HAL open science

Computational microscopy

Valentin Debarnot

▶ **To cite this version:**

Valentin Debarnot. Computational microscopy. Computer Aided Engineering. Université Paul Sabatier - Toulouse III, 2020. English. NNT : 2020TOU30156 . tel-03146497

HAL Id: tel-03146497

https://theses.hal.science/tel-03146497v1

Submitted on 19 Feb 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*

Présentée et soutenue le *10/12/2020* **par :**

Valentin Debarnot

Microscopie computationnelle

JURY

Kristian Bredies	Professeur	Rapporteur
Yuejie Chi	Associate professor	Examinatrice
Loïc Denis	Maitre de Conférence	Rapporteur
Charles Dossal	Professeur	Examinateur
Jérôme Idier	Directeur de recherche	Examinateur
Thomas Mangeat	Ingénieur de Recherche	Directeur de thèse
Anne Sentenac	Directrice de recherche	Examinatrice
Pierre Weiss	Chargé de Recherche	Directeur de thèse

École doctorale et spécialité :
MITT : Domaine Mathématiques : Mathématiques appliquées
Unité de Recherche :
Institut des Technologies Avancées en Sciences du Vivant (ITAV)
Directeur(s) de Thèse :
Thomas Mangeat et *Pierre Weiss*
Rapporteurs :
Kristian Bredies et *Loïc Denis*

Remerciements

S'il n'y avait que deux personnes à remercier, ce serait elles, le fantastique duo formé par mes deux brillants directeurs de thèse : Pierre et Thomas. Merci à vous deux pour cet environnement de travail idéal, où l'on croule sous les idées et les projets toujours plus intéressants. Je remercie tout particulièrement Pierre, pour sa présence quotidienne. De l'ITAV aux sentiers pyrénéens, j'ai appris bien plus de choses que je n'aurais pu l'imaginer. Un grand merci également à Thomas, ton énergie débordante et ta fougue ont toujours su apporter un souffle nouveau aux différents projets.

I would like to thank my two reviewers, Kristian Bredies and Loïc Denis, for their proofreading, their time, and their valuable feedback. I am very honored to have them on my jury. I would like to express my deepest thanks to the examiners, Yuejie Chi, Charles Dossal, Jérôme Idier, and Anne Sentenac, for accepting this invitation and François Malgouyres for his almost participation in the jury.

Un grand merci à Léo, compagnon de thèse parti trop tôt, pour ses conseils éclairés (à la lueur de la lanterne) et son expérience. Un grand merci également à Manu qui a su rapidement se montrer essentiel. Ses nombreuses relectures, commentaires et conseils avisés ont fortement contribué à faire évoluer ce manuscrit dans le bon sens. Je souhaite également remercier Paul, pour sa présence et ses travaux préliminaires sur le sujet.

La PRIMO team a maintenant troqué le préfixe 'dé-' pour 'ex-':'(. Mais cette triste disparition administrative n'enlève en rien les qualités de ses membres : Alban, Corbi, Frédéric et Jonas. Un remerciement tout particulier à Corbi, qui n'a pas le droit à sa traduction personnalisée puisqu'il est maintenant bilingue. Présent de Toulouse jusqu'au col des crêpes, il sait se montrer toujours accompagné de bonnes intentions (und gutes Weihnachtsplätzchen). On pourrait presque rajouter Renaud dans la Primo team, lui qui avait le droit à son rendez-vous hebdomadaire dans le bureau Shakira. Je remercie particulièrement ce dernier pour ses conseils avisés sur un grand nombre de questions importantes. Ce manuscrit est probablement pour moi la seule opportunité d'apparaître sur le renommé blog[1] dont j'ai ouïe dire qu'il était un des acteurs.

Je tiens maintenant à remercier les membres (et les anciens) de l'ITAV et plus largement du CPP : Aurélie, Corine, Childerick, Fabien, Grég, Jacques, Julia, Laetitia, Lise, Marine, Mél, Odile, Remy (J pour les intimes), Sophie,

[1] `http://je-mattarde.com/`

Victoria, Zoely. Une mise en valeur toute particulière est nécessaire pour Camille et sa capacité à résoudre nos problèmes plus vite que l'éclair. Merci également à Bernard et Valérie pour leur accueil dans le laboratoire.

Ces trois années sont passées extrêmement vite, et je pensais passer plus de temps au CBI et au LBMCP. Cependant, le temps d'un bref détour, j'ai eu la chance de discuter et de travailler avec quelques uns d'entre eux : Aude, Martine, Sylvain, Tong, Vincent. Je tiens également à remercier les membres d'Abbelight pour leur apport de données expérimentales.

Je tiens à remercier les membres du GMM de l'INSA de m'avoir accueilli tout au long de ces trois années, et particulièrement à Aude, Béatrice, Cathy, Clément, Florent, Jean-Yves, Jessica, Lorick et Sandrine.

Je me dois également de remercier Aurélien et Sébastien pour m'avoir accueilli durant quelques mois, le temps de m'initier aux joies des problèmes de bandits manchots.

Je voudrais finir par un remerciement plus général, aux amis qui sont présents depuis la maternelle ou depuis l'INSA, à la famille et à mon généraliste Adrien. Une pensée particulière pour ceux qui ne pourront pas lire ces quelques lignes, mamie Odile, papy Frédéric et papy Jean.

Les meilleurs pour la fin, Maman (en espérant qu'elle n'ait pas fait un malaise pensant que je l'ai oublié), Papa, pour leur soutien infaillible en toute circonstance, un immense merci.

Ces derniers mots vont maintenant à celle qui partage mon quotidien depuis longtemps, et mon pain depuis quelques mois. Un grand merci pour son soutien inconditionnel. マリーバ, あなたは美しいです.

Résumé

Les travaux présentés de cette thèse visent à proposer des outils numériques et théoriques pour la résolution de problèmes inverses en imagerie. Nous nous intéressons particulièrement au cas où l'opérateur d'observation (e.g. flou) n'est pas connu. Les résultats principaux de cette thèse s'articulent autour de l'estimation et l'identification de cet opérateur d'observation.

Une approche plébiscitée pour estimer un opérateur de dégradation consiste à observer un échantillon contenant des sources ponctuelles (microbilles en microscopie, étoiles en astronomie). Une telle acquisition fournit une mesure de la réponse impulsionnelle de l'opérateur en plusieurs points du champ de vue. Le traitement de cette observation requiert des outils robustes pouvant utiliser rapidement les données rencontrées en pratique. Nous proposons une boîte à outils qui estime un opérateur de dégradation à partir d'une image contenant des sources ponctuelles. L'opérateur estimé à la propriété qu'en tout point du champ de vue, sa réponse impulsionnelle s'exprime comme une combinaison linéaire de fonctions élémentaires. Cela permet d'estimer des opérateurs invariants (convolutions) et variants (développement en convolution-produit) spatialement. Une spécificité importante de cette boîte à outils est son caractère automatique : seul un nombre réduit de paramètres facilement accessibles permettent de couvrir une grande majorité des cas pratiques. La taille de la source ponctuelle (e.g. bille), le fond et le bruit sont également pris en compte dans l'estimation. Cet outil se présente sous la forme d'un module appelé PSF-Estimator pour le logiciel *Fiji*, et repose sur une implémentation parallélisée en *C++*.

En réalité, les opérateurs modélisant un système optique varient d'une expérience à une autre, ce qui, dans l'idéal, nécessite une calibration du système avant chaque acquisition. Pour pallier à cela, nous proposons de représenter un système optique non pas par un unique opérateur de dégradation, mais par un sous-espace d'opérateurs. Cet ensemble doit permettre de représenter chaque opérateur généré par un microscope. Nous introduisons une méthode d'estimation d'un tel sous-espace à partir d'une collection d'opérateurs de faible rang (comme ceux estimés par la boîte à outils PSF-Estimator). Nous montrons que sous des hypothèses raisonnables, ce sous-espace est de faible dimension et est constitué d'éléments de faible rang. Dans un second temps, nous appliquons ce procédé en microscopie sur de grands champs de vue et avec des opérateurs variant spatialement. Cette mise en œuvre est possible grâce à l'utilisation de méthodes complémentaires pour traiter des images réelles (e.g. le fond, le bruit,

la discrétisation de l'observation).

La construction d'un sous-espace d'opérateurs n'est qu'une étape dans l'étalonnage de systèmes optiques et la résolution de problèmes inverses. Il est alors nécessaire de pouvoir identifier un élément du sous-espace à partir d'une image. Dans cette thèse, on donne un cadre mathématique à ce problème d'identification d'opérateur dans le cas où l'image originale est constituée de sources ponctuelles. Des conditions pratiques découlent de ces travaux, permettant de mieux comprendre quel est le cadre dans lequel on peut identifier un opérateur. Nous illustrons en pratique comment cette étude théorique permet de résoudre des problèmes de défloutage aveugle réels.

Malheureusement, l'hypothèse selon laquelle l'image originale est composée de sources ponctuelles n'est pas toujours valide. Dans le cas d'une image arbitraire, trouver des conditions pratiques sous lesquelles un opérateur peut être estimé reste essentiellement un problème ouvert. Dans cette thèse, nous proposons l'utilisation d'un réseau de neurones pour aborder ce problème. Nous montrons comment mettre en place une telle méthode à l'aide des outils introduits tout au long de cette thèse.

Finalement, dans une dernière partie, nous proposons un recueil de différents problèmes abordés en parallèle des travaux précédents, mais dont le sujet principal s'écarte du fil conducteur qu'est la résolution de problèmes inverses aveugles. Le premier de ces travaux est la réalisation d'un environnement numérique intitulé *Biolapse*, qui automatise le traitement d'images en Biologie. Cet ensemble de codes vise à détecter automatiquement des cellules contenues dans une image et à les classifier en fonction de leur état dans le cycle cellulaire. Cet outil repose principalement sur des outils récents d'apprentissage machine. Dans un second travail, nous introduisons une méthode originale pour résoudre des problèmes inverses décrits par une équation de diffusion non-linéaire (loi de Beer-Lambert). Nous montrons comment à partir de différentes observations de la même scène sous des angles différents, on peut estimer les inconnues du problème. Finalement nous proposons une analyse du problème de Graetz qui modélise les phénomènes de convection-diffusion dans des tuyaux à largeur constante.

Résumé pour les non-scientifiques

Ce manuscrit de thèse concerne l'étude et l'amélioration des images obtenues avec des microscopes. Le microscope est un instrument qui permet de voir des objets extrêmement petits. Son fonctionnement est très proche de celui de la loupe. Lorsque l'on cherche à lire un texte écrit en petit avec une loupe, il faut la positionner convenablement. Si la loupe est trop loin, le texte va apparaître flou. Si la loupe est trop près, on ne voit que quelques lettres sans grossissement. Dans les deux cas, on met des heures à lire le texte en question. Un microscope fonctionne exactement de la même manière. Le problème en microscopie, c'est que l'on travaille avec des objets bien plus petits, et donc le moindre écart (inclinaison, mouvement, changement de lumière ou de température) à la position idéale de la loupe va produire du flou.

En pratique, les biologistes utilisent le microscope pour prendre une photo des cellules qu'ils cherchent à étudier, et une image floue apparaît sur l'ordinateur. C'est là que mes travaux rentrent en jeu. Le but final est de proposer des programmes informatiques qui vont retirer le flou de l'image obtenue afin de

voir au mieux les éléments de la cellule. Pour établir de tels programmes informatiques, j'utilise des outils mathématiques pour représenter le processus de formation d'une image. De façon schématique, on a alors l'image nette (le biologiste met ses cellules sous le microscope) qui passe dans une série de formules mathématiques, et l'on obtient l'image floue. L'idée est alors d'inverser ces formules mathématiques. Un problème est que ces équations ne sont que des approximations de la réalité. Il y a tellement d'effets en jeu que nous ne sommes pas capables de tout modéliser. Dans cette thèse, je me suis alors intéressé à comment coupler le formalisme mathématique avec des images provenant directement du microscope pour améliorer ce processus d'inversion.

Au terme de ces trois années, j'ai développé différentes méthodes améliorant le retrait du flou de certains microscopes. Un aspect important de cette thèse est la certification des méthodes. Grâce à des études mathématiques, je peux garantir quand est ce que mes méthodes vont produire des résultats corrects ou erronés. J'ai également développé des outils basés sur les réseaux de neurones (outil de l'intelligence artificielle), mais dans ce cas, aucune certification n'est disponible. C'est pour certifier ce type d'approche que la recherche en mathématiques est essentielle. Une autre force de cette approche basée sur les mathématiques est que les outils développés pour la microscopie peuvent également être utilisés dans d'autres domaines tels que le spatial ou la photographie.

Abstract

The contributions of this thesis are numerical and theoretical tools for the resolution of blind inverse problems in imaging.

We first focus in the case where the observation operator is unknown (e.g. microscopy, astronomy, photography). A very popular approach consists in estimating this operator from an image containing point sources (microbeads or fluorescent proteins in microscopy, stars in astronomy). Such an observation provides a measure of the impulse response of the degradation operator at several points in the field of view. Processing this observation requires robust tools that can rapidly use the data. We propose a toolbox that estimates a degradation operator from an image containing point sources. The estimated operator has the property that at any location in the field of view, its impulse response is expressed as a linear combination of elementary estimated functions. This makes it possible to estimate spatially invariant (convolution) and variant (product-convolution expansion) operators. An important specificity of this toolbox is its high level of automation: only a small number of easily accessible parameters allows to cover a large majority of practical cases. The size of the point source (e.g. bead), the background and the noise are also taken in consideration in the estimation. This tool, coined PSF-Estimator , comes in the form of a module for the *Fiji* software, and is based on a parallelized implementation in *C++*.

The operators generated by an optical system are usually changing for each experiment, which ideally requires a calibration of the system before each acquisition. To overcome this, we propose to represent an optical system not by a single operator (e.g. convolution blur with a fixed kernel for different experiments), but by subspace of operators. This set allows to represent all the possible states of a microscope. We introduce a method for estimating such a subspace from a collection of low rank operators (such as those estimated by the toolbox PSF-Estimator). We show that under reasonable assumptions, this subspace is low-dimensional and consists of low rank elements. In a second step, we apply this process in microscopy on large fields of view and with spatially varying operators. This implementation is possible thanks to the use of additional methods to process real images (e.g. background, noise, discretization of the observation).

The construction of an operator subspace is only one step in the resolution of blind inverse problems. It is then necessary to identify the degradation operator in this set from a single observed image. In this thesis, we provide a mathematical

framework to this operator identification problem in the case where the original image is constituted of point sources. Theoretical conditions arise from this work, allowing a better understanding of the conditions under which this problem can be solved. We illustrate how this formal study allows the resolution of a blind deblurring problem on a microscopy example.

Unfortunately, the hypothesis that the original image is composed of point sources is not always valid. Considering arbitrary images, finding practical conditions under which an operator can be estimated remains an open problem. In this thesis, we propose the use of neural networks to tackle this problem. We show how to implement such a method using the tools introduced throughout this thesis.

Finally, in a last part, we propose a collection of different problems studied in parallel of the other works. Their focus deviates from the resolution of blind inverse problems. The first of these works is the implementation of a software called *Biolapse*, which automates the processing of biological images. This set of codes allows to automatically detect cells contained in an image and classify them according to their state in the cell cycle. This tool is mainly based on recent machine learning developments and convolutional neural networks. In a second work, we introduce an original method to solve inverse problems described by a nonlinear diffusion equation (Beer-Lambert's law). We show how to estimate attenuation coefficients from pairs of images illuminated from different sides. Finally, we study the Graetz problem, which models convection-diffusion equations in constant width tubes.

Abstract for non-scientists

This thesis manuscript concerns the study and improvement of images obtained with microscopes. The microscope is an instrument that allows us to see extremely small objects. Its operating mode is very close to that of a magnifying glass. When one tries to read a text written in small with a magnifying glass, it must be positioned correctly. If the magnifying glass is too far away, the text will appear blurred, if the magnifying glass is too close, one sees only a few letters without magnification and it takes hours to read the text. A microscope works in similar way. The problem with microscopy is that we work with much smaller objects, and therefore the smallest deviation (tilt, movement, change of light or temperature) at the ideal position of the magnifying glass will produce blur.

In practice, biologists use the microscope to take a picture of the cells they are trying to study, and a blurred image appears on the computer. That's where my work comes in. games. The final goal is to propose computer programs that will remove the blur from the obtained image in order to see the elements of the cell as well as possible. To establish such computer programs, I use mathematical tools to represent the process of image formation. In a schematic way, we then have the sharp image (the biologist puts his cells under the microscope) which passes through a series of mathematical formulas, and we obtain the blurred image. The idea is then to invert these mathematical formulas. One problem is that these equations are only approximations of reality. There are so many effects at play that we are not able to model everything. In this thesis, I am interested in how to couple mathematical formalism with observations from the microscope to improve this inversion process.

At the end of these three years, I am able to provide different methods to improve the blur removal of some microscopes. An important aspect of this thesis is the certification of the methods. Thanks to to mathematical studies, I can guarantee when my methods will produce correct or incorrect results. I have also developed tools based on neural networks (artificial intelligence tool), but in this case, no certification is available. It is to certify this type approaches that mathematical research is essential. Another strength of this mathematically based approach is that the tools developed for microscopy can also be used in other fields such as space or photography.

Avant-propos

Les parties d'introduction et de conclusion de cette thèse sont rédigées en français. Les autres parties sont rédigées en anglais car elles reposent sur des travaux publiés, ou en cours de publication, dans des revues ou conférences internationales. Un résumé de ces chapitres est systématiquement présent en anglais et en français.

Chaque chapitre correspond à un travail publié ou en cours de publication. Certains éléments ont pu être déplacés de façon à faciliter la lecture du manuscrit. Cependant, les chapitres de ce document sont rédigés dans le but de pouvoir être lu indépendamment. Une conséquence est que certains éléments peuvent apparaître à plusieurs endroits, notamment au niveau des introductions des différents chapitres. Une seconde conséquence est que les notations peuvent différer d'un chapitre à l'autre.

La Partie I est composée de deux sous-parties : le Chapitre 1 qui présente une introduction générale du sujet de cette thèse, et le Chapitre 2 qui expose quelques notions sur l'optique en microscopie et les développements en convolution-produit. La Partie II est composée de trois sous-parties et traite du problème de calibration de systèmes optiques. Cette partie est constituée de résultats mathématiques ainsi que d'illustrations pratiques dans le cas de la microscopie à fluorescence. La Partie III est composée de trois sous-parties et traite de l'identification d'opérateurs dans les problèmes inverses aveugles. Nous y présentons des résultats à la fois théoriques et pratiques. La Partie IV est composée de trois sous-parties et présente divers travaux réalisés avant et pendant la thèse, dont le sujet d'étude ne sert pas à la résolution de problèmes inverses aveugles. La Partie V conclut le manuscrit en proposant une discussion et un résumé des activités scientifiques conduites durant cette thèse.

Table des matières

I Introduction **1**

1 Introduction **3**

2 Quelques notions utiles **17**

II Subspace of operators **33**

3 Estimating an impulse responses subspace **37**

4 A scalable estimator of sets of integral operators **73**

5 Estimating a subspace of space-varying blur operators from microbeads image **97**

III Blind inverse problems **115**

6 Blind inverse problems with isolated spikes **119**

7 The blind sparse+smooth (BSS) algorithm **149**

8 Deepblur: blind identification of space variant PSF **157**

IV Miscellaneous problems **171**

9 Biolapse toolbox: automating biological image analysis. **173**

10 Multiview Attenuation Estimation and Correction **187**

V Closing remarks **213**

11 Discussion **215**

VI Bibliographie **219**

12 Mes références **221**

13 Autres références **223**

A Semi infinite generalized Graetz problem **247**

Première partie

Introduction

Chapitre 1

Introduction

Contents

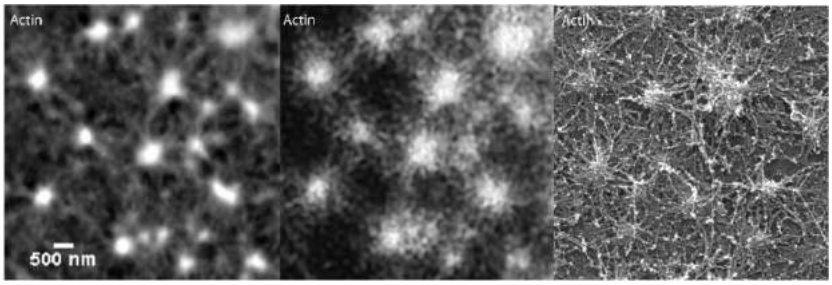
1.1 Introduction **3**
1.2 Contributions **12**
1.2.1 Théorique 14
1.2.2 Méthodologique 14
1.3 Organisation du manuscrit **15**

1.1 Introduction

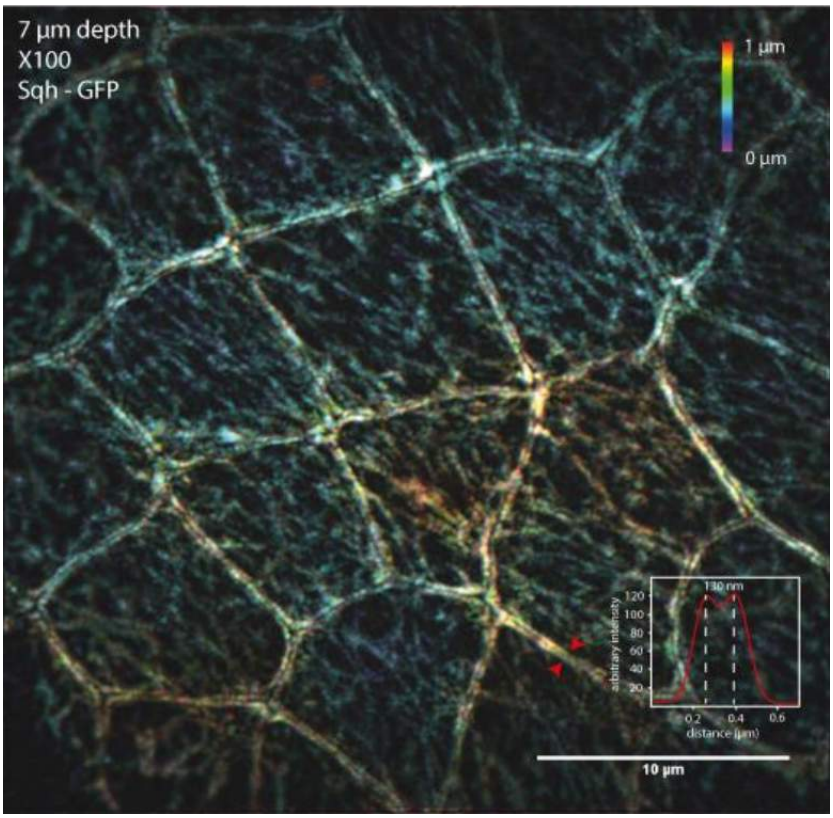
Les systèmes optiques sont apparus plusieurs siècles avant J-C. Les premières traces de lentilles optiques remontent à l'époque de l'Égypte antique [SS87], où de tels objets étaient utilisés comme loupes. Cet effet d'agrandissement, que permettent les lentilles optiques, a été étudié bien plus tardivement en mathématiques, notamment par Euclide au 3e siècle avant J-C et par René Descartes dans son ouvrage intitulé "La Dioptrique" au 17e siècle. À la même période, les premiers systèmes optiques apparaissent entre les mains des scientifiques : le mathématicien italien Galilée construit la première lunette astronomique, mais aussi probablement l'un des premiers microscopes ; Antoni van Leeuwenhoek démocratise l'utilisation des microscopes en biologie [Lee08]. Suite à cela s'enchaînent des siècles d'améliorations continues des différents systèmes optiques (microscopes, télescopes, etc), avec les contributions de quelques-uns des scientifiques les plus célèbres tels que Kepler, Newton, Huygens, Fraunhofer, Fresnel, Maxwell, Abbe, etc. Au début du 20e siècle, il existe alors une diversité de systèmes optiques plus grande que ne peut l'accueillir cette brève introduction.

La motivation première de cette thèse vient de la microscopie à fluorescence. Cette technique introduite par Herschel [Her45] en 1845, est aujourd'hui un outil clé dans le domaine de la biologie [SP03 ; Söd+08 ; Lan+06]. L'utilisation de cette technologie a connu un regain d'intérêt après 1962, avec la découverte de la protéine fluorescente verte (GFP en anglais pour *green fluorescent protein*) [SJS62], récompensée du prix Nobel de Chimie en 2008. Les applications se sont alors principalement tournées vers la compréhension du fonctionnement cellulaire.

Les études du poisson-zèbre ou de la drosophile, par exemple, ont permis de mieux comprendre certains processus biologiques tels que le développement embryonnaire ou la réponse immunitaire de ces espèces [Kel+08 ; EMS00]. En guise d'illustration, la Figure 1.1b reprend des résultats très récents qui permettent d'étudier la concentration de myosine dans les cellules (protéine responsable des mécanismes de mouvements cellulaires), au cours du temps sur une drosophile.



(a) Reconstruction du réseau d'actine au sein de macrophages pour trois modalités de microscopie. De gauche à droite : microscopie à localisation de molécule unique dSTORM [Gus+16], microscopie à illumination structurée [Man+20 ; Idi+17], et microscopie électronique à balayage.



(b) Reconstruction 3D d'un réseau de myosine dans un thorax de drosophile. Coupe dorsale. Microscopie à illumination structurée [Man+20 ; Idi+17]

FIGURE 1.1: Reconstructions obtenues en utilisant différentes modalités de microscopes.

Ces avancées spectaculaires sont possibles grâce aux développements simultanés des composants matériels (tels que les caméras, les objectifs, les miroirs déformables, etc), et aux outils informatiques, notamment grâce à l'augmentation des capacités de calculs. On s'intéresse dans cette thèse au domaine de la microscopie computationnelle, qui consiste à développer des outils numériques permettant d'extraire un maximum d'informations à partir d'acquisitions de microscope, rendant ainsi possible l'obtention d'images super-résolues telles que dans la Figure 1.1. Les systèmes d'imagerie modernes reposent souvent sur un traitement des données par des algorithmes de reconstruction. C'est notamment le cas de la ptychographie [HH70 ; Mar+13], l'imagerie par résonance magnétique [Laz+17 ; Jun+09], la microscopie à localisation de molécule unique [HW94 ; CGI96 ; Bet+06 ; HGM06 ; RBZ06], la super-résolution en microscopie par illumination structurée [HC99 ; Gus00 ; Man+20 ; Idi+17] ou encore l'imagerie par contraste de phase [Zer42 ; Pop11]. Cette approche computationnelle se retrouve dans toutes les étapes d'acquisitions et d'analyse d'images. Il est par exemple possible avec ces méthodes, d'aider à la détection d'exoplanètes [Fla+20], de corriger les acquisitions IRM [Col+13], ou même d'optimiser des méthodes d'acquisition en IRM [Leb+19]. Cela se caractérise bien souvent par une approche de co-conception, qui cherche à développer conjointement le système et les méthodes de reconstruction numériques [Bon+09 ; Bou+15 ; TW15]. Cette liste de références n'est bien évidemment pas représentative de la grande diversité de travaux utilisant des méthodes numériques.

L'objectif principal de cette thèse concerne l'amélioration de la résolution des acquisitions en microscopie à fluorescence grâce au développement de méthodes numériques.

Le gain en résolution est un des plus grands défis du domaine de la microscopie. Le microscope confocal et les microscopes de super-résolution (illumination structurée, localisation de molécule unique) sont deux exemples modernes qui ont permis de mieux comprendre certains phénomènes biologiques complexes grâce à une meilleure résolution [Bea+14 ; Wan+16 ; Bou+17 ; Dau+19].

Plus formellement, on modélise l'observation y d'un échantillon u par la relation :

$$y = \mathcal{P}(\text{Ш}(H(u))), \tag{1.1}$$

où $H : \mathcal{U} \to \mathcal{C}_0(\mathbb{R}^d)$ est un opérateur intégral avec $\mathcal{U}$ un sous-espace vectoriel d'images, $\text{Ш} : \mathcal{C}_0(\mathbb{R}^d) \to \mathbb{R}^m$ est un opérateur d'échantillonnage tel que $\text{Ш}(v)[i] = v(x_i)$ pour tout $v \in \mathcal{C}(\mathbb{R}^d)$ et pour tout $1 \leq i \leq m$, et $\mathcal{P}$ est une perturbation déterministe ou aléatoire.

Le théorème de Nyquist-Shannon [Nyq28 ; Sha48] nous indique qu'une fonction continue dont la transformée de Fourier est supportée sur un disque de rayon F_c, peut-être reconstruite exactement en l'échantillonnant à la fréquence $2F_c$. Cela nous permet de choisir l'opérateur d'échantillonnage Ш en fonction du système optique de façon à collecter toute l'information disponible. Cependant, la résolution d'un système optique est limitée par le phénomène de diffraction de la lumière. La réponse impulsionnelle du système produit une tâche, appelée aussi PSF (*Point Spread Function*, fonction d'étalement en français) empêchant ainsi la séparation de deux points trop proches. Plus formellement, ces systèmes agissent comme des filtres passe-bas. Le critère de Rayleigh donne une borne

de la distance à laquelle peuvent se trouver deux points pour être correctement séparés. Cette distance minimale de résolution d_{res} est donnée dans le cas d'un microscope à immersion (à fluorescence) par :

$$d_{res} = 0.61 \frac{\lambda}{\mathrm{NA}}, \tag{1.2}$$

où λ est la longueur d'onde de la lumière collectée, et NA l'ouverture numérique. Pour un microscope à champ large standard, la distance de résolution est de l'ordre de 200nm. La formule de l'équation (1.2) est basée sur la théorie de l'optique de Fourier [Goo05] dans le cas où la réponse impulsionnelle du système est donnée par la tache de Airy. Les systèmes optiques conventionnels (e.g. champ large, confocal) sont limités par la diffraction. Cette limite ne peut pas être franchie sans hypothèse supplémentaire. On peut néanmoins améliorer la qualité des images par déconvolution ou débruitage, mais les fréquences au-delà d'une certaine fréquence, appelée *fréquence de coupure*, ne sont pas retrouvées (sans hypothèses supplémentaire).

La limite de résolution donnée par le critère de Rayleigh, qui repose sur le fait qu'on observe deux sources simultanément, peut-être fortement réduite si l'on suppose cette fois que l'on observe seulement une seule source. C'est le cadre des techniques de microscopie à localisation de molécule unique [RBZ06 ; Hua+08]. On peut dans ce cas distinguer des structures éloignées de l'ordre de 20nm, voir Figure 1.2. Cette limite peut aussi être franchie en utilisant une illumination structurée, comme c'est le cas dans la Figure 1.1 avec une illumination structurée aléatoire. Ces techniques collectent en général un nombre plus important de données en faisant varier les paramètres du système : l'illumination en illumination structurée, ou l'activation séquentielle des molécules en SMLM. L'acquisition et la reconstruction sont considérées conjointement dès la conception de la technique d'imagerie.

Pour atteindre les limites de résolution en microscopie à localisation de molécule unique, il est nécessaire d'avoir :

1. une connaissance précise du système optique, et plus particulièrement de l'opérateur de flou H,
2. un algorithme de restauration pour estimer u à partir de l'équation (1.1).

Ces deux éléments sont en fait indispensables pour tous les systèmes d'acquisitions mentionnés précédemment (ptychographie, imagerie par résonance magnétique, etc). Prenons l'exemple de la microscopie à localisation de molécule unique où un *challenge* a vu le jour pour comparer les différentes méthodes de reconstruction. Dans le récent rapport sur les résultats de ce *challenge* [Sag+19], les auteurs font remarquer qu'un modèle précis de PSF semble nécessaire pour atteindre les meilleurs résultats en 3D.

Dans cette thèse, on se concentre particulièrement sur l'estimation de l'opérateur H qui décrit les effets de diffraction et qui est responsable de la perte de résolution. Nous porterons une attention particulière à une modélisation précise d'un opérateur H tout en conservant de bonnes propriétés numériques. Ces deux points semblent essentiels pour gagner en résolution et pour mettre en œuvre des algorithmes de reconstruction efficaces.

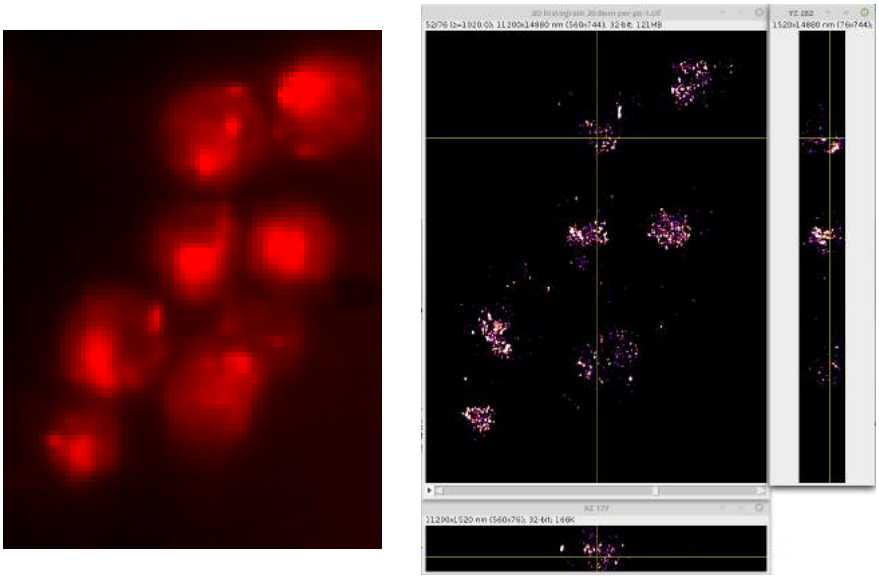


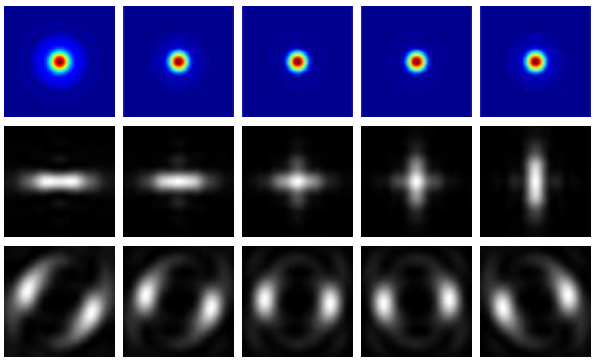
FIGURE 1.2: Image d'un même échantillon obtenu au microscope champ large standard (gauche) et en utilisant le procédé de localisation de molécule unique (droite). La taille du pixel est de 160nm à gauche, et 20nm à droite.

La grande majorité des travaux en optique repose sur un modèle d'opérateur H invariant spatialement. Cela se modélise naturellement par l'action d'un filtre de convolution h, tel que :

$$H(x) = h \star x. \tag{1.3}$$

Un modèle très utilisé consiste à prendre h comme étant la tâche de Airy [Air35], voir Figure 1.3a. La tâche de Airy correspond au motif de diffraction de la lumière lorsque celle-ci traverse une ouverture circulaire. Ce motif intervient donc dans les systèmes optiques équipés de lentilles circulaires, tels que les microscopes, les télescopes ou les appareils photo. Cela explique que ce modèle soit largement utilisé pour déflouter les images obtenues lorsque aucune PSF expérimentale n'est disponible. C'est par exemple le cas avec les logiciels *Hyugens* et *DeconvolutionLab2* [Sag+17b]. Dans le cas de la microscopie à localisation de molécule unique en 3D, le filtre h est modélisé par une forme variant avec la 3e dimension z (la profondeur). Cela permet d'encoder une information tridimensionnelle dans une mesure bidimensionnelle. Les PSFs les plus communes sont l'astigmatique [Hua+08] ou la double hélice [Pav+09]. La Figure 1.3a montre les formes théoriques de ces deux familles de PSFs. La Figure 1.3b provient d'une acquisition de microscope où la PSF du système est astigmatique.

En pratique, le modèle stationnaire n'est valable que pour de petits champs de vue. La Figure 1.4a montre un exemple d'acquisition obtenue avec un grand champ de vue. On remarque que le flou présent sur l'image semble plus ou moins important en fonction de la zone que l'on considère. Une façon plus précise de mettre en avant ce phénomène consiste à imager des microbilles placées dans le champ de vue, c'est l'expérience réalisée sur la Figure 1.4b. Le flou généré par une bille, si elle est assez petite comme dans la Figure 1.4b, nous donne accès à la réponse impulsionnelle du système localement. On observe ici que



(a) Réponse impulsionnelle théorique. La première ligne correspond à la tache de Airy, la seconde à un motif astigmatique, et la troisième ligne correspond à un motif double-hélice. La profondeur varie de gauche à droite de -200nm à +200nm, par pas de 100nm.

(b) Réponse impulsionnelle astigmatique expérimentale. La profondeur varie de gauche à droite de -500nm à +500nm, par pas de 250nm. Les paramètres expérimentaux (longueur d'onde, ouverture numérique, etc) sont différents des simulations précédentes. Cela explique la différence d'échelle entre les figures. Le pixel est de taille 160nm et la bille de taille 100nm.

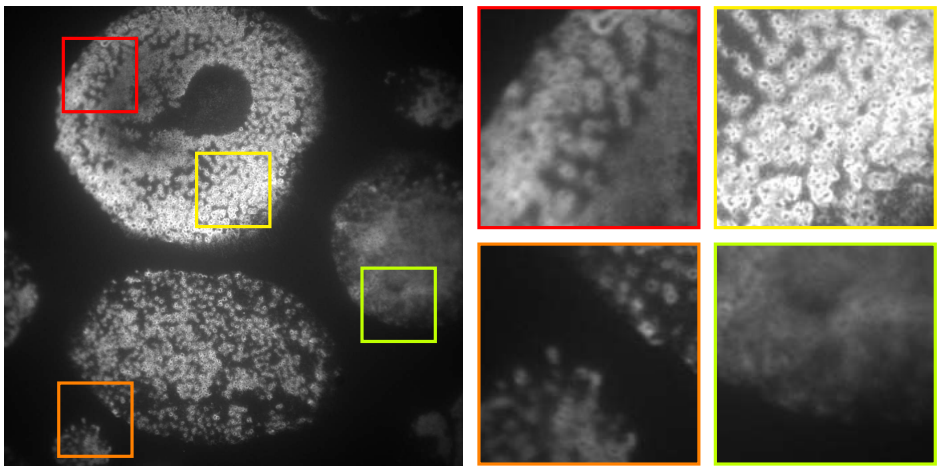
FIGURE 1.3: Exemple de réponses impulsionnelles simulées et expérimentales en microscopie.

cette PSF varie spatialement. Ces deux expériences ont été réalisées sur le même microscope champ large, l'une après l'autre.

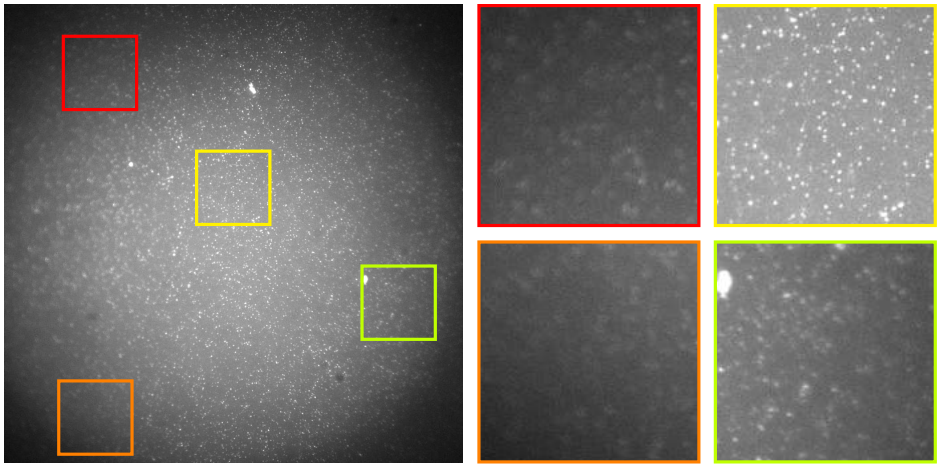
Un aspect important de cette thèse est la prise en compte des variations spatiales dans l'opérateur de flou H.

Ces variations spatiales sont liées à différents phénomènes physiques, tels que :

- L'échantillon est placé entre lame et lamelle, mais celles-ci sont légèrement bombées.
- L'échantillon occupe un espace tri-dimensionnel, il ne vit pas exactement dans un plan 2D aligné avec l'objectif.
- La lumière d'excitation n'est pas parfaitement uniforme. On peut le constater sur la Figure 1.4b.
- Les effets de polarisation, qui sont difficiles à mesurer pour un objectif à grande ouverture numérique.
- Les vibrations mécaniques (caméras, platine, etc).



(a) Acquisition d'un podosome.



(b) Acquisition de microbilles.

FIGURE 1.4: Exemple d'acquisitions sur de grands champs de vue. Les acquisitions ont été réalisées sur un microscope grand champ à immersion, avec une ouverture numérique de 1.4. Les images sont de taille 2048×2048 pixels, couvrant un champ de vue de 80µm.

- Les fluctuations thermiques (mouvement brownien) des émetteurs même lorsqu'ils sont fixés.

Cependant, ces effets peuvent souvent être négligés pour de petits champs de vue, ce qui rend le modèle stationnaire populaire en pratique. Le fait de ne pas prendre en compte ces variations spatiales peut conduire à l'échec des algorithmes de reconstruction sur de grands champs de vue. C'est particulièrement vrai en microscopie à localisation de molécule unique, où la localisation des molécules repose sur un modèle précis de la réponse impulsionnelle du système. La majorité des expériences pratiques sont alors réalisées en utilisant seulement une petite partie du champ de vue. Un scénario typique consiste à utiliser un patch de taille 128×128 sur une image pouvant atteindre 512×512 pixels. Cela représente

une réduction d'un facteur 16. Ce niveau de troncature est nécessaire pour bénéficier des méthodes de l'état de l'art qui supposent que le flou est invariant spatialement [Man+20 ; Sag+19].

Prendre en compte les variations spatiales est donc indispensable pour utiliser au mieux les avancées technologiques en microscopie. Cela soulève plusieurs questions importantes :

- Comment coder cet opérateur de flou H ? Le modèle stationnaire, qui repose sur le calcul d'une convolution, peut-être réalisé efficacement grâce à la FFT (Fast Fourier Transform). De plus, l'opérateur H est dans ce cas représenté par son filtre de convolution h, qui fait au plus la même taille que l'image. Ces propriétés ne sont pas nécessairement conservées lorsque l'on considère des opérateurs plus avancés.

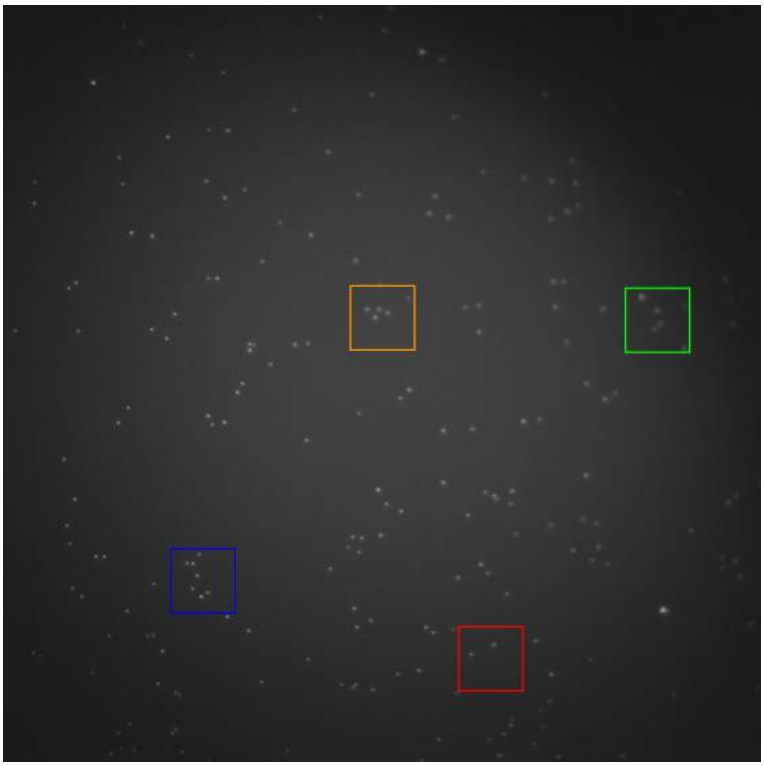
- Comment estimer cet opérateur de flou H ? Dans le cas du modèle stationnaire, une façon de faire est d'imager une ou plusieurs microbilles, à n'importe quel endroit du champ de vue, comme c'est le cas pour l'acquisition de la Figure 1.3b. Dans le cas non-stationnaire, on peut reproduire une expérience similaire en utilisant une image de microbilles, voir Figure 1.5. On a alors la réponse du système localement, en chaque bille, mais il reste à étendre cette information sur tout le champ de vue.

Nous supposons que l'opérateur H est linéaire et que toutes les réponses impulsionnelles peuvent être exprimées dans une même base.

Sous ces hypothèses, différents auteurs [FR05b ; Hir+11 ; Den+15] ont montré que l'opérateur H pouvait être approché par un développement en convolution-produit (*product-convolution expansion* en anglais). Paul Escande et Pierre Weiss [EW17] ont analysé mathématiquement ces approximations. Un exemple de développement en convolution-produit reproduisant le flou généré par un microscope sur une image de microbilles est affiché sur la Figure 1.5c. On remarque que cette décomposition permet de reproduire les variations spatiales. Nous introduisons formellement les développements en convolution-produit dans le chapitre suivant, voir Section 2.5.

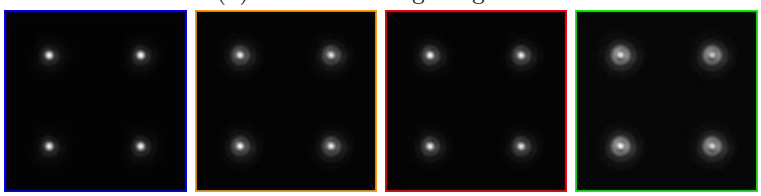
Dans cette thèse, nous exploiterons les séries en convolution-produit pour diverses applications en microscopie. Nous nous référons à la thèse de Paul Escande [Esc16] pour l'étude des propriétés d'approximation de cette décomposition et les différentes implémentations envisageables.

Il est possible d'estimer un opérateur sous forme de développement en convolution-produit à partir d'une image de microbilles. C'est le sujet de Chapitre 3. On a alors accès à l'état du système à un instant donné. Cependant, l'opérateur de flou H change d'une expérience à une autre. Cela est lié à beaucoup de paramètres tels que les changements de température dans la pièce, le mouvement des éléments optiques, le changement de position de la lamelle contenant l'échantillon, etc. Pour toutes ces raisons, le système optique doit être étalonné à chaque acquisition. Cette opération est longue et en général impossible. La Figure 1.6 rend compte des variations temporelles sur un microscope grand champ. Ces écarts sont particulièrement importants sur la figure 1.6b. Cela est



(a) Image 2304 × 2304 de microbilles avec un microscope champ large. La taille du pixel est de 43nm.

(b) Zoom de l'image originale.



(c) Zoom d'un opérateur convolution-produit estimé à partir de l'image de microbilles ci-dessus.

FIGURE 1.5: Image de microbilles de 100nm de diamètre et estimation de l'opérateur de flou. Des informations plus précises sur le système optique utilisé sont disponibles dans la Section 5.3.1.

dû au fait qu'un nouvel élément de contrôle de température au niveau du miroir déformable a été ajouté. Cet élément présentait un défaut de construction.

Nous proposons de caractériser un système optique non pas par un unique opérateur H, mais par une famille d'opérateurs $\mathcal{H}$. Cette famille prend en compte les variations du système optique.

L'utilisation d'une famille $\mathcal{H}$ nous semble indispensable pour modéliser au mieux les effets intervenants sur un microscope. Malheureusement, cette modélisation a une conséquence importante sur la résolution des problèmes inverses à résoudre. On est maintenant confronté à des problèmes de type :

$$\text{Retrouver } u \in \mathcal{U} \text{ et } H \in \mathcal{H} \text{ tel que } y = \mathcal{P}\left(Ш(H(u))\right). \tag{1.4}$$

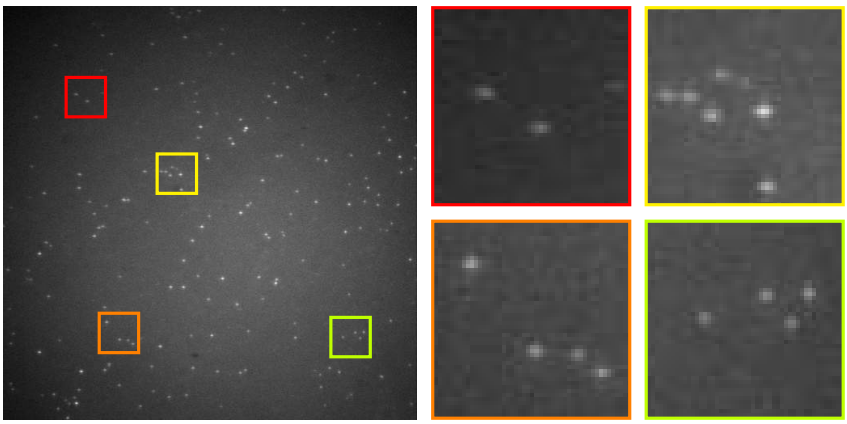
On rentre alors dans la catégorie des problèmes inverses aveugles. [1] Ce domaine nous semble aujourd'hui essentiellement ouvert. Plusieurs méthodes empiriques ont montré de bon résultats [CW98 ; YK99 ; RC08 ; SJA08 ; Lev+09 ; CL09 ; XJ10 ; Xu+11 ; Sou+12 ; Mou+15 ; Mou16]. La majorité de ces travaux résolvent le Problème (1.4) en introduisant des termes de pénalisations sur l'image ou l'opérateur (qui est une convolution dans la grande majorité des méthodes). Les méthodes d'apprentissage machine ont aussi montré de très bons résultats pratiques [EPF14 ; Xu+14 ; Sch+15 ; Zha+17]. Elles manquent cependant de bases mathématiques solides pour le moment. Récemment, des garanties théoriques ont commencé à émerger [ARR13 ; CSV13 ; JKS17 ; Li+19 ; KS19 ; DC20]. Ces travaux s'appuient sur la méthode de *lifting* qui consiste à transformer le Problème bilinéaire (1.4) en un problème linaire, au prix d'une augmentation de la dimension du problème. Malgré des efforts pour rendre ces méthodes utilisables pour des problèmes réels [Li+19 ; BB19], on observe une disparité entre les méthodes utilisées en pratique et les méthodes avec des garanties mathématiques. En particulier, les conditions sous lesquelles les garanties de reconstruction s'appliquent sont très restrictives et souvent irréalistes. L'image est supposée être tirée aléatoirement dans un sous-espace $\mathcal{U}$ de faible dimension. Cette hypothèse de connaissance d'un sous-espace de faible dimension dans lequel vit l'image est en général bien trop forte. D'autres travaux étudient la structure des problèmes inverses aveugles [BB18]. Dans cette optique, plusieurs groupes [LLB16a ; KK17] cherchent à caractériser les conditions sous-lesquelles la solution du Problème (1.4) est identifiable.

Nous tenterons d'apporter quelques contributions à la compréhension et la résolution de problèmes inverses aveugles en microscopie.

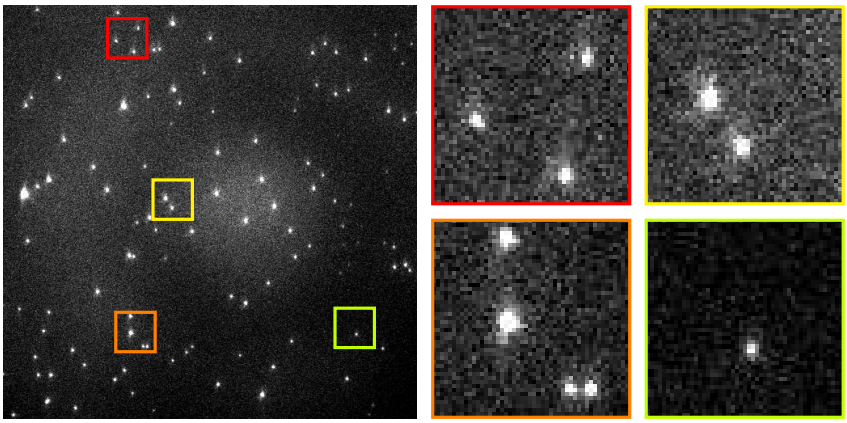
1.2 Contributions

Dans cette partie, nous présentons les éléments clés des différentes contributions de cette thèse.

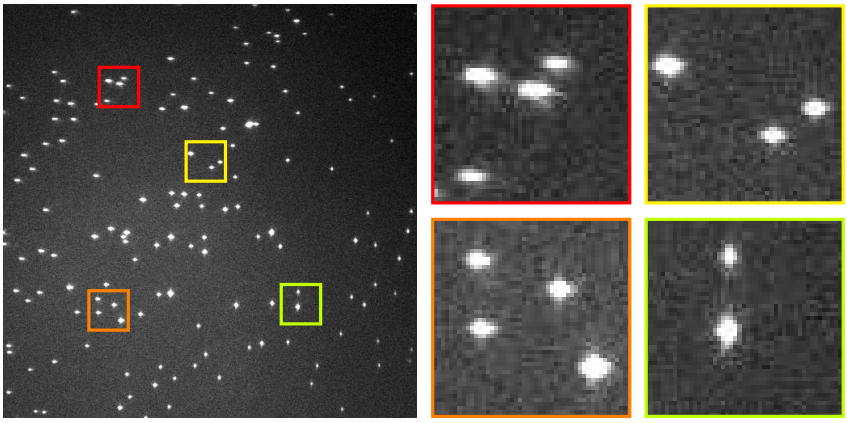
[1] Quelques auteurs [Mug+01] parleraient ici de restauration myope (plutôt qu'aveugle) car on possède tout de même quelques informations sur le système optique. Cette distinction entre défloutage aveugle et myope nous semble cependant un peu arbitraire, car on dispose toujours d'information a priori sur la réponse impulsionnelle d'un système optique.



(a) Image de microbilles acquise en mars 2018.



(b) Image de microbilles acquise en septembre 2018.



(c) Image de microbilles acquise en janvier 2019.

FIGURE 1.6: Images de microbilles acquises sur plusieurs mois. Le système d'acquisition est un microscope grand champ équipé d'un miroir déformable utilisé pour produire de l'astigmatisme. Les images sont de taille 512×512 pixels. Ces acquisitions ont été réalisées par Sylvain Cantaloube.

1.2.1 Théorique

Dans un premier temps, nous listons les éléments plutôt théoriques ou abstraits.

Estimation de sous-espace linéaire d'opérateurs : Nous proposons une méthode pour estimer un sous-espace de tenseurs de faible rang, qui approche simultanément un ensemble d'opérateurs intégraux. Cette estimation peut être considérée comme une généralisation de méthode de décomposition des tenseurs, ce qui n'a jamais été utilisé dans ce contexte. Nous établissons des garanties de stabilité dans différents régimes de bruit aléatoire.

Cadre mathématique pour l'identification d'opérateur : Nous établissons un cadre formel garantissant l'identification d'opérateur dans le cas où l'on observe son action sur une ou plusieurs sources ponctuelles. Cette étude propose un niveau d'abstraction permettant de traiter de nombreux cadres pratiques pas encore explorés. Les garanties de reconstruction sont vérifiables sur des applications réelles et minimales.

Restriction à un cône convexe d'opérateur : Conjointement à l'utilisation d'un sous-espace d'opérateurs, nous proposons de construire un sous-ensemble convexe permettant de réduire l'espace de recherche tout en conservant certaines propriétés des opérateurs d'intérêts (positivité, décroissance des coefficients). Nous fournissons des garanties théoriques sur cet estimateur et quelques résultats numériques.

Atténuation multi-vues : L'estimation du coefficient d'atténuation d'un échantillon peut être réalisée par l'utilisation de différents systèmes d'acquisitions (e.g. lidar, rayon X). Nous proposons une nouvelle approche basée sur l'observation d'un échantillon sous quelques angles différents. Il repose sur la résolution d'un problème inverse non-linéaire. Nous proposons une approche de calcul efficace avec des garanties théoriques pour le résoudre.

1.2.2 Méthodologique

Dans un second temps, nous listons les éléments plutôt méthodologiques.

SIFT (Scale Invariant Feature Transform) vectoriel : La détection de formes au sein d'une image est un problème courant. Nous avons proposé une méthode d'extraction de caractéristiques à partir d'un dictionnaire de formes admissibles. La spécificité de cette approche réside dans sa non-dépendance en la taille des objets et en son automatisation, réduisant au maximum le nombre de paramètres à fournir. Cette approche peut être vue comme une généralisation des SIFT (scale invariant feature detectors) qui ont été popularisés dans les années 2000.

PSF-Estimator : Nous proposons une boîte à outils permettant à partir d'une image de microbilles, d'estimer un sous-espace vectoriel de faible dimension permettant de capturer toutes les réponses impulsionnelles d'un système optique. Cet outil se présente sous la forme d'un module pour le logiciel *Fiji*, et repose sur

une implémentation parallélisée en *C++*. Il possède également de nombreuses caractéristiques le rendant robuste à une large variété d'images expérimentales.

Boîte à outils pour l'estimation de sous-espaces d'opérateurs : Nous montrons comment estimer un sous-espace vectoriel d'opérateurs en microscopie à partir d'images de microbilles. Cette méthode fait appel à divers outils rendant robuste l'estimation sur des données réelles. Cela permet notamment la calibration fine d'un microscope, rendant possible la résolution de problèmes inverses aveugles. Elle permet également de créer des bases de données d'opérateurs modélisant précisément les systèmes optiques.

Identification d'opérateur à partir d'images quelconques : Le problème d'identification d'opérateur à partir d'une image quelconque est essentiellement ouvert. Nous proposons d'utiliser les résultats précédents d'estimation de sous-espaces d'opérateurs pour construire un réseau de neurones réalisant cette tache. L'utilisation des résultats précédents permet d'entraîner de façon rapide et robuste un tel réseau, sans nécessité de collecter de grandes quantités de données expérimentales.

Biolapse : Les microscopes haut-débit permettent de collecter une grande quantité de données. Le traitement de ces acquisitions reste un problème compliqué souvent réalisé manuellement. Nous proposons une boîte à outils permettant d'automatiser la détection de cellules dans ce type d'acquisitions. Les cellules extraites sont ensuite classifiées en fonction de leur état dans le cycle cellulaire. Cette approche permet de réaliser des statistiques robustes sur les images obtenues. Biolpase repose sur des méthodes de machine learning. Nous proposons également une interface graphique permettant l'entraînement et l'utilisation de ces méthodes par les non-experts.

1.3 Organisation du manuscrit

Le chapitre suivant introduit des notations utiles que nous allons utiliser tout au long du manuscrit. En particulier, on y introduit quelques notions de microscopie et de modélisation, ainsi qu'un paragraphe sur l'approximation d'opérateurs linéaires.

Dans la Partie II, nous introduisons les outils et les méthodes permettant d'estimer un sous espace d'opérateurs de faible dimension. Nous présentons dans le Chapitre 3 une méthode d'estimation d'opérateurs à partir d'images contenant des sources ponctuelles. Dans le Chapitre 4, nous introduisons un cadre mathématique pour estimer un sous-espace d'opérateurs. Nous étudions les propriétés de cet estimateur dans un cadre non limité à celui de la microscopie. Dans le Chapitre 5 nous utilisons les résultats du Chapitre 4 pour estimer un sous-espace de faible dimension décrivant au mieux un système optique donné.

Dans la Partie III, nous proposons différentes approches pour résoudre des problèmes inverses aveugles en microscopie. Dans le Chapitre 6, nous proposons un cadre théorique pour l'identification d'opérateurs linéaires dans le cas d'un signal contenant des sources ponctuelles. Dans le Chapitre 8, nous donnons un cadre plus pratique avec l'utilisation des réseaux de neurones convolutifs pour identifier un opérateur de flou à partir d'acquisitions arbitraires. Dans le

Chapitre 7, nous combinons les résultats des chapitres précédents pour résoudre un problème de défloutage aveugle en microscopie.

Finalement, dans la Partie IV, nous présentons d'autres travaux réalisés avant ou au cours de cette thèse, mais non directement liés à notre motivation principale qu'est la résolution de problèmes inverses aveugles en microscopie.

La Partie V vient conclure le manuscrit. Elle contient une discussion sur les travaux présentés, les pistes explorées et celles restant à explorer.

Chapitre 2

Quelques notions utiles

Contents

2.1 Notations **17**
2.2 Contexte académique **18**
2.3 Microscopie à fluorescence **18**
2.3.1 Microscopie champ large 20
2.3.2 Microscopie à localisation de molécules uniques . . 21
2.3.3 Résumé 25
2.4 Modèle de formation de l'image **25**
2.4.1 Modélisation physique 25
2.4.2 Limite du modèle physique 28
2.4.3 Résumé 29
2.5 Approximation d'opérateurs linéaires : décomposition en convolution-produit **29**
2.5.1 Développement en convolution-produit 29
2.5.2 D'autres types d'approximations 31
2.5.3 Résumé 31

2.1 Notations

La majorité des chapitres de cette thèse ont été écrits indépendamment les uns des autres, cela pour permettre au lecteur de lire seulement quelques parties. La conséquence est que les notations peuvent être variables d'un chapitre à l'autre. Nous ferons attention à préciser celles-ci si nécessaire. Néanmoins, lorsque rien n'est spécifié, nous utiliserons les notations introduites ci-dessous.

Les termes anglais qui ne peuvent pas être traduits ou qui sont volontairement laissés en anglais sont écrits en italique (dans les chapitres en français). Dans une bonne partie du manuscrit, la police d'écriture grasse fait référence à des vecteurs, des matrices ou des fonctions à valeur vectorielle. La police d'écriture standard fait référence à des scalaires ou des fonctions. La $i$ème valeur d'un vecteur $\mathbf{x}$ est notée x_i ou $\mathbf{x}[i]$. La norme ℓ^p d'un vecteur $\mathbf{x}$ est noté $\|\mathbf{x}\|_p$. La

valeur de la fonction f est $f(x)$ et sa norme ℓ^p est notée $\|f\|_p$. La masse de Dirac à la position $\mathbf{x} \in \mathbb{R}^d$ est notée $\delta_{\mathbf{x}}$. Les notations I, J, K, L, M et N font référence soit à des ensembles d'entiers allant de 1 à $|I|$, $|J|$, $|K|$, $|L|$, $|M|$ et $|N|$ ou simplement à un nombre de composantes. L'une ou l'autre de ces notations seront précisées dans les chapitres les utilisant. Le produit scalaire sur tous les espaces est noté $\langle \cdot \, , \, \cdot \rangle$. Le produit tensoriel entre deux vecteurs $\mathbf{a} \in \mathbb{R}^n$ et $\mathbf{b} \in \mathbb{R}^m$ est défini par $(a \otimes b)[i,j] = a[i]b[j]$, pour tout $1 \leq i \leq n$ et $1 \leq j \leq m$. On note $\odot$ le produit terme à terme (produit de Hadamard) et $\star$ le produit de convolution. On note $L^2(\Omega)$ l'espace des fonctions de carré intégrale, où $\Omega \subset \mathbb{R}^d$. On note $H^s(\Omega)$ l'espace de Hilbert des fonctions de $L^2(\Omega)$ dont les s premières dérivées appartiennent $L^2(\Omega)$. Une famille de vecteur $(e_i)_{1 \leq i \leq I}$ est appelée base orthogonale (*orthogonal basis* en anglais) lorsque l'on a la relation suivante : $\langle e_i, e_j \rangle = 0$ si $i \neq j$ et $\langle e_i, e_i \rangle = 1$.

2.2 Contexte académique

Comme nous l'avons vu en introduction, pour exploiter pleinement le potentiel des nouvelles technologies optiques, il faut prendre en compte le fait que le flou induit par le système varie spatialement. Ces problématiques se posent dans beaucoup de laboratoires, et notamment au laboratoire de biologie cellulaire et moléculaire (LBCMCP) et au Centre de biologie integrative (CBI) de Toulouse, où Thomas Mangeat (directeur de cette thèse) est en charge de la plateforme d'imagerie. Dans le même temps, Paul Escande et Pierre Weiss (directeur de cette thèse) à l'institut des technologies avancées en sciences du vivant (ITAV) ont proposé des outils pouvant aider à mieux traiter certains de ces problèmes. Cette thèse est une continuation des travaux de Paul Escande et Pierre Weiss [Esc16] et fait le lien avec les questions pratiques amenées par Thomas Mangeat, liées à la reconstruction d'images dans différents systèmes d'acquisition en microscopie. La grande majorité des exemples donnés dans cette thèse s'appliquent naturellement à la microscopie à fluorescence, bien que les méthodes n'y soient souvent pas limitées.

2.3 Microscopie à fluorescence

La fluorescence est la capacité d'un objet à émettre des photons (lumière d'émission) après avoir absorbé des photons de plus haute énergie (lumière d'excitation). La lumière d'émission et d'excitation n'ont pas la même longueur d'onde, i.e. ont des couleurs différentes. La différence entre leur longueur d'onde est connue comme le déplacement de Stokes.

La microscopie à fluorescence est principalement utilisée en biologie et c'est dans ce domaine qu'apparaissent la majorité des nouvelles technologies. Cette technique repose sur la propriété de fluorescence de certaines molécules/protéines, ce qui permet l'étude ciblée des zones d'intérêts au sein de cellules, d'embryons, etc.

Le principe de la microscopie à fluorescence est d'éclairer un échantillon biologique composé de molécules fluorescentes introduites par les expérimentateurs. Les molécules fluorescentes vont alors émettre une lumière d'émission, de couleur différente de la lumière d'excitation. Cette lumière est collectée et isolée de la

lumière d'excitation par un jeu de lentilles et de miroirs (dichroïque). Ce principe est schématisé sur la Figure 2.1. Dans cet exemple, la lumière d'émission est en trait plein rouge et la lumière d'excitation est en pointillé bleu. Nous décrivons l'action des différents éléments optiques pour la configuration d'un microscope champ large dans le paragraphe suivant.

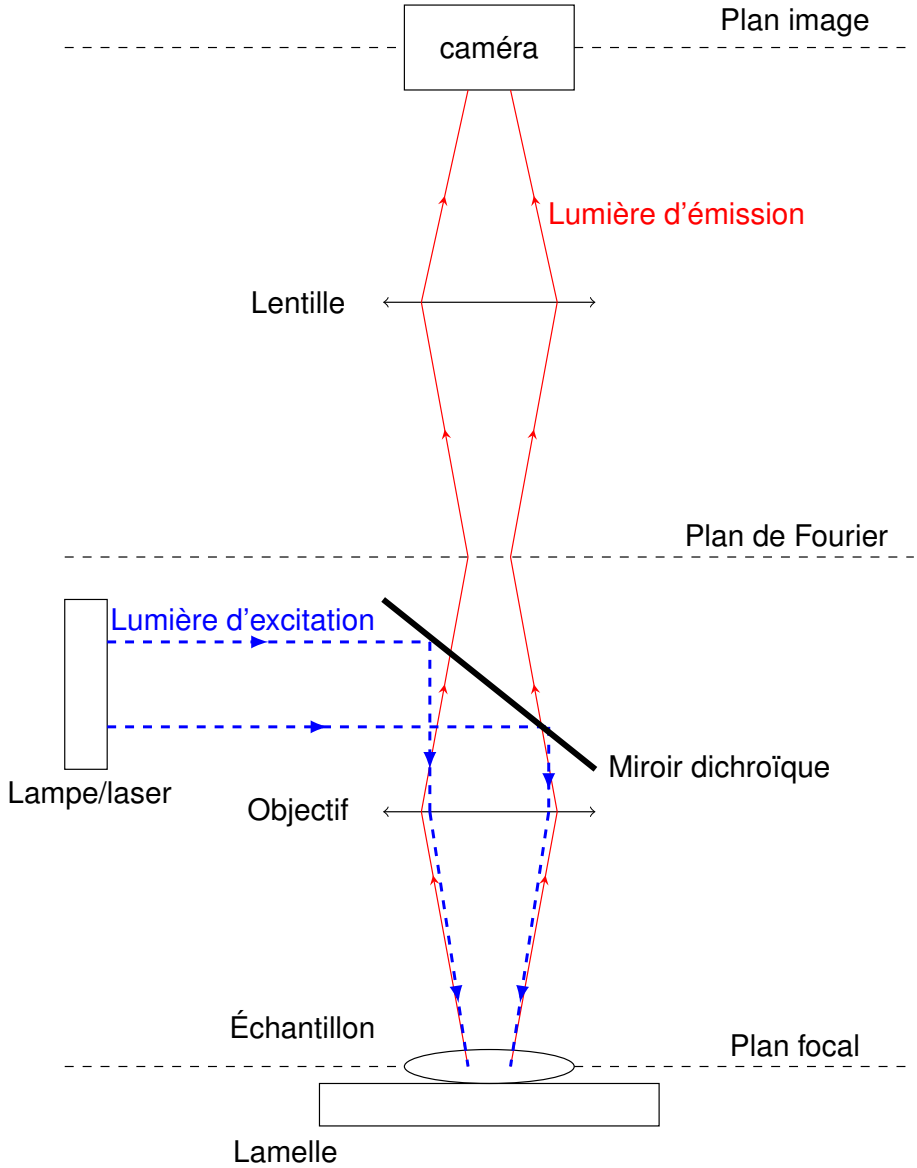


FIGURE 2.1: Principe schématique du fonctionnement d'un microscope à fluorescence (confocal).

Un élément commun à tous les microscopes est l'objectif. C'est l'élément principal du microscope. Son rôle est de collecter la lumière d'émission pour constituer une image plus grande de l'échantillon fluorescent. Le grandissement

d'un objectif est le rapport entre la taille physique de l'objet imagé et la taille de son image sur le capteur. Cela peut aller d'un objectif $\times 10$ à un objectif $\times 100$. L'objectif laisse passer la lumière par un orifice circulaire. Comme nous allons le voir dans les prochains paragraphes, la limite de résolution d'un objectif est inversement proportionnelle à l'ouverture numérique. Cette quantité se définit à l'aide de deux quantités : n_I l'indice de réfraction du milieu dans lequel se propagent les rayons lumineux entre l'échantillon fluorescent et l'objectif, et α l'angle maximum des rayon lumineux capté par l'objectif, voir Figure 2.2. On définit alors l'ouverture numérique NA (numerical aperture) par :

$$\mathrm{NA} = \mathrm{n}_I \sin(\alpha),$$

Cela correspond à la capacité d'un objectif à collecter les rayons lumineux les plus éloignés. Cette quantité intervient dans la modélisation physique d'un système optique que nous présentons dans la Section 2.4

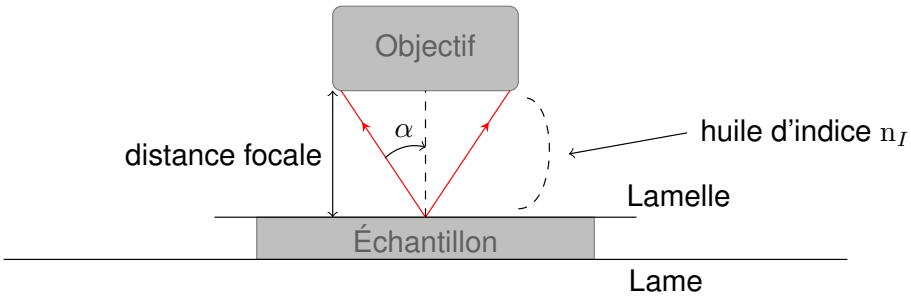


FIGURE 2.2: Fonctionnement d'un objectif et calcul de l'ouverture numérique.

2.3.1 Microscopie champ large

Il existe un grand nombre de microscopes à fluorescence : confocal, multi-confocal (*spinning disk*), multi-photonique (ou bi-photon), à feuille de lumière, etc. Un des systèmes les plus répandus est le microscope à champ large. Son fonctionnement relativement simple est décrit ci-dessous :

1. L'échantillon biologique à imager doit être préparé pour permettre aux structures que l'on veut observer d'être fluorescentes. Un des marquages les plus courants est le marquage par immunofluorescence. Cela consiste à utiliser un anticorps dirigé contre la molécule d'intérêt. Cet anticorps est couplé à un fluorochrome. L'exemple le plus répandu est peut-être celui du GFP (*Green Fluorescent Protein*), qui a été récompensé par le prix Nobel de chimie en 2008 pour sa découverte et ses applications [SJS62 ; Cha+94 ; Tsi98]. D'autres marquages alternatifs existent tels que le RFP (*Red Fluorescent Protein*) ou le DAPI (*Di Aminido Phenyl lndo*) particulièrement utilisé pour marquer l'ADN.

2. La lumière d'excitation, dont la longueur d'onde dépend de la protéine fluorescente utilisée, est envoyée sur l'échantillon. Le trajet de la lumière est décrit sur la Figure 2.1 par la lumière de couleur bleue. Elle va dans un premier temps rencontrer un miroir dichroïque. Cet élément a la propriété

de pouvoir réfléchir la lumière ayant une certaine longueur d'onde, et de laisser passer la lumière ayant une autre longueur d'onde. Ce miroir permet de réfléchir la lumière d'excitation et de laisser passer la lumière d'émission. La lumière d'excitation passe donc par l'objectif pour se focaliser au niveau de l'échantillon, en illuminant l'entièreté de la zone d'intérêt. Le plan de focalisation est appelé le plan objet ou plan focal. En pratique, la lumière d'excitation est envoyée par une lampe ou un laser. Cela permet de contrôler de façon assez précise la quantité de photons envoyée, leur longueur d'onde, ainsi que la largeur du faisceau.

3. Les protéines fluorescentes vont alors émettre des photons d'une longueur d'onde différente : c'est lumière d'émission. Cette lumière va suivre le chemin décrit par la Figure 2.1. La lumière émise va alors passer par l'objectif, puis va traverser le miroir dichroïque, puis une seconde lentille. Le plan de focalisation entre l'objectif et la deuxième lentille est appelé plan de Fourier. La caméra est ensuite placée au plan focal de la deuxième lentille, aussi appelé plan image.

4. La caméra est composée d'une multitude de capteurs photosensibles qui convertissent les photons collectés en un signal électrique. Ce signal est ensuite converti en un signal numérique que l'on peut visualiser comme une image.

Le microscope à champ large a le mérite de permettre une acquisition rapide avec un minimum de réglages. Dans une configuration classique, on peut imager un échantillon de 80µm sur 2048 × 2048 pixels en quelques millisecondes.

2.3.2 Microscopie à localisation de molécules uniques

Résolution d'un microscope L'image d'une source ponctuelle par un microscope n'est pas un point, mais une tâche de diffraction. Cette tache de diffraction est aussi appelée PSF (point spread function), fonction d'étalement, ou réponse impulsionnelle. En l'absence d'aberrations et de masque de phase, la PSF du système est la tache d'Airy, voir Figure 2.3.

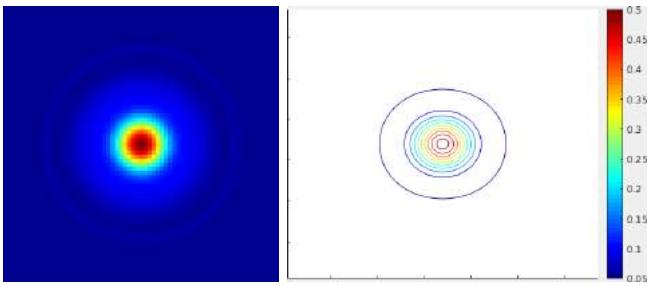
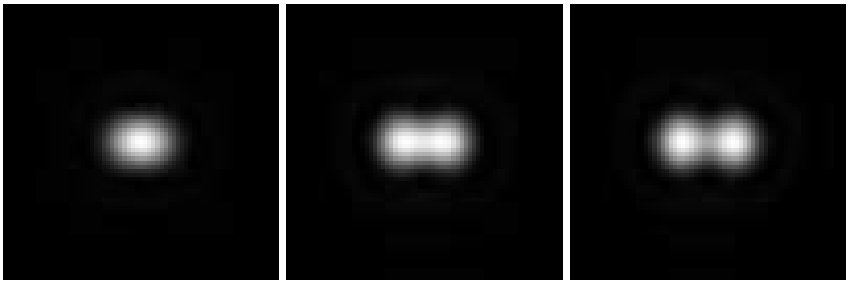


FIGURE 2.3: Tâche de Airy et ses lignes de niveaux.

La tache de Airy est donnée par :

$$h(r) = I_0 \left(\frac{2J_1(r)}{r} \right)^2 ,$$

où r est proportionnel à l'écartement du centre de la tache de diffraction, J_1 la fonction de Bessel du premier genre d'ordre un et I_0 l'intensité de la tache de diffraction au centre du disque. Le pouvoir de résolution d'un système optique peut être défini comme la distance minimale à laquelle peuvent se trouver deux objets que l'on est capable de différencier. Dans le cas où la PSF est donnée par la tache d'Airy, on peut définir la limite de résolution en fonction des paramètres du microscope. La forme de la tache d'Airy peut être modélisée en utilisant les équations de la Maxwell comme nous le verrons dans la Section 2.4. La distance entre son centre et son premier minimum est environ égal à $0.61\frac{\lambda}{\mathrm{NA}}$, où λ est la longueur d'onde, et NA l'ouverture numérique. Le critère de Rayleigh, introduit par Lord Rayleigh en 1896, définit la limite de résolution comme la demi-largeur de la tache d'Airy. Deux points sont donc considérés comme distinguables s'ils se trouvent à une distance supérieure à $d_{res} = 0.61\frac{\lambda}{\mathrm{NA}}$. La Figure 2.4 montre comment deux points apparaissent théoriquement en fonction de leur distance l'un à l'autre.



(a) Deux points séparés par $0.5d_{res}$.

(b) Deux points séparés par $0.75d_{res}$.

(c) Deux points séparés par d_{res}.

FIGURE 2.4: Simulation de la réponse d'un système optique à deux points plus ou moins séparés l'un de l'autre.

Super-résolution par localisation de molécule unique On parle de super-résolution lorsque l'on réussit à séparer deux objets ou structures étant séparées d'une distance inférieure à la limite de résolution définie par le critère de Rayleigh. Une des techniques les plus populaires pour réaliser cela est la localisation de molécule unique. Le principe est le suivant : observer une molécule fluorescente à la fois et estimer sa position de façon précise. Le critère de Rayleigh conjecture qu'on ne peut pas séparer deux points trop proches l'un de l'autre. Cependant, dans l'exemple de la Figure 2.4, on aurait été capable de donner la position précise de chacun des points si on les avait observés séquentiellement plutôt que simultanément. C'est le principe de la microscopie à localisation de molécule unique, on observe les molécules fluorescentes d'un échantillon biologique une par une. Le gain de résolution est alors significatif, permettant de passer d'une résolution de l'ordre de 200nm à une résolution pouvant atteindre quelques dizaines de nanomètres [RBZ06 ; Hua+08], voir Figure 1.2.

D'un point de vue pratique, cela consiste à envoyer la lumière d'excitation de façon très brève sur l'échantillon qui a subi un marquage fluorescent particulier. Il est possible d'activer seulement un faible nombre de molécules séquentiellement grâce à l'utilisation de protéines fluorescentes photo-activables. L'hypothèse sous-

jacente est que la probabilité d'avoir deux molécules proches qui émettent des photons en même temps est presque nulle. On observe un grand nombre d'images comme celles de la Figure 2.5, où seulement un petit nombre de molécules émettent des photons. Il faut donc réaliser un très grand nombre d'images pour voir suffisamment d'information. La reconstruction de la Figure 1.2 a nécessité 5000 acquisitions similaires à celle de la Figure 2.5.

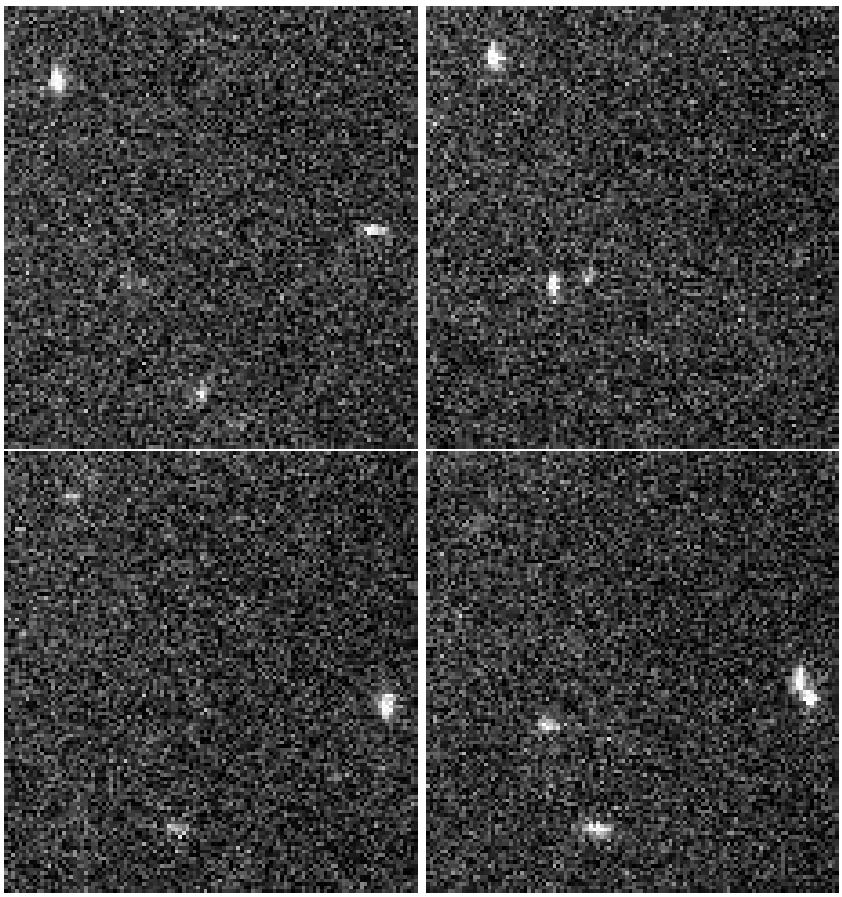


FIGURE 2.5: Exemple d'acquisition de microscopie à localisation de molécule unique. Les images sont de tailles 128×128.

Cette technique fut développée indépendamment par différents groupes, et est connue sous plusieurs noms : PALM (photo-activated localization microscopy) [Bet+06], STORM (stochastic optical reconstruction microscopy) [RBZ06], et fluorescent PALM [HGM06]. Cette technique fut récompensée par le prix Nobel de chimie [MLB14] en 2014. Toutes ces méthodes font partie de la même famille de microscopie à localisation de molécule unique, ou SMLM (single molecule localization microscopy).

Reconstruction tri-dimensionnelle Nous n'avons parlé que de résolution latérale jusqu'à présent. Cependant, les échantillons biologiques sont des structures tri-dimensionnelles. Le principe de super-résolution par localisation de

molécule peut être étendu en 3D au prix d'un contrôle fin du système optique. Pour permettre de localiser une molécule en 3D, on peut modifier le système pour que sa réponse impulsionnelle change en fonction de la profondeur à laquelle se trouve la molécule. On a alors à faire à des PSFs avec un motif astigmatique, en double-hélice, etc. Ces formes sont illustrées sur la Figure 1.3a.

Le contrôle de la forme de la PSF est réalisé en modifiant le front d'onde de la lumière d'émission. Il existe principalement deux techniques pour réaliser cela. La première approche consiste à introduire des éléments optiques actifs qui vont modifier le front d'onde de la lumière d'excitation. On introduit par exemple un miroir déformable sur le parcours de la lumière. Un miroir déformable est constitué d'un grand nombre de petits miroirs qui bougent de façon à introduire la déformation souhaitée. Les optiques actives modifient le front d'onde de manière à reproduire un motif donné. La deuxième technique consiste à placer un masque de phase au plan de Fourier. Cette méthode permet d'obtenir des PSFs plus variées qu'en utilisant de l'optique adaptative. On peut notamment introduire des discontinuités dans la phase pour obtenir des PSF en forme de double hélice [Pav+09], voir Figure 1.3a. Certaines approches proposent même d'estimer un masque de phase idéal pour la localisation des molécules [Neh+]. Le processus de création d'un masque de phase est complexe et coûteux. Il nécessite d'être réalisé en salle blanche.

Limite de résolution En introduisant des PSFs différentes de la tache d'Airy et en imageant séquentiellement un sous-ensemble de molécules, le critère de Rayleigh perd tout son sens. De façon générique, on peut définir un critère de résolution en microscopie à localisation de molécule unique en utilisant l'inégalité de Cramer-Rao. Cette définition permet de donner la limite théorique de résolution de ce type de modalité en fonction de la PSF utilisée et du bruit aléatoire présent sur l'image.

On se place dans le cadre idéal d'application de la microscopie à localisation de molécule unique, à savoir lorsqu'une seule source ponctuelle est présente. On note θ la vraie position de cette source, et $\hat{\theta}$ un estimateur de θ. Sous l'hypothèse que cet estimateur est non-biaisé, l'inégalité de Cramer-Rao donne :

$$\mathrm{Cov}(\hat{\theta}) \geq I^{-1}(\theta),$$

avec $\mathrm{Cov}(\hat{\theta})$ la matrice de variance-covariance de l'estimateur $\hat{\theta}$, et $I^{-1}(\theta)$ l'inverse de la matrice d'information de Fisher. L'inégalité $A \geq B$ pour deux matrices A et B signifie que $A - B$ est semi-définie positive.

La matrice d'information de Fisher peut être calculée dès lors que l'on connaît la réponse impulsionnelle du système. Nous référons à l'article de 2005 de Ram, Ward et Ober qui ont introduit cette méthode [RWO05] pour déterminer la précision de localisation en microscopie à molécule unique.

La connaissance de cette borne inférieure est utile non seulement pour estimer la limite de résolution que l'on ne pourra pas dépasser, mais aussi pour déterminer une zone d'incertitude autour de la position estimée des molécules. Cela permet de déterminer la taille des points dans les images reconstruites, voir Figure 1.2. Les systèmes les plus précis permettent d'atteindre une précision théorique de l'ordre de la dizaine de nanomètre.

Limite actuelle des techniques de microscopie à localisation de molécules uniques La microscopie à localisation de molécule unique offre un fort gain de résolution, d'un facteur 10 dans certains cas, pour un microscope champ large. En revanche, les besoins en ressources de calcul et de mémoire sont significativement plus importants. De plus, la localisation des molécules fluorescentes repose sur une connaissance précise de la PSF du système optique. La majorité des méthodes cherchent à corréler la PSF (estimée ou théorique) avec les observations [Kec+13 ; BZ17 ; Li+18c]. La quasi-totalité des méthodes reposent sur le fait que la PSF ne varie pas dans le champ de vue, limitant alors les acquisitions à des champs de vue réduit (là où la PSF ne varie pas trop). Nous référons à [Sag+19] pour une revue de différentes méthodes sur un même jeu de données.

La modélisation du flou par un opérateur variant spatialement permettrait en théorie, de considérer des champs de vue beaucoup plus grands (de l'ordre de 10 à 50 fois plus grand). Plusieurs problèmes majeurs se posent :

- La collecte de milliers d'images pour reconstruire l'image est très coûteuse en mémoire. Considérer des images 10 à 50 fois plus grandes représente alors un réel enjeu informatique.
- L'hypothèse de stationnarité de la PSF a l'avantage que son action peut être approchée par une convolution, rendant alors les algorithmes itératifs rapides à utiliser. L'utilisation d'un opérateur de flou variable ne sera possible que si son action peut être implémentée efficacement.
- La PSF du système peut simplement être estimée en imageant une bille. Une procédure de calibration similaire peut être effectuée en imageant plusieurs billes placées dans le champ de vue. C'est l'objet du Chapitre 3.

2.3.3 Résumé

L'approche standard en microscopie consiste à supposer que la réponse impulsionnelle du système ne varie pas spatialement. Nous avons vu que cette hypothèse est vraie seulement pour de petits champs de vue, limitant alors les acquisitions à de plus petites régions que ce que permettent les caméras actuelles. La prise en compte de ces variations spatiales soulève plusieurs problèmes que l'on se propose d'étudier dans cette thèse. A savoir, l'implémentation efficace (en temps de calcul et en mémoire) et l'estimation d'opérateurs de flou variable.

2.4 Modèle de formation de l'image

La microscopie computationnelle repose sur la modélisation précise des phénomènes physiques décrivant le processus de formation de l'image au sein du microscope. Dans cette partie, nous introduisons les équations de la théorie scalaire de la diffraction et discutons leurs avantages et leurs limites.

2.4.1 Modélisation physique

La lumière est une onde électromagnétique et se propage suivant les équations de Maxwell. Les champs électriques et magnétiques sont des composantes vectorielles, décrites par leur amplitude et leur direction (polarisation) en tout

point de l'espace. Cependant, dans beaucoup de systèmes optiques, on peut faire l'hypothèse que toutes les composantes de ces champs vectoriels se comportent de la même façon, on parle alors d'approximation scalaire, voir [Goo05] Chapitre 3. Cette hypothèse néglige le couplage entre les différentes composantes des champs vectoriels notamment au niveau du bord du domaine. Ce modèle n'a du sens que si les conditions suivantes sont respectées :

- La longueur d'onde de la lumière observée doit être petite devant la taille de l'objet de diffraction (la lentille de l'objectif).
- Le champ de diffraction doit être observé assez loin de l'objet de diffraction (approximation de Fraunhofer).

Ces hypothèses sont valides pour la majorité des microscopes à fluorescence.

On se place plus particulièrement dans le cadre donné par l'approximation de Fraunhofer, qui décrit le comportement des ondes dans un champ éloigné de la source de diffraction. En combinant ces approximations et le principe de Fresnel-Huygens, on obtient que la fonction d'étalement du système, ou PSF (Point Spread Function) est donnée à une constante près par le module au carré de la transformée de Fourier du champ électrique $E(\,\cdot\,|x_0, y_0, z_0)$ (aussi appelé fonction pupille). On a donc une relation entre la PSF dans le plan image et le champ électrique dans le plan pupille. À partir d'une source ponctuelle placée au point (x_0, y_0, z_0), la PSF du système s'exprime comme suit [Goo05 ; PSM17] :

$$K(x', y'|x_0, y_0, z_0) \propto \left|\mathcal{F}\left(E(\,\cdot\,|x_0, y_0, z_0)\right)(x', y')\right|^2, \tag{2.1}$$

où $\mathcal{F}$ est la transformée de Fourier, et (x', y') est le point d'observation. La fonction $K(\,\cdot\,,\,\cdot\,|x_0, y_0, z_0)$ est la réponse impulsionnelle au point (x_0, y_0, z_0) de l'opérateur de flou du système optique.

En pratique, la valeur du champ électrique E dans le plan pupille (ou plan de Fourier) n'est pas accessible puisque l'on mesure seulement des intensités dans le plan image. Dans un système idéal, le champ électrique prend la forme suivante :

$$E(x, y) \propto \rho(x, y) \exp\left(i\Phi(x, y)\right) \exp\left(i\Phi_{ax}(x, y; z_0)\right) \exp\left(i\Phi_{lat}(x, y; x_0, y_0)\right),$$

où i est le nombre complexe tel que $i^2 = -1$, et Φ est la phase au plan de Fourier, voir le schéma de la Figure 2.1. La fonction ρ est définit pour tout $\mathbf{x} \in \mathbb{R}^2$ par :

$$\rho(\mathbf{x}) = \begin{cases} 1 & \text{si } \|\mathbf{x}\|_2 \leq \frac{\mathrm{NA}}{\lambda}, \\ 0 & \text{autrement,} \end{cases}$$

où NA est l'ouverture numérique de l'objectif et λ la longueur d'onde. La fonction Φ_{ax} est définie par :

$$\Phi_{ax}(x, y; z_0) = 2\pi z_0 \sqrt{\left(\frac{\mathrm{n}_I}{\lambda}\right)^2 - (x^2 + y^2)},$$

où n_I est l'indice de réfraction du milieu dans lequel est l'échantillon. Ce terme prend en compte le fait que le point source est en z_0 et non au plan focal ($z = 0$). La fonction Φ_{lat} est un terme de translation de la phase qui prend en compte que l'objet positionné en (x_0, y_0, z_0) n'est pas nécessairement le long de l'axe optique, et est défini par :

$$\Phi_{lat}(x, y; x_0, y_0) = \frac{2\pi}{\lambda f}(x x_0 + y y_0),$$

où f est la distance entre le plan focal du système optique et l'objectif.

La phase Φ est généralement exprimée dans la base formée par les polynômes de Zernike. Ces polynômes forment une base orthonormale sur le disque unité. La fonction Φ s'exprime alors par :

$$\Phi(x, y) = \sum_{i \in \mathcal{I}} c_i Z_i(x, y),$$

où Z_i est le i-e polynôme de Zernike (en utilisant l'indexation de Noll [Nol76]). Les trois premiers polynômes de Zernike n'influencent pas la forme de la PSF, et sont souvent écartés. On utilise alors $\mathcal{I} = \{4, \ldots, I\}$, avec $I \geq 4$. L'intérêt d'utiliser les polynômes de Zernike réside dans leur interprétation physique. Pour ne citer que les plus rencontrés, les polynômes 5 et 6 vont participer à obtenir une PSF astigmatique (allongée verticalement et horizontalement), et les polynômes 7 et 8 vont participer à obtenir une PSF en forme de Coma. Si tous les coefficients c_i sont identiquement nuls, on retrouve la tache de Airy.

Les coefficients des différents polynômes de Zernike correspondent à un développement des erreurs de phase par rapport à un objectif idéal. Les premiers polynômes correspondent à des aberrations optiques classiques (astigmatisme, coma, aberration sphérique). Différents polynômes et leur PSF associée sont affichés sur la Figure 2.6.

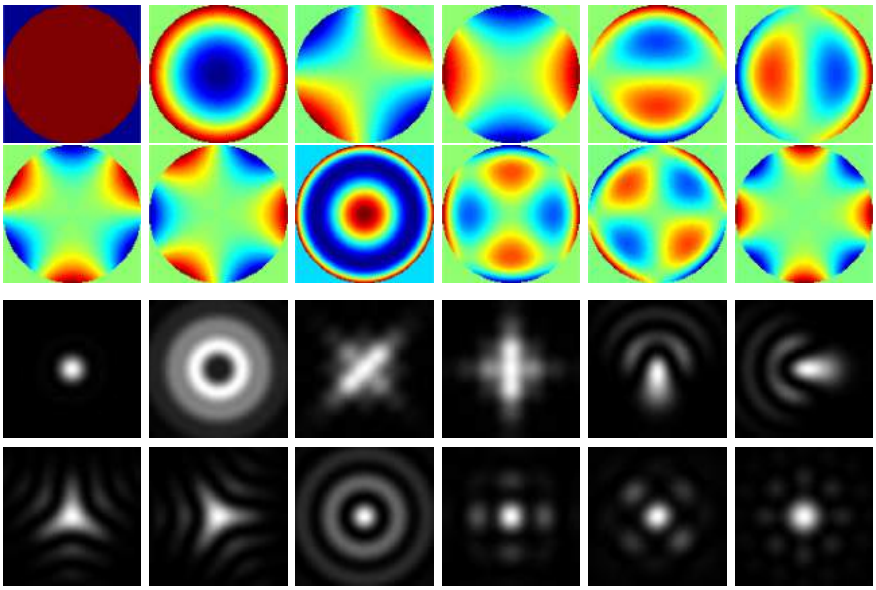


FIGURE 2.6: Polynôme de Zernike sur le cercle unité et PSF associée en utilisant l'équation (2.1). Les polynômes affichés sont $(Z_i)_{i \in \mathcal{I}}$ pour $\mathcal{I} = \{3, \ldots, 14\}$ en utilisant l'indexation de Noll.

Finalement, l'image obtenue peut être décrite par la relation suivante entre la PSF $K(x', y'|x_0, y_0, z_0)$ d'un objet lumineux en position (x_0, y_0, z_0) et la densité de fluorophores $u(x_0, y_0, z_0)$ émis par l'objet que l'on cherche à imager :

$$v(x', y') = \int K(x', y'|x_0, y_0, z_0) u(x_0, y_0, z_0) \mathrm{d}x_0 \mathrm{d}y_0 \mathrm{d}z_0.$$

Une propriété importante qui apparaît avec cette formule (et qui a été utilisée dans les approximations utilisées) est la dépendance linéaire en l'échantillon. Cette propriété de linéarité, c'est-à-dire que la réponse du système à plusieurs stimuli est identique à la somme des réponses à chaque stimulus, contribue à la simplification des équations de la physique.

