Phase Retrieval with Sparsity Constraints

Dissertation

zur Erlangung des mathematisch-naturwissenschaftlichen Doktorgrades
"Doctor rerum naturalium"
der Georg-August-Universität Göttingen

im Promotionsprogramm *Mathematical Sciences*
der Georg-August University School of Sciences (GAUSS)

vorgelegt von
Stefan Loock
aus Stade

Göttingen, 2016

Betreuungsausschuss

Prof. Dr. Gerlind Plonka-Hoch [1]
Prof. Dr. D. Russell Luke [1]

Mitglieder der Prüfungskommission

Referentin:	Prof. Dr. Gerlind Plonka-Hoch [1]
Korreferent:	Prof. Dr. D. Russell Luke [1]
Weitere Mitglieder:	PD Dr. Timo Aspelmeier [2]
	Prof. Dr. Dorothea Bahns [3]
	Prof. Dr. Thorsten Hohage [1]
	Prof. Dr. Max Wardetzky [1]

[1] Institut für Numerische und Angewandte Mathematik,
[2] Institut für Mathematische Stochastik,
[3] Mathematisches Institut,
Fakultät für Mathematik und Informatik, Georg-August-Universität Göttingen

Tag der mündlichen Prüfung: 7. Juni 2016

The true work of the mathematician is not experienced until the later parts of graduate school, when the student is challenged to create knowledge in the form of a novel proof. It is common to fill page after page with an attempt, the seasons turning, only to arrive precisely where you began, empty-handed — or to realize that a subtle flaw of logic doomed the whole enterprise from its outset. The steady state of mathematical research is to be completely stuck. It is a process that Charles Fefferman of Princeton, himself a onetime math prodigy turned Fields medalist, likens to "playing chess with the devil." The rules of the devil's game are special, though: The devil is vastly superior at chess, but, Fefferman explained, you may take back as many moves as you like, and the devil may not. You play a first game, and, of course, "he crushes you." So you take back moves and try something different, and he crushes you again, "in much the same way." If you are sufficiently wily, you will eventually discover a move that forces the devil to shift strategy; you still lose, but — aha! — you have your first clue.

Gareth Cook, New York Times Magazine [23]

Acknowledgments

Over the course of the last four years, I had the joy to work on this thesis in the pleasant atmosphere of the Institute for Numerical and Applied Mathematics at the Georg-August Universität Göttingen.

I am greatly indebted to my advisor Gerlind Plonka-Hoch for providing supervision and support throughout all this time. I owe many thanks to my co-advisor Russell Luke for his time and the interest he takes in my work. I am grateful for many inspiring conversations with them and their insightful views on many different aspects of my research and beyond. Moreover, I would like to thank both my advisors for also acting as referees for this thesis.

Furthermore, I would like to express many thanks to my colleagues in the "Research Group for Mathematical Signal and Image Processing" who provided a warm, welcoming atmosphere, a continuous coffee supply, and many interesting discussions.

This thesis would not have been possible without the generous financial support of the collaborative research center SFB755 "Nanoscale Photonic Imaging" which also provided for many interesting workshops and fantastic travel opportunities. In this regard, I would like to express my special gratitude to Tim Salditt whose enthusiasm in bringing together different scientific disciplines is exceptional.

The last years would not have been as interesting and joyful as they were without the friendship of many people. I would like to especially mention Sirach Lotz, Malte Rösner, Malte Schüler, and Bruno Schyska. They made my time in Bremen more cheerful, comforting, and inspiring than I could have ever hoped for. I am furthermore very thankful for the friendship of Jan Preibisch and Kai Feldhusen.

I am eternally grateful to my parents, Susann and Otmar Loock, for their understanding, advice, and never-ending support. Their hard work and dedication provided me with incredible role models and gave me the freedom to pursue my interests.

Special thanks are due to Anna Wolf whose determination was inspiring and set an example for me. It was her encouragement, support, and love which made me stick it out in the end. Thank you!

Contents

1. Introduction **1**
1.1. Notation 3
1.2. Outline of the Thesis 4

2. The Mathematical Model **7**
2.1. Maxwell's Equations 7
2.2. Scalar Diffraction Theory 9
2.3. Helmholtz Equation 9
2.4. Fresnel-Kirchhoff Diffraction 13
2.5. Rayleigh-Sommerfeld Diffraction 16
2.6. Fresnel Approximation 18
  2.6.1. Equivalent Representations of the Fresnel Transform 20
  2.6.2. Properties of the Fresnel Transform 22
2.7. Fraunhofer Diffraction 23
2.8. The Imaging Model 24
2.9. Experimental Setup 25
2.10. The Noise Model 26

3. Frames **29**
3.1. Wavelets 29
  3.1.1. Smoothness and Vanishing Moments 31
  3.1.2. Discrete Wavelets 32
  3.1.3. Filter Banks 38
  3.1.4. Wavelets in Two Dimensions 40
3.2. Shearlets 42
  3.2.1. Construction of Shearlets 43
  3.2.2. Compactly Supported Shearlets 46
  3.2.3. Discrete Shearlets 47

3.2.4. Numerical Implementation 53

4. Iterative Algorithms for Phase Retrieval 59

4.1. Projection Algorithms and Feasibility Formulation 60
4.2. New *a priori* Conditions Based on Sparsity 65
4.3. Convergence Results for Exact RAAR-(T, γ, ℓ_1) in the Convex Setting . 74
4.3.1. Convergence of RAAR-($\mathrm{Id}, \gamma, \ell_1$) for $\beta_k \equiv 1$ 74
4.3.2. Fixed-Points of RAAR-($\mathrm{Id}, 1, \iota_A$) 79
4.4. A Proximity Operator for Tight Frames 81
4.5. Boundedness and Cesàro-Convergence of RAAR-(T, γ, ℓ_1) 87
4.6. Generalizations to other Threshold Functions 93
4.7. Numerical Implementation 96

5. Numerical Evaluation 99

5.1. Discrete Shearlet Transform and Soft-Thresholding 99
5.2. Range Constraint 101
5.3. Simulating the Object Transmission Function 101
5.4. Details on the Numerical Implementation 102
5.4.1. Error Measures 103
5.5. Numerical Results for the Reconstruction of Simulated Data 104
5.5.1. Exact Data 104
5.5.2. Poisson Data 118
5.6. Numerical Results using Smooth-Hard Shrinkage 127
5.7. Applicability to Experimental Data 131

6. Conclusion 133

A. The Mathematical Model 137

B. Fundamental Results from Convex Analysis 141

Bibliography 151

Curriculum Vitae 161

List of Figures

2.1. Graphical illustration of Fresnel-Kirchhoff diffraction 13
2.2. Graphical illustration of the occurring angles in Fresnel-Kirchhoff diffraction 15
2.3. Graphical illustration of Rayleigh-Sommerfeld diffraction 17
2.4. Diffraction setup in rectangular coordinates 19

3.1. Mexican hat wavelet for $\sigma = 1$ and its Fourier transform 30
3.2. Flowchart illustrating two decomposition steps from the wavelet transform (top) and the inverse transform (bottom). 39
3.3. Induced tiling of the frequency plane $\widehat{\mathbb{R}}^2$ for classical band-limited, cone-adapted shearlets 45
3.4. Illustration of filters used for the discrete shearlet transform: (a) maximally flat, symmetric 9-tap low-pass filter, (b) magnitude of its Fourier transform, (c) magnitude of the Fourier transform of the maximally flat 2D fan filter. 56

4.1. Schematic flowchart describing the Gerchberg-Saxton algorithm, based on the original figure from [39]. 62

5.1. Illustration of different thresholding operations and the influence of the smoothing parameter α occurring in smooth-hard shrinkage. 100
5.2. Decay of the root-mean-squared error of a real-valued object for different constraints on exact data with starting guess (a) $x^{(0)} = \mathcal{D}_\tau^{-1} m$ and (b) $x^{(0)} = m$ 105
5.3. Two different initial guesses for a real-valued object with exact data where in (a) $x^{(0)} = \mathcal{D}_\tau^{-1} m$ and in (b) $x^{(0)} = m$ 106
5.4. Reconstructions of a real-valued object from exact measurements for different constraints with starting guess $x^{(0)} = \mathcal{D}_\tau^{-1} m$. 107

5.5. Reconstructions of a real-valued object from exact measurements for different constraints with starting guess $x^{(0)} = m$. 108
5.6. Amplitude object with estimated support (dashed line) used for numerical simulation of the object transmission function 109
5.7. Exact measurements of an amplitude object and initial guesses for phase and amplitude 110
5.8. Decay of the root-mean-squared error of an amplitude object for different constraints on exact data for (a) phase and (b) amplitude 110
5.9. Reconstructions of an amplitude object from exact measurements for different constraints 111
5.10. Phase object with estimated support (dashed line) used for numerical simulation of the object transmission function 112
5.11. Exact measurements of the phase object and initial guesses for phase and amplitude 113
5.12. Decay of the root-mean-squared error of a phase object for different constraints on exact data for (a) phase and (b) amplitude 113
5.13. Reconstructions of a phase object from exact measurements for different constraints 114
5.14. Mixed object with estimated support (dashed line) used for numerical simulation of the object transmission function 115
5.15. Exact measurements of a mixed object and initial guesses for phase and amplitude 116
5.16. Decay of the root-mean-squared error of a mixed object for different constraints on exact data for (a) phase and (b) amplitude 116
5.17. Reconstructions of a mixed object from exact measurements for different constraints 117
5.18. Poisson measurements of the amplitude object and initial guesses for phase and amplitude 119
5.19. Decay of the root-mean-squared error of an amplitude object for different constraints on Poisson data for (a) phase and (b) amplitude 119
5.20. Reconstructions of an amplitude object from Poisson measurements for different constraints 120
5.21. Poisson measurements of the phase object and initial guesses for phase and amplitude 121

5.22. Decay of the root-mean-squared error of a phase object for different constraints on Poisson measurements for (a) phase and (b) amplitude . 122
5.23. Reconstructions of a phase object from Poisson measurements for different constraints 123
5.24. Poisson measurements of a mixed object and initial guesses for phase and amplitude 124
5.25. Decay of the root-mean-squared error of a mixed object for different constraints on Poisson measurements for (a) phase and (b) amplitude . 125
5.26. Reconstructions of a mixed object from Poisson measurements for different constraints 126
5.27. Soft-thresholding compared to smooth-hard shrinkage for different parameters 127
5.28. Poisson measurements of a phase object and initial guesses for phase and amplitude 129
5.29. Decay of the root-mean-squared error of a phase object for different constraints comparing soft-thresholding and smooth-hard shrinkage of shearlet coefficients on Poisson measurements for (a) phase and (b) amplitude 129
5.30. Reconstructions of a phase object from Poisson measurements for different constraints comparing the reconstruction quality using soft-thresholding or smooth-hard shrinkage of shearlet coefficient 130

List of Tables

3.1. An example shearlet system structure computed using ShearLab3D 1.1 and MATLAB 2013a 55
3.2. Filter coefficients for the maximally flat symmetric 9-tap filter 56

5.1. Parameters for the numerical simulation of the object transmission function for amplitude objects with exact data 110
5.2. Parameters for numerical simulation of phase objects with exact data . 113
5.3. Parameters for the numerical simulation of the object transmission function for mixed objects with exact data 115
5.4. Parameters for the numerical simulation of the object transmission function for amplitude objects with Poisson data 118
5.5. Parameters for the numerical simulation of the object transmission function for phase objects with Poisson noise 121
5.6. Parameters for the numerical simulation of the object transmission function for mixed objects with Poisson noise 124
5.7. Parameters for the numerical simulation of the object transmission function for phase objects with Poisson noise using smooth-hard thresholding 128

1. Introduction

This thesis deals with methods on how to solve the phase retrieval problem which can be stated, in the discrete setting, as follows: Given a unitary mapping $U : \mathbb{C}^{d_1 \times d_2} \to \mathbb{C}^{d_1 \times d_2}$ and measurements $m \in \mathbb{R}_+^{d_1 \times d_2}$, find $x \in \mathbb{C}^{d_1 \times d_2}$ such that

$$|Ux|_\circ = m \tag{1.1}$$

where $|\cdot|_\circ$ denotes the point-wise modulus. For a treatment of the continuous phase retrieval problem, we refer to [74].

The phase retrieval problem arises in many applications, mainly in experimental physics. These applications include crystallography, astronomy, and electron microscopy. This thesis will be concerned with the latter, especially the microscopy using hard x-rays in the near-field regime. This includes x-rays with wavelengths on the scale of nanometers.

Traditionally, the phase retrieval problem was solved using algorithms that were called *iterative transform algorithms* such as the Gerchberg-Saxton algorithm [39], the error reduction algorithm [37], and the hybrid input-output (HIO) algorithm [38]. All these algorithms use projections onto measurement and constraint sets in an alternating fashion in order to determine a feasible solution. The connection of these iterative transform algorithms to existing projection algorithms was first drawn in [63], where it was recognized that the error-reduction algorithm is an instance of the method of alternating projections, first notably used in [106].

An overview of the related projection algorithms can be found in [7], the aforementioned similarities are further studied in [6]. The reference [8] introduces new variants based on a mathematical analysis of a convex variant of the underlying problem. In practice, relaxed versions such as the relaxed averaged alternating reflections (RAAR) algorithm [75] have been successfully used. For the RAAR algorithm, convergence results can be found in [75] for the convex case and in [76] for the non-convex setting. Recent results on the alternating projections method and the Douglas-Rachford algorithm for the non-convex setting are obtained in [65, 64, 51, 52, 50].

