection of a volume into the weighted sum along one axis of the projection of each volume plane orthogonal to said axis. Each projection is thus turned into a homography. More precisely, when $|\cos\theta| > |\sin\theta|$, we use axis y that is closest to the optical axis; otherwise, axis x is used. This ensures that all homographies are invertible. Without loss of generality, we now consider that $t_x = t_y = 0$ as these translations do not change the sampling issues.

Let us consider a summation along axis y: coordinates (u, v) of the projection of any point (x, y_0, z) of the volume coronal plane $y = y_0$ onto the detector plane are given by

$$\big(su, sv, s\big)^\top = \mathbf{P}\,\big(x, y_0, z, 1\big)^\top = \mathbf{H}_{y_0}\,\big(x, z, 1\big)^\top \tag{6.5}$$

$$\text{with } \mathbf{H}_{y_0} = \begin{pmatrix} p_{1,1} & 0 & p_{1,2}y_0 + p_{1,4} \\ p_{2,1} & p_{2,3} & p_{2,2}y_0 + p_{2,4} \\ p_{3,1} & 0 & p_{3,2}y_0 + p_{3,4} \end{pmatrix} = \begin{pmatrix} h_{1,1} & 0 & h_{1,3} \\ h_{2,1} & h_{2,2} & h_{2,3} \\ h_{3,1} & 0 & h_{3,3} \end{pmatrix},$$

so that

$$\left\{\begin{array}{rcl} s(x) & = & h_{3,1}x + h_{3,3} \\ u & = & h_1(x) = \frac{h_{1,1}x+h_{1,3}}{s(x)} \\ v & = & h_2(x,z) = \frac{h_{2,1}x+h_{2,2}z+h_{2,3}}{s(x)}. \end{array}\right. \tag{6.6}$$

The projection of plane $y = y_0$ is thus a resampling by 2D homography $\mathbf{H}_{y_0}$ which, in our tomographic case, displays a resampling in v that is a 1D magnification between v and z of factor $h_{2,2}/s(x)$.
The resampling in u is a 1D homography of x only, corresponding to a flat-detector fan-beam geometry. The projective relationship between (x, y_0, z) and (u, v) can be equivalently defined using another matrix $\mathbf{H}_{y_0}^{-1}$ whose structure is the same as the one of $\mathbf{H}_{y_0}$, as

$$\big(tx, tz, t\big)^\top = \mathbf{H}_{y_0}^{-1} \big(u, v, 1\big)^\top \tag{6.7}$$

where $t > 0$, and

$$\left\{\begin{array}{rcl} x & = & h_1^{-1}(u) \\ z & = & h_2^{-1}(u,v). \end{array}\right. \tag{6.8}$$

Let L be the number of voxels in the volume and K be the number of detector cell measurements acquired in a conic geometry with a flat panel detector. Applying this pipeline to all homographies $\mathbf{H}_s$ and $\tilde{\mathbf{H}}_s^{-1}$ deduced from projection matrix $\mathbf{P}$ gives rise to backprojection matrix $\mathbf{B}_s \in \mathbb{R}^{L\times K}$ for AR and forward projection matrix $\mathbf{R}_s$ for IR.
We now present two resampling approaches for a degenerate case of 1D homography with a fixed magnification factor resulting in a 1D magnification.

6.3 Resampling for a 1D magnification

6.3.1 B-splines

First, we recall some basic notions and notation for spline interpolation [193]. The convolution product is $(f * g)(\cdot) = \int_{-\infty}^{+\infty} f(t)g(\cdot - t)dt$ for functions f and g in $L_2(\mathbb{R})$, the Hilbert space of measurable, square-integrable functions from $\mathbb{R}$ to $\mathbb{R}$. The convolution product is also defined for discrete signals $a \in \ell_2$ and $b \in \ell_1$ as, for $k \in \mathbb{Z}$, $(b * a)(k) = \sum_{\ell\in\mathbb{Z}} b(\ell)a(k-\ell)$ where ℓ_2 (resp. ℓ_1) is the space of square summable (resp. summable) sequences.
Polynomial splines are piecewise polynomials that satisfy specific continuity constraints to interpolate or approximate a given sequence $a \in \ell_2$. The generic space of polynomial splines of order n is denoted S_1^n, where the superscript n refers to the degree of the polynomial segments and where the subscript represents the spacing between the knots (i.e., the joining points of the polynomial segments). More precisely, S_1^n is the subset of functions of L_2 that are of class C^{n-1} (i.e., continuous functions with continuous derivatives up to order $n-1$).
B-splines are the atoms for constructing spline representations because they offer the best cost-performance trade-off, a benefit well documented in the literature [213]. One theoretical explanation for this superior performance is that the B-spline of degree $n \in \mathbb{N}$, denoted by $\beta^n : \mathbb{R} \to \mathbb{R}$, is the shortest and smoothest function that allows for the reproduction of polynomials of degree n. The magnification of the centered B-spline of order n is defined by $\beta_\Delta^n = \beta^n(\cdot/\Delta)$ and, most importantly, is itself a centered B-spline of order n. Let us define $b^n \in \mathbb{R}^{\mathbb{Z}}$ as the discrete B-spline of order n, obtained by sampling β^n at integer values, i.e. $b^n(\ell) = \beta^n(\ell)$ for $\ell \in \mathbb{Z}$.

The set of shifted B-splines $\{\beta^n(x-k), k \in \mathbb{Z}\}$ constitutes a basis of S_1^n.
In particular, we have

$$(\forall \tau \in \mathbb{R}) \quad \beta^0(\tau) = \begin{cases} 1 & \text{if } |\tau| < \frac{1}{2} \\ 0 & \text{otherwise,} \end{cases}, \qquad \beta^1(\tau) = \begin{cases} 1-|\tau| & \text{if } |\tau| < 1 \\ 0 & \text{otherwise.} \end{cases} \tag{6.9}$$

The simplest B-spline, of order 0, leads to nearest neighbor interpolation, while order 1 corresponds to linear interpolation. Notably, B-splines functions can be constructed recursively from

$$\beta^n(x) = \beta^{n-1} * \beta^0(x) \tag{6.10}$$

which states that a B-spline of order n can be generated by convolving β^0 $(n+1)$ times with itself. Both the support length and the smoothness of B-splines increase with the order. In the limit, B-splines converge to the Gaussian function. From (6.10), we see that all B-splines are positive, symmetric, and have an integral equal to one. The support and the approximation order of these functions is one more than their degree.
Any signal $s \in S_1^n$ can be expressed as

$$s(x) = \sum_{k \in \mathbb{Z}} c(k) \beta^n(x-k), \tag{6.11}$$

where $c(j)$ is the associated sequence of B-spline coefficients of s such that

$$(\forall z \in \mathbb{Z}) \quad s(k) = b^n * c(k). \tag{6.12}$$

In (6.12), b^n is a finite impulse response (FIR) filter.
Similarly, the B-splines coefficients of s can be obtained from the signal samples as

$$(\forall k \in \mathbb{Z}) \quad c(k) = b^{-n} * s(k). \tag{6.13}$$

In (6.13), b^{-n} is an all-pole infinite impulse response filter (IIR) which can be conveniently implemented using the fast recursive technique in [214].
The spline formalism extends Shannon's sampling theory and provides a unifying view of continuous/discrete signal processing using least-squares approximation. The multiresolution property of splines makes them prime candidates for constructing multiresolution pyramids and wavelets. We now show how splines have been used for image resizing by a non-integer factor.

6.3.2 Continuous-to-discrete (C-D) approach

Let $\Delta > 0$ and let $a(\Delta \cdot)$ be the continuous magnified version of a 1D continuous signal $a : \mathbb{R} \to \mathbb{R}$.
The resampling task we consider here consists of computing from N_J uniformly spaced samples of a with sampling step 1, N_I samples that are therefore uniformly spaced by Δ. A resampling step $\Delta > 1$ thus corresponds to a downsampling (reduction) while a sampling step $\Delta < 1$ is an upsampling (enlargement), the magnification factor being equal to $1/\Delta$. We denote $I = \{1, \ldots, N_I\}$ and $J = \{1, \ldots, N_J\}$ the respective sets of indices. The goal is to compute a reduction/enlargement a_Δ on the same axis as a from the vector $\mathbf{a} \in \mathbb{R}^{N_J}$ of known values $(a(j))_{j \in J}$ of a using B-splines expansions.
In the C-D approach, from the discrete set of data points, we aim to find a discrete set of B-splines coefficients $(c(j))_{j \in J} \in \mathbb{R}^{N_J}$ which parameterizes a under the constraint that the evaluation of a at the sampling points yields the same value as the data themselves.

Then, the C-D approach assumes that it is possible to compute discrete data values of a at any abscissa from its continuous B-splines representation.
Given values $(a(j))_{j\in J}$, we build a representation of a on the space spanned by $\{\beta^n(\cdot - j) \mid j \in J\}$ given by

$$a(x) = \sum_{j\in J} c(j)\beta^n(x - j), \tag{6.14}$$

The magnified signal is then summarized by the set of discrete samples $\mathbf{a_\Delta} = (a(\Delta i))_{i\in I}$ derived by applying the resampling transform on the set of coordinates of the initial data points

$$(\forall i \in I) \qquad a(\Delta i) = \sum_{j\in J} c_\Delta(j)\beta^n(\Delta i - j). \tag{6.15}$$

Altogether C-D resampling is a two-step procedure. The first step is fitting a continuous representation to the data points $(a(j))_{j\in J}$. This results in the inversion of (6.14): coefficients $c(j)$ are found by deconvolving the sampled function $a(j)$ with the factorized filter b^{-n}. The second step computes the arbitrarily located samples of $a(\Delta\cdot)$ according to (6.15).

6.3.3 Continuous-to-continuous (C-C) approach

The C-C approach has been advocated by [215] to minimize information loss during resizing. Given values $(a(j))_{j\in J}$, we assume that our representation of a in (6.14) still holds. Instead of deriving directly signal samples from the continuous representation of a, the C-C approach relies on an intermediate continuous expansion a_Δ which represents the magnified signal on the same scale as a. This representation is provided onto the space spanned by functions $\{\beta^n_\Delta(\cdot - \Delta i) \mid i \in I\}$ given by

$$a_\Delta(x) = \sum_{i\in I} c_\Delta(i)\beta^n_\Delta(x - \Delta i). \tag{6.16}$$

In contrast to the C-D approach, the C-C approach relates the two continuous representations (6.16) and (6.14). The vector of coefficients $\mathbf{c_\Delta} = (c_\Delta(i))_{i\in I} \in \mathbb{R}^{N_I}$ is determined by minimizing the norm $\|a_\Delta - a\|_{L_2(\mathbb{R})}$ and thus satisfies normal equations [144], which are expressed in matrix form as

$$\mathbf{T}\mathbf{c_\Delta} = \mathbf{\Xi}\mathbf{c}, \tag{6.17}$$

where $\mathbf{\Xi} = (\Xi_{i,j})_{(i,j)\in I\times J} \in [0, +\infty[^{N_I\times N_J}$ and $\mathbf{T} = (T_{i,i'})_{(i,i')\in I^2} \in [0, +\infty[^{N_I\times N_I}$ are such that,[1]

$$\Xi_{i,j} = \int_{-\infty}^{+\infty} \beta^n(x - j)\beta^n_\Delta(x - \Delta i)dx = \xi^{n,n}_\Delta(j - \Delta i) \tag{6.18}$$

with, for every $\theta \in \mathbb{R}$,

$$\xi^{n,n}_\Delta(\theta) = (\beta^n * \beta^n_\Delta)(\theta) \tag{6.19}$$

and, for every $(i, i') \in I^2$,

$$T_{i,i'} = \int_{-\infty}^{+\infty} \beta^n_\Delta(x - \Delta i)\beta^n_\Delta(x - \Delta i')dx = \beta^{2n+1}_\Delta((i - i')\Delta) = b^{2n+1}(i - i'). \tag{6.20}$$

$\mathbf{\Xi}$ is a cross-correlation matrix containing the correlations of functions β^n_Δ and β^n according to the relative positions of the samples. In our context of projective geometry,

[1] The functions β^n with $n \in \mathbb{N}$ are even.

these cross-correlations will be interpreted as "footprints". Note that the general scaling property of the convolution of B-splines implies that

$$(\forall \theta \in \mathbb{R}) \quad \xi_{\Delta}^{n,n}(\theta) = \Delta \xi_{\frac{1}{\Delta}}^{n,n}(\theta/\Delta). \tag{6.21}$$

This means that the magnification of step Δ and the inverse magnification of step $1/\Delta$ result in the same footprint, up to a normalization factor.
The matrix $\mathbf{T}$ is a Gram matrix, hence symmetric and semi-definite positive, independent of Δ. It is also Tœplitz so that its inverse can be implemented by means of digital filters. Finally, $\mathbf{a_\Delta} = (a_\Delta(i))_{i \in I}$ is given by

$$\mathbf{a_\Delta} = \mathbf{\Lambda} \mathbf{T}^{-1} \mathbf{\Xi} \mathbf{c}, \tag{6.22}$$

where $\mathbf{\Lambda} = (\Lambda_{i,i'})_{(i,i') \in I^2}$ is such that,

$$(\forall (i,i') \in I^2) \quad \Lambda_{i,i'} = b^n(i - i'). \tag{6.23}$$

The discrete convolution form of (6.22) is

$$(\forall i \in I) \quad a_\Delta(i) = \left(b^n * (b^{2n+1})^{-1} * (\mathbf{\Xi c})\right)(i). \tag{6.24}$$

Remark 6.3.3.1 This routine can be directly applied to image magnification by implementing separable magnifications along each direction resulting in successive 1D processing along the rows and the columns of an image.

Remark 6.3.3.2 In general, the C-C solution results in a higher-order interpolation than the one of C-D. In (6.20), we see that the C-C solution corresponds to a polynomial spline interpolation of degree $2n + 1$ (i.e., twice the order) [2].

6.4 Proposed magnification-driven approach: an extension of C-C

In the following, we extend the C-C approach for dealing with homographies found in C-arm CBCT. First, we allow the signal of known samples to be approximated with B-splines of order m, possibly different from the order n of the output. We present our approach in 1D before extending it to 2D.
We now consider that one line $f(x)$ of the volume and one line of the projector $p(u)$ are related by the 1D homography h such that, for every $x \in \mathcal{X} =]-h_{3,3}/h_{3,1}, +\infty[$,

$$u = h(x) = \frac{h_{1,1}x + h_{1,3}}{h_{3,1}x + h_{3,3}}. \tag{6.25}$$

This defines a bijective mapping from $\mathcal{X}$ to $\mathcal{U} = h(\mathcal{X})$. Backprojection generates f from p and projection generates p from f, as shown in Figure 6.2, according to

$$p(u) = p \circ h(x) = \frac{f(x)}{|h'(x)|}, \tag{6.26}$$

[2] The C-D and C-C solutions are identical asymptotically for the bandlimited case because $\mathrm{sinc} * \mathrm{sinc}(x) = \mathrm{sinc}(x)$

or

$$f(x) = f \circ h^{-1}(u) = |h'(h^{-1}(u))|p(u), \tag{6.27}$$

where h' denotes the derivative of h. These relations ensure the expected conservation of matter density through the integral identity:

$$\int_{\mathcal{U}} p(u)du = \int_{\mathcal{X}} f(x)dx. \tag{6.28}$$

Now N_I (resp. N_J) refers to the number of samples along u (resp. x) and $I = \{1, \ldots, N_I\}$ (resp. $J = \{1, \ldots, N_J\}$) are the associated set of indices. Let $(u_i)_{i\in I}$ be the locations of values $(p_i)_{i\in I}$ of discrete signal $\mathbf{p} = (p(u_i))_{i\in I}$. Let $(x_j)_{j\in J}$ be the locations of the observed values $(f_j)_{j\in J}$ of f, giving rise to the discrete signal $\mathbf{f} = (f(x_j))_{j\in J}$.

6.4.1 Projector

As an extension of (6.21), let us define function $\xi_\Delta^{m,n}$ such that

$$(\forall \theta \in \mathbb{R}) \quad \xi_\Delta^{m,n}(\theta) = \beta^m * \beta_\Delta^n(\theta). \tag{6.29}$$

We first assume that f belongs to the space spanned by $\{\beta^m(\cdot - x_j) \mid j \in J\}$ i.e.,

$$f(x) = \sum_{j\in J} c(j)\beta^m(x - x_j), \tag{6.30}$$

where $\mathbf{c} = (c(j))_{j\in J}$ is the associated set of B-spline coefficients of f.
Unlike the magnification case, the homography of a centered B-spline is not a B-spline in general. The magnification-driven approach consists therefore in approximating the homography of the centered B-spline by its magnification. We thus use the absolute value $|h'(x)|$ of the derivative of h at x, which provides the continuous change in sampling rate from x to u induced by h. Furthermore, we note that $|(h^{-1})'(u)| = 1/|h'(x)|$. For the approximation to be valid, coefficient $h_{3,1}$ must be small enough to make the variation of the magnification factor negligible over the support of the B-splines. We have then $|(h^{-1})'(z)| \simeq 1/|h'(z)| \simeq h_{1,1}/h_{3,3}$. Let $\Delta_i = 1/|h'(h^{-1}(u_i))|$ be the local sampling step in an open neighborhood $\mathcal{V}(u_i)$ of u_i with $i \in I$. This defines a vector of resampling parameters $\mathbf{\Delta} = (\Delta_i)_{i\in I}$. Hence, when $u \in \mathcal{V}(u_i)$, (6.27) yields

$$p(u) = \frac{f(h^{-1}(u))}{|h'(h^{-1}(u))|} \simeq \Delta_i f(h^{-1}(u)). \tag{6.31}$$

Let $p_{\mathbf{\Delta}}$ be an approximation of p on the same axis as f such that, for every $i \in I$,

$$p(u_i) = \Delta_i \, p_{\mathbf{\Delta}}(h^{-1}(u_i)) \tag{6.32}$$

(see Figure 6.4). We assume that $p_{\mathbf{\Delta}}$ can be decomposed onto a family of nonuniform B-splines of order $n \in \mathbb{N}$ as

$$p_{\mathbf{\Delta}}(x) = \sum_{i\in I} s_{\Delta_i}(i) \, \beta_{\Delta_i}^n(x - h^{-1}(u_i)). \tag{6.33}$$

The optimal coefficients $\mathbf{s}_{\mathbf{\Delta}}^* = (s_{\Delta_i}(i))_{i\in I}$ satisfy the normal equations according to

$$\mathbf{G}\mathbf{s}_{\mathbf{\Delta}}^* = \mathbf{F}\mathbf{c}, \tag{6.34}$$

where $\mathbf{F} = (F_{i,j})_{(i,j)\in I\times J} \in [0,+\infty[^{N_I\times N_J}$ and $\mathbf{G} = (G_{i,l})_{(i,l)\in I^2} \in [0,+\infty[^{N_I\times N_I}$ are such that, for every $(i,l) \in I^2$ and $j \in J$,

$$F_{i,j} = \int_{-\infty}^{+\infty} \beta^m(x - x_j)\beta^n_{\Delta_i}(x - h^{-1}(u_i))dx = \xi^{m,n}_{\Delta_i}(x_j - h^{-1}(u_i)) \tag{6.35}$$

and

$$\begin{aligned} G_{i,l} &= \int_{-\infty}^{+\infty} \beta^n_{\Delta_l}(x - h^{-1}(u_l))\beta^n_{\Delta_i}(x - h^{-1}(u_i))dx = \beta^n_{\Delta_l} * \beta^n_{\Delta_i}(h^{-1}(u_l) - h^{-1}(u_i)) \\ &= \Delta_i \xi^{n,n}_{\Delta_l/\Delta_i}\left(\frac{h^{-1}(u_l) - h^{-1}(u_i)}{\Delta_i}\right) = \Delta_l \xi^{n,n}_{\Delta_i/\Delta_l}\left(\frac{h^{-1}(u_i) - h^{-1}(u_l)}{\Delta_l}\right). \end{aligned} \tag{6.36}$$

$\mathbf{F}$ contains the footprints of functions $(\beta^n_{\Delta_i}(\cdot - h^{-1}(u_i)))_{i\in I}$ and (resp. over) $(\beta^m(\cdot - x_j))_{j\in J}$ (see Figure 6.3). Note that the Gram matrix $\mathbf{G}$ is not Tœplitz anymore and that its diagonal elements are

$$(\forall i \in I) \quad G_{i,i} = \Delta_i \xi^{n,n}_1(0) = \Delta_i b^{2n+1}(0). \tag{6.37}$$

Finally, the expression of vector $\mathbf{p}$ is derived from (6.32):

$$\mathbf{p} = \mathrm{Diag}(\boldsymbol{\Delta})\tilde{\boldsymbol{\Lambda}}\mathbf{s}^*_{\boldsymbol{\Delta}} \tag{6.38}$$

where $\mathrm{Diag}(\boldsymbol{\Delta})$ is the diagonal matrix whose diagonal is equal to vector $\boldsymbol{\Delta}$ and $\tilde{\boldsymbol{\Lambda}} = (\tilde{\Lambda}_{i,l})_{(i,l)\in I^2} \in [0,+\infty[^{N_I\times N_I}$ is such that,

$$\tilde{\Lambda}_{i,l} = \beta^n_{\Delta_i}\left(h^{-1}(u_l) - h^{-1}(u_i)\right). \tag{6.39}$$

By combining (6.34) and (6.38), we have

$$\mathbf{p} = \mathrm{Diag}(\boldsymbol{\Delta})\tilde{\boldsymbol{\Lambda}}\mathbf{G}^{-1}\mathbf{F}\mathbf{c}, \tag{6.40}$$

where $\mathbf{G}^{-1}$ denotes the pseudo-inverse of $\mathbf{G}$. Note that the above equality holds exactly provided that model (6.30)-(6.33) is perfectly satisfied, which is obviously an approximation in practice.

Remark 6.4.1.1 For every $(n,m) \in \mathbb{N}^2$, the support of β^n_Δ is $]-(n+1)\Delta/2,(n+1)\Delta/2[$ and, for every $\Delta > 0$ the support of function $\xi^{m,n}_\Delta$ is thus equal to $]-(m+1+(n+1)\Delta)/2,(m+1+(n+1)\Delta)/2[$. This implies that most elements of matrices $\mathbf{G}$, $\mathbf{F}$, $\tilde{\boldsymbol{\Lambda}}$, $\mathbf{G}'$, $\mathbf{F}'$, and $\tilde{\boldsymbol{\Lambda}}'$ are zero, giving them a band structure. For example, according to (6.36), for every $(i,l) \in I^2$, if

$$|h^{-1}(u_l) - h^{-1}(u_i)| \geqslant (n+1)\frac{\Delta_i + \Delta_l}{2}, \tag{6.41}$$

then $G_{i,l} = 0$ and, if

$$|h^{-1}(u_l) - h^{-1}(u_i)| \geqslant (n+1)\frac{\Delta_i}{2}, \tag{6.42}$$

then $\tilde{\Lambda}_{i,l} = 0$.

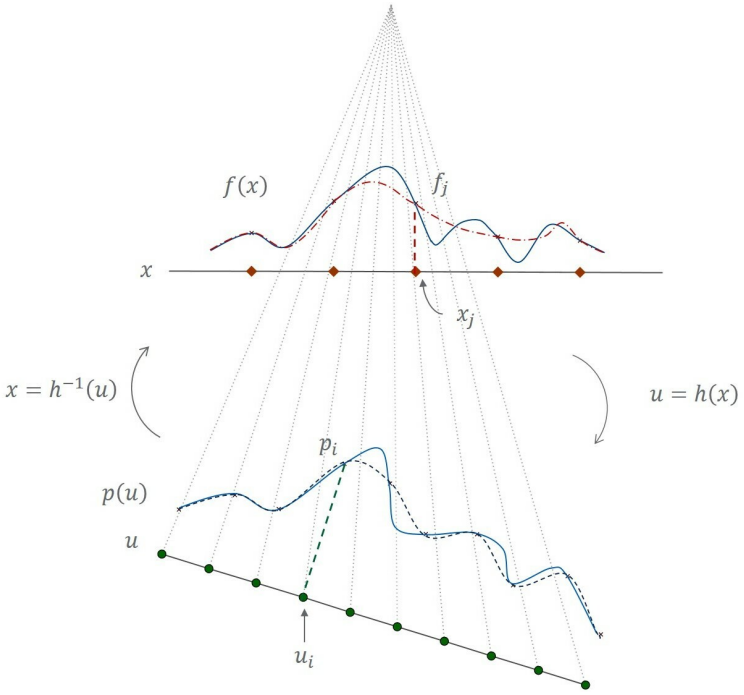


Figure 6.2: Example of a signal (solid line) and its B-splines approximations in the volume and in the projections (dashed line).

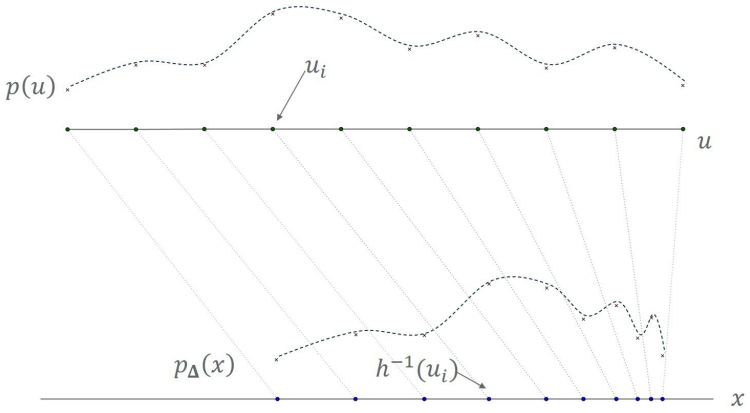


Figure 6.3: Least-square resampling of $f(x)$ on a basis of non-uniform B-splines centered on $(h^{-1}(u_i))_{i \in I}$.

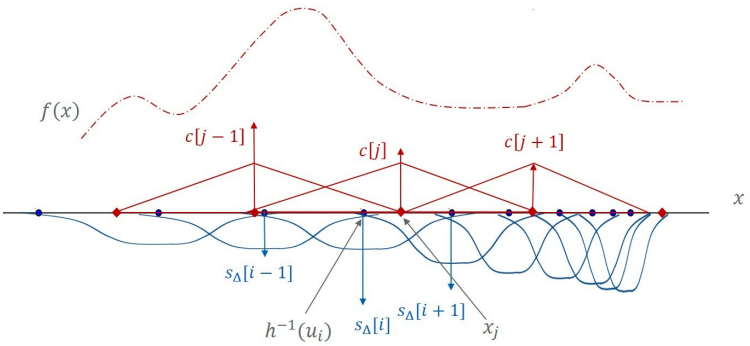


Figure 6.4: Construction of $p_{\Delta}(x)$ from $p(u)$ on the same axis as $f(x)$.

6.4.2 Backprojector

Let us now see how to retrieve $\mathbf{f}$ from $\mathbf{p}$ in the backprojection stage. Here, we assume that the continuous projection p can be decomposed as

$$p(u) = \sum_{i \in I} c'(i) \beta^n(u - u_i), \tag{6.43}$$

where $\mathbf{c'} = (c'(i))_{i \in I}$ is the associated set of B-spline coefficients. A new vector of resampling parameters $\boldsymbol{\Delta'} = (\Delta'_j)_{j \in J}$ is defined such that $\Delta'_j = |h'(x_j)|$, the sampling step in $\mathcal{V}(x_j)$, an open neighborhood of x_j. Thus,

$$f(x_j) = |h'(x_j)| p(h(x_j)) = \Delta'_j \, p(h(x_j)). \tag{6.44}$$

Let $f_{\boldsymbol{\Delta'}}$ be an approximation of f on the same axis as p and such that, for every $j \in J$,

$$f(x_j) = \Delta'_j f_{\boldsymbol{\Delta'}}(h(x_j)). \tag{6.45}$$

We now assume that $f_{\boldsymbol{\Delta'}}$ is the projection of p onto the vector space generated by $\{\beta^m_{\Delta'_j}(\cdot - h(x_j)) \mid j \in J\}$, which leads to the following relation:

$$\mathbf{f} = \mathrm{Diag}(\boldsymbol{\Delta'}) \tilde{\boldsymbol{\Lambda}}' (\mathbf{G'})^{-1} \mathbf{F'} \mathbf{c'}, \tag{6.46}$$

where matrices $\mathbf{F'} \in [0, +\infty[^{N_J \times N_I}$, $\mathbf{G'} \in [0, +\infty[^{N_J \times N_J}$, and $\tilde{\boldsymbol{\Lambda}}' \in [0, +\infty[^{N_J \times N_J}$ are such that, for every $(j, l) \in J^2$ and $i \in I$,

$$F'_{j,i} = \xi^{n,m}_{\Delta'_j}(u_i - h(x_j)), \quad G'_{j,l} \;= \Delta'_j \xi^{m,m}_{\Delta'_l/\Delta'_j}\left(\frac{h(x_l) - h(x_j)}{\Delta'_j}\right), \quad \tilde{\Lambda}'_{j,l} = \beta^m_{\Delta'_l}\left(h(x_l) - h(x_j)\right). \tag{6.47}$$

6.4.3 Choice of the magnification factors

When comparing (6.40) and (6.46), we note that the main modeling difference lies in the set of magnification factors. Given the scaling property

$$(\forall \theta \in \mathbb{R}) \quad (1/\Delta) \xi^{m,n}_{\Delta}(\theta) = \xi^{n,m}_{1/\Delta}(\theta/\Delta), \tag{6.48}$$

we remark that

$$F'_{i,j} = \xi^{n,m}_{\Delta'_j}(h(x_j) - u_i) = \Delta'_j \xi^{m,n}_{1/\Delta'_j}((h(x_j) - u_i)/\Delta'_j). \tag{6.49}$$

If $u_i \simeq h(x_j)$ with $j \in J$ (with $N_J \leqslant N_I$), then

$$\Delta'_j \simeq \frac{1}{\Delta_i}. \tag{6.50}$$

This means that one could use the sampling steps $\boldsymbol{\Delta}^{-1} = (1/\Delta_j)_{j \in J}$ instead of $\boldsymbol{\Delta'}$ and thus, according to (6.48), (6.35), and (6.47),

$$\begin{aligned} F'_{j,i} = \frac{1}{\Delta_i} \xi^{m,n}_{\Delta_i}\big(\Delta_i (h(x_j) - u_i)\big) &\simeq \frac{1}{\Delta_i} \xi^{m,n}_{\Delta_i}\big(\Delta_i \, h'(h^{-1}(u_i)) \, (x_j - h^{-1}(u_i)\big) \\ &\simeq \frac{1}{\Delta_i} \xi^{m,n}_{\Delta_i}(x_j - h^{-1}(u_i)) = \frac{1}{\Delta_i} F_{i,j}. \end{aligned} \tag{6.51}$$

In this case, the projection and backprojection steps would share the same interpolation model. The sampling steps may be close but are different since they cannot be defined at the same locations (i.e., there exists no bijection between the set of locations $(x_j)_{j \in J}$ and the set of locations $(u_i)_{i \in I}$).

Remark 6.4.3.1 For the projection step, the sampling steps $(\Delta_i)_{i\in I}$ are the derivative of h^{-1} at the sampling points $(u_i)_{i\in I}$. When $n = m = 0$, piecewise constant approximations are performed for each signal which matches the description made by geometric models that compute the footprints between pixels and detector bins based on the locations of their edges. It is straightforward to compute sampling steps δ_i from these edge locations. We define the set of segments of center $h^{-1}(u_i)$ and width δ_i by setting

$$\delta_1 = \delta_2 = h^{-1}(u_2) - h^{-1}(u_1) \tag{6.52}$$

and, for every $i \in \{2, \dots, N_I - 1\}$,

$$h^{-1}(u_{i+1}) - h^{-1}(u_i) = \frac{\delta_{i+1} + \delta_i}{2}. \tag{6.53}$$

In this way, given that $(h^{-1}(u_{i+1}) - h^{-1}(u_i))_{1\leqslant i\leqslant N_I-1}$ is a sequence of increasing steps, the interval $[h^{-1}(u_1), h^{-1}(u_{N_I})]$ is partitioned in intervals $[h^{-1}(u_1), h^{-1}(u_1) + \delta_1/2]$, $(]h^{-1}(u_i) - \delta_i/2, h^{-1}(u_i) + \delta_i/2])_{2\leqslant i\leqslant N_I-1}$, and $]h^{-1}(u_{N_I}) - \delta_I/2, h^{-1}(u_{N_I})]$. By construction when $l > i$

$$\frac{h^{-1}(u_l) - h^{-1}(u_i)}{\delta_l} = \frac{1 + \delta_i/\delta_l}{2} + \frac{\sum_{k=i+1}^{l-1} \delta_k}{\delta_l} \geqslant \frac{1}{2}(1 + \frac{\delta_i}{\delta_l}) \geqslant \frac{1}{2}. \tag{6.54}$$

When δ_i is substituted for Δ_i in (6.31), it follows from Remark 6.4.1.1, that $\mathbf{G} = \mathrm{Diag}(\boldsymbol{\Delta})$, $\tilde{\boldsymbol{\Lambda}} = b^n(0)\,\mathrm{Id}_{N_I}$, where Id_{N_I} denotes the identity matrix of size $N_I \times N_{I_1}$ and (6.40) leads to

$$\mathbf{p} = b^n(0)\mathbf{F}\mathbf{c}. \tag{6.55}$$

This provides an alternative way of setting the magnification factors.

6.4.4 Approximation for fast implementation

For projection, in order to reduce the computation burden related to the inversion of matrix $\mathbf{G}$, we propose to approximate this matrix by a surrogate matrix $\tilde{\mathbf{G}} \in \mathbb{R}^{N_I \times N_I}$ in (6.40).
For every $(i, l) \in I^2$ for which (6.41) is not satisfied, we will make the assumption that

$$h^{-1}(u_l) \simeq h^{-1}(u_i) + \frac{1}{h'(h^{-1}(u_i))}(u_l - u_i) = h^{-1}(u_i) \pm \Delta_i(l - i). \tag{6.56}$$

Based on these approximations, (6.36), and Remark 6.4.1.1, we will define $\tilde{\mathbf{G}} = (\tilde{G}_{i,l})_{(i,l)\in I^2}$ as follows:

- if $|u_l - u_i| \leqslant n$, then

$$\tilde{G}_{i,l} = \sqrt{\Delta_i \Delta_l}\xi_1^{n,n}(l - i) = \sqrt{\Delta_i \Delta_l} b^{2n+1}(l - i), \tag{6.57}$$

- otherwise $\tilde{G}_{i,l} = 0$.

In particular, $\tilde{G}_{i,i} = G_{i,i}$ and we can write $\tilde{\mathbf{G}} = \mathrm{Diag}(\boldsymbol{\Delta})^{1/2}\mathbf{T}_n\,\mathrm{Diag}(\boldsymbol{\Delta})^{1/2}$ where $\mathbf{T}_n$ is the Tœplitz matrix previously encountered for the magnification case with B-splines of order n. The resulting approximate vector of B-spline coefficients then reads as

$$\tilde{\mathbf{s}}^*_{\boldsymbol{\Delta}} = \mathrm{Diag}(\boldsymbol{\Delta})^{-1/2}\mathbf{T}_n^{-1}\,\mathrm{Diag}(\boldsymbol{\Delta})^{-1/2}\mathbf{F}\mathbf{c}. \tag{6.58}$$

Remark 6.4.4.1 (i) Since multiplication by a Tœplitz matrix is equivalent to discrete convolution with suitable boundary conditions, the components $(\tilde{s}^*_{\mathbf{\Delta},i})_{i\in I}$ of vector $\tilde{s}^*_{\mathbf{\Delta}}$ are given by

$$(\forall i \in I) \quad \tilde{s}^*_{\mathbf{\Delta},i} = \frac{1}{\sqrt{\Delta_i}} \sum_{l=1}^{N_I} \frac{1}{\sqrt{\Delta_l}} (\mathbf{Fc})_l \, (b^{2n+1})^{-1}(i-l), \tag{6.59}$$

where

$$(\forall l \in I) \quad (\mathbf{Fc})_l = \sum_{j=1}^{N_J} c(j) \, \xi^{m,n}_{\Delta_i}(h^{-1}(u_l) - x_j). \tag{6.60}$$

(ii) Applying (6.56) for every couple $(i,l) \in I^2$ that does not satisfy (6.41) and using (6.39), leads to a rougher approximation where $\tilde{\mathbf{\Lambda}}$ is replaced by $\mathbf{\Lambda}$, which was also introduced in the magnification case with B-splines of order n. Here again, $\Lambda_{i,i} = \tilde{\Lambda}_{i,i}$.

Then, the components of vector $\mathbf{p}$ in (6.40) are approximated by the following discrete convolution

$$(\forall i \in I) \quad \tilde{p}_i = \Delta_i \, (b^n * \tilde{s}^*_{\mathbf{\Delta}})(i). \tag{6.61}$$

The same simplifications apply for backprojection: we define surrogate matrix $\tilde{\mathbf{G}}' \in \mathbb{R}^{N_J \times N_J}$ such that, for $(j,l) \in J^2$ such that $|x_l - x_j| \leqslant m$,

$$\tilde{G'}_{j,l} = \sqrt{\Delta'_j \Delta'_l} \, b^{2m+1}(l-j). \tag{6.62}$$

As above, vector $\mathbf{f}$ in (6.46) can be approximated by

$$\tilde{\mathbf{f}} = \mathrm{Diag}(\mathbf{\Delta'}) \tilde{\mathbf{\Lambda}}' \, \mathrm{Diag}(\mathbf{\Delta'})^{-1/2} \mathbf{T}_m^{-1} \, \mathrm{Diag}(\mathbf{\Delta'})^{-1/2} \mathbf{F}' \mathbf{c}', \tag{6.63}$$

where the inversion performed by $\mathbf{T}_m^{-1}$ can be implemented by filtering with $(b^{2m+1})^{-1}$.

In our context, low order splines corresponding to $(n,m) \in \{0,1\}^2$ and $s = m+n+1 \in \{1,2,3\}$ are used. Order 0 indeed provides a good model of the sampling process performed at the physical detector level, while order 1 corresponds to the most common linear interpolation used in signal/image processing. In the implementation, the three main practical aspects are the explicit evaluation of the sampling kernel $\xi^{m,n}_{\Delta}$, and the multiplication by the inverse of the Gram matrix (i.e., $\tilde{\mathbf{G}}$ for projection). No prefiltering is needed to compute the B-spline coefficients which are equal to the pixel values.
We derive from (6.29) explicit formulas for correlation functions $\xi^{0,0}_{\Delta}$, $\xi^{1,0}_{\Delta}$ ($\xi^{0,1}_{\Delta}$ being deduced by using (6.48)), and $\xi^{1,1}_{\Delta}$:

- Case 1

$$(\forall \theta \in \mathbb{R}) \quad \xi^{0,0}_{\Delta}(\theta) = \begin{cases} \min(1,\Delta) & \text{if} \quad |\theta| < a_1 \\ a_2 - |\theta| & \text{if} \quad a_1 \leqslant |\theta| < a_2 \\ 0 & \text{if} \quad |\theta| \geqslant a_2 \end{cases} \tag{6.64}$$

with $a_1 = \frac{|\Delta - 1|}{2}$, and $a_2 = \frac{\Delta+1}{2}$.

- Case 2

$$(\forall \theta \in \mathbb{R}) \quad \xi^{1,0}_{\Delta}(\theta) = \begin{cases} c_{k,0} + c_{k,1}|\theta| + c_{k,2}|\theta|^2 & \text{for } |\theta| \in [a_{k-1}, a_k[\\ & \text{and } k \in \{1,2\} \\ 0 & \text{otherwise} \end{cases} \tag{6.65}$$

with $a_0 = 0$, $a_1 = |\frac{\Delta}{2} - 1|$, $a_2 = \frac{\Delta}{2} + 1$, and expressions for $(c_{k,0}, c_{k,1}, c_{k,2})$ given in Table 6.1. Calculation details can be found in the Appendix.

Interval	$c_{k,0}$	$c_{k,1}$	$c_{k,2}$
$\lvert\theta\rvert < a_1$			
if $\Delta \leqslant 2$ and $\lvert\theta\rvert < \Delta/2$	$\Delta - \Delta^2/4$	0	-1
if $\Delta \leqslant 2$ and $\lvert\theta\rvert \geqslant \Delta/2$	Δ	$-\Delta$	0
if $\Delta > 2$	1	0	0
$a_1 \leqslant \lvert\theta\rvert < a_2$			
if $\lvert\theta\rvert \geqslant \Delta/2$	$(\Delta^2 + 4\Delta + 4)/8$	$-1 - \Delta/2$	$1/2$
if $\lvert\theta\rvert < \Delta/2$	$(-\Delta^2 + 4\Delta + 4)/8$	$-1 + \Delta/2$	$-1/2$

Table 6.1: B-spline correlation function parameters for case 2

- Case 3 The expression of $\xi_\Delta^{1,1}(\theta)$ is given in Table II. in [215].

