HAL
open science

Algorithmic solutions toward applications of compressed sensing for optical imaging

Yoann Le Montagner

To cite this version:

Yoann Le Montagner. Algorithmic solutions toward applications of compressed sensing for optical imaging. Signal and Image processing. Télécom ParisTech, 2013. English. NNT : 2013ENST0065 . tel-00950365v2

HAL Id: tel-00950365

https://pastel.hal.science/tel-00950365v2

Submitted on 24 Jul 2015

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.

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

TELECOM ParisTech

2013-ENST-0065

EDITE ED 130

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

Télécom ParisTech

Spécialité “ Traitement du signal et des images ”

présentée et soutenue publiquement par

Yoann LE MONTAGNER

le 12 novembre 2013

Solutions algorithmiques pour des applications d’acquisition parcimonieuse en bio-imagerie optique

Directeurs de thèse : **Elsa ANGELINI** et **Jean-Christophe OLIVO-MARIN**

Jury

Mme. Elsa ANGELINI, Maître de conférence, Télécom ParisTech	Co-directrice
Mme. Caroline CHAUX, Chargée de recherche, Aix-Marseille Université	Examinatrice
M. Rémi GRIBONVAL, Directeur de recherche, INRIA Rennes	Rapporteur
M. Jean-Christophe OLIVO-MARIN, Professeur, Institut Pasteur	Co-directeur
Mme. Béatrice PESQUET-POPESCU, Professeur, Télécom ParisTech	Examinatrice
M. Dimitri VAN DE VILLE, Professeur, Université de Genève / EPFL	Rapporteur

T H È S E

Télécom ParisTech
école de l’Institut Mines Télécom – membre de ParisTech
46, rue Barrault – 75634 Paris Cedex 13 – Tél. + 33 (0)1 45 81 77 77 – www.telecom-paristech.fr

Télécom ParisTech – Institut Mines-Télécom
École doctorale EDITE de Paris – ED 130
Mémoire enregistré sous le numéro 2013-ENST-0065

T H È S E

pour obtenir le grade de

Docteur de Télécom ParisTech
Spécialité « Traitement du signal et des images »

présentée et soutenue publiquement par

Yoann Le Montagner

le 12 novembre 2013

Algorithmic solutions toward applications of compressed sensing for optical imaging

Thèse dirigée par Elsa Angelini et Jean-Christophe Olivo-Marin,
préparée à l'Institut Pasteur au sein de l'Unité d'Analyse d'Images Quantitative

JURY

Elsa Angelini	Télécom ParisTech	Co-directrice
Caroline Chaux	Aix-Marseille Université	Examinatrice
Rémi Gribonval	INRIA Rennes	Rapporteur
Jean-Christophe Olivo-Marin	Institut Pasteur	Co-directeur
Béatrice Pesquet-Popescu	Télécom ParisTech	Examinatrice
Dimitri Van De Ville	Université de Genève/EPFL	Rapporteur

Télécom ParisTech – Institut Mines-Télécom
École doctorale EDITE de Paris – ED 130
Mémoire enregistré sous le numéro 2013-ENST-0065

T H È S E

pour obtenir le grade de

Docteur de Télécom ParisTech
Spécialité « Traitement du signal et des images »

présentée et soutenue publiquement par

Yoann Le Montagner

le 12 novembre 2013

Solutions algorithmiques pour des applications d'acquisition parcimonieuse en bio-imagerie optique

Thèse dirigée par Elsa Angelini et Jean-Christophe Olivo-Marin,
préparée à l'Institut Pasteur au sein de l'Unité d'Analyse d'Images Quantitative

JURY

Elsa Angelini	Télécom ParisTech	Co-directrice
Caroline Chaux	Aix-Marseille Université	Examinatrice
Rémi Gribonval	INRIA Rennes	Rapporteur
Jean-Christophe Olivo-Marin	Institut Pasteur	Co-directeur
Béatrice Pesquet-Popescu	Télécom ParisTech	Examinatrice
Dimitri Van De Ville	Université de Genève/EPFL	Rapporteur

Table of contents

Table of contents **5**

Acknowledgments **9**

Abstract **11**

Résumé **13**

Notations **15**

General introduction **19**

I Introduction on CS theory **21**
- I.1 A few definitions . . . 22
  - I.1.1 Signals and images . . . 22
  - I.1.2 Sparsity and compressibility . . . 23
- I.2 Compressed sensing theoretical results . . . 25
  - I.2.1 Recovering sparse data from incomplete measurements . . . 25
  - I.2.2 Restricted isometry property . . . 26
  - I.2.3 Partial unitary transforms . . . 27
  - I.2.4 Sparse representations and dictionaries . . . 28
  - I.2.5 Block sparsity and total variation . . . 30
- I.3 Application of compressed sensing for imaging devices . . . 31
  - I.3.1 Magnetic resonance imaging . . . 31
  - I.3.2 Digital holography . . . 32
  - I.3.3 Single-pixel camera . . . 34
  - I.3.4 Schlieren deflectometry . . . 35

II Reconstruction through convex optimization **37**
- II.1 CS reconstruction formulations . . . 38
  - II.1.1 Convex optimization formulations: classical form, BPDN and LASSO 38
  - II.1.2 Alternative approach: orthogonal matching pursuit . . . 40
- II.2 Convex optimization algorithms . . . 41
  - II.2.1 SOCP methods . . . 41
  - II.2.2 NESTA . . . 43
  - II.2.3 RecPF . . . 44

II.2.4 SPGL1 45
II.3 Compared performance 46
II.3.1 Methodology 46
II.3.2 Simulation results and reconstruction time 48
II.4 Conclusion 48

III Sampling parameters in the Fourier space 51
III.1 Sampling strategies in the Fourier space 52
III.1.1 Introduction 52
III.1.2 Existing Fourier sampling strategies 52
III.2 Numerical evaluation of an optimal sampling rate 56
III.2.1 Random uniform sampling 56
III.2.1.1 Problematic 56
III.2.1.2 Simulations on isotropic shapes 57
III.2.1.3 Optimal sampling rate and sparsity 60
III.2.1.4 Optimal sampling rate and shape factor 60
III.2.1.5 Putting things together 61
III.2.2 Random Gaussian sampling 63
III.2.3 Realistic images 65
III.3 Conclusion 67

IV Video sampling 69
IV.1 CS applied to video signals 70
IV.1.1 Acquisition model and problem formulation 70
IV.1.2 Existing sparsity models adapted to video signals 71
IV.1.3 Reconstruction using 3D total variation 73
IV.1.3.1 Three-dimensional total variation 73
IV.1.3.2 Mean background correction 76
IV.1.4 Comparative numerical simulations 77
IV.1.4.1 Methodology 77
IV.1.4.2 Data fidelity and reconstruction artifacts 78
IV.1.4.3 Sampling rate gain over frame-by-frame reconstruction . . . 80
IV.2 Non-linear acquisition and phase retrieval 81
IV.2.1 Non-linear versus linear optical Fourier measurements 81
IV.2.2 Translation invariance issue and problem formulation 82
IV.2.3 Phase retrieval reconstruction 83
IV.2.3.1 General framework 83
IV.2.3.2 Projection operator on the data set 85
IV.2.3.3 Hybrid total variation 86
IV.2.3.4 Projection operator on the regularization set 87
IV.2.3.5 Overall reconstruction algorithm 94
IV.2.4 Numerical simulations 96
IV.2.4.1 Methodology 96

IV.2.4.2 Qualitative and quantitative results 96
IV.2.4.3 Weight map in the hybrid total variation 98
IV.3 Conclusion 100

V Using CS as a denoising method? 101
V.1 Introduction 102
V.1.1 Denoising background 102
V.1.2 Denoising via aggregation of multiple CS reconstructions 104
V.1.3 SURE and parameter estimation 106
V.1.4 Chapter outline 107
V.2 Mixed Poisson-Gaussian noise model 107
V.2.1 Generalized unbiased risk estimators 107
V.2.2 Poisson noise and associated PURE estimator 108
V.2.3 Mixed Poisson-Gaussian noise 108
V.2.4 Unbiased risk estimator for the MPG model 110
V.3 Stochastic evaluation of the Poisson-Gaussian URE 111
V.3.1 Why is a deterministic evaluation of PG-URE impossible? 111
V.3.2 Evaluation of the first-order derivative term 112
V.3.3 Evaluation of the second-order derivative term 113
V.3.4 Empirical PG-URE estimator 114
V.3.5 Variance with respect to the random perturbations 115
V.4 Numerical validation and application 117
V.4.1 Simulation goals and process 117
V.4.2 Influence of the amplitude parameters ϑ and □ 119
V.4.2.1 Parameter ϑ 119
V.4.2.2 Parameter □ 122
V.4.3 Optimization of the denoising parameters θ driven by PG-URE . . 124
V.5 Conclusion 127
V.A Derivation of the PG-URE estimator 127
V.B Optimal perturbation for the second-order derivative term 128
V.B.1 Expression of V_{δ} 129
V.B.2 Optimal probability distribution 131

Conclusion and perspectives 133

Résumé francophone détaillé 135
A Introduction à la théorie du CS 137
A.1 Définitions et formalisme mathématique 137
A.2 Exemples d'applications d'imagerie utilisant le CS 139
B Reconstruction par optimisation convexe 140
C Paramètres d'échantillonnage dans le plan de Fourier 141
C.1 Stratégies d'échantillonnage dans le plan de Fourier 141
C.2 Évaluation numérique d'un taux d'échantillonnage optimal 143
D Échantillonnage vidéo 144

D.1 Application du CS aux signaux vidéos 144
D.2 Acquisitions non-linéaires et reconstruction de phase 146
E Vers une méthode de débruitage CS ? 149
E.1 Débruitage par agrégation de reconstructions CS multiples 149
E.2 Estimateur de risque applicable au bruit mixte poisson-gaussien . 151
Conclusion 153

List of publications **155**

Bibliography **157**

Acknowledgments

First, I would like to express my sincere gratitude to my PhD advisors Elsa Angelini, associate professor at the department of signal and image processing of Télécom ParisTech, and Jean-Christophe Olivo-Marin, head of the quantitative image analysis unit at Institut Pasteur for their advice and encouragements, and for the confidence they granted me along the three years it took to prepare this thesis.

My sincere thanks are also due to the other jury members: Rémi Gribonval, senior research scientist at INRIA, and Dimitri Van De Ville, tenure-track assistant professor at University of Geneva and at École Polytechnique Fédérale de Lausanne, who have accepted to devote some of their time to review this manuscript; Caroline Chaux, research scientist at Aix-Marseille Université, and Béatrice Pesquest-Popescu, professor at the department of signal and image processing of Télécom ParisTech, who have accepted to be part of my PhD jury. I also thank the members of my "mid-term evaluation" jury, Marco Cagnazzo, associate professor at the department of signal and image processing of Télécom ParisTech, and Rémi Gribonval, for their valuable advice and remarks.

I would like to thank my – past and present – colleagues and friends of the quantitative image analysis group at Institut Pasteur, Fabrice de Chaumont, Stéphane Dallongeville, Alexandre Dufour, Nicolas Hervé, Thibault Lagache, Timothée Lecomte, Marie-Anne Lin, Marcio Marim, Vannary Meas-Yedid, François Orieux, Praveen Pankajakshan, Sorin Pop, Thomas Provoost, as well as all the numerous trainees who worked in the laboratory. I hope that furious chess game sessions will still be held at the coffee break after I leave the lab! I would also like to thank the people I met at Télécom ParisTech during my frequent visits, especially Juan Pablo de la Plata and Sonia Dahdouh.

This work has been funded by the École Doctorale de l'École Polytechnique (EDX) and by the Institut Pasteur, and I would like to grant these institutions with my gratitude for having made this PhD possible.

Finally, my last thanks go to my parents, family and friends, for everything else.

Abstract

In the past few years, the mathematical theory of compressed sensing (CS) has emerged as a new tool in the image processing field, leading to some progress in surpassing the limits stated by the Nyquist sampling theory. In particular, the CS theory establishes that a signal (image, video, etc.) can be reconstructed from a relatively small subset of non-adaptive linear random measurements, assuming that it presents a sparse structure. As this hypothesis actually holds for a large number of natural images, several imaging applications have already benefited from this theory in various aspects.

The goal of the present PhD work is to investigate how the CS theory – and more generally the ideas and methods developed in relation with sparse signal reconstruction problematics – can be used to design efficient optical sensing devices with high spatial and temporal resolution for biological imaging applications. We first investigate some practical issues related to the post-processing stage required by CS acquisition schemes, and to the selection of sampling parameters. We then examine how CS can benefit to video sampling applications. Finally, with the application of CS methods for denoising tasks in mind, we focus on the error estimation issue in image denoising problems for low-light microscopy applications.

Résumé

Ces dernières années, la théorie mathématique de l'échantillonnage compressé (*compressed sensing*, CS) a émergé en tant que nouvel outil en traitement d'images, permettant notamment de dépasser certaines limites établies par la théorie de l'échantillonnage de Nyquist. En particulier, la théorie du CS établit qu'un signal (une image, une séquence vidéo, etc.) peut être reconstruit à partir d'un faible nombre de mesures linéaires non-adaptatives et aléatoires, pourvu qu'il présente une structure parcimonieuse. Dans la mesure où cette hypothèse se vérifie pour une large classe d'images naturelles, plusieurs applications d'imagerie ont d'ores-et-déjà bénéficié à des titres divers des résultats issus de cette théorie.

Le but du travail doctoral présent est d'étudier comment la théorie du CS – et plus généralement les idées et méthodes en relation avec les problèmes de reconstruction de signaux parcimonieux (*sparse*) – peuvent être utilisés pour concevoir des dispositifs d'acquisition optiques à haute-résolution spatiale et temporelle pour des applications en imagerie biologique. Nous étudions tout d'abord quelques questions pratiques liées à l'étape de reconstruction nécessairement associée aux systèmes d'acquisition exploitant le CS, ainsi qu'à la sélection des paramètres d'échantillonnage. Nous examinons ensuite comment le CS peut être utilisé dans le cadre d'applications d'échantillonnage de signaux vidéo. Enfin, avec dans l'idée l'utilisation dans des problèmes de débruitage de méthodes inspirées du CS, nous abordons la question de l'estimation d'erreur dans les problèmes de débruitage d'images acquises en conditions de faible luminosité, notamment dans le cadre d'applications de microscopie.

Notations

We recapitulate here some of the notations and conventions used along this manuscript.

Mathematical sets and entities

- $\mathbb{N}$: set of natural integers, including 0.
- $\mathbb{Z}$: ring of relative integers.
- $\mathbb{R}$: field of real numbers.
- $\mathbb{C}$: field of complex numbers.
- $\mathbb{R}^*$: real numbers excluding 0. $\mathbb{N}^*$, $\mathbb{Z}^*$, $\mathbb{C}^*$, defined accordingly.
- $\mathbb{R}^+$: positive real numbers including 0.
- $\mathbb{E}^d$ with $d \in \mathbb{N}^*$: set of the d-tuples of elements of a set $\mathbb{E}$.
- $\mathbb{E}^{m \times n}$ with $m, n \in \mathbb{N}^*$: set of matrices with m lines and n columns with entries in $\mathbb{E}$. Unless otherwise mentioned, elements of $\mathbb{E}^m$ are assimilated to elements of $\mathbb{E}^{m \times 1}$ (column matrices).
- $\mathcal{F}(\Omega \to \mathbb{E})$: sets of functions defined over a domain Ω and taking values in a set $\mathbb{E}$. Ω is typically a subset of $\mathbb{R}^d$ (continuous domain) or a subset of $\mathbb{Z}^d$ (discrete domain).
- $[\![a, b]\!]$ with $a, b \in \mathbb{Z}$: interval of all relative integers n such that $a \leq n \leq b$.

Arithmetic and miscellaneous notations

- $|\Omega|$: number of elements of the finite set Ω.
- $\lfloor t \rfloor$ with $t \in \mathbb{R}$ ("floor of t"): the largest relative integer smaller than or equal to t.
- $\lceil t \rceil$ with $t \in \mathbb{R}$ ("ceil of t"): the smallest relative integer larger than or equal to t.
- $(a \bmod N)$ with $a \in \mathbb{Z}$, $N \in \mathbb{N}^*$: remainder of the Euclidian division of a by N, i.e. unique r such that there exists $q \in \mathbb{Z}$ with $a = q \times N + r$.
- $C_n^p = \frac{n!}{p!(n-p)!}$: binomial coefficient.
- $\bar{z}$ with $z \in \mathbb{C}$: complex conjugate of z.

Linear algebra

- x_k, f_k: k^{th} coefficient of a vector $x \in \mathbb{C}^N$ (typically a rasterized image) or a function f returning a vector.
- $\|x\|_p$ with $p \in \mathbb{N}^* \cup \{+\infty\}$: l_p-norm of a vector $x \in \mathbb{C}^N$. Formally:

$$\|x\|_p = \left(\sum_k |x_k|^p\right)^{\frac{1}{p}} \quad \text{for any } p \in \mathbb{N}^* \qquad\qquad \|x\|_\infty = \max_k |x_k|$$

- $\|x\|_0$: number of non-zero coefficients (also called l_0-pseudo-norm) of a vector $x \in \mathbb{C}^N$.
- $\langle x | y \rangle = \sum_k \overline{x_k} \cdot y_k$: canonical inner product between two vectors $x, y \in \mathbb{C}^N$.
- $x \odot y \in \mathbb{C}^N$ with $x, y \in \mathbb{C}^N$: pointwise product (also called Hadamard product) between two vectors x and y.
- $\mathbf{0}, \mathbf{1}$: constant vectors with all entries equal to 0 or 1 respectively.
- $e_k \in \mathbb{C}^N$ with $k \in [0, N-1]$: k^{th} vector of the canonical basis of $\mathbb{C}^N$; all its entries are 0, except the kth one, equal to 1.
- $\mathrm{Id} \in \mathbb{C}^{N \times N}$: identity matrix of $\mathbb{C}^{N \times N}$; all its entries are 0, except the N ones on the main diagonal, equal to 1.
- W^* with $W \in \mathbb{C}^{M \times N}$: adjoint of the matrix W. If W has real-valued entries, W^* is simply its transpose.
- $\|W\| = \sup_x \frac{\|Wx\|_2}{\|x\|_2}$: operator norm of $W \in \mathbb{C}^{M \times N}$.
- $\mathrm{Tr}(W) = \sum_k w_{k,k}$: trace of a square matrix $W \in \mathbb{C}^{N \times N}$ whose entries are denoted as $w_{k,l}$ $(k, l \in [0, N-1])$.

Differential calculus

If $f \in \mathcal{F}(\mathbb{R}^N \to \mathbb{R})$, the gradient of f at point $x \in \mathbb{R}^N$ is denoted as $\nabla f(x) \in \mathbb{R}^N$, and its Hessian matrix at x as $\nabla^2 f(x) \in \mathbb{R}^{N \times N}$ (assuming that f is regular enough for these objects to exist). Formally:

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_0}(x) \\ \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_{N-1}}(x) \end{bmatrix} \qquad \nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_0^2}(x) & \frac{\partial^2 f}{\partial x_0 \partial x_1}(x) & \cdots & \frac{\partial^2 f}{\partial x_0 \partial x_{N-1}}(x) \\ \frac{\partial^2 f}{\partial x_1 \partial x_0}(x) & \frac{\partial^2 f}{\partial x_1^2}(x) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_{N-1}}(x) \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_{N-1} \partial x_0}(x) & \frac{\partial^2 f}{\partial x_{N-1} \partial x_1}(x) & \cdots & \frac{\partial^2 f}{\partial x_{N-1}^2}(x) \end{bmatrix}$$

The usual notation $\frac{\partial}{\partial x_k}$ is used to denote the kth partial derivative of a function depending on a variable whose "natural symbol" is $x \in \mathbb{R}^N$.

Random variables

- $\mathbb{E}[X]$: expected value of a random variable X.
- $\mathrm{Var}[X]$: variance of a random variable X.
- $\mathcal{N}(\mu, \Sigma)$ with $\mu \in \mathbb{R}^N$ and $\Sigma \in \mathbb{R}^{N \times N}$ a symmetric positive matrix: probability distribution of Gaussian vectors with mean μ and covariance matrix Σ.
- $\mathcal{P}(\lambda)$ with $\lambda \in \mathbb{R}^N$: probability distribution of vectors composed of N independent entries, with the kth entry following a Poisson law of parameter λ_k for all k.

General introduction

Nowadays, microscopy techniques play an increasing role in the development and advances in modern biological science, which requires increasing imaging capabilities in terms of depth penetration, optical resolution, acquisition speed, sensitivity, etc. To tackle these challenging issues, several microscopy imaging modalities have been developed in the last ten to twenty years: two-photon excitation microscopy [Denk90] permits to observe samples at very high depths, structured illumination microscopy (SIM) [Gustafsson00] or single-molecule imaging techniques (PALM/STORM) [Betzig06, Rust06] allow to obtain spatial resolutions beyond the diffraction limit, selected plane illumination microscopy (SPIM) [Huisken04] enables fast acquisition and "3D+Time" imaging of living samples. In terms of signal processing, these imaging techniques produce very large sets of data, due to their increased resolution and/or to the multi-dimensional nature of the acquired images. Handling such large sets of data may raise difficulties, and imposes strong technical constraints on the design of the acquisition systems. Using smart sensing techniques stemming from the compressed sensing (CS) theory, we believe that these constraints can be relaxed by reducing the number of samples that need to be acquired to reconstruct these large optical microscopy images.

In this thesis, we propose to study how the compressed sensing theory can benefit to optical imaging, with in mind the design of efficient optical microscopy systems. More precisely, following the approach initiated by Marcio Marim's PhD work [Marim11a], we focus on the study of Fourier-based compressed sensing: in such acquisition model, the imaged scene is observed through an optical set-up whose role is to implement an optical Fourier transform [Goodman96], and an array of photo-electric transducers properly positioned downstream to this optical set-up is in charge of the actual acquisition task. The organisation of the manuscript reflects the different problematics tackled during this PhD work.

In chapter I, we present the mathematical theory of compressed sensing. We introduce the CS formalism and notions, and recall some of the main theoretical results obtained in the CS framework. We finally illustrate the interest of this theory by presenting four examples of imaging applications that benefited from CS results or whose design were directly inspired by them.

Chapter II is dedicated to the reconstruction issues raised by CS acquisition schemes. We present how the CS reconstruction problem may be formulated in practice, focusing in

particular on convex optimization formulations. We then propose a review of the existing algorithmic methods that solve these convex optimization problems, by presenting the general characteristics and principles of these optimization algorithms, and comparing their performances in practical situations.

In chapter III, we discuss the different parameters associated to the sensing operation in Fourier-based CS, namely the position in the Fourier domain where samples should be acquired (i.e. the sampling strategy), and the number of such samples (i.e. the sampling rate). We first review the works addressing the determination of the best sampling strategy, and show that, currently, answers to this problem remain mostly based on empirical observations, in spite of recently released theoretical works on this issue. We also investigate the incidence of the choice of a sampling rate on the efficiency of the CS acquisition and reconstruction scheme, and the artifacts observed in reconstructed images in the context of two representative sampling strategies, namely uniform and Gaussian random sampling.

In chapter IV, we study how Fourier-based CS can be applied to video sensing and reconstruction applications. We first consider the case of a video to be reconstructed from partial Fourier measurements acquired on each of its frames, focusing in particular on the sparsity models to use for efficient video reconstruction: we compare several existing sparsity models, and introduce a new one based on 3D total variation, which improve the quality of the reconstructed sequences. We then switch to a non-linear acquisition model – beyond the "pure" CS framework – in which only the modulus of the Fourier transform of the signal would be acquired: for this different reconstruction problem, we show that we can exploit the same sparsity properties exhibited by video sequences than the ones used in the linear acquisition scenario to implement a "phase-retrieval-like" reconstruction procedure.

Finally, chapter V focuses on the design of an estimator of the mean squared error in denoising problems, in a context of a mixed Poisson-Gaussian noise model, that is relevant to model the noise present in low-light microscopy applications. Although this work goes off the general point of this thesis, it was originally motivated as part of the extension of CS denoising methods, as proposed by [Marim09, Marim11a]. We however derive a practical formulation for our PG-URE estimator that make this tool usable "out of the box" with almost any existing denoising algorithm. In particular, we present some examples of denoising parameter optimizations involving standard denoising methods and phantom test images, and show that our framework leads to results similar than the ones obtained using an oracle-based approach.

Chapter I

Introduction on CS theory

Compressed sensing (CS) is a theory that has emerged and developed over the last ten years, based on the seminal works of Candès, Romberg and Tao [Candès06a] on the one hand, and Donoho [Donoho06] on the other hand; the goal of this theory is to study a class of inverse problems involving signals that have a sparse structure. To be more precise, the problem tackled by CS consists in recovering a signal of interest $x \in \mathbb{C}^N$ from a vector of observations $y = \Phi x \in \mathbb{C}^M$ constituted by linear projections of this signal x, the number M of scalar projections being significantly smaller than the size N of the signal ($M \ll N$): in this context, the linear operator $\Phi \in \mathbb{C}^{M \times N}$ is called the *measurement operator*. In order to remove the indeterminacy due to the small size of the observation vector y, some assumptions have to be made on the structure of the signal of interest x to recover: in the case of CS, this consists in assuming that x has a sparse representation in some known basis or dictionary Ψ, i.e. there exists a vector $s \in \mathbb{C}^L$ such that $x = \Psi s$ whose most of the coefficients are zero. In this case, $\Psi \in \mathbb{C}^{N \times L}$ is called the *sparsity basis* or *sparsity dictionary*.

The goal of this introductory chapter is to give a brief overview of this theory. We start by describing the general formalism used for signals and images throughout this manuscript, and by stating some more specific definitions about the notion of *sparsity* as it is at the heart of the CS theory. We will then review some of the important theoretical results that were established in the theory of CS, including the related precursory works. Finally, we will conclude by presenting some imaging applications that have benefited from CS or that were developed subsequently to the emergence of this theory.

I.1 A few definitions **22**

I.1.1 Signals and images 22

I.1.2 Sparsity and compressibility 23

I.2 Compressed sensing theoretical results **25**

I.2.1 Recovering sparse data from incomplete measurements 25

I.2.2 Restricted isometry property 26
I.2.3 Partial unitary transforms 27
I.2.4 Sparse representations and dictionaries 28
I.2.5 Block sparsity and total variation 30

I.3 Application of compressed sensing for imaging devices 31
I.3.1 Magnetic resonance imaging 31
I.3.2 Digital holography 32
I.3.3 Single-pixel camera 34
I.3.4 Schlieren deflectometry 35

I.1 A few definitions

I.1.1 Signals and images

In this manuscript, signals are modeled as elements of a function set $\mathcal{F}(\Omega, \mathbb{E})$, where Ω and $\mathbb{E}$ characterize the different types of signals. In particular:

- 2D images are modeled as elements of $\mathcal{F}(\Omega, \mathbb{R})$ with $\Omega \subset \mathbb{R}^2$ (continuous modeling) or $\Omega \subset \mathbb{Z}^2$ (discrete modeling, more common in this manuscript);
- 3D images are modeled as elements of $\mathcal{F}(\Omega, \mathbb{R})$ with $\Omega \subset \mathbb{R}^3$ or $\Omega \subset \mathbb{Z}^3$;
- video sequences of 2D images (also denoted as 2D+T signals) are modeled in the same way, except that one of the dimensions of the domain Ω is particularized as the time dimension;
- multi-channel images are modeled as elements of $\mathcal{F}(\Omega, \mathbb{R}^c)$ where $c \in \mathbb{N}^*$ is the number of channels.

We use bold blue font to denote signals and functions that return signals (example: "let $\mathbf{x} \in \mathcal{F}(\Omega, \mathbb{R})$ be an image", or "let $\mathbf{f}$ be a denoising operator"), and bold red font for linear operators between signal spaces.

When Ω is a finite set – in particular when Ω is a bounded subset of $\mathbb{Z}^d$ – the sets $\mathcal{F}(\Omega, \mathbb{E})$ and $\mathbb{E}^{|\Omega|}$ are isomorphic: an isomorphic mapping between these two sets is then characterized by a bijection $\phi : [0, |\Omega| - 1] \to \Omega$, i.e. an ordered list of all the elements of Ω. As the actual chosen bijection ϕ does not matter in general for ideas and demonstrations developed in this manuscript, we use either $\mathbf{x} \in \mathcal{F}(\Omega, \mathbb{E})$ or equally $\mathbf{x} \in \mathbb{E}^{|\Omega|}$ to characterize a signal of the corresponding type. Individual components of this signal $\mathbf{x}$ are denoted:

- either as $\mathbf{x}_k$ with $k \in [0, |\Omega| - 1]$ when $\mathbf{x}$ is seen as an element of $\mathbb{E}^{|\Omega|}$,
- or as $\mathbf{x}(u)$ with $u \in \Omega$ when $\mathbf{x}$ is seen as an element of $\mathcal{F}(\Omega, \mathbb{E})$

Obviously, x_k and $x[u]$ refer to the same element when $u = \phi(k)$.

Finally, for signals x that are defined on a bounded domain $\Omega \subset \mathbb{Z}^d$, it is convenient in some situations to consider extensions of x over the whole grid $\mathbb{Z}^d$. In particular, for $\Omega = [0, n_1 - 1] \times [0, n_2 - 1] \times \cdots \times [0, n_d - 1] \subset \mathbb{Z}^d$, the extension of x using *periodic boundary conditions* is defined as follows:

$$x[u_1, \ldots, u_d] = x[u_1 \bmod n_1, \ldots, u_d \bmod n_d] \qquad \text{for all } u = (u_1, \ldots, u_d) \in \mathbb{Z}^d \qquad \text{(I–1)}$$

where the same notation is kept for both x and its extended version.

I.1.2 Sparsity and compressibility

Definition I–1 (Sparsity) *A vector $x \in \mathbb{C}^N$ is said to be S-sparse (with $0 \leq S \leq N$) if it has at most S non-zero coefficients. The minimal value S for which x is S-sparse is denoted as $\|x\|_0$, and is called the l_0-norm*[1] *of x. Finally, x is said to be sparse if $\|x\|_0 \ll N$.*

To be more concrete, $\|x\|_0$ denotes the number of non-zero coefficients of x, and x is said to be sparse if most of its coefficients are zero. By extension, we will say that a vector $x \in \mathbb{C}^N$ is sparse in a dictionary $\Psi \in \mathbb{C}^{N \times L}$ if there exists a sparse vector $s \in \mathbb{C}^L$ such that $x = \Psi s$: in this case, s is said to be a *sparse representation* of x in the *sparsity dictionary* Ψ. An example of this situation is illustrated in Fig. I–1. In general, the matrix Ψ will be assumed to be full-rank; then, $\Psi \in \mathbb{C}^{N \times L}$ will be denoted as a *sparsity basis* if $L = N$, and as a *sparsity redundant dictionary* if $L \gg N$. The case $L \ll N$ (under-determined dictionary) is less common, as such type of dictionary does not allow to represent all the signals of $\mathbb{C}^N$.

It should be noted that, except in the trivial situations corresponding to $S = 0$ or $S = N$, the subset of the S-sparse signals of $\mathbb{C}^N$ is not a vector subspace, but rather a union of $\mathbb{C}_N^S$ subspaces, each of them of dimension S. As a consequence, this space is not closed for the addition: the sum of two S-sparse vectors might not be S-sparse.

Another important concept in relation with sparsity is the notion of *compressibility*, also denoted as *weak sparsity* by some authors. Informally, a compressible vector $x \in \mathbb{C}^N$ can be thought as a vector that can be approximated by a well-chosen sparse vector $\tilde{x}$. More precisely, we will say that $x \in \mathbb{C}^N$ is compressible if there exists a sparse vector $\tilde{x} \in \mathbb{C}^N$ such that the order of magnitude of the approximation error $(x - \tilde{x})$ is significantly smaller than the one of x, where these orders of magnitude are measured with an appropriate metric. An example of a compressible 1D signal is presented in Fig I–2.

In order to formalize the concept of compressibility in a more rigorous way, [Candès06b] proposes the following definition:

[1] Here, the denomination "norm" is abusive, in that the functional $\|\cdot\|_0$ does not verify all the properties usually required for a norm in a vector space: in particular, $\|\cdot\|_0$ is not positive-homogeneous. However, the term "l_0-norm" is very common, and we will use it for convenience in the rest of the manuscript.

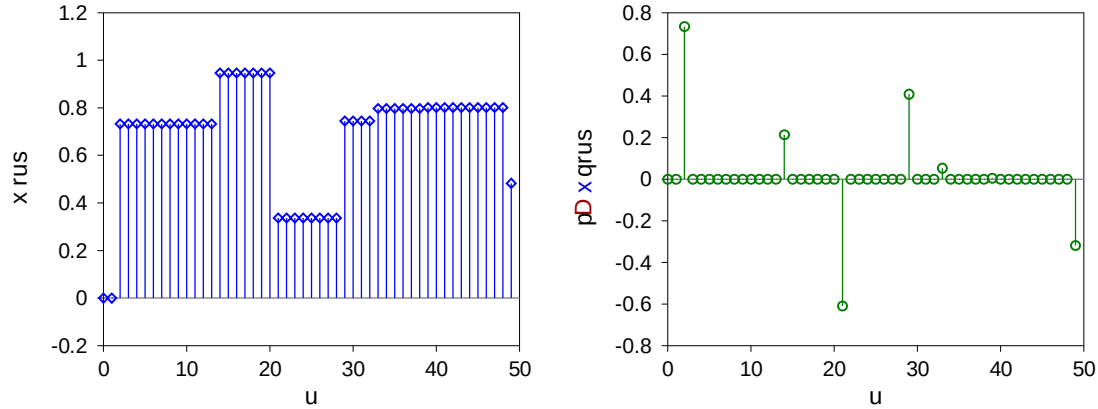


Figure I–1: Example of a piecewise constant signal x (left chart), i.e. a signal whose discrete derivative Dx (defined here as $pDx qrus = x rus - x ru-1s$, right chart) is sparse.

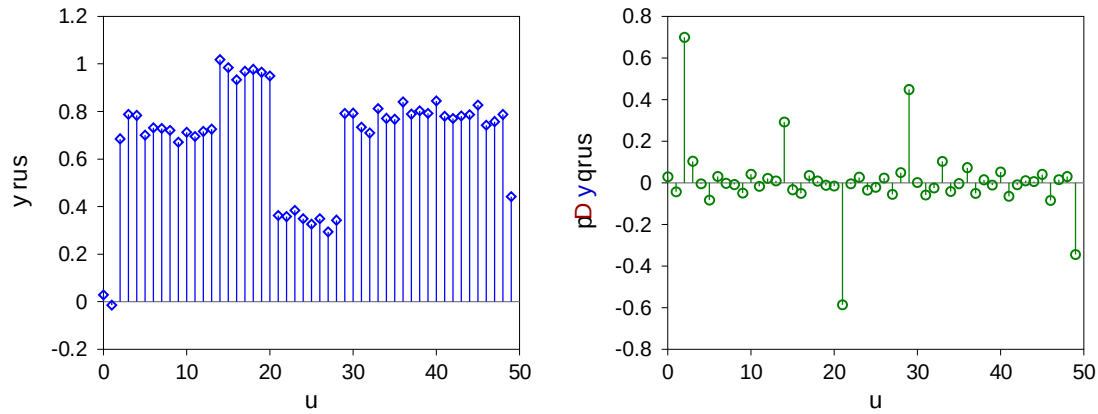


Figure I–2: Example of a signal y whose discrete derivative Dy is compressible: as most of the coefficients of Dy are close to zero, sparse approximations of this signal can be obtained by actually setting these coefficients to zero, keeping only those having a significant order of magnitude. An example of such sparse approximation is the signal Dx presented in Fig. I–1.

Definition I–2 (Compressibility, weak sparsity) *A vector $x \in \mathbb{C}^N$ is said to be compressible if, for some value $r > 1$, its components obey the following decreasing power law, i.e. there exists a constant C_r (depending only on r) such that:*

$$\left|x_{p(k)}\right| \leq C_r \cdot k^{-r} \qquad \text{for all } k \in [0, N-1] \tag{I–2}$$

where the sequence $\left(x_{p(0)}, x_{p(1)}, \ldots, x_{p(N-1)}\right)$ represents the components of x sorted in decreasing order with respect to their magnitudes: $\left|x_{p(0)}\right| \geq \left|x_{p(1)}\right| \geq \cdots \geq \left|x_{p(N-1)}\right|$.

This formal definition I–2 does meet the intuitive one given above. Indeed, if $x \in \mathbb{C}^N$ is a vector for which (I–2) holds, for any S with $1 \leq S \leq N$ we can construct a S-sparse vector $\tilde{x}^S \in \mathbb{C}^N$ by setting to zero all the components of x except the S largest ones. Then, the l_1-norm of the approximation error $\left\|x - \tilde{x}^S\right\|_1$ can be bounded as follows:

$$\|x - \tilde{x}^S\|_1 \leq \frac{C_r}{r - 1} \cdot \frac{1}{S^{r-1}} \tag{I–3}$$

which shows that $\tilde{x}^S$ is a good approximation for x, provided that C_r is not too large. This latter requirement on C_r is somehow legitimized by the fact that, in a vector space of finite dimension, the definition I–2 does hold for any vector x, as it is always possible to find a finite constant C_r large enough to be compatible with (I–2): the formal definition I–2 actually captures the intuitive notion of compressibility only if C_r is not allowed to take extremely large values. More generally, it can be shown that $\tilde{x}^S$ is the S-sparse vector that minimizes the approximation error $\|x - \tilde{x}^S\|_p$ for any of the l_p-norm with $1 \leq p \leq \infty$: in that, $\tilde{x}^S$ is referred as the best S-sparse approximation of x.

I.2 Compressed sensing theoretical results

I.2.1 Recovering sparse data from incomplete measurements

The idea that signals with an underlying sparse structure can be efficiently measured and processed emerged about ten years ago. Among the works that initiated what would become the theory of compressed sensing, we mention [Donoho01], which studied the problem of recovering the underlying structure of a signal obtained as a superposition of a few Dirac atoms together with a few sine wave atoms. The authors showed that the decomposition of such a 1D signal in terms of a sum of Dirac and sinusoid atoms is unique, assuming that the number S_d of Dirac atoms and the number S_s of sinusoid atoms are far smaller than the size N of the signal (more specifically: $S_d + S_s < \sqrt{\frac{N}{2}}$). They extended their work to what they called mutually incoherent bases, and derived some similar conditions about the uniqueness of the decomposition of a signal in terms of sparse linear combination of atoms taken from a pair of such mutually incoherent orthonormal bases. The authors described this property of mutual incoherence between two bases as the fact that *"no nonzero signal can have a sparse representation in both bases simultaneously"*, and showed that this property holds for many pairs of bases (Dirac and sinusoids, wavelets and sinusoids, wavelets and ridgelets, etc.).

These results were then extended in parallel by [Donoho03] and [Gribonval03], who studied the problem of seeking sparse solutions $x \in \mathbb{C}^N$ to the system of linear equations $y = \Phi x$, where $y \in \mathbb{C}^M$ is a vector of observations and $\Phi \in \mathbb{C}^{M \times N}$ a given dictionary. Such type of solution x could be obtained as a minimizer of the following optimization problem:

$$\underset{x \in \mathbb{C}^N}{\arg\min} \|x\|_0 \text{ subject to } y = \Phi x \tag{P_0}$$

In general, solving (P_0) requires to seek the smallest subset of columns of Φ – among all the 2^N possible subsets – such that there exists a linear combination of these columns equal to y: the complexity of this problem grows exponentially with N, making it intractable even for small values of this parameter. Therefore, pursuing a *convex relaxation* approach

already proposed in previous works on sparsity, the authors suggested to replace the l_0-norm by the l_1-norm, turning (P_0) into a convex optimization problem (P_1), that can be handled in a more practical manner:

$$\underset{x \in \mathbb{C}^N}{\arg\min} \|x\|_1 \text{ subject to } y = \Phi x \qquad (P_1)$$

These works demonstrated an important result about the strategy for seeking sparse solutions to linear equations, which can be presented as follows: given a vector $y \in \mathbb{C}^M$ and a dictionary $\Phi = [\varphi_1 \ \varphi_2 \ \cdots \ \varphi_N] \in \mathbb{C}^{M \times N}$ where the vectors φ_k represent the columns of Φ, if there exists a vector $x \in \mathbb{C}^N$ such that $y = \Phi x$, and if the following holds:

$$\|x\|_0 < \left(\frac{1}{2}\right)\left(1 + \frac{1}{\mu(\Phi)}\right) \quad \text{where} \quad \mu(\Phi) = \max_{k \neq l} \frac{|\langle \varphi_k, \varphi_l \rangle|}{\|\varphi_k\|_2 \cdot \|\varphi_l\|_2} \qquad \text{(I–4)}$$

then x is the unique solution to both (P_0) and (P_1). In other words, this result means that if y is indeed the result of a linear combination of a sufficiently small number of columns of Φ, then:

- first, solving (P_0) does permit to identify this linear combination;
- second, the strategy consisting in solving (P_1) instead of (P_0) is relevant, as the solutions of these two problems are equal.

I.2.2 Restricted isometry property

In the mid-2000, Candès, Romberg and Tao extend the ideas previously developed by [Donoho03] and [Gribonval03] in a serie of papers [Candès05b, Candès06a, Candès06c, Candès06b]. However, compared to previous works, Candès, Romberg and Tao establish some theoretical results about the reconstruction of sparse signals in a framework relying on hypotheses that are more consistent with situations encountered in practical image sensing applications. More specifically, the new framework relaxes the following hypotheses:

1. the observation vector may be inaccurate to some extent (for instance, it can be degraded by some noise sources),
2. the signals to be reconstructed do not need to be strictly sparse, but rather compressible.

To obtain their results, these authors introduce in [Candès05b] the *restricted isometry constant* associated to a linear operator.

Definition I–3 (Restricted isometry property) *Given a linear operator $\Phi \in \mathbb{C}^{M \times N}$ and an integer $S \in \mathbb{N}^*$, the S-restricted isometry constant associated to Φ is the smallest scalar $\delta_S \in \mathbb{R}$ such that:*

$$(1 - \delta_S)\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1 + \delta_S)\|x\|_2^2 \qquad \textit{for all } x \in \mathbb{C}^N \quad \textit{with} \quad \|x\|_0 \leq S \qquad \text{(I–5)}$$

A constant δ_S close to 0 means that the operator Φ behaves approximately like an isometry for S-sparse inputs, i.e. it almost preserves the norms of these vectors. More generally, the smaller the constants δ_S associated to Φ, the larger the class of signals that can be recovered by solving either (P_0) or (P_1).

This latter property is stated in more formal ways in several papers; for instance, we recall here a result taken from [Candès08]: given a signal of interest $x \in \mathbb{C}^N$ (not necessarily sparse), a measurement operator $\Phi \in \mathbb{C}^{M \times N}$, and an observation vector $y = \Phi x + b$ degraded by an unknown additive noise b such that $\|b\|_2 \leq \epsilon$, an estimator $\hat{x}$ of x is defined as the solution of a convex optimization problem, as follows[2]:

$$\hat{x} = \underset{x \in \mathbb{C}^N}{\arg\min} \|x\|_1 \text{ subject to } \|\Phi x - y\|_2 \leq \epsilon \tag{I–6}$$

Then, if $\delta_{2S} < \sqrt{2} - 1$ for some S, the following inequality holds, that establishes an upper bound on the error committed when estimating x by $\hat{x}$:

$$\|\hat{x} - x\|_2 \leq A \cdot \epsilon + \frac{B}{\sqrt{S}} \|x - \tilde{x}^S\|_1 \tag{I–7}$$

where A and B are two positive constants depending only on δ_{2S}, and $\tilde{x}^S$ is the best S-sparse approximation of x, i.e. the vector obtained by setting to zero all the components of x except the S largest ones (same definition than in Sec. I.1.2).

It can be noted that, in the inequality (I–7), the two terms involved in the upper bound are related to the imperfect characteristics of the "real-world" signals, already mentioned above:

1. the noise that affects the observation vector y (term $A \cdot \epsilon$),
2. the non-sparseness of the signal to recover (term $\frac{B}{\sqrt{S}} \|x - \tilde{x}^S\|_1$).

In particular, this second term involving $\|x - \tilde{x}^S\|_1$ is likely to be very small if x is compressible, as explained in Sec. I.1.2.