A different approach to solve the phase retrieval problem uses the transport of intensity equation

$$\nabla_{(x,y)} \cdot \Big(I(x,y,z)\nabla_{(x,y)}\phi(x,y,z)\Big) = -\frac{\partial I(x,y,z)}{\partial z}, \qquad \psi(x,y,z) \;=\; \sqrt{I(x,y,z)}\mathrm{e}^{\mathrm{i}\phi(x,y,z)} \tag{1.2}$$

to obtain the phase by numerically solving a PDE. This was originally proposed in [104] with recent application to experimental x-ray data, and it serves as a starting guess for iterative transform algorithms in [57]. Here, the right-hand-side $-{}^{\partial I(x,y,z)}/{}_{\partial z}$ is approximated by finite differences, using multiple intensity measurements of $I(x,y,z)$ at different positions $z \pm \Delta z$. Further similar approaches can be found in [47, 46, 20, 110].

Another widespread and promising approach is to use regularized Gauss-Newton methods to solve the inverse problem of phase retrieval. For recent results we refer to [99] for the one-dimensional problem and to [54, 109] for the general setting with data misfit terms for Poisson noise. The latter methods perform very well when being used with experimental data, see [81, 82].

Recently, compressed sensing techniques were applied to the phase retrieval problem as well, see [87, 90, 18]. Since the problem is non-linear and the corresponding minimization problem non-convex, lifting schemes are applied in order to linearize the problem. While these methods promise unique recovery and global convergence, due to the lifting into a higher dimensional setting, the algorithmic complexity is squared in comparison to the lower-dimensional model. This renders these methods unusable for most practical problems. Furthermore, a randomization of the measurement process is often necessary in order to achieve the restricted isometry property of the measurement matrix with high probability which is essential to show the equivalence of the non-convex to a convex minimization problem.

Further research includes wavelet methods applied to the phase retrieval problem in different settings. In [101], wavelets were used together with the transport of intensity equation. Moreover, wavelet methods have been employed in different phase retrieval settings. In [2], the assumption is used that a low-pass version of the signal is known, which is, of course, a very strong assumption. A wavelet Wiener filtering has been applied in [62] in order to improve reconstruction results. Here, one determines optimal shrinkage coefficients based on the assumption that the signal is corrupted by additive Gaussian noise. In [108], a wavelet transform with an application adapted generator was chosen in order to improve on traditional wavelet approaches. The authors of [89] used an orthogonal matching pursuit (OMP) in combination with projection methods to obtain sparse solutions of the phase retrieval problem. An iterative nonlinear method using Tikhonov regularization was performed in [27] which then was combined using an orthogonal wavelet basis which was further investigated in [28, 29].

The multi-resolution properties of the wavelet transform were also employed in [67, 66, 68] in the context of digital holography.

However, wavelets are not optimally suited for sparse representations of two-dimensional objects such as images with singularities along curves. This drawback will be addressed in this thesis. First, we embed the approach of sparsity into the framework of projection methods and show how the often used soft-threshold operator can be interpreted as a proximity operator using tight frames. We will furthermore use shearlets, a representation system that almost optimally (up to logarithmic factors) approximates cartoon-like images.

This motivates our approach where, on the one hand, we develop an algorithm based on the well understood RAAR algorithm using a generalization of the projections. On the other hand, we will employ shearlets which are optimally suited for the purpose of sparse approximation of images. First numerical simulations together with a discussion on the convergence behavior in the discrete setting were published in [72]. The applicability of this method to experimental data in x-ray imaging has been shown in [93].

1.1. Notation

While the mathematical model and part of the wavelet theory is formulated in the continuum, the main part of this thesis will focus on the discrete setting. Infinite dimensional Hilbert spaces will be denoted by $\mathcal{H}$. However, most of the time, the underlying vector spaces will be finite dimensional. We denote by $\mathbb{E}$ an arbitrary finite dimensional Hilbert space which may be over the field of real or complex numbers.

We will use matrices as a representation for discrete images which we denote by $x \in \mathbb{E}$, i.e., we may have $x \in \mathbb{E} = \mathbb{R}^{d_1 \times d_2}$ or $x \in \mathbb{E} = \mathbb{C}^{d_1 \times d_2}$. However, in most cases images can also be vectorized without problems such that we may write $\mathbb{E} = \mathbb{R}^{d_1 \cdot d_2} = \mathbb{R}^d$ or similarly $\mathbb{E} = \mathbb{C}^{d_1 \cdot d_2} = \mathbb{C}^d$. We will denote the j-th component of such an image or vector $x \in \mathbb{E}$ by $x[j]$ where on the other hand j is implied to be a tuple $j = (j_1, j_2)$ if

x is an image. We will only use the extended notation $x[j_1, j_2]$ if neccessary.

We denote by $|\cdot|_\circ$ the point-wise modulus of a vector or matrix and by $\odot$ the point-wise product of vectors and matrices.

We denote the space of proper, lower semi-continuous, convex functions $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ by $\Gamma_0(\mathbb{R}^d)$.

The *difference set* $B - A$ is defined as

$$B - A := \{b - a \mid a \in A, b \in B\}$$

which follows the convention of the *Minkowski sum* of two sets. Similarly, for any vector $g \in \mathbb{R}^d$ and set $A \subset \mathbb{R}^d$ we define

$$A - g := \{a - g \mid a \in A\}.$$

Further notation will be introduced when needed.

1.2. Outline of the Thesis

This thesis is organized as follows. Chapter 2 introduces the mathematical model that is used for the operator U in (1.1). Following the derivation given in [13, 42], we start with Maxwell's equations and use assumptions on the medium as well as approximations such as scalar diffraction theory to derive the Helmholtz equation. Using a specific setup and further approximations, the Fresnel transform is derived as an approximation to the near field as well as a formula for the far field that is very similar to the Fourier transform. We further introduce the imaging model and provide more details from the experimental setup as well as a motivation for the noise model.

Chapter 3 introduces the shearlet transform, following the exposition in [60], which is used in the algorithm. First, a brief summary on continuous and discrete wavelets is given based on [79]. The fast wavelet transform using wavelet filter banks is described and a two-dimensional tensor-product approach outlines how wavelets can be used in image processing. Based on the presented wavelet theory, the construction of compactly supported shearlet frames is explained. Shearlet frames were first introduced in [61]. Starting with cone-adapted, band-limited shearlets used in [61, 45], we finally discuss compactly supported shearlets with non-separable generators as proposed in [69, 70, 60]. We furthermore describe the discretization scheme of the wavelet functions and the shearing operator and give examples for wavelet and scaling functions following the examples in [60]. The chapter concludes with a short exposition of the numerical implementation.

In Chapter 4 we introduce iterative algorithms for phase retrieval following [6, 7]. We briefly review the most common projection algorithms and give some references to convergence results. Based on the RAAR algorithm which was developed in [75, 76], we propose a new algorithm for phase retrieval with sparsity constraints using soft-thresholding of frame coefficients. We prove results for the convergence behavior in the finite dimensional, discrete setting. These results are published in [72]. Furthermore we prove a result for the convex case using the Douglas-Rachford algorithm. We discuss a problem that arises when soft-thresholding is used with frames instead of unitary transforms. This problem is very common in image processing applications and was already discussed in [35]. The main contribution on this problem is that we show that the soft-thresholding of tight frame coefficients is the proximity operator of a lower continuous, convex function. We discuss the implications and further thresholding functions such as smooth-hard shrinkage. The chapter closes with details on the numerical implementation of the projection and proximity operators.

We numerically evaluate the proposed method in Chapter 5. Using the introduced tools, we perform reconstructions using simulated data of different types of objects on exact data as well as data corrupted with Poisson noise. For real-valued objects, the results are published in [72]. We further evaluate the method on amplitude, phase, and mixed objects using soft-thresholding. The results from this new method are compared to existing methods and shown to outperform them. Furthermore, we compare the results using smooth-hard shrinkage against the results using soft-thresholding on a phase object. The application to experimental data using soft-thresholding is shown in [93].

Chapter 6 summarizes the thesis, discusses open questions, and provides an outlook for future research.

2. The Mathematical Model

This chapter introduces the basic ideas of scalar diffraction theory that results as an approximation to Maxwell's equations. In order to derive the diffraction integral given by Fresnel, the first approximation is from vectorial to scalar diffraction theory. This simplification leads to the Helmholtz equation and the integral theorem by Helmholtz and Kirchhoff. Based on this, Kirchhoff's formulation of diffraction theory gives some insight into diffraction on a plane. Finally, the perspective by Rayleigh and Sommerfeld, using a different choice of Green's function, results in the diffraction formula named after them. In practice, Fresnel and Fraunhofer diffraction play an important role. We discuss these two approximations in the end of the chapter. This chapter is based on [41, 42, 74, 77] and [13] as well as [91] and [92].

2.1. Maxwell's Equations

Maxwell's equations describe the propagation of electromagnetic waves, e.g., light. By $E = \left(E_x, E_y, E_z\right)$ we denote the electric field and by $H = \left(H_x, H_y, H_z\right)$ the magnetic field, respectively. Furthermore, ε describes the dielectricity and μ is the magnetic permeability of the medium. The constants ε_0, μ_0 are the corresponding vacuum constants. Maxwell's equations are given by

$$\nabla \times E = -\mu \frac{\partial H}{\partial t} \tag{2.1}$$

$$\nabla \times H = \varepsilon \frac{\partial E}{\partial t} \tag{2.2}$$

$$\nabla \cdot \varepsilon E = 0 \tag{2.3}$$

$$\nabla \cdot \mu H = 0. \tag{2.4}$$

Using assumptions on the medium, we will derive the wave equations describing the propagation of the wave.

Consider a dielectrically isotropic medium, i.e., the dielectric properties are independent of the direction of the polarization of the wave. A medium is said to be *homogenous* if ε is spatially constant, and *non-dispersive* if ε does not depend on the wavelength λ. We further assume that wave propagation happens in the vacuum, i.e., $\mu = \mu_0$ and $\varepsilon = \varepsilon_0$. Using these assumptions, we can derive a wave equation for the electric field E. The derivation for the magnetic field follows using analogous steps. Applying the rotation operator $\nabla\times$ to the first Maxwell equation (2.1), we obtain

$$\nabla \times (\nabla \times E) = \nabla \times \left(-\mu \frac{\partial H}{\partial t}\right) = -\nabla^2 E,$$

where we used

$$\nabla \times (\nabla \times E) = \nabla (\nabla \cdot E) - \nabla^2 E$$

as well as the absence of charge, i.e., $\nabla \cdot E = 0$ in (2.3). Furthermore, for a linear, isotropic and non-dispersive homogenous medium, we can change the order of differentiation with respect to space and time and obtain[1]

$$-\nabla^2 E = -\mu \frac{\partial}{\partial t} (\nabla \times H) .$$

Using the second Maxwell equation (2.2) we obtain the wave equation for the electric field

$$\nabla^2 E - \mu\varepsilon \frac{\partial^2 E}{\partial t^2} = 0. \tag{2.5}$$

We define the propagation velocity in vacuum c and the refractive index n by

$$c = \frac{1}{\sqrt{\mu_0 \varepsilon_0}}, \qquad n = \sqrt{\varepsilon/\varepsilon_0},$$

i.e.,

$$\frac{n^2}{c^2} = \frac{\varepsilon \mu_0 \varepsilon_0}{\varepsilon_0} = \varepsilon_0 \mu_0$$

---

[1] Here, ∇^2 denotes the vector Laplacian which is defined for a vector field $A : \mathbb{R}^n \to \mathbb{R}^n$ by $\nabla^2 A = \nabla (\nabla \cdot A) - \nabla \times (\nabla \times A)$. This, in cartesian coordinates, is the same as $\nabla^2 A = \left(\Delta A_x, \Delta A_y, \Delta A_z\right)$ where Δ denotes the (scalar) Laplacian and A_x, A_y, A_z are the components of A. We use the different notation Δ and ∇^2 to indicate weather the Laplacian acts on a scalar or a vector field.

since $\varepsilon = \varepsilon_0$ and $\mu = \mu_0$. Plugging this into the wave equation (2.5), we obtain

$$\nabla^2 E - \frac{n^2}{c^2}\frac{\partial^2 E}{\partial t^2} = 0.$$

In complete analogy one derives the wave equation for the magnetic field H and obtains that both E and H fulfill the same wave equation.

2.2. Scalar Diffraction Theory

Since the electric and magnetic fields fulfill the same wave equation, every component of E and H satisfies the scalar wave equation

$$\Delta u(x,t) - \frac{n^2}{c^2}\frac{\partial^2 u(x,t)}{\partial t^2} = 0, \tag{2.6}$$

with $u(x,t) : \mathbb{R}^3 \times \mathbb{R} \to \mathbb{C}$ and where $u(x,\cdot)$ represents an arbitrary component of E or H. In this case, scalar and vectorial diffraction theory are the same. In the case of inhomogeneous media, i.e., $\varepsilon = \varepsilon(x)$, we would obtain mixed derivatives of the refractive index n with respect to the position what would lead to a coupling of the different components of the field. Thus, every component of the field would satisfy a different wave equation.

A coupling effect also occurs when posing boundary conditions on bounded domains – the coupling of electric and magnetic field on the boundary of the domain even occurs in homogenous media.

In our model, we consider the diffraction of light by an aperture. The coupling occurs due to the interaction of light and matter at the boundary of the aperture. The coupling effects only reach some wavelengths into the aperture, therefore if the size of the aperture $\mathcal{A}$ is much larger than the wavelength, this coupling can be neglected. We will therefore focus on scalar diffraction theory as it is sufficiently accurate to model the behavior of light in diffractive imaging.