Figure 6.5 displays the 1D kernels. Their width increases with both Δ and the approximation order. By design, our method makes use of all sampling points therefore the computation cost is proportional to the total number of samples $N_I + N_J$. On the contrary, destination-driven interpolation skips samples when $\Delta > 1$. Thanks to the property (6.48) of $\xi_\Delta^{m,n}$, the footprints can always be computed using $\Delta < 1$ so that the computation complexity only depends on the order of the splines. The weighted sums will require between two samples for $s = 1$ and four samples when $s = 3$. The computation of $\xi_\Delta^{1,1}$ is complex with several tests to handle. Efficiency relies on using pre-computed look-up tables, with a trade-off between the sizes of the tables and the desired numerical precision.
Finally, when numerical simplifications presented in Remark 6.4.4.1 are implemented, multiplication by $\mathbf{\Lambda}$ and by $\tilde{\mathbf{G}}^{-1}$ reduces to applying the identity except when $n = 1$ where $\tilde{\mathbf{G}}^{-1}$ is the direct cubic filter $(b^3)^{-1}$. Its complexity is proportional to the number of output samples N_J, hence it is faster at downsampling. This filter is classically applied according to [216] but one can resort to more efficient implementations, for instance on GPU [33] or using FIR filters [220].

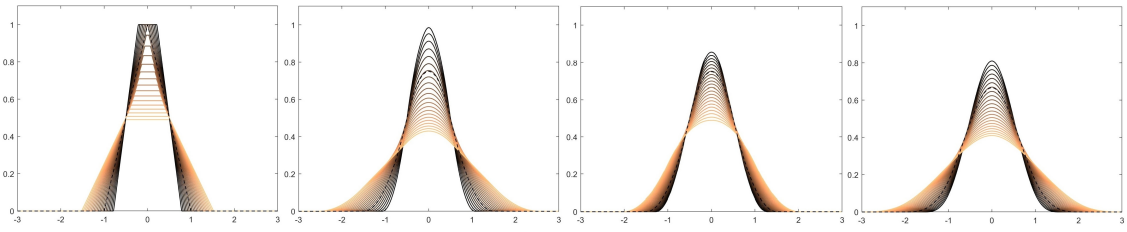


Figure 6.5: Normalized spline correlation kernels when $\Delta \in [0.5, 2]$. From left to right: $\frac{1}{\Delta}\xi_\Delta^{0,0}$, $\frac{1}{\Delta}\xi_\Delta^{0,1}$, $\frac{1}{\Delta}\xi_\Delta^{1,0}$, $\frac{1}{\Delta}\xi_\Delta^{1,1}$. For $\Delta = 1$, the kernels reduce to B-splines of order 2 (second and third), B-splines of order 1 (first) and B-splines of order 3 (fourth) plotted in dashed lines.

6.4.5 Resampling with a 2D representation

We recall that the resampling associated with $\mathbf{H}_{y_0}$ ((6.6)) is a combination of a 1D homography with a 1D magnification in z when x is fixed. Thus the presented optimal resampling for 1D magnifications and homographies is sufficient to perform this operation. We now investigate the potential of a 2D approach to the problem. We require the following conservation of 2D integrals:

$$\int_{\mathcal{U}\times\mathcal{V}} p(u,v)dudv = \int_{\mathcal{X}\times\mathcal{Z}} f(x,z)dxdz, \tag{6.66}$$

where $\mathcal{U}\times\mathcal{V}$ and $\mathcal{X}\times\mathcal{Z}$ are suitable domains of integration. It follows from (6.6) that

$$(\forall (u,v)\in\mathcal{U}\times\mathcal{V}) \quad p(u,v) = f\big(h_1^{-1}(u), h_2^{-1}(u,v)\big)\,|\det \mathbf{J}_{\mathbf{H}_{y_0}^{-1}}(u,v)|, \tag{6.67}$$

and the Jacobian $\mathbf{J}_{\mathbf{H}_{y_0}^{-1}}(u,v)$ is given by

$$\mathbf{J}_{\mathbf{H}_{y_0}^{-1}}(u,v) = \begin{pmatrix} \frac{\partial h_1^{-1}(u)}{\partial u} & 0 \\ \frac{\partial h_2^{-1}(u,v)}{\partial u} & \frac{\partial h_2^{-1}(u,v)}{\partial v} \end{pmatrix}, \tag{6.68}$$

so that $|\det \mathbf{J}_{\mathbf{H}_{y_0}^{-1}}(u,v)| = |\frac{\partial h_1^{-1}(u)}{\partial u}\frac{\partial h_2^{-1}(u,v)}{\partial v}|$.
We now assume that p and f can be decomposed in 2D. For the projection task, f is decomposed as follows:

$$f(x,z) = \sum_{j_2\in J_2}\sum_{j_1\in J_1} c(j_1,j_2)\,\beta^m(x-x_{j_1})\beta^m(z-z_{j_2}). \tag{6.69}$$

The two-dimensional representation $p_{\mathbf{\Delta}}$ of the resampling of p is defined by

$$p_{\mathbf{\Delta}}(x,z) = \sum_{i_1\in I_1}\sum_{i_2\in I_2} s_{\mathbf{\Delta}}(i_1,i_2)\,\beta^n_{\Delta_{1,i_1}}(x-h_1^{-1}(u_{i_1}))\beta^n_{\Delta_{2,i_1}}(z-h_2^{-1}(u_{i_1},v_{i_2})) \tag{6.70}$$

where $\mathbf{\Delta} = (\Delta_{1,i_1}\Delta_{2,i_1})_{i_1\in I_1}$ is the vector whose components are the products of the diagonal elements of the Jacobian matrix which describes the continuous change of sample rate in x (resp. z) along u (resp. v), i.e. $\Delta_{1,i_1} = |h_1^{-1\prime}(u_{i_1})|$ and $\Delta_{2,i_1} = \frac{\partial h_2^{-1}(u_{i_1},v_{i_2})}{\partial v}$. We thus use the same values that would appear with successive 1D processing, while verifying

$$p(u_{i_1},v_{i_2}) = |\det \mathbf{J}_{\mathbf{H}_{y_0}^{-1}}(u_{i_1},v_{i_2})| p_{\mathbf{\Delta}}(h_1^{-1}(u_{i_1}), h_2^{-1}(u_{i_1},v_{i_2})). \tag{6.71}$$

Let us define the vector $\mathbf{c} = (c(j_1,j_2))_{j_1\in J_1, j_2\in J_2}$ (resp. $\mathbf{s}_{\mathbf{\Delta}} = (s_{\mathbf{\Delta}}(i_1,i_2))_{i_1\in I_1, i_2\in I_2}$) whose components have been indexed according to $(j_1-1)N_{J_2}+j_2$ (resp. $(i_1-1)N_{I_1}+i_2$). The normal equations in 2D are expressed in matrix form as

$$\mathbf{G}\mathbf{s}_{\mathbf{\Delta}} = \mathbf{F}\mathbf{c}, \tag{6.72}$$

where $\mathbf{F}\in[0,+\infty[^{N_{I_1}N_{I_2}\times N_{J_1}N_{J_2}}$ and $\mathbf{G}\in[0,+\infty[^{N_{I_1}N_{I_2}\times N_{I_1}N_{I_2}}$ are such that, for every $i=(i_1-1)N_{I_2}+i_2$, $j=(j_1-1)N_{J_2}+j_2$ and $l=(l_1-1)N_{I_2}+l_2$ with $(i_1,l_1)\in I_1^2$, $(i_2,l_2)\in I_2^2$, $j_1\in J_1$, and $j_2\in J_2$,

$$\begin{aligned} F_{i,j} &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \beta^m(x-x_{j_1})\beta^n_{\Delta_{1,i_1}}(x-h_1^{-1}(u_{i_1}))\beta^m(z-z_{j_2})\beta^n_{\Delta_{2,i_1}}(z-h_2^{-1}(u_{i_1},v_{i_2}))dxdz \\ &= \xi^{m,n}_{\Delta_{1,i_1}}(x_{j_1}-h_1^{-1}(u_{i_1}))\xi^{m,n}_{\Delta_{2,i_1}}(z_{j_2}-h_2^{-1}(u_{i_1},v_{i_2})) \end{aligned} \tag{6.73}$$

and

$$G_{j,l} = \left(\beta^n_{\Delta_{2,l_1}} * \beta^n_{\Delta_{2,i_1}}(h_2(u_{i_1}, v_{i_2}) - h_2(u_{l_1}, v_{l_2}))\right)\left(\beta^n_{\Delta_{1,l_1}} * \beta^n_{\Delta_{1,i_1}}(h_1^{-1}(u_{i_1}) - h_1^{-1}(u_{l_1}))\right)$$
$$= \Delta_{2,i_1}\Delta_{1,i_1}\xi^{n,n}_{\Delta_{2,l_1}/\Delta_{2,i_1}}\left(\frac{h_2^{-1}(u_{i_1}, v_{i_2}) - h_2^{-1}(u_{l_1}, v_{l_2})}{\Delta_{2,j_1}}\right)\xi^{n,n}_{\Delta_{1,l_1}/\Delta_{1,i_1}}\left(\frac{h_1^{-1}(u_{i_1}) - h_1^{-1}(u_{l_1})}{\Delta_{1,i_1}}\right). \tag{6.74}$$

Since h_1^{-1} does not depend on v, separability of $\mathbf{F}$ is achieved as

$$\mathbf{F} = \begin{bmatrix} (F_1)_{1,1}\mathbf{F}_2^1 & \dots & (F_1)_{N_{I_1},1}\mathbf{F}_2^{N_{I_1}} \\ \vdots & \dots & \vdots \\ (F_1)_{1,N_{J_1}}\mathbf{F}_2^1 & \dots & (F_1)_{N_{I_1},N_{J_1}}\mathbf{F}_2^{N_{I_1}} \end{bmatrix} = \begin{bmatrix} (\mathbf{F}_1)_{1,*}\otimes\mathbf{F}_2^1 & \dots & (\mathbf{F}_1)_{N_{I_1},*}\otimes\mathbf{F}_2^{N_{I_1}} \end{bmatrix}, \tag{6.75}$$

where $\otimes$ denotes the Kronecker product. Matrix $\mathbf{F}_1 \in [0,+\infty[^{N_{I_1}\times N_{J_1}}$ is such that, for every $i_1 \in I_1$ and $j_1 \in J_1$,

$$(F_1)_{i_1,j_1} = \xi^{m,n}_{\Delta_{1,i_1}}(x_{j_1} - h_1^{-1}(u_{i_1})). \tag{6.76}$$

For every $i_1 \in I_1$, $(\mathbf{F}_1)_{i_1,*}$ denotes the i_1-th row of $\mathbf{F}_1$ and matrix $\mathbf{F}_2^{i_1} \in [0,+\infty[^{N_{I_2}\times N_{J_2}}$ is such that, for every $i_2 \in I_2$ and $j_2 \in J_2$,

$$(F_2{}^{i_1})_{i_2,j_2} = \xi^{m,n}_{\Delta_{2,i_1}}(z_{j_2} - h_2^{-1}(u_{i_1}, v_{i_2})). \tag{6.77}$$

Then, vector $\mathbf{p} = (p(u_{i_1}, v_{i_2}))_{i_1\in I_1, i_2\in I_2} \in \mathbb{R}^{N_{I_1}N_{I_2}}$ (whose components are indexed according to $(i_1-1)N_{I_2}+i_2$) is expressed as

$$\mathbf{p} = (\mathrm{Diag}(\boldsymbol{\Delta})\otimes \mathrm{Id}_{N_{I_2}})\tilde{\boldsymbol{\Lambda}}\mathbf{G}^{-1}\mathbf{F}\mathbf{c}, \tag{6.78}$$

where matrix $\tilde{\boldsymbol{\Lambda}} \in [0,+\infty[^{N_{I_1}N_{I_2}\times N_{I_1}N_{I_2}}$ is such that, for every $i = (i_1-1)N_{I_2}+i_2$ and $l = (l_1-1)N_{I_2}+l_2$, with $(i_1,l_1) \in I_1^2$, $(i_2,l_2) \in I_2^2$,

$$\tilde{\Lambda}_{i,l} = \beta^n_{\Delta_{1,l_1}}\left(h_1^{-1}(u_{i_1}) - h_1^{-1}(u_{l_1})\right)\beta^n_{\Delta_{2,l_1}}\left(h_2^{-1}(u_{i_1}, v_{i_2}) - h_2^{-1}(u_{l_1}, v_{l_2})\right). \tag{6.79}$$

Since we use the same low order B-splines, for every $(i_1,l_1) \in I_1^2$ for which

$$|h_1^{-1}(u_{i_1}) - h_1^{-1}(u_{l_1})| \leqslant (n+1)\frac{\Delta_{1,i_1}+\Delta_{1,l_1}}{2}, \tag{6.80}$$

we can again assume that u_{i_1} and u_{l_1} are close enough so that $\frac{\partial h_2^{-1}}{\partial v}(u_{i_1}, v_{i_2}) \simeq \frac{\partial h_2^{-1}}{\partial v}(u_{l_1}, v_{l_2})$, leading to

$$\Delta_{2,i_1}\xi^{n,n}_{\Delta_{2,l_1}/\Delta_{2,i_1}}\left(\frac{h_2^{-1}(u_{i_1}, v_{i_2}) - h_2^{-1}(u_{l_1}, v_{l_2})}{\Delta_{2,i_1}}\right) \simeq \sqrt{\Delta_{2,i_1}\Delta_{2,l_1}}\beta^{2n+1}(v_{i_2} - v_{l_2}). \tag{6.81}$$

In matrix form, this translates to the following approximation:

$$\mathbf{G} \simeq \mathbf{G}_1 \otimes \tilde{\mathbf{G}}_2, \tag{6.82}$$

where the elements of $\mathbf{G}_1 \in [0,+\infty[^{N_{I_1}\times N_{I_1}}$ and $\tilde{\mathbf{G}}_2 \in [0,+\infty[^{N_{I_2}\times N_{I_2}}$ are, for every $(i_1,l_1) \in I_1^2$,

$$(G_1)_{i_1,l_1} = \Delta_{1,i_1}\sqrt{\Delta_{2,i_1}\Delta_{2,l_1}}\xi^{n,n}_{\Delta_{1,l_1}/\Delta_{1,i_1}}\left(\frac{h_1^{-1}(u_{i_1}) - h_1^{-1}(u_{l_1})}{\Delta_{1,i_1}}\right) \tag{6.83}$$

and, for every $(i_2, l_2) \in I_2^2$,

$$(\tilde{G}_2)_{i_2,l_2} = \beta^{2n+1}\left(v_{i_2} - v_{l_2}\right). \tag{6.84}$$

It can be noticed that both $\mathbf{G}_1$ and $\tilde{\mathbf{G}}_2$ are symmetric matrices. Likewise $\tilde{\mathbf{\Lambda}}$ can be approximated as

$$\tilde{\mathbf{\Lambda}} \simeq \tilde{\mathbf{\Lambda}}_1 \otimes \mathbf{\Lambda}_2, \tag{6.85}$$

where the elements of $\tilde{\mathbf{\Lambda}}_1 \in [0, +\infty[^{N_{I_1} \times N_{I_1}}$, $\mathbf{\Lambda}_2 \in [0, +\infty[^{N_{I_2} \times N_{I_2}}$ are, for every $(i_1, l_1) \in I_1^2$ and $(i_2, l_2) \in I_2^2$,

$$(\tilde{\Lambda}_1)_{i_1,l_1} = \beta^n_{\Delta_{1,l_1}}\left(h_1^{-1}(u_{i_1}) - h_1^{-1}(u_{l_1})\right), \quad (\Lambda_2)_{i_2,l_2} = \beta_1^n(v_{i_2} - v_{l_2}). \tag{6.86}$$

Finally, $\mathbf{p}$ can be derived as

$$\begin{aligned}\mathbf{p} &\simeq (\mathrm{Diag}(\mathbf{\Delta}) \otimes \mathrm{Id}_{N_{I_2}})(\tilde{\mathbf{\Lambda}}_1 \otimes \mathbf{\Lambda}_2)(\mathbf{G}_1^{-1} \otimes \tilde{\mathbf{G}}_2^{-1})\mathbf{F}\mathbf{c} \\ &= \left((\mathrm{Diag}(\mathbf{\Delta})\tilde{\mathbf{\Lambda}}_1\mathbf{G}_1^{-1}) \otimes (\mathbf{\Lambda}_2\tilde{\mathbf{G}}_2^{-1})\right)\mathbf{F}\mathbf{c}.\end{aligned} \tag{6.87}$$

As long as the magnification of the B-splines provides a good enough approximation of the change of sampling rates induced by the homography, the 2D solution is separable into 1D computations. Note that (6.87) shares the same footprint as would be obtained by applying our 1D resampling approach separately on each row and column.
The backprojection task uses the reverse geometric transforms

$$\begin{cases} u = h_1(x) \\ v = h_2(x, z). \end{cases} \tag{6.88}$$

The magnification factors are then chosen equal to $\Delta'_{1,j_1} = |h'_1(x_{j_1})|$ and $\Delta'_{2,j_1} = \frac{\partial h_2(x_{j_1}, z_{j_2})}{\partial z} = h_{22}/s(x_{j_1})$ where $j_1 \in J_1$ and $j_2 \in J_2$. Hereinafter, we will denote $\mathbf{H}_s$ the discretized homographic transform involved in backprojection implemented as (6.87) and $\tilde{\mathbf{H}}_s^{-1}$ the discretized homographic transform involved in projection for $s = m+n+1$.

Remark 6.4.5.1 Up to now, the cone-beam geometry has been assumed to have axis v aligned with axis z. When a 2D rotation within the detector plane can make the axes parallel again, it can be computed within the same framework of centered B-splines and separable 1D processing [217] and can be merged with the steps of rebinning or rectification.
Otherwise, in the practical case of the vibrations of a C-arm system, the rotations that break the parallelism at each angle are small. They cannot be ignored without degrading the resolution, but they can be neglected in the definition of the set of magnifications. We therefore now assume that the null elements, $h_{3,2}$ and $h_{1,2}$ of $\mathbf{H}_{y_0}$ in (6.5) are replaced by small nonzero values. In this case $\frac{\partial h_1^{-1}}{\partial v}$ no longer vanishes. We use the diagonal elements of the Jacobian, considered a sufficient description of the local magnifications while keeping the correct projection matrix to compute the sampling point's locations. This yields magnification factors $(\Delta_{1,i_1,i_2}, \Delta_{2,i_1,i_2})_{i_1 \in I_1, i_2 \in I_2}$. In this case, the elements of $\mathbf{F}$ are

$$F_{i,j} = \xi^{m,n}_{\Delta_{1,i_1,i_2}}(x_{j_1} - h_1^{-1}(u_{i_1}, v_{i_2}))\xi^{m,n}_{\Delta_{2,i_1,i_2}}(z_{j_2} - h_2^{-1}(u_{i_1}, v_{i_2})) \tag{6.89}$$

and $(\mathrm{Diag}(\mathbf{\Delta}) \otimes \mathrm{Id}_{N_{I_2}})$ has to be replaced by the diagonal matrix whose i-th diagonal elements for $i = (i_1 - 1)N_{I_2} + i_2$ with $i_1 \in I_1$ and $i_2 \in I_2$ is equal to $|\det \mathbf{J}_{\mathbf{H}_{y_0}^{-1}}(u_{i_1}, v_{i_2})|$. Note that the separability of $\mathbf{F}$ and the scaling diagonal matrix no longer hold. We can neglect $h_{3,2}$ for matrices $\mathbf{G}$ and $\tilde{\mathbf{\Lambda}}$ to resort to the same surrogate matrices as in the ideal case. Assuming small rotations, the gradient $\frac{\partial h_1^{-1}}{\partial v}$ is small, and using such a model is expected to outperform linear interpolation that ignores magnifications.

6.5 Revisiting current data resampling strategies

Within our formalism, several conventional projection models can be revisited and their limitations highlighted.

6.5.1 Distance-driven interpolation

Similar to our approach, the DD model [72] captures both sides of the sampling process, at the voxel and detector bin levels, but from the perspective of a geometrical discretization, which is not specific to flat panel detector and therefore not relying on projection matrices and homographies. For planar parallel or fan-beam geometry, or when separability holds, the 1D version is used as follows: voxels and bins are located by their edges on their respective axis. These locations are mapped according to the system geometry onto a common axis. Interpolation between one voxel and one bin is computed as the length of the overlapping segment footprints of the voxel and the bin over this axis as shown in Figure 6.6 for a flat-panel detector. Under this choice, the scheme is neither destination

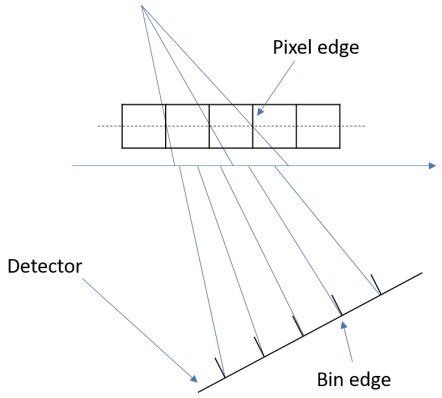


Figure 6.6: 1D Distance-Driven

nor source driven, rendering it equally adequate for projection and backprojection. This results in a matched pair (up to normalization factors) for this particular axis.
Remarkably, for projection, for every $i \in I$ and $j \in J$,

$$\xi^{0,0}_{\Delta_i}(h^{-1}(u_i) - x_j) = \int_{-\infty}^{+\infty} \beta^0_1(\tau)\beta^0_{\Delta_i}(h^{-1}(u_i) - x_j - \tau)d\tau = \int_{x_j-\frac{1}{2}}^{x_j+\frac{1}{2}} \beta^0_{\Delta_i}(h^{-1}(u_i) - \tau)d\tau. \tag{6.90}$$

The above integral is equal to the intersection of the support of β^0_1 centered at x_j and of length 1, with that of $\beta^0_{\Delta_i}$ centered at $h^{-1}(u_i)$ and of length Δ_i, which is the quantity at the core of the DD scheme. The function $\beta^0_{\Delta_i}(h^{-1}(u_i) - \cdot)$ can be viewed as the projection of the detector bin centered at u_i, that is a segment centered at $h^{-1}(u_i)$ with length Δ_i. As a result, in the context of flat panel cone-beam geometry, the DD scheme can be expressed in our framework with $n = m = 0$ with a slightly different set of magnification factors $\boldsymbol{\Delta}^{\mathrm{DD}}$ as described in Remark 6.4.3.1. In particular, our choice of setting $\boldsymbol{\Delta}^{\mathrm{DD}} = \boldsymbol{\Delta}'$ corresponds to the case when the intermediary axis is chosen as the detector axis while setting $\boldsymbol{\Delta}^{\mathrm{DD}} = \boldsymbol{\Delta}$ corresponds to magnifications at the voxel axis. In the following, we shall not discuss this degree of freedom of the method, we will instead consider the DD with any intermediate axis and associated set of magnifications as being an instance of our approach when $n = m = 0$.

6.5.2 Destination-driven interpolation

The standard Joseph ray-driven projection model and voxel-driven backprojection model are instances of C-D resampling. C-D backprojection with homography matrices $\mathbf{H}$ is straightforward: from arbitrary location (x_{j_1}, z_{j_2}), the corresponding coordinates onto the detector (u, v) are computed according to (6.5). Then interpolation takes place in the projection space. Likewise, C-D forward projection based on $\mathbf{H}^{-1}$ steps through destination locations, i.e. every bin center (u_{i_1}, v_{i_2}) and finds the corresponding set of voxels in the volume that map into the output. Setting (u_{i_1}, v_{i_2}) into (6.5) yields the equation of the line that goes from location (u_{i_1}, v_{i_2}) to the focal point S. Numerical integration of volume f is then performed over this line (or ray) by means of interpolation in the volume space. The most common instance of this approach consists in using bilinear interpolation. For 1D projection, this amounts to plugging $h^{-1}(u_i)$ in (6.30) with $m = 1$. Our framework offers an alternative interpretation of this approach. In (6.40) if we set $n = m = 0$ and define set $\mathbf{\Delta}$ with constant sampling step 1 in the normal equations (6.34), then the C-D projection footprint matrix $\mathbf{F}$ is such that, for every $i \in I$ and $j \in J$,

$$F_{i,j} = \xi_1^{0,0}(x_j - h^{-1}(u_i)) = \beta^1(x_j - h^{-1}(u_i)) \tag{6.91}$$

while the C-D backprojection footprint matrix $\mathbf{F}'$ in (6.46) is

$$F'_{j,i} = \xi_1^{0,0}(u_i - h(x_j)) = \beta^1(u_i - h(x_j)). \tag{6.92}$$

Even though both $\mathbf{F}$ and $\mathbf{F}'$ use β^1, they rely on different representations so that they are far from being the transpose of each other. This teases out the *magnification-agnostic* nature of such projection and backprojection models, thus explaining their limitations in terms of adjoint and noise handling as we will show further on in our numerical experiments.

6.5.3 Data downsampling

As we outlined in Chapter 2 (section 6.2), in the current medical practice, the reconstructed slices are of size 512×512 while X-ray flat panels with a pitch of 200μ can deliver many more samples, yielding a small ratio α/t_z in (6.2) and (6.3). Therefore, the downsampling factor is large. The C-D projector thus oversamples the volume, yielding modelization errors when applying the adjoint during reconstruction, while the C-D backprojector misses samples, thus not making use of the full X-ray dose. To limit downsampling, reconstruction needs to be performed on small voxels. However, it must also comply with the clinical constraints of fast reconstruction and limited storage capability in favor of downsampling. Besides, in IR, using smaller voxels increases L and degrades the conditioning of $\mathbf{R}_s$ because the angular sampling becomes insufficient. Since it is impossible to oversample the data in the angular direction, prior downsampling is often performed. This pre-processing induces a loss of information and forces optimizing image and data representations. A prior rebinning[3] of the data to larger pixels by block-averaging is commonly used. Rebinning trades resolution loss for noise and aliasing reduction. Resizing with a non-necessarily integer factor is less restrictive, but bilinear interpolation modifies the noise properties by introducing correlations. This is detrimental to statistical reconstruction methods directly modeling the detector's statistically independent bin measurements. One expected outcome of using convolutions of B-splines is to provide less alteration of the noise properties to allow one to keep a simple noise model

[3]The term *rebinning* refers to using constant magnifications of integer factors at the detector level

after data resampling, whether after a magnification by a real factor or a rectification homography. Our model embeds the magnifications factors, hence alleviating the need for rebinning. With better noise handling, IR allows for improved resolution. Note again that a B-spline of order 0 at the detector level is a faithful resolution degradation model, but a large cubic voxel model may not be appropriate for representing a higher resolution volume. Our approach shows how to introduce a higher level of precision for the solution (potentially higher than the precision of the DD).
Another downsampling routine that may be used for reconstruction is *rectification* [184]. Rectification relies on the observation that, for any paired homographies $(\mathbf{H}_{y_0}, \mathbf{H}_{y_1})$ derived from projection matrix $\mathbf{P}$, each one can be deduced from the other by a magnification. The decomposition of the projection into a composition of homographies can thus be simplified into computing a single rectification homography derived from, e.g., $\mathbf{H}_{y_0}$, to which 2D magnification $\mathbf{H}_{y_1}\mathbf{H}_{y_0}^{-1}$ is applied to obtain $\mathbf{H}_{y_1}$. As magnifications present a better computation layout than homographies, rectification has been introduced to get faster FP-BP pairs for both FBP and IR. Resampling based on the convolution of B-splines is an obvious candidate for both steps. The last benefit of IR is to reduce the undersampling artifacts appearing, for instance, in CBCT with a circular orbit or with a limited number of projections [130]. Missing projections are obviously independent of the interpolation scheme within a projection; in that case, an accurate model minimizes interpolation errors [151]. This is easier to achieve using a virtual rectified detector with 2D magnifications where the order of the B-spline sets the compromise between speed and precision. We see that there is no one-fit-for-all discretization, but B-splines adapt remarkably well to one's various needs.

6.6 Application

6.6.1 Experiments

We tested our magnification-driven interpolation scheme for cone-beam projection using orders $(n, m) \in \{0, 1\}^2$ with $m \geqslant n$. The destination-driven FP and bp based on linear interpolation taken as a reference, are denoted by $\mathbf{R}_r$ and $\mathbf{B}_r$. The corresponding homographic transforms are denoted by $\tilde{\mathbf{H}}_r^{-1}$ and $\mathbf{H}_r$. In subsection 6.4.4, we proposed various simplified implementations of FP and BP. Henceforth, we will use a label a, b, or c to specify the chosen implementation. For backprojection, implementation a consists in computing $\mathbf{p}$ according to (6.87) using inversion of tridiagonal matrices [150,175]. Implementation b consists in substituting $\operatorname{Diag}(\boldsymbol{\Delta})^{-1/2}\mathbf{T}_n^{-1}\operatorname{Diag}(\boldsymbol{\Delta})^{-1/2}$ for $\mathbf{G}_1^{-1}$ in (6.87). Based on implementation b, implementation c consists in further replacing $\tilde{\boldsymbol{\Lambda}}_1$ by $\boldsymbol{\Lambda}_1$. It can be noticed that given the range values of (n, m), (6.87) then becomes

$$\mathbf{p} = \big((\operatorname{Diag}(\boldsymbol{\Delta})^{1/2}\mathbf{T}_n^{-1}\operatorname{Diag}(\boldsymbol{\Delta})^{-1/2}) \otimes \tilde{\mathbf{G}}_2^{-1}\big)\mathbf{F}\mathbf{c}. \tag{6.93}$$

Note that with c, post-filtering through operator $\tilde{\mathbf{G}}_2^{-1}$ in (6.93) is completely independent of the homography step. By linearity, it can thus be performed in a single global pass, after summation of each transformed plane, on the resulting projections (or on the volume for backprojection).

6.6.1.1 Simulation scenarios

An ideal cone-beam geometry is considered; it is made of 600 projection matrices covering a 360° circular acquisition. By ideal, we mean that the trajectory of the source point is strictly planar and that the optical axis always crosses the center of rotation and hits the center of the detector, which gives $t_x = t_y = u_0 = v_0 = 0$. All projection matrices are thus of the form

$$\mathbf{P} = \begin{pmatrix} \alpha\cos\theta & \alpha\sin\theta & 0 & 0 \\ 0 & 0 & \alpha & 0 \\ -\sin\theta & \cos\theta & 0 & t_z \end{pmatrix}. \tag{6.94}$$

The detector size is such that the data is never truncated. Isosampling is defined as $t_z = \alpha = 1500$ voxels. The set of magnification factors for approximating the homographies is defined at the destination level, namely at the detector level for $\mathbf{R}_s$ and at the volume level for $\mathbf{B}_s$. Recall that these operators rely on the repeated use of operators $\mathbf{H}_s$ and $\tilde{\mathbf{H}}_s^{-1}$ for various matrices $\mathbf{H}$. We also considered FP and BP $\mathbf{V}_s$ corresponding to the discretization of the associated rectified virtual geometry made of magnifications only. The latter case serves as a baseline since it is known that magnification-driven interpolation fulfills optimality conditions for magnifications [215].

6.6.1.2 Tasks

Projection for the whole orbit involves a mix of homographies, from simple magnification when the detector is aligned with the volume to the worst case for a Cartesian grid at $\pi/4$. The impact of such homographies is over coronal and sagittal slices. Therefore, we first tested our different implementations of $\mathbf{H}_s$ and $\tilde{\mathbf{H}}_s^{-1}$ for the matrix $\mathbf{H}$ deduced from $\theta = \pi/4$ and $y = 0$ as

$$\mathbf{H} = \begin{pmatrix} \alpha & \alpha\sin\pi/4 & 0 \\ 0 & \alpha & 0 \\ -\sin\pi/4 & 0 & t_z \end{pmatrix}. \tag{6.95}$$

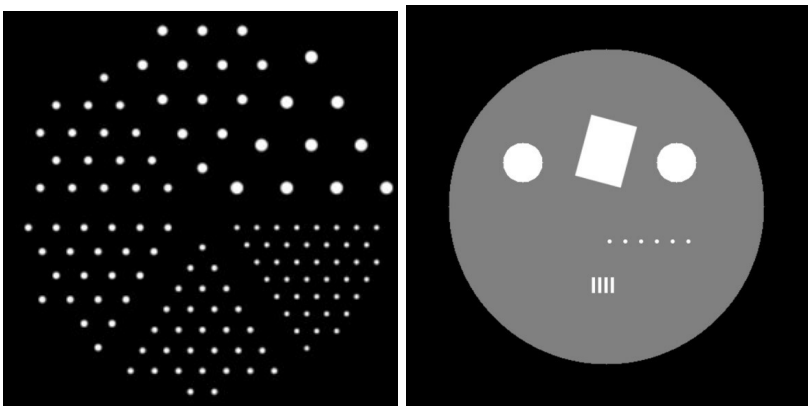
We tested resampling steps $\delta_H = \alpha/t_z \in \{1, 2, 3.5\}$, where $\delta_H = 1$ means isosampling and $\delta_H > 1$ means that the voxel size is chosen δ_H times bigger than isosampling, and that downsampling by a factor δ_H is performed.
Projectors and backprojectors themselves were evaluated through tasks of analytical and iterative reconstructions. For analytical reconstruction, each model $\mathbf{R}_r$ and $\mathbf{R}_s$ was successively used to simulate the projection of a vertical edge at isosampling followed by FDK reconstruction with the corresponding operator $\mathbf{B}_r$ or $\mathbf{B}_s$. At backprojection, all projection matrices were rotated by angle atan(1/16) to yield a slanted edge in the reconstructed image so that the edge is sampled with 16 sub-voxel shifts. For iterative reconstruction, the pair $(\mathbf{R}_s, \mathbf{R}_s^\top)$ is employed. For a given forward model $\mathbf{R}_s$, IR was the result of minimizing the following objective function:

$$\Psi(\mathbf{f}) = \frac{1}{2}\|\mathbf{R}_s\mathbf{f} - \mathbf{p}\|^2 + \frac{\beta}{2}\|\mathbf{f}\|_2^2. \tag{6.96}$$

Parameter β, set to 2×10^{-2}, ensures the strong convexity of the cost function Ψ, and thus the uniqueness of the minimizer. The initial value for volume $\mathbf{f}$ was set to the zero vector. The minimization problem was solved by simple gradient descent. Vector $\mathbf{p}$ is the projection of the 512×512 image displayed in Figure 6.7b with $\mathbf{R}_r$ performing an 8-times oversampling followed by an 8×8 bin averaging.

6.6.1.3 Image quality metrics

We compared the models in terms of bias (accounting for the loss of spatial resolution or presence of artifacts) and noise propagation. For visual assessment of bias, we used two simulated images: a "wire" image made of cylinders of varying diameters and fixed intensity set to 100 Figure 6.7a, and a phantom Figure 6.7b containing sharp geometrical shapes (two cylinders, one rectangle, six wires, and one line pair pattern). Spatial resolution was assessed by computing the modulation transfer functions (MTF) of the FDK-reconstructed slanted edge. Bias was evaluated for IR as the root mean square error (RMSE) with respect to the ground truth over uniform ROIs.
For noise propagation, we used ensemble statistics: for two different operations (homography $\mathbf{H}_s$ and backprojection $\mathbf{B}_s$), N statistical replicates of noise were processed by the same operation so that the mean, standard deviation (STD), and signal-to-noise ratio (SNR), taken as the ratio of mean to standard deviation, can be computed at each pixel of the output after processing. The 2D Noise Power Spectrum (NPS) can also be computed and averaged radially over a circular region of interest. The noise, added on each operation input, was always independent and identically distributed, zero-mean and Gaussian with variance set to 1 for each test of $\mathbf{H}_s$ and to 10^3 for tasks involving $\mathbf{B}_s$.



(a) Resolution image

(b) Phantom for iterative reconstruction (axial slice)

Figure 6.7: Reference images

6.6.1.4 C-arm CBCT data

An exemplary real-data case is also studied. A CBCT acquisition of a quality assurance phantom, containing a resolution section with bar patterns, was obtained on a GE Healthcare C-arm system with a circular orbit of 200° sampled by 607 projections. The detector bin size was 0.2 mm. The distance from the focal spot to the detector is 1295 mm which yields $\alpha = 1295/0.2 = 6475$. The distance from the focal spot to the center of rotation is 820 mm. The voxel size at isosampling is 0.127 mm ($t_z = 820/0.127 = 6456$). A 512×512 image with this voxel size yields only a field of view of 65 mm. We, therefore, compared the performance of interpolation for the voxel size at isosampling and for a four times bigger field of view by increasing the voxel size to 0.508 mm. For the latter case, we compared FDK reconstructions with the following three options: i) direct backprojection of the original data ($\alpha = 1295/0.2$, $t_z = 820/0.508$), ii) rectification of the original data ($\alpha = 1295/0.200$, $t_z = 820/0.127$) followed by backprojection in rectified geometry ($\alpha = 820/0.127$, $t_z = 820/0.508$) and iii) bin averaging by a factor 4, designated

by operator $\mathbf{A}$, followed by direct backprojection ($\alpha = 1295/0.800$, $t_z = 820/0.508$). We compared the reconstruction of the bar pattern of 8 line pairs per mm on an axial slice 7.62 mm away from the central slice. The system perfectly resolves this pattern when the voxel size is equal to 0.127 mm, but it is degraded for a voxel size of 0.508 mm. The amount of degradation induced by the large voxel size depends only on the interpolation, not the system. In such real-data conditions, (6.6) is not rigorously satisfied, hence we fall in the situation described in Remark 6.4.5.1.

6.6.2 Results

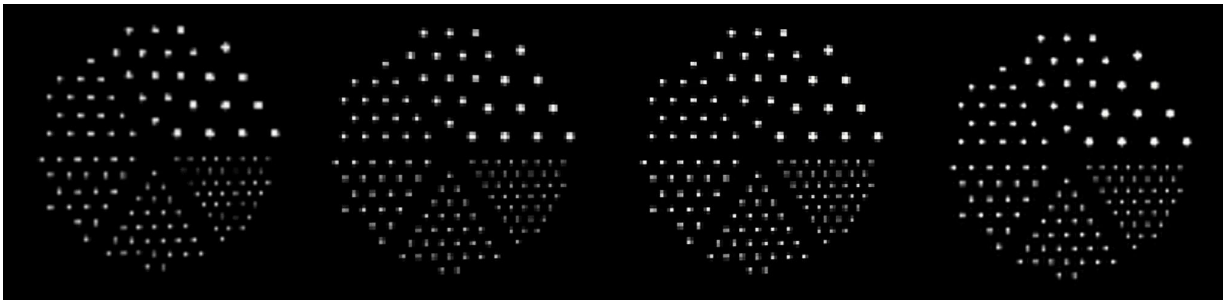
6.6.2.1 Homography

Table 6.2 reports measures of RMSE associated to implementations b and c with respect to implementation a after applying either homography $\tilde{\mathbf{H}}_s^{-1}$ or $\mathbf{H}_s$ with our splines of order $s = m + n + 1$ on the wire image degraded by noise. First, for all models, the errors are very low. For $\delta_H > 1$, $\mathbf{H}_s$ corresponds to a downsampling while $\tilde{\mathbf{H}}_s^{-1}$ performs an upsampling. We see almost no error when performing downsampling with our lowest order model. Moreover, implementations b and c are on average 1.15 and 1.4 times faster than implementation a for the considered homographies. Based on these facts, we now focus on implementation c and drop the corresponding index.

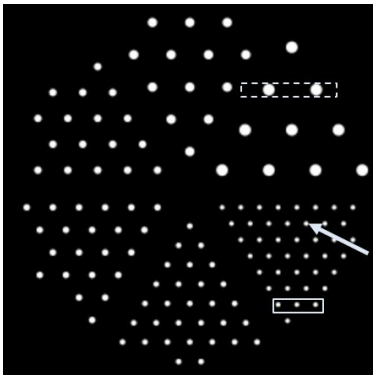
	$\mathbf{H}_1$		$\mathbf{H}_2$		$\mathbf{H}_3$		$\tilde{\mathbf{H}}_1^{-1}$		$\tilde{\mathbf{H}}_2^{-1}$		$\tilde{\mathbf{H}}_3^{-1}$	
	b	c	b	c	b	c	b	c	b	c	b	c
$\delta_H = 1$	0	0	5.2	22	4.6	20	53	53	49	56	46	53
$\delta_H = 2$	0	0	37	54	35	52	10	10	27	31	26	30
$\delta_H = 3.5$	0.1	0.1	25	25	24	25	1.8	1.8	0.2	2.9	0.2	2.4

Table 6.2: Comparison of implementations b and c with respect to implementation a in terms of RMSE ($\times 10^{-3}$)

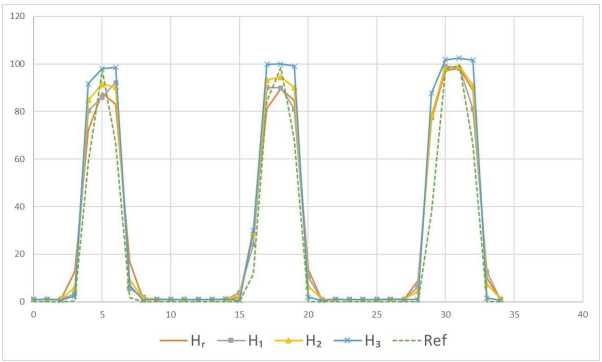
Figure 6.8a shows the wire images obtained after performing a homography with $\delta_H = 3.5$ followed by the inverse homography. One can notice that the image obtained with linear interpolation i.e. $\tilde{\mathbf{H}}_r^{-1}\mathbf{H}_r$ lacks some wires as indicated by the arrow on Figure 6.8b. This highlights the issue of nonstationarity with linear interpolation. With the proposed magnification-driven interpolation, all wires are visible. The images obtained with $\tilde{\mathbf{H}}_1^{-1}\mathbf{H}_1$ have a patchy look typical of 0-order B-spline models. The highest-order B-spline model gives the images with the least distortion at the price of small undershoots. These differences between models are less apparent for lower downsampling factors as illustrated for $\delta_H = 1$ by Figure 6.8c showing the profile through the three bottom-right wires (solid box) on Figure 6.8b. The profiles with $\tilde{\mathbf{H}}_2^{-1}\mathbf{H}_2$ and $\tilde{\mathbf{H}}_3^{-1}\mathbf{H}_3$ show higher resolution than $\tilde{\mathbf{H}}_r^{-1}\mathbf{H}_r$ and $\tilde{\mathbf{H}}_1^{-1}\mathbf{H}_1$.