2.4.2 Limite du modèle physique

En pratique, la définition théorique de la PSF par l'équation (2.1) est souvent imprécise à cause d'aberrations optiques difficiles à prendre en compte. Les paramètres sont souvent connus approximativement, et leur estimation reste difficile. Une façon de capturer ces aberrations est de chercher une décomposition du terme de phase Φ dans la base des polynômes de Zernike, qui reproduit les observations. Cette approche porte le nom d'estimation de phase (*phase retrieval* en anglais). La dépendance en les coefficients dans la base des polynômes de Zernike n'est pas linéaire, ce qui rend le problème relativement compliqué. Des résultats théoriques récents font leur apparition, notamment avec des idées de *lifting* convexe [CSV13 ; LLB16a ; BB19]. Cependant, on observe toujours une disparité entre ces méthodes qui reposent sur des hypothèses souvent difficile à vérifier en pratique, et les méthodes utilisées en microscopie qui proposent d'estimer les paramètres du modèle par un maximum de vraisemblance [PSM17 ; Ari+18]. L'apprentissage machine a aussi été utilisé pour proposer de résoudre ce problème [SL20b]. Cette approche d'estimation de phase présente l'avantage de pouvoir se rattacher à une interprétation physique. D'autres types de dégradations peuvent apparaître et aussi être estimées simultanément tels que le mauvais alignement latéral du masque de phase placé au plan de Fourier [PSM17], ou le fond dû à l'autofluorescence localement autour de la PSF [Ari+18]. Dans toutes les situations, l'incertitude de l'amplitude de la PSF et son mauvais alignement peuvent introduire des erreurs supplémentaire. Dans le cas ou plusieurs PSFs estimées sont utilisées dans une analyse en composantes principales, c'est biais ont des conséquences importantes [Deb+20a].

Les aberrations capturées par l'estimation de la phase de la fonction pupille peuvent être arbitraires, et donc nécessiter un nombre important de polynômes de Zernike. L'étude de l'identifiabilité de tous les paramètres décrivant un système et leur unicité s'avère bien souvent impossible.

Finalement, les approximations effectuées pour aboutir à la théorie de la diffraction scalaire sont des sources d'incohérences avec les observations. Une amélioration possible du modèle consiste à considérer la théorie de la diffraction vectorielle, qui contrairement à la théorie de la diffraction scalaire, ne suppose pas que le champ électrique (qui est une composante vectorielle) est isotrope dans toutes les directions de l'espace. Cette hypothèse permet de simplifier très largement les équations de Maxwell, et de se ramener à des formules explicites. Sans cette hypothèse, il faut alors résoudre les équations de Maxwell sur l'ensemble du domaine [Shi+20]. Cela peut se faire numériquement avec des logiciels de résolution d'équations aux dérivées partielles, mais à un coût très élevé en terme de ressources de calculs. L'utilisation de ce modèle permet notamment la prise en compte de l'orientation du dipôle ou de sa rotation[Bac+14]. Ces phénomènes sont importants dans beaucoup de situations [LBM13].

2.4.3 Résumé

Au prix de certaines approximations acceptables dans beaucoup de situations, on peut simplifier les équations de Maxwell pour décrire la relation entre la densité de fluorophores de l'échantillon et l'observation. Cela nécessite la connaissance précise des paramètres du système optique. Cependant, les différentes approximations combinées aux incertitudes expérimentales peuvent conduire à un modèle erroné. Les différents paramètres physiques peuvent alors être estimés à partir d'observations, mais leur dépendance est bien souvent non-linéaire et le problème inverse associé est difficile à résoudre. De plus, la modélisation d'un système optique par une simple fonction pupille ne permet pas de prendre en compte les variations spatiales de la réponse impulsionnelle. Cette approche est également limitée aux processus de formation d'images linéaires. Lorsque l'échantillon est composé de plusieurs couches par exemple, ce modèle n'est plus valable, voir chapitre 10.

2.5 Approximation d'opérateurs linéaires : décomposition en convolution-produit

Un objectif pratique de cette thèse est le traitement d'images de grande taille. Typiquement, en microscopie, les images peuvent aller de 512×512 à 500000×500000 pixels (en histopathologie par exemple, en utilisant un système de scanner [Ngu+18]) pour des acquisitions bidimensionnelles. Comme nous l'avons vu dans le paragraphe précédent, une hypothèse raisonnable est de considérer l'opérateur de flou comme linéaire. Dans une version discrétisée du problème, un opérateur linéaire $H : \mathbb{R}^n \to \mathbb{R}^n$ peut être représenté par une matrice de taille $n \times n$. Pour une image de taille $n = 512 \times 512$, on doit alors travailler avec une matrice (le noyau de son opérateur) de taille 512^4, ce qui nécessite 500 Go avec un stockage en double précision. Cette approche brutale n'est pas envisageable dans des situations pratiques.

Sous des hypothèses de régularité sur l'opérateur H, on peut approcher sa SVIR (*Space Varying Impulse Response*) par une décomposition de faible rang : ce sont les développements en convolution-produit. Cette décomposition a la propriété d'être économique en ressource de calcul et en mémoire, en plus de présenter des propriétés d'approximation intéressantes pour une classe d'opérateurs proches de ceux rencontrés en optique. Dans cette partie, nous présentons brièvement ces résultats.

2.5.1 Développement en convolution-produit

Soit $H : L^2(\Omega) \to L^2(\Omega)$ un opérateur linéaire, défini à partir de son noyau K par :

$$H(u)(x) = \int_\Omega K(x,y)u(y)\mathrm{d}y, \tag{2.2}$$

pour tout $u \in L^2(\Omega)$, avec $\Omega \subset \mathbb{R}^d$. On peut alors définir la réponse impulsionnelle variant spatialement, ou SVIR (*Space Varying Impulse Response*), comme une version translatée du noyau :

$$S(x,y) \overset{\text{def.}}{=} K(x+y,y), \forall (x,y) \in \Omega^2.$$

La réponse impulsionnelle (ou PSF pour *point spread function*) du système en un point $y \in \Omega$ du domaine, est la fonction $S(\,\cdot\,, y)$. En pratique, les réponses impulsionnelles d'un système optique sont relativement similaires au sein du domaine d'acquisition. Une hypothèse raisonnable est de supposer qu'elles peuvent toutes être approchées par un petit nombre d'éléments communs. Cela est capturé par l'hypothèse suivante.

Hypothèse 2.1: Décomposition des PSFs

On suppose qu'il existe une famille $(u_k)_k \in L^2(\Omega)^K$ telle que pour tout $y \in \Omega$:

$$S(\,\cdot\,, y) \approx \sum_{k=1}^{K} v_k(y) u_k(\,\cdot\,),$$

avec $v_k(y) \in \mathbb{R}$.

L'approximation de S dans l'équation (2.2), donne lieu à la formule suivante :

$$\begin{aligned} H(u)(x) &\approx \int_\Omega \sum_{k=1}^{K} v_k(y) u_k(x-y) u(y) \mathrm{d}y \\ &= \sum_{k=1}^{K} u_k \star (v_k \odot u), \end{aligned} \tag{2.3}$$

où $\star$ est l'opérateur de convolution, et $\odot$ le produit de Hadamard. Cette décomposition, appelée approximation en convolution-produit (car on effectue un produit suivi d'une convolution), se traduit par *product-convolution* en anglais. On note H_K le convolution-produit d'ordre K tel que :

$$H_K(u) = \sum_{k=1}^{K} u_k \star (v_k \odot u), \forall u \in L^2(\Omega). \tag{2.4}$$

Ce type de décomposition connait de nombreuses variantes dans la littérature [FR05b ; HB11 ; MP12 ; Den+15 ; Esc16]. Nous renvoyons le lecteur intéressé à ces articles pour plus d'informations. Nous énonçons maintenant les propriétés qui rendent la décomposition (2.4) intéressante pour les applications en microscopie.

Une **première propriété** qui découle de l'hypothèse 1 est que la SVIR de l'opérateur H est bien approchée par un tenseur de faible rang. On a alors :

$$S \approx \sum_{k=1}^{K} u_k \otimes v_k,$$

où $\otimes$ est l'opérateur de produit tensoriel.

La **deuxième propriété** importante vient de la décomposition (2.4). On voit ainsi qu'appliquer un développement en convolution-produit d'ordre K nécessite le calcul de K convolutions et de K multiplications par une matrice diagonale. Le coût est donc multiplié par K par rapport à une convolution usuelle.

La **troisième propriété** importante est le contrôle de l'erreur entre un opérateur dont la réponse impulsionnelle varie de façon lisse et son approximation par un convolution-produit. Cette propriété est capturée par le théorème suivant.

Theorem 2.5.1. *Taux d'approximation de développement en convolution-produit [EW17] Soit $\Omega \subset \mathbb{R}^d$ Soit $H : L^2(\Omega) \to L^2(\Omega)$ un opérateur linéaire et S sa SVIR. Supposons que pour tout $x \in \Omega$ la fonction $S(x, \cdot) \in H^s(\Omega)$. Supposons également que $\|S(x, \cdot)\|_{H^s(\Omega)}$ soit uniformément borné en x, i.e :*

$$\sup_{x\in\Omega} \|S(x, \cdot)\|_{H^s(\Omega)} \leq C < +\infty.$$

Alors

$$\|H - H_K\|_{\mathrm{HS}} = \mathcal{O}\left(\kappa^{d/2} K^{-s/d}\right),$$

avec κ tel que $S(x, y) = 0$ pour tout $|x| > \kappa/2$. Cette erreur d'approximation ne peut pas être améliorée uniformément sur les opérateurs dont la SVIR vérifie les hypothèses ci-dessus.

L'hypothèse de restriction de support de la fonction $S(\cdot, y)$ pour tout $y \in \Omega$ assure que l'on puisse décomposer cette fonction sur une base de faible dimension. Cette hypothèse peut être relâchée si l'on se place sous l'Hypothèse 1 : toute fonction suffisamment régulière peut être bien approchée par un petit nombre d'élément de base (e.g. les fonctions à bande limitée).

La **quatrième propriété** des développements en convolution-produit est leur possible interprétation physique. Le famille $(u_k)_k$ est une base dans laquelle s'expriment les réponses impulsionnelles, on peut parler "*d'eigen-PSF*". La famille $(v_k)_k$ quant à elle, décrit les variations spatiales. Dans le cas d'un opérateur invariant spatialement, la famille $(v_k)_k$ serait simplement composée d'éléments constants. La Figure 2.7 présente le cas d'une décomposition avec $K = 2$ éléments :

$$H(u) = u_1 * (v_1 \odot u) + u_2 * (v_2 \odot u), \tag{2.5}$$

où u est un peigne de Dirac.

2.5.2 D'autres types d'approximations

La littérature concernant l'approximation d'opérateurs linéaires est très vaste.

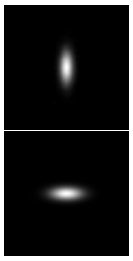
Il existe trois grandes familles de décompositions :

- l'approximation du noyau K par une structure de faible rang (dans des bases d'ondelettes ou à l'aide de $\mathcal{H}$-matrice par exemple),
- l'approximation de la SVIR par une structure de faible rang,
- l'utilisation d'approximation spécifique à une application précise.

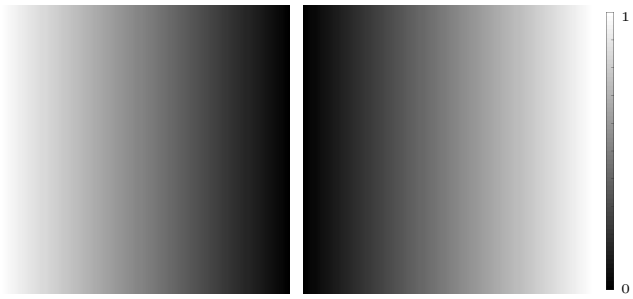
Une présentation de ces différentes méthodes ainsi qu'une liste complète de références peuvent être trouvées dans la thèse de Paul Escande [Esc16].

2.5.3 Résumé

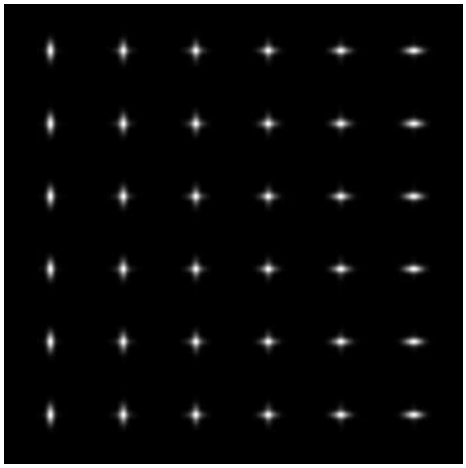
La réponse impulsionnelle d'un système optique varie généralement dans le champ de vue, mais souvent de façon lisse. Mathématiquement, un tel opérateur est bien approché par une décomposition en convolution-produit. Cette approximation a l'avantage de posséder de très bonnes propriétés numériques et mathématiques, rendant alors possible leur utilisation sur des signaux de grandes dimensions.



(a) Famille $\{u_1, u_2\}$ de taille 64×64.



(b) Famille $\{v_1, v_2\}$ de taille 512×512.



(c) Convolution-produit (2.5) appliqué à une grille de points uniformément répartis sur le champ de vue.

FIGURE 2.7: Exemple de convolution-produit avec $K = 2$.

Part II

Subspace of operators

Français : Cette partie est dédiée à l'estimation d'opérateur et de sous-espace d'opérateur. Nous introduisons dans un premier temps une boîte à outils permettant de réaliser cette estimation à partir d'une image de microbilles. Dans un second chapitre, nous présentons un cadre formel pour construire un estimateur de sous-espace d'opérateurs structurés. Dans le troisième chapitre, nous illustrons cet estimateur de sous-espace sur un exemple concret de microscopie.

English: This part is dedicated to operator and subspace estimation. We first introduce a toolbox to perform this estimation from an image containing micro-beads. In a second chapter, we present a formal framework to construct a structured operator subspace estimator. In the third chapter, we illustrate this subspace estimator on a concrete microscopy example.

Chapter 3

Estimating an impulse responses subspace

Contents

3.1 Introduction **38**
3.1.1 Contributions 40
3.2 Mathematical and algorithmic foundations **42**
3.2.1 Image formation model 42
3.2.2 Initialization 45
3.2.3 Estimation algorithm 51
3.3 Numerical results **53**
3.3.1 Simulation data and dictionary learning 53
3.3.2 Microscopy data 64
3.4 Conclusion **67**
3.5 Appendices **70**
3.5.1 Regularizing the PSF family 70
3.5.2 Handling the bead function 70
3.5.3 Why a Gaussian fitting implicitly assumes centered PSFs? 71
3.5.4 Estimating space variations 72

Résumé : *Ce chapitre décrit une méthode d'estimation d'une PSF, ou d'un sous-espace vectoriel de faible dimension capturant un ensemble de PSFs, à partir de l'observation d'une ou plusieurs images contenant des microbilles. Cette approche se présente actuellement sous la forme d'un ensemble de scripts Matlab et C++ qui sont fortement automatisés, permettant de traiter des données variées avec un minimum d'expertise. Le but final de ces travaux est la distribution d'une méthode robuste et accessible par le biais d'un module Fiji. L'approche proposée ici présente différentes caractéristiques, telles que la possibilité d'estimer une PSF variant spatialement, la prise en compte de la taille de la bille, la possibilité d'utiliser plusieurs images pour estimer une même PSF (ou une même famille), la possibilité de traiter des acquisitions aussi bien 2D que 3D, ou une estimation*

jointe de la forme de la PSF et de la position de cette dernière sur un domaine continu.

Abstract: *This chapter describes a method for estimating a PSF, or a PSF subspace, based on the observation of one or more images containing micro-beads. This approach currently takes the form of a set of Matlab and C++ scripts that are highly automated, allowing the processing of various data with a minimum of expertise. The final goal of this work is the distribution of this method through a Fiji plugin. The method presented here has various features, such as the possibility to estimate a spatially varying PSF, taking into account the bead's width, the possibility to use several images to estimate a single PSF (or a subspace), the possibility to process both 2D and 3D acquisitions, a joint estimation of the shape of the PSF and its position on continuous domain.*

The initialization procedure is strongly inspired by the preprint [Deb+20a]:
Debarnot, V., Escande, P., Mangeat, T., & Weiss, P. (2020). Learning low-dimensional models of microscopes.
The toolbox and the refined PSF extraction is based on a current work in collaboration with Daniel Sage, Emmanuel Soubies, Thomas Mangeat, and Pierre Weiss.

Contents

3.1 Introduction **38**
3.1.1 Contributions 40
3.2 Mathematical and algorithmic foundations **42**
3.2.1 Image formation model 42
3.2.2 Initialization 45
3.2.3 Estimation algorithm 51
3.3 Numerical results **53**
3.3.1 Simulation data and dictionary learning 53
3.3.2 Microscopy data 64
3.4 Conclusion **67**
3.5 Appendices **70**
3.5.1 Regularizing the PSF family 70
3.5.2 Handling the bead function 70
3.5.3 Why a Gaussian fitting implicitly assumes centered PSFs? 71
3.5.4 Estimating space variations 72

3.1 Introduction

The characterization of the point spread function (PSF) of an imaging system is a half-century old problem that has become increasingly important as traditional systems have evolved toward modern computational imaging techniques. Motivated for several decades by the study of the properties (*e.g.,* resolution limit) of conventional imaging systems[Hop55; Sto69; CJB11] and the improvement of deconvolution techniques,[Aga+89; HSA90; CGI96; WSS01; AK00;

Sou14] it has been recently revitalized with the advent of PSF-engineering for three-dimensional single molecule localization microscopy (3D-SMLM) [Hua+08; Pav+09; Jue+08; She+14; Sag+19]. The very principle of 3D-SMLM is based on the optical design of PSFs whose shapes allow to infer, through numerical computations, the 3D positions of single fluorescent molecules from their emission patterns. As a consequence, it is crucial to obtain an accurate numerical model of the PSF in order to reach the promised nanoscale resolution [Mlo+18; Xu+20]. It is noteworthy to mention that, beyond 3D-SMLM, the quality of the PSF model is also an essential component of other off-the-shelf super-resolution microscopy techniques such as stimulated emission depletion (STED) microscopy[HW94] and structured illumination microscopy (SIM) [HC99; Gus00; Gus+08]. Overall, accurate characterization of the PSF plays a central role in the performance of any optical imaging system, if only to assess the quality and alignment of optical elements, and to monitor the calibration of the system over time [MC10; CJB11; TMK14].

Sophisticated theoretical PSF models,[Hop55; Sto69; McC64; GL91] including certain type of aberrations such as layered refractive index mismatches,[GL91] can be derived from diffraction theory [BW13; Goo05]. However, not only do they correspond to ideal imaging conditions that are never encountered in practice, but they also depend on physical quantities such as wavelengths, refractive indices, temperatures, and other system parameters, which may not be (accurately) known at the time of acquisition. These limitations compromise the direct use of these theoretical models in numerical reconstruction algorithms.

The standard practice is then to estimate PSFs from experimental measurements. The most commonly used protocol for this purpose is the imaging of point-source objects such as stars in astronomy [SCU16] and fluorescent microspheres (or single molecules) in microscopy[Aga+89; HSA90; CJB11] as they provide a direct measure of the impulse response of the system (*i.e.,* PSF). Assuming that this response is spatially invariant and that the imaged objects are sub-resolved, analytical models, going from simple Gaussian models[ZZO07] to non-trivial physical PSF models,[Pan+09; Kir+13] can be fit to these measured PSFs in order to estimate unknown parameters. The properties of Gaussian models are appealing to deploy fast algorithms. However, they are only accurate to approximate the main lobe of the PSF of specific aberration-free systems [ZZO07]. Although more accurate, existing theoretical PSF models derived from diffraction theory cannot represent the variety of aberrations encountered in real optical systems. They are therefore also rarely used in practice. A more flexible way to characterize the PSF is through the pupil function, see Section 2.4.

Estimating the phase of the pupil function is often reduced to find the decomposition of the phase using Zernike polynomials [Han+04; QPP12; Liu+13; Xu+20]. Corrective terms can be added to take into account the misalignment of the bead along the optical axis or the fact that the sample is not exactly at the focal plane [Ari+18; PSM17]. The advantage of this approach lies in the fact that the Zernike coefficients make it possible to cling to a physical interpretation of the optical aberrations (astigmatism, coma, etc) [She+14]. However, some restoration algorithms require the use of PSF in the spatial domain (e.g. off-the-grid single molecule localization) and it is therefore necessary to use a physical model, that possibly limits the range of possible PSFs. Phase retrieval remains a mathematically difficult problem. Recent approaches allow theoretically to solve this problem [CSV13; CSV13; ARR13], at the price of a high numerical

cost and with hypotheses possibly difficult to verify. This type of approach is only rarely used in microscopy or astrophysics.

The other possibility remains to estimate the shape of the PSF directly in the spatial domain [Ber11; Sou+12; Mbo+15; Mou+15; Ngo+16; BZ17; Li+18a]. Contrarily to phase estimation, this makes it possible to take into account other kind of aberrations of the optical system. In reality, the response of a bead by the optical system is not simply its convolution by the local PSF due to the optics[AK00]. In a more complete model, we should take into account, among other things, the effects of sensor integration. It is known that the value of a pixel is not simply due to the contribution of the light over this pixel (which would result in integration over a square). There are at least two reasons for that: the sensitivity of the sensor is not uniform, and there is a diffusion effect between sensors [Hol+95; Lau99]. These effects are one example among others [Die+15]. Estimating what we named here a PSF, in the spatial domain, actually captures the contribution of all these effects.

Imaging beads remains the simplest method to obtain an indirect measure of the PSF. The observation necessarily depends on the size of the bead. Taking it into account in the model is therefore essential to avoid any bias. This is particularly important for large bead size, typically larger than the pixel size [Han+04; YSG06; Pan+09].

Another phenomenon present on all systems is the spatial variation of PSF. This effect, which is often neglected for practical reasons, or because the optical system allows it, is nevertheless present and can be important, especially over large fields of view [Ber11; Ngo+16].

3.1.1 Contributions

In this work, we present a method for estimating a family of PSFs from one or more images containing micro-beads. The estimation procedure is currently developed using *Matlab* and *C++* functions, with the objective of releasing it shortly as a Fiji (Java) plugin. This problem has been studied for a long time now and Table 3.1 summarizes different existing approaches and their main specifications. We highlight the following features:

- "Iterative estimation of positions": specifies whether the position of the beads is estimated in conjunction with the PSF shape.
- "Bead size": specifies whether the size of the bead is taken into account.
- "Space varying PSF": specifies whether spatial variation models of PSF can be used.
- "Regularizing basis": specifies the regularization to be imposed on the estimated PSFs (if there is one).
- "Spatial estimation/phase retrieval": specifies whether the PSF is estimated in the spatial or frequency domain.
- "2D/3D": specifies if the estimate takes into account 2D and/or 3D data.
- "Toolbox": specifies whether a toolbox is available.

The symbol ✓ means that the reference has the corresponding feature and the absence of symbol means that the reference does not have this feature.

Reference	**Iterative estimation of positions**	**Bead size**	**Space varying PSF**	**Regularizing basis**	**Spatial estimation/ phase retrieval**	**2D/3D**	**Toolbox**
[Han+04]		✓		Zernike	phase	2D/3D	
[YSG06]		✓			spatial	2D/3D	
[ZZO07]				Gaussian model	spatial	2D/3D	
[Pan+09]		✓		scalar diffraction model	spatial	2D/3D	
[Ber11]			✓	learned basis	spatial	2D	shell
[QPP12]			✓	Zernike	phase	2D/3D	
[Kir+13]	✓			Gibson-Lanni	spatial	3D	
[Liu+13]				Zernike	phase	2D/3D	
[She+14]			✓	Zernike	phase	2D/3D	
[Ngo+16]			✓	learned basis	sp/phase	2D/3D	code available
[BZ17]	✓			spline	spatial	2D/3D	Python
[PSM17]				Zernike	phase	2D/3D	
[Ari+18]	✓			Zernike	phase	2D/3D	Fiji
[Li+18a]				learned from scalar theory	spatial	2D/3D	Matlab
[Li+18c]	✓			spline	spatial	2D/3D	Matlab
[Xu+20]	✓			Zernike	phase	2D/3D	
[Tur+20]			✓	learned from data	spatial	2D	Fiji & Matlab
PSF-Estimator	✓	✓	✓	learned from scalar theory	spatial	2D/3D	Matlab

Table 3.1: References and their main features regarding PSF extraction.

3.2 Mathematical and algorithmic foundations

3.2.1 Image formation model

We are interested in the setting where we have access to one or more images containing point sources, obtained from an optical system. Examples of images obtained with a wide-field microscope (3D with multifocus), and 3D-SPIM (Selective Plane Illumination Microscopy) are presented in Fig. 3.12 a) and Fig. 3.14 a).

More formally, we suppose that we observe $|J|$ sources, placed at the unknown positions $\mathbf{x}_j \in \mathbb{R}^d$, and with unknown amplitude $w_j \in \mathbb{R}$. The measurement function is then given by

$$\mathbf{y} = \mathcal{A}\left(\sum_{j=1}^{|J|} w_j \delta_{\mathbf{x}_j}\right),$$

where $\mathcal{A}$ is an unknown operator. In the following, we assume that the operator $\mathcal{A}$ is linear. This means that its action on the sum of several sources is equal to the sum of its action on each individual source. Then, the observation model becomes

$$y = \sum_{j=1}^{|J|} w_j \mathcal{A}(\delta_{\mathbf{x}_j}).$$

The blur operator $\mathcal{A}$ is often endowed with smoothness properties. In this work, we make the hypothesis that for any points in the field of view, the impulse responses of this operator can be expressed in a common subspace. We let $(h_l)_{1\leq l\leq L}$ denote a family of elementary functions generating this subspace, and we have

$$y = \sum_{j=1}^{|J|} w_j \sum_{l=1}^{|L|} \gamma_l[j] h_l(\ \cdot\ - \mathbf{x}_j),$$

where the vector $\boldsymbol{\gamma}_j \overset{\text{def.}}{=} (\gamma_l[j])_l \in \mathbb{R}^{|L|}$ gives the decomposition of the PSF at the point $\mathbf{x}_j$ in the subspace generated by the functions $(h_l)_l$. The most popular model corresponds to $L = 1$. It allows to estimate the PSF of a stationary system corresponding to a convolution model. Setting $L > 1$ allows to encompass space varying kernels, variations in the axial direction, or other exotic models arising in the recent trend of PSF engineering. The space variations of the system PSF are then encoded through the vectors $\boldsymbol{\gamma}_j$ [Den+15; EW17; Deb+20a]. The main assumption behind model (3.1) is that L should be small. In other words, the impulse response of the system belong to a low dimensional subspace that is characterized by the functions h_l. This assumption is challenged numerically in Figure 3.10. In this experiment, we show that a subspace containing 158 elements is enough to capture more than 99% of the energy of about 9000 PSFs randomly generated using scalar diffraction theory.

There is a first important ambiguity that appears in this formulation: we can shift the PSF h_l by an arbitrary vector $e \in \mathbb{R}^d$, and shift the position of the points in the opposite direction, and still produce the same observation. This ambiguity makes the problem ill-posed. This can be taken into account by constraining the family of elementary PSFs $(h_l)_l$ to live in a $|M|$-dimensional subspace. For instance, we can force the firstorder moment of the PSF (its center of mass) to be null. In the following, we let $U = [u_1, \ldots, u_{|M|}]$ denote the functions that span this $|M|$-dimensional subspace. The elementary PSFs are now given by $h_l = U\mathbf{c}_l$, where the unknown is now $\mathbf{C} \in \mathbb{R}^{|M|\times|L|}$. We can also impose the PSFs to be band-limited by taking U as the first functions of the Fourier basis, see Figure 3.9, or to live in a subspace generated by a large collection of admissible PSFs previously obtained by acquisition or simulation, see Figure 3.10. This gives rise to the following formation model

$$y = \sum_{j=1}^{|J|} w_j \sum_{l=1}^{|L|} \gamma_l[j] U\mathbf{c}_l(\ \cdot\ - \mathbf{x}_j).$$

In microscopy, we never observe the impulse response of the system, but rather the action of the system on a bead of **known size**. We add this information to our observation model by convolving the elementary PSFs by a function b describing the bead intensity profile. In our experiments, we will simply use the indicator of a ball of known radius denoted R_b. Arbitrary profiles could be taken into account although it is not implemented in the toolbox. Finally, the observation model becomes

$$y = \sum_{j=1}^{|J|} w_j \sum_{l=1}^{|L|} \gamma_l[j] (U\mathbf{c}_l \star b)(\ \cdot\ - \mathbf{x}_j),$$

where $\star$ stands for the convolution.

This observation model is available numerically only after a discretization. We let $\mathbf{I} = \mathbf{I}_1 \cdots \times \cdots \mathbf{I}_d \subset \mathbb{Z}^d$ denote a set that contains the indexes of the voxels and $\mathbf{\Delta} = (\Delta_1 \cdots \Delta_d)^T \in \mathbb{R}^d_{>0}$ denote a vector formed out of the sampling steps of the acquisition system in each dimension (voxel size). For example, for a two-dimensional image composed of $N \times N$ pixels, the standard setting where the sampling is realized on a unitary Cartesian grid corresponds to $\mathbf{I}_1 = \mathbf{I}_2 = \{1, \ldots, N\}$ and $\mathbf{\Delta} = (1/N, 1/N)$. An additive noise is added to the observation model, which finally states as:

$$\mathbf{y}[\mathbf{i}] = \sum_{j=1}^{|J|} \sum_{l=1}^{|L|} \boldsymbol{\gamma}_j[l](U\mathbf{c}_l \star b)(\mathbf{\Delta} \odot \mathbf{i} - \mathbf{x}_j) + \mathbf{n}[\mathbf{i}], \ \forall \mathbf{i} \in \mathbf{I}, \tag{3.1}$$

where $\odot$ denotes the Hadamard (component-wise) product.

PSF discretization: Since only discrete quantities can be handled on a computer, we need a discrete representation of the continuous functions $h_l = U\mathbf{c}_l$ and b. To that end, we assume that there exists a vector $\mathbf{h}_l \in \mathbb{R}^{|\mathbf{K}|}$ such that

$$h_l = \Phi(\mathbf{h}_l) \stackrel{\text{def.}}{=} \sum_{\mathbf{k} \in \mathbf{K}} \mathbf{h}_l[\mathbf{k}] \varphi\left(\ \cdot \ - (\mathbf{\Delta}/r) \odot \mathbf{k}\right), \tag{3.2}$$

where $\varphi \in L_2(\mathbb{R}^d)$ is a generating function, $r > 0$ allows to vary the resolution of the estimated basis, and $\mathbf{K} \subset \mathbb{Z}^d$ describes the set of indices of the discrete function $\mathbf{h}_l$. Hence, the h_l are represented in a basis made of shifted versions of φ that lie on a grid that is r-times finer than the acquisition grid. For the generating function φ, we consider a *cubic convolution* kernel [Key81]. The motivations behind this choice are twofold. First it allows for fast computations and leads to an easy interpretation of the coefficients $\mathbf{h}_l$: contrarily to cubic splines, the value $\mathbf{h}_l[\mathbf{k}]$ directly encodes the value of the continuous function $h_l(\mathbf{\Delta} \odot \mathbf{k}/r)$. In addition, despite being slightly less accurate than cubic splines interpolation, it simplifies the numerical analysis significantly since no linear systems have to be solved while still possessing good approximation properties (convergence to the original function as fast as $\|\mathbf{\Delta}/r\|^3$). Note that the convolution between the function h_l and the bead b can be done efficiently using the fast Fourier transform, see Appendix 3.5.2 for the technicalities.

In the following, we let $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_{|M|}]$ denote the discrete matrix associated to U such that $U\mathbf{c} = \Phi(\mathbf{U}\mathbf{c})$ for all $\mathbf{c} \in \mathbb{R}^{|M|}$, assuming that it exists.

Forward model: The unknowns we wish to evaluate from an observation $\mathbf{y}$ are the matrix $\mathbf{\Gamma} = (\gamma_{l,j})$ that characterizes the decomposition of each impulse response observed in the image $\mathbf{y}$, the coefficients $\mathbf{C} = (\mathbf{c}_l)_l$ which describe the basis functions (h_l) within the regularization basis $\mathbf{U}$ and the micro-beads positions $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_L)$. In our algorithms, we will make heavy use of the mapping $\mathcal{M}$, which allows to synthesize an image given a triplet $(\mathbf{\Gamma}, \mathbf{X}, \mathbf{C})$. It is defined by

$$\begin{aligned} \mathcal{M}: \quad \mathbb{R}^{|\mathbf{L}| \times |\mathbf{J}|} \times \mathbb{R}^{d \times |\mathbf{J}|} \times \mathbb{R}^{|\mathbf{M}| \times |\mathbf{L}|} &\to \mathbb{R}^{|\mathbf{I}|} \\ (\mathbf{\Gamma}, \mathbf{X}, \mathbf{C}) &\mapsto \left(\sum_{j \in \mathbf{J}} \sum_{l \in \mathbf{L}} \gamma_{l,j} [\Phi(\mathbf{U}\mathbf{c}_l) \star b](\mathbf{\Delta} \odot \mathbf{i} - \mathbf{x}_j) \right)_{\mathbf{i} \in \mathbf{I}} . \end{aligned} \tag{3.3}$$

Note that this mapping is not injective since $\mathcal{M}(\alpha\mathbf{\Gamma}, \mathbf{X}, \frac{\mathbf{C}}{\alpha}) = \mathcal{M}(\mathbf{\Gamma}, \mathbf{X}, \mathbf{C})$ for any $\alpha > 0$. This phenomenon is well known in blind inverse problems and corresponds to an ambiguity between the intensity of the micro-beads and the scaling of the PSF intensity. One way to resolve this ambiguity is to normalize the vectors $\mathbf{c}_l$: $\|\mathbf{c}_l\|_p = 1$, for any $p \geq 1$. This type of normalization is taken into account by forcing $\mathbf{c}_l$ to belong to some abstract set $\mathcal{C}$. We only require that the projection onto this set is numerically feasible in reasonable time. In PSF-Estimator , we propose to use $\mathcal{C} = \mathbb{R}^{|M|}$ and $\mathcal{C} = \{\mathbf{c} \in \mathbb{R}^{|M|}, \|\mathbf{c}\|_p = 1\}$ with $p = 2$.

Assuming a Gaussian noise model on $\mathbf{n}$, a natural minimization problem to recover the triplet $(\mathbf{\Gamma}, \mathbf{X}, \mathbf{C})$ consists in solving:

$$\underset{\substack{\mathbf{x}_j \in \mathbb{R}^d, j \in \mathbf{J} \\ \mathbf{c}_l \in \mathcal{C}, l \in \mathbf{L} \\ \gamma_{l,j} \in \mathbb{R}, l \in \mathbf{L}, j \in \mathbf{J}}}{\operatorname{argmin}} \quad \frac{1}{2}\|\mathcal{M}(\mathbf{\Gamma}, \mathbf{X}, \mathbf{C}) - \mathbf{y}\|_2^2. \tag{3.4}$$

This problem is nonlinear and nonconvex. The model is bilinear in the couple $(\mathbf{\Gamma}, \mathbf{C})$ and nonlinear in the variable $\mathbf{X}$. Therefore, we cannot expect to solve it globally, and we will resort to standard local nonlinear programming techniques with a careful initialization. The alternating minimization scheme used in PSF-Estimator to solve Problem (3.4) is described in the next paragraph.

3.2.2 Initialization

The initialization procedure consists in two steps: i) we estimate the positions of the beads present on the image, ii) we estimate the parameters of the different patches to form an initial subspace $(\mathbf{h}_l)$. The full initialization procedure is summarized in Algorithm 1.

Algorithm 1 Automatic initialization

INPUT: PSF visible width D_{PSF}, observation $\mathbf{y}$, a basis $\mathbf{U}$.
OUTPUT: Initial guess $(\mathbf{\Gamma}_0, \mathbf{X}_0, \mathbf{C}_0)$ for the optimizer (see next section) (3.4) and a scaling factor for the basis $\mathbf{U}$.

```
procedure
    Estimate X using Algorithm 2
    Scale the basis U
    Estimate C by projecting patches centered in (x_j) on the span of U
    Estimate Γ by solving a J × J linear system.
end procedure
```

Localizations estimation

Given an image containing impulse responses, we aim at **automatically** estimating their positions. The only parameter essential to the following procedure is the PSF visible width D_{PSF}, that is a rough estimation of the diameter of a disc that captures most of the PSF's energy. The entire procedure is summarized in Algorithm 2.

Algorithm 2 Automatic detection of beads positions

INPUT: PSF visible width D_{PSF}, observation $\mathbf{y}$.
OUTPUT: List of J positions $\widehat{\mathbf{x}}_j$ of admissible patches.
1: **procedure**
2: Estimate the background by solving Problem (3.5).
3: Use the homemade vectorial SIFT (Scale Invariant Feature Transform) to perform a rough automatic PSF detection.
4: Select a subset of the detected PSF using the greedy maximum weighted independent set algorithm 3.
5: Reject patches too close to each other or too close to the edges.
6: Refine the position estimation.
7: **end procedure**

Background estimation: The background is not modeled in Equation (3.3), we rely on the procedure described in this paragraph to take it into account. Getting a correct estimate of the background is a critical step since not accounting for it would strongly bias the PSFs estimates and their amplitudes. The background might have different sources, the most common one being the auto-fluorescence of the sample or the cover-slip. It may vary spatially, but we assume that it varies slowly comparatively to the PSFs.

We therefore describe the background as a linear combination of low-degree polynomials and a few thin-plate functions centered on a coarse Cartesian grid. We let $\mathbf{B} \in \mathbb{R}^{|\mathbf{I}| \times P}$ denote the matrix containing the P vectorized thin-plate splines and polynomial and solve the following robust fitting variational problem:

$$\min_{\beta \in \mathbb{R}^P} \|\mathbf{B}\beta - \mathbf{y}_v\|_1 , \tag{3.5}$$

where $\mathbf{y}_v \in \mathbb{R}^{|\mathbf{I}|}$ is the vectorized version of $\mathbf{y}$. This problem is solved using an ADMM algorithm. The idea is that the ℓ^1-fitting should be able to distinguish between the slowly varying background component and the additional impulse responses which are more localized in space. In Figure 3.1, we display the result of this approach on three different microscopy images.

Rough PSF detection and scale estimate with a vectorial SIFT: In order to detect the PSFs, we designed a methodology which is similar in spirit to the SIFT (Scale Invariant Feature Transform) detector [Low04], which provides a scale invariant feature extractor.

An input of the algorithm is a discrete basis $\mathbf{U}$, which allows to efficiently encode the PSFs. This basis can be thought of as the principal components sorted by importance of a vast family of realistic impulse responses observed offline. With this tool at hand, a natural approach to detect the PSFs is to find the maxima of correlation with the first components.

Unfortunately, this approach proved to be insufficient to tackle all the examples we came across *automatically*. The PSF width (e.g. the FWHM) is approximately given by the parameter D_{PSF} set by the user. However, this is only meant to represent a rough estimate. In addition, the PSFs may vary spatially and have different widths across the field of view. Hence, there is a necessity to find a spatial scaling of the dictionary $\mathbf{U}$. In order to detect the

possible beads, we propose to compute the energy contained in ℓ^2-normalized patches of the image on the first elements of the family $\mathbf{U}$ (we take 5 elements by default), with different spatial scales. This can be achieved efficiently with Fast Fourier Transforms and outputs a set of maps (a 3D cube in 2D and a 4D cube in 3D). We then localize the local maxima of this map and only keep the ones with values exceeding a certain threshold which depends on the number of dictionary elements used. This approach returns a rough estimate of possible bead positions. An example is given in Figure 3.2a.

Aggregation of maxima: The previous step returns too many maxima, out of which we need to find a good sub-selection. We propose to aggregate them by calculating the *maximum weighted independent set of the graph* formed by these points.

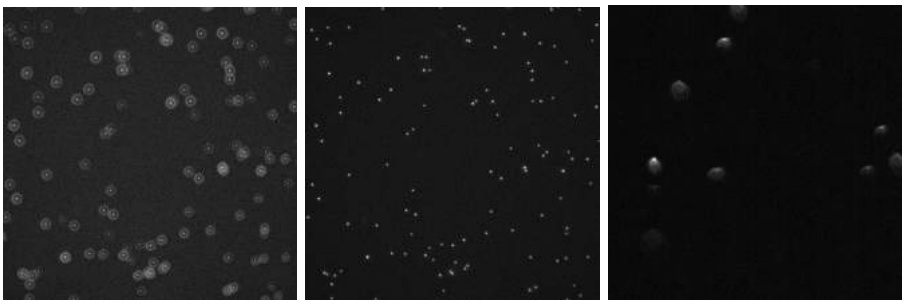
More formally, the previous step outputs a set of $|S'|$ local maximum, at position $(x_s)_{s\in S'}$ and with amplitude $(w_s)_{s\in S'}$. We aim at finding a subset $S \subset S'$ of 'good' maxima (e.g. having one point per cluster). This problem is well known in graph theory and is in general NP-hard. We propose to solve it approximately using a greedy implementation proposed in Algorithm 3. We display an experimental result in Figure 3.2.

Algorithm 3 Greedy maximum weighted independent set

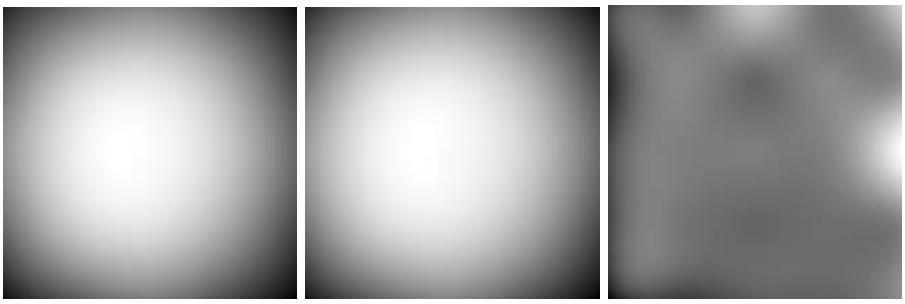
INPUT: PSF visible width D_{PSF}, the set of local maxima S', their amplitude $(w_s)_{s\in S'}$ and their position $(x_s)_{s\in S'}$.
OUTPUT: Set of admissible maxima $S \subset S'$.
Initialization:
$S = \emptyset$, $\tilde{S} = S'$.
procedure
  while $\tilde{S}$ non-empty **do**
    Find $s_{max} \in \operatorname{argmax}_{s\in\tilde{S}} w_s$.
    Add s_{max} to S.
    Set $\tilde{S} = \tilde{S} \setminus \{s \in \tilde{S} \text{ such that } \|x_s - x_{s_{max}}\|_2 \leq D_{PSF}\}$
  end while
end procedure

This original procedure produces remarkably stable and accurate results with a single - easy to tune - parameter.

(a) Original noisy images. Contrast have been stretched for better visualization.

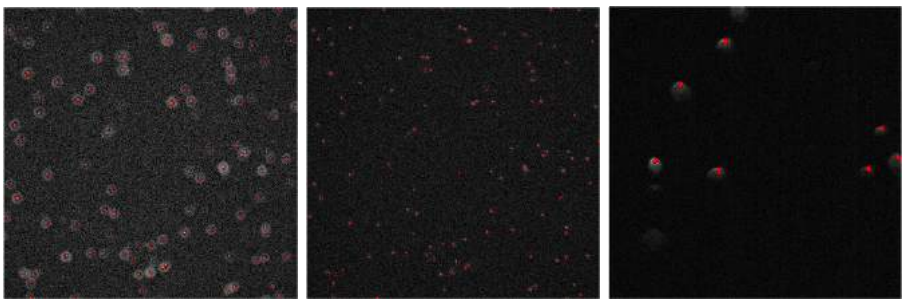


(b) Images with estimated background removed. Contrast have been stretched for better visualization.

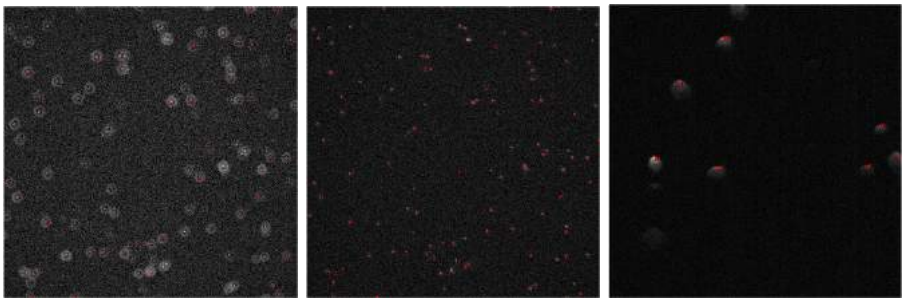


(c) Estimated background.

Figure 3.1: Background extraction procedure using an ℓ^1 fitting algorithm. We illustrate the behavior of this procedure on three microscopy images (first row). From left to right: 1024×1024 image of micro-beads (pixel size of 140nm and beads of size 100nm) acquired using a wide-field microscope ($NA = 1.49$) at 400nm of the focal plane, same image but acquired at the focal plane, and image of micro-beads acquired using a SPIM microscope ($NA = 0.8$) (pixel size of 108nm and beads of size 100nm).



(a) Detection using threshold $\tau = 0.6$ with matrix $\mathbf{V}$.



(b) Detection outputs by the greedy maximum weighted independent set procedure.

Figure 3.2: Illustration of the PSF detection procedure. The maxima are detected in the images in Figure 3.1.

Rejection of bad patches: The analytical expression of a PSF yields functions that are not compactly supported. However, they decay quite fast at infinity (an Airy pattern decays as $1/|\mathbf{x}|^3$) and the information brought by the image is dominated by noise far away from the PSF center. To avoid patches containing more than one PSF or PSFs too close to the image boundaries, we select the subset of PSFs with a center being at least $D_{PSF}/2$ pixels away from the image boundary and D_{PSF} pixels from the other centers. At the end of this procedure, we obtain a set of patches containing isolated PSFs.

Shift and re-sampling: The estimation of the bead center might be inaccurate by the previous estimation. In addition, there is no reason for the micro-beads to be perfectly centered on a pixel. To obtain a better PSF centering, we find the maximums of correlation with a continuous Gaussian function. This allows us to re-interpolate the patches around this center using a bi-cubic interpolation.

The caveats of our approach when dealing with non-centered PSFs, is that we will find shifted PSFs. This can result in slight image deformations when applying deblurring methods for instance. We believe that this is not so important for image interpretation, but this might become an issue for more advanced problems which require registration such as image colocalization.

Unfortunately, there is an intrinsic ambiguity related to the problem of PSF estimation: we obtain an equivalent result by shifting the PSF and the underlying object by an opposite vector. To avoid this ambiguity, it is possible to add an

assumption such as: "the center of mass of the PSF is located at the origin". This hypothesis is implicit when finding the maximums of correlation with a sufficiently wide Gaussian, as explain in Appendix 3.5.3.

PSF coefficients, weights and scaling the basis

Algorithm 2 outputs a list of admissible PSFs. It remains to compute their decomposition $\mathbf{C}$ into the basis given by $\mathbf{U}$, and to determine their weights.

Let $(\widehat{\mathbf{x}}_j)_{1\leq j\leq J}$ denote the J positions of the beads given by Algorithm 2, and let $\boldsymbol{\omega}_j$ denote the patch of the observation $\mathbf{y}$ of size D_{PSF} in each direction and centered in $\widehat{\mathbf{x}}_j$.

Local background removal The background estimated in the first procedure only provides a rough estimate. In practice, we have noticed that an accurate estimation of the background on each patch significantly improves the quality of the reconstruction method.

We assume that the PSF is dominated by noise in a domain χ made of pixels outside a disk of radius D_{PSF} from the patch center in $\widehat{\mathbf{x}}$. In this region, the image is therefore considered to be dominated by background and noise only. Assuming that the background is a smooth component, we fit a low degree polynomial to these pixels. This amounts to solving the following quadratic problem:

$$\inf_{\boldsymbol{\lambda}\in\mathbb{R}^P} \frac{1}{2}\|\boldsymbol{\omega} - \mathbf{M}\boldsymbol{\lambda}\|^2_{\ell^2(\chi)}, \tag{3.6}$$

where $\mathbf{M} = [\mathbf{m}_1, \ldots, \mathbf{m}_p] \in \mathbb{R}^{|\boldsymbol{\omega}_j|\times P}$ is a matrix containing the sampled low degree monomials $\mathbf{m}_p$ and $\boldsymbol{\lambda}$ represents the coefficients of the polynomial. This problem boils down to a low dimensional linear system which can be solved with a linear conjugate gradient algorithm. Letting $\boldsymbol{\lambda}^\star$ denote the solution of this problem, the estimated background is simply $\mathbf{M}\boldsymbol{\lambda}^\star$. In practice, we simply use polynomials of order 2. In 2D, this yields the value $P = 6$ for the monomials 1, x, y, xy, x^2, y^2. We let $\bar{\boldsymbol{\omega}}_j$ denote the j-th patch where the estimated local background has been removed.

In figure 3.3, we display results obtain with this procedure on small patches.

Scaling the dictionary As mentioned in the previous paragraph, if the family $\mathbf{U}$ is a dictionary of PSFs, the spatial scale must be adapted to the observation. In PSF-Estimator , we propose to determine the scale that makes the family $\mathbf{U}$ best suited to the set of patches $(\bar{\boldsymbol{\omega}}_j)_j$. To this end, we simply select the scaling parameter that minimizes the error between the set of patches and their projection on the rescaled family.

PSF coefficients initialization Finally, for all background free patches $(\bar{\boldsymbol{\omega}}_j)_j$, the initial coefficient $(c_j)_j$ are given by the projection of the patch onto the rescaled family $\mathbf{U}$. In the case where $\mathcal{C}$ is a non-trivial set, the initial coefficients are projected on it.

Weights initialization The initial weights $(\gamma_j)_j$ can simply be estimated by solving a $J \times J$ linear system. Given $(\mathbf{X}_0, \mathbf{C}_0) \in \mathbb{R}^{d\times|\mathbf{J}|} \times \mathbb{R}^{|\mathbf{M}|\times|\mathbf{L}|}$, we aim to

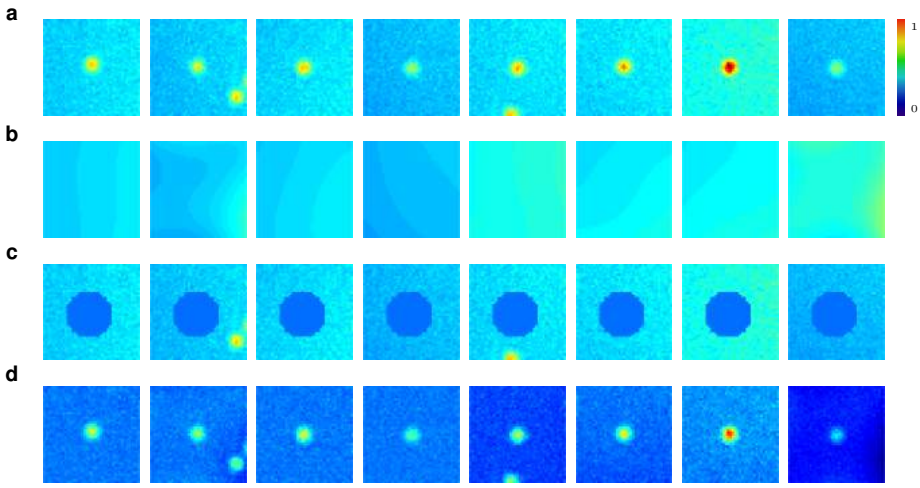


Figure 3.3: **Local estimation of the background.** **a**, Original PSFs degraded by Gaussian random noise and smooth background. **b**, Estimated background by polynomials of order 3. **c,** Image used to estimate the background (outside the disk). **d,** PSFs when estimated background is removed. Remark that the procedure is robust to other PSFs in the area used to estimate the background.

solve

$$\min_{\gamma_{l,j}, l\in\mathbf{L}, j\in\mathbf{J}} \frac{1}{2}\|\mathcal{M}(\mathbf{\Gamma}, \mathbf{X}_0, \mathbf{C}_0) - \mathbf{y}\|_2^2.$$

This corresponds to Problem (3.8) in the following.

3.2.3 Estimation algorithm

Problem (3.4) is convex in each of its variables, but non-convex in the triplet $(\mathbf{\Gamma}, \mathbf{X}, \mathbf{C})$. Fortunately, the previous procedure turns out to provide a robust and accurate automatic initialization of $(\mathbf{\Gamma}, \mathbf{X}, \mathbf{C})$. This initialization seems to lead to a basin of attraction of a satisfying minimum of the energy (3.4) when using an alternating optimization scheme. The entire procedure is summarized in Algorithm 4 and coined PSF-Estimator. Each sub-problem is detailed below:

- Given $(\mathbf{\Gamma}, \mathbf{X}) \in \mathbb{R}^{|\mathbf{L}|\times|\mathbf{J}|} \times \mathbb{R}^{d\times|\mathbf{J}|}$, solve

$$\min_{\mathbf{c}_l\in\mathcal{C}, l\in\mathbf{L}} \frac{1}{2}\|\mathcal{M}(\mathbf{\Gamma}, \mathbf{X}, \mathbf{C}) - \mathbf{y}\|_2^2. \tag{3.7}$$

  In the case where the constraint set $\mathcal{C} = \mathbb{R}^{|M|}$, i.e. no normalization is enforced, this amounts to solving a linear system, which is numerically efficient using a conjugate gradient algorithm. In the case where the constraint set $\mathcal{C}$ is more general, we use a projected gradient descent presented in Algorithm 5. We use a backtracking line search, i.e. we aim at finding the steepest descent direction for the k-th iterate by solving:

$$\tau_{\mathbf{C}} = \operatorname*{argmin}_{\tau\in\mathbb{R}} \frac{1}{2}\|\mathcal{M}(\mathbf{\Gamma}, \mathbf{X}, \mathbf{C}_k - \tau\nabla_{\mathbf{C}}\mathcal{M}(\mathbf{\Gamma}_k, \mathbf{X}_k, \mathbf{C}_k)) - \mathbf{y}\|_2^2,$$

Algorithm 4 PSF-Estimator

INPUT: PSF visible width D_{PSF}
Initialization:
$k = 1$,
Compute $(\mathbf{\Gamma}_k, \mathbf{X}_k, \mathbf{C}_k)$ using Algorithm 1.
procedure
  while cost function decreases **do**
    Given $(\mathbf{\Gamma}_k, \mathbf{X}_k)$, compute $\mathbf{C}_{k+1}$ by solving Problem (3.7),
    Given $(\mathbf{X}_k, \mathbf{C}_{k+1})$, compute $\mathbf{\Gamma}_{k+1}$ by solving Problem (3.8),
    Given $(\mathbf{\Gamma}_{k+1}, \mathbf{C}_{k+1})$, compute $\mathbf{X}_{k+1}$ by solving Problem (3.9),
    $k = k + 1$,
  end while
end procedure

with solution given by

$$\tau_{\mathbf{C}} = \frac{\langle \mathcal{M}(\mathbf{\Gamma}, \mathbf{X}, \nabla_{\mathbf{C}} \mathcal{M}(\mathbf{\Gamma}, \mathbf{X}, \mathbf{C}_k)), \mathcal{M}(\mathbf{\Gamma}, \mathbf{X}, \mathbf{C}_k) - y \rangle}{\| \mathcal{M}(\mathbf{\Gamma}, \mathbf{X}, \nabla_{\mathbf{C}} \mathcal{M}(\mathbf{\Gamma}, \mathbf{X}, \mathbf{C}_k)) \|_2^2},$$

assuming the denominator is non-zero.

- Given $(\mathbf{X}, \mathbf{C}) \in \mathbb{R}^{d \times |\mathbf{J}|} \times \mathbb{R}^{|\mathbf{M}| \times |\mathbf{L}|}$, solve

$$\min_{\gamma_{l,j}, l \in \mathbf{L}, j \in \mathbf{J}} \frac{1}{2} \| \mathcal{M}(\mathbf{\Gamma}, \mathbf{X}, \mathbf{C}) - \mathbf{y} \|_2^2. \tag{3.8}$$

  The operator $\mathcal{M}$ being linear in the input $\mathbf{\Gamma}$, this amounts to solve a linear system, which is done using conjugate gradient algorithm.

- Given $(\mathbf{\Gamma}, \mathbf{C}) \in \mathbb{R}^{|\mathbf{L}| \times |\mathbf{J}|} \times \mathbb{R}^{|\mathbf{M}| \times |\mathbf{L}|}$, solve

$$\min_{\mathbf{x}_j \in \mathbb{R}^d, j \in \mathbf{J}} \frac{1}{2} \| \mathcal{M}(\mathbf{\Gamma}, \mathbf{X}, \mathbf{C}) - \mathbf{y} \|_2^2. \tag{3.9}$$

  We use a gradient descent presented in Algorithm 6 with backtracking line search. The descent step size is computed similarly to Algorithm 5.

Algorithm 5 Solving Problem (3.7) with projected gradient descent with a backtracking line search

INPUT: $(\mathbf{\Gamma}, \mathbf{X})$, number of iteration $nit_{\mathbf{C}}$.
OUTPUT: approximated solution $\mathbf{C}_{nit_{\mathbf{C}}+1}$ of Problem (3.7).

1: **procedure**
2: **for** $k = \{1, \cdots, nit_{\mathbf{C}}\}$ **do**
3: Line search:

$$\tau_{\mathbf{C}} = \frac{\langle \mathcal{M}(\mathbf{\Gamma}, \mathbf{X}, \nabla_{\mathbf{C}}\mathcal{M}(\mathbf{\Gamma}, \mathbf{X}, \mathbf{C}_k)), \mathcal{M}(\mathbf{\Gamma}, \mathbf{X}, \mathbf{C}_k) - y\rangle}{\|\mathcal{M}(\mathbf{\Gamma}, \mathbf{X}, \nabla_{\mathbf{C}}\mathcal{M}(\mathbf{\Gamma}, \mathbf{X}, \mathbf{C}_k))\|_2^2},$$

4: Descent step: $\widehat{\mathbf{C}} = \mathbf{C}_k - \tau_{\mathbf{C}}\nabla_{\mathbf{C}}\mathcal{M}(\mathbf{\Gamma}, \mathbf{X}, \mathbf{C}_k)$
5: Projection: $\mathbf{C}_{k+1} = \Pi_{\mathcal{C}}\left(\widehat{\mathbf{C}}\right)$
6: $k = k + 1$
7: **end for**
8: **end procedure**

Algorithm 6 Solving Problem (3.9) with gradient descent with a backtracking line search

INPUT: $(\mathbf{\Gamma}, \mathbf{C})$, number of iteration $nit_{\mathbf{X}}$.
OUTPUT: approximated solution $\mathbf{X}_{nit_{\mathbf{X}}+1}$ of Problem (3.9).

1: **procedure**
2: **for** $k = \{1, \cdots, nit_{\mathbf{X}}\}$ **do**
3: Line search:

$$\tau_{\mathbf{X}} = \frac{\langle \mathcal{M}(\mathbf{\Gamma}, \nabla_{\mathbf{X}}\mathcal{M}(\mathbf{\Gamma}, \mathbf{X}_k, \mathbf{C}), \mathbf{C}), \mathcal{M}(\mathbf{\Gamma}, \mathbf{X}_k, \mathbf{C}) - y\rangle}{\|\mathcal{M}(\mathbf{\Gamma}, \nabla_{\mathbf{X}}\mathcal{M}(\mathbf{\Gamma}, \mathbf{X}_k, \mathbf{C}), \mathbf{C})\|_2^2},$$

4: Descent step: $\mathbf{X}_{k+1} = \mathbf{X}_k - \tau_{\mathbf{X}}\nabla_{\mathbf{X}}\mathcal{M}(\mathbf{\Gamma}, \mathbf{X}_k, \mathbf{C})$
5: $k = k + 1$
6: **end for**
7: **end procedure**

3.3 Numerical results

The methodology described in the previous sections and summarized in Algorithm 4 is based on many non-trivial elements. In this part, we first try to highlight the most important features using simulated data. In a second part, we show the merits of the method on microscopy data.

3.3.1 Simulation data and dictionary learning

We simulate realistic 3D PSFs using the scalar theory of diffraction introduced in Chapter 2. For the sake of readability, we briefly recall the main equations of this model.

The PSF of a point source placed at position (x_0, y_0, z_0) in the field of view is given by [Goo05; PSM17]:

$$K(x', y'|x_0, y_0, z_0) \propto \left|\mathcal{F}\left(E(\ \cdot\ |x_0, y_0, z_0)\right)(x', y')\right|^2, \tag{3.10}$$

where $\mathcal{F}$ is the Fourier transform, and $E(\,\cdot\,|x_0, y_0, z_0)$ is the electric field at position (x_0, y_0, z_0). The latter is given by

$$E(x, y) \propto \rho(x, y) \exp\left(i\Phi(x, y)\right) \exp\left(i\Phi_{ax}(x, y; z_0)\right),$$

where i is a complex number such that $i^2 = -1$, and Φ is the phase at the Fourier plane. The function ρ is defined by:

$$\rho(\mathbf{x}) = \begin{cases} 1 & \text{if } \|\mathbf{x}\|_2 \leq \frac{\text{NA}}{\lambda}, \\ 0 & \text{otherwise,} \end{cases}$$

where NA is the numerical aperture of the camera and λ is the wavelength. The function Φ_{ax} is given by:

$$\Phi_{ax}(x, y; z_0) = 2\pi z_0 \sqrt{\left(\frac{n_I}{\lambda}\right)^2 - (x^2 + y^2)},$$

where n_I is the refractive index of the medium containing the sample. This function takes into account that the bead is not necessarily placed in the focal plane ($z = 0$).

Unlike the equations presented in Chapter 2, we do not use the function Φ_{lat}. In this chapter, the lateral displacement is handle by the estimated locations Xd.

The phase term Φ is expressed using Zernike polynomials:

$$\Phi(x, y) = \sum_{i \in \mathcal{I}} c_i Z_i(x, y),$$

where Z_i is the i-th polynomials (using Noll indexes [Nol76]). To generate the observations, we use $\mathcal{I} = \{5, 6, 7, 8\}$, which corresponds to deformations name respectively oblique and vertical astigmatism, and vertical and horizontal coma. We also use $n_I = 1.57$, NA $= 1.4$ and $\lambda = 600$nm.

Dictionary learning In all the numerical simulations, we use the PSF adapted dictionary presented in Appendix 3.5.1. We let vary the numerical aperture in $\{0.8, 1.4\}$, the refractive index in $\{1.51, 1.57\}$, the wavelength (in nanometers) in $\{560, 680, 800\}$, the value of the Zernike polynomials (indexes $\{5, 6, 7, 8\}$) in $\{-0.4, 0.4\}$. The dictionary is estimated by keeping eigenvectors of this family that allows to preserve more than 99% of the ℓ_2 energy. This experiment is presented in Figure 3.10.

Simulated image Finally, the observed image v is given by the following relation:

$$v(x', y') = \int K(x', y'|x_0, y_0, z_0) u(x_0, y_0, z_0) \mathrm{d}x_0 \mathrm{d}y_0 \mathrm{d}z_0.$$

In our simulations, the original image u is made of J fluorescent beads at position $\{\mathbf{x}_j \in \mathbb{R}^d\}_{j=1}^{|J|}$, with radius $R_b > 0$, amplitude α_j. The intensity profile of the bead is given by $b \in L_2(\mathbb{R}^d)$. We have

$$u = \sum_{j=1}^{|J|} \alpha_j b(\,\cdot\, - \mathbf{x}_j).$$

We choose $b(x) = \begin{cases} 1 & \text{if } \|x\|_2 \leq R_b \\ 0 & \text{otherwise} \end{cases}$. We assume that R_b is the same for each bead.

To generate a realistic observation, we add a background estimated from one real image of micro-beads with a wide-field microscope. Each fluorescent bead has an amplitude α_j that is proportional to the background intensity, and to the volume of the bead of radius R_b, leading to more intense observation for large radii. Finally, the resulting image is degraded by Poisson noise and additional Gaussian noise. Because the intensity of the fluorescent beads are proportional to their size, experiments with smaller beads leads to noisier images.

Remark 3.3.1. *We are aware that scalar diffraction theory is not necessarily the correct model to efficiently reproduce all optical systems (e.g. large numerical aperture). However, it is relatively simple to implement numerically, unlike the vectorial model of diffraction theory and allows to generate reasonable PSFs.*

Taking into account the bead size and estimating the position

In the first experiment, we illustrate the importance of considering the bead size and to estimate the bead positions jointly with the PSF shape. We generate observations using an invariant Airy PSF, with bead size varying from 30nm to 390nm between the different experiments (the pixel size of the observation is 100nm). We display the full width at half maximum (FWHM) on the estimated PSFs in Figure 3.4 b). We compare the estimated PSFs using four different methods:

- "No. size + pos. esti.", we use Algorithm 4 but given the smallest bead size (30 nm). We highlight the damage due to considering the bead smaller than it really is. The FWHM is poorly recovered as soon as the true bead size increase. The estimated PSF tends to be larger than the true one.

- "size + pos. init.", we use Algorithm 4 with the exact bead size, but without updating the positions, i.e. using the initial one. We highlight the benefit of jointly estimating the PSF shape and refining the positions of the beads. Not refining the PSFs localization tends to create degenerate shapes whatever the bead size.

- "size + true pos.", we use Algorithm 4 with the exact bead size, but starting with the true positions, and thus without updating them. This shows the best result that one can hope when using this approach. The FWHM is exactly retrieved, and the estimated PSF suffers from deformation only in the noisier regime where the beads size is 30nm, and thus, small amount of photon is available.

- "PSF-Estimator ", we use Algorithm 4 with the exact bead size, without adjustment on the method. As soon as the signal to noise ratio increases, the estimated PSF becomes relevant, and the FWHM is well retrieved.

The results are reported in Figure 3.4.

The full width at half maximum of the true PSF is 9 pixels. In Figure 3.4, we see that the proposed approach is able to produce a fine estimation of the PSF,

even when the underlying bead is large. Not surprisingly, when the true bead is small, not taking it into account doesn't produce a too large error. However, for larger beads, e.g. 510nm, the estimated full width at half maximum becomes larger. Similar effects also appear when the positions of the beads are frozen after initialization. This experiment shows that taking into account the bead size precisely and refining the estimated locations is particularly important for large beads.

Influence of the number of observed PSFs

In Figure 3.5, we propose to explore the influence of the number of point-sources on the quality of the estimate. The results suggest that only a small number of beads (< 5) is sufficient to obtain satisfactory results, and that using more beads tends to decrease the error only marginally.

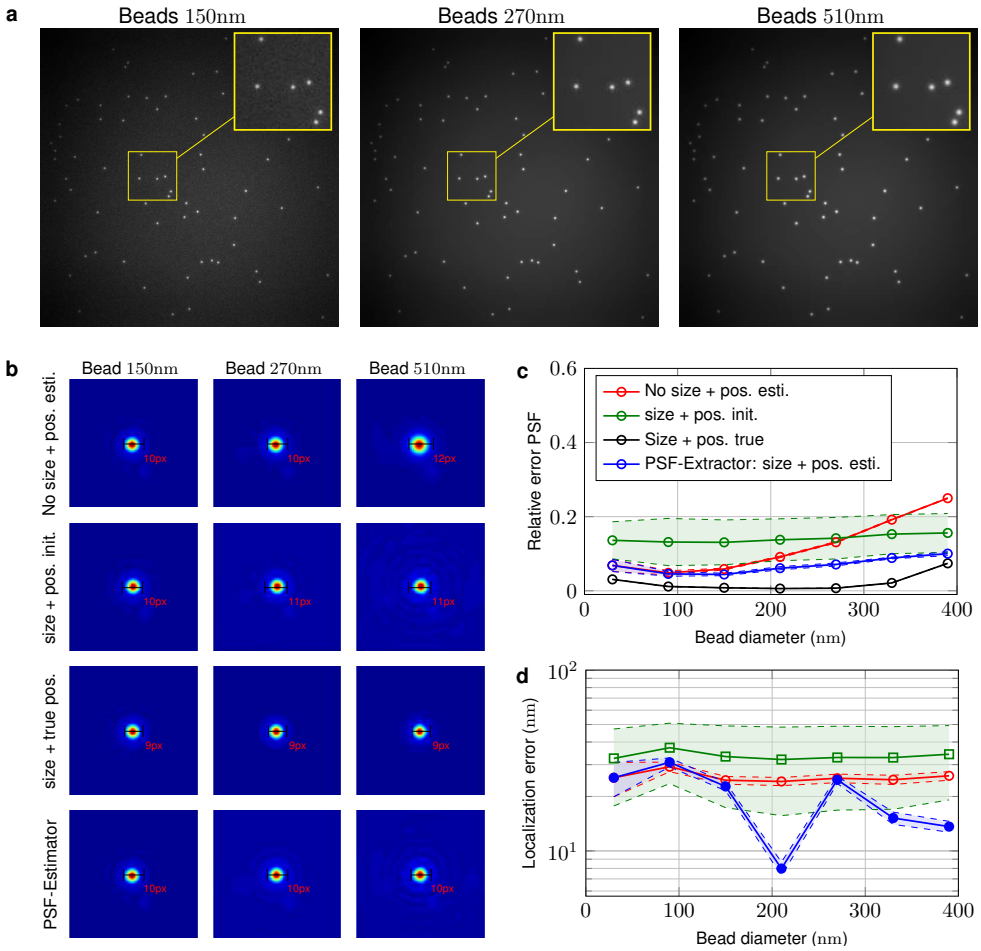


Figure 3.4: **Taking the size of the beads into account.** **a**, Simulated images of size 512×512 pixels, with pixel size of 100nm. The PSF is an Airy shape with pixel size of PSF 30nm. The added background has been estimated on a real wide-field microscope. The size of the beads is varying from 30 to 390nm. **b**, PSF reconstructed with different methods and for different beads sizes. *No size+pos. esti.* is the full method given that the bead size is 30nm, *size + pos. init.* if the optimization method where the positions of the beads are frozen after initialization and the true size of the bead is used, *size + pos. true.* if the optimization method where the positions of the beads are frozen and taken as the true one and the true size of the bead is used, and *PSF-Estimator* is the full method with the true size of the bead and where the positions of the beads are estimated. **c**, Relative error between the estimated PSF and the true PSF according to the bead diameter. **d**, Localization error between the different method for different bead diameter. recall that one pixel in the observation is 100nm.

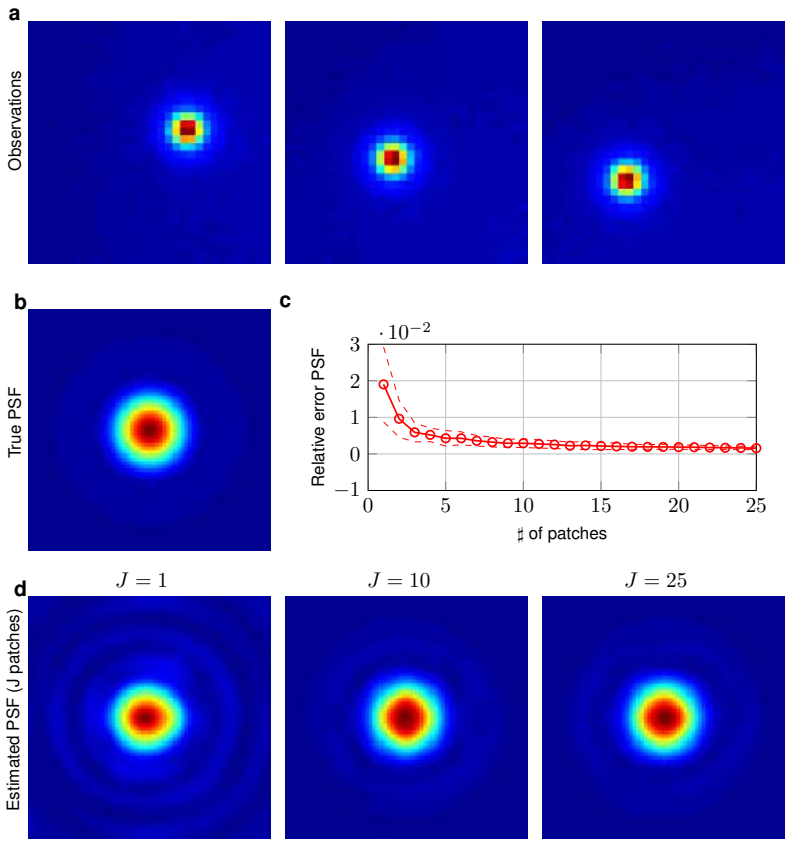


Figure 3.5: **Influence of the number of observed PSFs. a**, Simulated images of size 128×128 pixels containing a single bead, with pixel size of 100nm. The PSF is an Airy shape with pixel size of PSF 30nm. The added background has been simulated using third order polynomials. The size of the beads is 270nm. **b**, True PSF. **c**, Relative error between the true PSFs and the estimation with the number of observed PSFs varying from 1 to 25. **d**, Estimated PSFs using $J = 1, 10, 25$ observed PSFs.

PSF-Estimator allows to learn microscope specific subspace of PSFs

We now use 3D models of PSFs: Airy, astigmatism, coma and double-helix. We simulate 2D images by taking the central slice of beads randomly placed in a 3D volume degraded by the 3D PSF. This results in different 2D PSF shapes on the field on view, see Figure 3.6 a) & d) and Figure 3.7 a) & d). Using PSF-Estimator on these images, we need to carefully choose the number of elementary PSFs $|L|$. If too little elements are chosen, the family might not be able to reproduce accurately every slice of the 3D PSF. We estimate this value based on the simulated PSFs. In the case of Airy PSFs, we report that a minimum of three elements allows to capture more than 99% of the energy of the 3D PSF (these eigen-elements do not necessarily lie in the subspace spanned by the dictionary). This number increases for more complex shapes, going to four for astigmatism, four for coma and six for double-helix.

In Figure 3.6 and Figure 3.7, we show the reconstruction depending on the

estimation results with several choices of parameter $|L|$. In this experiment, the size of the beads is known, and we use the procedure implemented by PSF-Estimator. Figure 3.6 reports the results for an Airy and a coma PSF, and Figure 3.7 deals with astigmatism and double-helix PSF shapes. Let us emphasize that the double-helix PSFs are not part of the family used to generate the regularizing dictionary. Yet, the later is able to accurately reproduce them.

The results are similar in all situations. When the number of elements in the family is $L = 1$, the different slices of the 3D shape of the PSF cannot be represented simultaneously. An averaged shape is then estimated. When the number of elements increases, the different slices can be expressed in this family. In the experiments of Figures 3.6 and 3.7, we obtain satisfactory visual results starting from $L = 3$ elements.

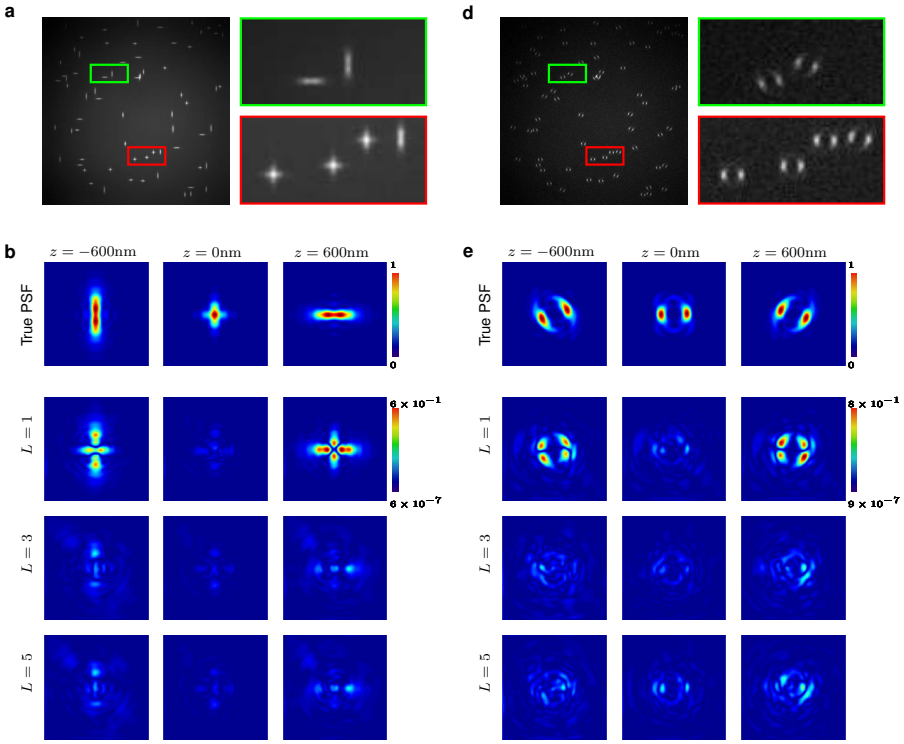


Figure 3.6: **Estimating a 3D PSF. a & c**, Simulated observation, micro-beads are located at random on a 3D volume. On panel **a**, the PSF is astigmatism, on panel **c**, the PSF is double-helix. Astigmatism: four elements allow to capture more than 99% of the energy. Double-helix: six elements are enough to capture 99% of the energy of the 3D PSF. **b and d**, validation of the estimations. The true PSF for three different z positions is displayed. We compute the pixel-wise relative error (divided by the maximum value of the true PSF) for different number of basis elements L. We use the jet colormap.

PSF-Estimator allows to estimate spatially varying PSFs

Spatial variations of the PSF can be of several natures. For 2D acquisitions, this may be due to the observation of different slices of a 3D PSF (see previous

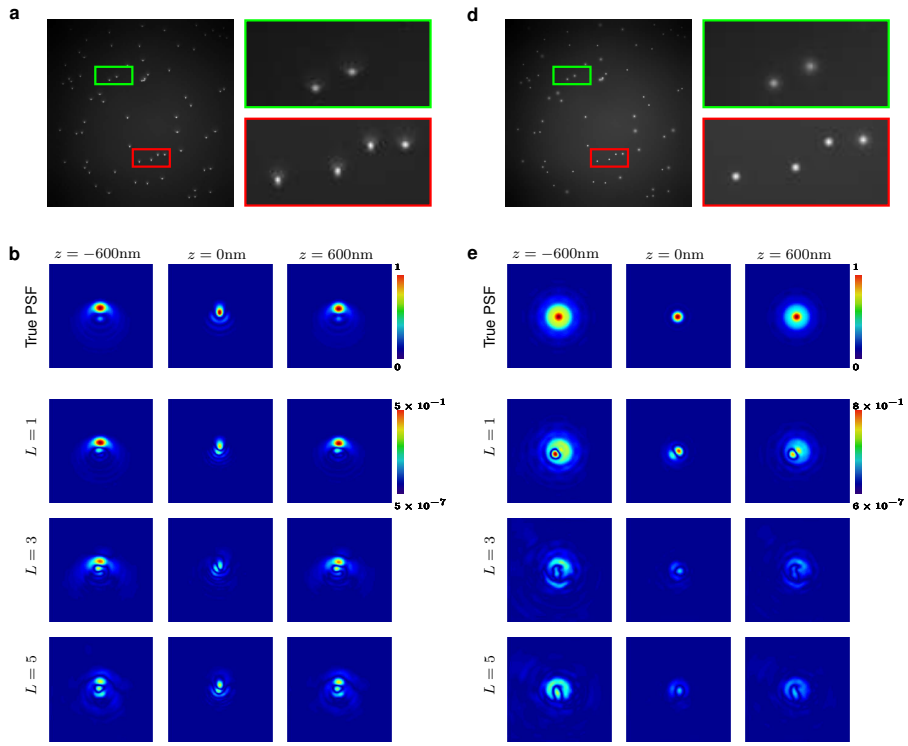


Figure 3.7: **Estimating a 3D PSF. a & c**, Simulated observation, micro-beads are located at random on a 3D volume. On panel **a**, the PSF is coma, on panel **c**, the PSF is Airy. Coma: four elements allow to capture more than 99% of the energy. Airy: three elements are enough to capture 99% of the energy of the 3D PSF. **b and d**, validation of the estimations. The true PSF for three different z positions is displayed. We compute the pixel-wise relative error (divided by the maximum value of the true PSF) for different number of basis elements L. We use the jet colormap.

experiments), or it may be due to a 2D PSF that varies in the field (see Chapter 2 for the reasons leading to this phenomenon).

In the latter case, it is often reasonable to assume that the variations of the PSF are smooth in the field of view. In this case, interpolation techniques can be used to extend the local PSF information, obtained from the beads, over the entire field of view. The mathematical procedure is detailed in Appendix 3.5.4.

We simulate astigmatic PSFs that smoothly vary from left to right. This experiment reproduces what happens when we observe beads placed on a coverslip not parallel to the focal plane. We randomly placed beads in the 2D field of view, and use PSF-Estimator to estimate a PSF family with $|L| = 5$ elements. We then use the procedure described in Appendix 3.5.4 to estimate the space variations. The PSFs are rather well retrieved, except on the edges where some defects appear. We believe that this is unavoidable since we don't have access to sampled PSFs at these positions, and since interpolations techniques poorly managed to extend values outside the sampled domain. Results are reported in Figure 3.8.

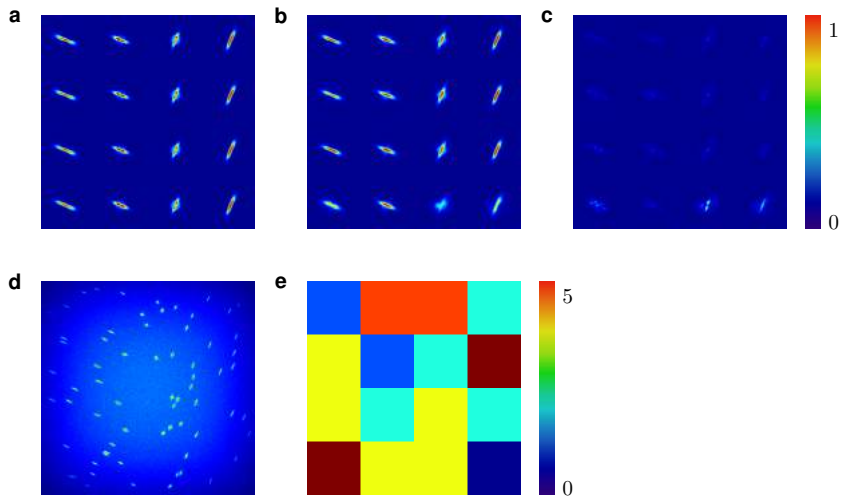


Figure 3.8: **Estimation of spatially variant PSFs. a**, Original PSFs at different positions of the field of view. **b**, Estimated PSFs using the toolbox and the interpolation procedure described in Appendix 3.5.4. **c,** Relative difference between the true PSFs **a** and the estimation **b**. **d,** Original image used to estimate the PSFs. **e,** Number of PSFs used in each area, it goes from 0 to 5.

Regularizing the PSF family

There are numerous ways to enforce desired properties on the estimated PSFs in Problem (3.4). Hereafter, we illustrate two different approaches, regularity constraint and PSF adapted dictionary, even though in all the numerical experiments we use the latter. More details on the procedure are presented in Appendix 3.5.1.

Regularity constraint: In Figure 3.9 a), we display the basis obtained using the 793 lower frequencies given by the low frequency basis elements of the 2D-discrete cosine transform. We also illustrate the capacity of this basis to reproduce standard PSFs shapes in Figure 3.9 b), using these 793 elements.

PSF adapted dictionary: In Figure 3.10 a), we learn a subspace of PSFs based on 9216 physically realistic PSFs, generated using scalar diffraction theory. The learned subspace is able to reproduce efficiently most PSFs, even double-helix that are not used to generate it. We keep 158 eigen-elements, allowing to capture more than 99% of the ℓ_2 energy of the simulated family.

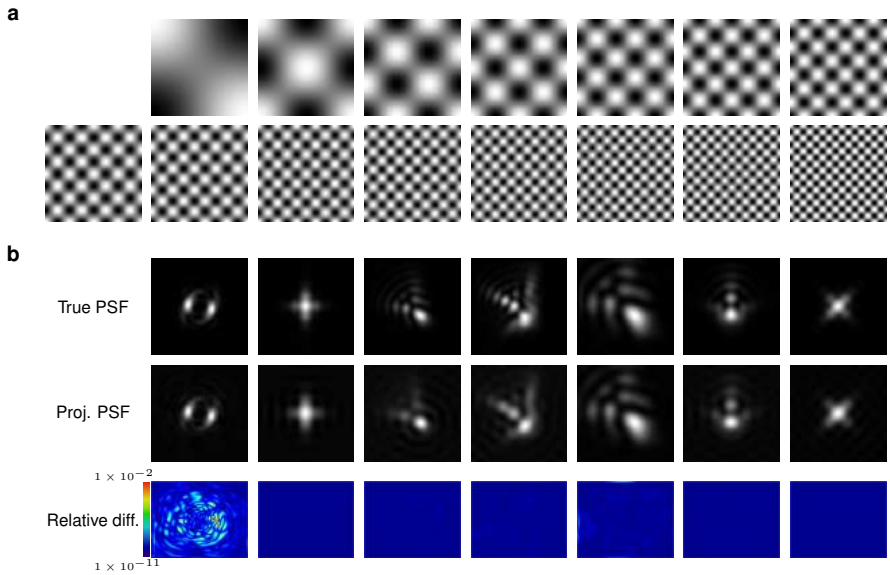


Figure 3.9: **Regularizing the PSF family using the discrete cosine transform, with 793 elements.** a) Elements of the basis corresponding to low-frequencies. b) Example of PSFs, their projection onto the basis, and the pixel-wise relative difference between them (in norm ℓ_∞).

The ability of this dictionary to reproduce simulated PSFs depends very little on the number of elements used to generate it. This is the experience of Figure 3.11. We generate a family of $100,000$ different PSFs, and calculate the error made by projecting this complete family into a dictionary where only a fraction of this data-set was used to construct it. The PSFs are generated by using a fine discretization of the values of the Zernike coefficients and the physical parameters such as the numerical aperture, the refractive index or the wavelength. The estimated dictionary is composed of enough eigen-elements to capture 95% of the energy of the PSFs used to build it (a small fraction of the $100,000$ PSFs). Remark that using a family composed of more than 3000 elements doesn't improve much the capability of the dictionary to reproduce other PSFs.

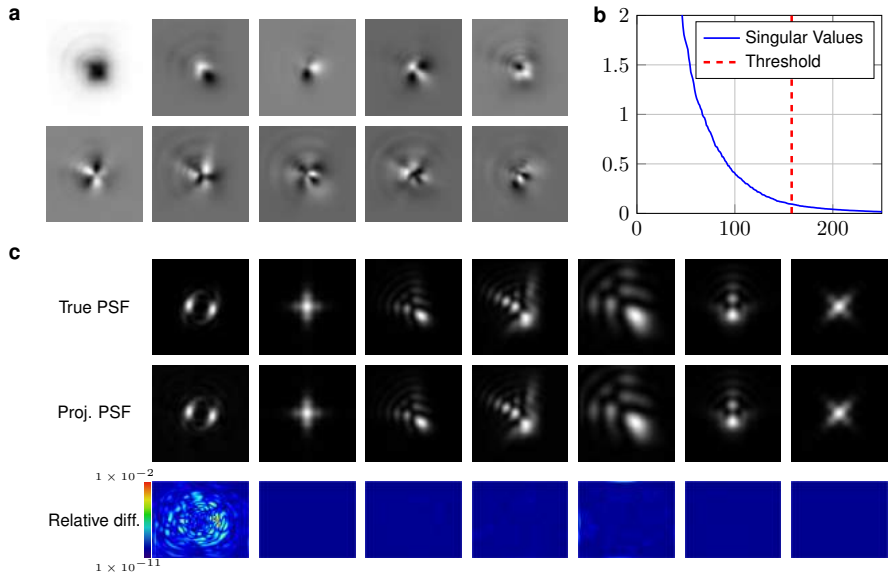


Figure 3.10: **Regularizing the PSF family using a learned dictionary.** **a**, First elements of the orthogonal basis obtained from a principal component analysis (PCA) of 9216 different PSFs, generated from three different types: Airy, coma and astigmatism, with different numerical apertures, wavelengths, and refractive indexes. Some of them are display in **c**. **b**, Rescaled singular values by decreasing order. The 158 first elements of the PCA capture 99% of the energy, meaning that in average, 158 elements are enough to reproduce any PSF with 99% accuracy. **c**, Example of PSFs, their projection onto the basis and the pixel-wise relative difference between them (in norm ℓ_∞).

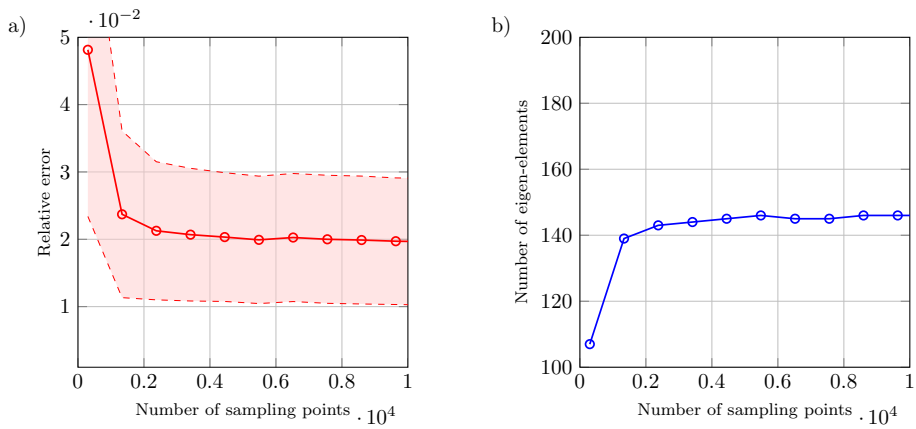


Figure 3.11: **Number of elements needed to compute the dictionary.** **a**, relative error made by projecting the 100,000 PSFs on the dictionary learned with a few portion of the full family (sampling points). **b**, Number of eigen-elements used to capture more than 95% of the ℓ_2 energy of the PSFs used for the estimation of the dictionary.

3.3.2 Microscopy data

The results presented in this section are preliminary. They aim to give a foretaste of what the toolbox will be able to do in its final version, which should hopefully be ready before the last version of this manuscript.

Multifocus wide-field microscopy

In this paragraph, we apply the toolbox PSF-Estimator to images from a wide-field microscopy produced by *Abbelight*. We used a X83 Olympus wide-field fluorescence microscope with a $100X$ 1.5NA objective. The illumination is produced by the SAFe 360 (Abbelight technology). The optical system is equipped with a deformable mirror used to produce astigmatism. The observations are captures by two cameras Orca Fusion Hamamatsu with $50ms$ exposure time, producing images of 1024×1024 pixels. We use Oxxius laser engine with excitation of 660nm, and emission of 680nm. We collected 20 stacks of a perfectly plane mono-layer of far-red micro-spheres of 40nm, with axial range of 72nm. There is no refractive index mismatch between the cover-slide and the immersion oil, allowing to avoid spherical aberrations.

We display several planes in Fig. 3.12 a).

We use PSF-Estimator to estimate a blur operator of this experiment based on one stack of images as presented in Fig. 3.12 a). We use $|L| = 3$ to allow to capture possible spatial variations of PSFs, but, as expected by the practitioners, the PSFs is mostly the same in the whole field of view. We display the estimated operator computed by interpolating the values of the coefficients of the observed PSFs. This is shown on Fig. 3.12 b). This phenomenon is confirmed by the interpolated coefficients maps displayed in Fig. 3.12 c) that are almost flat.

We now take a closer look at the beads used in PSF-Estimator, and display the observation and the estimation in Fig. 3.13. We display the orthogonal view of arbitrary 3D crops of the original image after the local background removal, and the associated estimated PSFs convolved with the bead. The background-free observed PSF should be a noisy version of the estimation. This experiment gives the opportunity to insist on the importance of background subtraction. A cheap estimation will produce a bias in the estimation, leading to a larger estimated PSF. On the contrary, an accurate background estimation allows to capture some features present in the data. For instance, we note the non-symmetry of the PSF along the optical axis (especially in the y direction). Here again, we observe that the PSF doesn't seem to vary in the field of view.

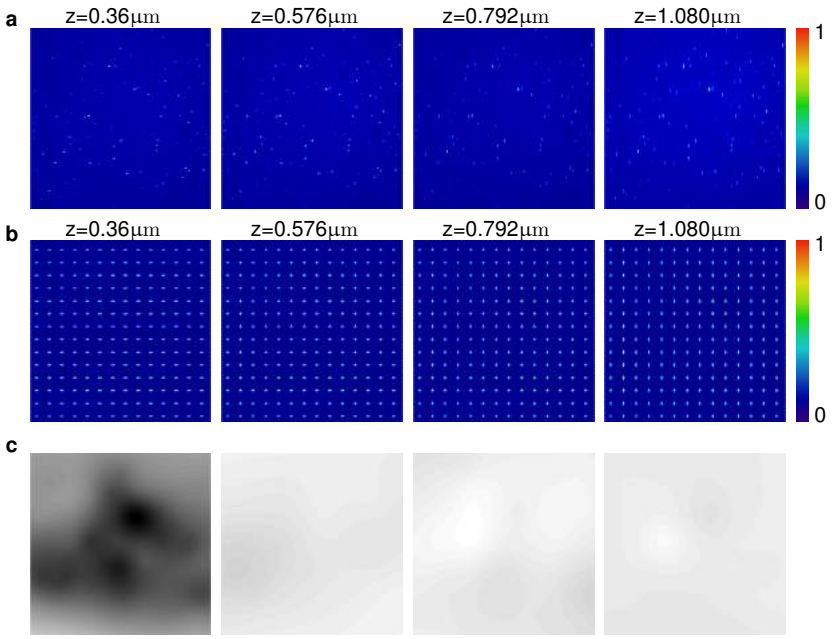


Figure 3.12: **Multifocus wide-field microscopy.** **a**, Images of same micro-beads at several positions along the optical axis. The optical system is equipped with a deformable mirror used to induce astigmatism. The contrasts have been stretched for a better visualization. **b**, Estimated operator using the toolbox PSF-Estimator. It estimates a 3D-PSF for each bead from the original image. The final operator is obtained using thin-plate interpolation to estimate the PSF at location where there is no micro-beads. **c**, Interpolation maps that determine the value of the PSF coefficients for every point of the field of view. Here, we have $|L| = 4$ elementary PSFs, and therefore L coefficients map.

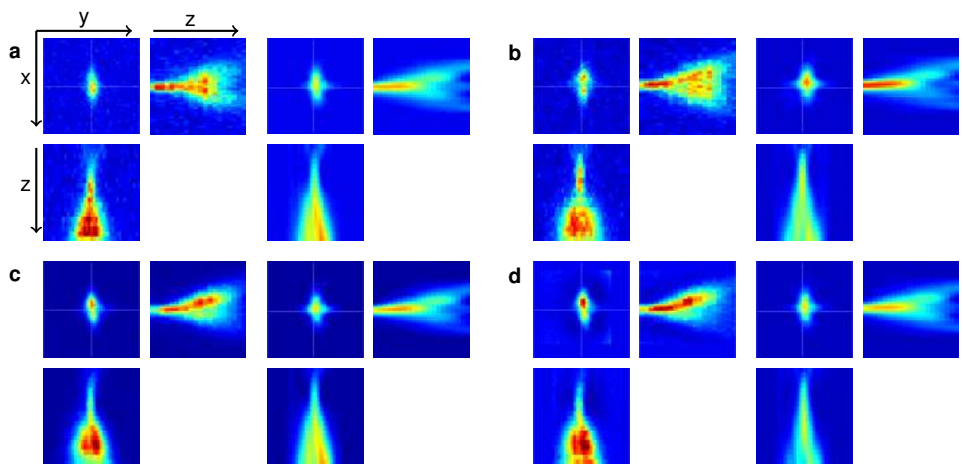


Figure 3.13: **Multifocus wide field microscopy. a, b, c,d**, Orthogonal view of arbitrary 3D PSFs. Each sub-image is composed of the (background-free) observed PSF (left), and the (noise-free, background-free) estimated PSF using PSF-Estimator (right). The estimated PSF is projected to ensure positive values. The orthogonal view allows to visualize the 3D PSF along every axis: (x,y) top-left, (x,z) top-right, and (z,y) bottom-left. The six images of each panel are displayed with the same colorbar to ease the comparisons.

3D SPIM

We now use PSF-Estimator with a 3D-SPIM. Martine Cazales realized acquisitions using the dual side illumination SPIM Z1 from Zeiss. It produces 3D images by scanning the 3D volume with two light sheets. This produces images of $1920 \times 1920 \times 90$ pixel, with pixel size of 228nm in the illumination plane, and 480nm along the axial axis. We image beads of 100nm placed in agarose gel tube. In contrast to the previous experiment, we now scan a larger domain along the optical axis, leading to PSFs fully contained in the acquired volume.

We display in Fig. 3.14 a) images of the observation at different positions along the optical axis. Similarly to the previous experiment, we display the estimated blur operator on the full field of view, see Fig. 3.14 b). Contrarily to the previous setting, less beads are available. In particular, a very small number of beads are used by PSF-Estimator in the bottom-left part of the image. This explains why the full operator doesn't seem realistic in the full field of view. Nevertheless, the spatial variation induced by this system seems rather smooth, or even non-existent, as evidenced by Fig. 3.14 c).