2.3. Helmholtz Equation

An important step is the simplification using assumptions on the model. The x-rays are assumed to only have one wavelength, i.e., they are *monochromatic*. For position x

and time t we therefore can write the scalar field as

$$u\,(x,t) = A(x)\cos\left[2\pi\nu t - \Phi(x)\right]$$

where $A(x)$ denotes the amplitude, ν the optical frequency and $\Phi(x)$ is a phase function. Furthermore, we define the *phasor* $U(x) : \mathbb{R}^3 \to \mathbb{C}$ by

$$U(x) := A(x)\exp\left[\mathrm{i}\Phi(x)\right].$$

We are now able to write the scalar field as the real part of a product of two functions, one as a function of position and one as a function of time, i.e.,

$$u(x,t) = \operatorname{Re}\left\{U(x)\exp\left[-2\pi\mathrm{i}\nu t\right]\right\}. \tag{2.7}$$

We derive the Helmholtz equation by plugging (2.7) into the scalar wave equation (2.6). This yields

$$\Delta\left(\operatorname{Re}\left\{U(x)\exp\left[-2\pi\mathrm{i}\nu t\right]\right\}\right) - \frac{n^2}{c^2}\frac{\partial^2}{\partial t^2}\left(\operatorname{Re}\left\{U(x)\exp\left[-2\pi\mathrm{i}\nu t\right]\right\}\right) = 0.$$

Interchanging the differential operators with the Re operation leads to

$$\operatorname{Re}\left\{\Delta\left(U(x)\exp\left[-2\pi\mathrm{i}\nu t\right]\right)\right\} - \operatorname{Re}\left\{\frac{n^2}{c^2}\frac{\partial^2}{\partial t^2}U(x)\exp\left[-2\pi\mathrm{i}\nu t\right]\right\} = 0.$$

Since $U(x)$ does not depend on the time t, the second partial derivative w.r.t. t gives a factor $(-2\pi\mathrm{i}\nu)^2 = 4\pi^2\nu^2$. Setting $k = {}^{2\pi n\nu}/_{c} = {}^{2\pi}/_{\lambda}$ where $\lambda = {}^{c}/_{n\nu}$, we obtain

$$\operatorname{Re}\left\{\Delta U(x)\exp\left[-2\pi\mathrm{i}\nu t\right]\right\} + k^2\operatorname{Re}\left\{U(x)\exp\left[-2\pi\mathrm{i}\nu t\right]\right\} = 0.$$

Therefore, $U(x)$ in (2.7) fulfills the time independent equation

$$\left(\Delta + k^2\right)U(x) = 0 \tag{2.8}$$

which is known as the homogeneous Helmholtz equation. The next step is to express $U(x)$ in terms of the values of a boundary integral. Therefore, we will make use of Green's identity which is a consequence of Gauß' divergence theorem, cf. [102, Theorem 15K] and [103, Section 7.2].

Corollary 2.3.1 (Green's Identity). *Let $\Omega \subset \mathbb{R}^n$ be compact with piecewise smooth boundary $S := \partial\Omega$, and $U, G \in C^2(\Omega)$. Then*

$$\int_\Omega [U(x)\Delta G(x) - G(x)\Delta U(x)] \,\mathrm{d}x = -\int_S \left[U(x)\frac{\partial G(x)}{\partial n} - G(x)\frac{\partial U(x)}{\partial n}\right] \mathrm{d}\sigma \qquad (2.9)$$

where $\partial/\partial n$ denotes the partial derivative with respect to the inward *normal to S.*

Suppose that U and G fulfill the homogeneous Helmholtz equation (2.8). Therefore, it follows that

$$U(x)\Delta G(x) - G(x)\Delta U(x) = U(x)\left(-k^2 G\right) - G(x)\left(-k^2 U(x)\right) = 0$$

and hence

$$\int_\Omega [U(x)\Delta G(x) - G(x)\Delta U(x)] \,\mathrm{d}x = \int_S \left[U(x)\frac{\partial G(x)}{\partial n} - G(x)\frac{\partial U(x)}{\partial n}\right] \mathrm{d}\sigma = 0. \qquad (2.10)$$

We chose the auxiliary function

$$G_0(x;p) = \frac{\mathrm{e}^{\mathrm{i}k|x-p|}}{|x-p|}$$

that has a singularity in $x = p$. This, indeed, does not fulfill the homogeneous Helmholtz equation but instead satisfies

$$\left(\Delta + k^2\right) G_0(x;p) = -4\pi\delta(x-p), \qquad (2.11)$$

see Lemma A.1.1.
In order to apply (2.10), we consider the equation on the modified domain $\Omega_\varepsilon := \Omega \setminus \mathbb{B}_\varepsilon(p)$ where $x \neq p$, thus the right hand side in (2.11) vanishes and we can apply (2.10). This leads to a new boundary $\partial\Omega_\varepsilon = S \cup S_\varepsilon$ where $S := \partial\Omega$ and $S_\varepsilon := \partial\mathbb{B}_\varepsilon(p)$ and we will study the limit as $\varepsilon \to 0$. In other words, we have

$$\begin{aligned}\int_{\Omega_\varepsilon} [U(x)\Delta G_0(x;p) - G_0(x;p)\Delta U(x)] \,\mathrm{d}x &= \int_{S\cup S_\varepsilon} \left[U(x)\frac{\partial G_0(x;p)}{\partial n} - G_0(x;p)\frac{\partial U(x}{\partial n}\right] \mathrm{d}\sigma \\ &= 0,\end{aligned}$$

i.e.,

$$\int_S \left[U(x)\frac{\partial G_0(x;p)}{\partial n} - G_0(x;p)\frac{\partial U(x)}{\partial n} \right] \mathrm{d}\sigma = -\int_{S_\varepsilon} \left[U(x)\frac{\partial G_0(x;p)}{\partial n} - G_0(x;p)\frac{\partial U(x)}{\partial n} \right] \mathrm{d}\sigma$$

We will now show that the right-hand-side is equal to $-4\pi U(p)$. For the partial derivative ${}^{\partial G_0}/_{\partial n}$ on S_ε we have [2]

$$\begin{aligned} \frac{\partial G_0(x;p)}{\partial n} &= \frac{\partial}{\partial n}\frac{\mathrm{e}^{\mathrm{i}k|x-p|}}{|x-p|} = \left(\mathrm{i}k - \frac{1}{|x-p|}\right)\frac{\mathrm{e}^{\mathrm{i}k|x-p|}}{|x-p|}\cos\left(n, x-p\right) \\ &= \left(\frac{1}{|x-p|} - \mathrm{i}k\right)\frac{\mathrm{e}^{\mathrm{i}k|x-p|}}{|x-p|} \end{aligned}$$

since $\cos\left(n, x-p\right) = -1$ on S_ε and therefore

$$\begin{aligned} &-\int_{S_\varepsilon} \left[U(x)\frac{\partial G_0(x;p)}{\partial n} - G_0(x;p)\frac{\partial U(x)}{\partial n} \right] \mathrm{d}\sigma \\ &= -\int_{S_\varepsilon} \left[U(x)\frac{\mathrm{e}^{\mathrm{i}k|x-p|}}{|x-p|}\left(\frac{1}{|x-p|} - \mathrm{i}k\right) - \frac{\mathrm{e}^{\mathrm{i}k|x-p|}}{|x-p|}\frac{\partial U(x)}{\partial n} \right] \mathrm{d}\sigma \end{aligned}$$

Expressing the integrand in spherical coordinates yields

$$\begin{aligned} &-\int_{S_\varepsilon} \left[U(x)\frac{\mathrm{e}^{\mathrm{i}k|x-p|}}{|x-p|}\left(\frac{1}{|x-p|} - \mathrm{i}k\right) - \frac{\mathrm{e}^{\mathrm{i}k|x-p|}}{|x-p|}\frac{\partial U(x)}{\partial n} \right] \mathrm{d}\sigma \\ &= -\int_{[0,\pi]\times[0,2\pi)} \left[U(x)\frac{\mathrm{e}^{\mathrm{i}k\varepsilon}}{\varepsilon}\left(\frac{1}{\varepsilon} - \mathrm{i}k\right) - \frac{\mathrm{e}^{\mathrm{i}k\varepsilon}}{\varepsilon}\frac{\partial U(x)}{\partial n} \right] \varepsilon^2 \sin\vartheta \,\mathrm{d}\vartheta\,\mathrm{d}\varphi. \end{aligned}$$

Recall that ε is the radius of the ball located at p. Since U and its partial derivatives were assumed to be continuous and Ω_ε is a bounded domain, we can change the integration with the limit. Therefore,

$$-\int_{[0,\pi]\times[0,2\pi)} \lim_{\varepsilon\searrow 0}\left[U(x)\mathrm{e}^{\mathrm{i}k\varepsilon} - \varepsilon \mathrm{i}kU(x)\mathrm{e}^{\mathrm{i}k\varepsilon} - \varepsilon\mathrm{e}^{\mathrm{i}k\varepsilon}\frac{\partial U(x)}{\partial n}\right] \sin\vartheta\,\mathrm{d}\vartheta\,\mathrm{d}\varphi = -4\pi U(p)$$

since the second and third term vanish in the limit and $\lim_{\varepsilon\to 0}\mathrm{e}^{\mathrm{i}k\varepsilon} = 1$ while the

---

[2] In this setting, n denotes the outward normal vector on S_ε and $x-p$ is the vector pointing from the center of the ball $\mathbb{B}_\varepsilon(p)$ onto the point x on S_ε. Hence, the vectors are parallel but with opposed directions, hence $\angle(n, x-p) = \pi$. Note that here n is outward with respect to S_ε but inward with respect to the modified domain Ω_ε as required by Corollary 2.3.1.

integration of $\sin\vartheta$ over $[0,\pi]\times[0,2\pi)$ contributes 4π. For $\varepsilon\to 0$ we further have $x\to p$ and hence $U(x)\to U(p)$ due to continuity. We therefore established the relationship

$$-4\pi U(p)=\int_S\left[U(x)\frac{\partial G_0(x;p)}{\partial n}-G_0(x;p)\frac{\partial U(x)}{\partial n}\right]\mathrm{d}\sigma$$

or rewritten

$$U(p)=\frac{1}{4\pi}\int_S\left[\frac{\mathrm{e}^{\mathrm{i}k|x-p|}}{|x-p|}\frac{\partial U(x)}{\partial n}-U(x)\frac{\partial}{\partial n}\left(\frac{\mathrm{e}^{\mathrm{i}k|x-p|}}{|x-p|}\right)\right]\mathrm{d}\sigma \tag{2.12}$$

which is the *integral theorem of Helmholtz and Kirchhoff*, cf. [42].

2.4. Fresnel-Kirchhoff Diffraction

As the next step, we derive the Fresnel-Kirchhoff diffraction formula. Therefore we consider a monochromatic wave starting from a point source at p_0. We are interested in the behavior when the wave hits an opaque screen with a small opening $\mathcal{A}$. We want to derive an expression of the wave at the point p. Therefore, we assume that the opening is large compared to the wavelength but small compared to the propagation distance from p_0 to $\mathcal{A}$ as well as from $\mathcal{A}$ to p. In our setting, we have $S=\mathcal{A}\cup\mathcal{B}\cup\mathcal{C}$ with $\mathcal{A},\mathcal{B},\mathcal{C}$ as in Figure 2.1.

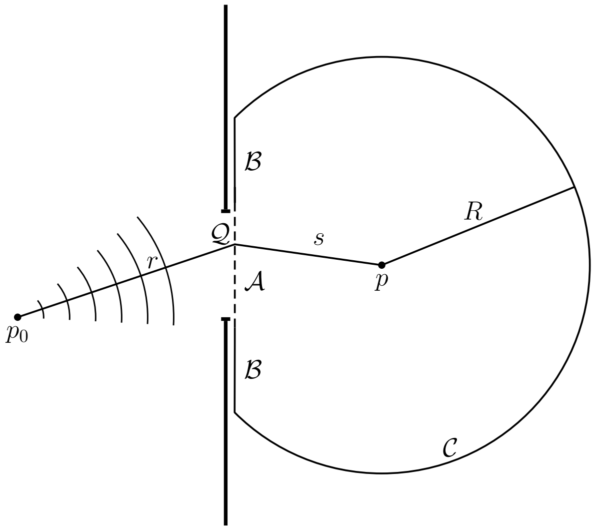


Figure 2.1.: Graphical illustration of Fresnel-Kirchhoff diffraction

We start to discuss the behavior on C. On C we have $\left|x-p\right| = R$ for all $x \in C$ and therefore $G_0(x;p) = {}^{\exp(\mathrm{i}kR)}/_{R}$. In order to imply that we are considering $G_0(x;p)$ on C we write $G_0(R)$ instead. For the derivative we obtain

$$\frac{\partial G_0(R)}{\partial n} = \left(\mathrm{i}k - \frac{1}{R}\right)\frac{\mathrm{e}^{\mathrm{i}kR}}{R} \approx \mathrm{i}kG(R)$$

for large R.[3,4] This leads to the approximation

$$\int_C \left[G_0(R)\frac{\partial U(x)}{\partial n} - U(x)\left(\mathrm{i}kG_0(R)\right)\right]\mathrm{d}\sigma = \int_{\mathcal{M}'} G_0(R)\left(\frac{\partial U(x)}{\partial n} - \mathrm{i}kU(x)\right)R^2\,\mathrm{d}\varphi\,\mathrm{d}\vartheta$$

where $\mathcal{M}' \subset [0,\pi]\times[0,2\pi)$ is the part of the solid angle covering C. Since $G_0(R) = {}^{\mathrm{e}^{\mathrm{i}kR}}/_{R}$ on C, $|RG_0|$ is uniformly bounded on C. We further use the *Sommerfeld radiation condition* from [100], i.e.,

$$\lim_{R\to\infty} R\left(\frac{\partial U(x)}{\partial n} - \mathrm{i}kU(x)\right) = 0. \tag{2.13}$$

This guarantees that the integral over C vanishes for sufficiently large R. In order to handle the integrals over $\mathcal{A}$ and $\mathcal{B}$, we use *Kirchhoff's boundary conditions*. These are

$$U(x) = U_i(x), \qquad \frac{\partial U(x)}{\partial n} = \frac{\partial U_i(x)}{\partial n} \qquad \text{on } \mathcal{A} \tag{2.14}$$

$$U(x) = 0, \qquad \frac{\partial U(x)}{\partial n} = 0, \qquad \text{on } \mathcal{B} \tag{2.15}$$

where

$$U_i(x;p_0) = \frac{A\mathrm{e}^{\mathrm{i}k|x-p_0|}}{\left|x-p_0\right|}, \qquad \frac{\partial U_i(x;p_0)}{\partial n} = \frac{A\mathrm{e}^{\mathrm{i}k|x-p_0|}}{\left|x-p_0\right|}\left(\mathrm{i}k - \frac{1}{\left|x-p_0\right|}\right)\cos\left(n, x-p_0\right).$$

Here $(n,\, x-p_0)$ denotes the angle between n and $x-p_0$ and U_i is the *incident field* with amplitude A, see Figure 2.2.