I.2.3 Partial unitary transforms

One class of measurement operators Φ that is encountered in several CS imaging applications (see I.3) is the class of *partial unitary transforms* (also denoted as *randomly subsampled unitary transforms*). Such measurement operator $\Phi \in \mathbb{C}^{M \times N}$ is constructed by selecting M rows from a unitary matrix $U \in \mathbb{C}^{N \times N}$. Formally, this means that $\Phi = \Sigma U$, where $\Sigma \in \{0, 1\}^{M \times N}$ is a selection matrix, with exactly one non-zero entry per line and at most one non-zero entry per column (the columns corresponding to the selected rows of U). This property entails that $\Phi\Phi^* = Id$, and $\Phi^*\Phi = U^*(\Sigma^*\Sigma)U$, where $\Sigma^*\Sigma$

[2] In this formulation, the scalar parameter ϵ is supposed to be known. Prior knowledge on the probability distribution of the noise component b is often required in practice to set this parameter (see Sec. II.1.1 for more details).

is a diagonal matrix. Common examples in imaging applications include partial Fourier transforms and partial Hadamard transforms (see Sec. I.3). It is also worth mentioning that the associated unitary transforms encountered in practical applications often come with fast algorithms for computing their product against a vector $x \in \mathbb{C}^N$.

For such measurement operators Φ, [Candès07] shows the following result: given a signal $x \in \mathbb{C}^N$ such that $\|x\|_0 \leq S$ and partial unitary transform $\Phi \in \mathbb{C}^{M \times N}$ built by selecting in a uniform random manner M rows from $U \in \mathbb{C}^{N \times N}$ (unitary) such that the following inequality holds:

$$M \geq C \cdot \mu^2(U) \cdot S \cdot N \log(N) \quad \text{with} \quad \mu(U) = \max_{k,l} |u_{k,l}| \tag{I–8}$$

where C is a (small) numerical constant and $\mu(U)$ is defined as the largest magnitude among the entries $u_{k,l}$ of the matrix U. Then, if $y = \Phi x$, the program (P_1) recovers the original signal x from the measurement vector y with an overwhelming probability[3].

Here, the coefficient $\mu(U)$ is somehow a measure of the ability of U to be a "good" sensing basis for sparse signals: in the most favorable cases (when U is chosen such that $\mu(U) = N^{-1/2}$), the result by [Candès07] predicts that the minimal number of measurements required to reconstruct a S-sparse signal of size N is about $S \log(N)$, which is relatively small[4]; on the contrary, in the worst cases (when $\mu(U) = 1$), the necessary number of measurements rockets. It can be noted that, if $U = \Phi\Psi$ where Φ and Ψ are two orthonormal matrices, and if $\varphi_0, \varphi_1, \ldots, \varphi_{N-1}$ and $\psi_0, \psi_1, \ldots, \psi_{N-1}$ denote respectively the rows of Φ and the columns of Ψ, then:

$$\mu(U) = \max_{k,l} |\langle \varphi_k | \psi_l \rangle| \tag{I–9}$$

which matches the definition of the mutual coherence measure between the bases Φ and Ψ introduced in [Donoho01] (see I.2.1).

I.2.4 Sparse representations and dictionaries

In practical situations, dealing with images that are sparse in their canonical representation basis is quite unusual. A more relevant hypothesis consists in assuming that vectorized images $x \in \mathbb{C}^N$ have sparse representations $s \in \mathbb{C}^L$ in some appropriate dictionaries $\Psi \in \mathbb{C}^{N \times L}$, i.e. $x = \Psi s$ (see Sec. I.1.2). How the dictionary Ψ is actually chosen depends on the application, and more specifically on the underlying image formation model and on the properties of the studied images: several options have been proposed, such as wavelet basis or undecimated wavelet frames for piecewise regular images [Starck04], curvelet frames for piecewise regular images with regular discontinuities [Candès04], *ad hoc* dictionaries

[3] Please note that this formulation differs from the one given in [Candès07], as a different normalization is used for U in this document.

[4] With less that S measurements, the original signal could never be recovered, even if an oracle could predict the position of its non-zero coefficients.

constructed using automatic learning approaches [Duarte-Carvajalino09, Gleichman11], etc.

To extend the reconstruction through convex optimization framework to deal with sparsity dictionaries, two different approaches may be considered, as proposed by [Elad07]:

- The first one consists in changing the optimization variable in the reconstruction problem (I–6), replacing x with its representation s to make sparse. This approach, denoted by [Elad07] as *synthesis*, leads to the following reconstruction problem:

$$\hat{s} = \underset{s \in \mathbb{C}^L}{\arg\min} \|s\|_1 \text{ subject to } \|\Phi\Psi s - y\|_2 \leq \varepsilon \qquad (\mathrm{P}_{\mathrm{synthesis}})$$

  The estimate $\hat{x}$ of the signal of interest is then recovered as $\hat{x} = \Psi\hat{s}$.

- The second approach, known as *analysis*, keeps x as the optimization variable of the reconstruction problem, but modifies the optimized objective function. The estimate $\hat{x}$ of the signal of interest is then obtained as:

$$\hat{x} = \underset{x \in \mathbb{C}^N}{\arg\min} \|\Psi^{\mathrm{inv}} x\|_1 \text{ subject to } \|\Phi x - y\|_2 \leq \varepsilon \qquad (\mathrm{P}_{\mathrm{analysis}})$$

  where the operator $\Psi^{\mathrm{inv}} \in \mathbb{C}^{L \times N}$ – denoted as the *analysis operator*[5] – transforms x into a representation vector s such that $x = \Psi s$ and s is sparse for the signal to recover.

Obviously, both the analysis and the synthesis approaches are equivalent in the case where $L = N$, i.e. when Ψ is indeed a basis. In this situation, the analysis operator Ψ^{inv} is equal to the actual inverse of Ψ, i.e. $\Psi^{\mathrm{inv}} = \Psi^{-1}$.

However, the situation is more complicated when using redundant dictionaries Ψ, i.e. when $L \geq N$. [Elad07] proposes a comparison of the analysis and synthesis reconstruction for general inverse problems, and shows that in this case the behavior of these two reconstruction methods may significantly diverge. However, the authors do not advocate for one formulation compared to the other, although they point out that the analysis problem is likely to be easier to solve than the synthesis problem when Ψ is a highly redundant dictionary (i.e. $L \gg N$), since in this case x (i.e. the optimization variable of $(\mathrm{P}_{\mathrm{analysis}})$) belongs to a vector space whose dimension is much smaller than the one of the vector space to which s belongs (i.e. the optimization variable of $(\mathrm{P}_{\mathrm{synthesis}})$). An illustration of the differences between analysis and synthesis approaches in the context of image restoration problems is proposed in [Chaari09].

An extension of the restricted isometry property framework to signals that are sparse in a redundant dictionary Ψ was proposed in [Candès10]. More precisely, this work focuses on the case where Ψ is a tight frame[6], which covers several types of dictionaries used

[5] Unless otherwise specified, the notation Ψ^{inv} does not refer to the inverse of Ψ (Ψ may not even be invertible).

[6] $\Psi = \begin{bmatrix} \psi_0 & \psi_1 & \cdots & \psi_{L-1} \end{bmatrix} \in \mathbb{C}^{N \times L}$ is a tight frame if there exists a real constant $\alpha \geq 0$ such that

in practical applications: curvelet frames, redundant wavelets, or any concatenation of orthonormal bases. In this situation, the authors show that a signal $x \in \mathbb{C}^N$ having a sparse or compressible representation in such dictionary Ψ can be recovered from a vector of linear measurements $y = \Phi x + b$ (same notations than in Sec. I.2.2) by solving the analysis problem ($\mathcal{P}_{\text{analysis}}$) with $\Psi^{\text{inv}} = \Psi^{*}$: up to a modified version of the restricted isometry property (see definition I–3), they demonstrate that the error $\|\hat{x} - x\|_2$ between the solution $\hat{x}$ of this problem and the true sought signal can be bound as in (I–7).

I.2.5 Block sparsity and total variation

For image reconstruction tasks, several works such as [Candès06a, Kim09, Marim11a] propose to replace the ℓ_1-norm used in the objective function of ($\mathcal{P}_{\text{analysis}}$) by the total variation ($\mathcal{TV}$) of the image x:

$$\hat{x} = \underset{x \in \mathbb{C}^N}{\arg\min} \|x\|_{\mathcal{TV}} \text{ subject to } \|\Phi x - y\|_2 \leq \varepsilon \qquad (\mathcal{P}_{\mathcal{TV}})$$

where $\|x\|_{\mathcal{TV}}$ is defined as follows, for a 2D image $x \in \mathcal{F}(\Omega, \mathbb{C})$ defined on a domain $\Omega \subset \mathbb{Z}^2$:

$$\|x\|_{\mathcal{TV}} = \sum_{(u,v) \in \Omega} \sqrt{|(D_h x)(u,v)|^2 + |(D_v x)(u,v)|^2} \qquad \text{(I–10)}$$

where D_h and D_v represent the horizontal and vertical discrete derivative operators: the most common implementations assume that $(D_h x)(u,v) = x(u+1,v) - x(u,v)$ and similarly for D_v, but other finite difference schemes can be considered.

The effect of the $\mathcal{TV}$ driven reconstruction problem ($\mathcal{P}_{\mathcal{TV}}$) is to enforce sparsity on the gradient of the sought image x, which corresponds to assuming that x obeys a piecewise constant model (also denoted as the *cartoon* model). As a consequence, the reconstructed image $\hat{x}$ exhibits in general sharp edges and well-contrasted objects.

More generally, the gradient sparsity enforced by $\mathcal{TV}$ minimization in ($\mathcal{P}_{\mathcal{TV}}$) can be seen as a special case of *block sparsity* (also named as *group sparsity* or *structured sparsity* by certain authors). This notion was introduced by [Yuan06] and then developed by several authors (see for instance [Stojnic09, Eldar09b, Bach12]) in order to refine the sparsity models used in CS as well as in other signal processing problems. Indeed, if the ℓ_1-norm used in the objective function of the inverse problems ($\mathcal{P}_{\text{analysis}}$) and ($\mathcal{P}_{\text{synthesis}}$) does enforce sparsity, it does not account for the fact the set of non-zero coefficients of sparse representations s corresponding to typical signals of interest x often exhibit some particular structures.

To make up for these limitations, the above mentioned works introduce a notion of mixed $\ell_{1,2}$-norm over the space $\mathbb{C}^L$ of considered sparse representations: more precisely, given a partition $\{\omega_1, \ldots, \omega_G\}$ of the integer interval $[0, L-1]$ – i.e. a family of subsets

$\Psi\Psi^{*} = \alpha \mathrm{Id}$, or equivalently if $\sum_{k=0}^{L-1} |\langle x, \psi_k \rangle|^2 = \alpha \|s\|_2^2$ for all $s \in \mathbb{C}^L$. The tight frame is denoted as normalized if $\alpha = 1$.

$\omega_g \subset \{0, \ldots, L-1\}$ such that $\bigcup_{g=1}^{G} \omega_g = \{0, \ldots, L-1\}$ and $\omega_g \cap \omega_{g'} = \emptyset$ for any pair (g, g') with $g \neq g'$ – and a vector $s \in \mathbb{C}^L$, the mixed $l_{1,2}$-norm of s is defined as:

$$\|s\|_{1,2} = \sum_{g=1}^{G} \sqrt{\sum_{k \in \omega_g} |s_k|^2} \tag{I–11}$$

It can be shown that substituting the l_1-norm with this mixed $l_{1,2}$-norm in either ($P_{analysis}$) or ($P_{synthesis}$) leads to block-sparse signals, i.e. signals whose non-zero coefficients of the representation vector s are grouped over a small number of sets ω_g that compose the partition used in the definition (I–11): the number of non-zero coefficients within these few "active blocks" can however be quite large. The design of the partition $(\omega_1, \ldots, \omega_G)$ depends on the expected relations between the coefficients s_k of the sparse representation.

It is worth mentioning that several alternative definitions of mixed norms similar to (I–11) exist: one could for instance relax the requirement $\omega_g \cap \omega_{g'} = \emptyset$ to allow overlapping blocks, or substitute the "l_2 part" in (I–11) by any l_p-norm to shape the distribution of the non-zero coefficients inside the blocks. See [Bach12] and references therein for more details on these extensions, and for instance [Gramfort09] for an example of a practical application – signal reconstruction from magneto- and electro-encephalography measurements – that makes use of mixed norms.

I.3 Application of compressed sensing for imaging devices

I.3.1 Magnetic resonance imaging

Magnetic resonance imaging (MRI) is probably one of the first imaging modalities that have benefited from CS theoretical results. One reason for that is that, as noted by [Lustig08], *"MRI obeys two key requirements for successful application of CS"*:

1. typical medical images have compressible representations in appropriate domains, either wavelet or gradient (see for instance [Lustig07, Huang12]);
2. the transducers of MRI scanners (i.e. the antennas) measure a physical signal that is by essence a Fourier transform of the actual image of interest.

While the first of these two properties is not particularly related to MRI (compressible representations can be found for almost every class of natural images), the second one is indeed very specific. More precisely, the signal $y(t)$ collected by the coils at time point t of a MRI acquisition has the following form:

$$y(t) = \int x(\vec{r}) \exp\left(-2i\pi \vec{k}(t) \cdot \vec{r}\right) d^3r \tag{I–12}$$

where $x(\vec{r})$ is a 3D signal proportional to the spatially varying physical quantity that is to be imaged (typically the proton density in the tissue of a patient). Then, the measured

data $y(t)$ appears as a sample of the 3D spatial Fourier transform of the signal of interest corresponding to the spatial frequency $\vec{k}(t)$. The set of all spatial frequencies $\vec{k}(t)$ visited during the acquisition forms a sampling trajectory in the Fourier domain (i.e. the k-*space* in MRI terminology) of the signal of interest, and this trajectory is necessarily continuous (see [Wright97, Lustig08] for more details). The sampling trajectory is an adjustable parameter of the acquisition device.

The trajectory in the k-space has to be designed to satisfy a trade-off between two contradictory objectives: on the one hand, it has to be as short as possible, as the length of the trajectory conditions the total scanning time; on the other hand, the number of collected Fourier samples has to be large enough to allow the recovery of the signal of interest with minimal artifacts and sufficient spatial resolution. Traditional strategies for designing the sampling trajectories propose to follow straight lines distributed over a Cartesian grid in the k-space: while the reconstruction process corresponding to this strategy is particularly straightforward (it consists in a simple discrete inverse Fourier transform), it results in particularly long sampling paths, and thereby long scanning times. Increasing the step between two sampled lines in the k-space could reduce the acquisition time, but this strategy introduces aliasing artifacts or reduces the spatial resolution of the reconstructed image.

However, these sampling strategies do not take advantage of the underlying sparsity properties of the sampled signal, and this is where compressed sensing comes into play: it has been shown that exploiting these properties in an analysis reconstruction scheme ($P_{analysis}$) allows to significantly reduce the number of collected Fourier samples without degrading the quality of the reconstructed MRI image (see [Lustig07, Lustig08] and references therein): a 5 to 10-fold acceleration is reported in [Lustig07] in the case of some real *in vivo* applications, without significant loss of information.

I.3.2 Digital holography

Using CS techniques for digital holographic imaging applications was proposed by several authors: see for instance [Brady09, Marim10, Marim11b, Rivenson11] and references therein. These applications have in common to perform the sampling operation in the Fresnel domain of the signal of interest. The Fresnel transform characterizes the free propagation of an electromagnetic wave in an isotropic homogeneous non-dispersive medium, such as the air (see [Goodman96]). As an example of this type of application, a more detailed description of the set-up proposed in [Marim11b] is provided in what follows.

In this work, the authors introduce an off-axis compressed holographic optical set-up using a Mach-Zehnder interferometer (see Fig. I–3). In this set-up, a coherent radiation emitted by a laser is split into two beams, that follow different paths: the first one, the *object beam*, is used to illuminate a transparent planar object of interest, which transmits a diffracted light field E; the second beam E_{LO} (the *reference beam*) bypasses the object of interest, and is made to interfere with the transmitted light field E at the recording

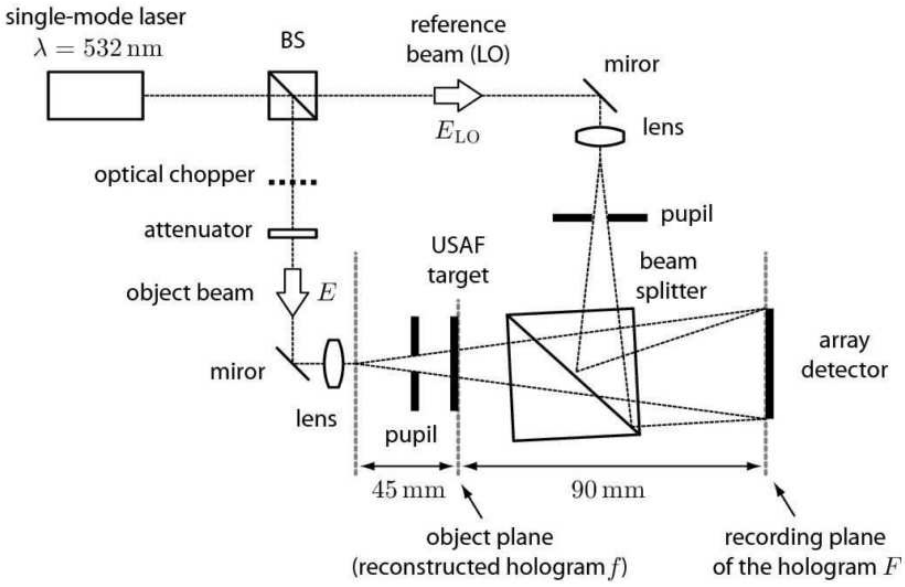


Figure I–3: Off-axis holography optical set-up proposed by [Marim11b] (the picture is reproduced from this publication).

plane. The interference pattern is then collected by a CCD or CMOS array detector. The measured image corresponding to the intensity I of the interfered light fields encodes the following signal:

$$I = |E + E_{\mathrm{LO}}|^2 = \underbrace{|E_{\mathrm{LO}}|^2}_{\text{Zero-order}} + \underbrace{|E|^2}_{\text{Real image}} + \underbrace{E \cdot \overline{E_{\mathrm{LO}}}}_{\text{Twin image}} + \underbrace{\overline{E} \cdot E_{\mathrm{LO}}}_{\text{Twin image}} \tag{I–13}$$

In an off-axis scheme, the object beam E and the reference beam E_{LO} reach the detector plane with different incidence angles, which makes the three components of the measured image I – namely the zero-order, the real image and the twin image – appear as separated in terms of spectral content: an appropriate band-pass filter applied to I permits to extract the real image component $y = E \cdot \overline{E_{\mathrm{LO}}}$. Finally, it can be shown that the transmission map x of the imaged object – which is the actual signal of interest – and the measured real image component y extracted from I are related through an optical Fresnel transform:

$$y(\xi, \eta) = \iint x(u, v) \exp\left(\frac{i\pi}{\lambda d}\left((u - \xi)^2 + (v - \eta)^2\right)\right) du\, dv \tag{I–14}$$

where λ is the wavelength of the radiation emitted by the laser, (u, v) and (ξ, η) are the spatial coordinates respectively in the object plane and in detector plane, and d is a length parameter characteristic of the set-up (see [Gross07, Cuche99] for more details about the off-axis holographic set-up).

When the imaged object presents some appropriate sparsity properties, [Marim11b] proposes to reduce the number of collected samples over the CCD/CMOS array detector: using a U.S. Air Force target as the object, the authors demonstrate that well-resolved

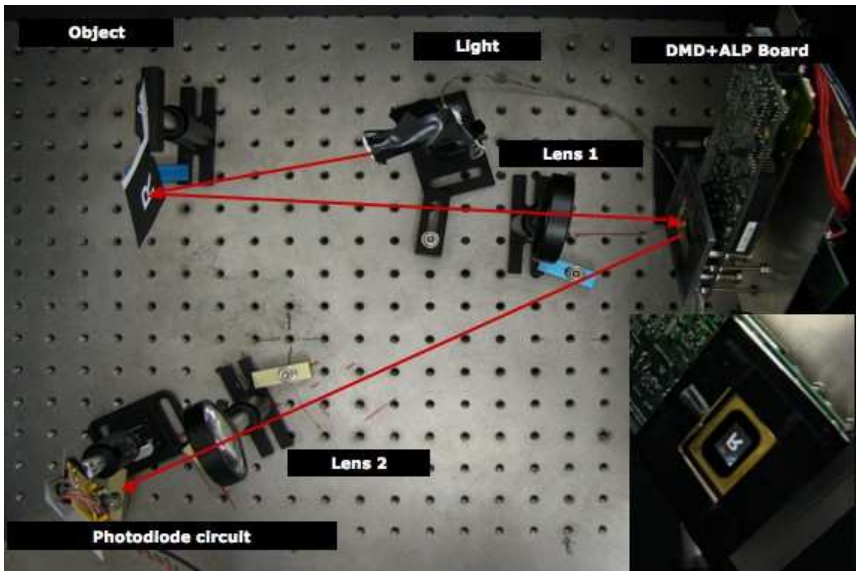


Figure I–4: Single-pixel camera set-up proposed by [Takhar06, Duarte08] (the picture is reproduced from the latter). The image of the object of interest is focused by the lens 1 on the digital micro-mirror device (DMD), which reflects it to the photodiode sensor.

transmission maps can be recovered using only 9% to 19% of all available pixels on the sensor. Up to a technical adaptation of the detector design that would avoid to acquire the remaining 81% to 91% pixel values, such compressed sensing scheme could theoretically speed up the acquisition operation, leading to faster CCD/CMOS sensors.

I.3.3 Single-pixel camera

In [Takhar06, Duarte08], an innovative image acquisition set-up – denoted as the *single-pixel camera* or *one-pixel camera* – is introduced as a proof of concept for a camera based on a single photodiode and implementing CS imaging. The motivation for such type of acquisition device is to design cameras that could operate in wavelength domains for which building arrays of sensors is technically unfeasible or highly expensive.

The principle of the single-pixel acquisition set-up is conceptually quite simple (see Fig. I–4): the observed object of interest is focused through a lens on a *digital micro-mirror device* (DMD), which reflects its image to the photodiode sensor back through another lens. The key ingredient here is the digital micro-mirror device: this instrument consists in an array of tiny mirrors; the orientation of each of these mirrors can be individually switched between two states. Thus, each facet of the DMD receives a small spatial fraction of the object of interest image (i.e. a pixel), and either reflects it or not toward the photodiode. Then, a full acquisition sequence consists into measuring the signal intensity collected by the photodiode for several configurations of the facets. Formally, the vector of collected samples y P R^M is related to the signal of interest x P R^N by the equation:

$$y = \Phi x \quad \text{with} \quad \Phi = \begin{bmatrix} \varphi_{0,0} & \varphi_{0,1} & \cdots & \varphi_{0,N-1} \\ \varphi_{1,0} & \varphi_{1,1} & \cdots & \varphi_{1,N-1} \\ \vdots & \vdots & & \vdots \\ \varphi_{M-1,0} & \varphi_{M-1,1} & \cdots & \varphi_{M-1,N-1} \end{bmatrix} \in \{0,1\}^{M \times N} \tag{I–15}$$

where each coefficient $\varphi_{k,l}$ characterizes the state of the lth facet during the acquisition of the kth sample: $\varphi_{k,l} = 1$ if it is oriented to reflect the image of the object toward the photodiode, and $\varphi_{k,l} = 0$ otherwise.

In this set-up, the acquisition time is directly related to the number M of samples that are necessary to retrieve the signal of interest x, as samples have to be acquired sequentially. In a naive approach, M would be set equal to the number N of pixels of the acquired image, so that Φ could be chosen as an invertible matrix (typically $\Phi = Id$ in a one by one pixel scan strategy). However, up to some appropriate sparsity assumption on the imaged object, the CS theory demonstrates that the number M of measurements can be significantly reduced while still allowing accurate reconstruction of the image x. Using a random acquisition matrix Φ and a reconstruction scheme enforcing sparsity of the Haar wavelet coefficients of x, [Takhar06] shows reconstructed images corresponding to ratios $\frac{M}{N}$ varying from 40% to 66%.

I.3.4 Schlieren deflectometry

Schlieren deflectometry is an imaging modality that aims at visualizing and measuring the deflection undergone by a light beam when it crosses a section of a thin transparent object. This type of measures can then be used to characterize some properties of the studied object, such as the curvature of its surface, or the distribution of its refractive index.

In [Sudhakar13], the authors propose to use results from the CS theory to improve the performance of a Schlieren deflectometer device. The principle of the corresponding optical set-up is described in Fig. I–5: on one side, the object is illuminated by a light source that undergoes a spatial modulation, the modulation pattern being controlled by a spatial light modulator (SLM); on the other side of the transparent object, the deflected light crosses a telecentric system and is collected by a standard CMOS/CCD array sensor. The goal of the telecentric system is to filter light emerging from the transparent object so that only light beams parallel to the optical axis of the system can actually reach the CMOS/CCD sensor: thus, each pixel p of the sensor collects the light that emerges at one particular point A_p of the surface of the transparent object. Then, the light intensity measured by the sensor pixel p appears as an inner product between the modulation image formed by the SLM device and a map $x_p \in \mathbb{R}^N$ that characterizes the deflecting properties of the surface of the object at point A_p: this map x_p is the signal of interest to recover. Similarly to the single-pixel camera, a full acquisition sequence consists in probing the light intensity on the sensor for several modulation patterns. Then, the relation between the

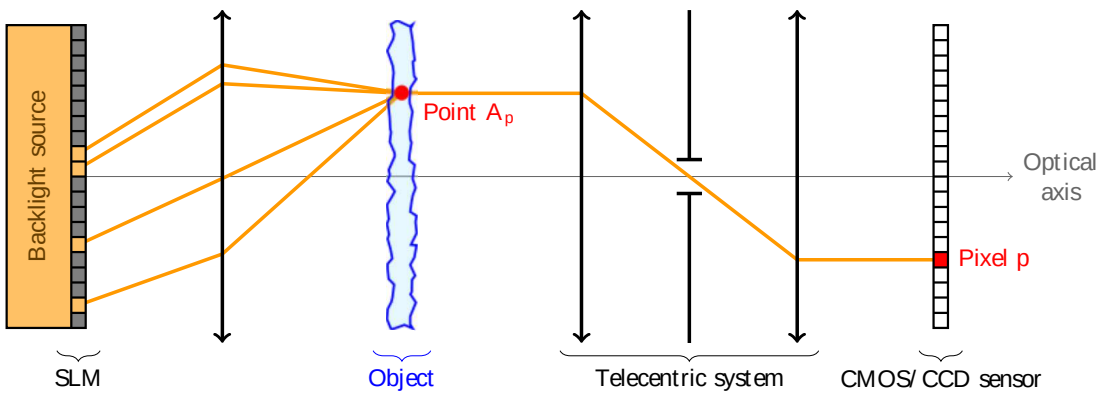


Figure I–5: Schlieren deflectometer set-up proposed by [Sudhakar13]. A thin transparent object is illuminated by a structured pattern controlled by a SLM device. The light received at a point A_p on the object surface is deflected toward the right. The telecentric system (made of two lenses and a pinhole) filters out the light beams that are not parallel to the optical axis. Finally, the digital sensor measures the output signal; thanks to the pinhole, each pixel p of the sensor collects only the light that emerges from the object at point A_p.

signal of interest $x_p \in \mathbb{R}^N$ and the vector of collected measures $y_p \in \mathbb{R}^M$ is the following:

$$y_p = \Phi x_p \quad \text{with} \quad \Phi \in \mathbb{R}^{M \times N} \tag{I–16}$$

where the kth row of the matrix Φ characterizes the modulation pattern formed by the SLM during the acquisition of the kth sample. The non-negativity of the entries of Φ accounts for the physical constraints imposed by the SLM device.

As noted in [Sudhakar13], each pixel p of the CMOS/CCD sensor used in this set-up behaves like a single-pixel camera for the deflecting map x_p that characterizes the surface of the object at point A_p, all the maps x_p being acquired in parallel. Then, as in the case of the single-pixel camera, the contribution of the CS theory to this deflectometer set-up is to allow the reduction of the number M of modulations patterns that are necessary to accurately estimate the deflecting maps x_p, by taking advantage of their sparsity properties. In [Sudhakar13], the authors present some deflecting maps acquired with a matrix Φ whose rows are made of vectors of the Hadamard basis, and reconstructed by enforcing a sparsity constraint on their Daubechies $9/7$ wavelet coefficients. Additionally, as proposed by [Puy12], the authors introduce a random modulation by ± 1 of the columns of Φ, to increase the incoherence between the sensing and the sparsity bases. The authors present some results obtained with compression ratios $\frac{M}{N} = 2.5\%$ and $\frac{M}{N} = 10\%$: these reconstructed deflecting maps capture the main features of an ideal map acquired with $\frac{M}{N} = 100\%$, although some significant differences can be noticed. These reconstruction errors are justified in [Sudhakar13] by the fact that the measurements collected by the Schlieren deflectometer set-up are quite noisy.

Chapter II

Practical resolution of CS reconstruction problems through convex optimization

The compressed sensing theory basically states that large signals $x \in \mathbb{C}^N$ can be recovered from a relatively small number M of linear measurements $y = \Phi x \in \mathbb{C}^M$, under some appropriate hypothesis. In practice, this property is used to design sensing devices with improved characteristics in terms of acquisition speed or sensor simplicity (see Sec. I.3). However, the price to pay for these improved properties that benefit the sensing devices is that the raw collected samples must undergo a heavy post-acquisition numerical processing so that the actual signal of interest can be recovered. Being able to perform this post-acquisition processing efficiently is therefore a crucial issue to make CS acquisition devices usable in practical situations.

In this chapter, we present the general reconstruction approaches used to post-process CS acquired data. As these reconstruction procedures are often formulated as convex optimality equations, we also review some of the algorithmic solutions that exist to solve the underlying optimization programs. Finally, we present a comparison of these algorithmic solutions, based on empirical evaluations of their performance in image reconstruction problems involving real microscopy images arising from biological applications.

II.1 CS reconstruction formulations **38**

II.1.1 Convex optimization formulations: classical form, BPDN and LASSO . 38

II.1.2 Alternative approach: orthogonal matching pursuit 40

II.2 Convex optimization algorithms **41**

II.2.1 SOCP methods 41

II.2.2 NESTA 43

II.2.3 RecPF 44

II.2.4 SPGL1 45

II.3 Compared performance **46**

II.3.1 Methodology 46

II.3.2 Simulation results and reconstruction time 48

II.4 Conclusion **48**

II.1 CS reconstruction formulations

II.1.1 Convex optimization formulations: classical form, BPDN and LASSO

In chapter I, we introduced the notion of CS reconstruction problem, namely the problem of reconstructing a signal $x \in \mathbb{C}^N$ that is sparse in some sense from a vector of linear measurements $y = \Phi x \in \mathbb{C}^M$. [Candès06b, Candès10] and related papers propose to formulate the CS reconstruction as an optimality search problem, and provide some theoretical reconstruction guaranties and error bounds for this formulation using the RIP framework (see Sec. I.2.2).

More precisely, the works presented in Sec. I.2 propose to address the CS reconstruction through one of the optimization problems (I–6), ($\mathcal{P}_{\text{analysis}}$), ($\mathcal{P}_{\text{synthesis}}$), or ($\mathcal{P}_{TV}$), depending on the actual sparsity properties of the signal to reconstruct. The general form of these optimization problems is the following[1]:

$$\underset{x \in \mathbb{C}^N}{\arg\min}\ f(x) \text{ subject to } \|\Phi x - y\|_2 \leq \eta \qquad \text{(CCSR)}$$

where the objective function $f(x)$ is a convex sparsity-promoting function, with typically $f(x) = \|x\|_1$ in the case of a synthesis reconstruction ($\mathcal{P}_{\text{synthesis}}$) or for signals that are sparse in the canonical basis of $\mathbb{C}^N$. In this formulation, the scalar parameter $\eta \geq 0$ controls the trade-off between the fidelity to the measurements and the desired level of sparsity. Its value is related to the noise level that affects the measurement vector y: more specifically, [Candès06b] defines this parameter such that η^2 is an upper bound of the noise power that corrupts y. In the case where y is corrupted by a white additive Gaussian noise of standard deviation σ, a common heuristic is to select $\eta = \sigma\sqrt{M + 2\sqrt{2M}}$ (see [Becker11]).

The (CCSR) problem belongs to the class of constrained convex problems, meaning that the optimized variable x is forced into a subset of the whole subspace $\mathbb{C}^N$, denoted as the *feasible domain*, implicitly defined by the constraint $\|\Phi x - y\|_2 \leq \eta$ in the case

[1] As this problem does not seem to have a consensual denomination, we will refer to it as the *Classical CS Reconstruction* problem, or (CCSR).

of (CCSR). As it is generally assumed that implementing an algorithmic solver for constrained optimization problems is more difficult than for unconstrained ones, some authors (see for instance [Lustig07, Marcia08, Provost09]) favor the following unconstrained formulation – denoted as *Basis Pursuit De-Noising* in most of the publications – for practical applications:

$$\underset{x \in \mathbb{C}^N}{\arg\min} \frac{1}{2} \|\Phi x - y\|_2^2 + \lambda f(x) \qquad \text{(BPDN)}$$

where $f(x)$ is the same sparsity-promoting function than for (CCSR), and where the scalar parameter $\lambda \geq 0$ plays the same role as τ in (CCSR).

It is worth mentioning that both problems (CCSR) and (BPDN) are equivalent in the following sense (a proof of this equivalence is given in [Weiss08], Theorem 2.7):

1. For a fixed parameter $\tau \geq 0$, let $\hat{x}_\tau^{CCSR}$ be a solution of (CCSR). Then, there exists a value $\lambda' \geq 0$ of the parameter λ for which $\hat{x}_\tau^{CCSR}$ is also a solution of (BPDN).
2. Reciprocally, for a fixed parameter $\lambda \geq 0$, if $\hat{x}_\lambda^{BPDN}$ denotes a solution of (BPDN), then there exists $\tau' \geq 0$ for which $\hat{x}_\lambda^{BPDN}$ is also a solution of (CCSR).

However, in general, these λ' and τ' depend on the other entities involved in the (CCSR) and (BPDN) problems, in particular the measurement matrix Φ and the observation vector y. In other words, for a given value of τ (resp. λ), there is no general method to determine a value λ' (resp. τ') that would make the solutions of both (CCSR) and (BPDN) be identical for any set of acquired samples y.

For this reason, while [Candès06b] and following papers provide some theoretical guaranties on the reconstruction error between an estimator $\hat{x}_\tau^{CCSR}$ obtained by solving (CCSR) and the "true" signal to recover, as far as we know such kind of result does not exist for the reconstruction formulation (BPDN). Therefore, we generally prefer to use the constrained formulation (CCSR) whenever possible.

Finally, let us mention a third convex problem that is related to (CCSR) and (BPDN), and known as *Least Absolute Shrinkage and Selection Operator*:

$$\underset{x \in \mathbb{C}^N}{\arg\min} \|\Phi x - y\|_2 \text{ subject to } f(x) \leq \tau \qquad \text{(LASSO)}$$

The (LASSO) formulation introduces a scalar parameter τ whose role is similar to the one of the parameters τ and λ defined above, i.e. τ controls the trade-off between the sparsity level of the solution and its fidelity to the measurements y; (LASSO) is also equivalent to both (CCSR) and (BPDN) in the same sense than mentioned above. This formulation can be preferred to (CCSR) and (BPDN) when some information about the sparsity level of the sought signal is available prior to the reconstruction: for instance, if $f(x) = \|x\|_1$ and if in the case of a particular problem an upper bound of the l_1-norm of the signal to reconstruct can be determined, then using the reconstruction formulation (LASSO) with τ set to this upper bound can be considered. However, as for (BPDN), there is no theoretical guaranties about the reconstruction error achieved using (LASSO).

function OMP(y, Φ)
  $p \leftarrow 0$, $\Lambda_0 \leftarrow \emptyset$, $x_0 \leftarrow 0$, $r_0 \leftarrow y$
  repeat
    $p \leftarrow p + 1$
    $k_p \leftarrow \arg\max_{k \in [0, N-1]} |h_k|$ where $h = \Phi^* r_p$ ▷ Identification step
    $\Lambda_p \leftarrow \Lambda_{p-1} \cup \{k_p\}$
    $x_p \leftarrow \arg\min_x \{y - \Phi x\}_2$ subject to $\mathrm{Supp}(x) \subset \Lambda_p$ ▷ Update step
    $r_p \leftarrow y - \Phi x_p$
  until stop condition
  return x_p
end function

Figure II–1: OMP algorithm to recover a signal $x \in \mathbb{C}^N$ from a vector $y = \Phi x \in \mathbb{C}^M$. The vector x to recover is supposed to be sparse in its canonical basis, i.e. $\|x\|_0 \ll N$.

II.1.2 Alternative approach: orthogonal matching pursuit

Although convex optimization is the original approach proposed for the CS reconstruction problem, other algorithmic formulations have been proposed to tackle it, based on the *orthogonal matching pursuit* (OMP) algorithm (see [Pati93, Tropp07]) or derived methods (see for instance [Needell09, Dai09]).

A description of the OMP algorithm is provided in Fig. II–1. The principle of this reconstruction procedure is to iteratively identify the support of the signal to reconstruct. At each iteration p of its main loop, the algorithm maintains:

1. a set $\Lambda_p \subset [0, N-1]$, which is an estimate of the support of the signal to recover;
2. an estimate x_p of the signal to recover, constructed such that its support lies in Λ_p;
3. a residual $r_p = y - \Phi x_p$.

The key point of the algorithm consists in selecting the new k_p element to add to the support estimate Λ_p so that the corresponding basis vector e_{k_p} best correlates to the current measurement residual; in other words, the new k_p element is selected as $\arg\max_k \langle \Phi e_k | r_{p-1} \rangle$: this selection is denoted as the *identification step*. The signal estimate x_p is then updated to minimize the energy of the corresponding residual r_p, while satisfying the support constraint $\mathrm{Supp}(x_p) \subset \Lambda_p$. The algorithm terminates when a certain stopping condition holds, which consists generally in requesting $\|x_p\|_0 = S$ where S is a targeted sparsity level, or $\|r_p\|_2 \leq \varepsilon$, where ε is a parameter controlling the tolerable residual energy.

One of the main interests of the OMP is that a full implementation of the method can be achieved very simply: as noted by [Davenport10b], translating the pseudo-code in Fig. II–1 into a Matlab® program requires approximately the same number of code lines. Compared to the programs required to solve the optimization problems arising in the convex relaxation approaches (see Sec. II.2), the OMP implementation is indeed much more straightforward. Moreover, if Φ is a partial unitary transform (see Sec. I.2.3) with an

associated fast computation algorithm, it appears that the execution of one OMP iteration can be achieved very efficiently: for instance, in the case of a partial Fourier transform, the OMP iteration has an algorithmic complexity of $O(N \log N)$. Finally, theoretical reconstruction guaranties have been obtained for the OMP algorithm: see for instance [Davenport10b] and references therein.

However, the overall algorithmic complexity to recover a S-sparse signal using the OMP algorithm is S times the complexity of one iteration: in the case of imaging applications, the additional S factor may be quite large, making the OMP algorithm rather inefficient. Moreover, similarly to the synthesis formulation ($P_{synthesis}$) used in convex relaxation, the OMP algorithm has to operate in the sparsity domain of the signal of interest if the latter is not sparse in its canonical basis: this potentially entails a performance issue if the signal to recover is sparse in a highly redundant dictionary. OMP reconstruction is also less flexible than convex optimization, since integrating a sparsity constraint such as the 2D total variation (I–10) – for which no synthesis formulation is available – is not feasible (as far as we known, there is no thing such as a "total-variation driven OMP reconstruction algorithm"). For these reasons, we mostly focus on CS reconstruction through convex optimization in this manuscript.

II.2 Convex optimization algorithms

We focus now on the CS reconstruction formulations based on convex optimization (mostly (CCSR)), and propose a short review of the algorithmic solutions that have been designed to solve such convex optimization problems. We will assume in this section that the studied signals are real-valued, as convex optimization solvers are generally presented in this context[2].

II.2.1 SOCP methods

Solving (CCSR) is indeed a challenging task, at least for two reasons:

- the sparsity-promoting objective function $f(x)$ involves one or several non-smooth terms (for instance $\|\cdot\|_1$, $\|\cdot\|_{TV}$ (I–10), $\|\cdot\|_{1,2}$ (I–11));
- the space $\mathbb{C}^N$ (or $\mathbb{R}^N$) in which the sought signal is defined is very large in the case of imaging applications (typically $N \geq 10^6$).

The first issue can be tackled by recasting (CCSR) into an appropriate form. For instance, for $f(x) = \|x\|_{TV}$ where the 2D total variation semi-norm $\|\cdot\|_{TV}$ is defined as in (I–10), (CCSR) can be recast into the following convex form, denoted as a *second-order cone*

[2] In the presented methods, complex-valued signals can often be handled as real-valued signals with twice more components, i.e. by encoding separately the real and the imaginary parts of their components.

program (see [Boyd04])[3]:

$$\underset{\substack{x \in \mathbb{R}^N \\ n \in \mathbb{R}^N}}{\arg\min} \sum_k n_k \text{ subject to } \begin{cases} (D_h x)_k^2 + (D_v x)_k^2 \leq n_k^2 \text{ and } n_k \geq 0 \text{ for all } k \\ \|\Phi x - y\|_2^2 \leq \varepsilon^2 \end{cases} \tag{II–1}$$

Compared to (CCSR) with $f(x) = \|x\|_{TV}$, both the objective function and the feasible domain of (II–1) are smooth. However, the price to pay for this regularization is that the optimized variable is now the tuple (x, n), which belongs to a space whose dimension is $2N$, i.e. twice bigger than for the original problem (CCSR): the second issue mentioned above becomes therefore even more challenging!

This approach consisting in recasting (CCSR) into a second-order cone program has been implemented in a CS-dedicated Matlab® toolbox called l_1-magic [Candès05a], released at the same time as the early theoretical CS papers and by the same authors (see Sec. I.2.2). The associated method proposed in [Candès05a] to solve the second-order cone program (II–1) can be summarized as follows:

1. First, the constrained problem (II–1) is transformed into a sequence of unconstrained problems using a *log-barrier* method, i.e. by injecting each inequality constraint as a logarithmic penalty into the objective function:

$$\underset{X \in \mathbb{R}^{2N}}{\arg\min}\, h_p(X) \quad \text{with} \quad X = (x, n) \quad \text{and} \quad h_p(X) = \sum_k n_k - \frac{1}{\alpha_p} \sum_l \log(-g_l(x, n)) \tag{II–2}$$

   where each of the functions g_l represents one of the inequality constraint that defines the feasible domain of (II–1) ($g_l(x, n) \leq 0$ if (x, n) is a feasible point), and $(\alpha_p)_{p \in \mathbb{N}}$ is an increasing sequence of positive scalars. If $\hat{X}_p$ denotes the solution of (II–2) corresponding to the log-barrier parameter α_p, it can be shown that $\lim_{p \to \infty} \hat{X}_p = \hat{X}$ where $\hat{X}$ is the solution to (II–1).

2. Then, for each value α_p of the log-barrier parameter, (II–2) is solved using Newton's method: a sequence of estimates $(X_{p,q})_{q \in \mathbb{N}}$ is constructed sequentially such that, for all q, $X_{p,q+1}$ is the point that minimizes the second-order Taylor approximation of the function h_p at point $X_{p,q}$:

$$X_{p,q+1} = \underset{X \in \mathbb{R}^{2N}}{\arg\min}\, h_p(X_{p,q}) + (X - X_{p,q})\nabla h_p(X_{p,q}) + \frac{1}{2}\left(X - X_{p,q}\right)^T \nabla^2 h_p(X_{p,q})(X - X_{p,q}) \tag{II–3}$$

   It can be shown that $\lim_{q \to \infty} X_{p,q} = \hat{X}_p$.

3. Problem (II–3) has an algebraic solution: $X_{p,q+1} = -\left(\nabla^2 h_p(X_{p,q})\right)^{-1} \nabla h_p(X_{p,q})$; however, as the dimension of the problem gets very large in the case of imaging

[3] Other types of objective functions $f(x)$ may lead to different formulations, but for common choices of $f(x)$ such as $\|x\|_1$, $\|\Psi^* x\|_1$, $\|x\|_{1,2}$, the resulting problem is also a second-order cone program, or even a linear program in some special situations (for instance: $f(x) = \|x\|_1$, $\Phi = 0$ and x real-valued). Here, we choose to describe the principle of the method in the case $f(x) = \|x\|_{TV}$ instead of in the general case, for the sake of simplicity.

applications, direct evaluation of the inverse of the Hessian matrix $\nabla^2 h_p(x_{p,q})$ is not feasible. Therefore, [Candès05a] proposes to use a conjugate-gradient method (see [Hesteness52]) to evaluate iteratively the solution $x_{p,q+1}$ to the quadratic problem (II–3).