In Fig. 3.15, we take a closer look at the observed and estimated PSFs. We see that this experiment seems more challenging for PSF-Estimator than the previous one since the shape of the PSFs are not perfectly captured. This phenomenon is certainly due to the regularizing family $\mathbf{U}$ that is not able to reproduce perfectly the observation. We observe that this is feature key to obtain accurate results. For now we only use a very simple family in 3D constructed using Zernike polynomials. A more precise model specific to SPIM acquisition should improve the results significantly. Recall that the estimated PSFs in Fig. 3.15 are a noise-free and background-free estimation of the background-free observed PSFs.

3.4 Conclusion

We presented a robust and versatile method for estimating a PSF, or a family of PSFs, from one or more micro-bead images. This method exists currently as a set of *Matlab* and *C++* functions, but is intended to evolve quickly to a *Fiji* plugin, in order to allow an easy use in microscopy. We have shown from the various experiments that a fine estimate of the physical effects, such as the bead size, the background estimation, the joint estimation of the position and shape of the PSFs, is essential. An accurate PSF estimation tool is now essential to accurately calibrate an optical system. This tool will now serve as a starting point for the next chapters which consist in construction a subspace of blur operator for a given microscope.

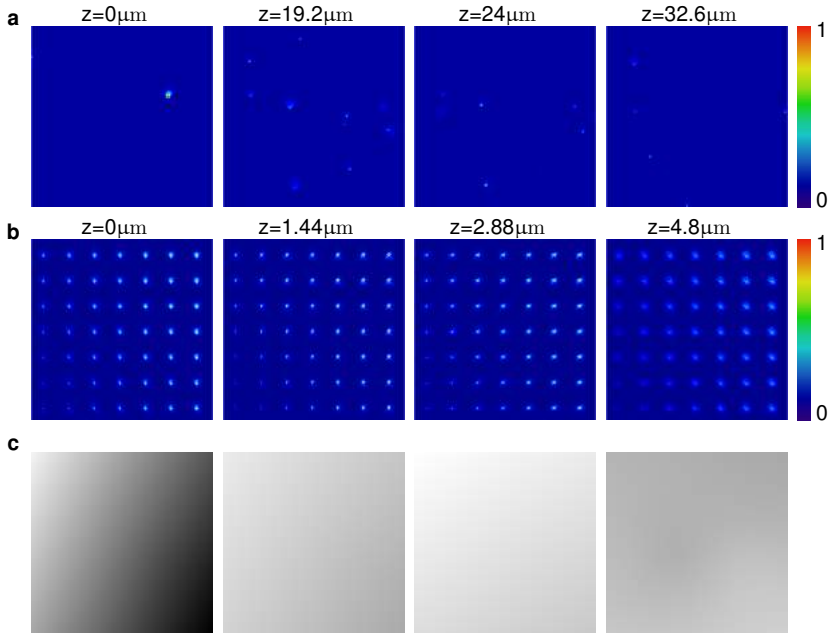


Figure 3.14: **3D SPIM.** **a**, Images of micro-beads placed into a 3D volume and observed using a 3D SPIM. The contrasts have been stretched for a better visualization. **b**, Estimated operator using the toolbox PSF-Estimator. It estimates a 3D-PSF for each bead from the original image. We finaly use thin-plate interpolation to estimate the variation of the coefficients of the PSFs on the 2D-plane orthogonal to the optical axis (we neglect variations along the optical axis). **c**, Interpolation maps that determine the value of the PSF coefficients for every point of the 2D-plane orthogonal to the optical axis. Here, we have $L = 3$ elementary PSFs, and therefore L coefficients maps.

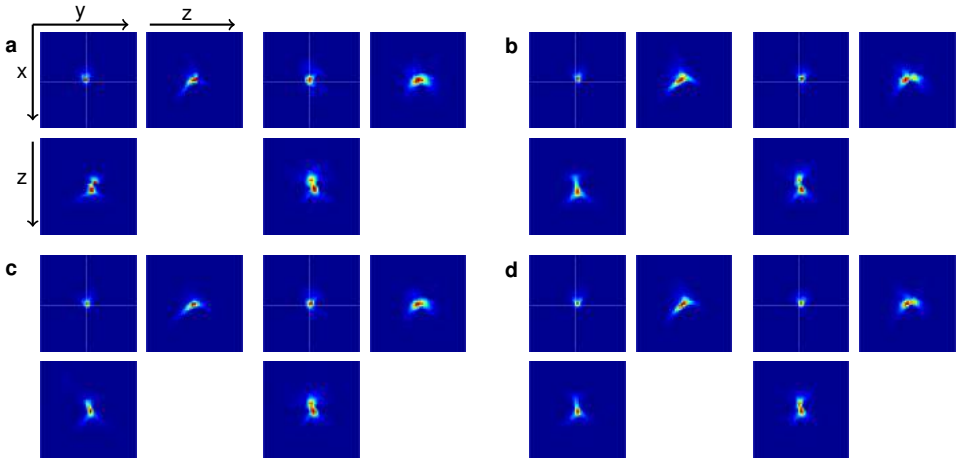


Figure 3.15: **3D SPIM. a, b, c,d**, Orthogonal view of arbitrary 3D PSFs. Each sub-image is composed of the (background-free) observed PSF (left), and the (noise-free, background-free) estimated PSF using PSF-Estimator (right). The estimated PSF is projected to ensure positive values. The orthogonal view allows to visualize the 3D PSF along every axis: (x,y) top-left, (x,z) top-right, and (z,y) bottom-left. The six images of each panel are displayed with the same colorbar to ease the comparisons.

3.5 Appendices

3.5.1 Regularizing the PSF family

The proposed estimation in Problem (3.4) is ill-posed. Hence, the choice of the basis $\mathbf{U}$ is critical. We discuss a few possibilities below.

Moment constraints

Our problem generalizes blind-deconvolution for three important reasons: i) the operator might be space varying, ii) the Dirac mass positions live off-the-grid and iii) we estimate a subspace rather than a single PSF. The blind-deconvolution problem is known as a very challenging issue. In particular there exist two ambiguities that cannot be resolved: i) the PSF can always be multiplied by a constant and the signal by its inverse and ii) the PSF can be shifted, and the signal shifted by in the opposite direction. To avoid those two issues, we can work on the affine subspace $\bar{\mathcal{H}}$ of $\mathcal{H}$ defined by

$$\bar{\mathcal{H}} = \left\{ h \in \mathcal{H}, \int_\Omega h(\mathbf{x})\, d\mathbf{x} = 1, \int_\Omega \mathbf{x} h(\mathbf{x}) d\mathbf{x} = \mathbf{0} \right\}. \tag{3.11}$$

In that case, we can define $\mathbf{U}$ as the projector on the affine subspace with $d+1$ fixed moments.

Regularity constraints

Another popular idea in the literature is to use Fourier based regularizers (e.g. the L^2-norm of the Fourier coefficients), which promote the smoothness of the solution. For a practical viewpoint, we can define $\mathbf{U}$ as the low-frequency elements of the discrete cosine transform, possibly intersected with the moment constraints above. We display such an example in Fig. 3.9.

PSF adapted dictionaries

Probably the most efficient way to regularize is by learning the basis $\mathbf{U}$ by keeping the principal components of a collection of simulated PSFs.

In Fig. 3.10, we show an example of such a learning process. We generated a set of 3 different types of 3D PSFs (Airy, coma, astigmatism), with different numerical apertures, wavelengths and refractive indices. We then generated 512 3D PSFs by varying the numerical aperture NA, the wavelength, the refractive index and the pupil function in a large range of physically admissible values. Each 3D PSF is composed of 18 slices along the axial direction, resulting in 9216 2D PSFs. We extracted some slices of the 3D PSFs at random and displayed them in Fig. 3.10. We then performed a principal component analysis (PCA) of 9216 different slices, which yielded the orthogonal basis displayed in Fig. 3.10. In this example, the 158 first elements of the PCA capture 99% of the energy. It means that the average approximation error is less than 1%.

3.5.2 Handling the bead function

The function b which describes the micro-bead needs to be discretized. Let $\mathbf{\Lambda_k}$ denote the sampling grid in which is defined the bead b and the PSF h, let $\mathbf{K}_b$

denote the domain supporting the bead b, and let $\mathbf{K}_h$ the domain supporting the PSF h. Here, we propose to model it as a discrete measure of the form

$$b = \sum_{\mathbf{k} \in \mathbf{K}_b} \mathbf{b}[\mathbf{k}] \delta_{\mathbf{\Lambda k}}. \tag{3.12}$$

With this choice, we have

$$h \star b = \varphi \star \mu(\mathbf{h}, \mathbf{K}_h) \star \mu(\mathbf{b}, \mathbf{K}_b),$$

where

$$\mu(\mathbf{h}, \mathbf{K}_h) := \sum_{\mathbf{k} \in \mathbf{K}_h} \mathbf{h}[\mathbf{k}] \delta_{\mathbf{\Lambda k}}.$$

We can then show that

$$\mu(\mathbf{h}, \mathbf{K}_h) \star \mu(\mathbf{b}, \mathbf{K}_b) = \sum_{\mathbf{k} \in \mathbf{K}_c} \mathbf{c}[\mathbf{k}] \delta_{\mathbf{\Lambda k}},$$

where $\mathbf{c} = \mathbf{h} \star \mathbf{b}$ and where $\mathbf{K}_c$ is the domain describing the support of the convolution. Therefore, with this discretization choice, we have to evaluate a *discrete convolution*, which can be performed efficiently using Fast Fourier Transforms. Overall, we see that

$$\Phi(\mathbf{h}_l) \star b = \Phi(\mathbf{h}_l \star \mathbf{b}),$$

where $\mathbf{h}_l = \mathbf{U}\mathbf{c}_l$ and we now use a discrete convolution instead of a continuous domain convolution.

Notice that discretizing a function by a discrete measure is bad in general since we do not have convergence of the discretization in L^2 for instance. However, this effect is not critical here since we afterwards convolve the result with φ, mapping everything to L^2 anyway.

3.5.3 Why a Gaussian fitting implicitly assumes centered PSFs?

To illustrate our claim, we work with $d = 1$ to ease to readability, but the analysis extends to $\mathbb{R}^d$. We define the k-th order moment of a function h as

$$m_k = \int_{\mathbb{R}} t^k h(t) \mathrm{d}t.$$

The center of mass of the real PSF is defined by $\bar{t} = \frac{m_1}{m_0}$. Now, assume that we estimate the center $\hat{t}$ of the observed PSF h as the maximizer of the correlation function c with a function g defined as

$$c(x) = \int h(t) g(t - x) \mathrm{d}t.$$

If g is radially symmetric, and decays slowly and smoothly to 0, we can write, up to a scaling, that

$$g(t) \simeq 1 - \frac{\gamma}{2} t^2$$

locally around 0 (i.e. at the locations where the PSF is large) for some constant $\gamma \propto g''(0)$. Then we can write that

$$c(x) \simeq m_0 - \frac{\gamma}{2}[x^2 m_0 + m_2 - 2x m_1].$$

The maximum of c is located at the point $\bar{x}$ where $c'(\bar{x}) = 0$, i.e. $\bar{x} \simeq \frac{m_1}{m_0} = \bar{t}$. As a conclusion, we see that any algorithm that defines the center of a PSF by correlation with a function, which is radial, stricly concave at 0 and sufficiently wide (s.t. the second order approximation is correct on the PSF support) will lead to recovering the approximate center of mass of the observed PSF.

3.5.4 Estimating space variations

Thin-plate approximation Once a family $(\mathbf{h}_l)_{1 \leq l \leq L}$, with $\mathbf{h}_l = \mathbf{U}\mathbf{c}_l$, is computed (see section 3.2.2), it is possible to project each noisy patch on this family to get a low dimensional representation of the selected PSFs. Let $(\bar{\mathbf{h}}_l)_l$ denote such an orthogonal family formed using the collection $(\mathbf{h}_l)_l$.

We aim to find a vectorial function $\alpha = (\alpha_l)_l : \mathbb{R}^d \to \mathbb{R}^L$ that describes how the coefficients of the PSFs vary in the subspace spanned by $(\bar{\mathbf{h}}_l)_l$. Let $\mathbf{p}_j = \sum_{l=1}^{L} \gamma_{j,l} \mathbf{h}_l$ denote the j-th estimated PSF at the position $\mathbf{x}_j$ output by algorithm 4. This provides a value $\beta_{l,j} = \langle \mathbf{p}_j, \bar{\mathbf{h}}_l \rangle$. In order to estimate the space variations, we can use surface fitting techniques on the set $(\mathbf{x}_j, \beta_{l,j})_{1 \leq j \leq J}$ to get an approximation of the function α.

There exist numerous surface fitting techniques. Following the numerical experiment conducted in [GCM13], it seems that the use of radial basis function [Buh03] is significantly more efficient than other approaches in the context of astronomy. We therefore resort to this technique.

Radial basis functions approximation can be interpreted as a variational problem in the framework of Reproducible Kernel Hilbert Spaces. In this context, the estimators $\hat{\alpha}_l$ of α_l can be expressed as

$$\hat{\alpha}_l = \underset{\alpha \in H^2(\mathbb{R}^2)}{\operatorname{argmin}} \frac{1}{2} \sum_{j=1}^{|J|} w_j |\alpha(\mathbf{x}_j) - \beta_{l,j}|^2 + \frac{\eta}{2} |\alpha|_{H^2}^2, \tag{3.13}$$

where $|\alpha|_{H^2}^2 \overset{\text{def.}}{=} \langle \Delta\alpha, \Delta\alpha \rangle_{L^2(\mathbb{R}^d)}$ and where $\eta > 0$ is a parameter that allows to trade off the proximity to the samples $\beta_{l,j}$ for the smoothness of the surface. In order to balance the importance of each PSF in the approximation, the weights w_j are chosen equal to the area of the Voronoï cell associated to each location $\mathbf{x}_j$. Imposing $\alpha \in H^2(\mathbb{R}^d)$ limits the approach to $d \leq 3$, otherwise point evaluation is not possible anymore. However, this is not a limitation for the applications in microscopy.

The solution of (3.13) is known to be a thin-plate spline [Pin85] and can be computed by solving a $(M+3) \times (M+3)$ linear system.

Chapter 4

A scalable estimator of sets of integral operators

Résumé : *L'objectif principal de ce travail est d'estimer un sous-espace vectoriel d'opérateurs de faible dimension afin d'améliorer l'identifiabilité des problèmes inverses aveugles. Nous proposons une méthode pour estimer un sous-espace $\widehat{\mathcal{H}}$ de tenseurs de faible rang, qui approche simultanément un ensemble d'opérateurs intégraux. Cet estimateur peut être considéré comme une généralisation de méthode de décomposition des tenseurs, ce qui n'a jamais été utilisé dans ce contexte. En outre, nous proposons de construire un sous-ensemble convexe de $\widehat{\mathcal{H}}$ afin de réduire davantage l'espace de recherche. Nous fournissons des garanties théoriques sur les estimateurs et quelques résultats numériques.*

Abstract: *The main objective of this work is to estimate a low dimensional subspace of operators in order to improve the identifiability of blind inverse problems. We propose a scalable method to find a subspace $\widehat{\mathcal{H}}$ of low-rank tensors that simultaneously approximates a set of integral operators. The method can be seen as a generalization of tensor decomposition models, which was never used in this context. In addition, we propose to construct a convex subset of $\widehat{\mathcal{H}}$ in order to further reduce the search space. We provide theoretical guarantees on the estimators and a few numerical results.*

This chapter is based on the publication [DEW19]:
Debarnot, V., Escande, P., & Weiss, P. (2019). A scalable estimator of sets of integral operators. Inverse Problems, 35(10), 105011.

Contents

4.1 Introduction **74**
4.1.1 Application examples 75
4.1.2 Contributions 75
4.1.3 Related works 76
4.2 Notation **76**
4.3 Operator representations **77**

4.3.1 Low-rank approximations 77
4.3.2 Product-convolution expansions 77
4.3.3 Hierarchical matrices 78
4.3.4 A general setting 78
4.4 Subspace estimation **78**
4.4.1 The algorithm 79
4.4.2 Theoretical guarantees 84
4.5 Subset estimation and projection **88**
4.5.1 Convex hull estimator 88
4.5.2 A projection algorithm 89
4.6 Numerical experiments **90**
4.6.1 Approximation rate and computation times 90
4.6.2 Blind deblurring 92
4.7 Conclusion **95**

4.1 Introduction

In many measurement devices, a signal v_0 living in some Hilbert space $\mathcal{B}_n$ of dimension n is probed indirectly using an operator $H_0 : \mathcal{B}_n \to \mathcal{B}_m$, where $\mathcal{B}_m$ is a Hilbert space of dimension m[1]. This yields a measurement vector $u_0 \in \mathcal{B}_m$ defined by

$$u_0 = f(H_0 v_0),$$

where f is some perturbation of the measurements (e.g. additive noise, modulus for phase retrieval, quantization,...). Solving an inverse problem consists in recovering an approximation $\hat{v}$ of the signal v_0 using the measurements u_0.

When the operator H_0 is known, many efficient solutions are now available. Unfortunately, in many cases, only a crude estimate of H_0 is available or it is even completely unknown. This is the field of bilinear or blind inverse problems. In that case, finding a reasonable approximation is far more involved. Significant theoretical progresses have been achieved in the last fews years though, [GCD12; ARR14; LLB16b; LLB17; Li+18b; LS18].

One of the key ideas behind these methods is the principle of lifting. To apply it, it is common to assume that the operator H_0 and the signal v_0 live in known low dimensional vector spaces of operators $\mathcal{H} = \text{span}(P_1, \dots, P_{|S|})$ and signals $\mathcal{Q} = \text{span}(q_1, \dots, q_{|T|})$. Then, we can write that $H_0 = P\alpha_0$ and that $v_0 = Q\beta_0$ for some $\alpha_0 \in \mathbb{R}^{|S|}$ and some $\beta_0 \in \mathbb{R}^{|T|}$. Under those assumptions, the blind inverse problem is simplified to finding a pair of vectors $(\alpha, \beta) \in \mathbb{R}^{|S|} \times \mathbb{R}^{|T|}$ and the measurement associated to the pair can be written as

$$u_0 = (P\alpha)(Q\beta) = \sum_{s \in S, t \in T} \alpha_s \beta_t w_{s,t},$$

with $w_{s,t} = P_s q_t$, $S = \{1, \dots, |S|\}$ and $T = \{1, \dots, |T|\}$. This last expression only depends on the outer product $\alpha\beta^T$, allowing to lift the problem to the

[1] In all this chapter, we assume that the operators are defined in finite dimensional spaces. An extension to infinite dimensional Hilbert spaces is feasible but requires additional discretization procedures. We decided to skip this aspect to clarify the exposition.

matrix space $\mathbb{R}^{|S|\times|T|}$. A typical way to attack the blind inverse problem is then to solve the following optimization problem:

$$\min_{M\in\mathbb{R}^{|S|\times|T|},\mathrm{rank}(M)=1} \frac{1}{2}\|\mathcal{W}M - y\|_2^2, \tag{4.1}$$

where $\mathcal{W} : M \mapsto \sum_{s\in S,t\in T} M_{s,t}w_{s,t}$. Various relaxations and algorithms can then be used to solve the lifted problem (4.1) and come with strong theoretical guarantees. We refer the interested reader to the above mentioned papers.

A critical issue to apply these techniques is the knowledge of the subspaces $\mathcal{H}$ and $\mathcal{Q}$. In this chapter, we will focus on the estimation of the subspace $\mathcal{H}$ from a sampling set of operators $(H_l)_{l\in L}$ in $\mathcal{C} \subset \mathcal{H}$.

The interest is that determining a low dimensional set of operators with a small volume can significantly ease the problem of operator identification in blind inverse problems. While our primary motivation lies in the field of inverse problems, this problem can also be understood as a generic problem of approximation theory.

4.1.1 Application examples

Space varying blur An example that will be used in our numerical experiments is the case of space varying blurs in wide field microscopy. In this imaging modality, the blur varies spatially due to multiple effects such as scattering or defocus for instance. The possible family of blurs may vary depending on factors such as the focal screw, the temperature (which changes the refractive index of the immersion oil), small tilts with respect to the focal plane and many other parameters that are hard to model from a mathematical point of view. It is possible to collect a family of operators $(H_l)_{l\in L}$ by observing fluorescent microbeads in a slide under various conditions and by using operator interpolation techniques such as [BEW17].

Magnetic Resonance Imaging (MRI) In MRI, the traditional observation model simply states that the Fourier transform values of the image are observed. The reality is far more complex and complete image formation models comprise many unknowns such as inhomogeneities of the main magnetic field or spatial sensitivities [Fes10]. To apply the proposed methodology to this device, the idea would be to first run many calibration scans to recover a list of operators $(H_l)_{l\in L}$ and then build a reduced model from this set.

Diffusion equations In many applications such as electrical impedance tomography [CIN99], the operators H_l are given implicitly as solutions of partial differential equations (PDEs). For instance diffusion equations, which are widespread in applications, are of the form $\mathrm{div}(c_l\nabla u) = v$, where c_l is a space varying diffusion coefficient that may change depending on external parameters. The application that maps v to u can be written as a linear integral operator H_l.

4.1.2 Contributions

The simplest approach to find a low dimensional vector space of operators $\mathcal{H}$ is to apply a principal component analysis (PCA) on the set of vectorized operators

$(H_l)_{l\in L}$. This approach is optimal in the sense of the Hilbert-Schmidt norm, but infeasible in practice. For instance, space varying blurring operators acting on small $2D$ images of size 1000×1000 can be encoded as matrices H_l of size $10^6 \times 10^6$, which can hardly be stored since each of them contains 9 Tera octets of data.

In this work, we therefore work under the assumption that the operators can be well approximated by low-rank tensors up to an invertible transformation. This hypothesis is reasonable for many applications of interest. For instance, it includes product-convolution expansions [EW17] and hierarchical matrices [Hac15] as special cases. We then provide an estimator $\widehat{\mathcal{H}}$ of the subspace of operators $\mathcal{H}$ with an upper-bound of its rate of approximation. In addition, we propose to construct an estimator $\widehat{\mathcal{C}}$ of $\mathcal{C}$, as the convex hull (in a matrix space) of the operators $(H_l)_{l\in L}$ projected onto $\widehat{\mathcal{H}}$. To make further use of this convex hull, we propose a fast projection algorithm on $\widehat{\mathcal{C}}$. We finally provide various numerical examples to highlight the strengths of the approach and its scalability.

4.1.3 Related works

To the best of our knowledge, the overall objective of this work is new, even though most of the individual tools that we combine together are well established. A related idea can be found in the literature of PDEs, where reduced order bases [MP89; RHP07] or their variants [CCS14] allow to solve families of PDEs efficiently. However, the objective there is to approximate the solutions of a PDE (usually linear) and not the associated operator. This is a significant difference, since approximating the operator (and its adjoint) allows to use the rich collection of nonlinear regularizers commonly used in the field of inverse problems to find regularized solutions.

4.2 Notation

In all the chapter, I, J, K and L are the sets of integers ranging from 1 to $|I|$, $|J|$, $|K|$ and $|L|$. We assume that $u \in \mathcal{B}_m$ is defined over a set X of cardinality m. We let $u(x)$ denote the value of u at $x \in X$. Similarly, we assume that $Hu \in \mathcal{B}_n$ is defined over a set Y. The set of linear operators from $\mathcal{B}_m$ to $\mathcal{B}_n$ is denoted Ξ. An operator $H \in \Xi$ can either refer to an operator or its matrix representation in an arbitrary orthogonal basis. The entries in the matrix representation will be denoted $H(x, y)$. The Frobenius norm of H is defined by $\|H\|_F := \sqrt{\mathrm{tr}(H^*H)}$. It is invariant by orthogonal transforms. The scalar products over all spaces will be denoted by $\langle \cdot \, , \, \cdot \rangle$.

The tensor product between two vectors $a \in \mathcal{B}_n$ and $b \in \mathcal{B}_m$ is defined by $(a \otimes b)(x, y) = a(x)b(y)$. The notation $\odot$ stands for the element-wise (Hadamard) product and if X has a group structure and $a_1, a_2 \in \mathcal{B}_m$, $a_1 \star a_2$ denotes the convolution product between a_1 and a_2.

Let $E = (e_i)_{i\in I}$ denote a family of elements in $\mathcal{B}_m$. The same notation will also apply to the matrix $E = [e_1, \ldots, e_{|I|}]$ and to the subspace $E = \mathrm{span}(e_i, i \in I)$. Let $W = (w_k)_{k\in K}$ denote a family of vectors with an SVD of the form $W = U\Sigma V^T$ with $U = [u_1, \ldots, u_n]$, then the truncated SVD, denoted $\mathrm{SVD}_{|I|}$,

by keeping $|I|$ elements is defined by:

$$\mathrm{SVD}_{|I|}(w_k, k \in K) \stackrel{\text{def.}}{=} [u_1, \ldots, u_{|I|}],$$

i.e. the $|I|$ left singular vectors associated to the largest singular values.

We let $\Delta_{N-1} = \{x \in \mathbb{R}^N,\ \sum_{i=1}^N x_i = 1, x_i \geq 0\}$ denote the simplex of dimension $N-1$. We let $\mathcal{K}_d$ denote the set of compact and convex sets of $\mathbb{R}^d$ with non empty interior. The Hausdorff distance between C_1 and C_2 is defined by $\mathcal{D}(C_1, C_2) = \inf\{\epsilon \geq 0\ :\ C_1 \subset C_2 + \epsilon B(0,1), C_2 \subset C_1 + \epsilon B(0,1)\}$, where $B(0,1)$ is the unit Euclidean ball. Let $(X_n)_{n\in\mathbb{N}}$ be sequence of random variables and $(t_n)_{n\in\mathbb{N}}$ denote a sequence of real numbers, the notation $X_n = \mathcal{O}_{\mathbb{P}}(t_n)$ means that for any $\epsilon > 0$, there exists $M > 0$ and $N > 0$ such that $\mathbb{P}\left(|X_n/t_n| > M\right) < \epsilon$ for all $n > N$. We let $\mathbb{E}(X)$ denote the expectation of a random variable X.

4.3 Operator representations

A critical requirement in this work is that the operators H_l can be approximated by (local) low-rank tensors, up to a linear transform. This need comes from the fact that arbitrary operators have no chance of being i) computed efficiently in large scale applications and ii) approximated efficiently by low dimensional subspaces. We describe a few possible decompositions below.

4.3.1 Low-rank approximations

The simplest assumption is to state that every operator H_l is well approximated by a low-rank tensor of the form $H_l = \sum_{k\in K} \alpha_{k,l} \otimes \beta_{k,l}$, with $|K| \ll \min(m,n)$. Unfortunately, many observation operators met in practice are concentrated along their diagonal, making this assumption unrealistic.

4.3.2 Product-convolution expansions

Product-convolution expansions are a family of decompositions that were analyzed recently in [EW17]. They can be defined whenever $X = Y$ and X possesses a group structure. It amounts to assuming that

$$H_l(u) = \sum_{k\in K} \alpha_{k,l} \star (\beta_{k,l} \odot u). \tag{4.2}$$

This decomposition can be computed efficiently using fast Fourier transforms.

To understand its link with the low-rank assumption, it is handy to introduce the *space varying impulse response* (SVIR) of H_l defined by $S_l(x,y) = H_l(x+y, y)$. One can show that the SVIR of an operator S_l of the form (4.2) can be written as $S_l = \sum_{k\in K} \alpha_{k,l} \otimes \beta_{k,l}$. Hence, assuming that H_l can be approximated by a product-convolution expansion is equivalent to saying that its SVIR is nearly low-rank.

This assumption covers many practical applications. For instance, a sufficient condition for an operator H_l to be exactly approximated using this decomposition is that all the *impulse responses* $(S_l(\ \cdot\ , y))_{y\in Y}$ of the operators H_l can be simultaneously encoded in the basis $\mathrm{span}(\alpha_{k,l}, k \in K)$.

4.3.3 Hierarchical matrices

Hierarchical matrix approximations [Beb08; Hac15], are another popular method to approximate linear operators. It amounts to assuming that $H_l = \sum_{k\in K} \alpha_{k,l} \otimes \beta_{k,l}$, where $|K|$ is not necessarily small compared to m and n, but where most of the elements $\alpha_{k,l}$ and $\beta_{k,l}$ have a small support, allowing for fast matrix-vector products. It can be shown that many practical applications are well suited to those approximations. It is particularly popular in the fields of PDEs and some inverse problems. In addition, related approximations such as fast multipole methods [BG97] or wavelet expansions [BCR91; EW15] also fit this formalism.

4.3.4 A general setting

Overall, the most generic assumption on H_l can be formulated as follows.

Assumption 4.3.1. *There exists a left invertible linear mapping $\mathcal{R} : \Xi \to \Xi$ such that each sample $H_l \in \Xi$ satisfies:*

$$S_l = \mathcal{R}(H_l) = \sum_{k\in K} \alpha_{k,l} \otimes \beta_{k,l},$$

where for all $l \in L$, the sets $(\alpha_{k,l})_k \in \mathcal{A}$ and $(\beta_{k,l})_k \in \mathcal{B}$, where $\mathcal{A}$ and $\mathcal{B}$ are subspaces of $\mathcal{B}_m^{|K|}$ and $\mathcal{B}_n^{|K|}$ respectively.

Introducing the operator $\mathcal{R}$ allows to encompass the usual low-rank assumption by taking $\mathcal{R} = \mathbf{I}$, but also the product-convolution expansions: going from the SVIR to the matrix representation can be expressed through an operator $\mathcal{R} : \Xi \to \Xi$ that shifts each column of H_l. The spaces $\mathcal{A}$ and $\mathcal{B}$ allow to incorporate support constraints, which are used for many decompositions such as the hierarchical matrices.

The final objective of this work is to estimate a subspace $\mathcal{H}$ and a set $\mathbb{C}$. In fact, we will rather estimate $\mathcal{H}_\mathcal{R} = \mathcal{R}\mathcal{H}$ and $\mathbb{C}_\mathcal{R} = \mathcal{R}\mathbb{C}$, which is equivalent since $\mathcal{R}$ is assumed to be left-invertible. In order to lighten the notation, we will skip the multiplication by $\mathcal{R}$ in the rest of the chapter.

4.4 Subspace estimation

In this section we provide an efficient and robust method to estimate the vector space of operators $\mathcal{H}$. We look for an estimator $\widehat{\mathcal{H}}$ of $\mathcal{H}$ with a tensor product structure:

$$E \otimes F \overset{\text{def.}}{=} \operatorname{span}(e_i \otimes f_j, (e_i)_{i\in I} \in \mathcal{E}_{|I|}, (f_j)_{j\in J} \in \mathcal{F}_{|J|}),$$

where the sets $\mathcal{E}_{|I|}$ and $\mathcal{F}_{|J|}$ can be thought of as:

- The set of orthogonal families of cardinality $|I|$ and $|J|$ defined by

$$\mathcal{E}_{|I|} = \{(e_i)_{i\in I} \in \mathcal{B}_m^{|I|}, \|e_i\|_2 = 1, \langle e_i, e_{i'}\rangle = \delta_{i,i'}\} \tag{4.3}$$

  and

$$\mathcal{F}_{|J|} = \{(f_j)_{j\in J} \in \mathcal{B}_n^{|J|}, \|f_j\|_2 = 1, \langle f_j, f_{j'}\rangle = \delta_{j,j'}\}. \tag{4.4}$$

- The set of orthogonal families of cardinality $|I|$ and $|J|$ with support constraints.
- Additional knowledge on the operators, such as non-negativity, can possibly be added.

We impose a tensor product structure so that every operator living in $\widehat{\mathcal{H}}$ can be evaluated rapidly. The sets $\mathcal{E}_{|I|}$ and $\mathcal{F}_{|J|}$ do not necessarily coincide with the sets $\mathcal{A}$ and $\mathcal{B}$, since it could be interesting to change the structure of the operators that are given as input to the algorithm.

The principle of our approach is to find a structured low-dimensional basis of operators that allows to approximate simultaneously all the sampled representations $(S_l)_{l\in L}$. This principle can be expressed through a variational problem, as follows:

$$(\widehat{E}, \widehat{F}) \stackrel{\text{def.}}{=} \underset{\substack{(e_i)_{i\in I}\in\mathcal{E}_{|I|}\\ (f_j)_{j\in J}\in\mathcal{F}_{|J|}}}{\operatorname{argmin}} \phi(E, F), \tag{4.5}$$

with

$$\phi(E, F) \stackrel{\text{def.}}{=} \frac{1}{2}\sum_{l\in L} \|\Pi_{E\otimes F}(S_l) - S_l\|_F^2,$$

where $\Pi_{E\otimes F}$ is the projection onto the tensor product space $E \otimes F$.

4.4.1 The algorithm

Problem (4.5) appears to be rather complicated due to the product structure of the search space. However, minimizing in $E \in \mathcal{E}_{|I|}$ for $F \in \mathcal{F}_{|J|}$ fixed and minimizing in $F \in \mathcal{F}_{|J|}$ for $E \in \mathcal{E}_{|I|}$ amounts to computing two singular value decompositions. This motivates the use of the alternating minimization procedure presented in Algorithm 7.

Algorithm 7 Alternating Least Squares (ALS)

Approximatively solve: Problem (4.5)
INPUT: $(S_l)_{l\in L}$, subspace constraints $\mathcal{E}_{|I|}$ and $\mathcal{F}_{|J|}$, initial guess (E_0, F_0).
1: **procedure**
2: Initialization: $t = 0$.
3: **while** stopping criterion not satisfied **do**
4: $E_{t+1} = \operatorname{argmin}_{E\in\mathcal{E}_{|I|}} \phi(E, F_t)$.
5: $F_{t+1} = \operatorname{argmin}_{F\in\mathcal{F}_{|J|}} \phi(E_{t+1}, F)$.
6: $t = t + 1$
7: **end while**
8: Return $\widehat{\mathcal{H}}_L = E_t \otimes F_t$.
9: **end procedure**

This algorithm is tightly related to common methods found in the field of tensor decompositions. In the particular case where $\mathcal{E}_{|I|}$ and $\mathcal{F}_{|J|}$ are sets of orthogonal families of cardinality $|I|$ and $|J|$, Problem (4.5) coincides exactly with the Tucker2 model. This decomposition was first introduced by Tucker in [Tuc66]. It was then reinvented independently and given several names such as tensor

PCA, 2DSVD, GLRAM, common component analysis, or tensor decompositions [Tuc66; DY05; Ye05; WBB11]. We refer to the review papers [KB09; Com14] for more insight on tensor decompositions. Computing this decomposition is a complex nonconvex problem, but the most standard approach to solve it takes the algorithmic form provided in Algorithm 7. It does not converge to the global minimizer in general and only provides approximate solutions. However, it is observed that it usually yields estimates close to the global minimizer in practice with a properly chosen initialization.

Orthogonal constraints

In this section, we detail the algorithm, when the spaces $\mathcal{E}_{|I|}$ and $\mathcal{F}_{|J|}$ denote the set of orthogonal families of cardinality $|I|$ and $|J|$ respectively.

Initialization The initialization of Algorithm 7 is of major importance since Problem (4.5) is non convex. We suggest using the High Order Singular Value Decomposition (HOSVD) [DDV00] in order to initialize the algorithm. This can be seen as a generalization of the SVD for tensors. As discussed in [KB09], this popular method provides a good starting point for an alternating algorithm.

From a variational point of view, the principle of the HOSVD consists in solving the following problems:

$$E_0 = \underset{E \in \mathcal{E}_{|I|}}{\operatorname{argmin}} \frac{1}{2} \sum_{l \in L} \|S_l - \sum_{k \in K} \Pi_E(\alpha_{k,l}) \otimes \beta_{k,l}\|_F^2, \tag{4.6}$$

$$F_0 = \underset{F \in \mathcal{F}_{|J|}}{\operatorname{argmin}} \frac{1}{2} \sum_{l \in L} \|S_l - \sum_{k \in K} \alpha_{k,l} \otimes \Pi_F(\beta_{k,l})\|_F^2, \tag{4.7}$$

i.e. to find the subspace E (resp. F) that captures most of the energy.

We will show below that we can leverage the specific low-rank structure of the operators S_l to evaluate the HOSVD rapidly. We let $A_l = [\alpha_{1,l}, \dots, \alpha_{|K|,l}]$ and $B_l = [\beta_{1,l}, \dots, \beta_{|K|,l}]$. We also diagonalize $A_l^T A_l \in \mathbb{R}^{|K| \times |K|}$ and $B_l^T B_l \in \mathbb{R}^{|K| \times |K|}$ as

$$A_l^T A_l = \Psi_{A_l} \Lambda_l \Psi_{A_l}^T \text{ and } B_l^T B_l = \Psi_{B_l} \Sigma_l \Psi_{B_l}^T$$

with $\Sigma_l = \operatorname{diag}(\sigma_{1,l}^2, \dots, \sigma_{|K|,l}^2)$ and $\Lambda_l = \operatorname{diag}(\lambda_{1,l}^2, \dots, \lambda_{|K|,l}^2)$.

Lemma 4.4.1 (Higher Order Singular Value Decompositon (HOSVD))**.** *Let* $\tilde{A}_l = A_l \Psi_{B_l} = [\tilde{\alpha}_{1,l}, \dots, \tilde{\alpha}_{|K|,l}]$ *and* $\tilde{B}_l = B_l \Psi_{A_l} = [\tilde{\beta}_{1,l}, \dots, \tilde{\beta}_{|K|,l}]$. *We have*

$$E_0 = \mathrm{SVD}_{|I|}(\sigma_{k,l} \tilde{\alpha}_{k,l}, k \in K, l \in L)$$

and

$$F_0 = \mathrm{SVD}_{|J|}(\lambda_{k,l} \tilde{\beta}_{k,l}, k \in K, l \in L)$$

We display in Figure 4.1 the dimensions of the tensor different elements.

Proof. We concentrate on E_0 only, since the proof for F_0 is similar. The first

$$\alpha_{k,l} = \begin{pmatrix} \\ \vdots \\ \\ \end{pmatrix} \in \mathbb{R}^{m\times 1} \qquad \beta_{k,l} = \begin{pmatrix} \\ \vdots \\ \\ \end{pmatrix} \in \mathbb{R}^{n\times 1}$$

Vectorized impulse responses

$$\boxed{A_l = (\alpha_{k,l})_k \in \mathbb{R}^{m\times|K|}}$$

Vectorized factors

$$\boxed{B_l = (\beta_{k,l})_k \in \mathbb{R}^{n\times|K|}}$$

$$A_l^T A_l = \Psi_{A_l}\Lambda_l\Psi_{A_l}^T \in \mathbb{R}^{|K|\times|K|} \qquad B_l^T B_l = \Psi_{B_l}\Sigma_l\Psi_{B_l}^T \in \mathbb{R}^{|K|\times|K|}$$

$$\Sigma_l = \mathrm{diag}(\sigma_{1,l}^2,\dots,\sigma_{|K|,l}^2) \in \mathbb{R}^{|K|\times|K|} \qquad \Lambda_l = \mathrm{diag}(\lambda_{1,l}^2,\dots,\lambda_{|K|,l}^2) \in \mathbb{R}^{|K|\times|K|}$$

$$A_l\Psi_{B_l} = \left[\tilde{\alpha}_{1,l},\dots,\tilde{\alpha}_{|K|,l}\right] \in \mathbb{R}^{m\times|K|} \qquad B_l\Psi_{A_l} = \left[\tilde{\beta}_{1,l},\dots,\tilde{\beta}_{|K|,l}\right] \in \mathbb{R}^{n\times|K|}$$

$$[(\sigma_{k,l}\tilde{\alpha}_{k,l})_{k,l}] = E\tilde{\Sigma}_E V_E \in \mathbb{R}^{m\times|K||L|} \qquad \left[(\lambda_{k,l}\tilde{\beta}_{k,l})_{k,l}\right] = F\tilde{\Sigma}_F V_F \in \mathbb{R}^{n\times|K||L|}$$

$$E_0 = |I| \text{ first column of } E,\ E_0 \in \mathbb{R}^{m\times|I|} \qquad F_0 = |J| \text{ first column of } F,\ F_0 \in \mathbb{R}^{n\times|J|}$$

Figure 4.1: Summary of the dimensions of the different elements involved in Lemma 4.4.1.

argument is to notice that Problem (4.6) is equivalent to

$$\begin{aligned}
E_0 &= \operatorname*{argmin}_{E\in\mathcal{E}_{|I|}} \frac{1}{2}\sum_{l\in L}\|S_l - \sum_{k\in K}\Pi_E(\alpha_{k,l})\otimes\beta_{k,l}\|_F^2 \\
&= \operatorname*{argmin}_{E\in\mathcal{E}_{|I|}} \frac{1}{2}\sum_{l\in L}\|\sum_{k\in K}\alpha_{k,l}\otimes\beta_{k,l} - \sum_{k\in K}\Pi_E(\alpha_{k,l})\otimes\beta_{k,l}\|_F^2 \\
&= \operatorname*{argmin}_{E\in\mathcal{E}_{|I|}} \frac{1}{2}\sum_{l\in L}\|\sum_{k\in K}(\alpha_{k,l} - \Pi_E(\alpha_{k,l}))\otimes\beta_{k,l}\|_F^2 \\
&= \operatorname*{argmin}_{E\in\mathcal{E}_{|I|}} \frac{1}{2}\sum_{l\in L}\|\sum_{k\in K}\Pi_{E^\perp}(\alpha_{k,l})\otimes\beta_{k,l}\|_F^2 \\
&= \operatorname*{argmax}_{E\in\mathcal{E}_{|I|}} \frac{1}{2}\sum_{l\in L}\|\sum_{k\in K}\Pi_E(\alpha_{k,l})\otimes\beta_{k,l}\|_F^2.
\end{aligned}$$

where $E^\perp$ is the orthogonal complementary of E.

Expanding the squared Frobenius norm leads to:

$$\begin{aligned}
E_0 &= \operatorname*{argmax}_{E\in\mathcal{E}_{|I|}} \frac{1}{2}\sum_{\substack{l\in L\\ k_1\in K\\ k_2\in K}}\langle\Pi_E(\alpha_{k_1,l})\otimes\beta_{k_1,l}, \Pi_E(\alpha_{k_2,l})\otimes\beta_{k_2,l}\rangle \\
&= \operatorname*{argmax}_{E\in\mathcal{E}_{|I|}} \frac{1}{2}\sum_{\substack{l\in L\\ k_1\in K\\ k_2\in K}}\langle\Pi_E(\alpha_{k_1,l}), \Pi_E(\alpha_{k_2,l})\rangle\langle\beta_{k_1,l}, \beta_{k_2,l}\rangle \\
&= \operatorname*{argmax}_{E\in\mathcal{E}_{|I|}} \frac{1}{2}\sum_{l\in L}\langle\Pi_E(A_l)^T\Pi_E(A_l), B_l^T B_l\rangle.
\end{aligned}$$

Recalling that $\tilde{A}_l = A_l \Psi_{B_l} = [\tilde{\alpha}_{1,l}, \dots, \tilde{\alpha}_{|K|,l}]$ and $B_l^T B_l = \Psi_{B_l} \Sigma_l \Psi_{B_l}^T$, this leads to:

$$\begin{aligned} E_0 &= \operatorname*{argmax}_{E \in \mathcal{E}_{|I|}} \frac{1}{2} \sum_{l \in L} \langle \Psi_{B_l}^T \Pi_E(A_l)^T \Pi_E(A_l) \Psi_{B_l}, \Sigma_l \rangle \\ &= \operatorname*{argmax}_{E \in \mathcal{E}_{|I|}} \frac{1}{2} \sum_{l \in L} \sum_{k \in K} \sigma_{k,l}^2 \|\Pi_E(\tilde{\alpha}_{k,l})\|_2^2 \\ &= \operatorname*{argmax}_{E \in \mathcal{E}_{|I|}} \frac{1}{2} \sum_{l \in L} \sum_{k \in K} \|\Pi_E(\sigma_{k,l} \tilde{\alpha}_{k,l})\|_2^2 \\ &= \mathrm{SVD}_{|I|}(\sigma_{k,l} \tilde{\alpha}_{k,l}, k \in K, l \in L). \end{aligned}$$

□

Lemma 4.4.1 shows that the computational cost of this initialization is dominated by the computation of two singular value decompositions: the first matrix is of size $m \times |L||K|$ and the second is of size $n \times |L||K|$. Depending on the cardinality $|L||K|$, this can be achieved either with standard linear algebra routines, or with randomized SVDs [HMT11]. In the applications that we consider here, n and m would typically be very large, while the number of samples $|L|$ and the rank of the tensors $|K|$ are expected to be small. In that situation, the computation can be performed even for very large scale applications.

Apart from being computable, the HOSVD presents additional advantages: the cost function can be controlled by the tail of the square singular values and running the alternating least squares on top of this initialization procedure ensures that the cost function will not increase above this upper-bound [DDV00]. In addition, the ranks $|I|$ and $|J|$ of the decomposition can be chosen automatically according to the decay of the singular values in the HOSVD.

The partial optimization problems The ALS algorithm requires to solve the two following partial optimization problem

$$\operatorname*{argmin}_{(e_i)_{i \in I} \in \mathcal{E}_{|I|}} \phi(E, F_t), \tag{4.8}$$

and

$$\operatorname*{argmin}_{(f_j)_{j \in J} \in \mathcal{F}_{|J|}} \phi(E_{t+1}, F), \tag{4.9}$$

where $E_t = [e_{t,1}, \dots, e_{t,|I|}]$ and $F_t = [f_{t,1}, \dots, f_{t,|J|}]$ are the output of Algorithm 7 at iteration $t \geq 0$. Solving the two subproblems requires the computation of two SVDs as in the previous section.

Lemma 4.4.2 (Partial optimization problem (4.8) and (4.9))**.** *Let* $\tilde{A}_l = A_l(B_l^T F_t) = [\tilde{\alpha}_{1,l}, \dots, \tilde{\alpha}_{|J|,l}]$ *and* $\tilde{B}_l = B_l(A_l^T E_{t+1}) = [\tilde{\beta}_{1,l}, \dots, \tilde{\beta}_{|I|,l}]$. *For all* $t > 0$ *we have*

$$E_{t+1} = \mathrm{SVD}_{|I|}(\tilde{\alpha}_{j,l}, j \in J, l \in L)$$

and

$$F_{t+1} = \mathrm{SVD}_{|J|}(\tilde{\beta}_{i,l}, i \in I, l \in L)$$

Proof. We concentrate on E_{t+1} only, since the proof for F_{t+1} is similar.

The projection $\Pi_{E\otimes F_t}(S_l)$ of the operator S_l onto the subspace $E \otimes F_t$ can be expressed as follows

$$\begin{aligned}
\Pi_{E\otimes F_t}(S_l) &= \sum_{k\in k} \Pi_E(\alpha_{k,l}) \otimes \Pi_{F_t}(\beta_{k,l}) \\
&= \sum_{k\in K}\sum_{i\in I} \langle \alpha_{k,l}, e_i\rangle e_i \otimes \sum_{j\in J} \langle \beta_{k,l}, f_{t,j}\rangle f_{t,j} \\
&= \sum_{i\in I}\sum_{j\in J} \left\langle \sum_{k\in K} \langle \beta_{k,l}, f_{t,j}\rangle \alpha_{k,l}, e_i \right\rangle e_i \otimes f_{t,j} \\
&= \sum_{j\in J} \Pi_E(\tilde{\alpha}_{j,l}) \otimes f_{t,j}.
\end{aligned}$$

Replacing this expression in (4.8), leads to solve the problem (4.6) again, with the difference that the second factors $(f_{t,j})$ form an orthogonal family. This allows to avoid the diagonalization step of Lemma 4.4.1:

$$\begin{aligned}
E_{t+1} &= \operatorname*{argmax}_{E\in\mathcal{E}_{|I|}} \frac{1}{2}\sum_{l\in L} \| \sum_{j\in J} \Pi_E(\tilde{\alpha}_{j,l}) \otimes f_{t,j}\|_F^2 \\
&= \operatorname*{argmax}_{E\in\mathcal{E}_{|I|}} \frac{1}{2}\sum_{l\in L}\sum_{j\in J} \|\Pi_E(\tilde{\alpha}_{j,l})\|_F^2 \\
&= \mathrm{SVD}_{|I|}(\tilde{\alpha}_{j,l}, j\in J, l\in L).
\end{aligned}$$

□

Hierarchical matrices

The results presented in the previous paragraph can readily be applied to the case of hierarchical decompositions. To this end, let $(\mathcal{T}_p)_{p\in P}$ denote a block-partition of $X \times Y$ [Hac15]:

- each $\mathcal{T}_p$ has a product structure: $\mathcal{T}_p = X_p \times Y_p$ for some $X_p \subset X$ and $Y_p \subset Y$.
- $\mathcal{T}_{p_1} \cap \mathcal{T}_{p_2} = \emptyset$ if $p_1 \neq p_2$.
- $X \times Y = \cup_{p\in P}\mathcal{T}_p$.

We assume that the subspaces $\mathcal{A}$ and $\mathcal{B}$ defining the operators S_l (see Assumption 4.3.1) encode support constraints. For each $l \in L$, the k-th tensor $\alpha_{k,l} \otimes \beta_{k,l}$ should satisfy:

$$\exists p_k \in P, \mathrm{supp}\,(\alpha_{k,l} \otimes \beta_{k,l}) \subseteq \mathcal{T}_{p_k}.$$

In order to apply the proposed ideas, we can first define two vectors of ranks $(q_p)_{p\in P}$ and $(r_p)_{p\in P}$ and generate an estimate $\widehat{\mathcal{H}}$ of $\mathcal{H}$ of the form

$$\widehat{\mathcal{H}} = \sum_{p\in P} E_p \otimes F_p,$$

with $\dim(E_p) = q_p$ and $\dim(F_p) = r_p$.

The estimation of the subspaces $\widehat{E}_p$ and $\widehat{F}_p$ can then be achieved with the same methodology as the one described for orthogonal matrices. In this setting, we can use HOSVD algorithm for each sub-blocks, this implies computing $|P|$ SVDs with lower dimensional matrices (depending of the size of support).

Non-negative decompositions

A common choice of family is the set of non-negative vectors, that is

$$\mathcal{E}_{|I|} = \{e \in \mathbb{R}_+^m, \|e\|_2 = 1\}^{|I|}$$

and

$$\mathcal{F}_{|J|} = \{f \in \mathbb{R}_+^n, \|f\|_2 = 1\}^{|J|},$$

where $\mathbb{R}_+^m$ denotes the set of nonnegative vectors of $\mathbb{R}^m$. Problems of the form (4.5) can then be solved with approaches such as [BD97; MHA08; Cic+07]. We do not explore this possibility further in this work.

4.4.2 Theoretical guarantees

We are now ready to establish the theoretical guarantees of the estimator $(\widehat{E}, \widehat{F})$ under additional assumptions on the sampling model.

Assumption 4.4.1 (Sampling model)**.** *The operators S_l are i.i.d. realizations of a random operator S with $\|S\|_F \leq r$ almost surely. Let*

$$\Phi(E, F) \stackrel{\text{def.}}{=} \frac{1}{2}\mathbb{E}\left(\|\Pi_{E\otimes F}(S) - S\|_F^2\right).$$

We assume that:

$$\inf_{(E\in\mathcal{E}_{|I|}, F\in\mathcal{F}_{|J|})} \Phi(E, F) = r^2\kappa(I, J). \tag{4.10}$$

The scaling in r^2 in equation (4.10) is natural: if the random operator S is scaled by a constant factor, so will the approximation error. The bound (4.10) provides the best achievable estimate of subspace.

Arbitrary bounded errors

We let $(\widehat{E}_L, \widehat{F}_L)$ denote the solution of (4.5). In practice, we do not directly observe the operator S_l, but only an approximate version of it that we denote S_l^K. The approximation S_l^K verifies Assumption 4.3.1 and we can only estimate the approximation error. This is capture by the following assumption.

Assumption 4.4.2 (Approximation error)**.** *The operators S_l^K satisfy the following inequality : $\|S_l^K - S_l\|_F \leq \kappa(K)\|S_l\|_F$ with $\kappa(K) \leq 1$.*

Theorem 4.4.3. *Assume that Assumptions 4.4.1 and 4.4.2 hold, then:*

$$\mathbb{P}\left(\Phi(\widehat{E}_L, \widehat{F}_L) \leq 6r^2 \max\left(\kappa(K), \kappa(I, J)\right)\right) \geq 1-2\exp\left(-8|L|\max\left(\kappa(K), \kappa(I, J)\right)^2\right).$$

We first discuss the consequences of this Theorem 4.4.3 prior to detailing its proof. In case the relative approximation error $\kappa(K)$ is too large w.r.t. to $\kappa(I,J)$ there will be no guarantee to reach $\Phi(E^*,F^*)$ since the best achievable error will be of the order $r^2\kappa(K)$. This bound can be achieved with probability $1-\delta$ by choosing $|L| = \frac{\log(2/\delta)}{8\kappa(K)^2}$. However, when the approximation gets finer i.e. $\kappa(K) < \kappa(I,J)$, the estimator $(\widehat{E}_L, \widehat{F}_L)$ becomes as good as possible up to a constant. This bound can be achieved with probability $1-\delta$ by choosing $|L| = \frac{\log(2/\delta)}{8\kappa(I,J)^2}$.

Proof. We let

$$\Phi_L(E,F) \stackrel{\text{def.}}{=} \frac{1}{2|L|}\sum_{l\in L}\left(\|\Pi_{E\otimes F}(S_l) - S_l\|_F^2\right)$$

and

$$\Phi_L^K(E,F) \stackrel{\text{def.}}{=} \frac{1}{2|L|}\sum_{l\in L}\left(\left\|\Pi_{E\otimes F}(S_l^K) - S_l^K\right\|_F^2\right).$$

Step 1. We first control the bias term as follows

$$|\Phi_L^K(E,F) - \Phi_L(E,F)| \le 3r^2\kappa(K)/2. \tag{4.11}$$

Let $G = E\otimes F$ and $G^\perp$ denote its orthogonal complementary with respect to the Frobenius inner-product over the space of operators. We let $D_l = S_l^K - S_l$ and notice that $\|D_l\|_F \le \kappa(K)\|S_l\|_F$ by Assumption 4.4.2. Now, we can decompose S_l as $S_l = S_l^G + S_l^{G^\perp}$ and S_l^K as $S_l^K = S_l^G + S_l^{G^\perp} + D_l^G + D_l^{G^\perp}$. This leads to

$$\|\Pi_G(S_l) - S_l\|_F^2 = \|S_l^{G^\perp}\|_F^2$$

and

$$\|\Pi_G(S_l^K) - S_l^K\|_F^2 = \|S_l^{G^\perp} + D_l^{G^\perp}\|_F^2.$$

So that

$$\begin{aligned}
&\left|\|\Pi_G(S_l^K) - S_l^K\|_F^2 - \|\Pi_G(S_l) - S_l\|_F^2\right| \\
&= \left|2\langle S_l^{G^\perp}, D_l^{G^\perp}\rangle + \|D_l^{G^\perp}\|_F^2\right| \\
&\le 2r^2\kappa(K) + r^2\kappa(K)^2 \le 3r^2\kappa(K).
\end{aligned}$$

By summing this inequality over $l\in L$, we get the inequality (4.11).

Step 2. As in the previous step, we let $G = E\otimes F$. We show here that

$$\mathbb{P}(|\Phi_L - \Phi| \ge t) \le 2\exp\left(-\frac{8|L|t^2}{r^4}\right). \tag{4.12}$$

Let us introduce the random variable $Z_l = \|\Pi_G(S_l) - S_l\|_F^2 = \|S_l^{G^\perp}\|_F^2$. We have $\mathbb{E}(Z_l) = \Phi$ and by Assumption 4.4.1, we have $Z_l \in [0, r^2]$. Let $X \stackrel{\text{def.}}{=} \sum_{l\in L}(Z_l - \mathcal{E}(Z_l))$. We have $X/(2|L|) = \Phi_L - \Phi$ and Hoeffding's inequality [BLM13, Thm 2.8] ensures that for all $t>0$ the random variable X satisfies

$$\mathbb{P}(|X| \ge t) \le 2\exp\left(-\frac{2t^2}{|L|r^4}\right).$$

Step 3. We are now ready to conclude the proof. We have

$$|\Phi_L^K - \Phi| \leq |\Phi_L^K - \Phi_L| + |\Phi_L - \Phi|.$$

The problem (4.10) has at least one solution denoted $(E^\star, F^\star)$. Indeed, the finite dimensional vector spaces E and F can be parameterized by $|I|$ and $|J|$ unit vectors. The tensor product of $|I||J|$ unit balls is a compact set and the function Φ is continuous, ensuring the existence of a minimizer. We get:

$$\begin{aligned}\Phi(\widehat{E}_L, \widehat{F}_L) &\leq \Phi_K^L(\widehat{E}_L, \widehat{F}_L) + |\Phi_L^K - \Phi_L| + |\Phi_L - \Phi| \\ &\leq \Phi_K^L(E^\star, F^\star) + 3/2r^2\kappa(K) + |\Phi_L - \Phi| \\ &\leq \Phi(E^\star, F^\star) + 3r^2\kappa(K) + 2|\Phi_L - \Phi|.\end{aligned}$$

Using the inequality (4.12), we get for all $t > 0$:

$$\mathbb{P}\left(\Phi(\widehat{E}_L, \widehat{F}_L) \leq r^2\left(\kappa(I,J) + 3\kappa(K)\right) + 2t\right) \geq 1 - 2\exp\left(-\frac{8|L|t^2}{r^4}\right).$$

The first part of the theorem is obtained by selecting $t = r^2 \max(\kappa(K), \kappa(I,J))$. □

Random errors

The bound in Theorem 4.4.3 may look a bit disappointing since it is impossible to reach the absolute best error $r^2\kappa(I,J)$. This is due to the fact that the approximation errors $D_l = S_l^K - S_l$ can be adversarial and create a bias in the estimation. If we add randomness assumptions on these errors, the situation can improve. We illustrate it below with random isotropic errors.

Theorem 4.4.4. *Suppose that assumptions 4.4.1 and 4.4.2 hold. Assume furthermore that the errors D_l have an isotropic distribution with $\mathbb{E}(\|D_l\|_F^2) = R^2$ and $\|D_l\|_F^2 \leq \kappa^2(K)r^2$ almost surely then:*

$$\mathbb{P}\left(\Phi(\widehat{E}_L, \widehat{F}_L) \leq r^2(\kappa(I,J) + \epsilon)\right) \geq 1 - 8\exp\left(-\frac{2|L|\epsilon^2}{(6\kappa(K)+1)^2}\right).$$

Theorem 4.4.4 shows that under isotropic random approximation errors, the estimator $(\widehat{E}_L, \widehat{F}_L)$ can become arbitrarily good. This bound can be achieved with probability $1-\delta$ by choosing $|L| = \frac{(6\kappa(K)+1)^2}{2\epsilon^2}\log\left(\frac{8}{\delta}\right)$.

Proof. We let,

$$\Phi_L(E,F) \stackrel{\text{def.}}{=} \frac{1}{2|L|}\sum_{l\in L}\left(\|\Pi_{E\otimes F}(S_l) - S_l\|_F^2\right)$$

and

$$\Phi_L^K(E,F) \stackrel{\text{def.}}{=} \frac{1}{2|L|}\sum_{l\in L}\left(\left\|\Pi_{E\otimes F}(S_l^K) - S_l^K\right\|_F^2\right) - \frac{R^2}{mn}(mn - |I||J|)R^2,$$

where m and n are the dimension of the space $\mathcal{B}_m$ and $\mathcal{B}_n$ respectively.

The difference compared to the previous proof is that we can now bound $|\Phi_L^K - \Phi_L|$ by a quantity that vanishes with the number of observations L. For any pair of subspaces (E, F), we have:

$$\mathbb{P}\left(\left|\Phi_L^K(E,F) - \Phi_L(E,F)\right| \geq t\right) \leq 2\exp\left(-\frac{|L|t^2}{18r^4\kappa(K)^2}\right).$$

To prove this statement, let $G = E \otimes F$. We get:

$$|\Phi_L^K(E,F) - \Phi_L(E,F)| = \left|\frac{1}{2|L|}\sum_{l\in L}\left(2\langle S_l^{G^\perp}, D_l^{G^\perp}\rangle + \|D_l^{G^\perp}\|_F^2\right) - \frac{R^2}{mn}(mn - |I||J|)\right|.$$

Letting $Z_l^{G^\perp} = 2\langle S_l^{G^\perp}, D_l^{G^\perp}\rangle + \|D_l^{G^\perp}\|_F^2$, we get $\mathbb{E}(Z_l^{G^\perp}) = (mn - |I||J|)R^2/(mn)$ since D_l is isotropic. Indeed, $\mathbb{E}\left(2\langle S_l^{G^\perp}, D_l^{G^\perp}\rangle\right) = 0$ since $\mathbb{E}(D_l) = 0$, and by letting $\Pi_{G^\perp}$ denote the projection onto $G^\perp$ we get:

$$\begin{aligned}
&\mathbb{E}\left(\|D_l^{G^\perp}\|_F^2\right) \\
&= \mathbb{E}\left(\operatorname{tr}\left(D_l^T \Pi_{G^\perp}^T \Pi_{G^\perp} D_l\right)\right) \\
&= \mathbb{E}\left(\operatorname{tr}\left(\Pi_{G^\perp} D_l D_l^T \Pi_{G^\perp}^T\right)\right) = \operatorname{tr}\left(\Pi_{G^\perp}\mathcal{E}(D_l D_l^T)\Pi_{G^\perp}^T\right) \\
&= \frac{R^2}{mn}\operatorname{tr}\left(\Pi_{G^\perp}^T \Pi_{G^\perp}\right) = \frac{R^2}{mn}(mn - |I||J|)
\end{aligned}$$

Noticing that $|Z_l^{G^\perp}| \leq r^2(2\kappa(K) + \kappa(K)^2) \leq 3r^2\kappa(K)$, we can use Hoeffding's inequality:

$$\mathbb{P}\left(\left|\frac{1}{2|L|}\sum_{l\in L} Z_l^{G^\perp} - \mathcal{E}(Z_l^{G^\perp})\right| \geq t\right) \leq 2\exp\left(-\frac{2|L|t^2}{9r^4\kappa(K)^2}\right).$$

We are now ready to conclude the proof. Similarly to the previous proof, we get:

$$\Phi(\widehat{E}_L, \widehat{F}_L) \leq \Phi_L^K(E^\star, F^\star) + |\Phi_L^K(\widehat{E}_L, \widehat{F}_L) - \Phi_L(\widehat{E}_L, \widehat{F}_L)| + |\Phi_L(\widehat{E}_L, \widehat{F}_L) - \Phi(\widehat{E}_L, \widehat{F}_L)|.$$

Using a union bound argument (given a set of events $(A_i)_{i\in\mathbb{N}}$, we have $\mathbb{P}\left(\cup_{i\in\mathbb{N}} A_i\right) \leq \sum_{i\in\mathbb{N}} \mathbb{P}(A_i)$), we get:

$$\mathbb{P}\left(\Phi(\widehat{E}_L, \widehat{F}_L) \leq \Phi_L^K(E^\star, F^\star) + t + t'\right) \geq \\ 1 - 2\exp\left(-\frac{2|L|t^2}{9r^4\kappa(K)^2}\right) - 2\exp\left(-\frac{8|L|t'^2}{r^4}\right).$$

Using another union bound argument, we get:

$$\mathbb{P}\left(\Phi(\widehat{E}_L, \widehat{F}_L) \leq \Phi(E^\star, F^\star) + 2t + 2t'\right) \geq \\ 1 - 4\exp\left(-\frac{2|L|t^2}{9r^4\kappa(K)^2}\right) - 4\exp\left(-\frac{8|L|t'^2}{r^4}\right).$$

By taking $t = 6\kappa(K)t'$ we get

$$\mathbb{P}\left(\Phi(\widehat{E}_L, \widehat{F}_L) \leq \Phi(E^\star, F^\star) + t'(12\kappa(K) + 2)\right) \geq 1 - 8\exp\left(-\frac{8|L|t'^2}{r^4}\right).$$

Letting $\epsilon > 0$ and setting $t' = \frac{r^2\epsilon}{12\kappa(K)+2}$, we get:

$$\mathbb{P}\left(\Phi(\widehat{E}_L, \widehat{F}_L) \leq \Phi(E^\star, F^\star) + \epsilon r^2\right) \geq 1 - 8\exp\left(-\frac{2|L|\epsilon^2}{(6\kappa(K)+1)^2}\right).$$

Finally, given $\delta > 0$, we can select $|L| = \frac{(6\kappa(K)+1)^2}{2\epsilon^2}\log\left(\frac{8}{\delta}\right)$ so that the following holds true:

$$\mathbb{P}\left(\Phi(\widehat{E}_L, \widehat{F}_L) \leq \Phi(E^\star, F^\star) + r^2\epsilon\right) \geq 1 - \delta.$$

This concludes the proof. □

4.5 Subset estimation and projection

4.5.1 Convex hull estimator

In this section, we show that we can use the collection of observed operators to restrict further the space of admissible elements by computing its convex hull. In particular, it is a key element of several blind inverse approaches proposed in Chapter 7 and Chapter 8. We will show that this convex set has good estimation properties.

We assume that $(S_l)_{l\in L}$ are i.i.d. realizations of the random operator S. We assume that the distribution of S is uniform over a convex, compact and non-empty set $\mathcal{C}$. Letting Π_L denote the projector onto $\widehat{\mathcal{H}}_L = \widehat{E}_L \otimes \widehat{F}_L$, we propose to construct an estimate $\widehat{\mathcal{C}}_L^{K,\Pi}$ of $\mathcal{C}$, by taking the convex hull of the projected and observed operators

$$\widehat{\mathcal{C}}_L^{K,\Pi} \overset{\text{def.}}{=} \operatorname{conv}(\Pi_L(S_l^K), l \in L).$$

We can only expect $\widehat{\mathcal{C}}_L^{K,\Pi}$ to approximate $\Pi_L(\mathcal{C})$, and not $\mathcal{C}$ directly, since some information is lost by the projection. The following proposition summarizes the rate of convergence of $\widehat{\mathcal{C}}_L^{K,\Pi}$.

Proposition 4.5.1. *Under the assumptions 4.4.1 and 4.4.2, we get the following result*

$$\mathcal{D}(\widehat{\mathcal{C}}_L^{K,\Pi}, \Pi_L(\mathcal{C})) \leq r\kappa(K) + \mathcal{O}_{\mathbb{P}}\left(\left(\frac{\ln|L|}{|L|}\right)^{\frac{1}{\alpha}}\right),$$

where $\mathcal{D}$ denotes the Hausdorff distance between sets and

- *$\alpha = d$ if $\mathcal{C}$ is a polytope,*
- *$\alpha = \frac{d+1}{2}$ if $\mathcal{C}$ has C^3 boundary and positive curvature everywhere.*

Proof. Step 1. The difficult part of this inequality is the rightmost term, which is due to [Bru17] (Theorem 11). With our notation, his main result states that $\mathcal{D}(\widehat{\mathcal{C}}_L, \mathcal{C}) = \mathcal{O}_{\mathbb{P}}\left(\left(\frac{\ln|L|}{|L|}\right)^{\frac{1}{\alpha}}\right)$, where $\widehat{\mathcal{C}}_L = \operatorname{conv}(S_l, l \in L)$.

Step 2. In order to obtain our result, we first observe that since Π_L is a projection, it is also a contraction and $\mathcal{D}(\Pi_L(\widehat{\mathcal{C}}_L), \Pi_L(\mathcal{C})) \leq \mathcal{D}(\widehat{\mathcal{C}}_L, \mathcal{C})$.

Step 3. Now, let $\widehat{\mathcal{C}}_L^K \overset{\text{def.}}{=} \operatorname{conv}(S_l^K, l \in L)$. We have

$$\mathcal{D}(\widehat{C}_L, \widehat{C}_L^K) \leq r\kappa(K) \tag{4.13}$$

Indeed, the distance function $d_{\widehat{C}_L^K}(H) = \inf_{H' \in \widehat{C}_L^K} \|H - H'\|_2$ is convex. Hence, the problem $\sup_{H \in \widehat{\mathcal{C}}_L} d_{\widehat{C}_L^K}(H)$ appearing in the definition of the Hausdorff distance consists of finding the maximum of a convex function over a convex set. Hence the maximum is attained at an extremal point of $\widehat{C}_L^K$, i.e. at a point S_l^K. All those points satisfy $\|S_l^K - S_l\|_F \leq r\kappa(K)$, hence $\sup_{H \in \widehat{\mathcal{C}}_L} d_{\widehat{C}_L^K}(H) \leq r\kappa(K)$. A similar reasoning on the other part of the distance yields the inequality (4.13). Since Π_L is a contraction, we also get $\mathcal{D}(\Pi_L(\widehat{C}_L), \Pi_L(\widehat{C}_L^K)) \leq r\kappa(K)$.

Step 4. To conclude, we use the fact that the Hausdorff distance satisfies the triangle inequality. Hence:

$$\begin{aligned}\mathcal{D}(\widehat{\mathcal{C}}_L^{K,\Pi}, \Pi_L(\mathcal{C})) &\leq \mathcal{D}(\widehat{\mathcal{C}}_L^{K,\Pi}, \Pi_L(\widehat{\mathcal{C}}_L)) \\ &\quad + \mathcal{D}(\Pi_L(\widehat{\mathcal{C}}_L), \Pi_L(\mathcal{C})) \\ &\leq r\kappa(K) + \mathcal{O}_{\mathbb{P}}\left(\left(\frac{\ln|L|}{|L|}\right)^{\frac{1}{\alpha}}\right).\end{aligned}$$

□

Remark 4.5.1. *There are different ways to control the distance between sets. Another possibility is to use the Nikodym metric, i.e. the relative difference of volume between $\widehat{\mathcal{C}}_L^{K,\Pi}$ and $\Pi_L(\mathcal{C})$. For this metric, it can be shown that the convex hull estimator is a minimax operator (i.e. that it is optimal uniformly on the class of convex bodies) and we also obtain a convergence rate of the form $\mathcal{O}_{\mathbb{P}}(|L|^{-2/d+1})$ for a convex set $\mathcal{C}$ with C^3 boundary and positive curvature everywhere.*

Remark 4.5.2. *Proposition 4.5.1 only characterizes the asymptotic behavior of this estimator. This result should be taken carefully since the constants in the $\mathcal{O}_{\mathbb{P}}$ depend on the geometry of the convex set $\mathcal{C}$. In particular, the sharper the corners of $\mathcal{C}$, the larger the constant.*

4.5.2 A projection algorithm

In what follows, we let $\widehat{\mathcal{C}} = \widehat{\mathcal{C}}_L^{K,\Pi}$ to simplify the notation. In the framework of blind inverse problems (see equation (4.1)), the knowledge of the convex set $\widehat{\mathcal{C}}$ may lead to the resolution of variational problems of the form

$$\min_{H \in \widehat{\mathcal{C}}, u \in \mathcal{W}} \frac{1}{2}\|Hu - y\|_2^2. \tag{4.14}$$

A critical tool to solve (4.14) is a projection operator $\Pi_{\widehat{\mathcal{C}}}$ onto the set $\widehat{\mathcal{C}}$. For instance, it would allow using a projected gradient descent. Let $H \in \Xi$ and $S = \mathcal{R}(H)$. The projection is defined as follows:

$$\Pi_{\widehat{\mathcal{C}}}(S) = \operatorname*{argmin}_{\lambda \in \Delta_{|L|}} \frac{1}{2}\|M\lambda - S\|_F^2, \tag{4.15}$$

where $M : \lambda \to \sum_{l \in L} \lambda_l \Pi_L(S_l^K)$ and $\Delta_{|L|} = \{x \in \mathbb{R}^{|L|}, \sum_{i=1}^{|L|} x_i = 1, x_i \geq 0\}$.

Depending on the number of samples $|L|$, different algorithms can be used to solve (4.15). For small $|L|$, interior point methods [NN94] are an excellent

candidate, since they lead to high precision solutions in short computation times. For larger $|L|$, they become intractable and it is then possible to use lighter, but less precise first order solutions. We detail such an approach below.

First, we let $\tau = 1/\|M^*M\|_F$. This quantity can be computed using a power method for instance. We can then use the accelerated proximal gradient [Nes13] descent described in Algorithm 8.

Algorithm 8 Projection onto convex hull of operators

INPUT: $\Pi_L(S_l^K)$, S, initial guess $\lambda_0 \in \Delta_{|L|}$.
OUTPUT: Projection of S onto $\widehat{\mathcal{C}}$.

```
procedure
    for k = 1, 2, ..., k_end do
        λ̃_k = Π_{Δ_|L|} (λ_k − τM*(Mλ_k − S))
        λ_{k+1} = λ̃_k + (k−1)/(k+2) (λ̃_k − λ̃_{k−1})
    end for
    Return Mλ_{k_end}
end procedure
```

The projection on the $(|L|-1)$-dimensional simplex can be computed in linear time and Algorithm 8 ensures that the cost function decays as $O(1/k^2)$. The matrix M^*M can be precomputed with a numerical complexity in $O(|L|^2(|I|^2n + |J|^2m))$. The product M^*S can also be computed efficiently, for operators S given in a tensor form. This is for instance the case if $S \in \widehat{\mathcal{H}}_L$.

4.6 Numerical experiments

In this section we illustrate the previous methods with a few numerical examples.

4.6.1 Approximation rate and computation times

The setting

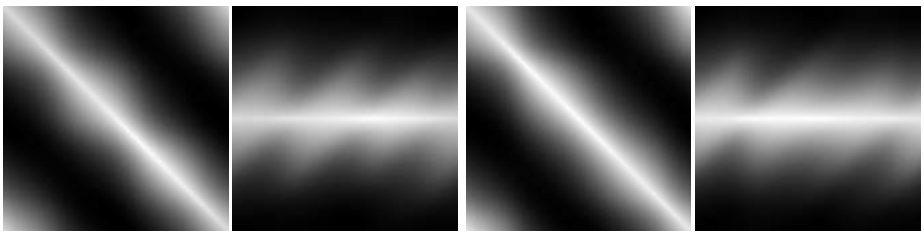
We start with a one dimensional diffusion equation as introduced in Section 4.1.1. Our main aim here is to illustrate the computational complexity of the approach.We take $\mathcal{B}_n = \mathcal{B}_m = \mathbb{R}^n$ with $n = m$. We define the operator ∇ with forward finite differences and homogeneous Neumann boundary conditions. The divergence operator $\operatorname{div} = -\nabla^*$, where ∇^* is the adjoint of ∇.

We wish to find a family of estimators of the mapping $f \mapsto u$ for the following equation

$$\operatorname{div}(c\nabla u) = f, \forall f \in \mathbb{R}^n,$$

and for diffusion coefficients $c \in \mathbb{R}^n$ living in a subset Ω of nonnegative vectors. We assume that we can access $|L|$ observations of c, denoted c_l for $l \in L$. We let

$$\begin{aligned} H_l : \quad & \mathbb{R}^n \mapsto \mathbb{R}^n \\ & f \mapsto (\operatorname{div}(c_l\nabla))^+ f \end{aligned}$$



(a) Kernel operator 1. (b) SVIR operator 1. (c) Kernel operator 2. (d) SVIR operator 2.

Figure 4.2: Kernel and SVIR of two different inverse diffusion operators.

denote the operators of interest, where + denotes the pseudo-inverse. In our simulations we consider diffusion coefficients c_l of the form:

$$c_l(x) = 3 + \sum_{p \in P} w_{l,1}(p) \cos(2\pi p x) + w_{l,2}(p) \sin(2\pi p x), \forall x \in \mathbb{R}^n,$$

with $w_{l,1}, w_{l,2}$ taken uniformly at random in the $|P| - 1$-dimensional simplex $\Delta_{|P|-1}$. We assume that the operators H_l are given in a product-convolution form, or equivalently that their SVIR S_l can be written as $S_l = \sum_{k \in K} \alpha_{k,l} \otimes \beta_{k,l}$. In our numerical experiments, we compute the factors $\alpha_{k,l}$ and $\beta_{k,l}$ using a SVD of S_l. This is feasible since we work in 1D. The number of factors in the SVD is set to $|K| = 20$ which is enough to capture 97% percent of the energy on average. Two instances of operators H_l and their SVIR S_l are displayed in Figure 4.2.

Description of the approaches

Given $|I|$ and $|J|$, our aim is to find two families $(e_i)_{i \in I} \in \mathcal{E}_{|I|}$ and $(f_j)_{j \in J} \in \mathcal{F}_{|J|}$, with $\mathcal{E}_{|I|}$ and $\mathcal{F}_{|J|}$ defined as the sets of orthonormal families, see equations (4.3) and (4.4). We compare four approaches to estimate the subspace $\mathcal{H}$.

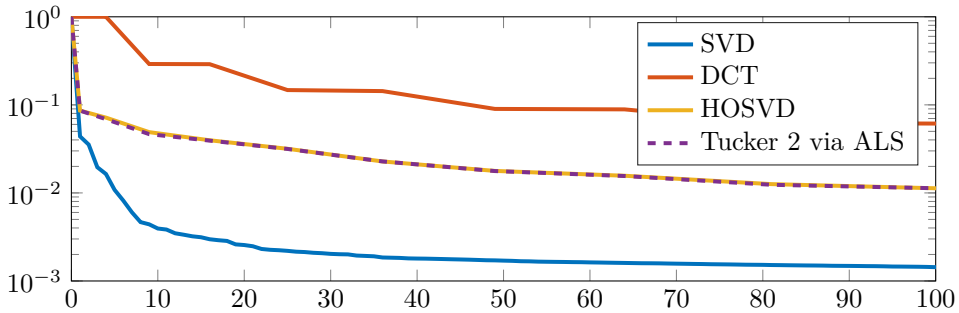
- SVD: We concatenate the vectorized representation of H_l in a matrix M. The family $(e_i)_{i \in I}$ is set to be the first $|I|$ left-eigenvectors, and the family $(f_j)_{j \in J}$ is set to be the first $|J|$ right-eigenvectors of M. This approach is optimal in terms of Frobenius norm but can only be applied because we work in a low dimensional 1D setting.

- DCT: We set e_i and f_j as the first elements of the discrete cosine transform, i.e. $e_i(x) = \cos(\pi/(n-1)ix)$ and $f_j(y) = \cos(\pi/(m-1)iy)$ with n corresponding to the number of elements in the discretization. The family $(e_i \otimes f_j)_{i \in I, j \in J}$ is in tensor product form and it is orthogonal, which allows making very fast computations.

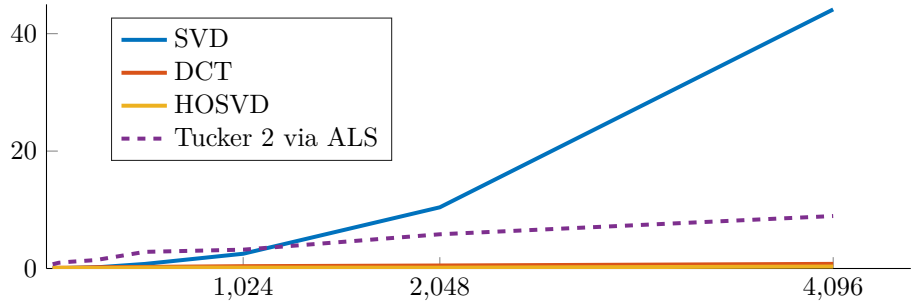
- HOSVD: implements the decomposition in equations (4.6) and (4.7).

- ALS: use the Alternating Least Square Algorithm 7 with 15 iterations and the HOSVD as an initialization.

We first compare the four different methods in terms of their approximation quality for $|L| = 50$ observations. We evaluate the average relative projection error defined by $\mathbb{E}\left(\frac{\|H - \Pi_{\widehat{\mathcal{H}}}(H)\|_F}{\|H\|_F}\right)$. It can be evaluated through a Monte-Carlo



(a) Relative approximation error versus the dimension $|I||J|$ of each basis.



(b) Computation time in seconds versus the dimension n of the problem (the operators $(H_l)_{l\in L}$ are of size $n \times n$).

Figure 4.3: Numerical behavior for 1D operators.

simulation. Figure 4.3a shows the relative error for the different methods and various sizes $|I|$ with $|I| = |J|$.

The approximation rate given by the SVD is upper-bounded by the approximation properties of the considered family of operators. This is an illustration of Theorem 4.4.3 which describes the behavior of the approximation rate in terms of the constants $\kappa(I, J)$ and $\kappa(K)$. In this example, we distinguish two regimes: when $|I||J| < |L|$ the approximation rate is bounded by the constant $\kappa(I, J)$, and when $|I||J| \geq |L|$, the approximation rate is bounded by the constant $\kappa(K)$.

Computing times

We now examine the computational time for each method in Figure 4.3b.

The efficiency of the SVD has to be balanced by its important computational time. It becomes completely impractical on a usual workstation when the dimension n of the space $\mathcal{B}_n$ is larger than 10^5. We also observe that using the ALS algorithm instead of the HOSVD leads to negligible gains, despite a significantly higher computational burden. The runtime is basically proportional to the number of iterations.

4.6.2 Blind deblurring

In this section we apply the proposed method of subspace estimation to solve a blind deblurring problem. We use simulated operators and grayscale images.

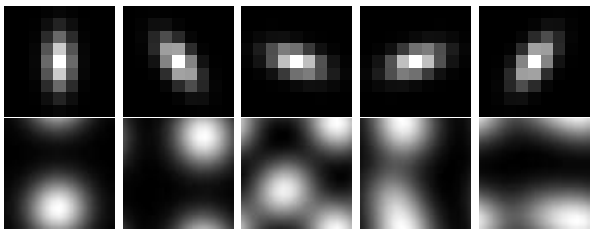


Figure 4.4: Examples of factors. Top: the full collection of $(\alpha_k)_{k\in K}$. Bottom: the factors $(\beta_{k,1})_{k\in K}$.

The setting We let $\mathcal{B}_n = \mathcal{B}_m = \mathbb{R}^{n\times n}$ with $n = 64$ and set $\mathcal{A} = \mathbb{R}^{n\times n}$ and $\mathcal{B} = \mathbb{R}^{n\times n}$. We generate random space varying impulse responses of the form

$$S_l = \sum_{k=1}^{|K|} \alpha_k \otimes \beta_{k,l}.$$

In the following, we set $|K| = 5$, let $\theta_k = \frac{\pi k}{6}$ and set for all $k \in K$ and all $(x_1, x_2) \in \{1, \dots, n\}^2$

$$\alpha_k(x_1, x_2) = \exp\Big(- \frac{(\cos(\theta_k)x_1 - \sin(\theta_k)x_2)^2}{8} - \frac{(\sin(\theta_k)x_1 + \cos(\theta_k)x_2)^2}{2}\Big).$$

This corresponds to anisotropic Gaussian functions rotated differently. We generate the maps $\beta_{k,l}$ as follows. For each $l \in L$:

1. We generate a matrix of $\mathbb{R}^{n\times n}$ where each element is a uniform random number in $[0, 1]$, independent of the others.

2. We compute a discrete convolution of this random matrix with an isotropic Gaussian blur. We then rescale it in $[0, 1]$, producing a discrete random field $f_l \in [0, 1]^{n\times n}$.

3. We then partition the domain Ω into $|K|$ sets $(\omega_{k,l})_{k\in K}$ defined as

$$\omega_{k,l} = f_l^{-1}\left([(k-1)/|K|, k/|K|]\right).$$

4. Finally, the factors $\beta_{k,l}$ are defined as the indicators of $\omega_{k,l}$ convolved with a Gaussian kernel.

We display the matrices elements α_k and $\beta_{k,1}$ in Figure 4.4. The elements α_k and $\beta_{k,1}$ are defined as matrices of size $n \times n$. To apply the results of previous sections, we simply consider vectorized version of these matrix (either with column first of row first convention). A summary of the dimension involved is display in Figure 4.1.

The output of our algorithm With those definitions, we get a list of random product-convolution operators H_l defined by

$$H_l u = \sum_{k\in K} \alpha_k \star (\beta_{k,l} \odot u), \forall u \in \mathbb{R}^{n\times n}.$$

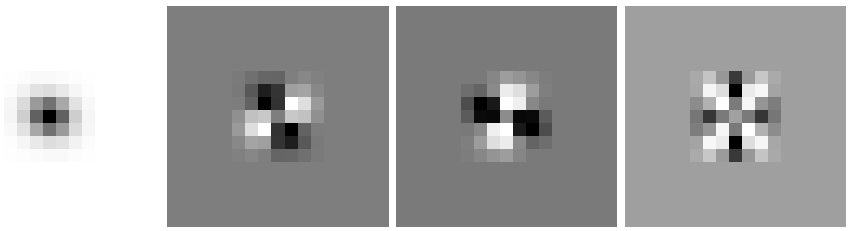


Figure 4.5: Learned family $(e_i)_{i\in I}$.

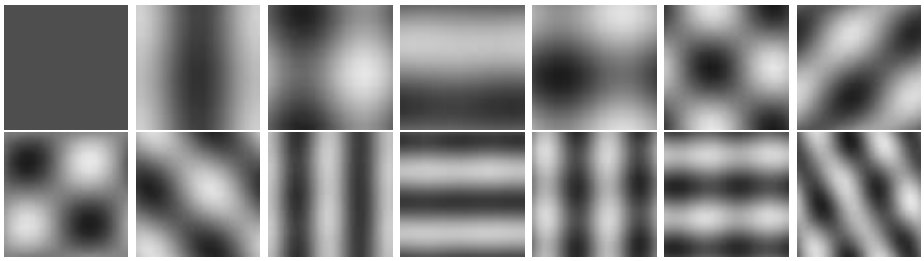


Figure 4.6: Learned family $(f_j)_{j\in J}$.

From the collection of $(S_l)_{l\in L}$ we can use the initialization of Algorithm 7 to estimate a subspace $\widehat{\mathcal{H}}_{I,J}$. In this paragraph we index the estimator by I and J.

The families $(e_i)_{i\in I}$ and $(f_j)_{j\in J}$ produced by the initialization of Algorithm 7 are displayed in Figure 4.5 and 4.6. The family $(e_i)_{i\in I}$ is an orthogonalization of the family α_k. The family $(f_j)_{j\in J}$ is quite similar to the first elements of a Fourier basis. This is to be expected since the functions $\beta_{k,l}$ are smooth and Fourier bases optimally encode smooth function spaces, see e.g. [Pin85].

A blind-deblurring experiment Using the notation of the introduction, we set $|S| = |I||J|$ and let $P_s = e_i \otimes f_j$ for $s = (i, j)$ denote the elementary operators constituting the subspace $\widehat{\mathcal{H}} = \text{span}(P_1, \ldots, P_{|S|})$. We let $Q \in \mathbb{R}^{m\times|T|}$ denote a matrix with columns $(q_t)_{t\in T}$ corresponding to elements of the discrete Haar wavelet basis with $|T| = 274$. We let

$$\mathcal{Q} = \{Q\beta, \beta \in \mathbb{R}^{|T|}\}$$

denote the subspace containing the images of interest. We let $\beta_0 \in \mathbb{R}^{|T|}$ denote the coefficients of the true image in the subspace $\mathcal{Q}$, and $H_0 \in \mathcal{H}$ the true operator that we want to recover. Finally we let

$$u_0 = H_0 v_0 + \eta,$$

where η is an additive white Gaussian noise. We display the true image v_0 in Figure 4.7a and the blurry-noisy image u_0 in Figure 4.7b.

We wish to solve the following bilinear inverse problem

$$\min_{v\in\mathcal{Q}, H\in\widehat{\mathcal{H}}} \|Hv - u_0\|_2^2. \tag{4.16}$$

Let $q \in \mathbb{R}^{|T|}$ and $p \in \mathbb{R}^{|S|}$, the mapping $(p, q) \mapsto [P_1, \ldots, P_{|S|}]pQq$ is bilinear. Then, there exists a linear mapping $\mathcal{W} : \mathbb{R}^{|T|\times|S|}$ such that $\mathcal{W}(pq^t) =$

$[P_1, \ldots, P_{|S|}]pQq$, where q^t denote the transposition of the vector q. Using the lifting and convex relaxation techniques described in [ARR14] leads to

$$\min_{Z \in \mathbb{R}^{|T| \times |S|}} \|Z\|_\star + \frac{\lambda}{2} \|\mathcal{W}(Z) - u_0\|_2^2 \tag{4.17}$$

where $\lambda > 0$ is a regularization parameter. For a matrix Z, $\|Z\|_\star$ denotes the nuclear norm, i.e. the sum of the singular values of Z. This convex program is solved using an accelerated proximal gradient method.

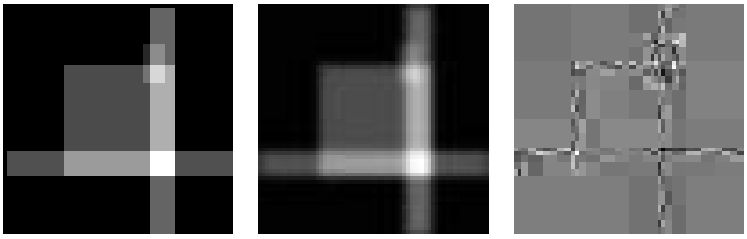
We let the algorithm run until the cost function stops decreasing. In Figure 4.7, we compare the reconstructed images with three different estimations $\widehat{\mathcal{H}}$ of $\mathcal{H}$:

- When $\widehat{\mathcal{H}} = \mathcal{H}$, we use the full subspace to solve (4.17), this yields the result in Figure 4.7c. It takes 390 seconds to solve the problem and we obtain a SNR of -3.0dB. The reason for this failure is that the dimension of the subspace is too large, making it impossible to identify the true image.
- When $\widehat{\mathcal{H}} = \widehat{\mathcal{H}}_{I,J}$ with $|I| = 5$ and $|J| = 30$, i.e we use subspace of dimension 150 to solve (4.17), we obtain the result in Figure 4.7d. The computing times are divided by three (120 seconds) compared to the full subspace $\widehat{\mathcal{H}} = \mathcal{H}$. More importantly, the method succeeds to recover the sharp image and we obtain a SNR of 26.7dB.
- When $\widehat{\mathcal{H}} = \widehat{\mathcal{H}}_{I,J}$ with $|I| = 5$ and $|J| = 8$, we subspace of dimension 40 to solve (4.17), leading to the results in Figure 4.7e. This time, the computing times decay to 31 seconds, which is 12 times faster than the case $\widehat{\mathcal{H}} = \mathcal{H}$. We also obtain a good result with a SNR of 26.2dB.

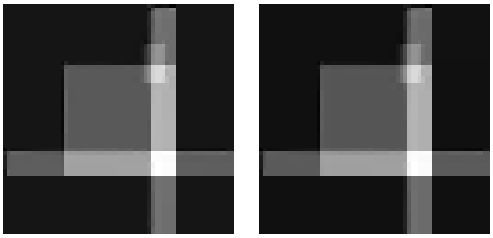
4.7 Conclusion

We presented a scalable approach to estimate a low dimensional subspace of linear operators from a sampling set of operators expressed as low rank tensors. This formalism covers several commonly used operator structures such as hierarchical matrices, wavelet expansions or product-convolutions. An important application lies in the field of blind inverse problems where the prior knowledge of a low dimensional subspace can make it possible to identify both the signal and the operator.

The principle outlined in this work changes the way a device is calibrated: instead of characterizing it through a single operator, we propose to describe all its potential states. For instance, in microscopy, variations of temperature change the refraction indexes and hence the associated measurement operators. With the proposed approach, we can capture these variations and hence use more precise models for image reconstruction. We hypothesize that the proposed formalism can improve reconstructions for many other practical problems.



(a) Original. (b) Blurry-noisy. (c) $\widehat{\mathcal{H}} = \mathcal{H}$, 390 seconds.



(d) $\widehat{\mathcal{H}} = \widehat{\mathcal{H}}_{5,30}$,120 seconds. (e) $\widehat{\mathcal{H}} = \widehat{\mathcal{H}}_{5,8}$, 31 seconds.

Figure 4.7: Blind deblurring experiment with different estimated subspaces. 4.7c: full dimensional search space. 4.7d: $|I| = 5$ and $|J| = 30$. 4.7e: $|I| = 5$ and $|J| = 8$. Reducing the search space makes the problem identifiable and reduces the computing times.

Chapter 5

Estimating a subspace of space-varying blur operators from microbeads image

Résumé : *Nous proposons une procédure robuste et efficace numériquement pour calibrer les microscopes à fluorescence à partir d'images de microbilles. Les algorithmes utilisées présentent de nombreux aspects originaux. Tout d'abord, ils permettent d'estimer des flous variants spatialement, ce qui est une caractéristique essentielle pour les grands champs de vue. Deuxièmement, nous proposons une approche pour l'étalonnage : au lieu de décrire un système optique par un seul opérateur, nous suggérons de faire varier les conditions d'imagerie (température, plan focal, éléments actifs) pour obtenir des images indirectes de ses différents états. Nos algorithmes permettent alors de représenter les réponses du microscope comme un ensemble d'opérateurs convexe et de faible dimension. Nous montrons sur un microscope champ-large que cette nouvelle approche améliore considérablement l'estimation. Cette approche est une étape essentielle vers la résolution efficace des problèmes inverses aveugles. Dans le Chapitre 7, nous illustrons le potentiel de l'approche en concevant une procédure originale pour le défloutage aveugle des images de sources ponctuelles et montrons une amélioration massive par rapport aux logiciels commerciaux.*

Abstract: *We propose accurate and computationally efficient procedures to calibrate fluorescence microscopes from micro-beads images. The designed algorithms present many original features. First, they allow to estimate space-varying blurs, which is a critical feature for large fields of views. Second, we propose a novel approach for calibration: instead of describing an optical system through a single operator, we suggest to vary the imaging conditions (temperature, focus, active elements) to get indirect observations of its different states. Our algorithms then allow to represent the microscope responses as a low-dimensional convex set of operators. This approach is deemed as an essential step towards the effective resolution of blind inverse problems. In chapter 7, we illustrate the potential of the approach by designing an original procedure for blind image deblurring of point sources and show a massive improvement compared to*

commercial software.

This chapter is based on the publication [Deb+20a]:
Debarnot, V., Escande, P., Mangeat, T., & Weiss, P. (2020). Learning low-dimensional models of microscopes, IEEE Transactions on Computational Imaging.

Contents

5.1 Introduction **98**
5.2 Operator estimation **100**
5.2.1 Notation 100
5.2.2 Preliminaries 100
5.2.3 Estimating a single operator 102
5.2.4 Estimating a subspace of operators 103
5.2.5 Implementation details 105
5.3 Results **106**
5.3.1 Data-sets 106
5.3.2 Estimating operators 107
5.4 Discussion **111**
5.5 Conclusion **112**
5.6 Appendices **113**
5.6.1 PSFs selection and processing 113
5.6.2 Estimating space variations 113

5.1 Introduction

Many recent breakthroughs in optics pertain to the field of computational microscopy: computers play a critical role to generate images. This evolution allowed to observe objects with unprecedented contrasts, temporal/spatial resolutions or gave access to new quantitative features. To name a few examples, let us mention Single Molecule Localization Microscopy (SMLM), Structured Illumination Microscopy (SIM), Total Internal Reflection Fluorescence microscopy (TIRF) or Stimulated Emission Depletion (STED) microscopy.

A common prerequisite for these techniques is the design of an accurate mathematical model of the optical system. This step is critical since the generation of images usually relies on an explicit or implicit inversion of this model. The advent of these microscopes therefore makes it more and more important to finely characterize their transfer function.

By far, the dominant models in the image processing literature are space invariant systems: the point spread function (PSF) is identical wherever in space. While simplifying the theoretical, numerical and experimental aspects, this assumption is however often unrealistic. Following [BW13], the space invariance is approximately valid only under very restrictive assumptions. There is clear theoretical and experimental evidence that the variations of the PSF need to be taken into acount along the optical axis [PC04; SF07] and the lateral axis for large numerical apertures (see for instance Fig. 5.3a). Neglecting this aspect can

have dramatic consequences. For instance, it was shown in [Zhe+13; Die+15] that this approximation can severely damage the reconstruction of images in single molecule localization imaging, with localization errors of more than 20% for a displacement of less than 200nm. The effects would be even more stringent for large fields of views which are a current challenge with the improved quality of sCMOS detectors. These model mismatches can significantly downgrade the performance of all computational microscopy systems and it is hence critical to finely estimate the optical response of the system.

Existing works A well spread approach to describe the response of an imaging system consists in using Fourier optics [Goo05]. This theory provides a nice description of the system through the pupil function of the objective. In this domain, it is possible to derive mathematical models of space variant PSFs [Die+15; Ari+18; Yan+19] and to infer the parameters of these models (e.g. Zernike coefficients) from experimental data. There are however two limitations to these approaches. First, they are often based on parameters such as the numerical aperture, the wavelengths, or many other physical quantities. In particular, the models become significantly more complex for large numerical apertures [BW13; GL89]. The more parameters, the more precise the model, but the harder it becomes to finely characterize them experimentally. In addition, some active components such as micro-mirrors introduce additional perturbations which cannot be easily modeled or inferred. The second problem comes from numerical considerations: the dependency between the model and its parameters is non-linear, which inevitably leads to non-convex estimation procedures, leading to local minima and additional inaccuracies.