Let us briefly comment on the assumptions made here. The first assumption (2.14) implies that the values of U and ${}^{\partial U}/_{\partial n}$ on $\mathcal{A}$ are invariant under the presence of the screen. The second assumption (2.15) simply means that the field and its derivative

---

[3] A priori, it is not completely clear what *"for large R"* means. For sake of brevity, we assume R in practice to be so large that this approximation leads to negligible errors.

[4] In this setting, opposed to above, the inward normal n on C faces into the same direction as $x-p$ and we therefore have $\cos(n, x-p) = 1$.

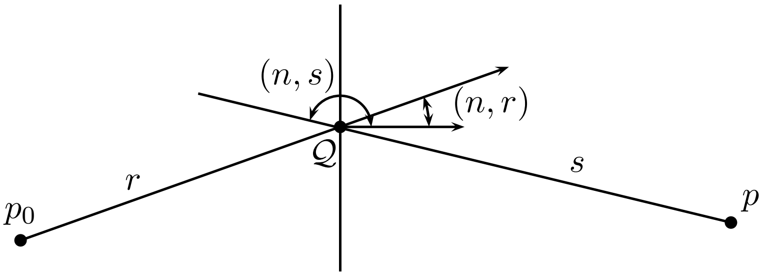


Figure 2.2.: Graphical illustration of the occurring angles in Fresnel-Kirchhoff diffraction

in normal direction directly behind the screen, i.e. on $\mathcal{B}$, vanish. To conclude, we can write the field at p as

$$U(p) = \frac{1}{4\pi} \int_{\mathcal{A}} \left[U_i(x) \frac{\partial G_0(x;p)}{\partial n} - G_0(x;p) \frac{\partial U_i(x)}{\partial n} \right] \mathrm{d}\sigma. \tag{2.16}$$

We now use (2.16) and apply all the a priori information that we gathered in the last section. Let us briefly recall the functions U_i and G_0 and their normal derivatives.

$$\begin{aligned} U_i(x;p_0) &= \frac{A\mathrm{e}^{\mathrm{i}k|x-p_0|}}{|x-p_0|}, & \frac{\partial U_i(x;p_0)}{\partial n} &= \frac{A\mathrm{e}^{\mathrm{i}k|x-p_0|}}{|x-p_0|} \left(\mathrm{i}k - \frac{1}{|x-p_0|} \right) \cos(n, x-p_0) \\ G_0(x;p) &= \frac{\mathrm{e}^{\mathrm{i}k|x-p|}}{|x-p|}, & \frac{\partial G_0(x;p)}{\partial n} &= \left(\mathrm{i}k - \frac{1}{|x-p|} \right) \frac{\mathrm{e}^{\mathrm{i}k|x-p|}}{|x-p|} \cos(n, x-p) \end{aligned}$$

Plugging this into (2.16) yields

$$\begin{aligned} U(p) = \frac{1}{4\pi} \int_{\mathcal{A}} & \frac{A\mathrm{e}^{\mathrm{i}k|x-p_0|}}{|x-p_0|} \left(\mathrm{i}k - \frac{1}{|x-p|} \right) \frac{\mathrm{e}^{\mathrm{i}k|x-p|}}{|x-p|} \cos(n, x-p) \, \mathrm{d}\sigma \\ & - \frac{\mathrm{e}^{\mathrm{i}k|x-p|}}{|x-p|} \frac{A\mathrm{e}^{\mathrm{i}k|x-p_0|}}{|x-p_0|} \left(\mathrm{i}k - \frac{1}{|x-p_0|} \right) \cos(n, x-p_0) \, \mathrm{d}\sigma. \end{aligned}$$

Writing $r := x - p_0$ and $s := x - p$ and neglecting the terms $1/|r|$ and $1/|s|$ in the normal derivatives yields

$$\begin{aligned} U(p) &= \frac{1}{4\pi} \int_{\mathcal{A}} \frac{A\mathrm{e}^{\mathrm{i}k|r|}}{|r|} \mathrm{i}k \frac{\mathrm{e}^{\mathrm{i}k|s|}}{|s|} \cos(n, s) - \frac{A\mathrm{e}^{\mathrm{i}k|r|}}{|r|} \frac{\mathrm{e}^{\mathrm{i}k|s|}}{|s|} \mathrm{i}k \cos(n, r) \, \mathrm{d}\sigma \\ &= \frac{A\mathrm{i}k}{4\pi} \int_{\mathcal{A}} \frac{\mathrm{e}^{\mathrm{i}k(|s|+|r|)}}{|sr|} \left[\cos(n, s) - \cos(n, r) \right] \, \mathrm{d}\sigma \end{aligned}$$

Substituting $k = {}^{2\pi}/_{\lambda}$ and using $\mathrm{i} = {}^{-1}/_{\mathrm{i}}$ finally yields

$$U(p) = \frac{A}{\mathrm{i}\lambda} \int_{\mathcal{A}} \frac{\mathrm{e}^{\mathrm{i}k(|r|+|s|)}}{|rs|} \left[\frac{\cos(n,r) - \cos(n,s)}{2} \right] \mathrm{d}\sigma. \tag{2.17}$$

This representation given by (2.17) is known as *Fresnel-Kirchhoff diffraction formula*. Note that, while practically relevant, the assumed boundary conditions are theoretically questionable. Suppose a two-dimensional potential function vanishes together with its normal derivative on any arbitrary line segment. Then by [58, Theorem 6.7], this function is identically zero in the whole plane. This contradicts the result established here.

The next section will establish an equally convenient expression for the solution whilst circumventing this problem.

2.5. Rayleigh-Sommerfeld Diffraction

In order to avoid the aforementioned inconsistencies we will work with a slightly modified model. The assumptions on U and ${}^{\partial U}/_{\partial n}$ vanishing on $\mathcal{B}$ can be circumvented with the following setup. We consider two point charges at the positions p_0 and p_1, cf. Figure 2.3. It can be shown that the Green's function

$$G(x; p_0, p_1) = \frac{\mathrm{e}^{\mathrm{i}k\left|x-p_0\right|}}{\left|x - p_0\right|} - \frac{\mathrm{e}^{\mathrm{i}k\left|x-p_1\right|}}{\left|x - p_1\right|} \tag{2.18}$$

solves the Helmholtz equation

$$\left(\Delta + k^2\right) G(x; p_0, p_1) = 4\pi \left(\delta(x - p_0) - \delta(x - p_1)\right)$$

in the right half-space $\Omega := \{(x_1, x_2, z) \in \mathbb{R}^3 \mid z \geq 0\}$ using similar arguments as in Lemma A.1.1 and

$$G(x; p_0, p_1) \equiv 0 \qquad \text{on} \left\{(x_1, x_2, z) \in \mathbb{R}^3 \mid z = 0\right\}$$

since $\left|x - p_0\right| = \left|x - p_1\right|$ for all $x \in \{x = (x_1, x_2, z) \in \mathbb{R}^3 \mid z = 0\}$. Assuming the radiation condition (2.13) one can, similarly to the section before, show that the integral over the infinitely large hemisphere C vanishes. Since G vanishes on the plane where $z = 0$,

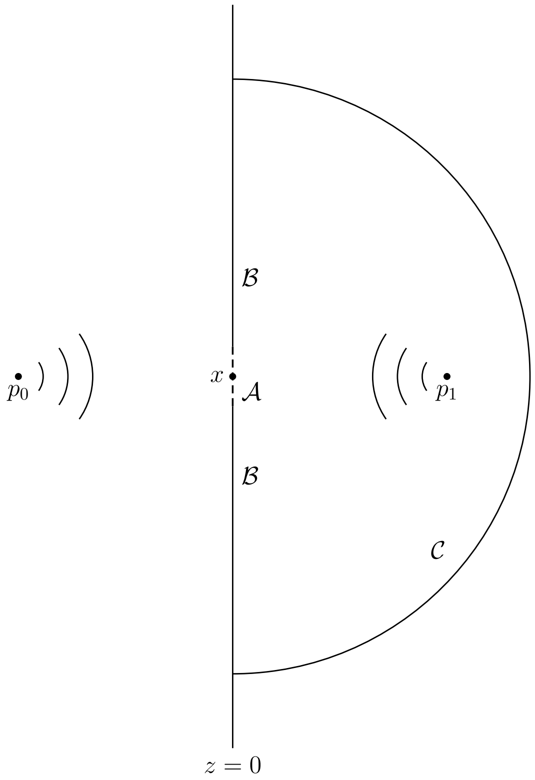


Figure 2.3.: Graphical illustration of Rayleigh-Sommerfeld diffraction

we obtain the approximation

$$U(p) = -\frac{1}{4\pi}\int_{\mathcal{A}} U(x)\,\frac{\partial G(x;p_0,p_1)}{\partial n}\,\mathrm{d}\sigma, \tag{2.19}$$

the *first Rayleigh-Sommerfeld solution*, cf. [13, Section 8.11].

We now explicitly calculate the normal derivative of G and, using the assumption that $\left|x-p_0\right| \gg \lambda$, simplify the integral in (2.19). Similar calculations as for G_0 yield

$$\frac{\partial G(x;p_0,p_1)}{\partial n} = \cos(n,p_0)\left(\mathrm{i}k - \frac{1}{\left|x-p_0\right|}\right)\frac{\mathrm{e}^{\mathrm{i}k\left|x-p_0\right|}}{\left|x-p_0\right|} - \cos(n,p_1)\left(\mathrm{i}k - \frac{1}{\left|x-p_1\right|}\right)\frac{\mathrm{e}^{\mathrm{i}k\left|x-p_1\right|}}{\left|x-p_1\right|}$$

where on $\mathcal{A}$ we have $x - p_0 = p_1 - x$. This implies

$$\cos(n,p_0) = -\cos(n,p_1)$$

and with $\left|x-p_0\right| = \left|x-p_1\right|$ on $\mathcal{A}$ we obtain

$$\frac{\partial G(x;p_0,p_1)}{\partial n} = 2\cos(n,p_0)\left(\mathrm{i}k - \frac{1}{\left|x-p_0\right|}\right)\frac{\mathrm{e}^{\mathrm{i}k|x-p_0|}}{\left|x-p_0\right|}.$$

Neglecting the term $1/|x-p_0|$ for $\left|x-p_0\right| \gg \lambda$ yields

$$\frac{\partial G(x;p_0,p_1)}{\partial n} = 2\mathrm{i}k\cos(n,p_0)\frac{\mathrm{e}^{\mathrm{i}k|x-p_0|}}{\left|x-p_0\right|}.$$

Finally, we can write the integral representation (2.19) as

$$U_1(p) := \frac{1}{\mathrm{i}\lambda}\int_{\mathcal{A}} U(x)\frac{\mathrm{e}^{\mathrm{i}k|x-p_0|}}{\left|x-p_0\right|}\cos(n,p_0)\,\mathrm{d}\sigma \tag{2.20}$$

for $\left|x-p_0\right| \gg \lambda$.

2.6. Fresnel Approximation

Next, we express the setup in rectangular coordinates. Based on this, we establish the integral representation we will work with using another approximation. The precise setup is illustrated in Figure 2.4. Note that in the literature the caption of the axes may be reversed ($x_1 \leftrightarrow \xi_1, x_2 \leftrightarrow \xi_2$). However, since the object is considered in spatial domain and the propagated object in Fresnel or Fourier domain, this notation seems more natural. In these coordinates the angle θ is given by $\cos\theta = z/|r_{01}|$ where r_{01} is the vector from p_0 to p_1. Any point p_0 in the object plane possesses the coordinates $p_0 = (x_1, x_2, 0)$ and any point p_1 in the image plane the coordinates $p_1 = (\xi_1, \xi_2, z)$. Plugging this into the representation (2.20) results in

$$U_1(\xi_1,\xi_2) = \frac{z}{\mathrm{i}\lambda}\int_{\mathcal{A}} U(x_1,x_2)\frac{\mathrm{e}^{\mathrm{i}k|r_{01}|}}{|r_{01}|^2}\,\mathrm{d}x_1\,\mathrm{d}x_2$$

where

$$|r_{01}| := \sqrt{z^2 + (\xi_1 - x_1)^2 + (\xi_2 - x_2)^2} = z\sqrt{1 + \left(\frac{\xi_1 - x_1}{z}\right)^2 + \left(\frac{\xi_2 - x_2}{z}\right)^2}.$$

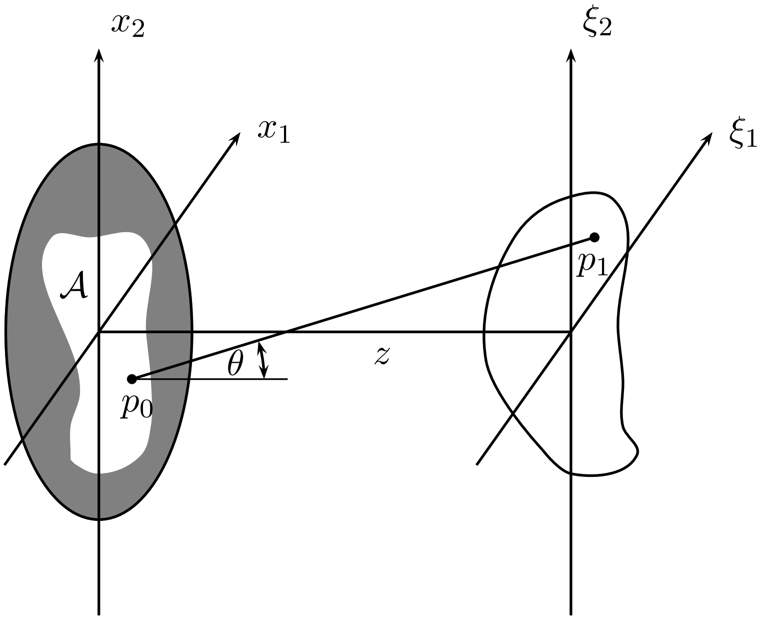


Figure 2.4.: Diffraction setup in rectangular coordinates

For a real number $|b| < 1$ we can express the square-root of $1+b$ by its Taylor expansion

$$\sqrt{1+b} = 1 + \frac{1}{2}b + \mathcal{O}\left(b^2\right).$$

Using only first order terms in the expansion we can approximate $|r_{01}|$ by

$$|r_{01}| \approx z\left[1 + \frac{1}{2}\left(\frac{\xi_1 - x_1}{z}\right)^2 + \frac{1}{2}\left(\frac{\xi_2 - x_2}{z}\right)^2\right]. \tag{2.21}$$