(a) Output wire slice using $\delta_H = 3.5$. From left to right: $\tilde{\mathbf{H}}_r^{-1}\mathbf{H}_r$, $\tilde{\mathbf{H}}_1^{-1}\mathbf{H}_1$, $\tilde{\mathbf{H}}_2^{-1}\mathbf{H}_2$, $\tilde{\mathbf{H}}_3^{-1}\mathbf{H}_3$.



(b) Input resolution image



(c) Plots through three wires (solid box in Figure 6.8b) using $\delta_H = 1$

Figure 6.8: Assessment of resolution for direct and inverse homography

Figure 6.9 shows the SNR images obtained from $N = 200$ replicate homographies $\tilde{\mathbf{H}}_s^{-1}\mathbf{H}_s$ and $\tilde{\mathbf{H}}_r^{-1}\mathbf{H}_r$ for $\delta_H = 2.025$. The window widths (WW) and window levels (WL) are set independently for each image. We see that noise correlation appears along the columns where the homography is a magnification, while along the rows, the varying local magnification factor of the homography induces a complex pattern. These correlations vanish as the B-spline order grows. The mean and standard deviation of these SNR images are reported in Table 6.3.

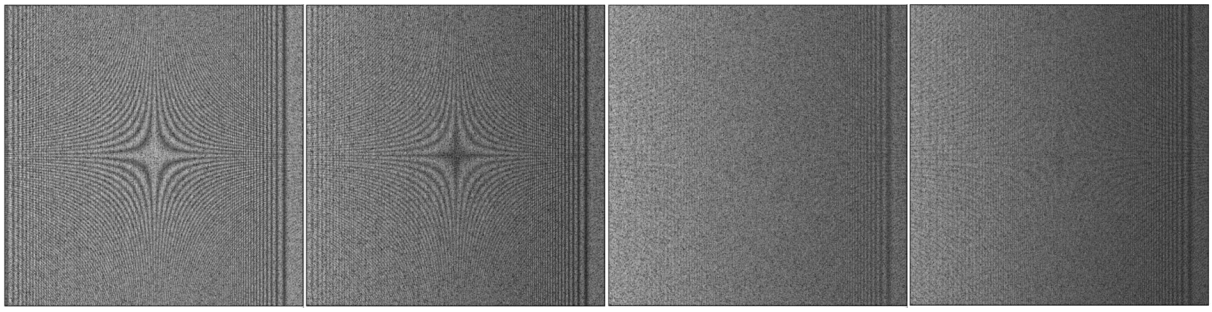
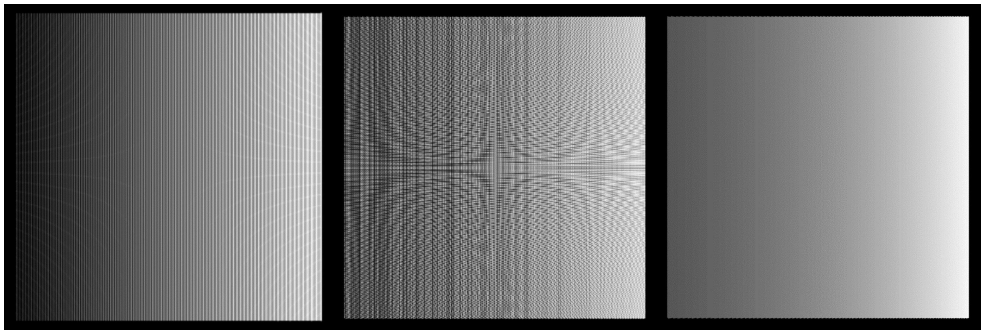


Figure 6.9: SNR images using $\delta_H = 2$. From left to right: $\tilde{\mathbf{H}}_r^{-1}\mathbf{H}_r$, $\tilde{\mathbf{H}}_1^{-1}\mathbf{H}_1$, $\tilde{\mathbf{H}}_2^{-1}\mathbf{H}_2$, $\tilde{\mathbf{H}}_3^{-1}\mathbf{H}_3$.

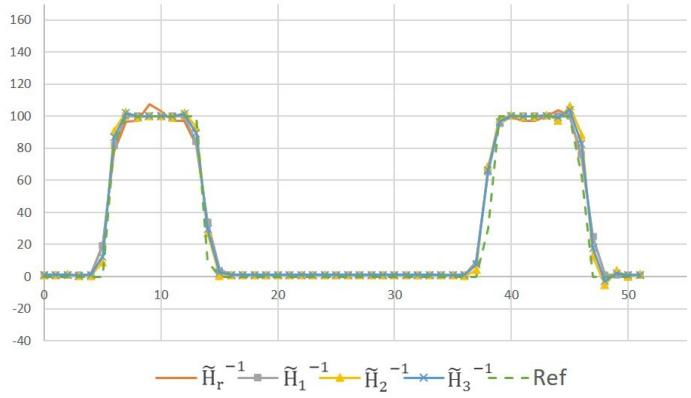
	$\tilde{\mathbf{H}}_r^{-1}\mathbf{H}_r$	$\tilde{\mathbf{H}}_1^{-1}\mathbf{H}_1$	$\tilde{\mathbf{H}}_2^{-1}\mathbf{H}_2$	$\tilde{\mathbf{H}}_3^{-1}\mathbf{H}_3$
$\delta_H = 1$	29 ± 7.8	28 ± 7.9	25 ± 4.5	21 ± 3.7
$\delta_H = 2$	29 ± 7.8	33 ± 8.5	27 ± 6.5	23 ± 6.8
$\delta_H = 3.5$	29 ± 7.8	48 ± 13	42 ± 12	40 ± 13

Table 6.3: Mean $\pm$ standard deviation of the SNR image generated by $\tilde{\mathbf{H}}_s^{-1}\mathbf{H}_s$

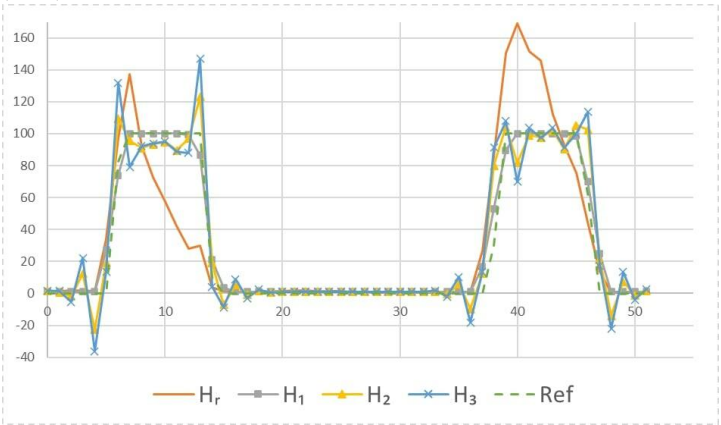
First of all, linear interpolation has the same mean SNR no matter the change in the sampling step, as only two samples are always considered. In contrast, with magnification-driven models, the mean SNR grows with the downsampling factor as more samples get involved. Model $\tilde{\mathbf{H}}_1^{-1}\mathbf{H}_1$ gives a higher mean SNR than linear interpolation when $\delta_H > 1$ and an equivalent mean SNR for smaller δ_H. Models $\tilde{\mathbf{H}}_3^{-1}\mathbf{H}_3$ and $\tilde{\mathbf{H}}_2^{-1}\mathbf{H}_2$ display higher mean SNR than linear interpolation for large downsampling factors and lower ones for lower downsampling factors. Understandably, the reduction presented above of the image distortions for $s > 1$ is associated with noisier images. The noise is compensated by a superior resolution, resulting in SNR increases.
Regarding the evaluation of the adjoint scheme, Figure 6.10 compares the interpolation models for operation $\mathbf{H}_s^\top\mathbf{H}_s$ and $(\tilde{\mathbf{H}}_s^{-1})^\top\tilde{\mathbf{H}}_s^{-1}$ with $\delta_H = 2$. On Figure 6.10a, one can see that, since linear interpolation always uses two samples no matter the magnification factor, applying $\mathbf{H}_r^\top\mathbf{H}_r$ or $(\tilde{\mathbf{H}}_r^{-1})^\top\tilde{\mathbf{H}}_r^{-1}$ to a constant image results in artificial high-frequency patterns that are eliminated with our models. The model order has no impact on a constant image. Regarding noise propagation, $(\tilde{\mathbf{H}}_r^{-1})^\top\tilde{\mathbf{H}}_r^{-1}$ yields a higher SNR (25 ± 4.8) than $(\tilde{\mathbf{H}}_s^{-1})^\top\tilde{\mathbf{H}}_s^{-1}$ (20 ± 4.5 for $s = 1$, 20 ± 3.5 for $s = 2$, 19 ± 3.6 for $s = 3$). In all scenarios, for higher-order models, the SNR images, however, display reduced correlation patterns. Figure 6.10b and Figure 6.10c show two profiles taken at large wires of the resolution image for $\mathbf{H}_s^\top\mathbf{H}_s$ and $(\tilde{\mathbf{H}}_s^{-1})^\top\tilde{\mathbf{H}}_s^{-1}$. Oscillations patterns are mostly visible with schemes based on linear interpolation and are more pronounced with $\mathbf{H}_r^\top\mathbf{H}_r$. Small overshoots are noticeable at the border of the wires with $\mathbf{H}_2^\top\mathbf{H}_2$ and $\mathbf{H}_3^\top\mathbf{H}_3$ that do not appear using any model $(\tilde{\mathbf{H}}_s^{-1})^\top\tilde{\mathbf{H}}_s^{-1}$.



(a) Uniformity. From left to right: $(\tilde{\mathbf{H}}_r^{-1})^\top \tilde{\mathbf{H}}_r^{-1}$, $\mathbf{H}_r^\top \mathbf{H}_r$, others: $(\tilde{\mathbf{H}}_s^{-1})^\top \tilde{\mathbf{H}}_s^{-1}$ and $\mathbf{H}_s^\top \mathbf{H}_s$. Grayscale: 0-50



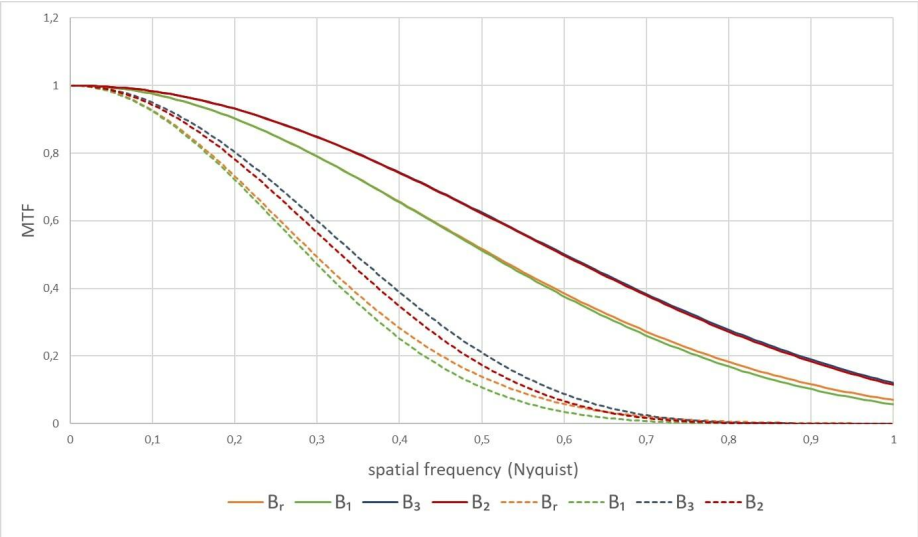
(b) Plots through two large wires (dotted box of Figure 6.8b) with $(\tilde{\mathbf{H}}_s^{-1})^\top \tilde{\mathbf{H}}_s^{-1}$



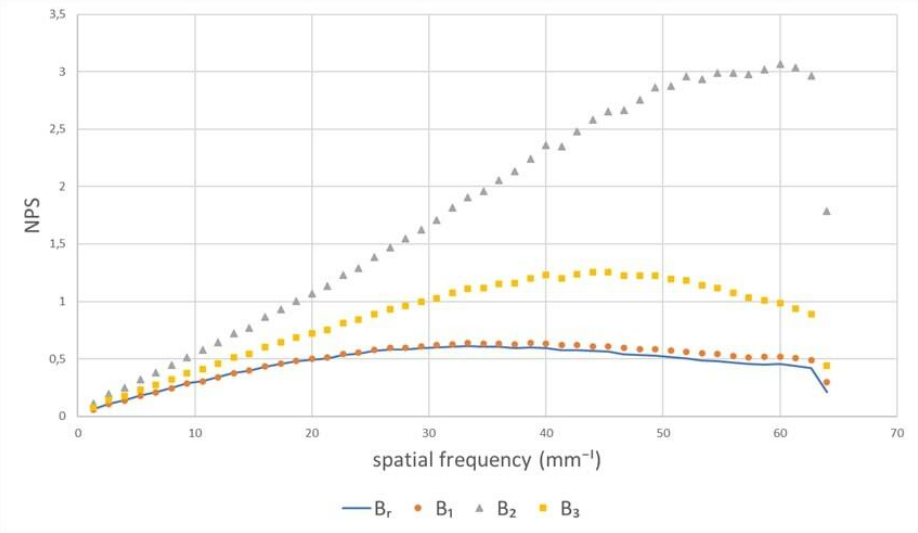
(c) Plots through two large wires (dotted box of Figure 6.8b) with $\mathbf{H}_s^\top \mathbf{H}_s$

Figure 6.10: Evaluation of direct and adjoint homography with $\delta_H = 2$.

6.6.2.2 Projector and backprojector



(a) MTF curves obtained with $t_z = \alpha$ (solid lines) and with $t_z = 2\alpha$ (dashed lines)



(b) 1D radial NPS obtained with $t_z = \alpha$

Figure 6.11: Frequency Analysis

Figure 6.11a displays the MTF curves obtained for FDK reconstruction at iso-sampling and with $t_z = 2\alpha$ for the direct geometry. All curves were normalized to 1 at zero frequency. At iso-sampling, the MTF curves for $\mathbf{B}_r$ and $\mathbf{B}_1$ are superimposed. Above, one can see the curves obtained with $\mathbf{B}_2$ and $\mathbf{B}_3$ that are also superimposed. When $t_z = 2\alpha$, model $\mathbf{B}_3$ outperforms $\mathbf{B}_2$ while the MTF for linear interpolation is slightly higher than that of $\mathbf{B}_1$. For the rectified geometry, $\mathbf{V}_s$ provides the same MTF as their counterparts $\mathbf{B}_s$ (curves not shown).
Figure 6.11b displays the radial NPS curves obtained from the replicate FDK reconstructions at iso-sampling. The positive slope of these curves results from the ramp filtering. Models $\mathbf{B}_r$ and $\mathbf{B}_1$ behave similarly and correlate more the noise than higher-order models.

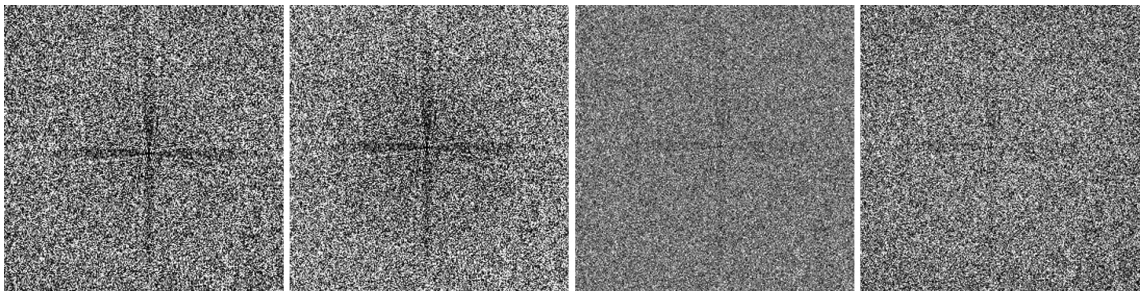
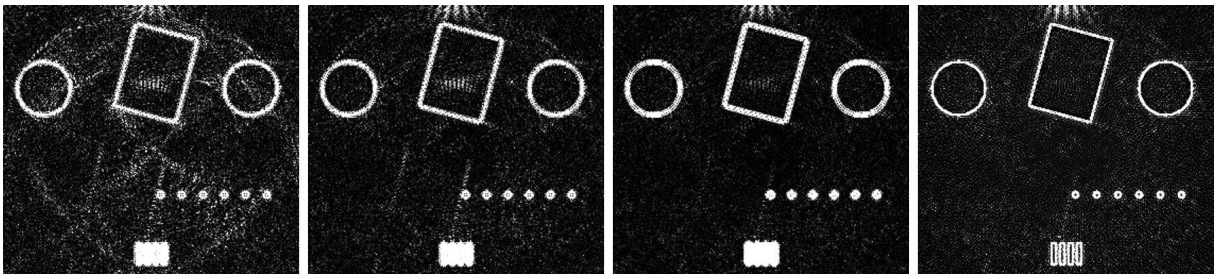


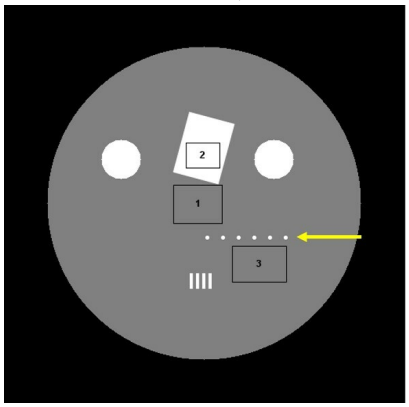
Figure 6.12: SNR images after FDK reconstruction of a uniform cylinder. From left to right, $\mathbf{B}_r$ (WL=128, WW=18), $\mathbf{B}_1$ (WL=122, WW=18), $\mathbf{B}_2$ (WL=63, WW=18) and $\mathbf{B}_3$ (WL=100, WW=18).

Figure 6.12 shows the SNR images obtained from $N = 200$ replicate FDK reconstructions of a uniform cylinder with $\mathbf{B}_r$ and $\mathbf{B}_s$ at iso-sampling. The window level was chosen as the mean value of each cylinder image (i.e., each mean SNR), while the window width was kept constant. Uniform SNR images are expected. We see that models $\mathbf{B}_r$ and $\mathbf{B}_1$ have similar mean SNR and display a small streak pattern, which is reduced with models $\mathbf{B}_2$ and $\mathbf{B}_3$. Model $\mathbf{B}_3$ appears as the best compromise between a uniform SNR and a high global SNR level.

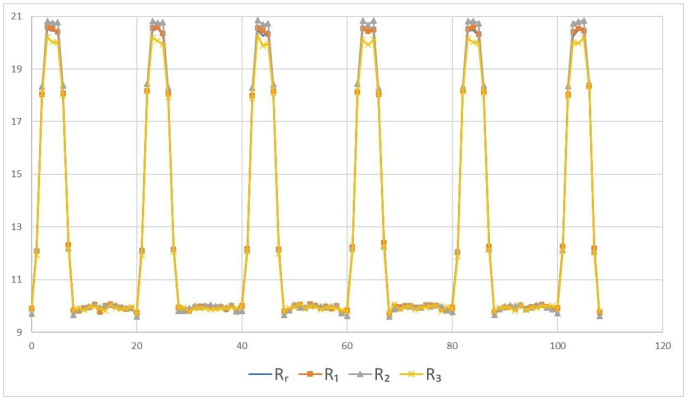
Figure 6.13a shows a zoom on the error images between the ground truth and IR after 300 iterations for the central slice using projectors $\mathbf{R}_s$ and $\mathbf{R}_r$. Edge distortions and aliasing patterns are visible in all images. However, they are more pronounced with $\mathbf{R}_r$. The strength of these artifacts was quantified as the mean RMSE over a union of three ROIs where they have the strongest effect. The RMSE errors are 0.74%, 0.59%, 0.56% and 0.53% of the background for projectors $\mathbf{R}_r$, $\mathbf{R}_1$, $\mathbf{R}_2$, and $\mathbf{R}_3$ respectively. Projector $\mathbf{R}_r$ yielded the highest error while projector $\mathbf{R}_3$ led to the lowest one. Plots of the horizontal profiles along the central wire are presented in Figure 6.13c. Peaks have a greater intensity for $\mathbf{R}_r$, $\mathbf{R}_1$, and $\mathbf{R}_2$ than the ground truth intensity (equal to 20), while for $\mathbf{R}_3$ the intensity is correct. The iterative process inverts the discretization errors of the projector, which yields a stronger unwanted deconvolution. The profiles thus show that higher-order projectors induce fewer deconvolution biases because they rely on a more accurate representation of the signal. This means that magnification-driven interpolation can lead to reduced edge artifacts with respect to linear interpolation.



(a) Error images between reconstructions and phantom (WL=0.15, WW=0.30). From left to right: $\mathbf{R}_r$, $\mathbf{R}_1$, $\mathbf{R}_2$, $\mathbf{R}_3$



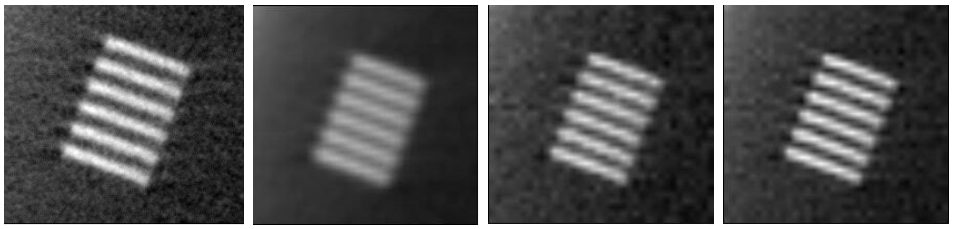
(b) Phantom with three ROIs



(c) Profiles through the six wires pointed by the arrow of Figure 6.13b

Figure 6.13: Iterative reconstruction of the simulated geometric phantom after 300 iterations

6.6.2.3 Real data



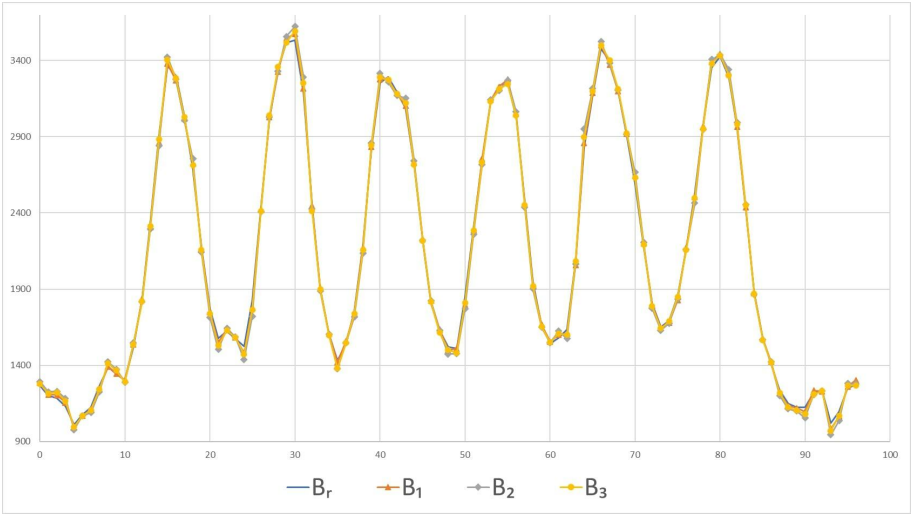
(a) $\mathbf{B}_r$ ($t_z = \alpha$) (b) $\mathbf{B}_r\mathbf{A}$ ($t_z = 4\alpha$) (c) $\mathbf{B}_r$ ($t_z = 4\alpha$) (d) $\mathbf{B}_3$ ($t_z = 4\alpha$)

Figure 6.14: C-arm CBCT reconstruction of a quality assurance phantom. Displayed ROI centered on the bar pattern of 8 lp/cm (WL = 1200, WW = 2000).

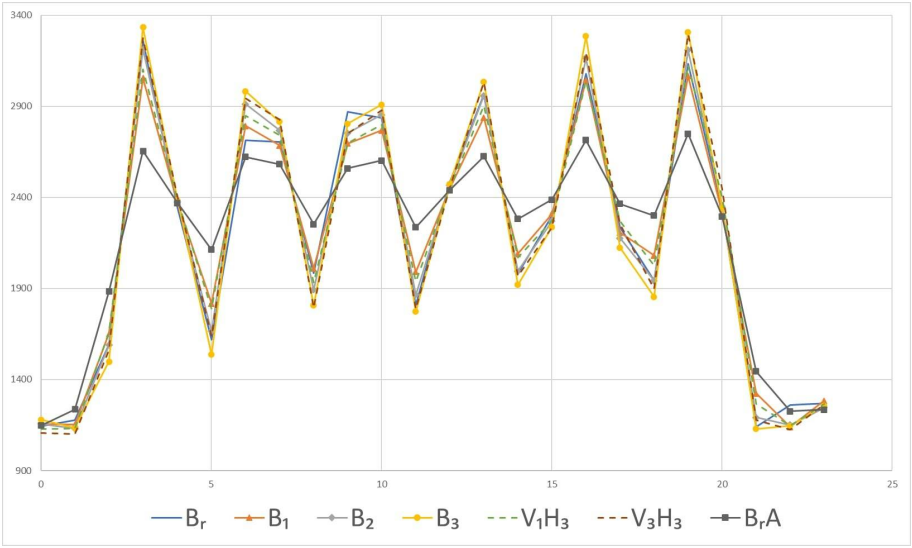
Figure 6.14 shows the reconstructions of the bar pattern of 8 line pairs at iso-sampling and with a downsampling factor of 4. Figure 6.15 shows the profiles through the resolution bars according to different options and interpolation models.
All models offer similar performance when reconstructing at iso-sampling. With a downsampling of 4, the worst case is the standard approach of detector rebinning followed by linear interpolation. A first improvement is obtained by substituting the rebinning step with rectification. The highest resolution is obtained using the native geometry with $\mathbf{B}_2$ and $\mathbf{B}_3$. With $\mathbf{B}_r$, the issue of non-stationarity from peak to peak is again visible.
Table 6.4 shows that bin averaging of the data with $\mathbf{A}$ before backprojecting with linear interpolation $\mathbf{B}_r$ achieved the best noise performance but at the price of a strong loss of resolution. Linear interpolation without bin averaging $\mathbf{B}_r$ yielded the highest noise level. In contrast, model $\mathbf{B}_1$, which led to a resolution very close to that provided by $\mathbf{B}_r$, was associated with the second lowest RMSE, which is twice lower than the RMSE obtained with $\mathbf{B}_r$. Using a first rectification homography $\mathbf{H}_s$ followed by our spline models in rectified geometry $\mathbf{V}_s$ achieved an intermediate compromise between noise and resolution for a reduced computational complexity. Using model $\mathbf{V}_3\mathbf{H}_3$ barely decreased spatial resolution compared to performing a direct reconstruction with $\mathbf{B}_3$ while gaining noise uniformity (RMSE decrease of 20 sHU).
All these observations show that the behavior of magnification-driven interpolation assessed in simulations extends to real data. Moreover, despite a non-ideal geometry, requiring a non-separable interpolation and approximate local magnifications, our approach still resulted in improvements over C-D methods with linear interpolation.

$\mathbf{B}_r$	$\mathbf{B}_1$	$\mathbf{B}_2$	$\mathbf{B}_3$	$\mathbf{B}_r\mathbf{A}$	$\mathbf{V}_1\mathbf{H}_3$	$\mathbf{V}_3\mathbf{H}_3$
151	76	136	129	43	85	110

Table 6.4: Mean RMSE of the four uniform ROIs



(a) $t_z = \alpha$



(b) $t_z = 4\alpha$

Figure 6.15: Profiles through the bar pattern of 8 lp/cm shown in Figure 6.14

6.6.3 Discussion

The modeling of magnification, performed through optimal expansions over B-splines of varying widths, provides a new framework for computing homographies found in flat-panel cone-beam FP and BP. The proposed framework generalizes current separately developed approaches of magnification-agnostic signal resampling on the one hand and geometric discretization on the other. The first approach works well enough for common image processing tasks, especially when using high-order polynomial interpolation. However, in the practice of clinical X-ray imaging, nearest neighbor (for rebinning) and linear interpolations are preferred for AR. The reasons are related to limited computation and storage resources, and the constraint of full X-ray dose utilization, so spatial resolution is sacrificed to noise reduction through the use of large or rebinned acquisition pixels. We here provide an array of interpolation kernels to fully benefit from small acquisition pixels with improved data resizing or rectification steps and improved projection inner steps. Our method displays reasonably simple computations through interpolation kernels of order up to 3 and enhanced noise uniformity.
We do not claim that resolution can be improved and noise decreased simultaneously: the proposed kernels may either keep the noise or blur the signal, but they do not randomly lose information, as may happen when downsampling with a magnification-agnostic linear interpolation. At iso-sampling, the current state-of-the-art linear and DD interpolations have similar performances.
The shortcomings of the adjoint of the destination-driven FP are induced by the lack of magnification modeling in the interpolation process. The alternative footprint approach has been introduced to overcome this issue. Again, dose and computation constraints have made the DD model a better tool for IR than alternative classical tools of image processing. Put into our framework, we get several advantages. First, the DD appears to oscillate between linear interpolation and nearest-neighbor interpolation depending on the magnification factor, so its preferential use is at iso-sampling. Second, its computation can be simplified through the use of the convolution of 0-order B-splines with respect to the complex logic of sorting the edges of the voxels and pixels on an intermediate axis. Third, it can be improved by slightly higher-order kernels, and it can be associated with rebinning and rectification within a single modeling framework.

Improvements in clinical systems are expected from our framework. First, the increase in resolution achieved on a real acquisition of a quality assurance phantom will translate to clinical exams for linear AR. Secondly, regarding IR, let us recall that working with finer voxels than that of iso-sampling increases the computation load. Instead, a higher-order model for the volume side allows the compression of the information held by many small pixels on the detector side. We consider it important to provide an optimized link between a 0-order acquisition sampling model and a higher-order reconstruction model through either resizing or rectification, to perform reconstruction at isosampling. Accordingly, through the use of operators with increased symmetry, possibly based on virtual detectors where the resampled data can still be modeled as uncorrelated, our method displays features necessary for faster convergence and computation, much desirable in clinical practice.

6.7 Conclusion

In this chapter, we introduced a new *magnification-driven* interpolation framework for tomography. It leverages a resizing scheme based on families of B-splines of varying widths to account for the magnifications introduced by the homographies found in flat-panel cone-beam projection. Focusing on magnification was a key path toward improving the modeling of a cone-beam projector and its adjoint. An interpolation kernel set was derived that allows novel forward and backward projection pairs. These kernels balance spatial resolution versus noise and yield better noise uniformity. The benefits with respect to standard FP/BP models based on linear interpolation appeared more significant when downsampling frames acquired by the small pixels of X-ray flat-panel detectors: full dose usage is guaranteed, while linear interpolation randomly misses data. Magnification-driven interpolation is perfectly adapted to downsampling high-resolution data at the detector level through simple magnification or rectification that further provides simpler and faster computations. In our experiments, the tested kernels were of order up to 3. Such a choice results in reasonable computations, which, by taking advantage of separability, translate well to highly-parallel computing architectures. The lowest order model reduces to the DD interpolation, for which we provided new insight and computational scheme.

7 | Decomposition method for DTV with application to needle reconstruction from limited-angle acquisitions

7.1 Introduction

As mentioned in Chapter 2, flat-panel C-arm systems provide real-time 2D imaging to navigate therapeutic devices during minimally invasive vascular or percutaneous procedures. This chapter investigates iterative reconstruction methods in the context of interventions involving metallic needles.
During such interventions, the patient is positioned to optimize the real-time 2D visualization of the device and its trajectory. Although a CBCT scan could be performed to precisely assess the position of the device according to the planned trajectory, repeated acquisitions would increase the X-ray dose received by the patient. It would also require multiple repositionings of the patient because of kinematic constraints - due to the patient size or additional medical equipment - that are incompatible with a 200° short-scan rotation. An alternative to 2D imaging could be rotating over a smaller angular amplitude to reduce the number of required position changes and the number of projections, and thus patient dose.
In this chapter, we consider a series of acquisitions, the first of which is acquired around a 200° rotation. The first acquisition enables a precise reconstruction of the anatomical tissues, while subsequent shorter acquisitions focus on monitoring the needles. We assume that the fusion of the first pre-operative reconstruction with subsequent reconstructions is possible through registration techniques and allows the visualization of both the needles and the anatomy. As a result, we solely focus on reconstructing needles over a background from limited-angle acquisitions shorter than 200°.
Iterative reconstruction methods based on TV regularization have proven useful to reconstruct both interventional devices [74], and human tissues from a reduced number of projections [130]. However, these methods fail when reducing the angular amplitude of the acquisition. In 2D, the original isotropic TV, proposed in [187], penalizes the sum of the ℓ_2 norms of the image partial derivatives in the vertical and horizontal directions equally. It cannot recover the edges along directions not sampled by the limited angular amplitude. Only edges and details tangent to the projection directions are recovered [178]. For piecewise constant geometrical objects, successful results have been obtained with the anisotropic total variation (ATV). ATV assigns different weights to the image's vertical and horizontal partial derivatives. This strategy allows for considering the angular range as an additional prior information [225]. Non-convex potentials approximating the ℓ_0 pseudo-norm, combined with reweighting strategies, have also been explored within the ATV approach. However, due to non-convexity, it is not clear that

the resulting optimization methods converge to a global minimum [225, 232, 237]. Recently, ATV constrained formulations (instead of regularization-based ones) have been successfully used for the reconstruction of complex patterns from limited-angle acquisitions [244]. The anisotropic regularizer proposed in [17] is particularly suited to thin objects like needles because it emphasizes one specific direction. Hereinafter, we call this directional total variation (DTV). DTV relies on estimating the gradient norm along one selected direction that is not necessarily aligned with the pixel grid. Applications on denoising and reconstructing images of fiber materials have been successful [126]. Additionally, DTV with a spatially varying direction and strength [149,166,242], that includes higher-order derivatives [126, 165], has been proposed for applications involving vessels and fingerprints, which exhibit more complex directional patterns. In this chapter, we consider the simple geometric shape of needles that is very sparse and can be reconstructed from a limited-angle acquisition. We here allow for more than one direction for needles and assume a superposition of them over a non-sparse anatomical background. We adopt an image decomposition approach that applies DTV over multiple directions for needles and TV to approximate the background. Decomposition was first proposed for texture-geometry decomposition [12] and has also been applied to CT imaging to decompose the reconstruction into three components of the object, sub-sampling artifacts and noise [133].
This chapter is organized as follows. In section 7.2, we propose an iterative reconstruction method in 2D. Because of their simpler geometry, needles benefit from a directional *a priori* that can be easily embedded in iterative methods. In particular, we adapt the 2D DTV regularization of [17] for incorporating directional information and the decomposition method to apply different directional constraints on separate components selectively, as well as to exclude the anatomical background. Then, in section 7.3, we show how we can modify the 2D method to be used in 3D cone-beam geometry with a low computational cost.

7.2 Two-dimensional case with a 2D regularization

7.2.1 Method

Let $H \in \mathbb{R}^{M\times N}$ be the discretized model of forward projection adapted to a limited-angle acquisition, $x \in \mathbb{R}^N$ the unknown attenuation image, and $y \in \mathbb{R}^M$ the log-transform of the data measured by the detector. To estimate the attenuation map, we consider the sum of a least-squares data fidelity term and a convex regularizer g embedding this prior information, in particular, sparsity and direction:

$$\underset{x\in\mathbb{R}^N}{\text{minimize}} \ \frac{1}{2}\|y - Hx\|_F^2 + g(x), \tag{7.1}$$

where $F \in \mathbb{R}^{M\times M}$ is a symmetric positive definite matrix. We now discuss choices for g.

Directional total variation

DTV enforces the prior that the object is piecewise constant and follows one main direction. For an image $x \in \mathbb{R}^N$, its DTV can be defined as

$$\mathrm{DTV}_\Omega(x) = \sum_{n=1}^{N} \|(\nabla_\Omega x)_n\|_1 = \|\Lambda R_\theta(\nabla x)\|_{1,1}, \tag{7.2}$$

where $\Omega = \{\theta, s\}$, $\nabla_\Omega \in \mathbb{R}^{2\times N}$ contains two directional derivatives at pixel n, $\Delta_n^\theta x$ and $\Delta_n^{\theta+\pi/2} x$, parameterized by an angle $\theta \in [0°, 180°[$, and a "stretching factor" $s > 0$ for anisotropy, i.e.

$$
\begin{aligned}
(\nabla_\Omega x)_n &= \begin{pmatrix} \Delta_n^\theta x \\ s\Delta_n^{\theta+\pi/2} x \end{pmatrix} = \Lambda R_\theta \begin{pmatrix} \Delta_n^{\mathrm{h}} x \\ \Delta_n^{\mathrm{v}} x \end{pmatrix} \\
&= \begin{pmatrix} 1 & 0 \\ 0 & s \end{pmatrix} \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \Delta_n^{\mathrm{h}} x \\ \Delta_n^{\mathrm{v}} x \end{pmatrix},
\end{aligned} \tag{7.3}
$$

with $\Delta_n^{\mathrm{h}} \in \mathbb{R}^N$, $\Delta_n^{\mathrm{v}} \in \mathbb{R}^N$ the horizontal and vertical discrete gradient operators at location n respectively. The discrete gradient operators can be obtained by applying a forward finite difference scheme with zero boundary conditions.
Given that a set of needles makes a very sparse image, we add an ℓ_1 penalty so that $g(x) = g_\Omega(x) = \rho \mathrm{DTV}_\Omega(x) + \alpha\|x\|_1 + i_{[0,+\infty[^N}(x)$, $(\alpha, \rho) \in]0, +\infty[^2$ in (7.1).

Image decomposition

With a non-sparse background, the lack of data due to the low angular amplitude cannot be compensated, and the problem does not have a sparse solution. Decomposing x into a linear combination of several components restores sparsity in all components that can then be recovered from the limited data. A component is thus defined by its specific sparsity prior. A different sparse approximation is used to direct the interfering background into a single component. Then, instead of estimating the sum directly, we simultaneously solve the minimization problem for each of these components.
Here we decompose x into the anatomical background component x_{B}, penalized with TV, and $I \in \mathbb{N}$ directional components $x_{\Omega_{\mathrm{i}}}$, penalized with DTV of direction $\theta_i \in [0°, 180°[$ and stretching parameter $s_i > 0$ such that

$$
x = x_{\mathrm{B}} + \sum_{i=1}^{I} x_{\Omega_i}, \tag{7.4}
$$

where $\Omega_i = \{\theta_i, s_i\}$, $i \in \{1, \ldots, I\}$.
Altogether, we must solve the following convex problem:

$$
\underset{x_{\mathrm{B}}, (x_{\Omega_i})_{i=1}^I \in \mathbb{R}^N}{\text{minimize}} \ \frac{1}{2}\|y - H(x_{\mathrm{B}} + \sum_{i=1}^{I} x_{\Omega_i})\|_F^2 + \sum_{i=1}^{I} g_{\Omega_i}(x_{\Omega_i}) + g_{\mathrm{TV}}(x_{\mathrm{B}}), \tag{7.5}
$$

where $g_{\mathrm{TV}}(x_{\mathrm{B}}) = \beta\|\nabla x_{\mathrm{B}}\|_{1,2} + i_{[0,+\infty[^N}(x_{\mathrm{B}})$, $\beta \in]0, +\infty[$.
Note that each directional component x_{Ω_i} can capture a needle or a group of needles of about the same direction. The decomposition is expected because each regularization function captures features targeted to one component.