Instead of relying on a physical model it is possible to directly estimate the optical response from the data. We will follow this approach in this chapter. Imaging fluorescent micro-beads in a cover-slide in 2D or a cylinder of agarose in 3D gives a partial idea of the system by providing an access to a few sampled and noisy scattered PSFs. This information can be used to estimate a space invariant system by averaging multiple micro-beads [BA96; Ber13; Li+18c]. When the images are aliased it is even possible to obtain a super-resolved estimation [EH01; Mbo+15]. It becomes more delicate to estimate a space variant system. A few researchers - especially in the field of astrophysics [1] - have addressed this issue. The general framework is the following: a parameterized PSF model is designed either from physics equations or from the data itself. The observed PSFs are then interpolated to cover the whole field of view. In [GCM13; Cha+12], the authors propose to decompose the PSFs over a low dimensional basis and to interpolate the coefficients using thin-plate splines. A subset of the authors proved that this method was minimax optimal in [BEW19a] and we will propose a refined version in this chapter. It is also possible to use more advanced interpolation methods, using matrix factorization techniques [Ngo+16], which share similarities with the proposed approach or optimal transport [Mbo+15], but this method would not scale computationally to the large field of views considered in here.

Our contribution While there now exists a solid theoretical and algorithmic framework to estimate space varying optical responses of optical systems, these

[1] In astrophysics, the PSFs variations may be due to weak gravitational lensing and reveal distant massive galaxies.

methods are often tested on synthetic data that do not reproduce all the complexity of real microscopy images. For instance, the estimation of a PSF requires a very careful treatment of the background and of the noise. Its interpolation requires specific care to avoid obtaining unrealistic results far from the observed responses. Our first objective is to provide *precise estimation algorithms adapted to real data emanating from fluorescence microscopy.* In particular, we propose an algorithm to promote realistic PSFs encoding properties such as nonnegativity or smoothness properties by estimating a conical hull of projection coefficients. This approach shares similarities with [Ngo+16], but differs in that the features of realistic PSFs are directly learned from the data rather than defined as priors. The second and arguably most original contribution is a *new way to calibrate an optical system by learning all its possible states.* Instead of trying to estimate a single operator to describe the microscope, we propose to learn a whole family of possible states by varying the experimental conditions, following the recent theoretical work [DEW19]. In Chapter 7, we illustrate how the proposed approach can help to solve a blind deblurring problem in microscopy.

5.2 Operator estimation

5.2.1 Notation

In all this chapter, bold fonts refer to vectors, matrices or vectorial functions while regular fonts refer to scalar numbers or functions. The i-th value of a vector $\mathbf{x}$ is denoted either x_i or $\mathbf{x}[i]$. The ℓ^p norm of a vector $\mathbf{x}$ is denoted $\|\mathbf{x}\|_p$. The value of a function f is $f(x)$ and its ℓ^p-norm is denoted $\|f\|_p$. The delta Dirac function at a position $\mathbf{x} \in \mathbb{R}^d$ is denoted $\delta_{\mathbf{x}}$. In all the chapter, the integers I, J, K, M and N refer to a number of components described in Table 5.1. A family of vectors $(\mathbf{x}_i)_{1 \leq i \leq I}$ is said to be orthogonal if we have

$$\langle \mathbf{x}_i, \mathbf{x}_j \rangle = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise} \end{cases}.$$

Table 5.1: Notation.

Symbol	Meaning
I	size of PSF basis
J	size of space variations basis
K	number of observed microbeads images
M	number of observed microbeads
N	number of pixels of an image

5.2.2 Preliminaries

A space varying blurring operator $H : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ can be seen as a linear integral operator mapping an image u to the degraded image Hu through the

formula

$$Hu(\mathbf{x}) = \int_{\mathbb{R}^d} L(\mathbf{x}, \mathbf{y}) u(\mathbf{y})\, d\mathbf{y}. \tag{5.1}$$

The function $L(\,\cdot\,,\,\cdot\,)$ is called *kernel* of the operator. It describes the impulse response of the system at every location $\mathbf{z} \in \mathbb{R}^d$ of the image domain since:

$$(H\delta_{\mathbf{z}})(\,\cdot\,) = L(\,\cdot\,, \mathbf{z}). \tag{5.2}$$

The PSF $S(\,\cdot\,, \mathbf{z})$ of the system at $\mathbf{z}$ is defined as the impulse response centered at 0, i.e.

$$S(\,\cdot\,, \mathbf{z}) = L(\,\cdot\, - \mathbf{z}, \mathbf{z}). \tag{5.3}$$

The function S is called the space varying impulse response of the system. This work is based on two important assumptions.

Assumption 5.2.1 (PSF approximation)**.** *Every PSF in the field of view is well approximated by its projection over a low-dimensional orthogonal basis* $(h_i)_{1 \le i \le I}$*, i.e.*

$$S(\,\cdot\,, \mathbf{z}) \simeq \sum_{i=1}^{I} \langle S(\,\cdot\,, \mathbf{z}), h_i \rangle h_i. \tag{5.4}$$

This assumption is valid both from a theoretical and an empirical viewpoint. It is indeed well known that any smooth function can be well approximated by its projection on a low dimensional subspace. Typical bases include splines or low frequency Fourier atoms [Pin85]. In practice, we can also construct the family (h_i) by computing the principal component analysis of a family of sampled PSFs. The numerical experiments performed in this manuscript reveal that for our imaging systems, as little as $I = 5$ elements are enough to capture all possible PSFs accurately. In addition, restricting the PSFs to live on a low dimensional subspace is an efficient method to denoise them as will be illustrated in the numerical section.

Assumption 5.2.2 (PSF variations)**.** *Letting*

$$\alpha_i(\mathbf{x}) := \langle S(\,\cdot\,, \mathbf{x}), h_i \rangle \tag{5.5}$$

denote the i-th coefficient of the PSF at $\mathbf{x} \in \mathbb{R}^d$*, we assume that* α_i *varies slowly in space.*

This hypothesis means that the PSFs vary smoothly in space. It can be substantiated experimentally using arrays of micro-beads for instance, see e.g. [Zhe+13; Die+15; Hei+13].

Under Assumptions 5.2.1 and 5.2.2, a space varying operator is completely characterized by the pair $(\alpha_i, h_i)_{1 \le i \le I}$. Of interest, this representation of the operator also leads to fast numerical computations using a structure called *product-convolution*. This decomposition has been developed and improved for the last two decades [PC04; PP15; Ari+10; Den+15; FR05a; NO98]. Its precise approximation rates have been studied in [EW17] and we refer to the previous references for more insight on these structures. A key property of this decomposition is the following.

Proposition 5.2.1 (Product-convolution [EW17; Den+15; FR05a]). *Assume that a blurring operator H has a space varying impulse response S defined by the tensor:*

$$S(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{I} \alpha_i(\mathbf{x}) h_i(\mathbf{y}),$$

then for any image u, we have

$$Hu = \sum_{i=1}^{I} h_i \star (\alpha_i u),$$

where the symbol $\star$ stands for the convolution operator.

Hence, the numerical complexity of computing a space-varying operator is just I times the one of a convolution, which can be achieved efficiently using Fast Fourier Transforms for instance in a way independent of the PSF size.

5.2.3 Estimating a single operator

Under Assumptions 5.2.1 and 5.2.2, the problem of estimating the operator reduces to recovering the low dimensional bases (h_i) and (α_i), or at least their discretized counterparts $(\mathbf{h}_i)$ and $(\boldsymbol{\alpha}_i)$. In this paragraph, we describe the general principle of the estimation of a blurring operator from a single image of fluorescent micro-beads. The general process is described in Fig. 5.1.

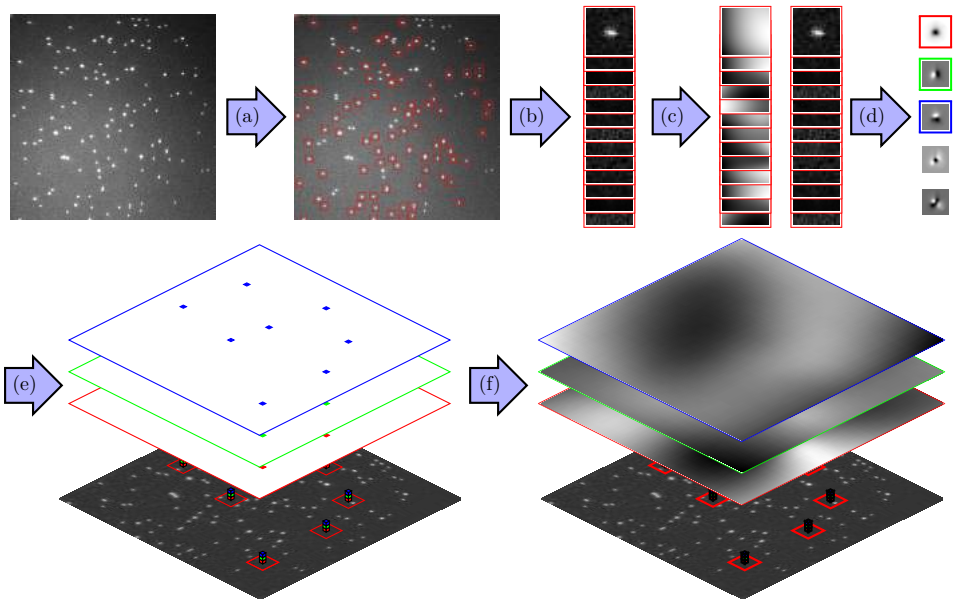


Figure 5.1: Structure of the algorithm for single operator estimation. (a) Background removal procedure. See Chapter 3.2.2. (b) Selection of well isolated PSFs. See Chapter 3.2.2. (c) Extraction of relevant PSF patches. (d) Principal Component Analysis of the PSFs to find a low dimensional basis. (e) Projection of each selected PSF on the low-dimensional basis. (f) Interpolation of the PSFs coefficients using radial basis functions and correction to ensure admissible PSFs.

The first part of the pipeline aims to estimate a family that approximate well the PSFs from the data. In particular, the automatic detection of beads are presented in Algorithm 2 in Chapter 3.2.2. We briefly review the operations hereafter, but the technicalities can be found in Chapter 3.2.2.

The first step consists in extracting the most relevant PSFs in the form of small patches (see Chapter 3.2.2). Then the background is estimated and removed on each patch independently to avoid biases in the PSFs estimation. A principal component analysis is then performed to estimate the basis $(\mathbf{h}_i)$ (see Section 5.6.1). Each PSF is projected on this basis and the resulting coefficients are interpolated spatially to provide an estimate of the functions (α_i), which can then be discretized as $(\boldsymbol{\alpha}_i)$ (see Section 5.6.2). All those steps are subtle and need to be performed carefully to obtain precise estimates. The technical details reported in the appendices are therefore of great importance. While revising the manuscript, we realized that the astronomical software [BA96; Ber13] was actually proposing many similar ideas for a different purpose.

As an output of the algorithm, the pair $(\mathbf{h}_i, \boldsymbol{\alpha}_i)_{1\leq i\leq I}$ provides a complete description of the operator, since we know an approximation of the PSFs at each image location. The integer I is a user provided parameter.

5.2.4 Estimating a subspace of operators

Motivation A microscope produces different transfer functions depending on physical parameters that can be hard to control. Typical examples include temperature variations, focal screws, small tilts of optical elements, surface flatness of cover-slides, slight variations of a spatial light modulator rest state,... In those conditions, capturing a single operator (as proposed in the previous section) to describe the microscope might lead to model mismatches and reconstruction errors. In this section, we propose an alternative approach where we aim at learning a *family of realistic operators* that capture all the possible states of a microscope. The principle and the mathematical foundations behind this approach (statistical properties and fields of application) were recently established by a subset of the authors in [DEW19]. We refer the interested reader to this paper for more details. We provide a simplified description below.

Principle The first requirement to apply this technique is to image stacks of fluorescent micro-beads (in 2D or 3D) under multiple conditions. This process can be automatized when using advanced optical tables with motorized stages and thermostatic chambers. An alternative is to probe only the "extreme" conditions (e.g. highest and lowest realistic temperatures and tilts). After this experimental process is achieved, we have access to a set of images $(\mathbf{u}_k)_{1\leq k\leq K}$. The idea of our estimation procedure is to apply the following procedure:

1. For each image $\mathbf{u}_k$, extract the most relevant PSF patches and remove the background (see Chapter 3.2.2).

2. Apply a principal component analysis to the set of patches over multiple images and keep I principal components.

3. Apply a z-score test to discard the patches that are likely outliers (e.g. multiple PSFs in a patch).

4. Reapply a principal component analysis to better estimate the principal components.

5. For each image $\mathbf{u}_k$ and each coefficient i, interpolate the coefficient maps $\alpha_{i,k}$ (see Section 5.6.2). This interpolation process is subtle: in particular we provide a novel method to learn features such as nonnegativity, or the natural decay of coefficients on the PSF basis.

6. Apply a randomized principal component analysis [HMT11] to the whole set of sampled interpolation maps $(\boldsymbol{\alpha}_{i,k})_{1\leq i\leq I, 1\leq k\leq K}$. It is often necessary to apply a randomized SVD [2] here since the interpolation maps $\boldsymbol{\alpha}_{i,k}$ are typically large images.

7. Keep the J largest principal components $(\mathbf{a}_j)_{1\leq j\leq J}$.

8. Project each interpolation map $\boldsymbol{\alpha}_{i,k}$ onto the basis $(\mathbf{a}_j)_{1\leq j\leq J}$, to obtain the matrices $\boldsymbol{\Gamma}_k \in \mathbb{R}^{I\times J}$ defined by

$$\boldsymbol{\Gamma}_k[i,j] = \langle \boldsymbol{\alpha}_{i,k}, \mathbf{a}_j \rangle.$$

The output of this process is two orthogonal bases $(\mathbf{h}_i)_{1\leq i\leq I}$ (which describe the PSFs compactly) and $(\mathbf{a}_j)_{1\leq j\leq J}$ (which describe the PSFs variations compactly) as well as a set of matrices $(\boldsymbol{\Gamma}_k)_{1\leq k\leq K}$ in $\mathbb{R}^{I\times J}$ (which describe the operators associated to each image $\mathbf{u}_k$). The operator $\mathbf{H}_k$ associated to the k-th input image $\mathbf{u}_k$ is then defined for all $\mathbf{u}$ by:

$$\mathbf{H}_k\mathbf{u} = \sum_{1\leq i\leq I}\sum_{1\leq j\leq J} \boldsymbol{\Gamma}_k[i,j]\mathbf{h}_i \star (\mathbf{a}_j \odot \mathbf{u}).$$

Reducing the family of admissible operators The subspace of operators $\mathcal{H}$ that compactly describes the possible operators is defined by

$$\mathcal{H} \stackrel{\text{def.}}{=} \operatorname{span}(\mathbf{L}_{i,j}, 1\leq i\leq I, 1\leq j\leq J)$$

where $\mathbf{L}_{i,j}\mathbf{u} \stackrel{\text{def.}}{=} \mathbf{h}_i \star (\mathbf{a}_j \odot \mathbf{u})$ is a simple product-convolution operator. The dimension of $\mathcal{H}$ is $I \times J$.

However, all operators in the subspace $\mathcal{H}$ are not plausible. For instance, all PSFs are nonnegative, which is often a critical feature to avoid ringing artifacts. It is possible to further restrict the family of operators as follows. Assuming that all the extreme conditions have been explored, we can construct the *convex hull* of the coefficients $\boldsymbol{\Gamma}_k$:

$$\begin{aligned}\mathcal{C} &\stackrel{\text{def.}}{=} \operatorname{conv}(\boldsymbol{\Gamma}_k, 1\leq k\leq K)\\ &\stackrel{\text{def.}}{=} \left\{\sum_{k=1}^{K}\lambda_k\boldsymbol{\Gamma}_k, \lambda_k\geq 0, \sum_{k=1}^{K}\lambda_k = 1\right\}.\end{aligned}$$

[2] A standard way to compute principal components requires computing a singular value decomposition. To retrieve the k first components, for an $m\times n$ matrix, the complexity is $O(mnk)$, which is intractable for large scale computations. On its side, the randomized SVD provides a certified approximate solution with a complexity in $O(\log(k)mn)$ and requires significantly less memory.

The quality of the estimate $\mathcal{C}$ with respect to the number of observations K was studied in [DEW19]. If all the sampled PSFs are nonnegative, then any conical combination is nonnegative too and imposing the matrix $\Gamma \in \mathcal{C}$ will therefore preserve this property. Another important feature can be preserved: the coefficients $\mathbf{\Gamma}_k[i,j]$ follow a distribution that decays in average with i and j, since they correspond to eigenvectors of decreasing importance. The set $\mathcal{C}$ also captures this property, resulting in more realistic PSFs.

5.2.5 Implementation details

Normalizing the operators Of importance, let us mention that the procedure described previously suffers from a well known identifiability issue. Since the micro-beads intensity is usually unknown, the operator can be estimated only up to multiplicative constant. To address this problem, it is possible to replace the matrices $\mathbf{\Gamma}_k$ by a normalized version

$$\bar{\mathbf{\Gamma}}_k = \mathbf{\Gamma}_k / \sum_{i,j} \mathbf{\Gamma}_k[i,j],$$

and to replace the convex hull by the conic hull

$$\mathcal{C} = \mathrm{cone}(\mathbf{\Gamma}_k, 1 \leq k \leq K) \stackrel{\text{def.}}{=} \left\{ \sum_{k=1}^{K} \lambda_k \mathbf{\Gamma}_k, \lambda_k \geq 0 \right\}.$$

Normalizing the PSFs The PSFs need to be normalized in different ways. First, they need to be registered at a subpixel accuracy. To do so, we propose a method described in Chapter 3.2.2 that amounts to assuming that their center of mass is located at the origin.

Second, it may also happen that the micro-beads are not perfectly identical and have different fluorescence levels. In that case it is important to normalize the PSF patches (after background removal) by imposing that they sum to 1. By doing so, the operators will be estimated without accounting for the variations of intensity that they may induce due to non uniform illumination. This effect can still be captured by assuming that the loss of intensity is proportional to the background. It then suffices to multiply the normalized patches by the background estimate.

We also normalize the ℓ^2-norm of the PSFs prior to computing the principal component analysis, in order to give the same importance to every PSF.

Selecting the subspace sizes The subspace sizes I and J are the two values that a user needs to provide in order to estimate the subspace. If the number I (related to the subspace of PSFs) is too small, then the PSFs will be badly reproduced, while a value that is too large will result in noisy operators (the so-called over-fitting in machine learning). Similarly, the number J captures the variations of the PSFs and has to be chosen with caution. Finally, we would like I and J to be as small as possible to reduce the computing times: the cost of applying a product-convolution operator is directly proportional to I.

The simplest way to choose I is to test different values on a subset of representative PSFs and keep the lowest value that leads to a visually decent reconstruction of the PSFs. The same can be done with J. In practice, we

observed that the values $I = 5$ and $J = 5$ faithfully reproduce the operators from a perceptual point of view.

Another possibility is to apply recent results in statistics [GD14] that provide a simple and optimal way (under a Gaussian noise assumption) to choose I and J. The rule consists in keeping the principal components associated to a singular value larger than $2.858 \cdot \sigma_{med}$, where σ_{med} is the median of the set of singular values. This procedure requires computing the set of all singular values to evaluate the median. In practice, this is possible only for the PSF patches which are low dimensional.

2D vs 3D PSF models All the proposed algorithms are implemented in 2D, but their extension to 3D is straightforward. From a practical point of view, the estimation of a 3D operator requires to image uniformly scattered microbeads in a medium such as an agarose gel. We do not report experimental results for this problem in this chapter.

5.3 Results

In this section, we test the proposed algorithms against 2 different data-sets: the first one is simulated while the other comes from a wide-field microscope. We start with the estimation of a single operator and of a subspace of operators.

5.3.1 Data-sets

Simulation

We generate several product-convolution operators by designing a collection of admissible PSFs and space variations. The collection of PSFs is obtained by taking all the slices of a 3D astigmatic PSF $h(x_1, x_2, z)$ where z denotes the variations in the optical axis direction. The expression of the 2D PSF at a distance z from the focal plane (used in [Ari+18] for instance) is given by:

$$h(\mathbf{x}, z) \propto \left| \int_D (\exp(-2i\pi(\langle \boldsymbol{\xi}, \mathbf{x} \rangle) E(\boldsymbol{\xi}, z))\, d\boldsymbol{\xi} \right|^2, \tag{5.6}$$

where $\mathbf{x} = (x_1, x_2)$, $\boldsymbol{\xi} = (\xi_1, \xi_2)$, D is a disk of radius NA/λ and E is the electric field at the pupil plane given by

$$E(\boldsymbol{\xi}, z) = \exp\left(2i\pi\left(\sum_i c_i Z_i(\boldsymbol{\xi}) + z\sqrt{\frac{n^2}{\lambda^2} - |\boldsymbol{\xi}|^2}\right)\right). \tag{5.7}$$

Here, Z_i denotes the i-th Zernike polynomial, NA is the numerical aperture, n is the refractive index of the immersion oil and λ is the emission wavelength. Note that we neglected the lateral displacements, which is a crude approximation for a large numerical apertures.

We then produce a collection of 3D PSFs by varying the parameters of the model. Each PSF is produced by taking random values of the parameters $NA \in [1.1, 1.4]$, $\lambda \in [550, 650]$, $n \in [1.41, 1.61]$. The parameters c_i have been fixed to reproduce an astigmatic PSF. In addition, we model spatial variations by varying the depth z in the lateral direction. We use random polynomials of low

degree to generate various depth profiles. An additive background is generated with a smooth Gaussian random process, and the final image is degraded with Poisson noise, see Fig. 5.2a for a simulation example.

Remark 5.3.1. *The simulation model is almost the same as in Chapter 3.3.1 and is based on the scalar diffraction theory presented in Chapter 2.4.*

Experimental data

In all experiments, we used a perfectly plane mono-layer of 100nm diameter micro-beads. There is no refractive index mismatch between the cover-slide and the immersion oil, allowing to avoid spherical aberrations. We used a wide-field fluorescence microscope with a $\times 100$ objective lens (CFI SR APO 100XH NA 1,49 DT 0,12 Nikon) mounted on a Nikon Eclipse Ti-E and a Hamamatsu sCMOS camera (ORCA FLASH4.0 LT). A lens with 1.5 magnification (this is an additional magnification available on Nikon Eclipse Ti-E) is used to obtain 43nm pixel size on the image plane. A 200nm Z interval was acquired on each image. We use SPECTRA X light engine with excitation of 633nm, and emission of 670nm. Micromanager software was used for the acquisition software. This produces images of 2304×2304 pixels. We collected 18 stacks of fluorescent micro-beads, each one is 8µm thick and is composed of 21 z-stacks. We keep only the 5 central slices since the beads are too degraded when far away from the focal plane. This amounts to a total of 90 images and more than 9700 2D PSFs. We display one image in Fig. 5.3a.

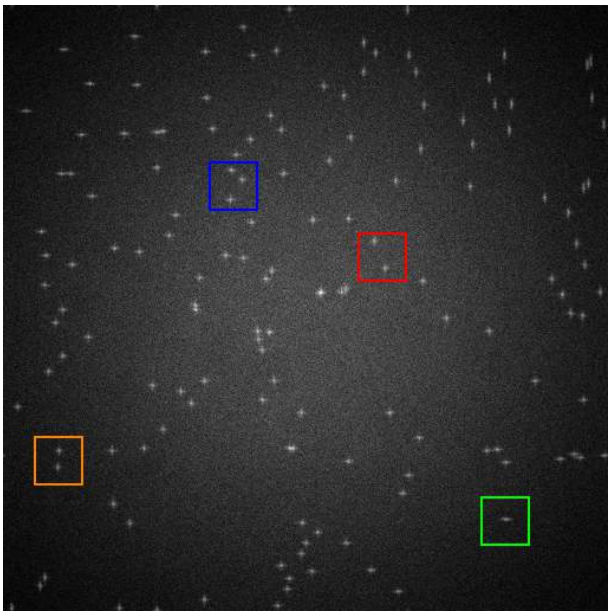
5.3.2 Estimating operators

In this section we illustrate some features of the proposed methods by estimating a single operator and a subspace of operators from the image of micro-beads generated by the previously described microscopes. Each experiment is performed on a workstation equipped with Intel Xeon E5 and a GPU card Nvidia Tesla K20c from 2012 (2019 technologies are expected to be 4 times faster).

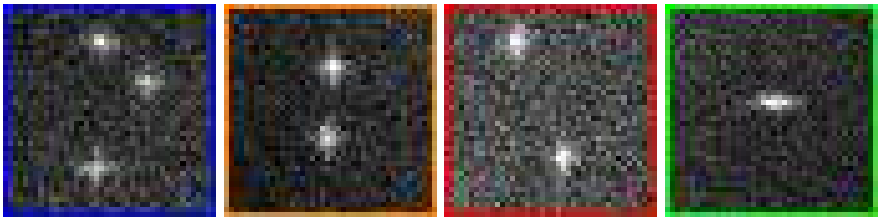
Simulation

We apply the proposed estimation procedures both for a single operator and for a subset of operators. The computing time for a single operator is 15 seconds when estimating 3 principal elements for the PSFs and 3 principal components to describe the coefficients variations. To estimate the subspace of simulated operators, we used 50 different micro-beads images. The computing times increased to 500 seconds (i.e. 10 second per image). The results are displayed in Fig. 5.4. Of importance, notice that the results obtained with the subspace of operators are based on micro-beads images generated with *different* operators and in particular different PSFs. Despite this higher variety of possible shapes, the method is able to automatically infer the common patterns and to achieve better denoising and estimation performance thanks to the redundancy.

The estimated operators can be visualized by applying them to a Dirac comb, see Fig. 5.2c and 5.2d. To compare the quality of reconstruction, we simply evaluate a rescaled version ℓ^2 distance between the resulting images and the



(a) Simulated micro-beads image.

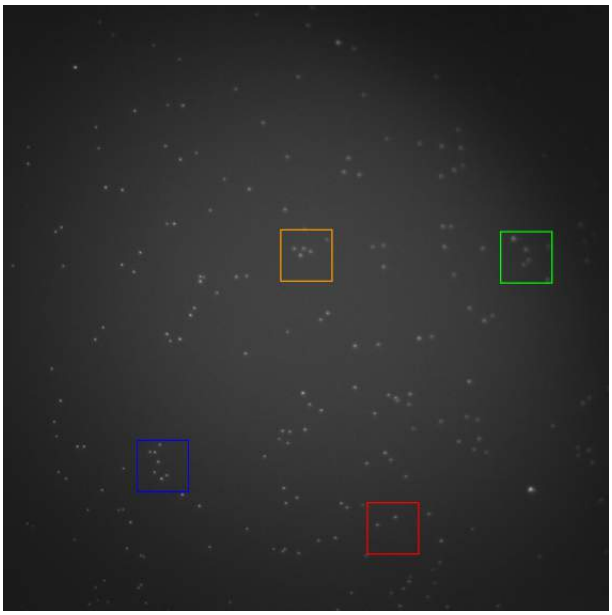


(b) Zooms on a few PSFs in Fig. 5.2a.

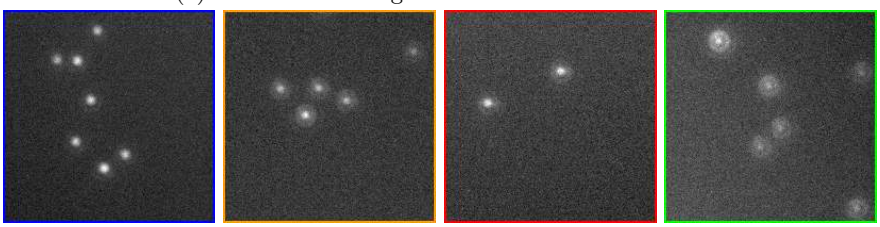
(c) Operator estimated from the image in Fig. 5.2a.

(d) Operator estimated from a family of 50 images.

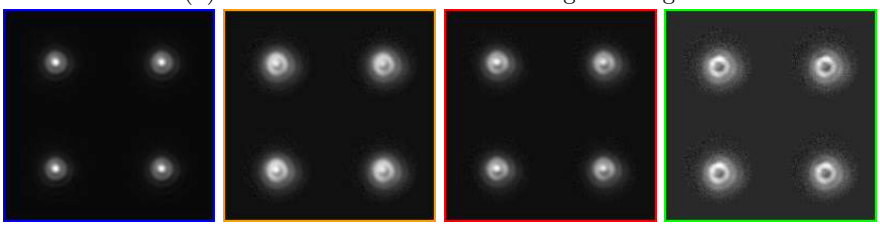
Figure 5.2: Simulation experiment: synthesized operators are applied to randomly scattered Dirac masses in Fig. 5.2a (512×512 pixels). We then test the estimation procedure using a single micro-bead image in Fig. 5.2c and from 50 images in Fig. 5.2d. Observe that the estimation from a family is far less noisy.



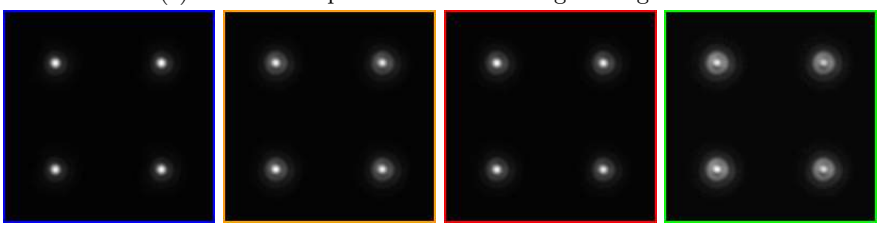
(a) 2304 × 2304 image of fluorescent micro-beads



(b) Zooms on a few PSFs of the original image.

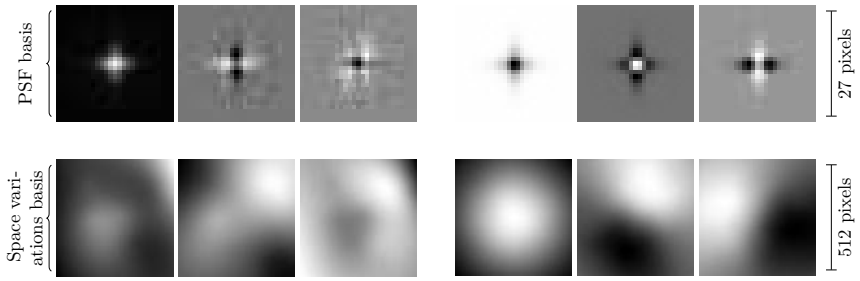


(c) Estimated operator from the image in Fig. 5.3a.



(d) Operator estimated from a family of 18 images.

Figure 5.3: Image of micro-beads taken with a wide field microscope and estimation results. The contrasts have been stretched for a better visualization. Similarly to the simulation example in Fig. 5.2, the operator seems far better reconstructed using the family of images.



(a) Single image estimation. (b) Subspace estimation

Figure 5.4: Estimation on simulated operators. We compare the estimation with a single operator (left) and with a set of 50 randomly sampled operators (right). Top: the PSF basis $(h_i)_{1\leq i\leq 3}$. Bottom: the space variations basis $(a_j)_{1\leq j\leq 3}$. Observe that the PSF basis is significantly less noisy and that the space variations are significantly smoother when estimating over the set of 50 images.

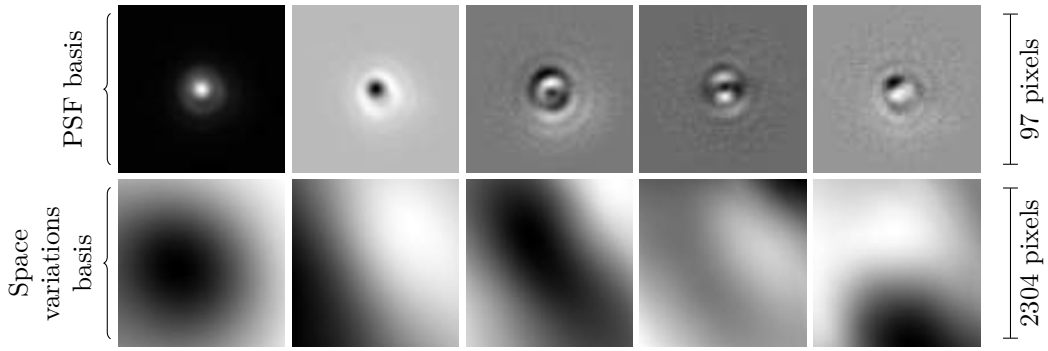
true one. It is compulsory to rescale the distance, since there is an inherent ambiguity between the micro-beads intensity and the microscope gain.

Estimating a subspace of operators rather than a single operator improves the quality of the reconstruction allowing to go from 50% to 15% of relative distance between the estimation and the ground truth. While this figure is not per se impressive, the PSFs family is no longer corrupted with noise, and the coefficients maps family seems smoother. It is likely that other metrics would better reflect this fact.

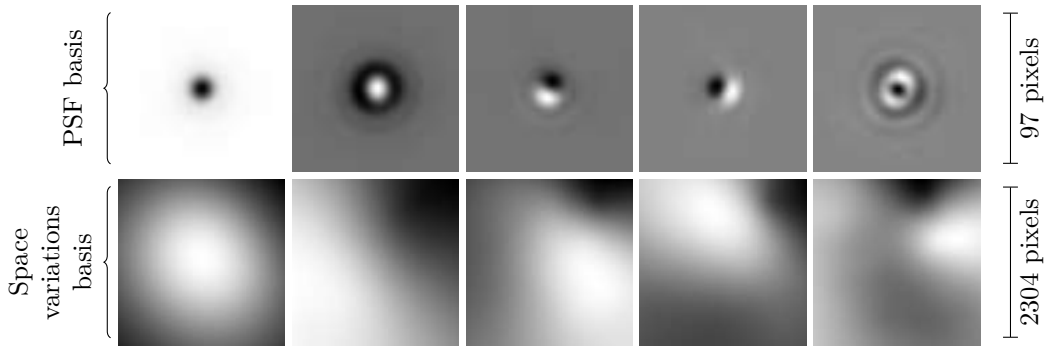
Wide-field microscopy

We estimate a single operator and a whole subspace of product-convolution operators based on the 2304×2304 images from the data-set from wide-field microscopy. We set $I = 5$ and $J = 5$. Estimating a single operator takes about 150 seconds using 120 PSFs, while estimating the whole subspace takes about 3 hours (i.e 2 minutes per operator) using 9700 PSFs. These computing times are remarkable given the computer features and that the complete dataset takes about 5Gb to store.

The PSFs and space variations bases are displayed in Fig. 5.5 and the estimated operators are displayed in Fig. 5.3. Similarly to the previous section, we observe that the basis and operators obtained using a large set of images is significantly less noisy. While the principal components beyond 3 contain a significant amount of noise for the single image, the 5-th component of the subspace approach still seems to contain useful geometrical features. The improvement of the coefficients maps is harder to evaluate since the corresponding convolution kernels have changed. Overall, learning the subspace of operators led to a significantly improved reconstruction of the operator with no visible residual noise remaining in Fig. 5.5. A large part of the improvement comes from the fact that more PSFs are observed and that the noise can be averaged out. The second reason is that the estimates of space variations are less sensitive to errors and turn out to be smoother.



(a) Operator estimation with a single image containing 94 micro-beads, see Fig. 5.3a. Top: the PSF basis $(h_i)_{1\leq i\leq 5}$. Bottom: the space variations basis $(a_j)_{1\leq j\leq 5}$.



(b) Subspace estimation with 90 images and 9700 PSFs.

Figure 5.5: Learning the PSF and space variations bases for a standard wide-field microscope.

5.4 Discussion

In this section, we discuss some limitations and possible extensions of the proposed approach.

Centering the PSFs This question has already been addressed in Chapter 3.

The detection of PSFs works by finding the maximum of correlation of a Gaussian with the images of micro-beads. The implicit assumption behind this procedure is that the PSFs have a center of mass located at the origin. Unfortunately, this hypothesis is wrong for some abberrations such as coma. In that case, the proposed method will result in operators that - in addition to blur - produce slight deformations of the image. Unfortunately, without prior assumption on the PSF center, it is impossible to resolve the ambiguity between the micro-bead position and the center of mass of the PSF. In general, we can therefore expect slight distortions of the images with the proposed approach.

Physicality of the PSFs The proposed methodology is able to reproduce some features of real point spread functions such as nonnegativity and natural decay of the coefficients in the PSF basis, thanks to the projection step on

the conical hull of observed operators. This feature is important and original. When looking at the PSFs inferred by our algorithms, see e.g. Fig. 5.3, we can however see that they are not entirely satisfactory. For instance, the dark rings that can be seen on real diffraction patterns are not reproduced accurately. At this stage we do not know whether it is possible to obtain them with purely data driven approach as the one proposed here since they are not visible on the acquired images, which suffer from numerous artifacts such as noise, quantization, sampling, and background addition.

2D versus 3D All the proposed algorithms and examples have been implemented in 2D, but their extension to 3D is straightforward. From a practical point of view, the estimation of a 3D operator requires to image scattered microbeads in a medium such as an agarose gel. This would allow to characterize the optical response of other types of microscopes such as confocal microscopes or light sheet fluorescence microscopes.

A limitation of the proposed approach for 2D microscopy is cases where the variations of the PSF in depth are important and the object is really 3D. In that case, the microscope response should be modelled as an operator mapping 3D functions to 2D images and we should infer 3D PSFs from 2D slices, which is significantly harder than what we did here. For instance, the method does not apply to 3D super-resolution microscopy.

If this is not the case, the method presented in the Chapter 3 can be used.

5.5 Conclusion

We proposed a set of fine algorithms to learn a set of product-convolution representations of optical responses in fluorescence microscopy. One of the main originality is to estimate a subspace of operators to capture the whole diversity of possible space-varying blurs of a given microscope. This is in sharp contrast with existing approaches which simply characterize the microscope by a single PSF, or - at best - by a single spatially variant operator.

An important outcome of this work is that it strongly improves the identifiability in blind-inverse problems such as blind deblurring or blind super resolution. These arguably constitute two of the most challenging issues in computational imaging. For instance, recent theoretical progresses based on lifting techniques [ARR13] require the prior knowledge of a low dimensional subspace. We illustrate this fact in Chapter 7 with an original blind deblurring approach coined BSS to efficiently solve this problem when imaging point sources with a smooth background.

Future works will consist in extending the existing codes to 3D images, providing an open-source toolbox together with realistic responses of microscopes. This will enable the optics and signal processing communities to test their algorithms against realistic operators.

We expect the proposed work to have far reaching applications ranging from the metrology of imaging systems to new advanced microscopy methods such as supercritical angle localization microscopy, metal enhanced fluorescence, polarization microscopy. All these applications require highly accurate models which are currently unavailable for large fields of view.

5.6 Appendices

5.6.1 PSFs selection and processing

We use results from Chapter 3.2.2 to automatically detect isolated beads. This initialization procedure is rather advanced, and we refer to Chapter 3 for a complete discussion on the matter. In a nutshell, we extract a patch that contains the observed PSF around each detected bead. We then perform a principal component analysis to obtain an optimal representation basis. Depending on the number of sampled PSFs, we can use a standard singular value decomposition or a randomized one [HMT11].

5.6.2 Estimating space variations

Thin-plate approximation Once a basis $(\mathbf{h}_i)_{1\leq i\leq I}$ is computed (see section 5.6.1), it is possible to project each noisy patch on this basis to get a low dimensional representation of the selected PSFs. This provides an estimate $\beta_{i,m} = \langle \mathbf{p}_m, \mathbf{h}_i \rangle$ of the values $\alpha_i(\mathbf{z}_m)$. In order to estimate the space variations, we can use surface fitting techniques on the set $(\mathbf{z}_m, \beta_{i,m})_{1\leq m\leq M}$ to get an approximation of the functions α_i.

There exist numerous surface fitting techniques. Following the numerical experiments led in [GCM13], it seems that the use of radial basis function [Buh03] is significantly more efficient than other approaches in the context of astronomy. We therefore resort to this technique.

Radial basis functions approximation can be interpreted as a variational problem in the framework of Reproducible Kernel Hilbert Spaces. In this context, the estimators $\hat{\alpha}_i$ of α_i can be expressed as

$$\hat{\alpha}_i = \underset{\alpha\in H^2(\mathbb{R}^2)}{\operatorname{argmin}} \frac{1}{2}\sum_{m=1}^{M} w_m|\alpha(\mathbf{z}_m) - \beta_{i,m}|^2 + \frac{\eta}{2}|\alpha|^2_{H^2}, \tag{5.8}$$

where $|\alpha|_{H^2} \overset{\text{def.}}{=} \langle \Delta\alpha, \Delta\alpha\rangle_{L^2(\mathbb{R}^d)}$ and where $\eta > 0$ is a parameter that allows to trade off the proximity to the samples $\beta_{i,m}$ for the smoothness of the surface. In order to balance the importance of each PSF in the approximation, the weights w_m are chosen equal to the area of the Voronoï cell associated to each location $\mathbf{z}_m$.

The solution of (5.8) is known to be a thin-plate spline [Pin85] and can be computed by solving a $(M+3)\times(M+3)$ linear system.

In what follows, we will let $\boldsymbol{\alpha}_i$ or $\hat{\boldsymbol{\alpha}}_i$ denote a version of α_i sampled on a Euclidean grid.

Enforcing realistic PSFs There is no reason for the thin-plate approximation method to generate realistic PSFs everywhere in the field of view. Indeed, the coefficients are interpolated *independently* of each other while there exists strong dependencies between them. In practice we observed that the previous method was not good at extrapolating the PSFs outside of the convex hull of the sampled PSFs. Important features like positivity for instance might be lost far away from the sampled PSFs. To avoid this effect, we propose an original framework below.

The procedure for estimating the PSF basis $(\mathbf{h}_i)$ also yields the projected coefficients $(\beta_{i,m})_{1\leq i\leq I, 1\leq m\leq M}$. We propose to define the set of *admissible* PSFs

coefficients as

$$\mathcal{B} \stackrel{\text{def.}}{=} \text{cone}((\boldsymbol{\beta}_{i,m})_i, 1 \leq m \leq M) \subset \mathbb{R}^I.$$

This roughly amounts to say that admissible PSFs correspond to the conic hull of the already observed and denoised PSFs. Taking the conic hull seems natural: if a PSF is in the set, all its scaled versions by a non-negative factor also belong to the set. Let $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_I) \in \mathbb{R}^{I \times N}$ denote a - possibly infeasible - estimate of interpolation map. We can generate a feasible one $\hat{\boldsymbol{\alpha}}$ by projection: $\hat{\boldsymbol{\alpha}}[n] \stackrel{\text{def.}}{=} \Pi_{\mathcal{B}}(\boldsymbol{\alpha}[n])$ for all $1 \leq n \leq N$.

The projection algorithm Let $\mathbf{B} = [\boldsymbol{\beta}_{\cdot,1}, \ldots, \boldsymbol{\beta}_{\cdot,M}] \in \mathbb{R}^{I \times M}$ denote the matrix of observed coefficients. Projecting the coefficients $\boldsymbol{\alpha} \in \mathbb{R}^I$ of a PSF onto $\mathcal{B}$ amounts to solving the following convex variational problem:

$$\inf_{\boldsymbol{\lambda} \in \mathbb{R}^M, \boldsymbol{\lambda} \geq 0} \frac{1}{2} \|\mathbf{B}\boldsymbol{\lambda} - \boldsymbol{\alpha}\|_2^2.$$

It can be solved using an accelerated projected gradient descent [Nes18]. If the constraint $\boldsymbol{\alpha} \in \mathcal{B}$ is incorporate directly in Problem (5.8), then the functions $(\hat{\alpha}_i)_i$ are not necessarily thin-plate spline anymore, which tends to significantly complicate the problem.

Unfortunately, if M is very large, applying matrix-vector products with B for every pixel $n \leq N$ becomes untractable and we need to simplify the cone $\mathcal{C}$. Following [KSK13], we propose to select a small subset of the columns of $\mathbf{B}$ using a simple greedy algorithm. We start with a matrix $\hat{\mathbf{B}}$ containing a single vector equal to the average of the columns of $\mathbf{B}$. We then update it by iteratively adding the column in $\mathbf{B}$ which maximizes the angle with the current conic hull of the columns in $\hat{\mathbf{B}}$. We stop when the angle is below a given threshold. In our experiments with $M = 14000$ PSFs and $I = 5$, we could obtain a very good approximation of the hull with only 20 components instead of 14000, making the projection algorithm relevant even for very large images.

Part III

Blind inverse problems

Français : Cette partie est dédiée à la résolution de problèmes inverse aveugle. L'hypothèse essentielle dans tous les travaux présentés ici est la connaissance d'une famille d'opérateur permettant de bien capturer les opérateurs recherchés. Cette hypothèse a longuement été étudiée dans la partie précédente. Cette partie est décomposée en trois chapitres. Nous commençons par présenter un cadre théorique garantissant l'identification d'un opérateur à partir de mesures ponctuelles de celui-ci. Dans un second temps, nous proposons un réseau de neurones pour résoudre le problème d'identification d'opérateur dans le cas où l'image observée ne serait pas seulement constituée de points. Dans le dernier chapitre, nous proposons une méthode de defloutage aveugle incorporant les différents outils introduits dans cette thèse.

English: This part is dedicated to solving blind inverse problem. The main assumption in all the work presented here is the knowledge of a family of operators that captures well the desired operators. This hypothesis has been studied in the previous part. The following is divided into three chapters. We begin by presenting a theoretical framework that guarantees the identification of an operator based on a point measurement of itself. In a second step, we propose a neural network to solve the problem of operator identification in the case where the observed image is not only composed of spikes. In the last chapter, we propose a blind deblurring method incorporating the different tools introduced in this thesis.

Chapter 6

Blind inverse problems with isolated spikes

Résumé : *Supposons qu'un opérateur intégral inconnu vivant dans un sous-espace connu soit observé indirectement, en évaluant son action sur un quelques masses de Dirac à des endroits inconnus. Ces informations sont-elles suffisantes pour retrouver de manière stable l'opérateur et les positions des réponses impulsionnelles ? Nous étudions cette question et y répondons positivement dans le cadre d'hypothèses techniques réalistes. Nous illustrons le bien-fondé de cette théorie sur deux problèmes difficiles de l'imagerie optique : la super-résolution aveugle et déconvolution. Cela fournit une approche simple, pratique et théorique pour résoudre ces problèmes qui résistent depuis longtemps.*

Abstract: *Assume that an unknown integral operator living in some known subspace is observed indirectly, by evaluating its action on a few Dirac masses at unknown locations. Is this information enough to recover the operator and the impulse responses locations stably ? We study this question and answer positively under realistic technical assumptions. We illustrate the well foundedness of this theory on two challenging optical imaging problems: blind super-resolution and deconvolution. This provides a simple, practical and theoretically grounded approach to solve these long resisting problems.*

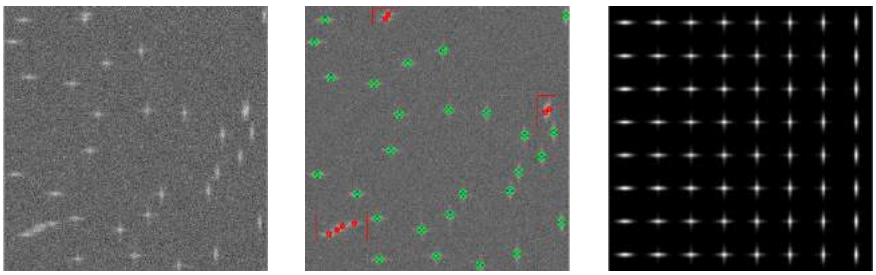
This chapter is going to be submitted soon. It is join work with Pierre Weiss.

Contents

6.1 Introduction **120**
6.1.1 Related works 121
6.1.2 Our contribution 122
6.2 Preliminaries **123**
6.2.1 Notation 123
6.2.2 Assumptions 123
6.2.3 Some intuition 124
6.3 Main results **127**

6.3.1 The case of known weights 127
6.3.2 The case of unknown weights 131
6.4 Applications **134**
6.4.1 Convolution operators with known weights 134
6.4.2 Product-convolution operators and unknown weights 136
6.4.3 A 2D experiment 138
6.5 Proofs **139**
6.5.1 Proofs of the propositions from Section 6.2.3 139
6.5.2 Proof of Theorem 6.3.1 140
6.5.3 Proof of Theorem 6.3.2 141
6.5.4 Proof of Proposition 6.3.3 142
6.5.5 Proof of Proposition 6.3.4 143
6.5.6 Proof of Theorem 6.3.5 144
6.5.7 Proof of Theorem 6.3.6 146
6.5.8 Proof of Theorem 6.3.7 147

6.1 Introduction



(a) (b) (c)

Figure 6.1: A sketch of the contribution: (a) a noisy image of the action of an unknown operator on a few Dirac masses, (b) detection of isolated spikes, (c) an operator estimate applied to a Dirac comb. The results in (b) and (c) were obtained using the algorithms proposed in this chapter, see Section 6.4.3 for the technical details.

To motivate this chapter, let us start with a concrete problem in imaging. In Figure 6.1a, we simulated an image of fluorescent proteins observed with an optical microscope. Assume that an algorithm is able to recover the proteins locations at a sub-pixel accuracy from this image. By taking thousands of such images and stacking the protein locations, it is possible to break the diffraction limit and to construct an image with a resolution of the order of a dozen of nanometers. This principle was awarded the 2014 Nobel prize in chemistry [Bet+06; MK89].

From a mathematical viewpoint, this problem can be modelled as follows. Let $\bar{\mu} = \sum_{n=1}^{N} \bar{w}_n \delta_{\bar{x}_n} \in \mathcal{M}(\mathbb{R}^d)$ denote a Radon measure that encodes the protein locations $(\bar{x}_n)$ and their intensity $(\bar{w}_n)$. Assume that this measure is observed indirectly through a linear regularizing operator $\bar{A} : \mathcal{M}(\mathbb{R}^d) \to \mathcal{C}_0^0(\mathbb{R}^d)$:

$$y_m = (\bar{A}\bar{\mu})(z_m) + b, \tag{6.1}$$

where $y \in \mathbb{R}^M$ is the observed data, $(z_1, z_2, \ldots, z_M)$ denotes a set of sampling locations in $\mathbb{R}^D$ and $b \in \mathbb{R}^M$ is some additive noise.

Numerous approaches have been developped over the years to recover the positions $(\bar{x}_n)$ from the measurements y. We refer the interested reader to the summaries of the super-resolution challenges [Sag+15; Sag+19] for more insight on the possible approaches. The main hurdles to solve this problem are the following:

- The number of measurements M can be huge, making it essential to design computationally efficient methods.
- The weights $(\bar{w}_n)$ are usually unknown.
- It is important to work off-the-grid to avoid biases in the location estimation.
- The proteins can sometimes be aggregated in clusters, resulting in a difficult disentanglement of their individual locations.
- Most importantly for this chapter: the operator $\bar{A}$ is often only partially known, making it important to estimate both the positions and weights $(\bar{w}_n, \bar{x}_n)$, but also the operator $\bar{A}$ itself.

The main objective of this work is to design certified methods, which are able to cope with the above difficulties. This work will strongly rely on the following unformal assumption:

Assumption 6.1.1. *A few Dirac masses are sufficiently separated from the others, so that the resulting impulse responses can be sensed* independently.

6.1.1 Related works

When the operator $\bar{A}$ is known, recovering $\bar{\mu}$ is a challenging problem, since the inverse problem is ill-posed and infinite dimensional. Specifying prior assumptions on the signal $\bar{\mu}$ to certify its approximate recovery is essential [Sch+09]. A few mathematical breakthroughs were achieved in the recent past.

Off-the-grid total variation minimization with a known operator In [CF14; DP15], the authors proposed to recover the individual point sources by solving a generalization of the basis pursuit to an infinite dimensional setting. They showed that the recovery was stable given that the spikes were sufficiently separated. In [DDP17], the authors showed that the separation was not needed, provided that the weights $(\bar{w}_n)$ were positive. From a numerical perspective, the solution of this problem can be found rather efficiently using techniques of semi-infinite programming [BP13; Den+19; FGW20]. This type of approach is currently amongst the best competitors when a high density of proteins is used.

Gridded lifting for an unknown operator Assume that the operator $\bar{A}$ is unknown but lives in a *known finite dimensional subspace* $\mathcal{A}$. Also assume that the positions $(\bar{x}_n)$ are known, but that the weights $(\bar{w}_n)$ are unknown. Under these hypotheses, an elegant solution to recover $\mathcal{A}$ and $\bar{\mu}$ was proposed by Ahmed

et al in [ARR13] based on a trick called lifting. This approach allows to tackle bilinear problems of the form

$$\inf_{A\in\mathcal{A},u\in\mathcal{U}} \frac{1}{2}\|Au-y\|_2^2, \tag{6.2}$$

where $\mathcal{U}$ is a finite dimensional subspace of signals, by transforming the bilinear problem into a linear one restricted to rank-1 matrices. This nonconvex constraint can then be relaxed to a convex one by using the nuclear norm. This approach can be guaranteed to stably estimate $(\bar{w}_n)$ and $\bar{A}$ under rather stringent assumptions. The assumptions were relaxed in a series of works [ARR13; LS15; Chi16; JKS17; AD18]. One important achievement was to allow to handle sparsity constraints over a fixed grid instead of subspace constraints. This is particularly relevant for the considered setting.

Off-the-grid lifting In [Chi16], Y. Chi showed that the lifting trick could also be used when the positions $(\bar{x}_n)$ live off-the-grid, $D = 1$ and the operators in $\mathcal{A}$ are convolution operators. The approach was then extended to the 2 dimensional setting for convolution operators in [SD18]. In [Che+20a], an alternative formulation was proposed based on the Hankel lifting for convolution operators in 1D. This approach is elegant but is currently restricted to convolution operators, while it is important in many applications to consider space variant systems. In addition, we will see that a convex relaxation may not be the most efficient approach from a practical viewpoint in the numerical experiments.

6.1.2 Our contribution

Our main contribution in this work is to propose a simple estimation method that strongly relies on Assumption (6.1.1). The proposed methodology offers many significant advantages:

- We can work with near arbitrary subspaces of operators $\mathcal{A}$.
- We work under a general linear sampling model with arbitray linear forms.
- The proposed theory doesn't require a grid.
- Our theory is rather simple and leads to recovery conditions that can be checked in advance (for some of them) or a posteriori (for some others).
- The proposed theory provides answers to alternative questions such as the ability to recover the position of Dirac masses from simple correlation algorithms.
- We propose an original study of the stability to noise under a white Gaussian noise assumption on b. This requires analyzing the suprema of continuous Gaussian processes and chaos.
- The proposed framework - though strongly dependent on Assumption 6.1.1 - is still realistic for various applications. For instance, we recently proposed a set of algorithms to estimate the subspace of operators $\mathcal{A}$ in [DEW19; Deb+20a].
- Most importantly, the proposed algorithms are simple to implement and efficient in practice.

6.2 Preliminaries

All the proofs in this chapter are postponed to the appendix.

6.2.1 Notation

Throughout the chapter $\mathcal{M}(\mathbb{R}^D)$ will denote the set of Radon measures, i.e. the dual of the set $\mathcal{C}_0^0(\mathbb{R}^D)$ of continuous functions vanishing at infinity. For $\mu \in \mathcal{M}(\mathbb{R}^D)$ and $u \in \mathcal{C}_0^0(\mathbb{R}^D)$, we let $\langle \mu, u \rangle \in \mathbb{R}$ denote the value of the linear form μ on u. We also let $\mu \star u$ denote the convolution product between μ and u defined for all $x \in \mathbb{R}^D$ by $(\mu \star u)(x) = \langle \mu, u(x - \ \cdot\ ) \rangle$.

In all the chapter, the notation $\langle\ \cdot\ ,\ \cdot\ \rangle$ will also refer to the usual scalar product on the vector space $\mathbb{R}^N$, where $N \in \mathbb{N}$ and for $u \in \mathbb{R}^N$, $\|u\|_2$ will denote the ℓ^2-norm of u defined by $\|u\|_2^2 = \langle u, u \rangle$. For two matrices M_1, M_2 in $\mathbb{R}^{M \times N}$, the notation M_1^T will stand for the transpose of M_1 and $\langle M_1, M_2 \rangle_F = \mathrm{Tr}(M_1^T M_2)$ will denote the Frobenius scalar product.

We let $\langle\ \cdot\ ,\ \cdot\ \rangle_{L^2(\mathbb{R}^D)}$ denote the usual scalar product of $L^2(\mathbb{R}^D)$. For a compact and symmetric set $\Omega \subset \mathbb{R}^D$, we let $PW(\Omega)$ denote the Paley-Wiener set of bandlimited functions on Ω, i.e. the set of functions in $L^2(\mathbb{R}^D)$ that have a Fourier transform that vanishes outside Ω.

A family of functions $(e_i)_{1 \leq i \leq I}$ is said to be orthogonal if we have

$$\langle \mathbf{x}_i, \mathbf{x}_j \rangle = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise} \end{cases}.$$

6.2.2 Assumptions

All our results will be established under the following two assumptions on the family of operators.

Assumption 6.2.1 (The operators' structure)**.** *We assume that the family of observation operators $\mathcal{A} = \mathrm{span}\{A_1, \ldots, A_I\}$ is a subspace of linear operators from $\mathcal{M}(\mathbb{R}^D)$ to $C_0^0(\mathbb{R}^D)$.*

For any $A \in \mathcal{A}$, there exists a vector $\gamma = (\gamma_i) \in \mathbb{R}^I$ such that for any $\mu \in \mathcal{M}(\mathbb{R}^D)$:

$$A\mu = A(\gamma)\mu \stackrel{\text{def.}}{=} \sum_{i=1}^{I} \gamma_i A_i \mu \tag{6.3}$$

The next assumption describes the general sampling model considered in this work.

Assumption 6.2.2 (The observation model)**.** *Let $(\nu_m)_{1 \leq m \leq M}$ in $\mathcal{M}(\mathbb{R}^D)$ denote a collection of M linear forms on $C_0^0(\mathbb{R}^D)$. Let $(\bar{x}_n)_{1 \leq n \leq N}$ denote a collection of N points in $\mathbb{R}^D$ and $(\bar{w}_n)_{1 \leq n \leq N}$ denote N weights.*

We assume that we are given the $N \times M$ measurements

$$y_{n,m} \stackrel{\text{def.}}{=} \bar{w}_n \langle \nu_m, A(\bar{\gamma}) \delta_{\bar{x}_n} \rangle + b_{n,m}. \tag{6.4}$$

In what follows, we will let $y_n = (y_{n,m})_m$ denote the measurement vector in $\mathbb{R}^M$ associated to the n-th Dirac mass $\delta_{\bar{x}_n}$. The observation model 6.2.2 allows to describe nearly any sampling device. For instance the traditional pointwise sampling would consist in choosing $\nu_m = \delta_{z_m}$, where $(z_m)_{1\leq m\leq M}$ is a set of sampling locations. The critical element in this assumption is that the impulse responses are observed *independently from each other*.

Under Assumptions 6.2.1 and 6.2.2, the impulse response of an operator $A(\gamma)$ at a location $z \in \mathbb{R}^D$ is given by $A(\gamma)\delta_z = \sum_{i=1}^I \gamma_i A_i \delta_z$. Letting

$$(E(z))_{m,i} \stackrel{\text{def.}}{=} \langle \nu_m, A_i \delta_z \rangle, \tag{6.5}$$

we can rewrite equation (6.4) compactly as

$$y_{n,m} = (E(\bar{x}_n)\bar{\alpha}_n)_m + b_{n,m}, \tag{6.6}$$

with $\bar{\alpha}_n = \bar{w}_n \bar{\gamma}$. The matrix-valued function $E : \mathbb{R}^D \to \mathbb{R}^{M\times I}$ will play an essential role in our analysis. Some of our results will depend on two additional hypotheses.

Assumption 6.2.3 (Identifiability of the operator)**.** *For all $x \in \mathbb{R}^D$, the mapping $E(x) : \mathbb{R}^I \to \mathbb{R}^M$, is such that:*

$$\sigma_- \mathbf{I} \preccurlyeq E^*(x)E(x) \preccurlyeq \sigma_+ \mathbf{I} \tag{6.7}$$

with $0 < \sigma_- \leq \sigma_+ < +\infty$. In what follows, we let $\kappa \stackrel{\text{def.}}{=} \frac{\sigma_+}{\sigma_-}$.

This assumption will be useful to guarantee that an operator can be stably estimated once the location of a Dirac mass is known. Throughout the chapter, we let

$$R(x) \stackrel{\text{def.}}{=} \operatorname{Ran}(E(x)) \tag{6.8}$$

denote the subspace of possible measurements for an impulse response located at $x \in \mathbb{R}^D$ and $\Pi_{R(x)}$ denote the orthogonal projector onto the range $R(x)$. Another important technical assumption to guarantee the stability of the recovery of the Dirac locations is the following.

Assumption 6.2.4 (Identifiability of the Dirac masses location)**.** *The mapping E satisfies the following inequality for any pair $x, \bar{x} \in \mathbb{R}^D$*

$$\|\Pi_{R(x)}\Pi_{R(\bar{x})}\|_{2\to 2} \leq 1 - \phi(\|x - \bar{x}\|_2) \tag{6.9}$$

for some nondecreasing function $\phi : \mathbb{R}_+ \to [0, 1]$ with $\phi(0) = 0$ and $\phi(t) > 0$ for $t > 0$.

This assumption will allow to guarantee the stable recovery of the Dirac masses locations. This can be understood informally as follows. Take two locations $x \neq \bar{x}$ in $\mathbb{R}^D$. Then, the two ranges $R(x)$ and $R(\bar{x})$ do not contain two identical elements. Hence, the knowledge of a measurement of the form $E^*(\bar{x})\bar{\gamma}$ different from 0 should be enough to perfectly recover $\bar{x}$.

6.2.3 Some intuition

Before stating our main results, we provide some intuition on the meaning of Assumption 6.2.3 and Assumption 6.2.4.

An injectivity condition As we will see in the following, Assumptions 6.2.3 and 6.2.4 taken together state that the mapping $(x, \gamma) \mapsto E(x)\gamma$ is injective on $(\mathbb{R}^D \times \mathbb{R}^I \backslash \{0\})$, which is a necessary condition to guarantee the identifiability of a position and an operator from a *single* measurement. For instance, it implies that - for any x - the subspace $\mathrm{span}(A_i\delta_x, 1 \leq i \leq I)$ does not contain two elements that are shifted versions of each other. This hypothesis is essential to discard the standard ambiguity in blind deconvolution related to the fact that the signal and the convolution kernel can be shifted in opposite directions and still yield the same measurement vector, see e.g. [LLB17].

A correlation condition Assumption 6.2.4 allows to control the correlation between measurements of an impulse response at x with an operator $A(\gamma)$ and another at $\bar{x}$ with an operator $A(\bar{\gamma})$. Indeed, we obtain using Cauchy-Schwarz inequality:

$$\begin{aligned}\langle E(x)\gamma, E(\bar{x})\bar{\gamma}\rangle &= \langle \Pi_{R(x)}E(x)\gamma, \Pi_{R(\bar{x})}E(\bar{x})\bar{\gamma}\rangle = \langle \Pi_{R(\bar{x})}\Pi_{R(x)}E(x)\gamma, E(\bar{x})\bar{\gamma}\rangle \\ &\leq \|\Pi_{R(\bar{x})}\Pi_{R(x)}\|_{2\to 2}\|E(x)\gamma\|_2\|E(\bar{x})\bar{\gamma}\|_2\end{aligned}$$

A geometric condition The quantity $\|\Pi_{R(x)}\Pi_{R(\bar{x})}\|_{2\to 2}$ is closely related to the principal angle between the subspaces $R(x)$ and $R(\bar{x})$. To realize this, let us recall that the principal angle between two subspaces U and V of a Hilbert space is defined by

$$\angle(U, V) = \arccos \max_{\substack{u\in U, v\in V \\ u\neq 0, v\neq 0}} \frac{\langle u, v\rangle}{\|u\|_2\|v\|_2}. \tag{6.10}$$

Proposition 6.2.1. *We have*

$$\|\Pi_{R(x)}\Pi_{R(\bar{x})}\|_{2\to 2}^2 \leq \cos\left(\angle\left(R(x), R(\bar{x})\right)\right).$$

Convolution operators In this paragraph, we aim at providing some insights on Assumptions 6.2.3 and 6.2.4 for the particular case of convolution operators. We will work under the following assumption.

Assumption 6.2.5. *We assume that we are given an orthogonal* [1] *family* $(e_i)_{1\leq i\leq I}$ *of functions in* $PW(\Omega)$*. The operators* A_i *are convolutions with the filters* e_i*, i.e.* $A_i\mu = e_i \star \mu$ *for* $\mu \in \mathcal{M}(\mathbb{R}^D)$*.*

The linear forms ν_m *describing the sampling device correspond to a Shannon sampler, i.e.* $\nu_m = \delta_{z_m}$*, where the positions* z_m *correspond to a Cartesian grid with a grid-size smaller than* $\frac{2\pi}{\mathrm{diam}(\Omega)}$*.*

Under Assumption 6.2.5, any $(u, v) \in PW(\Omega)^2$ satisfy

$$\langle u, v\rangle_{L^2(\mathbb{R}^D)} = \sum_{m\in\mathbb{N}} u(z_m)v(z_m), \tag{6.11}$$

which is a variant of the Shannon-Nyquist theorem, see e.g. [Mal99, Thm 3.5].

Proposition 6.2.2 (Operator identifiability for convolution operators)**.** *Under Assumption* (6.2.5)*, we have* $E(z)^*E(z) = \mathbf{I}_I$*, hence Assumption 6.2.3 is satisfied with* $\sigma_- = \sigma_+ = 1$*.*

[1] The orthogonality is not a strong assumption, since any family can be orthogonalized.

Proposition 6.2.3 (Location identifiability for convolution operators)**.** *Let $\mathcal{M} : \mathbb{R}^D \to \mathbb{R}^{I\times I}$ denote the following cross-correlation matrix-valued function:*

$$[\mathcal{M}(x-x')]_{i,i'} \stackrel{\text{def.}}{=} \langle e_i(\ \cdot\ -x), e_{i'}(\ \cdot\ -x')\rangle_{L^2(\mathbb{R}^D)} \tag{6.12}$$

Under Assumption (6.2.5)*, we have*

$$\|\Pi_{R(x)}\Pi_{R(x')}\|_{2\to 2} = \|\mathcal{M}(x-x')\|_{2\to 2}. \tag{6.13}$$

Proposition 6.2.3 shows that the condition (6.9) characterizes the speed of decay of a cross-correlation matrix. For instance, consider the simplest case $I = 1$, corresponding to a convolution with a known filter e_1. Then (6.9) simply measures how fast the auto-correlation function of e_1 decays away from 0. Intuitively, this information is central to derive stability results for algorithms that estimate the Dirac locations by finding correlation maxima. This statement will be made precise in Theorem 6.3.2.

Product-convolution operators To encode space varying operators, we now turn to product-convolution expansions [EW17]. These decompositions allow to represent compactly most linear integral operators arising in applications. For the sake of the current chapter, we will work under the simplifying assumptions below.

Assumption 6.2.6 (Product-convolution expansion)**.** *We assume that we are given an orthogonal family $(e_k)_{1\le k\le K}$ of bandlimited functions in $PW(\Omega)$, and another orthogonal family $(f_l)_{1\le l\le L}$ of functions in $L^2(\mathbb{R}^D)\cap C_0^0(\mathbb{R}^D)$.*

The family of observation operators $\mathcal{A}$ is a subspace of product-convolution expansions from $\mathcal{M}(\mathbb{R}^D)$ to $C_0^0(\mathbb{R}^D)$ defined as follows. For any $A\in\mathcal{A}$, there exists a matrix $\gamma = (\gamma_{k,l}) \in \mathbb{R}^{K\times L}$ such that for any $\mu\in\mathcal{M}(\mathbb{R}^D)$:

$$A\mu = A(\gamma)\mu \stackrel{\text{def.}}{=} \sum_{k=1}^{K}\sum_{l=1}^{L} \gamma_{k,l} e_k \star (f_l \odot \mu). \tag{6.14}$$

Similarly to Assumption 6.2.5, we assume that a Shannon sampler is used. Letting $i = (k,l)$, this implies that $(E(z))_{i,m} = f_l(z_m)e_k(z_m - z)$.

Let us mention that the blurring operators appearing in optics can be represented very efficiently using this structure [FR05a]. In addition, we recently showed how a subspace of product-convolution operators $\mathcal{A}$ could be constructed in practice in optical imaging [BEW19b; DEW19; Deb+20a].

Proposition 6.2.4. *Under Assumptions 6.2.1, 6.2.2 and 6.2.6, we have*

$$\|\Pi_{R(x)}\Pi_{R(x')}\|_{2\to 2} = \|\mathcal{M}(x-x')\|_{2\to 2}. \tag{6.15}$$

However, for $L\ge 2$, Assumption 6.2.3 is not valid: the mapping $E(z)$ is not injective for any z.

As a consequence of this proposition, we will see that the identification of a product-convolution operator with $K\ge 2$ is possible only under the condition $N\ge L$, i.e. by observing multiple impulse responses.

6.3 Main results

A significant difficulty in the problem of operator estimation comes from the fact that the weights $\bar{w}_n$ are possibly unknown. This issue has been carefully analyzed in a series of recent works, leading to a better understanding of the strengths and limitations of the idea of convex lifting and relaxation [ARR13; Li+19; JKS17; LLB17]. The second difficulty comes from the fact that the positions $\bar{x}_n$ are unknown. This issue received less attention in the literature and we will first focus on this, by assuming that the weights $\bar{w}_n \neq 0$ are known. We will then turn to the case of unknown weights.

6.3.1 The case of known weights

For all $1 \leq n \leq N$, any measurement y_n of the form (6.6) can be written as

$$y_n = E(\bar{x}_n)\bar{\alpha}_n + b_n, \tag{6.16}$$

where $\bar{x}_n \in \mathbb{R}^D$ is an unknown location that we wish to recover and $\bar{\alpha}_n = \bar{w}_n\bar{\gamma}$ is a vector colinear to the unknown operator parameterization $\bar{\gamma}$. Our aim in this section is to design an estimate $\hat{X} = (\hat{x}_1, \ldots, \hat{x}_N)$ of $\bar{X} = (\bar{x}_1, \ldots, \bar{x}_N)$ and $\hat{\gamma}$ of $\bar{\gamma}$ and certify their proximity, despite the noise term $b_n \in \mathbb{R}^M$. Letting $X = (x_1, \ldots, x_n) \in \mathbb{R}^{D\times N}$ and

$$F(\gamma, X) \stackrel{\text{def.}}{=} \frac{1}{2}\sum_{n=1}^{N} \|\bar{w}_n E(x_n)\gamma - y_n\|_2^2 = \frac{1}{2}\sum_{n=1}^{N} F_n(\gamma, x), \tag{6.17}$$

a natural approach to achieve this goal is to solve the following nonlinear least-square problem

$$(\hat{\gamma}, \hat{X}) \stackrel{\text{def.}}{=} \operatorname*{argmin}_{X\in\mathbb{R}^{D\times N}, \gamma\in\mathbb{R}^{I\times N}} F(\gamma, X). \tag{6.18}$$

Characterizing the solutions

Theorem 6.3.1. *A global minimizer* $(\hat{\gamma}, \hat{X})$ *of Problem* (6.18) *can be obtained in a two step procedure.*

First, estimating the positions in (6.18) *boils down to solving the following maximum correlation problem*

$$\hat{x}_n \in \operatorname*{argmax}_{x\in\mathbb{R}^D} \langle \Pi_{R(x)} y_n, y_n\rangle, \tag{6.19}$$

independently for all $1 \leq n \leq N$.

Second, the vector $\hat{\gamma}$ *is given by any solution of the following linear system system:*

$$\left(\sum_{n=1}^{N} \bar{w}_n^2 E^*(\hat{x}_n)E(\hat{x}_n)\right)\gamma = \sum_{n=1}^{N} \bar{w}_n E^*(\hat{x}_n)y_n. \tag{6.20}$$

Assumption 6.2.3 is sufficient to ensure that $\hat{\gamma}$ *is unique with a single observation.*

Theorem 6.3.1 shows that under realistic assumptions, the blind inverse Problem (6.18) can be solved exactly in a two step procedure. The first step consists in finding independently N maxima of nonconvex functions in $\mathbb{R}^D$.

This is not a trivial task, but it can be implemented with standard methods from nonlinear programming (gradient or Newton-like method) starting from a sufficiently good initialization. Let us mention that this procedure is one of the most popular approach to estimate the positions [Sag+15]. The second step consists in solving a small dimensional linear system of equations.

Remark 6.3.1. *In the proposed formulation, two implicit regularization terms are used: i) we look for a single Dirac mass and ii) the operator live in a known subspace. If the dimension I of the subspace is large, the proposed methodology might fail: as I increases, so does* $\dim(R(x))$. *For instance in the case of convolution operators (see Assumption 6.2.5), we have* $\dim(R(x)) = I$. *If $I = M$, the correlation function* (6.19) *is constant and there is no hope to recover $\bar{x}$.*

A possible solution for this problem is to add a weighted ℓ^2-regularization on γ of the form $\frac{1}{2}\sum_{i=1}^{I}\theta_i\gamma_i^2$, where θ_i are weights adapted to the problem at hand. Most of the theory developed in this chapter applies to this setting as well. The main difference is that this regularization introduces a bias in the operator estimate.

Stability of the location estimates

As seen in Theorem 6.3.1, recovering the location of each Dirac location x_n can be achieved independently. The following theorem shows that this estimation is robust despite the noise term b_n.

Theorem 6.3.2 (Stability of the Dirac locations)**.** *Let $y_{0,n} = E(\bar{x}_n)\bar{\gamma}$ denote the noiseless measurements and assume that $\|b_n\|_2 \leq \theta\|y_{0,n}\|_2$ with $\theta < \frac{\sqrt{6}}{2} - 1 \simeq 0.225$. Then, under Assumption 6.2.4, the following inequality holds:*

$$\|\hat{x}_n - \bar{x}_n\|_2 \leq \phi_+^{-1}\left(2\theta^2 + 4\theta\right), \tag{6.21}$$

where $\phi_+^{-1}(t) = \inf\{s \text{ s.t. } \phi(s) \geq t\}$ is the quantile function of ϕ (in particular $\phi_+^{-1} = \phi^{-1}$ is ϕ is bijective).

The above bound can be quite pessimistic for large M, but - apart for the constant $\frac{\sqrt{6}}{2} - 1$ - it cannot be improved for an arbitrary noise term b_n. In the case of random noise however, we can likely obtain a much better probabilistic control using concentration inequalities. We focus on the case $N = 1$ observed impulse response, which allows us to discard the indices n. The noiseless part of forward model (6.16) now becomes $y_0 = E(\bar{x})\bar{\alpha}$. We let $b \in \mathbb{R}^M$ denote the additive noise and we let $\hat{x}$ denote the estimated position given by solving

$$\hat{x} \in \operatorname*{argmax}_{x\in\mathbb{R}^d} \left\langle \Pi_{R(x)}(y_0 + b), y_0 + b \right\rangle.$$

Proposition 6.3.3. *Assume that $b \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ is white Gaussian noise of variance σ^2. Define the following two random processes*

$$\Delta_1(x) \overset{\text{def.}}{=} \langle \Pi_{R(x)}y_0, \Pi_{R(x)}b\rangle \text{ and } \Delta_2(x) \overset{\text{def.}}{=} \frac{1}{2}\|\Pi_{R(x)}(b)\|_2^2. \tag{6.22}$$

Then under Assumption 6.2.4

$$\|\hat{x} - \bar{x}\|_2 \leq \phi_+^{-1}\left(\frac{2\mathrm{Ampl}(\Delta_1 + \Delta_2)}{\|y_0\|_2^2}\right), \tag{6.23}$$

where $\mathrm{Ampl}(f) \overset{\text{def.}}{=} \sup_{x\in\mathbb{R}^D} f(x) - \inf_{x\in\mathbb{R}^D} f(x)$.

This proposition reveals that the critical quantity to control, to evaluate the localization error is the ammplitude of the random process $\Delta_1 + \Delta_2$. Obtaining tight analytical bounds for this is a difficult problem in general. Hopefully the following proposition shows that it can be evaluated efficiently using numerical procedures.

Proposition 6.3.4. *We have* $\mathrm{Ampl}(\Delta_1 + \Delta_2) \leq \mathrm{Ampl}(\Delta_1) + \mathrm{Ampl}(\Delta_2)$. *In addition, the random variable* $Z_1 = \mathrm{Ampl}(\Delta_1)$ *is sub-Gaussian:*

$$\mathbb{P}(|Z_1 - \bar{Z}_1| \geq t) \leq 2\exp\left(-t^2/(8\sigma^2\|y_0\|_2^2)\right)$$

and the random variable $Z_2 = \mathrm{Ampl}(\Delta_2)$ *is sub-exponential:*

$$\mathbb{P}(|Z_2 - \bar{Z}_2| \geq t) \leq 2\exp\left(-Ct/\sigma\right),$$

where $\bar{Z}_1$ *and* $\bar{Z}_2$ *are the expectations of* Z_1 *and* Z_2 *and* C *is a universal constant.*

This proposition has two consequences. First, we see that the deviation around the mean of Z_1 scales as $\sigma\|y_0\|_2$ and the deviation around the mean of Z_2 scales as σ. Second, Hoeffding [Ver18, Thm 2.6.1] and Bernstein [Ver18, p. 2.8.1] inequalities imply that computing an empirical average of Z_1 and Z_2 will converge rapidly to the true means $\bar{Z}_1$ and $\bar{Z}_2$. Hence, it is possible to obtain a precise numerical estimate using an empirical average and we know that the probability that the variables deviate from the means by more than $\sigma(\|y_0\|_2 + 1)$ is negligible.

Unfortunately, the averages $\bar{Z}_1$ and $\bar{Z}_2$ are difficult to compute in general. Hence, the above proposition can only be used to estimate average deviations with a computer. The following proposition provides upper-bounds for $\bar{Z}_1$ and $\bar{Z}_2$ under additional regularity assumptions in the 1D setting.

Theorem 6.3.5. *Assume that* $D = 1$, *and that:*

- *The range* $R(x)$ *is constant for* $|x| > 1$.
- *the mapping* $x \mapsto \Pi_{R(x)}$ *is Lipschitz continuous:*

$$\|\Pi_{R(x)} - \Pi_{R(x')}\|_{2\to 2} \leq L\|x - x'\|_2.$$

- *The following inequality holds for* x, x' *in* $[-1, 1]$ *(this can be seen as a specification of Assumption 6.2.4):*

$$\|\Pi_{R(x)}\Pi_{R(x')}\|_{2\to 2} \leq \frac{1}{1 + \|x - x'\|_2^\alpha} \quad \text{with } \alpha > 1. \tag{6.24}$$

Then, we have:

$$\bar{Z}_1 \leq C\sigma\|y_0\|_2 L^{\frac{\alpha}{\alpha+1}} \quad \text{and} \quad \bar{Z}_2 \leq \sigma^2\left(C' + C''L\sqrt{I}\right),$$

for some absolute constants C, C' *and* C''.

Proposition 6.3.3 and Theorem 6.3.5 improve Theorem 6.3.2 massively. A sufficient condition for the bound (6.21) to be informative is that $\mathrm{Ampl}(\Delta_1) + \mathrm{Ampl}(\Delta_2) \leq 2\|y_0\|_2^2$, which holds true for $\sigma \lesssim \min\left(\frac{\|y_0\|_2}{L\sqrt{I}}, \frac{\|y_0\|_2}{L^{\alpha/(\alpha+1)}}\right)$. Here each noise component can have an amplitude of the order of the signal's ℓ^2-norm! Previously, it was the whole ℓ^2-norm of the noise which had to be less than the signal's norm. This surprising phenomenon is actually observed in practice, with a good localization despite a huge amount of noise.

Stability of the operator estimates

Finally, it is possible to guarantee the closeness between $\bar{\gamma}$ and $\hat{\gamma}$ under an additional Lipschitz regularity assumption on E. Here, we work with $N = 1$ observed impulse response. A similar result can be obtained for arbitrary N, but we want to emphasize that the estimation is possible and stable with *a single* observation.

Theorem 6.3.6 (Stability of the operator estimate with a single observation)**.** *Assume that $N = 1$ and that E is $\sqrt{\sigma_+}L_E$-Lipschitz continuous*[2]*:*

$$\|E(x) - E(x')\|_{2\to 2} \leq \sqrt{\sigma_+}L_E\|x - x'\|_2 \quad \textit{for all} \quad (x, x') \in \mathbb{R}^D \times \mathbb{R}^D. \tag{6.25}$$

Then, under Assumption 6.2.3, we have

$$\frac{\|\hat{\gamma} - \bar{\gamma}\|_2}{\|\bar{\gamma}\|_2} \leq \kappa^{3/2}\frac{\|b_1\|_2}{\|y_{0,1}\|_2} + \epsilon_2(\hat{x}) \tag{6.26}$$

with

$$\epsilon_2(\hat{x}) \leq C\kappa^{5/2}L_E\|\hat{x} - \bar{x}\|_2\left(1 + \frac{\|b_1\|_2}{\|y_{0,1}\|_2}\right) + o_{\hat{x}\to\bar{x}}\left(\|\hat{x} - \bar{x}\|_2^2\right),$$

for some absolute constant C.

Together with Theorem 6.3.2, this last result ensures that $\hat{\gamma} \to \bar{\gamma}$ when the noise level $\|b_1\|_2$ vanishes. This means that we can stably recover an operator when observing a single impulse response.

Unfortunately, Assumption 6.2.3 is not always met in practical situations of interest as outlined in Proposition 6.2.4. In that case, observing multiple impulse responses $N > 1$ can still make a stable estimation possible.

Theorem 6.3.7 (Stability of the operator estimate with multiple observations)**.** *Given $X = (x_1, \ldots, x_n)$, let $\bar{w}_- = \min_n |\bar{w}_n|$, $\bar{w}_+ = \max_n |\bar{w}_n|$ and*

$$\mathcal{C}(X) \stackrel{\text{def.}}{=} \sum_{n=1}^{N} \bar{w}_n^2 E^*(x_n)E(x_n).$$

Let $\tilde{\sigma}_- = \bar{w}_-^2\hat{\sigma}_-$ and $\tilde{\sigma}_+ = \bar{w}_+^2\hat{\sigma}_+$ and assume that

$$\tilde{\sigma}_-\mathbf{I} \preccurlyeq \mathcal{C}(\hat{X}) \preccurlyeq \tilde{\sigma}_+\mathbf{I}. \tag{6.27}$$

Similarly to Theorem 6.3.6, assume that E is $\sqrt{\tilde{\sigma}_+}L_E$-Lipschitz continuous and let $\tilde{\kappa} = \frac{\tilde{\sigma}_+}{\tilde{\sigma}_-}$.

Then we have

$$\frac{\|\hat{\gamma} - \bar{\gamma}\|_2}{\|\bar{\gamma}\|_2} \leq \tilde{\kappa}^{3/2}\frac{\|B\|_2}{\|Y_0\|_2} + \epsilon_2(\hat{X}) \tag{6.28}$$

with

$$\epsilon_2(\hat{X}) = O_{\hat{X}\to\bar{X}}\left(\tilde{\kappa}^{5/2}L_E\|\bar{X} - \hat{X}\|_2\left(1 + \frac{\|Y_0\|_2 + \|B\|_2}{\|Y_0 + B\|_2}\right)\right).$$

[2] The scaling in $\sqrt{\sigma_+}$ is natural considering Assumption 6.2.3.

Assumption (6.27) is a geometrical condition intertwining the locations of the Dirac masses and the properties of the observation mapping E. It can be hard to verify in advance. However it only requires computing the $I \times I$ matrix $\mathcal{C}(\hat{X})$, which can be achieved once $\hat{X}$ has been evaluated. The stable estimation of $\hat{X}$ on its side only depends on Assumption 6.2.4, which can be verified in advance and can be satisfied independently of Assumption 6.2.3. Hence, Theorem 6.3.7 actually yields a constructive result to guarantee the stable recovery of an operator with the following approach:

- If Assumption 6.2.4 is satisfied and the noise level is low, estimate $\hat{X}$.
- Evaluate the condition number $\tilde{\kappa}$ of $\mathcal{C}(\hat{X})$.
- If $\tilde{\kappa}$ is sufficiently low, $\hat{\gamma}$ provides a good estimate of $\bar{\gamma}$.

6.3.2 The case of unknown weights

If the weights $\bar{w}_n$ are unknown, the previously described approaches cannot be used directly. It is then tempting to solve the following nonlinear least squares problem:

$$\inf_{\substack{w \in \mathbb{R}^N, \gamma \in \mathbb{R}^I \\ X \in \mathbb{R}^{N \times D}}} J(w, \gamma, X) \tag{6.29}$$

where

$$J(w, \gamma, X) \stackrel{\text{def.}}{=} \frac{1}{2} \sum_{n=1}^{N} J_n(w_n, \gamma, x_n) \text{ with } J_n(w_n, \gamma, x_n) \stackrel{\text{def.}}{=} \frac{1}{2} \|w_n E(x_n)\gamma - y_n\|_2^2.$$

In what follows, we let $(\hat{w}, \hat{\gamma}, \hat{X})$ denote any minimizer of (6.29).

A bilinear inverse problem

Theorem (6.3.1) shows that computing the locations estimate $\hat{X}$ can be achieved independently of the weights $\hat{w}$ and of the operator $\hat{\gamma}$. Minimizing J with respect to (w, γ) for a fixed $X = \hat{X}$ is a *bilinear* inverse problem. It received a considerable attention lately, with numerous progress both on the necessary and sufficient conditions to guarantee the recovery [ARR13; JKS17; LLB17; AD18; KS19], on the optimal stability to noise [Che+20b], and on the numerical aspects through convex lifting [BB19] or local optimization [AMS09; BST14; CJ19; TA20; Zhu+18; Li+19]. We briefly explain below how these results can be applied to the proposed setting.

Let $\hat{I}_n \stackrel{\text{def.}}{=} \dim(R(\hat{x}_n))$ and $\hat{I} = \sum_{n=1}^{N} \hat{I}_n$. Using a singular value decomposition, we can decompose $E(\hat{x}_n)$ as

$$E(\hat{x}_n) = \hat{U}_n \hat{V}_n^*, \tag{6.30}$$

where $\hat{U}_n \in \mathbb{R}^{M \times \hat{I}_n}$ is orthogonal in the sense that $\hat{U}_n^* \hat{U}_n = \mathbf{I}$ and $\hat{V}_n \in \mathbb{R}^{I \times \hat{I}_n}$

contains orthogonal colums. Hence, letting $c_n \stackrel{\text{def.}}{=} \hat{U}_n^* y_n$, we obtain

$$\begin{aligned}\underset{w\in\mathbb{R}^N,\gamma\in\mathbb{R}^I}{\operatorname{argmin}} J(w,\gamma,\hat{X}) &= \underset{w\in\mathbb{R}^N,\gamma\in\mathbb{R}^I}{\operatorname{argmin}} \frac{1}{2}\sum_{n=1}^{N} \|\hat{U}_n\hat{V}_n^* w_n\gamma - y_n\|_2^2 \\ &= \underset{w\in\mathbb{R}^N,\gamma\in\mathbb{R}^I}{\operatorname{argmin}} \frac{1}{2}\sum_{n=1}^{N} \|\hat{V}_n^* w_n\gamma - c_n\|_2^2.\end{aligned}$$

Letting $\hat{\mathcal{B}} : \mathbb{R}^N \times \mathbb{R}^I \to \mathbb{R}^{\hat{I}}$ denote the following bilinear mapping:

$$\hat{\mathcal{B}}(w,\gamma) \stackrel{\text{def.}}{=} \begin{pmatrix} \hat{V}_1^* w_1\gamma \\ \vdots \\ \hat{V}_N^* w_N\gamma \end{pmatrix} \text{ and } c \stackrel{\text{def.}}{=} \begin{pmatrix} c_1 \\ \vdots \\ c_N \end{pmatrix}, \tag{6.31}$$

we can rewrite J more compactly as $J(w,\gamma,\hat{X}) = \frac{1}{2}\|\hat{\mathcal{B}}(w,\gamma) - c\|_2^2$, and hence:

$$\underset{w\in\mathbb{R}^N,\gamma\in\mathbb{R}^I}{\operatorname{argmin}} J(w,\gamma,\hat{X}) = \underset{w\in\mathbb{R}^N,\gamma\in\mathbb{R}^I}{\operatorname{argmin}} \frac{1}{2}\|\hat{\mathcal{B}}(w,\gamma) - c\|_2^2. \tag{6.32}$$

Notice that the dimension M, which might be huge in applications, completely disappeared from this formulation.

Global injectivity conditions

Recovering w and γ is possible only up to a multiplicative constant since

$$J(tw,\gamma/t,\hat{X}) = J(w,\gamma,\hat{X}) \text{ for all } t \neq 0.$$

Now, consider the noiseless setting $B = 0$ and assume that the locations are perfectly recovered: $\hat{X} = \bar{X}$. In that situation, a necessary condition to recover $(\bar{w},\bar{\gamma})$ modulo the above scaling ambiguity is that there exists a unique pair (w,γ) with $\|w\|_2 = 1$ such that $\hat{\mathcal{B}}(w,\gamma) = c$. To the best of our knowledge, deriving conditions to ensure this local injectivity condition received little attention in the literature.

In [KK17; LLB17], the authors study a more stringent *global* injectivity condition of the form

$$\forall c \in \mathbb{R}^{\hat{I}}, \exists \text{ a unique } (w,\gamma) \text{ with } \|w\|_2 = 1 \text{ s.t. } \hat{\mathcal{B}}(w,\gamma) = c. \tag{6.33}$$

Their main result states that a *necessary* condition for $\hat{\mathcal{B}}$ to be globally injective is that

$$\hat{I} \geq 2(N+I) - 4, \tag{6.34}$$

which provides a rule on how to choose the number of measurements N. In addition, they prove that almost every bilinear mapping $\hat{\mathcal{B}}$ with respect to the Lebesgue measure is globally injective provided that the inequality (6.34) holds. In the proposed setting, there are three limitations to this result. First, the operator $\hat{\mathcal{B}}$ that appears in our formulation possesses a peculiar structure which may well fall in a set of 0 measure. Second, we observed that the condition (6.34) could be violated significantly in practice and that stable recovery still occurred for a particular c. Third, the result does not certify that a low complexity algorithm can actually recover the factors.

Optimization of the factors

Solving (6.32) can be achieved using local optimization over each factor w and γ [BST14; Zhu+18; Li+19]. A simple approach consists in using an alternate minimization between the factors as outlined in Algorithm 9. Notice that every

Algorithm 9 Alternating minimization

Require: Initial guess: $w_1 \in \mathbb{R}^N$.
Require: Iteration number K.
  for all $k = 1 \to K-1$ **do**
    $\gamma_{k+1} = \underset{\gamma \in \mathbb{R}^I}{\operatorname{argmin}} \frac{1}{2} \|\hat{\mathcal{B}}(w_k, \gamma) - c\|_2^2$.
    $w_{k+1} = \underset{w \in \mathbb{R}^N}{\operatorname{argmin}} \frac{1}{2} \|\hat{\mathcal{B}}(w, \gamma_{k+1}) - c\|_2^2$.
  end for
  return (w_K, γ_K).

step of the algorithm can be performed efficiently since the dimensions of the problem are significantly reduced. This approach can be certified to recover a stable estimate $(\hat{w}, \hat{\gamma})$ of $(\bar{w}, \bar{\gamma})$ provided that a clever initialization is used [Zhu+18; Li+19]. Sufficient recovery guarantees are for instance provided when the bilinear mapping $\hat{\mathcal{B}}$ is chosen at random. This method also allows to easily incorporate constraints (e.g. nonnegativity) in the factors, which can sometimes allow a significantly improved reconstruction. In all our numerical experiments, we will use the spectral initialization from [Li+19] as a starting point.

Optimization over rank-1 matrices

The bilinear mapping $\hat{\mathcal{B}}(w, \gamma)$ can be rewritten as a linear mapping $\hat{\mathcal{L}}$ on the rank-1 outer product $T = w\gamma^T : \hat{\mathcal{B}}(w, \gamma) = \hat{\mathcal{L}}(T)$. Hence, we have:

$$\inf_{w \in \mathbb{R}^N, \gamma \in \mathbb{R}^I} J(w, \gamma, \hat{X}) = \inf_{T \in \mathbb{R}^{N \times I}, \operatorname{rank}(T)=1} \frac{1}{2} \|\hat{\mathcal{L}}(T) - c\|_2^2. \tag{6.35}$$

The interest of the right-hand side in equation (6.35) compared to the left hand side is that the scaling ambiguity is discarded. Letting $\mathcal{T}$ denote the set of rank-1 matrices, this alternative formulation can be solved using a projected gradient descent described in Algorithm 10. The notation $\Pi_{\mathcal{T}}$ stands for the projection

Algorithm 10 Projected gradient descent

Require: Initial guess: $T \in \mathbb{R}^{N \times I}$.
Require: Iteration number K.
  Compute $\tau = \frac{1}{\|\hat{\mathcal{L}}\|_{2\to 2}^2}$ using a power iteration.
  for all $k = 1 \to K-1$ **do**
    $T_{k+1} = \Pi_{\mathcal{T}} \Big(T_k - \tau \hat{\mathcal{L}}^*(\hat{\mathcal{L}}(T_k) - c) \Big)$.
  end for
  Decompose $T_K = w_K \gamma_K^*$.
  return (w_K, γ_K).

onto the set of rank-1 matrices. It boils down to keep the first eigen-element of the singular value decomposition of the matrix to project.

To the best of our knowledge, this formulation received no attention yet in the literature and we will provide numerical comparisons in the next section. Again, we will use the spectral initialization from [Li+19] as a starting guess for this algorithm.

Convex relaxation using the nuclear norm

Finally, a popular method [ARR13; AD18; BB19; Chi16] is a convex relaxation using the nuclear norm. The usual convex relaxation of the nonconvex problem (6.35) is the following:

$$\inf_{T\in\mathbb{R}^{N\times I},\hat{\mathcal{L}}(T)=c} \|T\|_* \quad \text{or} \quad \inf_{T\in\mathbb{R}^{N\times I}} \frac{1}{2}\|\hat{\mathcal{L}}(T)-c\|_2^2 + \lambda\|T\|_*, \tag{6.36}$$

where $\lambda > 0$ is a regularization parameter and $\|\cdot\|_*$ is the nuclear norm, i.e. the sum of the singular values of T. This convex function over the space of matrices is well known to promote low-rank solutions since the extreme points of the associated unit ball are the rank-1 matrices [Boy+19a]. The stable recovery of the tensor $\bar{w}\bar{\gamma}^T$ has been established under rather stringent conditions based on random subspace assumptions [ARR13; AD18]. Experimentally, the method seems to provide satisfactory results under much weaker conditions.

From a numerical perspective, Problem (6.36) can be solved using a diversity of proximal algorithms, such as an accelerated proximal gradient descent or a Douglas-Rachford algorithm [CP11]. We do not further detail these algorithms, which are well documented in the literature.

6.4 Applications

The aim of this section is to illustrate the proposed theory using simple 1D examples and to explain the setting of the 2D experiment in Figure 6.1.

6.4.1 Convolution operators with known weights

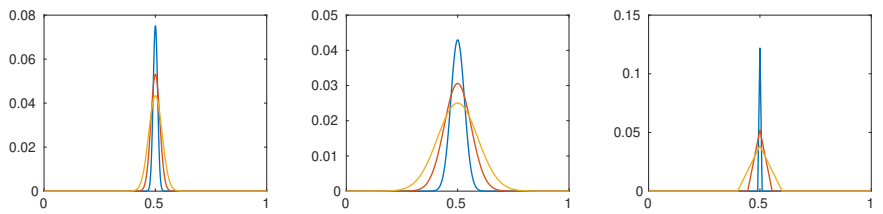
We start with an illustration of Theorem 6.3.2 using convolution operators only. We focus on the case of pointwise sampling on $[0,1]$, by setting $\nu_m = \delta_{z_m}$, with $z_m = m/M$ for $m \in \{1,\dots,M\}$. Notice that this case also covers the case of product-convolution operators since the ranges $R(x)$ of convolution and product-convolution operators are identical.

The families of operators

We consider three families of convolution operators $\mathcal{A}_1$, $\mathcal{A}_2$ and $\mathcal{A}_3$ differing by the choice of the convolution filters.

Family $\mathcal{A}_1$ is defined through a set of convolution operators A_i with Gaussian filters (e_i) defined by:

$$e_i(x) = \exp(-x^2/(2\sigma_i^2)) \text{ with } \sigma_i = 1e^{-2}\cdot\frac{i-1}{I-1} + 3e^{-2}\cdot\left(1-\frac{i-1}{I-1}\right).$$



(a) Filters in family $\mathcal{A}_1$ (b) Filters in family $\mathcal{A}_2$ (c) Filters in family $\mathcal{A}_3$

Figure 6.2: The different families of convolution filters used in Section 6.4.1.