We now apply two different approximations. For $|r_{01}|^2$ in the denominator only the constant order term is taken, i.e., $|r_{01}|^2 \approx z^2$, and hence this yields

$$U_1\left(\xi_1, \xi_2\right) \approx \frac{1}{\mathrm{i}\lambda z} \int_{\mathcal{A}} U(x_1, x_2)\, \mathrm{e}^{\mathrm{i}k|r_{01}|}\, \mathrm{d}x_1\, \mathrm{d}x_2 \tag{2.22}$$

Since errors in the approximation of the exponent have more severe consequences, we apply formula (2.21) to approximate $|r_{01}|$ there. In the exponential term we obtain

$$\begin{aligned}\exp\left(\mathrm{i}k\,|r_{01}|\right) &\approx \exp\left(\mathrm{i}k\left[z + \frac{z}{2}\left(\frac{\xi_1 - x_1}{z}\right)^2 + \frac{z}{2}\left(\frac{\xi_2 - x_2}{z}\right)^2\right]\right)\\ &= \exp\left(\mathrm{i}kz\right)\exp\left(\mathrm{i}k\frac{z}{2}\left[\left(\frac{\xi_1 - x_1}{z}\right)^2 + \left(\frac{\xi_2 - x_2}{z}\right)^2\right]\right)\end{aligned}$$

$$= \exp(\mathrm{i}kz)\exp\left(\frac{\mathrm{i}k}{2z}(\xi_1 - x_1)^2 + (\xi_2 - x_2)^2\right).$$

Plugging this into (2.22) gives

$$U_1(\xi_1, \xi_2) \approx \frac{\mathrm{e}^{\mathrm{i}kz}}{\mathrm{i}\lambda z}\int_{\mathcal{A}} U(x_1, x_2)\exp\left(\mathrm{i}\frac{k}{2z}\left[(\xi_1 - x_1)^2 + (\xi_2 - x_2)^2\right]\right)\mathrm{d}x_1\,\mathrm{d}x_2, \tag{2.23}$$

which is the *Fresnel diffraction integral*. The scope of this approximation is called *near field* or *Fresnel regime*.

Definition 2.6.1 (Fresnel transform). *We denote the* Fresnel transform $\mathcal{D}_\tau : L^2(\mathbb{R}^2) \to L^2(\mathbb{R}^2)$ *of a function* $f \in L^2(\mathbb{R}^2)$ *by* $\mathcal{D}_\tau[f]$ *where* $x = (x_1, x_2)$, $\xi = (\xi_1, \xi_2)$. *We have the representation*

$$\tilde{f}_\tau = \mathcal{D}_\tau[f](\xi) = \frac{1}{\tau}\int_{\mathbb{R}^2} f(x)\exp\left(\frac{\mathrm{i}\pi}{\tau^2}|x - \xi|^2\right)\mathrm{d}x \tag{2.24}$$

where $k = {}^{2\pi}/_{\lambda}$ *and* $\tau = \sqrt{\lambda z}$.

Note that our definition of the Fresnel transform is equivalent to the Fresnel diffraction integral up to the multiplicative constant $-\mathrm{i}\mathrm{e}^{\mathrm{i}kz}$. For a discussion on the accuracy of the approximations we refer to [42, p. 68].

2.6.1. Equivalent Representations of the Fresnel Transform

For numerical purposes it may be convenient to express the Fresnel transform in terms of the Fourier transform. Therefore we first define the Fourier transform.

Definition 2.6.2 (Fourier transform). *Let* $f \in L^2(\mathbb{R}^d)$. *We define the* Fourier transform $\hat{f}$ *of* f *by*

$$\hat{f}(\omega) := \mathcal{F}[f](\omega) = \int_{\mathbb{R}^n} f(x)\,\mathrm{e}^{-2\pi\mathrm{i}\langle x,\omega\rangle}\,\mathrm{d}x.$$

Remark 2.6.3. Classically, the Fourier transform is defined for integrable functions but can be extended for square integrable functions. We refer to [80, Section 2.2.2] for a proof. The *inverse Fourier transform of $\hat{f}$*, denoted by f, is given by

$$f(x) = \mathcal{F}^{-1}[\hat{f}](x) = \int_{\mathbb{R}^n} \hat{f}(\omega) \mathrm{e}^{2\pi \mathrm{i}\langle x,\omega\rangle}\, \mathrm{d}\omega.$$

A proof of this result is given in [80, Theorem 2.1]. ○

In order to obtain a better understanding of the Fresnel transform, we will derive different representations of it. Therefore we first expand the term in the exponential

$$(\xi_1 - x_1)^2 + (\xi_2 - x_2)^2 = \xi_1^2 + x_1^2 + \xi_2^2 + x_2^2 - 2x_1\xi_1 - 2\xi_2 x_2$$

or respectively

$$\mathrm{e}^{(\xi_1-x_1)^2+(\xi_2-x_2)^2} = \mathrm{e}^{\xi_1^2+\xi_2^2}\mathrm{e}^{x_1^2+x_2^2}\mathrm{e}^{-2(x_1\xi_1+x_2\xi_2)}.$$

We can now factor out the first factor and rewrite the Fresnel transform in terms of the Fourier transform of a modulated function

$$\mathcal{D}_\tau\left[f\right](\xi) = \frac{\mathrm{e}^{\frac{\mathrm{i}k}{2z}\left(\xi_1^2+\xi_2^2\right)}}{\tau} \int_{\mathbb{R}^2} \left[f(x)\, \exp\left[\frac{\mathrm{i}k}{2z}\left(x_1^2 + x_2^2\right)\right]\right] \exp\left(-\frac{2\pi \mathrm{i}}{\tau^2}\left\langle x,\xi\right\rangle\right) \mathrm{d}x. \tag{2.25}$$

To see this, we use $\mathrm{i}k/2z = 2\pi\mathrm{i}/\lambda z = \mathrm{i}\pi/\tau^2$ and $x_1^2 + x_2^2 = |x|^2$ and obtain

$$\mathcal{D}_\tau\left[f\right](\xi) = \frac{\mathrm{e}^{\frac{\mathrm{i}\pi|\xi|^2}{\tau^2}}}{\tau} \int_{\mathbb{R}^n} \left[f(x)\, \mathrm{e}^{\frac{\mathrm{i}\pi|x|^2}{\tau^2}}\right] \mathrm{e}^{\frac{-2\pi\mathrm{i}\langle x,\xi\rangle}{\tau^2}}\, \mathrm{d}x.$$

Substituting $y = x/\tau^2$ yields $\mathrm{d}x = \tau^2\, \mathrm{d}y$ and hence

$$\mathcal{D}_\tau\left[f\right](\xi) = \tau \mathrm{e}^{\frac{\mathrm{i}\pi|\xi|^2}{\tau^2}} \int_{\mathbb{R}^2} \left[f(\tau^2 y)\, \mathrm{e}^{\mathrm{i}\pi\tau^2|y|^2}\right] \mathrm{e}^{-2\pi\mathrm{i}\langle y,\xi\rangle}\, \mathrm{d}y.$$

We can now write (2.25) using the Fourier transform as

$$\mathcal{D}_\tau\left[f\right](\xi) = \tau \mathrm{e}^{\frac{\mathrm{i}\pi|\xi|^2}{\tau^2}} \mathcal{F}\Big(f(\tau^2 y)\, \mathrm{e}^{\mathrm{i}\pi\tau^2|y|^2}\Big). \tag{2.26}$$

In the presentation of the properties of the Fresnel diffraction integral we will follow [66, Chapter 2]. However, these properties are also presented in [42, Chapter 5] with references to the original work. We start by noting that we can write (2.24) in the form

$$\mathcal{D}_\tau\left[f\right](\xi) = (f \star K_\tau)(\xi) = \int_{\mathbb{R}^2} f(x)\, K_\tau(x-\xi)\,\mathrm{d}x$$

where

$$K_\tau(x) = \frac{1}{\tau}\exp\left(\frac{\mathrm{i}\pi}{\tau^2}\,|x|^2\right).$$

Furthermore, we note that the kernel K_τ is separable with

$$\frac{1}{\tau}\exp\left(\frac{\mathrm{i}\pi}{\tau^2}\,|x|^2\right) = \frac{1}{\tau}\exp\left(\frac{\mathrm{i}\pi}{\tau^2}\left[x_1^2 + x_2^2\right]\right) = \underbrace{\frac{1}{\sqrt{\tau}}\exp\left(\frac{\mathrm{i}\pi}{\tau^2}x_1^2\right)}_{=:k_{\sqrt{\tau}}(x_1)}\underbrace{\frac{1}{\sqrt{\tau}}\exp\left(\frac{\mathrm{i}\pi}{\tau^2}x_2^2\right)}_{=:k_{\sqrt{\tau}}(x_2)}.$$

2.6.2. Properties of the Fresnel Transform

Most important in our setting is how to compute the inverse transform $\mathcal{D}_\tau^{-1}$.

Proposition 2.6.4 (Inverse Fresnel transform). *The inverse Fresnel transform of a function $\tilde{f} \in L^2(\mathbb{R}^2)$ is given by*

$$\mathcal{D}_\tau^{-1}\left[\tilde{f}\right](x) = \frac{1}{\tau}\int_{\mathbb{R}^2} \tilde{f}(\xi)\exp\left(-\frac{\mathrm{i}\pi}{\tau^2}\,|x-\xi|^2\right)\mathrm{d}\xi. \tag{2.27}$$

Proof. We prove this for the 1D-case where $\frac{1}{\tau}$ is replaced by $\frac{1}{\sqrt{\tau}}$, $x = (x_1, x_2)$ by $x = x_1$, $\xi = (\xi_1, \xi_2)$ by $\xi = \xi_1$, and $f(x) = f(x_1, x_2)$ by $f(x_1)$ since the kernel is separable. We simply apply the forward and the inverse transform and show that we retrieve the original function. In order to make things more readable, we write y instead of x for the variable that appears for the inverse transform.

$$\begin{aligned}
\mathcal{D}_{\sqrt{\tau}}^{-1}\left[\mathcal{D}_{\sqrt{\tau}}\left[f\right]\right](y) &= \frac{1}{\tau}\int_{\mathbb{R}}\int_{\mathbb{R}} f(x)\mathrm{e}^{\mathrm{i}\pi(x-\xi)^2/\tau}\,\mathrm{d}x\ \mathrm{e}^{-\mathrm{i}\pi(\xi-y)/\tau}\,\mathrm{d}\xi \\
&= \frac{1}{\tau}\int_{\mathbb{R}}\int_{\mathbb{R}} f(x)\mathrm{e}^{\mathrm{i}\pi x^2/\tau}\mathrm{e}^{-2\mathrm{i}\pi x\xi/\tau}\,\underbrace{\mathrm{e}^{\mathrm{i}\pi\xi^2/\tau}\mathrm{e}^{-\mathrm{i}\pi\xi^2/\tau}}_{=1}\,\mathrm{e}^{-\mathrm{i}\pi y^2/\tau}\mathrm{e}^{2\mathrm{i}\pi y\xi/\tau}\,\mathrm{d}x\,\mathrm{d}\xi \\
&= \frac{1}{\tau}\int_{\mathbb{R}}\left(\int_{\mathbb{R}}\left[f(x)\mathrm{e}^{\mathrm{i}\pi x^2/\tau}\right]\mathrm{e}^{-2\mathrm{i}\pi x\xi/\tau}\,\mathrm{d}x\right)\mathrm{e}^{2\mathrm{i}\pi\xi y/\tau}\mathrm{e}^{-\mathrm{i}\pi y^2/\tau}\,\mathrm{d}\xi
\end{aligned}$$

$$\begin{aligned}
&\overset{\tau\xi'=\xi}{=} \frac{1}{\tau}\int_{\mathbb{R}}\left(\int_{\mathbb{R}}\left[f(x)\mathrm{e}^{\mathrm{i}\pi x^2/\tau}\right]\mathrm{e}^{-2\mathrm{i}\pi x\xi}\,\mathrm{d}x\right)\mathrm{e}^{2\mathrm{i}\pi\xi y}\mathrm{e}^{-\mathrm{i}\pi y^2/\tau}\,\tau\,\mathrm{d}\xi\\
&= \mathrm{e}^{-\mathrm{i}\pi y^2/\tau}\int_{\mathbb{R}}\mathcal{F}\left\{f(x)\mathrm{e}^{\mathrm{i}\pi x^2/\tau}\right\}\mathrm{e}^{2\pi\mathrm{i}\xi y}\,\mathrm{d}\xi\\
&= \mathrm{e}^{-\mathrm{i}\pi y^2/\tau}\mathcal{F}^{-1}\left[\mathcal{F}\left\{f(x)\mathrm{e}^{\mathrm{i}\pi x^2/\tau}\right\}\right] = \mathrm{e}^{-\mathrm{i}\pi y^2/\tau}f(y)\mathrm{e}^{\mathrm{i}\pi y^2/\tau} = f(y).
\end{aligned}$$

Hence, $\mathcal{D}^{-1}_{\sqrt{\tau}}\mathcal{D}_{\sqrt{\tau}} = \mathrm{Id}$. The 2D-case follows equivalently. □

Proposition 2.6.5 (Properties). *Let* $f, \tilde{f}_\tau \in L^2(\mathbb{R}^2)$ *with* $\tilde{f}_\tau(\xi) := \mathcal{D}_\tau[f](\xi)$. *Then the following properties hold:*

(i) Duality: $\overline{f(x)} = \mathcal{D}_\tau\left[\overline{\mathcal{D}_\tau[f]}\right](x)$

(ii) Translation: $\mathcal{D}_\tau[f(\cdot - t)](\xi) = \tilde{f}_\tau(\xi - t)$

(iii) Dilation: $\mathcal{D}_\tau\left[f\left(\frac{\cdot}{s}\right)\right](\xi) = \tilde{f}_{\tau/s}(\xi/s)$

Proof. (i) follows from the definition of $\mathcal{D}_\tau$ and its inverse, (ii) follows since the Fresnel transform is a convolution, and (iii) follows by calculation. □

Remark 2.6.6. The characterization of the Fresnel transform as a convolution and the explicit expression for the Fourier transform of the kernel simplifies the computation of the Fresnel transform numerically. Therefore, the numerical effort reduces to the computation of fast Fourier transforms and point-wise multiplications. However, in certain situations it may be important how to sample the kernel to avoid artifacts. For further details, we refer to [105].