The method proposed by [Candès05a] to solve (II–1) thus consists in an iterative procedure involving three levels of nested loops: the performance obtained with this scheme is rather poor, especially when used for 2D image reconstruction. Therefore, specialized algorithms were subsequently developed starting in the late-2000s to handle the CS reconstruction problem in a more efficient way.

II.2.2 NESTA

One of these specialized solutions is the NESTA algorithm, introduced in [Becker11]. This algorithm is based on the general framework developed in [Nesterov07] for the minimization of composite objective functions. It addresses the constrained problem (CCSR) with either $f(x) = \|x\|_{TV}$ or $f(x) = \|\Psi^* x\|_1$ without specific requirements on the sparsity matrix Ψ, and consists in an accelerated gradient descent with back-projection on the feasible set.

More precisely, the NESTA algorithm proceeds in two steps:

1. First, the non-smooth objective function $f(x)$ is approximated by a smooth function. This step takes advantage of the fact that the targeted functions $f(x)$ can be written as[4]:

$$f(x) = \max_{z \in Q} \langle z, W x \rangle \tag{II–4}$$

   where $W \in \mathbb{R}^{L \times N}$ and Q is a convex subset of $\mathbb{R}^L$. This type of function $f(x)$ belongs to a larger class of functions, introduced by [Nesterov04] and denoted as *max functions* by [Weiss08]; as a max function, [Nesterov04] shows that $f(x)$ can be approximated by a function $f_\mu(x)$ defined as follows:

$$f_\mu(x) = \max_{z \in Q} \langle z, W x \rangle - \frac{\mu}{2} \|z - z_0\|_2^2 \tag{II–5}$$

   where $\mu > 0$ is a scalar parameter, and $z_0 \in Q$. In [Nesterov04], it is shown that the error $|f(x) - f_\mu(x)|$ can be bounded uniformly by a constant proportional to μ: the smaller this parameter, the better the approximation; moreover, $f_\mu(x)$ is Lipschitz-differentiable with a Lipschitz constant $L_\mu = \frac{1}{\mu} \|W\|^2$, and its gradient has an explicit expression (see [Nesterov04] or [Weiss08] for more details).
2. The second step consists in solving (CCSR) where $f(x)$ is replaced with $f_\mu(x)$. To proceed, [Becker11] proposes to use an improved version of the well-known gradient descent method with back-projection on the feasible set at each step of the descent (see [Levitin66]): this improved version was introduced in [Nesterov07], which

[4] For instance, for $f(x) = \|\Psi^* x\|_1$, $W = \Psi^*$ and $Q = \{u \text{ such that } \|u\|_\infty \leq 1\}$ (l_∞ unit ball).

demonstrates that modifying the descent direction to account not only for the direction of the local gradient but also for the gradient directions encountered in the previous steps speeds up the convergence of the algorithm. A generic description of this improved gradient descent method with back-projection is given in [Weiss08] (Algorithm 4.3).

A potentially expensive step of the NESTA algorithm is the evaluation of the projection operator Π on the feasible set – which has to be performed twice per descent step – defined as follows:

$$\Pi(x) = \underset{z \in \mathbb{R}^N}{\arg\min} \|z - x\|_2 \text{ subject to } \|\Phi z - y\|_2 \leq \delta \tag{II–6}$$

However, [Becker11] shows that, in the case where the sensing operator Φ is such that $(\Phi^* \Phi)^2 = \Phi^* \Phi$ (i.e. $\Phi^* \Phi$ is a linear projector), the problem (II–6) has an algebraic solution that can be evaluated efficiently[5]:

$$\Pi(x) = \left(Id - \frac{\lambda}{1+\lambda}\Phi^* \Phi\right)(x + \lambda \Phi^* y) \quad \text{with} \quad \lambda = \max\left(0, \frac{1}{\delta}\|\Phi x - y\|_2 - 1\right) \tag{II–7}$$

In particular, the expression (II–7) does not involve any matrix inversion. Moreover, the condition $(\Phi^* \Phi)^2 = \Phi^* \Phi$ does hold for the matrices Φ whose row vectors form an orthogonal family, which includes in particular the partial unitary transforms (see Sec. I.2.3).

II.2.3 RecPF

In [Yang10], the RecPF specialized algorithm was introduced to solve the (BPDN)-formulated CS reconstructions problem, in the case where the two following additional hypotheses hold:

1. $f(x)$ must be either $\|x\|_{TV}$, $\|\Psi^* x\|_1$ with Ψ a tight frame, or a linear combination of both,
2. the sampling matrix Φ must be a partial Fourier transform[6].

This second requirement makes the RecPF algorithm only applicable for imaging modalities where the sensing operations occurs in the Fourier domain of the signal of interest (for instance, MRI, see Sec. I.3.1). However, this extreme specialization allows the use of several tricks and optimizations, which result in a very fast resolution method.

For the sake of simplicity, we focus here on the case where $f(x) = \|x\|_{TV}$. Then, the principle of the RecPF algorithm is to solve the following problem:

[5] Please note that the notations used in (II–7) are slightly different than those used in [Becker11]. In particular, the Lagrange multiplier λ in (II–7) corresponds to $\frac{1}{L\mu}$ times the Lagrange multiplier λ defined in [Becker11]. Our definition of λ allows an extra-simplification in the expression of the projection operator Π.

[6] The authors mention that their RecPF algorithm can be adapted to also handle the case where Φ is a partial cosine transform. Their Matlab® implementation of the RecPF algorithm however only supports partial Fourier transforms.

$$\underset{\substack{x \in \mathbb{R}^N \\ d_h \in \mathbb{R}^N \\ d_v \in \mathbb{R}^N}}{\arg\min} \frac{1}{2}\|\Phi x - y\|_2^2 + \lambda \sum_k \sqrt{(d_h)_k^2 + (d_v)_k^2} + \frac{\beta}{2}\left(\|d_h - D_h x\|_2^2 + \|d_v - D_v x\|_2^2\right) \quad \text{(II–8)}$$

where λ is the scalar parameter introduced in (BPDN), D_h and D_v are the derivation operators introduced in (I–10), and $\beta \geq 0$ is a new scalar parameter, whose ideal value is as large as possible[7]: indeed, one can observe that (BPDN) with $f(x) = \|x\|_{TV}$ and (II–8) become equivalent when $\beta \to +\infty$. The method proposed by [Yang10] to solve the unconstrained problem (II–8) consists then in performing alternated minimizations of the objective function with respect to the (d_h, d_v) variables for a fixed x on the one hand, and with respect to x for fixed (d_h, d_v) on the other hand:

- The minimization with respect to (d_h, d_v) is straightforward, as the objective function of (II–8) is separable in each of the pair of variables $((d_h)_k, (d_v)_k)$: this step has therefore a complexity of $O(N)$ operations.
- The minimization with respect to x consists in finding the minimum of a quadratic function, which is equivalent to inverting the following system:

$$\left(\Phi^* \Phi + \beta D_h^* D_h + \beta D_v^* D_v\right) x = \Phi^* y + \beta D_h^* d_h + \beta D_v^* d_v \quad \text{(II–9)}$$

  Solving this system requires to invert the $N \times N$ matrix $\left(\Phi^* \Phi + \beta D_h^* D_h + \beta D_v^* D_v\right)$, which is potentially a very expensive operation; however, it appears that this matrix is diagonal in the Fourier basis as:

  - both D_h and D_v are convolution operators,
  - Φ is required to be a partial Fourier transform (see in Sec. I.2.3 the decomposition available for $\Phi^* \Phi$ in this case).

  The inversion of (II–9) becomes therefore trivial, and the complexity of this operation is dominated by the cost of the Fourier transforms involved in the change of basis, i.e. $O(N \log N)$.

II.2.4 SPGL1

The SPGL1 algorithm introduced in [Van Den Berg08] proposes an original approach to solve the (CCSR) problem, in the case where $f(x) = \|x\|_1$. As no other particular hypothesis is required on Φ, the SPGL1 algorithm can handle either signals that are sparse in their canonical basis, or signals that can be recovered by following a synthesis approach ($P_{synthesis}$); however, it cannot be adapted to cover the TV minimization case (P_{TV}). The key idea of the SPGL1 algorithm is to solve several instances of the (LASSO) problem for different values of the parameter τ introduced by this formulation, until the corresponding solution $\hat{x}_\tau^{LASSO}$ of (LASSO) is such that the equality $\|\Phi \hat{x}_\tau^{LASSO} - y\|_2 = \sigma$ holds: indeed,

[7] The principle of the RecPF algorithm is actually very similar to the one of the Alternating Direction Method of Multipliers (ADMM) algorithm. See for instance [Combettes11] (Algorithm 6.4) and references therein for more details about this type of methods.

it can be easily demonstrated that, if $f(x)$ is positive-homogeneous, then any solution $\hat{x}$ of (CCSR) saturates the underlying constraint $\|\Phi x - y\|_2 \leq \varepsilon$ (i.e. $\|\Phi \hat{x} - y\|_2 = \varepsilon$), except in the trivial case where the zero vector 0 is solution.

More precisely, if $\chi(\tau)$ denotes the optimal value of the objective function in (LASSO) (i.e. $\chi(\tau) = \|\Phi \hat{x}_\tau^{LASSO} - y\|_2$), the approach proposed by [Van Den Berg08] consists in finding a root τ^* of the equation $\chi(\tau) = \varepsilon$: the authors show that the underlying $\hat{x}_{\tau^*}^{LASSO}$ is then the solution of (CCSR). They first demonstrate several properties of the function χ, in particular that it is differentiable and that its derivative has an explicit expression involving the solution $\hat{x}_\tau^{LASSO}$ of (LASSO): thanks to these properties, the authors propose to use the Newton's root finding method (see for instance [Press07]) to estimate the actual value τ^* that solves $\chi(\tau) = \varepsilon$. To proceed, several instances of the (LASSO) problem have to be solved: the method proposed by the authors for this task consists in gradient descent with back-projection on the feasible set at each step of the descent (see [Levitin66, Weiss08]). A key issue that conditions the efficiency of this approach is the choice of the method used to evaluate those projections on the feasible set, or in other words the method used to evaluate the following operator Π:

$$\Pi(x) = \underset{z \in \mathbb{R}^N}{\arg\min} \|z - x\|_2 \text{ subject to } \|z\|_1 \leq \tau \tag{II–10}$$

However, as noted by [Van Den Berg08], it turns out that (II–10) has an explicit solution, which is evaluated by applying the soft-thresholding to the input x, with a threshold value that can be computed in $O(N \log N)$ basic operations.

II.3 Compared performance

We have compared the performance of the NESTA, RecPF and SPGL1 algorithms in reconstructing various biological images from simulated Fourier CS measurements. This work was presented in the conference papers [Le Montagner11a] and [Le Montagner11b].

II.3.1 Methodology

We tested the three reconstruction algorithms on a set of seven biological images with various characteristics in terms of biological content, noise level, and size. For each image, we generated a vector of Fourier measurements, by selecting 15% of their Fourier coefficients. This selection of Fourier coefficients was performed randomly, following a so-called *Gaussian sampling strategy* (see Sec. III.1.2).

Two regularization energies were used for each tested image: minimization of the TV semi-norm, and minimization of the l_1-norm of the Daubechies-4 wavelet coefficients. For the TV-based reconstructions, we tested the NESTA and RecPF algorithms, while for the l_1-based reconstructions, we tested NESTA and SPGL1. The stop conditions were set

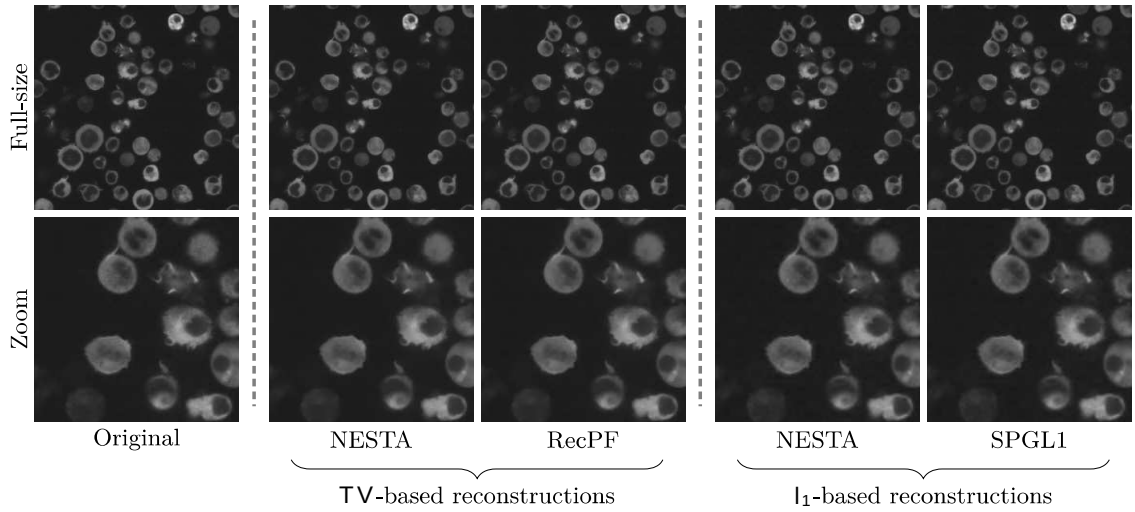


Figure II–2: Example of reconstruction results obtained for the *Lymphocytes T* test image, in the conditions described in Sec. II.3.1. For a given regularization energy, the reconstructed images present no visual differences whatever the algorithm actually used.

in the same manner for all the three algorithms, using a criteria based on the relative variation between two consecutive iterates.

In each configuration, we evaluated:

- the final value of the objective function ($\|x\|_{TV}$ or $\|\Psi^{\circ} x\|_1$),
- the number of iterations required for convergence,
- the execution time.

The notion of iteration is obviously algorithm dependent: it is one gradient descent step for NESTA, one pair of minimizations with respect to (d_h, d_v) and with respect to x for RecPF, and one gradient descent step occurring in one (LASSO) sub-problem in the case of SPGL1. For each of these three algorithms, one iteration has a $O(N \log N)$ complexity, although their actual cost in terms of computation time may be different. The number of iterations however gives an idea of the practical convergence speed of the algorithms, that is independent of the computational power of the computer used for the simulations.

All the simulations were performed using Matlab®, with implementations of the algorithms provided by their respective authors. A particular procedure was adopted to handle RecPF, as this algorithm solves (BPDN) instead of (CCSR): for each simulation, we adjusted the λ parameter so that the solution $\hat{x}_{\lambda}^{BPDN}$ returned by RecPF is such that $\|\Phi\hat{x}_{\lambda}^{BPDN} - y\|_2 = \epsilon \pm 2\%$, where ϵ is the parameter involved in the data term constraint in (CCSR); the reported reconstruction times do not take into account this parameter adjustment procedure. This point illustrates however one of the drawback of RecPF – and more generally of all (BPDN)-based CS reconstruction methods, which is that adjusting the parameter λ involved in (BPDN) is not straightforward, contrary to the parameter ϵ involved in (CCSR) (see Sec. II.1.1).

II.3.2 Simulation results and reconstruction time

An example of reconstruction results obtained for one of the test images is presented in Fig. II–2, while more comprehensive quantitative results are presented in Fig. II–3.

A first remark that can be drawn about the results presented in Fig. II–3 is that the tested algorithms reach similar levels for the respective objective functions being minimized. As the underlying problems are convex, this is actually an expected result: there is no such thing as local minima that could trap the algorithms here. A visual inspection of the reconstructed images do confirm that, for a given regularization energy, they do not present significant differences whatever the algorithm actually used.

On the contrary, it can be noticed that large differences exist between the algorithms in terms of computation time. Although the exact ratios between these computation times depend on the processed images, it can be observed that:

- first, RecPF is faster than NESTA by a factor 15 to 20 for the two smallest tested images (about $400\hat{}\ 400$ pixels), and this factor tends to increase for larger images;
- then, NESTA is faster than SPGL1 by a factor of at least 2, while this factor can be significantly larger with some large images, for which SPGL1 seems to converge very slowly (the number of iterations necessary to converge is also unexpectedly high).

This latest observation is concordant with some results presented in [Becker11], where the authors notice that SPGL1 can be very fast in certain circumstances (even faster than NESTA), but this execution time could vary significantly depending on the input signal, even for signals with identical size; on the contrary, the execution time of the NESTA algorithm is quite stable for a given input signal size.

The advantage of RecPF over NESTA can be explained by the deeper degree of specialization of the former, which is limited to partial Fourier transform sensing matrices Φ. It is also worth mentioning that the Matlab® implementations of these two algorithms use different mechanisms:

- RecPF makes use of the MEX function features provided by Matlab®, i.e. part of the code is written in C language and compiled, which potentially speeds up its execution,
- on the contrary, NESTA is written in pure Matlab® code.

Drawing final conclusions between the comparative execution time of these two algorithms would require to use similar implementation languages.

II.4 Conclusion

In this chapter, we detailed the post-processing operations that have to be carried on to recover a signal of interest for raw data acquired using a CS-based strategy. We reviewed

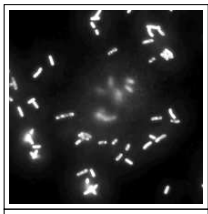
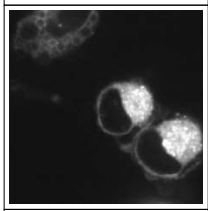
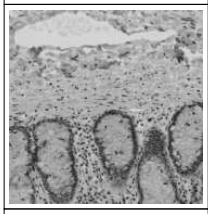
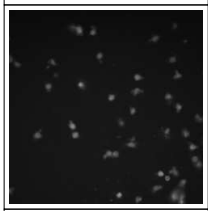
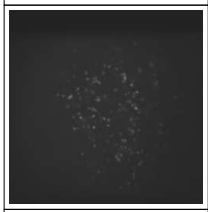
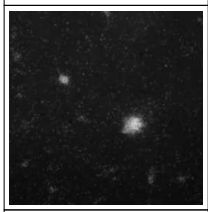
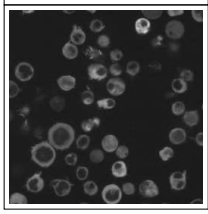
		TV-based reconstructions		l_1-based reconstructions	
		NESTA	RecPF	NESTA	SPGL1
	425 ˆ 425 pixels	Obj. fun.: 2230 #iter.: 136 Time: 14 sec.	Obj. fun.: 2254 #iter.: 26 Time: 1 sec.	Obj. fun.: 3977 #iter.: 73 Time: 10 sec.	Obj. fun.: 3985 #iter.: 283 Time: 47 sec.
	453 ˆ 453 pixels	Obj. fun.: 1115 #iter.: 171 Time: 23 sec.	Obj. fun.: 1169 #iter.: 17 Time: 1 sec.	Obj. fun.: 3154 #iter.: 69 Time: 12 sec.	Obj. fun.: 3157 #iter.: 80 Time: 24 sec.
	716 ˆ 716 pixels	Obj. fun.: 13.6 ˆ 10^3 #iter.: 86 Time: 23 sec.	Obj. fun.: 13.7 ˆ 10^3 #iter.: 15 Time: 1 sec.	Obj. fun.: 35.4 ˆ 10^3 #iter.: 63 Time: 21 sec.	Obj. fun.: 35.4 ˆ 10^3 #iter.: 276 Time: 132 sec.
	960 ˆ 960 pixels	Obj. fun.: 1384 #iter.: 206 Time: 78 sec.	Obj. fun.: 1486 #iter.: 23 Time: 2 sec.	Obj. fun.: 6475 #iter.: 63 Time: 38 sec.	Obj. fun.: 6400 #iter.: 90 Time: 119 sec.
	992 ˆ 992 pixels	Obj. fun.: 815 #iter.: 198 Time: 91 sec.	Obj. fun.: 981 #iter.: 14 Time: 2 sec.	Obj. fun.: 13.1 ˆ 10^3 #iter.: 63 Time: 46 sec.	Obj. fun.: 13.0 ˆ 10^3 #iter.: 577 Time: 987 sec.
	1024 ˆ 1024 pixels	Obj. fun.: 3315 #iter.: 159 Time: 75 sec.	Obj. fun.: 3504 #iter.: 26 Time: 3 sec.	Obj. fun.: 10.2 ˆ 10^3 #iter.: 66 Time: 47 sec.	Obj. fun.: 10.1 ˆ 10^3 #iter.: 3103 Time: 5823 sec.
	1024 ˆ 1024 pixels	Obj. fun.: 4962 #iter.: 179 Time: 82 sec.	Obj. fun.: 5177 #iter.: 20 Time: 2 sec.	Obj. fun.: 15.6 ˆ 10^3 #iter.: 68 Time: 47 sec.	Obj. fun.: 15.7 ˆ 10^3 #iter.: 1693 Time: 3176 sec.

Figure II–3: Quantitative results obtained with the presented algorithms for seven test images reconstructed in the conditions described in Sec. II.3.1. In each case, the table shows the final value reached on the objective function ("Obj. fun."), the number of iterations required for convergence ("#iter."), and the execution time ("Time").

and assessed some of the algorithmic solutions that exist to perform these post-processing operations: we illustrated on a set of experiments that these solutions present large differences in terms of performance. In the perspective of our work on CS, the NESTA algorithm shows an interesting trade-off between flexibility and execution speed. Besides, the understanding of the internal machinery involved in this algorithm enables to modify and adapt it to meet some specific requirements: the integration of 3D total variation (see Chap. IV) is an example of such a modification that we have made on the algorithm.

Chapter III

Effect of the sampling parameters in the Fourier space

When undergoing reconstruction through a compressed sensing scheme, images are affected by various artifacts that degrade them in different ways and lead to detail loss. The RIP framework provides an upper bound of the global l_2 reconstruction error (I–7), but this result does not account for the nature of the artifacts introduced in the reconstructed image $\hat{\mathsf{x}}$ due to the lack of samples. When the sampling is performed in the Fourier space of the image of interest, two sampling parameters may influence these artifacts: first, the number of Fourier samples that are acquired (or equivalently the sampling rate); second, the sampling strategy, i.e. the position in the Fourier space where – for a given budget of measurements – the samples are chosen to be acquired.

In this chapter, we discuss how these two parameters affect the CS reconstructed images. We first review the different existing sampling strategies, before focusing more thoroughly on two of them (the uniform and the Gaussian sampling strategies). Using simulations involving some test images, we analyze how variations of the number of acquired samples affect the reconstruction in these two cases. We conclude by extending our results to real microscopy images.

III.1 Sampling strategies in the Fourier space **52**

III.1.1 Introduction 52

III.1.2 Existing Fourier sampling strategies 52

III.2 Numerical evaluation of an optimal sampling rate **56**

III.2.1 Random uniform sampling 56

III.2.1.1 Problematic 56

III.2.1.2 Simulations on isotropic shapes 57

III.2.1.3 Optimal sampling rate and sparsity 60

III.2.1.4 Optimal sampling rate and shape factor 60

III.2.1.5 Putting things together 61
III.2.2 Random Gaussian sampling 63
III.2.3 Realistic images 65

III.3 Conclusion **67**

III.1 Sampling strategies in the Fourier space

III.1.1 Introduction

CS theoretical results predict that sparse signals can be recovered from partial Fourier data, assuming that a sufficient number of measurements is available with respect to the sparsity level and the size of the signal to recover. Results presented in [Candès07] (see Sec. I.2.3) propose a quantitative upper bound on this necessary number of measurements, in the case where samples are acquired at uniform random positions in the Fourier space.

However, for CS image acquisition, a simple experiment shows that uniform Fourier sampling is sub-optimal. In Fig. III–1, we present the reconstruction results obtained for the well-known Shepp-Logan phantom image based on two different sampling strategies of the Fourier space: starting from a noisy version x_0 of Shepp-Logan degraded with an additive Gaussian noise (left image in Fig. III–1), we generated two measurement vectors $y_u = \Phi_u x_0$ and $y_g = \Phi_g x_0$, with Φ_u and Φ_g two partial Fourier transforms corresponding respectively to a uniform and a Gaussian sampling mask. From these measurement vectors, two reconstructed images $\hat{x}_u$ and $\hat{x}_g$ were obtained by solving the TV reconstruction scheme (P_{TV}). Although the number of samples was identical in both masks, the reconstructed images $\hat{x}_u$ and $\hat{x}_g$ are completely different: while the Gaussian sampling leads to a reconstruction $\hat{x}_g$ that is almost identical to the original Shepp-Logan, $\hat{x}_u$ exhibits significant artifacts, with all but the largest structures lost.

This experiment illustrates the fact that all Fourier samples seem not to carry the same amount of information: in particular, having a higher sampling density in the low-frequency area of the Fourier domain improves dramatically the quality of the reconstruction.

III.1.2 Existing Fourier sampling strategies

Several sampling strategies have been proposed to perform acquisitions in the Fourier space, either based on empirical observations as in Sec. III.1.1, or recently as the result of more theoretical works (see below). As Fourier sampling finds an application with magnetic resonance imaging (see Sec. I.3.1), it appears that the proposed sampling strategies are generally designed to meet the requirements associated to this imaging modality, especially the need to have continuous sampling paths (see [Chauffert13a]). Among the works that address this issue, we can mention the following papers:

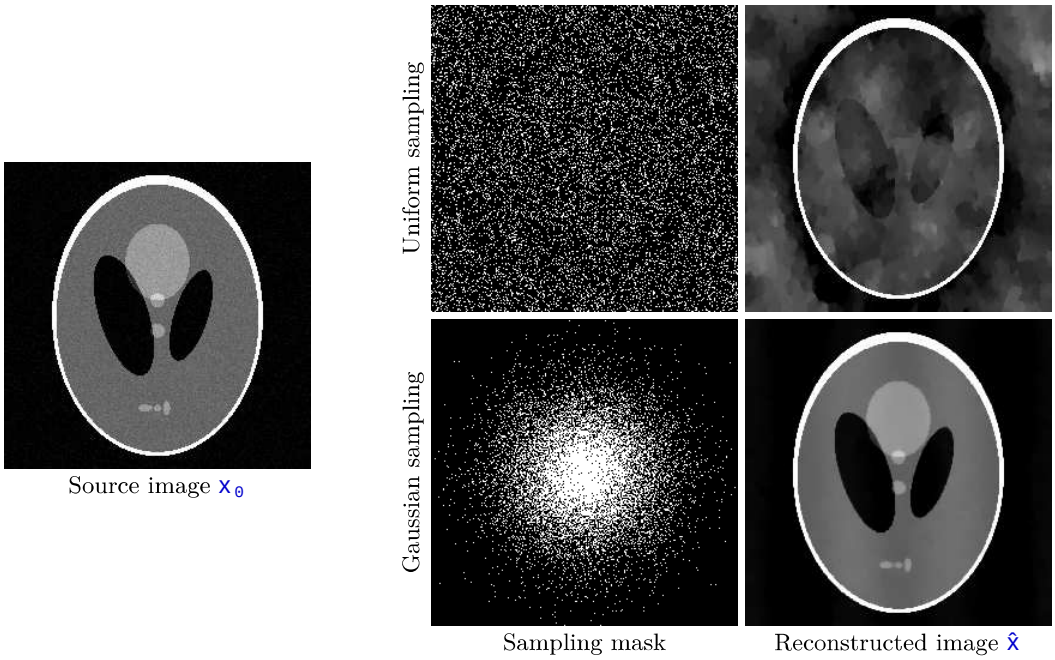


Figure III–1: Two reconstructions of the same noisy version of the Shepp-Logan phantom image, using two different sampling strategies in the Fourier domain. The reconstructions were performed using the TV minimization scheme (P_{TV}). In both cases, the number of Fourier samples used is the same (15% of the number of pixels of the input image), but their location is different: the sampling masks show (in white) the positions in the Fourier space that were sampled.

- In [Candès06a], the authors use a star-shaped sampling pattern when demonstrating the practical recovery capabilities of CS schemes applied to images (see Fig. III–2, example (d)).
- [Lustig07] suggests that the sampling density should be scaled *"according to a power of the distance from the origin"*. In [Chauffert13b], this formulation is interpreted as the fact that the probability $\pi(\bar{\omega})$ to sample the Fourier coefficient corresponding to the spatial frequency $\bar{\omega}$ is given by:

$$\pi(\bar{\omega}) = p_0 \left(1 - \frac{\|\bar{\omega}\|_2}{\omega_{max}}\right)^{\alpha} \tag{III–1}$$

  where $\alpha \geq 0$ is a parameter controlling the spread of the sampling probability distribution (the larger α, the more the distribution is concentrated close to the center of the Fourier space), ω_{max} is the maximum amplitude value for the spatial frequencies $\bar{\omega}$, and $p_0 \geq 0$ is tuned according to the targeted overall sampling rate. [Lustig07] suggests to select α between 1 and 6 based on empirical observations. We will refer to this family of sampling strategies as the *polynomial sampling strategies* (see Fig. III–2, examples (b) and (f)).
- In a context of 3D MRI, [Kim09] proposes a non-uniform sampling pattern having a small fully sampled area close to the k-space center, while the remaining

high-frequency region is sampled in a uniform random manner (see Fig. III–2, example (e)).

- [Wang10] presents a comparative study of several sampling strategies, including the star-shaped sampling pattern and various spiral-shaped sampling patterns. Based on a theoretical study on the energy repartition in the Fourier domain of 2D images that are sparse in the wavelet domain (which however is not shown to be directly related to the efficiency of CS sampling and reconstruction schemes), the authors also introduce a non-uniform random strategy with an exponential decay of the local sampling density with respect to the distance from the origin. In this latter sampling strategy, the probability $\pi(\omega)$ of sampling the Fourier coefficient ω is given by:

$$\pi(\omega) = \exp\left(-\frac{\|\omega\|_2}{\rho}\right)^{\alpha} \tag{III–2}$$

  where $\alpha \geq 0$ controls the spread of the sampling distribution and $\rho \geq 0$ is tuned according to the targeted overall sampling rate. [Wang10] advocates for $\alpha = 3.5$ based on empirical observations. While we haven't performed a thorough optimization of this parameter, we observed that setting $\alpha = 2$ – which corresponds to a more spread distribution than $\alpha = 3.5$ – also leads to interesting results. We will refer to this family of strategies as the *exponential sampling strategies*, and as the *Gaussian sampling strategy* in the particular case $\alpha = 2$ (see Fig. III–2, examples (c) and (g)).

- In [Puy11], relying on some theoretical results from [Rauhut10], the authors suggest that the search for an optimal sampling strategy should be driven by the minimization of a modified version of mutual coherence coefficient (I–9) between the sampling and the sparsity bases. More precisely, if π_k denotes the probability of sampling the kth Fourier coefficient, they propose to construct the optimal map of sampling probabilities π^* as:

$$\pi^* = \arg\min_{\pi} \max_{k,l} \frac{|\langle \varphi_k, \psi_l \rangle|}{\sqrt{\pi_k}} \quad \text{subject to} \quad \begin{cases} 0 \leq \pi_k \leq 1 & \text{for all } k \\ \sum_k \pi_k = M \end{cases} \tag{III–3}$$

  where $\varphi_0, \varphi_1, \ldots, \varphi_{N-1}$ are the rows of the sampling matrix Φ (the Fourier atoms), $\psi_0, \psi_1, \ldots, \psi_{N-1}$ are the columns of the sparsity basis Ψ in which the signal to recover is assumed to be sparse, and M is the targeted number of samples. The authors propose a heuristic to solve the non-convex optimization problem (III–3), and present some global reconstruction success rates on synthetic signals obtained with different sampling strategies: as expected, the sampling strategy generated according to (III–3) shows a significant improvement over the uniform sampling, but remains slightly less efficient than the empirically derived strategies advocated by [Lustig07]. Moreover, [Puy11] do not present neither examples of the sampling masks obtained with their method, nor examples of reconstructed images.

- Finally, [Chauffert13b] proposes a two-step sampling strategy: first, a region close

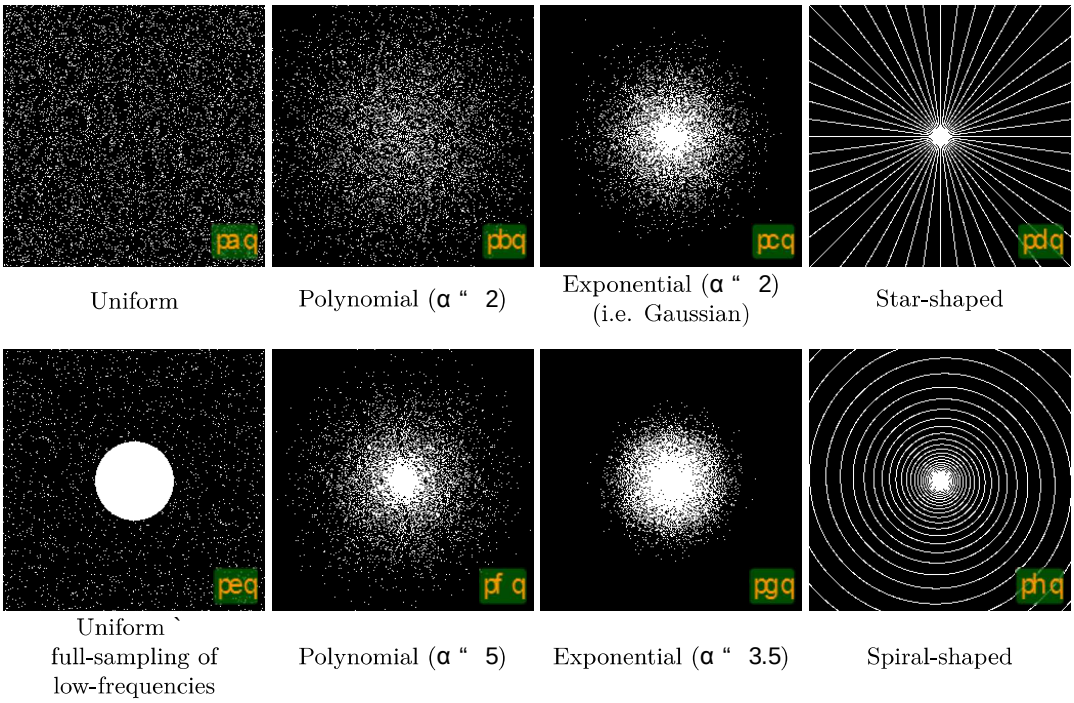
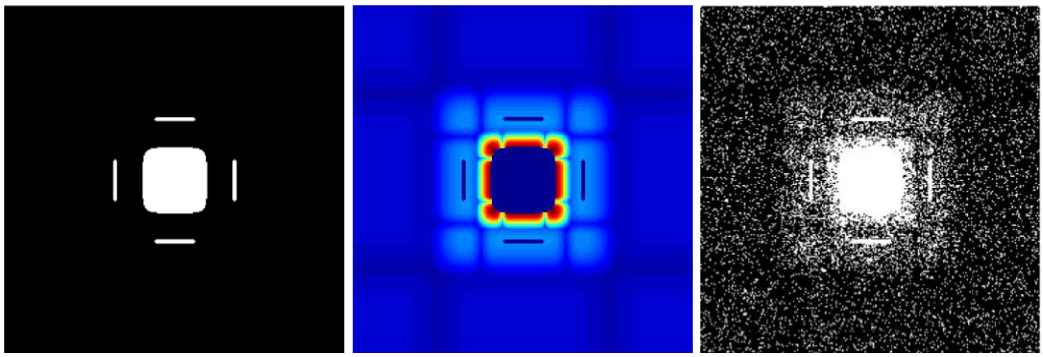


Figure III–2: Example of sampling locations in the Fourier space with different sampling strategies (the sampled locations are represented in white). In each of the eight presented examples, the number of sampled positions is the same, and corresponds to a overall sampling rate of 10%.

to the center of the k-space center is fully sampled (similar to what is proposed by [Kim09]); second, the high-frequencies are sampled according to a certain probability distribution $\pi^{\prime}$ (see Fig. III–3). This probability distribution $\pi^{\prime}$ is derived from theoretical results presented in [Rauhut10] (similarly to the approach followed in [Puy11], although the results from [Rauhut10] are not exploited in the same way in both works) to be optimal to sample in the Fourier domain 2D images that are sparse in the wavelet domain. With the application of their framework for the MRI modality in mind, the same authors present in [Chauffert13a] a heuristic to design a continuous sampling trajectory from a random distribution of sampling locations.

Contrary to their predecessors that mostly rely on empirical observations, the two latest works [Puy11] and [Chauffert13b] base their searches for an optimal sampling strategy on theoretical works, mostly [Rauhut10]. However, the result is not completely satisfactory: indeed, as noticed by the authors themselves, their optimal sampling strategy is outperformed by the polynomial sampling strategies from [Lustig07] – especially the one obtained for $\alpha = 5$ – when experimented via simulations on real MRI images. One hypothesis pushed forward by the authors of [Chauffert13b] to explain this suboptimal performance is that their root hypothesis, i.e. the fact that MRI images are sparse in a wavelet basis, is too simplistic to characterize for the properties of this type of data, which disrupts the theoretical derivation of the sampling density map. They suggest in



Fully sampled low-frequency region | Probability distribution $\pi^{\prime}$ used to select the high-frequency coefficients | Example of sampling pattern as proposed by [Chauffert13b]

Figure III–3: The sampling strategy proposed by [Chauffert13b] can be decomposed into two steps. First, a region close to the center of the k-space is fully sampled (cf. left image, the white area). Second, additional high-frequency coefficients are sampled, selected randomly according to the probability distribution $\pi^{\prime}$ (cf. middle image: red points correspond to higher probabilities of selection). An example of the sampling mask obtained with such strategy is presented on the right image. These images are reproduced from [Chauffert13b].

particular that the level of sparsity shown by the wavelet coefficients of these images may depend on the considered wavelet sub-band, with the majority of the non-zero coefficients presumingly concentrated in the coarse-scale sub-bands, and that modifying their analysis to account for this property would lead to better results.

Indeed, this review shows that finding a sampling strategy in the Fourier domain that would be optimal for all natural images (or for sub-classes of natural images) is still an open question. The only consensual characteristic among the existing sampling strategies mentioned above is the need to allocate more samples to the low-frequency area of the Fourier space.

III.2 Numerical evaluation of an optimal sampling rate

III.2.1 Random uniform sampling

III.2.1.1 Problematic

An issue related to the determination of an optimal sampling strategy is the evaluation of the appropriate number of measurements to be performed to sample a given signal. The theoretical result (I–8) from [Candès07] provides a sufficient condition on this number M of measurements, depending on the size N of the signal, its sparsity level S, and quantities characterizing the measurement process and the sparsity basis or dictionary (see Sec. I.2.3). However, this is only a sufficient condition, and it may happen that a signal is accurately

reconstructed even if (I–8) does not hold. In addition, this result neither describe nor quantify the visual quality of an image sampled and reconstructed from a reduced number of measurements.

We propose here to explore empirically how the number of measurements – or equivalently the sampling rate τ, defined as $\tau = \frac{M}{N}$ – affects the performance of the CS reconstruction, first in the case of the uniform sampling strategy. To proceed, we conducted the following numerical simulations: from a given known original image x_0, we generated several sets of Fourier measurements, each of these sets corresponding to a given sampling rate τ and acquired according to a uniform sampling strategy. Then, from each of these sets of measurements, we computed the solution $\hat{x}$ of the TV reconstruction scheme (P_{TV}), and measured a reconstruction error as follows:

$$\mathrm{RecErr} = \frac{\|\hat{x} - x_0\|_2}{\|x_0 - \mu_0 1\|_2} \tag{III–4}$$

where μ_0 is equal to the mean value of x_0. This reconstruction error measure is indeed proportional to the root mean squared error measure.

When building a set of measurements, we always sample the central Fourier coefficient (equal to μ_0), which otherwise could not be recovered by the TV minimization scheme. Our sampling patterns also obey a central symmetry invariance, to be coherent with the Hermitian symmetry property exhibited by the Fourier transform of real-valued images. With these settings, thanks to the normalization factor $\|x_0 - \mu_0 1\|_2$, we ensure that $\mathrm{RecErr}(\tau)$ tends to 1 when $\tau \to 0$.

III.2.1.2 Simulations on isotropic shapes

Our first simple experiment consists in reconstructing an elementary image composed of a single centered circular white object on a black background (this would represent for example a single cell visualized in fluorescence microscopy). We studied the evolution of the reconstruction error between the CS reconstructed image $\hat{x}$ and the original image x_0 as a function of the sampling rate τ. Since the reconstruction error for a given value of the sampling rate depends on the actual location of the samples in the Fourier space, each reconstruction was re-run ten to twenty times with different sampling patterns, and the median error value is reported.

We present some detailed results obtained for a one-disk image with a radius $\rho = 22$ pixels on Fig. III–4 (the image size is 256×256 pixels). The corresponding curve $\mathrm{RecErr} = f(\tau)$ shows three distinct domains:

- for small values of τ, the reconstruction error is constant and high: in this domain, the number of Fourier samples is too low to achieve a correct CS reconstruction, and the solution computed from (P_{TV}) is roughly unstructured;

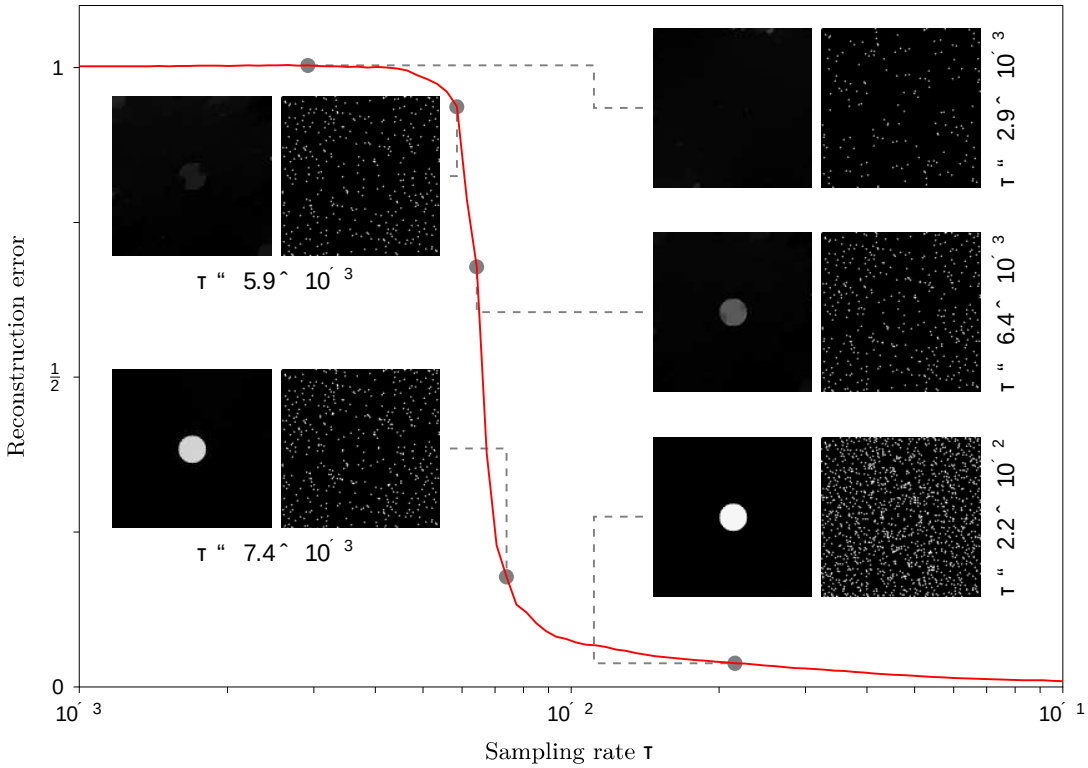


Figure III–4: Evolution of the reconstruction error of the same one-disk image with radius $\rho = 22$ pixels for a sampling rate τ in the range $[10^{-3}, 10^{-1}]$, and five reconstructed images obtained for different values of τ. Each reconstructed image (left thumbnails) is presented with its associated Fourier sampling mask (right thumbnails). The reconstructed image obtained for $\tau = 2.2 \times 10^{-3}$ is almost identical to the original image x_0.