Using this family in a blind deconvolution problem allows to identify the variance of a Gaussian convolution filter. Gaussian convolution filters are amongst the most popular simplified point spread function models in microscopy.

Family $\mathcal{A}_2$ is also defined using Gaussian convolution filters, but the standard deviation ranges in $[3e^{-2}, 9e^{-2}]$ instead of $[1e^{-2}, 3e^{-2}]$.

Family $\mathcal{A}_3$ is defined with less regular convolution filters. Let $\psi(x) = (1 - |x-1|)_+$ denote the hat function.

$$e_i = \psi(x \cdot \sigma_i) \text{ where } \sigma_i = 2e^{-2} \cdot \frac{i-1}{I-1} + 2e^{-1} \cdot \left(1 - \frac{i-1}{I-1}\right).$$

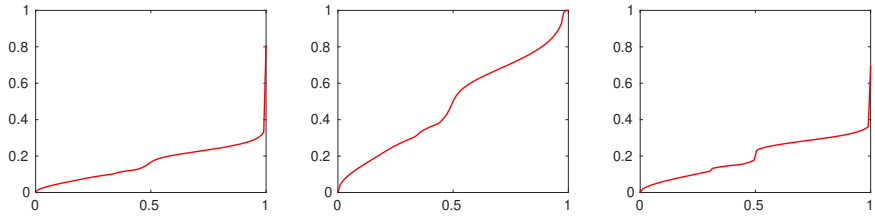
In all settings we set $I = 3$. The filters corresponding to each family are displayed in Figure 6.2. We then orthogonalize the filters using a singular value decomposition on a very fine grid. This leads to a new family of orthogonal filters $(e_i^{\perp})$ which will be used in all experiments to satisfy Assumption 6.2.5.

The inverse functions ϕ^{-1}

As stated in Theorem 6.3.2, the critical element to guarantee a stable recovery of the locations $\bar{x}_m$ is the function ϕ and its inverse, which characterizes the angle between the ranges $R(x)$ and $R(x')$. To evaluate this function, we first sample the function $\|\Pi_{R(0)}\Pi_{R(k\Delta x)}\|_{2\to 2}$ on a fine grid. We store the result in the vector $\phi_0(k) \overset{\text{def.}}{=} 1 - \|\Pi_{R(0)}\Pi_{R(k\Delta x)}\|_{2\to 2}$ with a sampling step Δx. This function is not necessarily nondecreasing. Hence, we find the closest nondecreasing function by solving an isotonic regression problem of the form:

$$\inf_{\phi} \frac{1}{2}\|\phi - \phi_0\|_2^2 \text{ with } \phi_{k+1} - \phi_k \geq 0 \text{ and } \phi \geq \phi_0.$$

This problem is convex and can be solved using the CVX library [GB14] for instance. We use the solution $\hat{\phi}$ of this problem in place of ϕ in Assumption 6.2.4. The inverse filters are displayed in Figure 6.3. The stability to noise is dependent on the speed of ascent of ϕ_+^{-1}. As can be seen by comparing the two Gaussian families, the smallest the filter, the slower the ascent. Hence, very localized impulse responses should be easier to detect with a good accuracy than larger ones. Also notice that the regularity of the convolution kernels seem to have little importance since the inverses $\phi_{1,+}^{-1}$ and $\phi_{3,+}^{-1}$ behave roughly similarly in terms of speed of ascent.



(a) $\phi_{1,+}^{-1}$ (b) $\phi_{2,+}^{-1}$ (c) $\phi_{3,+}^{-1}$

Figure 6.3: The corresponding inverse functions $\phi_{i,+}^{-1}$ (in red) for $i \in \{1, 2, 3\}$, for the different convolution systems.

Stability of the locations

Finally, we study the robustness of the estimation to noise. To this end, we compute the empirical average of the error $\mathbb{E}(|\hat{x} - \bar{x}|)$ for various noise levels and realizations. The expectation is estimated by averaging 100 noise realizations. We fix $\bar{\gamma}$ once for all. We use white Gaussian noise, i.e. $b_n \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$, with $\sigma = \theta \|y_0\|_2 / \sqrt{M}$ and $\theta \in [0, 2]$. Figure 6.4 shows the resuting signals with $M = 100$ for the noise levels $\theta \in \{0, 1, 2\}$ and each family. Notice that $\theta = 1$ and $\theta = 2$ correspond to rather extreme noise levels. We will see that the localization accuracy is surprisingly good in spite of this challenging setting.

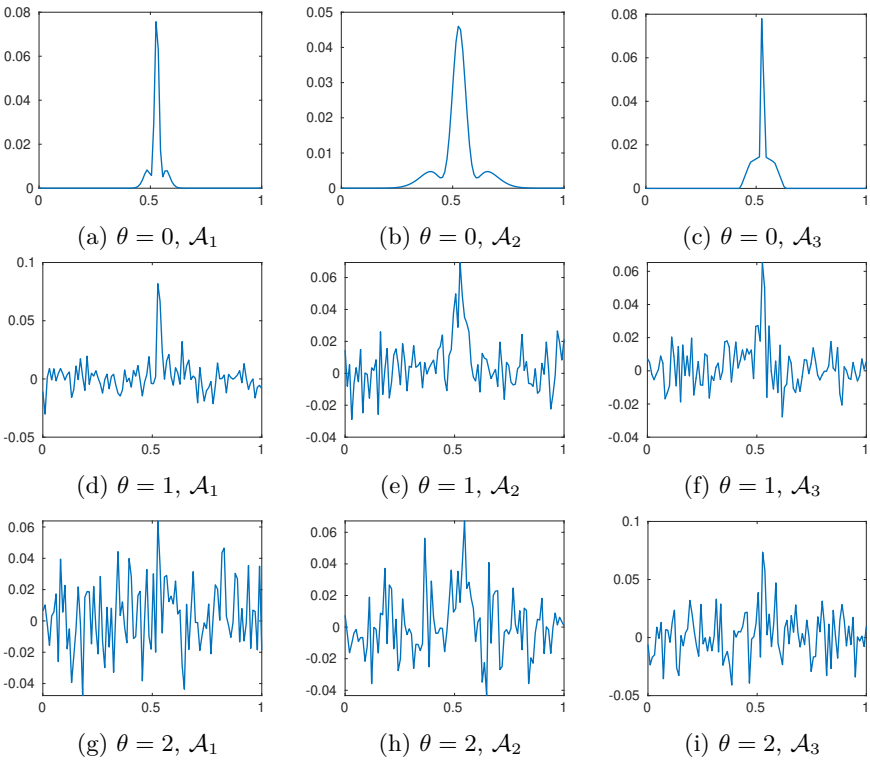
The empirical estimate of $\mathbb{E}(|\hat{x} - \bar{x}|)$ is displayed with respect to the noise level σ in Figure 6.5. The error is displayed as a proportion of a pixel. For instance, an error of 0.1 means that the localization was accurate at a tenth of a pixel. Hence, we can expect a super-resolution effect for precisions below 0.5.

To end this experiment, we evaluate $\hat{\gamma}$ for all experiments and display the relative error $\frac{\|\hat{\gamma} - \bar{\gamma}\|_2}{\|\bar{\gamma}\|_2}$ for all families of operators. Letting $\bar{h} = \sum_{i=1}^{I} \bar{\gamma}_i e_i^{\perp}$ and $\hat{h} = \sum_{i=1}^{I} \hat{\gamma}_i e_i^{\perp}$ denote the true convolution filter and the estimated one, notice that we have $\frac{\|\bar{h} - \hat{h}\|_{L^2(\mathbb{R}^D)}}{\|\bar{h}\|_{L^2(\mathbb{R}^D)}} = \frac{\|\hat{\gamma} - \bar{\gamma}\|_2}{\|\bar{\gamma}\|_2}$, since the family $(e_i^{\perp})$ is orthogonal. In Figure 6.6, we see that the reconstruction errors for any family of convolution filters behave really similarly. In particular, the errors using the family $\mathcal{A}_1$ and the family $\mathcal{A}_2$ are nearly identical. This might come as a surprise since the localization errors were significantly higher for the family $\mathcal{A}_2$, which is a scaled version of $\mathcal{A}_1$. This fact can be explained by the fact that the Lipschitz constant L_E in Theorem 6.3.6 is inversely proportional to the scaling of the Gaussian, which compensates for the localization errors.

6.4.2 Product-convolution operators and unknown weights

The objective of this section is to compare the different algorithms described in Section 6.3.2 for the specific case of 1D product-convolution operators described in Assumption 6.2.6. In this experiment, we work on a grid and set $\hat{X} = \bar{X}$ since the objective is not to assess the localization performance, but rather the ability to solve a bilinear inverse problem.

We use the pointwise sampling model $\nu_m = \delta_{z_m}$ with $z_m = 10 \cdot m/M$ and $M = 1000$. This corresponds to a uniform sampling of the interval $[0, 10]$. For the filters (e_k), we use the family of Gaussian convolution kernels $\mathcal{A}_1$ with $K = 3$. We



(a) $\theta = 0$, $\mathcal{A}_1$ (b) $\theta = 0$, $\mathcal{A}_2$ (c) $\theta = 0$, $\mathcal{A}_3$

(d) $\theta = 1$, $\mathcal{A}_1$ (e) $\theta = 1$, $\mathcal{A}_2$ (f) $\theta = 1$, $\mathcal{A}_3$

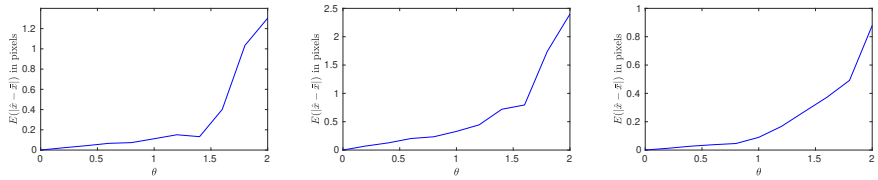
(g) $\theta = 2$, $\mathcal{A}_1$ (h) $\theta = 2$, $\mathcal{A}_2$ (i) $\theta = 2$, $\mathcal{A}_3$

Figure 6.4: Examples of measurements vectors y for different noise levels and each family.

set the vectors f_l as smooth random Gaussian processes by convolving a random vector with distribution $\mathcal{N}(0, \mathbf{I}_M)$ with a Gaussian filter of large variance.

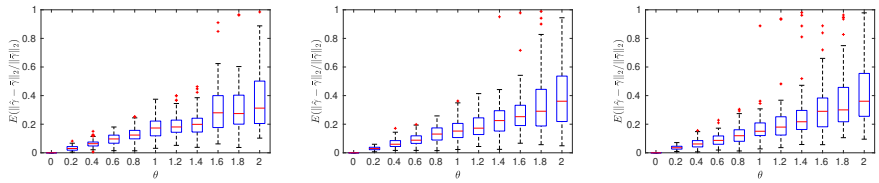
Once the family of operators is defined through the pairs of families (e_k) and (f_l), we can sample operators at random in this family by setting $\gamma \sim \mathcal{N}(0, \mathbf{I}_I)$. In Figure 6.7, we visualize a set of operators indirectly by applying them to a Dirac comb with 4 spikes. As can be seen, the operators are space-varying. Their impulse responses belong to $\text{span}(e_k, k \in \{1, \ldots, K\})$.

To assess the performance of the different algorithms, we evaluate the percentage of perfect recovery results with various values of L and N. We run the algorithms 100 times with random locations for the N spikes $\bar{x}_n$, with random weights $\bar{w}_n$ and with a random family (f_l). The recovery results are displayed in Figure 6.8. For the considered families, the nuclear norm relaxation performs very poorly, suggesting that the relaxation approaches suggested both for discrete and gridless problems might not be the best competitor. In comparison, the alternate minimization (Algorithm 9) with the spectral initialization from [Li+19] and the seemingly novel projected gradient descent (Algorithm 10) perform satisfactorily for a good range of values of L and N. Between both, the projected gradient descent seems to provide better results for a wider range of parameters. A theoretical analysis of this idea might be worth an exploration. Unfortunately, no algorithm is able to succeed systematically. This might be related to the fact that the random positions $(\bar{x}_n)$ are badly located position for instance.



(a) Family $\mathcal{A}_1$ (b) Family $\mathcal{A}_2$ (c) Family $\mathcal{A}_3$

Figure 6.5: Average localization error $\mathbb{E}(|\hat{x}-\bar{x}|)$ as a fraction of a pixel for different noise levels $\theta \in [0,1]$ and the three families $\mathcal{A}_1$, $\mathcal{A}_2$ and $\mathcal{A}_3$.



(a) $\theta = 0$ (b) $\theta = 1$ (c) $\theta = 2$

Figure 6.6: Boxplots of the relative errors $\frac{\|\hat{\gamma}-\bar{\gamma}\|_2}{\|\bar{\gamma}\|_2}$ using a single measurement for various noise levels.

In this setting, the condition for global injectivity (6.34) reads:

$$N \geq \frac{2KL-4}{K-2},$$

i.e. $N \geq 6L-4$ for $K=3$. We see the shortcomings of this rule in this experiment: perfect recovery does not always occur when this condition is satisfied, because the condition does not certify the success of an algorithm. And the algorithms manage to recover some operators when this condition is not met. However, it is clear that a necessary condition for identifiability is $N \geq L$, since otherwise, even the problem with known weights cannot be identified.

6.4.3 A 2D experiment

To end up this chapter, we briefly describe how the experiment from Figure 6.1 was carried out. We generated a family of product-convolution operators with astigmatic impulse responses as follows. We set the family (e_k) as anisotropic Gaussian vectors with $K=8$. We also set the family (f_l) as the monomials 1, x and y, resulting in $L=3$ basis elements to describe the space variations.

We launched the maximum correlation algorithm to locate the beads positions in Figure 6.1b. The average localization error is 0.015 pixel, despite a significant amount of additive Gaussian noise. We then discarded by hand the locations that were too close from each other (red stars). Notice that this part can be easily automatized by thresholding the minimal distance between adjacent locations. We kept the other locations (blue stars) to estimate the operator, resulting in a total of $N=27$ impulse responses with a slightly inaccurate localization. We then used this information to recover the operator. Here we assumed that the weight $(\bar{w}_n)$ were known and all equal to 1. The relative error between the operator estimated and the true one is $6e^{-3}$. Here, we measured the distance

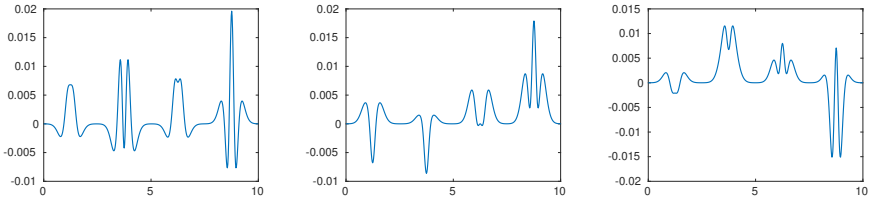
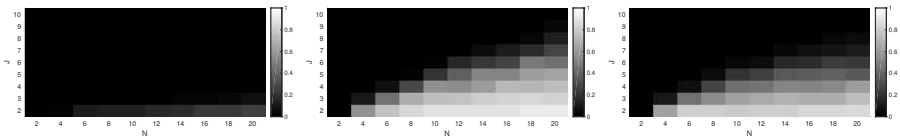


Figure 6.7: Examples of random product-convolution operators for a fixed family (e_k) and (f_l) and 3 random realizations of γ.



(a) Lifting & nuclear norm (b) Projected gradient (c) Alternate minimization

Figure 6.8: Percentages of perfect recovery results for the different algorithms.

between operators with the Hilbert-Schmidt norm. The whole process takes less than a second on a usual personal computer with Matlab.

Acknowledgments

The authors wish to thank Landry Duguet for his initial exploration of the problem during a one month internship. P. Weiss thanks T. Rezk for fruitful discussions.

6.5 Proofs

6.5.1 Proofs of the propositions from Section 6.2.3

Proof of Proposition 6.2.1

Proof. We have:

$$
\begin{aligned}
&\|\Pi_{R(x)}\Pi_{R(\bar{x})}\|_{2\to 2}^2 \\
&= \sup_{y\in\mathbb{R}^M, \|y\|_2=1} \|\Pi_{R(x)}\Pi_{R(\bar{x})}y\|_2^2 = \sup_{y\in\mathbb{R}^M, \|y\|_2=1} \langle \Pi_{R(x)}\Pi_{R(\bar{x})}y, \Pi_{R(x)}\Pi_{R(\bar{x})}y\rangle \\
&= \sup_{y\in\mathbb{R}^M, \|y\|_2=1} \langle \Pi_{R(x)}\Pi_{R(\bar{x})}y, y\rangle = \sup_{y\in\mathbb{R}^M, \|y\|_2=1} \langle \Pi_{R(\bar{x})}y, \Pi_{R(x)}y\rangle \\
&= \sup_{y\in R(x), \|y\|_2=1} \langle \Pi_{R(\bar{x})}y, y\rangle \leq \sup_{\substack{y\in R(x), \bar{y}\in R(\bar{x}) \\ \|y\|_2=1, \|\bar{y}\|_2=1}} \langle \bar{y}, y\rangle = \cos\left(\angle\left(R(x), R(\bar{x})\right)\right).
\end{aligned}
$$

□

Proof of Proposition 6.2.2

Proof. We have $[E(z)]_{m,i} = \langle e_i(\ \cdot\ - z), \delta_{z_m}\rangle = e_i(z_m - z)$. Hence

$$[E^*(z)E(z)]_{i,i'} = \sum_{m\in\mathbb{N}} e_i(z_m - z)e_{i'}(z_m - z) = \begin{cases} 1 & \text{if } i = i' \\ 0 & \text{otherwise,} \end{cases}$$

where we used (6.11) to obtain the last equality. □

Proof of Proposition 6.2.3

Proof. By Assumption 6.2.5, we have $\Pi_{R(x)} = E(x)E^*(x)$. Then, by definition, we have

$$\begin{aligned} \|\Pi_{R(x)}\Pi_{R(x')}\|_{2\to 2} &= \|E(x)E^*(x)E(x')E^*(x')\|_{2\to 2} \\ &= \|E^*(x)E(x')E^*(x')\|_{2\to 2} \\ &= \|E(x')E^*(x')E(x)\|_{2\to 2} \\ &= \|E^*(x')E(x)\|_{2\to 2} \\ &= \|\mathcal{M}(x - x')\|_{2\to 2}, \end{aligned}$$

using the fact that $E^*(x)E(x) = \mathbf{I}_I$. □

Proof of Proposition 6.2.4

Proof. The first part of the proof is trivial: the range $R(x)$ of $E(x)$ is unchanged.

As for the second part, it suffices to realize that E contains columns which are colinear. □

6.5.2 Proof of Theorem 6.3.1

Proof. For a fixed $X \in \mathbb{R}^{D\times N}$, any optimal coefficient matrix $\gamma(X)$ is characterized by the linear $I \times I$ system

$$\left(\sum_{n=1}^{N} \bar{w}_n^2 E^*(x_n)E(x_n)\right)\gamma = \sum_{n=1}^{N} \bar{w}_n E^*(x_n)y_n. \tag{6.37}$$

Letting

$$\mathcal{E}(X) = \begin{pmatrix} \bar{w}_1 E(x_1) \\ \vdots \\ \bar{w}_N E(x_N) \end{pmatrix} \text{ and } Y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix},$$

this system can be rewritten compactly as

$$\mathcal{E}^*(X)\mathcal{E}(X)\gamma = \mathcal{E}^*(X)Y.$$

Letting $\mathcal{E}^+(X)$ denote the Moore-Penrose pseudo-inverse of $\mathcal{E}(X)$, we can for instance choose $\gamma(X) = \mathcal{E}^+(X)Y$, leading to

$$\begin{aligned} G(X) &\overset{\text{def.}}{=} F(\gamma(X), X) = \frac{1}{2}\|[\mathcal{E}^+(X)\mathcal{E}^*(X) - \mathbf{I}]Y\|_2^2 = \frac{1}{2}\|[\Pi_{\mathrm{Ran}(\mathcal{E}(X))} - \mathbf{I}]Y\|_2^2 \\ &= \frac{1}{2}\|Y\|_2^2 - \frac{1}{2}\|\Pi_{\mathrm{Ran}(\mathcal{E}(X))}Y\|_2^2. \end{aligned}$$

by Pythagorean's theorem.

Hence, minimizing G amounts to maximizing the quadratic form

$$H(X) \stackrel{\text{def.}}{=} \frac{1}{2}\langle \Pi_{\mathrm{Ran}(\mathcal{E}(X))} Y, Y\rangle = \frac{1}{2}\sum_{n=1}^{N} \langle \Pi_{\mathrm{Ran}(E(x_n))} y_n, y_n\rangle,$$

which boils down to N independent subproblems. Under Assumption 6.2.3, the mapping $E(x)$ is injective for all $x \in \mathbb{R}^D$. Hence, $E^*(x)E(x)$ is a positive definite matrix and so is $\mathcal{E}^*(X)\mathcal{E}(X)$. This and the fact that $\bar{w}_n \neq 0$ justifies that $\gamma(X)$ is unique for any points configuration X. □

6.5.3 Proof of Theorem 6.3.2

We start with a simple lemma.

Lemma 6.5.1. *Let $D \in \mathbb{N}$, $f : \mathbb{R}^D \to \mathbb{R}$ and $\epsilon : \mathbb{R}^D \to \mathbb{R}$. Define $g \stackrel{\text{def.}}{=} f + \epsilon$ and assume that the following sets are non-empty*

$$\hat{X} = \operatorname*{argmax}_{x\in\mathbb{R}^D} g(x) \quad \text{and} \quad \bar{X} = \operatorname*{argmax}_{x\in\mathbb{R}^D} f(x).$$

Further assume that $\|\epsilon\|_\infty \leq \eta$ for some $\eta > 0$ and that there exists an increasing function $\varphi : \mathbb{R} \to \mathbb{R}_+$ such that

$$f(x) \leq f(\bar{x}) - \varphi(\|x - \bar{x}\|_2), \ \forall x \in \mathbb{R}^D. \tag{6.38}$$

Then $\bar{X} = \{\bar{x}\}$ is a singleton and any $\hat{x} \in \hat{X}$ satisfies $\|\hat{x} - \bar{x}\|_2 \leq \varphi_+^{-1}(2\eta)$.

Proof. By inequality (6.38) and strict monotonicity of ϕ, we have $f(\bar{x}) > f(x)$ for all $x \neq \bar{x}$. Hence $\bar{x}$ is the unique maximizer of f. We have

$$g(\hat{x}) \geq g(\bar{x}) = f(\bar{x}) - \epsilon(\bar{x}) \geq f(\bar{x}) - \eta. \tag{6.39}$$

In addition, for any $x \in \mathbb{R}^D$, we have

$$\begin{aligned} g(x) &= f(x) + \epsilon(x) \leq f(x) + \eta \\ &\leq f(\bar{x}) - \varphi(\|x - \bar{x}\|_2) + \eta. \end{aligned}$$

For any x such that $\|x - \bar{x}\|_2 > \varphi_+^{-1}(2\eta)$, we have $g(x) < f(\bar{x}) - \eta$, and thus $g(x) < g(\bar{x})$, which implies that $x \neq \hat{x}$. The contraposition is that any $\hat{x} \in \hat{X}$ satisfies $\|\hat{x} - \bar{x}\|_2 \leq \varphi_+^{-1}(2\eta)$. □

Proof. i) Let $y_{0,n} = \bar{w}_n E(\bar{x}_n)\bar{\gamma} = y_n - b_n$ denote the noiseless measurements and let $F_{0,n}(x, \gamma) = \frac{1}{2}\|\bar{w}_n E(x)\gamma - y_{0,n}\|_2^2$. We have $F_{0,n}(\bar{x}_n, \bar{\gamma}) = 0$. Let $\gamma_0(x) \in \operatorname{argmin}_{\gamma\in\mathbb{R}^I} F_{0,n}(x, \gamma)$ denote any minimizer (for instance the one given by the pseudo-inverse). Now, define $G_{0,n}(x) \stackrel{\text{def.}}{=} F_{0,n}(x, \gamma_0(x))$. We have $G_{0,n}(x) = \frac{1}{2}\left(\|y_{0,n}\|_2^2 - \|\Pi_{R(x)} y_{0,n}\|_2^2\right)$, by an argument similar to the one in the proof of Theorem 6.3.1. By Assumption 6.2.4, $R(x) \cap R(\bar{x}) = \{0\}$ for $x \neq \bar{x}$. Hence, $\|\Pi_{R(x)} y_{0,n}\|_2^2 < \|y_{0,n}\|_2^2$ for $x \neq \bar{x}$ and $\bar{x}_n$ is the unique minimizer of $G_{0,n}$. Therefore, the function

$$H_{0,n}(x) \stackrel{\text{def.}}{=} \frac{1}{2}\|\Pi_{R(x)} y_{0,n}\|_2^2 = \frac{1}{2}\left\langle \Pi_{R(x)} y_{0,n}, y_{0,n}\right\rangle,$$

also admits a unique maximizer in $\bar{x}_n$. Overall, we see that under Assumption 6.2.4, $G_{0,n}$ admits a unique minimizer equal to $\bar{x}_n$. Under the additional Assumption 6.2.3, $F_{0,n}$ admits a unique solution $(\bar{x}_n, \bar{\gamma})$.

ii) Now, let $F_n(x,\gamma) = \frac{1}{2}\|\bar{w}_n E(x)\gamma - y_{0,n}\|_2^2$, $\gamma(x)$ denote any minimizer of F_n w.r.t. γ, $G_n(x) = F_n(x,\gamma(x))$ and $H_n(x) \overset{\text{def.}}{=} \frac{1}{2}\langle \Pi_{R(x)} y_n, y_n\rangle$. Let $\hat{x}_n$ denote any maximizer of H_n and assume for now that we manage to obtain a bound of the form $|H_n(x) - H_{0,n}(x)| \leq \eta$ for some $\eta \geq 0$. We have

$$\begin{aligned} H_{0,n}(x) &= \frac{1}{2}\left\langle \Pi_{R(x)} y_{0,n}, y_{0,n}\right\rangle = \frac{1}{2}\left\langle \Pi_{R(x)} y_{0,n}, \Pi_{R(\bar{x}_n)} y_{0,n}\right\rangle \\ &= \frac{1}{2}\left\langle \Pi_{R(\bar{x}_n)}\Pi_{R(x)} y_{0,n}, y_{0,n}\right\rangle \leq \frac{1}{2}\left\|\Pi_{R(\bar{x}_n)}\Pi_{R(x)}\right\|_{2\to 2}\|y_{0,n}\|_2^2 \\ &\leq \frac{1}{2}\left(1 - \phi(\|x - \bar{x}_n\|_2)\right)\|y_{0,n}\|_2^2 = H_{0,n}(\bar{x}_n) - \frac{1}{2}\phi(\|x-\bar{x}_n\|_2)\|y_{0,n}\|_2^2. \end{aligned}$$

by Assumption 6.2.4. Hence, we can use Lemma 6.5.1 with $f(x) = H_{0,n}(x)$, $g(x) = H_n(x)$ and $\varphi(r) = \frac{1}{2}\phi(r)\|y_{0,n}\|_2^2$. This allows us to conclude that

$$\|\hat{x}_n - \bar{x}_n\|_2 \leq \phi_+^{-1}\left(\frac{4\eta}{\|y_{0,n}\|_2^2}\right). \tag{6.40}$$

iii) The last remaining point is to control $\|H_{0,n} - H_n\|_\infty$. We have

$$\begin{aligned} H_n(x) &= \frac{1}{2}\langle \Pi_{R(x)}\left(y_{0,n} + b_n\right), y_{0,n} + b_n\rangle \\ &= H_{0,n}(x) + \langle \Pi_{R(x)} y_{0,n}, b_n\rangle + \frac{1}{2}\|\Pi_{R(x)}(b_n)\|_2^2. \end{aligned}$$

Hence, using Cauchy-Schwarz inequality, we obtain for all x

$$|H_n(x) - H_{0,n}(x)| \leq \|\Pi_{R(x)} y_{0,n}\|_2 \|\Pi_{R(x)} b_n\|_2 + \frac{1}{2}\|\Pi_{R(x)}(b_n)\|_2^2.$$

Using the facts that $\|\Pi_{R(x)} b_n\|_2 \leq \|b_n\|_2$ and that $\|b_n\|_2 \leq \theta\|y_{0,n}\|_2$, we obtain

$$\|H_n - H_{0,n}\|_\infty \leq \|y_{0,n}\|_2^2\left(\theta + \frac{1}{2}\theta^2\right).$$

For the inequality (6.40) to make sense, we need $4(\theta + \frac{1}{2}\theta^2) \leq 1$, which is equivalent to $\theta < \frac{\sqrt{6}}{2} - 1$ and Theorem 6.3.2 is proven. □

6.5.4 Proof of Proposition 6.3.3

Proof. We remind that $N = 1$ and that we discard the indices n. This proposition derives from point iii) in the previous proof. The inequality (6.40) can be improved as

$$\|\hat{x} - \bar{x}\|_2 \leq \phi_+^{-1}\left(\frac{4\inf_{c\in\mathbb{R}}\|H - H_0 - c\|_\infty}{\|y_0\|_2^2}\right), \tag{6.41}$$

since a constant term does not affect the location of a minimizer. We have

$$\Delta(x) \overset{\text{def.}}{=} H(x) - H_0(x) = \langle \Pi_{R(x)} y_0, \Pi_{R(x)} b\rangle + \frac{1}{2}\|\Pi_{R(x)}(b)\|_2^2.$$

It is therefore natural to set $c = \frac{1}{2}\left(\sup_{x\in\mathbb{R}^D}\Delta(x) - \inf_{x\in\mathbb{R}^D}\Delta(x)\right)$, to minimize in the infinite norm in Problem (6.41). □

6.5.5 Proof of Proposition 6.3.4

Proof. First notice that

$$\begin{aligned}\mathrm{Ampl}(\Delta_1+\Delta_2) &= \sup_{x\in\mathbb{R}^D} \Delta_1(x)+\Delta_2(x) - \inf_{x\in\mathbb{R}^D} \Delta_1(x)+\Delta_2(x)\\ &\le \sup_{x\in\mathbb{R}^D} \Delta_1(x) + \sup_{x\in\mathbb{R}^D} \Delta_2(x) - \left(\inf_{x\in\mathbb{R}^D} \Delta_1(x) + \inf_{x\in\mathbb{R}^D} \Delta_2(x)\right)\\ &= \mathrm{Ampl}(\Delta_1)+\mathrm{Ampl}(\Delta_2).\end{aligned}$$

We will treat the two random processes Δ_1 and Δ_2 separately.

i) Consider the function $f_1(b) \stackrel{\text{def.}}{=} \sup_{x\in\mathbb{R}^D}\langle \Pi_{R(x)}y_0, b\rangle$ and define the random variable $V_1^+ = f_1(b)$ with mean $\bar{V}_1^+$. Similarly, define $V_1^- = \inf_{x\in\mathbb{R}^D}\langle \Pi_{R(x)}y_0, b\rangle$. In addition, notice that $Z_1 = \mathrm{Ampl}(\Delta_1) \le V_1^+ - V_1^-$. We first show that f_1 is Lipschitz continuous. We have

$$\begin{aligned} f_1(b+\epsilon) &= \sup_{x\in\mathbb{R}^D}\langle \Pi_{R(x)}y_0, b+\epsilon\rangle \le \sup_{x\in\mathbb{R}^D}\langle \Pi_{R(x)}y_0, b\rangle + \|y_0\|_2\|\epsilon\|_2\\ f_1(b+\epsilon) &= \sup_{x\in\mathbb{R}^D}\langle \Pi_{R(x)}y_0, b+\epsilon\rangle \ge \sup_{x\in\mathbb{R}^D}\langle \Pi_{R(x)}y_0, b\rangle - \|y_0\|_2\|\epsilon\|_2\end{aligned}$$

Hence, $|f_1(b) - f_1(b+\epsilon)| \le \|y_0\|_2\|\epsilon\|_2$ and f_1 is $\|y_0\|_2$-Lipschitz continuous. Using a Gaussian logarithmic Sobolev inequality [BLM13, Thm 5.6], we obtain that V_1^+ is sub-Gaussian with

$$\mathbb{P}\left(|V_1^+ - \bar{V}_1^+| \ge t\right) \le 2\exp(-t^2/(2\sigma^2\|y_0\|^2)).$$

The same result holds for V_1^-. Finally, the sum of two dependent sub-Gaussian variables with parameters σ_1 and σ_2 is sub-Gaussian with a sub-Gaussian parameter smaller than $\sigma_1+\sigma_2$, so that

$$\mathbb{P}\left(|Z_1 - \bar{Z}_1| \ge t\right) \le 2\exp(-t^2/(8\sigma^2\|y_0\|^2)).$$

ii) Now, define the random variable $Y_2^+ \stackrel{\text{def.}}{=} \sup_{x\in\mathbb{R}^D}\|\Pi_{R(x)}b\|_2$ and $V_2^+ \stackrel{\text{def.}}{=} \frac{1}{2}(Y_2^+)^2$. The function $b \mapsto \sup_{x\in\mathbb{R}^D}\|\Pi_{R(x)}b\|_2$ is 1-Lipschitz continuous. Hence using a Gaussian logarithmic Sobolev inequality again, we obtain that Y_2^+ is sub-Gaussian with

$$\mathbb{P}\left(|Y_2^+ - \bar{Y}_2^+| \ge t\right) \le 2\exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Using [Ver18, Lemma 2.7.6], we conclude that Z_2^+ is sub-exponential and satisfies

$$\mathbb{P}\left(|V_2^+ - \bar{V}_2^+| \ge t\right) \le 2\exp(-Ct/\sigma),$$

for an absolute constant C. We can make a similar proof for the random variable $Y_2^- \stackrel{\text{def.}}{=} \inf_{x\in\mathbb{R}^D}\|\Pi_{R(x)}b\|_2$ and conclude as in the previous proof. □

6.5.6 Proof of Theorem 6.3.5

Controlling $\bar{Z}_1$. Here, we wish to control the supremum of the centered Gaussian process Δ_1. A traditional approach to bound it consists in computing Dudley's entropy integral, see e.g. [BLM13, Corollary 13.2]. To this end, we introduce the pseudo-metric:

$$d(x,x') \stackrel{\text{def.}}{=} \sqrt{\mathbb{E}\left((\Delta_1(x)-\Delta_1(x'))^2\right)}. \tag{6.42}$$

Let $B(c,\delta)=\{x\in\mathbb{R}^D, d(c,x)\leq\delta\}$ denote a ball of radius δ centered at c with respect to d. Let $S\subset\mathbb{R}$ denote a set and define the covering number $N(\delta,S)$ as the minimum number of δ-balls needed to cover S. We then have

$$\mathbb{E}\left(\sup_{x\in\mathbb{R}}\Delta_1(x)\right)\leq 12\int_0^{\eta/2}\sqrt{\log(N(u,\mathbb{R}))}\,du \text{ with } \eta=\inf_{t\in\mathbb{R}}\sup_{t'\in\mathbb{R}}d(t,t'). \tag{6.43}$$

Using that $y_0\in R(\bar{x})$, we have

$$\begin{aligned} d(x,x')^2 &= \mathbb{E}\left((\Delta_1(x)-\Delta_1(x'))^2\right)\\ &= \mathbb{E}\left((\langle\left(\Pi_{R(x)}-\Pi_{R(x')}\right)\Pi_{R(\bar{x})}y_0,b\rangle)^2\right)\\ &\leq \sigma^2\|y_0\|_2^2\|\left(\Pi_{R(x)}-\Pi_{R(x')}\right)\Pi_{R(\bar{x})}\|_{2\to2}^2. \end{aligned}$$

Thus

$$\begin{aligned} \|\left(\Pi_{R(x)}-\Pi_{R(x')}\right)\Pi_{R(\bar{x})}\|_{2\to2} &\leq \|\Pi_{R(x)}\Pi_{R(\bar{x})}\|_{2\to2}+\|\Pi_{R(x')}\Pi_{R(\bar{x})}\|_{2\to2}\\ &\leq \frac{2}{1+\min\left(\|x-\bar{x}\|_2,\|x'-\bar{x}\|_2\right)^\alpha} \end{aligned}$$

This leads to

$$d(x,x')\leq\frac{2\sigma\|y_0\|_2}{1+\min\left(\|x-\bar{x}\|_2,\|x'-\bar{x}\|_2\right)^\alpha} \tag{6.44}$$

and the Lipschitz continuity also implies that

$$d(x,x')\leq\sigma\|y_0\|_2L\|x-x'\|_2. \tag{6.45}$$

In what follows, we let

$$\tilde{d}(x,x')\stackrel{\text{def.}}{=}\sigma\|y_0\|_2\min\left(L\|x-x'\|_2,\frac{2}{1+\min\left(\|x-\bar{x}\|_2,\|x'-\bar{x}\|_2\right)^\alpha}\right) \tag{6.46}$$

and $\tilde{N}$ denote the corresponding covering number. The inequality $d(x,x')\leq\tilde{d}(x,x')$ implies that $N(\delta,S)\leq\tilde{N}(\delta,S)$ for all S and δ.

Without loss of generality, we assume that $\bar{x}=0$. Now, let $c=\left(\frac{2\sigma\|y_0\|_2}{\delta}\right)^{1/\alpha}$ and remark that the decay condition (6.44) implies that all x with $|x|\geq c$ belong to the ball $B(c,\delta)$. Hence

$$\tilde{N}(\delta,\mathbb{R})\leq 1+\tilde{N}(\delta,[-c,c]).$$

The second inequality (6.45) implies that the δ-balls have a diameter no smaller than $\frac{2\delta}{L\sigma\|y_0\|_2}$. Hence, the interval $[-c,c]$ is covered by at most $\left\lceil 2c\cdot\frac{L\sigma\|y_0\|_2}{2\delta}\right\rceil$ δ-balls, leading to

$$N(\delta,\mathbb{R})\leq\tilde{N}(\delta,\mathbb{R})\leq 2^{\frac{1}{\alpha}}L\left(\frac{2\sigma\|y_0\|_2}{\delta}\right)^{\frac{1}{\alpha}+1}+2.$$

We have $\eta = \inf_{t\in\mathbb{R}} \sup_{t'\in\mathbb{R}} d(t,t') \leq 2\sigma\|y_0\|_2$. Hence:

$$\mathbb{E}\left(\sup_{x\in\mathbb{R}} \Delta_1(x)\right)$$
$$\leq 12 \int_0^{\sigma\|y_0\|_2} \sqrt{\log\left(2^{1/\alpha} L \left(\frac{\sigma\|y_0\|_2}{u}\right)^{\frac{1}{\alpha}+1} + 2\right)} du.$$

For all $u \in [0, \sigma\|y_0\|_2]$, we have

$$2^{1/\alpha} L \left(\frac{\sigma\|y_0\|_2}{u}\right)^{\frac{1}{\alpha}+1} + 2 \leq \left(1 + \frac{2}{2^{\frac{1}{\alpha}} L}\right) 2^{\frac{1}{\alpha}} L \left(\frac{\sigma\|y_0\|_2}{u}\right)^{\frac{1}{\alpha}+1}.$$

Hence

$$\mathbb{E}\left(\sup_{x\in\mathbb{R}} \Delta_1(x)\right) \leq 12 \int_0^{\sigma\|y_0\|_2} \sqrt{\log\left[\left(2^{\frac{1}{\alpha}} L + 2\right)\left(\frac{\sigma\|y_0\|_2}{u}\right)^{\frac{1}{\alpha+1}}\right]} du$$
$$\leq 6\sqrt{\pi}\sigma\|y_0\|_2 \sqrt{\frac{\alpha+1}{\alpha}} \left(2^{\frac{1}{\alpha}} L + 2\right)^{\frac{\alpha}{\alpha+1}},$$

where we skipped the elementary (but ugly) technical details to obtain the last bound. To conclude, we use the fact that $\bar{Z}_1 \leq 2\mathbb{E}\left(\sup_{x\in\mathbb{R}} \Delta_1(x)\right)$. □

Controlling $\bar{Z}_2$. The fact that $R(x)$ is constant outside $[-1,1]$ allows to restrict our study to this interval. Similarly to the proof of Proposition 6.3.4, consider the random variable

$$Y_2^+ = \sup_{x\in[-1,1]} \|\Pi_{R(x)} b\|_2.$$

We have $\mathbb{E}(\|\Pi_{R(x)} b\|_2) \leq \sigma\sqrt{I}$ for all $x \in [-1,1]$. By the Lipschitz continuity assumption on $x \mapsto \Pi_{R(x)}$, we can use the natural distance $d(x,x') = \sigma L\|x - x'\|_2$, leading to the following upper-bound on the covering number $N(\delta, [-1,1]) \leq \frac{L\sigma}{\delta} + 1$. Using Dudley's entropy integral, we obtain

$$\mathbb{E}(Y_2^+) - \sigma\sqrt{I} \leq 12 \int_0^{L\sigma/2} \sqrt{\log\left(\frac{L\sigma}{u} + 1\right)} du \leq CL\sigma$$

for some absolute constant C. Therefore

$$(\bar{Y}_2^+)^2 \leq C^2 L^2 \sigma^2 + \sigma^2 I + 2CL\sigma^2\sqrt{I}.$$

We want to control

$$\bar{V}_2^+ = \frac{1}{2}\mathbb{E}\left(\sup_{x\in[-1,1]} \|\Pi_{R(x)} b\|_2^2\right) = \frac{1}{2}\mathbb{E}\left((Y_2^+)^2\right)$$

The random variable Y_2^+ is sub-Gaussian with parameter σ by the proof of Proposition 6.3.4, so that

$$\mathrm{Var}(Y_2^+) = \mathbb{E}\left(Y_2^+ - \bar{Y}_2^+\right)^2 = \mathbb{E}((Y_2^+)^2) - (\bar{Y}_2^+)^2 \leq C''\sigma^2,$$

where the last inequality is due to [Ver18, Prop. 2.5.2, ii)]. We have

$$\bar{V}_2^+ = \frac{1}{2}\left(\mathbb{E}((Y_2^+)^2)\right) \leq \frac{1}{2}\left((\bar{Y}_2^+)^2 + C''\sigma^2\right) \leq \sigma^2\left(I + C'' + C^2L^2 + 2CL\sqrt{I}\right).$$

To control the average amplitude of Δ_2, we still need to control $\bar{V}_2^- = \mathbb{E}\left(\inf_{x\in[-1,1]} \|\Pi_{R(x)}b\|_2^2\right)$. To this end, let $Y_2^- \overset{\text{def.}}{=} \inf_{x\in[-1,1]} \|\Pi_{R(x)}b\|_2$. Using the same approach as before, we obtain:

$$\bar{Y}_2^- \geq \sigma\sqrt{I} - CL\sigma \quad \Rightarrow \quad (Y_2^-)^2 \geq \sigma^2\left(I + C^2L^2 - 2CL\sqrt{I}\right)$$

and

$$(V_2^-)^2 = (\bar{Y}_2^-)^2 + \mathrm{Var}(Y_2^-) \geq \sigma^2\left(I + C^2L^2 - 2CL\sqrt{I}\right).$$

This leads to $\bar{Z}_2 \leq \sigma^2\left(C'' + 4CL\sqrt{I}\right)$. □

6.5.7 Proof of Theorem 6.3.6

Proof. Let $P(x) \overset{\text{def.}}{=} E^*(x)E(x)$. By definition, we have

$$\hat{\gamma} = P(\hat{x})^{-1}E^*(\hat{x})(y_{0,1} + b_1) \tag{6.47}$$

We have $E(\hat{x}) = E(\bar{x}) + \Delta$ with $\|\Delta\|_{2\to2} \leq \sqrt{\sigma_+}L_E\|\hat{x} - \bar{x}\|_2$. Hence $P(\hat{x}) = P(\bar{x}) + \Delta'$ with

$$\|\Delta'\|_{2\to2} \leq 2L_E\|\hat{x} - \bar{x}\|_2\sqrt{\sigma_+}\|E(\bar{x})\|_{2\to2} + \sigma_+L_E^2\|\hat{x} - \bar{x}\|_2^2. \tag{6.48}$$

The linear system to recover $\hat{\gamma}$ is then

$$(P(\bar{x}) + \Delta')\hat{\gamma} = (E^*(\bar{x}) + \Delta)\,(y_{0,1} + b_1) = E^*(\bar{x})y_{0,1} + \delta \tag{6.49}$$

with $\delta = \Delta(y_{0,1} + b_1) + E^*(\bar{x})b_1$. Under Assumption 6.2.3, the unique solution of $P(\bar{x})\gamma = E^*(\bar{x})y_{0,1}$ is $\bar{\gamma}$. If $\hat{x}$ is sufficiently close to $\bar{x}$, we have $\|\Delta'\|_{2\to2} < \sigma_-$ and $\|P(\bar{x})^{-1}\Delta'\|_{2\to2} < 1$. We can now use standard results of linear algebra, see e.g. [Tyr12, p. 3.6], to obtain that

$$\begin{aligned}
&\frac{\|\hat{\gamma} - \bar{\gamma}\|_2}{\|\bar{\gamma}\|_2} \\
&\leq \frac{\|P(\bar{x})\|_{2\to2}\|P(\bar{x})^{-1}\|_{2\to2}}{1 - \|P(\bar{x})^{-1}\Delta'\|_{2\to2}}\left(\frac{\|\Delta'\|_{2\to2}}{\|P(\bar{x})\|_{2\to2}} + \frac{\|\delta\|_2}{\|E^*(\bar{x})y_{0,1}\|_2}\right) \\
&\leq \frac{\sigma_+}{\sigma_-}\frac{1}{\left(1 - \frac{\|\Delta'\|_{2\to2}}{\sigma_-}\right)}\left(\frac{\|\Delta'\|_{2\to2}}{\sigma_-} + \frac{\|\Delta\|_{2\to2}(\|y_{0,1}\|_2 + \|b_1\|_2) + \sqrt{\sigma_+}\|b_1\|_2}{\sqrt{\sigma_-}\|y_{0,1}\|_2}\right)
\end{aligned}$$

Letting $\hat{x} \to \bar{x}$, we have $\frac{1}{\left(1 - \frac{\|\Delta'\|_{2\to2}}{\sigma_-}\right)} \sim 1 + \frac{\|\Delta'\|_{2\to2}}{\sigma_-}$.

Using that

$$\|\Delta'\|_{2\to2} \leq 2L_E\sqrt{\sigma_+}\|E(\bar{x})\|_{2\to2}\|\hat{x} - \bar{x}\|_2 + o_{\hat{x}\to\bar{x}}\left(\|\hat{x} - \bar{x}\|_2^2\right),$$

and

$$\|\Delta\|_{2\to2} \leq L_E\sqrt{\sigma_+}\|\hat{x} - \bar{x}\|_2,$$

we have

$$
\begin{aligned}
&\frac{\sigma_+}{\sigma_-}\frac{1}{\left(1-\frac{\|\Delta'\|_{2\to 2}}{\sigma_-}\right)}\left(\frac{\|\Delta'\|_{2\to 2}}{\sigma_-}+\frac{\|\Delta\|_{2\to 2}(\|y_{0,1}\|_2+\|b_1\|_2)+\sqrt{\sigma_+}\|b_1\|_2}{\sqrt{\sigma_-}\|y_{0,1}\|_2}\right)\\
&=\frac{\sigma_+}{\sigma_-}\left(1+\frac{\|\Delta'\|_{2\to 2}}{\sigma_-}+o_{\hat{x}\to\bar{x}}\left(\|\hat{x}-\bar{x}\|_2^2\right)\right)\Big(\frac{\|\Delta'\|_{2\to 2}}{\sigma_-}\\
&+\frac{\|\Delta\|_{2\to 2}(\|y_{0,1}\|_2+\|b_1\|_2)+\sqrt{\sigma_+}\|b_1\|_2}{\sqrt{\sigma_-}\|y_{0,1}\|_2}\Big)\\
&\le \kappa\|\hat{x}-\bar{x}\|_2\Big(\frac{2L_E\sqrt{\sigma_+}\|E(\bar{x})\|_{2\to 2}}{\sigma_-}+\frac{L_E\sqrt{\sigma_+}}{\sqrt{\sigma_-}}(1+\frac{\|b_{0,1}\|_2}{\|y_{0,1}\|_2})\\
&+\sqrt{\kappa}\frac{\|b_{0,1}\|_2}{\|y_{0,1}\|_2}2L_E\frac{\sqrt{\sigma_+}\|E(\bar{x})\|_{2\to 2}}{\sigma_-}\Big)+\kappa^{3/2}\frac{\|b_1\|_2}{\|y_{0,1}\|_2}+o_{\hat{x}\to\bar{x}}\left(\|\hat{x}-\bar{x}\|_2^2\right)\\
&\le C\kappa^{5/2}L_E\|\hat{x}-\bar{x}\|_2\left(1+\frac{\|b_1\|_2}{\|y_{0,1}\|_2}\right)+o_{\hat{x}\to\bar{x}}\left(\|\hat{x}-\bar{x}\|_2^2\right)
\end{aligned}
$$

since $\kappa \ge 1$, and where C denote a universal constant. Let

$$
\epsilon_2 \overset{\text{def.}}{=} C\kappa^{5/2}L_E\|\hat{x}-\bar{x}\|_2\left(1+\frac{\|b_1\|_2}{\|y_{0,1}\|_2}\right)+o_{\hat{x}\to\bar{x}}\left(\|\hat{x}-\bar{x}\|_2^2\right),
$$

this concludes the proof. □

6.5.8 Proof of Theorem 6.3.7

Proof. Let

$$
\mathcal{E}(X)=\begin{pmatrix}\bar{w}_1E(x_1)\\ \vdots\\ \bar{w}_NE(x_N)\end{pmatrix},\ Y_0=\begin{pmatrix}y_{0,1}\\ \vdots\\ y_{0,N}\end{pmatrix}\ \text{and}\ B=\begin{pmatrix}b_1\\ \vdots\\ b_N\end{pmatrix}.
$$

By definition, we have

$$
\hat{\gamma}=\mathcal{C}(\hat{X})^{-1}\left(\mathcal{E}^*(\hat{X})Y_0+\mathcal{E}^*(\hat{X})B\right) \tag{6.50}
$$

We have

$$
\begin{aligned}
\|\mathcal{E}(\hat{X})-\mathcal{E}(\bar{X})\|_{2\to 2}^2 &\le \sum_{n=1}^N \bar{w}_n^2\|E(\hat{x}_n)-E(\bar{x}_n)\|_{2\to 2}^2\\
&\le w_+^2\hat{\sigma}_+L_E^2\|\hat{X}-\bar{X}\|_2^2\\
&\le \tilde{\sigma}_+L_E^2\|\hat{X}-\bar{X}\|_2^2.
\end{aligned}
$$

Hence, we have $\mathcal{E}(\hat{X}) = \mathcal{E}(\bar{X}) + \Delta$ with $\|\Delta\|_{2\to 2} \le \sqrt{\tilde{\sigma}_+}L_E\|\hat{X}-\bar{X}\|_2$ and $\mathcal{C}(\bar{X}) = \mathcal{C}(\hat{X}) + \Delta'$ with

$$
\|\Delta'\|_{2\to 2} \le 2\|\Delta\|_{2\to 2}\|\mathcal{E}(\hat{X})\|_{2\to 2}+\|\Delta\|_{2\to 2}^2. \tag{6.51}
$$

Under Assumption (6.27), the linear system $\mathcal{C}(\hat{X})\gamma = \mathcal{E}^*(\hat{X})(Y_0 + B)$ admits a unique solution $\hat{\gamma}$. Under Assumption (6.27) again, and given that $\hat{X}$ is sufficiently close to $\bar{X}$ so that $\|\Delta'\|_{2\to 2} < \tilde{\sigma}_-$, the linear system $\mathcal{C}(\bar{X})\gamma = \mathcal{E}^*(\bar{X})Y_0$ admits $\bar{\gamma}$ as a unique solution. In addition $\|\mathcal{C}(\hat{X})^{-1}\Delta'\|_{2\to 2} < 1$.

Let $\delta = \mathcal{E}^*(\hat{X})(Y_0 + B) - \mathcal{E}^*(\bar{X})Y_0$ denote the residual of the right hand-side term between the two previous linear systems. We have

$$\|\delta\|_2 \leq \|\Delta\|_{2\to 2}\|Y_0\|_2 + \sqrt{\tilde{\sigma}_+}\|B\|_2,$$

provided that $\|\mathcal{E}(\hat{X})\|_{2\to 2} \leq \sqrt{\tilde{\sigma}_+}$.

We can now use standard results of linear algebra, see e.g. [Tyr12, p. 3.6], to obtain that

$$\begin{aligned}
&\frac{\|\hat{\gamma} - \bar{\gamma}\|_2}{\|\bar{\gamma}\|_2} \\
&\leq \frac{\|\mathcal{C}(\hat{X})\|_{2\to 2}\|\mathcal{C}(\hat{X})^{-1}\|_{2\to 2}}{1 - \|\mathcal{C}(\hat{X})^{-1}\Delta'\|_{2\to 2}} \left(\frac{\|\Delta'\|_{2\to 2}}{\|\mathcal{C}(\hat{X})\|_{2\to 2}} + \frac{\|\delta\|_2}{\|\mathcal{E}^*(\hat{X})\,(Y_0 + B)\,\|_2} \right) \\
&\leq \frac{\tilde{\sigma}_+}{\tilde{\sigma}_-} \frac{1}{\left(1 - \frac{\|\Delta'\|_{2\to 2}}{\tilde{\sigma}_-}\right)} \left(\frac{2\sqrt{\tilde{\sigma}_+}\|\Delta\|_{2\to 2} + \|\Delta\|_{2\to 2}^2}{\tilde{\sigma}_-} + \frac{\|\Delta\|_{2\to 2}\|Y_0\|_2 + \sqrt{\tilde{\sigma}_+}\|B\|_2}{\sqrt{\tilde{\sigma}_-}\|Y_0 + B\|_2} \right) \\
&= \frac{\tilde{\sigma}_+}{\tilde{\sigma}_-} \frac{1}{\left(1 - \frac{\|\Delta'\|_{2\to 2}}{\tilde{\sigma}_-}\right)} \left(\frac{2\sqrt{\tilde{\sigma}_+}\|\Delta\|_{2\to 2} + \|\Delta\|_{2\to 2}^2}{\tilde{\sigma}_-} + \frac{\|\Delta\|_{2\to 2}\|Y_0\|_2}{\sqrt{\tilde{\sigma}_-}\|Y_0 + B\|_2} + \frac{\sqrt{\tilde{\sigma}_+}}{\sqrt{\tilde{\sigma}_-}} \frac{\|B\|_2}{\|Y_0 + B\|_2} \right)
\end{aligned}$$

Similarly to the proof of Theorem 6.3.6, by letting $\hat{X} \to \bar{X}$, we can decompose previous equation in two terms:

$$\begin{aligned}
&\frac{\|\hat{\gamma} - \bar{\gamma}\|_2}{\|\bar{\gamma}\|_2} \\
&\leq \tilde{\kappa}^{3/2} \frac{\|B\|_2}{\|Y_0\|_2} + \epsilon_2(\hat{X}),
\end{aligned}$$

where

$$\epsilon_2(\hat{X}) \stackrel{\text{def.}}{=} C\tilde{\kappa}^{5/2} L_E \|\bar{X} - \hat{X}\|_2 \left(1 + \frac{\|Y_0\|_2 + \|B\|_2}{\|Y_0 + B\|_2}\right) + o_{\hat{X}\to\bar{X}}\left(\|\bar{X} - \hat{X}\|_2^2\right),$$

for some absolute constant C. □

Chapter 7

The blind sparse+smooth (BSS) algorithm

Résumé : *Dans ce chapitre, nous présentons une méthode de défloutage aveugle sur un exemple de microscopie. La résolution de ce problème est possible grâce à la combinaison des différents résultats présentés dans les chapitres précédents.*

Abstract: *In this chapter, we present a blind deblurring method on an example of microscopy. The resolution of this problem is possible thanks to the combination of the different results presented in the previous chapters.*

Contents

7.1 Method **149**
7.1.1 Identifying the operator 150
7.1.2 Sparse+Smooth-deblurring 151
7.2 Real-life experiment **152**

This chapter is the natural application of the subspace estimation presented in Chapter 5. It introduced the BSS algorithm originaly published in [Deb+20a]:
Debarnot, V., Escande, P., Mangeat, T., & Weiss, P. (2020). Learning low-dimensional models of microscopes, IEEE Transactions on Computational Imaging.

7.1 Method

Here, we propose a method called BSS-deblurring, where BSS stands for Blind Sparse+Smooth. Given a blurred image $\mathbf{u}_0$, the method provides an estimate of the associated operator and a deblurred image. It consists of two separate steps: first the operator is estimated using isolated micro-beads in the image. This estimate is then used inside an original variational problem.

7.1.1 Identifying the operator

The following approach can be seen as a direct application of the identification theory presented in Chapter 6. We use an additional heuristic to incorporate a regularization on the known subspace of operators.

We assume that the user is able to select P patches out of $\mathbf{u}_0$ supported on $(\omega_p)_{1\leq p\leq P}$ that contain isolated micro-beads. The patches $\mathbf{p}_p$ are assumed to be well centered and without background, which can be achieved using results of Chapter 3.

Our aim is to estimate a discrete operator $\mathbf{H} : \mathbb{R}^N \to \mathbb{R}^N$ of the form:

$$\mathbf{H}\mathbf{u} = \sum_{1\leq i\leq I} \sum_{1\leq j\leq J} \mathbf{\Gamma}[i,j]\mathbf{h}_i \star (\boldsymbol{\alpha}_j \odot \mathbf{u})$$

where the pairs of orthogonal bases $((\mathbf{h}_i)_i, (\boldsymbol{\alpha}_j)_j)$ are known and the coefficients $(\mathbf{\Gamma}[i,j])_{i,j}$ are unknown. These notations are borrowed from Chapter 5 and we let the reader to refer to it for more details on how to obtain these quantities.

By projecting the patch $\mathbf{p}_p$ onto the PSFs basis $(\mathbf{h}_i)_i$, we obtain the coefficient $c_{p,i} = \langle \mathbf{p}_p, \mathbf{h}_i \rangle$ which is a noisy estimate of the interpolation map $\boldsymbol{\alpha}_i$ at the position $\mathbf{z}_p$:

$$\mathbf{h}\delta_{\mathbf{z}_p} = \sum_{1\leq i\leq I} \sum_{1\leq j\leq J} \mathbf{\Gamma}[i,j]\boldsymbol{\alpha}_i(\mathbf{z}_p)\mathbf{h}_i \approx \sum_{1\leq i\leq I} c_{p,i}\mathbf{h}_i.$$

We propose to identify $\mathbf{\Gamma}$ by solving the following bi-linear inverse problem:

$$\underset{\substack{\mathbf{\Gamma}\in\mathcal{C} \\ \langle \mathbf{\Gamma},\mathbf{\Gamma}_0\rangle=1 \\ \mathbf{g}\in\mathbb{R}_+^P}}{\operatorname{argmin}} \frac{1}{2} \sum_{1\leq p\leq P} \left\| \sum_{1\leq i\leq I} \sum_{1\leq j\leq J} g_p \mathbf{\Gamma}_{i,j}\boldsymbol{\alpha}_j - c_{p,i} \right\|_2^2, \tag{7.1}$$

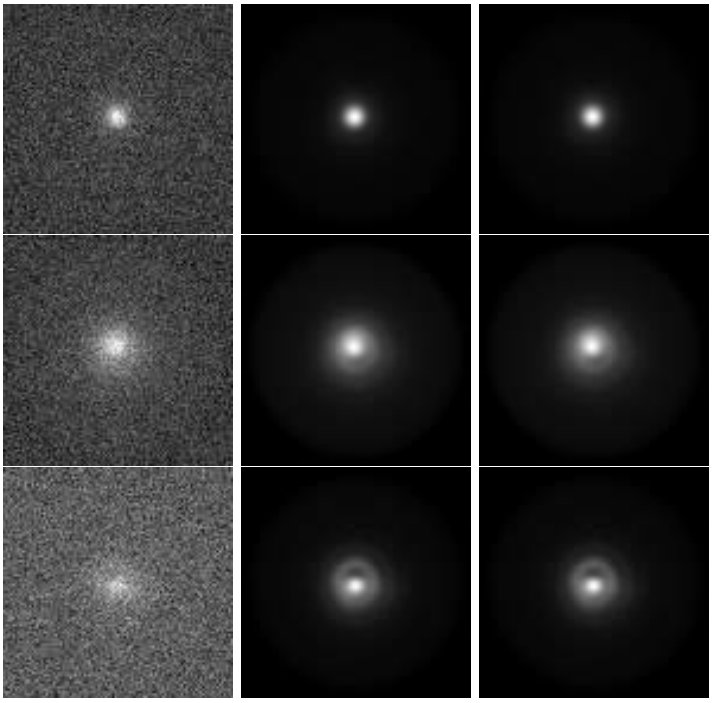
where g_p is the unknown amplitude of the bead at position z_p and $\mathcal{C}$ is the known conical hull of the sampled operators. The additional linear constraint $\langle \mathbf{\Gamma}, \mathbf{\Gamma}_0 \rangle = 1$ is related to an intrinsic identifiability problem in blind deblurring problems: the operator can be multiplied by a constant factor and the signal by its inverse, leading to the same result. Letting $\mathbf{\Gamma}_0$ denote a reference vector in $\mathcal{C}$, we can avoid this caveat. A nice geometrical choice consists in choosing the so-called *circumcenter* of the cone [HS10b]. We do not discuss this choice further since the proposed method is in essence heuristic.

We solve this problem using an alternating minimization algorithm: we first solve the problem w.r.t. $\mathbf{g}$ with fixed $\mathbf{\Gamma}$ and then solve the problem w.r.t. $\mathbf{\Gamma}$ with fixed $\mathbf{g}$. The individual minimization problems are convex and can be solved with accelerated projected gradient descents.

If the amplitudes $\mathbf{g}$ were known, the problem (7.1) would boil down to a constrained least square problem of size $I \times J$. Since each patch $\mathbf{p}_p$ yields I coefficients, the condition $P \geq J$ should be enough to ensure the identification. It is remarkable that such a low value (typically 5) is enough to identify the operator! Higher values of P would however make the method more robust to noise.

To validate the proposed approach, we use the simulation example presented in Chapter 5. We refer to it for further details. We randomly select an operator

in the conical hull $\mathcal{C}$, and apply it to a grid of 25 Dirac masses (for a field of view of 2304×2304 pixels). A significant amount of noise is added to the image and the true locations of the beads are randomly perturbed (with Gaussian random variable of variance 0.5). We then estimate the operator and display the result in Fig. 7.1.



(a) SNR 1.5dB (b) (c) SNR 17.6dB

Figure 7.1: Identification of an operator in simulated images. a) Noisy crops used for identification. b) Impulse responses of the true operator at some locations. c) Estimated operator. Notice that the method is able to denoise the PSFs very efficiently.

7.1.2 Sparse+Smooth-deblurring

Let $\mathbf{u}_0 \in \mathbb{R}^N$ denote a blurry image and $\mathbf{H} : \mathbb{R}^N \to \mathbb{R}^N$ a discrete integral operator. We aim at deblurring an image composed of a sparse component $\mathbf{u}_1$ (e.g. scattered micro-beads), and a smooth component $\mathbf{u}_2$ (e.g. auto-fluorescent background). In order to recover $\mathbf{u}_1$ and $\mathbf{u}_2$, we propose to solve the following original variational problem:

$$\inf_{\mathbf{u}_1 \in \mathbb{R}^N_+, \mathbf{u}_2 \in \mathbb{R}^N_+} \frac{1}{2}\|\mathbf{H}(\mathbf{u}_1 + \mathbf{u}_2) - y\|_2^2 + \gamma_1 \|\mathbf{u}_1\|_1 + \gamma_2 \|\Delta \mathbf{u}_2\|_2^2.$$

The term $\frac{1}{2}\|\mathbf{H}(\mathbf{u}_1 + \mathbf{u}_2) - y\|_2^2$ is the data fitting term, the term $\gamma_1 \|\mathbf{u}_1\|_1$ promotes the sparsity of the $\mathbf{u}_1$ component and the term $\gamma_2 \|\Delta \mathbf{u}_2\|_2^2$ promotes the smoothness of $\mathbf{u}_2$. The non-negative parameters γ_1 and γ_2 allow to balance each term and have been tuned manually so as to obtain a visually pleasant

result. This problem can be solved efficiently using accelerated proximal gradient descent algorithm [Nes18].

This formulation has already been proposed in the setting of blind deblurring [Mou+15]. The main difference concerns the blur identification. In the BSS algorithm, the blur operator is only estimated once. This is only possible since we use the stronger assumption that we know a low dimensional subspace $\mathcal{C}$.

To test the proposed algorithm, we evaluate its performance on synthetic micro-beads images with a spatially *invariant* Gaussian PSF with variance $\sigma = 10^{-2}$. We add a second order polynomial to simulate the background and random white Gaussian noise. The blurry-noisy image is displayed in Fig. 7.2a. The value of the proposed methodology comes from two distincts features: a more accurate model of microscope and a better deblurring model with the Sparse+Smooth prior. To disentangle the respective contributions of each aspect, we conduct two experiments.

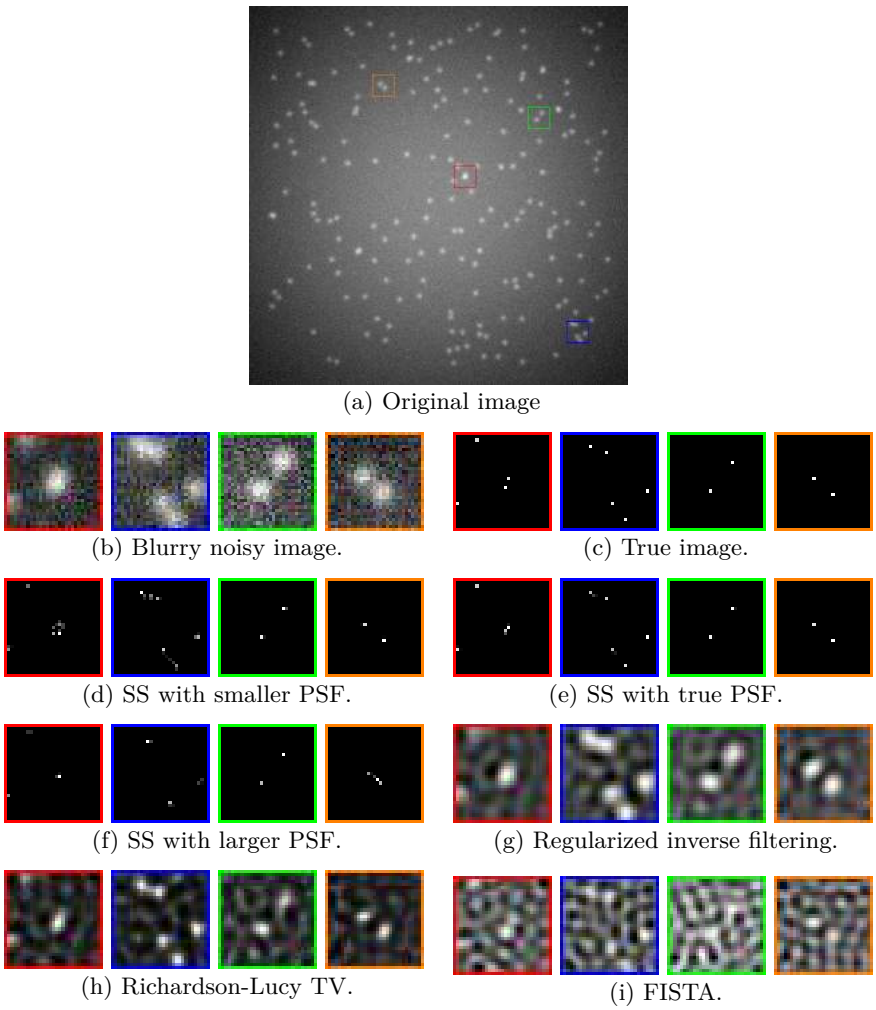
We first show the impact of an accurate model. We apply the Sparse+Smooth algorithm with a PSF smaller than the true one ($\sigma = 0.5 \times 10^{-2}$) in Fig. 7.2d, with the true PSF ($\sigma = 10^{-2}$) in Fig. 7.2e, and with a PSF larger than the true one ($\sigma = 2 \times 10^{-2}$) in Fig. 7.2d. As can be seen in Fig. 7.2e, 7.2f and 7.2d, only the algorithm run with the correct PSF is able to correctly localize all sources, even when they are close together. This shows the importance of describing the microscope response accurately.

Second, we compare the Sparse+Smooth deblurring algorithm with other standard approaches implemented in *DeconvolutionLab2* [Sag+17a] using the true PSF ($\sigma = 10^{-2}$). We have selected the following popular methods: the regularized inverse filtering in Fig. 7.2g, the Richardson-Lucy TV algorithm in Fig. 7.2h, and FISTA algorithm with ℓ^1 penalization of the Haar wavelet coefficients in Fig. 7.2g. In all experiments we tuned the parameters manually so as to obtain the best results from a perceptual point of view. The Sparse+Smooth algorithm seems to be by far the preferable approach to detect point sources over a smooth background.

7.2 Real-life experiment

Image deblurring is a technique than can lead to significant improvements of image resolution and quality. In most acquisitions, this technique is however neglected since it requires strong skills in optics, image processing and computer science. In particular, the prior calibration of a microscope is critical: model mismatches can lead to dramatic performance losses and oftentimes lead biologists to prefer using the raw images. In this paragraph, we show that the methodology proposed in Chapter 5 of learning a whole family of operators to describe the microscope allows to avoid a precise calibration before each experiment and therefore significantly eases the application of a deblurring algorithm.

The key observation is that identifying an operator from a single degraded image becomes rather easy when the operator depends linearly on a small number of parameters. In particular, if we know beforehand that the degraded image contains a few point sources, we show in previous paragraph how a simple constrained least squares problem allows to recover the operator. We then design an original non-blind deblurring algorithm with a known operator for sparse + smooth images. The overall algorithm is called BSS for Blind Sparse + Smooth



(a) Original image

(b) Blurry noisy image.

(c) True image.

(d) SS with smaller PSF.

(e) SS with true PSF.

(f) SS with larger PSF.

(g) Regularized inverse filtering.

(h) Richardson-Lucy TV.

(i) FISTA.

Figure 7.2: Validation of the Sparse+Smooth deblurring algorithm on a synthesized image. We first challenge the algorithm with inaccurate models of PSFs (the exact PSF, a smaller one and a larger one). As can be seen, only the true PSF produces near exact results, but the algorithm behaves nicely even with some inaccuracies. We also compare the Sparse+Smooth deblurring algorithm with three popular methods proposed in *DeconvolutionLab2* [Sag+17a]. Here, the proposed model performs significantly better.

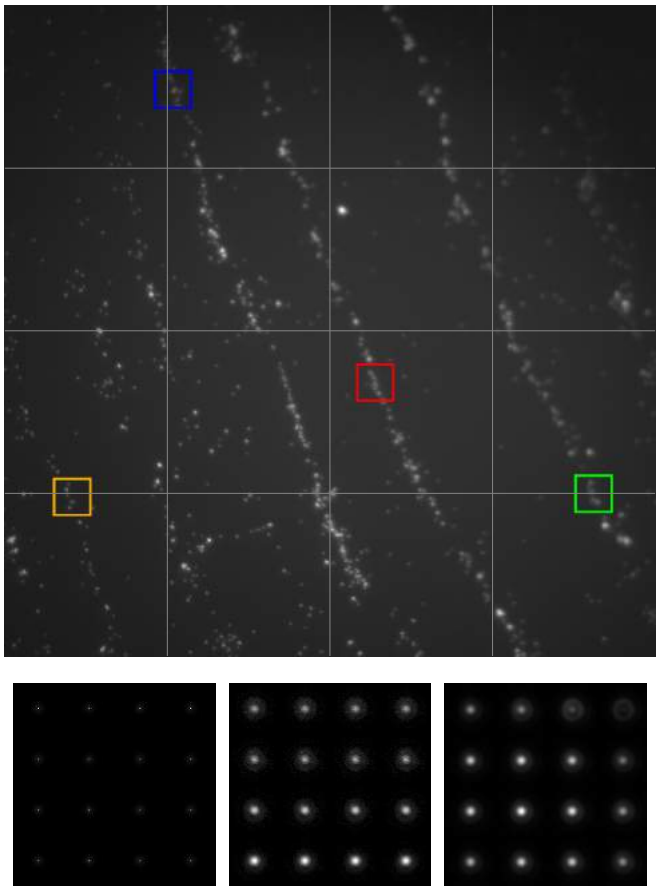
deblurring. The idea of using two steps algorithms to perform blind deblurring based on reduced models was already explored in [Hir+11; Why+12] for the specific case of motion deblurring. Our approach however differs significantly in the way we model the blur, in the estimation and deblurring process.

To assess the proposed methodology, we test the proposed algorithm on a real image of fluorescent micro-beads aligned along filament like structures. We perform this experiment on an image obtained with the same wide-field microscope as the one used to collect the real data in Chapter 5. The blurry image in Fig. 7.3 is acquired at a distance of 400nm from the focal plane. It is possible to compare the image at the focal plane in Fig. 7.4, (ii) with the image

used for the deblurring experiment in Fig. 7.4, (i).

We first compare the estimation of the PSF from the image in Fig. 7.3, using different approaches. In Fig. 7.3, (i), (ii), (iii) we show 16 equidistant PSFs estimated using 3 different approaches. In (i), we used the Matlab code *deconv_blind* based on an alternate minimization of two quadratic criteria. We set 20 iterations and initialized the method with a PSF size of 21×21. The PSF size is clearly underestimated. In Fig. (ii), we partitioned the image into 4×4 patches. Within each patch, we estimated the PSF by averaging multiple isolated PSFs as is usually recommended, see e.g. `http://python-microscopy.org/doc/PSFExtraction.html`. In average, we could only use two PSFs within each patch since the bead density is high and only isolated PSFs can be used. Therefore, the PSFs are still noisy, and their shape seems inaccurate, especially on the top-left corner. In Fig. (iii), we show the output of our blind identification algorithm. The PSFs are denoised and it seems that we can better reproduce the first ring of the PSF, though this effect cannot be quantified in this experiment.

We then propose comparisons with different deblurring algorithms and show the result on some patches in Fig. 7.4. In (i), we show the image used as an input of for the deblurring. In (ii), we show the image at the focal plane. In (iii), we show the result of the function *deconv_blind* from Matlab. Here we assumed that the blur was piecewise constant on each of the 16 patches. In (iv), we show the result obtained with the software Huygens Professional version 19.04 (Scientific Volume Imaging, The Netherlands). The choice of Huygens software is motivated by its wide use among research facilities. It allows to perform a patch-wise deblurring of the full image. Here, we used the 4x4 patch decomposition in Fig. 7.3, (ii). The reconstruction is displayed in Fig. 7.4, (iii). We also conduct a second comparison with the open-source software *DeconvolutionLab2* [Sag+17a]. Again, this plugin is unable to identify the blur and we feed it with the operator estimated in Fig. 7.3, (ii). We use the Richardson-Lucy with total-variation regularization, which provides the most satisfactory results, see Fig. 7.2. The reconstruction is displayed in Fig. 7.4, (v). Finally, we show the output of the proposed Blind Sparse+Smooth algorithm in Fig. (vi). The Matlab *deconv_blind* approach clearly outputs unsatisfactory results (here we show the best achievable result by manual tuning of the parameters). The Huygens software was used with the default parameters. We input a 2D PSF extended in 3D to account for the defocus, specify 200 iterations with the CMLE algorithm. We also use the *DeconvolutionLab2* software with a Richardson-Lucy algorithm regularized with the total variation. Both algorithms identify single molecules, but also produce a significant amount of ringing. Finally, the output of our blind deblurring algorithm in Fig. 7.4, is really convincing. It rather faithfully reproduces the image obtained at the focal plane in Fig. 7.4, (ii) with an even better resolution. Observe that this is a really challenging setting: the input image has a significant amount of noise and while the PSFs on the left part are rather small, their diameter is about 40 pixels on the right of the field of view.



(i) (ii) (iii)

Figure 7.3: Large 2048×2048 crop of the original image taken at 400nm from the focal plane. The contrasts have been stretched for a better visualization. In (i), (ii) and (iii), we show the PSFs estimated using different approaches at the centers of the 4×4 uniform tiling of the image represented in gray. In (i), we used the *deconv_blind* algorithm from Matlab. In (ii), we averaged isolated PSFs within each patch. In (iii), we show the output of our blind identification algorithm.

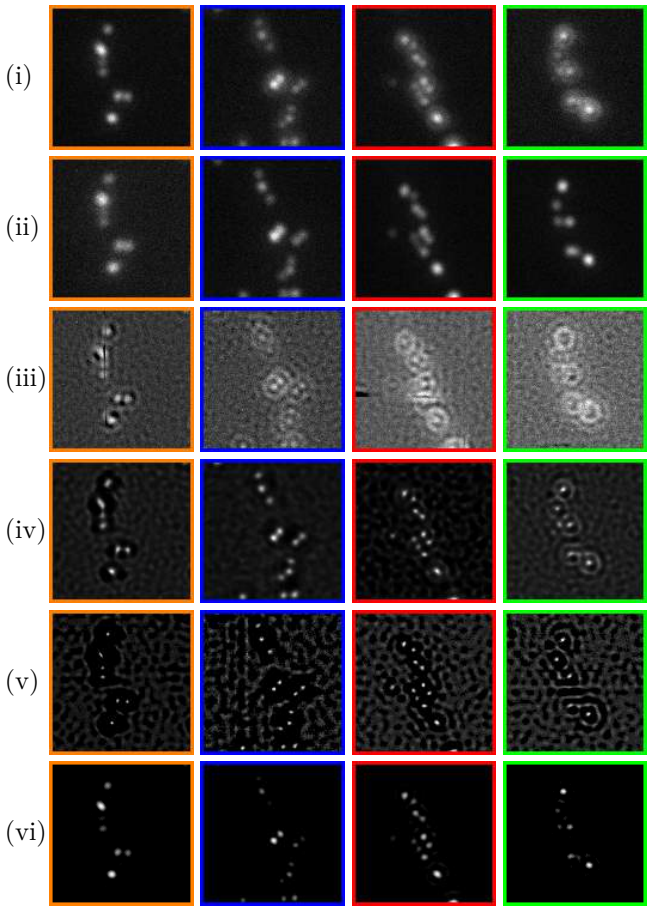


Figure 7.4: Zooms on the output of different deblurring algorithms. (i) image used for the deblurring experiment. (ii) image at the focal plane. (iii) output of the Matlab *deconv_blind* algorithm. (iv) output of the Huygens software. (v) output of *DeconvolutionLab2* with a Richardson-Lucy TV algorithm. (vi) output of our Blind Sparse+Smooth algorithm.

Chapter 8

Deepblur: blind identification of space variant PSF

Résumé : *Nous proposons une méthode basée sur l'apprentissage machine pour estimer les opérateurs de flou en utilisant uniquement des images floues. Nous proposons une procédure d'apprentissage pratique utilisant la connaissance d'un sous-espace à basse dimension d'opérateurs. Nous montrons que si une collection d'opérateurs admissibles est disponible, nous pouvons restreindre ce sous-espace à un ensemble convexe pour accélérer et renforcer la procédure d'entraînement. La méthode proposée fonctionne en utilisant une architecture de réseau neuronal convolutif, qui permet une estimation rapide. La motivation principale de cette approche est de résoudre des problèmes inverses aveugles ou de calibrer automatiquement les systèmes optiques. Enfin, nous illustrons la performance du réseau sur des problèmes de defloutage.*

Abstract: *We propose a machine-learning based method to estimate blur operators using only blurry images. We provide a practical training procedure using the knowledge of a low dimensional subspace of operators. We show that if a collection of admissible operators is available, we can restrict this subspace to a convex set to speed up and strengthen the training procedure. The proposed method performs well using rather simple neural-network architecture, which allows fast estimation. The main motivation behind this approach is to solve blind inverse problems or automatically calibrate optical systems. We finally illustrate the performance of the network on deblurring problems.*

This chapter is not published yet. We plan to first submit it in a conference and if time allows it, extend this work to a journal publication. It is joint work with Pierre Weiss.

Contents

8.1 Introduction **158**
8.1.1 Contributions 158
8.1.2 Existing works 159
8.2 The proposed method **160**

8.2.1 Operator parameterization 161
8.2.2 Operator sampling 161
8.2.3 Examples 163
8.2.4 Neural network 163
8.2.5 Training data set and numerical implementation . . 165
8.3 Numerical experiments **165**
8.4 Conclusion **170**

8.1 Introduction

Optical systems, although increasingly efficient, produce deteriorated observations induced by the various optical components. These degradations can be prejudicial to the interpretation of the results. Reconstruction algorithms are used to extract meaningful information from the acquisitions, but require to know the acquisition operator precisely. This assumption is crucial to produce accurate reconstruction [Die+15].

In many situations, it is reasonable to suppose that the acquisition operator is known. It can be calibrated beforehand [Des+14], or with the help of a theoretical model [GL89; Goo05]. This estimate can then be used in standard inverse problems solvers. A wide range of methods are available, leading to different algorithm depending on the application [CP11; AÖ17].

Unfortunately, in numerous applications the forward operator is known only approximately, or even completely unknown. We can then try to estimate the original image and the forward operator *simultaneously* based on the degraded image. This is the field of blind inverse problems. A few heuristic methods perform suprisingly well despite the challenging setting [CW98; YK99; Lev+09; Xu+11]. In addition, recent theoretical progresses were achieved recently, allowing to better understand situations where the problem can actually be solved and when the identification is not possible [ARR13; CSV13]. The theory has grown over the years, with progresses both on the conditions for exact recovery guarantees and for the numerical aspects [LLB17; KS19; BB19; Li+19].

These methods, however, suffer from two flaws. First, the reconstruction guarantees only hold under stringent assumptions that are not realistic in practice. Second, the numerical cost is often prohibitive, despite a significant improvement over the years [BB19].

The recovery theories often require that a low-dimensional of operators $\mathcal{H}$ is available. Im microscopy, we recently showed that this subspace could be learned efficiently [Deb+20a]. We also argue that this procedure presents significant advantages to calibrate a microscope, than using a single operator. The aim of this paper is to show that the knowledge of a this subspace can also be used to train a neural networking learning the blur operator from a single image. This procedure provides an efficient tool to evaluate the quality of an optical system or to solve blind inverse problems.

8.1.1 Contributions

Assume that we observe an image $y \in \mathbb{R}^m$ generated by the following forward model:

$$y = H_0(u_0) + \eta,$$

where $u_0 \in \mathcal{U}$ is the original signal, $\mathcal{U}$ is a subspace of images, H_0 is an unknown blur operator living in some subspace $\mathcal{H}$, and η is an additive random perturbation.

Solving a blind inverse problem consists in estimating H_0 and u_0 based on the measurements y. In this work, we focus only in the first step of the 2-step method:

1. Find an estimate $\hat{H} \in \mathcal{H}$ of the original forward operator H_0 based on the observation y.

2. Find an estimate $\hat{u} \in \mathcal{U}$ by solving a standard inverse problem with forward operator $\hat{H}$.

The reason for focusing on the first step only is that there is a rich literature to solve standard inverse problems with a wide range of efficient solvers. In this work, we focus on the first step, which is coined *operator identification*.

A key assumption in this work is that the operator subspace $\mathcal{H}$ is low-dimensional. An important additional ingredient to improve the training procedure is to restrict the range of operators to a convex cone. This constraint has the advantage of preserving the key properties of the admissible operators (e.g. nonnegativity of the impulse responses), while reducing the size of the search space. Under these assumptions, we show that a neural network can solve the operator identification problem both accurately and rapidly. We illustrate this method on numerical simulations and provide a blind deblurring example on realistic simulation with spatially varying blur.

8.1.2 Existing works

Solving blind inverse problems is a challenging issue that received a lot of attention in the literature since the 1970's [Can76; AD88]. Throughout this PhD, we tested many exsting approaches. Surprisingly and sadly, none of them revealed sufficient for the applications we had in mind for one or more of the following reasons:

- The method operates only under too restrictive assumptions (e.g. the underlying image is constituted of text or point sources,...). A typical example is the theory we developed in Chapter 6.

- The computing times are definitely not compatible with large scale problems. We had to stop algorithms downloaded on the web after two days of computation for a single image on a 40 cores computer.

- The operator estimate turned out to be biased and the results were unsatisfactory. This last issue is definitely the most important one and, sadly, it turns out that *no method* was able to perform satisfactorily even for relatively simple problems.

The standard approach that has prevailed for years is to estimate the original signal and the blur operator using an alternating minimization procedure. Under smoothness assumptions on the image and the operator, each sub-problem boils down to solve a standard inverse problem. This regularity assumption is imposed through regularization as total variation [CW98; PF14], a Gaussian

model assumption [Fer+06; CZF10] or Bayesian hypotheses [Lev+09]. Detection of structure, such as edges, is also a popular approach [Sun+13; DW08], which - in our opinion - is critical for the success of the approach.

Blind inverse problems are present in nearly any scientific domain. However, computer vision community has been particularly productive in this theme. Motion blurs are the standard model, which arise when taking a photograph while moving are probably the main reason for this interest. For each point of the field of view, the impulse response of a motion blur can be characterized by two values: an angle describing the direction of the blur and its length. This assumption is similar to the low-dimensional hypothesis of $\mathcal{H}$ made in this work [Sun+15; Sch+15; Gon+17]. This approach also allows the estimation of spatially varying blurs, either by decomposing the domain into a patch or by interpolating the different coefficients on the field of view [Sun+15; Gon+17; Cou+13]. In this domain, space varying blurs are referred to as "non-uniform" blurs. Of importance, let us mention that motion blurs are simpler to treat that the blurs appearing in microscopy: the Fourier transform of a measure supported on a 1D curve vanishes much slowlier that the diffraction blurs considered in this PhD. Hence, the high frequencies are less attenuated and can be restored more efficiently.

In the last 5 years, machine learning approaches have begun to emerge and outperform older methods. The learning approaches can be divided in two categories. The first category concerns methods that directly estimate the reconstructed image from the observation [APS19; NCF17; NHM17]. These approaches are harder to compare to the present chapter since our goal is mainly to estimate the blur operator. The second category of approaches produce an estimation of both the blur operator and the original image [Sun+15; Sch+15; Gon+17]. They only consider motion blur that are not well captured by a low dimensional subspace, thus making the comparison with these methods difficult. Some works also focus on the identification of invariant motion blur operator [Kra+06; Cha16], without proposing an associated reconstruction method as it is the case in this paper.

The idea closest to the one proposed here is two recently published papers by Shajkofci and Liebling [SL20b; SL20a]. The authors use a neural network to estimate the parameters of a blur operator from various images. The first difference with this work is that their operators are exclusively convolutions. The spatial variations are then processed by splitting the observation domain in patches where the blur is assumed invariant. The second difference is that we consider a subspace $\mathcal{H}$. The linear dependency in the search parameters allows us to ensure the injectivity of the representation, meaning that two distinct parameters will necessarily lead to two different observations (up to the scaling ambiguity).

8.2 The proposed method

Let $\mathcal{U} \subset \mathbb{R}^n$ denote a space of admissible images and let $\mathcal{H} \subset \{H : \mathcal{U} \to \mathbb{R}^m, H \text{ linear}\}$ denote the space of admissible forward operators. Then, the observation $\mathbf{y} \in \mathbb{R}^m$ is given by the action of $H_0 \in \mathcal{H}$ on the original image $u_0 \in \mathcal{U}$, leading to

$$\mathbf{y} = H_0(u_0) + \eta, \tag{8.1}$$

where η is an additive random perturbation. In this paper, we aim at producing an estimate $\hat{H}$ of H_0 based only on the observation y. In this section, we introduce the proposed approach and discuss the choice of the sets $\mathcal{U}$ and $\mathcal{H}$.

8.2.1 Operator parameterization

The collection of admissible operators $\mathcal{H} \subset \{H : \mathcal{U} \to \mathbb{R}^m, H \text{ linear}\}$ can be rather arbitrary. In this work, we work under the following assumption.

Assumption 8.2.1 (Subspace of operators)**.** *We are given a collection of $K \geq 1$ linearly independent elementary operators $H_k \in \{H : \mathcal{U} \to \mathbb{R}^m, H \text{ linear}\}$. The subspace $\mathcal{H}$ of admissible operators is then defined by*

$$\mathcal{H} = \{\text{span}(H_k, 1 \leq k \leq K)\}.$$

Hence an operator $H \in \mathcal{H}$ can be parameterized uniquely by a vector $\boldsymbol{\gamma} \in \mathbb{R}^K$: $H = H(\boldsymbol{\gamma}) \overset{\text{def.}}{=} \sum_{k=1}^{K} \boldsymbol{\gamma}[k] H_k$.

This assumption allows to model a large collection of space varying operators. Without further assumption, taking any arbitrary vector $\boldsymbol{\gamma} \in \mathbb{R}^K$ might produce a non-realistic operator (e.g. with negative values). For instance, the point spread functions usually have nonnegative values. Not complying with this property can result in unpleasant image reconstruction artefacts such as ringing. To avoid this, we will make use of the following assumption.

Assumption 8.2.2 (A conical hull in $\mathcal{H}$)**.** *We are given a collection $(\boldsymbol{\gamma}_p)_{1 \leq p \leq P}$ of P admissible operator parameterizations. Assuming that this set represents the set of operators sufficiently densely, we can construct the conical hull of the coefficients $\boldsymbol{\gamma}_p$:*

$$\mathcal{C} \overset{\text{def.}}{=} \text{cone}\left(\boldsymbol{\gamma}_p, 1 \leq p \leq P\right), \tag{8.2}$$

which we will use as a proxy to describe the set of all admissible operators.

The previous assumption can also be seen as an accurate *convex regularizer*. For instance, if all the sampled operators are nonnegative (i.e. have nonnegative impulse responses), then any conical combination is non-negative too. Imposing the vector $\boldsymbol{\gamma} \in \mathcal{C}$ will therefore preserve this property. Another regularization effect appears if the value of the coefficients in $\boldsymbol{\gamma}$ follow a specific distribution, e.g. decay of the coefficients $\boldsymbol{\gamma}$ if they are obtained using a principal component analysis. The set $\mathcal{C}$ will also capture this effect, resulting in a thin cone in some directions of space.

Letting $\boldsymbol{\gamma}_0 \in \mathcal{C}$ and $u_0 \in \mathcal{U}$, the observation model now becomes

$$\mathbf{y} = \sum_{k=1}^{K} \boldsymbol{\gamma}_0[k] H_k(u_0) + \eta. \tag{8.3}$$

8.2.2 Operator sampling

In order to train a neural network, we will need to sample operators at random within $\mathcal{C}$. Sampling this high dimensional set is a bit trickier than it looks at first sight. For instance, we could imagine generating vectors $\alpha \in \mathbb{R}^P$ uniformly at random on the $(P-1)$-dimensional simplex Δ^{P-1} and defining a random

operator as $H = \sum_{p=1}^{P} \alpha_p H(\boldsymbol{\gamma}_p)$. Unfortunately, if P is large, this would result in the fact that the extreme rays of the cone $\mathcal{C}$ are not explored: the probability that α is sparse is very low. To avoid this flaw, we propose two complementary solutions:

i) Simplify the cone by keeping only extreme rays (or a subset of extreme rays).

ii) Do not draw the coefficients α uniformly, but rather favor sparse distributions.

Cone simplification As for point i), let us mention that simplifying a conical hull is a difficult problem related to nonnegative matrix factorization. We refer the interested reader to [VRR17] for instance. In this work, we propose a simple greedy algorithm described in Algorithm 11. We recall that the circumcenter $\bar{\boldsymbol{\gamma}}$

Algorithm 11 A cone simplification algorithm

Require: A set of vectors $(\boldsymbol{\gamma}_p)$.
Require: A number of extreme rays N.
Find the circumcenter [HS10a] $\bar{\boldsymbol{\gamma}}$ of the conical hull cone$(\boldsymbol{\gamma}_p, 1 \leq p \leq P)$. This is a convex programming problem.
Initialize the set of kept rays to $\mathcal{R} = \{\bar{\boldsymbol{\gamma}}\}$.
Initialize the set of rays to explore to $\mathcal{S} = \{\boldsymbol{\gamma}_p, 1 \leq p \leq P\}$.
for all $n = 0 \rightarrow N-1$ **do**
Find the vector in $\mathcal{S}$ that maximizes the minimum angle with the rays in $\mathcal{R}$.
Remove this vector from $\mathcal{S}$ and add it to $\mathcal{R}$.
end for
return $\mathcal{R} \backslash \{\bar{\boldsymbol{\gamma}}\}$ and the simplified set of parameterizations cone$(\mathcal{R})$.

of the conical hull cone$(\boldsymbol{\gamma}_p, 1 \leq p \leq P)$ is the unit vector that is center of the largest possible ball inside the conic set [HS10a].

Sampling the operators at random By Caratheodory's theorem, any point within a conic hull can be expressed a conical combinations of the K extreme rays of the cone. Hence, a possibility to sample the cone $\mathcal{R}_N$ is to pick random conical combinations of the elements in $\mathcal{R}_N$. The difficulty is how to choose the distribution on the simplex. In this work, we propose a heuristic method to favor the extreme rays. This is captured by Algorithm 12.

Algorithm 12 Sampling procedure in $\mathcal{C}$

Require: The subspace dimension K.
Draw a random integer J uniformly in $\{1, \ldots, K\}$.
Pick J integers $\{z_1, \ldots, z_J\}$ at random on the set $\{1, \ldots, |\mathcal{R}_N|\}$.
Draw a random vector $\boldsymbol{\beta}$ uniformly distributed on the $(J-1)$-dimensional simplex.
return Coefficient vector $\sum_{j=1}^{J} \boldsymbol{\beta}[j] \boldsymbol{\gamma}_{z_j}$.

8.2.3 Examples

We provide some practical examples of models that meet Assumption 8.2.1.

Convolution: The convolutional model is omnipresent in the literature of image processing. It is used in computer vision to model motion blur, in microscopy or astronomy to model optical blur (locally). For instance, a rough model of the system is a set of Gaussian convolution filters with various variances [ZZO07; BSZ12]. Such a collection is enough to construct a subspace as in Assumption 8.2.1. The restriction to the conical hull in Assumption 8.2.2 allows to avoid oscillations in the filters.

Product-convolution: The convolution model can only capture **space invariant** operators. One way to overcome this limitation is by using product-convolution expansion [Den+15; EW17; Deb+20a]. The subspace $\mathcal{H}$ is then defined using $K = IJ$ elementary operators:

$$H_{i,j}u = e_i * (f_j \odot u), \forall u \in \mathcal{U},$$

where $(e_i)_i$ and $(f_j)_j$ are two orthogonal bases which respectively describe the PSFs and the PSFs variations in the field of view. The estimation of such families $(e_i)_i$ and $(f_j)_j$ is possible using either a collection of sampled operators [DEW19], or several images of micro-beads [Deb+20a].

Parametric model: In microscopy, several models of PSFs (point spread function, i.e. impulse response of the operator) are defined based on a non-linear combination of optical parameters [GL89; Goo05]. It is often reasonable to assume that the PSFs generated by such models can be well expressed in a common low-dimensional subspace [Deb+20a], thus verifying Assumption 8.2.1. Computing the elementary elements $(H_k)_k$ boils down to computing a principal component analysis on the collection of PSFs.

8.2.4 Neural network

Recall that we aim at solving the following problem:

$$\text{Find } \hat{H} \in \mathcal{H} \text{ an estimate of } H_0 \text{ based on the observation } \mathbf{y}. \tag{8.4}$$

It can be recast as finding a mapping $\mathcal{M} : \mathbb{R}^m \to \mathbb{R}^K$, such that $\hat{H} = \mathcal{M}(\mathbf{y})$. The recent literature illustrated that some neural network architectures provide an efficient parametric family to approximate such mappings [SL20b; GOW19; BR19]. In the following, we let $\theta \in \Theta$ denote the weights of a neural network and we will use the notation $\mathcal{M}$ or $\mathcal{M}_\theta$ to describe the neural net.

In our experiments, we chose $\mathcal{M}_\theta$ to be the popular Resnet convolutional neural network (CNN) [He+16]. It is shown in Figure 8.1. The parameter space Θ characterizes the values of the filters for the convolution, the value of the affine transformation for the fully connected layer, etc... The total number of parameters for this network is about 25×10^6.