Optimization algorithm

We compare two optimization methods for solving (7.5): FISTA and CV, presented in Chapter 3 (subsection 3.3.3).

FISTA:

Let $z = \begin{bmatrix} x_{\mathrm{B}}^\top & x_{\Omega_1}^\top & \dots & x_{\Omega_I}^\top \end{bmatrix}^\top \in \mathbb{R}^{(I+1)N}$. Let $\tilde{H} = \Pi H$ and $\tilde{F} = \Pi F$ where $\Pi \colon M \mapsto \begin{bmatrix} M & \dots & M \end{bmatrix} \in \mathbb{R}^{((I+1)L)\times L}$. A first reformulation of (7.5) is

$$\underset{z\in\mathbb{R}^{(I+1)N}}{\text{minimize}} \ \frac{1}{2}\|y - \tilde{H}z\|_{\tilde{F}}^2 + h(z), \tag{7.6}$$

with $h \colon z \mapsto \sum_i^I g_{\Omega_i}(x_{\Omega_i}) + g_{\mathrm{TV}}(x_{\mathrm{B}})$.
Let a be a positive real number such that $a > 2$. The k-th iteration of FISTA applied to (7.6) reads

$$\begin{cases} \beta_k = k/(k+1+a) \\ \tilde{z}^k = z^k + \beta_k(z^k - z^{k-1}) \\ z^{k+1} = \mathrm{prox}_{\tau h}(\tilde{z}^k - \tau\tilde{H}^\top\tilde{F}(\tilde{H}\tilde{z}^k - y)) \end{cases}, \tag{7.7}$$

where $z^0 \in \mathbb{R}^{(I+1)N}$.
Thanks to the separability of h in each component of z, we derive an update rule for each map as

$$\begin{cases} \beta_k = k/(k+1+a) \\ \tilde{x}_{\mathrm{B}}^k = x_{\mathrm{B}}^k + \beta_k(x_{\mathrm{B}}^k - x_{\mathrm{B}}^{k-1}) \\ \text{For } i \in \{1,\dots,I\}\text{:} \\ \qquad \tilde{x}_{\Omega_i}^k = x_{\Omega_i}^k + \beta_k(x_{\Omega_i}^k - x_{\Omega_i}^{k-1}) \\ \tilde{x}^k = H^\top F\left(H(\tilde{x}_{\mathrm{B}}^k + \sum_i^I \tilde{x}_i^k) - y\right) \\ x_{\mathrm{B}}^{k+1} = \mathrm{prox}_{\tau g_{\mathrm{TV}}}(\tilde{x}_{\mathrm{B}}^k - \tau\tilde{x}^k) \\ \text{For } i \in \{1,\dots,I\}\text{:} \\ \qquad x_{\Omega_i}^{k+1} = \mathrm{prox}_{\tau g_{\Omega_i}}(\tilde{x}_{\Omega_i}^k - \tau\tilde{x}^k) \end{cases}. \tag{7.8}$$

The sequence $(x_{\mathrm{B}}^{k+1}, (\tilde{x}_{\Omega_i}^k)_{i=1}^I)_{k\in\mathbb{N}}$ produced by Algorithm (7.8) converges to a solution to (7.5) for

$$0 < \tau \leqslant \frac{1}{|||\tilde{H}^\top\tilde{F}\tilde{H}|||} = \frac{1}{(I+1)|||H^\top F H|||}. \tag{7.9}$$

The proximity operators of g_{TV} and g_{Ω_i} do not have a closed form, hence they are both approximated using inner iterations of the DFB algorithm, presented in Chapter 3 (section 3.3), with warm-restart. In particular, for DTV and $\tilde{x} \in \mathbb{R}^N$, $\hat{x} = \mathrm{prox}_{\tau g_{\Omega_i}}(\tilde{x})$ is estimated using the following sub-iteration:

$$\begin{cases} x^n = \mathrm{proj}_{[0,+\infty[^N}(\tilde{x} - \nabla_\Omega^\top u^n) \\ u^{n+1} = \mathrm{proj}_{\|\cdot\|_{\infty,\infty}\leqslant\tau\rho}(u^n + \gamma\nabla_\Omega x^n) \end{cases}, \tag{7.10}$$

where $\gamma < 2/|||\nabla_\Omega|||^2$, $u^0 \in \mathbb{R}^{2N}$ and $\|u^n\|_{\infty,\infty}$ is the maximum value of the 2 components of $u^n \in \mathbb{R}^{2N}$.

CV:

A second way of reformulating (7.5) leads to

$$\underset{z\in\mathbb{R}^{(I+1)N}}{\operatorname{argmin}} \; g(z) + f(Kz) + \frac{1}{2}\|\tilde{H}z - y\|_F^2, \tag{7.11}$$

where

$$\begin{aligned}
&(\forall z = \begin{bmatrix} x_{\mathrm{B}}^\top & x_{\Omega_1}^\top & \dots & x_{\Omega_I}^\top \end{bmatrix}^\top \in \mathbb{R}^{(I+1)N}) \qquad g(z) = i_{[0,+\infty[^{(I+1)N}}(z) + \sum_{i=1}^{I} \alpha \|x_{\Omega_i}\|_1, \\
&(\forall u = \begin{bmatrix} u_{\mathrm{B}}^\top & u_{\Omega_1}^\top & \dots & u_{\Omega_I}^\top \end{bmatrix}^\top \in \mathbb{R}^{(I+1)2N}) \qquad f(u) = \beta\|u_{\mathrm{B}}\|_{1,2} + \sum_{i=1}^{I} \rho \|u_{\Omega_i}\|_{1,1}, \\
&K = \begin{bmatrix} \nabla & \nabla_{\Omega_1} & \dots & \nabla_{\Omega_I} \end{bmatrix}.
\end{aligned} \tag{7.12}$$

The k-th iteration of CV applied to (7.11) reads

$$\begin{cases} z^{k+1} = \operatorname{prox}_{\tau g}(z^k - \tau(\tilde{H}^\top F(\tilde{H}z^k - y) + K^* u^k)) \\ u^{k+1} = \operatorname{prox}_{\sigma f^*}(u^k + \sigma K(2z^{k+1} - z)) \end{cases}, \tag{7.13}$$

where $z^0 \in \mathbb{R}^{(I+1)N}$ and $u^0 \in \mathbb{R}^{2(I+1)N}$.
The convergence of Algorithm (7.13) is guaranteed for $\tau, \sigma > 0$ such that

$$\tau\left(\sigma|||K|||^2 + |||\tilde{H}^\top \tilde{F}\tilde{H}|||\right) < 1, \tag{7.14}$$

where we note that

$$\sigma|||K|||^2 + |||\tilde{H}^\top \tilde{F}\tilde{H}||| \leqslant \sigma(|||\nabla|||^2 + I|||\nabla_{\Omega_1}|||^2) + (I+1)|||H^\top F H|||. \tag{7.15}$$

Component-wise, the update (7.13) becomes

$$\begin{cases}
v^{k+1} = H^\top F\left(H(x_{\mathrm{B}}^k + \sum_i^I x_i^k) - y\right) \\
x_{\mathrm{B}}^{k+1} = \operatorname{proj}_{[0,+\infty[^N}(x_{\mathrm{B}}^k - \tau(v^{k+1} + \nabla^\top u_B^k)) \\
\text{For } i \in \{1,\dots,I\}: \\
\qquad x_{\mathrm{i}}^{k+1} = \operatorname{proj}_{[0,+\infty[^N}(\operatorname{prox}_{\tau\alpha\|\cdot\|_1}(x_{\mathrm{i}}^k - \tau(v^{k+1} + \nabla_{\Omega_i}^\top u_{\theta_i}^k))) \\
v^{k+1} = \nabla(2x_{\mathrm{B}}^{k+1} - x_{\mathrm{B}}^k) + \sum_i^I \nabla_{\Omega_i}(2x_{\mathrm{i}}^{k+1} - x_{\mathrm{i}}^k) \\
u_{\mathrm{B}}^{k+1} = \operatorname{proj}_{\|\cdot\|_\infty \leqslant \beta}(u_{\mathrm{B}}^k + \sigma v^{k+1}) \\
\text{For } i \in \{1,\dots,I\}: \\
\qquad u_{\mathrm{i}}^{k+1} = \operatorname{proj}_{\|\cdot\|_\infty \leqslant \rho}(u_{\mathrm{i}}^k + \sigma v^{k+1})
\end{cases} \tag{7.16}$$

Contrary to Algorithm (7.8), Algorithm (7.16) does not require sub-iterations.

7.2.2 Application

Simulations

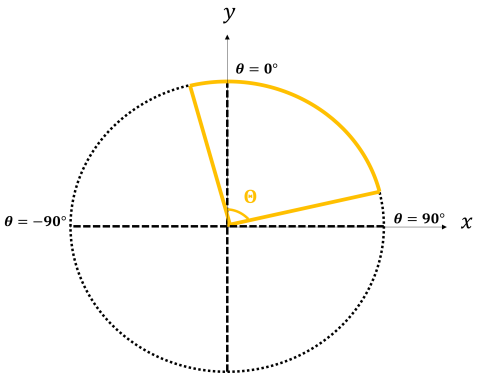


Figure 7.1: A scanning configuration collecting data over limited angular range Θ.

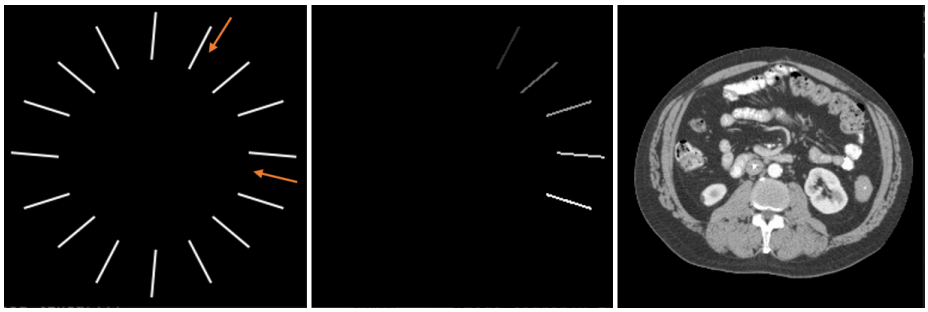


Figure 7.2: Reference images. From left to right: Phantom (A) with needles of intensity 3500 sHU, Image with needles of growing intensity from 3000 sHU up to 5000 sHU, Anatomical background [1800-2200 sHU].

We carried out simulations in parallel geometry (Figure 7.1): an X-ray source and a 1D detector rotate over a circular arc Θ.
We considered two numerical phantoms on a 256×256 grid. Phantom (A) is purely geometric and represents a set of needles of intensity 3500 sHU covering 8 directions (5°, 27.5°, 50°, 72.5°, 95°, 107.5°, 130°, 152.5°) as shown in Figure 7.2. Phantom (B) is the sum of an axial CT slice of an abdomen (see Figure 7.2) with a subset of needles of varying intensity (3000-5000 sHU).
A needle is said to be within the scanning arc if the projection data contain its bull's eye view. We computed simulated data of these phantoms over a circular arc of amplitude $\theta \in [29°, 95°]$ (indicated by the arrows in Figure 7.2) so that the projection data contains the bull's eye view of three needles. I.i.d. Gaussian noise of mean 0 and standard deviation 50 was then added to the projections. The angular sampling was uniform with a step of 2°. Reconstruction with TV regularization was taken as a baseline. FBP reconstruction followed by thresholding of the intensity was added to the comparison.
First, we analyzed the performance of our DTV decomposition method (7.5) for the reconstruction of a subset of the needles of Phantom (A). We performed a reconstruction using four DTV of direction $\{5°, 27.5°, 72.5°, 107.5°\}$ (i.e., $I = 4$). Then we reduced the angular density. Finally, we added two directional components ($I = 6$) $\{130°, 152.5°\}$ to model (7.5) so that all the directions of Phantom (A) were covered. We compared the

convergence of CV and FISTA (with 100 inner iterations of DFB). Then we showed the method's applicability to the more complex case of background and needles of different intensities by reconstructing Phantom (B). This time, a set of $I = 3$ directions was used: $\{27.5°, 72.5°, 107.5°\}$. In all these simulations, the needles have the same size, so we used the same stretching parameter $s = 0.001$. Needles with the same intensity have the same regularization parameters ρ and α. The matrix F was chosen as the ramp filter (see Chapter 3), which provides an approximate inversion of $HH^\top$. The TV and DTV parameters then became thresholds that are homogeneous to the sHU intensity values of the image.

Background-free needles

Figure 7.3 shows the reconstructions of Phantom (A) with FBP, (7.1) with TV ($\beta = 50$) and (7.5) with DTV decomposition ($\rho = 50$, $\alpha = 1$). First, we see that with FBP, only a partial reconstruction of the three needles within the scanning arc is achieved. Figure 7.4 displays the four reconstructed directional components. Both DTV and TV methods lead to similar reconstructions for the three needles in the scanning arc. For the two needles of direction close to the starting value of Θ, TV yields a partial recovery only. In contrast, the DTV decomposition method fully recovers 12 out of 16 needles because their directions were sufficiently close to the imposed *a priori* directions. The four remaining missing needles show no recovery without *a priori* directional information. Reconstruction of the 12 needles is the same with the DTV decomposition method when doubling the angular sampling step, which shows the robustness of the method to a varying angular density (see Figure 7.5).
Figure 7.6 shows that the last four needles can be recovered by adding two more directional components ($I = 6$) to (7.5).
Regarding the comparison of the performance of our two optimization algorithms for the DTV decomposition problem, Figure 7.7 displays the evolution of the PSNR associated with the iterates of (7.8) and (7.16). We see that FISTA converges faster than CV in terms of the number of iterations, as expected by the improved convergence rate of FISTA. Figure 7.8 shows the estimated directional maps for the case $I = 4$ at 2500 and 5000 iterations with FISTA, while Figure 7.9 shows the same maps obtained after 2500, 5000, and 50000 iterations of CV. We see that the directional maps are the same for CV after 5000 iterations while they are already well-estimated after 2500 iterations of FISTA.

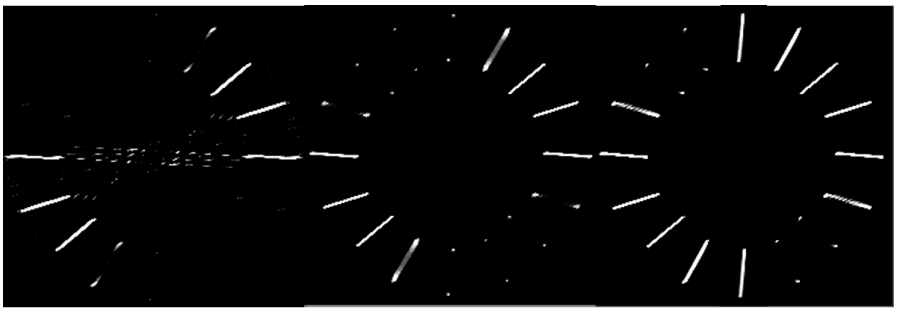


Figure 7.3: Reconstructed images for $\theta \in [29°, 95°]$. From left to right: FBP, (7.1) with TV and (7.5) with DTV decomposition ($I = 4$).

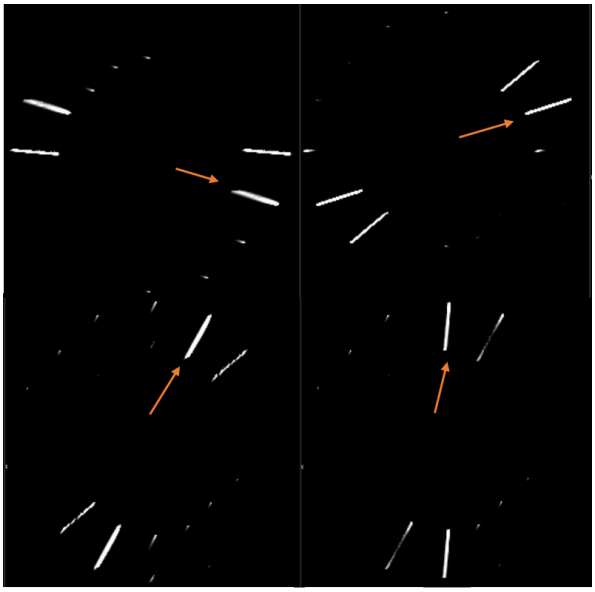


Figure 7.4: Directional components obtained with the DTV decomposition method ($I = 4$) for $\theta \in [29°, 95°]$ on Phantom (A). Top: from left to right, $\theta_1 = 107.5°$, $\theta_2 = 72.5°$. Bottom: from left to right, $\theta_3 = 27.5°$, $\theta_4 = 5°$.

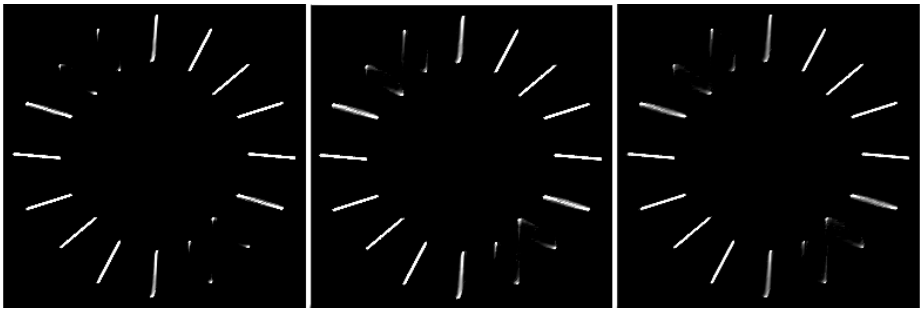


Figure 7.5: Reconstructions obtained with the DTV decomposition method ($I = 4$) for $\theta \in [29°, 95°]$ when decreasing the angular density. Sum of all directional components obtained with the DTV decomposition method on Phantom (A) for an angular step of, from left to right, 2°, 3° and 4°.

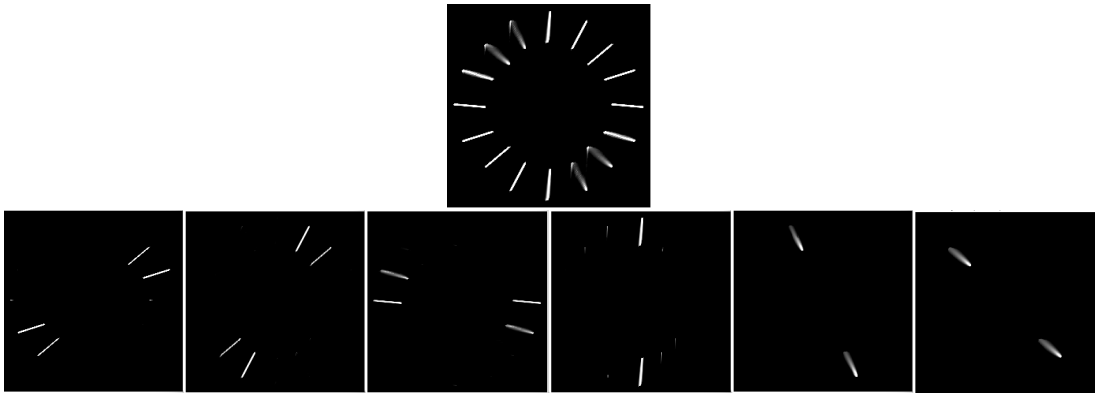


Figure 7.6: Reconstruction obtained with the DTV decomposition method for $\theta \in [29°, 95°]$ when using a directional component for all possible directions ($I = 6$). Top: sum of all reconstructed components. Bottom: directional components obtained with the DTV decomposition method on Phantom (A). From left to right, $\theta_1 = 107.5°$, $\theta_2 = 72.5°$, $\theta_3 = 27.5°$, $\theta_4 = 5°$, $\theta_5 = 152.5°$, $\theta_4 = 130°$.

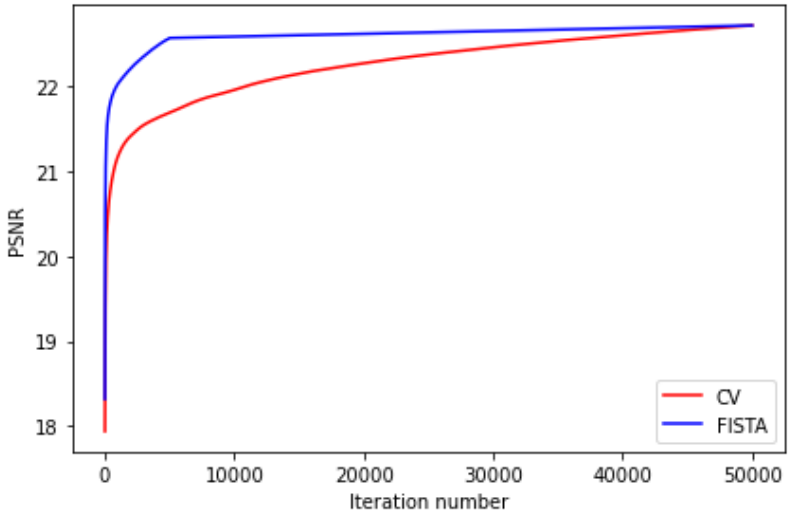


Figure 7.7: PSNR associated to the sum of all reconstructed maps produced by FISTA and CV with respect to Phantom (A) for $\theta \in [29°, 95°]$ and $I = 4$.

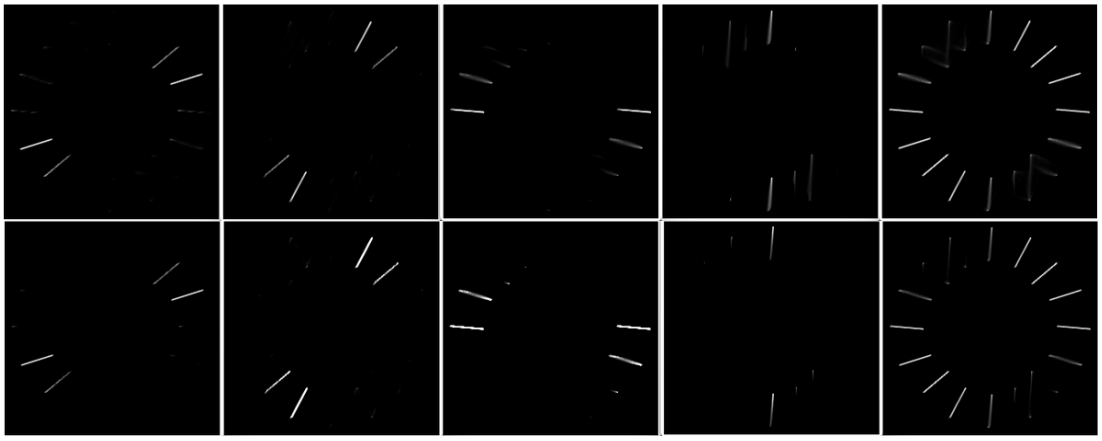


Figure 7.8: Estimates of the components of Phantom (A) for $\theta \in [29°, 95°]$ with FISTA and $I = 4$. From top to bottom: 2500, 5000 iterations. From left to right, $\theta_1 = 107.5°$, $\theta_2 = 72.5°$, $\theta_3 = 27.5°$, $\theta_4 = 5°$, sum of all components.

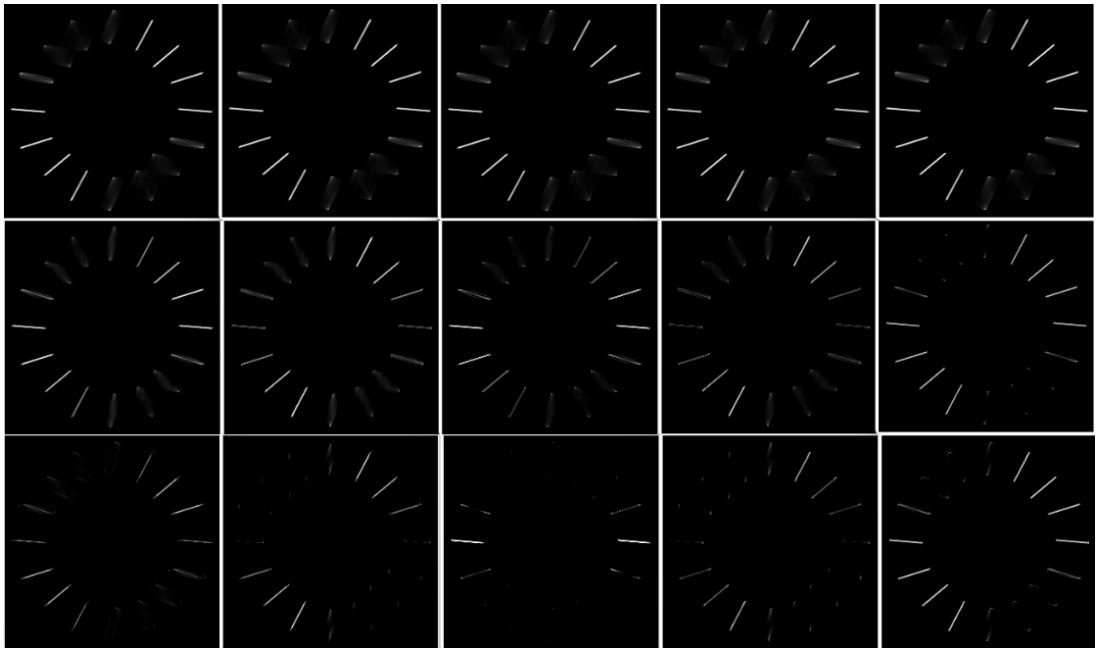


Figure 7.9: Estimates of the components of Phantom (A) for $\theta \in [29°, 95°]$ with CV and $I = 4$. From top to bottom: 2500, 5000, and 50000 iterations. From left to right, $\theta_1 = 107.5°$, $\theta_2 = 72.5°$, $\theta_3 = 27.5°$, $\theta_4 = 5°$, sum of all components.

Needles with background

Figure 7.10 shows the reconstruction of the needles of Phantom (B) with FBP, (7.1) with TV and (7.5) with DTV decomposition (sum of all needle maps). First, as expected, the anatomical background cannot be recovered with sufficient quality with a sparse prior in this limited angle setting. The needles reconstructed with FBP are distorted, and the intensity values are not retrieved. With a simple TV regularization, only the three needles within the scanning arc remain after thresholding, whereas five are recovered with the DTV decomposition method. Figure 7.11 shows that the decomposition method coupled with directional information separates the three sets of needles from the background map.

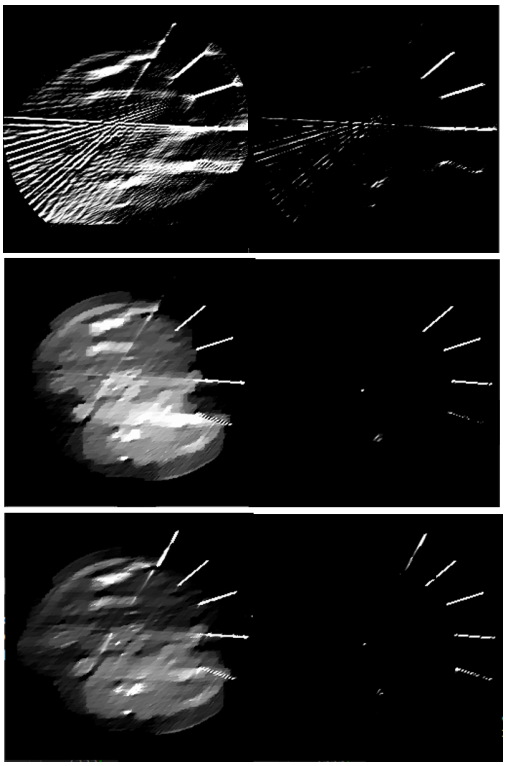


Figure 7.10: Reconstruction of needles of Phantom (B) in the presence of a background for $\theta \in [29°, 95°]$. Left: needles with background. Right: needles map. From top to bottom: FBP, (7.1) with TV and (7.5) with DTV decomposition.

Figure 7.11: Reconstructed needles maps obtained with the DTV decomposition method on Phantom (B). From left to right: x_{Ω_1} ($\theta_1 = 107.5°$), x_{Ω_2} ($\theta_2 = 27.5°$), x_{Ω_3} ($\theta_3 = 72.5°$).

7.3 Three-dimensional case with a 1D regularization

7.3.1 Method

In cone-beam geometry, applying a 2D DTV within each axial slice does not capture the main direction of the needles when they deviate from the tomographic plane. We now extend the previous model to this case.

Limitations of the 2D DTV formulation

In (7.3), the gradient in an arbitrary direction θ, ∇_θ is estimated from a linear combination of the vertical Δ^{v} and horizontal Δ^{h} differences operators. This estimation has two main limitations.
First, a straightforward extension of the previous DTV decomposition method to an arbitrary direction in a 3D volume includes three finite difference operators along with the three directions of the reference frame. Using a 3D formulation of the DTV regularization to enhance a 1D structure is computationally heavy.
Second, the estimation of the directional gradient is poor at $\theta \in \{\pm 45°\}$ while it is the most precise at $\theta \in \{0°, 90°\}$, i.e., along the axis of the Cartesian grid. Such poor discretization indicates that the rate of recovery of a needle by our optimization algorithms solving (7.5) depends on the needle's direction. We illustrate this point by considering two different starting and ending angular positions to acquire the projections of Phantom (A) while keeping the angular amplitude fixed. More precisely, we choose $\theta \in [-35°, 15°]$ (Case 1) and $\theta \in [35°, 85°]$ (Case 2). Similar to our previous experiments, we computed simulated projections and added i.i.d Gaussian noise of mean 0 and standard deviation 50. This time, we focused on the reconstruction of two maps: a map containing all well-sampled needles ($\theta_1 = -14°$ for Case 1 and $\theta_1 = 61°$ for Case 2) and a map containing a needle with a direction close to the end of the scanning arc ($\theta_2 = -50°$ for Case 1 and $\theta_2 = 95°$ for Case 2).
Figure 7.12 shows the reconstruction of the two directional maps for both acquisitions after 2500 iterations of FISTA. We see that the needle whose direction is closest to that of the grid is recovered before the needle whose direction is closest to a diagonal.

Building on these limitations, we propose to change the reconstruction problem (7.5) so that the directional filtering is enabled by a 1D regularization.

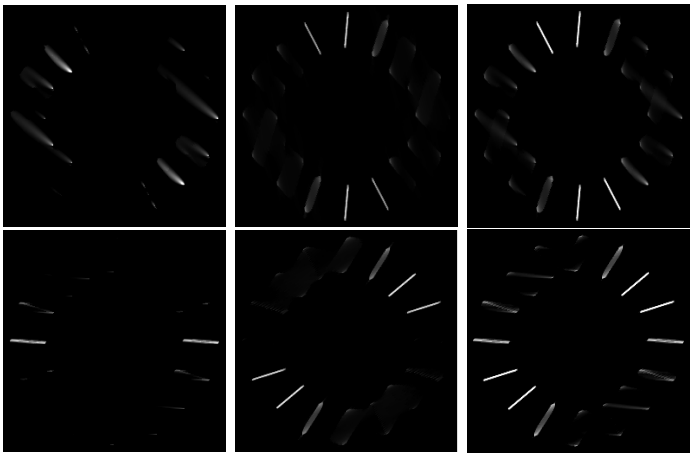


Figure 7.12: Two reconstructed needle maps of Phantom (B) for different scanning trajectories with a 2D DTV regularization. Top: $\theta \in [-35°, 15°]$. Bottom: $\theta \in [35°, 85°]$. From left to right: component of direction θ_1, component of direction θ_2, sum of the two components.

Changing the reconstruction grid for each directional component

In Chapter 6 (section 6.2), we have seen that the cone-beam geometry can be described by a set of 3×4 projection matrices. More specifically, the matrix of extrinsic parameters $\mathbf{P}_\mathrm{e}$ relates the reconstruction frame to a 3×3 rotation matrix with coefficients $(r_{i,j})_{1\leqslant i,j\leqslant 3}$ followed by three translations $(t_x,\, t_y,\, t_z) \in \mathbb{R}^3$, one along each rotated axis to reach the position of the source of the acquisition system with respect to the world coordinate system:

$$\mathbf{P}_\mathrm{e} = \begin{pmatrix} r_{1,1} & r_{1,2} & r_{1,3} & t_x \\ r_{2,1} & r_{2,2} & r_{2,3} & t_y \\ r_{3,1} & r_{3,2} & r_{3,3} & t_z \end{pmatrix}. \tag{7.17}$$

A rotation of the reconstruction frame around the rotated $x-$axis is performed by right-multiplying $\mathbf{P}_e$ with a 4×4 matrix such that

$$\mathbf{R}_{\theta_1} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\theta_1 & -\sin\theta_1 & 0 \\ 0 & \sin\theta_1 & \cos\theta_1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{7.18}$$

Likewise a rotation around the rotated y- and z- axis is achieved using the matrices

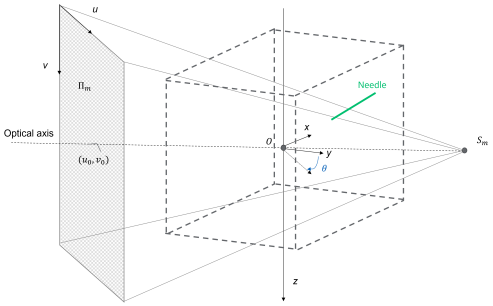
$$\mathbf{R}_{\theta_2} = \begin{pmatrix} \cos\theta_2 & 0 & \sin\theta_2 & 0 \\ 0 & 1 & 0 & 0 \\ -\sin\theta_2 & 0 & \cos\theta_2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{7.19}$$

and

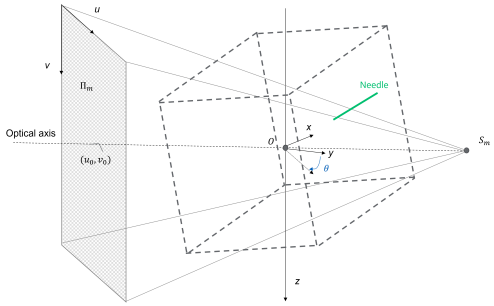
$$\mathbf{R}_{\theta_3} = \begin{pmatrix} \cos\theta_3 & \sin\theta_3 & 0 & 0 \\ \sin\theta_3 & \cos\theta_3 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{7.20}$$

We now introduce different projectors associated to rotated reconstruction grids. Let $\Omega_i = \{\theta_1^i, \theta_2^i, \theta_3^i\}$. H is the projector obtained from $\mathbf{P}_e$ while H_{Ω_i} is the projector obtained from $\mathbf{P}_e\mathbf{R}_{\theta_1^i}\mathbf{R}_{\theta_2^i}\mathbf{R}_{\theta_3^i}$. We define different projectors $(H_{\Omega_i})_{i=1}^I \in \mathbb{R}^{I(M\times N)}$ for each

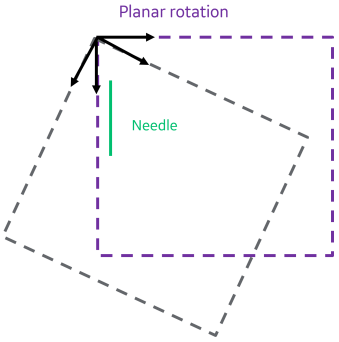
directional component. The set of angles Ω_i is selected such that the ratio between the sampling step along each direction θ_j^i ($j \in \{1, 2, 3\}$) and the sampling step along the new vertical or horizontal direction is closest to one. Ω_i then contains the minimal rotations required to align the needle map with one of the grid's axis. The pipeline is summarized in Figure 7.13.



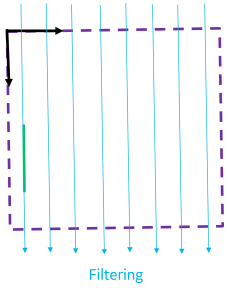
(a) Initial C-arm CBCT geometry ($\mathbf{P}_e$)



(b) 1st rotation of the reconstruction frame



(c) 2nd rotation to align the needle with the vertical axis



(d) The proximity operator of g_{1D} performing a 1D filtering

Figure 7.13: Pipeline for replacing DTV regularization by a 1D directional regularization. Two rotations are performed to change the reconstruction frame. The first rotation makes the needle parallel to a plane of type $z = z_0$ and the second aligns the needle with the axis y of the new frame.

We can then consider the following optimization problem:

$$\underset{x_{\mathrm{B}},(x_{\Omega_i})_{i=1}^I \in \mathbb{R}^N}{\text{minimize}} \ \frac{1}{2}\|y - Hx_{\mathrm{B}} - \sum_{i=1}^{I} H_{\Omega_i} x_{\Omega_i}\|_F^2 + \sum_{i=1}^{I} g_{1D}(x_{\Omega_i}) + g_{\mathrm{TV}}(x_{\mathrm{B}}), \tag{7.21}$$

where, for $x \in \mathbb{R}^N$, $g_{1D}(x) = i_{[0,+\infty[^N}(x) + \alpha\|x\|_1 + \rho\|\Delta^v x_{\omega_i}\|_1$ if x follows the vertical direction in the new reconstruction frame or $g_{1D}(x) = i_{[0,+\infty[^N}(x) + \alpha\|x\|_1 + \rho\|\Delta^h x_{\omega_i}\|_1$ if x follows the horizontal direction.

Optimization algorithm

The FISTA algorithm now yields

$$\begin{cases} \beta_k = k/(k+1+a) \\ \tilde{x}_{\mathrm{B}}^k = x_{\mathrm{B}}^k + \beta_k(x_{\mathrm{B}}^k - x_{\mathrm{B}}^{k-1}) \\ \text{For } i \in \{1,\dots,I\}\text{:} \\ \qquad \tilde{x}_{\Omega_i}^k = x_{\Omega_i}^k + \beta_k(x_{\Omega_i}^k - x_{\Omega_i}^{k-1}) \\ y^k = F(H\tilde{x}_{\mathrm{B}}^k + \sum_i^I H^{\theta_i}\tilde{x}_i^k - y) \\ x_{\mathrm{B}}^{k+1} = \mathrm{prox}_{\tau g_{\mathrm{TV}}}(\tilde{x}_{\mathrm{B}}^k - \tau H^\top y^k) \\ \text{For } i \in \{1,\dots,I\}\text{:} \\ \qquad x_{\Omega_i}^{k+1} = \mathrm{prox}_{\tau g_{1D}}(\tilde{x}_{\Omega_i}^k - \tau(H^{\theta_i})^\top y^k) \end{cases}, \tag{7.22}$$

where $x_{\mathrm{B}}^0 \in \mathbb{R}^N$ and $\forall i \in \{1,\dots,I\}$, $x_{\Omega_i}^0 \in \mathbb{R}^N$. Algorithm (7.22) trades a complex regularization scheme for a simpler and more precise scheme with additional forward projections and backprojections, one for each component (total of 2 in Algorithm (7.8) versus $2(I+1)$ in Algorithm (7.22)).

7.3.2 Application

Before testing our alternative DTV decomposition method (7.21) on 3D data, we provide a first illustration of its performance compared to that of the one relying upon a 2D DTV (7.5) on the simulated case presented in Figure 7.12 (subsection 7.3.1).
Figure 7.14 shows the equivalent reconstructed components obtained with (7.21) for Case 1 after 2500 iterations of FISTA. The first needle map along θ_1 is thus associated with the projector H_{Ω_1} where $\Omega_1 = \{0°, 0°, -\theta_1\}$ while the second is associated with the projector H_{Ω_2} where $\Omega_2 = \{0°, 0°, -\theta_2\}$. Both projectors align the map directions with the vertical axis (see Figure 7.14). Comparing the last column of Figure 7.12 (Top) and Figure 7.14, we see that pre-aligning the needle of direction θ_2 out of the scanning arc with the horizontal axis allows for a faster reconstruction than when the 2D DTV is used.

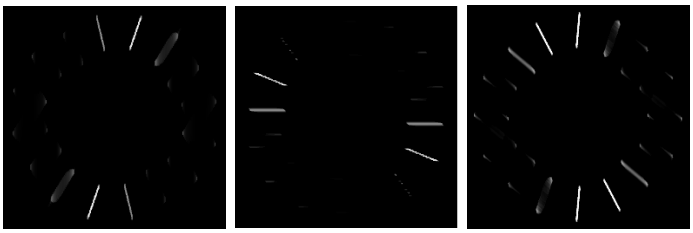


Figure 7.14: Reconstructed needle maps of Phantom (B) for $\theta \in [-35°, 15°]$ with a 1D regularization. From left to right: component of direction θ_1, component of direction θ_2, sum of rotated components.

We now demonstrate how our DTV decomposition method performs on two CBCT acquisitions for biopsies. Contrary to our previous simulations, noise is not the only source of data corruption in real data. Other physical effects, such as beam-hardening, are associated with metallic needles. Our DTV decomposition method was thus tested against these data inconsistencies.
We provided, as a comparison, the FDK reconstructions for $\theta \in [0°, 200°]$ and for limited-angle scans with a shorter angular amplitude. Note that, when $\theta \in [0°, 200°]$, Parker's weights were included in FDK. FDK reconstructions were performed on a grid of size $N = 512 \times 512 \times 512$. Hereinafter, the *a priori* set of directions for each needle map was estimated from a first FDK reconstruction using the complete acquired measurements over 200°.