- for high values of τ, the reconstruction error is also almost constant at a level close to zero: in this domain, the sampling rate is sufficient to perform an exact reconstruction of the original image from the subset of Fourier coefficients that are actually sampled;
- between these two constant domains, there is a narrow area around a transition sampling rate τ^* where the reconstruction error decreases from the high plateau to almost zero.

A set of similar experiments reproduced with different values of the disk radius ρ produce similar results, except that the transition between the two domains where $\mathrm{RecErr} = f(\rho, \tau)$ is constant does not occur for the same transition sampling rate τ^* (see Fig. III–5).

The value τ^* of the sampling rate, as it somehow separates the domain where reconstruction is possible from the domain where it is not, can be interpreted as an empirical measure of the sampling threshold that is defined in a theoretical manner in (I–8). The fact that this threshold is drastically modified depending on the input image – which can be observed on the curves $\mathrm{RecErr} = f(\rho, \tau)$ presented in Fig. III–5 – reflects on the variation of the underlying sparsity coefficient S.

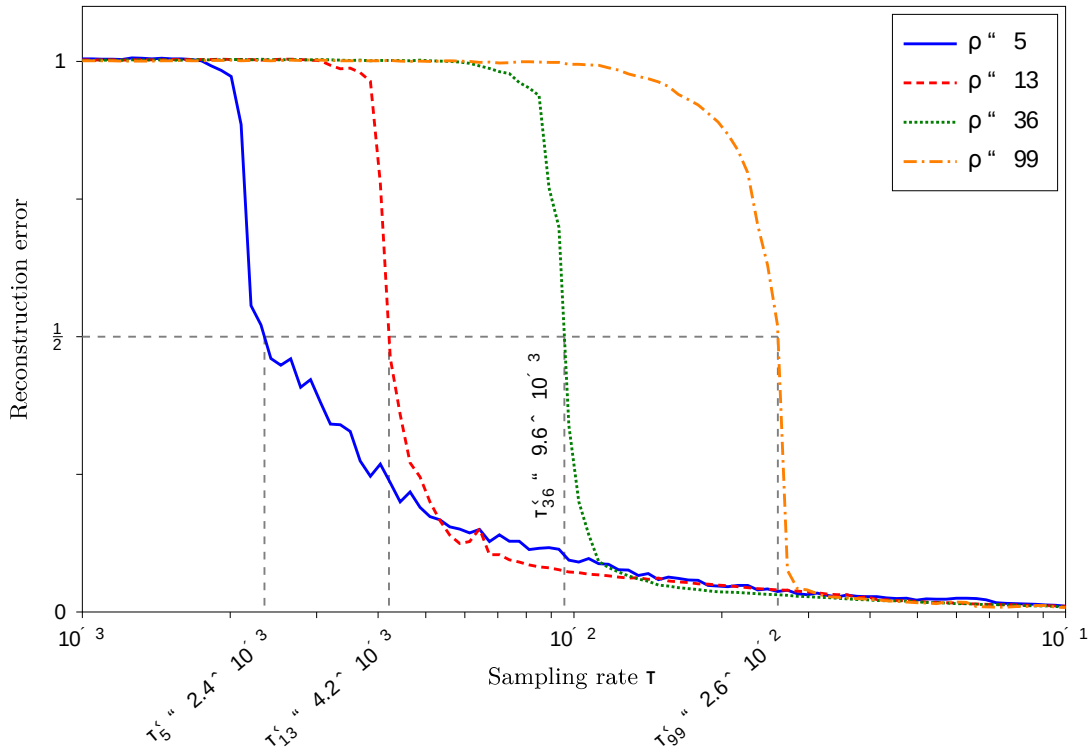


Figure III–5: Evolution of the reconstruction error for four one-disk images with various radius ρ, and the associated critical sampling rates τ^* extracted from the curves.

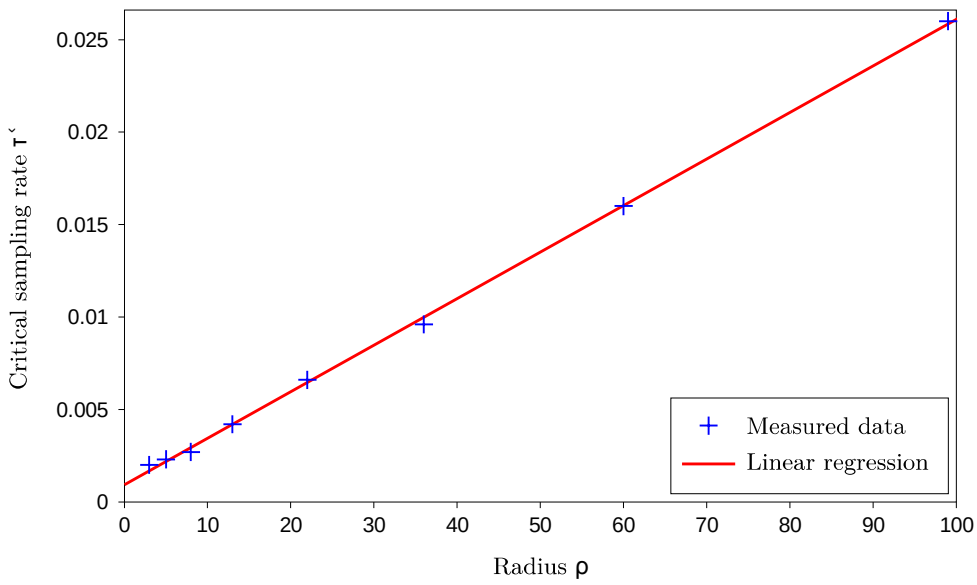


Figure III–6: Transition sampling rate τ^* for eight values of ρ. A linear regression on these data confirms that, for the single-disk images, τ^* obeys a linear increasing law with respect to ρ.

There are several possible definitions of the actual value $\tau^{\star}$ from the curve $\mathrm{RecErr} = f(\tau)$ in Fig. III–4: one could define $\tau^{\star}$ such that $\mathrm{RecErr}(\tau^{\star}) = 1/2$ ($1/2$ being the mean value between the two constant levels of the curve), or decide that $\tau^{\star}$ is the point where the first derivative of the function takes its maximal absolute value (as the function seems to have an inflexion point in the transition domain); however, as long as the transition domain is sufficiently narrow, all these definitions are likely to be equivalent. For the sake of simplicity, we define here $\tau^{\star}$ such that $\mathrm{RecErr}(\tau^{\star}) = 1/2$ in our simulations (see Fig. III–5); this definition does not depend on the spread of the transition domain.

III.2.1.3 Optimal sampling rate and sparsity

When performing a CS reconstruction using the optimization problem (P_{TV}), we know that the underlying *a priori* hypothesis on the input image is that it has a sparse gradient. In the case of our simple binary images, the number of non-zero gradient coefficients is approximately equal to the perimeter of the object. Then, together with (I–8), the transitional sampling rate $\tau^{\star}$ should be an increasing linear function of the perimeter of the object. Therefore, in the case of our one-disk images, $\tau^{\star}$ should increase linearly with the disk radius ρ.

In order to check this hypothesis, we computed the transition sampling rate $\tau^{\star}$ for eight values of the disk radius ρ. Results in Fig. III–6 confirm that $\tau^{\star}$ obeys a linear increasing law with respect to ρ, hence corroborating empirically the theoretical relation (I–8).

III.2.1.4 Optimal sampling rate and shape factor

We also investigated the dependency of the transitional sampling rate $\tau^{\star}$ with respect to the shape factor of the imaged object. Equation (I–8) suggests that $\tau^{\star}$ depends only on the number S of non-zero gradient coefficients, that is related to the perimeter of the object but not to its area or its shape factor. Therefore, two objects with the same perimeter should have the same transitional sampling rate $\tau^{\star}$, even if the first is isotropic (for example, a disk) while the second has a spatial dimension much larger than the others (for example, an elongated ellipse).

To validate this hypothesis, we followed a similar approach than for the one-disk test images, but we replaced the disks with ellipses of constant perimeter and various eccentricities. By varying the ratio γ between the half minor axis and the half major axis from 1 (circle) to almost 0 (flat shape), we tested shapes with different spatial and frequency characteristics; on the other hand, by setting a constant perimeter, we have maintained a constant TV-based sparsity measure for all the test images. The reconstruction error curves and the associated $\tau^{\star}$ are presented on Fig. III–7.

Although the four illustrated curves do not perfectly overlap, the associated transitional sample rates are distributed in a narrow domain, approximately $[10^{-2}, 1.7 \times 10^{-2}]$. Moreover, this analysis neglects all the effects due to the fact that our test shapes are

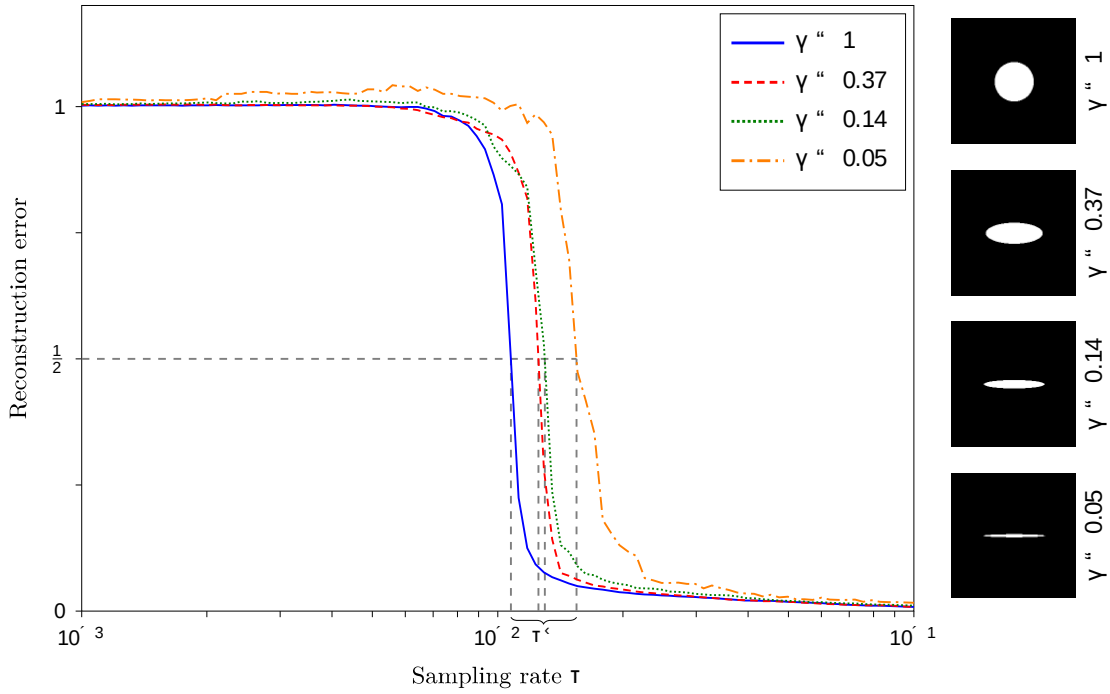


Figure III–7: Evolution of the reconstruction error for four ellipses of constant perimeter $P = 250$ pixels but various shape factor γ, and position of the associated critical sampling rates τ^*. The corresponding test images x_0 are presented on the right.

not drawn in a continuous domain but on a Cartesian grid instead; in particular, our hypothesis that the number of non-zero gradient coefficients is approximately equal to the perimeter of the object is no longer valid for small disks or very flat ellipses. This certainly explains why the four critical sampling rates τ^* are not strictly identical.

III.2.1.5 Putting things together

Results presented in sections III.2.1.3 and III.2.1.4 show that the critical sampling rate τ^* associated to a binary image composed of one elliptical object is proportional to the perimeter of the object, but does not depend on its shape factor. This is in agreement with the relation (I–8) stated by [Candès07], which expresses that the minimal number of measurements needed to reconstructed an image through a CS recovery scheme is proportional, for a given number of pixels N, to its sparsity coefficient S. When TV regularization is used to reconstruct binary images, this coefficient S is equal to the length of the boundary between the two binary domains.

Going further, we can study the following question: if the binary image is now composed of several elliptical objects, can we still observe a linear dependency between the critical sampling rate and the length of the boundary separating the objects from the background, which is equal to the sum of the perimeters of all the objects? Then, is it possible to predict a suitable sampling rate for CS reconstruction from an *a priori* knowledge of

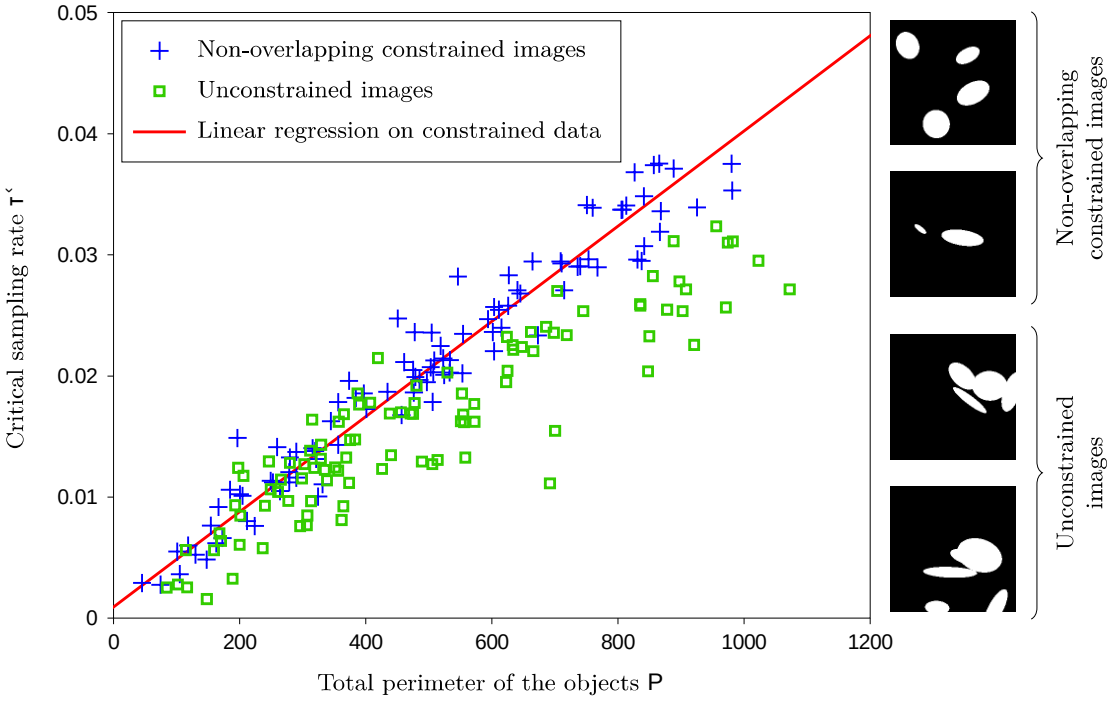


Figure III–8: Points (P, τ') computed on random binary images composed of ellipses generated with and without enforcing a non-overlapping constraint on them. Examples of such test images x_0 are presented on the right.

some geometric characteristics of the imaged objects, from which the size of the boundary between the two domains could be computed? An application example where it would make sense to formulate such knowledge can be found in the biological imaging field: if we image a sample containing a fixed number of cells with non-elastic membranes, a reasonable prediction about the size of the interface between the cells and the medium can be formulated, as this size would not change much over time.

To answer this question, we carried out the following experiment: we generated some test images containing a random number of ellipses, each of them having random perimeter, eccentricity and orientation. Then, for each of these test images, we computed the associated critical sampling rate τ' as well as the total perimeter of the objects P. Results are reported in Fig. III–8.

To be more precise, we carried out two series of experiments:

- First, we prevent ellipses from overlapping and from touching the edges of the image; this constraint ensures that the length of the boundary between the objects and the image background is actually equal to the sum P of the perimeters of the ellipses. The result of a linear regression computed on the points (P, τ') collected from these experiments is presented on Fig. III–8: even if we can observe that some data (P, τ') deviate from the predicting model, the general trend of this linear model is relevant. The encountered deviations might be due to the uncertainty in the measure of τ'.

- Then, we remove the non-overlapping and non-edge-touching constraints; which corresponds to simulation conditions closer to real-life applications. The counterpart of this relaxation is that we no longer ensure that the length of the boundary in the image is equal to P: actually, P will be greater than the length of the boundary between the two binary domains of the image. Therefore, as observed on Fig. III–8, the linear predicting model computed with the non-overlapping shapes provides an upper bound on the critical sampling rate for the unconstrained images.

Thus, from a practical point of view, given a prior knowledge on the geometric parameter P, we can predict which sampling rate is suitable for CS reconstruction, although this prediction will be pessimistic if the imaged objects overlap.

III.2.2 Random Gaussian sampling

The notion of optimal sampling rate is obviously related to the sampling strategy used to define the position of the measurements in the Fourier space. In all previous simulations, we used a random uniform sampling strategy for the sake of simplicity. However, studying how variations of the sampling rate affects the reconstructed images in the case of non-uniform sampling strategies is also worth of interest, and can be carried out similarly.

To proceed, we reproduced the one-disk test image experiments presented in Sec. III.2.1; however, instead of allocating measurements in a random uniform manner, we chose to allocate them according to the Gaussian sampling strategy, which is taken in a first approximation as a model for the other low-frequency biased sampling strategies presented in Sec. III.1.2. The corresponding curves $\mathrm{RecErr} = f(\tau, \rho)$ obtained for four different values of the disk radius ρ are presented on Fig. III–9.

Compared to the results presented on fig. III–5, the profile of the curves $\mathrm{RecErr} = f(\tau, \rho)$ is dramatically modified with the Gaussian sampling strategy. Indeed, for the single-disk images with $\rho = 3$ and $\rho = 5$, we can identify at least three domains where the reconstruction error is quite stable, and two transitional domains in between. The two extremal stable domains (which correspond approximately to $\tau \leq 10^{-3}$ and $\tau \geq 10^{-1}$) are similar to the stable domains that we observe with random uniform sampling: they correspond respectively to a complete failure of the CS recovery scheme, and to a perfect reconstruction of the original image. On the contrary, the intermediate stable domain observed with the Gaussian random sampling cannot be related to phenomena observed with uniform random sampling; when working with sampling rates in this range, the reconstructed images look very similar to the original one, but have blurred edges: for these reconstructions, the CS recovery procedure works well but induces a low-pass filtering effect on the reconstructed image. This low-pass filtering effect is linked to the Gaussian sampling strategy which allows very few measurements in the high frequency areas of the Fourier space.

Another particularity of the Gaussian sampling strategy is that small objects become harder to retrieve than large ones (meaning that a higher sampling rate is required to

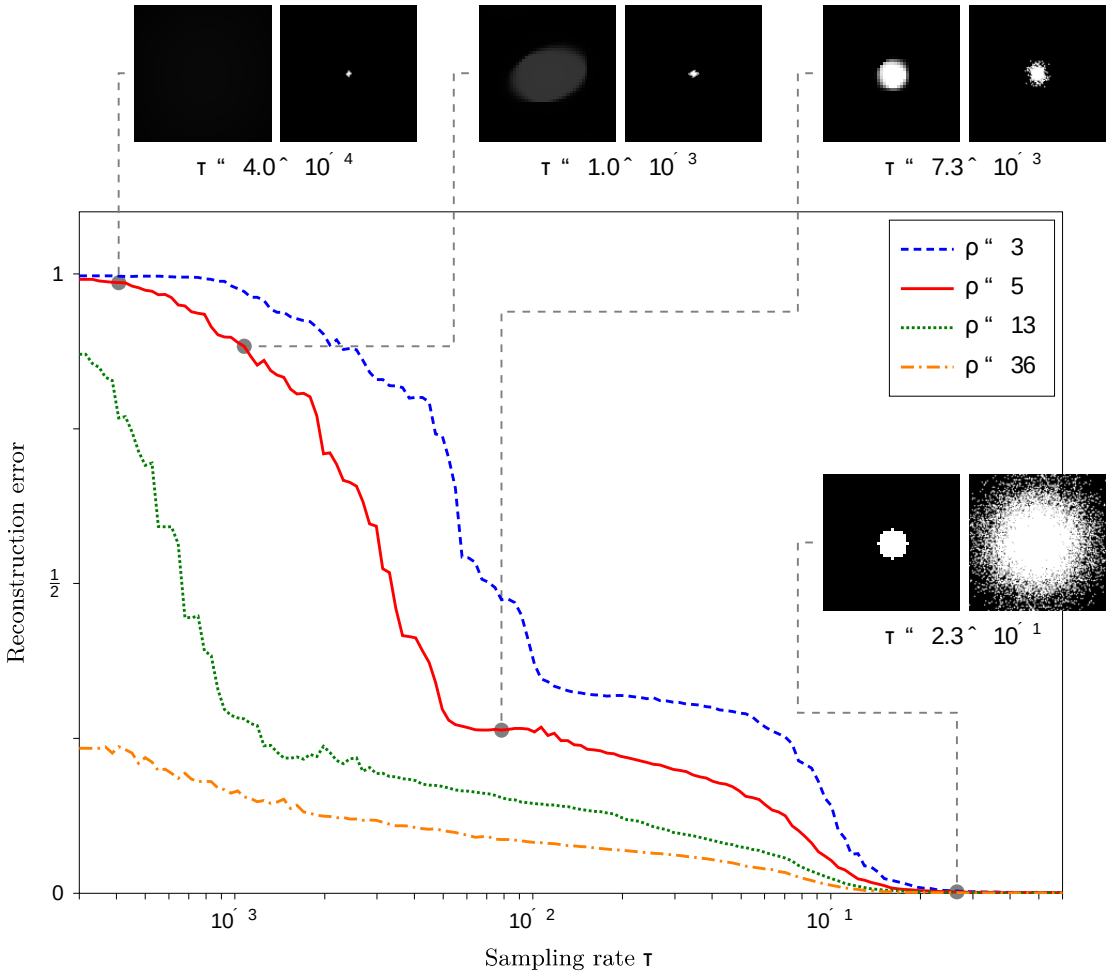


Figure III–9: Evolution of the reconstruction error for four one-disk images with various radius ρ, when sampled with a Gaussian sampling mask. We present also four zoomed reconstructions of the $\rho = 5$ one-disk test image (left thumbnails), with their associated sampling masks (right thumbnails). The reconstructed image obtained for $\tau = 2.3 \times 10^{-1}$ is almost identical to the original image.

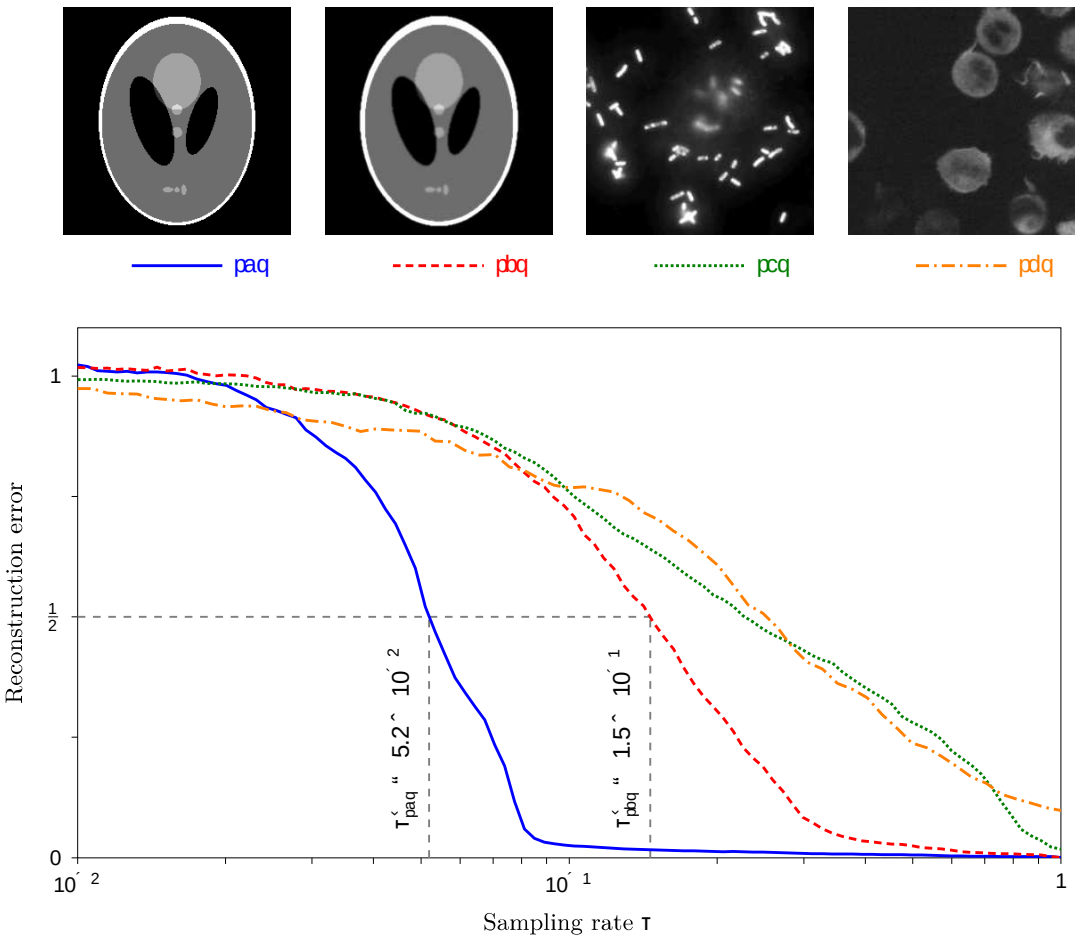


Figure III–10: Evolution of the reconstruction error for the four test images presented on the top of the figure, sampled according to a random uniform sampling strategy. These images are, from left to right, the Shepp-Logan test image paq, a blurred version of Shepp-Logan pbq, and two real biological images, *Shigella* pcq and *Lymphocytes T* pdq.

reconstruct the small-disk images), which is the inverse trend of what is observed with uniform random sampling. This is well understandable, as the Fourier transform of large objects is concentrated closer to the center of the Fourier space. The fact that CS recovery schemes together with Gaussian sampling strategy perform better on input signal that have a larger sparsity coefficient S is however paradoxical.

III.2.3 Realistic images

So far, our simulations were carried out on simple test images with low complexity compared to what is encountered in real ones. Real images have textures, may contain details at various scales, and are often subject to degradation during the acquisition process, resulting in blur and noise. Compared to what is observed on simplistic test images such as those in Figs. III–4, III–7 and III–8, all these phenomena induce a dramatic increase of the

number of non-zero coefficients that are necessary to represent the images. Actually, most of the real images are not even sparse at all, but only compressible, making such notion of sparsity level meaningless. To understand the influence of the sampling rate in such situation, we carried out CS reconstruction simulations on four realistic images, following the same protocol than above (see Sec. III.2.1), and using uniform Fourier sampling. The four realistic test images are:

- the well-known Shepp-Logan phantom image, with strictly piecewise constant structures and sharp edges;
- the Shepp-Logan phantom image degraded with a small Gaussian blur (using a standard deviation of 1 pixel length for the filter, while the image is $256 \hat{} \ 256$ pixels);
- a fluorescence microscopy image of *Shigella* bacteria;
- a microscopy image of *Lymphocytes T* cells, presenting a high level of noise.

Results of the simulations on these images are presented on Fig. III–10.

For both Shepp-Logan images, the error curves look very similar to the one obtained for the disks and the ellipses in Figs. III–5 and III–7: they present two stable domains where the CS reconstruction process respectively fails and succeeds, and a narrow transition area between these domains where the relative reconstruction error falls from almost 1 to almost 0 as τ increases. However:

- compared to what was observed with disks and ellipses, the transition between the failure and the success domains occurs at a much higher sampling rate (approximately 10^{-1}, instead of 10^{-2}); moreover, the spread of this transition domain is larger;
- even if the two Shepp-Logan images are very similar (the blurring effect applied on the second image is moderate as it does not lead to any loss of details), both the position and the spread of the transition domain are dramatically changed due to the blur.

Therefore, defining a critical sampling rate for these images still makes sense, even if the accuracy obtained on the corresponding measured values is poorer than for the simple binary images. Together with the theoretical relation (I–8), this can be explained by the fact that the Shepp-Logan images still have strict underlying sparse structures (in terms of gradient), even if the corresponding sparsity level S becomes much larger as soon as a blurring filter is applied.

On the contrary, in the case of the *Shigella* and *Lymphocytes T* images, the reconstruction error never reaches a stable low level; the reconstruction error associated to the *Lymphocytes T* image does not even reach 0 (even for $\tau = 100\%$, i.e. fully sampled reconstructions), because of its high level of noise. As these images are not strictly sparse, increasing the amount of available information (i.e. the number of measurements) always improves the fidelity of the reconstruction; therefore, defining a critical sampling rate

based on reconstruction errors for these images is not as relevant as in the case of binary sparse images.

III.3 Conclusion

In this chapter, we tackled the issues related with how sampling can and should be performed in the Fourier space. We first presented the existing sampling strategies, emphasizing that none of them has been proved to outperform the others in practice, in spite of theoretical studies carried on this question. We also investigated the influence of the sampling rate on the reconstructed images in the case of two particular sampling strategies (uniform and Gaussian sampling), showing that different reconstruction regimes exist depending on the value of this parameter and on the content of the reconstructed data. In the case of simple binary images sampled according to a uniform random strategy, we identified an optimal sampling rate separating the two domains corresponding respectively to perfect reconstruction and failure; we showed that the value of this optimal sampling rate could be predicted based on geometric characteristics of the sampled image that may be inferred prior to the acquisition. We also discussed the effects induced by a low-frequency favoring sampling strategy (the Gaussian sampling strategy) and the existence of an associated critical sampling rate in this case, and presented some results obtained with realistic images.

Chapter IV

Video sampling

Compared to 2D images, processing 2D+T video signals leads to particular problems related to the large size of this type of data. However, the counter part of this large size is that natural video signals are in general highly redundant, which allows them to undergo important compression ratio without noticeable degradations. Formally, this property can be exploited to represent the 2D+T video signals in a highly sparse or compressible manner, making this type of signals good candidates for being acquired as advocated by the compressed sensing theory.

In this chapter, we investigate how the CS framework can be adapted to video acquisition problems. We first consider a frame-by-frame linear acquisition model in the Fourier domain of the signal, and discuss the relevance of several sparsity models that could be used to drive the reconstruction of the whole video sequence. Then, we switch to a non-linear acquisition model – beyond the "pure" CS framework – in which only the modulus of the Fourier transform of the signal would be acquired: by exploiting sparsity properties similar to the one used in the linear acquisition case, we demonstrate the feasibility of a phase retrieval reconstruction procedure applied to video signals.

IV.1 CS applied to video signals **70**

IV.1.1 Acquisition model and problem formulation 70

IV.1.2 Existing sparsity models adapted to video signals 71

IV.1.3 Reconstruction using 3D total variation 73

IV.1.3.1 Three-dimensional total variation 73

IV.1.3.2 Mean background correction 76

IV.1.4 Comparative numerical simulations 77

IV.1.4.1 Methodology 77

IV.1.4.2 Data fidelity and reconstruction artifacts 78

IV.1.4.3 Sampling rate gain over frame-by-frame reconstruction 80

IV.2 Non-linear acquisition and phase retrieval **81**

IV.2.1 Non-linear versus linear optical Fourier measurements 81
IV.2.2 Translation invariance issue and problem formulation 82
IV.2.3 Phase retrieval reconstruction 83
IV.2.3.1 General framework 83
IV.2.3.2 Projection operator on the data set 85
IV.2.3.3 Hybrid total variation 86
IV.2.3.4 Projection operator on the regularization set 87
IV.2.3.5 Overall reconstruction algorithm 94
IV.2.4 Numerical simulations 96
IV.2.4.1 Methodology 96
IV.2.4.2 Qualitative and quantitative results 96
IV.2.4.3 Weight map in the hybrid total variation 98

IV.3 Conclusion **100**

IV.1 CS applied to video signals

The work on linear acquisition and CS reconstruction of video sequences developed in this section was presented in the conference paper [Le Montagner12].

IV.1.1 Acquisition model and problem formulation

We focus on the following problem: a signal of interest $X \in \mathbb{C}^{N \times T}$ composed of T successive 2D frames $x_t \in \mathbb{C}^N$ $(0 \leq t \leq T-1)$ is measured through a linear memoryless operator Φ, resulting in a vector $Y \in \mathbb{R}^M$ of observations[1]. Formally:

$$\underbrace{\begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{T-1} \end{bmatrix}}_{Y} = \underbrace{\begin{bmatrix} \varphi_0 & & & \\ & \varphi_1 & & \\ & & \ddots & \\ & & & \varphi_{T-1} \end{bmatrix}}_{\Phi} \underbrace{\begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{T-1} \end{bmatrix}}_{X} \tag{IV–1}$$

Using an appropriate CS reconstruction scheme ($\mathcal{P}_{analysis}$) or ($\mathcal{P}_{synthesis}$), the goal is then to recover X from Y.

The *memoryless* notion means that Y is accumulated from T sub-vectors $y_t \in \mathbb{C}^{m_t}$ of observations, with each y_t depending only on a given frame x_t; this measurement model corresponds to a 2D sensing device that would record and stack frame information sequentially. In such acquisition mode, temporal redundancy between 2D frames enables to

[1] In this chapter, capital letters denote 2D+T (or 2D+T related) signals and entities, while lower-case letters are reserved to objects with no temporal dimensions.

decrease the sampling rate compared to what is necessary to reconstruct them individually. Depending on the actual sensing device, the spare measurements could then be re-allocated in order to improve the temporal resolution of the system.

In this work, we focus on the case where the blocks φ_t are partial Fourier transforms, although the results presented in what follows might be extended to other types of sensing operators. The motivation for studying this type of Fourier-based acquisition model is that it can be used as a basis to design optical imaging devices working according to the following principle:

1. the imaged scene is observed through an optical set-up whose role is to implement an optical Fourier transform of the corresponding 2D image (see [Goodman96] for details on how this can be achieved),
2. this optical Fourier transform is focused on a plane array of photo-electric transducers (such as a CCD or CMOS array) in charge of the actual measurement operation.

From the CS theory, we learn that a small subset of the Fourier coefficients is sufficient to recover the 2D imaged scene. The goal is to design the array of photo-electric sensors involved in this acquisition scheme in such a way that:

- first, it allows to use only a configurable subset of sensors for a given acquisition,
- second, it takes advantage (for instance in terms of speed or energy efficiency) of being operated in such partial acquisition mode instead of having all its sensors "enabled" when acquiring an image.

The improved sensing capabilities of such type of sensing lead to a CS imaging system that would theoretically be able to over-perform the usual CCD or CMOS cameras.

The algebraic consequence of the memoryless measurement hypothesis is that the operator Φ is block-diagonal. In [Park11], the authors demonstrate that restricted isometry inequalities (I–5) do hold for such type of operator Φ with small constants δ_S when the blocks φ_t are random matrices with entries following a Gaussian distribution. However, to obtain this result, additional constraints have to be applied on the class of signals for which (I–5) is required to hold: besides the sparsity constraint, the authors require that the energy $\|x_t\|_2^2$ of each frame is proportional to the number of measurements m_t allocated to the corresponding sensing operator φ_t. Based on some empirical observations, we believe that such kind of result could also be established for other types of blocks φ_t, such as partial unitary transforms (see Sec. I.2.3). From a practical viewpoint, assuming that the frame energy remains almost constant over time and using the same number of measurements for each frame leads to satisfactory results (see Sec. IV.1.4).

IV.1.2 Existing sparsity models adapted to video signals

The sparsity model put on the signal of interest X conditions the form of the reconstruction scheme to use to recover X from Y, especially the sparsity basis Ψ in the case

of reconstruction by synthesis ($\mathcal{P}_{synthesis}$), or the objective function in the case of reconstruction by analysis ($\mathcal{P}_{analysis}$). For 2D natural images, several sparsity models exist (wavelets, curvelets, total variation), but few results have been established so far for joint reconstructions of time-correlated 2D images, i.e. 2D+T sequences.

In [Wakin06], the authors propose to assimilate the 2D+T signal $\mathbf{X}$ to a 3D signal, and to reconstruct it by enforcing a sparsity constraint on its 3D wavelet coefficients; formally, this approach leads to a synthesis reconstruction ($\mathcal{P}_{synthesis}$) where a 3D wavelet basis is used for the matrix Ψ. Although it is a natural generalization of the 2D case, this approach does not take into account the fact that the objects appearing in a 2D+T sequence might have very anisotropic spatio-temporal shape, while wavelets are best suited for isotropic objects.

In [Park09], the authors introduce a multi-scale video reconstruction framework, which relies on the idea of increasingly refining the spatial scale of the estimated signal: at each step, the algorithm exploits information obtained from coarser estimates to reduce the temporal redundancies and to estimate motion. However, although presenting some promising results, this method requires to adapt the measurement protocol in order to get some information about the coarse versions of the signal. Such modification is possible with the single-pixel camera, which is the acquisition device targeted by the work [Park09]. However, it cannot be easily extended to other CS imaging modalities.

In [Marcia08], the authors propose to perform a joint reconstruction of sequences of K-consecutive frames (where $K \geq 2$ is a predefined parameter) in the following way: given a basis $\psi \in \mathbb{C}^{N \times N}$ in which each frame has sparse or nearly sparse representation – typically a 2D wavelet basis – they define the following NK-square matrices:

$$\mathbf{B}_K = \begin{bmatrix} \psi & & & \\ -\psi & \psi & & \\ \vdots & & \ddots & \\ -\psi & & & \psi \\ \psi & & & \psi \end{bmatrix} \qquad \mathbf{C}_K = \begin{bmatrix} \psi & & & & \\ -\psi & \psi & & & \\ \vdots & \vdots & \ddots & & \\ -\psi & \psi & \cdots & \psi & \\ \psi & \psi & \cdots & \psi & \psi \end{bmatrix} \tag{IV–2}$$

Then, they propose to use either $\mathbf{B}_K$ or $\mathbf{C}_K$ as the dictionary in a l_1-synthesis reconstruction scenario ($\mathcal{P}_{synthesis}$). The underlying idea is to exploit the temporal redundancy existing in the video sequence by reconstructing the difference between frames instead of the frames themselves. More precisely, with the $\mathbf{C}_K$ matrix, the l_1-synthesis enforces sparsity on the coefficients of the vectors $\psi^{-1}(\mathbf{x}_t - \mathbf{x}_{t-1})$ – i.e. the 2D wavelet coefficients of the difference $(\mathbf{x}_t - \mathbf{x}_{t-1})$ between each frame $\mathbf{x}_t$ and its predecessor if ψ is a 2D wavelet basis. The $\mathbf{B}_K$ matrix behaves similarly: sparsity is enforced on the coefficients of the vectors $\psi^{-1}(\mathbf{x}_t - \mathbf{x}_{t_0})$ where $\mathbf{x}_{t_0}$ represents the first frame of the group of K frames being reconstructed. These "frame difference-based" approaches lead to interesting results (especially the $\mathbf{C}_K$-based one), but introduce some reconstruction artifacts that will be discussed more thoroughly in Sec. IV.1.4.

Finally, in [Trocan13] (also in the series of conference papers [Trocan10a, Trocan10b, Trocan10c]), the authors propose a multistage video reconstruction framework that includes motion estimation and compensation heuristics to improve the quality of the reconstructed video sequences. The problem tackled in those works is actually more general than the one we consider here, in that the authors propose to reconstruct in a joint manner several video signals acquired simultaneously from the same scene with different cameras (this scenario being denoted as "multiview" acquisition), each camera measuring a vector of CS data for each frame (potential applications of this problem include in particular video surveillance and stereoscopy). More precisely, the reconstruction principle presented in [Trocan13] consists in the following steps:

1. for each vector of CS measurements acquired by each camera, reconstruct the corresponding 2D frame by solving a CS reconstruction problem;
2. then, refine iteratively the reconstructed frames as follows:
   (a) construct for each frame of each camera a prediction based on the estimates obtained for the adjacent frames at step 1 and motion prediction heuristics,
   (b) adjust these predictions to fit the CS measurements.

The second step of this reconstruction procedure may be repeated several times to achieve a targeted level of reconstruction fidelity.

The motivation for the first step of this procedure, which consists in independent 2D CS reconstructions, is to provide the motion prediction algorithms involved in step 2 with estimates (even imperfect) of the video signals to reconstruct, as those algorithms cannot directly deal with raw CS data. However, as step 1 and step 2 are not directly related, other reconstruction methods could be used to achieve this initialization: indeed, step 2 – the main contribution of the paper [Trocan13] – can be considered as a post-processing method to improve CS reconstructions, rather than as a reconstruction method by itself. For this reason, we did not include this method in the comparative work presented in the next section.

IV.1.3 Reconstruction using 3D total variation

IV.1.3.1 Three-dimensional total variation

As suggested by [Marcia08], considering frame-to-frame differences could be an interesting starting point to exploit temporal redundancies existing in 2D+T signals. However, one should notice that the significant non-zero coefficients are not randomly distributed in a typical consecutive frame difference. Indeed, if x_t and x_{t-1} are two consecutive frames in a video sequence, then the coefficients of the difference $(x_{t-1} - x_t)$ with large magnitudes – i.e. the displacement and deformation fronts of the objects shown by the video – are mostly located close to the edges shown in the frames x_t and x_{t-1} (see Fig. IV–1).

To be more formal, the sparsity model that we propose to use for the reconstruction of 2D+T video sequences is composed of the following hypotheses:

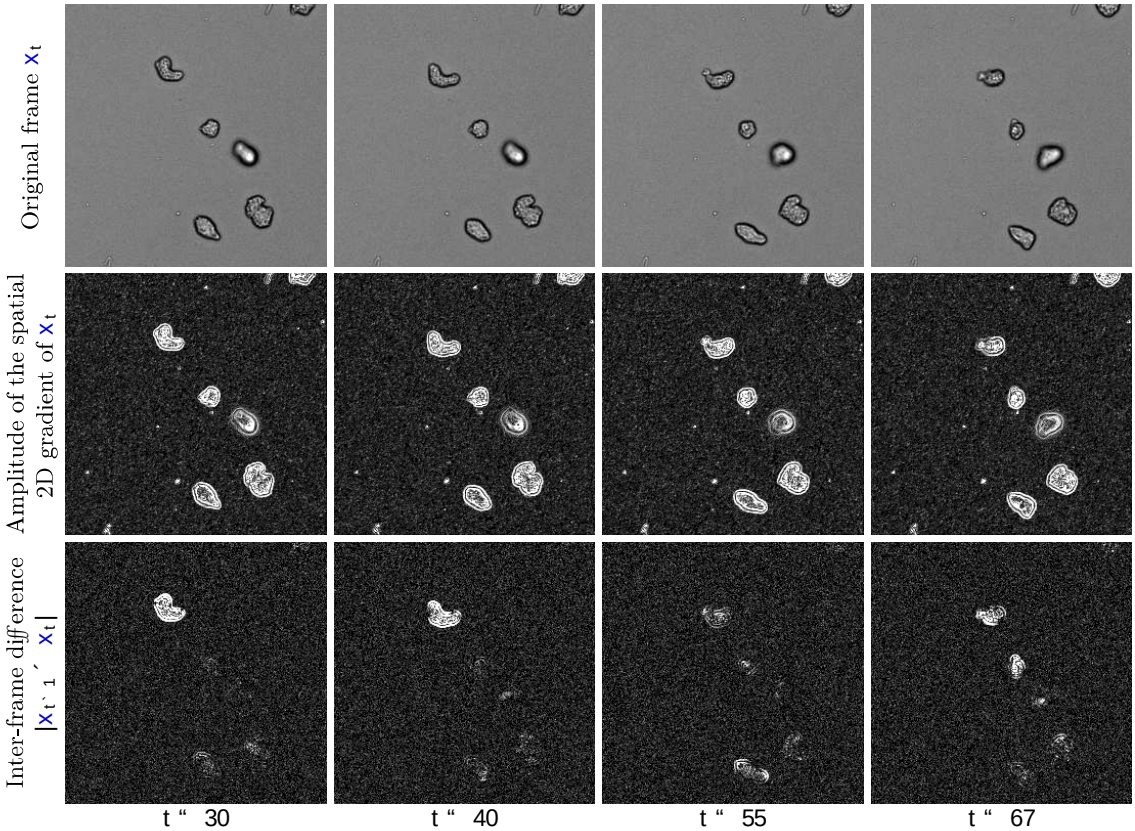


Figure IV–1: Example of a microscopy video sequence showing amoeba cells (top row). The middle row shows the amplitude of the spatial 2D gradient of each frame (white pixels denoting large gradient amplitudes), while the bottom row shows the amplitude of the difference between two consecutive frames (white pixels denoting large differences). One can notice that the locations where large inter-frame difference is observed correspond to locations where the amplitude of the spatial 2D gradient is also significant.