2.7. Fraunhofer Diffraction

In the last sections we derived an approximate representation of the solution in the near field. In this section we will, based on these facts, derive an even simpler representation that will turn out to be valid in the *far field* or *Fraunhofer regime*.

As observed in (2.25) the Fresnel diffraction integral can be rewritten in terms of the Fourier transform of the modulated function

$$f(x)\,\exp\left[\frac{\mathrm{i}k}{2z}\left(x_1^2 + x_2^2\right)\right].$$

If we further assume that

$$z \gg \frac{k\left(x_1^2 + x_2^2\right)_{\max}}{2} = \frac{k \operatorname{diam}\left(\operatorname{supp}(f)\right)^2}{2}$$

where $\operatorname{supp}(f) = \overline{\{x = (x_1, x_2) \in \mathbb{R}^2 \mid f(x) \neq 0\}}$ and $\operatorname{diam}(X) = \sup\left\{\left|x - y\right| : x, y \in X\right\}$ it follows that the quadratic phase factor almost vanishes and hence

$$f(x) \exp\left[\frac{\mathrm{i}k}{2z}\left(x_1^2 + x_2^2\right)\right] \approx f(x).$$

Therefore the Fraunhofer approximation is given (up to multiplicative phase factors) by the Fourier transform with scaled frequencies of f, i.e.,

$$\mathcal{D}_\tau^{\mathrm{Fra}}\left[f\right](\xi) = \frac{\mathrm{e}^{\frac{\mathrm{i}k}{2z}\left(\xi_1^2+\xi_2^2\right)}}{\tau} \int_{\mathbb{R}^2} f(x) \exp\left(-\frac{2\pi\mathrm{i}}{\tau^2}\left\langle x, \xi\right\rangle\right) \mathrm{d}x.$$

A rescaling of the model (cf. [77, Section 3.1.4]) leads to the new, idealized model that the measurements taken are the point-wise moduli of the Fourier transform of that function.

2.8. The Imaging Model

In the last sections we derived an expression for the propagation of the wave in free space from the object plane to the imaging plane. In applications like coherent x-ray imaging, the measurements are taken typically on some sort of digital image sensor. The measurements that can be taken are the point-wise moduli of the propagated wave that usually are complex numbers in each point.

Therefore, our measurements can be written as

$$m(\xi) = \left|\mathcal{D}_\tau\left[f\right](\xi)\right| \quad \text{or} \quad m(\xi) = \left|\mathcal{D}_\tau^{\mathrm{Fra}}\left[f\right](\xi)\right|,$$

respectively. In practice, we will discretize f and $\mathcal{D}_\tau$ on a rectangular domain with a regular grid. The discretized image f will be a matrix $f \in \mathbb{C}^{d_1 \times d_2}$ and we consider the discrete transform $\mathcal{D}_\tau : \mathbb{C}^{d_1 \times d_2} \to \mathbb{C}^{d_1 \times d_2}$. Then the matrices fulfilling the measurements can be written as

$$M = \left\{f \in \mathbb{C}^{d_1 \times d_2} \mid \left|\mathcal{D}_\tau\left[f\right](\xi)\right| = m(\xi),\ \forall \xi \in \Omega\right\}$$

where $\Omega = \{1, \dots, d_1\} \times \{1, \dots, d_2\}$ and the modulus is taken element-wise. The task is to recover f from those measurements. This problem is, in general, ill-posed, since all matrices with point-wise same modulus as $m(\xi)$ fulfill the measurements independently of their phase. A typical approach to overcome this difficulty is to regularize the problem using different types of *a priori* information. Furthermore, the measured data may not be exact. Thus a solution that both is in M and fulfills the additional a priori information may not exist. In this case one may solve a relaxed problem where the distance of f to the set M is small and not necessarily $f \in M$. In Chapter 4 we will introduce iterative phase retrieval methods as a tool to solve such problems. Furthermore, we will study different types of relaxations concerning the set M, the modulus function $|\cdot|$, and the projection algorithms. Before that, we discuss a suitable model for experimental phase retrieval data.

2.9. Experimental Setup

Since experimental details are beyond the scope of this work, we only briefly mention the idea behind the image formation process as derived in, e.g., [91]. The Fresnel transform derived in the previous sections is only valid for the wave propagation in *free space*, i.e. in the absence of matter. However, in the object plane the wave hits a physical object, namely the specimen. It can be shown that in the presence of matter, the wave U_ω fulfills the equation

$$\left(\Delta^2 + k^2 n_\omega^2(x)\right) U_\omega(x) = 0$$

where $x = (x_1, x_2, z)$ and $n_\omega(x) = \sqrt{\varepsilon(x)/\varepsilon_0}$ is called *frequency-dependent refractive index* which can be written as

$$n_\omega(x) = 1 - \delta(x) + \mathrm{i}\beta(x),$$

cf. [91], and where $\delta(x) > 0$ and $\beta(x) > 0$ may also depend on the wavelength. In some applications, those may be of the order of 10^{-5} to 10^{-8} which would legitimize the approximation of the squared refractive index

$$n_\omega^2(x) \approx 1 - 2\delta(x) + 2\mathrm{i}\beta(x).$$

It can be shown, cf. [92], that $\delta(x)$ relates to an induced phase shift while $\beta(x)$ reflects the absorption in the material. The so called *projection approximation* is then given by

$$U_\omega(x_1, x_2, 0) \approx \mathrm{e}^{\mathrm{i}k\tau} U_\omega(x_1, x_2, -\tau) \exp\left(\int_{-\tau}^{0} \delta(x_1, x_2, z) - \mathrm{i}\beta(x_1, x_2, z)\,\mathrm{d}z\right) \tag{2.28}$$

where $U_\omega(x, y, -\tau)$ is the incident wave and $U_\omega(x, y, 0)$ the exit wave in the object plane after passing an object of thickness τ. This can be shown to be a reasonable approximation in many physical applications, cf. [91, 92]. Defining the *illumination function*

$$P(x_1, x_2) := U_\omega(x_1, x_2, -\tau)$$

and the *object transmission function*

$$O(x_1, x_2) := \exp\left(-k\int_{-\tau}^{0} \beta(x_1, x_2, z)\,\mathrm{d}z - \mathrm{i}k\int_{-\tau}^{0} \delta(x_1, x_2, z)\,\mathrm{d}z\right),$$

formula (2.28) becomes

$$U_\omega(x_1, x_2, 0) \approx \mathrm{e}^{\mathrm{i}k\tau} P(x_1, x_2) O(x_1, x_2).$$

While most functions arriving from an experimental setup will be of this form, our discussion will be as general as possible. Hence, we will not employ the knowledge about this special form specifically. The results therefore will be valid for a more general class of functions. Nonetheless, since the central theme of this thesis is the recovery of information about the specimens from phase retrieval data, this type of function is of course of special interest. Before we study sparsifying transforms in the next chapter, the following section briefly introduces the common noise model that appears in the aforementioned applications.

2.10. The Noise Model

One important ingredient when dealing with experimental data is the noise model. This section treats the derivation of the Poisson noise model used in phase retrieval. In our reasoning we follow [42, Chapter 3].

The noise model is based on on three fundamental hypotheses, cf. [41]. With a slight abuse of notation, in this section τ denotes a real-number describing a time interval.

Fact 2.10.1. *I. For an arbitrarily small time interval Δt, the probability of a* single impulse *occurring in the time interval $[t, t + \Delta t]$ is equal to the product of Δt and a real-nonnegative function $\lambda(t)$, i.e.,*

$$\mathbb{P}\left(1, t, t + \Delta t)\right) = \lambda(t)\Delta t. \tag{2.29}$$

II. For an arbitrarily small time interval $[t, t + \Delta t]$, the probability of more than one impulse *occurring in the time interval $[t, t + \Delta]$ is negligibly small, i.e., there are no multiple events,*

$$\mathbb{P}\left(0, t, t + \Delta t\right) = 1 - \lambda(t)\Delta t. \tag{2.30}$$

III. The impulses in non-overlapping *time intervals are statistically independent.*

We are interested in the probability that K impulses occur in a certain time interval $t + \tau$. Hence, we assume that $\mathbb{P}$ is differentiable and derive an ordinary differential equation whose solution will be the noise model. Consider the time interval $[t, t + \tau + \Delta\tau]$ and the probability that K impulses occur. Based on Fact 2.10.1, there are only two possibilities, since we excluded multiple events:

1. There are K impulses occurring in the time interval $[t, t + \tau]$ and no impulses occurring in the time interval $[t + \tau, t + \tau + \Delta\tau]$.

2. There are $K - 1$ impulses occurring in the time interval $[t, t + \tau]$ and one impulse occurring in the time interval $[t + \tau, t + \tau + \Delta\tau]$.

Applying (2.29) and (2.30) to this observation yields

$$\begin{aligned}\mathbb{P}(K, t + \tau + \Delta\tau) &= \mathbb{P}(K, t, t + \tau) \cdot \mathbb{P}(0, t + \tau, t + \tau + \Delta\tau) \\ &\qquad + \mathbb{P}(K - 1, t, t + \tau) \cdot \mathbb{P}(1, t, t + \tau, t + \tau + \Delta\tau) \\ &= \mathbb{P}(K, t, t + \tau) \cdot (1 - \lambda(t + \tau)\Delta\tau) + \mathbb{P}(K - 1, t, t + \tau) \cdot \lambda(t + \tau)\Delta\tau.\end{aligned}$$

After rearranging the terms and dividing by $\Delta\tau$ we obtain

$$\frac{\mathbb{P}(K, t, t+\tau+\Delta\tau) - \mathbb{P}(K, t, t+\tau)}{\Delta\tau} = \lambda(t+\tau)\left[\mathbb{P}(K-1, t, t+\tau) - \mathbb{P}(K, t, t+\tau)\right].$$

Assuming differentiability we arrive at the differential equation

$$\frac{\mathrm{d}\mathbb{P}(K, t, t+\tau)}{\mathrm{d}\tau} = \lambda(t+\tau)\left[\mathbb{P}(K-1, t, t+\tau) - \mathbb{P}(K, t, t+\tau)\right]. \tag{2.31}$$

The solution for (2.31) is given by

$$\mathbb{P}(K, t_1, t_2) = \frac{\left[\int_{t_1}^{t_2} \lambda(t)\,\mathrm{d}t\right]^K}{K!} \exp\left[-\int_{t_1}^{t_2} \lambda(t)\,\mathrm{d}t\right],$$

see Fact A.1.3 for the proof.

Our interpretation of this result is as follows: The measurements are taken using a digital sensor that measures the number of photons in each individual pixel. The true image to be measured would be the expectation value of the underlying probability distribution. Since we are only able to measure a finite time interval, the data obtained is intrinsically incomplete.

3. Frames

In this chapter we introduce the multivariate wavelet-based transforms that we will utilize in the phase retrieval problem as sparsifying transforms. In order to understand shearlets, we begin by introducing the one-dimensional wavelet transform and explain the digital realization using wavelet filter banks. In order to be able to introduce the discrete shearlet transform, we sketch how to perform a fast wavelet transform in two-dimensions using tensor-product wavelets.

3.1. Wavelets

There are several ways to introduce wavelets and a lot of literature on that topic, see, e.g., [25, 80, 84, 94]. We will follow the path of [80], but take a few shortcuts when appropriate.