First, we looked at a volume that contains a thick needle that belongs to the tomographic plane. Figure 7.15 shows a 2D view of the projections. We compared the reconstructions of the volume from three acquisitions: the first was such that $\theta \in [0°, 200°]$, the second was such that $\theta \in [20°, 123°]$, and the last was such that $\theta \in [97°, 200°]$. The angular density was constant for each acquisition (1 view every 0.68°).
We applied our DTV decomposition method for $I = 1$ for all three acquisitions with $\Omega_1 = \{\theta_1^1, \theta_2^1, \theta_3^1\} = \{0°, 0°, 36°\}$. The initialization of the needle map was chosen as the FDK reconstruction while the background map was initialized to zero. Contrary to our 2D simulations in subsection 7.2.2, the projections are now truncated. We thus reconstructed each component of the volume $x_B, x_{\Omega_1} \in \mathbb{R}^N$ on a slightly extended grid such that $N = 576 \times 576 \times 512$. We performed 300 iterations of FISTA and 80 inner iterations of DFB (with warm restart). Regularization parameters were chosen as $\beta = 5$, $\alpha = 10$ and $\rho = 50$.
Figure 7.16 shows the same transaxial slice where the needle appears in FDK reconstructions from the three acquisitions. We see that the intensity of the needle is distributed in a fan shape across the transaxial slice for $\theta \in [97°, 200°]$ and only the tip of the needle

Figure 7.15: 2D view of the projection data (Case 1).

is visible. In contrast, when the starting and ending angles are shifted by 77°, the needle is perfectly visible.

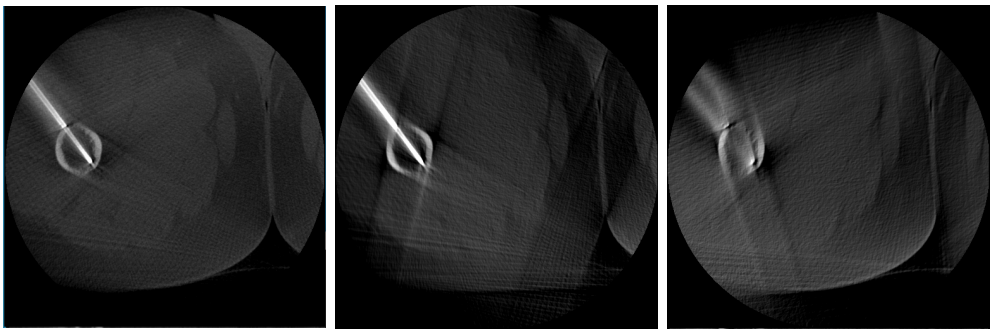


Figure 7.16: Transaxial slices obtained from FDK-reconstructed volumes for different acquisition amplitudes. From left to right: $\theta \in [0°, 200°]$, $\theta \in [20°, 123°]$, and $\theta \in [97°, 200°]$.

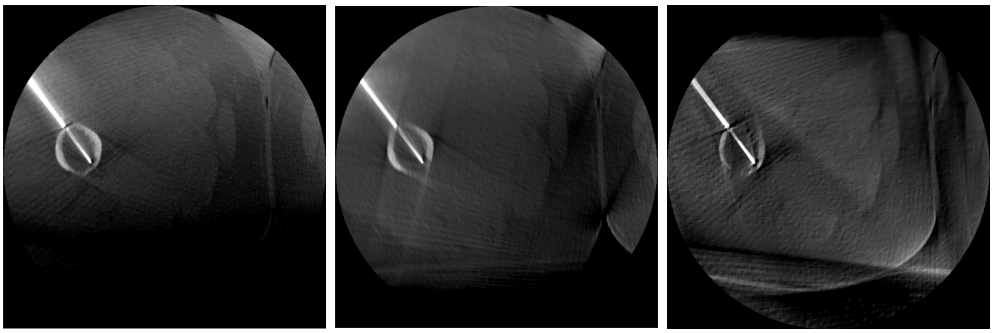


Figure 7.17: Transaxial slices obtained from reconstructions with our DTV decomposition method for the acquisitions considered in Figure 7.16.

Figure 7.17 shows the same slices obtained from the superposition of the rotated directional and background components given by the DTV decomposition method after 300 iterations. For the superposition to be possible, all components must be represented in the same reference frame. The reconstruction frame of the anatomical background component corresponds to the reference frame used in Figure 7.16. The directional components were therefore rotated so they could be summed with the background. We see that the needle is recovered for all angular ranges. Figure 7.18 shows the different reconstructed maps for the angular range $\theta \in [97°, 200°]$ which have been successfully

estimated and separated.

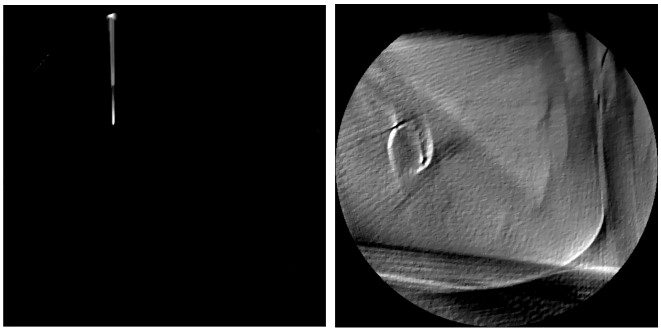


Figure 7.18: Reconstructed components obtained with our DTV decomposition method for $\theta \in [97°, 200°]$. From left to right: rotated needle map, background map.

We investigated a second case where three thin needles slightly deviate from the tomographic plane (see Figure 7.19).

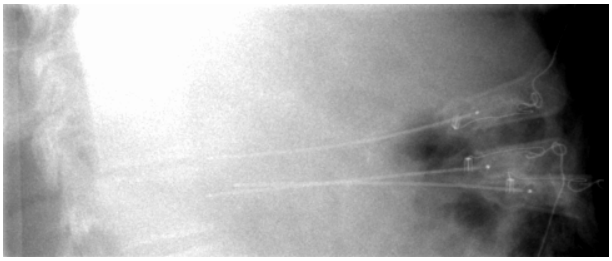


Figure 7.19: 2D view of the projection data (Case 2).

As in the first case, we show the transaxial slices where two needles appear (Figure 7.20) obtained using FDK with $\theta \in [0°, 200°]$, $\theta \in [27°, 82°]$, $\theta \in [86°, 200°]$. The maximum value of a voxel stack with a thickness of 10 voxels in the axial direction was taken. Again when the view minimizing the projected size of the needle is not sampled, we see that the needles cannot be reconstructed with FDK.
The needles are approximately parallel so we used a unique directional map ($I = 1$) with $\Omega_1 = \{\theta_1^1, \theta_2^1, \theta_3^1\} = \{0°, -6°, -38°\}$ for applying our DTV decomposition method. The size of the reconstruction grid, the initialization strategy, and the regularization parameters were the same as in the first case. We performed 1000 iterations of FISTA and 80 inner iterations of DFB. Figure 7.21 shows the reconstructions obtained with our method (the two components and their sum) for $\theta \in [0°, 200°]$, $\theta \in [27°, 82°]$, $\theta \in [86°, 200°]$. As expected, our method successfully interpolates along the direction of the needles. Figure 7.22 shows the different reconstructed maps for the less favorable angular range $\theta \in [97°, 200°]$.
Note that we used a higher number of iterations of FISTA in the second case than in the first case because the method had not converged after 300 iterations.

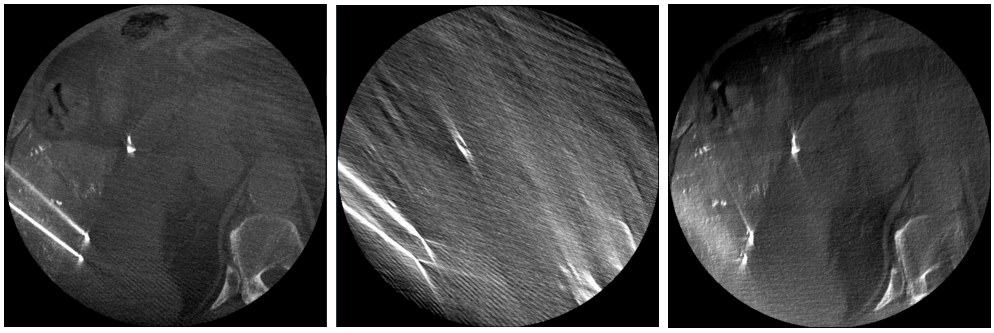


Figure 7.20: Transaxial slices (MIP with a thickness of 10 voxels) of three FDK-reconstructed volumes from different angular amplitudes and the same angular density. From left to right: $\theta \in [0°, 200°]$, $\theta \in [27°, 82°]$, $\theta \in [86°, 200°]$.

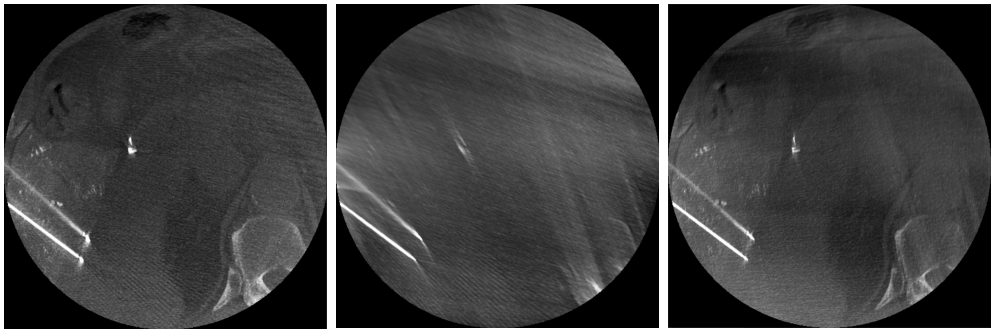


Figure 7.21: Transaxial slices (MIP with a thickness of 10 voxels) obtained from reconstructions with our DTV decomposition method for the acquisitions considered in Figure 7.20.

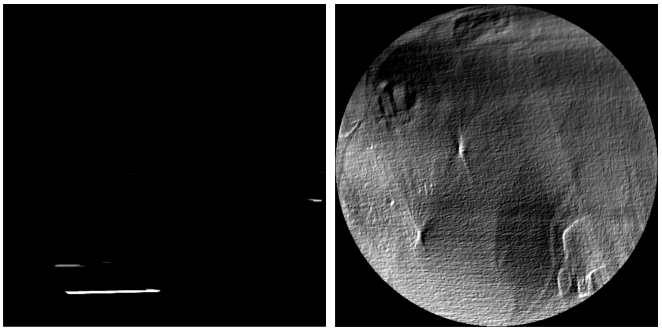


Figure 7.22: Reconstructed components obtained with our DTV decomposition method for $\theta \in [86°, 200°]$. From left to right: rotated needle map, background map.

7.4 Conclusion

In this chapter, we exploited the simple geometric shape of percutaneous needles, which is very sparse and follows one main direction for iterative reconstruction. We introduced a decomposition method that allows several DTV regularizations - and thus several directional *a priori* to be considered at once - and treats the anatomical background separately. The potential of our DTV decomposition method for reconstructing geometrical objects from small scanning arcs was confirmed in our results, where needles are recovered even when their bull's eye view is not sampled. We showed that the FISTA algorithm was more effective than the Condat-Vũ algorithm for solving our reconstruction problem. Note that an inertial variant of the Condat-Vũ algorithm has been proposed in [142]. Still, since one application of $H^\top FH$ takes significantly longer than one application of $\nabla^\top \nabla$, FISTA remains interesting in terms of computing time as it requires fewer forward and backward projections. We also demonstrated that the 2D formulation of the DTV could reduce to a 1D formulation when changing the reconstruction grid for each needle component through different rotations of the projection matrices associated with the projector. This reformulation is especially convenient for the cone-beam geometry as it is more computationally efficient and bypasses the angular sensitivity of the original DTV discretization. Our study highlighted that the benefit of DTV over TV depends on the direction of the needle with respect to the directions covered by the scanning arc. In the case where the most critical projections (around the bull's eye view) are collected, our method performed as well as TV. However, we argue that choosing a convenient scanning arc is not always easy. First, the trajectory may not be available on the system because it has not been pre-calibrated. Then, when the trajectory has been calibrated, various equipment surrounding the patient could collide with the C-arm, depending on the intervention. Our DTV decomposition method thus offers a robust alternative to TV, shifting the issue of developing a flexible calibration for the systems to that of computation time. To reduce the computational time of our method, fast non-iterative algorithms for computing the proximity of the 1D DTV regularization [65] should be investigated. Finally, we should also devise an automated way of roughly selecting the direction of each needle map from the projection data; the direction being here given as prior knowledge.

8 | Deep unfolding of the DBFB Algorithm with application to ROI imaging with limited angular density

8.1 Introduction

8.1.1 Challenges of ROI imaging

As outlined in Chapter 2, in most interventions involving CBCT imaging, only a small region of the patient is of interest. On a C-arm, the problem of reconstructing a ROI from a set of truncated projections is also combined with low angular sampling. In this chapter, we will consider the setup of using IR for reconstructing ROIs from angularly under-sampled and truncated tomographic acquisitions.
Many works have been published on this subject; it is known that with only a few additional data [181], the reconstruction can be greatly improved. However, hereinafter, we will consider cases where such data is unavailable.
To implement IR, one must choose the reconstruction grid, i.e., the support of the reconstructed area. When the grid matches the support of the ROI, the data will not agree with the reprojection of the ROI. When the reconstruction grid includes the support of the entire object [235], the reconstruction becomes computationally expensive and less stable due to the increase of unknowns for the same amount of data. Truncated data only allows a rough estimation of the exterior anatomical background, which can never be clinically acceptable. Thus, one more practical solution is to consider an intermediate smaller grid size with a "margin" outside of the ROI [163]. This achieves a faster and more stable ROI reconstruction in general. Yet, when dense objects such as metallic cables or needles are outside the reconstruction grid, the reprojection of the extended ROI contains high-frequency errors. Moreover, when too few projections are used for reconstruction, such objects suffer from aliasing, and additional streak artifacts can degrade the reconstructed ROI [130]. Another approach is to reconstruct the entire object with large voxels and then subtract the reprojection of the exterior from the data before reconstructing only the ROI from the subtracted data [109,247]. This approach produces a low-frequency approximation of the exterior of the ROI. However, such an approximation is poor in the presence of dense objects in the exterior of the ROI, and unwanted high-frequency content remains after subtraction that must again be dealt with.
In Chapter 3 (section 3.5), we highlighted that convolutional neural networks are an attractive alternative to IR due to their increased expressivity and fast inference. CNNs, particularly the U-net [186], have already been used for removing sub-sampling artifacts in reconstructions obtained from analytical methods for both non-truncated [111, 116] and truncated data [110]. However, there are concerns about the lack of guarantees and capacity for generalization of post-processing CNNs, because these networks do not en-

sure data consistency [198]. Deep unfolding methods circumvent this issue by offering a way to include *a priori* information in a neural network. They have been applied to many fields of imaging, such as denoising [224], deblurring [24], MRI reconstruction [233], and CT reconstruction from few-view data [3]. Unfolding consists of untying each iteration of an optimization algorithm, defining a set of learnable parameters, and training each iteration (or layer) in an end-to-end manner. Some authors allow the unfolding network to learn the optimization algorithm hyperparameters [24] as well as linear operators in the regularization, such as convolution kernels in ISTA-net [243]. Others use CNNs to replace proximity operators as in PD-net [3]. Deep unfolding networks automatically inherit from the feedback mechanism of IR for data consistency.
This chapter presents a deep learning architecture called Unfolded Reweighted Dual Block Forward Backward (U-RDBFB). The method exploits the framework of deep unfolding for accelerating the convergence, ensuring data consistency, and optimizing the parameters involved both in our cost function (e.g., adjoint of the linear operators) and our optimization algorithm (e.g., step sizes).
This chapter is divided into three sections. Section 8.2 presents our choices for the data fidelity and regularization terms and introduces a convergent iterative algorithm to minimize the resulting cost function. We then explain how this algorithm is unfolded into a deep learning architecture. This is followed by experiments (section 8.3), results (section 8.4), and discussions (section 8.5).

8.1.2 Problem formulation

We consider a 1D detector array of $B \in \mathbb{N}$ bins rotating around an object. The detector is too short to measure the projections of the entire object; its size defines a circular ROI we aim to reconstruct. Let $S \in \mathbb{N}$ be the number of projection angles. The vector of truncated sub-sampled tomographic data is $y \in \mathbb{R}^T$ with $T = B\,S$.
A reconstruction of the object attenuation map in the ROI can be obtained by considering a model of the form:

$$H\overline{x}_{\mathrm{G}} = y + n, \tag{8.1}$$

where $n \in \mathbb{R}^T$ accounts for some acquisition noise, $\overline{x}_{\mathrm{G}} \in \mathbb{R}^L$ is the scanned image restricted to a grid G with support larger than the ROI but smaller than that of the entire object, and $H \in \mathbb{R}^{T\times L}$ is the projector that models projection over this intermediary grid G. Operator H contains a subset of the columns of the projector on the entire space, or equivalently it corresponds to setting a subset of the columns of the complete projector, corresponding to the pixels outside of G, to zero.
When the grid G corresponds to the ROI, (8.1) assumes that the image values are 0 outside of the ROI. This assumption is not necessarily true for truncated data, so (8.1) does not hold. Hereafter, we suppose that the grid is extended beyond the ROI so that no assumption is made about the values outside the ROI.
We find an estimate of $\overline{x}_{\mathrm{G}}$ by computing a minimizer of a penalized cost function consisting of the sum of a data fidelity term f involving H and y, and a regularization term r, as

$$\underset{x\in\mathbb{R}^L}{\operatorname{argmin}} \;\; f(x) + r(x). \tag{8.2}$$

8.2 A deep unfolding network based on semi-local TV regularization and a Cauchy data fidelity

8.2.1 Iterative reconstruction

8.2.1.1 Cost function

Cauchy data fidelity:

When objects with high gradients (metallic wires or needles) do not belong to the reconstruction grid G, the error between the data and the reprojection of the estimate over G contains outliers at the projections of those high gradients. Angular sub-sampling of these outliers leads to streaks originating from these objects, i.e., from outside the grid. This means that the data should not be trusted equally but through a statistical analysis different from measurement noise. To avoid the streaks, we propose to decrease the influence of the largest errors between y and the reprojection of $H\overline{x}_{\mathrm{G}}$ using an M-estimator.
Here, we focus on the Cauchy estimator ϕ, which is a redescending M-estimator i.e., its derivative decreases to zero on $]-\infty,\kappa]\cap[\kappa,+\infty[$. It is defined as

$$(\forall \zeta \in \mathbb{R}) \qquad \phi(\zeta) = \frac{\beta\kappa^2}{2} \ln\left(1+\left(\frac{\zeta}{\kappa}\right)^2\right), \tag{8.3}$$

where $\beta > 0$ is a weighting term and $\kappa > 0$ monitors the sensitivity to outliers: the lower κ, the lower the influence of the outliers.
A graphical comparison of the Cauchy function (8.3) and the quadratic function $\phi(\cdot) = \frac{\beta}{2}(\cdot)^2$ is displayed in Figure 8.1 for $\beta = 1$ and various values of κ.

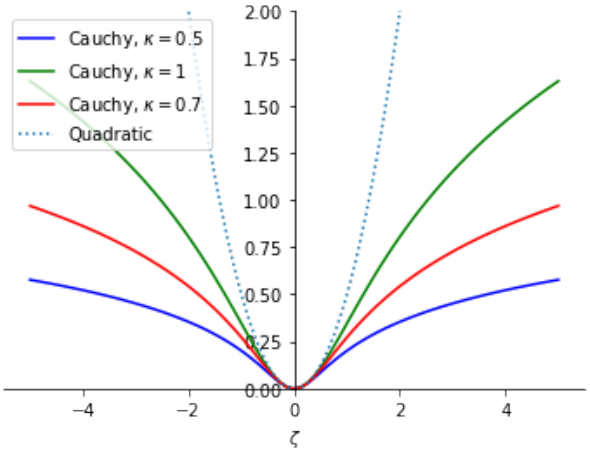


Figure 8.1: Comparison between the Cauchy and the quadratic functions.

Our data fidelity term f reads

$$(\forall x \in \mathbb{R}^L) \quad f(x) = g(Hx - y), \tag{8.4}$$

with

$$(\forall z = (z_t)_{1\leqslant t\leqslant T} \in \mathbb{R}^T) \qquad g(z) = \sum_{t=1}^{T} \phi(z_t). \tag{8.5}$$

Semi-local total variation:

Semi-local Total Variation regularizations (STV) [123] extend TV in a neighborhood of pixels indexed in $\Lambda_J = \{-J, \dots, J\} \setminus \{0\}$:

$$\begin{aligned}(\forall x \in \mathbb{R}^L)& \\ r_{\mathrm{STV}}(x) &= \sum_{j=1}^{J} \sum_{\ell=1}^{L} \alpha_{j,\ell} \sqrt{(x - V_j x)_\ell^2 + (x - V_{-j} x)_\ell^2} \\ &= \sum_{j=1}^{J} r_j(\nabla_j x). \end{aligned} \tag{8.6}$$

Hereinabove $\ell \in \{1, \dots, L\}$ is the spatial index and $V_j, V_{-j} \in \mathbb{R}^{L \times L}$ are shift operators as shown in Figure 8.2 for $j \in \{1, \dots, J\}$ and $J = 6$. Moreover, for every $j \in \{1, \cdots, J\}$, we define $\nabla_j = \begin{bmatrix} V_j^\top & V_{-j}^\top \end{bmatrix}^\top \in \mathbb{R}^{2L \times L}$ and, for every $z = (z_1, z_2) \in \mathbb{R}^{2L}$, $h_j(z) = \sum_{\ell=1}^{L} \alpha_{j,\ell} \sqrt{(z_1)_\ell^2 + (z_2)_\ell^2}$.
Parameters $(\alpha_{j,\ell})_{1 \leqslant j \leqslant J, 1 \leqslant \ell \leqslant L}$ are nonnegative weights that can be chosen to vary spatially, so making STV adaptive to the spatial contents [98]. We recover the standard TV regularization for constant values of these parameters and $J = 1$.

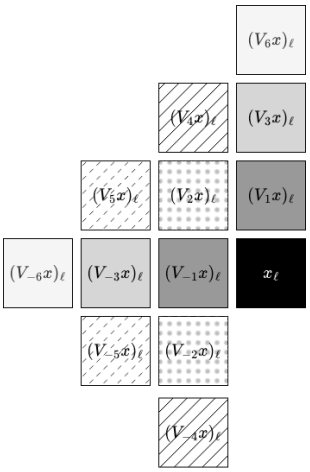


Figure 8.2: Shift operators $(V_j)_{j \in \Lambda_6}$ applied to a given pixel position ℓ

We add a nonnegativity constraint on the pixel values and a quadratic term $\frac{1}{2}\|x\|_M^2 = x^\top M x$ to the STV regularization. Here, matrix $M = \mathrm{diag}((m_\ell)_{\ell=1}^L) \in \mathcal{S}_L^+$ is such that, for every $\ell \in \{1, \dots, L\}$, $m_\ell = 1$ if the ℓ-th entry x_ℓ of vector x belongs to the ROI, and $m_\ell = \xi > 1$ otherwise. Thus $M \in \mathcal{S}_L^+$ and acts as a mask, limiting high values outside of the ROI.

Altogether, our regularization function in (8.2) reads

$$(\forall x \in \mathbb{R}^L) \quad r(x) = \sum_{j=1}^{J} r_j(\nabla_j x) + \frac{1}{2}\|x\|_M^2 + \iota_{[0,+\infty[^L}(x). \tag{8.7}$$

In this work, STV provides extra capacity compared to TV for learning. The Cauchy fidelity term has been used in ultrasound imaging [160] and in CT imaging [121] for mitigating the ring artifacts that appear due to defective detector bins only.

8.2.1.2 Minimization algorithm

Reweighting for non-convex data fidelity:

Given our choices for r and f, Problem (8.2) becomes

$$\underset{x\in\mathbb{R}^L}{\operatorname{argmin}}\ g(Hx-y)+\sum_{j=1}^{J} r_j(\nabla_j x)+\frac{1}{2}\|x\|_M^2+\iota_{[0,+\infty[^L}(x). \tag{8.8}$$

Because of the non-convexity of g, we adopt an iterative reweighting strategy where Problem (8.8) is replaced by a sequence of surrogate convex problems built following a majoration principle.
Let ϕ be given by (8.3). It was shown in [51] that, for every $\overline{\zeta}\in\mathbb{R}$, the following convex quadratic function $\tilde{\phi}(\cdot,\overline{\zeta})$, defined for every $\zeta\in\mathbb{R}$ as

$$\tilde{\phi}(\zeta,\overline{\zeta})=\phi(\overline{\zeta})+\beta\frac{(\zeta-\overline{\zeta})\overline{\zeta}}{1+(\overline{\zeta}/\kappa)^2}+\frac{\beta}{2}\frac{(\zeta-\overline{\zeta})^2}{(1+(\overline{\zeta}/\kappa)^2)}, \tag{8.9}$$

is a tangent majorant approximation to ϕ at $\overline{\zeta}$, that is

$$(\forall\zeta\in\mathbb{R})\ \ \tilde{\phi}(\zeta,\overline{\zeta})\geqslant\phi(\zeta)\quad\text{ and }\quad\tilde{\phi}(\overline{\zeta},\overline{\zeta})=\phi(\overline{\zeta}). \tag{8.10}$$

This allows us to deduce a tangent majorant function $\tilde{g}$ of function g at any point $\overline{z}\in\mathbb{R}^T$: $(\forall z\in\mathbb{R}^T)$,

$$\begin{aligned}\tilde{g}(z,\overline{z})&=\sum_{t=1}^{T}\phi(z_t,\overline{z}_t)\\&=g(\overline{z})+\beta\,\operatorname{diag}\left(\left(\frac{\overline{z}_t}{1+(\overline{z}_t/\kappa)^2}\right)_{t=1}^{T}\right)(z-\overline{z})\\&+\frac{\beta}{2}(z-\overline{z})^\top\operatorname{diag}\left(\left(\frac{1}{1+(\overline{z}_t/\kappa)^2}\right)_{t=1}^{T}\right)(z-\overline{z})\geqslant g(z).\end{aligned}$$

Finally, for every $\overline{x}\in\mathbb{R}^L$, we set

$$(\forall x\in\mathbb{R}^L)\quad\tilde{f}(x,\overline{x})=\tilde{g}(Hx-y;H\overline{x}-y), \tag{8.11}$$

that satisfies $\tilde{f}(x,\overline{x})\geqslant g(Hx-y)=f(x)$. Given this majoration, the iterative reweighting strategy approximates the solution to (8.8) by the estimate produced by Algorithm 1, where

$$(\forall(x,\overline{x})\in(\mathbb{R}^L)^2)\quad Q(x,\overline{x})=\tilde{f}(x,\overline{x})+r(x). \tag{8.12}$$

is a convex surrogate cost function.
$Q(\cdot,\overline{x})$ can be rewritten as

$$Q(x,\overline{x})=\iota_{[0,+\infty[^L}(x)+h_0(B_0x;B_0\overline{x})+h_1(B_1x)+\frac{1}{2}\|x\|_M^2, \tag{8.13}$$

where $\forall(x,\overline{x})\in(\mathbb{R}^L)^2$, with

$$B_0=H\in\mathbb{R}^{T\times L},\quad B_1=\begin{bmatrix}\nabla_1^\top&\cdots&\nabla_j^\top\end{bmatrix}^\top\in\mathbb{R}^{2JL\times L} \tag{8.14}$$

and, for every $\overline{x} \in \mathbb{R}^L$,

$$h_0(\cdot; B_0\overline{x}) = \tilde{g}(\cdot - y; B_0\overline{x} - y)$$
$$h_1(B_1 x) = \sum_{j=1}^{J} r_j(\nabla_j x). \tag{8.15}$$

Algorithm 1 Iterative reweighting strategy for Problem (8.8)

Require: Number of iterations $K \in \mathbb{N}^*$, $x_0 \in \mathbb{R}^L$
  for $k = 0$ to $K - 1$ **do**
    Define majorant function $Q(x; \overline{x})$ using (8.13)

$$x_{k+1} = \underset{x \in \mathbb{R}^L}{\operatorname{argmin}} \; Q(x, x_k) \tag{8.16}$$

  end for
Ensure: x_K approximating the solution to (8.8)

Let $(x_k)_{k\in\mathbb{N}}$ be generated by Algorithm 1. The cost sequence value (defined from (8.8)) monotonically converges.

Dual block coordinate forward-backward algorithm:

The k-th iteration of Algorithm 1 requires to solve (8.16), which amounts to minimizing the function $Q(\cdot, \overline{x})$, with $\overline{x}$ equals to the current iterate x_k. Since $M \in \mathcal{S}_L^+$, $Q(\cdot, \overline{x})$ is strongly convex for every $\overline{x} \in \mathbb{R}^L$. The minimization (8.16) is hence well-defined, with a unique solution that can be conveniently obtained using the dual forward-backward algorithm [177]. An accelerated and light version of this algorithm is its block coordinate version (DBFB) [1], which allows for accessing the proximity operators of h_0 and h_1 separately.
Algorithm 2 describes $N \in \mathbb{N}^*$ iterations of DBFB. The output $\boldsymbol{x}_N$ generated by DBFB with input x_k then defines x_{K+1} in Algorithm 1. For every $n \in \{1, \ldots, N\}$, DBFB updates the main primal variable $\boldsymbol{x}_n$ as well as two dual variables $z_n^0 \in \mathbb{R}^T$ and $s_n^1 \in \mathbb{R}^{2JL}$, associated to the data fidelity (data step (D)) or the regularization (regularization step (R)) terms, respectively. Each dual variable is activated (or not) at iteration n according to a binary variable ε_n.
When $N \to \infty$, the DBFB sequence $(\boldsymbol{x}_n)_{n\in\mathbb{N}}$ converge to the solution to (8.16) under the following assumptions on the algorithm parameters:

$$\begin{cases} \sigma \geqslant |||B_0 M^{-1} B_0^\top |||, \\ (\forall j \in \{1, \ldots, J\}) \quad \tau_j \geqslant |||\nabla_j M^{-1} \nabla_j^\top |||, \\ \gamma_n \in [\epsilon, 2 - \epsilon] \text{ with } \epsilon \in]0, 1] \\ (\exists M \in \mathbb{N} \setminus \{0, 1\})(\forall n \in \mathbb{N}) \quad 0 < \sum_{n'=n}^{n+M-1} \varepsilon_{n'} < M. \end{cases} \tag{8.17}$$

The first three assumptions are stepsize range conditions. The last one means that each step (D) and (R) is performed at least once, every M iterations.

Algorithm 2 DBFB algorithm to solve (8.16) with $x_k = \overline{x} \in \mathbb{R}^L$

Require: Number of iterations $N \in \mathbb{N}^*$, tangent point $\overline{x} \in \mathbb{R}^L$, initial dual variables $z_0^0 \in \mathbb{R}^T, (\forall j \in \{1, \dots, J\})\ z_0^j \in \mathbb{R}^{2L}$ with constant J defined in (8.6), operators B_0 and B_1 defined in (8.14), stepsizes $(\sigma, \tau_1, \dots, \tau_J) \in]0, +\infty[^{J+1}$.

$s_0^1 = (z_0^1, \dots, z_0^J)$
$\Sigma = \mathrm{diag}(\tau_1, \dots, \tau_J)$
$w_0 = -M^{-1}(B_0^\top z_0^0 + B_1^\top s_0^1)$
For $n = 0, 1 \dots, N$
  $(z_n^1, \dots, z_n^J) \equiv s_n^1$
  $\boldsymbol{x}_n = \mathrm{proj}_{[0,+\infty[^L}(w_n)$
  Select $\varepsilon_n \in \{0, 1\}$ and $\gamma_n \in]0, +\infty[$
  If $\varepsilon_n = 0$ **(D)**
    $\tilde{z}_n^0 = z_n^0 + \gamma_n \sigma^{-1} B_0 \boldsymbol{x}_n$
    $z_{n+1}^0 = \tilde{z}_n^0 - \gamma_n \sigma^{-1} \mathrm{prox}_{\gamma_n^{-1}\sigma h_0(\cdot; B_0\overline{x})}(\gamma_n^{-1}\sigma \tilde{z}_n^0)$
    $w_{n+1} = w_n - M^{-1} B_0^\top (z_{n+1}^0 - z_n^0)$
    $s_{n+1}^1 = s_n^1$
  If $\varepsilon_n = 1$ **(R)**
    $\tilde{s}_n^1 = s_n^1 + \gamma_n \Sigma^{-1} B_1 \boldsymbol{x}_n$
    $s_{n+1}^1 = \tilde{s}_n^1 - \gamma_n \Sigma^{-1} \mathrm{prox}_{\gamma_n^{-1}\Sigma h_1}(\gamma_n^{-1}\Sigma \tilde{s}_n^1)$
    $w_{n+1} = w_n - M^{-1} B_1^\top (s_{n+1}^1 - s_n^1)$
    $z_{n+1}^0 = z_n^0$

Ensure: $\boldsymbol{x}_N$ approximating the minimizer of $Q(\cdot, \overline{x})$.

Step (D) involves the calculation of the proximity operator $\mathrm{prox}_{\gamma_n^{-1}\sigma h_0(\cdot, B_0\overline{x})}$, which has a closed-form [16, Example 24.2], for $(\forall (z, \overline{z}) \in (\mathbb{R}^T)^2)$,

$$\begin{aligned}
&\mathrm{prox}_{\gamma_n^{-1}\sigma h_0(\cdot, B_0\overline{z})}(z) = \mathrm{prox}_{\gamma_n^{-1}\sigma \tilde{g}(\cdot - y, \overline{z} - y)}(z), \\
&= y + \mathrm{prox}_{\gamma_n^{-1}\sigma \tilde{g}(\cdot, \overline{z} - y)}(z - y) \\
&= \left(y_t + \frac{z_t - y_t}{1 + \beta\gamma_n^{-1}\sigma\big(1 + (\overline{z}_t - y_t)^2/\kappa^2\big)^{-1}} \right)_{t=1}^{T}. \qquad (8.18)
\end{aligned}$$

Step (R) requires calculating the proximity operator of h_1 scaled by parameter $\gamma \in]0, +\infty[$. It also has a closed form: for $s = (s_1, \dots, s_J) \in \mathbb{R}^{2JL}$, $\mathrm{prox}_{\gamma h_1}(s) = \Big(\mathrm{prox}_{\gamma r_j}(s_j)\Big)_{j=1}^{J}$, where, for every $z = (z_1, z_2) \in \mathbb{R}^{2L}$,

$$\mathrm{prox}_{\gamma r_j}(z) = \left(\max\left\{0, 1 - \frac{\gamma \alpha_{j,l}}{\|z_\ell\|_2}\right\} z_\ell \right)_{\ell=1}^{L}, \qquad (8.19)$$

where, for every $\ell \in \{1, \dots, L\}$, $z_\ell = ((z_1)_\ell, (z_2)_\ell) \in \mathbb{R}^2$.

The overall iterative strategy for approximating the solution to (8.8) consists of applying Algorithm 1, where, for every $k \in \mathbb{N}^*$, $N_k \in \mathbb{N}^*$ iterations of Algorithm 2 are used as an inner solver with $\overline{x} = x_k$ to compute x_{k+1} in (8.16). We call the resulting iterations reweighted DBFB (RDBFB) algorithm.

In the context of CT, iterative reweighted algorithms usually involve surrogates to the regularization term [225, 237] rather than to the data fidelity, as done here.

8.2.2 Unfolded reconstruction

Hereafter, we present a deep neural network, designated as U-RDBFB (Unfolded Reweighted DBFB), by unfolding all the steps of RDBFB. Specifically, the network mimics the application of K iterations of Algorithm 1, as K main layers, each of them grouping $N_k \in \mathbb{N}$ iterations of Algorithm 2. This yields an architecture with $\sum_{k=0}^{K-1} N_k$ layers in total.

8.2.2.1 From RDBFB iterations to U-RDBFB layers

The deep unfolding paradigm recasts every step of Algorithm 2 as one neural network layer: step (D) becomes $\mathcal{L}_D$ ($\varepsilon_n = 0$) and step (R) becomes $\mathcal{L}_R$ ($\varepsilon_n = 1$). It requires truncating the number of layers drastically. To optimize the depth of our network, we will introduce two modifications of the steps (D) and (R) to construct the corresponding layers.

Throughout this thesis, we have highlighted that replacing an adjoint operator with a surrogate has the potential to accelerate convergence or improve the solution of an IR method [196, 238, 240]. In CT, a frequently encountered operator is the ramp filter F, which satisfies $FHH^\top \approx \mathrm{Id}$. Thus, to improve conditioning and allow larger values for the step sizes associated with layer $\mathcal{L}_D$, hence a lower number of such a layer, we replace B_0 in Algorithm 2 with FH.
By setting $\nu_{n,0} = \gamma_n \sigma^{-1}$ and by using the relation

$$\mathrm{prox}_{\nu_{n,0} h_0(\cdot; FH\overline{x})} = \mathrm{prox}_{\nu_{n,0}\tilde{g}(\cdot; H\overline{x}-y)}(\cdot - Fy) + Fy,$$

we define layer $\mathcal{L}_\mathrm{D}$ as

$$\begin{aligned}
&\text{Data layer } (\mathcal{L}_\mathrm{D}):\\
&\left\lfloor
\begin{array}{l}
\boldsymbol{x}_n = \mathrm{proj}_{[0,+\infty[^L}(w_n)\\
u_n = z_n^0 + \nu_{n,0} F(H\boldsymbol{x}_n - y)\\
z_{n+1}^0 = u_n - \nu_{n,0}^{-1}\,\mathrm{prox}_{\nu_{n,0}\tilde{g}(\cdot; FH\overline{x}-y)}(\nu_{n,0} u_n)\\
w_{n+1} = w_n - M^{-1} H^\top (z_{n+1}^0 - z_n^0)\\
z_{n+1}^j = z_n^j \quad (\forall j \in \{1,\dots,J\}).
\end{array}
\right.
\end{aligned} \tag{8.20}$$

Similarly, for step (R), we unfold by replacing the adjoint of the regularization operator B_1 with $\tilde{B}_1 = \begin{bmatrix} \tilde{\nabla}_1^\top & \cdots & \tilde{\nabla}_J^\top \end{bmatrix}^\top$. Setting, for every $j \in \{1,\dots,K\}$, $\nu_{n,j} = \gamma_n \tau_j^{-1}$ yield the following regularization layer $\mathcal{L}_\mathrm{R}$:

$$\begin{aligned}
&\text{Regularization layer } (\mathcal{L}_\mathrm{R}):\\
&\left\lfloor
\begin{array}{l}
\boldsymbol{x}_n = \mathrm{proj}_{[0,+\infty[^L}(w_n)\\
\text{For } j \in \{1,\dots,J\}\\
\quad \left\lfloor
\begin{array}{l}
(\forall \ell \in \{1,\dots,L\})\\
\quad (z_{n+1}^j)_\ell = \dfrac{\left(z_n^j + \nu_{n,j}\nabla_j \boldsymbol{x}_n\right)_\ell}{\max\left\{1, \left\| \left(z_n^j + \nu_{n,j}\nabla_j \boldsymbol{x}_n\right)_\ell \right\|_2 / \alpha_{j,\ell}\right\}}
\end{array}
\right.\\
w_{n+1} = w_n - M^{-1}\sum_{j=1}^J \tilde{\nabla}_j (z_{n+1}^j - z_n^j)\\
z_{n+1}^0 = z_n^0.
\end{array}
\right.
\end{aligned} \tag{8.21}$$

Note that $\mathcal{L}_R$ does not involve $\overline{x}$.