1. The spatial 2D gradient of each frame x_t is sparse (see Sec. I.2.5). We refer to this property as *intra-frame sparsity*.
2. The difference $(x_{t+1} - x_t)$ between two consecutive frames is sparse. We refer to this property as *inter-frame sparsity*.
3. The significant non-zero coefficients of the spatial 2D gradient of a given frame x_t – that correspond to the edges of the objects and structures shown in x_t – and the significant non-zero coefficients in $(x_{t+1} - x_t)$ – that correspond to the displacement and deformation fronts – are mostly located at the same positions.

To account for these properties, we introduce the three-dimensional total variation functional (TV-3D), defined as follows:

$$\|X\|_{\text{TV-3D}} = \sum_{t=0}^{T-2} \sum_{(u,v)\in\Omega} \sqrt{|(D_h x_t)(u,v)|^2 + |(D_v x_t)(u,v)|^2 + |(x_{t+1} - x_t)(u,v)|^2} \tag{TV-3D}$$

where D_h and D_v represent the horizontal and vertical discrete derivative operators operating on rasterized 2D images (as in (I–10)), and $\Omega \subset \mathbb{Z}^2$ is the spatial domain on which the frames x_t are defined. Thanks to this functional, we define an estimator $\hat{X}$ of the signal of interest X from the measurement vector Y as a solution of the following analysis reconstruction problem:

$$\hat{X} = \underset{X \in \mathbb{C}^{NT}}{\arg\min} \|X\|_{\text{TV-3D}} \text{ subject to } \|\Phi X - Y\|_2 \leq \varepsilon \tag{IV–3}$$

where Φ is the measurement operator defined in (IV–1), and the parameter $\varepsilon \geq 0$ accounts for the inaccuracy of the measurement vector Y induced by various phenomenons involved in the measurement process, such as noise and quantization.

The reason why minimizing $\|X\|_{\text{TV-3D}}$ enforces the first two sparsity properties mentioned above (i.e. intra- and inter-frame sparsity) stems from the following inequalities, that can be easily derived from the definition of the 3D total variation (TV-3D):

$$\max\left(\sum_t \|x_t\|_{TV}, \sum_t \|x_{t+1} - x_t\|_1\right) \leq \|X\|_{\text{TV-3D}} \leq \sum_t \|x_t\|_{TV} + \sum_t \|x_{t+1} - x_t\|_1 \tag{IV–4}$$

where $\|\cdot\|_{TV}$ is the 2D total variation as defined in (I–10). Indeed, minimizing $\|X\|_{\text{TV-3D}}$ leads to small values of both the cumulated 2D TV of all the frames of the sequence $\sum_t \|x_t\|_{TV}$ and the cumulated l_1-norm of all the frame to frame differences $\sum_t \|x_{t+1} - x_t\|_1$, and reciprocally. Moreover, from the concavity property of the square root function, it can be shown that:

$$\sum_t \|x_t\|_{TV} + \sum_t \|x_{t+1} - x_t\|_1 \leq \sqrt{2}\,\|X\|_{\text{TV-3D}} \tag{IV–5}$$

and that this inequality is tight if and only if, for all $t \in [0, T-2]$ and all $(u,v) \in \Omega$, the following holds:

$$\sqrt{|(D_h x_t)(u,v)|^2 + |(D_v x_t)(u,v)|^2} = |(x_{t+1} - x_t)(u,v)| \tag{IV–6}$$

In other words, for given values of $\sum_t \|x_t\|_{TV}$ and $\sum_t \|x_{t+1} - x_t\|_1$ – which can be thought as measures of respectively the intra-frame and inter-frame sparsity – the 3D total variation is minimal when, at each spatial point (u,v) and time point t, the amplitude of the local spatial gradient $\sqrt{|(D_h x_t)(u,v)|^2 + |(D_v x_t)(u,v)|^2}$ is equal to amplitude of the local frame-to-frame difference $|(x_{t+1} - x_t)(u,v)|$: this explains the relation between the minimization of $\|X\|_{TV\text{-}3D}$ and the third sparsity property enforced on 2D+T video sequences mentioned above. This relation can also be explained by interpreting $\|X\|_{TV\text{-}3D}$ as a particular mixed $l_{1,2}$-norm (see Sec. I.2.5) operating on a linear transform of X that would stack its discrete derivatives in the horizontal, vertical and temporal directions.

IV.1.3.2 Mean background correction

There are some situations where the difference $(x_{t+1} - x_t)$ between two consecutive frames is not sparse at all, even if x_{t+1} and x_t are well-correlated. For microscopy applications, this includes the case when the global illumination of the observed scene changes over time. To make TV-3D regularization robust to this phenomenon, we reformulate the reconstruction scheme (IV–3) as follows:

$$\underset{X \in \mathbb{C}^{N \times T}}{\arg\min} \|X - A\|_{TV\text{-}3D} \text{ subject to } \|\Phi X - Y\|_2 \leq \square \tag{IV–7}$$

where A is a sequence whose frames a_t ($0 \leq t \leq T - 1$) are defined by $a_t = \mu_t 1$, with μ_t representing the mean intensity value of the tth frame in the original signal of interest.

The sequence A, or equivalently the mean value μ_t of each frame, has to be estimated prior to the resolution of (IV–7) from the vector of observations Y. As $\mu_t = \frac{1}{N} \langle 1 | x_t \rangle$, if each measurement operator φ_t contains a row proportional to 1, the values μ_t can directly be read from the vector of observations Y; this is for example the case when the φ_t are partial Fourier transforms for which the sampling pattern is designed such that the central Fourier coefficient (the one corresponding to the constant basis vector) is always sampled. If φ_t does not contain a row proportional to 1, μ_t can still be estimated using the framework developed in [Davenport10a]; according to the results presented in that paper, μ_t, being a linear function of the signal of interest, can be estimated as $\hat{\mu}_t = \frac{1}{N} \langle \varphi_t 1 | y_t \rangle$. This property can be understood as a consequence of the restricted isometry property (see Sec. I.2.2): assuming that φ_t has small RIP constants, this operator behaves like an isometry, meaning in particular that it preserves inner products between vectors:

$$\hat{\mu}_t = \frac{1}{N} \langle \varphi_t 1 | y_t \rangle = \frac{1}{N} \langle \varphi_t 1 | \varphi_t x_t \rangle \approx \frac{1}{N} \langle 1 | x_t \rangle = \mu_t \tag{IV–8}$$

Actually, the property (IV–8) requires additional conditions on φ_t to hold, that are precisely defined and thoroughly justified in [Davenport10a] (Theorem 4).

Finally, as a practical remark, it can be noticed that the optimization scheme (IV–7) is actually equivalent to the TV-3D driven scheme (IV–3) up to the variable change $X^1 = X \cdot A$. Then, (IV–7) can be solved in practice with the usual dedicated CS solvers, assuming that they can handle the 3D total variation (either natively or after being adapted). In practice, we use the NESTA algorithm [Becker11] to solve (IV–3) and (IV–7): the modification required to make this algorithm handle TV-3D are quite straightforward as 2D TV is natively supported.

IV.1.4 Comparative numerical simulations

IV.1.4.1 Methodology

We compared the proposed TV-3D-based regularization methods with other existing reconstruction formulations, including:

- l_1-synthesis using a 3D wavelet transform (see [Wakin06]), using either the Haar wavelet (as suggested by the authors) or the Dauchechies-4 orthogonal wavelet (DB4);
- l_1-synthesis using the B_K and C_K dictionaries (see [Marcia08]), with a block size of $K = 4$ or $K = 20$ frames, and a Dauchechies-4 wavelet transform as the 2D dictionary ψ.

To assess the improvement offered by 3D reconstruction methods thanks to temporal redundancies, we also provide the results obtained with frame-by-frame reconstruction, using either TV or Dauchechies-4 wavelet regularization.

For this evaluation, we used three video test sequences:

- *Amoeba*, sized $256 \times 256 \times 80$ (height $\times$ width $\times$ number of frames), which is a microscopy sequence of moving and deforming amoeba cells;
- *Foreman*, sized $288 \times 352 \times 80$, which is a widely-used test sequence in the signal processing community, representing a talking person over a non-stabilized background;
- *Disks 1*, sized $256 \times 256 \times 80$, which is a synthetic sequence representing moving disks with random gray levels, sizes (diameters between 5 and 25 pixels) and speeds (the distance travelled by one disk is about 1 to 3 pixels between two consecutive frames). The boundaries of these disks is also blurred by Gaussian kernels with various radius, in order to simulate different conditions of focus. We designed this synthetic sequence so that it breaks the underlying model corresponding to 3D total variation regularization; more precisely, the gray level of the background oscillates quickly between two values, simulating rapid variations of the global illumination.

For each of these test sequences, we generated a vector Y of observations by concatenating partial Fourier measurements of each frame. The selection of the Fourier coefficients was performed according to a random uniform strategy, with different realizations of this random strategy for each frame. The sampling patterns were also constrained to obey a

Reconstruction method	PSNR (dB)		
	Amoeba	*Foreman*	*Disks*
Frame-by-frame, TV regularization	42.5	16.2	26.6
Frame-by-frame, wavelets (DB4) regularization	38.3	12.7	15.5
3D total variation	46.8	27.6	22.0
3D total variation with background correction	46.8	26.8	38.9
3D wavelets (Haar) [Wakin06]	45.2	20.8	18.6
3D wavelets (DB4) [Wakin06]	45.3	21.1	15.4
B_4 + 2D wavelets (DB4) [Marcia08]	30.7	17.4	17.2
C_4 + 2D wavelets (DB4) [Marcia08]	43.8	17.8	18.1
B_{20} + 2D wavelets (DB4) [Marcia08]	43.0	20.7	17.9
C_{20} + 2D wavelets (DB4) [Marcia08]	45.8	23.6	18.2

Figure IV–2: Reconstruction error (expressed as a PSNR) between the original sequences *Amoeba*, *Foreman* and *Disks 1* and the corresponding reconstructed sequences obtained using various regularization methods. The proposed regularization methods are highlighted with bold font.

central symmetry invariance, to be coherent with the Hermitian symmetry property exhibited by the Fourier transform of real-valued frames. We used an arbitrary set sampling rate of 10% for both *Amoeba* and *Disks 1*, and 20% for *Foreman* to handle the higher level of complexity exhibited by this sequence.

Then, we reconstructed each of the test sequences from the corresponding measurement vector Y with all the considered reconstruction methods. We assessed the reconstruction fidelity of the algorithms for each test sequence by measuring the peak signal-to-noise ratio (PSNR) between the input and the reconstructions (see Fig. IV–2); visual qualitative evaluation of the artifacts was also performed (see Figs. IV–3, IV–4, and IV–5).

IV.1.4.2 Data fidelity and reconstruction artifacts

In terms of PSNR, the proposed TV-3D-based methods obtain the best reconstruction results, although the improvement over the other best performing methods (C_{20} [Marcia08] or 3D wavelets [Wakin06] regularizations) is not dramatic in most cases (1 dB for *Amoeba*, about 4 dB for *Foreman*). However, this measurement does not reflect the gain in terms of visual perception brought by the two TV-3D-based methods.

Indeed, compared to the wavelet-based regularization methods, TV-3D tends to produce sequences with very sharp edges, without the oscillatory patterns typically present close to the edges in 3D wavelet reconstructed sequences. TV-3D reconstructions also do not have the following problems typically encountered with B_K and C_K-based estimators:

- B_K dictionaries tend to produce estimators where all the K frames belonging to a given group are similar (the gray level of a given pixel is almost piecewise constant over time), resulting in a jerky effect when switching from one group to the next.
- C_K reconstructed sequences display precognition and trailing artifacts, meaning that the reconstructed frame corresponding to time t contains patterns belonging to the

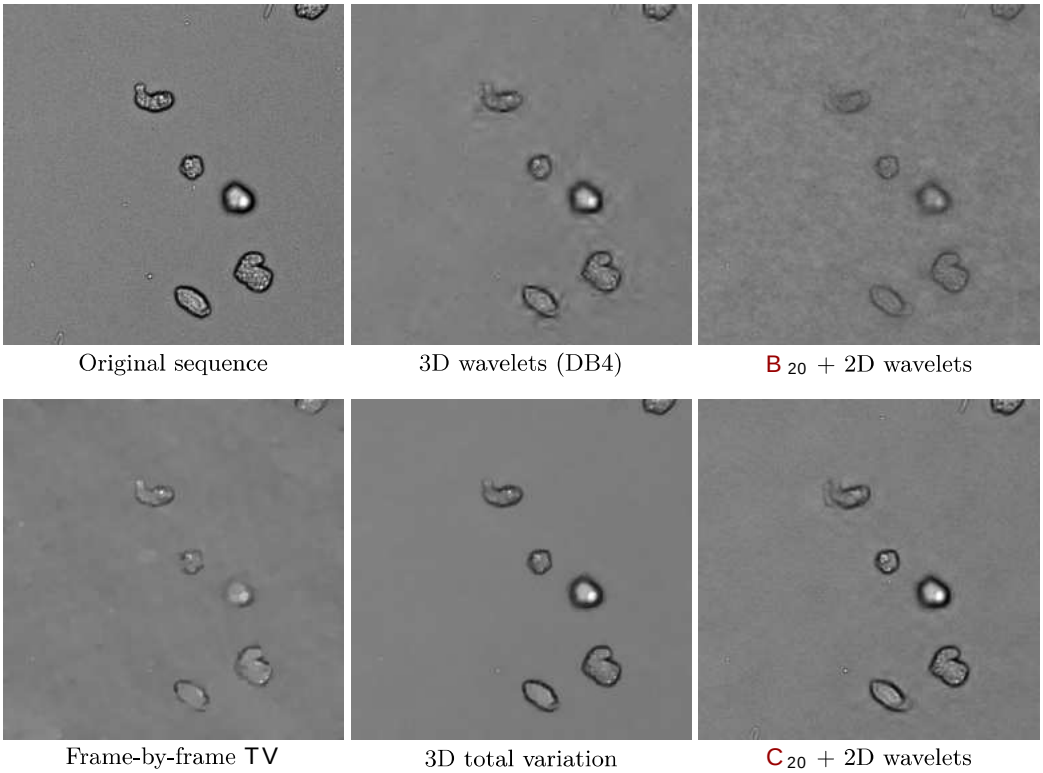


Figure IV–3: Reconstruction results obtained for the test sequence *Amoeba* (frame t " 50) with various regularization methods.

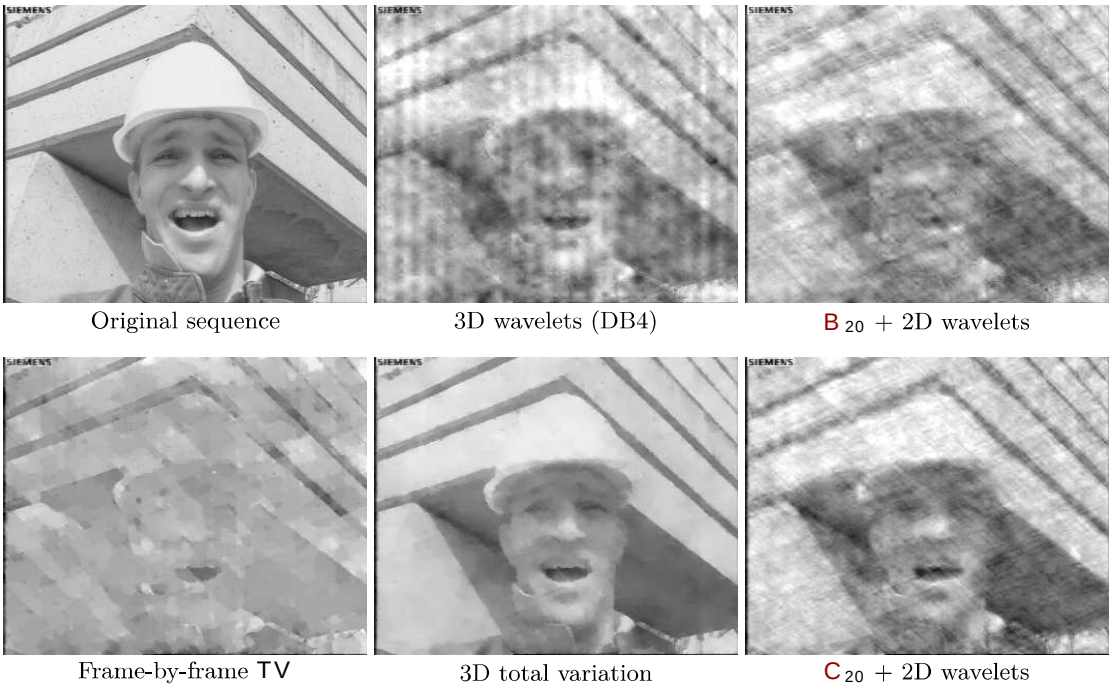


Figure IV–4: Reconstruction results obtained for the test sequence *Foreman* (frame t " 23) with various regularization methods.

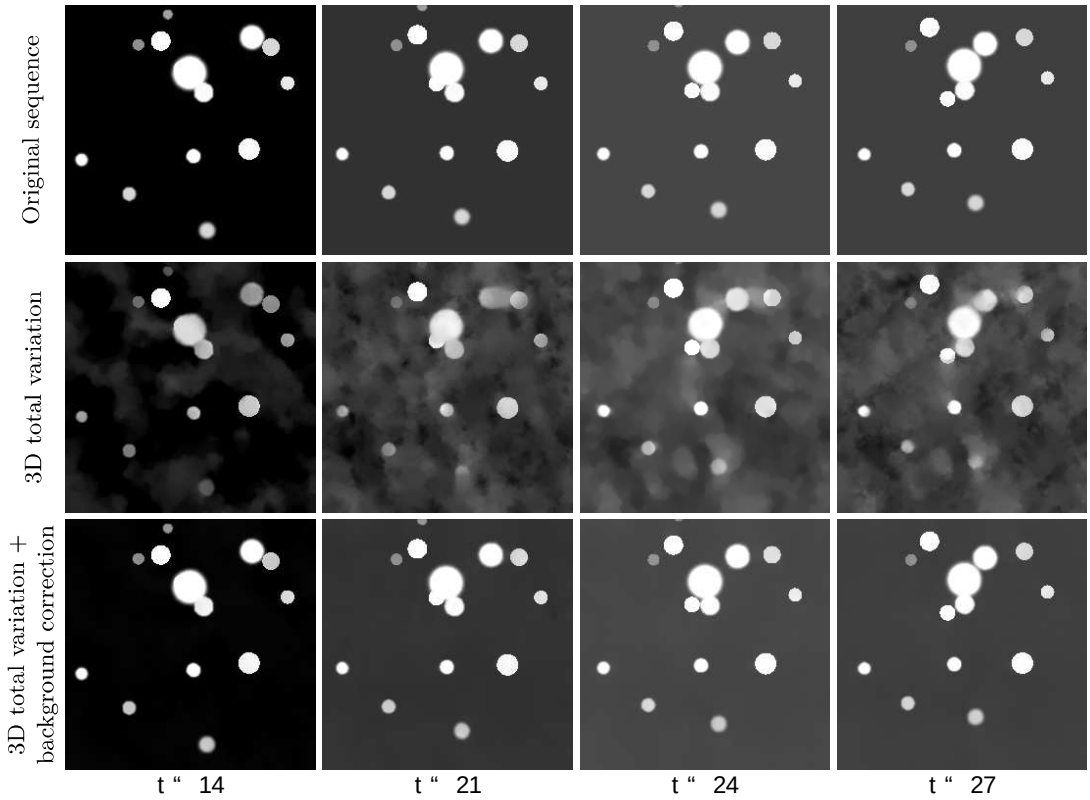


Figure IV–5: Reconstruction results obtained for the test sequence *Disks 1* with the two proposed methods. The result obtained using the TV-3D with background correction regularization (bottom row) is visually identical to the original sequence (top row).

frames t` 1, t´ 1, t` 2, t´ 2, etc. This is particularly noticeable close to the moving or deforming objects.

Finally, for most of the sequences, the simple 3D total variation estimator is very similar to its TV-3D with background correction counterpart, both in terms of PSNR and visual quality. The only exception is the *Disks 1* sequence, which was designed on purpose to challenge the TV-3D reconstruction: since the difference between two consecutive frames is non-zero at almost every pixel, the corresponding hypothesis on which the TV-3D estimator – as well as many other estimators, especially those using the B_K and C_K dictionaries – relies on does not hold. Using the TV-3D regularization term with background correction tackles this issue, leading to a result almost identical to the original in the case of the *Disks 1* sequence (see Fig. IV–5).

IV.1.4.3 Sampling rate gain over frame-by-frame reconstruction

To quantify the sampling rate gain provided by the TV-3D-based reconstruction methods over simple frame-by-frame reconstructions, we reproduced the simulations performed on our test sequences with several values of the sampling rate, following the same methodology

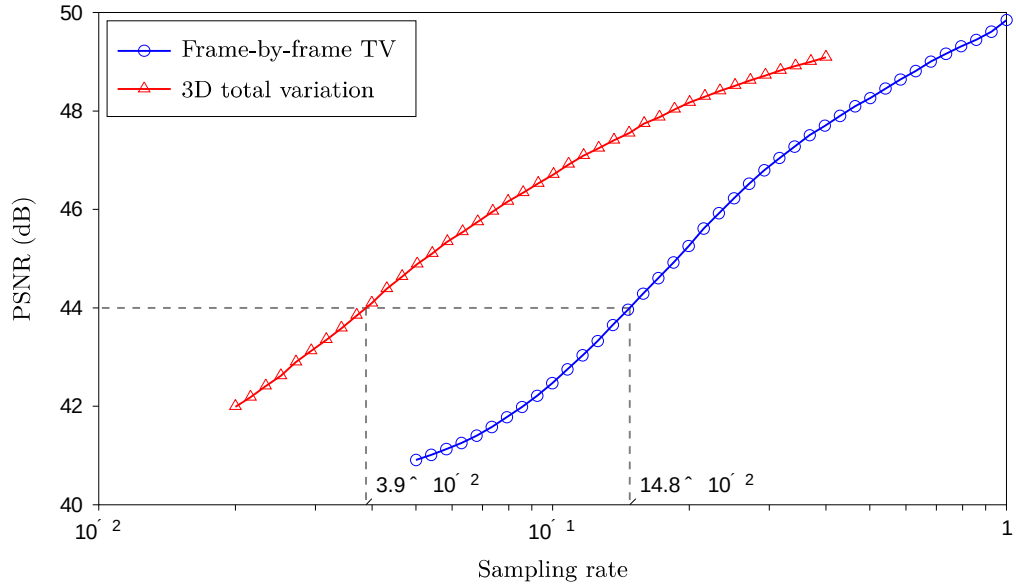


Figure IV–6: Trade-*off* curves between sampling rate and reconstruction error for the *Amoeba* sequence, depending on the reconstruction method. To achieve reconstruction with a given error bound, the TV-3D regularization method needs three to four times less measurements than its 2D frame-by-frame counterpart.

than described in Sec. IV.1.4.1. We then evaluated the evolution of the reconstruction error (measured as a PSNR) as a function of the sampling rate: results obtained for the *Amoeba* sequence are presented in Fig. IV–6.

We observed that the sampling rate corresponding to a given level of fidelity of the reconstructed sequence with respect to the original data is generally 3 to 4 times smaller with the TV-3D-based reconstruction than with the frame-by-frame TV reconstruction, that do not exploit the temporal redundancies of the sequences; this ratio tends to decrease when the PSNR increases. One should also mention that this result does not depend on whether TV-3D reconstruction with background correction or TV-3D alone is considered, except in the case of the *Disks 1* sequence, for which TV-3D reconstruction alone completely fails.

IV.2 Non-linear acquisition and phase retrieval

The work developed in this section – about non-linear acquisition in the Fourier domain and reconstruction of video sequences using a phase retrieval methodology – was published in the conference papers [Le Montagner13b] and [Le Montagner13c].

IV.2.1 Non-linear versus linear optical Fourier measurements

In Sec. IV.1, we studied a problem of video recovery, aiming at reconstructing a 2D+T sequence from samples of the 2D Fourier transform of each of its frames. The motivation

for studying this problem is to conceive optical CS cameras based on an optical Fourier transform and a specific photo-electric sensor array device (see Sec. IV.1.1).

However, although measuring the complex Fourier transform of a scene can be implemented optically (see [Goodman96]), it remains challenging as photo-electric transducers such as a CCD or CMOS array return output signals that correspond to a measure of the energy of the incident photons, which is independent of the phase of the corresponding electro-magnetic wave. Measuring this phase relies on more complex optical set-ups (holography set-ups for instance, see [Marim11b] and Sec. I.3.2) which are not always compatible with the experimental arrangement. On the contrary, measuring the sole magnitude of the complex Fourier transform relaxes the constraints put on the optical part of our optical CS camera scheme.

Formally, the acquisition model corresponding to sole magnitude measurement paradigm is the following:

$$y_t = |\varphi_t x_t| \qquad \text{for all } t \tag{IV–9}$$

where $x_t \in \mathbb{R}^N$ is the tth frame of the acquired sequence, $y_t \in \mathbb{R}^{m_t}$ is the vector of measures acquired at time point t, φ_t is a partial Fourier transform, and $|\cdot|$ stands for the pointwise modulus. Contrary to the "phase-aware" acquisition model (IV–1), that states that the measured data are related to the signal of interest through a linear relation and therefore falls into the general framework of compressed sensing, the new acquisition model (IV–9) is non-linear, calling for a completely different reconstruction strategy, denoted as *phase retrieval*.

IV.2.2 Translation invariance issue and problem formulation

Rejecting the phase information of the Fourier transform during the measurement process (IV–9) makes two images equal up to a translation with periodic boundary conditions indistinguishable, due to the properties of the Fourier transform. Therefore, a necessary condition to make the recovery of an image x_t based on measures y_t acquired as (IV–9) possible is to inject some prior knowledge about the location of the structures and objects shown in the image.

In a context of video acquisition and reconstruction, such prior knowledge can be provided by the frames adjacent to the frame being reconstruct. In what follows, we propose a reconstruction scheme that operates recursively on consecutive frames: starting from an initial key-frame x_0 assumed to be available, we reconstruct the following frame x_1 using its partial Fourier modulus data y_1 and the key-frame x_0; the reconstruction process is then iterated to the next frame, to reconstruct x_2 using y_2 and x_1, etc. The global video reconstruction problem is then recast into a sequence of frame reconstruction problems, defined as:

$$\text{Find } x_t \text{ such that } \begin{cases} y_t = |\varphi_t x_t| \text{ (up to noise and measurement inaccuracy)} \\ x_t \text{ is compatible with } x_{t-1} \end{cases} \tag{IV–10}$$

where the compatibility condition between x_t and its predecessor x_{t-1} will be formalized in Sec. IV.2.3.3.

It is worth noting that this step-by-step reconstruction procedure differs from the reconstruction scheme used in the context of linear measurements (see Sec. IV.1), in which all the frames of the video sequence are reconstructed in a joint manner. It also requires to have a key-frame x_0 to initialize the recursive procedure: such key-frame has to be acquired in a different manner than the other frames of the sequence, which impacts the design of an imaging set-up that would implement such acquisition strategy. However, the acquisition of such additional data is somehow unavoidable, as one has to break the translation invariance mentioned above.

IV.2.3 Phase retrieval reconstruction

In this section, we introduce the phase retrieval reconstruction algorithm used to solve (IV–10). For the sake of simplicity, we assume here that the frame index t is fixed, and we drop the corresponding subscripts: x (previously x_t) will denote the frame being reconstruct, a (previously x_{t-1}) its predecessor (assumed to be known when x_t is reconstructed), and similarly for y (previously y_t) and φ (previously φ_t).

IV.2.3.1 General framework

The problem of recovering a signal from the modulus of its Fourier transform, known as the phase retrieval problem, has been studied for a long time: this reconstruction technique is used for instance for X-ray microscopy applications in crystallography (see [Fienup82, Miao99]). To recover a signal $x \in \mathbb{R}^N$ from a measurement vector y defined as in (IV–9), [Fienup82] propose an iterative algorithm based on alternated projections over two subsets of $\mathbb{R}^N$:

- The *data set* $D_{y,\epsilon}$, that contains all the candidates x that correspond to the measured samples, up to a certain tolerance ϵ that depends on the noise that affects these measurements:

$$D_{y,\epsilon} = \left\{ x \in \mathbb{R}^N \text{ such that } \|y - |\varphi x|\|_2 \leq \epsilon \right\} \tag{IV–11}$$

- The *regularization set* R that corresponds to all the signals that meet certain prior conditions which are known to be true for the actual solution. For crystallography applications, R typically consists in all the 2D images that are supported on a given subset of pixels.

Then, an estimator $\hat{x}$ of the solution is obtained as a limit of alternated projections over the two sets $D_{y,\epsilon}$ and R:

$$\hat{x} = \left(\Pi_R \circ \Pi_{D_{y,\epsilon}} \circ \Pi_R \circ \cdots \circ \Pi_{D_{y,\epsilon}} \right)(x_{init}) \tag{IV–12}$$

where $\circ$ is the composition operator, x_{init} is an initial guess of the solution, and $\Pi_{D_{y,\square}}$ and Π_R stand respectively for the projection operators over $D_{y,\square}$ and R:

$$\Pi_{D_{y,\square}}(x) = \underset{z}{\arg\min} \|z - x\|_2 \text{ subject to } z \in D_{y,\square} \tag{IV–13}$$

and similarly for Π_R. It is shown in [Fienup82] that the sequence of estimators produced by the iterations (IV–12) converges toward the intersection of $D_{y,\square}$ and R.

Other works have refined the algorithm (IV–12), mostly to improve the convergence rate (see for instance [Bauschke03, Luke05, Marchesini07] and references therein). However, to the best of our knowledge, most of the existing phase retrieval algorithms rely on the construction of a sequence of estimators that converges toward the intersection of two sets, one characterizing the valid signals with respect to the measurements, and the other one the prior information available on the solution. We however mention three works that are exceptions with respect to this general approach:

- In [Candès11], the authors propose to transform the non-linear modulus constraint $y = |\varphi x|$ into a linear one, by lifting up the unknown variable from $\mathbb{R}^N$ to a higher dimension space. More precisely, if the rows of the matrix φ are denoted as $\varphi_0^\circ, \varphi_1^\circ, \ldots, \varphi_{N-1}^\circ$, the set of non-linear equations $\{y_k = |\langle \varphi_k, x\rangle|\}_{k=0,\ldots,N-1}$ that expresses the compatibility of the signal x with the vector of measurements y is recast into the following equivalent formulation:

$$y_k^2 = \mathrm{Tr}(\varphi_k \varphi_k^\circ x x^\circ) \qquad \text{for all } k \tag{IV–14}$$

  where $\mathrm{Tr}(\cdot)$ is the matrix trace operator[2]. The authors in [Candès11] propose to change the unknown variable, switching from the column vector $x \in \mathbb{R}^N$ to the symmetric positive matrix $X = xx^\circ \in \mathbb{R}^{N \times N}$: this modification makes the equations (IV–14) linear, allowing convex formulations of the phase retrieval problem. However, the price to pay for this transformation is that the unknown variable is now defined in a N^2 dimension space: for imaging applications where N might be quite large ($N = 6.5 \times 10^4$ in the numerical examples presented in Sec. IV.2.4, which corresponds to $N^2 = 4.3 \times 10^9$), the approach proposed by [Candès11] is mostly inapplicable due to performance and memory issues.

- In [Waldspurger13], the set of non-linear equations $y = |\varphi x|$ is turned into a phase completion problem, in which the authors propose to search for not only the signal of interest x, but also a "phase vector" $u \in \mathbb{C}^N$ (meaning that each component u_k of u has a unit modulus) such that $y \times u = \varphi x$. More precisely, the approach proposed in [Waldspurger13] consists in solving the following problem[3]:

[2] This re-writing trick involving the trace operator actually dates back to [Balan09], where it is presented in a quite intricate manner.

[3] The analysis proposed is [Waldspurger13] is stated for complex-valued signals $x \in \mathbb{C}^N$, while in Sec. IV.2 we chose to formulate the phase-retrieval problem for real-valued signals $x \in \mathbb{R}^N$ for the sake of clarity. However, most of the results presented here can be extended to complex-valued signals.

$$\underset{x,u \in \mathbb{C}^N}{\arg\min} \|\varphi x - y \circ u\|_2^2 \text{ subject to } \forall k \; |u_k| = 1 \tag{IV–15}$$

It can be noticed that the x-minimization sub-problem occurring in (IV–15) has an algebraic solution, namely $x = \varphi^{\dagger} \cdot (y \circ u)$ where $\varphi^{\dagger}$ denotes the Moore-Penrose pseudo-inverse of φ. By substituting this solution into (IV–15), the authors recast the problem to solve into the following one:

$$\underset{u \in \mathbb{C}^N}{\arg\min} \langle u, H u \rangle \text{ subject to } \forall k \; |u_k| = 1 \tag{IV–16}$$

where $H \in \mathbb{C}^{N \times N}$ is a positive Hermitian matrix depending on φ. Then, the method proposed by the authors to solve (IV–16) makes use of a variable change similar to the one used in [Candès11] (see the above point), replacing $u \in \mathbb{C}^N$ with $U = u u^* \in \mathbb{C}^{N \times N}$, to "convexify" the underlying problem. As such, this approach suffers from the same limitations as the one proposed in [Candès11] when dealing with large-dimension signals as encountered in imaging applications.

- In the paper [Shechtman13] – that was released simultaneously to our work [Le Montagner13b] – the authors tackled a problem similar to ours, that is video reconstruction based on measurements of the Fourier transform modulus of each frame composing the sequence. In their approach, the difference map between two consecutive frames is assumed to be sparse, and this is this difference map that is actually reconstructed. More precisely, the support of this difference map is iteratively discovered by a procedure that alternatively:

  1. for a given support, finds the difference map that best matches the Fourier modulus measurements,
  2. updates the support to improve this matching.

  This approach however supposes that the size of the support of the difference map between two consecutive frames is known: we do not see how to infer this information in the context of our problem.

In our case, the prior information is quite different from the one available in crystallography applications (i.e. support constraints). We propose in the next section to therefore adapt the alternated projection scheme (IV–12) to our video reconstruction problem by formulating a suitable set of prior hypotheses and deriving the corresponding regularization set $\mathcal{R}$.

IV.2.3.2 Projection operator on the data set

Before moving to the prior hypothesis and the regularization set issue, let us state some remarks about the data set $\mathcal{D}_{y,\square}$ and its associated projection operator. One of the key points that makes the alternated projection scheme (IV–12) efficient is that, although the set $\mathcal{D}_{y,\square}$ is not convex in general, its associated projection operator $\Pi_{\mathcal{D}_{y,\square}}$ can be evaluated explicitly in a very efficient manner, as demonstrated in what follows.

First, let us rewrite the expression (IV–13) with the Fourier transforms $x^1 = \mathcal{F} x$ and $z^1 = \mathcal{F} z$ of the involved variables x and z ($\mathcal{F} \in \mathbb{C}^{N \times N}$ denoting the Fourier transform operator):

$$\Pi_{D_{y,\square}}(x) = \mathcal{F}^{\circ} \hat{z}^1 \quad \text{with} \quad \hat{z}^1 = \underset{z^1}{\arg\min} \left\| z^1 - x^1 \right\|_2^2 \text{ subject to } \left\| y - \check{\Sigma} z^1 \right\|_2 \leq \delta \tag{IV–17}$$

where $\Sigma \in \{0,1\}^{m_t \times N}$ is the selection operator associated to the partial Fourier transform φ (see Sec. I.2.3). Via a permutation of the rows of z^1, this vector can be expressed as a vertical concatenation of two column vectors $z^s \in \mathbb{C}^{m_t}$ and $z^d \in \mathbb{C}^{(N - m_t)}$, with $z^s = \Sigma z^1$: these two sub-vectors correspond to coefficients of z^1 that are respectively selected and discarded by Σ. Using similar notations for x^1, it follows that:

$$\hat{z}^1 = \begin{pmatrix} \hat{z}^s \\ \hat{z}^d \end{pmatrix} \quad \text{with} \quad \left(\hat{z}^s, \hat{z}^d\right) = \underset{z^s, z^d}{\arg\min} \left\| z^s - x^s \right\|_2^2 + \left\| z^d - x^d \right\|_2^2 \text{ subject to } \left\| y - |z^s| \right\|_2 \leq \delta \tag{IV–18}$$

From the expression of the optimization problem in (IV–18), it clearly results that $\hat{z}^d = x^d$, as the minimization with respect to the variable z^d is unconstrained. We then rewrite the remaining complex-valued vectors in a polar form: $z^s = r \cdot \exp(i \cdot \theta)$ and[4] $x^s = \rho \cdot \exp(i \cdot \phi)$, leading to:

$$\hat{z}^s = \left(\hat{r} \cdot \exp\left(i \cdot \hat{\theta}\right)\right) \quad \text{with} \quad \left(\hat{r}, \hat{\theta}\right) = \underset{r, \theta}{\arg\min} \sum_k \left(r_k^2 + \rho_k^2 - 2 \cdot r_k \cdot \rho_k \cdot \cos(\theta_k - \phi_k)\right)$$
$$\text{subject to } \left\| y - r \right\|_2 \leq \delta \text{ and } r_k \geq 0 \text{ for all } k \tag{IV–19}$$

From the positivity of the modulus values r_k and ρ_k, it results that $\hat{\theta} = \phi$, meaning that $\hat{z}^s$ and x^s must have the same phase. The remaining optimization with respect to the variable r is a quadratic constraint problem. Using the Karush-Kuhn-Tucker conditions (see for instance [Boyd04]), it can be shown that:

$$\hat{r} = \frac{1}{1 + \lambda}(\rho + \lambda y) \quad \text{with} \quad \lambda = \max\left(0, \frac{1}{\delta}\left\|\rho - y\right\|_2 - 1\right) \tag{IV–20}$$

All together, the expressions (IV–17) to (IV–20) lead to a closed-form expression for the projection operator $\Pi_{D_{y,\square}}$. The algorithmic complexity of the evaluation of this expression is $\mathcal{O}(N \log N)$, this cost being dominated by the computation of Fourier transforms.

IV.2.3.3 Hybrid total variation

The sparsity hypothesis assumed on the video sequences to reconstruct is the same here than in the case of the linear measurement context studied in Sec. IV.1: indeed, these

[4] The notation $z^s = r \cdot \exp(i \cdot \theta)$ expresses a pointwise relation, i.e. $z_k^s = r_k \cdot \exp(i \cdot \theta_k)$ for all components k.

hypotheses are supposed to describe intrinsic properties of the sampled sequences, which implies that they do not depend on the measurement process. We therefore assume that the three sparsity properties introduced in Sec. IV.1.3.1 still hold for the current phase retrieval problem.

To enforce these properties in the current context, we introduce the following functional, denoted as *hybrid total variation*:

$$\|x\|_{hTV,w,a} = \sum_{(u,v)\in\Omega} w(u,v)\sqrt{(D_h x(u,v))^2 + (D_v x(u,v))^2 + (x - a(u,v))^2} \qquad (hTV)$$

where x and a are two 2D images defined on the spatial discrete domain $\Omega \subset \mathbb{Z}^2$, D_h and D_v are the discrete derivative operators defined as in (TV-3D), and w is a weight map defined on the domain Ω such that $0 \leq w(u,v) \leq +\infty$ for all the points $(u,v) \in \Omega$. In a first approach, w can be though as a uniform map ($w = 1$), although the computations associated to the hybrid total variation will be conducted for generic weight maps w. The role of this parameter and the way it is set will be specified in Sec. IV.2.4.3.

The functional (hTV) is related to the 3D total variation (TV-3D), in that if X represents a 2D+T video sequence composed of T frames x_t ($0 \leq t \leq T-1$), then the following relation holds:

$$\|X\|_{TV\text{-}3D} = \sum_{t=0}^{T-1} \|x_t\|_{hTV,1,x_{t-1}} \qquad (IV–21)$$

The hybrid total variation of a frame x_t (with the parameter a set to x_{t-1}) can therefore be thought as the contribution of this frame to the 3D total variation of the whole 2D+T sequence. As a consequence, the hybrid total variation inherits the properties of the 3D total variation in terms of sparsity enforcement: minimizing this functional do select frames with the sparsity characteristics stated in Sec. IV.1.3.1.

Finally, we define the regularization set R involved in the alternated projection scheme (IV–12) as a level set of the hybrid total variation (hTV):

$$R_{w,a,\tau} = \left\{ x \in \mathbb{R}^N \text{ such that } \|x\|_{hTV,w,a} \leq \tau \right\} \qquad (IV–22)$$

The newly introduced parameter $\tau \geq 0$ becomes an input prior to the reconstruction problem or can be set adaptively during the reconstruction process, as proposed below.

IV.2.3.4 Projection operator on the regularization set

Using $R_{w,a,\tau}$ as a regularization set for the reconstruction requires to be able to evaluate efficiently the projection operator $\Pi_{R_{w,a,\tau}}$, which implies to solve several instances of the following problem:

$$\Pi_{R_{w,a,\tau}}(x) = \underset{z\in\mathbb{R}^N}{\arg\min} \|z - x\|_2 \text{ subject to } \|z\|_{hTV,w,a} \leq \tau \qquad (IV–23)$$

To solve this convex optimization problem, we use an algorithm derived from the total variation projection method presented in [Fadili11]. The idea behind this method is to recast the constrained problem (IV–23) into an equivalent unconstrained problem, and to solve the latter using a gradient descent method with Nesterov acceleration. We detail this approach in the following paragraphs:

1. we start by introducing some of the notations and lemmas needed to describe the method,
2. then we derive unconstrained formulation equivalent to (IV–23),
3. we detail how the proximal operator associated to the weighted ℓ_∞-norm defined in the first step – which is used in the algorithm solving the unconstrained formulation – is evaluated,
4. finally, we present the algorithm used to solve this unconstrained formulation.

Preliminary notations and lemmas First, let us introduce a few notations. In what follows, $\mathcal{F}_d$ ($d \in \mathbb{N}^*$) denotes the vector space whose objects are obtained as the concatenation of d elements of $\mathbb{R}^N$. For instance, for $d = 3$ such concatenation is represented as $(x, x^1, x^2) \in \mathcal{F}_3$ for any $x, x^1, x^2 \in \mathbb{R}^N$. Conversely, if $f \in \mathcal{F}_d$, then $f_{(p)} \in \mathbb{R}^N$ (with $p \in [\![0, d-1]\!]$) denotes the pth vector that composed the object f: for instance, $(x, x^1, x^2)_{(1)} = x^1$. By extension, $f_{(p,q)} \in \mathcal{F}_2$ is defined as $(f_{(p)}, f_{(q)})$ for any $f \in \mathcal{F}_d$. We also provide the vector space $\mathcal{F}_d$ with an inner product and a ℓ_2-norm, defined canonically as:

$$\langle f | g \rangle = \sum_{p=0}^{d-1} \langle f_{(p)}, g_{(p)} \rangle \qquad \text{for all } f, g \in \mathcal{F}_d \tag{IV–24}$$

and $\|f\|_2 = \sqrt{\langle f | f \rangle}$ for all $f \in \mathcal{F}_d$.

By focusing more specifically on the vector space $\mathcal{F}_3$, we introduce the notions of weighted ℓ_1 and ℓ_∞-norms on this space[5]. For any $w \in \mathbb{R}^N$ with $0 \leq w_k \leq +\infty$ and $f \in \mathcal{F}_3$, we define:

$$\begin{aligned} \|f\|_{1,w} &= \sum_k w_k \sqrt{\left(f_{(0)}\right)_k^2 + \left(f_{(1)}\right)_k^2 + \left(f_{(2)}\right)_k^2} \\ \|f\|_{\infty,w} &= \max_k w_k \sqrt{\left(f_{(0)}\right)_k^2 + \left(f_{(1)}\right)_k^2 + \left(f_{(2)}\right)_k^2} \end{aligned} \tag{IV–25}$$

where we remind that $\left(f_{(p)}\right)_k$ refers to the kth component of the pth vector that composed the object $f \in \mathcal{F}_3$. One can easily check that the norm properties do hold for both $\|\cdot\|_{1,w}$ and $\|\cdot\|_{\infty,w}$: in particular, the requirement that the components of the weight map w are non-zero ensures that these functionals are definite. Also, please note that, although $\mathcal{F}_3$ is isomorphic to $\mathbb{R}^{3N}$, $\|\cdot\|_{1,w}$ and $\|\cdot\|_{\infty,w}$ are different from the usual ℓ_1 and ℓ_∞-norms defined on $\mathbb{R}^{3N}$, even in the case of the uniform weight map (i.e. $w = 1$). One can also notice that the hybrid total variation (hTV) can be expressed as a weighted ℓ_1-norm:

[5] These definitions could easily be extended to other $\mathcal{F}_d$ spaces and ℓ_p-norms, but we do not need such extensions for our proof.

$$\|x\|_{\mathrm{hTV},w,a} = \|(x - a, D_h x, D_v x)\|_{1,w} \tag{IV–26}$$

Finally, we introduce the linear operator $\nabla : \mathbb{R}^N \to F_2$, defined as $\nabla x = (D_h x, D_v x)$. The adjoint of $(-\nabla)$ is denoted as Div, and its expression is $\mathrm{Div} f = -D_h^{\circ} f_{(p,0,q)} - D_v^{\circ} f_{(p,1,q)}$ for any $f \in F_2$.