Problem (8.4) now turns to finding a parameter $\theta \in \Theta$, such that $\mathcal{M}_\theta(\mathbf{y}) \approx \boldsymbol{\gamma}_0$, for any observation $\mathbf{y}$ given by (8.3). We let $(\mathbf{y}_l, \boldsymbol{\gamma}_l)$ denote $L \geq 1$ observations and associated operator decompositions. Given this data set, we aim at finding

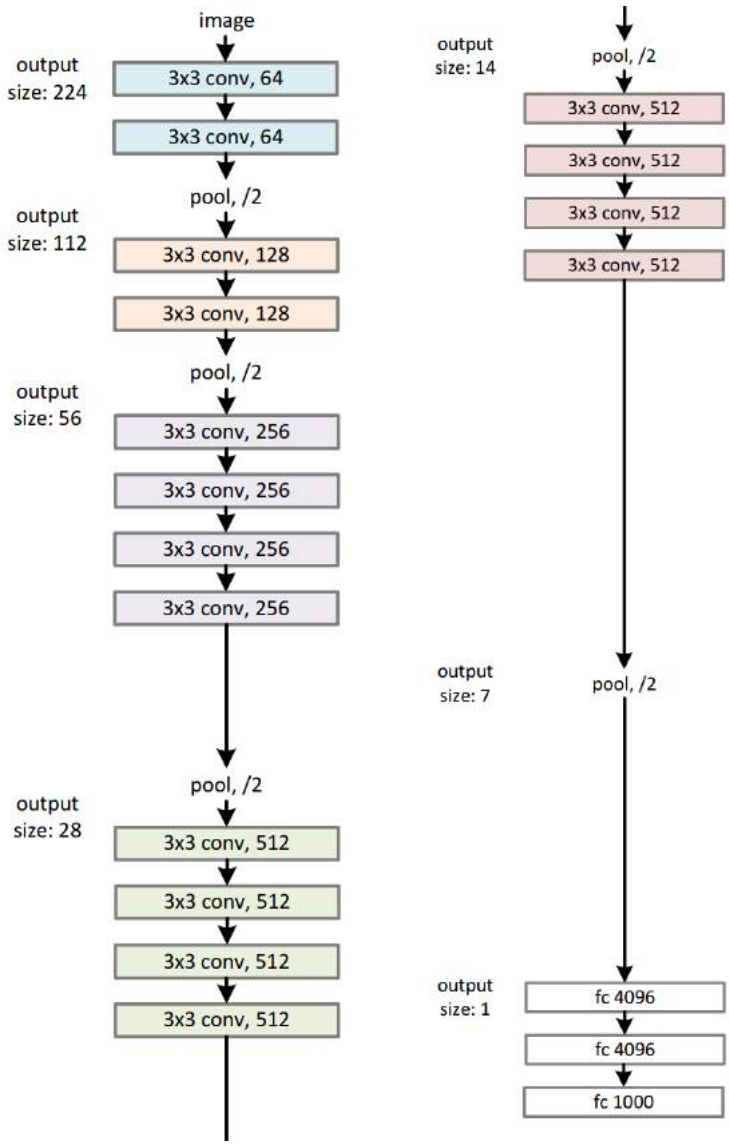


Figure 8.1: Resnet architecture given in [He+16]

the parameter $\theta \in \Theta$ that best fits the data set. Using an ℓ^2 discrepancy, this leads to the following optimization problem:

$$\inf_{\theta \in \Theta} \frac{1}{L} \sum_{l=1}^{L} \|\mathcal{M}_\theta(\mathbf{y}_l) - \boldsymbol{\gamma}_l\|_2^2 \,. \tag{8.5}$$

It can be solved using *ADAM* algorithm [KB14] implemented within the *Pytorch* library [Pas+19].

8.2.5 Training data set and numerical implementation

The training data set $(\mathbf{y}_l, \boldsymbol{\gamma}_l)_l$ is crucial in the estimation of an accurate neural network. If L is too small, the model will perform poorly on new acquisitions. The key feature for good performance of the neural network is then to construct a data set large enough to capture most possible acquisitions.

Based on the image formation model in Equation (8.3), generating a pair $(\mathbf{y}_l, \boldsymbol{\gamma}_l)$ simply consists in getting one operator parameter $\boldsymbol{\gamma} \in \mathcal{C}$ following the distribution described previously, and one image $u \in \mathcal{U}$. We will describe the set of images in the numerical experiments.

Numerical implementation To assess the method, we generate two collections of pairs $(\mathbf{y}_l, \boldsymbol{\gamma}_l)_l$. The first one is used to train the method, while the second one is used to test the method. The images constituting each set are selected at random with no repetition from one set to the other. This ensures that the method doesn't work well only because it perfectly fit the set $\mathcal{U}$. During the test procedure, the coefficients $\boldsymbol{\gamma}_l$ are drawn from the same distribution as in the training phase.

Figure 8.2 summarizes the proposed approach.

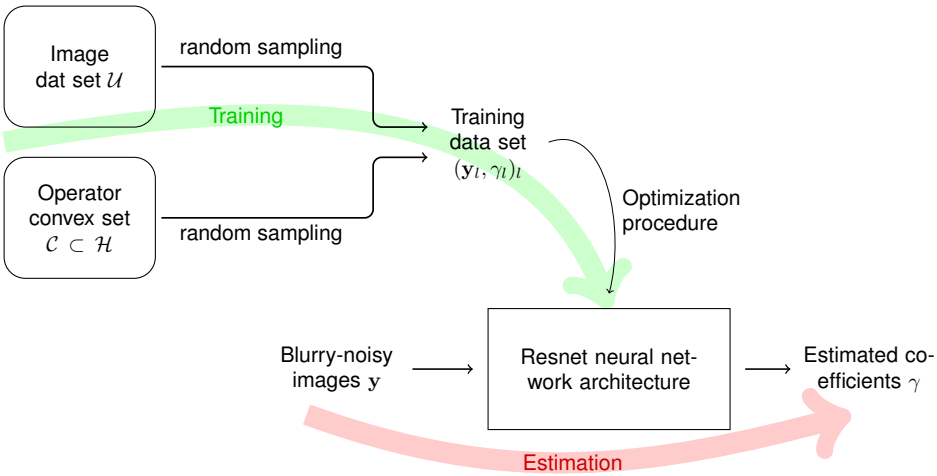


Figure 8.2: Workflow of method. It is composed of two distinct steps: the training (green) and the estimation (red).

8.3 Numerical experiments

In this section we propose a numerical illustration of the proposed approach. We focus on two examples: a convolution operator applied to simulated images, and a space variant operator applied to natural images. In this section, we work with $n = m = 128 \times 128$.

Identify convolution blur In this first numerical illustration, we focus on a very simple setting. The original images contained in the set $\mathcal{U}$ are random

diffomorphisms of a Cartesian grid, see Figure 8.4. This particular choice is motivated by the recent literature [Fer+06; Sun+13]. Therein, the authors show that the structure inside images is often a key feature to ease the identifiability of the blur kernel.

The subspace $\mathcal{H}$ is constructed using Gaussian convolution filters. We generate a collection of $P = 1000$ matrices $(e_p)_p$ defined by

$$e_p(x, y) = \exp(-x^2/(2\sigma_{1,p}^2) - y^2/(2\sigma_{2,p}^2)) \text{ with } \sigma_{1,p}, \sigma_{2,p} \in [1, 4].$$

The subspace $\mathcal{H}$ is then obtained by computing the principal component analysis of the family $(e_p)_p$ and keeping $K = 5$ eigen-elements. The conical hull $\mathcal{C}$ is then computed by projecting the collection $(e_p)_p$ onto $\mathcal{H}$.

Finally, we train a convolutional neural network with 100 iterations of the *ADAM* algorithm. The training and testing data set are respectively composed of $L = 2048$ and $L_{test} = 512$ elements generated using the same distribution. Two independent sets are generated for each iteration of the learning optimization procedure. The batch size in *ADAM* algorithm is 64. The learning rate is 10^{-3}.

We compare the Resnet architecture on two scenarios: estimating the standard deviation parameter of the blur (coined 'Gaussian' method), as proposed in [SL20b], or estimating the coefficients $\boldsymbol{\gamma}$ in the convex set $\mathcal{C}$ (coined 'subspace' method). The results are reported in Figure 8.3. We display the relative Hilbert-Schmidt distance between the true operator and its estimation. There is clearly a gap between the two methods. The proposed method requires more training steps to obtain the same accuracy than the Resnet architecture used to only estimated standard-deviation parameters. However, the difference is only of small magnitude (1% after 100 iterations). The advantage of our approach lies in its ability to estimate spatially varying blurs and in the robustness of the training procedure in practical applications using the convex set $\mathcal{C}$.

In Figure 8.4, we pick up some of the original and blurry images used in the testing procedure. We also display a magnification of the original convolution kernel and its estimation using our approach. A visual inspection shows that the kernel is rather well estimated. This claim is corroborated by Figure 8.3.

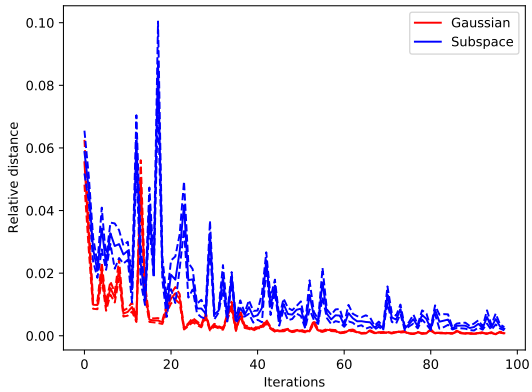


Figure 8.3: Relative Hilbert-Schmidt distance between the estimated and the true convolution with the testing data set. The red line corresponds to the method in [SL20b] and the blue line corresponds to the proposed approach. The dashed lines correspond to the standard-deviation around the mean value.

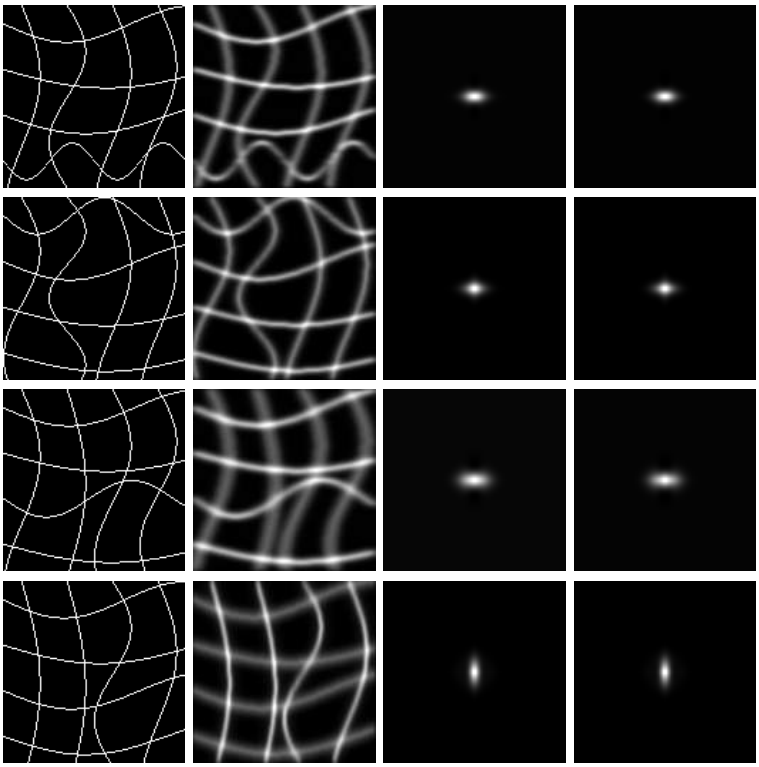


Figure 8.4: Example of the images processed by the method and its output with convolution operator on structured images. From left to right: original image, blurry image, true convolution filter, estimated convolution filter.

Identify spatially varying blur and deblurring In this experiment, we focus on a more realistic setting. We select $\mathcal{U}$ as a collection of images from the STL-10 data set [CNL11] available with the *Pytorch* library. We select 10000 random images for the training procedure, and 1000 images for the testing procedure. The batch size in *ADAM* algorithm is 64. The learning rate is 10^{-3}.

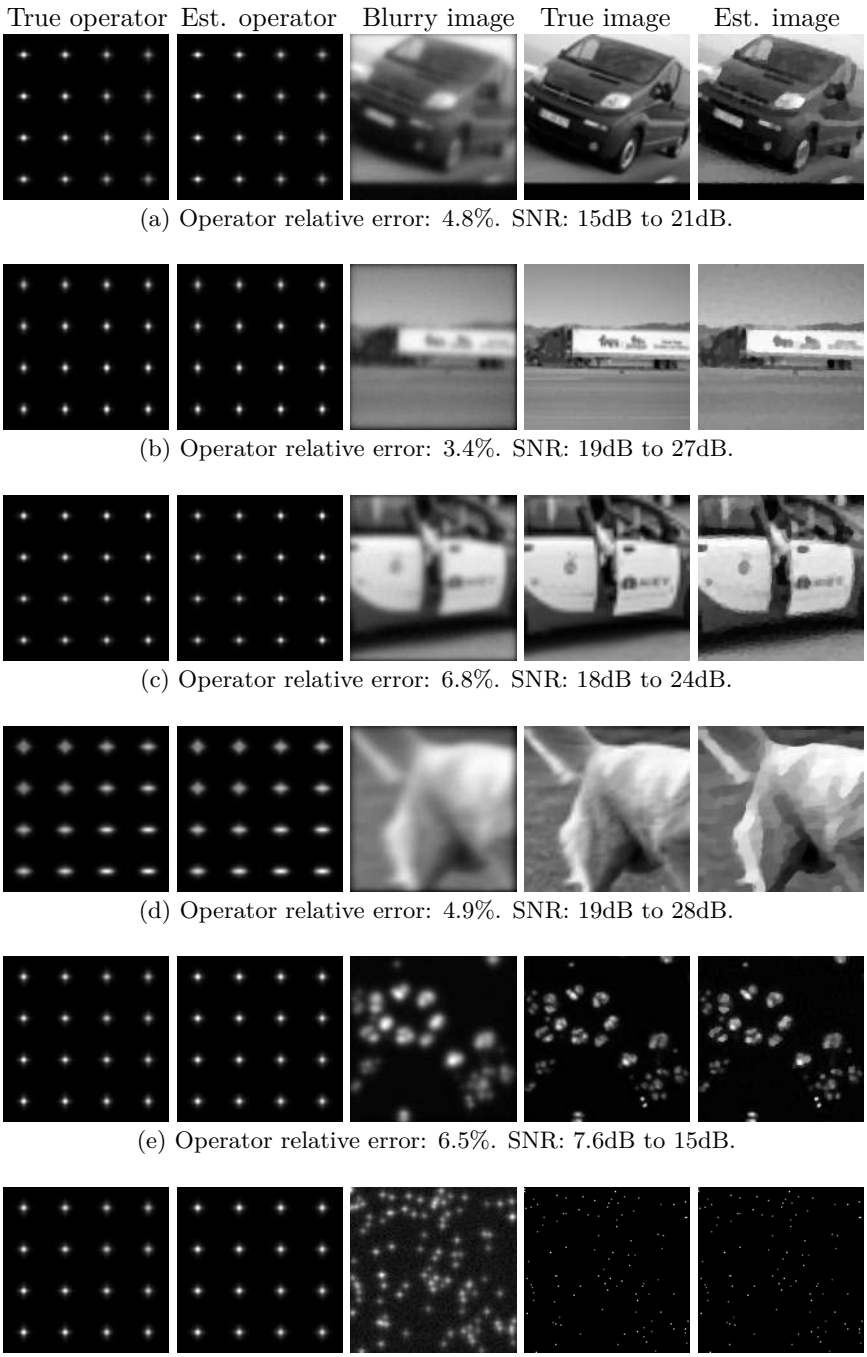
We consider operators given by the product-convolution expansion introduced in Section 8.2.3. We generate a collection of $P = 1000$ operators. They are generated using the same procedure as in Chapter 4.6.2. We do not detail further this step. Given this collection of P operators, we apply the subspace estimation method presented in Chapter 4. Roughly speaking, it boils down to compute two principal component analysis. The estimated subspace is described by $I = 5$ eigen-elements to express the PSFs, and $J = 4$ eigen-elements to express the space variations. Finally, this result in a subspace composed of $K = IJ = 20$ elements. Using once again methodology of Chapter 4, we construct the convex set $\mathcal{C}$ by taking the conical hull of the collection of operators $(H_p)_p$.

To illustrate the proposed methodology, we solve the following total variation deblurring problem with a proximal gradient descent algorithm:

$$\inf_{\mathbf{u}\in\mathbb{R}^m} TV(\mathbf{u}) + \frac{\lambda}{2}\|\hat{H}\mathbf{u} - \mathbf{y}\|_2^2, \tag{8.6}$$

where the operator $\hat{H}$ is the estimation output by the neural network. In Fig. 8.5, we display the true and estimated operators applied to a Dirac comb, and the result of the deblurring algorithm using the estimated operator with total variation regularization for various images. In Fig. 8.5a-d, we use natural images of the STL-10 test data-set. As expected, the network – that has been trained on similar images – outputs an accurate estimate of the true operator. Solving the deblurring Problem (8.6) leads to sharp results, and allows to recover hidden details. The average relative Frobenius norm between the true operator and the estimation produced by the network is less than 5% at the end of the training phase (on the test data-set).

In Fig. 8.5e and Fig. 8.5f, we observe how the trained network behaves on other images. In Fig. 8.5e, we use a real image of microscopy from the dataset [LSC12]. In Fig. 8.5f, we use a simulated image of single molecule localization microscopy, i.e. points sources at random locations. In both problems the network performs well and retrieves the operator with high accuracy. More importantly, solving the deblurring Problem (8.6) with the estimated operator allows to better discriminate the biological elements, such as the cells in Fig. 8.5e. In the last experiment, we replace the total variation regularization term by a ℓ_1 penalty on the signal. It allows to retrieve sparse elements and greatly improves the resolution of fluorescent proteins. Notice that these deblurring results are simply a proof of concept since better deblurring algorithms could be used (e.g. a variational network).



(a) Operator relative error: 4.8%. SNR: 15dB to 21dB.

(b) Operator relative error: 3.4%. SNR: 19dB to 27dB.

(c) Operator relative error: 6.8%. SNR: 18dB to 24dB.

(d) Operator relative error: 4.9%. SNR: 19dB to 28dB.

(e) Operator relative error: 6.5%. SNR: 7.6dB to 15dB.

(f) Operator relative error: 8.4%. SNR: -7dB to 0.64dB.

Figure 8.5: From left to right. True and estimated operator, blurry-noisy image, original image and estimated one by solving problem (8.6) with total variation regularization for (a)-(e) and ℓ_1 regularization for (f)). The SNR with the blurry and the estimated images is given in the sub-captions.

8.4 Conclusion

This paper is in line with recent works showing that neural networks can be successfully trained to identify blurring operators. This is a preliminary work, which still requires some attention. In particular, the following issues retain our attention:

- Study the robustness to new operators living outside the training set.
- Can the method scale to identify larger subspaces?
- Can we use smaller databases of images when dealing with specific imaging modalities where training images are scarce?
- Can we find simpler architectures, perhaps inspired by existing deterministic approaches? Among other advantages, this would allow to reduce the carbon footprint of the training procedure.
- Is there any theory which could explain the encouraging results observed so far?

Part IV

Miscellaneous problems

Chapter 9

Biolapse toolbox: automating biological image analysis.

Résumé : *Nous proposons une boîte à outils clé en main pour segmenter, suivre et classer les événements biologiques à partir d'images temporelles obtenues au microscope haut-débit. La tâche de segmentation est effectuée par un réseau de neurones. Son entraînement est rendu reproductible sur d'autres jeux de données. Le suivi des cellules dans le temps est effectué par un algorithme modulable basé sur des a priori physiques. L'algorithme de classification combine un réseau de neurones et un modèle de chaîne de Markov. Cela convient particulièrement aux applications biologiques où les états d'intérêt sont ordonnés dans le temps.* `Biolapse` *est accessible avec une interface graphique qui facilite l'étape d'apprentissage de l'utilisation des différents outils. Nous concluons en illustrant la simplicité d'utilisation de la boîte à outils* `Biolapse` *pour les non-spécialistes en apprentissage machine avec une application sur des cellules cancéreuses.*

Abstract: *We propose a turnkey toolbox to segment, track and classify biological events from a time-lapse image obtain on high-content microscope. The segmentation task is performed by a neural network. Its training is made reproducible on other data-set. The tracking is performed by a modular algorithm based on physical a priori on the image. The classification algorithm combines a neural network and a Markov chain model. This is particularly well suited for Biological applications where the possible states are ordered in time.* `Biolapse` *is accessible with a graphical interface that assists the training of the machine leaning algorithm and the use of them. In the last part we explicit the assets of* `Biolapse` *toolbox for non-machine-learning experts with a concrete application with living cells.*

This chapter is still under development; It has been introduced by the two proceeding papers [DL19; LDM20]:
Debarnot, V., Lebrat, L. (2019). Segmentation: a data driven approach though neural network. IEEE 16th International Symposium on Biomedical Imaging (ISBI).
Debarnot, V., Lebrat, L., Mangeat, T. (2020). Biolapse Toolbox. Quantitative BioImaging Society

Contents

9.1 Introduction **174**
9.1.1 Existing work 174
9.1.2 Contributions 175
9.2 Materials and methods **176**
9.2.1 Graphical User Interface 176
9.2.2 Segmentation 178
9.2.3 Tracking 180
9.2.4 Classification 183
9.3 Conclusion **185**

9.1 Introduction

Cell cycle phases turn to be central in the comprehension of many biological phenomena. It makes possible, among other things, to better understand numerous pathologies or biological processes. In particular, a promising line of research is the understanding of cancer evolution and possibly the development of drugs to stop this mechanism [Hal+11; Neu+10]. This enthusiasm is coupled with advances in microscopy. High throughput screening permits the acquisition of cells over a long period of time, thus capturing the different phenomena that occur during the cell cycle, and over a large field of view, thus capturing a significant number of events of interest. On the other hand, this inevitably conducts to a significant amount of data to analyze, requiring even more efficient and fast algorithms to extract the useful information. Low illumination is also very important to decrease the toxicity of the sample, but this leads to noisy images. We are therefore interested in robust algorithms for the automatic detection of cells from time-lapse images. Moreover, in order to keep only the data of interest, and thus reduce the memory cost, we aims to automatically classify the state of the extracted cells, e.g. G1, S, mitosis.

9.1.1 Existing work

Although each problem is specific, most algorithms treating time-lapse images, if not all, share the same structure. First, the images are processed with a reconstruction algorithm to correct some distortions induced by the microscope (blur, noise,...). In a second step, the images are segmented, i.e. the cells are separated from the background. Then each cell is tracked, this is assigning a unique identifier over time. Finally, the last step consists in associating to each

cell its state among several possible choices, for instance a phase of the cell cycle. In this work, we focus on the last three steps, namely segmentation, tracking and classification. These problems have been widely studied independently and adapted for biological applications several times.

Segmentation is present in the main biological image processing softwares: ImageJ/Fiji [Sch+12; FKW17; DCU13], Icy [De +12], Ilastik [Som+11], and others [Car+06; Hod+13]. The methods proposed in these softwares are based on well-established techniques such as watershed [VS91] or active contours [KWT88]. Recently, neural network algorithms have shown great performances in image processing, outperforming human performance in some situation [LJ16; WKP16; Deh+17]. However, these techniques are not fully understood from a theoretical point of view. This leads to different ways of using them, which can make them difficult for a non-specialist to implement. Tracking algorithms are based on more or less advanced methods, but are often dependent on the underlying application [Deb+05; Li+08; Möl+14; LL93]. Capturing the dynamics of objects of interest necessarily depends on the time sampling scheme, the speed at which the objects move, etc. The tracking algorithms share features with a classification algorithm [Thi+13] or are independently link to machine learning methods [Huh+10].

However, several recent approaches aim to bind these three steps into a common pipeline more or less dedicated to a particular application. Some approaches are dedicated to 3D images [DPW11]. They show the benefits on drosophila images. Specific method has been developed for phase contrast microscopy [Gra+17], where mitotic and non-mitotic cells are identified. The LineageMapper toolbox [Cha+16] is much more general and provide a complete toolbox with a similar tracking procedure than the proposed approach. It is based on the minimization of a physically based cost function. Few methods proposed to use Markov model to ensure time consistency [Zho+08], e.g. S phase should be preceded by G1 phase. We follow a similar approach in this work to ensure valid prediction. Other works propose methods that reflect the same general spirit, but with their own techniques and specifications [Wan+07; Che+13; Gul+14].

9.1.2 Contributions

The main originality of the proposed method is to provide a `Python` Toolbox that is intended to be modular. This allows it to be easily adapted for distinct applications, and avoids the need to develop a new toolbox for each experimental set-up. One other difference with part of the previous toolboxes is that we provide machine-learning based algorithms for the segmentation and classification procedures, allowing the algorithms to automatically adapt to the data-set provided during the training phase. We furnish assistance for the training of the machine learning algorithms with the use of interactive script or graphical interface unit. This approach minimizes the expertise needed to select features of the underlying objects. If a large enough training data-set is provided, the proposed method tend to be robust for a large family of similar problems. This remark was already made in [DPW11] about classification. They show that the support vector machine method does not express well from one data-set to another unless the algorithm has been trained on a data-set that mixes the two data-sets.

We make available `Biolapse`, a user-friendly and flexible `Python` toolbox

that implements **latest machine learning** algorithms, to **biologists**. `Biolapse` aims to make handy the use of neural network for segmentation, extraction of features, and classification. It also includes an intuitive Graphical Unit Interface (GUI) that eases the analysis of biological time-lapse. More precisely, `Biolapse` comes with the following main characteristics:

- Segmentation: we provide a `Python` script that implements a neural network to segment biological images. Three architectures are proposed and should cover most applications.
- Tracking: we propose a simple approach based on the minimization of a flexible cost function.
- Classification: we provide a graphical interface that helps the user to label data. This comes with a neural network trained to extract feature from the image, and a machine learning classification algorithm that predicts the state of each cells, imposing time ordering as in the cell cycle.
- Automatic extraction: once the learning part is completed, a graphical user interface allows the user to load time-lapse images and automatically extract and classify the cells contained in it.

The main objective of this toolbox is to bind all these techniques into a single user-friendly interface, and to make it plug-and-play for people with limited expertise in image analysis or computer science. An importance has been given to the modularity of the different parts, allowing the replacement of segmentation, tracking or classification algorithms by methods more adapted to a particular problem if needed.

9.2 Materials and methods

9.2.1 Graphical User Interface

We present `Biolapse` a `Python` toolbox that extracts and classifies cells within time-lapse images. The workflow of the method is based three main steps: segmentation, tracking and classification, see Figure 9.1.

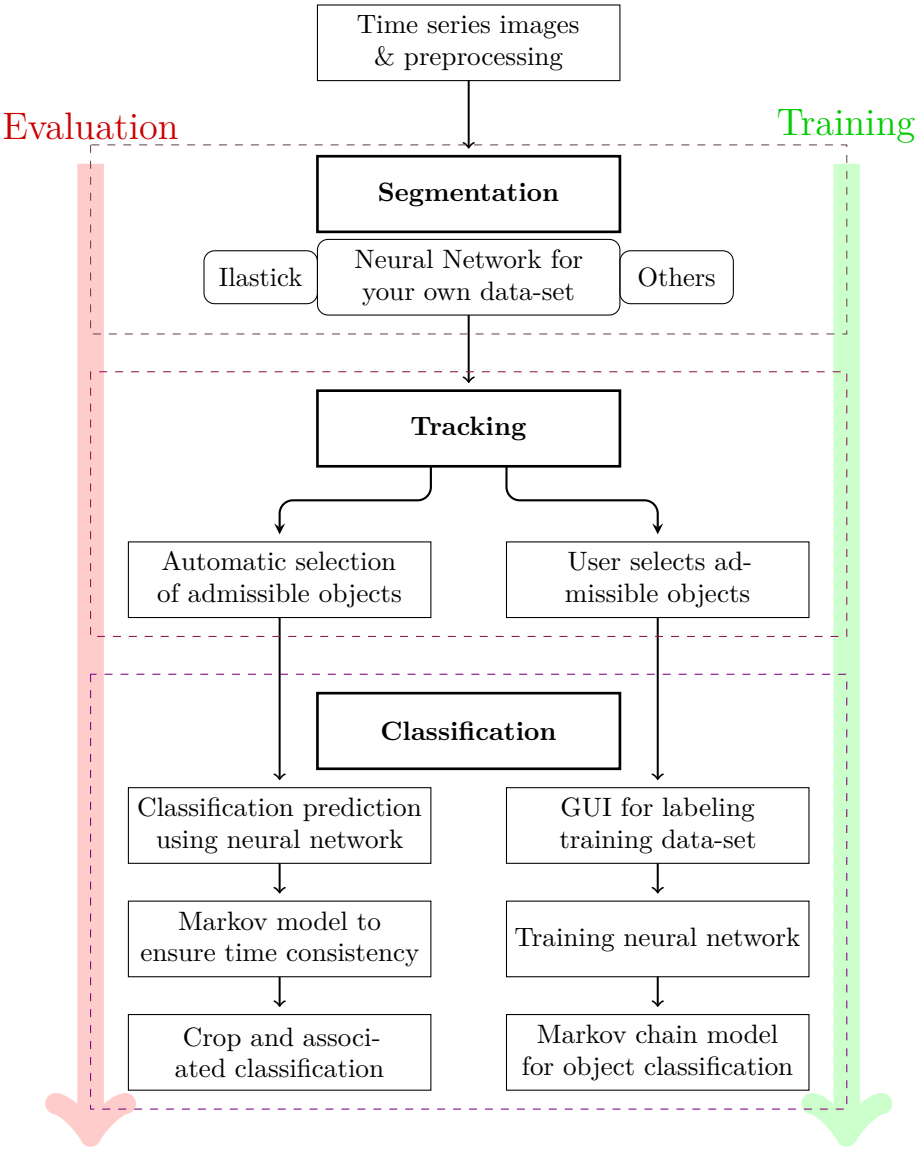


Figure 9.1: Workflow of Biolapse. The toolbox is divided into three distinct blocks: segmentation, tracking and classification. Notice that the workflow differs between training and evaluation.

The use of Biolapse is divided into two distinct phases: the training of the machine learning algorithms to obtain accurate results of the different tasks and the execution step. In the second part, `Biolapse` will aggregate all the algorithms presented in this chapter in order to extract and determine the state of the cells contained in a given time-series of images. Only few parameters related to the tracking and to the selection of eligible cells can be tuned. A screen-shot of the final interface is displayed in Figure 9.2.

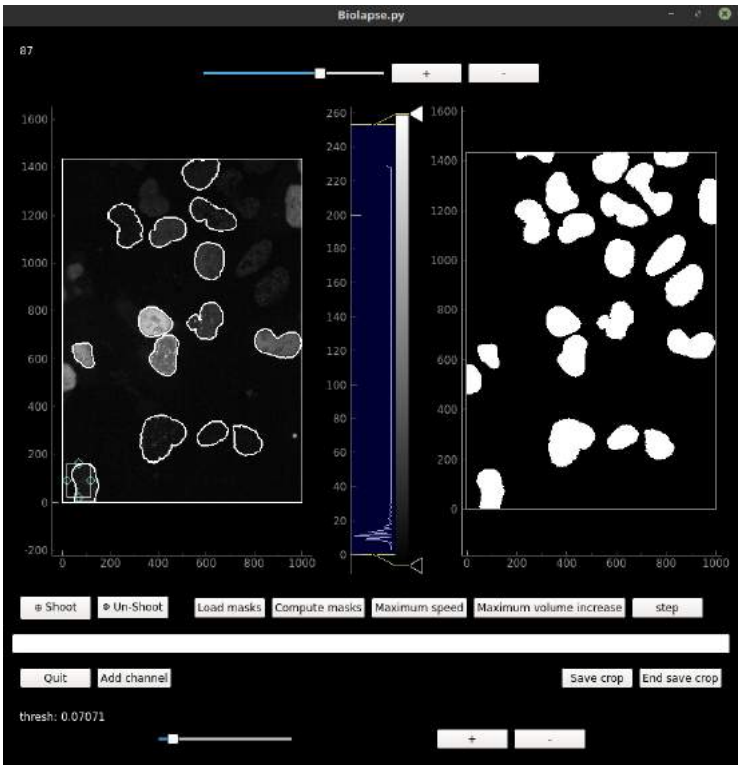


Figure 9.2: Screenshot of the `Biolapse` interface. On the left is the original image, and on the right is the associated segmented image. By clicking on the 'Shoot' button, the tracking and classification algorithm extract admissible cells (surrounded by white on the left) and predict the phase of the cell cycle of each cropped cells. The final prediction and the crops around cells is then saved for future statistical analysis.

The full toolbox, the experiments present in this chapter and the data-set used are available at `https://github.com/lebrat/Biolapse`. The toolbox is entirely implemented with *Python*, a free programming language, and, arguably, one of the most used tool in image processing and machine learning. We take particular care in writing documented codes that are quickly understandable and pliable to other applications.

9.2.2 Segmentation

Segmentation offers a wide range of methods, each with its advantages and disadvantages. In this work, we propose to use a neural network to perform this

task. This is motivated, among other things, by the performances of CNNs to address problems in image processing [KW13; Ben+13].

Another point in favor of neural networks is its running time. Once trained, a neural network can run segmentation of more than thousand of images per second. This is in contrast to the time required for training and the fact that the majority of operations are finely implemented in GPUs.

In `Biolapse` toolbox, three different segmentation neural networks are made available. Each of these has its own advantages detailed in the following paragraphs. In addition, we let the possibility to import segmentation masks, thus using other segmentation algorithms.

When to use neural network

The use of neural networks has become commonplace. However, some situations are not adapted to the use of such tools. We therefore believe it is important to recall the main characteristics of a problem that can lead to the use of such algorithms. If these conditions are not met, we recommend the use of other interactive toolbox that allows the user to generate a data-set such as *Ilastik* [Som+11].

The data-set: Exhaustiveness of the data-set is key, and the images shown during the training should be representative of the variety of images that the neural network will encounter during the production stage. The data-set should contain as few errors as possible, otherwise, the neural network will not be able to devise general rules for segmentation given the relevant patterns extracted on the data-set. From our experiment, incomplete or incorrect data-set causes significant damages to the neural network's ability to generalize out of the training data-set. Building an accurate data-set is thence a tedious task and should be considered as a time investment.

Time consistency: in this work, we focused on the particular problem of segmenting images extracted from a biological time-lapse. The integration of the temporal information in the neural network is of major importance to enhance its performance. In the toolbox, we propose two different methods that integrate this information.

Graphical Process Unit (GPU): The codes provided in `Biolapse` support GPU implementation [Aba+16].

Neural network architecture

`Biolapse` implements three different neural network infrastructures. For further technical details, we refer the reader to the corresponding papers. We compare these algorithms on a data-set made of real images. The data-set is composed of 143 time-series images with associated masks. We use 113 images as a training data-set and 30 images for testing. We then compare the three proposed architecture on this data-set and report it in Fig.9.4. Code required for the training of the different models as well as the data-set is made available in the `Github` directory[1] of `Biolapse`.

Unet: The Unet architecture is part of the wide family of autoencoder convolutional neural network [GBC16]. This type of network offers an alternation of elementary operations mainly composed of convolution and non-linearity. It

[1] `https://github.com/lebrat/Biolapse/blob/master/segmentation/`

generates a function that maps an image to a vector composed of a small number of elements (encoder part) called a feature vector, then the network reconstructs an image by reproducing the operations in reverse order (decoder part), see Figure 9.3. The Unet architecture is widely used in image processing for biological applications. The architecture exists both for 2D and 3D images [RFB15; Cic+16]. The Unet 2D operates on the frame of the biological movie, each frame is processed independently, and intrinsically it cannot fathom the sequential nature of the biological movie. In contrast, the Unet 3D takes in input 3D images. In our case, the third dimension is time and we feed in the frames in this network with their natural "chronological order". The only difference between these two networks is whether the time information is taken into account, the Fig.9.4 reveals potential "gain" in including this information.

LSTM Unet: Although 3D Unet architecture has undeniable advantages, its major deficiency is its memory complexity. To bypass this limitation we implement The LSTM (Long Short Term memory), another variation of the Unet architecture.

The LSTM architecture is introduced in [HS97] and is mostly applied in speech processing. The LSTM Unet is very similar to the 2D Unet unlike its 2D convolution layers include an LSTM unit. This network is then able to predict segmentation using the information of the past processed images. However, one slight flaw of this network is the direction in which this temporal information propagates: from beginning to end. In contrast, 3D-Unet uses the whole temporal sequence information to return its prediction.

Benchmark of the presented neural networks

We compare the three proposed networks using the following criteria:

- Performance evaluation: we use the Jaccard index, Dice index and accuracy metrics to compare our methods [CCH06]. All these metrics range between 0 and 1 where 1 is the best achievable score.
- The number of parameters is the number of trainable values within a neural network.
- Evaluation time: the frame per second (fps) is the number of 256×256 images processed by a computer per second. These numbers are given for a CPU architecture (Intel Xeon E5-2680).
- Training time: Number of hours need to train the network with an Nvidia Tesla K20c GPU card.

We also provide the pre-trained neural networks that can be used directly to segment cell images. However, we strongly recommend that the user retrains the network if a new data-set is used.

9.2.3 Tracking

Tracking algorithms very often rely on physical processes, and in many situations lead to efficient and inexpensive algorithms. Although some neural network approaches exist [Ber+16], they are often difficult to implement for a limited gain

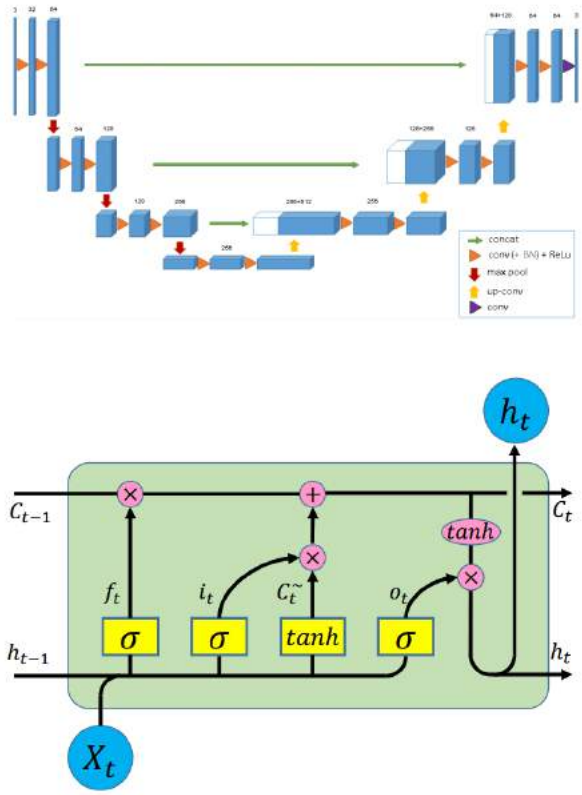


Figure 9.3: Unet architecture [Cic+16] and LSTM unit [Sha+16]. Variable C_t denotes the value of the cell, h_t denotes the hidden variable and σ some activation function (e.g. Relu).

in performance. This is particularly true under the assumption of a sufficiently large sampling rate (small time-step between images).

`Biolapse` implements a very intuitive algorithm, namely the similarity algorithm. It can be summarized in finding the association of the cells between the frame t and $t+1$ which minimizes certain criteria. In the next section we detail the selected criteria for our application. The reader's attention is drawn to the fact that this particular set of criteria may not be suitable for other applications. We therefore pay particular attention in making the presented method modular, by allowing other criteria to be added or subtracted subsequently.

Mathematical framework

`Biolapse` toolbox main goal is to reach swiftness: the tracking for a hundred time-lapse images should be performed within a few seconds. This quickness choice leads us to use similarity methods.

Given the i-th cell $C_i^t = \mathbb{1}_{\omega_i^t}$ at an time t, defined as the indicator of a sub-domain of the total image, we seek to find its next location at time $t+1$. Let I_t denote the number of cell in the time-lapse at time t. We estimate the best candidate among cells in image at time $t+1$ by solving the following optimization

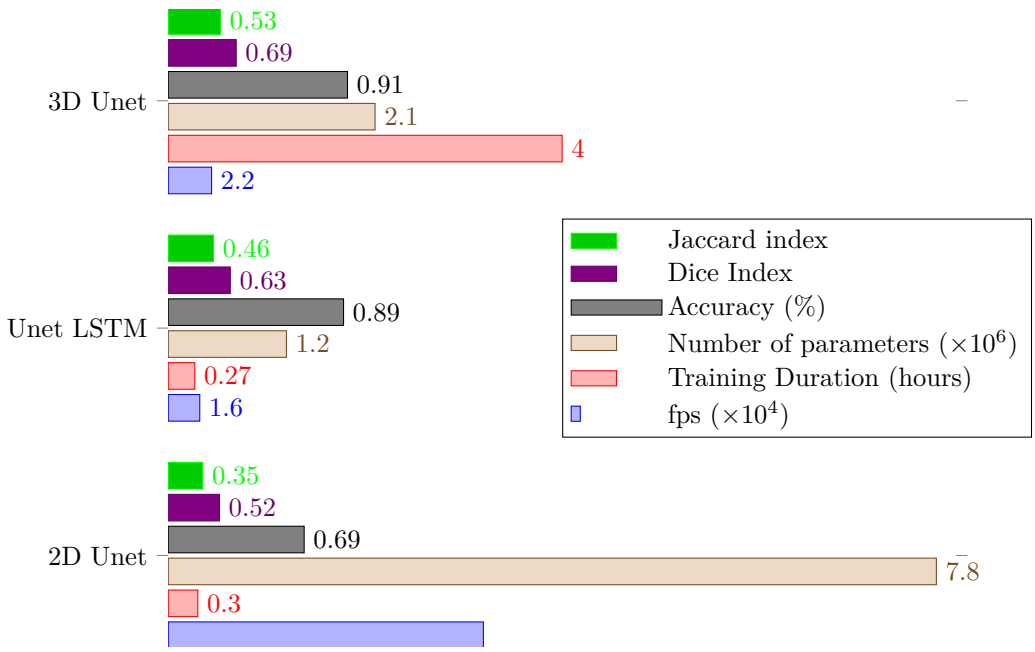


Figure 9.4: Performance of convolutional neural networks

problem

$$i \in \underset{1 \leq m \leq i_{t+1}}{\operatorname{argmin}} \sum_{k=1,n} \gamma_k \ell_k \left(C_i^t, C_m^{t+1}\right). \tag{9.1}$$

In our code we implement three measures of similarity that are :

1. Displacement of the center of mass: $\ell_1(\mathbb{1}_a, \mathbb{1}_b)$ is the square euclidean distance between the geometric centers of regions a and b.

2. Variation of the cell's area: $\ell_2(a, b)$ is the absolute value of the difference between the area of a and the area of b.

3. Variation of the turning angle:

$$\ell_3(a, b) = \frac{\langle f(a) f(b) \rangle}{\|f(a)\| \|f(b)\|},$$

where $f(a^t)$ returns the difference between the barycenter of a^t and the barycenter of a^{t-1}. This quantity is maximal if two cells move in the same direction.

The tracking procedure amounts to solve sequentially the problem (9.1) then to increment t to find the next cell. An insightful quantity is the similarity of a trajectory $(T_l)_{l \in 1..m}$ given by :

$$\sum_{t=1}^{T-1} \sum_{k=1,n} \gamma_k \ell_k \left(T_i^t, T_i^{t+1}\right).$$

The similarity measure of a trajectory is key for detecting breakdowns in the segmentation process (cells that are: disappearing, splitting or merging). Indeed, for the special case of merging, the trajectory's similarity will increase suddenly due to a huge variation of the barycenter and area of the considered cell. This method can be of course improved by adding more similarities terms in the Eq (9.1) but it increases the computation time. We empirically choose the three criteria described above which are offering a good trade-off between running time and robustness of the yielded trajectory.

Numerical implementation

In order to find the next best cell in the sense of Problem (9.1) one has to compute the similarity distance. The attentive reader has noticed that the presented method requires the computation of the barycenter for each cell. In our code, this is done with a linear time complexity using `openCV` connected component function. Hence the presented tracking method has a linear complexity with respect to the dimension of the input images.

9.2.4 Classification

Neural network are very efficient tool to determine the state a cell is [Xu+17; Eul+17]. This approach has the advantage of adapting naturally to the data used and does not require any physical expertise on the properties of the cells in each state. The main issue with this type of approach is that temporal information is not always taken into account, and even if it is, it is arduous to ensure the physical validity of the prediction.

For example, if the algorithm aims at determine if a cell is in phase G1, early S, mid S, late S or G2, we show in Fig. 9.5 some errors that may appear with an algorithm only returning a probability vector. There are two type of errors :

1. Anachronism errors, the probability of being in phase G1 and G2 are quite similar, one needs the knowledge of previous or later states.
2. Transition errors, state transition is not a sharp process: it can occur through a relatively long lapse and during several timestep. During these transitions the both state's characteristics coexist, this can lead to false predictions.

G1	G1	G1	G1	ES	ES	ES	ES	MS	MS	LS	LS	LS	LS	LS	G2

G1	G1	G2	G1	ES	ES	MS	ES	MS	MS	LS	LS	MS	LS	LS	G1

Figure 9.5: Above, sequence of states for a cells given by an expert (ground truth), states predicted by the neural network algorithm of `Biolapse` (below). In red are highlighted the errors.

To overcome this situation, we propose a two step algorithm: 1) a neural network is trained to predict the probability of each state base on a single crop image. There is no use of the time information. 2) A Markov Chain model is imposed to the previous prediction, ensuring the consistency of the prediction (e.g. phase order).

State probability

We use an encoder neural network algorithm to predict the probability of each state. It produces a mapping from a crop image to a vector of probability which indicates the likelihood to be in each state.

The main detriment for this method is the necessity to provide a learning data-set. This step can be time consuming and difficult to put into effect in many situations. `Biolapse` ease this task by providing a graphical user interface within the toolbox to label the cells of interest over time, see snapshot in Figure 9.6. This interface allows the user to quickly create a learning data-set with the appropriate format for the classification to perform well in `Biolapse`.

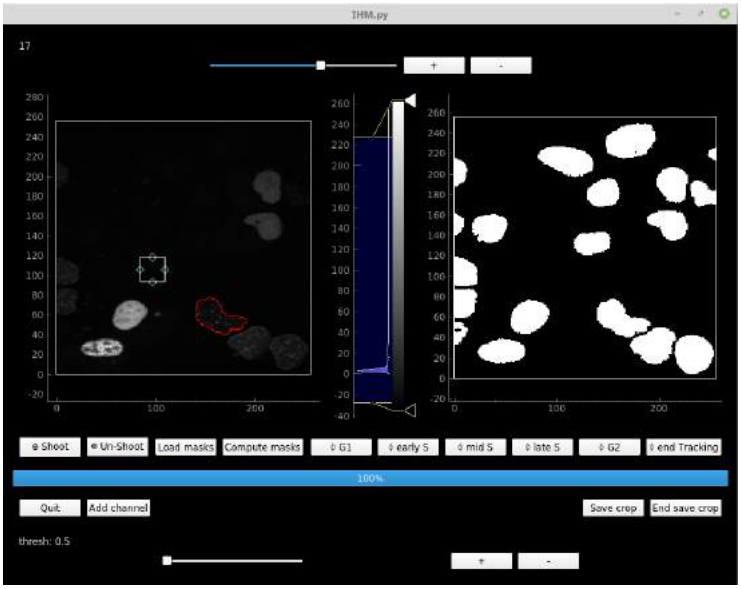


Figure 9.6: Snapshot of the interface used to label cells during the cell cycle. User load segmentation or compute it with the proposed neural network. Then the user select a cell by clicking on "Shoot", and determine the beginning of each phase by pushing "G1", 'early S", etc.

Time consistency

The last key element is to take into account the time information given by the biological sample. From a given state only few states can be reached. For instance, going back to the example where the state are G1, early S, mid S, late S and G2, the possible transitions are depicted in Figure 9.7, where the value on the edges represents the probability of the state's change.

For this purpose we model the possible succession of the states by a Markov chain. This imposes an order in the series of states and ensures the biological validity of the prediction. The computation of this Markov chain model is done using the Viterbi algorithm.

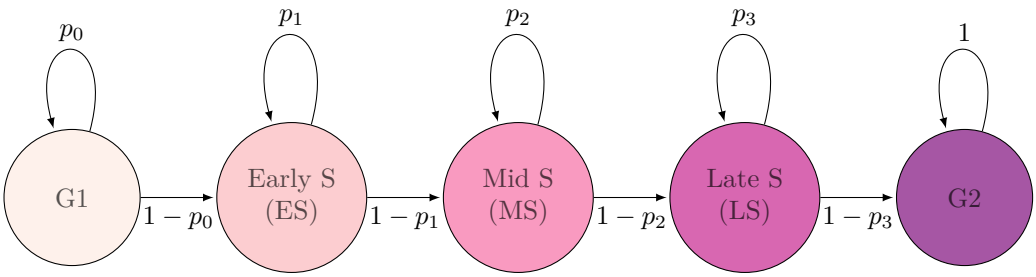


Figure 9.7: Example of state transition.

9.3 Conclusion

We presented `Biolapse`, a modular toolbox based on the *Python* programming language. This tool combines segmentation, object tracking and classification algorithms to provide a complete pipeline for extracting and classifying cells of interest within biological image film. The originality of this work is the contribution of a modular toolbox allowing the addition of new algorithms. In addition, the proposed algorithms are based on state of the art methods of machine learning and neural networks. They strongly depend on a training phase, which can be done within `Biolapse` on other data-sets. This makes it possible to rely on `Biolapse` algorithms relatively easily for new acquisitions, without expertise in machine learning and image processing.

Chapter 10

Multiview Attenuation Estimation and Correction

Résumé : *La mesure des coefficients d'atténuation est un problème fondamental qui peut être résolu par diverses techniques telles que les rayons X, la tomographie optique et le lidar atmoshpérique. Nous proposons une nouvelle approche basée sur l'observation d'un échantillon sous quelques angles différents. Ce principe peut être utilisé dans des dispositifs existants tels que le lidar ou divers types de microscopes à fluorescence. Il est basé sur la résolution d'un problème inverse non-linéaire. Nous proposons une approche de calcul spécifique pour le résoudre et montrons les mérites de l'approche sur des données simulées. Certains des outils développés présentent un intérêt indépendant. En particulier, nous proposons une méthode efficace pour corriger les défauts d'atténuation, de nouveaux estimateurs robustes pour l'équation du lidar ainsi que de nouveaux algorithmes efficaces pour calculer l'opérateur proximal de la fonction logsumexp en dimension 2.*

Abstract: *Measuring attenuation coefficients is a fundamental problem that can be solved with diverse techniques such as X-ray or optical tomography and atmospheric lidar. We propose a novel approach based on the observation of a sample from a few different angles. This principle can be used in existing devices such as lidar or various types of fluorescence microscopes. It is based on the resolution of a nonlinear inverse problem. We propose a specific computational approach to solve it and show the merits of the approach on simulated data. Some of the tools developed are of independent interest. In particular we propose an efficient method to correct attenuation defects, new robust solvers for the lidar equation as well as new efficient algorithms to compute the proximal operator of the logsumexp function in dimension 2.*

This chapter is based on the publication [DKW19]:
Debarnot, V., Kahn, J., & Weiss, P. (2019). Multiview Attenuation Estimation and Correction. Journal of Mathematical Imaging and Vision, 61(6), 780-797.

Contents

10.1 Introduction **188**
10.1.1 The basic principle 188
10.1.2 Contributions 189
10.2 Applications **190**
10.2.1 Lidar 190
10.2.2 Fluorescence microscopy 192
10.3 MAP estimator and numerical evaluation **192**
10.3.1 The discretized model 194
10.3.2 A Bayesian estimator 194
10.3.3 Making the problem convex 195
10.3.4 Density estimation with a fixed attenuation 196
10.4 Optimization methods **197**
10.4.1 Recovering the attenuation 197
10.4.2 Recovering the density 199
10.5 Additional comments **201**
10.5.1 Parameter selection 201
10.5.2 Computing times 201
10.5.3 Influence of attenuation and signal-to-noise ratio . 203
10.5.4 Toolbox 203
10.6 Comparison with existing work **205**
10.7 Conclusion & outlook **206**
10.8 Appendices **207**
10.8.1 Proof Proposition 10.3.1 207
10.8.2 Proof Proposition 10.3.2 208
10.8.3 Proximal operator of logsumexp in dimension 2 . . 208

10.1 Introduction

The ability to analyze the composition of gases in the atmosphere, the organization of a biological tissue, or the state of organs in the human body has invaluable scientific and societal repercussions. These seemingly unrelated issues can be solved thanks to a common principle: rays traveling through the sample are attenuated and this attenuation provides an indirect measurement of absorption coefficients. This is the basis of various devices such as X-ray and optical projection tomography [Nat86; Sha+02; Ver+14] or atmospheric lidar [Wei06]. The aim of this chapter is to provide an alternative approach based on the observation of the sample from a few different angles.

10.1.1 The basic principle

Let us provide a flavor of the proposed idea in an idealized 1D system. Assume that two measured (in opposite direction) signals u_1 and u_2 are formed according to the following model:

$$u_1(x) = \beta(x) \exp\left(-\int_0^x \alpha(t)\, dt\right) \text{ for } x \in [0,1] \tag{10.1}$$

and

$$u_2(x) = \beta(x) \exp\left(-\int_x^1 \alpha(t)\,dt\right) \text{ for } x \in [0,1]. \tag{10.2}$$

The function $\beta : [0,1] \to \mathbb{R}_+$ will be referred to as a density throughout the chapter. It may represent different physical quantities such as backscatter coefficients in lidar or fluorophore densities in microscopy. The function $\alpha : [0,1] \to \mathbb{R}_+$ will be referred to as the attenuation and may represent absorption or extinction coefficients. The signals u_1 and u_2 can be interpreted as measurements of the same scene under opposite directions. Equations (10.1) and (10.2) coincide with the Beer-Lambert law that is a simple model to describe attenuation of light in absorbing media. The question tackled in this chapter is: *can we recover both α and β from the knowledge of u_1 and u_2?*

Under a positivity assumption $\beta(x) > 0$ for all $x \in [0,1]$, the answer is straightforwardly positive. Setting $v(x) = \log\left(\frac{u_2(x)}{u_1(x)}\right)$, equations (10.1) and (10.2) yield:

$$v(x) = \int_0^x \alpha(t)\,dt - \int_x^1 \alpha(t)\,dt. \tag{10.3}$$

Therefore

$$\alpha(x) = \frac{1}{2}\frac{\partial}{\partial_x} v(x) \tag{10.4}$$

and

$$\beta(x) = \frac{u_1(x)}{\exp\left(-\int_0^x \alpha(t)\,dt\right)}. \tag{10.5}$$

Unfortunately, formulas (10.4) and (10.5) only have a theoretical interest: they cannot be used in practice since computing the derivative of a log of a ratio is extremely sensitive to noise and thus very unstable from a numerical point of view. We will therefore design a numerical procedure based on a Bayesian estimator to retrieve the density β and attenuation coefficients α in a stable and efficient manner. It is particularly relevant when the data suffer from Poisson noise.

10.1.2 Contributions

This chapter contains various contributions listed below.

- We show that it is possible to retrieve attenuation coefficients from multi-view measurements in different systems such as lidar, confocal or SPIM microscopes. Figure 10.1 summarizes the proposed idea. The attenuation, which is usually considered as a nuisance in confocal microscopy is exploited to measure the absorption coefficients. The algorithm successfully retrieves estimates of the density and attenuation from two attenuated and noisy images. Let us also mention that some researchers already proposed to measure absorption and correct attenuation by combining optical projection tomography and SPIM imaging [May+14]. The principle outlined here shows that much simpler optical setups (a traditional confocal microscope) theoretically allows estimating the same quantities.

- We propose novel Bayesian estimators for the density α and the attenuation β based on a Poisson noise modeling.

- The proposed estimators are solutions of a nonconvex problem. We show that exact solutions of the problem can be obtained by using a trick making the problem convex.
- The resulting convex program is challenging from a numerical point of view and involves functions that are uncommon in imaging. This leads us to develop an efficient algorithm to compute the proximal operator of the logsumexp function in dimension 2.
- The proposed estimators also seem to be novel for the standard mono-view inverse problem in lidar and for correcting attenuation defects under a Poisson noise assumption with multiple views.
- We perform a numerical validation of the proposed ideas on synthetic data, showing the well-foundedness of the approach. The validation of the method on specific devices is left as an outlook for future works.

We found the general principle stated above independently, but became aware of a few papers proposing similar concepts after finishing the manuscript. In atmospheric lidar, the idea was explored in the 1980's already [Kun87; HP88; CF10]. In confocal microscopy, the recent paper [Sch+13] proposes a setting very similar to the one proposed here. From a practical point of view, the early papers [Kun87; HP88; CF10] were based on simple Wiener filtering approaches. Our tests using these approaches on simulations led us to the conclusion that they were far too unstable and we will not report these results. On the other hand, the paper [Sch+13] presents a closely related framework: the authors use a maximum a posteriori estimator with a total variation regularizer, leading to a variational formulation of the problem. We will see however that its structure seems less amenable to an efficient numerical resolution. Our conclusion using the L-BFGS approach suggested in [Sch+13] is that unless a very good initial guess is provided, the method is unable to retrieve the attenuation, whereas our globally convergent approach is insensitive to the starting point.

10.2 Applications

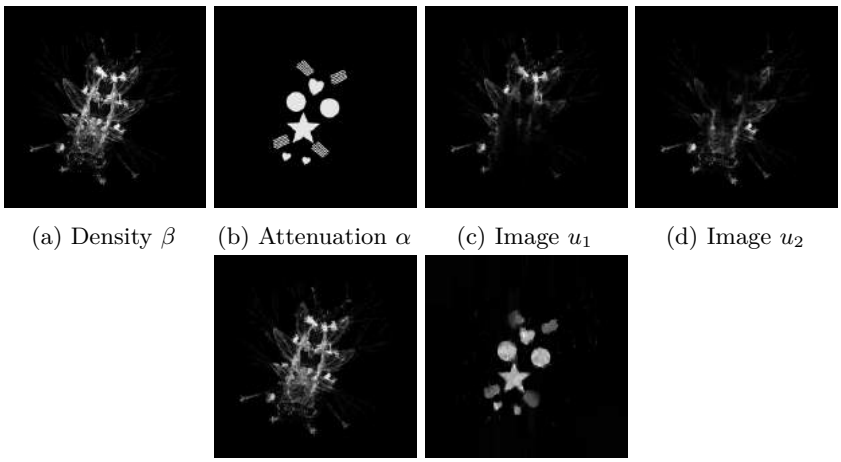
In this section, we show various applications where the methodology proposed in this chapter can be applied.

10.2.1 Lidar

In atmospheric lidar, an object (atmosphere, gas,...) is illuminated with a laser beam. Particles within the object reflect light. The time to return of the reflected light is then measured with a scanner. The received signal $u_1(x)$ is the backscattered mean power at altitude x for a specific wavelength. The density β corresponds to the backscattered coefficient, while α is called extinction coefficient. The equation relating u_1 to α and β is:

$$u_1(x) = \mathcal{P}\left(\frac{C}{x^2}\beta(x)\exp\left(-2\int_0^x \alpha(t)\,dt\right)\right), \tag{10.6}$$

where C is independent of x. The notation $\mathcal{P}(z)$ stands for a Poisson distributed random variable of parameter z. The Poisson distribution is a rather good



(a) Density β (b) Attenuation α (c) Image u_1 (d) Image u_2

(e) Estimated β SNR=22.4dB (f) Estimated α SNR=9.4dB

Figure 10.1: Illustration of the contribution. A sample (here an insect) has a fluorophore density β shown in Fig. 10.1a and an attenuation map α shown in Fig.10.1b. The two measured images u_1 and u_2 are displayed in Fig. 10.1c and 10.1d. As can be seen, they are attenuated differently (top to bottom and bottom to top) since the optical path is reversed. From these two images, our algorithm provides a reliable estimate of each map in Fig. 10.1e and 10.1f despite Poisson noise.

noise model in lidar, since measurements describe a number of detected photons. The term $\frac{C}{x^2}\beta(x)$ appears in the lidar equation (10.6) instead of simply β. The algorithm developed later will allow retrieving $\frac{C}{x^2}\beta(x)$ instead of β. This is not a problem since there is a direct known relationship between both.

Remark 10.2.1. *In Raman lidar, the coefficient β corresponds to the molecular density of the atmosphere, while α is the sum of extinction coefficients at different wavelengths. The theory developed herein also applies to this setting.*

When the backscatter coefficient β has a known analytical relationship with the extinction coefficient α, direct inversion is possible. A popular method is Klett's formula [Kle81] for instance. Alternative formula exist [ARW90] when the backscatter coefficient is known. The recent trend consists in using iterative methods coming from the field of inverse problems [Shc07; Por+08; Gar+16], leading to improved robustness. All these approaches crucially depend on a precise knowledge of the backscatter coefficient. This is a strong hypothesis that is often rough or unreasonable in practice.

To overcome this issue, a few authors proposed to use two opposite lidars and to retrieve the attenuation coefficients using equation (10.4) [Kun87; HP88; CF10]. The stability to noise was ensured by linear filtering of the input and output data. Our simulations using them did not yield satisfactory results and we will not report them.

10.2.2 Fluorescence microscopy

The principle proposed herein can also be applied to some fluorescence microscopes. This idea was already proposed in confocal microscopy [Sch+13]. Here we show that it can be extended to other microscopes such as 4π or selective plane illumination microscopes (SPIM).

All fluorescence microscopes share a common principle: a source of illumination excites fluorophores within the sample, which in turn emit some light. This light is collected with a camera. Both the illumination and emission light can be absorbed along its optical path, which results in inhomogeneities in the image contrasts.

Depending on the imaging device, the way light absorption distorts images can be different. We illustrate this with two synthetic examples in Fig. 10.3. In a confocal microscope, the illumination and emission light both travel in the same direction, creating unidirectional absorption. Two images suffering from opposite contrast losses can be obtained by rotating the sample or by using a 4-pi microscope [CC74; HS92], see Fig. 10.3a.

In the multi-view versions of the selective plane illumination microscopy SPIM (also called light sheet fluorescence microscopy) [HS07; Krz+12; Tom+12; Chh+15], the illumination and emission light travel in orthogonal directions, creating bi-directional contrast losses, see Fig. 10.3b.

The attenuation map α is wavelength dependent and to be precise, we should consider two attenuation maps α_i and α_e for the illumination and emission light respectively. In this work, we simply assume that the two are related through a linear relationship $\alpha = \alpha_e = \kappa\alpha_i$, where κ is a positive scalar. Under this assumption, the microscopes provide a set of images $(u_i)_{1\leq i\leq m}$, which can be modeled as follows:

$$u_i = \mathcal{P}\left(\beta \exp(-A_i\alpha)\right), \tag{10.7}$$

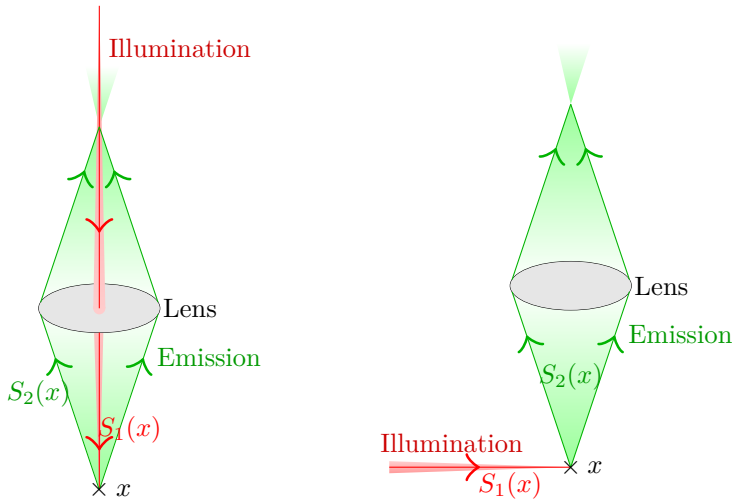
where A_i denotes a linear integral operator that depends on the geometry of the optical setup. Figure 10.2 provides a description of these operators in the case of a confocal microscope and of a multi-view SPIM microscope. For a point $x \in \mathbb{R}^d$, the expression of $(A_i\alpha)(x)$ is given by:

$$(A_i\alpha)(x) = \int_{S_1(x)} \alpha(y)\mathrm{d}y + \kappa \int_{S_2(x)} \alpha(y)\mathrm{d}y,$$

where $S_1(x)$ and $S_2(x)$ are the cones of light depicted in Fig. 10.2. In our numerical experiments, we use a simple version where the operators return line integrals (and not cone integrals). This amounts to assuming that the light rays are infinitely thin. However, the proposed approach may be extended to arbitrary geometries through the use of heavier linear algebra solvers.

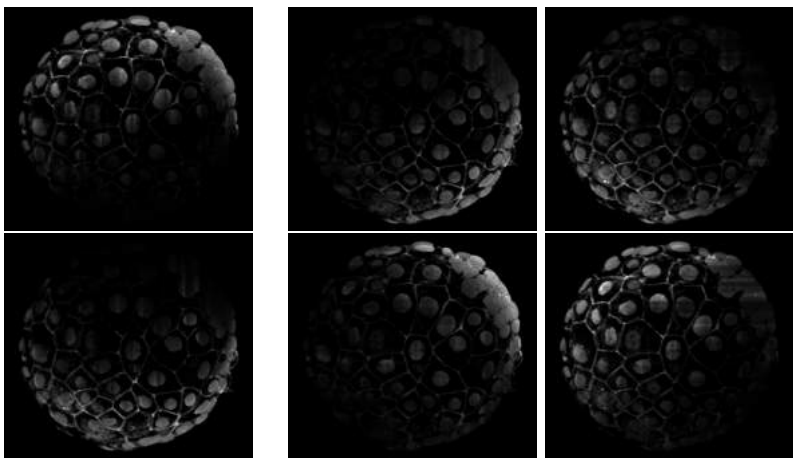
10.3 MAP estimator and numerical evaluation

In this section, we describe our numerical procedure completely. We start by providing a discrete version of our image formation model. Then, we design a Bayesian estimator of the attenuation map α and of the density β. We finally design an effective optimization algorithm to compute our statistical estimator.



(a) Confocal (b) Multi-SPIM

Figure 10.2: Path of the light from different microscopes.



(a) Confocal microscope (b) Multi-SPIM microscope

Figure 10.3: Simulated contrast loss in a slice of mouse embryo with a confocal microscope (left) and a multi-view light sheet fluorescence microscope (right).

10.3.1 The discretized model

The discrete model we consider in this chapter reads:

$$\begin{cases} u_1 & = \mathcal{P}\left(\beta \exp(-A_1\alpha)\right) \\ \vdots & \\ u_m & = \mathcal{P}\left(\beta \exp(-A_m\alpha)\right). \end{cases} \tag{10.8}$$

The signals u_i, β and α are assumed to be nonnegative and belong to $\mathbb{R}^n$, where $n = n_1 \dots n_d$ denotes the number of pixels and d is the space dimension. The value of a vector u_1 at location $i = (i_1, \dots, i_d)$ will be denoted either $u_1[i]$ or $u_1[i_1, \dots, i_d]$. The matrices A_i in $\mathbb{R}^{n\times n}$ are discretization of linear integral operators. In our numerical experiments, we use $m = 2$ views and the product A_1u_1 simply represents the cumulative sum of u_1 along one direction while the product A_2u_2 represents the cumulative sum of u_2 in the opposite direction. For instance, for a 1D signal, we set:

$$(A_1u)[i] = \sum_{j=1}^{i} u_1[j].$$

Therefore, the matrix A_1 has the following lower triangular shape:

$$A_1 = \begin{pmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \ddots & \dots & 0 \\ 1 & 1 & 1 & 1 & \dots & 1 \end{pmatrix} \tag{10.9}$$

We are now ready to design a Bayesian estimator of α and β from model (10.8).

10.3.2 A Bayesian estimator

The Maximum A Posteriori (MAP) estimators $\hat{\alpha}$ and $\hat{\beta}$ of α and β are defined as the maximizers of the conditional probability density:

$$\max_{\alpha\in\mathbb{R}^n, \beta\in\mathbb{R}^n} \mathbb{P}(\alpha, \beta | (u_i)_{1\le i\le m}).$$

By using the Bayes rule and a negative log-likelihood, this is equivalent to finding the minimizers of:

$$\min_{\alpha\in\mathbb{R}^n, \beta\in\mathbb{R}^n} -\log(\mathbb{P}((u_i)_{1\le i\le m}|\alpha, \beta)) - \log(\mathbb{P}(\alpha, \beta)).$$

Let us evaluate $\mathbb{P}((u_i)_{1\le i\le m}|\alpha, \beta)$. To this end, set

$$\lambda_j = \beta \exp(-(A_j\alpha)).$$

Since the distribution of a Poisson distributed random variable with parameter λ has the following probability mass function:

$$\mathbb{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!},$$

we get:

$$-\log\left(\mathbb{P}((u_i)_{1\leq i\leq m}|\alpha,\beta)\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} \lambda_j[i] - u_j[i]\log(\lambda_j[i]) + C,$$

where C is a value that does not depend on α and β. Next, we assume that α and β are *independent random vectors* with probability distribution functions of type:

$$\mathbb{P}(\alpha) \propto \exp(-R_\alpha(\alpha)) \text{ and } \mathbb{P}(\beta) \propto \exp(-R_\beta(\beta)),$$

where $R_\alpha : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ and $R_\beta : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ are regularizers describing properties of the density and attenuation maps. Overall, the optimization problem characterizing the MAP estimates reads:

$$\min_{\alpha\in\mathbb{R}^n,\beta\in\mathbb{R}^n} F(\alpha,\beta) \tag{10.10}$$

where

$$F(\alpha,\beta) = R_\alpha(\alpha) + R_\beta(\beta) + \left\langle \sum_{j=1}^{m} \beta\exp\left(-A_j\alpha\right) + u_j\left(A_j\alpha - \log(\beta)\right), \mathbb{1}\right\rangle \tag{10.11}$$

and $\mathbb{1}$ stands for the vector in $\mathbb{R}^n$ with all components equal to 1.

Remark 10.3.1. *For $m = 1$ view, the problem $\min_\alpha F(\alpha,\beta)$ allows recovering the attenuation knowing the density: this is the standard inverse problem met in lidar. To the best of our knowledge, the proposed formulation is novel for this problem.*

Remark 10.3.2. *The problem $\min_\beta F(\alpha,\beta)$ corresponds to correcting the attenuation on the density map. This is also a frequently met problem [RV91; RB93; Can+03; KLP04] and to the best of our knowledge, the proposed approach - based on the MAP principle - is original, though it bear resemblances with [Sch+13] for instance.*

10.3.3 Making the problem convex

Let us start by analyzing the convexity properties of the function F. To this end, let us introduce the function $G : \mathbb{R}^n \times \mathbb{R}^n_+ \to \mathbb{R}$ defined by:

$$G(\alpha,\beta) = \left\langle \sum_{j=1}^{m} \beta\exp\left(-A_j\alpha\right) + u_j\left(A_j\alpha - \log(\beta)\right), \mathbb{1}\right\rangle.$$

Notice that $F(\alpha,\beta) = G(\alpha,\beta) + R_\alpha(\alpha) + R_\beta(\beta)$. The following proposition provides the domain of convexity of G.

Proposition 10.3.1. *The function G*

- *is convex on each variable separately on $\mathbb{R}^n \times \mathbb{R}^n_+$.*
- *is non convex on $\mathbb{R}^n \times \mathbb{R}^n_+$.*

- *has a positive semidefinite Hessian on the (nonconvex) set:*

$$\left\{(\alpha,\beta)\in\mathbb{R}^n\times\mathbb{R}^n_+, \beta\sum_{j=1}^m \exp(-A_j\alpha)\leq\sum_{j=1}^m u_j\right\}. \tag{10.12}$$

This proposition is proved in the appendix 10.8.1. It shows that if R_α and R_β are convex functions, then F is convex in each variable separately. Unfortunately, it is nonconvex on the product space unless the regularizers R_α and R_β compensate for the nonconvexity. The main observation in this paragraph is that it is possible to find a global minimizer of F when R_α is a standard convex regularizer and R_β is the indicator function of the positive orthant. This property is related to the third item in Proposition 10.3.1.

Proposition 10.3.2. *Set*

$$R_\beta(\beta)=\iota_{\mathbb{R}^n_+}(\beta):=\begin{cases}0 & \text{if } \beta[i]\geq 0, \forall i\in\{1,\dots,n\},\\ +\infty & \text{otherwise.}\end{cases}$$

Then, if they exist, the solutions $(\hat{\alpha},\hat{\beta})$ *of problem* (10.10) *are given by:*

$$\hat{\alpha}\in\underset{\alpha\in\mathbb{R}^n}{\operatorname{argmin}} R_\alpha(\alpha)+\left\langle\sum_{j=1}^m u_j\left[A_j\alpha+\log\left(\sum_{i=1}^m\exp(-(A_i\alpha))\right)\right],\mathbb{1}\right\rangle \tag{10.13}$$

and

$$\hat{\beta}=\frac{\sum_{j=1}^m u_j}{\sum_{j=1}^m\exp\left(-A_j\hat{\alpha}\right)}. \tag{10.14}$$

In addition, Problem (10.13) *is convex if the regularizer* R_α *is convex.*

The expression (10.14) can be seen as a simple estimator of β knowing $(u_j)_{1\leq j\leq m}$ and α. Notice that it coincides exactly with the boundary of the set (10.12). The existence and uniqueness of minimizers can also be shown by adding assumptions on R_α, such as strict convexity. We do not study this question further, since at this point we state results for arbitrary regularizers.

The convexity of problem (10.13) is critical: it shows that *global* minimizers of (10.14) can likely be computed if the regularizer R_α is chosen wisely. This observation motivates solving (10.13) to get an estimate of $\hat{\alpha}$. The only problem is that the estimated density $\hat{\beta}$ is regularized mildly using the sole non-negativity assumption. Hence, we propose an additional denoising step in the next paragraph.

10.3.4 Density estimation with a fixed attenuation

In order to remove the noise from the density, we use once again a MAP estimator with a more advanced regularizer R_β, assuming that the true attenuation α is actually equal to $\hat{\alpha}$.

Following Section 10.3.2, we get that the estimator $\hat{\hat{\beta}}$ is given by:

$$\hat{\hat{\beta}}\in\underset{\beta\in\mathbb{R}^n_+}{\operatorname{argmin}} F(\hat{\alpha},\beta)=\underset{\beta\in\mathbb{R}^n_+}{\operatorname{argmin}} R_\beta(\beta)+ \tag{10.15}$$

$$\left\langle\sum_{j=1}^m\beta\exp\left(-A_j\hat{\alpha}\right)+u_j\left(A_j\hat{\alpha}-\log\left(\beta\right)\right),\mathbb{1}\right\rangle,$$

where we assumed that β is a random vector with probability distribution function of type:

$$\mathrm{p}(\beta) \propto \exp\left(-R_\beta(\beta)\right).$$

Once again, the global solution of this problem can be computed by choosing a sufficiently simple convex regularizer R_β.

Proposition 10.3.3. *Problem* (10.15) *is convex for a convex regularizer* R_β.

Proof. The proof derives directly from Proposition 10.3.1, first item. □

10.4 Optimization methods

10.4.1 Recovering the attenuation

We now delve into the numerical resolution of (10.13). First, we need to choose a convex regularizer R_α. In this chapter, we propose to simply use the total variation [ROF92] together with a non-negativity constraint, which is well known to preserve sharp edges. Its expression is given by:

$$R_\alpha(\alpha) = \lambda_\alpha \sum_{i=1}^{n} \|(\nabla\alpha)[i]\|_2 + \iota_{\mathbb{R}^n_+}(\alpha),$$

where $\nabla : \mathbb{R}^n \to \mathbb{R}^{dn}$ is a discretization of the gradient, $\lambda_\alpha \geq 0$ is a regularization parameter and $\iota_{\mathbb{R}^n_+}$ is the indicator of the nonnegative orthant. We will use the standard discretization proposed in [Cha04] in our numerical experiments.

Problem (10.13) is convex, but rather hard to minimize for various reasons listed below. First, the vectors α and β may be very high dimensional, preventing the use of an arbitrary black-box method. Second, the regularizer R_α is non differentiable. Third, the operators A_i have a spectral norm depending on the dimension n, preventing the use of gradient based methods since the Lipschitz constant of the gradient would be too high, see Proposition 10.4.1. Last, the proximal operator associated to the logsumexp function has no simple analytical formula.

Proposition 10.4.1. *Matrix* A_1 *in* (10.9) *satisfies* $\|A_1\|_{2\to 2} \gtrsim n$, *where* $\|\cdot\|_{2\to 2}$ *stands for the spectral norm.*

Proof.

$$\begin{aligned}\|A_1\|_{2\to 2}^2 &\geq \left\| A_1 \begin{pmatrix} 1/\sqrt{n} \\ \vdots \\ 1/\sqrt{n} \end{pmatrix} \right\|_2^2 \geq \frac{1}{n} \left\| \begin{pmatrix} 1 \\ 2 \\ \vdots \\ n \end{pmatrix} \right\|_2^2 \\ &\gtrsim \frac{n^3}{n} = n^2.\end{aligned}$$

□

A large number of splitting methods have been developed to solve problems of type (10.13), and we refer to the excellent review papers [CP11; CP16] for an

overview. It has been shown to perform well in presence of Poisson noise [FB10]. Among them, the Simultaneous Direction Method of Multipliers (SDMM), a variant of the ADMM [FG00; NWY10] is particularily adapted to the structure of our problem. This algorithm allows solving problems of type:

$$\min_{\alpha \in \mathbb{R}^n} g_1(L_1\alpha) + \ldots + g_m(L_m\alpha), \tag{10.16}$$

where functions $g_i : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ are convex closed and the operators $L_i : \mathbb{R}^n \to \mathbb{R}^{m_i}$ are linear and such that $Q = \sum_{i=1}^n L_i^T L_i$ is an invertible matrix.

To cast problem (10.13) into form (10.16), we use the following choices. We set $L_2 = c_2\nabla$

$$\begin{aligned} L_1 : \quad \mathbb{R}^n &\to \mathbb{R}^{2n} \\ \alpha &\mapsto c_1 \begin{pmatrix} A_1\alpha \\ A_2\alpha \end{pmatrix} \end{aligned}, \tag{10.17}$$

$$\begin{aligned} g_1 \quad : \mathbb{R}^{2n} &\to \mathbb{R} \cup \{+\infty\} \\ \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} &\mapsto \textstyle\sum_{i=1}^n \sum_{j=1}^2 u_j[i] \Big(z_j[i]/c_1 \\ &+ \log\Big(\textstyle\sum_{j=1}^2 \exp\ (-z_j[i]/c_1) \Big) \Big), \end{aligned}$$

and

$$g_2 \begin{pmatrix} z_1 \\ \vdots \\ z_d \end{pmatrix} = \frac{\lambda}{c_2} \sum_{i=1}^n \sqrt{z_1^2[i] + \ldots + z_d[i]^2},$$

$$L_3 = c_3 I_n \text{ and } g_3(z) = \iota_{\mathbb{R}^n_+}(z).$$

The numbers c_1, c_2, c_3 are positive constants allowing to accelerate the algorithm's convergence by balancing the relative importance of each term. This can also be seen as a simple diagonal preconditioner. In our numerical experiments, we set $c_1 = 1$ and tune c_2 and c_3 manually to accelerate convergence.

The SDMM then takes the algorithmic form described in Algorithm 13.

Algorithm 13 The SDMM algorithm to solve (10.16)

input: Nit, $\gamma > 0$, $(y_{i,0})_{1 \le i \le m} = 0_{\mathbb{R}^m}$, $(z_{i,0})_{1 \le i \le m} 0_{\mathbb{R}^m}$
for $k = 1$ to Nit **do**
  $x_k = Q^{-1} \sum_{i=1}^m L_i^T (y_{i,k} - z_{i,k})$
  for $i = 1$ to m **do**
    $s_{i,k} = L_i x_k$.
    $y_{i,k+1} = \mathrm{prox}_{\gamma g_i}(s_{i,k} + z_{i,k})$
    $z_{i,k+1} = z_{i,k} + s_{i,k} - y_{i,k+1}$
  end for
end for

In order to apply Algorithm 13, we need to compute the proximal operators of each function g_i, defined by:

$$\mathrm{prox}_{\gamma g_i}(z_0) = \underset{z \in \mathbb{R}^{m_i}}{\mathrm{argmin}}\, \gamma g_i(z) + \frac{1}{2}\|z - z_0\|_2^2.$$

The proximal operators of g_2 and g_3 have closed form solutions found in nearly all recent total variation minimization solvers. We refer to [NWY10] for instance. Unfortunately, the proximal operator of g_1 has no closed-form expression. In order to compute it, we propose using a non trivial Newton-based algorithm described in section 10.8.3. Finally, we need to evaluate matrix-vector products with Q^{-1}. This can be achieved using either a LU factorization or a conjugate gradient. The LU factorization is feasible if the operator acts independently on each column of the image, since the matrix then has a moderate size. It is not feasible in general for arbitrary integral operators due to the large image size. Hence, in our numerical experiments, we simply use a conjugate gradient (CG) algorithm. The precision of the resolution is fixed and the CG algorithm is initialized with the result at the previous iteration. In practice, we observe that 10 iterations are enough for the overall algorithm to converge.

To conclude this paragraph, we illustrate the results obtained by the described procedure in Fig. 10.4.

10.4.2 Recovering the density

In this paragraph, we focus on the resolution of problem (10.15). This amounts to simultaneously correcting the attenuation and denoising the resulting image. This is a rather simple inverse problem, but it seems original due to the noise statistics. A Poisson distributed variable multiplied by a positive constant different from 1 is not Poisson anymore. This makes the proposed algorithm similar, but different from existing approaches developed for Poisson noise in [DFS09; ST10] for instance.

A simple idea to regularize the problem is to use the total variation again, i.e. to set

$$R_\beta(\beta) = \lambda_\beta \sum_{i=1}^{n} \|(\nabla\beta)[i]\|_2.$$

Once again the resulting problem can be solved with the SDMM. Let us detail this procedure. Define

$$a[i] = \sum_{j=1}^{2} \exp(-(A_j\alpha)[i]) \text{ and } u[i] = \sum_{j=1}^{2} u_j[i].$$

The problem then reads:

$$\begin{aligned}
&\min_{\beta\in\mathbb{R}^n_+} \sum_{i=1}^{n} a[i]\beta[i] - u[i]\log(\beta[i]) + \lambda_\beta \sum_{i=1}^{n} \|(\nabla\beta)[i]\|_2 \\
&= \min_{\beta\in\mathbb{R}^n} f_1(L_1\beta) + f_2(L_2\beta),
\end{aligned}$$

with $L_1 = c_1 I_n$,

$$\begin{aligned}
f_1 \quad &: \mathbb{R}^n \quad \to \mathbb{R}\cup\{+\infty\} \\
&\quad\ z \quad\ \mapsto \iota_{\mathbb{R}^n_+}(z) + \tfrac{1}{c_1}\textstyle\sum_{i=1}^{n} a[i]z[i] - u[i]\log(z[i]),
\end{aligned}$$

$L_2 = c_2\nabla$ and

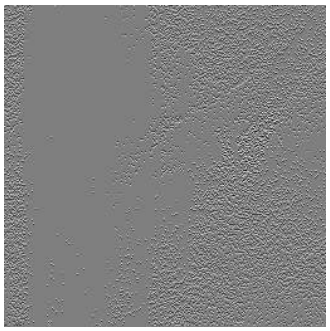
$$\begin{aligned}
f_2 \quad &: \mathbb{R}^{2n} \quad \to \mathbb{R} \\
&\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \quad \mapsto \tfrac{\lambda_\beta}{c_2}\textstyle\sum_{i=1}^{n} \sqrt{z_1[i]^2 + z_2[i]^2}.
\end{aligned}$$

(a) Density $\beta \in [0, 100]^n$

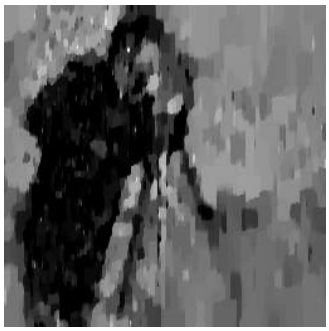
(b) Attenuation $\alpha \in [0, 0.03]^n$

(c) Image u_1

(d) Image u_2



(e) (10.14): SNR=-53.1dB



(f) TV: SNR=11.9dB

Figure 10.4: Illustration of the limits of a direct estimate of the attenuation coefficients and output of a warm start initialization. (10.4a) and (10.4b) are the original density and attenuation. (10.4c) and (10.4d) are the observed signals. (10.4e) is the direct density estimate (10.4). As can be seen, the formula yields useless results since it is completely unstable to noise. (10.4f) is the density estimate using the total variation solver with Algorithm 13. It allows recovering the main details of the cameraman, despite a significant amount of noise.

The proximal operators of f_2 is standard and we do not detail it here. The proximal operator of f_1 is provided below:

Proposition 10.4.2. *We have:*

$$\mathrm{prox}_{\gamma f_1}(z_0) = \frac{-(\gamma/c_1 a - z_0) + \sqrt{(\gamma/c_1 a - z_0)^2 + 4\gamma u}}{2}. \tag{10.18}$$

Proof. It suffices to write the first order optimality conditions of $\min_{z\geq 0} 1/2\|z - z_0\|_2^2 + a/c_1 z - u\log(z/c_1)$. This shows that z is the root of a second order polynomial. Its only positive root is given in (10.18). □

We show a typical result of total variation minimization in Fig. 10.5. Pamareter λ_β was chosen manually so as to maximize the SNR of the result.

10.5 Additional comments

10.5.1 Parameter selection

Data terms The two data term parameters are λ_α and λ_β. They specify the regularity of the attenuation and the density respectively. In all our experiments, we optimized them by trial and error. We observed experimentally, that similar results are obtained within a relatively large range, making a manual optimization quite easy. In addition, for a given measurement device, the same parameter is likely to be always the same, decreasing the interest of an automatized procedure such as SURE.

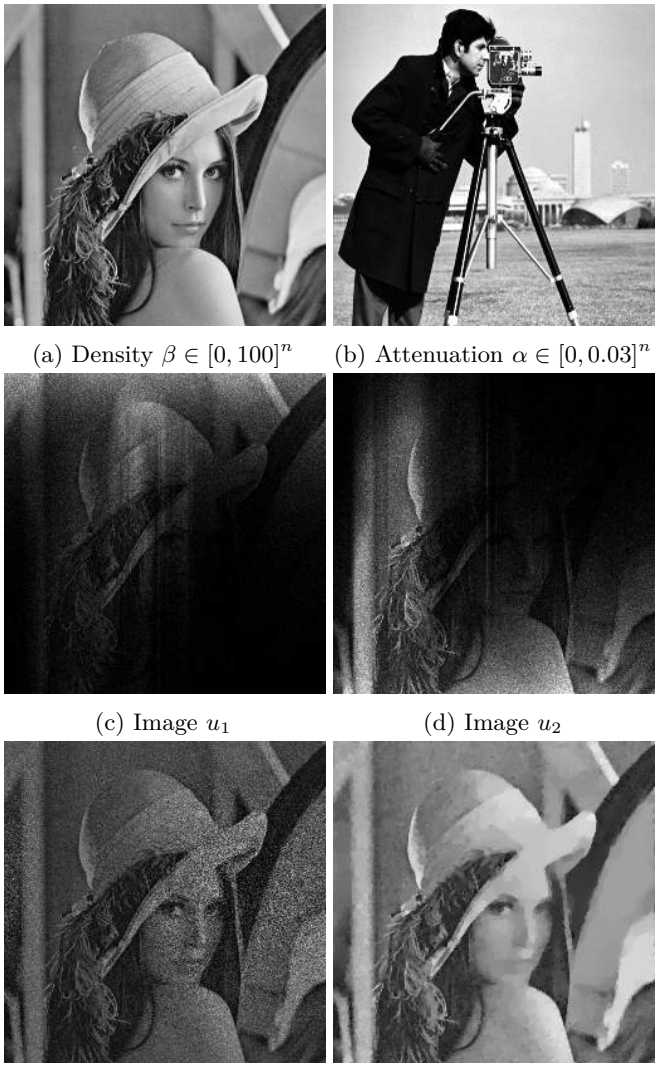
Algorithms parameters The optimization algorithms are based on the SDMM and their convergence rates depend a lot on the paramaters γ, c_1, c_2, c_3 and c_4. They may converge to a satisfactory solution rapidly (about 50 iterations) or slowly (more than 10000 iterations) depending on these choices. Unfortunately, we found no systematic method to choose them and also used a trial and error strategy in our numerical experiments. Our numerical experiments suggest that these parameters are suitable for a wide range of data (image size, maximum attenuation, image dynamics), so that the tedious tuning can be done once and for all for a given application.

10.5.2 Computing times

All the experiments of this chapter were performed on a laptop with an Intel i7 processor with 4-cores. The codes were written mostly in Matlab (natively parallel), with some parts written in C with OpenMP support.

The complexity of the proposed algorithms scale roughly linearly with the number of pixels n, as shown in Fig. 10.6. We observed that the number of iterations of the SDMM to reach a given relative accuracy remains the same whatever the size n, while the cost per iteration scales linearly with it (at least for the cumulative sum integral operators considered herein).

As can be seen on Fig. 10.6 the algorithm takes around 48 seconds for a 256×256 image. Out of these, 45 seconds are spent to recover the attenuation, while the 3 remaining are dedicated to correct the density.



(a) Density $\beta \in [0, 100]^n$ (b) Attenuation $\alpha \in [0, 0.03]^n$

(c) Image u_1 (d) Image u_2

(e) (10.14): SNR=12.9dB (f) TV: SNR=21.9dB

Figure 10.5: Recovering the density knowing the exact attenuation, with a non regularized estimator or a total variation solver. (10.5a) and (10.5b) are the original density and attenuation. (10.5c) and (10.5d) are the observed signals. (10.5e) is the direct density estimate (10.14). (10.5f) is the density estimate using a total variation solver.

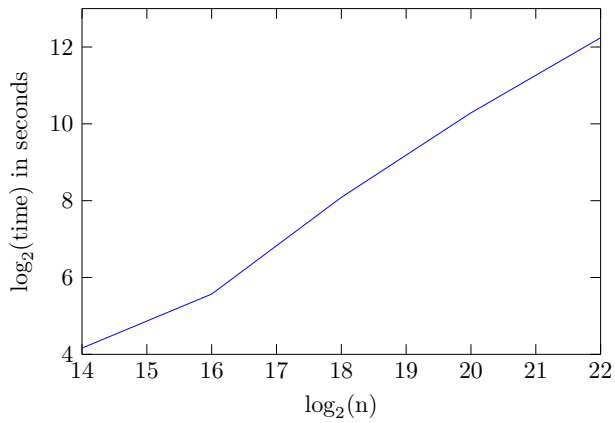


Figure 10.6: Time needed to compute the warm start estimate and correct the attenuation with respect to the number n of pixels (in log-log scale). A linear regression indicates that the slope is roughly equal to 1, showing a linear dependency with respect to the number of pixels.

All codes can be easily parallelized on a GPU. A speed-up of 100 can be expected on such an architecture, making the proposed methods suitable for large 2D or 3D images.

10.5.3 Influence of attenuation and signal-to-noise ratio

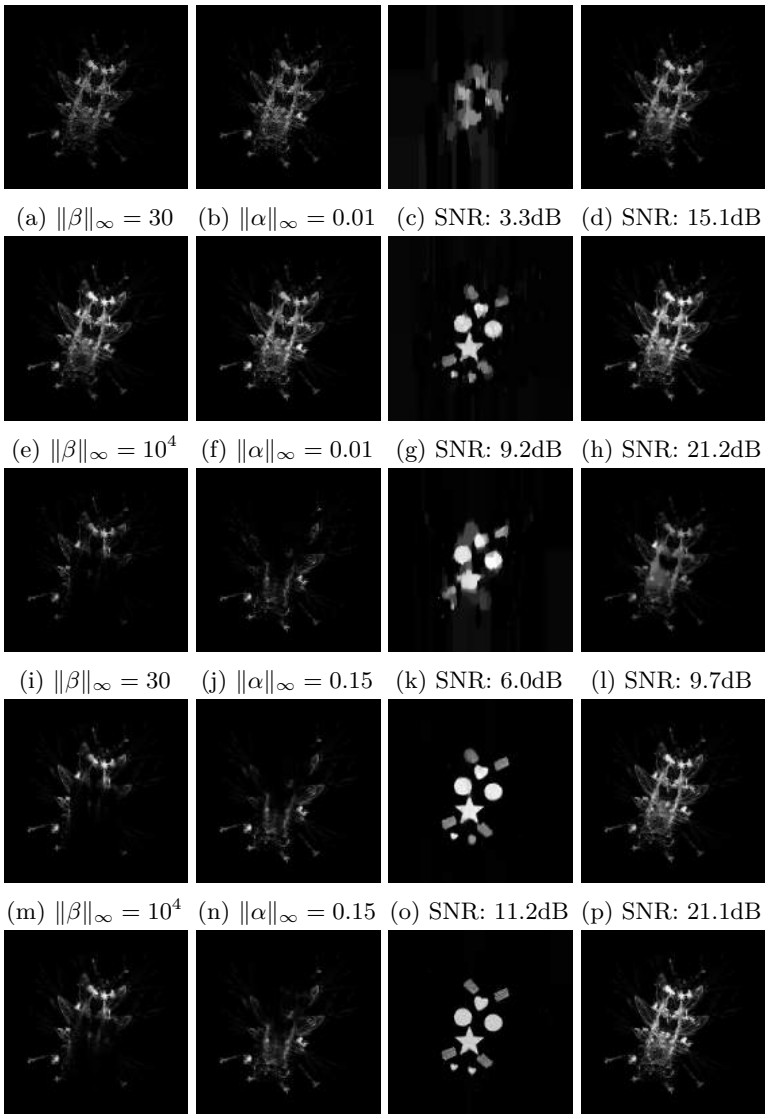
Two parameters strongly influence the ability to recover the attenuation and density: the signals dynamics (or signal-to-noise ratio) and the attenuation dynamics.

As the signal-to-noise ratio decreases, it becomes impossible to recover fine details. For instance, the fine stripes are not recovered in Fig. 10.1e, but they are recovered for signals with a much higher amplitude. In Fig. 10.7, it can indeed be verified that a high dynamics of 10^5 allows recovering most of the stripes. This experiment shows that highly sensitive EMCCD cameras should be preferred over more standard devices for this specific application.

The attenuation amplitude also plays a key role: if it is too low, then no attenuation can be detected. On the contrary, if it is too high, then the signals u_1 and u_2 will vanish too rapidly, making it impossible to evaluate the attenuation. This is illustrated in Fig. 10.7. It is remarkable that the algorithm manages to recover the attenuation partially for very low signal-to-noise ratio. In Fig. 10.7c, we observe that the attenuation is partially recovered with no more than 30 expected photons per pixel!

10.5.4 Toolbox

A Matlab toolbox containing all the main algorithms described in this chapter is provided on the website of the authors `https://www.math.univ-toulouse.fr/~weiss/` and on GitHub `https://github.com/pierre-weiss/MAEC`. The Lambert W function and the proximal operator of logsumexp have been implemented with C-mex files with OpenMP support for multicore acceleration. Demonstration scripts are available for testing.



(a) $\|\beta\|_\infty = 30$ (b) $\|\alpha\|_\infty = 0.01$ (c) SNR: 3.3dB (d) SNR: 15.1dB

(e) $\|\beta\|_\infty = 10^4$ (f) $\|\alpha\|_\infty = 0.01$ (g) SNR: 9.2dB (h) SNR: 21.2dB

(i) $\|\beta\|_\infty = 30$ (j) $\|\alpha\|_\infty = 0.15$ (k) SNR: 6.0dB (l) SNR: 9.7dB

(m) $\|\beta\|_\infty = 10^4$ (n) $\|\alpha\|_\infty = 0.15$ (o) SNR: 11.2dB (p) SNR: 21.1dB

(q) $\|\beta\|_\infty = 10^5$ (r) $\|\alpha\|_\infty = 0.1$ (s) SNR: 13.5dB (t) SNR: 44.3dB

Figure 10.7: Ability to recover the attenuation and density depending on the density and attenuation amplitude.