8.2.2.2 Total architecture

The total architecture of U-RDBFB, denoted $\mathcal{A}$, can be summarized as

$$\mathcal{A} = \mathcal{L}^{K-1} \circ \cdots \circ \mathcal{L}^0. \tag{8.22}$$

For $0 \leqslant k \leqslant K-1$, $\mathcal{L}^k$ corresponds to a sequence of N_k layers $\mathcal{L}_D$ or $\mathcal{L}_R$ and implements the following update:

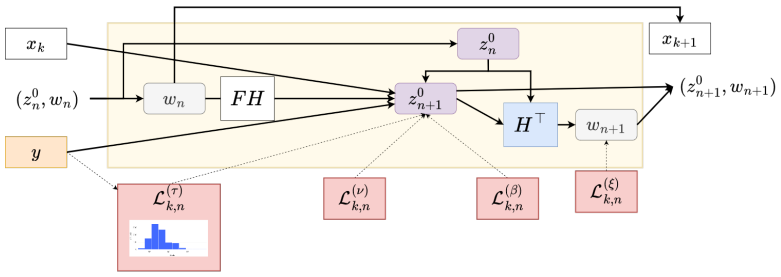
$$(z_{0,k+1}, x_{k+1}) = \mathcal{L}^k(z_{0,k}, x_k; \Theta_k), \tag{8.23}$$

where

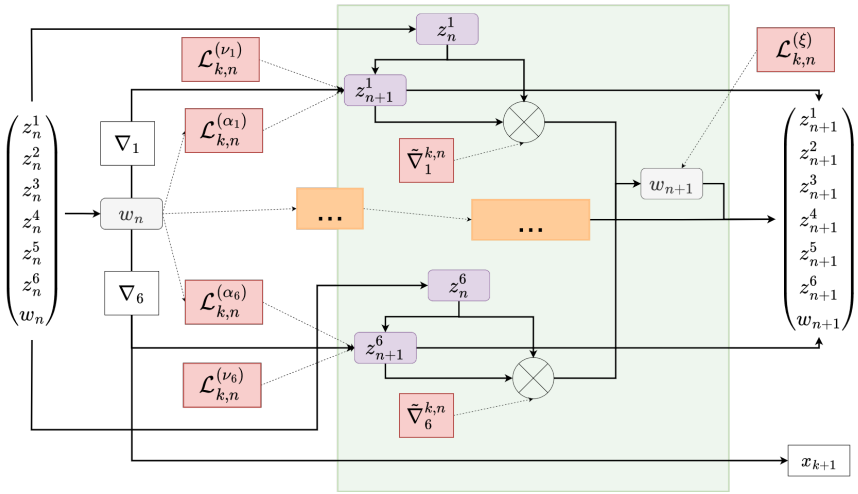
- x_k is the current reweighted estimate $\overline{x}$ in $\mathcal{L}_{\mathrm{R}}$-$\mathcal{L}_{\mathrm{D}}$ ($x_0 = H^\top F y$).
- x_{k+1} is the next reweighted estimate; it is equal to $\boldsymbol{x}_{N_k-1}$ given by the N_k-th layer $\mathcal{L}_{\mathrm{R}}$-$\mathcal{L}_{\mathrm{D}}$.
- $z_{0,k} \in \mathbb{R}^T \times (\mathbb{R}^{2L})^J$ is the initial value of the variables $(z_0^j)_{0 \leqslant j \leqslant J}$ for layers $\mathcal{L}_{\mathrm{R}}$-$\mathcal{L}_{\mathrm{D}}$ (for $k = 0$, $(z_0^j)_{j=1}^J$ are initialized to zero while z_0^0 is set to $-Fy$).
- $z_{0,k+1} \in \mathbb{R}^T \times (\mathbb{R}^{2L})^J$ is is equal to $(z^j_{N_{k-1}})_{0 \leqslant j \leqslant J}$ given by the N_k-th layer $\mathcal{L}_{\mathrm{R}}$-$\mathcal{L}_{\mathrm{D}}$
- Θ_k is the vector of trainable parameters. The parameters are layer-dependent, so we index them by k and n.
  For layer $\mathcal{L}_{\mathrm{D}}$, the parameters are those of the Cauchy function $(\beta_{k,n}, \kappa_{k,n})$, the one of the quadratic regularization $\xi_{k,n}$, and a single step size $\nu_{k,n,0}$.
  For layer $\mathcal{L}_{\mathrm{R}}$, the regularization parameters $(\alpha_{k,n,j,\ell})_{1 \leqslant j \leqslant J, 1 \leqslant \ell \leqslant L}$, $\kappa_{k,n}$, $\xi_{k,n}$, and step sizes $(\nu_{k,n,j})_{j=1}^J$ are learned as well as the surrogates $(\tilde{\nabla}_j^{k,n})_{j=1}^J$ to the adjoints of operators $(\nabla_j^\top)_{j=1}^J$.
  To infer all these parameters, we introduce learning modules $(\mathcal{L}_{k,n}^{(\theta)})_{n=0}^{N_k-1}$ for $\theta \in \Theta_k$.

Schematic views of layers $\mathcal{L}_{\mathrm{D}}$ and $\mathcal{L}_{\mathrm{R}}$ can be found in Figure 8.3a and Figure 8.3b, and a composition $\mathcal{A}$ of such layers is displayed in Figure 8.3c.
Here we propose using $K = 7$ in (8.22) with $N_k = 4$ for each $k \in \{0, \cdots, K-1\}$, resulting in a total of 28 layers:

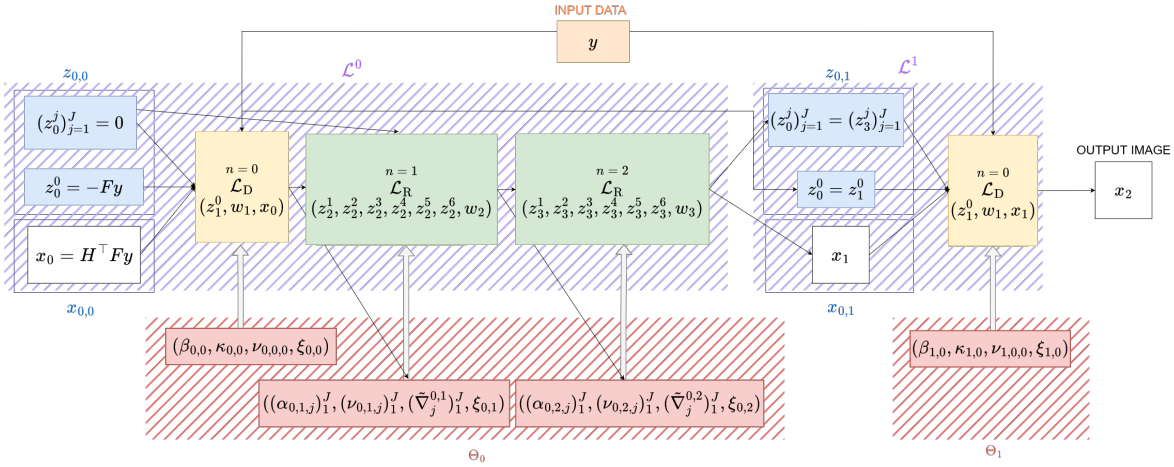
$$\mathcal{L}^0 = \cdots = \mathcal{L}^6 = (\mathcal{L}_R \circ \mathcal{L}_D)^2. \tag{8.24}$$



(a) Schematic view of a N_k-th layer $\mathcal{L}_D$ (8.20). The layer relies on $\overline{x} = x_k$, the k-th reweighted iterate to generate the next reweighted iterate x_{k+1}. The layer takes as inputs w_n, z_n^0 from the previous layers. The projections y are also used as input.



(b) Schematic view of a N_k-th layer $\mathcal{L}_R$ (8.21). The layer takes as inputs w_n, $(z_n^j)_1^J$ from the previous layer and generates the next reweighted iterate x_{k+1}. The update of parameters for $j \in \{2, \dots, 5\}$ is hidden in the orange block for the sake of readability. Only parameters α_j depend on the input.



(c) U-RDBFB in the case where $\mathcal{A} = \mathcal{L}^1 \circ \mathcal{L}_2^0$ where $\mathcal{L}^0 = (\mathcal{L}_R)^2 \circ \mathcal{L}_D$ and $\mathcal{L}_2^1 = \mathcal{L}_D$ (i.e., $K = 2$, $N_0 = 3$, $N_1 = 1$). Red blocks represent the hidden structures to infer all the parameters $\theta \in \Theta$. When $k = 1$, the dual variables of DBFB are initialized with the values of the dual variables at the end of the previous N_0 iterations of DBFB ($k = 0$).

Figure 8.3: Architecture of U-RDBFB

We now discuss our choices for $(\mathcal{L}_{k,n}^{(\theta)})_{n=0}^{N_k-1}$:

- step size for $\mathcal{L}_{\mathrm{D}}$: $\nu_{k,n,0} = \mathcal{L}_{k,n}^{(\nu)} = \mathrm{softplus}(a_{k,n})$ where $a_{k,n}$ is a learnable real-valued parameter.
- Parameters of the Cauchy function for $\mathcal{L}_{\mathrm{D}}$:
  - $\kappa_{k,n} = \mathcal{L}_{k,n}^{(\kappa)} = W_\kappa \,\mathrm{softplus}(c_{k,n})$ where $c_{k,n}$ is inferred from a fully connected layer, whose weights are shared across the U-RDBFB network, applied on a histogram of the absolute value of the filtered reprojection error, i.e., $H^\top(Fx_n - y)$. We implemented the learnable histogram layer proposed in [226], which is piecewise differentiable. More precisely, we built a cumulated histogram using 100 bins from 0 to the maximum value of the filtered reprojection error.
  - $\beta_{k,n} = \mathcal{L}_{k,n}^{(\beta)} = W_\beta \,\mathrm{softplus}(d_{k,n})$.
- Diagonal elements of M involved in (8.7) corresponding to the locations of pixels outside of the ROI for both $\mathcal{L}_{\mathrm{D}}$ and $\mathcal{L}_{\mathrm{R}}$: $\xi_{k,n} = \mathcal{L}_{k,n}^{(\xi)} = \mathrm{softplus}(e_{k,n})$ where $e_{k,n}$ is learned.
- step size for $\mathcal{L}_{\mathrm{R}}$: For every $j \in \{1, \dots, J\}$, $\nu_{k,n,j} = \mathcal{L}_{k,n}^{(\nu_j)} = W_\nu \,\mathrm{softplus}(b_{k,n,j})$.
- Parameters of the STV regularization for $\mathcal{L}_{\mathrm{R}}$: For every $j \in \{1, \dots, J\}$,

$$\begin{aligned}\alpha_{k,n,j} &= (\alpha_{k,n,j,l})_{\ell=1}^{L} = \mathcal{L}_{k,n}^{(\alpha_j)} \\ &= W_\alpha \,\mathrm{softplus}(A_{k,n} \circ \mathrm{relu} \circ B_{k,n}(\nabla_j x_k)),\end{aligned}$$

  where $A_{k,n}$ is a grouped convolution of 7 groups with size 3×3 kernels and $J = 7$ channels and $B_{k,n}$ is a grouped convolution of 14 groups with size 5×5 kernels and 14 channels.

In $\mathcal{L}_D$, initial values of $a_{k,n}$, $d_{k,n}$ are set to 1. In $\mathcal{L}_R$, initial values for $b_{k,n,j}$, $e_{k,n}$ are 1. Normalization scalars W_κ, W_ν, W_β and W_α are set to 10^{-5}, 10, 10 and 0.05 respectively.

For each layer $\mathcal{L}_R$, we also learn operators $(\tilde{L}_j^{k,n})_{j=1}^J$ which have the same support as $(L_j^{k,n})_{j=1}^J$.

8.2.2.3 Incremental training strategy

We circumvent the issue of optimizing the initial values of the parameters of our learning modules using an incremental training strategy, as sometimes advocated for when initializing the weights of recurrent neural networks [191]. The learning in each layer (k, n) $(n \geqslant 1)$ starts by considering all the previous layers from $(0, 0)$ to $(k, n-1)$ with their past trained parameters as an initialization. This means that an increasing number of layers is trained simultaneously. In the last step, all layers are trained end-to-end.

8.3 Experiments

We now describe our experimental setup. First, we illustrate the benefits of using a Cauchy-based data fidelity function. We compare the results of our RDBFB iterative approach to a simpler DBFB scheme minimizing the same cost function but with g in (8.8) replaced with the ℓ_2 norm.
Second, we comment on the improved performance brought by our unfolding strategy (learning of adjoints and parameters as well as the use of the ramp filter), and so U-RDBFB (subsection 8.2.2) is compared to the original RDBFB algorithm.
Third, we compare U-RDBFB to several state-of-the-art reconstruction methods and comment on the transfer of performance over different synthetic datasets.

8.3.1 Datasets

We used three datasets for evaluation.

8.3.1.1 Abdomen dataset

Our first dataset consists of 2D images obtained from 60 CT volumes of size 512×512×512 from the lower lungs to the lower abdomen of 60 patients, which were extracted from the public dataset CT Lymph Nodes from https://www.cancerimagingarchive.net/. These volumes correspond to fully sampled CT reconstructions. They were made isotropic by interpolating the axial slices. A total of 50 out of 512 slices were kept per volume. We randomly added intense metallic wires between 3000 and 5000 Hounsfield units (HU) of varying sizes on the axial slices. We shifted the HU values of the images by 1000 so that air is 0 HU and water is 1000 HU, using $a_{\text{tissue}} \mapsto (a_{\text{tissue}} - \mu_{\text{water}}) \times (1000/\mu_{\text{water}})$, where μ_{water} is the value of the attenuation coefficient of water equal to 0.017 mm^{-1} and $a_{\text{tissue}} \in \mathbb{R}^Q$ is the initial vector of attenuation values. Finally the 512×512 slices $\overline{x}_{\text{P}}$ were normalized between $[0, 1]$ (a value of 1 corresponding to an object of HU intensity equal to 5000).
To eliminate bias with respect to model discretization, projections were simulated for each slice of each volume in a 2D parallel geometry with a short detector of 600 bins (bin size equal to half a pixel size, i.e., 0.5 mm) and an angular density of 110 projections over 180° through to the projector $H_{\text{P}} \in \mathbb{R}^{512^2 \times (110 \times 600)}$. The projections were then rebinned by a factor 2 (operator $R \in \mathbb{R}^{300 \times 600}$). Noisy projections $y = (y_t)_{t=1}^T$ are computed as

$$(\forall t \in \{1, \dots, T\}) \quad y_t = \mu \log \left(\frac{I_0}{\mathcal{P}(I_0 \exp\left(-\mu (R H_{\text{P}} \overline{x}_{\text{P}})_t\right))} \right),$$

where we set $\mu = \mu_{\text{water}}/1000$, $I_0 = 10^4$, and, for some $\delta > 0$, $\mathcal{P}(\delta)$ denotes a realization of a Poisson law with mean δ. In this context, the ROI was a centered disk of diameter 300. The resulting pairs of axial slice/projections $(\overline{x}_{\text{P}}, y)$ were split into a training of 2500 pairs from a pool of 50 patients and a testing set of 500 pairs from 10 other patients.

8.3.1.2 Head dataset

We used a second dataset containing 2D images extracted from 10 CT high-dose brain reconstructions. These volumes are from the public repository 2016 Low Dose CT Grand Challenge from https://www.cancerimagingarchive.net/. After extracting 50 slices of size 512×512 per volume (pixel size of 0.5 mm), we performed the same processing as for the Abdomen dataset (addition of intense wires, normalization, projection, rebinning) for generating a testing set of 500 pairs of axial slices/projections $(\overline{x}_{\text{P}}, y)$.

8.3.1.3 Geometrical dataset

Our third dataset was created using the toolbox TomoPhantom [122]. 500 geometrical 2D piecewise-constant phantoms were randomly generated on a 512×512 grid and normalized between 0 and 1. Again, we performed the same processing as for the Abdomen dataset for generating a testing set of 500 pairs of axial slices/projections $(\overline{x}_{\mathrm{P}}, y)$.

8.3.2 Training details for U-RDBFB

Let $i \in \{1, \dots, I\}$ be the index covering all $I = 2500$ instances of the training set. The reconstruction grid (G) is a disk of diameter 400. Let $x^*_{\mathrm{G,i}} \in \mathbb{R}^L$ be the output of U-RDBFB for a given projection input y_i. Our network is thus designed to minimize $\sum_{i=1}^{I} \ell(C_{\mathrm{G}} x^*_{\mathrm{G},i}, C_{\mathrm{P}} \overline{x}_{\mathrm{P},i})$, where C_{G} is a cropping operator which extracts the ROI from the grid G, C_{P} is a cropping operator which extracts the ROI from the entire 512×512 grid, and ℓ is the loss retained for training the network. For all instances of the training set,

$$\ell(C_{\mathrm{G}} \cdot, C_{\mathrm{P}} \overline{x}_{\mathrm{P,i}}) = \frac{1}{I} \| C_{\mathrm{G}} \cdot - C_{\mathrm{P}} \overline{x}_{\mathrm{P,i}} \|^2, \tag{8.25}$$

corresponding to the MSE loss.
We implemented U-RDBFB following (8.24) in Pytorch, using a Tesla V100 32 Gb GPU. We used six epochs for training each layer $\mathcal{L}_{\mathrm{R}}$ and ten epochs for training each layer $\mathcal{L}_{\mathrm{D}}$; the only exception was the last layer, for which we used 20 epochs. The learning rate is decreased with a step decay by a factor of 0.99 from 10^{-2} every 4 epochs. The batch size for each epoch varied from 20 to 8 as the number of trained layers increased. We employed the toolbox TorchRadon [185] to include Pytorch-compatible parallel-beam tomographic operators in all architectures. Standard auto-differentiation tools can compute all necessary derivatives for backpropagation. The training procedure takes about one day and a half.

8.3.3 Competing methods

The quantitative metric used to assess the reconstruction quality of $C_{\mathrm{G}} x^*_{\mathrm{G,i}}$ is the PSNR. We also evaluate the reconstruction performance using the structural similarity index (SSIM), the PieApp value [174], and the Mean Absolute Error (MAE) of the difference between $C_{\mathrm{G}} x^*_{\mathrm{G,i}}$ and $C_{\mathrm{P}} \overline{x}_{\mathrm{P,i}}$.
We compare U-RDBFB with FBP, an iterative method, and four deep-learning methods that we describe hereinafter.

8.3.3.1 FBP

This analytical method consists of computing $H_{\mathrm{ROI}}^{\top} F y$, where $H_{\mathrm{ROI}} \in \mathbb{R}^{300^2 \times (110 \times 300)}$. As is commonly the case when applying FBP on truncated data, we extrapolated the projections prior to ramp filtering (using anti-symmetric padding).

8.3.3.2 RDBFB algorithm

For completeness, we perform comparisons with the iterative method proposed in subsection 8.2.1.2. For each reweighted iteration k, we used $N_k = 10$ DBFB iterations alternating between data and regularization steps (1:1 correspondence). For an easier manual tuning of the hyperparameters, instead of using $J = 6$ as in U-RDBFB, we set $J = 1$ so that STV reduces to TV and $\alpha_{1,\ell} \equiv \alpha_1$, for all $\ell \in \{1, \cdots, L\}$. The remaining

cost function parameters (ξ, κ, β) are selected by optimizing PSNR on the training set via a grid search.

8.3.3.3 Post-processing U-net

The third competing method is the CNN proposed in [111,116], which is a post-processing of FBP. It relies on a trained residual U-net, with a depth of 4 levels, filters of size 32, and batch normalization to improve the stability of training.

8.3.3.4 Preconditioned Neumann Network (PNN)

Our fourth competing method is a preconditioned Neumann network (PNN) initially introduced in [99] for MRI reconstruction. It builds on a method for solving Problem (8.2) with $f(x) = \frac{1}{2}\|Hx - y\|^2$. For a differentiable function r, the resulting minimizer reads

$$(H^\top H + \nabla r)x = H^\top y, \tag{8.26}$$

which can be rewritten as

$$(H^\top H + \lambda \,\mathrm{Id}\,)x + (\nabla r - \lambda \,\mathrm{Id}\,)x = H^\top y. \tag{8.27}$$

Setting $T_\lambda = (H^\top H + \lambda \,\mathrm{Id}\,)^{-1}$ yields

$$(\mathrm{Id}\, - \lambda T_\lambda + T_\lambda \nabla r)x = T_\lambda H^\top y. \tag{8.28}$$

Using the Neumann identity $B^{-1} = \sum_{n=0}^{\infty}(\mathrm{Id}\, - B)^n$, the authors derive the architecture of PNN with $N \in \mathbb{N}^*$ layers (see Figure 8.4)

$$(\lambda T_\lambda - T_\lambda \nabla r)^N \circ T_\lambda(H^\top y). \tag{8.29}$$

All instances of T_λ are applied approximately using an unrolling of 10 iterations of the conjugate gradient algorithm.
The operator $T_\lambda \nabla r$ is replaced by a U-net, denoted by Ψ, which has the same architecture as the aforementioned U-net without the residual connection. The weights of the U-net are shared for all layers. Following [99], no batch normalization is used. The inner U-net has a depth of 4, the learning rate is set to 10^{-4}, and the initial value for λ is 0.01. We choose $N = 3$. One feature of PNN compared to other deep unfolding networks is that it contains skip connections.

8.3.3.5 ISTA-net

Our fifth competing method is ISTA-net, derived from the work of [243]. ISTA-net is designed to solve Problem (8.2) for $f(x) = \frac{1}{2}\|Hx - y\|^2$, and $r(x) = \lambda\|Wx\|_1$ ($\lambda > 0$), where operator W is not known a priori but learned. W is an orthogonal linear operator in the initial ISTA algorithm, whose iteration reads

$$x_{n+1} = W^\top \,\mathrm{soft}\left(W(x_n - \tau H^\top(Hx_n - y), \lambda\tau\right), \tag{8.30}$$

where soft is the soft-thresholding operation and $\tau > 0$ is the gradient step size. In ISTA-net, the authors replace W and $W^\top$ by two decoupled nonlinear operators namely $A_n \circ \mathrm{ReLU} \circ B_n$ and $C_n \circ \mathrm{ReLU} \circ D_n$ (see Figure 8.5). The property of orthogonality of W is not imposed but favored during training by adding a term, weighted by $\chi \in]0, +\infty[$, penalizing the difference between $(C_n \circ \mathrm{ReLU} \circ D_n) \circ (A_n \circ \mathrm{ReLU} \circ B_n)x_n$ and x_n in the

loss function. Each A_n, B_n, C_n and D_n is a 2D convolutional operator. B_n and C_n are associated with a kernel of size 3×3 and 32 input and output channels; A_n has 1 input channel and 32 output channels and vice-versa for D_n. As suggested by the authors, we learn these convolutional operators as well as λ and τ, which are allowed to vary at each iteration.
Experiments are carried out with 10 layers, $\chi = 0.1$, x_0 is the FBP reconstruction, λ and τ are initialized to 0.1 and 0.01 respectively.

8.3.3.6 PD-net

The last competing method is the learned Primal-Dual (PD-net) introduced in [2] by unrolling the Primal-Dual Hybrid Gradient (PDHG) optimization algorithm [50]. The authors consider Problem (8.2) with a more generic data fidelity term $f(x) = G(Hx; y)$. They replace both the proximity operators of G and r in PDHG by residual CNN so that one layer n of their network reads

$$z_{n+1} = \mathrm{CNN}(z_n + \sigma H \tilde{x}_n; y) \tag{8.31}$$

$$x_{n+1} = \mathrm{CNN}(x_n - \tau H^\top z_{n+1}) \tag{8.32}$$

$$\tilde{x}_{n+1} = x_{n+1} + \gamma(x_{n+1} - x_n). \tag{8.33}$$

The CNNs act both in the image and projection domains. Furthermore, buffers of previous iterates of size $N_p \in \mathbb{N}$ in the primal domain (image) and of size $N_d \in \mathbb{N}$ in the dual domain (projection) are kept to enable the network to learn an acceleration. We used 9 layers, $N_d = N_p = 3$, and 32 filters in the convolutional layers. This network is illustrated in Figure 8.6.

	U-RDBFB	U-net	PNN	PD-net	ISTA-net
$\lvert\Theta\rvert$	2.3169×10^4	1.9278×10^6	1.9278×10^6	2.5470×10^5	1.7109×10^5

Table 8.1: Number of learnable parameters (Θ)

The competing networks were also trained with the MSE loss (using (8.25) for unfolding networks and a regularization term for ISTA-net weighted by χ) in a standard end-to-end manner. The number of epochs was chosen such that all networks have converged. Note that codes are publicly available for these networks. We re-implemented them in Pytorch, and kept the setting of the parameters advocated by the authors, except for PNN, for which we reduced the number of layers to 3 to obtain a stable behavior for training. The total number of parameters of each network is reported in Table 8.1.

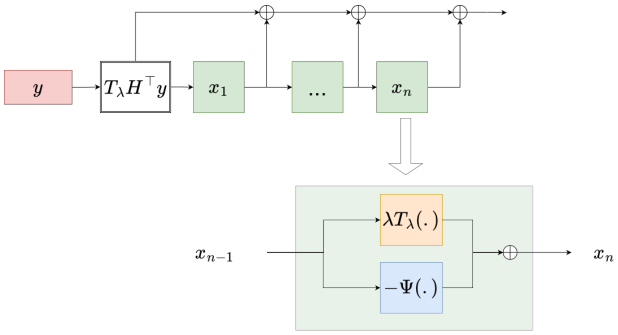


Figure 8.4: Architecture of PNN [99]: The network maps a linear function of the measurements $T_\lambda H^\top y$ to a reconstruction x_n by successive applications of an operator of the form $\lambda T_\lambda - \Psi$, while summing the intermediate outputs of each block. All instances of T_λ are replaced by an unrolling of 10 iterations of the conjugate gradient algorithm. Ψ is a trained network and the scale parameter λ is also trained.

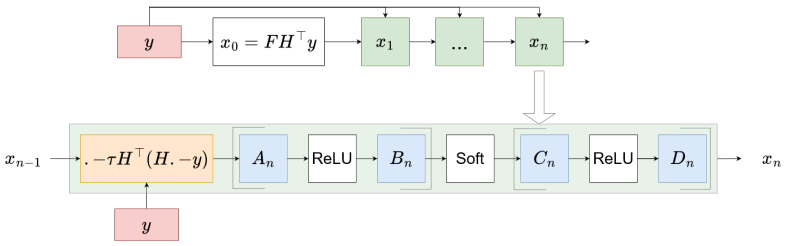


Figure 8.5: Architecture of ISTA-net [243]: Each layer is composed of a gradient step followed by the application of a nonlinear operator, which is the combination of two learnable linear convolutional operators (A_n, B_n) separated by a ReLU, a soft-thresholding operation and then two other learnable linear convolutional operators (C_n, D_n) separated by a ReLU. The property $(C_n \circ \text{ReLU} \circ D_n) \circ (A_n \circ \text{ReLU} \circ B_n) = \text{Id}$ is favored during training.

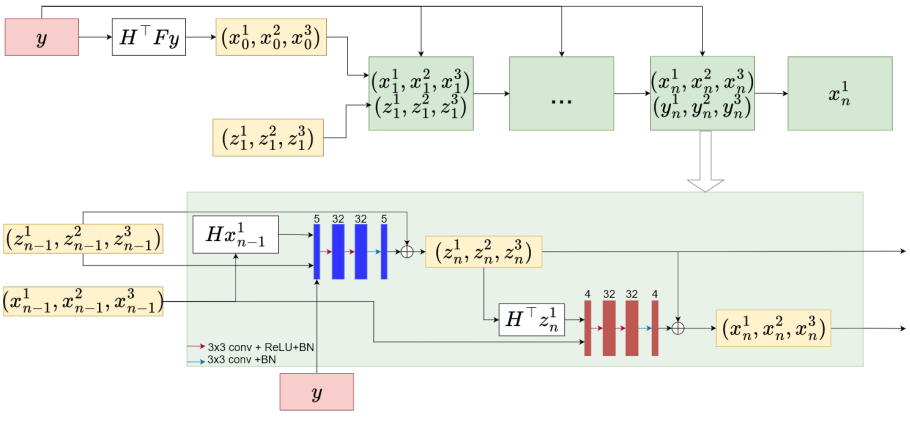


Figure 8.6: Architecture of PD-net [2]: The red and blue boxes represent the primal and dual networks, respectively. Buffers of 3 primal (x_n^1, x_n^2, x_n^3) and dual (z_n^1, z_n^2, z_n^3) estimates are used at each iteration. The initial primal estimates are set to the FBP reconstruction given by $H^\top Fy$, and the initial dual estimates are set to zero.

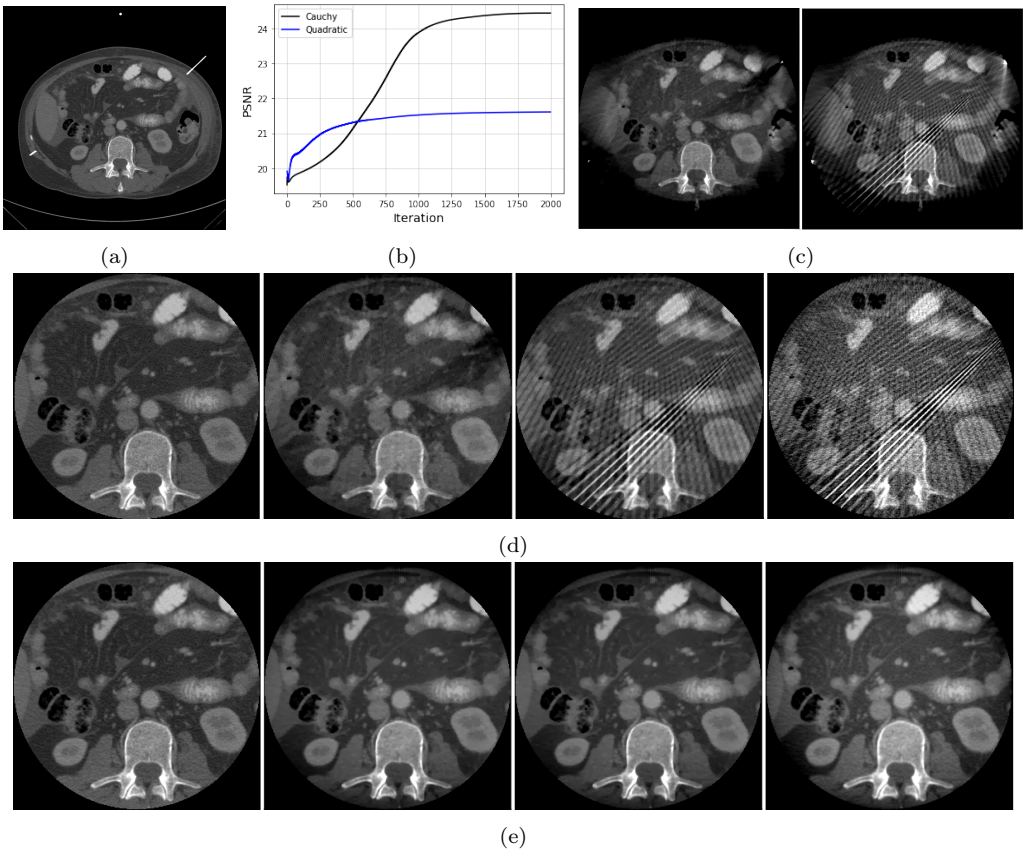


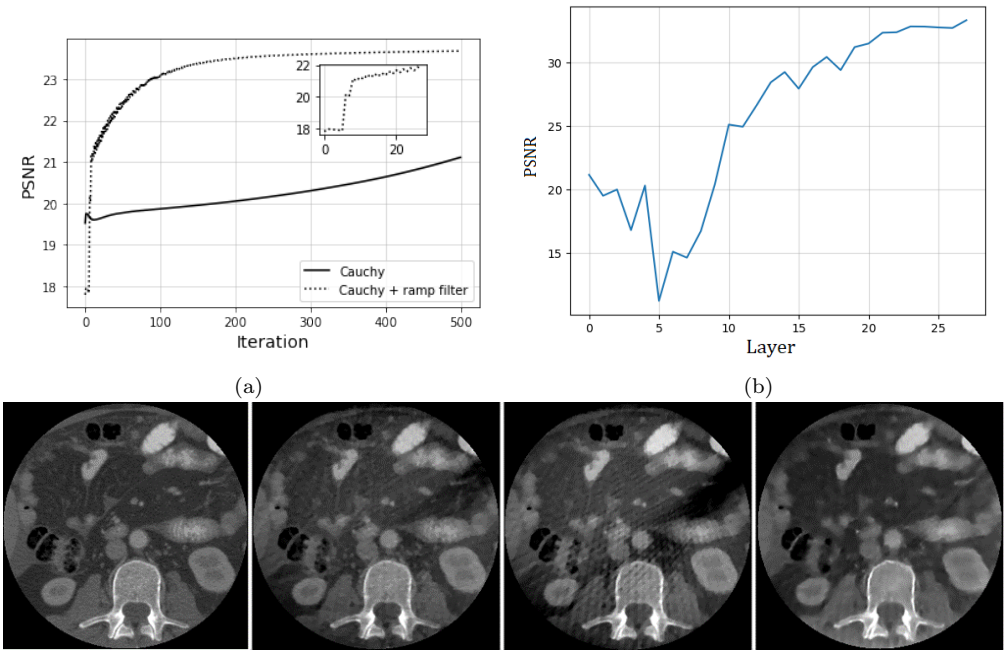
Figure 8.7: (a) Ground truth $\overline{x}_{\mathrm{P}}$. (b) Evolution of the PSNR along iterations using a Cauchy or quadratic data term for 110 projections. (c) Reconstructed extended ROIs using 110 projections, a Cauchy data fidelity, and a quadratic data fidelity. (d) Reconstructed ROIs using 110 projections. (e) Reconstructed ROIs using 600 projections. From left to right: Ground truth, reweighted DBFB with Cauchy fidelity, DBFB with quadratic fidelity, FBP.

8.4 Results

8.4.1 Assessing the benefits of the Cauchy fidelity term

Figure 8.7c shows the full reconstructed images on grid G of size 400×400 obtained using the DBFB algorithm with a quadratic fidelity term and the reweighed DBFB algorithm with a Cauchy fidelity on a test instance of the Abdomen dataset (shown in Figure 8.7a). Since the two data fidelity terms can be put into our optimization framework (Algorithm 1-Algorithm 2), the comparison is straightforward. Figure 8.7d shows the corresponding ROIs as well as the FBP reconstruction. The full image contains two intense objects out of the ROI and at the border of the reconstruction grid. In the solution obtained using the quadratic data fidelity term, the reduction of sub-sampling streaks is selective; only the streaks originating from objects within G have been eliminated in the ROI. When trading the quadratic term with a Cauchy term, as we proposed, the intensity of these streaks is reduced. This artifact reduction translates into an improvement of the PSNR as shown in Figure 8.7b. Figure 8.7e shows the ROIs obtained using the same reconstruction methods and grid size when increasing the number of projections from 110 to 600. The images now look identical and close to the ground truth. This observation highlights that, for relatively 'clean' data (no modeling of beam hardening

and scattering), the benefits of using a Cauchy fidelity over a quadratic fidelity emerge when data is sub-sampled.
We have shown that by using a regularized cost function with a Cauchy fidelity term and, thus, a more complex optimization framework, we can successfully reconstruct truncated data on a short grid, and that the reconstruction is at least as good as the one obtained with quadratic fidelity and even better when the data are sub-sampled.



(a) (b)

(c)

Figure 8.8: (a) Evolution of the PSNR along iterations using a Cauchy fidelity term with and without the ramp filter for the example of Fig Figure 8.7a. (b) Evolution of the PSNR along layers in U-RDBFB for the example of Fig Figure 8.7a. (c) Reconstructed ROIs. From left to right: Ground truth, Cauchy with ramp filter (500 iterations of the modified reweighted DBFB), Cauchy with ramp filter (28 iterations of the modified reweighted DBFB), U-RDBFB.

8.4.2 Comparing iterative RDBFB algorithm with U-RDBFB

Figure 8.8a shows the evolution of the PSNR along the iterations when inserting the ramp filter in the reweighted DBFB algorithm, re-tuning the regularization strength, and still using the same data. The PSNR stagnates around 300 iterations with the ramp filter while it stagnates around 12500 iterations without it (see Figure 8.7b). Thus applying the ramp filter on the reprojection error before backprojection can empirically accelerate convergence without degrading the solution (reconstructed ROI displayed in Figure 8.8c) in an early stopping scenario. This motivates our translation of a data iteration of DBFB to a data layer of U-RDBFB which embeds the ramp filter. It also provides empirical evidence that performance can be optimized by introducing mismatched adjoints without learning.
U-RDBFB also includes learned parameters, especially adjoints to the STV operators. It performs a total of 28 RDBFB iterations. Figure 8.8c compares the reconstruction ROI obtained with U-RDBFB, 500, and 28 iterations of the reweighed DBFB algorithm with the ramp filter. We see that after 28 iterations of the reweighed DBFB algorithm, there is a local offset near the intense object, and some streaks remain. On the contrary,

the image obtained with U-RDBFB does not contain these artifacts. It is similar to the ground truth and slightly smoother than the reconstruction after 500 iterations of the reweighed DBFB algorithm but with approximately 18 times fewer iterations.

8.4.3 Comparing U-RDBFB with deep learning methods on the Abdomen dataset

Metrics	U-RDBFB	U-net	PPN	PD-net	ISTA-net	RDBFB
PSNR	**38.7**	38.6	38.6	36.5	38.1	33.9
SSIM	**0.981**	0.972	0.974	0.956	0.975	0.903
MAE ($\times 10^{-3}$)	**3.54**	4.81	3.94	5.53	4.97	8.41
PieApp	**0.389**	0.501	0.603	0.599	0.614	0.653

Table 8.2: Quantitative assessment of the reconstructed ROIs. Mean values computed over the test set of the Abdomen dataset.

Metrics	U-RDBFB	U-net	PPN	PD-net	ISTA-net	RDBFB
PSNR	**31.7**	16.8	18.3	17.6	23.2	21.1
SSIM	**0.979**	0.903	0.845	0.765	0.956	0.908
MAE ($\times 10^{-3}$)	**2.17**	7.86	4.85	6.47	3.70	6.25
PieApp	**0.276**	0.881	0.941	0.984	0.650	0.476

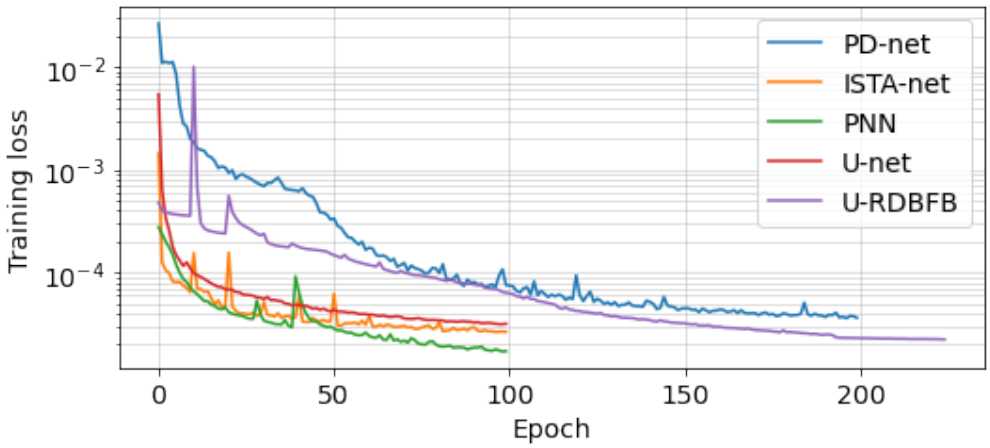
Table 8.3: Quantitative assessment of the reconstructed ROIs. Mean values computed over the test set of the Head dataset.

Metrics	U-RDBFB	U-net	PPN	PD-net	ISTA-net	RDBFB
PSNR	**27.3**	25.1	26.7	25.2	25.8	26.1
SSIM	**0.856**	0.656	0.736	0.736	0.817	0.848
MAE ($\times 10^{-3}$)	16.6	44.1	51.4	25.4	18.4	**16.5**
PieApp	**0.267**	1.166	1.255	1.324	0.894	0.158

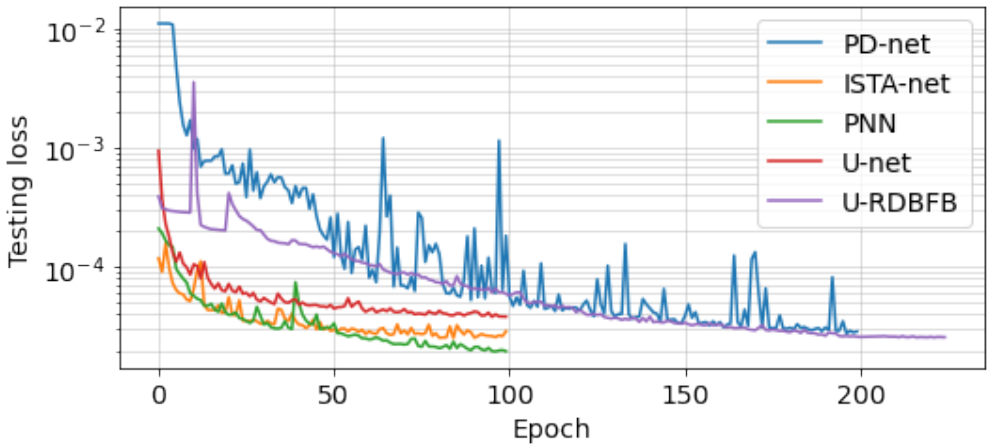
Table 8.4: Quantitative assessment of the reconstructed ROIs. Mean values computed over the testing set of the Geometrical dataset.