Thanks to these definitions, we can state the following properties (a justification of them is provided below):

Lemma IV–1 *For all $f, g \in F_3$:*

$$\langle f, g \rangle \leq \|f\|_{1,w} \, \|g\|_{\infty,\frac{1}{w}} \tag{IV–27}$$

where $\frac{1}{w} \in \mathbb{R}^N$ denotes the vector whose components are equal to $\frac{1}{w_k}$ pointwise. Moreover, for all $f \in F_3$, there exists $g \in F_3$ (with $g \neq 0$) that makes this inequality tight. Conversely, for all $g \in F_3$, there also exists a non-zero $f \in F_3$ such that (IV–27) *is tight.*

Lemma IV–2 (Dual norm of $\|\cdot\|_{1,w}$) *For all $g \in F_3$ and $\lambda \geq 0$:*

$$\lambda \|g\|_{\infty,\frac{1}{w}} = \sup_{f \in F_3} \langle f, g \rangle \quad \textit{subject to} \quad \|f\|_{1,w} \leq \lambda \tag{IV–28}$$

Lemma IV–3 (Legendre-Fenchel conjugate of $\|\cdot\|_{\infty,\frac{1}{w}}$) *For all $f \in F_3$:*

$$\sup_{g \in F_3} \left(\langle f, g \rangle - \tau \|g\|_{\infty,\frac{1}{w}} \right) = \begin{cases} 0 & \textit{if } \|f\|_{1,w} \leq \tau \\ +\infty & \textit{otherwise} \end{cases} \tag{IV–29}$$

The right-hand side of this equality is defined as the indicator function of the $\ell_{1,w}$-ball of radius τ, and is denoted as $\iota_{\|\cdot\|_{1,w} \leq \tau}(f)$.

Lemmas IV–2 and IV–3 are direct consequences of Lemma IV–1, whose proof is presented below. For more details about dual norms and Fenchel-Legendre conjugates, see for instance [Boyd04] and [Fadili10].

Proof of Lemma IV–1. For any $f, g \in F_3$, we have, thanks to Cauchy-Schwartz inequalities:

$$\langle f, g \rangle = \sum_k \left(f_{(p,0,q)_k} g_{(p,0,q)_k} + f_{(p,1,q)_k} g_{(p,1,q)_k} + f_{(p,2,q)_k} g_{(p,2,q)_k} \right) \leq \sum_k w_k \sqrt{f_{(p,0,q)_k}^2 + f_{(p,1,q)_k}^2 + f_{(p,2,q)_k}^2} \cdot \underbrace{\frac{1}{w_k} \sqrt{g_{(p,0,q)_k}^2 + g_{(p,1,q)_k}^2 + g_{(p,2,q)_k}^2}}_{\leq \|g\|_{\infty,\frac{1}{w}}}$$

This leads to the result $\langle f, g \rangle \leq \|f\|_{1,w} \|g\|_{\infty,\frac{1}{w}}$. Then, for any non-zero $f \in F_3$, it can be verified that this inequality is tight with g defined as follows:

$$\begin{pmatrix} g_{p0q,k} \\ g_{p1q,k} \\ g_{p2q,k} \end{pmatrix} = \begin{cases} \begin{pmatrix} f_{p0q,k} \\ f_{p1q,k} \\ f_{p2q,k} \end{pmatrix} & \text{if } \begin{pmatrix} f_{p0q,k} \\ f_{p1q,k} \\ f_{p2q,k} \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \\ \sqrt{\dfrac{w_k}{|f_{p0q}|^2 + |f_{p1q}|^2 + |f_{p2q}|^2}} \begin{pmatrix} f_{p0q,k} \\ f_{p1q,k} \\ f_{p2q,k} \end{pmatrix} & \text{otherwise} \end{cases} \tag{IV–30}$$

Conversely, for any non-zero $g \in \mathbb{F}_3$, the tight case is achieved with f defined as follows:

$$\begin{pmatrix} f_{p0q,k} \\ f_{p1q,k} \\ f_{p2q,k} \end{pmatrix} = \begin{cases} \begin{pmatrix} g_{p0q,k} \\ g_{p1q,k} \\ g_{p2q,k} \end{pmatrix} & \text{if } k = k_0 \\ \sqrt{\dfrac{1}{w_k \left(|g_{p0q}|^2 + |g_{p1q}|^2 + |g_{p2q}|^2\right)}} \begin{pmatrix} g_{p0q,k} \\ g_{p1q,k} \\ g_{p2q,k} \end{pmatrix} & \text{otherwise} \end{cases} \tag{IV–31}$$

where k_0 is such that $k_0 = \arg\max_k \frac{1}{w_k} \sqrt{|g_{p0q,k}|^2 + |g_{p1q,k}|^2 + |g_{p2q,k}|^2}$. One can also check that the objects g defined in (IV–30) and f defined in (IV–31) are such that $\|g\|_{\infty,\frac{1}{w}} = 1$ and $\|f\|_{1,w} = 1$ respectively. □

Derivation of the unconstrained formulation Thanks to the expression (IV–26), the value of the projection operator $R_{w,a,\tau}$ (IV–23) can be expressed as the optimum z^* of the following optimization problem, that involves an indicator function:

$$\inf_{z \in \mathbb{R}^N} \frac{1}{2}\|z - x\|_2^2 + \iota\left(\|(z - a, D_h z, D_v z)\|_{1,w} \leq \tau\right) \tag{IV–32}$$

Using Lemma IV–3, (IV–32) can be reformulated as follows:

$$\inf_{z \in \mathbb{R}^N} \sup_{f \in \mathbb{F}_3} \frac{1}{2}\|z - x\|_2^2 + \left\langle f_{p0q}, z - a \right\rangle + \left\langle f_{p1,2q}, \nabla z \right\rangle - \tau \|f\|_{\infty,\frac{1}{w}} \tag{IV–33}$$

Then, thanks to the convexity (respectively concavity) of the optimized function in (IV–33) with respect to z (respectively f), the order of the infimum and supremum operators can be switched (see for instance [Weiss08], Theorem 2.4 about this property). This leads to the following equivalent formulation, where the terms that do not depend on z have been isolated:

$$\sup_{f \in \mathbb{F}_3} \left\{ -\tau \|f\|_{\infty,\frac{1}{w}} - \left\langle f_{p0q}, a \right\rangle + \inf_{z \in \mathbb{R}^N} \left\{ \frac{1}{2}\|z - x\|_2^2 + \left\langle f_{p0q}, \mathrm{Div} f_{p1,2q} \right\rangle \right\} \right\} \tag{IV–34}$$

In (IV–34), the z minimization sub-problem is, for a fixed f, a quadratic problem, whose solution has a closed form: the minimum value is $\frac{1}{2}\|x\|_2^2 - \frac{1}{2}\|x - f_{p0q} - \mathrm{Div} f_{p1,2q}\|_2^2$, reached for $z = x - f_{p0q} - \mathrm{Div} f_{p1,2q}$. Finally, by substituting this minimal value in (IV–34), we obtain that the optimum of the initial constrained problem can be calculated

as follows:

$$\Pi_{R_{w,a,\tau}}(x) = z^{\star} = x - f^{\star}_{(0)} - \mathrm{Div} f^{\star}_{(1,2)}$$
$$\text{where} \quad f^{\star} = \underset{f \in F_3}{\arg\min}\, \tau \|f\|_{8,\frac{1}{w}} + \left\| f_{(0)} - a + \frac{1}{2}\left(x - f_{(0)} - \mathrm{Div} f_{(1,2)}\right)\right\|_2^2 \tag{IV–35}$$

As announced, contrary to the initial formulation (IV–23), the optimization problem in (IV–35) is unconstrained; however, the price to pay for this simplification is that the optimized variable belongs to the vector space F_3, whose dimension is three times larger than the dimension of the initial space R^N.

Proximal operator associated to the weighted l_8-norm Before presenting the method used to solve (IV–35), we need to prove a lemma about the weighted l_8-norm defined on F_3. More precisely, we need to prove that the function $f \mapsto \|f\|_{8,\frac{1}{w}}$ is *simple*. As defined in [Nesterov07, Weiss08], this property means that, for any $\lambda \geq 0$, the proximal operator[6] associated to the function $f \mapsto \lambda \|f\|_{8,\frac{1}{w}}$ has a closed form and can be evaluated efficiently; this proximal operator is defined as follows, for any $f \in F_3$:

$$\mathrm{Prox}_{\lambda\|\cdot\|_{8,\frac{1}{w}}}(f) = \underset{g \in F_3}{\arg\min}\, \lambda \|g\|_{8,\frac{1}{w}} + \frac{1}{2}\|g - f\|_2^2 \tag{IV–36}$$

To prove this assertion, let us reformulate the expression (IV–36). Using Lemma IV–2, this minimization problem can be recast as:

$$\inf_{g \in F_3} \sup_{\substack{h \in F_3 \\ \|h\|_{1,w} \leq \lambda}} \langle h | g \rangle + \frac{1}{2}\|g - f\|_2^2 \tag{IV–37}$$

Using arguments similar to the one stated above, we can switch the infimum and supremum operators in (IV–37), leading to the following equivalent expression:

$$\sup_{\substack{h \in F_3 \\ \|h\|_{1,w} \leq \lambda}} \inf_{g \in F_3} \langle h | g \rangle + \frac{1}{2}\|g - f\|_2^2 \tag{IV–38}$$

In this expression, the solution to the g minimization sub-problem has a closed form: the minimal value is $\frac{1}{2}\|f\|_2^2 - \frac{1}{2}\|h - f\|_2^2$, reached for $g = f - h$. By substituting this solution to (IV–38), we obtain the following expression[7] for the proximal operator (IV–36):

[6] See for instance [Moreau65, Combettes11] and references therein for a formal introduction to proximal operators and associated results.

[7] The expression (IV–40) of the proximal operator (IV–36) can also be obtained as a consequence of the Moreau decomposition property. This result expresses that, if ϕ and ψ are two real-valued convex functions defined on a Hilbert space H, if additionally $\psi(y) = \sup_{x \in H} \langle x | y \rangle - \phi(x)$ for all $y \in H$ (meaning that ψ is the Legendre-Fenchel conjugate of ϕ, and *vice versa*), then the following identity holds:

$$x = \mathrm{Prox}_{\phi}(x) + \mathrm{Prox}_{\psi}(x) \qquad \text{for all } x \in H \tag{IV–39}$$

Accurate details (such as the exact hypotheses required on ϕ and ψ) on this decomposition property can

$$\begin{aligned} &\mathrm{Prox}_{\lambda\|\cdot\|_{8,\frac{1}{w}}}(f) = g^\star = f - h^\star \\ &\text{where } h^\star = \underset{h \in F_3}{\arg\min}\, \|h - f\|_2 \text{ subject to } \|h\|_{1,w} \leq \lambda \end{aligned} \tag{IV–40}$$

The minimization problem appearing in (IV–40) consists in finding the projection of a vector f on a $l_{1,w}$-ball (which is a convex set): for the usual l_1-norm, it is a well-known result that this projection is obtained by applying a soft-thresholding transformation to f (see for instance [Van Den Berg08]). For our weighted l_1-norm, the result is similar, although the soft-thresholding transformation needs to be adapted to account for the weight vector w. In our case, the expression obtained for $h^\star$ is the following[8]:

$$\begin{bmatrix} h^\star_{(0,k)} \\ h^\star_{(1,k)} \\ h^\star_{(2,k)} \end{bmatrix} = \max\left(0, 1 - \frac{v \cdot w_k}{\sqrt{f_{(0,k)}^2 + f_{(1,k)}^2 + f_{(2,k)}^2}}\right) \begin{bmatrix} f_{(0,k)} \\ f_{(1,k)} \\ f_{(2,k)} \end{bmatrix} \quad \text{for all } k \tag{IV–41}$$

where $v \geq 0$ is a constant independent of k, whose value is to be determined according to the following rules:

- if $\|f\|_{1,w} \leq \lambda$, then $v = 0$;
- otherwise, v must be set such that $\|h^\star\|_{1,w} = \lambda$.

The remaining issue consists in determining the value of v when $\|f\|_{1,w} > \lambda$. To proceed, let us introduce $r_k = \frac{1}{w_k}\sqrt{f_{(0,k)}^2 + f_{(1,k)}^2 + f_{(2,k)}^2}$ for all k, and ϕ a permutation of $[0, N-1]$ that sorts the coefficients r_k in ascending order:

$$r_{\phi(0)} \leq r_{\phi(1)} \leq \cdots \leq r_{\phi(N-1)} \tag{IV–42}$$

With these notations, we can express the weighted l_1-norm of the vector $h^\star$ defined in (IV–41) as follows:

$$\|h^\star\|_{1,w} = \sum_{l=0}^{N-1} w_{\phi(l)}^2 \cdot \max\left(0, r_{\phi(l)} - v\right) \tag{IV–43}$$

Now, let us introduce the scalar s_k for all $k \in [0, N-1]$, that we define to be equal to the weighted l_1-norm of $h^\star$ if the parameter v were set to be equal to $r_{\phi(k)}$ in the expression (IV–43). This definition leads to the following expression for s_k:

be found for instance in [Combettes05], Lemma 2.10. Then, (IV–40) is a direct consequence of (IV–39) and Lemma IV–3.

[8] For the sake of clarity, we skip the detailed proof of this result, and just give some clues about it. The expression (IV–41) can be obtained by observing that, according to the equivalence property stated in Sec. II.1.1 (and also in [Weiss08], Theorem 2.7), there exists a constant v such that $h^\star$, as defined in (IV–40), is also solution of $h^\star = \arg\min_{h \in F_3} \frac{1}{2}\|h - f\|_2^2 + v\|h\|_{1,w}$; this latter problem is separable, and has a closed-form solution which is exactly (IV–41). The additional conditions used to determine the value of v results from a classical property of the projection operator on a set defined as the inverse image of a real interval by a continuous function.

$$s_k = \sum_{l=k+1}^{N-1} w_{\phi(l)}^2 \left(r_{\phi(l)} - r_{\phi(k)} \right) \qquad \text{for all } k \in [0, N-1] \tag{IV–44}$$

One can check from (IV–44) that the sequence $(s_k)_{k=0,\dots,N-1}$ is decreasing. Finally, we define the index k_0 as:

$$k_0 = \underset{k \in [0,N-1]}{\arg\min}\, k \text{ subject to } s_k \leq \lambda \tag{IV–45}$$

One can notice that k_0 is well-defined, as $s_{N-1} = 0 \leq \lambda$.

Let us recapitulate, in order to make things more concrete. Assuming $k_0 \geq 1$, we can state the following:

$$\left\{ \begin{array}{l} \|h^*\|_{1,w} \text{ if } v = r_{\phi(k_0)} = s_{k_0} \leq \lambda \\ \|h^*\|_{1,w} \text{ if } v = r_{\phi(k_0-1)} \end{array} \right. \tag{IV–46}$$

As we want to determine v such that $\|h^*\|_{1,w} = \lambda$, we can guess from (IV–46) that v must be set in the interval $[r_{\phi(k_0-1)}, r_{\phi(k_0)}]$. This intuition is indeed correct, and a careful verification shows that the property $\|h^*\|_{1,w} = \lambda$ do hold when the parameter v is defined as follows:

$$v = \begin{cases} r_{\phi(0)} - \dfrac{\|f\|_{1,w} - \lambda}{\|f\|_{1,w} - s_0} & \text{if } k_0 = 0 \\ r_{\phi(k_0-1)} - \dfrac{\lambda - s_{k_0}}{s_{k_0-1} - s_{k_0}} \left(r_{\phi(k_0)} - \dfrac{s_{k_0-1} - \lambda}{s_{k_0-1} - s_{k_0}} \right) & \text{if } k_0 \geq 1 \end{cases} \tag{IV–47}$$

One can notice that this proof is constructive, in that it can be used as a skeleton to implement an algorithm to evaluate the proximal operator (IV–36). The algorithmic complexity of such algorithm is $O(N \log N)$, the most expensive step being the sorting of the sequence $(r_k)_{k=0,\dots,N-1}$ that is necessary to evaluate the threshold parameter v.

Solving the unconstrained formulation The method we propose to evaluate the projection operator $\Pi_{R_{w,a,\tau}}$ consists then in solving the unconstrained minimization problem (IV–35). In this latter formulation, the function to minimize can be decomposed as a sum between:

- a smooth term ($f \mapsto \left\| f(0) - \left(\frac{1}{2} \|x - f(0) - \mathrm{Div} f(1,2)\|_2^2 \right) \right.$), that is Lipschitz-differentiable with a Lipschitz constant smaller than $1 + \|\mathrm{Div}\|^2$, which is itself smaller than 9 (see [Chambolle04, Fadili11] for details about determining an upper bound to $\|\mathrm{Div}\|^2$);
- a non-smooth term ($f \mapsto \tau \|f\|_{8,\frac{1}{w}}$), that is simple in the sense defined in [Nesterov07, Weiss08].

To minimize this function, we use the iterative accelerated gradient descent method proposed in [Nesterov07, Weiss08] (Algorithm 4.2 in [Weiss08]). The "accelerated" adjective refers to the fact that, if J_p denotes the value of the objective function after the iteration

p and $J^{\prime}$ the minimum of this objective function, then the method ensures that $p J_p - J^{\prime} q$ is smaller than a term proportional to $\frac{1}{p^2}$; in the case of a classical gradient descent, this bound would be only proportional to $\frac{1}{p}$, which results in a slower convergence.

Finally, all these results lead to an iterative scheme to evaluate the projection operator $\Pi_{R_{w,a,\tau}}$: each iteration is performed in $O(N \log N)$ operations, and the Nesterov acceleration ensures a quadratic convergence rate. This approach is much slower than what is needed to compute the other projection operator $\Pi_{D_{y,\square}}$ involved in the alternated projection scheme (IV–12). However, we observed that a careful initialization of the gradient descent provides significant speed up of the convergence (see the pseudo-code of the full algorithm in Fig. IV–7).

IV.2.3.5 Overall reconstruction algorithm

The proposed reconstruction algorithm is based on alternated projections of the iterated reconstructions over the data set $D_{y,\square}$ and the regularization set $R_{w,a,\tau}$, involving two scalar parameters $\square$ and τ. The parameter $\square$ controls the size of the data set $D_{y,\square}$, and is set proportional to the noise level that affects the measurements; we assume that this information is available prior to the reconstruction (in a real acquisition device, the noise level could be evaluated through a calibration step for instance). However, setting the parameter τ is not straightforward, as it is likely to be highly dependent on the image content. Therefore, we developed an adaptive heuristic to dynamically adjust this parameter during the iterative reconstruction process.

This dynamic adjustment process relies on the following observation: the alternated projection scheme (IV–12) produces a sequence of estimators that converge to the intersection $R_{w,a,\tau} \cap D_{y,\square}$, but this intersection is empty when τ is below a certain threshold $\tau^{\prime}$ (if the image a is not constant, the set $R_{w,a,\tau}$ itself is empty when $\tau = 0$). Therefore, the algorithm becomes non-convergent if $\tau < \tau^{\prime}$.

Based on this remark, we propose a reconstruction algorithm where τ is initialized at an arbitrary high value τ_0, and then reduced until the algorithm becomes non-convergent, as detailed in the pseudo-code in Fig. IV–7. The algorithm returns the result (denoted as $x_{candidate}$ in Fig. IV–7) obtained with the smallest value of τ that leads to convergence. One should notice that, compared to its mathematical definition (IV–23), the regularization set projector $\Pi_{R_{w,a,\tau}}$ takes here two input arguments (respectively x_{in} and f_{in}), and returns also two outputs (respectively x_{out} and f_{out}):

- x_{in} and x_{out} are respectively the vector being projected and the result of the projection (as defined in (IV–23)),
- f_{in} is used as the initialization of the gradient descent solving the auxiliary problem (IV–35),
- f_{out} refers to the object returned by this gradient descent (denoted as $f^{\prime}$ in (IV–35)).

```
function FRAME RECONSTRUCTION(x_init, τ_init, Δ_tolvar, α)
    p Ð 0
    τ Ð τ_init
    (x_0, f_0) Ð (x_init, 0)
    loop
        p Ð p ` 1                                          Ź increment the loop counter
        x_{p´1/2} Ð Π_{D_{y,□}}(x_{p´1})                   Ź projection on the data set
        (x_p, f_p) Ð Π_{R_{w,a,τ}}(x_{p´1/2}, f_{p´1})     Ź projection on the regularization set
        δ_p Ð ‖x_p ´ x_{p´1}‖_2 / ‖x_{p´1}‖_2              Ź relative variation of x
        if δ_p ă Δ_tolvar then
            x_candidate Ð x_p                              Ź save the current estimate of x
            τ Ð α ¨ τ          Ź reduce the bound τ (α is chosen such that 0 ă α ă 1)
        else if detect non-convergence then
            return x_candidate                             Ź return the previously saved estimate of x
        end if
    end loop
end function
```

Figure IV–7: Pseudo-code of the iterative reconstruction algorithm. The algorithm takes four arguments as input: x_{init} and τ_{init}, which are the initial values for the reconstructed frame and the hybrid total variation bound, and Δ_{tolvar} and α, that controls the general behavior of the algorithm. When the relative variation between successive iterates falls below Δ_{tolvar}, the bound τ is reduced by a factor α (chosen such that $0 \ll \alpha \ll 1$, and close to 1 in practice), until the algorithm becomes non-convergent. The second input and output arguments of the operator $\Pi_{R_{w,a,\tau}}$ are used to initialize the gradient descent loop that solves the auxiliary problem (IV–35), and to save the solution of this auxiliary problem.

Mathematically speaking, the value of the input argument f_{in} is not important: as $\Pi_{R_{w,a,\tau}}$ is defined as a convex optimization problem, its value does not depend on the initialization of the gradient descent used to evaluate it. However, by suggesting a "good" initialization point, this gradient descent converges to the optimum in fewer iterations, which dramatically reduces the computation time.

One of the challenging issues raised by the algorithm presented in Fig. IV–7 is to detect that the sequence of estimators does not converge for a given value of τ, since we do not have any result on the theoretical convergence rate of this sequence of estimators, To solve this issue, we developed an empirical approach based on the properties of the sequence (δ_p), which measures the relative variations between two consecutive iterates. More precisely, to detect whether the algorithm starts to diverge and therefore should be stopped at a given iteration p^*, we perform the following test:

1. linear regression over the truncated sequence of values of $(\log \delta_p)$ for $p^* - \Delta p + 1 \leq p \leq p^*$, where Δp is a fixed parameter, returning a slope of evolution s;
2. stop (i.e. decide that the current value of τ is too small for the algorithm to converge) if s is above a certain threshold s_{max}.

The proposed non-convergence test evaluates the mean variation of the sequence $\log \delta_p$ over a window of Δp samples: if this sequence increases at a rate higher than s_{max}, then we assume that the algorithm is diverging. Typical parameter values for this test are $\Delta p = 100$ iterations and $s_{max} = -10^{-4}$ per iteration. Finally, to improve the computation speed, we typically perform this test every 25 iterations only: as the linear regression is performed over a sliding window, the value of the resulting slope is not likely to change much from one iteration to the next, which justifies this approach.

IV.2.4 Numerical simulations

IV.2.4.1 Methodology

In order to validate the video reconstruction method based on partial Fourier modulus measurements presented in Sec. IV.2.3, we run this reconstruction method on numerically simulated measurements, generated from synthetic and real test sequences. We use two test video sequences here:

- *Disks 2*, sized $256 \times 256 \times 80$ (height$\times$ width$\times$ number of frames), which is a synthetic sequence representing disk shapes of random intensity levels and sizes (diameters between 5 and 25 pixels), and moving with random directions and speeds. The typical distance travelled by the disks between two frames is about 1 to 3 pixels. This sequence differs from the test sequence *Disks 1* used in Sec. IV.1.4 in that the intensity of its background is constant over time, and the boundaries of the disks are not blurred.
- *Amoeba*, sized $256 \times 256 \times 80$, which is a microscopy sequence of moving and stretching cells having similar sizes and speeds as in *Disks 2*. This sequence is the one already used in simulations involving linear Fourier measurements in Sec. IV.1.4.

Simulations were conducted on the *Disks 2* sequence using 15% of magnitude Fourier measurements; on *Amoeba*, we increased the sampling rate to 25% of Fourier samples to handle the more complex nature of the signal. In both cases, we assumed that the first frame of the sequence is a key-frame, i.e. is known prior to the reconstruction: we used an input to initialize the process, and then we progressively recovered all the following frames as described in Sec. IV.2.2.

IV.2.4.2 Qualitative and quantitative results

We first present results obtained with a weight map w set in a uniform manner in the hybrid total variation (hTV): the corresponding reconstructions obtained for the sequences *Disks 2* and *Amoeba* are presented on Fig. IV–8 and Fig. IV–9. Figure IV–10 also presents the evolution of the frame-by-frame reconstruction error (measured as the root mean squared error RMSE between the original and reconstructed frames) as a function of frame index t (i.e. the time) in the case of these two sequences.

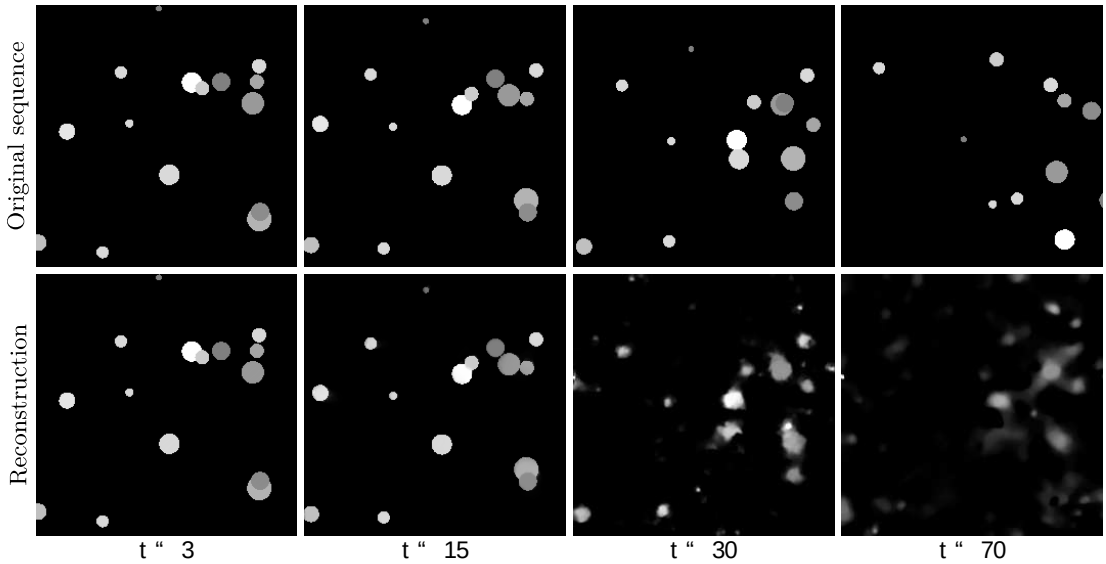


Figure IV–8: Reconstruction results obtained for the test sequence *Disks 2*, using 15% of Fourier modulus measurements, and one key-frame at t “ 0. We used a uniform weight map (w “ 1) in this example.

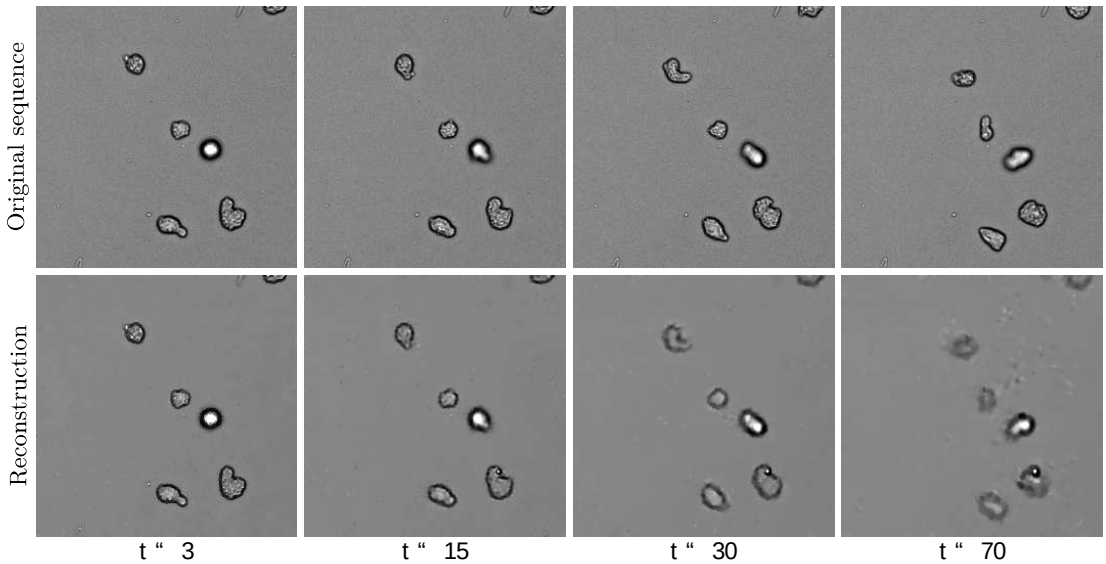


Figure IV–9: Reconstruction results obtained for the test sequence *Amoeba*, using 25% of Fourier modulus measurements, and one key-frame at t “ 0. We used a uniform weight map (w “ 1) in this example.

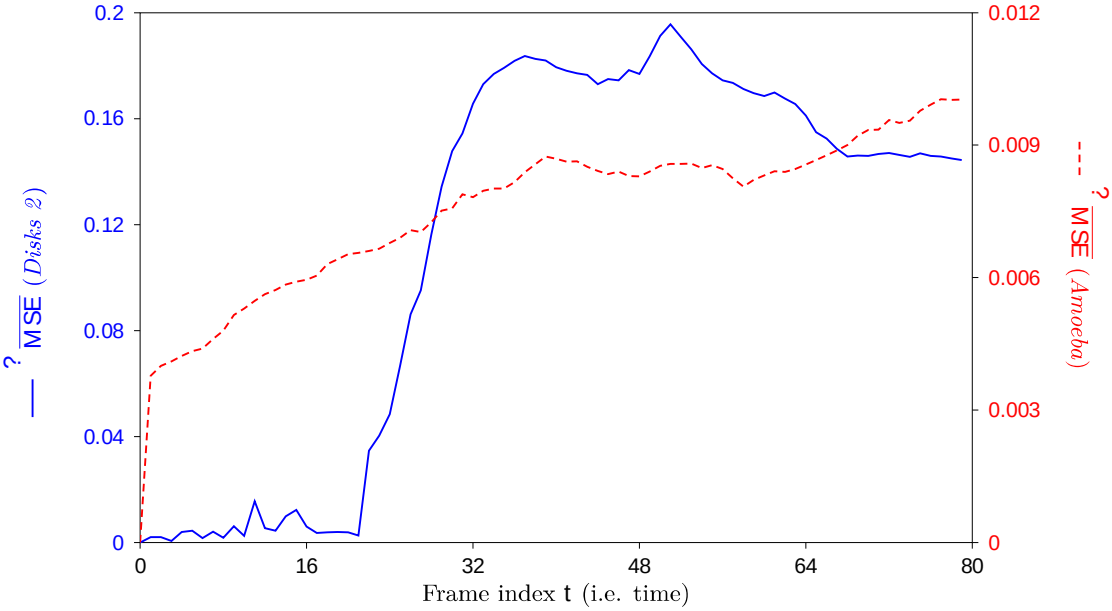


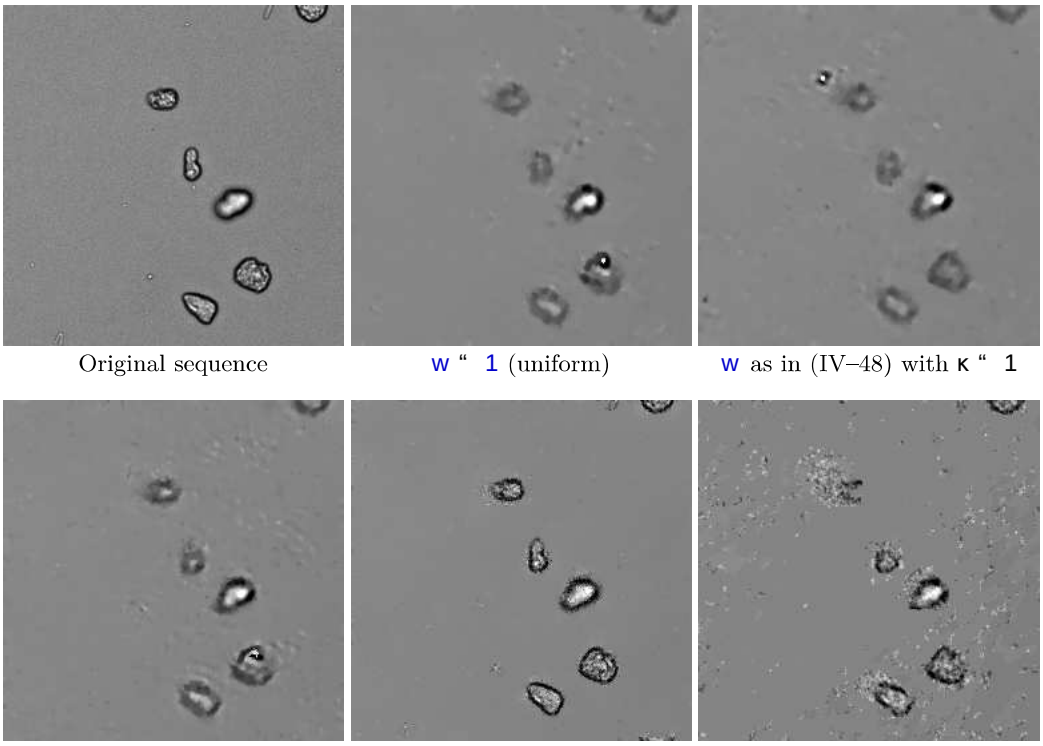
Figure IV–10: Frame-wise root mean squared error of the reconstructed video sequences *Disks 2* and *Amoeba* as a function of the frame index t (i.e. the time).

These results show that the distortions introduced by the reconstruction method increase with time, i.e. with the distance to the initial key-frame: while in both sequences the reconstructed frames for t ă 10 are quite similar to the original ones, errors become significant close to the end of the sequences, but exhibit different characteristics:

- For *Amoeba*, the RMSE increases progressively and quite regularly with time t, which is characteristic of error accumulation. Visually, this results in an increasing blurring effect.
- For *Disks 2*, the RMSE increases sharply at time t " 22, and then continues to grow over the next 10 frames, leading to a reconstruction that is completely inconsistent with the original sequence for t ą 30. This behavior is due to the fact that the algorithm outputs an erroneous reconstructed frame at t " 22, whose errors are then propagated. On the contrary, frames corresponding to t ă 22 were almost perfectly recovered. The reason explaining why the reconstruction fails at this particular point remains however unclear.

IV.2.4.3 Weight map in the hybrid total variation

Results presented in Sec. IV.2.4.2 were obtained with the weight map parameter w in the hybrid total variation (hTV) set to a uniform value (w " 1). However, this parameter can be refined with some prior hypothesis made on the reconstructed frames, by including some motion prediction heuristics for instance. More precisely, we tested an approach consisting in setting this parameter such that w ru, vs is small at the spatial positions



Original sequence | $w = 1$ (uniform) | w as in (IV–48) with $\kappa = 1$

w as in (IV–48) with $\kappa = 10$ | w as in (IV–48) with $\kappa = 100$ | w as in (IV–48) with $\kappa = 1000$

Figure IV–11: Frame $t = 70$ of the sequence *Amoeba* reconstructed with different parametrizations of the spatially varying weight maps w.

$(u, v) \in \Omega$ where we expect to observe edges in the reconstruction (see Sec. IV.2.3.3 for details about the notations).

We carried several reconstruction experiments in which the design of the weight map used to reconstruct the frame x_t were based on the gradient of the previous frame x_{t-1} in the sequence. In particular, we considered weight maps defined as follows:

$$w(u, v) = \exp\left(-\kappa \left(|D_h x_{t-1}(u, v)|^2 + |D_v x_{t-1}(u, v)|^2\right)\right) \tag{IV–48}$$

with $\kappa \geq 0$ a parameter to tune. The underlying assumption guiding this choice is that edges in the reconstructed frame are expected to be located close to the edges of the previous frame, which corresponds to a very simplistic motion prediction heuristic.

However, results presented in Fig. IV–11 show that – in spite of the simplicity of the motion prediction underlying hypothesis – a careful choice of the weight map can indeed reduce the reconstruction artifacts. In particular, in the sequence obtained with $\kappa = 100$, we were able to remove the blurring effect and contain the error accumulation phenomenon that otherwise dramatically degrades the frames close to the end of the test sequences (i.e. far from the key-frame). However, automatic calibration of the parameter κ remains challenging: for instance, all our attempts to reconstruct *Disks 2* using a non-uniform weight map resulted in a degradation of the reconstructed sequence, compared to what

we obtained with $w = 1$. Other forms of weight maps w were also tested, such as:

$$w(u,v) = \frac{1}{\eta + |D_h x_{t-1}(u,v)|^2 + |D_v x_{t-1}(u,v)|^2} \tag{IV–49}$$

with $\eta > 0$ a regularization parameter, but led to unsatisfactory results and numerical instability. Setting w in a robust manner is still an open question.

IV.3 Conclusion

In this chapter, we presented two sparsity-based video microscopy reconstruction methods, using random projections in the Fourier domain, exploiting either linear (amplitude and phase) or non-linear (amplitude only) measures. In the former case, the reconstruction scheme exploits general CS reconstruction results with a 3D total variation based reconstruction functional, where all the frames of the sequence are reconstructed in a joint manner. In the latter case, the reconstruction relies on an alternated projection algorithm inspired by previous phase retrieval techniques, and modified to account for the specific sparsity properties encountered in video microscopy sequences. This phase retrieval algorithm differs also from the TV-3D-based method in that the frames of the sequence are reconstructed recursively, starting from an initial key-frame assumed to be known, and recovering each frame using the result obtained for its predecessor in the sequence.

The results demonstrate that video reconstruction can be performed from partial Fourier measurements, opening the way for designing "compressed sensing" devices relying on optical Fourier transforms. However, numerical simulations show significant differences in terms of performances depending on whether the acquired samples include a Fourier phase information or not: in particular, the quality of the resulting sequences is much higher if phase information is available. Moreover, the phase retrieval reconstruction algorithm used in absence of such phase information is more sensitive to error accumulation during the reconstruction process due to its frame recursion behavior compared to the CS reconstruction, that recovers all the frames of the sequence in a joint manner. The phase retrieval algorithm also involves a large number of parameters that may be difficult to tune, and that may significantly impact the convergence rate of the overall method: obtaining theoretical results about this impact is still an open question, and would be definitively worth investigating.

Chapter V

Using CS as a denoising method?

The CS reconstruction operation that recovers the signals of interest from the partial measurements vectors can be formulated either as a convex optimization problem or as a greedy heuristic such as orthogonal matching pursuit (see Chap. II). In both cases, this reconstruction step enforces sparsity priors on the signal to reconstruct, which – as a side effect – tend to filter out the noisy component present in the measurements, as the latter violates the sparsity assumptions. In some preliminary work [Marim09], the authors exploited this characteristic to design an image denoising method based on the fusion of several CS reconstructed images, which appears to be efficient in denoising low-light microscopy images, whose noise component can be modeled as a mixture of an additive Gaussian and a Poisson model. The work presented in this chapter was carried out bearing in mind the study of denoising methods applicable to such low-light microscopy images, in particular the extension of the method proposed in [Marim09].

More precisely, the behavior and performance of denoising algorithms are governed by one or several parameters, whose optimal settings depend on the content of the processed image and the characteristics of the noise, and are generally designed to minimize the mean squared error (MSE) between the denoised image returned by the algorithm and a virtual *ground truth*. In this chapter, we introduce a new unbiased risk estimator (PG-URE) of the MSE applicable to a mixed Poisson-Gaussian noise model that unifies the widely used Gaussian and Poisson noise models in low-light microscopy applications. We propose a stochastic methodology to evaluate this estimator when little is known about the internal machinery of the considered denoising algorithm, and we analyze both theoretically and empirically the characteristics of the proposed PG-URE estimator. Finally, we evaluate the PG-URE-driven parametrization for three standard denoising algorithms, with and without variance stabilizing transforms, and different characteristics of the Poisson-Gaussian noise mixture. Beyond the application to CS denoising, we emphasize that this new tool can be used to optimize the parameters involved in any denoising algorithm, assuming that the mixed Poisson-Gaussian noise model holds for the processed images.

Finally, let us mention that most of the work presented in this chapter was proposed

as a journal paper [Le Montagner13d], and is currently undergoing peer review.

V.1 Introduction **102**
V.1.1 Denoising background . . . 102
V.1.2 Denoising via aggregation of multiple CS reconstructions . . . 104
V.1.3 SURE and parameter estimation . . . 106
V.1.4 Chapter outline . . . 107

V.2 Mixed Poisson-Gaussian noise model **107**
V.2.1 Generalized unbiased risk estimators . . . 107
V.2.2 Poisson noise and associated PURE estimator . . . 108
V.2.3 Mixed Poisson-Gaussian noise . . . 108
V.2.4 Unbiased risk estimator for the MPG model . . . 110

V.3 Stochastic evaluation of the Poisson-Gaussian URE **111**
V.3.1 Why is a deterministic evaluation of PG-URE impossible? . . . 111
V.3.2 Evaluation of the first-order derivative term . . . 112
V.3.3 Evaluation of the second-order derivative term . . . 113
V.3.4 Empirical PG-URE estimator . . . 114
V.3.5 Variance with respect to the random perturbations . . . 115

V.4 Numerical validation and application **117**
V.4.1 Simulation goals and process . . . 117
V.4.2 Influence of the amplitude parameters g and $\square$ . . . 119
V.4.2.1 Parameter g . . . 119
V.4.2.2 Parameter $\square$ . . . 122
V.4.3 Optimization of the denoising parameters θ driven by PG-URE . . . 124

V.5 Conclusion **127**

V.A Derivation of the PG-URE estimator **127**

V.B Optimal perturbation for the second-order derivative term **128**
V.B.1 Expression of V_{δ} . . . 129
V.B.2 Optimal probability distribution . . . 131

V.1 Introduction

V.1.1 Denoising background

Image denoising is one of the most studied problem in image processing. Many algorithms have been developed to tackle this issue, with various characteristics in terms of denoising

efficiency, applicability to different types of images and noise models, and running time. Among this large collection of available methods, we can single out the following families of algorithms:

Thresholding in a transformed domain. The general principle of this type of denoising algorithm is to apply a linear transform to the image in order to obtain a sparse representation of it, to threshold the obtained coefficients in a non-linear separable manner, and finally to revert the initial linear transform. This generic method can be instantiated with several types of sparsifying linear transforms, including orthogonal wavelets [Donoho94, Donoho95a] (which is the original method), translation-invariant wavelets [Coifman95], ridgelets [Candès99], curvelets [Candès04, Zhang08], etc. One of the key practical interest of this type of methods is that, assuming that the considered linear transform comes with a fast computation algorithm such as the fast discrete wavelet transform (which is generally the case), they can be applied very efficiently even on large 2D or 3D signals.