10.6 Comparison with existing work

In this section, we propose to compare our results with the approach suggested in [Sch+13]. In this paper, the authors concentrate on the case of confocal microscopy (i.e. two opposite views). They consider cone integrals to better model the light path. This is also possible with our approach, though we have not explored it yet. Following a maximum a posteriori principle, and using the notation of our work, they propose to minimize the following functional:

$$\min_{\alpha\in\mathbb{R}^n_+,\beta\in\mathbb{R}^n} \sum_{j=1}^{2} \left\| \frac{u_j - \beta\exp(-A_j\alpha)}{\sqrt{\gamma^2 u_j + \sigma^2}} \right\|_2^2 + R_\alpha(\alpha) + \mu\|\alpha\|_1. \tag{10.19}$$

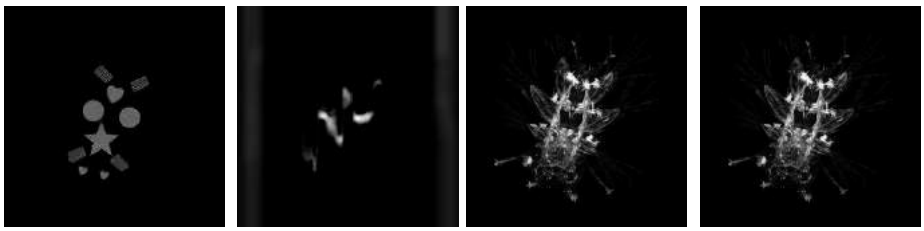
In this equation, $\lambda \geq 0$ and $\mu \geq 0$ are regularization parameters and $\gamma > 0$ and $\sigma > 0$ are parameters allowing to deal with a mixture of Poisson and Gaussian noise (in fact this energy is only an approximation of the MAP). In comparison with our work, the term $\mu\|\alpha\|_1$ is proposed to favor a black background.

The problem (10.19) is nonconvex and non differentiable. In order to minimize it, the authors propose to use smooth approximations of the functions $R_\alpha(\alpha)$ and $\|\alpha\|_1$ (to compute the derivative), and to use a limited-BFGS-B approach [Zhu+97], which allows to deal with bound constraints (the nonnegativity assumption). We reimplemented this algorithm with the line integrals used in the previous section and provide comparisons in Fig. 10.8. Since the problem is nonconvex the result depends on the initialization.

It takes around 30 seconds to run the limited-BFGS-B to solve the problem (10.19) on 256×256 images, which is on a par with the computing times required by our approach. The main differences with our approach are:

Parameters the energy (10.19) requires tuning 4 inter-dependent parameters: γ, σ, μ and the regularization parameter associated to the total variation. In comparison, our algorithm only requires tuning two *independent* parameters λ_α and λ_β, which is arguably much easier. That being said, we observed in our simulations that μ could be safely set to 0. A possible reason might be that the nonnegativity constraint already favors the background to be 0 [Boy+19b]. In addition, we expect that the parameters might be tuned once and for all for a given device, limiting the importance of this drawback. In our experiments, we discretized the parameter space finely, and kept the parameters leading to the highest SNR. This is of course feasible only when dealing with simulations.

Initialization we observed that the most important difference between the two approaches comes from the initialization. Problem (10.19) is nonconvex and the iterative descent algorithms may hence converge to different points depending on the starting point (α_0, β_0). In practice, we observed that this was a serious limitation of the approach in [Sch+13]. Figure 10.8 illustrates this point. We reproduce the experiment of Fig. 10.7, fourth line, corresponding to the favorable case $\|\beta\|_\infty = 10^4$ and $\|\alpha\|_\infty = 0.15$. On the left, we initialize the algorithm with the true attenuation and density maps, i.e. $\alpha_0 = \alpha$ and $\beta_0 = \beta$. The algorithm converges to a satisfactory estimate of the attenuation map. On the right, we initialized the algorithm with $\alpha_0 = \text{mean}(\alpha)$ and $\beta_0 = \frac{1}{2}(u_1 + u_2)$, which looks quite natural, since in



(a) Good initialization SNR:13.4dB. (b) Poor initialization SNR:-0.5dB. (c) Good initialization SNR: 141.2dB. (d) Poor initialization SNR:2.8dB.

Figure 10.8: Example of results obtained with the approach in [Sch+13]. Parameters selected to give the highest SNR. (a), (b): estimate of the attenuation map α. (c), (d): estimate of the density map β. The good initialization corresponds to the exact (unknown) maps that we would like to recover. The poor initialization corresponds to a realistic initialization that could be achieved in practice.

practice, nearly no information on the actual solution is available. As can be seen, the L-BFGS-B algorithm yields a very poor estimate of the true attenuation map. In comparison, the convex formulation proposed in this paper will converge to the same minimizer *whatever* the initialization. We believe that this is a distinctive trait and a real strength of the proposed approach.

Regularization there is no explicit regularization on the density map β in (10.19).

10.7 Conclusion & outlook

We proposed a robust and efficient approach based on a Bayesian estimator to recover attenuation and correct density from multiview measurements. This principle was already known in the field of lidar and solved with simple filtering approaches, while the algorithms proposed herein are based on a clear and versatile statistical framework. In confocal microscopy, Schmidt et al. [Sch+13] recently proposed a Bayesian formulation that shares many similarities with the proposed approach and applied it to real 3D data. The proposed approach differs in two regards: i) we consider a Poisson model for the noise, while [Sch+13] only uses an approximation of it and ii) we develop an algorithm that provably converges to the global minimizer of the cost function, while [Sch+13] is based on a nonlinear programming approach which leads to different results depending on the initialization. In practice, this distinctive feature seems to be of high importance, since good initial guesses are unavailable in the considered applications.

The approach seems promising for various devices such as lidar or some fluorescence microscopes. It is likely that its scope is much wider and we therefore provide a free Matlab toolbox on GitHub `https://github.com/pierre-weiss/MAEC`.

As a prospective, we plan to confront our algorithms with real data coming from lidar and microscopy. The total variation based algorithm to correct

attenuation defects is somewhat disappointing since it is unable to recover fine textures. A promising step would be to use more advanced nonlocal denoisers such as convolutional neural networks. To conclude, let us mention a serious limitation of the proposed approach: it is not so common to find a couple optical system-sample, where attenuation dominates scattering. We do not know at the present time how many applications can reasonably be modeled by equation (10.7). This question is central to precisely understand the strengths and limits of the proposed approach.

10.8 Appendices

10.8.1 Proof Proposition 10.3.1

Proof. The first item is obtained by direct inspection:

- for β fixed, $\alpha \mapsto \langle \exp(-A_j\alpha), \mathbb{1}\rangle$ is the composition of a linear operator with a convex function, hence it is convex. In addition $\alpha \mapsto \langle A_j\alpha, \mathbb{1}\rangle$ is a linear mapping.
- for α fixed, the first term in β is linear and $\beta \mapsto \langle -\log(\beta), \mathbb{1}\rangle$ is convex.

Let us now focus on the second and third points. The function G can be rewritten as a sum of functions:

$$G(\alpha, \beta) = \sum_{i=1}^{n} g_i\left((A_j\alpha[i])_{1\leq j\leq m}, \beta[i]\right),$$

where $g_i : \mathbb{R}^m \times \mathbb{R}_+ \to \mathbb{R}$ is defined as follows:

$$\begin{aligned}(x, y) \mapsto g_i(x, y) = \sum_{j=1}^{m} &\exp(-x[j])y \\ &+ u_j[i]\left(x[j] - \log(y)\right).\end{aligned}$$

To prove the convexity of G, it suffices to study the convexity of each function g_i. From now on, we skip the indices i to lighten the notation.

Let us analyze the eigenvalues of the Hessian H_g:

$$H_g(x, y) = \begin{pmatrix} \mathbf{diag}(()y\exp(-x)) & -\exp(-x) \\ -\exp(-x)^T & \sum_{j=1}^{m} u_j/y^2 \end{pmatrix}.$$

To study the positive semidefiniteness, let $(v, w) \in \mathbb{R}^{m+1}$, denote an arbitrary vector. We have:

$$\begin{aligned}&\left\langle \begin{pmatrix} v \\ w \end{pmatrix}, H_g(x, y)\begin{pmatrix} v \\ w \end{pmatrix}\right\rangle \\ &\qquad = \sum_{j=1}^{m} v[j]\exp(-x[j])\left(yv[j] - 2w\right) + w^2\frac{\sum_{j=1}^{m} u_j}{y^2}.\end{aligned}$$

In the case $y > \frac{\sum_{j=1}^m u_j}{\sum_{j=1}^m \exp(-x[j])}$ and $w \neq 0$, we get that:

$$\begin{aligned}
&\left\langle \begin{pmatrix} v \\ w \end{pmatrix}, H_g(x,y) \begin{pmatrix} v \\ w \end{pmatrix} \right\rangle \\
&< \sum_{j=1}^m v[j] \exp(-x[j]) \left(yv[j] - 2w\right) \\
&\quad + w^2 \frac{\sum_{j=1}^m \exp(-x[j])}{y} \\
&= \sum_{j=1}^m \exp(-x[j]) \left(yv[j]^2 - 2wv[j] + \frac{w^2}{y}\right) \\
&= \sum_{j=1}^m \exp(-x[j]) \left(v[j]\sqrt{y} - \frac{w}{\sqrt{y}}\right)^2
\end{aligned}$$

where the last expression is equal to 0 for the particular choice $v[j]y = w$, for all $1 \leq j \leq m$. This implies $\left\langle \begin{pmatrix} v \\ w \end{pmatrix}, H_g \begin{pmatrix} v \\ w \end{pmatrix} \right\rangle < 0$, which shows that function g is not convex in $\mathbb{R}^n \times \mathbb{R}^n_+$ and proves the second item.

The same argument with $y \leq \frac{\sum_{j=1}^m u_j}{\sum_{j=1}^m \exp(-x[j])}$ proves the third item. □

10.8.2 Proof Proposition 10.3.2

Proof. With this specific choice, it is easy to check that the optimality conditions of problem (10.10) with respect to variable β yield (10.14). By replacing this expression in (10.11), we obtain the optimization problem shown in equation (10.13).

Checking convexity of this problem can be done by simple inspection. The term $\langle u_j(A_j\alpha), \mathbb{1}\rangle$ is linear, hence convex. The term $\log\left(\sum_{j=1}^m \exp(-(A_j\alpha)[i])\right)$ is the composition of the convex logsumexp function with a linear operator, hence it is convex. □

10.8.3 Proximal operator of logsumexp in dimension 2

In this section, we propose a fast and accurate numerical algorithm based on Newton's method to solve the following problem:

$$\begin{aligned}
w &= \mathrm{prox}_{\gamma g_1}(z) \\
&= \underset{x \in \mathbb{R}^{2n}}{\mathrm{argmin}} \frac{1}{2}\|x - z\|_2^2 + \\
&\quad \gamma \sum_{i=1}^n \sum_{j=1}^2 u_j[i] \left[x_j[i] + \log\left(\sum_{j=1}^2 \exp(-x_j[i])\right)\right],
\end{aligned}$$

where $z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}$ and $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ are vectors in $\mathbb{R}^{2n}$. This problem may seem innocuous at first sight, but turns out to be quite a numerical challenge. The

first observation is that it can be decomposed as n *independent* problems of dimension 2 since:

$$w[i] = \underset{(x_1,x_2)\in\mathbb{R}^2}{\operatorname{argmin}} \frac{1}{2}(x_j - z_j[i])_2^2 + \gamma \sum_{j=1}^{2} u_j[i] \left[x_j + \log\left(\sum_{j=1}^{2} \exp(-x_j) \right) \right]. \tag{10.20}$$

To simplify the notation, we will skip the index i in what follows. The following proposition shows that our problem is equivalent to finding the proximal operator associated to the "logsumexp" function.

Proposition 10.8.1. *Define the logsumexp function* $\operatorname{lse}(x_1, x_2) = \log\left(\sum_{j=1}^{2} \exp(x_j)\right)$. *The solution of problem* (10.20) *coincides with the opposite of the proximal operator of* lse*:*

$$w[i] = -\underset{(x_1,x_2)\in\mathbb{R}^2}{\operatorname{argmin}} a\operatorname{lse}(x_1, x_2) + \frac{1}{2}((x_1 - y_1)^2 + (x_2 - y_2)^2) \tag{10.21}$$

$$= -\operatorname{prox}_{a\operatorname{lse}}(y_1, y_2), \tag{10.22}$$

where $a = \gamma(u_1 + u_2)$ *and* $y_j = \gamma u_j - z_j$.

Proof. The first order optimality conditions for problem (10.20) read

$$\begin{cases} \gamma u_1 - \frac{\gamma(u_1+u_2)\exp(-x_1)}{\exp(-x_1)+\exp(-x_2)} + x_1 - z_1 = 0 \\ \gamma u_2 - \frac{\gamma(u_1+u_2)\exp(-x_2)}{\exp(-x_1)+\exp(-x_2)} + x_2 - z_2 = 0. \end{cases} \tag{10.23}$$

By letting $a = \gamma(u_1 + u_2)$ and $y_j = \gamma u_j - z_j$, this equation becomes

$$\begin{cases} -\frac{a\exp(-x_1)}{\exp(-x_1)+\exp(-x_2)} + x_1 + y_1 = 0 \\ -\frac{a\exp(-x_2)}{\exp(-x_1)+\exp(-x_2)} + x_2 + y_2 = 0. \end{cases} \tag{10.24}$$

It now suffices to make the change of variable $x_i' = -x_i$ to retrieve the optimality conditions of problem (10.22)

$$\begin{cases} \frac{a\exp(x_1')}{\exp(x_1')+\exp(x_2')} + x_1' - y_1 = 0 \\ \frac{a\exp(x_2')}{\exp(x_1')+\exp(x_2')} + x_2' - y_2 = 0. \end{cases} \tag{10.25}$$

□

Remark 10.8.1. *To the best of our knowledge, this is the first attempt to find a fast algorithm to evaluate the prox of logsumexp. This function is important in many regards. In particular, it is a* C^∞ *approximation of the maximum value of a vector. In addition, its Fenchel conjugate coincides with the Shannon entropy restricted to the unit simplex. We refer to [Hir06, §3.2] for some details. The algorithm that follows has potential applications outside the scope of this work.*

We now design a fast and accurate minimization algorithm for problem (10.22) or equivalently, a root finding algorithm for problem (10.25). This algorithm

differs depending on whether $y_1 \geq y_2$ or $y_2 \geq y_1$. We focus on the case $y_1 \geq y_2$. The case $y_2 \geq y_1$ can be handled by symmetry.

Let $\lambda = \frac{\exp(x'_1)}{\exp(x'_1)+\exp(x'_2)}$ and notice that

$$\frac{\exp(x'_2)}{\exp(x'_1) + \exp(x'_2)} = 1 - \lambda.$$

Therefore (10.25) becomes:

$$\begin{cases} x'_1 = y_1 - a\lambda \\ x'_2 = y_2 - a(1-\lambda). \end{cases} \tag{10.26}$$

Hence

$$\frac{1-\lambda}{\lambda} = \exp(x'_2 - x'_1) = \exp(y_2 - y_1 - a)\exp(2a\lambda). \tag{10.27}$$

Taking the logarithm on each side yields [1]:

$$\log(1-\lambda) - \log(\lambda) = y_2 - y_1 - a + 2a\lambda. \tag{10.28}$$

We are now facing the problem of finding the root λ^* of the following function:

$$f(\lambda) = y_2 - y_1 - a + 2a\lambda - \log(1-\lambda) + \log(\lambda). \tag{10.29}$$

There are two important advantages for this approach compared to the direct resolution of (10.25). First, we have to solve a 1D problem instead of a 2D problem. More importantly, we directly constrain x' to be of form $x' = y - a\delta$, where δ lives on the 2D simplex.

Let us collect a few properties of function f. First, we have:

$$f'(\lambda) = 2a + \frac{1}{1-\lambda} + \frac{1}{\lambda} > 0, \forall \lambda \in (0,1). \tag{10.30}$$

Therefore, f is increasing on $(0,1)$. To use convergence results of Newton's algorithm, we need to compute f'' as well:

$$f''(\lambda) = -\frac{1}{\lambda^2} + \frac{1}{(1-\lambda)^2}. \tag{10.31}$$

Proposition 10.8.2. *If $y_1 \geq y_2$, then $x'_1 \geq x'_2$ and*

$$\max\left(\frac{1}{2}, \frac{1}{1+\exp(y_2 - y_1 + a)}\right) \leq \lambda^* \leq \frac{1}{1+\exp(y_2 - y_1)}. \tag{10.32}$$

Proof. The first statement can be proven by contradiction. Assume that $x'_2 > x'_1$, then equation (10.25) indicates that $y_2 > y_1$.

For the second statement, it suffices to evaluate f at the extremities of the interval since $f' > 0$. We get $f(1/2) = y_2 - y_1 \leq 0$ and $f\left(\frac{1}{1+\exp(y_2-y_1)}\right) = -a + \frac{2a}{1+\exp(y_2-y_1)} \geq 0$. □

[1] Applying the logarithm is important for numerical purposes. When $y_2 - y_1 - a$ is very small, the exponential cannot be computed accurately in double precision.

Proposition 10.8.3. *Set* $\lambda_0 = \frac{1}{1+\exp(y_2-y_1)}$. *Then, the following Newton's method*

$$\lambda_{k+1} = \lambda_k - \frac{f(\lambda_k)}{f'(\lambda_k)} \tag{10.33}$$

converges to the root λ^* *of* f*, with a locally quadratic rate.*

Proof. First notice that $f''(\lambda) \geq 0$ on the interval $[1/2, 1)$. Hence f'' is also positive on $I = [\lambda^*, \lambda_0]$. This ensures that

$$\lambda_0 \geq \lambda_1 \geq \ldots \geq \lambda^*. \tag{10.34}$$

We prove this assertion by recurrence. Notice that $\lambda_0 \geq \lambda^*$ by Proposition 10.8.2. Now, assume that $\lambda_k \geq \lambda^*$, then

$$f(\lambda_k) = f(\lambda^*) + \int_{\lambda^*}^{\lambda_k} f'(t)\, dt \leq f'(\lambda_k)(\lambda_k - \lambda^*). \tag{10.35}$$

Hence, $\lambda_k - \lambda^* \geq \frac{f(\lambda_k)}{f'(\lambda_k)}$ and $\lambda_{k+1} \geq \lambda^*$. In addition $\frac{f(\lambda_k)}{f'(\lambda_k)} \geq 0$ on I, so that $\lambda_{k+1} \geq \lambda_k$.

The sequence $(\lambda_k)_{k\in\mathbb{N}}$ is monotonically decreasing and bounded below, therefore it converges to some value $\lambda' \geq \lambda^*$. Necessarily $\lambda' = \lambda^*$, since for $\lambda' > \lambda^*$, $\frac{f(\lambda')}{f'(\lambda')} > 0$.

To prove the locally quadratic convergence rate, we just invoke the celebrated Newton-Kantorovich's theorem [Pol07; Ort68], that ensures local quadratic convergence if f'' is bounded in a neighborhood of the minimizer. □

Finally, let us mention that computing λ_0 on a computer is a tricky due to underflow problems: in double precision the command $1 + \exp(y_2 - y_1)$ will return 1 for $y_2 - y_1 < -37 \simeq \log(10^{-16})$. This may cause the algorithm to fail since f and its derivatives are undefined at $\lambda = 1$. In practice we therefore set $\lambda_0 = 1/(1 + \exp(y_2 - y_1)) - 10^{-16}$. Similarly, by bound (10.32), we get $\lambda^* = 1$ up to machine precision whenever $y_2 - y_2 - a < \log(10^{-16})$. Algorithm 14 summarizes all the ideas described in this paragraph.

An attentive reader may have remarked that the convergence of Newton's algorithm depends only on the difference $y(1) - y(2)$ and a. A shift of $y(1)$ and $y(2)$ by the same value does not change Newton's iteration. In Fig. 10.9, we show that the algorithm behaves very well for a wide range of parameters. For $y(1) - y(2)$ and a varying in the interval $[2^{-10}, 2^{20}]$, the algorithm never requires more than 18 iterations to reach machine precision and needs 2.8 iterations in average.

Algorithm 14 An algorithm to compute $\text{prox}_{a\text{lse}}(y_1, y_2)$ with machine precision

Input: $(y_1, y_2) \in \mathbb{R}^2, a \in \mathbb{R}_+$.
Output: $(x_1, x_2) = \text{prox}_{a\text{lse}(y_1,y_2)}$.
Set $\epsilon = 10^{-16}$.
if $y_1 \geq y_2$ **then**
  if $y_2 - y_1 + a < \log(\epsilon)$ **then**
    Set $\lambda = 1$.
  else
    Set $\lambda = \frac{1}{1+\exp(y_2-y_1)} - \epsilon$.
    Define $d(\lambda) = \frac{y_2-y_1-a+2a\lambda+\log(\lambda/(1-\lambda))}{2a+\frac{1}{\lambda(1-\lambda)}}$.
    while $|d(\lambda)| > \epsilon$ **do**
      Set $\lambda = \lambda - d(\lambda)$.
    end while
  end if
  Set $[x_1, x_2] = [y(1) - a\lambda, y(2) - a(1-\lambda)]$.
else if $y_1 < y_2$ **then**
  if $y_1 - y_2 + a < \log(\epsilon)$ **then**
    Set $\lambda = 1$.
  else
    Set $\lambda = \frac{1}{1+\exp(y_1-y_2)} - \epsilon$.
    Define $d = \frac{y_1-y_2-a+2a\lambda+\log(\lambda/(1-\lambda))}{2a+\frac{1}{\lambda(1-\lambda)}}$.
    while $|d| > \epsilon$ **do**
      Set $\lambda = \lambda - d(\lambda)$.
    end while
  end if
  Set $[x_1, x_2] = [y(1) - a(1-\lambda), y(2) - a\lambda]$.
end if

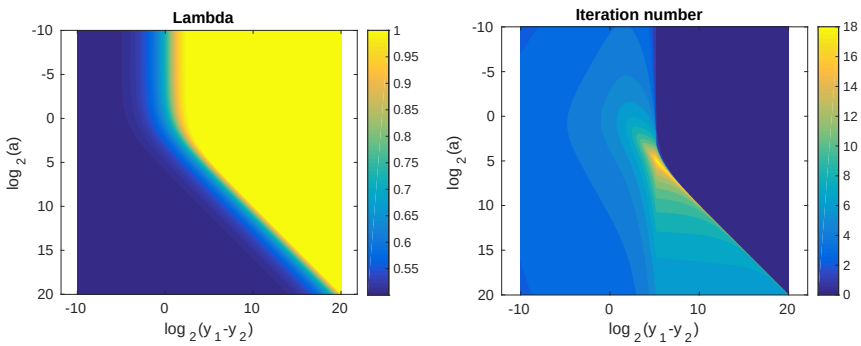


Figure 10.9: Performance evaluation for Newton's algorithm. Left: λ^* depending on a and $y_1 - y_2$. Right: number of iterations of Newton's method to reach machine precision.

Cinquième partie

Closing remarks

Chapitre 11

Discussion

11.1 Ouverture et perspectives

Les travaux présentés dans cette thèse s'articulent autour de la modélisation fine des systèmes optiques pour aider à la résolution de problèmes inverses aveugles. Un problème récurrent que nous avons rencontré dans la validation de nos approches était la construction d'exemples de simulations proches de ceux rencontrés en pratique. Il existe de nombreux travaux dans cette direction pour la construction des flous invariants spatialement [Kir+13 ; Sag+19], mais très peu pour des opérateurs variants spatialement. Au cours de cette thèse, nous avons collecté un grand nombre d'images de microscopie contenant des microbilles. Cela nous a permis d'apprendre des sous-espaces d'opérateurs variants spatialement sur différents systèmes optiques. Nous pensons que cette base de données, associée aux travaux présentés, peut avoir beaucoup de la valeur pour les groupes travaillant sur des systèmes similaires.

Toujours dans l'idée de fournir des outils pratiques à la communauté, les méthodes introduites dans ce manuscrit sont toutes disponibles sous *Matlab*, ou le seront très prochainement. Les algorithmes du Chapitre 3 sont actuellement portés sous le logiciel *Fiji*. Les autres méthodes bénéficieraient beaucoup d'un travail similaire. En effet, ces algorithmes reposent sur des concepts non-triviaux qui nécessitent une implémentation fine. Les scripts *Matlab* sont actuellement utilisables seulement par des personnes pratiquant le traitement d'images. La mise en place d'un module pour le logiciel *Fiji* par exemple, permettrait de rendre beaucoup plus accessible nos méthodes.

L'estimation fine des opérateurs intervenants en microscopie est un sujet qui nous tient à cœur comme en témoigne ce manuscrit. Mais, est-ce qu'une estimation fine des opérateurs de flous est réellement indispensable ? La question est légitime. Prenons pour exemple le cas d'une acquisition de microscopie à localisation de molécules uniques (SMLM) en 2D. Nous avons montré dans le Chapitre 6 qu'un simple maximum de corrélation avec le vrai opérateur de flou, permet d'estimer les positions des molécules. Poussons l'exemple un peu plus loin en prenant l'opérateur de flou comme une convolution avec un noyau Gaussien isotrope de variance σ. Une perturbation sur la variance σ de ce noyau

ne produira aucune erreur dans la localisation des molécules par un maximum de corrélation. Au contraire, si maintenant on utilise cet opérateur perturbé pour déflouter une image arbitraire, on risque de voir apparaître des artefacts (e.g. anneaux). Il serait alors intéressant de pouvoir caractériser l'effet que produit l'incertitude sur un opérateur dans la résolution des problèmes inverses présents en imageries.

Comme ce manuscrit le laisse suggérer, ces travaux de thèse ont été l'occasion de travailler sur beaucoup de projets plus ou moins différents. Il reste cependant encore un nombre conséquent de pistes que l'on souhaiterait explorer. L'une d'entre elles concerne les algorithmes de SMLM (*Single Molecule Localization Microscopy*). Nous avons proposé dans le Chapitre 6 une méthode pour retrouver précisément un signal composé de points sources. Cependant, les méthodes proposées sont limitées au cas de sources non-denses. Les méthodes récentes de super-résolution hors grille montrent que ce problème peut être résolu efficacement [Den+19]. Ces approches requièrent la connaissance de l'opérateur d'acquisition. Les sous-espaces d'opérateurs introduits dans cette thèse peuvent permettre de résoudre ce problème de façon aveugle. Cette piste est actuellement explorée en collaboration avec Alban Gossard et Pierre Weiss.

11.2 Supports financiers

La charge salariale de cette thèse est entièrement supportée par la Fondation pour la Recherche Médicale (FRM grant number ECO20170637521 to V.D). Cette thèse a également été soutenue par l'obtention du prix Kerialis (merci maman !). Les besoins de fonctionnement sont supportés par : l'ANR-17-CE23-0013-01 OMS (*Optimization on Measure Spaces*), l'ANR-3IA Artificial and Natural Intelligence Toulouse Institute, le GDR ISIS pour le projet FiMOSuReMi, le plan CANCER pour le projet MIMMOSA, et le défi IMAG'IN.

Activités scientifiques

Présentations

J'ai eu l'occasion de particper à différentes conférences durant ces dernières années et d'y présenter mes travaux.

Oraux

- OSA, Munich, juin 2019, Munich : Blind-Deblurring : Learning Based Approach
- Demystifying machine learning for microscopists, Toulouse, mars 2019 :
- ITWIST' 2018, novembre 2018 : A scalable estimator of sets of integral operators.
- OSA topical meetings Maths & Imaging, Orlando, juin 2018 : Learning and Exploiting Physics of Degradations.
- GdR MOA et MIA, Bordeaux, 2017 : Multiview Attenuation Estimation and Correction.

Posters

- QBI, January 2020, Oxford : High density 3D localization of fluorescent molecules.
- SPARS, juillet 2019, Toulouse : A scalable estimator of sets of integral operators.
- ISBI, avril 2019, Venise : Segmentation : a data driven approach through neural network.
- QBI, janvier 2019, Rennes : Calibrating a microscope by learning its diversity.
- Spars, Lisbonne, juin 2017 : Multiview Attenuation Computation and Correction
- Workshop Optimization and Learning, Toulouse, 2017.

Enseignement

J'ai effectué un contrat de doctorant en charge d'enseignement (DCE) à l'INSA de Toulouse durant mes trois années de thèse, de 2017 à 2020. Je suis intervenu dans les enseignements suivants :

- **Projet modélisation**, 3ème année MIC (équivalent L3), 35h.
  J'ai proposé des sujets sur les algorithmes de bandits manchots, la génération de variable aléatoire en machine et l'approximation en traitement d'images.

- **Projet en autonomie**, 2ème année MIC (équivalent L2), 15h.
  J'ai proposé un projet autour de la restauration d'image avec réseaux de neurones.

- **Travaux pratiques d'analyse numérique**, 2ème année IC (équivalent L2), 15h.
  J'ai encadré des TPs sur la résolution numérique d'équations aux dérivées partiels ordinaires.

- **Recherche bibliographie**, 4ème année GMM (équivalent M1), 2h30.
  Je suis intervenu pour présenter certains standards bibliographiques et comment réaliser une bibliographie avec *Latex*.

Sixième partie

Bibliographie

Chapitre 12

Mes références

12.0.1 Articles de journaux (international)

[Deb+20a] Valentin Debarnot, Paul Escande, Thomas Mangeat et Pierre Weiss. "Learning low-dimensional models of microscopes". In : *IEEE Transactions on Computational Imaging* (2020).

[DEW19] Valentin Debarnot, Paul Escande et Pierre Weiss. "A scalable estimator of sets of integral operators". In : *Inverse Problems* 35.10 (2019), p. 105011.

[Deb+18b] Valentin Debarnot, Jérôme Fehrenbach, Frédéric de Gournay et Léo Martire. "The Case of Neumann, Robin, and Periodic Lateral Conditions for the Semi-infinite Generalized Graetz Problem and Applications". In : *SIAM Journal on Applied Mathematics* 78.4 (2018), p. 2227-2251.

[DKW19] Valentin Debarnot, Jonas Kahn et Pierre Weiss. "Multiview Attenuation Estimation and Correction". In : *Journal of Mathematical Imaging and Vision* 61.6 (2019), p. 780-797.

12.0.2 Pré-publications

[Deb+20b] Valentin Debarnot, Thomas Mangeat, Daniel Sage, Emmanuel Soubies et Pierre Weiss. "PSF-Estimator". 2020.

[DW20a] Valentin Debarnot et Pierre Weiss. "Blind deblurring and super-resolution with isolated spikes". 2020.

[DW20b] Valentin Debarnot et Pierre Weiss. "DeepBlur: blind identification of space variant PSF". 2020.

12.0.3 Articles de conference (international)

[Deb+18a] Valentin Debarnot, Paul Escande, Thomas Mangeat et Pierre Weiss. "A scalable estimator of sets of integral operators". In : *Proceedings of iTWIST'18, Paper-ID: 2, Marseille, France.* 2018.

[Deb+19a] Valentin Debarnot, Paul Escande, Thomas Mangeat et Pierre Weiss. "A scalable estimator of space varying blurs: Application in super-resolution". In : *SPARS 19 Toulouse*. 2019.

[Deb+19b] Valentin Debarnot, Paul Escande, Thomas Mangeat et Pierre Weiss. "Blind-deblurring: learning based approach". In : *Mathematics in Imaging*. Optical Society of America. 2019, p. MM1D-3.

[DKW17] Valentin Debarnot, Jonas Kahn et Pierre Weiss. "Multiview Attenuation Computation and Correction". In : *SPARS 17 Lisbon*. 2017.

[DL19] Valentin Debarnot et Léo Lebrat. "Segmentation: a data driven approach though neural network". In : *IEEE 16th International Symposium on Biomedical Imaging (ISBI)*. 2019.

[DMW19] Valentin Debarnot, Thomas Mangeat et Pierre Weiss. "Calibrating a microscope by learning its diversity". In : *Quantitative BioImaging Society*. 2019.

[DMW20] Valentin Debarnot, Thomas Mangeat et Pierre Weiss. "Learning low-dimensional models of microscopes". In : *Quantitative BioImaging Society*. 2020.

[Esc+18] Paul Escande, Valentin Debarnot, Mauro Maggioni, Thomas Mangeat et Pierre Weiss. "Learning and Exploiting Physics of Degradations". In : *Mathematics in Imaging*. Optical Society of America. 2018, MTu2D-4.

[LDM20] Léo Lebrat, Valentin Debarnot et Thomas Mangeat. "Biolapse Toolbox". In : *Quantitative BioImaging Society*. 2020.

Chapitre 13

Autres références

[Aba+16] Martín Abadi et al. "Tensorflow: Large-scale machine learning on heterogeneous distributed systems". In : *arXiv preprint arXiv:1603.04467* (2016).

[AMS09] P-A Absil, Robert Mahony et Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

[AÖ17] Jonas Adler et Ozan Öktem. "Solving ill-posed inverse problems using iterative deep neural networks". In : *Inverse Problems* 33.12 (2017), p. 124007.

[Aga+89] David A Agard, Yasushi Hiraoka, Peter Shaw et John W Sedat. "Fluorescence microscopy in three dimensions". In : *Methods in cell biology*. T. 30. Elsevier, 1989, p. 353-377.

[AD18] Ali Ahmed et Laurent Demanet. "Leveraging diversity and sparsity in blind deconvolution". In : *IEEE Transactions on Information Theory* 64.6 (2018), p. 3975-4000.

[ARR13] Ali Ahmed, Benjamin Recht et Justin Romberg. "Blind deconvolution using convex programming". In : *IEEE Transactions on Information Theory* 60.3 (2013), p. 1711-1732.

[ARR14] Ali Ahmed, Benjamin Recht et Justin Romberg. "Blind deconvolution using convex programming". In : *IEEE Transactions on Information Theory* 60.3 (2014), p. 1711-1732.

[Air35] George Biddell Airy. "On the diffraction of an object-glass with circular aperture". In : *Transactions of the Cambridge Philosophical Society* 5 (1835), p. 283.

[APS19] Raied Aljadaany, Dipan K Pal et Marios Savvides. "Douglas-rachford networks: Learning both the image prior and data fidelity terms for blind image deconvolution". In : *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, p. 10235-10244.

[AK00] Jay Anderson et Ivan R King. "Toward High-Precision Astrometry with WFPC2. I. Deriving an Accurate Point-Spread Function". In : *Publications of the Astronomical Society of the Pacific* 112.776 (2000), p. 1360.

[ARW90] Albert Ansmann, Maren Riebesell et Claus Weitkamp. "Measurement of atmospheric aerosol extinction profiles with a Raman lidar". In : *Optics letters* 15.13 (1990), p. 746-748.

[Ari+10] Muthuvel Arigovindan, Joshua Shaevitz, John McGowan, John W. Sedat et David A. Agard. "A Parallel Product-Convolution approach for representing depth varying Point Spread Functions in 3D widefield microscopy based on principal component analysis". In : *Opt. Express* 18.7 (mar. 2010), p. 6461-6476. doi : 10.1364/OE.18.006461.

[Ari+18] Andrey Aristov, Benoit Lelandais, Elena Rensen et Christophe Zimmer. "ZOLA-3D allows flexible 3D localization microscopy over an adjustable axial range". In : *Nature communications* 9.1 (2018), p. 1-8.

[AD88] GR Ayers et J Christopher Dainty. "Iterative blind deconvolution method and its applications". In : *Optics letters* 13.7 (1988), p. 547-549.

[BZ17] Hazen P Babcock et Xiaowei Zhuang. "Analyzing single molecule localization microscopy data using cubic splines". In : *Scientific reports* 7.1 (2017), p. 1-9.

[BSZ12] Hazen Babcock, Yaron M Sigal et Xiaowei Zhuang. "A high-density 3D localization algorithm for stochastic optical reconstruction microscopy". In : *Optical Nanoscopy* 1.1 (2012), p. 1-10.

[Bac+14] Mikael P Backlund, Matthew D Lew, Adam S Backer, Steffen J Sahl et WE Moerner. "The role of molecular dipole orientation in single-molecule fluorescence microscopy and implications for super-resolution imaging". In : *ChemPhysChem* 15.4 (2014), p. 587-599.

[Bea+14] Jordan R Beach, Lin Shao, Kirsten Remmert, Dong Li, Eric Betzig et John A Hammer III. "Nonmuscle myosin II isoforms coassemble in living cells". In : *Current Biology* 24.10 (2014), p. 1160-1166.

[BG97] Rick Beatson et Leslie Greengard. "A short course on fast multipole methods". In : *Wavelets, multilevel methods and elliptic PDEs* 1 (1997), p. 1-37.

[Beb08] Mario Bebendorf. *Hierarchical matrices*. Springer, 2008.

[BB18] Robert Beinert et Kristian Bredies. "Non-convex regularization of bilinear and quadratic inverse problems by tensorial lifting". In : *Inverse Problems* 35.1 (2018), p. 015002.

[BB19] Robert Beinert et Kristian Bredies. "Tensor-Free Proximal Methods for Lifted Bilinear/Quadratic Inverse Problems with Applications to Phase Retrieval". In : *arXiv preprint arXiv:1907.04875* (2019).

[BR19] Chinmay Belthangady et Loic A Royer. “Applications, promises, and pitfalls of deep learning for fluorescence image reconstruction”. In : *Nature methods* (2019), p. 1-11.

[Ben+13] Yoshua Bengio, Li Yao, Guillaume Alain et Pascal Vincent. “Generalized denoising auto-encoders as generative models”. In : *Advances in neural information processing systems*. 2013, p. 899-907.

[Ber11] E Bertin. “Automated morphometry with SExtractor and PSFEx”. In : *Astronomical Data Analysis Software and Systems XX*. T. 442. 2011, p. 435.

[Ber13] Emmanuel Bertin. “PSFEx: Point Spread Function Extractor”. In : *ascl* (2013), ascl-1301.

[BA96] Emmanuel Bertin et Stephane Arnouts. “SExtractor: Software for source extraction”. In : *Astronomy and Astrophysics Supplement Series* 117.2 (1996), p. 393-404.

[Ber+16] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi et Philip HS Torr. “Fully-convolutional siamese networks for object tracking”. In : *European conference on computer vision*. Springer. 2016, p. 850-865.

[Bet+06] Eric Betzig et al. “Imaging intracellular fluorescent proteins at nanometer resolution”. In : *Science* 313.5793 (2006), p. 1642-1645.

[BCR91] Gregory Beylkin, Ronald Coifman et Vladimir Rokhlin. “Fast wavelet transforms and numerical algorithms I”. In : *Communications on pure and applied mathematics* 44.2 (1991), p. 141-183.

[BEW17] Jérémie Bigot, Paul Escande et Pierre Weiss. “Estimation of linear operators from scattered impulse responses”. In : *Applied and Computational Harmonic Analysis* (2017).

[BEW19a] Jérémie Bigot, Paul Escande et Pierre Weiss. “Estimation of linear operators from scattered impulse responses”. In : *Applied and Computational Harmonic Analysis* 47.3 (2019), p. 730-758. issn : 1063-5203.

[BEW19b] Jérémie Bigot, Paul Escande et Pierre Weiss. “Estimation of linear operators from scattered impulse responses”. In : *Applied and Computational Harmonic Analysis* 47.3 (2019), p. 730-758.

[BST14] Jérôme Bolte, Shoham Sabach et Marc Teboulle. “Proximal alternating linearized minimization or nonconvex and nonsmooth problems”. In : *Mathematical Programming* 146.1-2 (2014), p. 459-494.

[Bon+09] Pierre Bon, Guillaume Maucort, Benoit Wattellier et Serge Monneret. “Quadriwave lateral shearing interferometry for quantitative phase microscopy of living cells”. In : *Optics express* 17.15 (2009), p. 13080-13094.

[BW13] Max Born et Emil Wolf. *Principles of optics: electromagnetic theory of propagation, interference and diffraction of light*. Elsevier, 2013.

[BLM13] Stéphane Boucheron, Gábor Lugosi et Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[Bou+17] Anaïs Bouissou et al. "Podosome force generation machinery: a local balance between protrusion at the core and traction at the ring". In : *ACS nano* 11.4 (2017), p. 4028-4040.

[Bou+15] Nicolas Bourg, Céline Mayet, Guillaume Dupuis, Thomas Barroca, Pierre Bon, Sandrine Lécart, Emmanuel Fort et Sandrine Lévêque-Fort. "Direct optical nanoscopy with axially localized detection". In : *Nature Photonics* 9.9 (2015), p. 587-593.

[Boy+19a] Claire Boyer, Antonin Chambolle, Yohann De Castro, Vincent Duval, Frédéric De Gournay et Pierre Weiss. "On representer theorems and convex regularization". In : *SIAM Journal on Optimization* 29.2 (2019), p. 1260-1281.

[Boy+19b] Claire Boyer, Antonin Chambolle, Yohann De Castro, Vincent Duval, Frédéric De Gournay et Pierre Weiss. "On representer theorems and convex regularization". In : *SIAM Journal on Optimization* 29.2 (2019), p. 1260-1281.

[BP13] Kristian Bredies et Hanna Katriina Pikkarainen. "Inverse problems in spaces of measures". In : *ESAIM: Control, Optimisation and Calculus of Variations* 19.1 (2013), p. 190-218.

[BD97] Rasmus Bro et Sijmen De Jong. "A fast non-negativity-constrained least squares algorithm". In : *Journal of Chemometrics: A Journal of the Chemometrics Society* 11.5 (1997), p. 393-401.

[Bru17] Victor-Emmanuel Brunel. "Methods for estimation of convex sets". In : *https://arxiv.org/abs/1709.03137* (2017).

[Buh03] Martin D Buhmann. *Radial basis functions: theory and implementations*. T. 12. Cambridge university press, 2003.

[CJ19] Valerio Cambareri et Laurent Jacques. "Through the haze: a non-convex approach to blind gain calibration for linear random sensing models". In : *Information and Inference: A Journal of the IMA* 8.2 (2019), p. 205-271.

[Can+03] A Can, O Al-Kofahi, S Lasek, DH Szarowski, JN Turner et B Roysam. "Attenuation correction in confocal laser microscopes: a novel two-view approach". In : *Journal of microscopy* 211.1 (2003), p. 67-79.

[CF14] Emmanuel J Candès et Carlos Fernandez-Granda. "Towards a mathematical theory of super-resolution". In : *Communications on pure and applied Mathematics* 67.6 (2014), p. 906-956.

[CSV13] Emmanuel J Candes, Thomas Strohmer et Vladislav Voroninski. "Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming". In : *Communications on Pure and Applied Mathematics* 66.8 (2013), p. 1241-1274.

[Can76] Michael Cannon. "Blind deconvolution of spatially invariant image blurs with phase". In : *IEEE Transactions on Acoustics, Speech, and Signal Processing* 24.1 (1976), p. 58-63.

[Car+06] Anne E Carpenter et al. "CellProfiler: image analysis software for identifying and quantifying cell phenotypes". In : *Genome biology* 7.10 (2006), R100.

[Cha16] Ayan Chakrabarti. "A neural approach to blind motion deblurring". In : *European conference on computer vision*. Springer. 2016, p. 221-235.

[CZF10] Ayan Chakrabarti, Todd Zickler et William T Freeman. "Analyzing spatially-varying blur". In : *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE. 2010, p. 2512-2519.

[Cha+94] Martin Chalfie, Yuan Tu, Ghia Euskirchen, William W Ward et Douglas C Prasher. "Green fluorescent protein as a marker for gene expression". In : *Science* 263.5148 (1994), p. 802-805.

[Cha+16] Joe Chalfoun, Michael Majurski, Alden Dima, Michael Halter, Kiran Bhadriraju et Mary Brady. "Lineage mapper: A versatile cell and particle tracker". In : *Scientific reports* 6 (2016), p. 36984.

[Cha04] Antonin Chambolle. "An algorithm for total variation minimization and applications". In : *Journal of Mathematical imaging and vision* 20.1-2 (2004), p. 89-97.

[CP16] Antonin Chambolle et Thomas Pock. "An introduction to continuous optimization for imaging". In : *Acta Numerica* 25 (2016), p. 161-319.

[CGI96] Frederic Champagnat, Yves Goussard et Jerome Idier. "Unsupervised deconvolution of sparse spike trains using stochastic approximation". In : *IEEE Transactions on Signal Processing* 44.12 (1996), p. 2988-2998.

[CW98] Tony F Chan et Chiu-Kwong Wong. "Total variation blind deconvolution". In : *IEEE transactions on Image Processing* 7.3 (1998), p. 370-375.

[Cha+12] Chihway Chang et al. "Atmospheric point spread function interpolation for weak lensing in short exposure imaging data". In : *Monthly Notices of the Royal Astronomical Society* 427.3 (2012), p. 2572-2587.

[Che+20a] Jinchi Chen, Weiguo Gao, Sihan Mao et Ke Wei. "Vectorized Hankel Lift: A Convex Approach for Blind Super-Resolution of Point Sources". In : *arXiv preprint arXiv:2008.05092* (2020).

[Che+20b] Yuxin Chen, Jianqing Fan, Bingyan Wang et Yuling Yan. "Convex and Nonconvex Optimization Are Both Minimax-Optimal for Noisy Blind Deconvolution". In : *arXiv preprint arXiv:2008.01724* (2020).

[CIN99] Margaret Cheney, David Isaacson et Jonathan C Newell. "Electrical impedance tomography". In : *SIAM review* 41.1 (1999), p. 85-101.

[Che+13] Chung-Chuan Cheng, Tsu-Yi Hsieh, Jin-Shiuh Taur et Yung-Fu Chen. "An automatic segmentation and classification framework for anti-nuclear antibody images". In : *Biomedical engineering online* 12.1 (2013), S5.

[Chh+15] Raghav K Chhetri, Fernando Amat, Yinan Wan, Burkhard Höckendorf, William C Lemon et Philipp J Keller. "Whole-animal functional and developmental imaging with isotropic spatial resolution". In : *Nature methods* (2015).

[Chi16] Yuejie Chi. "Guaranteed blind sparse spikes deconvolution via lifting and convex optimization". In : *IEEE Journal of Selected Topics in Signal Processing* 10.4 (2016), p. 782-794.

[CCS14] Abdellah Chkifa, Albert Cohen et Christoph Schwab. "High-dimensional adaptive sparse polynomial interpolation and applications to parametric PDEs". In : *Foundations of Computational Mathematics* 14.4 (2014), p. 601-633.

[CL09] Sunghyun Cho et Seungyong Lee. "Fast motion deblurring". In : *ACM SIGGRAPH Asia 2009 papers*. 2009, p. 1-8.

[Cic+16] Ozgun Cicek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox et Olaf Ronneberger. "3D U-Net: learning dense volumetric segmentation from sparse annotation". In : *International conference on medical image computing and computer-assisted intervention*. Springer. 2016, p. 424-432.

[Cic+07] Andrzej Cichocki, Rafal Zdunek, Seungjin Choi, Robert Plemmons et Shun-Ichi Amari. "Non-negative tensor factorization using alpha and beta divergences". In : *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*. T. 3. IEEE. 2007, p. III-1393.

[CNL11] Adam Coates, Andrew Ng et Honglak Lee. "An analysis of single-layer networks in unsupervised feature learning". In : *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. 2011, p. 215-223.

[CJB11] Richard W Cole, Tushare Jinadasa et Claire M Brown. "Measuring and interpreting point spread functions to determine confocal microscope resolution and ensure quality control". In : *Nature protocols* 6.12 (2011), p. 1929-1941.

[Col+13] Guylaine Collewet, Jérôme Bugeon, Jérôme Idier, Stéphane Quellec, Benjamin Quittet, Mireille Cambert et Pierrick Haffray. "Rapid quantification of muscle fat content and subcutaneous adipose tissue in fish using MRI". In : *Food chemistry* 138.2-3 (2013), p. 2008-2015.

[CP11] Patrick L Combettes et Jean-Christophe Pesquet. "Proximal splitting methods in signal processing". In : *Fixed-point algorithms for inverse problems in science and engineering*. Springer, 2011, p. 185-212.

[Com14] Pierre Comon. "Tensors: a brief introduction". In : *IEEE Signal Processing Magazine* 31.3 (2014), p. 44-53.

[Cou+13] Florent Couzinie-Devy, Jian Sun, Karteek Alahari et Jean Ponce. "Learning to estimate and remove non-uniform image blur". In : *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2013, p. 1075-1082.

[CC74] Christoph Cremer et Thomas Cremer. “Considerations on a laser-scanning-microscope with high resolution and depth of field”. In : *Microscopica acta* (1974), p. 31-44.

[CCH06] William R Crum, Oscar Camara et Derek LG Hill. “Generalized overlap measures for evaluation and validation in medical image analysis”. In : *IEEE transactions on medical imaging* 25.11 (2006), p. 1451-1461.

[CF10] Juan Cuesta et Pierre H Flamant. “Lidar beams in opposite directions for quality assessment of Cloud-Aerosol Lidar with Orthogonal Polarization spaceborne measurements”. In : *Applied optics* 49.12 (2010), p. 2232-2243.

[DC20] Maxime Ferreira Da Costa et Yuejie Chi. “On the stable resolution limit of total variation regularization for spike deconvolution”. In : *IEEE Transactions on Information Theory* (2020).

[DW08] Shengyang Dai et Ying Wu. “Motion from blur”. In : *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2008, p. 1-8.

[Dau+19] Lise Dauban, Alain Kamgoué, Renjie Wang, Isabelle Léger-Silvestre, Frédéric Beckouet, Sylvain Cantaloube et Olivier Gadal. “Quantification of the dynamic behaviour of ribosomal DNA genes and nucleolus during yeast Saccharomyces cerevisiae cell cycle”. In : *Journal of structural biology* 208.2 (2019), p. 152-164.

[De +12] Fabrice De Chaumont et al. “Icy: an open bioimage informatics platform for extended reproducible research”. In : *Nature methods* 9.7 (2012), p. 690.

[DDV00] Lieven De Lathauwer, Bart De Moor et Joos Vandewalle. “A multilinear singular value decomposition”. In : *SIAM journal on Matrix Analysis and Applications* 21.4 (2000), p. 1253-1278.

[Deb+05] Olivier Debeir, Philippe Van Ham, Robert Kiss et Christine Decaestecker. “Tracking of migrating cells under phase-contrast video microscopy with combined mean-shift processes”. In : *IEEE transactions on medical imaging* 24.6 (2005), p. 697-711.

[Deh+17] Afshin Dehghan, Enrique G Ortiz, Guang Shu et Syed Zain Masood. “Dager: Deep age, gender and emotion recognition using convolutional neural network”. In : *arXiv preprint arXiv:1702.04280* (2017).

[DCU13] Ricard Delgado-Gonzalo, Nicolas Chenouard et Michael Unser. “Spline-based deforming ellipsoids for interactive 3D bioimage segmentation”. In : *IEEE Transactions on Image Processing* 22.10 (2013), p. 3926-3940.

[Den+15] Loïc Denis, Eric Thiébaut, Ferréol Soulez, Jean-Marie Becker et Rahul Mourya. “Fast approximations of shift-variant blur”. In : *International Journal of Computer Vision* 115.3 (2015), p. 253-278.

[DDP17] Quentin Denoyelle, Vincent Duval et Gabriel Peyré. “Support recovery for sparse super-resolution of positive measures”. In : *Journal of Fourier Analysis and Applications* 23.5 (2017), p. 1153-1194.

[Den+19] Quentin Denoyelle, Vincent Duval, Gabriel Peyré et Emmanuel Soubies. "The sliding frank-wolfe algorithm and its application to super-resolution microscopy". In : *Inverse Problems* (2019).

[Des+14] Hendrik Deschout, Francesca Cella Zanacchi, Michael Mlodzianoski, Alberto Diaspro, Joerg Bewersdorf, Samuel T Hess et Kevin Braeckmans. "Precisely and accurately localizing single emitters in fluorescence microscopy". In : *Nature methods* 11.3 (2014), p. 253.

[Die+15] Alex von Diezmann, Maurice Y Lee, Matthew D Lew et WE Moerner. "Correcting field-dependent aberrations with nanoscale accuracy in three-dimensional single-molecule localization microscopy". In : *Optica* 2.11 (2015), p. 985-993.

[DY05] Chris Ding et Jieping Ye. "2-dimensional singular value decomposition for 2d maps and images". In : *Proceedings of the 2005 SIAM International Conference on Data Mining*. SIAM. 2005, p. 32-43.

[DPW11] Tie Hua Du, Wee Choo Puah et Martin Wasser. "Cell cycle phase classification in 3D in vivo microscopy of Drosophila embryogenesis". In : *BMC bioinformatics*. T. 12. 13. BioMed Central. 2011, S18.

[DFS09] Francois-Xavier Dupé, Jalal M Fadili et Jean-Luc Starck. "A proximal iteration for deconvolving Poisson noisy images using sparse representations". In : *IEEE Transactions on Image Processing* 18.2 (2009), p. 310-321.

[DP15] Vincent Duval et Gabriel Peyré. "Exact support recovery for sparse spikes deconvolution". In : *Foundations of Computational Mathematics* 15.5 (2015), p. 1315-1355.

[EPF14] David Eigen, Christian Puhrsch et Rob Fergus. "Depth map prediction from a single image using a multi-scale deep network". In : *Advances in neural information processing systems*. 2014, p. 2366-2374.

[EH01] Michael Elad et Yacov Hel-Or. "A fast super-resolution reconstruction algorithm for pure translational motion and common space-invariant blur". In : *IEEE Transactions on Image Processing* 10.8 (2001), p. 1187-1193.

[EMS00] Monicia Elrod-Erickson, Smita Mishra et David Schneider. "Interactions between the cellular and humoral immune responses in Drosophila". In : *Current Biology* 10.13 (2000), p. 781-784.

[Esc16] Paul Escande. "Approximation and estimation of integral operators Applications to the restoration of images degraded by spatially varying blurs". Thèse de doct. 2016.

[EW15] Paul Escande et Pierre Weiss. "Sparse wavelet representations of spatially varying blurring operators". In : *SIAM Journal on Imaging Sciences* 8.4 (2015), p. 2976-3014.

[EW17] Paul Escande et Pierre Weiss. "Approximation of integral operators using product-convolution expansions". In : *Journal of Mathematical Imaging and Vision* 58.3 (2017), p. 333-348.

[Eul+17] Philipp Eulenberg, Niklas Köhler, Thomas Blasi, Andrew Filby, Anne E Carpenter, Paul Rees, Fabian J Theis et F Alexander Wolf. "Reconstructing cell cycle and disease progression using deep learning". In : *Nature communications* 8.1 (2017), p. 463.

[FKW17] Jérôme Fehrenbach, Bastien Kovac et Pierre Weiss. "FitEllipsoid: a fast supervised ellipsoid segmentation plugin". In : (2017).

[Fer+06] Rob Fergus, Barun Singh, Aaron Hertzmann, Sam T Roweis et William T Freeman. "Removing camera shake from a single photograph". In : *ACM SIGGRAPH 2006 Papers*. 2006, p. 787-794.

[Fes10] Jeffrey A Fessler. "Model-based image reconstruction for MRI". In : *IEEE Signal Processing Magazine* 27.4 (2010), p. 81-89.

[FB10] Mário AT Figueiredo et José M Bioucas-Dias. "Restoration of Poissonian images using alternating direction optimization". In : *IEEE transactions on Image Processing* 19.12 (2010), p. 3133-3145.

[Fla+20] Olivier Flasseur, Loïc Denis, Éric Thiébaut et Maud Langlois. "PACO ASDI: an algorithm for exoplanet detection and characterization in direct imaging with integral field spectrographs". In : *Astronomy & Astrophysics* 637 (2020), A9.

[FR05a] Ralf C. Flicker et Francois J. Rigaut. "Anisoplanatic deconvolution of adaptive optics images". In : *J. Opt. Soc. Am. A* 22.3 (mar. 2005), p. 504-513. doi : 10.1364/JOSAA.22.000504.

[FR05b] Ralf C Flicker et Francois J Rigaut. "Anisoplanatic deconvolution of adaptive optics images". In : *JOSA A* 22.3 (2005), p. 504-513.

[FGW20] Axel Flinth, Frédéric de Gournay et Pierre Weiss. "On the linear convergence rates of exchange and continuous methods for total variation minimization". In : *Mathematical Programming* (2020), p. 1-37.

[FG00] Michel Fortin et Roland Glowinski. *Augmented Lagrangian methods: applications to the numerical solution of boundary-value problems*. T. 15. Elsevier, 2000.

[Gar+16] Sara Garbarino, Alberto Sorrentino, Anna Maria Massone, Alessia Sannino, Antonella Boselli, Xuan Wang, Nicola Spinelli et Michele Piana. "Expectation maximization and the retrieval of the atmospheric extinction coefficients by inversion of Raman lidar data". In : *Optics Express* 24.19 (2016), p. 21497-21511.

[GD14] Matan Gavish et David L Donoho. "The optimal hard threshold for singular values is $4/\sqrt{3}$". In : *IEEE Transactions on Information Theory* 60.8 (2014), p. 5040-5053.

[GCM13] M Gentile, F Courbin et G Meylan. "Interpolating point spread function anisotropy". In : *Astronomy & Astrophysics* 549 (2013), A1.

[GL89] Sarah Frisken Gibson et Frederick Lanni. "Diffraction by a circular aperture as a model for three-dimensional optical microscopy". In : *JOSA A* 6.9 (1989), p. 1357-1367.

[GL91] Sarah Frisken Gibson et Frederick Lanni. "Experimental test of an analytical model of aberration in an oil-immersion objective lens used in three-dimensional light microscopy". In : *JOSA A* 8.10 (1991), p. 1601-1613.

[GOW19] Davis Gilton, Greg Ongie et Rebecca Willett. "Neumann Networks for Linear Inverse Problems in Imaging". In : *IEEE Transactions on Computational Imaging* (2019).

[Gon+17] Dong Gong, Jie Yang, Lingqiao Liu, Yanning Zhang, Ian Reid, Chunhua Shen, Anton Van Den Hengel et Qinfeng Shi. "From motion blur to motion flow: a deep learning solution for removing heterogeneous motion blur". In : *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, p. 2319-2328.

[GBC16] Ian Goodfellow, Yoshua Bengio et Aaron Courville. *Deep learning*. MIT press, 2016.

[Goo05] Joseph W Goodman. *Introduction to Fourier optics*. Roberts et Company Publishers, 2005.

[Gra+17] Joana Sarah Grah, Jennifer Alison Harrington, Siang Boon Koh, Jeremy Andrew Pike, Alexander Schreiner, Martin Burger, Carola-Bibiane Schönlieb et Stefanie Reichelt. "Mathematical imaging methods for mitosis analysis in live-cell phase contrast microscopy". In : *Methods* 115 (2017), p. 91-99.

[GB14] Michael Grant et Stephen Boyd. *CVX: Matlab software for disciplined convex programming, version 2.1*. 2014.

[GCD12] Rémi Gribonval, Gilles Chardon et Laurent Daudet. "Blind calibration for compressed sensing by convex optimization". In : *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE. 2012, p. 2713-2716.

[Gul+14] Jaza Gul-Mohammed, Ignacio Arganda-Carreras, Philippe Andrey, Vincent Galy et Thomas Boudier. "A generic classification-based method for segmentation of nuclei in 3D images of early embryos". In : *BMC bioinformatics* 15.1 (2014), p. 9.

[Gus00] Mats GL Gustafsson. "Surpassing the lateral resolution limit by a factor of two using structured illumination microscopy". In : *Journal of microscopy* 198.2 (2000), p. 82-87.

[Gus+08] Mats GL Gustafsson, Lin Shao, Peter M. Carlton, CJ Rachel Wang, Inna N. Golubovskaya, W. Zacheus Cande, David A. Agard et John W. Sedat. "Three-dimensional resolution doubling in wide-field fluorescence microscopy by structured illumination". In : *Biophysical Journal* 94.12 (juin 2008), p. 4957-4970.

[Gus+16] Nils Gustafsson, Siân Culley, George Ashdown, Dylan M Owen, Pedro Matos Pereira et Ricardo Henriques. "Fast live-cell conventional fluorophore nanoscopy with ImageJ through super-resolution radial fluctuations". In : *Nature communications* 7.1 (2016), p. 1-9.

[Hac15] Wolfgang Hackbusch. *Hierarchical matrices: algorithms and analysis*. T. 49. Springer, 2015.

[HB11] Saima Ben Hadj et Laure Blanc-Féraud. “Restoration method for spatially variant blurred images”. In : (2011).

[HMT11] Nathan Halko, Per-Gunnar Martinsson et Joel A Tropp. “Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions”. In : *SIAM review* 53.2 (2011), p. 217-288.

[Hal+11] Michael Halter et al. “Cell cycle dependent TN-C promoter activity determined by live cell imaging”. In : *Cytometry Part A* 79.3 (2011), p. 192-202.

[Han+04] Bridget M Hanser, Mats GL Gustafsson, DA Agard et John W Sedat. “Phase-retrieved pupil functions in wide-field fluorescence microscopy”. In : *Journal of microscopy* 216.1 (2004), p. 32-48.

[He+16] Kaiming He, Xiangyu Zhang, Shaoqing Ren et Jian Sun. “Deep residual learning for image recognition”. In : *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2016, p. 770-778.

[HH70] R Hegerl et W Hoppe. “Dynamic theory of crystalline structure analysis by electron diffraction in inhomogeneous primary wave field”. In : *Berichte Der Bunsen-Gesellschaft Fur Physikalische Chemie* 74.11 (1970), p. 1148.

[Hei+13] Felix Heide, Mushfiqur Rouf, Matthias B Hullin, Bjorn Labitzke, Wolfgang Heidrich et Andreas Kolb. “High-quality computational imaging through simple lenses”. In : *ACM Transactions on Graphics (TOG)* 32.5 (2013), p. 1-14.

[HC99] Rainer Heintzmann et Christoph G Cremer. “Laterally modulated excitation microscopy: improvement of resolution by using a diffraction grating”. In : *Optical Biopsies and Microscopic Techniques III.* T. 3568. International Society for Optics et Photonics. 1999, p. 185-196.

[HW94] Stefan W Hell et Jan Wichmann. “Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy”. In : *Optics letters* 19.11 (1994), p. 780-782.

[HS92] Stefan Hell et Ernst HK Stelzer. “Properties of a 4Pi confocal fluorescence microscope”. In : *JOSA A* 9.12 (1992), p. 2159-2166.

[HS10a] René Henrion et Alberto Seeger. “Inradius and circumradius of various convex cones arising in applications”. In : *Set-Valued and Variational Analysis* 18.3-4 (2010), p. 483-511.

[HS10b] René Henrion et Alberto Seeger. “On properties of different notions of centers for convex cones”. In : *Set-Valued and Variational Analysis* 18.2 (2010), p. 205-231.

[Her45] JFW Herschel. “On the epipolic dispersion of light, being a supplement to a paper entitled,on a case of superficial colour presented by a homogeneous liquid internally colourless”. In : *Philos Trans R Soc Lond* 135 (1845), p. 147-153.

[HGM06] Samuel T Hess, Thanu PK Girirajan et Michael D Mason. "Ultra-high resolution imaging by fluorescence photoactivation localization microscopy". In : *Biophysical journal* 91.11 (2006), p. 4258-4272.

[HSA90] Yasushi Hiraoka, John W Sedat et David A Agard. "Determination of three-dimensional imaging properties of a light microscope system. Partial confocal behavior in epifluorescence microscopy". In : *Biophysical journal* 57.2 (1990), p. 325-333.

[Hir06] Jean-Baptiste Hiriart-Urruty. "A note on the Legendre-Fenchel transform of convex composite functions". In : *Nonsmooth Mechanics and Analysis*. Springer, 2006, p. 35-46.

[Hir+11] Michael Hirsch, Christian J Schuler, Stefan Harmeling et Bernhard Schölkopf. "Fast removal of non-uniform camera shake". In : *2011 International Conference on Computer Vision*. IEEE. 2011, p. 463-470.

[HS97] Sepp Hochreiter et Jürgen Schmidhuber. "Long short-term memory". In : *Neural computation* (1997).

[Hod+13] Erlend Hodneland, Tanja Kögel, Dominik Michael Frei, Hans-Hermann Gerdes et Arvid Lundervold. "CellSegm-a MATLAB toolbox for high-throughput 3D cell segmentation". In : *Source code for biology and medicine* 8.1 (2013), p. 16.

[Hol+95] Jon A Holtzman et al. "The performance and calibration of WFPC2 on the Hubble Space Telescope". In : *Publications of the Astronomical Society of the Pacific* 107.708 (1995), p. 156.

[Hop55] HH Hopkins. "The frequency response of a defocused optical system". In : *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 231.1184 (1955), p. 91-103.

[Hua+08] Bo Huang, Wenqin Wang, Mark Bates et Xiaowei Zhuang. "Three-dimensional super-resolution imaging by stochastic optical reconstruction microscopy". In : *Science* 319.5864 (2008), p. 810-813.

[HP88] Herbert G Hughes et Merle R Paulson. "Double-ended lidar technique for aerosol studies". In : *Applied optics* 27.11 (1988), p. 2273-2278.

[Huh+10] Seungil Huh, Ryoma Bise, Mei Chen, Takeo Kanade et al. "Automated mitosis detection of stem cell populations in phase-contrast microscopy images". In : *IEEE transactions on medical imaging* 30.3 (2010), p. 586-596.

[HS07] Jan Huisken et Didier YR Stainier. "Even fluorescence excitation by multidirectional selective plane illumination microscopy (mSPIM)". In : *Optics letters* 32.17 (2007), p. 2608-2610.

[Idi+17] Jérôme Idier, Simon Labouesse, Marc Allain, Penghuan Liu, Sébastien Bourguignon et Anne Sentenac. "On the superresolution capacity of imagers using unknown speckle illuminations". In : *IEEE Transactions on Computational Imaging* 4.1 (2017), p. 87-98.

[Jue+08] Manuel F Juette, Travis J Gould, Mark D Lessard, Michael J Mlodzianoski, Bhupendra S Nagpure, Brian T Bennett, Samuel T Hess et Joerg Bewersdorf. "Three-dimensional sub–100 nm resolution fluorescence microscopy of thick samples". In : *Nature methods* 5.6 (2008), p. 527-529.

[Jun+09] Hong Jung, Kyunghyun Sung, Krishna S Nayak, Eung Yeop Kim et Jong Chul Ye. "k-t FOCUSS: a general compressed sensing framework for high resolution dynamic MRI". In : *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 61.1 (2009), p. 103-116.

[JKS17] Peter Jung, Felix Krahmer et Dominik Stöger. "Blind demixing and deconvolution at near-optimal rate". In : *IEEE Transactions on Information Theory* 64.2 (2017), p. 704-727.

[KWT88] Michael Kass, Andrew Witkin et Demetri Terzopoulos. "Snakes: Active contour models". In : *International journal of computer vision* 1.4 (1988), p. 321-331.

[KK17] Michael Kech et Felix Krahmer. "Optimal injectivity conditions for bilinear inverse problems with applications to identifiability of deconvolution problems". In : *SIAM Journal on Applied Algebra and Geometry* 1.1 (2017), p. 20-37.

[Kec+13] Adel Kechkar, Deepak Nair, Mike Heilemann, Daniel Choquet et Jean-Baptiste Sibarita. "Real-time analysis and visualization for single-molecule based super-resolution microscopy". In : *PLoS One* 8.4 (2013).

[Kel+08] Philipp J Keller, Annette D Schmidt, Joachim Wittbrodt et Ernst HK Stelzer. "Reconstruction of zebrafish early embryonic development by scanned light sheet microscopy". In : *science* 322.5904 (2008), p. 1065-1069.

[KLP04] C Kervrann, D Legland et L Pardini. "Robust incremental compensation of the light attenuation with depth in 3D fluorescence microscopy". In : *Journal of Microscopy* 214.3 (2004), p. 297-314.

[Key81] Robert Keys. "Cubic convolution interpolation for digital image processing". In : *IEEE transactions on acoustics, speech, and signal processing* 29.6 (1981), p. 1153-1160.

[KB14] Diederik P Kingma et Jimmy Ba. "Adam: A method for stochastic optimization". In : *arXiv preprint arXiv:1412.6980* (2014).

[KW13] Diederik P Kingma et Max Welling. "Auto-encoding variational bayes". In : *arXiv preprint arXiv:1312.6114* (2013).

[Kir+13] Hagai Kirshner, Franois Aguet, Daniel Sage et Michael Unser. "3-D PSF fitting for fluorescence microscopy: implementation and localization application". In : *Journal of microscopy* 249.1 (2013), p. 13-25.

[Kle81] James D Klett. "Stable analytical inversion solution for processing lidar returns". In : *Applied Optics* 20.2 (1981), p. 211-220.

[KB09] Tamara G Kolda et Brett W Bader. "Tensor decompositions and applications". In : *SIAM review* 51.3 (2009), p. 455-500.

[Kra+06] Felix Krahmer, Youzuo Lin, Bonnie McAdoo, Katharine Ott, Jiakou Wang, David Widemann et Brendt Wohlberg. "Blind image deconvolution: Motion blur estimation". In : (2006).

[KS19] Felix Krahmer et Dominik Stöger. "On the convex geometry of blind deconvolution and matrix completion". In : *arXiv preprint arXiv:1902.11156* (2019).

[Krz+12] Uros Krzic, Stefan Gunther, Timothy E Saunders, Sebastian J Streichan et Lars Hufnagel. "Multiview light-sheet microscope for rapid in toto imaging". In : *Nature methods* 9.7 (2012), p. 730-733.

[KSK13] Abhishek Kumar, Vikas Sindhwani et Prabhanjan Kambadur. "Fast conical hull algorithms for near-separable non-negative matrix factorization". In : *International Conference on Machine Learning*. 2013, p. 231-239.

[Kun87] Gerard J Kunz. "Bipath method as a way to measure the spatial backscatter and extinction coefficients with lidar". In : *Applied optics* 26.5 (1987), p. 794-795.

[Lan+06] Paul Lang, Karen Yeow, Anthony Nichols et Alexander Scheer. "Cellular imaging in drug discovery". In : *Nature Reviews Drug Discovery* 5.4 (2006), p. 343-356.

[Lau99] Tod R Lauer. "The Photometry of Undersampled Point-Spread Functions". In : *Publications of the Astronomical Society of the Pacific* 111.765 (1999), p. 1434.

[Laz+17] Carole Lazarus, Pierre Weiss, Nicolas Chauffert, Franck Mauconduit, Michel Bottlaender, Alexandre Vignaud et Philippe Ciuciu. "SPARKLING: Novel non-Cartesian sampling schemes for accelerated 2D anatomical imaging at 7T using compressed sensing". In : 2017.

[Leb+19] Léo Lebrat, Frédéric de Gournay, Jonas Kahn et Pierre Weiss. "Optimal transport approximation of 2-dimensional measures". In : *SIAM Journal on Imaging Sciences* 12.2 (2019), p. 762-787.

[Lee08] Antoni Van Leeuwenhoek. "V. Microscopical observations upon the tongue; in a letter to the Royal Society from Mr. Anthony Van Leeuwenhoek, FRS". In : *Philosophical Transactions of the Royal Society of London* 26.315 (1708), p. 111-123.

[Lev+09] Anat Levin, Yair Weiss, Fredo Durand et William T Freeman. "Understanding and evaluating blind deconvolution algorithms". In : *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2009, p. 1964-1971.

[LJ16] Daniel Lévy et Arzav Jain. "Breast mass classification from mammograms using deep convolutional neural networks". In : *arXiv preprint arXiv:1612.00542* (2016).

[LBM13] Matthew D Lew, Mikael P Backlund et WE Moerner. "Rotational mobility of single molecules affects localization accuracy in super-resolution fluorescence microscopy". In : *Nano letters* 13.9 (2013), p. 3967-3972.

[LL93] Frederic Leymarie et Martin D. Levine. “Tracking deformable objects in the plane using an active contour model”. In : *IEEE Transactions on Pattern Analysis & Machine Intelligence* 6 (1993), p. 617-634.

[Li+18a] Jizhou Li, Feng Xue, Fuyang Qu, Yi-Ping Ho et Thierry Blu. “On-the-fly estimation of a microscopy point spread function”. In : *Optics express* 26.20 (2018), p. 26120-26133.

[Li+08] Kang Li, Eric D Miller, Mei Chen, Takeo Kanade, Lee E Weiss et Phil G Campbell. “Cell population tracking and lineage construction with spatiotemporal context”. In : *Medical image analysis* 12.5 (2008), p. 546-566.

[Li+18b] Xiaodong Li, Shuyang Ling, Thomas Strohmer et Ke Wei. “Rapid, robust, and reliable blind deconvolution via nonconvex optimization”. In : *Applied and Computational Harmonic Analysis* (2018).

[Li+19] Xiaodong Li, Shuyang Ling, Thomas Strohmer et Ke Wei. “Rapid, robust, and reliable blind deconvolution via nonconvex optimization”. In : *Applied and computational harmonic analysis* 47.3 (2019), p. 893-934.

[LLB16a] Yanjun Li, Kiryung Lee et Yoram Bresler. “Identifiability in bilinear inverse problems with applications to subspace or sparsity-constrained blind gain and phase calibration”. In : *IEEE Transactions on Information Theory* 63.2 (2016), p. 822-842.

[LLB16b] Yanjun Li, Kiryung Lee et Yoram Bresler. “Optimal Sample Complexity for Blind Gain and Phase Calibration.” In : *IEEE Trans. Signal Processing* 64.21 (2016), p. 5549-5556.

[LLB17] Yanjun Li, Kiryung Lee et Yoram Bresler. “Identifiability in bilinear inverse problems with applications to subspace or sparsity-constrained blind gain and phase calibration”. In : *IEEE Transactions on Information Theory* 63.2 (2017), p. 822-842.

[Li+18c] Yiming Li et al. “Real-time 3D single-molecule localization using experimental point spread functions”. In : *Nature methods* 15.5 (2018), p. 367.

[LS15] Shuyang Ling et Thomas Strohmer. “Self-calibration and biconvex compressive sensing”. In : *Inverse Problems* 31.11 (2015), p. 115002.

[LS18] Shuyang Ling et Thomas Strohmer. “Self-Calibration and Bilinear Inverse Problems via Linear Least Squares”. In : *SIAM Journal on Imaging Sciences* 11.1 (2018), p. 252-292.

[Liu+13] Sheng Liu, Emil B Kromann, Wesley D Krueger, Joerg Bewersdorf et Keith A Lidke. “Three dimensional single molecule localization using a phase retrieved pupil function”. In : *Optics express* 21.24 (2013), p. 29462-29487.

[LSC12] Vebjorn Ljosa, Katherine L Sokolnicki et Anne E Carpenter. “Annotated high-throughput microscopy image sets for validation.” In : *Nature methods* 9.7 (2012), p. 637-637.

[Low04] David G Lowe. “Distinctive image features from scale-invariant keypoints”. In : *International journal of computer vision* 60.2 (2004), p. 91-110.

[MP89] Yvon Maday et Anthony T Patera. “Spectral element methods for the incompressible Navier-Stokes equations”. In : *IN: State-of-the-art surveys on computational mechanics (A90-47176 21-64). New York, American Society of Mechanical Engineers, 1989, p. 71-143. Research supported by DARPA.* 1989, p. 71-143.

[Mal99] Stéphane Mallat. *A wavelet tour of signal processing.* Elsevier, 1999.

[Man+20] Thomas Mangeat et al. “Super-resolved live-cell imaging using Random Illumination Microscopy”. In : *bioRxiv* (2020).

[Mar+13] Joanne Marrison, Lotta Räty, Poppy Marriott et Peter O’toole. “Ptychography–a label free, high-contrast imaging technique for live cells using quantitative phase information”. In : *Scientific reports* 3 (2013), p. 2369.

[MC10] Cédric Matthews et Fabrice P Cordelieres. “MetroloJ: an ImageJ plugin to help monitor microscopes’ health”. In : *ImageJ User & Developer Conference proceedings.* 2010.

[May+14] Jürgen Mayer, Alexandre Robert-Moreno, Renzo Danuser, Jens V Stein, James Sharpe et Jim Swoger. “OPTiSPIM: integrating optical projection tomography in light sheet microscopy extends specimen characterization to nonfluorescent contrasts”. In : *Optics letters* 39.4 (2014), p. 1053-1056.

[Mbo+15] FM Ngolé Mboula, J-L Starck, Samuel Ronayette, Koryo Okumura et Jérôme Amiaux. “Super-resolution method using sparse regularization for point-spread function recovery”. In : *Astronomy & Astrophysics* 575 (2015), A86.

[McC64] CW McCutchen. “Generalized aperture and the three-dimensional diffraction image”. In : *JOSA* 54.2 (1964), p. 240-244.

[MP12] David Miraut et Javier Portilla. “Efficient shift-variant image restoration using deformable filtering (Part I)”. In : *EURASIP Journal on Advances in Signal Processing* 2012.1 (2012), p. 100.

[Mlo+18] Michael J Mlodzianoski et al. “Active PSF shaping and adaptive optics enable volumetric localization microscopy through brain sections”. In : *Nature methods* 15.8 (2018), p. 583-586.

[MLB14] Leonhard Möckl, Don C Lamb et Christoph Bräuchle. “Super-resolved Fluorescence Microscopy: Nobel Prize in Chemistry 2014 for Eric Betzig, Stefan Hell, and William E. Moerner”. In : *Angewandte Chemie International Edition* 53.51 (2014), p. 13972-13977.

[MK89] William E Moerner et Lothar Kador. “Optical detection and spectroscopy of single molecules in a solid”. In : *Physical review letters* 62.21 (1989), p. 2535.

[Möl+14] Michael Möller, Martin Burger, Peter Dieterich et Albrecht Schwab. “A framework for automated cell tracking in phase contrast microscopic videos based on normal velocities”. In : *Journal of Visual Communication and Image Representation* 25.2 (2014), p. 396-409.

[MHA08] Morten Morup, Lars Kai Hansen et Sidse M Arnfred. “Algorithms for sparse nonnegative Tucker decompositions”. In : *Neural computation* 20.8 (2008), p. 2112-2131.

[Mou16] Rahul Kumar Mourya. “Contributions to image restoration: from numerical optimization strategies to blind deconvolution and shift-variant deblurring”. Thèse de doct. Lyon, 2016.

[Mou+15] Rahul Mourya, Loïc Denis, Jean-Marie Becker et Éric Thiébaut. “A blind deblurring and image decomposition approach for astronomical image restoration”. In : *2015 23rd European Signal Processing Conference (EUSIPCO)*. IEEE. 2015, p. 1636-1640.

[Mug+01] Laurent M Mugnier, Clélia Robert, Jean-Marc Conan, Vincent Michau et Sélim Salem. “Myopic deconvolution from wave-front sensing”. In : *JOSA A* 18.4 (2001), p. 862-872.

[NO98] James G Nagy et Dianne P O’Leary. “Restoring images degraded by spatially variant blur”. In : *SIAM Journal on Scientific Computing* 19.4 (1998), p. 1063-1082.

[NHM17] Seungjun Nah, Tae Hyun Kim et Kyoung Mu Lee. “Deep multi-scale convolutional neural network for dynamic scene deblurring”. In : *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, p. 3883-3891.

[Nat86] Frank Natterer. *The mathematics of computerized tomography*. T. 32. Siam, 1986.

[Neh+] Elias Nehme et al. “DeepSTORM3D: dense three dimensional localization microscopy and point spread function design by deep learning”. In : *Lateral* 20 (), p. 40.

[Nes13] Yu Nesterov. “Gradient methods for minimizing composite functions”. In : *Mathematical Programming* 140.1 (2013), p. 125-161.

[Nes18] Yurii Nesterov. *Lectures on convex optimization*. T. 137. Springer, 2018.

[NN94] Yurii Nesterov et Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*. T. 13. Siam, 1994.

[Neu+10] Beate Neumann et al. “Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes”. In : *Nature* 464.7289 (2010), p. 721.

[NWY10] Michael K Ng, Pierre Weiss et Xiaoming Yuan. “Solving constrained total-variation image restoration and reconstruction problems via alternating direction methods”. In : *SIAM journal on Scientific Computing* 32.5 (2010), p. 2710-2736.

[Ngo+16] F Ngolé, Jean-Luc Starck, Koryo Okumura, Jérôme Amiaux et P Hudelot. “Constraint matrix factorization for space variant PSFs field restoration”. In : *Inverse Problems* 32.12 (2016), p. 124001.

[Ngu+18] Hoai-Nam Nguyen, Vincent Paveau, Cyril Cauchois et Charles Kervrann. "A variational method for dejittering large fluorescence line scanner images". In : *IEEE Transactions on Computational Imaging* 4.2 (2018), p. 241-256.

[Nol76] Robert J Noll. "Zernike polynomials and atmospheric turbulence". In : *JOsA* 66.3 (1976), p. 207-211.

[NCF17] Mehdi Noroozi, Paramanand Chandramouli et Paolo Favaro. "Motion deblurring in the wild". In : *German conference on pattern recognition*. Springer. 2017, p. 65-77.

[Nyq28] Harry Nyquist. "Certain topics in telegraph transmission theory". In : *Transactions of the American Institute of Electrical Engineers* 47.2 (1928), p. 617-644.

[Ort68] James M Ortega. "The Newton-Kantorovich theorem". In : *The American Mathematical Monthly* 75.6 (1968), p. 658-660.

[Pan+09] Praveen Pankajakshan, Laure Blanc-Féraud, Zvi Kam et Josiane Zerubia. "Point-spread function retrieval for fluorescence microscopy". In : *International Symposium on Biomedical Imaging*. IEEE. 2009, p. 1095-1098.

[Pas+19] Adam Paszke et al. "Pytorch: An imperative style, high-performance deep learning library". In : *Advances in neural information processing systems*. 2019, p. 8026-8037.

[PP15] Nurmohammed Patwary et Chrysanthe Preza. "Image restoration for three-dimensional fluorescence microscopy using an orthonormal basis for efficient representation of depth-variant point-spread functions". In : *Biomedical optics express* 6.10 (2015), p. 3826-3841.

[Pav+09] Sri Rama Prasanna Pavani, Michael A Thompson, Julie S Biteen, Samuel J Lord, Na Liu, Robert J Twieg, Rafael Piestun et WE Moerner. "Three-dimensional, single-molecule fluorescence imaging beyond the diffraction limit by using a double-helix point spread function". In : *Proceedings of the National Academy of Sciences* 106.9 (2009), p. 2995-2999.

[PF14] Daniele Perrone et Paolo Favaro. "Total variation blind deconvolution: The devil is in the details". In : *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014, p. 2909-2916.

[PSM17] Petar N Petrov, Yoav Shechtman et WE Moerner. "Measurement-based estimation of global pupil functions in 3D localization microscopy". In : *Optics express* 25.7 (2017), p. 7945-7959.

[Pin85] Allan Pinkus. *N-widths in Approximation Theory*. T. 7. Springer Science & Business Media, 1985.

[Pol07] Boris T Polyak. "Newton's method and its use in optimization". In : *European Journal of Operational Research* 181.3 (2007), p. 1086-1096.

[Pop11] Gabriel Popescu. *Quantitative phase imaging of cells and tissues*. McGraw Hill Professional, 2011.

[Por+08] Pornsarp Pornsawad, Christine Böckmann, Christoph Ritter et Mathias Rafler. “Ill-posed retrieval of aerosol extinction coefficient profiles from Raman lidar data by regularization”. In : *Applied optics* 47.10 (2008), p. 1649-1661.

[PC04] Chrysanthe Preza et José-Angel Conchello. “Depth-variant maximum-likelihood restoration for three-dimensional fluorescence microscopy”. In : *JOSA A* 21.9 (2004), p. 1593-1601.

[QPP12] Sean Quirin, Sri Rama Prasanna Pavani et Rafael Piestun. “Optimal 3D single-molecule localization for superresolution microscopy with aberrations and engineered point spread functions”. In : *Proceedings of the National Academy of Sciences* 109.3 (2012), p. 675-679.

[RWO05] Sripad Ram, E Sally Ward et Raimund J Ober. “How accurately can a single molecule be localized in three dimensions using a fluorescence microscope?” In : *Imaging, Manipulation, and Analysis of Biomolecules and Cells: Fundamentals and Applications III*. T. 5699. International Society for Optics et Photonics. 2005, p. 426-435.

[RV91] Jean Paul Rigaut et Jany Vassy. “High-resolution three-dimensional images from confocal scanning laser microscopy. Quantitative study and mathematical correction of the effects from bleaching and fluorescence attenuation in depth.” In : *Analytical and quantitative cytology and histology/the International Academy of Cytology [and] American Society of Cytology* 13.4 (1991), p. 223-232.

[RB93] JBTM Roerdink et Miente Bakker. “An FFT-based method for attenuation correction in fluorescence confocal microscopy”. In : *Journal of microscopy* 169.1 (1993), p. 3-14.

[RC08] Pantaleón D Romero et Vicente F Candela. “Blind deconvolution models regularized by fractional powers of the Laplacian”. In : *Journal of Mathematical Imaging and Vision* 32.2 (2008), p. 181-191.

[RFB15] Olaf Ronneberger, Philipp Fischer et Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation”. In : *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015.

[RHP07] Gianluigi Rozza, Dinh Bao Phuong Huynh et Anthony T Patera. “Reduced basis approximation and a posteriori error estimation for affinely parametrized elliptic coercive partial differential equations”. In : *Archives of Computational Methods in Engineering* 15.3 (2007), p. 1.

[ROF92] Leonid I Rudin, Stanley Osher et Emad Fatemi. “Nonlinear total variation based noise removal algorithms”. In : *Physica D: Nonlinear Phenomena* 60.1 (1992), p. 259-268.

[RBZ06] Michael J Rust, Mark Bates et Xiaowei Zhuang. “Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM)”. In : *Nature methods* 3.10 (2006), p. 793.

[Sag+17a] D. Sage et al. “DeconvolutionLab2: An Open-Source Software for Deconvolution Microscopy”. In : *Methods—Image Processing for Biologists* 115 (fév. 2017), p. 28-41.

[Sag+19] Daniel Sage et al. "Super-resolution fight club: assessment of 2D and 3D single-molecule localization microscopy software". In : *Nature methods* 16.5 (2019), p. 387-395.

[Sag+15] Daniel Sage, Hagai Kirshner, Thomas Pengo, Nico Stuurman, Junhong Min, Suliana Manley et Michael Unser. "Quantitative evaluation of software packages for single-molecule localization microscopy". In : *Nature methods* 12.8 (2015), p. 717-724.

[Sag+17b] Daniel Sage et al. "DeconvolutionLab2: An open-source software for deconvolution microscopy". In : *Methods* 115 (2017), p. 28-41.

[Sch+09] Otmar Scherzer, Markus Grasmair, Harald Grossauer, Markus Haltmeier et Frank Lenzen. *Variational methods in imaging*. Springer, 2009.

[Sch+12] Johannes Schindelin et al. "Fiji: an open-source platform for biological-image analysis". In : *Nature methods* 9.7 (2012), p. 676.

[Sch+13] Thorsten Schmidt, Jasmin Dürr, Margret Keuper, Thomas Blein, Klaus Palme et Olaf Ronneberger. "Variational attenuation correction in two-view confocal microscopy". In : *BMC bioinformatics* 14.1 (2013), p. 366.

[Sch+15] Christian J Schuler, Michael Hirsch, Stefan Harmeling et Bernhard Schölkopf. "Learning to deblur". In : *IEEE transactions on pattern analysis and machine intelligence* 38.7 (2015), p. 1439-1451.

[SP03] Rajesh Babu Sekar et Ammasi Periasamy. "Fluorescence resonance energy transfer (FRET) microscopy imaging of live cell protein localizations". In : *The Journal of cell biology* 160.5 (2003), p. 629-633.

[Sha+16] Lei Sha, Baobao Chang, Zhifang Sui et Sujian Li. "Reading and thinking: Re-read lstm unit for textual entailment recognition". In : *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 2016, p. 2870-2879.

[SF07] Joshua W Shaevitz et Daniel A Fletcher. "Enhanced three-dimensional deconvolution microscopy using a measured depth-varying point-spread function". In : *JOSA A* 24.9 (2007), p. 2622-2627.

[SL20a] Adrian Shajkofci et Michael Liebling. "DeepFocus: a Few-Shot Microscope Slide Auto-Focus using a Sample Invariant CNN-based Sharpness Function". In : *arXiv preprint arXiv:2001.00667* (2020).

[SL20b] Adrian Shajkofci et Michael Liebling. "Spatially-Variant CNN-Based Point Spread Function Estimation for Blind Deconvolution and Depth Estimation in Optical Microscopy". In : *IEEE Transactions on Image Processing* 29 (2020), p. 5848-5861.

[SJA08] Qi Shan, Jiaya Jia et Aseem Agarwala. "High-quality motion deblurring from a single image". In : *Acm transactions on graphics (tog)* 27.3 (2008), p. 1-10.

[Sha48] Claude E Shannon. "A mathematical theory of communication". In : *Bell system technical journal* 27.3 (1948), p. 379-423.

[Sha+02] James Sharpe, Ulf Ahlgren, Paul Perry, Bill Hill, Allyson Ross, Jacob Hecksher-Sørensen, Richard Baldock et Duncan Davidson. “Optical projection tomography as a tool for 3D microscopy and gene expression studies”. In : *Science* 296.5567 (2002), p. 541-545.

[Shc07] Valery Shcherbakov. “Regularized algorithm for Raman lidar data processing”. In : *Applied optics* 46.22 (2007), p. 4879-4889.

[She+14] Yoav Shechtman, Steffen J Sahl, Adam S Backer et WE Moerner. “Optimal point spread function design for 3D imaging”. In : *Physical review letters* 113.13 (2014), p. 133902.

[Shi+20] Rui Shi, Norik Janunts, Christian Hellmann et Frank Wyrowski. “Vectorial physical-optics modeling of Fourier microscopy systems in nanooptics”. In : *JOSA A* 37.7 (2020), p. 1193-1205.

[SJS62] Osamu Shimomura, Frank H Johnson et Yo Saiga. “Extraction, purification and properties of aequorin, a bioluminescent protein from the luminous hydromedusan, Aequorea”. In : *Journal of cellular and comparative physiology* 59.3 (1962), p. 223-239.

[SS87] George Sines et Yannis A Sakellarakis. “Lenses in antiquity”. In : *American Journal of Archaeology* (1987), p. 191-196.

[Söd+08] Ola Söderberg, Karl-Johan Leuchowius, Mats Gullberg, Malin Jarvius, Irene Weibrecht, Lars-Gunnar Larsson et Ulf Landegren. “Characterizing proteins and their interactions in cells and tissues using the in situ proximity ligation assay”. In : *Methods* 45.3 (2008), p. 227-232.

[Som+11] Christoph Sommer, Christoph Straehle, Ullrich Koethe et Fred A Hamprecht. “Ilastik: Interactive learning and segmentation toolkit”. In : *2011 IEEE international symposium on biomedical imaging: From nano to macro*. IEEE. 2011, p. 230-233.

[Sou14] Ferréol Soulez. “A “learn 2D, apply 3D” method for 3D deconvolution microscopy”. In : *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2014, p. 1075-1078.

[SCU16] Ferréol Soulez, Frédéric Courbin et Michael Unser. “Back-propagating the light of field stars to probe telescope mirrors aberrations”. In : *Advances in Optical and Mechanical Technologies for Telescopes and Instrumentation II*. T. 9912. International Society for Optics et Photonics. 2016, p. 991277.

[Sou+12] Ferréol Soulez, Loïc Denis, Yves Tourneur et Éric Thiébaut. “Blind deconvolution of 3D data in wide field fluorescence microscopy”. In : *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2012, p. 1735-1738.

[ST10] Gabriele Steidl et Tanja Teuber. “Removing multiplicative noise by Douglas-Rachford splitting methods”. In : *Journal of Mathematical Imaging and Vision* 36.2 (2010), p. 168-184.

[Sto69] Per A Stokseth. “Properties of a defocused optical system”. In : *JOSA* 59.10 (1969), p. 1314-1321.

[SD18] Mohamed A Suliman et Wei Dai. "Blind two-dimensional super-resolution and its performance guarantee". In : *arXiv preprint arXiv:1811.02070* (2018).

[Sun+15] Jian Sun, Wenfei Cao, Zongben Xu et Jean Ponce. "Learning a convolutional neural network for non-uniform motion blur removal". In : *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, p. 769-777.

[Sun+13] Libin Sun, Sunghyun Cho, Jue Wang et James Hays. "Edge-based blur kernel estimation using patch priors". In : *IEEE International Conference on Computational Photography (ICCP)*. IEEE. 2013, p. 1-8.

[TMK14] Patrick Theer, Cyril Mongis et Michael Knop. "PSFj: know your fluorescence microscope". In : *Nature methods* 11.10 (2014), p. 981-982.

[Thi+13] Ketheesan Thirusittampalam, M Julius Hossain, Ovidiu Ghita et Paul F Whelan. "A novel framework for cellular tracking and mitosis detection in dense phase contrast microscopy images". In : *IEEE journal of biomedical and health informatics* 17.3 (2013), p. 642-653.

[TW15] Lei Tian et Laura Waller. "3D intensity and phase imaging from light field measurements in an LED array microscope". In : *optica* 2.2 (2015), p. 104-111.

[Tom+12] Raju Tomer, Khaled Khairy, Fernando Amat et Philipp J Keller. "Quantitative high-speed imaging of entire developing embryos with simultaneous multiview light-sheet microscopy". In : *Nature methods* 9.7 (2012), p. 755-763.

[TA20] Yann Traonmilin et Jean-Francois Aujol. "The basins of attraction of the global minimizers of the non-convex sparse spike estimation problem". In : *Inverse Problems* 36.4 (2020), p. 045003.

[Tsi98] Roger Y Tsien. *The green fluorescent protein*. 1998.

[Tuc66] Ledyard R Tucker. "Some mathematical notes on three-mode factor analysis". In : *Psychometrika* 31.3 (1966), p. 279-311.

[Tur+20] Raphaël Turcotte, Eusebiu Sutu, Carla C Schmidt, Nigel J Emptage et Martin J Booth. "Deconvolution for multimode fiber imaging: modeling of spatially variant PSF". In : *Biomedical Optics Express* 11.8 (2020), p. 4759-4771.

[Tyr12] Eugene E Tyrtyshnikov. *A brief introduction to numerical analysis*. Springer Science & Business Media, 2012.

[VRR17] Greg Van Buskirk, Benjamin Raichel et Nicholas Ruozzi. "Sparse approximate conic hulls". In : *Advances in Neural Information Processing Systems*. 2017, p. 2534-2544.

[Ver+14] KA Vermeer, J Mo, JJA Weda, HG Lemij et JF de Boer. "Depth-resolved model-based reconstruction of attenuation coefficients in optical coherence tomography". In : *Biomedical optics express* 5.1 (2014), p. 322-337.

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. T. 47. Cambridge university press, 2018.

[VS91] Luc Vincent et Pierre Soille. "Watersheds in digital spaces: an efficient algorithm based on immersion simulations". In : *IEEE Transactions on Pattern Analysis & Machine Intelligence* 6 (1991), p. 583-598.

[WSS01] Wes Wallace, Lutz H Schaefer et Jason R Swedlow. "A workingperson's guide to deconvolution in light microscopy". In : *Biotechniques* 31.5 (2001), p. 1076-1097.

[WBB11] Huahua Wang, Arindam Banerjee et Daniel Boley. "Common component analysis for multiple covariance matrices". In : *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2011, p. 956-964.

[Wan+07] Meng Wang, Xiaobo Zhou, Fuhai Li, Jeremy Huckins, Randall W King et Stephen TC Wong. "Novel cell segmentation and online SVM for cell cycle phase identification in automated microscopy". In : *Bioinformatics* 24.1 (2007), p. 94-101.

[Wan+16] Renjie Wang, Alain Kamgoue, Christophe Normand, Isabelle Léger-Silvestre, Thomas Mangeat et Olivier Gadal. "High resolution microscopy reveals the nuclear shape of budding yeast during cell cycle and in various biological states". In : *Journal of cell science* 129.24 (2016), p. 4480-4495.

[Wei06] Claus Weitkamp. *Lidar: range-resolved optical remote sensing of the atmosphere*. T. 102. Springer Science & Business, 2006.

[WKP16] Tobias Weyand, Ilya Kostrikov et James Philbin. "Planet-photo geolocation with convolutional neural networks". In : *European Conference on Computer Vision*. Springer. 2016, p. 37-55.

[Why+12] Oliver Whyte, Josef Sivic, Andrew Zisserman et Jean Ponce. "Non-uniform deblurring for shaken images". In : *International journal of computer vision* 98.2 (2012), p. 168-186.

[Xu+20] Fan Xu et al. "Three-dimensional nanoscopy of whole cells and tissues with in situ point spread function retrieval". In : *Nature Methods* 17.5 (2020), p. 531-540.

[XJ10] Li Xu et Jiaya Jia. "Two-phase kernel estimation for robust motion deblurring". In : *European conference on computer vision*. Springer. 2010, p. 157-170.

[Xu+11] Li Xu, Cewu Lu, Yi Xu et Jiaya Jia. "Image smoothing via L 0 gradient minimization". In : *Proceedings of the 2011 SIGGRAPH Asia Conference*. 2011, p. 1-12.

[Xu+14] Li Xu, Jimmy SJ Ren, Ce Liu et Jiaya Jia. "Deep convolutional neural network for image deconvolution". In : *Advances in neural information processing systems*. 2014, p. 1790-1798.

[Xu+17] Mengjia Xu, Dimitrios P Papageorgiou, Sabia Z Abidi, Ming Dao, Hong Zhao et George Em Karniadakis. "A deep convolutional neural network for classification of red blood cells in sickle cell anemia". In : *PLoS computational biology* 13.10 (2017), e1005746.

[Yan+19] T. Yan, C. J. Richardson, M. Zhang et A. Gahlmann. "Computational correction of spatially variant optical aberrations in 3D single-molecule localization microscopy". In : *Opt. Express* 27.9 (avr. 2019), p. 12582-12599. doi : 10.1364/OE.27.012582.

[Ye05] Jieping Ye. "Generalized low rank approximations of matrices". In : *Machine Learning* 61.1-3 (2005), p. 167-191.