Table 8.2 reports the performance of U-RDBFB compared to U-net and other deep unfolding networks on the testing set of the Abdomen dataset and Figure 8.9a-Figure 8.9b display the training and testing losses as a function of the number of epochs for all these networks. U-RDBFB performs, on average, better than the other unfolding networks (PNN, PD-net, and ISTA-net) and U-net for all considered metrics. We note that the peaks in the training and testing losses associated with U-RDBFB correspond to the addition of a new data layer during incremental training.
Figure 8.11a illustrates the reconstructed ROIs for four examples from the test set of the Abdomen dataset. The FBP reconstruction is also displayed as it is also the input of U-net. The figure confirms that U-RDBFB reduces streaks more effectively than the other unfolding networks. At first sight, the images produced by U-net have fewer artifacts than most deep unfolding networks. However, in the second-row and fourth-row images, U-net introduces an artificial dark background. This observation highlights that U-net can hallucinate structures under the sub-sampling streaks of the FBP input. Unfolding networks avoid these hallucinations; by simply alternating between U-net and several

consistency layers, PNN already minimizes this effect.
Figure 8.10b shows the complete reconstruction on grid G for all unfolding networks. In all cases, since the training loss acts on the ROI only, the exterior is always poorly reconstructed (with ISTA-net, it is very sparse).



(a)

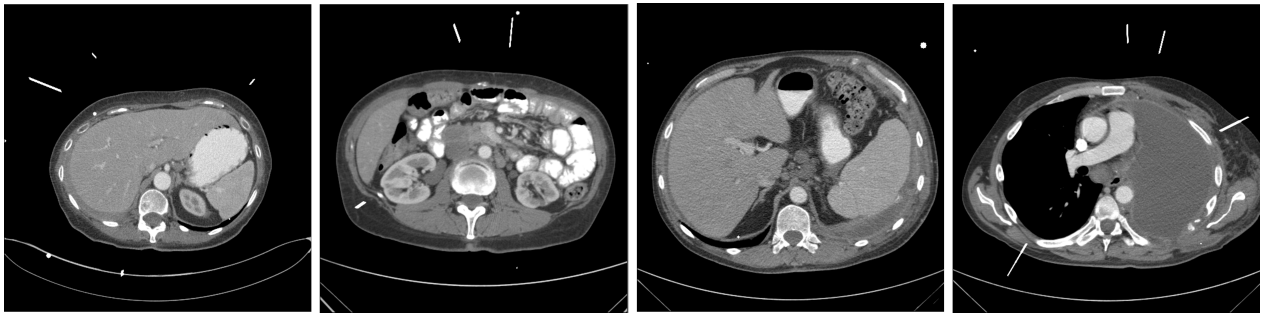


(b)

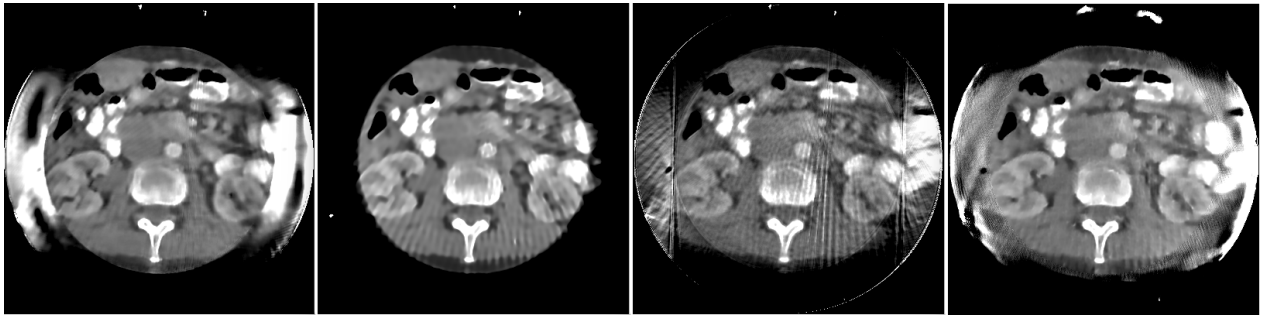
Figure 8.9: (a) MSE on the training set as a function of the epoch number. (b) MSE on the testing set as a function of the epoch number.

8.4.4 Changing the testing set

We now evaluate the generalization ability of the methods trained on the Abdomen dataset, using examples from the Head and Geometrical datasets. In Table 8.3 and Table 8.4, we report the performance of the trained networks when tested on the Head and the Geometrical datasets. U-RDBFB outperforms the competing unfolding networks for both datasets. ISTA-net is second-best on the Head dataset, and the iterative algorithm ranks second-best on the Geometrical dataset.
The four reconstructed images displayed in Figure 8.12a-Figure 8.12b confirm this trend. For the Head dataset, U-net performs poorly relative to U-RDBFB on all metrics except for SSIM, where the two methods are rather close. One explanation is that it introduces an offset in some images while limiting the streaks (cf second head image). However, when applied to the geometrical images, it yields unwanted background patterns that strongly degrade our metrics. Offsets and background artifacts are also visible with most unfolding networks, especially PNN and PD-net, except for U-RDBFB. We still note that the head images and the first geometrical image obtained with U-RBFB have a slightly patchy look, often characteristic of TV regularization. This shows that U-RDBFB retains the characteristics of the original optimization problem, avoiding the generation of unexpected content as is possible with U-net.

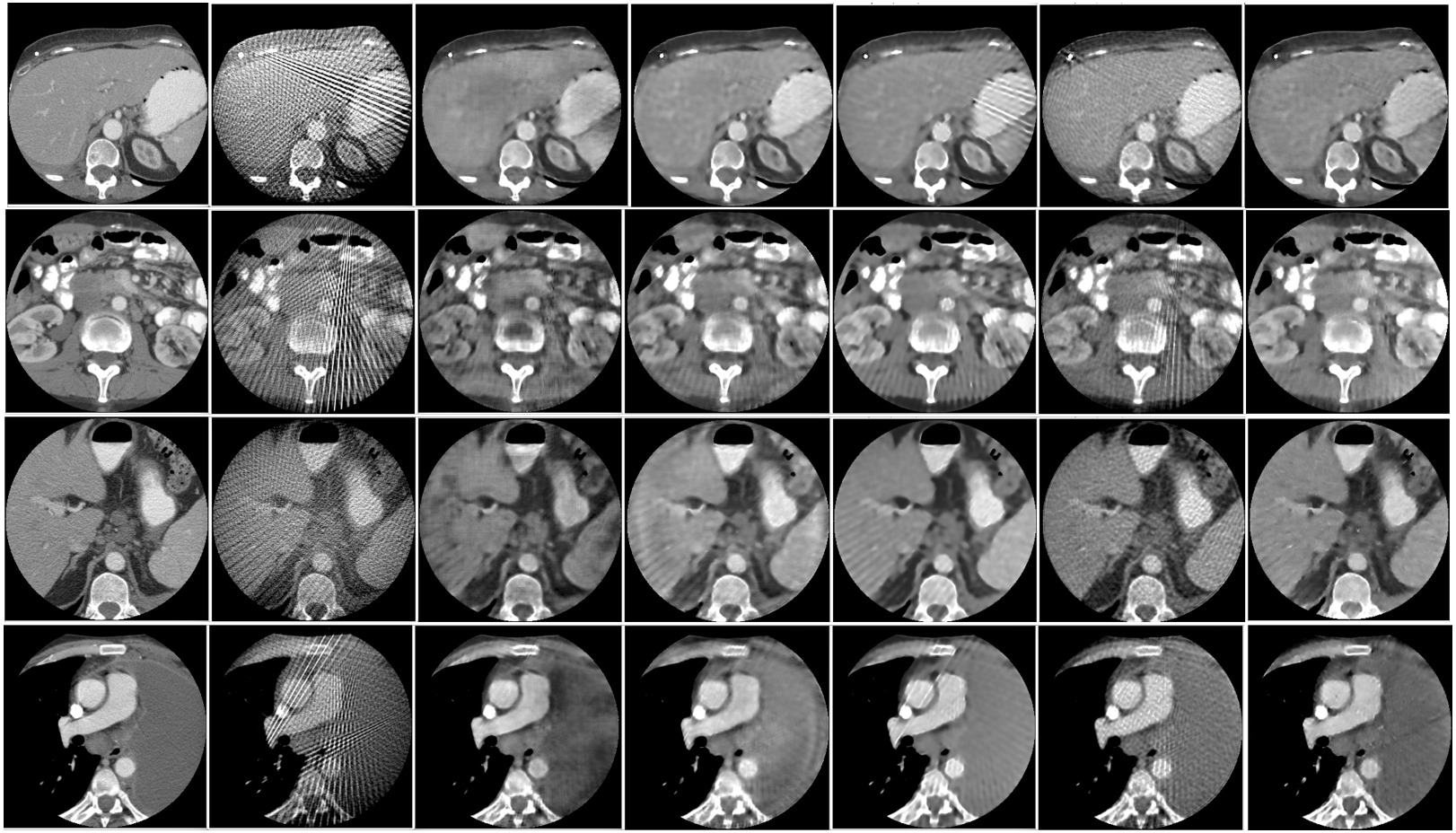


(a) Four ground truths of the testing set of the Abdomen dataset



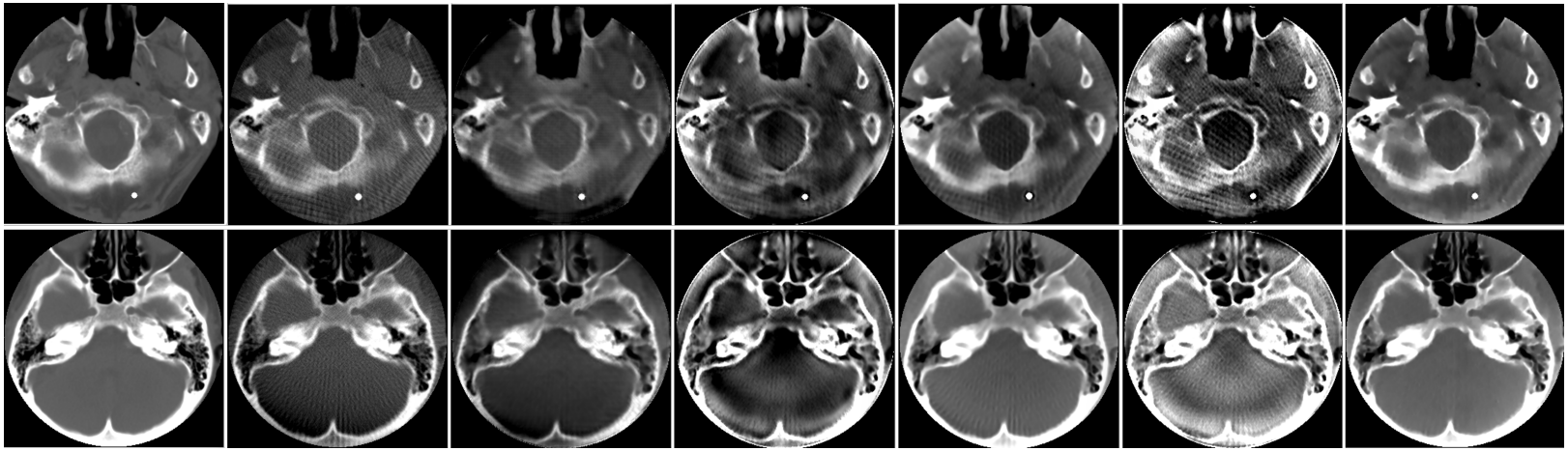
(b) Reconstructed images on grid G using PNN, ISTA-net, PD-net, and U-RDBFB.

Figure 8.10: Evaluation on the Abdomen dataset (1).

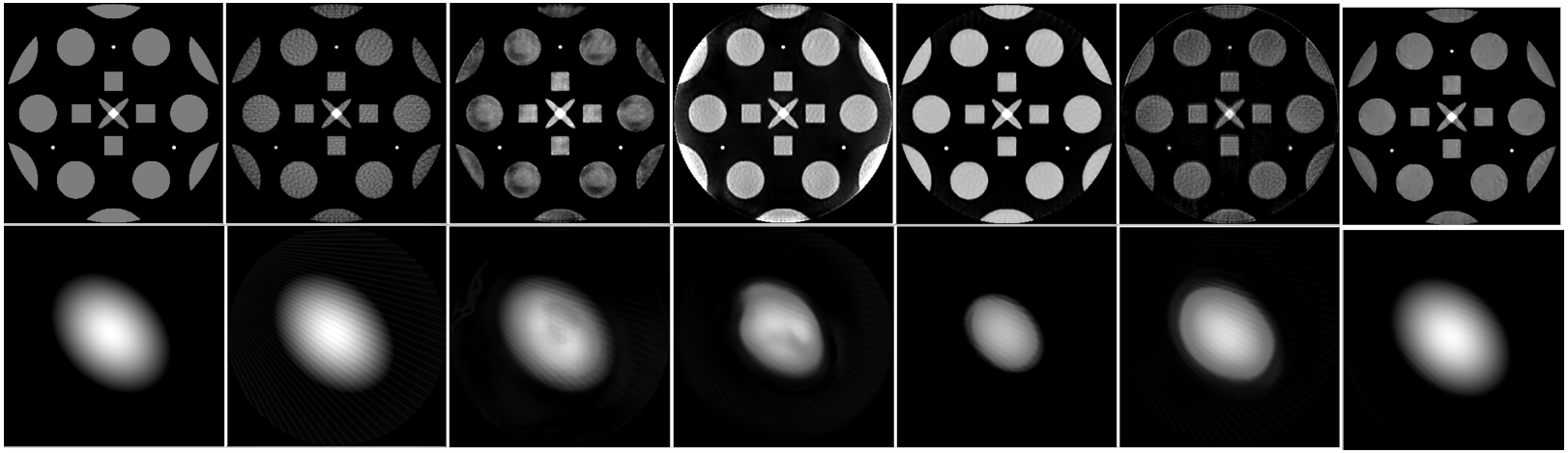


(a) Reconstructed ROIs using deep learning methods on four examples in the test set of the Abdomen dataset. From left to right: $\overline{x}_{\mathrm{ROI}}$, FBP, U-net, PNN, ISTA-net, PD-net, U-RDBFB

Figure 8.11: Evaluation on the Abdomen dataset (2).



(a) Reconstructed ROIs using different deep learning methods on four examples in the test set of the Head dataset. From left to right: $\overline{x}_{\mathrm{ROI}}$, FBP, U-net, PNN, ISTA-net, PD-net, U-RDBFB



(b) Reconstructed ROIs using different deep learning methods on four examples in the test set of the Geometrical dataset. From left to right: $\overline{x}_{\mathrm{ROI}}$, FBP, U-net, PNN, ISTA-net, PD-net, U-RDBFB

Figure 8.12: Evaluation on the Head and Geometrical datasets.

8.5 Conclusion

In this chapter, we proposed an iterative reweighted algorithm (RDBFB) where each inner optimization problem is solved using dual block coordinate forward-backward iterations, and an unfolded version of it (U-RDBFB), yielding a neural network for ROI reconstruction from a few measurements. These methods include a convex surrogate to a Cauchy data fidelity and a TV-based regularization to limit sub-sampling streaks originating from inside and outside the reconstruction grid.

Our results show that U-RDBFB outperforms its iterative counterpart as measured in PSNR, SSIM, and PieApp. It not only presents results similar to its iterative counterpart RDBFB in terms of streaks reduction, but it also recovers image details in a much lower number of iterations. We note that U-RDBFB leads to smoother images: the noise is reduced, but the resolution of the image is decreased. This could be explained by the fact that U-RDBFB is designed to minimize the ℓ_2 norm of the error with the noiseless ground truth, using very few regularization layers; it tends to selectively smooth some parts of the image to remove remaining artifacts. The smoothing also happens with other deep learning networks. Hence it is interesting to check other loss functions for training and investigate the relation between the number of layers (number of K, N_k and distribution of $\mathcal{L}_D$ and $\mathcal{L}_R$) and resolution.

All metrics agree that U-RDBFB improves upon learned post-processing U-net and other unfolding networks for our Abdomen dataset. The U-net was often associated with a high PSNR compared to the other reconstructions, but this was not always reflected by the PieApp metric. This can be explained by the learned post-processing being limited by the information content of the FBP input while the unfolding networks act directly with the information content of the data, which is greater than that of the FBP.

The computation time for U-RDBFB was about 200 ms in GPU for a 400×400 reconstruction grid. This is much faster than the iterative reconstruction, which, in our case, requires around 180 s after the regularization parameters have been selected, but slower than other deep learning methods (38 ms for U-net, 68 ms for PNN, 85 ms for PD-net, 74 ms for ISTA-net).

U-RDBFB contains fewer learnable parameters than all the other networks. Thanks to our incremental training strategy, training U-RDBFB was also found to be easier than other unfolding networks, such as PNN, whose stability highly depends on the initialization for parameter λ and the learning rate. Optimizing the architecture and, more precisely, the number of parameters of a neural network is key to transferring its performance to out-of-distribution examples, as shown on the Head and Geometrical datasets. This characteristic could help training from a low amount of data or even apply a network trained on synthetic data to real data. Generally, deep unfolding networks are introduced mainly to ensure data consistency through H and embed a fast optimization scheme for fast inference. Our results suggest that including additional a priori knowledge can further boost the performance of deep-learning-based techniques. Note that the structure of U-RDBFB was not extensively fine-tuned. Our results illustrate that the most straightforward choices work well in our context of ROI imaging from angularly sub-sampled data.

Our results also hinted that, even without learning, trading a quadratic fidelity for a Cauchy fidelity and including the ramp filter is still of interest to improve reconstruction and reduce the number of iterations.

Here, the Cauchy fidelity term aimed to discard the data incompatible with the a priori support of the object embedded in our IR criterion. Still, in other contexts, it could be

used to discard the data most degraded by physical effects such as noise and beam hardening strongly affecting the metal. Our work focused on 2D X-ray CT reconstruction on simulated data. We, however, acknowledge that our target remains the 3D cone-beam geometry.

9 | Conclusions

9.1 Contributions for CBCT reconstruction

In this thesis, we extended the use of fixed-point algorithms for C-arm CBCT reconstruction in interventional radiology. We considered two scanning configurations of clinical significance to reduce dose and scanning time: the limited-angle (Chapter 7) and the limited-density (Chapter 4, Chapter 5, Chapter 8) acquisitions. Iterative reconstruction methods are seldom used in clinical contexts despite compensating for missing data through *a priori* information and thus avoiding the necessity of expensive hardware counter-measures to under-sampling. CT practitioners have used "tricks" to improve the convergence and/or parameterization of these methods. However, these tricks often translate into replacing the hessian of the quadratic data fidelity term with an asymmetrical operator in the optimization algorithm used for reconstruction.
Throughout this thesis, we investigated ways of using iterative reconstruction methods without theoretical compromises in a clinical practice setting.

Our results in Chapter 4 and Chapter 5 proved that the convergence of a panel of proximal algorithms could be retained when the adjoint backprojector is replaced by a surrogate operator. A simple way to ensure convergence is to change the acquisition model by including an additional small quadratic term and tune the algorithm's parameters according to our convergence results. The magnification-driven interpolation proposed in Chapter 6 demonstrated that modeling the varying sampling variation over the Cartesian grid was key to improving the discretization model's symmetry for projection and backprojection. While the magnification-driven interpolation was developed for a cone-beam geometry with a flat panel detector, it can be adapted to other geometries. For instance, the native geometry of diagnostic CT with a curved detector cannot be described with projection matrices and homographies. As we showed for the distance-driven model, the local sampling steps can be estimated from a geometrical description, and the magnification-driven interpolation is applicable. Additionally, a cone-beam geometry with a very large source-to-detector distance is close to parallel geometry. One should not consider that the issue of magnification vanishes in that case. This issue vanishes only happens when the detector is parallel to the volume. However, sub-sampling arises as soon as the detector plane is not aligned with the Cartesian grid, and in this case, our framework maintains all of its advantages.
The results of these three chapters have thus either accounted for or made obsolete the use of tricks such as unmatched projector/backprojector.

Then, in Chapter 7, our DTV decomposition method allowed for reducing the acquisition trajectory as shown in two real CBCT cases of different contexts and the extreme case of acquisition not recording the bull's eye view. Although the latter is not a clinical constraint per se, the method also produced conclusive results. In practice, only a limited number of trajectories are pre-calibrated on the C-arm systems. Since calibrated trajectories may not record the optimal view, the robustness of our DTV decomposition

method allows for shifting from hardware constraints, i.e., having a generic calibration method for the C-arm system to computation time constraints.

Finally, inspired by accelerated reconstruction algorithms which vary the strength of the regularization along the iterations, in Chapter 8, we advanced the idea of finding an optimal untied parameterization further by unfolding a proximal algorithm for ROI reconstruction from limited-density acquisitions. Parameters of both the algorithm and cost function were learned. Thanks to the combined effect of a Cauchy fidelity term with a TV-based regularization, unmatched adjoints, and a learned parametrization, undersampling artifacts related to structures inside and outside of the reconstruction grid were successfully limited within a reduced number of iterations.

These last two chapters then offer ways to relax the matching between priors used for reconstruction and the type of data or images. The first uses decomposition to apply several regularizations that match some components of the images instead of one regularization for the superposition of all components. The second identifies and lowers the influence of problematic parts of the data to better comply with the prior.

9.2 Theoretical contributions

Initially motivated by practical constraints in C-arm CBCT imaging, this thesis has also led to theoretical contributions built on two observations.
First, iterative reconstruction methods have been designed to solve minimization problems. The formulation as a minimization problem is convenient because it offers a natural way of including *a priori* information, and the resulting solution is well-characterized. However, the closest ground truth estimation is not ensured to be the minimizer of any cost function.
Second, preconditioning is the most effective way of accelerating iterative algorithms, but its requirements appear too restrictive in many applications such as CT reconstruction. Therefore, decoupling matched operators, such as linear operators and their adjoints or a preconditioning metric and its inverse, in the iterations of a proximal algorithm gives rise to non-minimization equilibrium problems. By studying the behavior of a panel of proximal algorithms with such mathematical deviations, we have provided ways to include more information into these algorithms to accelerate them and/or reach a solution with minimum discrepancy. Although each proximal algorithm requires a separate study, a unique mathematical approach based on fixed-point theory has allowed the demonstration of their convergence. The results of Chapter 4 and Chapter 5 contribute to the shift in paradigm from minimization problems to equilibrium problems, which also arise in neural network architectures. They are in line with the recent efforts to design [190] and learn [15, 36, 37, 171] more expressive variants of well-known optimization schemes while keeping convergence guarantees. An important consequence is that taking unmatched adjoints and thus non-symmetrical surrogates of the hessian into account validates approximate inversion as an alternative to classical preconditioning, which is built, and therefore limited, to respect symmetry while approximate inversion does not.

9.3 Discussion

The considered unmatched algorithms were shown to converge to a fixed point different from the minimizer of a function. We characterized the fixed point as a perturbed solution at a certain distance from the original minimizer, but this characterization may still appear incomplete.
Our upper bounds on distance highlighted a trade-off between proximity to the minimizer and bias introduced by the regularization. To avoid using a strong regularization, large negative spectral values should be excluded from the spectrum of the linear operators involved in our convergence results.
Focusing on the magnifications arising in cone-beam geometry highlighted the reasons for the non-symmetry of pre-existing projectors and backprojectors based on linear interpolation. Our results from Chapter 4 and Chapter 5 with the magnification-driven interpolation could be combined, leading to different trade-offs between speed and symmetry and keeping lower values of κ. Using an approximate inverse of the projector instead of the backprojector requires a case-by-case evaluation. However, this is worth considering, as using FDK instead of the backprojection is more efficient and eases the parameterization.
Regarding the DTV decomposition method, a limited-angle acquisition lacks so much data that the generated fixed point, whether the minimum of a function or at a certain distance, depends on the relevance of the *a priori* knowledge to the real situation and can be very far from reality. There is, therefore, some doubt about the generalizability of the method to clinical practice. One very encouraging observation is that a single parameterization was sufficient in all cases, from the most obvious to the most difficult. However, if effective in all contexts, the DTV decomposition method induces a rate of recovery of the needles strongly dependent on the context. For both our results for limited-angle and limited-density acquisitions, we acknowledge that more extensive tests that include different sources of inconsistencies in the data should be performed to reach any clinical acceptance.

In the next section, we suggest several extensions of the aforementioned contributions.

9.4 Perspectives

Our analysis of mismatched forms of proximal algorithms could be further extended, as suggested hereinafter.

Extend the stability analysis of the algorithms under an adjoint mismatch to more general minimization problems:

The theoretical results obtained in Chapter 4 and Chapter 5 hold for penalized least-squares cost functions. Our considered form of the Combettes-Pesquet algorithm allows for using a Poisson data fidelity term, but then a quadratic regularization must be kept. Thus, it would be interesting to extend our analysis to more general forms of convex data fidelity terms arising from the Poisson modeling of the noise on pre-log data or based on the Huber function used in Chapter 8. In the current CT clinical practice, Poisson noise models are not used. However, with the advent of photon counting detectors, which cut off electronic noise, instead of energy-integrating detectors, Poisson models are likely to

become popular in the upcoming years for low-dose quantitative imaging. In CT, the use of efficient discretization rationales equivalently suited for projection and backprojection alleviates the need for using unmatched pairs of projection and backprojection. However, this motivation remains in other imaging modalities such as SPECT, where bypassing the modeling of the attenuation in the backprojection is beneficial to the convergence speed, and where the noise follows a Poisson distribution.
Since the choice of the regularization is critical to the reconstruction quality, more complex regularization schemes formulated as the sum of several functions (as in the semilocal total variation in Chapter 8) could also be interesting to consider. For instance, by using other algorithms for finding a zero of a sum of more than two maximally monotone operators [61] under an adjoint mismatch.

Improve the stability analysis of the algorithms under an adjoint mismatch:

In our analysis of PGA and primal-dual proximal algorithms, the error bounds on the distance between a solution of the original minimization problem and a fixed point of the new perturbed schemes are provided under strong convexity assumptions for the functions involved. A first improvement should be to relax the assumption of strong convexity, as done in our convergence results. Then, an extension of our results to the accelerated variants (with variable step sizes) of the primal-dual Chambolle-Pock and Combettes-Pesquet methods, which exhibit in these settings better convergence rates, could be conducted using results beyond fixed point theory as proposed in [140].

Consider other mismatched forms of the same primal-dual algorithms:

In our analysis of Chapter 5, we could have explored other ways of introducing a mismatch in the iterations of the primal-dual algorithms. For example, the Condat-Vũ algorithm is an instance of a preconditioned proximal gradient algorithm on a product space. Therefore, novel deviations from this scheme could be analyzed as an instance of our unmatched preconditioned PGA scheme of Chapter 4. In particular, the adjoint mismatched considered in [140] could be studied in such a way in the case of fixed step sizes.

We propose the following improvements and future leads related to our magnification-driven tomographic operators of Chapter 6.

Extend the magnification-driven approach to higher degree basis functions:

When the image expansion cannot be increased due to storage and time constraints, the form of the expansion elements for the volume must be optimized. In this thesis, we used exclusively centered B-splines of low orders, but alternative basis functions can provide additional flexibility between precision and computation. Splines of higher degrees could, for instance, be tested to quantify the loss in precision associated with the use of low-order splines. When using higher order B-splines, the formulation proposed in [153] for least-squares approximation might be more suitable. This formulation is based on a finite difference method applicable to splines of arbitrary degrees. Its main advantage is that it avoids computing the footprint function.

Produce more extensive tests of the magnification-driven framework on real clinical data:

We have shown the performance of the magnification-driven framework using piecewise constant numerical and physical phantoms. As a next step, the translation of resolution improvements obtained for analytical methods to clinical exams should be investigated, to be extended to iterative reconstruction in an early stopping scenario. However, we highlight that it might be difficult to directly link the single technical choice of the interpolation to any improvements in image quality that depend on the clinical task and sophisticated iterative reconstruction schemes. Deteriorated data may require a strong regularization which could compensate for the insufficiencies of the discrete forward projector. Our framework could also be interesting for high-resolution applications involving photon counting detectors which allow for the use of smaller bins thanks to the removal of septa from the detector's design.

Next, we suggest two possible future directions related to needle reconstruction from limited-angle acquisitions.

Accelerate the convergence of our decomposition method for needle reconstruction using deep unfolding:

The experiments performed in Chapter 7 demonstrated that between 300 and 1000 iterations of FISTA were needed to reconstruct needles whose bull's eye view is not sampled in the measurements and to interpolate the missing edges in the volume domain. We also observed that the number of iterations should be increased when the number of components increases. Translating our optimization algorithm to an unfolding architecture would be one next step for interpolating the missing edges more rapidly and tuning the regularization strength for each component. By embedding the projector and thus the scanning angular range into the network's architecture, we would expect the network to differentiate the directions of missing edges from that of visible edges, thus avoiding interpolation between random points.

Investigate the robustness of the decomposition approach with a variety of needle sizes and forms:

To reach clinical practice, we should further test our method with different needles, which vary in size and form (curved or thick) that influence the physical effects affecting needles.

Finally, building on the neural network architecture presented in Chapter 8, we propose the following research directions related to deep learning methods.

Investigate the benefits of embedding the projector and backprojector in the network's architecture:

Deep unfolding for reconstruction embeds the tomographic operators in the network's architecture. Other approaches, such as the plug-and-play approach, focus on learning the regularization in a standalone way. The learned regularization is then inserted into

a proximal algorithm. Even if the plug-an-play approach is likely to require more training data than the unfolding approach, we could evaluate and quantify the benefits of including the tomographic operators in the learning task. Such an evaluation would be conducted using task-based metrics with the aid of radiologists rather than perceptual image quality metrics, as radiologists might prefer some artifacts over others.

Evaluate the impact of acceleration techniques in U-RDBFB:

Following the work of [131] in the case of image denoising, we believe that studying the impact of accelerated schemes - through preconditioning or inertial steps - on learning performance would be useful to optimize the architecture of our unfolding approach.

Explore the robustness of architecture of type U-RDBFB:

First, we have not conducted any robustness analysis of our network: we assumed overall stability by staying close to the original stable iterative algorithm. Still, we should study the effects of a perturbation on the input of our architecture.
Second, to bridge the gap from our results to clinical applications, our network needs to be retrained and evaluated with changes in the acquisition pattern and detector settings (variation in the number of views, the number and size of detector bins), as well as the presence of other artifacts. We could investigate the trade-off between the number of measurements and the dose required to reach clinically satisfying images and the resolution/total number of layers trade-off. Deep unfolding has the potential to produce an accelerated reconstruction adapted to pre-corrected data by reproducing the pre-processing imaging chain to generate the datasets. However, we believe it is unlikely that unfolding schemes could replace all pre-corrections.

Reduce the memory footprint of U-RDBFB:

A significant impediment to the 3D adaptation of U-RDBFB is its memory footprint when performing end-to-end backpropagation. This memory footprint grows linearly with the depth of the network, and reducing the number of layers through acceleration techniques may not be sufficient to limit the memory footprint. Recent progress in decreasing the GPU memory footprint of deep unfolding methods can be found in [188], where the authors proposed to use invertible neural networks in the PD-net architecture of [3]. As outlined in [103], one of the key benefits of an invertible network is that the depth of the network can be increased while maintaining a constant memory footprint. The number of unfolded iterations in a learned iterative method can then be increased. Our current unfolding architecture could be easily translated into a Deep Equilibrium model, which economizes memory during training thanks to the Implicit Function Theorem. Note that the question of memory limitation also arises for post-processing CNNs. Since 3D convolutions have substantial storage requirements, "2.5"D architectures could be introduced to exploit spatial coherence in volumetric data.

List of Figures

List of Tables

Bibliography

[1] F. Abboud, E. Chouzenoux, J.-C. Pesquet, J.-H. Chenot, and L. Laborelli. Dual block-coordinate forward-backward algorithm with application to deconvolution and deinterlacing of video sequences. *Journal of Mathematical Imaging and Vision*, 59(3):415–431, 2017. 182

[2] J. Adler and O. Öktem. Learned Primal-Dual Reconstruction. *IEEE Transactions on Medical Imaging*, 37:1322–1332, 2018. 63, 191, 192, 212

[3] J. Adler and O. Öktem. Solving ill-posed inverse problems using iterative deep neural networks. *Inverse Problems*, 33(12):124007, 2017. 63, 178, 208

[4] E. Ahishakiye, M. Van Gijzen, J. Tumwiine, R. Wario, and J. Obungoloch. A survey on deep learning in medical image reconstruction. *Intelligent Medicine*, 1(3):118–127, 2021. 61

[5] A. Alotaibi, P. Combettes, and N. Shahzad. Solving coupled composite monotone inclusions by successive Fejér approximations of their Kuhn–Tucker set. *SIAM Journal on Optimization*, 24(4):2076–2095, Jan. 2014. 97

[6] J. Ambrose and G. Hounsfield. Computerized transverse axial tomography. *The British Journal of Radiology*, 46(542):148–149, Feb. 1973. 19, 45

[7] A. H. Andersen and A. C. Kak. Simultaneous Algebraic Reconstruction Technique (SART): A Superior Implementation of the Art Algorithm. *Ultrasonic Imaging*, 6(1):81–94, Jan. 1984. 45

[8] R. Anxionnat, S. Bracard, X. Ducrocq, Y. Trousset, L. Launay, E. Kerrien, M. Braun, R. Vaillant, F. Scomazzoni, A. Lebedinsky, and L. Picard. Intracranial aneurysms: clinical value of 3D digital subtraction angiography in the therapeutic decision and endovascular treatment. *Radiology*, 218(3):799–808, Mar. 2001. 27

[9] F. Arcadu, M. Stampanoni, and F. Marone. On the crucial impact of the coupling projector-backprojector in iterative tomographic reconstruction. *arXiv preprint arXiv:1612.05515*, 2016. 65

[10] S. Arridge, P. Maass, O. Öktem, and C. Schönlieb. Solving inverse problems using data-driven models. *Acta Numerica*, 28:1 – 174, 2019. 63

[11] A. Aspri, S. Banert, O. Öktem, and O. Scherzer. A data-driven iteratively regularized landweber iteration. *Numerical Functional Analysis and Optimization*, 41:1190 – 1227, 2018. 63

[12] J.-F. Aujol, G. Gilboa, T. F. Chan, and S. Osher. Structure-texture image decomposition—modeling, algorithms, and parameter selection. *International Journal of Computer Vision*, 67:111–136, 2006. 158

[13] S. Bai, Z. Kolter, and V. Koltun. Deep equilibrium models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 63

[14] S. Bai, V. Koltun, and Z. Kolter. Multiscale deep equilibrium models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 63

[15] S. Banert, J. Rudzusika, O. Öktem, and J. Adler. Accelerated forward-backward optimization using deep learning. *arXiv preprint arXiv:2105.05210*, 2021. 63, 204

[16] H. Bauschke and P. Combettes. *Convex Analysis And Monotone Operator Theory In Hilbert Spaces*. Springer, New York, 2nd edition, 2017. 49, 56, 57, 58, 72, 73, 75, 92, 99, 100, 102, 103, 104, 105, 112, 113, 114, 183

[17] I. Bayram and M. Kamasak. Directional total variation. *IEEE Signal Processing Letters*, 19:781–784, 2012. 84, 158

[18] J. Beaudry, P. Esquinas, and C.-C. Shieh. Learning from our neighbours: a novel approach on sinogram completion using bin-sharing and deep learning to reconstruct high quality 4D CBCT. In *Medical Imaging 2019: Physics of Medical Imaging*, volume 10948, pages 1025–1035. SPIE, 2019. 61

[19] S. Becker and J. Fadili. A quasi-newton proximal splitting method. *Advances in neural information processing systems*, 25, 2012. 54

[20] S. Becker, J. Fadili, and P. Ochs. On quasi-newton forward-backward splitting: Proximal calculus and convergence. *SIAM Journal on Optimization*, 29(4):2445–2481, 2019. 88

[21] J. Bect, L. Blanc-Féraud, G. Aubert, and A. Chambolle. A ℓ_1-unified variational framework for image restoration. In *Proceedings of the 8th European Conference on Computer Vision (ECCV 2004)*, volume 3024 of *Lecture Notes in Computer Science*, pages 1–13, May 2004. 54

[22] D. Benz, S. Ersözlü, F. L. A. Mojon, M. Messerli, A. Mitulla, D. Ciancone, D. Kenkel, J. Schaab, C. Gebhard, A. Pazhenkottil, P. Kaufmann, and R. Buechel. Radiation dose reduction with deep-learning image reconstruction for coronary computed tomography angiography. *European Radiology*, 32(4):2620–2628, Apr. 2022. 61

[23] M. Bertero, P. Boccacci, and C. De Mol. *Introduction to inverse problems in imaging*. CRC press, 1998. 19

[24] C. Bertocchi, E. Chouzenoux, M.-C. Corbineau, J.-C. Pesquet, and M. Prato. Deep unfolding of a proximal interior point method for image restoration. *Inverse Problems*, 2020. 63, 178

[25] J. Bian, J. Siewerdsen, X. Han, E. Sidky, J. Prince, C. Pelizzari, and X. Pan. Evaluation of sparse-view reconstruction from flat-panel-detector cone-beam CT. *Physics in Medicine and Biology*, 55(22):6575–6599, Nov. 2010. 49

[26] S. Bonettini, F. Porta, V. Ruggiero, and L. Zanni. Variable metric techniques for forward–backward methods in imaging. *Journal of Computational and Applied Mathematics*, 385:113192, 2021. 89

[27] F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. Sagastizábal. A family of variable metric proximal methods. *Mathematical Programming*, 68(1):15–47, 1995. 54, 88

[28] C. Bouman and K. Sauer. A generalized gaussian image model for edge-preserving map estimation. *IEEE Transactions on image processing*, 2(3):296–310, 1993. 48

[29] S. Boyd. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2010. 55

[30] R. I. Boţ, E. R. Csetnek, and A. Heinrich. A primal-dual splitting algorithm for finding zeros of sums of maximal monotone operators. *SIAM Journal on Optimization*, 23(4):2011–2036, Jan. 2013. 97

[31] R. I. Boţ, E. R. Csetnek, and D. Meier. Inducing strong convergence into the asymptotic behaviour of proximal splitting algorithms in Hilbert spaces. *Optimization Methods and Software*, 34(3):489–514, May 2019. 54

[32] L. M. Bregman. Finding the common point of convex sets by the method of successive projection. In *Doklady Akademii Nauk*, volume 162, pages 487–490. Russian Academy of Sciences, 1965. 54

[33] T. Briand and A. Davy. Optimization of image B-spline interpolation for gpu architectures. *Image Processing On Line*, 9:183–204, 2019. 136

[34] L. Briceño-Arias and P. Combettes. A monotone+skew splitting model for composite monotone inclusions in duality. *SIAM Journal on Optimization*, 21(4):1230–1250, 2011. 55, 58, 97

[35] L. Briceño-Arias and S. López. A projected primal–dual method for solving constrained monotone inclusions. *Journal of Optimization Theory and Applications*, 180(3):907–924, 2019. 97, 106

[36] T. Bubba, M. Galinier, M. Lassas, M. Prato, L. Ratti, and S. Siltanen. Deep neural networks for inverse problems with pseudodifferential operators: An application to limited-angle tomography. *SIAM Journal on Imaging Sciences*, 14(2):470–505, 2021. 63, 65, 204

[37] G. Buzzard, S. Chan, S. Sreehari, and C. Bouman. Plug-and-play unplugged: Optimization-free reconstruction using consensus equilibrium. *SIAM Journal on Imaging Sciences*, 11(3):2001–2020, 2018. 97, 204

[38] C. Caillaud and A. Chambolle. Error estimates for finite differences approximations of the total variation. Apr. 2020. 48

[39] E. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006. 48, 115

[40] Y. Censor, T. Elfving, G. Herman, and T. Nikazad. On Diagonally Relaxed Orthogonal Projection Methods. *SIAM Journal on Scientific Computing*, 30(1):473–504, Jan. 2008. 45

[41] Y. Censor, D. Gordon, and R. Gordon. BICAV: a block-iterative parallel algorithm for sparse systems with pixel-related weighting. *IEEE Transactions on Medical Imaging*, 20(10):1050–1060, Oct. 2001. 45

[42] A. Chambolle and C. Dossal. On the convergence of the iterates of FISTA. *Journal of Optimization Theory and Applications*, 166:25, 2015. 54

[43] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40(1):120–145, 2011. 57

[44] C. Chapdelaine, N. Gac, A.-M. Djafari, and E. Parra-Denis. New GPU implementation of separable footprint (SF) projector and backprojector : first results. In *The Fifth International Conference on Image Formation in X-Ray Computed Tomography*, pages 314–317, Salt Lake City, United States, May 2018. 123

[45] C. Chaux, P. L. Combettes, J.-C. Pesquet, and V. Wajs. A variational formulation for frame based inverse problems. *Inverse Problems*, 23(1):1495–1518, 2007. 48

[46] G. Chen and T. Rockafellar. Convergence rates in forward–backward splitting. *SIAM Journal on Optimization*, 7(2):421–444, 1997. 88

[47] H. Chen, Y. Zhang, Y. Chen, J. Zhang, W. Zhang, H. Sun, Y. Lv, P. Liao, J. Zhou, and G. Wang. LEARN: Learned experts' assessment-based reconstruction network for sparse-data CT. *IEEE Transactions on Medical Imaging*, 37(6):1333–1347, 2018. 63