Variational based methods. With this type of method, the denoised image is obtained as a minimizer of a functional, which is designed to enforce certain properties on the result. One of the most famous representative algorithm of this category is total variation filtering [Rudin92], which enforces a piecewise constant structure on the images, and whose formulation is the following:

$$\hat{x} = \arg\min_{x} \frac{1}{2}\|x - y\|_2^2 + \lambda \|x\|_{TV} \tag{V–1}$$

where y is the input image to denoise, $\hat{x}$ the output denoised image, $\|\cdot\|_{TV}$ is the 2D total variation semi-norm (I–10), and $\lambda \geq 0$ is a parameter adjusting the general behavior of the method: the higher λ is, the more noise will be removed, while for $\lambda \approx 0$, the output image will be constrained to match the input. Total variation filtering (V–1) is known to be very efficient in removing noise while preserving sharp edges in cartoon-like images. More generally, variational-based methods are very flexible and can easily be tuned to account for different types of image models (see for instance [Aelterman12, Zhu12]). It can also be noticed that several other denoising methods such as anisotropic diffusion [Perona90] can be seen as variational methods (see [Kawohl04]).

Patch-based non-local methods. This new category of denoising methods has been introduced by [Buades05], with the *non-local means* (NLM) denoising algorithm. The idea of this type of method is to exploit the spatial redundancy that exists in natural images and to compute the denoised value of a given pixel by taking into account the values of all "similar" pixels in the noisy image, including possibly pixels located at long distances (i.e. non-local). Formally, the denoised value for a given pixel k is defined as follows:

$$\hat{x}_k = \frac{1}{Z(k)} \sum_{l} w(k,l)\, y_l \tag{V–2}$$

where y is the noisy image, and where the summation index l visits all the pixels. In (V–2), the weight $w(k,l) \geq 0$ measures the similarity between the neighborhoods of pixels k and l, and $Z(k) = \sum_l w(k,l)$ is a normalization factor. This neighborhood similarity measure $w(k,l)$ is actually a parameter of the method; in [Buades05], the authors propose to define it as:

$$w(k,l) = \exp\left(-\frac{\|\pi(k) - \pi(l)\|_2^2}{h^2}\right) \tag{V–3}$$

where $\pi(k)$ represents a restriction of the whole noisy image y to a small window (a patch) around the pixel k, and $h > 0$ is a parameter adjusting the "denoising intensity" of the method, similarly to λ in (V–1). Several variants of NLM involving patches with different shapes have been proposed (see for example [Deledalle11]), although the most common implementation of NLM uses centered square patches.

Most of the state-of-the-art denoising algorithms [Elad06, Dabov07] consist in refinements of and crossings between these classical ones: for instance, BM3D [Dabov07] consists in looking for image patches that present similarities (as in [Buades05]), and then applying a thresholding operation on group of similar patches (in the manner of [Donoho95a]). One can refer to [Milanfar13] for a more comprehensive overview of filtering methods applied to denoising problems.

All these algorithms have in common that their behaviors is controlled by one or several parameters, whose optimal values are almost always dependent on the data being processed. More precisely, if y is the noisy image being observed, f_θ a denoising algorithm depending on a set of parameters θ, and $\hat{x} = f_\theta(y)$ the denoised image returned by the algorithm, it is often desirable to select θ such that it optimizes a similarity criteria between $\hat{x}$ and a *ground truth* noise-free image x. Several image similarity criteria exist, with various characteristics in terms of correlation to the human perception system (see for instance [Zhang12] for a detailed review of these criteria). In this chapter, we focus on the mean squared error (MSE), defined as follows:

$$\mathrm{MSE} = \frac{1}{N}\|f_\theta(y) - x\|_2^2 \tag{V–4}$$

where N is the size (i.e. the number of pixels) of the considered 2D images. This criterion is certainly not the best one with respect to the human perception system correlation issue, but its mathematical tractability makes it a valuable tool in image processing (see [Wang09]).

V.1.2 Denoising via aggregation of multiple CS reconstructions

Using ideas inspired by the CS theory for denoising tasks, as proposed by [Marim09], is justified by the two following remarks. First, in terms of frequency analysis, the energy of a noise-free natural image is mostly concentrated in the low-frequency area of its Fourier

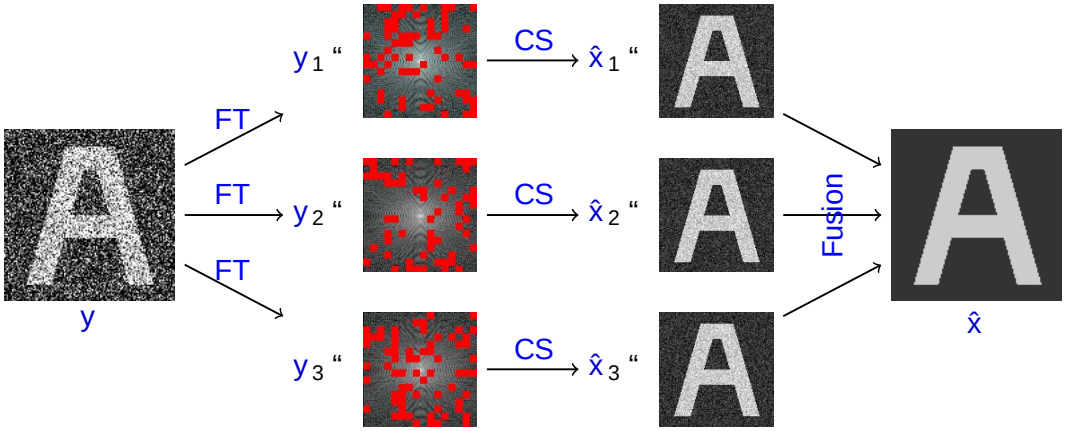


Figure V–1: Denoising scheme using several CS reconstructions. From a noisy image y, we generate several CS measurement vectors y_r, by taking the Fourier transform (FT) of y and rejecting a random subset of the Fourier coefficients. Then, each y_r is used to produce an estimator $\hat{x}_r$ of the original signal through a CS reconstruction scheme (P_{TV}) involving TV minimization. Finally, all the $\hat{x}_r$ are combined in a $\hat{x}$ estimator, with improved faithfulness properties.

domain, while a white noise has a uniform spectrum intensity: this implies that, in a noisy natural image, it is the high-frequency part of the image that is the most significantly affected by the noise, making the information coming from this part of the spectrum more inaccurate than its low-frequency counterpart. Second, the theory of CS states that a sparse or compressible signal can be recovered from a non-adaptive subset of linear noisy measurements.

Based on these two ideas, a CS-based denoising algorithm should follow the following workflow (see also Fig. V–1):

1. Generate several subsets of correlated measurements by considering the Fourier transform of the noisy image, and rejecting most of its – inaccurate – high frequency coefficients.
2. For each measurement subset y_r ($r = 1, \ldots, R$), compute an estimator $\hat{x}_r$ of the original non-noisy image through CS reconstruction (for instance (P_{TV})).
3. Combine all the $\hat{x}_r$ estimators in a proper way to produce an improved estimator $\hat{x}$ of the original noise-free image x.

In this global denoising scheme, the y_r can be considered as approximative and partially correlated observations of the original image y, and the $\hat{x}_r$ as partially correlated coarse estimators of the noise-free image. These $\hat{x}_r$ are aggregated thanks to a fusion operator g, designed such that the denoised image $\hat{x} = g(\hat{x}_1, \ldots, \hat{x}_R)$ presents improved faithfulness properties.

The denoising workflow presented in Fig. V–1 involves several parameters: the number R of intermediate CS reconstructions to fuse, the data fidelity bound $\square$ involved in the CS

reconstruction problem (P_{TV}), the method used to select the Fourier coefficients involved in the composition of the partial measurement vectors y_r, the method used to fuse the $\hat{x}_r$, etc.

A strategy to optimize these parameters could be to minimize a mean squared error criteria, as proposed above. However, except in special contexts such as simulations when the ground truth x is known, the MSE (V–4) is impossible to evaluate directly and cannot be used as an objective criteria for parameter optimization tasks. The unbiased risk estimator tools, among which SURE [Stein81, Donoho95b] is a well-known representative, aim at tackling this issue.

V.1.3 SURE and parameter estimation

Stein's unbiased risk estimator (SURE) [Stein81] is a well-known result in the statistics field, that has recently received a growing interest from the image processing community (see for instance [Donoho95b, Benazza-Benyahia05, Van De Ville09]).

The SURE estimator is built upon the hypothesis that the image y to denoise results from a ground truth x corrupted by a white additive Gaussian noise b:

$$y = x + b \quad \text{with} \quad b \sim \mathcal{N}\left(0, \sigma^2 \mathrm{Id}\right) \tag{V–5}$$

where the standard deviation parameter σ is assumed to be known. From this noise model, and given a denoising function[1] f, a similarity criteria SURE is defined as:

$$\mathrm{SURE} = \frac{1}{N}\|f(y) - y\|_2^2 - \sigma^2 + \frac{2\sigma^2}{N}\mathrm{Div}\, f(y) \tag{V–6}$$

where $\mathrm{Div}\, f(y) = \sum_k \frac{\partial f_k}{\partial y_k}(y)$ stands for the divergence of the function f. In [Stein81], the author showed that, up to some technical points[2], MSE and SURE have equal expected values over all the realizations of the random variable b: $E\{\mathrm{MSE}\} = E\{\mathrm{SURE}\}$. This means that, in practice, SURE is an estimator of the MSE similarity criteria, and can be taken as a surrogate. The empirical equality $\mathrm{SURE} \approx \mathrm{MSE}$ has been confirmed in various particular situations: see for instance [Ramani08, Van De Ville09].

A significant difference between MSE and SURE is that the latter does not depend on the ground truth x. As x is generally not available in real-life problems, this property dramatically increases the interest of SURE over MSE in practical applications. For instance, if $\theta_1, \ldots, \theta_K$ are K admissible parameter values for a denoising algorithm f_θ,

[1] From now on, we will drop the subscript θ from f_θ for the sake of readability, when no ambiguity is possible.

[2] For the following result to hold, f must be weakly differentiable, and its partial derivatives must fulfil $E\left\{\sum_k \left|\frac{\partial f_k}{\partial y_k}(y)\right|\right\} < \infty$. These technical conditions will always be assumed to be true, as well as all other requirements on the regularity of f that could be encountered in this work. Please note however that some realistic denoising functions f may not be even weak-differentiable: for instance, wavelet hard-thresholding [Donoho94] is not.

it is possible to select a "best-performing" value $\theta_{k^\star}$ in the sense of the MSE criterion as the one that minimizes $\mathrm{SURE}(\theta_k)$. Such selection is data-adaptive (it depends on y), and objective (it does not rely on human expert evaluation), opening the way to automated parameter estimation.

V.1.4 Chapter outline

The work presented in this chapter is built around the resolution of two issues that restrict in practice the use of SURE for automatic parameter tuning. First, SURE relies on the hypothesis of additive white Gaussian noise (V–5), which may not account for situations encountered in bio-imaging applications: for example, in this case, noise intensity may not be uniform in the whole image as assumed in (V–5), but rather depend on the presence of biological objects, and more generally on the value on the underlying signal (see [Starck98, Zhang08]). The extension of SURE to a more realistic mixed Poisson-Gaussian noise model is thus proposed in Sec. V.2, extending the work in [Luisier11].

The second limitation comes from the divergence term that appears in the expression of the SURE estimator (V–6). More precisely, the evaluation of the partial derivatives $\frac{\partial f_k}{\partial y_k}(y)$ is not a trivial task when the denoising algorithm f is not defined by a closed-form expression: such situations include variational-based algorithms (e.g. total variation minimization [Rudin92]) and diffusion methods (e.g. anisotropic diffusion [Perona90]). To tackle this issue, a methodology based on the introduction of small stochastic perturbations to y (similar to the one introduced by [Ramani08]) is proposed in Sec. V.3.

A numerical validation of the proposed framework is presented in Sec. V.4, along with several practical examples of parameter estimation.

V.2 Mixed Poisson-Gaussian noise model

V.2.1 Generalized unbiased risk estimators

The original SURE estimator [Stein81] (V–6) was designed around the Gaussian noise hypothesis (V–5). Other types of unbiased risk estimators have been derived since then to handle different noise models. It is worth noting that unbiased risk estimators can be refined to account for several phenomena that affect the image formation, beyond simple noise: see for instance [Vonesch08, Pesquet09, Eldar09a, Giryes11, Xue12] and references therein for applications of SURE-like estimators to deconvolution problems. An exhaustive review of the existing unbiased risk estimators applied to image restoration problems is however beyond the scope of the current work, and we focus here on pure denoising problems involving noise models encountered in microscopy imaging applications.

V.2.2 Poisson noise and associated PURE estimator

A usual noise model in bioimaging is the Poisson model, which is quite common in low-light fluorescence microscopy imaging, and more generally in imaging modalities that operate in low-signal conditions (see for instance [Starck98, Zhang08]). In this model, each observed pixel value y_k is assumed to be the result of a Poisson random process of intensity x_k, independent of the other pixels y_l. Formally:

$$y \sim \mathcal{P}(x) \tag{V–7}$$

A qualitative property of Poisson images is that the noise is signal dependent: its variance is higher on bright objects than in the dark background. This behavior is fundamentally different from what is modeled with the additive white Gaussian noise hypothesis (V–5), for which the noise intensity is uniform and independent of the value of the ground truth signal.

A *Poisson unbiased risk estimator* (PURE) of the MSE similarity criteria has been derived in [Luisier10] for the Poisson noise model (V–7):

$$\mathrm{PURE} = \frac{1}{N}\left(\|f(y)\|_2^2 - \|y\|_2^2 + 2\left\langle y, f^{(-1)}(y)\right\rangle + \langle x, 1\rangle - \langle y, y\rangle\right) \tag{V–8}$$

where the image-valued function $f^{(-1)}(y)$ is defined as $\left(f^{(-1)}\right)_k(y) = f_k(y + e_k)$ for all pixels k. For smooth functions f, this expression can be simplified using the following first-order Taylor approximation of $f^{(-1)}(y)$:

$$f^{(-1)}(y) \approx f(y) - \mathcal{J}f(y) \tag{V–9}$$

where the image-valued function $\mathcal{J}f(y)$ is defined as $(\mathcal{J}f)_k(y) = \frac{\partial f_k}{\partial y_k}(y)$ for all pixels k. Thanks to this Taylor approximation, (V–8) becomes:

$$\mathrm{PURE} = \frac{1}{N}\left(\|f(y) - y\|_2^2 - 2\langle y, \mathcal{J}f(y)\rangle - \langle x, 1\rangle - \langle y, y\rangle\right) \tag{V–10}$$

The terms $\left\langle y, f^{(-1)}(y)\right\rangle$ in (V–8) and $\langle y, \mathcal{J}f(y)\rangle$ in (V–10) play roles similar to the divergence term in SURE (V–6), in that they probe how small modifications of the observed image y impact the output of the denoising algorithm f. Their evaluation are subject to technical difficulties similar to those mentioned in Sec. V.1.4 for SURE.

V.2.3 Mixed Poisson-Gaussian noise

The Gaussian and Poisson noise models (V–5) and (V–7) do not individually account for the various phenomena involved with real image acquisition processes in fluorescence microscopy. Therefore, in the following, we consider a *mixed Poisson-Gaussian* (MPG) noise model, similar to the ones proposed in [Starck98, Zhang07b, Foi08, Delpretti08,

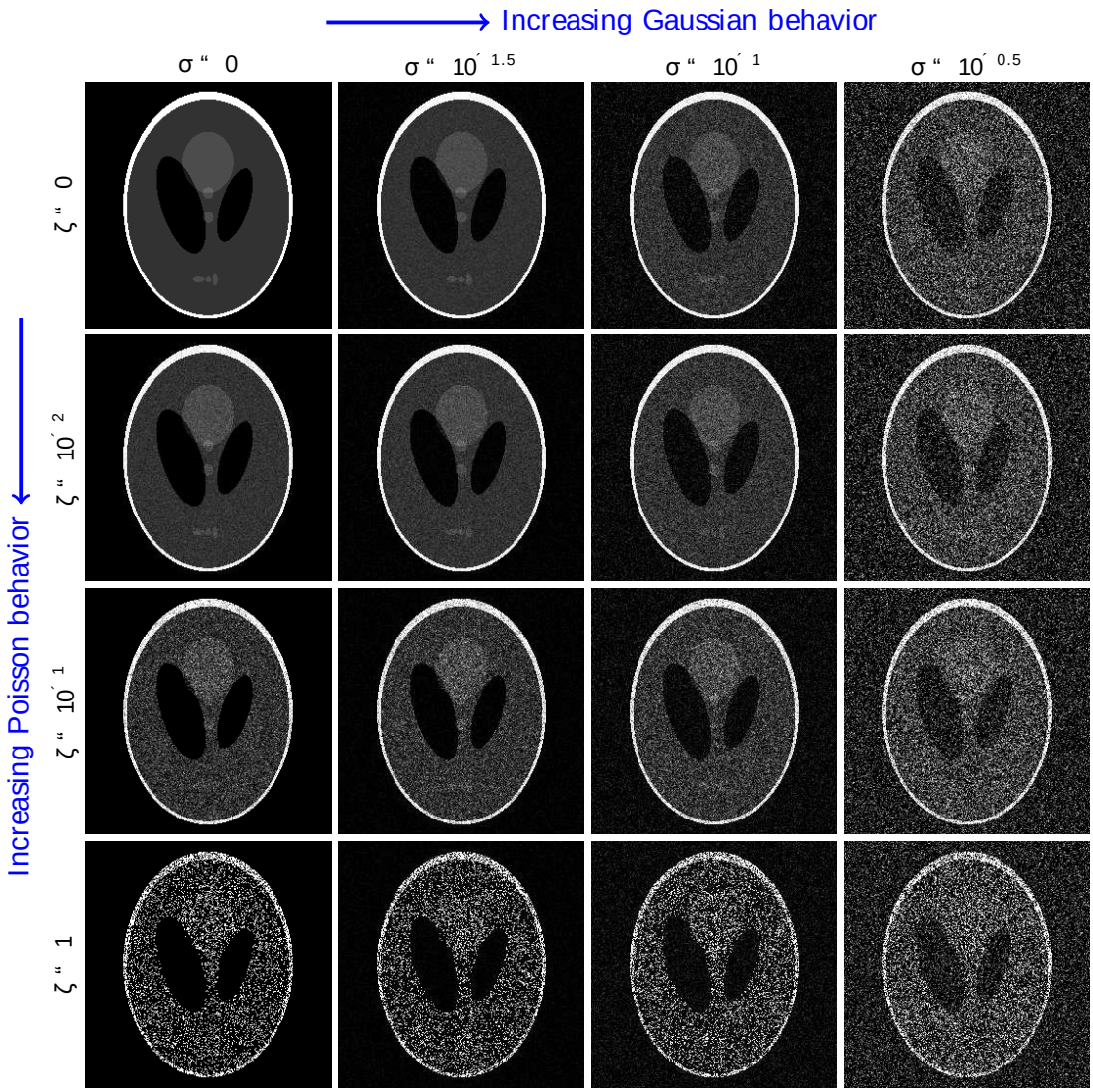


Figure V–2: Example of Shepp-Logan images y corrupted with the mixed Poisson-Gaussian noise model (V–11), for different values of the two parameters σ and ζ. For $\sigma = 0$ and $\zeta = 0$ (upper left), the image is identical to the the ground truth x (i.e. the original Shepp-Logan image).

Jezierska11, Jezierska13][3]:

$$y = \zeta z + b \quad \text{with} \quad \begin{cases} z \sim \mathcal{P}\left(\frac{x}{\zeta}\right) \\ b \sim \mathcal{N}\left(0, \sigma^2 \mathrm{Id}\right) \end{cases} \tag{V–11}$$

where z and b are two independent random variables, following respectively a Poisson and a Gaussian distribution. This noise model introduces two numerical parameters:

- $\sigma \geq 0$ is the standard deviation of b; the higher this parameter, the more the model (V–11) behaves like a pure Gaussian noise model.
- $\zeta \geq 0$ is the *gain* of the acquisition process[4]; the higher this parameter, the more Poisson-like is the behavior of the noise in (V–11).

It can be noted that the proposed MPG noise model (V–11) encompasses the classical Gaussian and Poisson noise models: setting $\zeta = 0$ and $\sigma > 0$ corresponds to the Gaussian noise model (V–5), while $\zeta = 1$ and $\sigma = 0$ leads to the Poisson noise model (V–7). Fig. V–2 shows examples of realisations of this noise model on the Shepp-Logan phantom image, for different values of the parameters ζ and σ.

In what follows, we will always assume that the values of the noise parameters σ and ζ are known. However, it is worth noting that estimating these parameters from a given noisy observation y is not trivial. In particular, as noticed in [Jezierska11, Jezierska12b, Jezierska13], the cumulant based approach advised in [Zhang07a] leads to unreliable estimates of the gain parameter ζ. This is due to the fact that this approach makes use of high-order empirical moments (order ≥ 3) evaluated on the noisy signal, which leads to numerical instability. As an alternative, [Jezierska11, Jezierska13] propose an expectation-maximization approach to address this parameter estimation issue, which provides more stable and reliable estimates.

V.2.4 Unbiased risk estimator for the MPG model

Extending the pioneer work in [Luisier11], we derive the *Poisson-Gaussian unbiased risk estimator* (PG-URE) of the MSE for the MPG model (V–11):

$$\text{PG-URE} = \frac{1}{N}\left(\|f(y)\|_2^2 - \|y\|_2^2 + 2\left\langle y, f^{(\zeta)}(y)\right\rangle - 2\sigma^2 \operatorname{Div} f^{(\zeta)}(y) - \zeta \mathbf{1}^T y - \sigma^2\right) \tag{V–12}$$

where the function $f^{(\zeta)}(y)$ is defined component-wise by:

[3]Please note that this type of model may also include a degradation matrix (see for instance [Jezierska12a]), to account for instance for the blurring introduced by the point-spread function of the acquisition system. However, as mentioned in Sec. V.2.1, we focus here on pure denoising problems, for the sake of simplicity.

[4]By convention, when $\zeta = 0$, the MPG model must be understood as $y = x + b$ (i.e. pure Gaussian noise (V–5)). This extension is motivated by the fact that the random variable ζz with $z \sim \mathcal{P}\left(\frac{x}{\zeta}\right)$ converges in law to x (deterministic value) when $\zeta \to 0$.

$$f^{(\zeta)}_k(y) = f_k(y + \zeta e_k) \tag{V–13}$$

The derivation of (V–12) and the proof that $\mathbb{E}_t\text{PG-UREu} = \mathbb{E}_t\text{MSEu}$ are given in appendix V.A, along with the technical conditions required on f for this result to hold. As for the Poisson model, if $f^{(\zeta)}(y)$ is replaced by its first-order Taylor expansion $f^{(\zeta)}(y) \approx f(y) + \zeta Bf(y)$, this leads to the following simplified expression of the PG-URE estimator:

$$\text{PG-URE} = \frac{1}{N}\left(\|f(y) - y\|_2^2 + 2\left(\zeta y + \sigma^2 \mathbf{1}\odot Bf(y)\right) + 2\sigma^2\zeta\left(\mathbf{1}\odot B^2 f(y)\right) + \zeta\langle \mathbf{1}, y\rangle - \sigma^2\right) \tag{V–14}$$

where the image-valued functions $Bf(y)$ and $B^2 f(y)$ are defined as:

$$(Bf)_k(y) = \frac{\partial f_k}{\partial y_k}(y) \qquad (B^2 f)_k(y) = \frac{\partial^2 f_k}{\partial y_k^2}(y) \qquad \text{for all pixels } k \tag{V–15}$$

It should be noted that this simplified expression (V–14) of PG-URE may significantly deviate from (V–12) in the case of large values of the gain parameter ζ, due to the Taylor approximation $f^{(\zeta)}(y) \approx f(y) + \zeta Bf(y)$. However, the numerical results presented in Sec. V.4 show that this deviation has no consequence in the range of gain values encountered in practice.

It can be verified that the expressions (V–12)-(V–14) of the PG-URE estimator are consistent with SURE (V–6) and PURE (V–8)-(V–10) for the special values of the parameters σ and ζ mentioned in Sec. V.2.3. They are also consistent with the unbiased risk estimator derived in [Luisier11] for a simpler mixed Poisson-Gaussian noise model that does not integrate a gain parameter ζ.

Again, evaluation of the terms involving $f^{(\zeta)}$, Bf or $B^2 f$ in (V–12)-(V–14) raises some technical difficulties: in the next section, we propose a non-deterministic method to handle them.

V.3 Stochastic evaluation of the Poisson-Gaussian unbiased risk estimator

V.3.1 Why is a deterministic evaluation of PG-URE impossible?

The expressions (V–12) and (V–14) define unbiased risk estimators of the MSE (V–4) under a mixed Poisson-Gaussian noise model hypothesis (V–11). These expressions do not involve non-accessible entities such as the ground truth x, making their numerical evaluation conceivable in practical settings. However, the terms involving $f^{(\zeta)}$, Bf or $B^2 f$ may not be directly computable, as explained below.

For instance, let us assume that the denoising algorithm f is modeled as a *black-box* process, meaning that we do not make any assumption on how f works internally, and

therefore that the only available "action" with f is to submit an input y and to retrieve an output $f(y)$. Then, due to its definition, a direct evaluation of $f^{(\zeta)}(y)$ would require to run f on N perturbed versions of the input y: $(y + \zeta e_k)$ for $k = 0$ to $N - 1$. As N represents the number of pixels in the input image, such direct evaluation would be computationally irrealistic even with images of reasonable size. The same argument holds for the terms Bf and $B^2 f$, that could be approximated through finite differences: for instance, the first order difference $\frac{1}{\epsilon}(f_k(y + \epsilon e_k) - f_k(y))$ for some small scalar parameter ϵ would provide a good approximation of the kth component of $Bf(y)$, but computing all the components of this term through this scheme would require to evaluate $f(y + \epsilon e_k)$ for $k = 0$ to $N - 1$, which is again irrealistic.

The method developed in the following sections bypasses these problems, thanks to the use of a stochastic scheme to evaluate the Taylor-expanded PG-URE estimator (V–14) in the context of the black-box denoising process mentioned above. One key advantage of this method is that the required number of evaluations of f – which is the most critical factor in terms of computation time – is small and does not depend on N.

V.3.2 Evaluation of the first-order derivative term

We first focus on the term involving the first-order partial derivatives of f in (V–14), namely $\left\langle (\zeta y + \sigma^2 1) \cdot Bf(y) \right\rangle$. The idea of the proposed method, which is a direct extension of the *Monte-Carlo SURE* approach proposed in [Ramani08], is to probe the behavior of f when applied on slightly modified versions of y, which are obtained by adding some well-chosen random perturbations to y.

Let us introduce a few notations: in what follows, $\epsilon > 0$ is a scalar parameter whose value is ideally as small as possible, δ is a random perturbation vector generated according to a probability distribution to be specified, and $\langle u, Bf(y) \rangle$ is the quantity to evaluate. For our particular problem, $u = (\zeta y + \sigma^2 1)$, but the method developed here to evaluate this term does not depend on the actual definition of the image u. In [Ramani08], the method is presented with $u = 1$, which corresponds to $\langle u, Bf(y) \rangle = \mathrm{Div} f(y)$.

First, assuming that f is continuously differentiable, we have:

$$f(y + \epsilon\delta) = f(y) + \epsilon \sum_l \delta_l \frac{\partial f}{\partial y_l}(y) + \epsilon\, r(\epsilon) \tag{V–16}$$

where $r(\epsilon)$ is some remainder that tends to 0 when $\epsilon \to 0$. From this Taylor expansion, it results that:

$$\lim_{\epsilon \to 0} \left\langle \delta, \frac{f(y + \epsilon\delta) - f(y)}{\epsilon} \right\rangle = \sum_{k,l} u_k \delta_k \delta_l \frac{\partial f_k}{\partial y_l}(y) \tag{V–17}$$

where each summation index k and l visits every components of the involved vectors.

Now, let us assume the following properties on the probability distribution of the random perturbation δ:

- the components δ_k of δ are independent,
- each δ_k has an expected value of 0 and a variance equal to 1.

Then, by considering the expected value[5] over the random variable δ on both sides of the equality (V–17), we obtain:

$$E_\delta\left[\lim_{\epsilon\to 0}\left\langle \delta, \frac{f(y+\epsilon\delta)-f(y)}{\epsilon}\right\rangle\right] = \sum_k u_k \frac{\partial f_k}{\partial y_k}(y) \tag{V–18}$$

Finally, up to some technical hypotheses (see [Ramani08] for more details) which are also important to derive the empirical formula (V–20), the expectation and the limit in (V–18) can be switched, leading to the final expression:

$$\lim_{\epsilon\to 0} E_\delta\left[\left\langle \delta, \frac{f(y+\epsilon\delta)-f(y)}{\epsilon}\right\rangle\right] = \langle u, \partial f(y)\rangle \tag{V–19}$$

Equation (V–19) shows that, by taking a parameter ϵ sufficiently small, the inner product $\langle u, \partial f(y)\rangle$ can be approximated by the expected value of the random variable $\frac{1}{\epsilon}\langle \delta, u\rangle\langle f(y+\epsilon\delta)-f(y)\rangle$. Moreover, as observed in [Ramani08], one realization of this random variable is likely to be sufficient to reach a reliable estimate of the expected value in the case of image processing applications (this point will be detailed in Sec. V.3.5). Therefore, we obtain the following empirical estimation formula for $\langle u, \partial f(y)\rangle$:

$$\langle u, \partial f(y)\rangle = \frac{1}{\epsilon}\left\langle \delta, u\right\rangle\left\langle f(y+\epsilon\delta)-f(y)\right\rangle \tag{V–20}$$

V.3.3 Evaluation of the second-order derivative term

A similar method can be proposed to evaluate the term involving the second-order partial derivatives of f in (V–14), namely $\left\langle v, \partial^2 f(y)\right\rangle$ with $v = 1$. Again, the method does not take advantage of the identity $v = 1$, motivating the use of a generic notation v.

We use here notations similar to those introduced in Sec. V.3.2. Then, assuming that f is continuously twice differentiable, a second-order Taylor expansion can be written as:

$$f(y+\epsilon\delta) = f(y) + \epsilon\sum_l \delta_l \frac{\partial f}{\partial y_l}(y) + \frac{\epsilon^2}{2}\sum_{l,m}\delta_l\delta_m \frac{\partial^2 f}{\partial y_l \partial y_m}(y) + \epsilon^2 r(\epsilon) \tag{V–21}$$

and similarly for $f(y-\epsilon\delta)$. By summing these two expansions, we obtain:

[5] In this section, we temporarily assume that y is deterministic. However, to be fully rigorous, what is considered here is not the expectation, but rather the conditional expectation given y. To avoid confusion, the latter is denoted with an additional subscript (E_δ), indicating the remaining source of randomness.

$$\lim_{\sigma \to 0} \left(\delta^T \left(\frac{f(y+\sigma\delta) - 2f(y) + f(y-\sigma\delta)}{\sigma^2} \right) \right) = \sum_{k,l,m} \delta_k \delta_l \delta_m \frac{\partial^2 f_k}{\partial y_l \partial y_m}(y) \tag{V–22}$$

In addition to the hypotheses made in Sec. V.3.2 for δ, we impose here the additional requirement that the third moment κ of the random variables δ_k is non-zero (which implies in particular that the corresponding probability distribution is non-symmetric). Then, the independence of the δ_k and their zero mean ensure that $\mathbb{E}_\delta[\delta_k \delta_l \delta_m]$ is always zero except when $k = l = m$, while $\mathbb{E}_\delta[\delta_k^3] = \kappa \neq 0$. Therefore, taking the expected value in (V–22) and switching it with the limit in the left-hand side leads to the following result:

$$\lim_{\sigma \to 0} \mathbb{E}_\delta \left[\left(\delta^T \left(\frac{f(y+\sigma\delta) - 2f(y) + f(y-\sigma\delta)}{\sigma^2} \right) \right) \right] = \kappa \left(\frac{\partial^2 f}{\partial y^2}(y) \right) \tag{V–23}$$

Finally, assuming that one realization of the random variable δ is sufficient to estimate the expected value in (V–23) (see Sec. V.3.5), we obtain the following empirical estimation formula for $\left(\frac{\partial^2 f}{\partial y^2}(y) \right)$:

$$\left(\frac{\partial^2 f}{\partial y^2}(y) \right) = \frac{1}{\sigma^2 \kappa} \left(\delta^T \left(\frac{f(y+\sigma\delta) - 2f(y) + f(y-\sigma\delta)}{\sigma^2} \right) \right) \tag{V–24}$$

V.3.4 Empirical PG-URE estimator

Using the results obtained in Sec. V.3.2 and V.3.3, we are now able to re-write the PG-URE estimator (V–14) without partial derivatives of f:

$$\begin{aligned} \text{PG-URE} = \frac{1}{N} \|f(y) - y\|_2^2 + \frac{\zeta}{N} \mathbf{1}^T y - \sigma^2 + \frac{2}{N} \left(\tilde{\delta}^T \left(\zeta y + \sigma^2 \mathbf{1} - \frac{f(y+\sigma\tilde{\delta}) - f(y)}{\sigma} \right) \right) \\ + \frac{2\sigma^2 \zeta}{N \sigma^2 \kappa} \left(\delta^T \left(f(y+\sigma\delta) - 2f(y) + f(y-\sigma\delta) \right) \right) \end{aligned} \tag{PG-URE}$$

This expression uses four parameters that are not related to the noise model, but that are introduced for computational purposes:

- $\tilde{\delta}$ is the random perturbation vector used to evaluate the term involving the first-order partial derivatives of f in (V–14). To fulfil the assumptions made in Sec. V.3.2, its components $\tilde{\delta}_k$ must be independent and identically distributed (i.i.d.) random variables with expected value 0 and variance 1. Several probability distributions with these properties can be used to generate the $\tilde{\delta}_k$, and we demonstrate that a binary distribution taking values -1 and 1 with probability $1/2$ each is optimal in the sense that it minimizes the variance of the PG-URE estimator with respect to the random variable $\tilde{\delta}$ (see Sec. V.3.5).
- δ is the random perturbation vector used to evaluate the second-order derivative term. δ is a random vector of i.i.d. components such that[6] $\mathbb{E}(\delta_k) = 0$, $\mathbb{E}(\delta_k^2) = 1$

[6] The constraint on the second moment of δ_k is not compulsory with respect to the methodology developed in Sec. V.3.3, but is rather a normalization convention.

and $E\left(\delta_k^3\right) = \kappa \neq 0$. Again, an optimum with respect to the variance of PG-URE (see Sec. V.3.5 for details) is reached if the δ are generated according to a binary distribution π, defined as:

$$\pi\left(\delta_k = -\sqrt{\frac{q}{p}}\right) = p \qquad \pi\left(\delta_k = \sqrt{\frac{p}{q}}\right) = q \qquad \text{with} \begin{cases} p = \frac{1}{2} + \frac{\kappa}{2}\left(\kappa^2 + 4\right)^{-1/2} \\ q = 1 - p \end{cases} \tag{V–25}$$

where κ is the third moment of the distribution π. The optimal value of κ may not be available in practical settings, and we set it to 1 in our experiments (we motivate this choice in appendix V.B).

- ϑ and ε are the amplitudes of the perturbations introduced to probe the partial derivatives of f. The values of these scalar parameters result from a compromise between 1) the fact that ϑ and ε must be chosen as small as possible to limit the approximation errors in the initial Taylor expansions (V–16) and (V–21), and 2) the finite precision of floating point calculators, which causes significant rounding errors when these parameters are too small. How these values should actually be set is discussed in Sec. V.4.

Finally, the computational complexity of evaluating the PG-URE estimator through the empirical formula (PG-URE) is $4C_f + O(N)$, where C_f is the computational complexity of applying the denoising algorithm f.

V.3.5 Variance of the empirical PG-URE estimator with respect to the random perturbations

In the expression (PG-URE) of the PG-URE estimator, the equality is mathematically proved in terms of expected value over the probability distribution of the two random vectors $\tilde{\vartheta}$ and $\tilde{\delta}$. In practice and similarly to what is proposed in [Ramani08], we evaluate the right-hand side of this expression with a single realization of each of these random variables, as we can assume that such evaluation is close to the expected value. Formally, the underlying assumption is that the standard deviation $\mathrm{Var}_{\tilde{\vartheta},\tilde{\delta}}\left(\text{PG-URE}\right)^{1/2}$ of the estimator (PG-URE) over the probability distribution of $\tilde{\vartheta}$ and $\tilde{\delta}$ is small with respect to its expected value.

Thanks to the independence of $\tilde{\vartheta}$ and $\tilde{\delta}$, the variance of PG-URE can be decomposed as follows:

$$\mathrm{Var}_{\tilde{\vartheta},\tilde{\delta}}\left(\text{PG-URE}\right) = \underbrace{\frac{1}{N^2}\mathrm{Var}_{\tilde{\vartheta}}\left(\sum_{k,l} a_{k,l}\tilde{\vartheta}_k\tilde{\vartheta}_l\right)}_{V_{\tilde{\vartheta}}} + \underbrace{\frac{1}{N^2\kappa^2}\mathrm{Var}_{\tilde{\delta}}\left(\sum_{k,l,m} b_{k,l,m}\tilde{\delta}_k\tilde{\delta}_l\tilde{\delta}_m\right)}_{V_{\tilde{\delta}}} + \tag{V–26}$$

where the notations $a_{k,l}$ and $b_{k,l,m}$ stand for:

$$a_{k,l} = 2\left(\zeta y_k - \sigma^2 \frac{\partial f_k}{\partial y_l}(y)\right) \qquad b_{k,l,m} = 2\sigma^2\zeta \frac{\partial^2 f_k}{\partial y_l \partial y_m}(y) \tag{V–27}$$

Let us focus on the term $V_{\tilde{\sigma}}$ in (V–26), which corresponds to the contribution of the perturbation $\tilde{\sigma}$ to the overall variance of the estimator. In what follows, the pth moment of the probability distribution associated to $\tilde{\sigma}$ will be denoted as $m_p = E\left(\tilde{\sigma}_k^p\right)$. By definition, $V_{\tilde{\sigma}}$ can be written as:

$$N^2 V_{\tilde{\sigma}} = \sum_{k,l,m,n} a_{k,l} a_{m,n} E_{\tilde{\sigma}}\left(\tilde{\sigma}_k \tilde{\sigma}_l \tilde{\sigma}_m \tilde{\sigma}_n\right) - \sum_{k,l} a_{k,k} a_{l,l} \tag{V–28}$$

Thanks to the independence of the $\tilde{\sigma}_k$ and the property $m_1 = 0$ introduced in Sec. V.3.2, the expected value $E_{\tilde{\sigma}}\left(\tilde{\sigma}_k \tilde{\sigma}_l \tilde{\sigma}_m \tilde{\sigma}_n\right)$ is 0 as soon as at least one of the indices k, l, m or n is different from all the others. Then, the remaining terms and the property $m_2 = 1$ lead to:

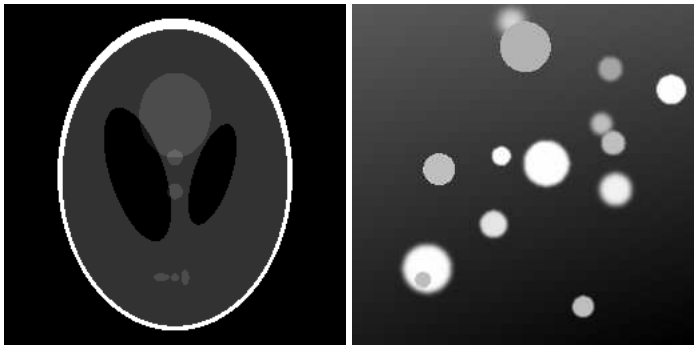
$$N^2 V_{\tilde{\sigma}} = m_4 \sum_k a_{k,k}^2 - \sum_{\substack{k,l \\ k \neq l}} a_{k,k} a_{l,l} - a_{k,l}^2 - a_{k,l} a_{l,k} - \sum_{k,l} a_{k,k} a_{l,l} \tag{V–29}$$

Up to additional simplifications and factorization, (V–29) leads to the following final expression of the term $V_{\tilde{\sigma}}$:

$$V_{\tilde{\sigma}} = \frac{m_4 - 1}{N^2} \sum_k a_{k,k}^2 - \frac{1}{2N^2} \sum_{k \neq l} \left(a_{k,l} - a_{l,k}\right)^2 \tag{V–30}$$

This expression (V–30) calls for two remarks:

1. As $V_{\tilde{\sigma}}$ should be made as small as possible to limit the variance of the PG-URE estimator, the probability distribution used to generate the $\tilde{\sigma}_k$ should be chosen so that m_4 is as small as possible. Yet, with the requirements $m_1 = 0$ and $m_2 = 1$, it can be shown that $m_4 \geq 1$ (see for instance [Akhiezer65]); the optimal value $m_4 = 1$ is obtained with a symmetric binary distribution taking values -1 and 1 with probability $1/2$ each. This justifies our proposition to use this probability distribution in Sec. V.3.4.
2. The second summation group (the one with two summation indices k and l) involves $N(N-1)$ terms (all the pairs $k, l = 1$ to N, except those with $k = l$), but most of the $(a_{k,l} - a_{l,k})^2$ terms are likely to be 0. Indeed, $a_{k,l}$ is proportional to $\frac{\partial f_k}{\partial y_l}(y)$, and the value of this partial derivative is likely to be insignificant when the indices k and l refer to pixels that are distant from each others: in particular, this is certainly true if f is a local denoising method. Furthermore, if we assume that the number of input pixels y_l that have a significant influence on the kth output pixel $f_k(y)$ is constant, then the number of non-zero $(a_{k,l} - a_{l,k})^2$ terms is proportional to N, making $V_{\tilde{\sigma}}$ proportional to $\frac{1}{N}$. As N is quite large in the case of images, $V_{\tilde{\sigma}}$, which represents the variance of the PG-URE estimator with respect to the perturbation $\tilde{\sigma}$, is likely to be very small: this justifies the assumption made in Sec. V.3.2 that only



Shepp-Logan *Disks 3*

Figure V–3: Test images used for the simulations (256×256 pixels, intensity range normalized to the interval $[0, 1]$).

one realization of this perturbation is sufficient to estimate the first-order partial derivatives of f involved in the computation of PG-URE.

The term $V_{\tilde{\delta}}$ corresponding to the contribution of the perturbation $\tilde{\delta}$ in (V–26) can also be expressed as a function of the coefficients $b_{k,l,m}$ and the moments $m_p = E\left(\tilde{\delta}_k^p\right)$, similarly to (V–30) (see appendix V.B). The obtained expression leads to conclusions similar to those drawn for $V_{\hat{\delta}}$, namely that $V_{\tilde{\delta}}$ is proportional to $\frac{1}{N}$ for reasonable denoising functions f, and that $V_{\tilde{\delta}}$ is minimal when $\tilde{\delta}$ is generated according to the binary probability distribution (V–25), for a particular value κ' of the parameter κ. Unfortunately, the optimal value κ' depends on the coefficients $b_{k,l,m}$ and consequently on the partial derivatives of f, whose values are by definition not available. Still, we noticed that the arbitrary setting $\kappa = 1$ leads to stable results (see Sec. V.4 and appendix V.B).

V.4 Numerical validation and application

V.4.1 Simulation goals and process

The expression (PG-URE) defines an unbiased risk estimator of the MSE under the mixed Poisson-Gaussian noise hypothesis (V–11). Sections V.3.4 and V.3.5 describe how the random perturbation $\hat{\delta}$ and $\tilde{\delta}$ involved in this PG-URE estimator are generated. However, we have not discussed yet on the values that should be attributed to the scalar parameters θ and τ. We propose to determine how these values should be set through numerical simulations; we will also make the most of these simulations to verify the empirical equality PG-URE $=$ MSE.

For the numerical simulations, we selected two phantom images (see Fig. V–3):

1. the well-known Shepp-Logan phantom, sized 256×256 pixels;

2. *Disks 3*, a synthetic test image also sized 256×256 pixels, representing several disks with random gray levels, sizes and boundary sharpness, over a non-uniform dark background.

All these images were normalized so that they are valued between 0 and 1. From each noise-free image x, we generated four noisy images y following the MPG model (V–11), with the following noise parameters:

- $\sigma = 10^{-1.5}$, $\zeta = 10^{-2}$ (this case is denoted as "low noise" in the following results);
- $\sigma = 10^{-1}$, $\zeta = 10^{-2}$ (denoted as "mostly Gaussian");
- $\sigma = 10^{-1.5}$, $\zeta = 10^{-1}$ (denoted as "mostly Poisson");
- $\sigma = 10^{-1}$, $\zeta = 10^{-1}$ (denoted as "high noise").