Formally, a *wavelet* is a normalized function $\psi \in L^2(\mathbb{R})$

$$\int_{-\infty}^{+\infty} \left|\psi(t)\right|^2 \,\mathrm{d}t = 1$$

with zero-average

$$\int_{-\infty}^{+\infty} \psi(t) \,\mathrm{d}t = 0.$$

The *wavelet transform* of a function $f \in L^2(\mathbb{R})$ at *time* u and on *scale* s is defined as

$$Wf(s,u) := \langle f, \psi_{s,u} \rangle = \int_{-\infty}^{+\infty} f(t) \,\frac{1}{\sqrt{s}}\, \psi^* \left(\frac{t-u}{s}\right) \mathrm{d}t$$

where

$$\psi_{s,u}(t) := \frac{1}{\sqrt{s}}\, \psi\left(\frac{t-u}{s}\right).$$

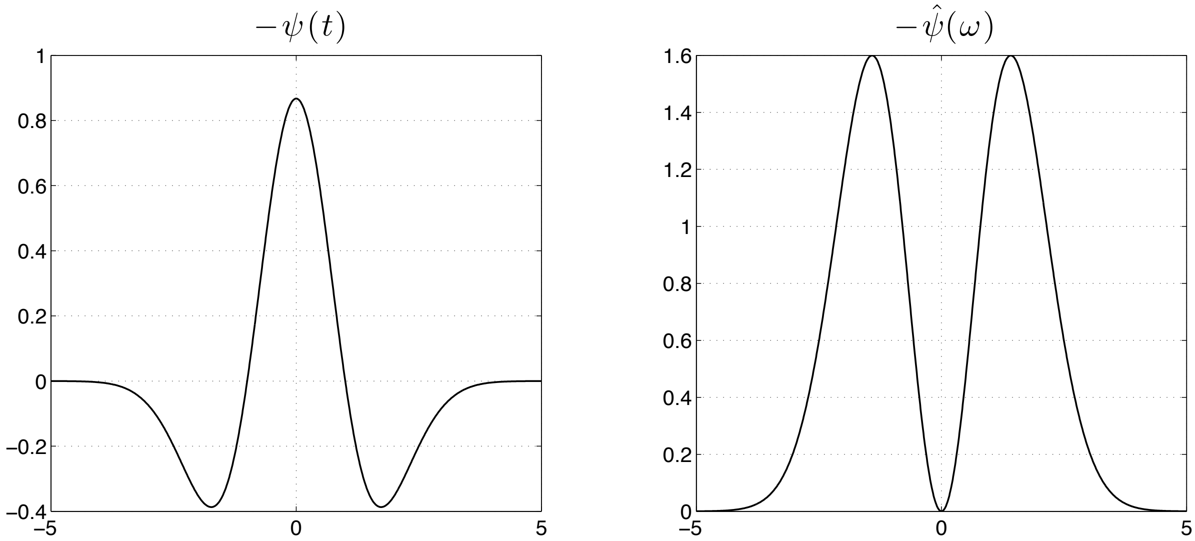


Figure 3.1.: Mexican hat wavelet for $\sigma = 1$ and its Fourier transform

A well-known example for a wavelet is the (normalized) *Mexican hat wavelet*, it is given by

$$\psi(t) = \frac{2}{\pi^{1/4}\sqrt{3\sigma}}\left(\frac{t^2}{\sigma^2} - 1\right)\exp\left(-\frac{t^2}{2\sigma^2}\right)$$

with Fourier transform

$$\widehat{\psi}(\omega) = \frac{-\sqrt{8}\sigma^{5/2}\pi^{1/4}}{\sqrt{3}}\omega^2 \exp\left(-\frac{\sigma^2\omega^2}{2}\right),$$

see Figure 3.1. The *admissibility condition*

$$0 < c_\psi := 2\pi \int_{\mathbb{R}} \frac{\left|\hat{\psi}(\omega)\right|^2}{|\omega|}\,\mathrm{d}\omega < \infty$$

ensures that every real-valued function $f \in L^2(\mathbb{R})$ has a wavelet representation

$$f(t) = \frac{1}{c_\psi}\int_0^{+\infty}\int_{-\infty}^{+\infty} Wf(s,u)\frac{1}{\sqrt{s}}\psi\left(\frac{t-u}{s}\right)\mathrm{d}u\frac{\mathrm{d}s}{s^2},$$

cf. [80, Theorem 4.3] which was originally proven in [15] and later independently in [43].

3.1.1. Smoothness and Vanishing Moments

In order to measure properties of the function f on which we perform the wavelet transform, the wavelet itself needs to satisfy certain properties such as smoothness and vanishing moments.

Definition 3.1.1 (Vanishing Moments). *The wavelet* $\psi \in L^2(\mathbb{R})$ *possesses* n vanishing moments *if*

$$\int_{\mathbb{R}} t^k \psi(t) \,\mathrm{d}t = 0$$

for all $0 \leq k \leq n$. *In other words,* ψ *is orthogonal to all polynomials of degree smaller than* n.

Definition 3.1.2 (Lipschitz Regularity, Definition 6.1 from [80]). *A function* f *is*

- pointwise Lipschitz regular of order $\alpha \geq 0$ at v *if there exists a constant* $K > 0$ *and a polynomial* p_v *of degree* $m = \lfloor \alpha \rfloor$ *such that*

$$\left|f(t) - p_v(t)\right| \leq K \left|t - v\right|^{\alpha} \qquad \forall\, t \in \mathbb{R}. \tag{3.1}$$

- uniformly Lipschitz regular of order α over $[a, b]$ *if it satisfies* (3.1) *for all* $v \in [a, b]$ *where* K *is independent of* v.

The Lipschitz regularity of f at v or over $[a, b]$ *is the supremum over all* α *such that* f *is Lipschitz* α.

Note that if f is uniformly Lipschitz regular of order $\alpha > m$ in a neighborhood of v then it is m times continuously differentiable in that neighborhood. The following theorem [80, Theorem 6.1] connects the smoothness of f with the decay of its Fourier transform.

Theorem 3.1.3 (Theorem 6.1 from [80]). *A function* f *is bounded and uniformly Lipschitz regular of order* α *over* $\mathbb{R}$ *if*

$$\int_{\mathbb{R}} \left|\widehat{f}(\omega)\right| (1 + \left|\omega\right|^{\alpha}) \,\mathrm{d}\omega < \infty.$$

Suppose now that f is uniformly Lipschitz regular of order α. Then by (3.1) we have

$$f(t) = p_v(t) + \varepsilon_v(t)$$

where $|\varepsilon_v(t)| \leq K|t-v|^{\alpha}$ and p_v is a polynomial of degree $m \leq \alpha$. If ψ has $n > \alpha$ vanishing moments, then it is orthogonal to all polynomials of degree smaller than or equal to m. Therefore, we have

$$\begin{aligned} Wf(s,u) &= \int_{\mathbb{R}} f(t)\psi\left(\frac{t-u}{s}\right) \mathrm{d}t \\ &= \int_{\mathbb{R}} \left(p_v(t) + \varepsilon_v(t)\right) \psi\left(\frac{t-u}{s}\right) \mathrm{d}t \\ &= \underbrace{\int_{\mathbb{R}} p_v(t)\psi\left(\frac{t-u}{s}\right) \mathrm{d}t}_{=0} + \int_{\mathbb{R}} \varepsilon_v(t)\psi\left(\frac{t-u}{s}\right) \mathrm{d}t \\ &= \int_{\mathbb{R}} \varepsilon_v(t)\psi\left(\frac{t-u}{s}\right) \mathrm{d}t \\ &= W\varepsilon_v(s,u). \end{aligned}$$

Therefore, the wavelet transform (using wavelets with a sufficient amount of vanishing moments) can measure the degree of the singularity of f.

3.1.2. Discrete Wavelets

In order to be able to implement the wavelet transform on a computer it is indispensable to discretize it. This is achieved by sampling the translation and scaling parameters in a suitable manner. In order to allow for a stable reconstruction, this sampling must be chosen such that the generated wavelets $\left\{\psi_{j,n}\right\}_{(j,n)\in\Lambda}$ form a frame for $L^2(\mathbb{R})$. Necessary and sufficient conditions on ψ, j, and n to achieve this requirement are given in [25, Proposition 3.3.2] as well as estimates on the frame bounds. Further it can be shown that if we sample only the scaling parameter on a dyadic grid $\left\{2^j\right\}_{j\in\mathbb{Z}}$ and there are $A, B > 0$ such that for all $\omega \in \mathbb{R} \setminus \{0\}$

$$A \leq \sum_{j=-\infty}^{+\infty} \left|\widehat{\psi}\left(2^j\omega\right)\right|^2 \leq B$$

then the resulting wavelets form a frame for $L^2(\mathbb{R})$, see [80, Theorem 5.11] and Definition 3.1.6. It is even possible to construct orthonormal bases of wavelets for $L^2(\mathbb{R})$ of the form

$$\left\{\psi_{j,n}(t) = \frac{1}{\sqrt{2^j}}\psi\left(\frac{t-2^jn}{2^j}\right)\right\}_{(j,n)\in\mathbb{Z}^2}.$$

In order to implement a fast wavelet transform using filter banks, the concept of *multiresolution analysis* is crucial. Therefore we need the notion of a Riesz basis that we introduce first. Since the term *frame* is related to the concept of Riesz bases, we introduce this definition here as well. We will use frames for the construction of shearlets in the subsequent sections.

Definition 3.1.4 (Riesz basis, Definition from [111]). *A sequence $\{f_n\}$ in a separable Hilbert space $\mathcal{H}$ is a* Riesz basis *if it can be obtained from an orthonormal basis by an invertible, bounded linear operator U, i.e. for all f_n it holds that*

$$f_n = Ue_n$$

where $U : \mathcal{H} \to \mathcal{H}$ and $\{e_n\}$ is an orthonormal basis for $\mathcal{H}$.

We will, however, work with a definition which can be shown to be equivalent to the one above, cf. [111, Theorem 9].

Theorem 3.1.5 (Riesz basis, cf. Theorem 9 from [111]). *A sequence $\{f_n\}$ in a separable Hilbert space $\mathcal{H}$ is a* Riesz basis *if and only if it is complete in $\mathcal{H}$ and there exist constants $0 < A \leq B < \infty$ such that for all $c \in \ell_2$ it holds that*

$$A\,||c||^2_{\ell_2} \leq \left\|\sum_{i\in\mathbb{Z}} c_i f_i\right\|^2_{\mathcal{H}} \leq B\,||c||^2_{\ell_2}\,. \tag{3.2}$$

Definition 3.1.6 (Frame, Definition 2.3 from [21]). *Let $\mathcal{H}$ be a Hilbert space. Then $\{f_n\}_{n\in\mathbb{Z}}$ is* a frame for *$\mathcal{H}$ if there exist constants $0 < A \leq B < \infty$ such that for all $x \in \mathcal{H}$ it* *holds that*

$$A \|x\|_{\mathcal{H}}^2 \leq \sum_{n \in \mathbb{Z}} \left| \langle x, f_n \rangle \right|^2 \leq B \|x\|_{\mathcal{H}}^2 .$$

The frame is called tight *if $A = B$ and* Parseval frame *if $A = B = 1$.*

We now introduce the multiresolution analysis and the corresponding approximation spaces. This concept was first introduced in [79] and [84].

Definition 3.1.7 (Definition 7.1 from [80]). *Let $V_j \subset L^2(\mathbb{R})$ for $j \in \mathbb{Z}$ be closed subspaces such that*

$$\forall (j,k) \in \mathbb{Z}^2 : \quad f(t) \in V_j \Longleftrightarrow f\left(t - 2^j k\right) \in V_j, \tag{3.3}$$

$$\forall j \in \mathbb{Z} : \quad V_{j+1} \subset V_j, \tag{3.4}$$

$$\forall j \in \mathbb{Z} : \quad f(t) \in V_j \Longleftrightarrow f\left(\frac{t}{2}\right) \in V_{j+1}, \tag{3.5}$$

$$\lim_{j \to +\infty} V_j = \bigcap_{j=-\infty}^{+\infty} V_j = \{0\}, \tag{3.6}$$

$$\lim_{j \to -\infty} V_j = \mathrm{cl}\left(\bigcup_{j=-\infty}^{+\infty} V_j \right) = L^2(\mathbb{R}). \tag{3.7}$$

$$\textit{There exists } \theta \textit{ such that } \{\theta(t-n)\}_{n \in \mathbb{Z}} \textit{ is a Riesz basis of } V_0. \tag{3.8}$$

The sequence $\left\{V_j\right\}_{j \in \mathbb{Z}}$ is called multiresolution analysis. *We refer to the spaces V_j as* approximation spaces.

For a discussion of these properties, we refer to [80, Chapter 7]. It is shown in [80, Theorem 7.1] that by using a function θ as in (3.8) it is possible to construct an orthonormal basis for each V_j by

$$\left\{ \phi_{j,n}(t) = \frac{1}{\sqrt{2^j}} \phi\left(\frac{t-n}{2^j}\right) \right\}_{(j,n) \in \mathbb{Z}^2} .$$

Here, the *scaling function* ϕ is defined in Fourier domain by

$$\widehat{\phi}(\omega) = \frac{\widehat{\theta}(\omega)}{\left(\sum_{k=-\infty}^{+\infty} \left|\widehat{\theta}(\omega + 2k\pi)\right|^2\right)^{1/2}}$$

where the Riesz basis property guarantees that the *autocorrelation symbol*

$$\sum_{k=-\infty}^{+\infty} \left|\widehat{\theta}(\omega + 2k\pi)\right|^2$$

is bounded away from zero and bounded from above. This orthonormalization trick always works for a given Riesz basis $\{\theta(t-n)\}_{n\in\mathbb{Z}}$ for V_0.

Using Definition 3.1.7 we see that $V_1 \subset V_0$ and $\phi\left(\frac{t}{2}\right) \in V_1$ since $\phi(t) \in V_0$ and therefore $\phi\left(\frac{t}{2}\right) \in V_0$. Since $\left\{\phi(t-n)\right\}_{n\in\mathbb{Z}}$ is an orthonormal basis of V_0 we can represent the dilated scaling function as

$$\frac{1}{\sqrt{2}}\phi\left(\frac{t}{2}\right) = \sum_{n=-\infty}^{+\infty} h\,[n]\,\phi(t-n)$$

where

$$h[n] := \left\langle \frac{1}{\sqrt{2}}\phi\left(\frac{t}{2}\right), \phi(t-n) \right\rangle. \tag{3.9}$$

The interpretation of $h[n]$ as a discrete filter will serve as the cornerstone for the fast wavelet transform using filter banks.[1] The theoretical justification is given in Theorem 3.1.8, cf. [80, Theorem 7.2] which originates from [79, 84]. This theorem shows a one-to-one connection between scaling functions and its corresponding discrete filter. It even demonstrates how to construct a scaling function if an appropriate filter is given.