[48] K. Chen. *Matrix preconditioning techniques and applications*, volume 19. Cambridge University Press, 2005. 54

[49] P. Chen, J. C. Huang, and X. Zhang. A primal-dual fixed point algorithm for convex separable minimization with applications to image restoration. *Inverse Problems*, 29(2):025011, 2013. 57

[50] J. Cheng, H. Wang, L. Ying, and D. Liang. Model learning: Primal dual networks for fast mr imaging. *arXiv preprint arXiv:1908.02426*, 2019. 191

[51] E. Chouzenoux, A. Jezierska, J.-C. Pesquet, and H. Talbot. A majorize-minimize subspace approach for ℓ_2-ℓ_0 image regularization. *SIAM Journal on Imaging Sciences*, 6(1):563–591, 2013. 181

[52] E. Chouzenoux, J.-C. Pesquet, and A. Repetti. Variable metric forward–backward algorithm for minimizing the sum of a differentiable function and a convex function. *Journal of Optimization Theory and Applications*, 162(1):107–132, 2014. 54

[53] R. Clackdoyle and M. Defrise. Tomographic reconstruction in the 21st century. *IEEE Signal Processing Magazine*, 27(4):60–80, 2010. 44

[54] N. Clinthorne, T. Pan, P.-C. Chiao, L. Rogers, and J. Stamos. Preconditioning methods for improved convergence rates in iterative reconstructions. *IEEE transactions on medical imaging*, 12(1):78–83, 1993. 46, 89

[55] P. Combettes, D. Dũng, and B. Vũ. Dualization of signal recovery problems. *Set-Valued and Variational Analysis*, 18(3):373–404, 2010. 84

[56] P. Combettes, D. Dũng, and B. C. Vũ. Proximity for sums of composite functions. *Journal of Mathematical Analysis and applications*, 380(2):680–688, 2011. 84

[57] P. Combettes and J.-C. Pesquet. Image restoration subject to a total variation constraint. *IEEE transactions on image processing*, 13(9):1213–1222, 2004. 93

[58] P. Combettes and J.-C. Pesquet. Primal-dual splitting algorithm for solving inclusions with mixtures of composite, lipschitzian, and parallel-sum type monotone operators. *Set-Valued and Variational Analysis*, 20:307–330, 2011. 58, 97

[59] P. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer, 2011. 19, 92

[60] P. Combettes and J.-C. Pesquet. Deep neural network structures solving variational inequalities. *Set-Valued and Variational Analysis*, 28:491–518, 09 2020. 73, 76, 103

[61] P. Combettes and J.-C. Pesquet. Fixed point strategies in data science. *IEEE Transactions on Signal Processing*, 69:3878–3905, 2021. 67, 206

[62] P. Combettes and B. Vũ. Variable metric forward–backward splitting with applications to monotone inclusions in duality. *Optimization*, 63(9):1289–1318, 2014. 54

[63] P. L. Combettes and J.-C. Pesquet. Proximal thresholding algorithm for minimization over orthonormal bases. *SIAM Journal on Optimization*, 18(4):1351–1376, 2007. 54, 66

[64] P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling and Simulation: A SIAM Interdisciplinary Journal*, 4:1164–1200, 2005. 54, 66, 89

[65] L. Condat. A direct algorithm for 1-d total variation denoising. *IEEE Signal Processing Letters*, 20(11):1054–1057, 2013. 176

[66] L. Condat. A primal–dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms. *Journal of Optimization Theory and Applications*, 158:460–479, 2013. 56, 57, 99

[67] L. Condat. Discrete total variation: new definition and minimization. *SIAM Journal on Imaging Sciences*, 10(3):1258–1290, Aug. 2017. 48

[68] L. Condat, D. Kitahara, A. C. Marcillo, and A. Hirabayashi. Proximal splitting algorithms: A tour of recent advances, with new twists. *arXiv: preprint arXiv:1912.00137*, 2020. 54, 108

[69] D. Dance. *Diagnostic Radiology Physics A handbook for teachers and students*. International atomic energy agency, 2014. 32

[70] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57(11):1413–1457, 2004. 54, 66

[71] B. De Man and S. Basu. Distance-driven projection and backprojection. In *2002 IEEE Nuclear Science Symposium Conference Record*, volume 3, pages 1477–1480 vol.3, 2002. 123

[72] B. De Man and S. Basu. Distance-driven projection and backprojection in three dimensions. *Physics in medicine and biology*, 49:2463–75, 07 2004. 60, 123, 140

[73] S. Deans. *The Radon transform and some of its applications.* Courier Corporation, 2007. 42

[74] C. Delmas, C. Riddell, Y. Trousset, E. Kerrien, M.-O. Berger, R. Anxionnat, and S. Bracard. Intra-operative 3D micro-coil imaging using subsampled tomographic acquisition patterns on a biplane C-arm system. In *Proceedings of the 4th International Conference on Image Formation in X-Ray Computed Tomography (CT meeting'16)*, Bamberg, Germany, July 2016. 157

[75] F. Deschamps, S. Solomon, R. Thornton, P. Rao, A. Hakime, V. Kuoch, and T. de Baere. Computed analysis of three-dimensional cone-beam computed tomography angiography for determination of tumor-feeding vessels during chemoembolization of liver tumor: a pilot study. *Cardiovascular and Interventional Radiology*, 33(6):1235–1242, Dec. 2010. 28

[76] Q. Ding, Y. Long, X. Zhang, and J. Fessler. Statistical image reconstruction using mixed poisson-gaussian noise model for x-ray ct. *arXiv preprint arXiv:1801.09533*, 2018. 47

[77] Y. Dong, P. C. Hansen, M. E. Hochstenbach, and A. N. R. Brogaard. Fixing nonconvergence of algebraic iterative reconstruction with an unmatched backprojector. *SIAM Journal on Scientific Computing*, 41(3):A1822–A1839, 2019. 65, 75, 77, 79

[78] D. Donoho. For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829, June 2006. 48

[79] D. Donoho and Y. Tsaig. Fast solution of ℓ_1 -norm minimization problems when the solution may be sparse. *IEEE Transactions on Information Theory*, 54(11):4789–4812, 2008. 55

[80] Y. Drori, S. Sabach, and M. Teboulle. A simple algorithm for a class of nonsmooth convex-concave saddle-point problems. *Operations Research Letters*, 43:209–214, 2015. 57

[81] T. Elfving and P. C. Hansen. Unmatched projector/backprojector pairs: perturbation and convergence analysis. *SIAM Journal on Scientific Computing*, 2018. 65

[82] T. Elfving, P. C. Hansen, and T. Nikazad. Convergence analysis for column-action methods in image reconstruction. *Numerical Algorithms*, 74(3):905–924, Mar. 2017. 45

[83] T. Eo, Y. Jun, T. Kim, J. Jang, H.-J. Lee, and D. Hwang. Kiki-net: cross-domain convolutional neural networks for reconstructing undersampled magnetic resonance images. *Magnetic Resonance in Medicine*, 80(5):2188–2201, 2018. 62

[84] H. Erdogan and J. Fessler. Ordered subsets algorithms for transmission tomography. *Physics in Medicine & Biology*, 44(11):2835, 1999. 48

[85] E. Esser, X. Zhang, and T. Chan. A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. *SIAM Journal on Imaging Sciences*, 3(4):1015–1046, Jan. 2010. 97

[86] D. Evangelista, E. Morotti, and E. L. Piccolomini. RISING a new framework for few-view tomographic image reconstruction with deep learning. *arXiv preprint arXiv:2201.09777*, 2022. 61

[87] L. F., Z. Y., J.-B. Thibault, B. De Man, M. McGaffin, and J. Fessler. Space-variant channelized preconditioner design for 3D iterative CT reconstruction. In *Proceedings of the 16th Virtual International Meeting on Fully 3D Image Reconstruction in Radiology and Nuclear Medicine*, pages 205–8, 2013. 89

[88] F. Xu and K. Mueller. A comparative study of popular interpolation and integration methods for use in computed tomography. In *Proceedings of the 3rd IEEE International Symposium on Biomedical Imaging: Nano to Macro, 2006.*, pages 1252–1255, Apr 2006. 60, 65, 115

[89] O. Faugeras. *Three-dimensional computer vision: a geometric viewpoint*. MIT press, 1993. 124

[90] L. A. Feldkamp, L. C. Davis, and J. W. Kress. Practical cone-beam algorithm. *Journal of the Optical Society of America A: Optics, Image Science, and Vision*, 1(6):612–619, Jun 1984. 44

[91] J. Fessler. Analytical tomographic image reconstruction methods. *Image Reconstruction: Algorithms and Analysis*, 66:67, 2009. 44

[92] J. Fessler and S. Booth. Conjugate-gradient preconditioning methods for shift-variant PET image reconstruction. *IEEE Transactions on Image Processing*, 8(5):688–699, 1999. 89

[93] C. Floyd. An artificial neural network for spect image reconstruction. *IEEE transactions on medical imaging*, 10(3):485–487, 1991. 62

[94] L. Fu, B. De Man, K. Zeng, T. Benson, Z. Yu, G. Cao, and J.-B. Thibault. A preliminary investigation of 3D preconditioned conjugate gradient reconstruction for cone-beam CT. In *Medical Imaging 2012: Physics of Medical Imaging*, volume 8313, pages 1051–1059. SPIE, 2012. 89

[95] L. Fu, T.-C. Lee, S. Kim, A. Alessio, P. Kinahan, Z. Chang, K. Sauer, M. Kalra, and B. De Man. Comparison between pre-log and post-log statistical models in ultra-low-dose CT reconstruction. *IEEE transactions on medical imaging*, 36(3):707–720, 2016. 47

[96] D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2(1):17–40, 1976. 55

[97] M. Ghani and C. Karl. Deep learning-based sinogram completion for low-dose CT. In *2018 IEEE 13th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, pages 1–5. IEEE, 2018. 61

[98] G. Gilboa and S. Osher. Nonlocal operators with applications to image processing. *Multiscale Modeling & Simulation*, 7(3):1005–1028, 2009. 180

[99] D. Gilton, G. Ongie, and R. Willett. Neumann networks for linear inverse problems in imaging. *IEEE Transactions on Computational Imaging*, 6:328–343, 2020. 63, 190, 192, 212

[100] D. Gilton, G. Ongie, and R. Willett. Deep equilibrium architectures for inverse problems in imaging. *IEEE Transactions on Computational Imaging*, 7:1123–1133, 2021. 63

[101] L. Gjesteby, B. D. Man, Y. Jin, H. Paganetti, J. Verburg, D. Giantsoudi, and G. Wang. Metal artifact reduction in CT: where are we after four decades? *IEEE Access*, 4:5826–5849, 2016. 59

[102] M. Goetz, M. Callstrom, J. W. Charboneau, M. Farrell, T. Maus, T. Welch, G. Wong, J. Sloan, P. Novotny, I. Petersen, R. Beres, D. Regge, R. Capanna, M. Saker, D. Grönemeyer, A. Gevargez, K. Ahrar, M. Choti, T. de Baere, and J. Rubin. Percutaneous image-guided radiofrequency ablation of painful metastases involving bone: a multicenter study. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 22(2):300–306, Jan. 2004. 30

[103] A. Gomez, M. Ren, R. Urtasun, and R. Grosse. The reversible residual network: Backpropagation without storing activations. *Advances in neural information processing systems*, 30, 2017. 208

[104] C. Graff and E. Sidky. Compressive sensing in medical imaging. *Applied optics*, 54(8):C23–C44, Mar. 2015. 48

[105] J. Gregor and J. Fessler. Comparison of sirt and sqs for regularized weighted least squares image reconstruction. *IEEE transactions on computational imaging*, 1(1):44–55, 2015. 89

[106] K. Gregor and Y. LeCun. Learning fast approximations of sparse coding. In *Proceedings of the 27th international conference on international conference on machine learning*, pages 399–406, 2010. 63

[107] H. Guo and X. Cui. Block-tridiagonal shift-variant preconditioner for iterative cone beam CT reconstruction. In *Proceedings of Fully3D*, Xian Shaaxi, China, 2017. 89

[108] H. Gupta, K. H. Jin, H. Nguyen, M. McCann, and M. Unser. Cnn-based projected gradient descent for consistent CT image reconstruction. *IEEE transactions on medical imaging*, 37(6):1440–1453, 2018. 63

[109] B. Hamelin, Y. Goussard, J.-P. Dussault, G. Cloutier, G. Beaudoin, and G. Soulez. Design of iterative ROI transmission tomography reconstruction procedures and image quality analysis. *Medical Physics*, 37(9):4577–4589, 2010. 177

[110] Y. Han and J. C. Ye. One network to solve all rois: Deep learning ct for any roi using differentiated backprojection. *Medical Physics*, 46(12):e855–e872, 2019. 177

[111] Y. Han, J. J. Yoo, and J. C. Ye. Deep residual learning for compressed sensing CT reconstruction via persistent homology analysis. *arXiv preprint arXiv:1611.06391*, 2016. 61, 177, 190

[112] Y. Huang, A. Preuhs, M. Manhart, G. Lauritsch, and A. Maier. Data extrapolation from learned prior images for truncation correction in computed tomography. *IEEE Transactions on Medical Imaging*, 40(11):3042–3053, 2021. 61

[113] A. Jalal, M. Arvinte, G. Daras, E. Price, A. Dimakis, and J. Tamir. Robust Compressed Sensing MRI with Deep Generative Priors. *Advances in Neural Information Processing Systems*, 34:14938–14954, 2021. 62

[114] E. Janssens, J. De Beenhouwer, M. Van Dael, T. De Schryver, L. Van Hoorebeke, P. Verboven, B. Nicolai, and J. Sijbers. Neural network hilbert transform based filtered backprojection for fast inline x-ray inspection. *Measurement Science and Technology*, 29(3):034012, 2018. 62

[115] J. Jelinek, M. Murphey, J. Welker, R. Henshaw, M. Kransdorf, B. Shmookler, and M. Malawer. Diagnosis of primary bone tumors with image-guided percutaneous biopsy: experience with 110 tumors. *Radiology*, 223(3):731–737, June 2002. 30

[116] K. Jin, M. T. McCann, E. Froustey, and M. Unser. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 26:4509–4522, 2017. 61, 177, 190

[117] J. Jorgensen, E. Sidky, and X. Pan. Quantifying Admissible Undersampling for Sparsity-Exploiting Iterative Image Reconstruction in X-Ray CT. *IEEE Transactions on Medical Imaging*, 32(2):460–473, Feb. 2013. 49

[118] P. Joseph and R. Spital. A method for correcting bone induced artifacts in computed tomography scanners. *Journal of computer assisted tomography*, 2(1):100–108, 1978. 39, 59

[119] A. Kak and M. Slaney. *Principles of Computerized Tomographic Imaging*. Society of Industrial and Applied Mathematics, 2001. 19

[120] J. Katsis, L. Roller, M. Lester, J. E. Johnson, R. J. Lentz, O. Rickman, and F. Maldonado. High accuracy of digital tomosynthesis-guided bronchoscopic biopsy confirmed by intraprocedural computed tomography. *Respiration*, 100:214 – 221, 2021. 27

[121] D. Kazantsev, F. Bleichrodt, T. van Leeuwen, A. Kaestner, P. Withers, K. J. Batenburg, and P. D. Lee. A novel tomographic reconstruction method based on the robust student's t function for suppressing data outliers. *IEEE Transactions on Computational Imaging*, 3(4):682–693, 2017. 180

[122] D. Kazantsev, V. Pickalov, S. Nagella, E. Pasca, and P. J. Withers. Tomophantom, a software package to generate 2d–4d analytical phantoms for ct image reconstruction algorithm benchmarks. *SoftwareX*, 7:150–155, 2018. 189

[123] S. Kindermann, S. Osher, and P. W. Jones. Deblurring and denoising of images by nonlocal functionals. *Multiscale Model. Simul.*, 4:1091–1115, 2006. 180

[124] N. Komodakis and J.-C. Pesquet. Playing with duality: An overview of recent primal-dual approaches for solving large-scale optimization problems. *IEEE Signal Processing Magazine*, 32(6):31–54, 2015. 97

[125] R. D. Kongskov and Y. Dong. Tomographic reconstruction methods for decomposing directional components. *Inverse Problems and Imaging*, 12(6):1429–1442, 2018. 83

[126] R. D. Kongskov, Y. Dong, and K. Knudsen. Directional total generalized variation regularization. *BIT Numerical Mathematics*, 59(4):903–928, May 2019. 158

[127] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951. 119

[128] M. Lagerwerf, D. Pelt, W. Palenstijn, and K. J. Batenburg. A Computationally Efficient Reconstruction Algorithm for Circular Cone-Beam Computed Tomography Using Shallow Neural Networks. *Journal of Imaging*, 6(12):135, Dec. 2020. 62

[129] K. Lange and R. Carson. EM reconstruction algorithms for emission and transmission tomography. *Journal of Computer Assisted Tomography*, 8(2):306–316, Apr. 1984. 47

[130] H. Langet, C. Riddell, A. Reshef, Y. Trousset, A. Tenenhaus, E. Lahalle, G. Fleury, and N. Paragios. Compressed-sensing-based content-driven hierarchical reconstruction: Theory and application to c-arm cone-beam tomography. *Medical Physics*, 42(9):5222–5237, 2015. 55, 59, 60, 142, 157, 177

[131] H. Le, N. Pustelnik, and M. Foare. The faster proximal algorithm, the better unfolded deep learning architecture? the study case of image denoising. *Proceedings of the 30th European Signal Processing Conference, EUSIPCO 2022*, 29 August - 2 September 2022. 208

[132] R. Lewitt and R. Bates. Image reconstruction: from projections: III. projection completion methods (theory). *Optik*, 50:189–204, 1978. 44

[133] J. Li, C. Miao, Z. Shen, G. Wang, and H.Yu. Robust frame based x-ray ct reconstruction. *Journal of Computational Mathematics*, 34:683–704, 2016. 158

[134] K. Liang, H. Yang, K. Kang, and Y. Xing. Improve angular resolution for sparse-view CT with residual convolutional neural network. In *Medical Imaging 2018: Physics of Medical Imaging*, volume 10573, page 105731K. International Society for Optics and Photonics, 2018. 61

[135] K. Liang, H. Yang, and Y. Xing. Comparision of projection domain, image domain, and comprehensive deep learning for sparse-view x-ray CT image reconstruction. In *arXiv preprint arXiv:1804.04289*, 2018. 61

[136] R. Liu, L. Fu, B. D. Man, and H. Yu. Gpu-based branchless distance-driven projection and backprojection. *IEEE Transactions on Computational Imaging*, 3(4):617–632, 2017. 123

[137] S.-C. Lo. Strip and line path integrals with a square pixel matrix: a unified theory for computational CT projections. *IEEE Transactions on Medical Imaging*, 7(4):355–363, Dec. 1988. 59

[138] Y. Long, J. A. Fessler, and J. M. Balter. 3D forward and back-projection for x-ray CT using separable footprints. *IEEE Transactions on Medical Imaging*, 29(11):1839–1850, Nov. 2010. 60, 123

[139] D. Lorenz, S. Rose, and F. Schöpfer. The randomized Kaczmarz method with mismatched adjoint. *BIT Numerical Mathematics*, 58:1079–1098, 2018. 65

[140] D. Lorenz and F. Schneppe. Chambolle-pock's primal-dual method with mismatched adjoint. *arXiv preprint arXiv:2112.00776*, 2022. 97, 114, 206

[141] D. Lorenz and N. Worliczek. Necessary conditions for variational regularization schemes. *Inverse Problems*, 29(7):075016, 2013. 93

[142] D. A. Lorenz and T. Pock. An inertial forward-backward algorithm for monotone inclusions. *Journal of Mathematical Imaging and Vision*, 51(2):311–325, 2015. 176

[143] I. Loris and C. Verhoeven. On a generalization of the iterative soft-thresholding algorithm for the case of non-separable penalty. *Inverse Problems*, 27:125007, 2011. 57, 107, 108

[144] D. Luenberger. *Optimization by vector space methods*. Series in decision and control. Wiley, New York, NY Chichester Weinheim, nachdr. edition, 1998. 128

[145] S. Lunz, O. Öktem, and C.-B. Schönlieb. Adversarial regularizers in inverse problems. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 8516–8525, Red Hook, NY, USA, Dec. 2018. Curran Associates Inc. 62

[146] Q. Luong and O. Faugeras. Self-calibration of a moving camera from point correspondences and fundamental matrices. *International Journal of Computer Vision*, 22(3):261–289, 1997. 125

[147] J. Ma, Z. Liang, Y. Fan, Y. Liu, J. Huang, W. Chen, and H. Lu. Variance analysis of x-ray CT sinograms in the presence of electronic noise background. *Medical physics*, 39:4051–4065, 2012. 47

[148] Y. Marnissi, E. Chouzenoux, A. Benazza-Benyahia, and J.-C. Pesquet. Majorize–minimize adapted metropolis–hastings algorithm. *IEEE Transactions on Signal Processing*, 68:2356–2369, 2020. 94

[149] O. Merveille, B. Naegel, H. Talbot, and N. Passat. nD variational restoration of curvilinear structures with prior-based directional regularization. *IEEE Transactions on Image Processing*, 28(8):3848–3859, 2019. 158

[150] G. Meurant. A review on the inverse of symmetric tridiagonal and block tridiagonal matrices. *SIAM Journal on Matrix Analysis and Applications*, 13(3):707–728, 1992. 142

[151] F. Momey, L. Denis, C. Burnier, E. Thiébaut, J.-M. Becker, and L. Desbat. Spline driven: high accuracy projectors for tomographic reconstruction from few projections. *IEEE Transactions on Image Processing*, 24(12):4715–4725, 2015. 142

[152] J.-J. Moreau. Fonctions convexes duales et points proximaux dans un espace hilbertian. *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences, Paris*, 255:2897–2899, 1962. 52

[153] A. Muñoz, T. Blu, and M. Unser. Least-squares image resizing using finite differences. *IEEE Transactions on image processing*, 10(9):1365–1378, 2001. 206

[154] F. Natterer. *The mathematics of computerized tomography*. Number 32 in Classics in applied mathematics. Society for Industrial and Applied Mathematics, Philadelphia, 2001. 19

[155] M. Nilchian, C. Vonesch, P. Modregger, M. Stampanoni, and M. Unser. Iterative fbp for improved reconstruction of x-ray differential phase-contrast tomograms. In *2013 IEEE 10th International Symposium on Biomedical Imaging*, pages 1260–1263. IEEE, 2013. 55

[156] J. Nocedal and S. Wright. *Numerical optimization*. Springer, 1999. 54

[157] F. Noo, R. Clackdoyle, and J. Pack. A two-step hilbert transform method for 2d image reconstruction. *Physics in Medicine & Biology*, 49(17):3903, 2004. 44

[158] J. Nuyts, B. D. Man, J. Fessler, W. Zbijewski, and F. Beekman. Modelling the physics in the iterative reconstruction for transmission computed tomography. *Physics in Medicine and Biology*, 58(12):R63–R96, jun 2013. 47

[159] M. Osborne, B. Presnell, and B. Turlach. A new approach to variable selection in least squares problems. *IMA journal of numerical analysis*, 20(3):389–403, 2000. 55

[160] N. Ouzir, A. Basarab, O. Lairez, and J.-Y. Tourneret. Robust optical flow estimation in cardiac ultrasound images using a sparse representation. *IEEE Transactions on Medical Imaging*, 38(3):741–752, 2019. 180

[161] K. M. P. C. Hansen, K. Hayami. GMRES methods for tomographic reconstruction with an unmatched back projector. *arXiv preprint arXiv:2110.01481*, 2022. 65

[162] W. J. Palenstijn, K. J. Batenburg, and J. Sijbers. Performance improvements for iterative electron tomography reconstruction using graphics processing units (GPUs). *Journal of structural biology*, 176(2):250–253, 2011. 93

[163] P. Paleo, M. Desvignes, and A. Mirone. A practical local tomography reconstruction algorithm based on a known sub-region. *Journal of Synchrotron Radiation*, 24(1):257–268, 2017. 80, 177

[164] X. Pan, E. Sidky, and M. Vannier. Why do commercial CT scanners still employ traditional, filtered back-projection for image reconstruction? *Inverse problems*, 25(12):123009, 2009. 10, 59, 61

[165] S. Parisotto, J. Lellmann, S. Masnou, and C.-B. Schönlieb. Higher-order total directional variation: Imaging applications. *SIAM Journal on Imaging Sciences*, 13(4):2063–2104, 2020. 158

[166] S. Parisotto and C.-B. Schönlieb. Total Directional Variation for Video Denoising. In *Lecture Notes in Computer Science*, pages 522–534. Springer International Publishing, 2019. 158

[167] D. Parker. Optimal short scan convolution reconstruction for fan beam ct. *Medical physics*, 9(2):254–257, 1982. 44

[168] D. Pelt and K. J. Batenburg. Fast tomographic reconstruction from limited data using artificial neural networks. *IEEE Transactions on Image Processing*, 22(12):5238–5251, 2013. 62

[169] D. Pelt and K. J. Batenburg. Improving filtered backprojection reconstruction by data-dependent filtering. *IEEE Transactions on Image Processing*, 23(11):4750–4762, 2014. 89

[170] D. Pelt and V. De Andrade. Improved tomographic reconstruction of large-scale real-world data by filter optimization. *Advanced structural and chemical imaging*, 2(1):1–14, 2016. 89

[171] J.-C. Pesquet, A. Repetti, M. Terris, and Y. Wiaux. Learning maximally monotone operators for image recovery. *SIAM Journal on Imaging Sciences*, 14(3):1206–1237, Jan. 2021. 204

[172] G. Poludniowski, N. Allinson, and P. Evans. Proton computed tomography reconstruction using a backprojection-then-filtering approach. *Physics in Medicine & Biology*, 59(24):7905, 2014. 89

[173] M. Powell, D. DiNobile, and A. Reddy. C-arm fluoroscopic cone beam CT for guidance of minimally invasive spine interventions. *Pain Physician*, 13(1):51–59, Feb. 2010. 28

[174] E. Prashnani, H. Cai, Y. Mostofi, and P. Sen. Pieapp: Perceptual image-error assessment through pairwise preference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1808–1817, 2018. 189

[175] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007. 142

[176] M. A. Pritchett, K. Bhadra, and J. Mattingley. Electromagnetic navigation bronchoscopy with tomosynthesis-based visualization and positional correction. *Journal of Bronchology and Interventional Pulmonology*, 28:10 – 20, 2020. 27

[177] N. Pustelnik, A. Benazza-Benhayia, Y. Zheng, and J.-C. Pesquet. *Wavelet-Based Image Deconvolution and Reconstruction*, pages 1–34. American Cancer Society, 2016. 63, 182

[178] E. T. Quinto. Tomographic reconstructions from incomplete data numerical inversion of the exterior radon transform. *Inverse Problems*, 4(3):867, 1988. 157

[179] S. Ramani and J. Fessler. A splitting-based iterative algorithm for accelerated statistical x-ray CT reconstruction. *IEEE transactions on medical imaging*, 31(3):677–688, 2011. 89

[180] E. Reid, L. Drummy, C. Bouman, and G. Buzzard. Multi-resolution data fusion for super resolution imaging. *IEEE Transactions on Computational Imaging*, 8:81–95, 2022. 97

[181] A. Reshef, C. Riddell, Y. Trousset, S. Ladjal, and I. Bloch. Dual-rotation c-arm cone-beam computed tomography to increase low-contrast detection. *Medical Physics*, 44(9):e164–e173, 2017. 44, 80, 177

[182] C. Riddell, H. Benali, and I. Buvat. Diffusion regularization for iterative reconstruction in emission tomography. *IEEE Transactions on Nuclear Science*, 51(3):712–718, 2004. 60

[183] C. Riddell, B. Bendriem, M. H. Bourguignon, and J.-P. Kernevez. The approximate inverse and conjugate gradient: non-symmetrical algorithms for fast attenuation correction in SPECT. *Physics in Medicine and Biology*, 40(2):269–281, feb 1995. 46

[184] C. Riddell and Y. Trousset. Rectification for cone-beam projection and backprojection. *IEEE Transactions on Medical Imaging*, 25(7):950–962, July 2006. 142

[185] M. Ronchetti. TorchRadon: Fast differentiable routines for computed tomography. *arXiv preprint arXiv:2009.14788*, 2020. 189

[186] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. 61, 177

[187] L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60:259–268, 1992. 48, 116, 157

[188] J. Rudzusika, B. Bajic, O. Öktem, C.-B. Schönlieb, and C. Etmann. Invertible Learned Primal-Dual. In *NeurIPS 2021 Workshop on Deep Learning and Inverse Problems*, 2021. 208

[189] Y. Saad. *Iterative methods for sparse linear systems*. SIAM, Philadelphia, 2nd ed edition, 2003. 45

[190] H. Sadeghi, S. Banert, and P. Giselsson. Forward-backward splitting with deviations for monotone inclusions. *arXiv preprint arXiv:2112.00776*, 2021. 204

[191] I. Safran and O. Shamir. On the quality of the initial basin in overspecified neural networks. In *International Conference on Machine Learning*, pages 774–782. PMLR, 2016. 187

[192] M. Schmidt, N. Roux, and F. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. *Advances in neural information processing systems*, 24, 2011. 89

[193] I. J. Schoenberg. Contributions to the problem of approximation of equidistant data by analytic functions. In C. de Boor, editor, *I. J. Schoenberg Selected Papers*, pages 3–57. Birkhäuser Boston, Boston, MA, 1988. 126

[194] L. A. Shepp and Y. Vardi. Maximum likelihood reconstruction for emission tomography. *IEEE transactions on medical imaging*, 1(2):113–122, 1982. 47

[195] R. Siddon. Fast calculation of the exact radiological path for a three-dimensional ct array. *Medical Physics*, 12(2):252–255, 1985. 59

[196] E. Sidky, P. C. Hansen, J. Jørgensen, and X. Pan. Iterative image reconstruction for CT with unmatched projection matrices using the generalized minimal residual algorithm. *arXiv preprint arXiv:2201.07408*, 2022. 65, 184

[197] E. Sidky, J. Jørgensen, and X. Pan. Convex optimization problem prototyping for image reconstruction in computed tomography with the chambolle–pock algorithm. *Physics in Medicine & Biology*, 57(10):3065, 2012. 59

[198] E. Sidky, I. Lorente, J. G. Brankov, and X. Pan. Do CNNs solve the CT inverse problem. *IEEE Transactions on Biomedical Engineering*, 68(6):1799–1810, 2020. 61, 178

[199] E. Sidky and X. Pan. Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization. *Physics in Medicine and Biology*, 53(17):4777–4807, Sept. 2008. 49

[200] J. Siewerdsen, W. Zbijewski, and J. Xu. Cone beam CT image quality. *Cone beam computed tomography*, 4:37–58, 2014. 39

[201] J. Solomon, P. Lyu, D. Marin, and E. Samei. Noise and spatial resolution properties of a commercially available deep learning-based ct reconstruction algorithm. *Medical Physics*, 47(9):3961–3971, 2020. 61

[202] S. Solomon and S. Silverman. Imaging in interventional oncology. *Radiology*, 257(3):624–640, Dec. 2010. 27

[203] R. Souza, M. Bento, N. Nogovitsyn, K. Chung, W. Loos, M. Lebel, and R. Frayne. Dual-domain cascade of U-nets for multi-channel magnetic resonance image reconstruction. *Magnetic resonance imaging*, 71:140–153, 2020. 62

[204] T. Strohmer and R. Vershynin. A Randomized Kaczmarz Algorithm with Exponential Convergence. *Journal of Fourier Analysis and Applications*, 15(2):262, Apr. 2008. 45

[205] C. Syben, B. Stimpel, K. Breininger, T. Würfl, R. Fahrig, A. Dörfler, and A. Maier. Precision learning: reconstruction filter kernel discretization. *arXiv preprint arXiv:1710.06287*, 2017. 89

[206] A. Tam, A. Mohamed, M. Pfister, P. Chinndurai, E. Rohm, A. Hall, and M. Wallace. C-arm cone beam computed tomography needle path overlay for fluoroscopic guided vertebroplasty. *Spine*, 35(10):1095–1099, May 2010. 30

[207] J.-B. Thibault, K. Sauer, C. Bouman, and J. Hsieh. A three-dimensional statistical approach to improved image quality for multislice helical CT. *Medical physics*, 34:4526–44, 12 2007. 48

[208] T. Tirer and R. Giryes. Back-projection based fidelity term for ill-posed linear inverse problems. *IEEE Transactions on Image Processing*, 29:6164–6179, 2020. 55

[209] T. Tirer and R. Giryes. On the convergence rate of projected gradient descent for a back-projection based objective. *SIAM Journal on Imaging Sciences*, 14(4):1504–1531, 2021. 55

[210] M. Tivnan, W. Wang, and J. Stayman. A preconditioned algorithm for model-based iterative CT reconstruction and material decomposition from spectral CT data. *arXiv preprint arXiv:2010.01371*, 2020. 89

[211] P. Trapp, J. Maier, M. Susenburger, S. Sawall, and M. Kachelrieß. Empirical scatter correction: Cbct scatter artifact reduction without prior information. *Medical Physics*, 49(7):4566–4584, 2022. 59

[212] H. K. Tuy. An inversion formula for cone-beam reconstruction. *SIAM Journal on Applied Mathematics*, 43(3):546–552, 1983. 38

[213] M. Unser. Splines: a perfect fit for signal and image processing. *IEEE Signal Processing Magazine*, 16(6):22–38, Nov. 1999. 126

[214] M. Unser, A. Aldroubi, and M. Eden. B-spline signal processing. ii. efficiency design and applications. *IEEE transactions on signal processing*, 41(2):834–848, 1993. 127

[215] M. Unser, A. Aldroubi, and M. Eden. Enlargement or reduction of digital images with minimum loss of information. *IEEE Transactions on Image Processing*, 4(3):247–258, March 1995. 123, 128, 136, 143

[216] M. Unser, A. Aldroubi, M. Eden, et al. Fast B-spline transforms for continuous image representation and interpolation. *IEEE Transactions on pattern analysis and machine intelligence*, 13(3):277–285, 1991. 136

[217] M. Unser, P. Thevenaz, and L. Yaroslavsky. Convolution-based interpolation for fast, high-quality rotation of images. *IEEE Transactions on Image Processing*, 4(10):1371–1381, 1995. 139

[218] W. van Aarle, W. J. Palenstijn, J. Cant, E. Janssens, F. Bleichrodt, A. Dabravolski, J. D. Beenhouwer, K. J. Batenburg, and J. Sijbers. Fast and flexible x-ray tomography using the ASTRA toolbox. *Opt. Express*, 24(22):25129–25147, Oct 2016. 115

[219] W. van Aarle, W. J. Palenstijn, J. De Beenhouwer, T. Altantzis, S. Bals, K. J. Batenburg, and J. Sijbers. The ASTRA toolbox: A platform for advanced algorithm development in electron tomography. *Ultramicroscopy*, 157:35–47, 2015. 93, 115

[220] B. Vrcej and P. Vaidyanathan. Efficient implementation of all-digital interpolation. *IEEE transactions on image processing*, 10(11):1639–1646, 2001. 136

[221] B. Vu. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Computational Mathematics*, 38:667–681, 2013. 56

[222] G. Wang, J. C. Ye, and B. De Man. Deep learning for tomographic image reconstruction. *Nature Machine Intelligence*, 2:737–748, 12 2020. 61

[223] L. Wang, A. Mohammad-Djafari, N. Gac, and M. Dumitru. Bayesian 3D X-ray Computed Tomography with a Hierarchical Prior Model for Sparsity in Haar Transform Domain. *Entropy*, 20(12):977, Dec. 2018. 61

[224] S. Wang, S. Fidler, and R. Urtasun. Proximal Deep Structured Models. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. 63, 178

[225] T. Wang, K. Nakamoto, H. Zhang, and H. Liu. Reweighted anisotropic total variation minimization for limited-angle ct reconstruction. *IEEE Transactions on Nuclear Science*, 64(10):2742–2760, 2017. 157, 158, 184

[226] Z. Wang, H. Li, W. Ouyang, and W. Wang. Learnable histogram: Statistical context features for deep neural networks. In *Proceedings of the 14th European Conference on Computer Vision (ECCV)*, Amsterdam, The Netherlands, 2016. 187

[227] B. Wible and G. Walker. *Diagnostic imaging: interventional procedures*. Elsevier Health Sciences, 2017. 19

[228] Z. Wu, T. Bicer, Z. Liu, V. De Andrade, Y. Zhu, and I. T. Foster. Deep Learning-based Low-dose Tomography Reconstruction with Hybrid-dose Measurements. In *2020 IEEE/ACM Workshop on Machine Learning in High Performance Computing Environments (MLHPC) and Workshop on Artificial Intelligence and Machine Learning for Scientific Applications (AI4S)*, pages 88–95, Nov. 2020. 61

[229] T. Wurfl, M. Hoffmann, V. Christlein, K. Breininger, Y. Huang, M. Unberath, and A. Maier. Deep learning computed tomography: Learning projection-domain weights from image domain in limited angle problems. *IEEE transactions on medical imaging*, 37(6):1454—1463, June 2018. 62

[230] Y. Xia, A. Maier, H. Hofmann, F. Dennerlein, K. Mueller, and J. Hornegger. Reconstruction from truncated projections in cone-beam CT using an efficient 1D filtering. In *Medical Imaging 2013: Physics of Medical Imaging*, volume 8668, pages 348–354. SPIE, 2013. 55, 95

[231] J. Xu and F. Noo. Convex optimization algorithms in medical image reconstruction—in the age of AI. *Physics in Medicine & Biology*, 67(7):07TR01, Apr. 2022. 63

[232] J. Xu, Y. Zhao, H. Li, and P. Zhang. An image reconstruction model regularized by edge-preserving diffusion and smoothing for limited-angle computed tomography. *Inverse Problems*, 35(8):085004, 2019. 158

[233] Y. Yang, J. Sun, H. Li, and Z. Xu. Deep ADMM-Net for Compressive Sensing MRI. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. 63, 178

[234] D. C. Youla and H. Webb. Image restoration by the method of convex projections: part 1 theory. *IEEE transactions on medical imaging*, 1(2):81–94, 1982. 45, 54

[235] H. Yu and G. Wang. Compressed sensing based interior tomography. *Physics in Medicine and Biology*, 54(9):2791, 2009. 177

[236] L. Yu, X. Liu, S. Leng, J. M. Kofler, J. C. Ramirez-Giraldo, M. Qu, J. Christner, J. G. Fletcher, and C. H. McCollough. Radiation dose reduction in computed tomography: techniques and future perspective. *Imaging in Medicine*, 1(1):65–84, 2009. 34

[237] W. Yu, C. Wang, and M. Huang. Edge-preserving reconstruction from sparse projections of limited-angle computed tomography using ℓ_0-regularized gradient prior. *Review of Scientific Instruments*, 88(4):043703, 2017. 158, 184

[238] G. L. Zeng. Counter examples for unmatched projector/backprojector in an iterative algorithm. *Chinese Journal of Academic Radiology*, 2019. 65, 184

[239] G. L. Zeng and G. T. Gullberg. A ray-driven backprojector for backprojection filtering and filtered backprojection algorithms. In *1993 IEEE Conference Record Nuclear Science Symposium and Medical Imaging Conference*, pages 1199–1201, Oct 1993. 115

[240] G. L. Zeng and G. T. Gullberg. Unmatched projector/backprojector pairs in an iterative reconstruction algorithm. *IEEE Transactions on Medical Imaging*, 19:548–555, 2000. 46, 60, 65, 184

[241] H. Zhang, J. Wang, D. Zeng, X. Tao, and J. Ma. Regularization strategies in statistical image reconstruction of low-dose x-ray ct: A review. *Medical physics*, 45(10):e886–e907, 2018. 48

[242] H. Zhang and Y. Wang. Edge adaptive directional total variation. *The Journal of Engineering*, 2013(11):61–62, 2013. 158

[243] J. Zhang and B. Ghanem. Ista-net: interpretable optimization-inspired deep network for image compressive sensing. pages 1828–1837, 2018. 63, 178, 190, 192, 212

[244] Z. Zhang, B. Chen, D. Xia, E. Y. Sidky, and X. Pan. Directional-tv algorithm for image reconstruction from limited-angular-range data. *Medical Image Analysis*, 70:102030, 2021. 158

[245] Z. Zhang, X. Liang, X. Dong, Y. Xie, and G. Cao. A sparse-view CT reconstruction method based on combination of densenet and deconvolution. *IEEE transactions on medical imaging*, 37(6):1407–1417, 2018. 61

[246] B. Zhu, J. Liu, S. Cauley, B. Rosen, and M. Rosen. Image reconstruction by domain-transform manifold learning. *Nature*, 555(7697):487–492, Mar. 2018. 62

[247] A. Ziegler, T. Nielsen, and M. Grass. Iterative reconstruction of a region of interest for transmission tomography. *Medical Physics*, 35(4):1317–1327, 2008. 177

[248] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005. 66