We selected six classical denoising algorithms, all dependent of a scalar parameter θ:

- Wavelet soft-thresholding [Donoho95a]:

$$f_\theta^{WSo}(y) = W^{-1} \cdot T_\theta(W \cdot y) \tag{V–31}$$

  where W is a 2D un-decimated wavelet transform (we used the Daubechies-4 orthogonal wavelet with 4 levels of decomposition), and T_θ is the component-wise soft-thresholding function, mapping each input wavelet coefficient w to $\mathrm{sign}(w) \cdot \max(0, |w| - \theta)$.
- TV minimization [Rudin92]:

$$f_\theta^{TV}(y) = \arg\min_x \|x\|_{TV} \text{ subject to } \|x - y\|_2 \leq \theta \tag{V–32}$$

  This constrained formulation is equivalent to the original unconstrained one (V–1) (see Chap. II), and we chose to use the former for practical reasons.
- Non-local means [Buades05]: $f_\theta^{NLM}(y)$ is defined component-wise as in (V–2). We used the similarity measure originally proposed in [Buades05], i.e. (V–3), with centered square patches of size 5×5 pixels.
- We derived three "stabilized" versions of these three denoising algorithms, for which we first applied a variance stabilization transform on the input image, to make the variance of the noisy pixel y_k independent of the ground truth value x_k, and therefore uniform over the whole image (see [Starck98, Zhang08]). Formally:

$$f_\theta^{S\text{-}WSo}(y) = S^{-1} \circ f_\theta^{WSo} \circ S(y) \tag{V–33}$$

  and similarly for $f_\theta^{S\text{-}TV}$ and $f_\theta^{S\text{-}NLM}$. The variance stabilization transform $S(y)$ is defined as:

$$S_k(y) = \frac{2}{\zeta}\,\mathrm{sign}(t)\sqrt{|t|} \quad \text{with} \quad t = \zeta y_k + \frac{3}{8}\zeta^2 + \sigma^2 \tag{V–34}$$

  In [Starck98], it is shown that, under the MPG hypothesis (V–11), $S_k(y)$ has a variance approximately equal to 1 (except for very low values of x_k, which correspond to an extremely low-light regime).

Finally, for each pair of tested noisy image and algorithm, and for several values of the corresponding parameter θ, we computed the denoised estimate $f_\theta(y)$ and the MSE (as we are using phantom test images, the ground truth is available), and we evaluated the estimator (PG-URE) with different values of the amplitude parameters ϑ and $\Box$. All simulations were performed with Matlab®, using double precision floating point arithmetic. The influence of ϑ and $\Box$ on the PG-URE estimator is studied in the next sections.

V.4.2 Influence of the amplitude parameters ϑ and $\Box$

To study how the parameters ϑ and $\Box$ affect the estimator (PG-URE), we decompose the latter into three terms, as $\text{PG-URE} = T_0 + T_1(\vartheta) + T_2(\Box)$, where:

$$
\begin{aligned}
T_0 &= \frac{1}{N}\|f(y) - y\|_2^2 - \frac{\zeta}{N}\langle 1, |y| - \sigma^2\rangle \\
T_1(\vartheta) &= \frac{2}{N\vartheta}\left\langle \sigma^2 \zeta y + \sigma^2 \left(f(y) + \vartheta f'(y)\right), \vartheta f(y) \right\rangle \\
T_2(\Box) &= -\frac{2\sigma^2\zeta}{N\Box^2}\left\langle \vartheta f(y) + \Box\vartheta - 2f(y) + f(y) + \Box\vartheta \right\rangle
\end{aligned}
\tag{V–35}
$$

In this decomposition, T_0 includes the contributions to PG-URE that do not depend on ϑ and $\Box$, while $T_1(\vartheta)$ and $T_2(\Box)$ represent respectively the contributions due to the first and second order partial derivatives of f. Figs. V–4 and V–5 present two examples of the evolution of T_0, $T_1(\vartheta)$ and $T_2(\Box)$ with respect to the denoising parameter θ, for different values of ϑ and $\Box$.

V.4.2.1 Parameter ϑ

Both graphs in Figs. V–4 and V–5 show that, although $T_1(\vartheta = 0.1)$ and $T_1(\vartheta = 1)$ have singular behaviors (the latter curve does not fit in the displayed range of the graph in Fig. V–5), $T_1(\vartheta)$ seems to converge to an asymptotic curve for smaller values of ϑ: indeed, for $\vartheta \leq 10^{-3}$, we can assume that $T_1(\vartheta)$ becomes almost independent of ϑ with a value close to the ideal one that would be obtained for $\vartheta \to 0$.

To confirm this assumption, we measured the term $T_1(\vartheta)$ for ϑ varying between 10^{-7} and 1 with samples geometrically spaced by a factor $10^{0.02}$ (i.e. $\vartheta = 10^{-7}, 10^{-6.98}, 10^{-6.96}, \ldots$), and for all the combinations of denoising algorithms, test images and noise parameter mentioned in Sec. V.4.1, with the denoising parameter θ set such that the MSE is minimal; the corresponding minimal value of the MSE is denoted as $\text{MSE}^\star$. We then measured the variability among the $T_1(\vartheta)$ values through the indicator ΔT_1, defined as:

$$
\Delta T_1 = \frac{1}{\text{MSE}^\star}\,\text{StdDev}_{\vartheta \in [\vartheta_{min}, \vartheta_{max}]}\, T_1(\vartheta)
\tag{V–36}
$$

where $\text{StdDev}_{\vartheta \in [\vartheta_{min}, \vartheta_{max}]}\, T_1(\vartheta)$ measures the empirical standard deviation of $T_1(\vartheta)$ for ϑ varying within a sub-range $[\vartheta_{min}, \vartheta_{max}]$ of the probed interval $[10^{-7}, 1]$. The values

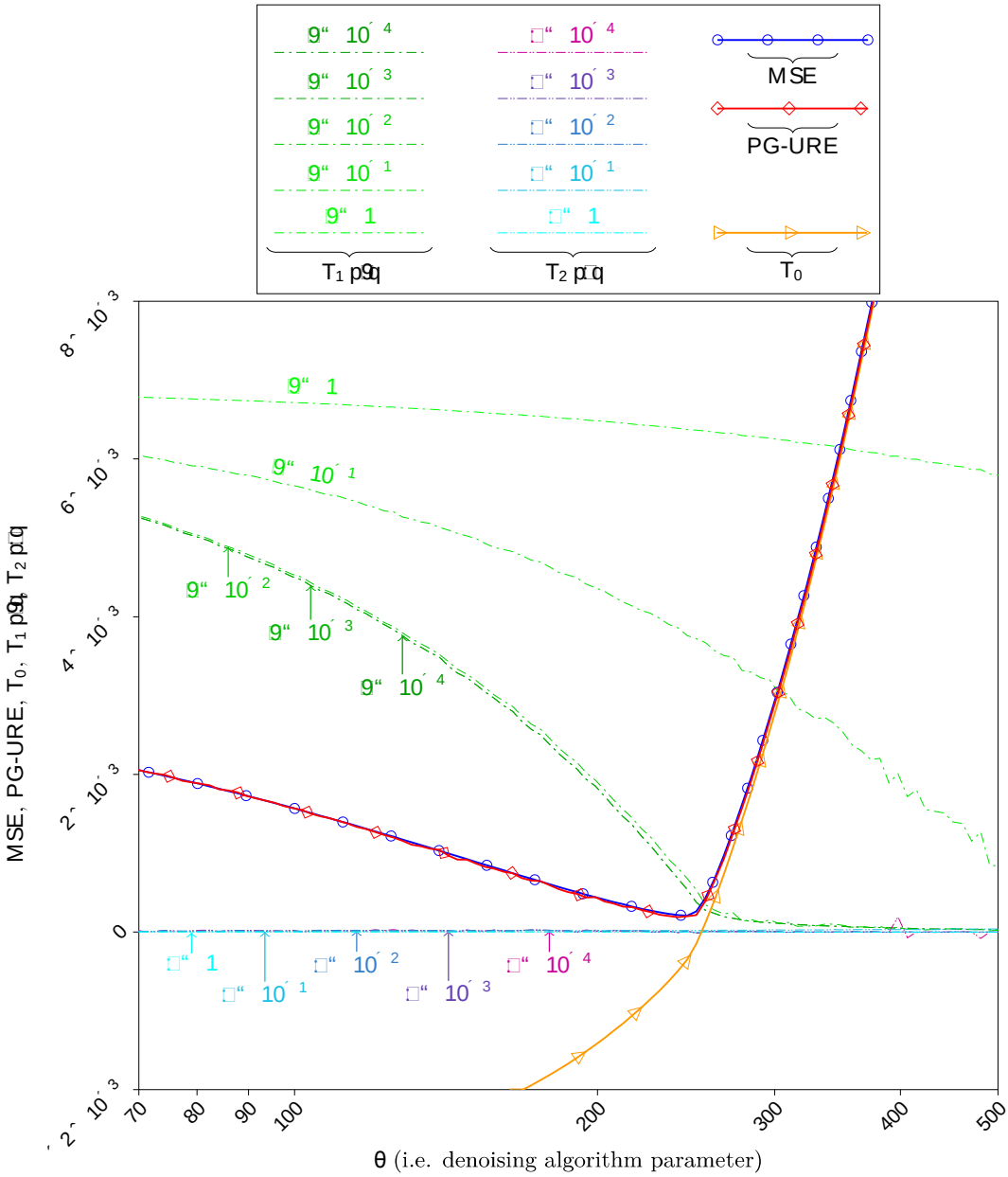


Figure V–4: Denoising of *Disks 3* ` "low noise", using the $f_\theta^{S\text{-}TV}$ algorithm (TV minimization together with variance stabilization transform). MSE and PG-URE values are plotted as functions of the denoising parameter θ, together with the individual PG-URE terms T_0, $T_1(\lambda)$ and $T_2(\mu)$ for several values of the parameters λ and μ. Only the PG-URE curve corresponding to $\lambda = 10^{-4}$ and $\mu = 10^{-2}$ is plotted.

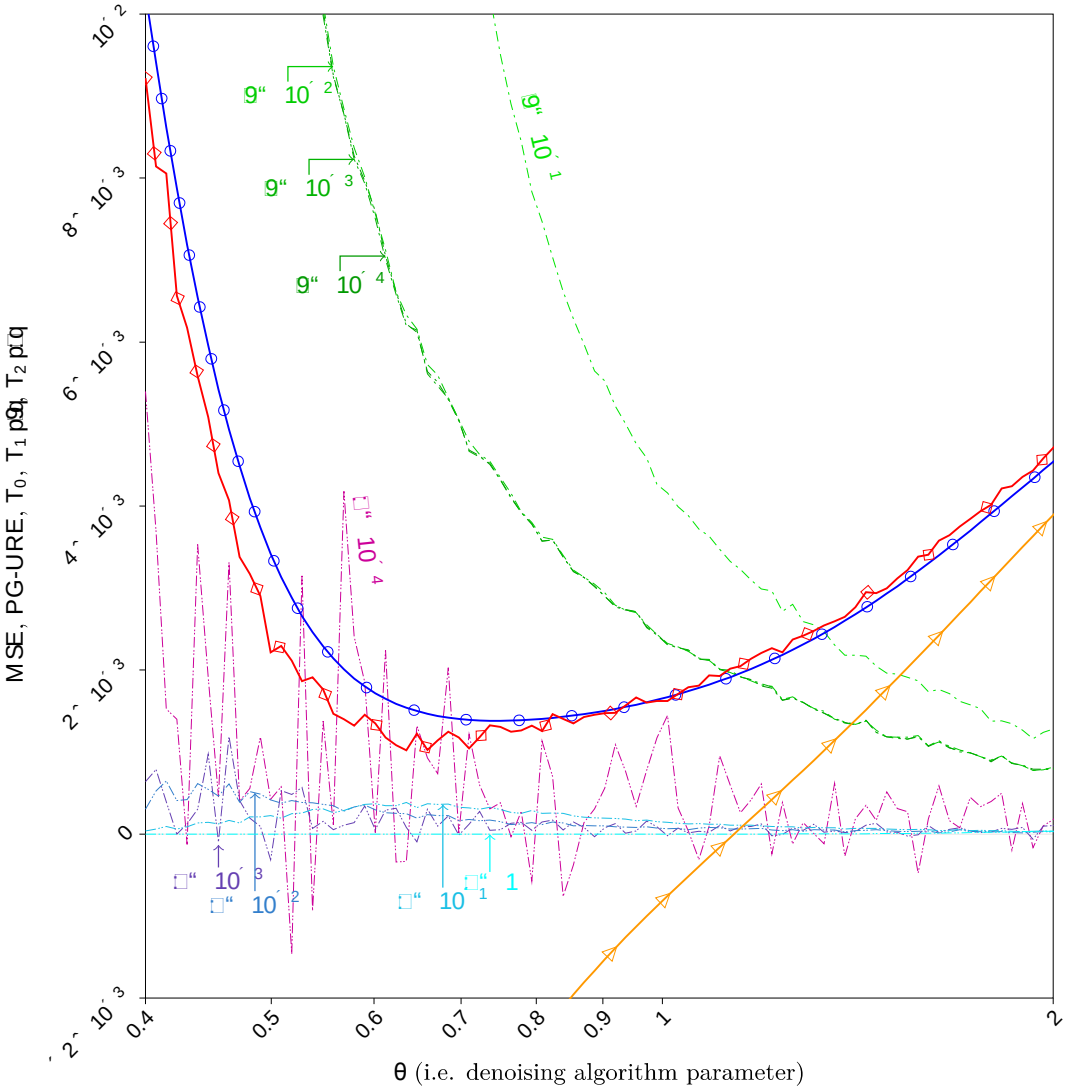


Figure V–5: Denoising of *Disks 3* ` "mostly Poisson" noise, using the $f_\theta^{\text{S-NLM}}$ algorithm (non-local means together with variance stabilization transform). Same representation and legend as in Fig. V–4.

	Shepp-Logan				*Disks 3*			
	Low noise	Mostly Gaussian	Mostly Poisson	High noise	Low noise	Mostly Gaussian	Mostly Poisson	High noise
f_θ^{WSo}	0.06%	0.07%	0.07%	0.06%	0.07%	0.09%	0.11%	0.11%
f_θ^{TV}	0.11%	0.04%	0.03%	0.11%	0.97%	0.34%	0.64%	0.53%
f_θ^{NLM}	0.01%	0.01%	0.01%	0.00%	0.02%	0.01%	0.00%	0.00%
$f_\theta^{\mathrm{S\text{-}WSo}}$	0.07%	0.10%	0.59%	0.30%	0.23%	0.26%	1.40%	0.52%
$f_\theta^{\mathrm{S\text{-}TV}}$	0.06%	0.10%	0.83%	0.46%	0.13%	0.09%	0.46%	0.47%
$f_\theta^{\mathrm{S\text{-}NLM}}$	0.05%	0.01%	0.66%	0.17%	0.02%	0.02%	0.59%	0.46%

Figure V–6: ΔT_1 obtained for $\lambda = 10^{-6}, 10^{-5.98}, \ldots, 10^{-3.02}, 10^{-3}$ (151 samples), given as percentages. The only value greater than 1% is highlighted in yellow.

obtained for ΔT_1 with $(\lambda_{min}, \lambda_{max}) = (10^{-6}, 10^{-3})$ are presented in Fig. V–6. Theses results show that the variability of $T_1(\lambda)$ induced by the choice of λ is very small compared to the MSE (the quantity to estimate): indeed, whatever the value chosen for λ in the range $(10^{-6}, 10^{-3})$, the value obtained for $T_1(\lambda)$ (and therefore for PG-URE) is constant. We therefore used in practice $\lambda = 10^{-4}$ in what follows.

It is important to note that this value depends on the normalization used for the intensity of the processed images: here, our images are valued between 0 and 1, but different normalizations would lead to different values. For instance, in the case of intensity normalized between 0 and 255, a correct setting is $\lambda = 255 \times 10^{-4}$. The floating point precision used for the computations may also have an influence, although this aspect is less critical for $T_1(\lambda)$ than for the second order term, as discussed in the next paragraph.

V.4.2.2 Parameter τ

We proceeded similarly to determine a satisfactory value for τ: we measured the term $T_2(\tau)$ for $\tau = 10^{-4}, 10^{-3.99}, \ldots, 10^{-0.02}, 10^{-0.01}, 1$, and for all the combinations of denoising algorithms, test images and noise parameters, with the denoising parameter θ set such that the MSE is minimal. The values obtained for $T_2(\tau)$ as functions of τ in six of these configurations are presented in Fig. V–7.

Contrary to what happens with the first order term, we did not observe a clear convergence of $T_2(\tau)$ to an asymptotic value when $\tau \to 0$: the curves $T_2(\tau)$ showed chaotic behaviors, with large and unpredicable oscillations when $\tau \geq 10^{-3}$. We interpret these behaviors as the consequence of rounding errors introduced by floating point operations involved when computing the term $T_2(\tau)$. More precisely, the latter involves a second-

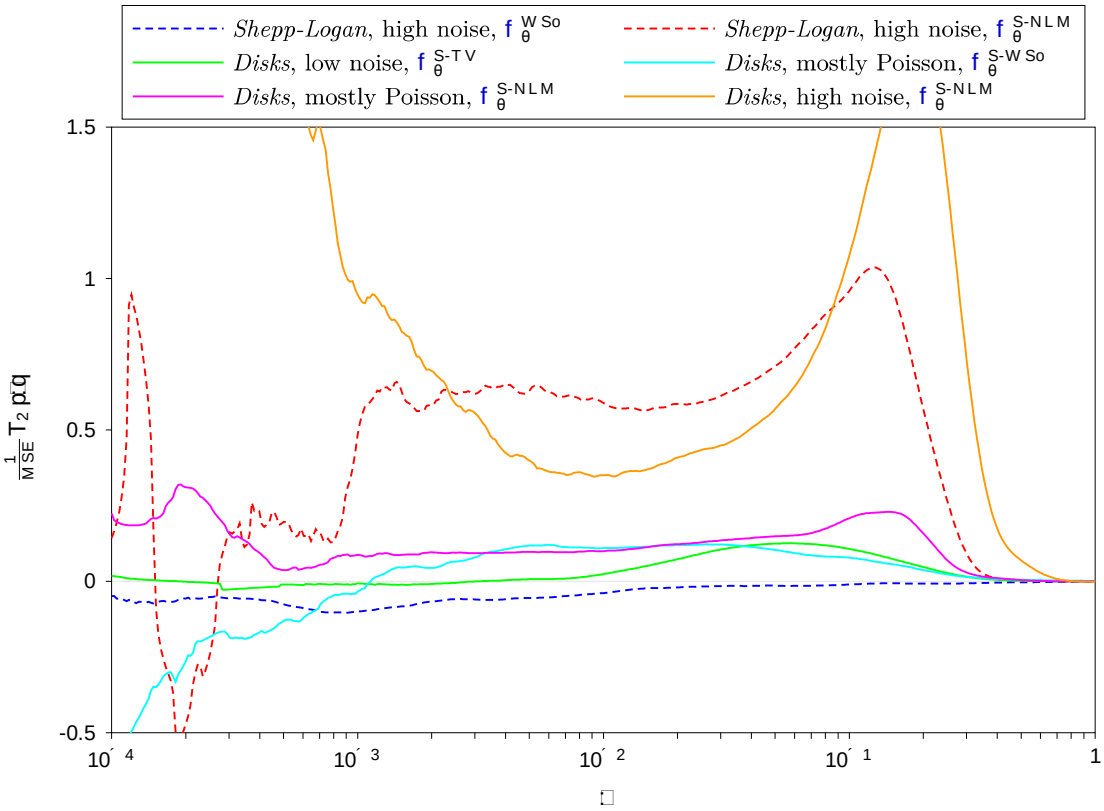


Figure V–7: Term $T_2(\lambda)$ as a function of λ for six of the tested combinations of test image, noise level and denoising algorithm (with in each case the parameter θ set such that the MSE is minimal). Each curve $T_2(\lambda)$ was normalized by the actual MSE measured for the corresponding tested combination.

order finite difference $f(y+\delta) - 2f(y) + f(y-\delta)$ whose order of magnitude might be significantly smaller than the ones of the individual terms $f(y+\delta)$ and $f(y)$; then, due to cancellation events (see [Goldberg91]), the error made when performing this operation is likely to be significant. A solution to avoid this problem could have been to increase the parameter λ, but in this case the assumption that $T_2(\lambda)$ is close to its theoretical limit obtained for $\lambda \to 0$ becomes erroneous: it appears that the trade-off between the need for λ to be small enough for the mathematical analysis derived in Sec. V.3 to be valid, and the need for λ to be large enough to avoid numerical rounding errors is much more tight for λ than for θ.

However, the curves on Fig. V–7 show that there seems to exist a narrow window around $\lambda = 10^{-2}$ where both requirements hold, leading to functions $T_2(\lambda)$ approximately constant. To validate this hypothesis, we introduce an indicator ΔT_2 as follows:

$$\Delta T_2 = \frac{1}{MSE} \cdot \mathrm{StdDev}_{\lambda \in [\lambda_{min}, \lambda_{max}]} T_2(\lambda) \tag{V–37}$$

where the empirical standard deviation is computed for λ varying in a sub-range of the

	Shepp-Logan				*Disks 3*			
	Low noise	Mostly Gaussian	Mostly Poisson	High noise	Low noise	Mostly Gaussian	Mostly Poisson	High noise
f_θ^{WSo}	0.2%	0.6%	0.1%	3.2%	0.6%	1.3%	0.3%	3.2%
f_θ^{TV}	0.4%	0.4%	0.0%	0.8%	1.3%	0.9%	0.4%	3.3%
f_θ^{NLM}	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
$f_\theta^{\text{S-WSo}}$	0.8%	0.6%	4.2%	4.6%	0.3%	1.5%	0.9%	4.5%
$f_\theta^{\text{S-TV}}$	3.2%	1.1%	12.2%	12.5%	5.5%	11.1%	8.6%	12.1%
$f_\theta^{\text{S-NLM}}$	2.6%	0.1%	2.1%	6.1%	0.5%	0.1%	2.3%	4.4%

Figure V–8: ΔT_2 given as percentages obtained for $5\times 10^{-3} \leq \sigma \leq 2\times 10^{-2}$ with geometric increments of $10^{0.01}$ (61 samples). Yellow cells contain values greater than 1%, while orange cells contain values greater than 10%.

probed interval. The values obtained for ΔT_2 with $[\sigma_{min}, \sigma_{max}] = [5\times 10^{-3}, 2\times 10^{-2}]$ are presented in Fig. V–8. These values show that the variability of $T_2(\sigma)$ (and therefore the variability of PG-URE) induced by the choice of σ represents less than 1% of the MSE to be estimated in more than half of the tested combinations. This variability seems to be mainly determined by the denoising algorithm: indeed, the value of $T_2(\sigma)$ is very stable in the case of f_θ^{NLM}, and on the contrary extremely dependent on σ in the case of $f_\theta^{\text{S-TV}}$. However, as other choices of intervals $[\sigma_{min}, \sigma_{max}]$ lead to poorer results for ΔT_2, we propose $\sigma = 10^{-2}$ as a reasonable compromise value for this parameter. Results presented in the next section show that this choice leads to an estimator PG-URE that can be successfully used to adaptively set the value of the parameter θ for each denoising algorithm.

Similarly to the case of the first order term, the setting for σ depends on the normalization used for the intensity of the processed images, and also on the floating point precision used for the computations.

V.4.3 Optimization of the denoising parameters θ driven by PG-URE

Finally, to evaluate the performance of the PG-URE estimator when used to optimize the parameter θ of the denoising algorithms, we performed the following simulations: for each combination of tested image, set of noise parameters, and denoising algorithm f_θ, we ran the denoising algorithm for several values of θ, and computed the resulting MSE and PG-URE values[7]; we then retained in each case the parameters θ^*_{MSE} and $\theta^*_{\text{PG-URE}}$ that minimize respectively the MSE and the PG-URE. The corresponding image $\hat{x}_{\text{PG-URE}} = f_{\theta^*_{\text{PG-URE}}}(y)$ represents the denoising result obtained by tuning the denoising parameter

[7] We selected $\gamma = 10^{-4}$ and $\sigma = 10^{-2}$ to evaluate PG-URE, as advised in Sec. V.4.2.

	Shepp-Logan				*Disks 3*			
	Low noise	Mostly Gaussian	Mostly Poisson	High noise	Low noise	Mostly Gaussian	Mostly Poisson	High noise
f_θ^{WSo}	0.4% 31 dB	1.4% 27 dB	0.9% 25 dB	0.1% 24 dB	0.3% 33 dB	0.1% 30 dB	3.5% 27 dB	0.0% 27 dB
f_θ^{TV}	1.6% 36 dB	12.3% 32 dB	1.3% 28 dB	2.8% 27 dB	0.0% 38 dB	15.4% 35 dB	0.0% 31 dB	0.0% 30 dB
f_θ^{NLM}	0.0% 35 dB	0.2% 32 dB	0.3% 27 dB	0.3% 26 dB	0.7% 39 dB	0.2% 36 dB	0.1% 29 dB	1.5% 29 dB
$f_\theta^{\text{S-WSo}}$	1.2% 30 dB	0.8% 27 dB	2.0% 23 dB	0.5% 22 dB	0.1% 33 dB	0.5% 30 dB	4.9% 26 dB	2.2% 24 dB
$f_\theta^{\text{S-TV}}$	0.0% 35 dB	2.8% 31 dB	0.7% 27 dB	6.5% 24 dB	4.1% 37 dB	11.3% 34 dB	15.5% 28 dB	56.1% 27 dB
$f_\theta^{\text{S-NLM}}$	0.4% 36 dB	1.4% 32 dB	10.7% 27 dB	0.9% 25 dB	0.5% 39 dB	0.5% 36 dB	8.4% 29 dB	6.3% 28 dB

Figure V–9: Δ Estim (V–38) given as percentages. Yellow cells contain values greater than than 5%, while orange cells contain values greater that 20%. PSNR values (in dB) obtained for $\hat{x}_{\text{MSE}}$ are reported, as a measure of the "best" denoising quality achievable using an oracle-based parametrization.

such that the PG-URE estimator is minimal – hence without using the ground truth – while $\hat{x}_{\text{MSE}} = f_{\theta^*_{\text{MSE}}}(y)$ corresponds to the denoised image obtained by selecting the best denoising parameter according to the MSE, following an oracle-based approach (hence not applicable for real denoising problems). We finally compared the differences between the two denoised images by measuring the following indicator:

$$\Delta\,\text{Estim} = \frac{\|\hat{x}_{\text{PG-URE}} - \hat{x}_{\text{MSE}}\|_2^2}{\|x - \hat{x}_{\text{MSE}}\|_2^2} \tag{V–38}$$

Here, Δ Estim relates the l_2 distance between the two denoised images to the l_2 distance between the ground truth x and the "best" denoised image, i.e. the one obtained by following the oracle based approach. The values measured for Δ Estim are presented in Fig. V–9, along with the peak signal-to-noise measure reached with $\hat{x}_{\text{MSE}}$ – defined as $\text{PSNR} = -10\log_{10}\frac{1}{N}\|x - \hat{x}_{\text{MSE}}\|_2^2$ – which assesses the "best" denoising quality achievable following the oracle-based parameter estimation approach. Four examples of pairs of denoised images $\hat{x}_{\text{MSE}}$ and $\hat{x}_{\text{PG-URE}}$ are also presented in Fig. V–10.

Although the best performing denoising parameters θ^*_{MSE} and $\theta^*_{\text{PG-URE}}$ selected by the MSE and the PG-URE are not always the same, Fig. V–9 shows that the distance between the corresponding denoised images is, in most cases, very small compared to the distance between the oracle-denoised image and the ground truth: the indicator Δ Estim is indeed smaller than 5% in 39 of the 48 tested configurations, which corresponds to differences

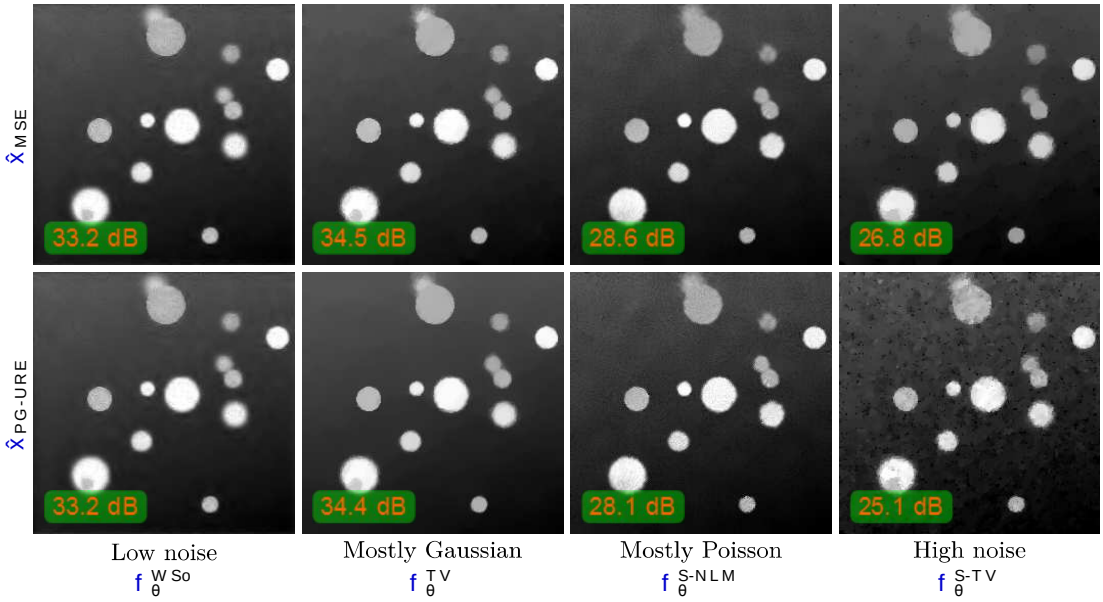


Figure V–10: Comparison between the denoised images $\hat{x}_{MSE}$ and $\hat{x}_{PG\text{-}URE}$ obtained for the original image *Disks 3*, with four different noise levels and denoising methods. PSNR values are also reported in the bottom left corner of each image.

between the denoised images that are visually unnoticeable. The visual similarity between the denoised images $\hat{x}_{MSE}$ and $\hat{x}_{PG\text{-}URE}$ obtained with these parameters is illustrated on four examples in Fig. V–10: in each of the three left-most columns – which correspond to situations with Δ Estim $\leq 20\%$ (either white or yellow cells in Fig. V–9) – the images $\hat{x}_{MSE}$ and $\hat{x}_{PG\text{-}URE}$ are indeed very similar. For all these cases, the PG-URE estimator therefore performed very well as a surrogate for the MSE value, while still being actually computable in real denoising problems, for which a ground truth is not available.

However, for the *Disks 3* image in the "high-noise" configuration and with the $f_\theta^{S\text{-}TV}$ algorithm (orange cell in Fig. V–9 and right-most column in Fig. V–10), we can clearly observe that the denoising task failed and did not return a reliable image. This is due to an inappropriate selection of the parameter θ value, itself derived from an erroneous estimation of the MSE with the empirical PG-URE estimate. Two scenarios can explain this erroneous estimation: drawing of a "bad" sample of the parameter $\square$ (Fig. V–8 shows that this configuration is one of the least favorable with respect to the indicator ΔT_2), and/or a realization of one of the random variables $\tilde{\sigma}^{\theta}$ or $\tilde{\delta}$ that makes the PG-URE estimator significantly deviate from its expected value. These scenarios correspond to the inherent risk taken with any stochastic Monte-Carlo type of method. One way to reduce this risk would be to draw several realizations of $\tilde{\sigma}^{\theta}$ or $\tilde{\delta}$ and average the corresponding values of $T_1(\theta)$ and $T_2(\theta)$, at the cost however of a higher computation time. Post-processing could also be proposed to detect failure of the denoising, or multiple runs could be performed to gauge the range of values obtained for the parameter being optimized, with detection of outliers.

V.5 Conclusion

In this chapter, we presented a new unbiased risk estimator (PG-URE) for general image denoising applications, in a context where the processed images are degraded following a mixed Poisson-Gaussian noise model. This model unifies the widely used Gaussian and Poisson noise models and is relevant to describe the degradations observed in bioimaging applications, in particular low-light fluorescence microscopy. We showed that the PG-URE estimator can be used as a surrogate for the usual mean squared error measure, although its evaluation does not require any knowledge about the noise-free version (i.e. the ground truth) of the image to denoise. We also proposed a practical formula (PG-URE) to evaluate this estimator when no specific knowledge on the partial derivatives of the denoising function f is available, making this framework usable "out of the box" with almost any available denoising algorithm.

We validated our approach through numerical simulations involving standard denoising algorithms and phantom test images. Relying on theses simulations, we discussed how to set the numerical parameters involved in PG-URE. We compared the results obtained when tuning the parameters θ of these standard denoising algorithms by minimizing the PG-URE estimator and the mean squared error, and showed that these two approaches lead to similar denoised images in most of the tested scenarios. This demonstrates the interest of the PG-URE estimator for practical applications, as MSE driven optimization is not applicable for real denoising problems.

Finally, although not carried out yet, we believe that this type of tool can benefit to the study and the improvement of the CS-based denoising method [Marim09] applied to low-light fluorescence microscopy images.

V.A Derivation of the PG-URE estimator

This appendix describes how the first definition (V–12) of the PG-URE estimator is obtained, and proves the equality $\mathbb{E}\{\text{PG-URE}\} = \mathbb{E}\{\text{MSE}\}$. This result could be derived quite directly from the work in [Luisier11], but we propose here a more intrinsic proof, relying on the two basic properties of the Gaussian and Poisson distributions that are mentioned below. Proofs of these lemmas can be found respectively in [Stein81] and [Peng75, Tsui82].

Lemma V–1 (Stein's lemma) *Let* $y = x + b$ *where* $x \in \mathbb{R}^N$ *is deterministic and* $b \sim \mathcal{N}(0, \sigma^2 \mathrm{Id})$. *Let* $\varphi : \mathbb{R}^N \to \mathbb{R}^N$ *be a weakly differentiable function such that* $\mathbb{E}\left\{\frac{\partial \varphi_k}{\partial y_k}(y)\right\} < +\infty$ *for all* k. *Then:*

$$\mathbb{E}\left\{\langle b, \varphi(y)\rangle\right\} = \sigma^2 \mathbb{E}\left\{\mathrm{Div}\,\varphi(y)\right\}$$

Lemma V–2 *Let* $z \in \mathbb{R}^N$ *such that* $z \sim \mathcal{P}(x)$ *(i.e. the components* z_k *are independent random variables following Poisson distributions of parameters* x_k*). Let* $\psi : \mathbb{R}^N \to \mathbb{R}^N$ *such that* $\mathbb{E}\{|\psi_k(z)|\} < +\infty$ *for all* k*. Then:*

$$\mathbb{E}\left\{\langle x, \psi(z)\rangle\right\} = \mathbb{E}\left\{\langle z, \psi^{(1)}(z)\rangle\right\}$$

Thanks to these results, we can state the following theorem:

Theorem V–3 *Let* $y = \zeta z + b$ *where* $b \sim \mathcal{N}(0, \sigma^2 \mathrm{Id})$ *and* $z \sim \mathcal{P}\left(\frac{x}{\zeta}\right)$ *(*b *and* z *independent). Let* $\varphi : \mathbb{R}^N \to \mathbb{R}^N$ *a weakly differentiable function such that* $\mathbb{E}\{|\varphi_k(y)|\} < +\infty$ *and* $\mathbb{E}\left\{\left|\frac{\partial \varphi_k}{\partial y_k}(y + \zeta e_k)\right|\right\} < +\infty$ *for all* k*. Then:*

$$\mathbb{E}\left\{\langle x, \varphi(y)\rangle\right\} = \mathbb{E}\left\{\langle y, \varphi^{(1)}(\zeta y)\rangle + \sigma^2 \operatorname{Div}\varphi^{(1)}(\zeta y)\right\}$$

Proof. We introduce the family of functions $\psi_b : \mathbb{R}^N \to \mathbb{R}^N$, defined by $\psi_b(z) = \varphi(\zeta z + b)$. Then:

$$\begin{aligned}
\mathbb{E}\left\{\langle x, \varphi(y)\rangle\right\} &= \mathbb{E}\left\{\zeta \mathbb{E}_z\left\{\left\langle \frac{x}{\zeta}, \psi_b(z)\right\rangle\right\}\right\} \\
&= \mathbb{E}\left\{\zeta \mathbb{E}_z\left\{\left\langle z, \psi_b^{(1)}(z)\right\rangle\right\}\right\} && \text{(cf. Lemma V–2)} \\
&= \mathbb{E}\left\{\left\langle y - b, \varphi^{(1)}(\zeta z + b)\right\rangle\right\} \\
&= \mathbb{E}\left\{\left\langle y, \varphi^{(1)}(y)\right\rangle - \mathbb{E}_b\left\{\left\langle b, \varphi^{(1)}(y)\right\rangle\right\}\right\} \\
&= \mathbb{E}\left\{\left\langle y, \varphi^{(1)}(y)\right\rangle - \sigma^2 \mathbb{E}_b\left\{\operatorname{Div}\varphi^{(1)}(y)\right\}\right\} && \text{(cf. Lemma V–1)} \\
&= \mathbb{E}\left\{\left\langle y, \varphi^{(1)}(y)\right\rangle + \sigma^2 \operatorname{Div}\varphi^{(1)}(y)\right\}
\end{aligned}$$

□

Finally, from the definition of the MSE (V–4), it can be noticed that:

$$\mathbb{E}\{\mathrm{MSE}\} = \frac{1}{N}\mathbb{E}\left\{\|f(y) - x\|_2^2 + 2\langle x, f(y) - y\rangle - \langle x, y\rangle\right\} \tag{V–39}$$

Theorem V–3 applied twice on this expression with $\varphi = f$ and $\varphi = \mathrm{Id}$ (the identity function) leads to the expected expression (V–12) of PG-URE. As previously mentioned, we assume that the regularity and expectation conditions of Theorem V–3 hold for f.

V.B Optimal perturbation for the second-order derivative term of the PG-URE estimator

In this appendix, we derive an algebraic expression for the contribution V_δ of the perturbation δ to the variance (V–26) of the PG-URE estimator. This expression uses only the

coefficients $b_{k,l,m}$ defined by (V–27), and the moments $m_p = E\left(\delta_k^p\right)$ of the probability distribution used to generate the components of δ. We finally derive the optimal conditions on these moments m_p for V_δ to be minimal.

V.B.1 Expression of V_δ

First, we introduce a few notations:

- $c_k = b_{k,k,k}$ for all pixel index k,
- $d_{k,l} = b_{k,k,l} + b_{k,l,k} + b_{l,k,k}$, for all $k \neq l$,
- $D_l = \sum_{k,k\neq l} d_{k,l}$ for all l,
- $e_{k,l,m} = b_{k,l,m} + b_{k,m,l} + b_{l,k,m} + b_{m,k,l} + b_{l,m,k} + b_{m,l,k}$ for all 3-tuple (k,l,m) with $k \neq l$, $k \neq m$ and $l \neq m$.

We also recall that $m_1 = 0$, $m_2 = 1$, $m_3 = \kappa$. Then, starting from the definition of V_δ, we have:

$$N^2\kappa^2 V_\delta = \mathrm{Var}_\delta\left[\sum_{k,l,m} b_{k,l,m}\delta_k\delta_l\delta_m\right] = \left(\sum_{i,j,k,l,m,n} b_{i,j,k}b_{l,m,n}E_\delta\left(\delta_i\delta_j\delta_k\delta_l\delta_m\delta_n\right) - \left(\sum_{k,l,m} b_{k,l,m}E_\delta\left(\delta_k\delta_l\delta_m\right)\right)^2\right)$$

As explained in Sec. V.3.3, $E_\delta\left(\delta_k\delta_l\delta_m\right) = 0$ except when $k = l = m$: this is due to the independence of the components of δ and to the property $m_1 = 0$. This leads to the immediate simplification of the expression above:

$$N^2\kappa^2 V_\delta = \underbrace{\sum_{i,j,k,l,m,n} b_{i,j,k}b_{l,m,n}E_\delta\left(\delta_i\delta_j\delta_k\delta_l\delta_m\delta_n\right)}_{S_6} - \kappa^2\sum_{k,l} c_k c_l \tag{V–40}$$

The same arguments can be used to simplify the sixfold sum S_6, as $E_\delta\left(\delta_i\delta_j\delta_k\delta_l\delta_m\delta_n\right) = 0$ as soon as one of the six indices is different from the others. Then, S_6 can be divided according to the four situations where $E_\delta\left(\delta_i\delta_j\delta_k\delta_l\delta_m\delta_n\right)$ is non-zero:

$$S_6 = m_6T_6 + m_4T_{4,2} + \kappa^2T_{3,3} + T_{2,2,2} \tag{V–41}$$

- T_6 includes the terms involved in S_6 for which all the six summation indexes are equal: obviously, we have $T_6 = \sum_k c_k^2$;
- $T_{4,2}$ groups together all the terms such that, among the six summation indices, there is one group of four equal indices on the one hand, and another group of two equal indices on the other hand (for instance: $i = j = l = n \neq k = m$);

- in the same way, $T_{3,3}$ includes all the terms such that the indices form two groups of three.
- finally, $T_{2,2,2}$ covers the situation where there are three pairs of equal indices.

A careful enumeration of the terms involved in these situations leads to the following expressions[8]:

$$\begin{aligned} T_{4,2} &= \sum_{k \neq l} d_{k,l}^2 + 2 \sum_k c_k D_k \qquad T_{3,3} = \sum_{k \neq l} c_k c_l + \sum_{k \neq l} d_{k,l} d_{l,k} \\ T_{2,2,2} &= \sum_k D_k^2 + \sum_{k \neq l} d_{k,l}^2 + \frac{1}{6} \sum_{k \neq l \neq m} e_{k,l,m}^2 \end{aligned} \tag{V–42}$$

By putting all things together, we finally obtain:

$$\begin{aligned} N^2 V_{\delta} = & \frac{m_6 + m_4^2 + \kappa^2}{\kappa^2} \sum_k c_k^2 + \frac{m_4 + \kappa^2 + 1}{\kappa^2} \sum_{k \neq l} d_{k,l}^2 + \frac{1}{\kappa^2} \sum_k (m_4 c_k + D_k)^2 \\ & + \frac{1}{2} \sum_{k \neq l} (d_{k,l} + d_{l,k})^2 + \frac{1}{6\kappa^2} \sum_{k \neq l \neq m} e_{k,l,m}^2 \end{aligned} \tag{V–43}$$

It can be verified that this expression (V–43) is indeed positive, as for any probability distribution with moments m_p the following Hankel matrix H_p is positive (see [Akhiezer65]):

$$H_p = \begin{bmatrix} 1 & m_1 & m_2 & \cdots & m_p \\ m_1 & m_2 & & \iddots & \vdots \\ m_2 & & \iddots & & \vdots \\ \vdots & \iddots & & & \vdots \\ m_p & \cdots & \cdots & \cdots & m_{2p} \end{bmatrix} \tag{V–44}$$

In our case, this implies:

$$m_6 + m_4^2 + \kappa^2 \geq 0 \quad \text{and} \quad m_4 + \kappa^2 + 1 \geq 0 \tag{V–45}$$

As in the case of V_{θ}, we can analyze the order of magnitude of the contribution V_{δ} to the variance of the PG-URE estimator. As explained in Sec. V.3.5, for reasonable denoising operators f, the second order derivative $\frac{\partial^2 f_k}{\partial y_l \partial y_m}(y)$ is likely to be zero, except when the pixels corresponding to the indexes k, l and m share some spatial proximity. As the $b_{k,l,m}$ are proportional to these second order derivatives, and due to their definition, we deduce that the number of non-zero coefficients $d_{k,l}$ and $e_{k,l,m}$ is proportional to N; the order of magnitude of V_{δ} is therefore proportional to $\frac{1}{N}$, as claimed in Sec. V.3.5.

[8] In what follows, the notation $\sum_{k \neq l \neq m}$ is used to represent a sum of terms index by tuples (k,l,m) with $k \neq l$, $l \neq m$, and $m \neq k$ (as if $\neq$ were a transitive relation).