[YSG06] Hongki Yoo, I Song et D-G GWEON. "Measurement and restoration of the point spread function of fluorescence confocal microscopy". In : *Journal of microscopy* 221.3 (2006), p. 172-176.

[YK99] Yu-Li You et Mostafa Kaveh. "Blind image restoration by anisotropic regularization". In : *IEEE Transactions on Image Processing* 8.3 (1999), p. 396-407.

[Zer42] Frits Zernike. "Phase contrast, a new method for the microscopic observation of transparent objects". In : *Physica* 9.7 (1942), p. 686-698.

[ZZO07] Bo Zhang, Josiane Zerubia et Jean-Christophe Olivo-Marin. "Gaussian approximations of fluorescence microscope point-spread function models". In : *Applied optics* 46.10 (2007), p. 1819-1829.

[Zha+17] Kai Zhang, Wangmeng Zuo, Shuhang Gu et Lei Zhang. "Learning deep CNN denoiser prior for image restoration". In : *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2017, p. 3929-3938.

[Zhe+13] Guoan Zheng, Xiaoze Ou, Roarke Horstmeyer et Changhuei Yang. "Characterization of spatially varying aberrations for wide field-of-view microscopy". In : *Optics express* 21.13 (2013), p. 15131-15143.

[Zho+08] Xiaobo Zhou, Fuhai Li, Jun Yan et Stephen TC Wong. "A novel cell segmentation method and cell phase identification using Markov model". In : *IEEE Transactions on Information Technology in Biomedicine* 13.2 (2008), p. 152-157.

[Zhu+97] Ciyou Zhu, Richard H Byrd, Peihuang Lu et Jorge Nocedal. "Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization". In : *ACM Transactions on Mathematical Software (TOMS)* 23.4 (1997), p. 550-560.

[Zhu+18] Z. Zhu, Q. Li, G. Tang et M. B. Wakin. "Global Optimality in Low-Rank Matrix Optimization". In : *IEEE Transactions on Signal Processing* 66.13 (2018), p. 3614-3628.

Annexe A

Semi infinite generalized Graetz problem

Résumé : *Le problème de Graetz est une équation de convection-diffusion dans un tuyau invariant le long d'une direction. La contribution du présent travail est de proposer une analyse mathématique de la condition limite de Neumann, Robin et périodique sur la limite d'un tuyau semi-infini. La solution dans l'espace 3D du problème original est réduite à des problèmes aux vecteurs propres dans la section 2D du tuyau. L'ensemble des solutions est décrit, sa structure dépend du type de condition limite et du signe de l'écoulement total du fluide. Cette analyse est la pierre angulaire des méthodes numériques permettant de résoudre le problème de Graetz dans les tuyaux finies, semi-infinies et les échangeurs de section arbitraire. Des exemples numériques illustrent les capacités de ces méthodes à fournir des solutions dans diverses configurations.*

Abstract : *The Graetz problem is a convection-diffusion equation in a pipe invariant along a direction. The contribution of the present work is to propose a mathematical analysis of the Neumann, Robin and periodic boundary condition on the boundary of a semi-infinite pipe. The solution in the 3D space of the original problem is reduced to eigenproblems in the 2D section of the pipe. The set of solutions is described, its structure depends on the type of boundary condition and of the sign of the total flow of the fluid. This analysis is the cornerstone of numerical methods to solve Graetz problem in finite pipes, semi infinite pipes and exchangers of arbitrary cross section. Numerical test-cases illustrate the capabilities of these methods to provide solutions in various configurations.*

This chapter is based on the publication [Deb+18b] :
Debarnot, V., Fehrenbach, J., de Gournay, F., & Martire, L. (2018). The Case of Neumann, Robin, and Periodic Lateral Conditions for the Semi-infinite Generalized Graetz Problem and Applications. SIAM Journal on Applied Mathematics, 78(4), 2227-2251.
This was part of a student project at INSA. With Léo Martire, we conducted some numerical experiments. Jérôme Fehrenbach and Frédéric de Gournay conducted the project, the mathematical analysis and some of the numerical experiments. They have been generous by including us to the list of authors. In the following, we display the article as published in SIAM journal.

THE CASE OF NEUMANN, ROBIN, AND PERIODIC LATERAL CONDITIONS FOR THE SEMI-INFINITE GENERALIZED GRAETZ PROBLEM AND APPLICATIONS*

VALENTIN DEBARNOT†, JÉRÔME FEHRENBACH‡, FRÉDÉRIC DE GOURNAY†, AND LÉO MARTIRE†

Abstract. The Graetz problem is a convection-diffusion equation in a pipe invariant along a direction. The contribution of the present work is to propose a mathematical analysis of the Neumann, Robin, and periodic boundary condition on the boundary of a semi-infinite pipe. The solution in the 3D space of the original problem is reduced to eigenproblems in the 2D section of the pipe. The set of solutions is described, its structure depending on the type of boundary condition and on the sign of the total flow of the fluid. This analysis is the cornerstone of numerical methods to solve the Graetz problem in finite pipes, semi-infinite pipes, and exchangers of arbitrary cross-section. Numerical test cases illustrate the capabilities of these methods to provide solutions in various configurations.

Key words. Graetz problem, numerical analysis, eigenvalue decomposition

AMS subject classifications. 65M70, 35Q79

DOI. 10.1137/17M1157507

1. Introduction.

1.1. Context. The seminal work of Graetz in the late 19th century addressed a stationary convection-diffusion problem inside an axisymmetrical cylindrical pipe [5], where the regime was supposed to be convection-dominated, which means that the longitudinal diffusion was neglected. It was the first contribution to the modeling of convective transport coupled with diffusion, with such important applications nowadays as the parallel convective exchangers involved in heating or cooling systems [16], haemodialysis [1], and heat exchangers [7]. The first extension to the Graetz problem, known as the "extended Graetz problem," takes into account longitudinal diffusion [10, 3, 18, 9]. Papoutsakis, Ramkrishna, and Lim in [12, 11] introduced a symmetric operator acting on a two-component space that solves the extended Graetz problem in axisymmetrical configurations. The so-called conjugated Graetz problem where multiple solid or fluid phases are taken into account was proposed in [13, 14] in the case of an axisymmetrical configuration. These successive models aimed at tackling more and more complex and realistic situations, and when only axisymmetrical configurations were considered, the equations boiled down to one-dimensional problems. The adaptation to parallel plate heat exchangers of these one-dimensional models, together with a parametric study, was proposed in [6]. The reader may also consult [2] for a review on the conjugated Graetz problem.

The work on nonaxisymmetrical configurations was initiated in [15] where the operator was proved to be self-adjoint with compact resolvent when Dirichlet boundary

*Received by the editors November 17, 2017; accepted for publication (in revised form) May 16, 2018; published electronically August 29, 2018.
http://www.siam.org/journals/siap/78-4/M115750.html

†Institut de Mathématiques de Toulouse (UMR 5219), Université de Toulouse, CNRS INSA, F-31077 Toulouse, France (valentindebarnot@gmail.com, frederic@degournay.fr, leo.martire@outlook.com).

‡Institut de Mathématiques de Toulouse (UMR 5219), Université de Toulouse, CNRS, UPS, F-31062 Toulouse Cedex 9, France (jerome.fehrenbach@math.univ-toulouse.fr).

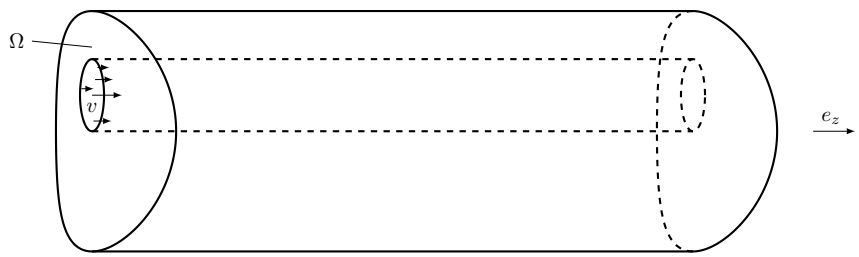


FIG. 1.1. *The domain $\Omega \times I$ where the Graetz problem is posed.*

conditions are applied on the boundary of the domain. In the case of a single fluid stream the negative eigenvalues correspond to downstream propagation, and positive eigenvalues to upstream propagation. The main novelty was that arbitrary geometries were addressed, and a detailed mathematical analysis of the Dirichlet problem was proposed. The authors of [15] called this the "generalized Graetz problem." Numerical methods for the approximation of this operator and error estimates were provided in [4].

The objective of the present work is to extend the work of [4] and provide explicit methods with general lateral boundary conditions, beyond the Dirichlet case. The cross-section of the domain has an arbitrary geometry and can incorporate different fluid domains, possibly with opposite signs of the velocity. The lateral boundary conditions that we address can be Dirichlet, Neumann, Robin, periodic, or a mixture of these different cases on different parts of the boundary. The periodic boundary conditions with rectangular or hexagonal cell are adapted to the analysis of micro-exchangers, where a design pattern is repeated.

1.2. Setting. In convection-dominated heat or mass transfer, we address the generalized Graetz problem, which occurs in a cylinder of arbitrary section Ω and of length I, possibly $I = \mathbb{R}^+$; see Figure 1.1. The diffusion coefficient is supposed to be invariant by translation along e_z, the axis of the cylinder. Similarly, the velocity vector v is supposed to be oriented in the direction of the axis of the cylinder, that is, $v = he_z$ with $h \in L^\infty(\Omega)$. The equation for the temperature T inside the domain is then

$$c\partial_{zz}T + \operatorname{div}(\sigma\nabla T) - h\partial_z T = 0 \text{ on } \Omega \times I, \tag{E}$$

with diffusion coefficients $c, \sigma > 0$ bounded in Ω with bounded inverse. In (E), it is implicitly supposed that the heat capacity C and the density ρ of the fluid satisfy $\rho C = 1$. If one has to handle several fluids with different physical properties, the choice of an adequate normalization leads to (E). In this case, at each point $x \in \Omega$, $h(x)$ represents the velocity multiplied by $\rho(x)C(x)$. For the sake of simplicity, in what follows we refer to h as the *velocity*.

The lateral homogeneous boundary conditions (LBC) may be of Neumann, Dirichlet, Robin, or periodic type, respectively, on Γ_N, Γ_D, Γ_R, $\Gamma_\sharp \subset \partial\Omega$ given by

$$\begin{cases} \sigma\nabla T \cdot n = 0 \text{ on } \Gamma_N \times I\text{: Neumann, and/or} \\ T = 0 \text{ on } \Gamma_D \times I\text{: Dirichlet, and/or} \\ \sigma\nabla T \cdot n + aT = 0 \text{ on } \Gamma_R \times I\text{: Robin, and/or} \\ T \text{ is periodic on } \Gamma_\sharp \times I\text{: periodic,} \end{cases} \tag{LBC}$$

where $a > 0$ in the Robin condition, and $\Gamma_\sharp$ must be tailored to support periodic conditions (e.g., Ω is the unit square, $\Gamma_\sharp = (\{x = 0\} \cup \{x = 1\}) \cap \partial\Omega$, and the boundary condition is $T(0, y) = T(1, y)$). As usual, the Γ's involved in the definition of the boundary condition must form a partition of $\partial\Omega$. Note that the Neumann (resp., Dirichlet) boundary conditions are degenerate cases of the Robin condition corresponding to $a = 0$ (resp., $a = +\infty$). The Inlet/Outlet boundary condition (I/OBC) is of Dirichlet and/or of Neumann type and is given by

$$\text{(I/OBC)} \qquad T = T_D \text{ on } \Omega_D \text{ and } \partial_z T = S_N \text{ on } \Omega_N \text{ with } \Omega_D \cup \Omega_N = \Omega \times \partial I.$$

In the case $I = \mathbb{R}^+$, we intentionally stay vague about the definition of ∂I; it is one of the results of this work to determine whether an Inlet/Outlet boundary condition is needed on $z = +\infty$.

A more realistic model in regimes of high velocities takes into account a viscosity term; see, e.g., [8], where a study in a microchannel including viscous effects and longitudinal conduction is performed. Our approach can also account for viscosity; details are presented in section 2.3.

1.3. Lax–Milgram. Note that the equation (E) is an elliptic equation with an additional convective term. It is possible to use the Lax–Milgram theorem [17] under the hypothesis that the Inlet/Outlet boundary condition is Dirichlet in the region where the flow is incoming. The following proposition states this more precisely.

PROPOSITION 1. *Let $I = [z_1, z_2]$ and $\omega_\pm = \{x \text{ s.t } \pm h(x) > 0\}$. If*

$$\omega_+ \times \{z_1\} \subset \Omega_D \text{ and } \omega_- \times \{z_2\} \subset \Omega_D$$

and if T_D and S_N are regular enough, then there exists a unique solution to (E) *with the boundary conditions* (LBC) *and* (I/OBC).

The proof is only sketched here for the sake of completeness. Denote by $\mathcal{X}$ the natural space of elements where the solution is sought, that is,

$$\mathcal{X} = \{T \in H^1(\Omega \times I) \text{ s.t. } T = 0 \text{ on } (\Gamma_D \times I) \cup \Omega_D \text{ and } T \text{ periodic on } \Gamma_\sharp \times I\}.$$

Nonhomogeneous Dirichlet boundary conditions of (I/OBC) are solved using a lift of T_D, still denoted T_D, that satisfies the lateral boundary conditions (LBC) with $\partial_z T_D = 0$ on Ω_N and denote

$$f_D = c\partial_{zz} T_D + \operatorname{div}(\sigma \nabla T_D) - h\partial_z T_D.$$

The change of unknown $\widetilde{T} = T - T_D$, where T solves (E) and (LBC), leads to the following variational formulation: find $\widetilde{T} \in \mathcal{X}$ such that for every $\phi \in \mathcal{X}$,

$$\underbrace{\int_{\Omega\times I} c\partial_z \widetilde{T} \partial_z \phi + \sigma \nabla \widetilde{T} \cdot \nabla \phi + h\partial_z \widetilde{T} \phi + \int_{\Gamma_R} a\widetilde{T}\phi}_{b(\widetilde{T},\phi)} + \underbrace{\int_{\Omega_N} S_N \phi - \int_{\Omega\times I} f_D \phi}_{\ell(\phi)} = 0.$$

The term $b(T, \phi)$ is bilinear in (T, ϕ) and continuous for the standard norm of $\mathcal{X}$; the term $\ell(\phi)$ is linear continuous if T_D and S_N are regular enough. It remains to study the coercivity of b.

$$b(T, T) = \int_{\Omega\times I} c\partial_z T.\partial_z T + \sigma \nabla T \cdot \nabla T + h\partial_z T.T = \int_{\Omega\times I} \|\nabla_{3\mathrm{D}} T\|_\kappa^2 + \frac{1}{2}\int_{\Omega\times I} h\partial_z(T^2),$$

where κ is a positive matrix with diagonal entries (σ, σ, c) in the basis (e_x, e_y, e_z). The first term is coercive. The second term is

$$\frac{1}{2}\int_\Omega hT^2|_{z=z_1}^{z=z_2} = \frac{1}{2}\int_{\Omega\times\{z_2\}} hT^2 - \frac{1}{2}\int_{\Omega\times\{z_1\}} hT^2.$$

It is nonnegative for all $T \in \mathcal{X}$ if and only if the Inlet/Outlet condition is of Dirichlet type at the boundary where the flow is entering the domain ($z = z_1$ if $h > 0$, and $z = z_2$ if $h < 0$).

1.4. Presentation of the paper. The objective of the present paper is to provide a general framework that allows one to solve (E) with any type of boundary condition beyond the case where the Lax–Milgram theorem can be used. Section 2 details the notation and the main properties of the operator involved in the solution, as well as the modifications required to take into account a viscosity term. The main results, namely, Theorems 4 and 5, are detailed in section 3, and their proofs are postponed to the appendix. In section 4 we solve the problem in a semi-infinite domain and show that, depending on the case, the temperature at infinity T_∞ can either be a free parameter of the problem or be imposed by the other condition. In section 5 we address the case of a domain of finite length, and numerical strategies are detailed in the different cases depending on the lateral boundary condition and on the Inlet/Outlet condition. Test cases are presented in section 6.

2. State of the art and position of the problem. Equation (E) may be interpreted as an evolution equation in the variable z if it is cast into

$$\partial_z \begin{pmatrix} \partial_z T \\ T \end{pmatrix} = \mathcal{A} \begin{pmatrix} \partial_z T \\ T \end{pmatrix} \text{ on } \Omega \times I, \text{ with } \mathcal{A}\begin{pmatrix} u \\ s \end{pmatrix} = \begin{pmatrix} hc^{-1}u - c^{-1}\operatorname{div}\sigma\nabla s \\ u \end{pmatrix}. \tag{2.1}$$

The goal of this section is to guide the reader to the analysis of (2.1) that was proposed in [4], to enlarge the frame to Neumann and periodic lateral boundary conditions, and to define the notation and state the results that will be used in the remainder of the paper. Since $\mathcal{A}$ is a symmetric operator with a compact resolvent, classical eigendecomposition leads to an explicit representation of the solution of (2.1) in the basis of eigenvectors (see, e.g., [17]).

Definition 2. *We say that* the constants are not controlled *when* $\Gamma_D \cup \Gamma_R = \emptyset$, *in other words when there is no Dirichlet or Robin condition on the lateral part of the boundary of the domain. The case where the constants are not controlled and in addition* $\int_\Omega h = 0$ *is called the* balanced case.

From an engineering point of view, the balanced case is a special instance of countercurrent configuration, where the integral of the velocities in the 2 directions have the same magnitude and the boundary of the domain Ω is perfectly insulating or periodic. As we prove in this section, the case where the constants are not controlled is a case where the constants are a solution of (E), and the balanced case is a case where $\mathcal{A}$ admits a nontrivial kernel.

2.1. Study of the operator $\mathcal{A}$. In this section we detail the Hilbert space, the scalar product, the kernel, the range, and the pseudoinverse of the symmetric operator $\mathcal{A}$.

Hilbert space and scalar product. First, introduce the space H that encodes the lateral boundary condition. When the constants are controlled, define

$$H = \{s \in H^1(\Omega), \text{ such that } s = 0 \text{ on } \Gamma_D \text{ and } s \text{ periodic on } \Gamma_\sharp\}.$$

If there is no Dirichlet or Robin boundary condition, hence no control on the constants, quotient by the constants and define

$$H = \{s \in H^1(\Omega)/\mathbb{R}, \text{ such that } s \text{ periodic on } \Gamma_\sharp\}.$$

Then define the Hilbert space $\mathcal{H}$ as

$$\mathcal{H} = \{(u,s) \,|\, u \in L^2(\Omega), s \in H\},$$

which is endowed with the scalar product

$$\big((u,s) \,|\, (u',s')\big)_{\mathcal{H}} = \int_\Omega cuu' + \sigma\nabla s \cdot \nabla s' + \int_{\Gamma_R} ass'.$$

The crucial step in showing that $\mathcal{H}$ is a Hilbert space is to show that the scalar product is definite. Setting $\big((u,s) \,|\, (u,s)\big)_{\mathcal{H}} = 0$ immediately gives $u = 0$ and $\nabla s = 0$, and hence s is a constant. If the constants are controlled, then $\Gamma_D \cup \Gamma_R \neq \emptyset$ and $s = 0$, whereas if the constants are not controlled, then s is a constant and $s = 0$ in H.

The domain of the operator $\mathcal{A}$ is

$$\mathcal{D}(\mathcal{A}) = \{(u,s) \in \mathcal{H}, u \in H^1(\Omega), \operatorname{div}(\sigma\nabla s) \in L^2(\Omega) + \text{boundary conditions (LBC)}\},$$

where the boundary conditions are $u \in H$, and where $\sigma\nabla s \cdot n$ is equal to 0 on Γ_N, is equal to $-as$ on Γ_R, and is periodic on $\Gamma_\sharp$. On $\mathcal{D}(\mathcal{A})$, the operator is symmetric, as we prove now. Let $\phi = (u,s)$ and $\phi' = (u',s') \in \mathcal{D}(\mathcal{A})$:

$$\begin{aligned}(\mathcal{A}\phi|\phi')_{\mathcal{H}} &= \int_\Omega (hu - \operatorname{div}\sigma\nabla s)u' + \sigma\nabla u \cdot \nabla s' + \int_{\Gamma_R} aus' \\ &= \int_\Omega huu' + \sigma\nabla u \cdot \nabla s' + \sigma\nabla u' \cdot \nabla s + \underbrace{\int_{\partial\Omega} (-\sigma\nabla s \cdot n)u' + \int_{\Gamma_R} aus'}_{(\mathbf{1})},\end{aligned}$$

and the term (**1**) is symmetric thanks to (LBC) on $\mathcal{D}(\mathcal{A})$.

Inverse of the Laplacian. Define the inverse of the Laplace operator as

$$u = \Delta_\sigma^{-1} f \text{ iff } \begin{cases} \operatorname{div}(\sigma\nabla u) = f, \text{ and} \\ u \in H \\ + \text{ boundary conditions,} \end{cases}$$

where the boundary conditions are $\sigma\nabla u \cdot n = 0$ on Γ_N and $\sigma\nabla u \cdot n + au = 0$ on Γ_R. If the constants are controlled, then Δ_σ^{-1} is well defined on $L^2(\Omega)$, whereas if there are only Neumann or periodic boundary conditions (no control of the constants), the operator Δ_σ^{-1} is only defined if $f \in L^2_m(\Omega)$, the subspace of $L^2(\Omega)$ with null average.

Kernel of $\mathcal{A}$. Following from the definition of $\mathcal{A}$ in (2.1), the kernel of $\mathcal{A}$ is the set of (u,s) in $\mathcal{D}(\mathcal{A})$ such that

$$u = 0 \text{ in } H \text{ and } hu - \operatorname{div}(\sigma\nabla s) = 0.$$

When the constants are controlled, both u and s are then equal to 0. When the constants are not controlled, since u is a constant, then $s = u\Delta_\sigma^{-1}h$ in Ω, which admits a solution if and only if $\int_\Omega h = 0$. To summarize, the kernel of $\mathcal{A}$ is

$$\mathcal{K}(\mathcal{A}) = \begin{cases} \operatorname{Vect}(\phi_0 = (1, \Delta_\sigma^{-1}h)) \text{ in the balanced case,} \\ \{0\} \text{ in the other cases.} \end{cases}$$

Range and inverse of $\mathcal{A}$. The range of $\mathcal{A}$, denoted $\mathcal{R}(\mathcal{A})$, is defined as the orthogonal of $\mathcal{K}(\mathcal{A})$ in $\mathcal{H}$, and the inverse of $\mathcal{A}$ is an operator from $\mathcal{R}(\mathcal{A})$ to $\mathcal{D}(\mathcal{A})$, defined as follows:

$$\forall \phi = (u,s) \in \mathcal{R}(\mathcal{A}),\ \mathcal{A}^{-1}\phi = \begin{cases} (s, \Delta_\sigma^{-1}(hs - cu)) \text{ if the constants are controlled,} \\ (s+k, \Delta_\sigma^{-1}(hs - cu + hk)), k \in \mathbb{R}, \text{ if not.} \end{cases}$$

When the constants are not controlled, the constant $k \in \mathbb{R}$ is chosen so that

$$\begin{cases} \int_\Omega hs - cu + hk = 0 \text{ in the nonbalanced case,} \\ (\mathcal{A}^{-1}\phi, \phi_0)_{\mathcal{H}} = 0 \text{ in the balanced case.} \end{cases}$$

It is easily checked that for all $\phi \in \mathcal{R}(\mathcal{A})$, $\mathcal{A}^{-1}\phi \in \mathcal{D}(\mathcal{A})$ and that $\mathcal{A}\mathcal{A}^{-1}\phi = \phi$. The operator $\mathcal{A}^{-1}$ is then symmetric (as a consequence of the symmetry of $\mathcal{A}$). Note also that in the balanced case one can also write

$$\forall \phi = (u,s) \in \mathcal{R}(\mathcal{A}), \quad \mathcal{A}^{-1}\phi = (s, \Delta_\sigma^{-1}(hs - cu)) + k\phi_0.$$

Eigenvalue decomposition of $\mathcal{A}$. The operator $\mathcal{A}^{-1}$ is a compact self-adjoint operator on $\mathcal{R}(\mathcal{A})$. To prove this let $\phi_n = (u_n, s_n)$ be a bounded sequence in $\mathcal{H}$. Then up to a subsequence it is a weakly convergent sequence and s_n converges strongly in $L^2(\Omega)$. Using the fact that Δ_σ^{-1} is a compact operator from L^2 to H finishes the proof. We denote by λ_i the nonzero ordered eigenvalues of $\mathcal{A}$ and by $\phi_i = (U_i, \lambda_i^{-1}U_i)$ the corresponding eigenvectors. By convention, λ_i is of the sign of i so that

$$-\infty \leftarrow \lambda_{-n} \leq \lambda_{-n-1} \leq \cdots \leq \lambda_{-1} < 0 < \lambda_1 \leq \cdots \leq \lambda_{n-1} \leq \lambda_n \to +\infty.$$

In the balanced case, we add to the family $(\phi_i)_i$ the vector $\phi_0 = (1, \Delta_\sigma^{-1}h)$, so that the Hilbert space $\mathcal{H}$ is the space spanned by the eigenvectors $(\phi_i)_{i\in\mathbb{Z}}$.

2.2. Solution of the evolution equation. The diagonalization of the operator $\mathcal{A}$ allows us to solve the evolution equation (E):

$$c\partial_{zz}T + \operatorname{div}(\sigma\nabla T) - h\partial_z T = 0 \text{ on } \Omega \times I. \tag{E}$$

Let $T \in C^1(I, L^2(\Omega)) \cap C^0(I, H)$ be a solution of this equation with corresponding lateral boundary conditions (LBC). If we denote $\phi : z \mapsto (\partial_z T(z), T(z))$ in $C^0(\mathcal{H})$, then equation (E) is equivalent to $\partial_z\phi = \mathcal{A}\phi$, and the solution ϕ is given by

$$\phi(z) = \sum_{i\in\mathbb{Z}} \frac{(\phi(0)|\phi_i)_{\mathcal{H}}}{\|\phi_i\|_{\mathcal{H}}^2} e^{\lambda_i z}\phi_i. \tag{2.2}$$

One can either identify the first coordinate and integrate with respect to z or identify the second coordinate and denote

$$\psi = \sum_{i\in\mathbb{Z}^*} \lambda_i^{-1} \frac{(\phi(0)|\phi_i)_{\mathcal{H}}}{\|\phi_i\|_{\mathcal{H}}^2}\phi_i,$$

to obtain

$$T(z) = \sum_{i\in\mathbb{Z}^*} (\psi|\phi_i)_{\mathcal{H}} U_i e^{\lambda_i z} \text{ if the constants are controlled, i.e., } \Gamma_D \cup \Gamma_R \neq \emptyset,$$

$$T(z) = \sum_{i\in\mathbb{Z}^*} (\psi|\phi_i)_{\mathcal{H}} U_i e^{\lambda_i z} + a_0 \text{ with } a_0 \in \mathbb{R} \text{ if } \Gamma_D \cup \Gamma_R = \emptyset \text{ and } \int_\Omega h \neq 0,$$

$$T(z) = \sum_{i\in\mathbb{Z}^*} (\psi|\phi_i)_{\mathcal{H}} U_i e^{\lambda_i z} + a_0 + a_1(z + \Delta_\sigma^{-1} h) \text{ in the balanced case,}$$

$$\text{with } a_0 \in \mathbb{R} \text{ and } a_1 = \frac{(\phi(0)|\phi_0)_{\mathcal{H}}}{\|\phi_0\|^2}.$$

If $\partial_z T$ and T are given at $z = 0$ such that $\phi(0) = (\partial_z T(0), T(0))$ belongs to $\mathcal{H}$, then ψ is uniquely determined. Moreover the constant a_0 is also determined by $T_{|z=0}$ (and also a_1 in the balanced case). We stress that this solution may not be defined everywhere; indeed the series on the right-hand side of (2.2) has to be convergent in some sense, and the convergence of the series for $z = 0$ is not sufficient to ensure the convergence for $z \neq 0$ due to the multiplication by $e^{\lambda_i z}$ for nonzero λ_i's. The set of initial datum ϕ that allows this series to exist is known as the set of compatible initial conditions for the Cauchy problem.

2.3. Including a viscous term. Let us consider the following modification of equation (E), where a viscous term is added:

$$c\partial_{zz} T + \operatorname{div}(\sigma \nabla T) - h\partial_z T = \mu |\nabla h|^2. \tag{2.3}$$

Proposition 3. *Let T be the solution of the Graetz equation with viscosity* (2.3)*. Then there exists an explicit change of unknown function that transforms the problem with viscosity into a problem without viscosity of the form* (E)*. Therefore the solution of problem* (2.3) *reduces to the solution of the original problem* (E)*.*

Proof. Once a particular solution $\widetilde{T}$ is found, the change of variable $\widehat{T} = T - \widetilde{T}$ transforms by linearity the problem with viscosity (2.3) into the problem without viscosity. We distinguish different cases, depending on if the constants are controlled or not, and in the case the constants are not controlled we treat separately the non-balanced and the balanced cases. In each case we provide an explicit particular solution $\widetilde{T}$.

(a) If the constants are controlled, a particular solution is given by

$$\widetilde{T} = \Delta_\sigma^{-1}(\mu|\nabla h|^2).$$

(b) If the constants are not controlled, in the nonbalanced case

$$\widetilde{T} = \alpha z + \Delta_\sigma^{-1}(\mu|\nabla h|^2 + \alpha h),$$

where $\alpha \in \mathbb{R}$ satisfies

$$\int_\Omega \left(\mu|\nabla h|^2 + \alpha h\right) = 0.$$

(c) If the constants are not controlled, in the balanced case the particular solution is given by $\widetilde{T} = \alpha(\frac{z^2}{2} + z\Delta_\sigma^{-1} h) + \Delta_\sigma^{-1}\gamma$, with $\alpha \in \mathbb{R}$ and $\gamma \in L^2(\Omega)$ such that

$$\begin{cases} \alpha \left(\displaystyle\int_\Omega c - h\Delta_\sigma^{-1} h\right) = \displaystyle\int_\Omega \mu|\nabla h|^2, \\ \gamma = \mu|\nabla h|^2 - \alpha(c - h\Delta_\sigma^{-1} h). \end{cases} \tag{2.4}$$

The choice of α ensures that γ has zero average so that $\Delta_\sigma^{-1}\gamma$ is well defined. Note that α is well defined since

$$\int_\Omega c - h\Delta_\sigma^{-1}h = \int_\Omega c + \int_\Omega \sigma|\nabla h|^2 > 0.$$

Note that the last term is equal to $\|\phi_0\|_{\mathcal{H}}^2$. □

3. Main decomposition theorem. In this section, the decomposition of a temperature field on the nonpositive eigenspace is studied. The result stated in Theorem 5 considers different cases depending on the control of constants and the sign of the total flow.

3.1. Notation and statement of the problem. The $D(\mathcal{A}^\alpha)$-norm, or "α-norm" for short, is defined by

$$\|\phi\|_\alpha^2 = \sum_{i\in\mathbb{Z}} \lambda_i^{2\alpha} \frac{(\phi_i|\phi)_{\mathcal{H}}^2}{\|\phi_i\|_{\mathcal{H}}^2} \quad \forall \phi \in \mathcal{H}.$$

The space $D(\mathcal{A}^\alpha)$ is the set of $\phi \in \mathcal{R}(\mathcal{A})$ whose α-norm is $< +\infty$. It is easy to check that $D(\mathcal{A}^1) = D(\mathcal{A})$ and that $D(\mathcal{A}^0) = \mathcal{R}(\mathcal{A})$. Define P as an orthogonal projection on $\mathcal{H}$:

$$\forall \phi = (u,s) \in \mathcal{H}, \quad P\phi = (u,0). \tag{3.1}$$

For any I subset of $\mathbb{Z}$ define π_I as the orthogonal projection:

$$\pi_I \phi = \sum_{i\in I} \frac{(\phi_i|\phi)_{\mathcal{H}}}{\|\phi_i\|_{\mathcal{H}}^2}\phi_i. \tag{3.2}$$

We denote $\pi_+ = \pi_{\mathbb{N}^*}$, $\pi_- = \pi_{-\mathbb{N}^*}$, $\pi_0 = \pi_{\{0\}}$, and $\mathcal{R}(\pi_I) = \pi_I(\mathcal{H})$.

The problem of decomposition of a temperature field on the nonpositive eigenspace is stated as follows:

For any $\phi \in \mathcal{H}$, find ψ such that

$$P\psi = P\phi \quad \textit{and} \quad \pi_+\psi = 0. \tag{3.3}$$

A similar problem of decomposition on the nonnegative eigenspace is obtained by replacing π_+ by π_-. All the results of the present section have a counterpart obtained by changing the sign of z.

3.2. Necessary and sufficient condition. In order to tackle problem (3.3), we first consider the following related problem:

$$\text{Find } \psi \in \mathcal{R}(\pi_-) \text{ such that } \pi_- P \pi_- \psi = \pi_- P\phi. \tag{3.4}$$

Indeed if ψ solves (3.3), then multiplying the equation by π_- and assuming that the kernel of $\mathcal{A}$ is reduced to the nullspace (which is true except in the balanced case), one derives (3.4). Such a problem admits a unique solution, given by the following theorem.

THEOREM 4. *The operator $\pi_- P \pi_-$ is invertible on $\mathcal{R}(\pi_-)$. Define B_- as the self-adjoint operator of $\mathcal{H}$:*

$$B_- \phi = \pi_- (\pi_- P \pi_-)^{-1} \pi_- \phi.$$

Moreover it holds that

$$\|B_- \phi\|_{\mathcal{H}} \leq C \|\pi_- \phi\|_{\mathcal{H}} \quad \text{and} \quad \|B_- \phi\|_{1/2} \leq C \|\pi_- \phi\|_{1/2}.$$

One can similarly define an operator B_+, obtained by replacing π_- by π_+.

The result is proved in [4] for the full-Dirichlet case, that is, $\Gamma_D = \partial\Omega$. The proof can be adapted without major changes to the case in consideration. It is reproduced in Appendix B for the convenience of the reader. Problem (3.3) is then solved in the next theorem.

THEOREM 5. *Let $\phi \in \mathcal{H}$, define $\Phi = (1, 0) \in \mathcal{H}$, and consider problem* (3.3) *of finding ψ a solution of*

$$P\psi = P\phi \quad \text{and} \quad \pi_+ \psi = 0. \tag{3.3}$$

- *If the constants are not controlled and $\int_\Omega h > 0$, there exists a solution if and only if*
$$(\Phi - PB_- \Phi | \phi)_{\mathcal{H}} = 0.$$
In this case, the solution is unique and given by $\psi = B_- P\phi$.
- *If the constants are not controlled and $\int_\Omega h = 0$ (balanced case), then $(\Phi - PB_- \Phi | \Phi)_{\mathcal{H}} \neq 0$ and there exists a unique solution given by*
$$\psi = B_- P\phi + \frac{(\Phi - PB_- \Phi | \phi)_{\mathcal{H}}}{(\Phi - PB_- \Phi | \Phi)_{\mathcal{H}}} (B_- \Phi - \phi_0).$$
- *In every other case $\psi = B_- P\phi$ is the unique solution.*

The proof of this result is given in Appendix C.

4. Resolution of the semi-infinite problem. In the semi-infinite problem, the equation is set on the cylinder $\Omega \times \mathbb{R}^+$; see Figure 4.1. Equation (E) becomes

$$\partial_{zz} T + \Delta T + h \partial_z T = 0 \text{ on } \Omega \times \mathbb{R}^+.$$

In this section we address different cases depending on the type of Inlet/Outlet condition, namely, either Dirichlet or Neumann.

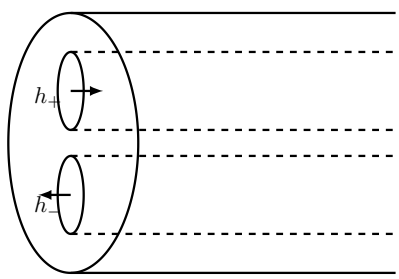


FIG. 4.1. *Example of a semi-infinite cylinder, with two fluid domains.*

In order to ensure uniqueness of the solution, we add the extra hypothesis that the temperature does not grow exponentially. We will say that the temperature has subexponential growth if and only if, for every $\lambda > 0$, $T(z) = o(e^{\lambda z})$ as z goes to $+\infty$.

4.1. Semi-infinite problem, Dirichlet Inlet/Outlet condition. We consider the Dirichlet (I/OBC) condition:

$$T_{|z=0} = T_0 \quad \text{on } \Omega. \tag{4.1}$$

Denote $\phi_{\mathcal{D}} = (T_0, 0) \in \mathcal{H}$.

Proposition 6. *Consider the Graetz problem* (E) *on the semi-infinite cylinder* $\Omega \times [0, +\infty)$*, with subexponential growth together with Dirichlet Inlet/Outlet condition* (4.1).

(a) *If the constants are controlled, then there exists a unique solution given by*

$$T = \sum_{i<0} e^{\lambda_i z} (B_- \phi_{\mathcal{D}} | \phi_i)_{\mathcal{H}} U_i.$$

In this case the temperature at infinity is 0.

(b) *If the constants are not controlled and* $\int_\Omega h \neq 0$*, then there exists a unique solution given by*

$$T = \sum_{i<0} e^{\lambda_i z} (B_-(\phi_{\mathcal{D}} - T_\infty \Phi) | \phi_i)_{\mathcal{H}} U_i + T_\infty,$$

where T_∞ *is an arbitrary constant in the case* $\int_\Omega h < 0$*, and* $T_\infty = (\Phi - PB_-\Phi | \phi_{\mathcal{D}})_{\mathcal{H}} (\Phi - PB_-\Phi | \Phi)_{\mathcal{H}}^{-1}$ *in the case* $\int_\Omega h > 0$*. In this case the temperature at infinity is the constant* T_∞*. Note that if* $\int_\Omega h > 0$*, the temperature at infinity is determined by* $\phi_{\mathcal{D}}$*, whereas in the case* $\int_\Omega h < 0$*, it is a free parameter of the problem.*

(c) *In the balanced case, the set of solutions is given by*

$$T(z) = \sum_{i<0} (B_-(\phi_{\mathcal{D}} - c_1 \Phi - c_2(\Delta_\sigma^{-1} h, 0)) | \phi_i)_{\mathcal{H}} U_i e^{\lambda_i z} + c_1 + c_2(z + \Delta_\sigma^{-1} h),$$

where c_2 *is an arbitrary constant and*

$$c_1 = (\Phi - PB_-\Phi | \phi_{\mathcal{D}} - c_2(\Delta_\sigma^{-1} h, 0))_{\mathcal{H}} (\Phi - PB_-\Phi | \Phi)_{\mathcal{H}}^{-1}.$$

In this case the temperature at infinity has the linear growth rate $T(z) \simeq c_2 z + (c_2 \Delta_\sigma^{-1} h + c_1) + o(z)$*. If the temperature is not allowed to have a linear growth rate, then* $c_2 = 0$ *and the temperature at infinity is* c_1*, which is determined by the initial conditions.*

Proof. We use the result of section 2.2 on the solution of the evolution equation; that is, T solves (E) if and only if there exist $\psi \in \mathcal{R}(\mathcal{A})$ and constants c_1 and c_2 such that

$$T(z) = \sum_{i \in \mathbb{Z}^*} (\psi | \phi_i)_{\mathcal{H}} U_i e^{\lambda_i z} + c_1 + c_2(z + \Delta_\sigma^{-1} h),$$

where $c_1 = c_2 = 0$ if the constants are controlled and $c_2 = 0$ in the nonbalanced case. The subexponential growth condition ensures that $\pi_+ \psi = 0$. The condition $(T(z = 0), 0) = \phi_{\mathcal{D}}$ yields

$$P\phi_{\mathcal{D}} = P\psi + c_1 \Phi + c_2(\Delta_\sigma^{-1} h, 0). \tag{4.2}$$

Using Theorem 5 leads to distinguishing the following cases:

(a) If the constants are controlled, then $c_1 = c_2 = 0$ and the equation $P\phi_{\mathcal{D}} = P\psi$ with $\pi_+\psi = 0$ has the unique solution $\psi = B_-\phi_{\mathcal{D}}$.

(b) In the nonbalanced case, $c_2 = 0$ and $P(\phi_{\mathcal{D}} - c_1\Phi) = P\psi$ together with $\pi_+\psi = 0$ imply $\psi = B_-(\phi_{\mathcal{D}} - c_1\Phi)$ without any additional assumption in the case $\int_\Omega h < 0$. In the case where $\int_\Omega h > 0$ the compatibility condition is

$$(\Phi - PB_-\Phi|\phi_{\mathcal{D}} - c_1\Phi) = 0, \text{ which gives } c_1 = (\Phi - PB_-\Phi|\phi_{\mathcal{D}})(\Phi - PB_-\Phi|\Phi)^{-1}.$$

(c) Finally, in the balanced case, let us fix an arbitrary value c_2. The conditions $\psi \in \mathcal{R}(\mathcal{A})$ and $\pi_+\psi = 0$ are equivalent to $\psi \in \mathcal{R}(\pi_-)$. Denoting $\widetilde{\Phi} = (\Delta_\sigma^{-1}h, 0)$ we recast (4.2) into

$$P\phi_{\mathcal{D}} = P\psi + c_1\Phi + c_2\widetilde{\Phi}, \qquad \psi \in \mathcal{R}(\pi_-),$$

or equivalently

$$P(\phi_{\mathcal{D}} - c_2\widetilde{\Phi}) = P(\psi + c_1\phi_0), \qquad \psi \in \mathcal{R}(\pi_-).$$

In view of Theorem 5 with $\phi = \phi_{\mathcal{D}} - c_2\widetilde{\Phi}$, there is a unique solution to the above equation given by

$$c_1 = (\Phi - PB_-\Phi|\phi_{\mathcal{D}} - c_2\widetilde{\Phi})_{\mathcal{H}}(\Phi - PB_-\Phi|\Phi)_{\mathcal{H}}^{-1}, \quad \psi = B_-(\phi_{\mathcal{D}} - c_1\Phi - c_2\widetilde{\Phi}). \qquad \square$$

4.2. Semi-infinite problem, Neumann Inlet/Outlet condition. We consider the Neumann (I/OBC) condition:

$$\partial_z T_{|z=0} = S_0. \tag{4.3}$$

Denote $\phi_{\mathcal{N}} = (S_0, 0) \in \mathcal{H}$.

PROPOSITION 7. *Consider the Graetz problem* (E) *on the semi-infinite cylinder* $\Omega\times[0,+\infty)$*, with subexponential growth together with Neumann Inlet/Outlet condition* (4.3).

(a) *If the constants are controlled, then there exists a unique solution given by*

$$T = \sum_{i<0} e^{\lambda_i z}(\mathcal{A}^{-1}B_-\phi_{\mathcal{N}}|\phi_i)_{\mathcal{H}}U_i.$$

In this case the temperature at infinity is 0.

(b) *If the constants are not controlled and* $\int_\Omega h \neq 0$*, then the following hold: If* $\int_\Omega h > 0$*, a solution always exists. If* $\int_\Omega h < 0$*, there exists a solution if and only if*

$$(\Phi - PB_-\Phi|\phi_{\mathcal{N}}) = 0.$$

When the solution exists, it is of the form

$$T = \sum_{i<0} e^{\lambda_i z}(\mathcal{A}^{-1}B_-(\phi_{\mathcal{N}} - T_\infty\Phi)|\phi_i)_{\mathcal{H}}U_i + T_\infty,$$

where the temperature at infinity T_∞ *is a free parameter of the problem.*

(c) *In the balanced case, the set of solutions is given by*

$$T(z) = \sum_{i<0}(\mathcal{A}^{-1}B_-(\phi_{\mathcal{N}} + c_2\Phi)|\phi_i)_{\mathcal{H}}U_i e^{\lambda_i z} + c_1 + c_2(z + \Delta_\sigma^{-1}h),$$

where c_1 is an arbitrary constant and c_2 is given by

$$c_2 = -(\Phi - PB_-\Phi|\phi_{\mathcal{N}})(\Phi - PB_-\Phi|\Phi)^{-1}.$$

In this case the temperature at infinity has the linear growth rate $T(z) \simeq c_2 z + (c_2\Delta_\sigma^{-1}h + c_1) + o(z)$.

Proof. We proceed as in the previous section. It follows from the result of section 2.2 on the solution of the evolution equation that T solves (E) if and only if there exist $\psi \in \mathcal{R}(\mathcal{A})$ and constants c_1 and c_2 such that

$$T(z) = \sum_{i\in\mathbb{Z}^*} (\psi|\phi_i)_{\mathcal{H}} U_i e^{\lambda_i z} + c_1 + c_2(z + \Delta_\sigma^{-1}h),$$

where $c_1 = c_2 = 0$ if the constants are controlled and $c_2 = 0$ in the nonbalanced case. Differentiating with respect to z, one finds

$$P\phi_{\mathcal{N}} = P\mathcal{A}\psi + c_2\Phi, \quad \pi_+\psi = 0. \tag{4.4}$$

Using Theorem 5 leads to distinguishing the following cases:

(a) If the constants are controlled, then $c_1 = c_2 = 0$ and (4.4) admits a unique solution $\mathcal{A}\psi = B_-\phi_{\mathcal{N}}$. The invertibility of $\mathcal{A}$ gives the result.

(b) If the constants are not controlled, then $c_2 = 0$. If $\int_\Omega h > 0$, there is always a solution $\mathcal{A}\psi$ to (4.4) and the operator $\mathcal{A}$ is invertible; hence there exists a unique solution ψ to (4.4) given by $\psi = \mathcal{A}^{-1}B_-\phi_{\mathcal{N}}$. The constant c_1 is then a free parameter of the problem. If $\int_\Omega h < 0$, then the condition for (4.4) to admit a solution is

$$(\Phi - PB_-\Phi|\phi_{\mathcal{N}}) = 0.$$

If this condition is met, by the invertibility of $\mathcal{A}$, $\psi = \mathcal{A}^{-1}B_-\phi_{\mathcal{N}}$ is the unique solution to (4.4) and c_1 is a free parameter of the problem.

(c) In the balanced case let c_2 be an arbitrary constant. It follows from Theorem 5 that $\mathcal{A}\psi$ satisfies (4.4) if and only if

$$\mathcal{A}\psi = B_-P\phi + \frac{(\Phi - PB_-\Phi|\phi)_{\mathcal{H}}}{(\Phi - PB_-\Phi|\Phi)_{\mathcal{H}}}(B_-\Phi - \phi_0)_{\mathcal{H}}, \quad \text{where } \phi = \phi_{\mathcal{N}} + c_2\Phi. \tag{4.5}$$

For ψ to exist, the right-hand side must belong to the range of $\mathcal{A}$, i.e., be orthogonal to ϕ_0. Performing the scalar product of the left-hand side of (4.5) with ϕ_0 and recalling that the range of B_- is orthogonal to $\mathcal{K}(\mathcal{A})$, we obtain the necessary condition

$$(\Phi - PB_-\Phi|\phi_{\mathcal{N}} + c_2\Phi)_{\mathcal{H}} = 0,$$

which is equivalent to

$$c_2 = -(\Phi - PB_-\Phi|\phi_{\mathcal{N}})(\Phi - PB_-\Phi|\Phi)^{-1}.$$

Conversely, if the above condition is met, then (4.5) admits a unique inverse in $\mathcal{R}(\mathcal{A})$ and c_1 is a free parameter of the problem. ∎

5. Resolution of the problem in a finite domain. We aim to solve the Graetz equation in a domain of finite length $\Omega \times [-L, L]$:

$$\begin{cases} c\partial_{zz}T + \operatorname{div}\sigma\nabla T - h\partial_z T = 0, & \Omega \times [-L, L], \\ \text{(LBC)}, & (\partial\Omega) \times [-L, L], \\ \text{Inlet/Outlet condition}, & \Omega \times \{-L, L\}, \end{cases} \tag{5.1}$$

where the Inlet/Outlet conditions can be of Neumann or Dirichlet type. According to section 2.2, the solutions may be sought in the form

$$T(z) = \sum_{i<0} (\psi|\phi_i) e^{\lambda_i (z+L)} U_i + \sum_{i>0} (\psi|\phi_i) e^{\lambda_i (z-L)} U_i + c_1 + c_2 (z + \Delta_\sigma^{-1} h),$$

with $c_1 = c_2 = 0$ if the constants are controlled and $c_2 = 0$ in the nonbalanced case.

The unknowns in this equation are $(\psi|\phi_i)$ for $i < 0$ and $i > 0$, plus possibly (depending on the case) c_1 and c_2. Note that $\sum_{i<0} (\psi|\phi_i) U_i = P\pi_- \psi$, and therefore if $P\pi_-\psi$ is known, it suffices to decompose this vector on the basis of $L^2(\Omega)$ given by $(U_i)_{i<0}$ to obtain the desired coefficients for $i < 0$. Similarly the coefficients $(\psi|\phi_i)$ for $i > 0$ are obtained by considering the coefficients of $P\pi_+\psi$ on the basis composed of the $(U_i)_{i>0}$. Therefore the unknowns to be determined are $P\pi_-\psi, P\pi_+\psi$ plus possibly c_1 and c_2.

Let X be the vector composed of all the unknowns. Then satisfying the Inlet/Outlet conditions amounts to solving a linear system for X. In the rest of this section we detail the linear system in each case, but first we focus on a linear operator involved in the system.

5.1. Study of the linear operator M. We define and study a linear operator that will be involved in the solution of the problem in a cylinder of length $2L$.

PROPOSITION 8. *Let $M_\pm$ be the operators from $\mathcal{R}(P)$ to $\mathcal{R}(P)$, and let M be given by*

$$M_\pm = P e^{\mp 2L\mathcal{A}} B_\pm \quad \text{and} \quad M = \begin{pmatrix} 0 & M_+ \\ M_- & 0 \end{pmatrix}.$$

Then the following hold:

(a) *There exists a constant C such that*

$$\|M\| \le C e^{-2\lambda L}, \text{ where } \lambda = \min(\lambda_1, -\lambda_{-1}).$$

As a consequence $\|M\| < 1$ for sufficiently large L.

(b) *If the constants are controlled, then for L positive sufficiently small, $\|M^2\| < 1$.*

(c) *It follows that $\mathrm{Id} + M$ is invertible on $\mathcal{R}(P) \times \mathcal{R}(P)$ for large L and for small positive L.*

Proof. (a) Since $M_+ = P e^{-2L\mathcal{A}} B_+$ we have

$$\|M_+\| \le \|B_+\| e^{-2L\lambda_1}.$$

A similar upper bound for M_- gives the result.

(b) Define

$$J(L) = \sup_{\|(\phi_1,\phi_2)\|=1} \|M^2(\phi_1, \phi_2)\| < 1.$$

Since $J(0) = 1$, it is sufficient to prove that $J'(0) < 0$. Note that

$$M^2 = \begin{pmatrix} M_+ M_- & 0 \\ 0 & M_- M_+ \end{pmatrix}.$$

Let us fix $\phi \in \mathcal{R}(P)$ and define $j(L) = \|M_+(L) M_-(L)\phi\|^2$. Then $j(0) = \|\phi\|^2$, and it remains to prove that $j'(0) \le -C\|\phi\|^2$ with a positive constant C independent of ϕ.

The derivative of j is

$$\begin{aligned} j'(L) &= (M_+'(L)M_-(L)\phi + M_+(L)M_-'(L)\phi | M_+(L)M_-(L)\phi) \\ &= (-2P\mathcal{A}e^{-2L\mathcal{A}}B_+Pe^{2L\mathcal{A}}B_-\phi + 2Pe^{-2L\mathcal{A}}B_+P\mathcal{A}e^{2L\mathcal{A}}B_-\phi | Pe^{-2L\mathcal{A}}B_+Pe^{2L\mathcal{A}}B_-\phi), \end{aligned}$$

hence

$$j'(0) = -2(P\mathcal{A}B_+\phi|\phi) + 2(P\mathcal{A}B_-\phi|\phi).$$

But since $P\phi = \phi$, $PB_+P = P$, and $P\mathcal{A} = \mathcal{A} + P\mathcal{A}P - \mathcal{A}P$ we have

$$\begin{aligned} (P\mathcal{A}B_+\phi|\phi) &= (P\mathcal{A}B_+\phi|B_+\phi) \\ &= ((\mathcal{A} + P\mathcal{A}P - \mathcal{A}P)B_+\phi|B_+\phi) \\ &= (\mathcal{A}B_+\phi|B_+\phi) + (P\mathcal{A}PB_+\phi|B_+\phi) - (\mathcal{A}PB_+\phi|B_+\phi) \\ &= (\mathcal{A}B_+\phi|B_+\phi) + (\mathcal{A}\phi|\phi) - (\mathcal{A}\phi|B_+\phi). \end{aligned}$$

This proves that

$$(P\mathcal{A}B_+\phi|\phi) = (\mathcal{A}\phi|B_+\phi) = \frac{1}{2}\left((\mathcal{A}B_+\phi|B_+\phi) + (\mathcal{A}\phi|\phi)\right).$$

Similarly we obtain that

$$(P\mathcal{A}B_-\phi|\phi) = \frac{1}{2}\left((\mathcal{A}B_-\phi|B_-\phi) + (\mathcal{A}\phi|\phi)\right).$$

As a summary we find that

$$\begin{aligned} j'(0) &= -(\mathcal{A}B_+\phi|B_+\phi) + (\mathcal{A}B_-\phi|B_-\phi) \\ &< \lambda_{-1}\|B_-\phi\|^2 - \lambda_1\|B_+\phi\|^2 < (\lambda_{-1} - \lambda_1)\|\phi\|^2. \end{aligned}$$

(c) The operator $\mathrm{Id} + M$ is invertible for large L by (a). Note that $M_\pm$ as endomorphism of $\mathcal{R}(P)$ are compact for $L > 0$ and equal to identity for $L = 0$. As a result $\mathrm{Id} + M$ is invertible for small $L > 0$ if and only if there is no eigenvector associated to the value -1. A sufficient condition for invertibility is then that M^2 does not admit 1 as eigenvalue, which is proved in (b) for L sufficiently small. ∎

5.2. The Dirichlet case. The different cases for Dirichlet Inlet/Outlet condition are summarized in the following.

PROPOSITION 9. *The Dirichlet Inlet/Outlet conditions* $T|_{z=-L} = T_{-L}$ *and* $T|_{z=L} = T_{+L}$ *are equivalent to the linear system*

$$ZX = b,$$

where Z, X, *and* b *are defined depending on the* (LBC)*; see Table* 5.1, *where we recall that* $\widetilde{\Phi} = (\Delta_\sigma^{-1}h, 0)$ *and we define* $\phi_{\pm L} = (T_\pm, 0)$, $u_- = \Phi - PB_-\Phi$, *and* $u_+ = \Phi - PB_+\Phi$.

Moreover, for sufficiently large L *this system is invertible.*

Note 1: Thanks to the Lax–Milgram theorem in 3D (see section 1.3), we know beforehand that there exists a unique solution to the system $ZX = b$.

Note 2: In the case when the constants are not controlled and $\int_\Omega h < 0$ it suffices to change the sign of z, or equivalently to replace the $_-$ by $_+$.

TABLE 5.1

	Constants controlled	Constants not controlled	
		nonbalanced ($\int_\Omega h > 0$)	balanced
Z	$\begin{pmatrix} \mathrm{Id} & M_+ \\ M_- & \mathrm{Id} \end{pmatrix}$	$\begin{pmatrix} \mathrm{Id} & M_+ & \Phi \\ M_- & \mathrm{Id} & \Phi \\ 0 & M_+^\star u_-^T & (\Phi\|u_-) \end{pmatrix}$	$\begin{pmatrix} \mathrm{Id} & M_+ & \Phi & -L\Phi+\tilde{\Phi} \\ M_- & \mathrm{Id} & \Phi & L\Phi+\tilde{\Phi} \\ 0 & M_+^\star u_-^T & (\Phi\|u_-) & (-L\Phi+\tilde{\Phi}\|u_-) \\ M_-^\star u_+^T & 0 & (\Phi\|u_+) & (L\Phi+\tilde{\Phi}\|u_+) \end{pmatrix}$
X	$\begin{pmatrix} P\pi_-\psi \\ P\pi_+\psi \end{pmatrix}$	$\begin{pmatrix} P\pi_-\psi \\ P\pi_+\psi \\ c_1 \end{pmatrix}$	$\begin{pmatrix} P\pi_-\psi \\ P\pi_+\psi \\ c_1 \\ c_2 \end{pmatrix}$
b	$\begin{pmatrix} \phi_{-L} \\ \phi_{+L} \end{pmatrix}$	$\begin{pmatrix} \phi_{-L} \\ \phi_{+L} \\ (\phi_{-L}\|u_-) \end{pmatrix}$	$\begin{pmatrix} \phi_{-L} \\ \phi_{+L} \\ (\phi_{-L}\|u_-) \\ (\phi_L\|u_+) \end{pmatrix}$

Proof. The (I/OBC) are equivalent to the following:

$$\text{(5.2)} \qquad \begin{cases} P\pi_-\psi = P\theta_- & \text{with } \theta_- = \phi_{-L} - e^{-2L\mathcal{A}}\pi_+\psi - c_1\Phi - c_2(-L\Phi+\tilde{\Phi}), \\ P\pi_+\psi = P\theta_+ & \text{with } \theta_+ = \phi_{+L} - e^{2L\mathcal{A}}\pi_-\psi - c_1\Phi - c_2(L\Phi+\tilde{\Phi}). \end{cases}$$

Combining

$$Pe^{-2L\mathcal{A}}\pi_+\psi = Pe^{-2L\mathcal{A}}B_+P\pi_+\psi = M_+P\pi_+\psi$$

and the similar version when the roles of $+$ and $-$ are interchanged with (5.2), we obtain the first two rows of the matrix Z.

(a) When the constants are controlled, $c_1 = c_2 = 0$ and (5.2) reads $ZX = b$.

(b) When the constants are not controlled and $\int_\Omega h > 0$, then $c_2 = 0$. Theorem 5 requires an additional compatibility condition to solve the first equation. This condition reads

$$(\Phi - PB_-\Phi|\theta_-) = 0,$$

which is the additional equation in the system $ZX = b$.

(c) In the balanced case, after the change of variable $\widetilde{\psi} = \pi_-\psi$, the first equation in (5.2), $P\pi_-\psi = P\theta_-$, is equivalent to

$$\begin{cases} P\widetilde{\psi} = P\theta_- \\ \pi_+\widetilde{\psi} = 0 \end{cases} \qquad \text{and} \quad (\widetilde{\psi}|\phi_0) = 0.$$

Theorem 5 gives an explicit expression for the solution of the system on the left, and the condition on the right becomes $(\theta_-|\Phi - PB_-\Phi) = 0$, which is the third row of the system $ZX = b$. The last row is obtained using the second equation in (5.2).

When L becomes large, M_+ and M_- are exponentially small, and in each case the matrix Z is asymptotic to an invertible matrix. The sole nonobvious case is the balanced case, where one can observe that the 2×2 lower right block is asymptotically equivalent to $\left(\begin{smallmatrix} (\Phi|u_-) & (-L\Phi|u_-) \\ (\Phi|u_+) & (L\Phi|u_+) \end{smallmatrix}\right)$, which has a determinant $2L(\Phi|u_-)(\Phi|u_+) \neq 0$. When L is large, Z can be rewritten as $Z = A + B$ with B small and A easily inverted. One can use a Neumann series strategy to solve $Zx = b$. ∎

TABLE 5.2

	Constants controlled	Constants not controlled	
		nonbalanced ($\int_\Omega h > 0$)	balanced
Z	$\begin{pmatrix} \mathrm{Id} & M_+ \\ M_- & \mathrm{Id} \end{pmatrix}$	$\begin{pmatrix} \mathrm{Id} & M_+ \\ M_- & \mathrm{Id} \\ 0 & M_+^\star u_-^T \end{pmatrix}$	$\begin{pmatrix} \mathrm{Id} & M_+ & \Phi \\ M_- & \mathrm{Id} & \Phi \\ 0 & M_+^\star u_-^T & (\Phi\|u_-) \\ M_-^\star u_+^T & 0 & (\Phi\|u_+) \end{pmatrix}$
X	$\begin{pmatrix} P\mathcal{A}\pi_-\psi \\ P\mathcal{A}\pi_+\psi \end{pmatrix}$	$\begin{pmatrix} P\mathcal{A}\pi_-\psi \\ P\mathcal{A}\pi_+\psi \end{pmatrix}$	$\begin{pmatrix} P\mathcal{A}\pi_-\psi \\ P\mathcal{A}\pi_+\psi \\ c_2 \end{pmatrix}$
b	$\begin{pmatrix} \phi_{-L} \\ \phi_{+L} \end{pmatrix}$	$\begin{pmatrix} \phi_{-L} \\ \phi_{+L} \\ (\phi_{-L}\|u_-) \end{pmatrix}$	$\begin{pmatrix} \phi_{-L} \\ \phi_{+L} \\ (\phi_{-L}\|u_-) \\ (\phi_L\|u_+) \end{pmatrix}$

5.3. The Neumann Inlet/Outlet case. The different cases for the Neumann Inlet/Outlet conditions are summarized in the following.

PROPOSITION 10. *The Neumann Inlet/Outlet conditions* $\partial_z T|_{z=-L} = S_{-L}$ *and* $\partial_z T|_{z=L} = S_{+L}$ *are equivalent to the linear system*

$$ZX = b,$$

where Z, X, *and* b *are defined depending on the* (LBC)*; see Table* 5.2*, where we define* $\phi_{\pm L} = (S_\pm, 0)$, $u_- = \Phi - PB_-\Phi$, *and* $u_+ = \Phi - PB_+\Phi$.

Note 1: When the constants are not controlled the value of c_1 is arbitrary. In these cases the linear systems are rectangular and the existence of the solution depends on a compatibility condition that expresses that b is in the range of Z.

Note 2: Once the quantities $P\mathcal{A}\pi_\pm\psi$ are known, then the $(\mathcal{A}\psi|\phi_i)$ for $i > 0$ and $i < 0$ can be computed as explained above, and $(\psi|\phi_i)$ is obtained by dividing by λ_i.

Proof. The (I/OBC) are equivalent to the following:

(5.3)
$$\begin{cases} P\mathcal{A}\pi_-\psi = P\theta_- & \text{with } \theta_- = \phi_{-L} - e^{-2L\mathcal{A}}\mathcal{A}\pi_+\psi - c_1\Phi - c_2(-L\Phi + \tilde{\Phi}), \\ P\mathcal{A}\pi_+\psi = P\theta_+ & \text{with } \theta_+ = \phi_{+L} - e^{2L\mathcal{A}}\mathcal{A}\pi_-\psi - c_1\Phi - c_2(L\Phi + \tilde{\Phi}). \end{cases}$$

A discussion similar to the Dirichlet case leads to the result. ∎

6. Numerical tests.

6.1. First test case: A domain of finite length. The section of the domain of the first test case is the square $\Omega = [-5, 5]^2$ with a circular fluid subdomain of radius 2 centered at the origin. The velocity and eigenvalues of the operator $\mathcal{A}$ are computed with $P1$ finite element methods on the mesh of Figure 6.1. The velocity has a parabolic profile (Poiseuille flow) with prescribed total flow $Q \in \{1, 10, 100, 1000\}$. The lateral boundary conditions are of Robin type with parameter a. The thermal conductivities are equal to $c = \sigma = 1$. In total 100 eigenvalues/eigenvectors of $\mathcal{A}$ are computed.

We first set $Q = 10$ and vary the Robin parameter a. When $a = 0$, one retrieves the Neumann case, and when $a = +\infty$, one retrieves the Dirichlet case. In order to emphasize this fact we plot in Figure 6.2 (left) the eigenvalues of smallest magnitude for different values of a. We also plot with dots the eigenvalues associated to the

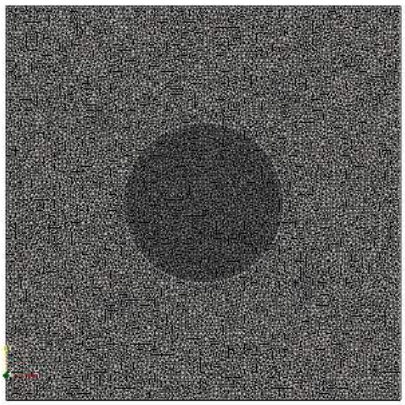


FIG. 6.1. *The mesh for the first test case is composed of* 13589 *vertices and* 26776 *triangles. The solid domain is in white and the fluid domain in gray.*

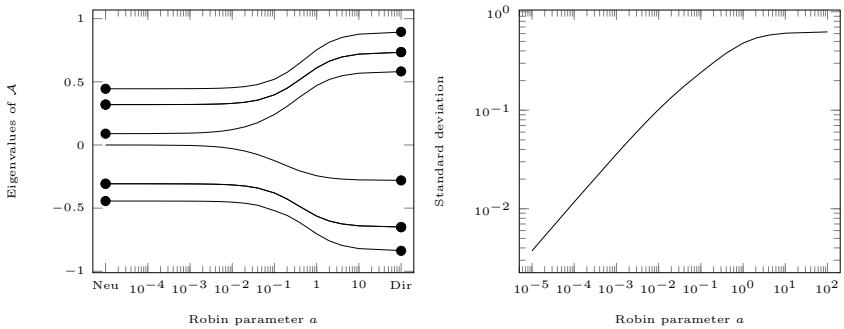


FIG. 6.2. *Left: evolution of the eigenvalues of $\mathcal{A}$ of smallest magnitude for varying parameter of the Robin lateral boundary condition. Eigenvalues for the Neumann (resp., Dirichlet) boundary conditions are shown as bullets on the left (resp., right) of the curves. Right: relative L^2 difference between the eigenvector with largest negative eigenvalue and its mean as the Robin parameter varies.*

Neumann problem (on the left of the curves) and the one associated to the Dirichlet case (on the right of the curves). The smooth transition from Neumann to Dirichlet as the Robin parameter varies is striking except for the fact that there exists an eigenvalue that goes to zero as a goes to zero even if the Neumann problem does not have zero as an eigenvalue. We claim that this behavior is consistent with the theory. First, zero is not an eigenvalue of the Neumann case since the total flow is nonzero (hence we are not in a balanced case even if the constants are not controlled). Second, we remark that the zero eigenvalue is the limit of a negative eigenvalue. Remember from Proposition 6 that it is always possible to decompose a temperature field on the set of negative eigenvectors in the Robin case (part a), but for the Neumann case it is necessary to add a constant (part b). In other words, in the Neumann case the constant must be added to the negative eigenvectors to obtain a Hilbert basis of H,

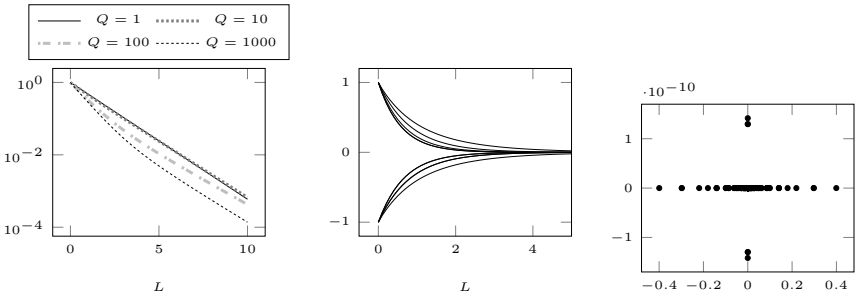


FIG. 6.3. *Eigenvalues of M for a Robin test case with $a = 1$. Left: evolution of the spectral radius for different values of the total flow Q. Center: the evolution of the five largest positive and first smallest negative eigenvalue for a total flow of $Q = 20$. Right: eigenvalues in the complex plane for $L = 1$ and $Q = 20$.*

while the set of positive eigenvectors form a Hilbert basis on their own. This explains why the constant emerges as the limit of a negative eigenvector; see Figure 6.2 (right), where the convergence of the eigenvector to the constant is numerically demonstrated.

In a second parametric study, we fix $a = 1$ and let both Q and L vary. First we plot the spectral radius of the matrix M defined in Proposition 8 versus the exchanger length L for the different values of the total flow Q in Figure 6.3 (left). Figure 6.3 (center) shows the evolution of the 5 smallest positive and 5 largest negative eigenvalues of M for a fixed total flow $Q = 20$. This test case shows that, apart from the case $L = 0$, the spectral radius of the matrix M is always smaller than one, so that the matrix $\mathrm{Id} + M$ is indeed always invertible. The exponential decrease for large L and the decrease at the origin follow from Proposition 8. Moreover, since the spectral radius of M is strictly smaller than one, a Neumann series strategy to solve

$$(\mathrm{Id} + M)^{-1} b = \sum_k (-M)^k b$$

is legitimate. In Figure 6.3 (right), the whole spectrum of M is shown in the complex plane. Although the spectrum seems real, we do not have mathematical proof of this fact.

6.2. Second test case: A periodic exchanger. The second test case consists of a heat exchanger with periodic boundary conditions. The whole device consists of one solid exchanger through which pass four tubes containing fluids. A cut along the middle of the exchanger is shown in Figure 6.4 (left), where the sign of the velocity of the fluid in the inner tubes is displayed. The fluids are assumed to obey a Poiseuille flow; the velocities are then quadratic in the radial coordinates of their corresponding tubes. The length of the exchanger is denoted L, the section of the exchanger is the square $[-4, 4]^2$, the radii of the inner tubes are fixed to 1, and the distance of the center of the inner tubes to the center of the exchanger is $\sqrt{2^2 + 2^2}$. The conductivity in both the fluid and solid parts is set to 1. The temperature is fixed for the four tubes with incoming flow (two at each side) on the exchanger: the warm temperature is set to $+1$, and the cold temperature to -1; see Figure 6.4 (left). In what follows, Q denotes the total flow of fluid in one tube. We glue together the different Graetz problems using the methodology developed in [4].

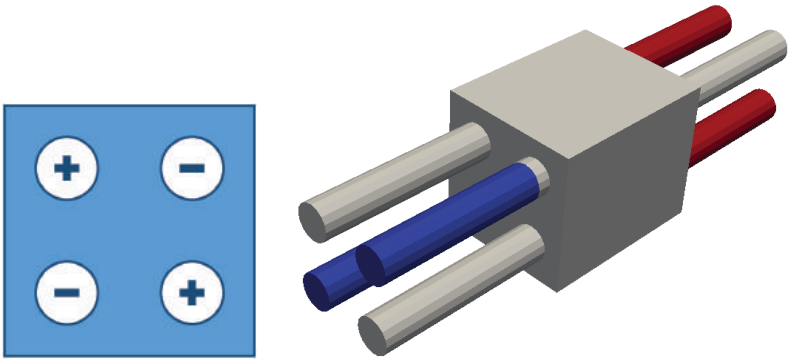


FIG. 6.4. *Geometry of the periodic exchanger. Left: a cut inside the exchanger with the sign of the fluid velocities. Right: a 3D representation of the exchanger. The tubes where the temperature is set at* ∞ *are colored according to their temperature.*

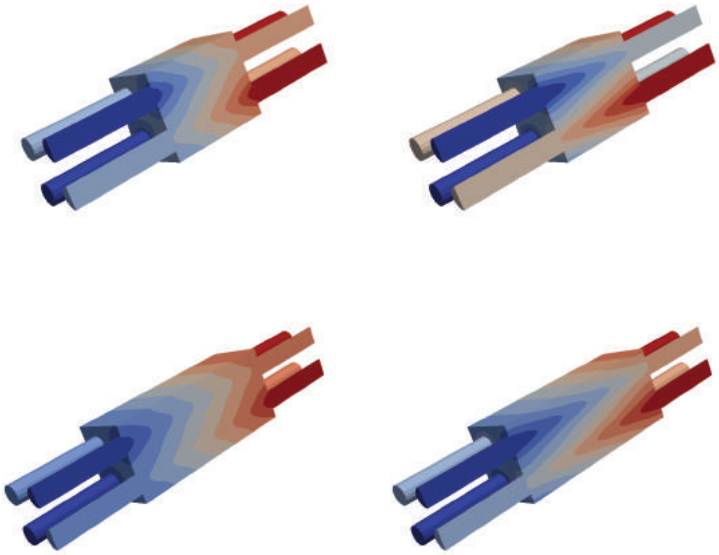


FIG. 6.5. *Four different solutions of the periodic exchanger, with different length and total flow. The length of the exchanger is set to* $L = 10$ *on top and* $L = 20$ *on bottom. The total flow is set to to* $Q = 10$ *on left and* $Q = 30$ *on right.*

In Figure 6.5, four solutions are shown for different values of the length L and the flow Q.

Figure 6.6 displays the efficiency and the total exchange for different values of

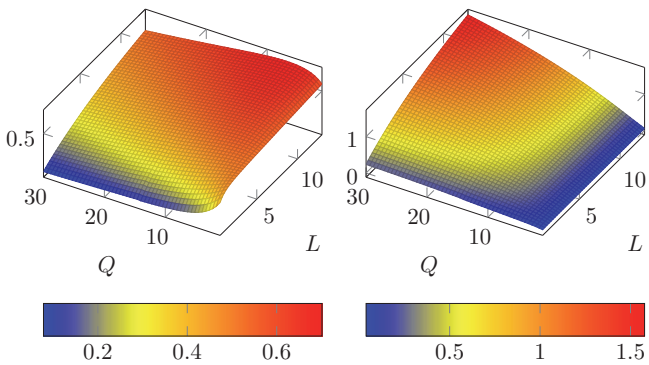


FIG. 6.6. *Values of the efficiency (left) and the exchange (right) for different values of the length L and the total flow Q of the periodic exchanger. L ranges from 0.5 to 13, and Q ranges from 1 to 30. Each direction has been sampled 50 times, for a total of 2500 exchanger computations.*

Q and L. For a tube containing fluid whose velocity is positive (resp., negative), the temperature at $-\infty$ (resp., $+\infty$) is set to 1 (resp., -1), and the efficiency of the exchanger is then defined by $-T_{+\infty}/T_{-\infty}$ (resp., $-T_{-\infty}/T_{+\infty}$), where $T_{\pm\infty}$ is the temperature at infinity. This efficiency is between -1 and 1. The exchange is simply the total amount of heat exchanged and is equal to Q times the efficiency. The aim of this test case is to document the fact that our method is able to deal with any boundary conditions and type of exchanger. It is well suited for parametric studies.

7. Conclusion. In the present work we have proposed a general framework dedicated to the resolution of the generalized Graetz problem in arbitrary geometry, involving any type of boundary conditions. The main novelty is the introduction of an insulating boundary condition (Neumann or periodic) that allows one to model realistic heat exchangers. Our study highlighted a special case that we call the balanced case, when $\int_\Omega h = 0$ (together with Neumann or periodic boundary condition) where the solution is different than in the general case. We have also proposed a number of numerical illustrations in various test cases.

Appendix A. Technical lemmas. We prove here results that will be used in what follows.

LEMMA 11. *For each $\phi = (u, s) \in \mathcal{D}(\mathcal{A}), \tilde{\phi} = (\tilde{u}, \tilde{s}) \in \mathcal{H}$, we have*

$$(\mathrm{Id} - P)\mathcal{A}(\mathrm{Id} - P)\phi = 0, \tag{A.1}$$

$$(P\mathcal{A}P\phi|\tilde{\phi})_{\mathcal{H}} = \int_\Omega hu\tilde{u} \leq \|h\|_{L^\infty(\Omega)}\|P\phi\|_{\mathcal{H}}\|P\tilde{\phi}\|_{\mathcal{H}}. \tag{A.2}$$

Proof. This results from elementary calculations using the definition of $\mathcal{A}$ (2.1), and the definition of P (3.1). ∎

LEMMA 12. *Let $\Phi = (1, 0) \in \mathcal{H}$. Let $\phi \in \mathcal{D}(A^{-1/2})$ such that $P\phi = \phi$. Then*

$$(\mathcal{A}^{-1}\phi|\phi)_{\mathcal{H}} = \left(\int_\Omega h\right)^{-1} (\phi|\Phi)^2_{\mathcal{H}} \text{ in the nonbalanced case,}$$

$$(\mathcal{A}^{-1}\phi|\phi)_{\mathcal{H}} = 0 \text{ in the “balanced” or “constant controlled” case.}$$

Proof. The expression of $\mathcal{A}^{-1}$ is given in section 2.1 for the various cases. Let $\phi \in \mathcal{D}(\mathcal{A}^{-1/2})$ such that $P\phi = \phi$; hence there exists $u \in L^2(\Omega)$ such that $\phi = (u, 0)$. If the constants are controlled, then $\mathcal{A}^{-1}\phi = (0, \Delta_\sigma^{-1}(-cu))$ and $(\mathcal{A}^{-1}\phi|\phi)_{\mathcal{H}} = 0$. If the constants are not controlled and $\int_\Omega h \neq 0$, then

$$\mathcal{A}^{-1}\phi = (k, \Delta_\sigma^{-1}(-cu + hk)) \text{ with } k \int_\Omega h = \int_\Omega cu,$$

and hence

$$(\mathcal{A}^{-1}\phi|\phi)_{\mathcal{H}} = \int_\Omega cku = \left(\int_\Omega h\right)^{-1} \left(\int_\Omega cu\right)^2.$$

Finally, in the balanced case, since $\phi \in \mathcal{D}(\mathcal{A}^{-1/2})$, then $(\phi|\phi_0)_{\mathcal{H}} = 0$ and $\mathcal{A}^{-1}\phi = (0, \Delta_\sigma^{-1}(-cu)) + k\phi_0$ exist and $(\mathcal{A}^{-1}\phi|\phi)_{\mathcal{H}} = 0$. ∎

LEMMA 13. *In the balanced case,*

$$(\Phi - PB_-\Phi|\Phi)_{\mathcal{H}} \neq 0.$$

Proof. Suppose the contrary and set $\theta = \Phi - PB_-\Phi$; we have $P\theta = \theta$, and by definition of B_-, we have $\pi_-\theta = 0$. Moreover, we have

$$\begin{aligned}(\theta|\phi_0)_{\mathcal{H}} &= (\Phi - PB_-\Phi|\phi_0)_{\mathcal{H}} = (P\Phi - PB_-\Phi|\phi_0)_{\mathcal{H}} = (\Phi - B_-\Phi|P\phi_0)_{\mathcal{H}} \\ &= (\Phi - B_-\Phi|P\Phi)_{\mathcal{H}} = (\Phi - PB_-\Phi|\Phi)_{\mathcal{H}} = 0.\end{aligned}$$

Lemma 12 ensures that $\mathcal{A}^{-1}\theta$ exists and that

$$(\mathcal{A}^{-1}\theta|\theta)_{\mathcal{H}} = 0.$$

Since θ belongs to $\mathcal{R}(\pi_+)$ and all the eigenvalues of $\mathcal{A}$ are positive on this space, this implies that $\theta = 0$ and then $\Phi = PB_-\Phi$. Hence there exists s such that $B_-\Phi = \big(1, s\big)$ and

$$(\mathcal{A}B_-\Phi|B_-\Phi)_{\mathcal{H}} = \big((h - \Delta_\sigma s, 0) \mid (1, s)\big)_{\mathcal{H}} = 0.$$

But $B_-\Phi$ belongs to $\mathcal{R}(\pi_-)$, and since all the eigenvalues of $\mathcal{A}$ are negative on $\mathcal{R}(\pi_-)$, $(\mathcal{A}B_-\Phi|B_-\Phi)_{\mathcal{H}} = 0$ implies that $B_-\Phi = 0$, which is in violation of $\Phi = PB_-\Phi$. Hence $(\Phi - B_-\Phi|\Phi)_{\mathcal{H}} \neq 0$. ∎

Appendix B. Proof of Theorem 4. Let $M \in \mathbb{N}^*$ and denote for short $\pi = \pi_{[\![-M,-1]\!]}$. The operator $\pi P\pi$ is a symmetric operator in a finite-dimensional space, and hence it is diagonalizable in an orthonormal basis. The first step is to prove that this operator is positive definite with a lower bound on its eigenvalues that is independent of M. Let ρ be an eigenvalue of $\pi P\pi$ and $\mathbf{v}$ an associated normalized eigenvector: $\pi P\pi\mathbf{v} = \rho\mathbf{v}$, $(\mathbf{v}|\mathbf{v})_{\mathcal{H}} = 1$ and $\pi\mathbf{v} = \mathbf{v}$. Since

$$\rho = (\pi P\pi\mathbf{v}|\mathbf{v})_{\mathcal{H}} = (P\pi\mathbf{v}|\pi\mathbf{v})_{\mathcal{H}} = (P\pi\mathbf{v}|P\pi\mathbf{v})_{\mathcal{H}} = \|P\mathbf{v}\|_{\mathcal{H}}^2 \leq \|\mathbf{v}\|_{\mathcal{H}}^2 = 1,$$

then $0 \leq \rho \leq 1$. Using (A.2) gives

$$|(P\mathcal{A}P\mathbf{v}|\mathbf{v})_{\mathcal{H}}| \leq \|h\|_{L^\infty(\Omega)}\|P\mathbf{v}\|_{\mathcal{H}}^2.$$

It follows from (A.1) that $((\mathrm{Id} - P)\mathcal{A}(\mathrm{Id} - P)\mathbf{v}|\mathbf{v})_{\mathcal{H}} = 0$ and $\pi\mathcal{A} = \mathcal{A}\pi$, and we have

$$(P\mathcal{A}P\mathbf{v}|\mathbf{v})_{\mathcal{H}} = (2\rho - 1)(\mathcal{A}\mathbf{v}|\mathbf{v})_{\mathcal{H}}.$$

Since $|(\mathcal{A}\mathbf{v}|\mathbf{v})_{\mathcal{H}}| = |\sum_{i\in I}\lambda_i \frac{(\mathbf{v}|\phi_i)^2_{\mathcal{H}}}{\|\phi_i\|^2_{\mathcal{H}}}| \geq |\lambda_{-1}|\|\mathbf{v}\|^2_{\mathcal{H}} = |\lambda_{-1}|$, we have

$$|\lambda_{-1}(2\rho-1)| \leq \|h\|_{L^\infty(\Omega)}\|P\mathbf{v}\|^2_{\mathcal{H}} = \|h\|_{L^\infty(\Omega)}\rho. \tag{B.1}$$

This in turn implies that $\rho \geq \frac{|\lambda_{-1}|}{2|\lambda_{-1}|+\|h\|_{L^\infty(\Omega)}}$, and hence there exists C independent of M such that

$$(\pi P\pi\phi|\phi)_{\mathcal{H}} \geq C\|\pi\phi\|_{\mathcal{H}} \quad \forall\phi\in\mathcal{H}(\mathcal{A}). \tag{B.2}$$

Since $\pi_-\phi$ is the strong $\mathcal{H}$-limit of $\pi\phi$ as M goes to infinity and the constant C does not depend on M, passing to the limit, we recover (B.2) with π replaced by π_-. The Lax–Milgram theorem applies and $\pi_-P\pi_-$ is a bijection from $\mathcal{R}(\pi_-)$ onto $\mathcal{R}(\pi_-)$ with a continuous inverse bounded by a constant in the $\mathcal{H}$-norm.

We turn our interest to the bound in the 1/2-norm of B_-. Let $\phi\in\mathcal{D}(\mathcal{A}^{1/2})$, and for any $M\in\mathbb{N}^*$ denote $\pi = \pi_{[-M,-1]}$, and let $\mathbf{v} = \pi B_-\phi$. We have $\pi\mathbf{v} = \mathbf{v}$ and $\mathbf{v}\in\mathcal{D}(\mathcal{A})$. Recalling (A.1) and $\pi\mathcal{A} = \mathcal{A}\pi$, we have

$$(P\mathcal{A}P\mathbf{v}|\mathbf{v})_{\mathcal{H}} = ((\mathcal{A}P + P\mathcal{A} - \mathcal{A})\mathbf{v}|\mathbf{v})_{\mathcal{H}} = 2(P\mathbf{v},\mathcal{A}\mathbf{v})_{\mathcal{H}} - (\mathcal{A}\mathbf{v},\mathbf{v})_{\mathcal{H}}.$$

Hence, since $\pi\mathbf{v} = \mathbf{v}$ and π is a projection on negative eigenvalues of $\mathcal{A}$ only, then $\|\mathbf{v}\|^2_{1/2} = -(\mathcal{A}\mathbf{v}|\mathbf{v})_{\mathcal{H}}$ and

$$\|\mathbf{v}\|^2_{1/2} = (P\mathcal{A}P\mathbf{v}|\mathbf{v})_{\mathcal{H}} - 2(P\mathbf{v}|\mathcal{A}\mathbf{v})_{\mathcal{H}} \leq \|h\|_{L^\infty(\Omega)}\|\mathbf{v}\|^2_{\mathcal{H}} + 2\|\pi P\mathbf{v}\|_{1/2}\|\mathbf{v}\|_{1/2}. \tag{B.3}$$

Using the bound on the $\mathcal{H}$-norm of B_-, we have

$$\|\mathbf{v}\|_{\mathcal{H}} = \|\pi B_-\phi\|_{\mathcal{H}} \leq C\|\pi_-\phi\|_{\mathcal{H}} \leq C\|\pi_-\phi\|_{1/2}. \tag{B.4}$$

We infer from (B.3) and (B.4) that $\|\mathbf{v}\|_{1/2} \leq C(\|\pi_-\phi\|_{1/2} + \|\pi P\mathbf{v}\|_{1/2})$. We let M go to infinity; then $\pi P\mathbf{v} = \pi P\pi B_-\phi$ goes to $\pi_-\phi$ and $\mathbf{v}$ goes to $B_-\phi$, and we obtain

$$\|B_-\phi\|_{1/2} \leq C\|\pi_-\phi\|_{1/2},$$

which finishes the proof.

Appendix C. Proof of Theorem 5.

First case: $\mathcal{K}(\mathcal{A}) = \{0\}$, i.e., every case but the balanced case. In this case the condition $\pi_+\psi = 0$ is then equivalent to $\psi = \pi_-\psi$.

After multiplication of (3.3) by $B_-\pi_-$, one obtains the following necessary condition for (3.3) to hold, which proves uniqueness:

$$\psi = B_-P\phi.$$

Denoting $\theta = PB_-P\phi - P\phi$, the question of the existence of the solution is reduced to determining under which condition $\theta = 0$.

We have $P\theta = \theta$, and Theorem 4 states that $\pi_-\theta = 0$. This implies that $\theta\in\mathcal{R}(\pi_+)$. The operator $\mathcal{A}^{-1}$ is symmetric positive definite on $\mathcal{R}(\pi_+)$ and induces the scalar product of the $-1/2$-norm. Lemma 12 states that if the constants are controlled, we have $(\mathcal{A}^{-1}\theta|\theta)_{\mathcal{H}} = 0$, and it follows that $\theta = 0$. This proves the result when the constants are controlled.

Assuming now that $\Gamma_D\cup\Gamma_R = \emptyset$ and $\int_\Omega h \neq 0$, Lemma 12 states that

$$\|\theta\|^2_{-1/2} = (\mathcal{A}^{-1}\theta|\theta)_{\mathcal{H}} = \left(\int_\Omega h\right)^{-1}(\theta|\Phi)^2_{\mathcal{H}}. \tag{C.1}$$

If $\int_\Omega h < 0$, the two terms have opposite signs, and hence both are zero. Then $\theta = 0$, and this proves the result for the case $\Gamma_D \cup \Gamma_R = \emptyset$ and $\int_\Omega h < 0$.

Let us assume now that $\int_\Omega h > 0$. Since changing the sign of λ amounts to studying the same problem where h is replaced by $-h$, we deduce from the case $\int_\Omega h < 0$ with $\phi = \Phi$ and the relation $P\Phi = \Phi$ that $PB_+\Phi = \Phi$. Since $\Phi \in \mathcal{D}(A^{1/2})$, it follows from Theorem 4 that $B_+\Phi \in \mathcal{D}(A^{1/2})$. Hence there exists an $s^\star \in H$ such that $B_+\Phi = (1, s^\star)$ and we have $\mathcal{A}B_+\Phi = (c^{-1}h - c^{-1}\operatorname{div}\sigma\nabla s^\star, 0)$. This proves

$$P\mathcal{A}B_+\Phi = \mathcal{A}B_+\Phi,$$

and a simple calculation proves that

$$(\mathcal{A}B_+\Phi|\Phi)_{\mathcal{H}} = (\mathcal{A}B_+\Phi|B_+\Phi)_{\mathcal{H}} = \int_\Omega h. \tag{C.2}$$

We then compute

$$(\Phi - PB_-\Phi|\mathcal{A}B_+\Phi)_{\mathcal{H}} = (\Phi - B_-\Phi|P\mathcal{A}B_+\Phi)_{\mathcal{H}} = (\Phi - B_-\Phi|\mathcal{A}B_+\Phi)_{\mathcal{H}} \underbrace{=}_{(\mathbf{1})} (\Phi|\mathcal{A}B_+\Phi)_{\mathcal{H}} \neq 0,$$

where the equality (**1**) is obtained by remarking that $\mathcal{A}B_+\Phi \in \mathcal{R}(\pi_+)$ and $B_-\Phi \in \mathcal{R}(\pi_-)$, which are orthogonal spaces. We then obtain $\Phi - PB_-\Phi \neq 0$.

It follows from (C.1) that

$$\begin{aligned}\|\theta\|^2_{-1/2} &= \left(\int_\Omega h\right)^{-1} (\theta|\Phi)^2_{\mathcal{H}} = \left(\int_\Omega h\right)^{-1} (\theta|PB_+\Phi)^2_{\mathcal{H}} \\ &= \left(\int_\Omega h\right)^{-1} (P\theta|B_+\Phi)^2_{\mathcal{H}} = \left(\int_\Omega h\right)^{-1} (\theta|B_+\Phi)^2_{\mathcal{H}}.\end{aligned}$$

Using that θ and $B_+\Phi$ belong to $\mathcal{R}(\pi_+)$ on which all the eigenvalues of $\mathcal{A}^{-1}$ are positive, the above equation implies

$$\|\theta\|^2_{-1/2} = \left(\int_\Omega h\right)^{-1} (\theta|\mathcal{A}B_+\Phi)^2_{-1/2}.$$

We recall that $\|\mathcal{A}B_+\Phi\|^2_{-1/2} = (\mathcal{A}B_+\Phi|B_+\Phi)_{\mathcal{H}} = \int_\Omega h$, and we obtain

$$\|\theta\|^2_{-1/2}\|\mathcal{A}B_+\Phi\|^2_{-1/2} = (\theta|\mathcal{A}B_+\Phi)^2_{-1/2},$$

which is an equality case in Cauchy–Schwarz inequality. This implies that θ and $\mathcal{A}B_+\Phi$ are collinear. Hence there exists some constant t such that

$$\theta = t\mathcal{A}B_+\Phi.$$

Performing the scalar product with Φ and using the fact that $(\mathcal{A}B_+\Phi|\Phi) \neq 0$, which follows from (C.2), we conclude that $t = 0$ (hence $\theta = 0$) if and only if $(\theta|\Phi) = 0$, which reads $(\phi|\Phi - PB_-\Phi)_{\mathcal{H}} = 0$.

Second case: $\mathcal{K}(\mathcal{A}) \neq \{0\}$, which is the balanced case. In the balanced case the kernel of $\mathcal{A}$ is $\mathbb{R}\phi_0$, where we recall from section 2.1 that $P\phi_0 = \Phi$. The condition $\pi_+\psi = 0$ is equivalent to the existence of $\alpha \in \mathbb{R}$ such that $\psi = \pi_-\psi + \alpha\phi_0$. The condition $P\psi = P\phi$ is thus equivalent to

$$P\phi = P\pi_-\psi + \alpha\Phi. \tag{C.3}$$

Necessary condition: After multiplying (C.3) by B_- one obtains

$$\pi_-\psi = B_-P\phi - \alpha B_-\Phi.$$

Replacing the expression of $\pi_-\psi$ in (C.3) yields the following necessary condition:

$$PB_-P\phi + \alpha\Phi - \alpha PB_-\Phi = P\phi,$$

which reads

$$\alpha(\Phi - PB_-\Phi) = P\phi - PB_-P\phi.$$

It follows from Lemma 13 that $(\Phi - PB_-\Phi|\Phi)_{\mathcal{H}} \neq 0$, and then it is necessary that

$$\alpha = \frac{(\Phi - PB_-\Phi|\phi)_{\mathcal{H}}}{(\Phi - PB_-\Phi|\Phi)_{\mathcal{H}}}.$$

ψ is uniquely determined by

$$\psi = B_-P\phi + \frac{(\Phi - PB_-\Phi|\phi)_{\mathcal{H}}}{(\Phi - PB_-\Phi|\Phi)_{\mathcal{H}}}(\phi_0 - B_-\Phi). \tag{C.4}$$

Conversely, if ψ is defined by (C.4), it is clear that $\pi_+\psi = 0$. Let $\theta = P\psi - P\psi$; it suffices to prove that $\theta = 0$ to ensure that ψ solves the problem.

$$(\theta|\phi_0)_{\mathcal{H}} = (\theta|\Phi)_{\mathcal{H}} = 0$$

by choice of α. A simple calculation shows that

$$\pi_-\theta = 0.$$

This proves that $\theta \in \mathcal{R}(\pi_+)$, where $\mathcal{A}^{-1}$ is a symmetric positive definite operator. It follows from Lemma 12 that $(\mathcal{A}^{-1}\theta|\theta)_{\mathcal{H}} = 0$ and hence $\theta = 0$. This finishes the proof.

REFERENCES

[1] C. Gostoli and A. Gatta, *Mass transfer in a hollow fiber dialyzer*, J. Membrane Sci., 6 (1980), pp. 133–148.

[2] A. Dorfman and Z. Renner, *Conjugate problems in convective heat transfer: Review*, Math. Probl. Eng., 2009 (2009), 927350.

[3] M. A. Ebadian and H. Y. Zhang, *An exact solution of extended Graetz problem with axial heat conduction*, Internat. J. Heat Mass Transfer, 32 (1989), pp. 1709–1717.

[4] J. Fehrenbach, F. de Gournay, C. Pierre, and F. Plouraboué, *The generalized Graetz problem in finite domains*, SIAM J. Appl. Math., 72 (2012), pp. 99–123, https://doi.org/10.1137/11082542X.

[5] L. Graetz, *Über die Wärmeleitungsfähigkeit von Flüssigkeiten*, Ann. Phys., 261 (1885), pp. 337–357.

[6] C.-D. Ho, H.-M. Yeh, and W.-S. Sheu, *An analytical study of heat and mass transfer through a parallel-plate channel with recycle*, Internat. J. Heat Mass Transfer, 41 (1998), pp. 2589–2599.

[7] J. Kragh, J. Rose, T. R. Nielsen, and S. Svendsen, *New counter flow heat exchanger designed for ventilation systems in cold climates*, Energy and Buildings, 39 (2007), pp. 1151–1158.

[8] H.-E. Jeong and J.-T. Jeong, *Extended Graetz problem including streamwise conduction and viscous dissipation in microchannel*, Internat. J. Heat Mass Transfer, 49 (2006), pp. 2151–2157.

[9] J. Lahjomri, A. Oubarra, and A. Alemany, *Heat transfer by laminar Hartmann flow in thermal entrance region with a step change in wall temperatures: The Graetz problem extended*, Internat. J. Heat Mass Transfer, 45 (2002), pp. 1127–1148.

[10] M. L. Michelsen and J. Villadsen, *The Graetz problem with axial heat conduction*, Internat. J. Heat Mass Transfer, 17 (1974), pp. 1391–1402.
[11] E. Papoutsakis, D. Ramkrishna, and H. C. Lim, *The extended Graetz problem with Dirichlet wall boundary conditions*, Appl. Sci. Res., 36 (1980), pp. 13–34.
[12] E. Papoutsakis, D. Ramkrishna, and H. C. Lim, *The extended Graetz problem with prescribed wall flux*, AIChE J., 26 (1980), pp. 779–787.
[13] E. Papoutsakis, D. Ramkrishna, and H.-C. Lim, *Conjugated Graetz problems. Part* 1: *General formalism and a class of solid-fluid problems*, Chem. Engrg. Sci., 36 (1981), pp. 1381–1391.
[14] E. Papoutsakis, D. Ramkrishna, and H.-C. Lim, *Conjugated Graetz problems. Part* 2: *Fluid-fluid problem*, Chem. Engrg. Sci., 36 (1981), pp. 1393–1399.
[15] C. Pierre and F. Plouraboué, *Numerical analysis of a new mixed formulation for eigenvalue convection-diffusion problems*, SIAM J. Appl. Math., 70 (2009), pp. 658–676, https://doi.org/10.1137/080736442.
[16] D. P. Sekulić and R. K. Shah, *Fundamentals of Heat Exchanger Design*, John Wiley and Sons, 2003.
[17] W. Rudin, *Functional Analysis*, Internat. Ser. Pure Appl. Math., McGraw-Hill, 1991.
[18] B. Weigand, M. Kanzamar, and H. Beer, *The extended Graetz problem with piecewise constant wall heat flux for pipe and channel flows*, Internat. J. Heat Mass Transfer, 44 (2001), pp. 3941–3952.