Theorem 3.1.8 (Theorem 7.2 from [80]). *Given an integrable scaling function* $\phi \in L^2(\mathbb{R})$ *such that* $\left\{\phi(t-n)\right\}_{n\in\mathbb{Z}}$ *is an orthonormal basis of* V_0 *and the corresponding filter as defined*

---

[1] Note that whenever we write the argument in square brackets, i.e. $h[n]$, we mean a discrete filter. In this case $\widehat{h}(\omega)$ is the Fourier *series* of h. For a function $\phi(t)$ we denote by $\widehat{\phi}(\omega)$ the Fourier *transform* of ϕ.

in (3.9)*, the Fourier series* $\hat{h}(\omega) = \sum_{n\in\mathbb{Z}} h[n]\,\mathrm{e}^{-\mathrm{i}\omega n}$ *of* $h[n]$ *satisfies*

$$\left|\widehat{h}(\omega)\right|^2 + \left|\widehat{h}(\omega+\pi)\right|^2 = 2, \qquad \forall\omega \in \mathbb{R} \tag{3.10}$$

$$\widehat{h}(0) = \sqrt{2}. \tag{3.11}$$

On the other hand, given a 2π*-periodic function* $\widehat{h}$ *that is* C^1 *in a neighborhood of* $\omega = 0$ *which satisfies* (3.10)*,* (3.11) *and*

$$\inf_{\omega\in[-\pi/2,\pi/2]} \left|\widehat{h}(\omega)\right| > 0$$

then

$$\widehat{\phi}(\omega) = \prod_{p=1}^{+\infty} \frac{\widehat{h}\,(2^{-p}\omega)}{\sqrt{2}}$$

is the Fourier transform of a scaling function $\phi \in L^2(\mathbb{R})$*.*

Definition 3.1.9. *A discrete filter* $h[n]$ *that satisfies* (3.10) *is called a* conjugate mirror filter.

We have seen that the approximation spaces V_j are subspaces of $L^2(\mathbb{R})$ and for each V_j we have an orthonormal basis. Hence, we can approximate any function $f \in L^2(\mathbb{R})$ by elements from V_j with the orthogonal projection

$$P_{V_j} f = \sum_{n=-\infty}^{+\infty} \left\langle f, \phi_{j,n} \right\rangle \phi_{j,n}.$$

In order to reconstruct a signal from a wavelet transform we have to take care of the *details* that are lost on each scale. Therefore, denote by W_j the orthogonal complement of V_j in V_{j-1}, i.e.,

$$V_{j-1} = V_j \oplus W_j. \tag{3.12}$$

In other words, we have

$$P_{V_{j-1}} f = P_{V_j} f + P_{W_j} f, \tag{3.13}$$

hence, we need to construct orthonormal bases for the spaces W_j as well. Consequently, we will refer to the spaces W_j as *detail spaces*. The following theorem, cf. [79, 84], gives precise instructions on how to construct a wavelet using a scaling function ϕ and the corresponding conjugate mirror filter.

Theorem 3.1.10 (Theorem 7.3 from [80]). *Let ϕ be a scaling function and h the corresponding conjugate mirror filter. Let ψ be the function whose Fourier transform satisfies*

$$\widehat{\psi}(\omega) = \frac{1}{\sqrt{2}}\widehat{g}\left(\frac{\omega}{2}\right)\widehat{\phi}\left(\frac{\omega}{2}\right)$$

with the 2π-periodic function

$$\widehat{g}(\omega) = \mathrm{e}^{-\mathrm{i}\omega}\,\overline{\widehat{h}(\omega + \pi)}.$$

Let us denote

$$\psi_{j,n}(t) = \frac{1}{\sqrt{2^j}}\psi\left(\frac{t - 2^j n}{2^j}\right).$$

For any scale 2^j, $\left\{\psi_{j,n}\right\}_{n\in\mathbb{Z}}$ is an orthonormal basis for W_j. For all scales, $\left\{\psi_{j,n}\right\}_{(j,n)\in\mathbb{Z}^2}$ is an orthonormal basis for $L^2(\mathbb{R})$.

The proof, which we will omit here and just refer to [80, p. 320–323], also shows that $\widehat{g}$ is of the form $\widehat{g}(\omega) = \sum_{n\in\mathbb{Z}} g[n]\mathrm{e}^{-\mathrm{i}\omega n}$ with the Fourier coefficients

$$g[n] = \left\langle \frac{1}{\sqrt{2}}\psi\left(\frac{t}{2}\right), \phi(t - n)\right\rangle$$

and $g[n]$ itself can be written as

$$g[n] = (-1)^{1-n}\,h[1 - n].$$

Therefore, we also have a corresponding filter for the wavelets. The filter $h[n]$ can be seen as a low-pass filter that provides a low-resolution approximation while $g[n]$ complementarily is a high-pass filter that maintains the details.

3.1.3. Filter Banks

The inner products $\left\langle f, \phi_{j,n} \right\rangle$ and $\left\langle f, \psi_{j,n} \right\rangle$ can be realized as convolutions of f with the scaling function ϕ and the wavelet ψ respectively. Filter banks use this property to implement a fast transform that can compute the wavelet transform on each scale with a complexity that is linear in the length of the signal. The orthogonality of both bases in V_j as well in W_j – or the scaling function and wavelets – is an important fact implying conjugate mirror filters. We have seen in (3.12) and (3.13) that we can successively decompose an approximation $P_{V_j} f$ of f into a coarser approximation $P_{V_{j+1}} f$ and a detail function $P_{W_{j+1}} f$. The coefficients in the corresponding expansions are therefore given by

$$a_j[n] := \left\langle f, \phi_{j,n} \right\rangle \quad \text{and} \quad d_j[n] := \left\langle f, \psi_{j,n} \right\rangle \tag{3.14}$$

since $\left\{\phi_{j,n}\right\}_{n\in\mathbb{Z}}$ is an orthonormal basis of V_j and $\left\{\psi_{j,n}\right\}_{n\in\mathbb{Z}}$ is an orthonormal basis of W_j. We denote $\bar{x}[n] := x[-n]$ and the upsampled sequence by

$$\check{x}[n] := \begin{cases} x[n], & \text{for } n = 2p \\ 0, & \text{for } n = 2p+1. \end{cases}$$

The following theorem first appeared in [79] and [84] and states that it is possible to calculate the coefficients $a_j[n], d_j[n]$ using discrete convolutions.

Theorem 3.1.11 (Theorem 7.7 from [80]). *The coefficients in* (3.14) *can be calculated using discrete convolutions with the filter sequences* $h[n]$ *and* $g[n]$*, i.e.,*

$$\begin{aligned} a_j[n] &= \sum_{n=-\infty}^{+\infty} h[n-2p]\, a_j[n] = a_j \star \bar{h}[2p], \\ d_j[n] &= \sum_{n=-\infty}^{+\infty} g[n-2p]\, a_j[n] = a_j \star \bar{g}[2p]. \end{aligned}$$

Furthermore, the original signal can be successively reconstructed by

$$\begin{aligned} a_j[p] &= \sum_{n=-\infty}^{+\infty} h[p-2n]\, a_{j+1}[n] + \sum_{n=-\infty}^{+\infty} g[p-2n]\, d_{j+1}[n] \\ &= \check{a}_{j+1} \star h[p] + \check{d}_{j+1} \star g[p]. \end{aligned}$$

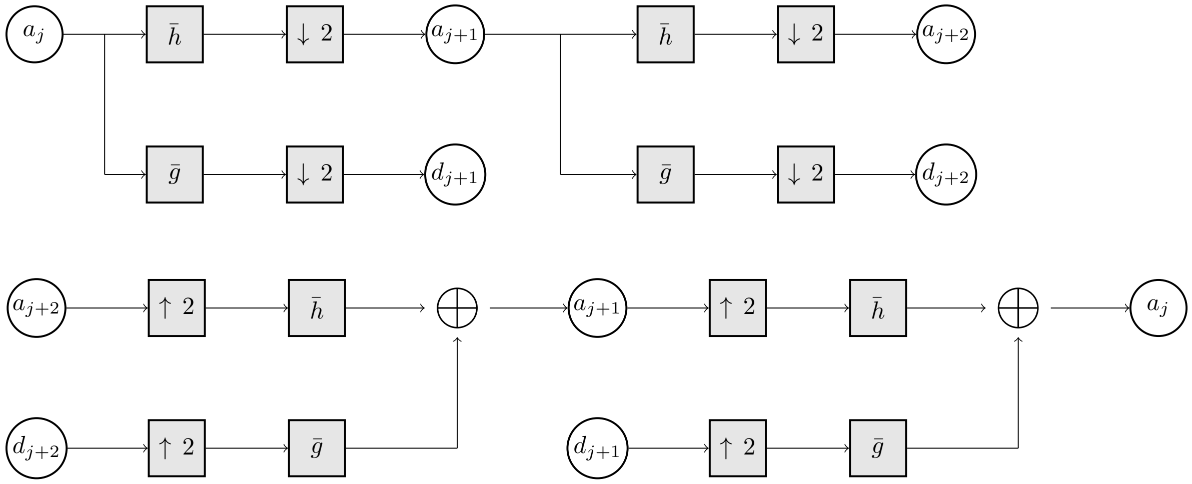


Figure 3.2.: Flowchart illustrating two decomposition steps from the wavelet transform (top) and the inverse transform (bottom).

Remark 3.1.12. In practice, we will work with compactly supported wavelets (and scaling functions) and therefore the discrete convolutions will be finite sums of a certain length. The complexity of the transform depends only linearly on the signal lengths as well as on the lengths of the filter sequences h and g. Daubechies wavelets are known to be optimal in the sense that, given a certain number of vanishing moments, they have the shortest possible support, see [80, Chapter 7.2.3]. ○

Figure 3.2 illustrates two steps of the forward and the inverse transform. The symbol $\downarrow 2$ means a downsampling by a factor of two, i.e., we drop every second entry and obtain a vector of half length. The symbol $\uparrow 2$ means introducing zeros in every second component which we interpreted as an upsampling, i.e., we obtain a vector of twice the length. The symbols $\bar{h}$ and $\bar{g}$ stand for the convolution with the corresponding filter sequences. In each decomposition level, a coefficient a_j is thus decomposed into a coarser approximation a_{j+1} using a low-pass filter and the corresponding detail part d_{j+1} using a high-pass filter. The new approximation coefficient a_{j+1} is then again decomposed into a_{j+2} and d_{j+2}. The wavelet transform therefore consists of an approximation vector a_L and all vectors with detail coefficients $\left(d_j\right)_{1\leq j\leq L}$.

The inverse transform follows accordingly. Here, we first insert zeros, expand the vector and then perform the convolution with low- and high-pass filters. By adding the result, we obtain a_{j+1} from a_{j+2} and d_{j+2}. This procedure can be iterated until we receive our original signal, see Figure 3.2.

3.1.4. Wavelets in Two Dimensions

In order to analyze images, we need to apply a two-dimensional wavelet transform. The most obvious approach is to use tensor-product wavelets. We will see that the construction and the fast wavelet transform can be applied for more than one dimension.

We start by extending the multiresolution analysis to cover the two-dimensional case, cf. [80].

Definition 3.1.13 (Separable multiresolution approximation). *Let* $\left\{V_j\right\}_{j\in\mathbb{Z}}$ *be a multiresolution of* $L^2(\mathbb{R})$*. Then a* separable multiresolution *for* $L^2(\mathbb{R}^2)$ *is given by*

$$\left\{V_j^2\right\}_{j\in\mathbb{Z}} := \left\{V_j \otimes V_j\right\}_{j\in\mathbb{Z}}.$$

Remark 3.1.14. It can be shown that we obtain an orthonormal basis for V_j^2 by

$$\left\{\phi_{j,n}^2(x_1,x_2) = \phi_{j,n_1}(x_1)\phi_{j,n_2}(x_2) = \frac{1}{2^j}\phi\left(\frac{x_1-2^jn_1}{2^j}\right)\phi\left(\frac{x_2-2^jn_2}{2^j}\right)\right\}_{n\in\mathbb{Z}^2},$$

cf. [80, Theorem A.3]. ○

In order to construct two-dimensional wavelets, we introduce the detail spaces W_j^2 as the orthogonal complements to the approximation spaces V_j^2 in V_{j-1}^2, i.e.

$$V_{j-1}^2 = V_j^2 \oplus W_j^2.$$

The following theorem [80, Theorem 7.24] shows how to construct orthonormal bases for the detail spaces.

Theorem 3.1.15 (Theorem 7.24 from [80]). *Let* ϕ *be a scaling function in* $L^2(\mathbb{R})$ *and* $\psi \in L^2(\mathbb{R})$ *the corresponding wavelet generating an orthonormal basis for* $L^2(\mathbb{R})$*. We define*

$$\begin{aligned}
\psi^1(x) &:= \phi(x_1)\psi(x_2),\\
\psi^2(x) &:= \psi(x_1)\phi(x_2),\\
\psi^3(x) &:= \psi(x_1)\psi(x_2),
\end{aligned}$$

and for $1 \le k \le 3$, *we consider*

$$\psi^k_{j,n}(x) := \frac{1}{2^j}\psi^k\left(\frac{x_1 - 2^j n_1}{2^j}, \frac{x_2 - 2^j n_2}{2^j}\right).$$

Then

$$\left\{\psi^1_{j,n}, \psi^2_{j,n}, \psi^3_{j,n}\right\}_{n\in\mathbb{Z}^2}$$

is an orthonormal basis for W^2_j *and*

$$\left\{\psi^1_{j,n}, \psi^2_{j,n}, \psi^3_{j,n}\right\}_{(j,n)\in\mathbb{Z}\times\mathbb{Z}^2}$$

is an orthonormal basis for $L^2(\mathbb{R}^2)$.

The approximation and detail coefficients are therefore given by

$$a_j[n] = \left\langle f, \phi^2_{j,n}\right\rangle, \quad d^k_j[n] = \left\langle f, \psi^k_{j,n}\right\rangle \quad \text{for } 1 \le k \le 3.$$

For the product of two one-dimensional filters $y[m], z[m]$ we denote the product filter by $yz[n] = y[n_1]z[n_2]$, and $n := [n_1, n_2]$. Furthermore, denote $\overline{y}[m] := y[-m]$ as before. Given the conjugate mirror filters $g[m], h[m]$ associated with a wavelet ψ (or a scaling function ϕ respectively), we obtain

$$\begin{aligned}
a_{j+1}[n] &= a_j \star \overline{h}\,\overline{h}[2n],\\
d^1_{j+1}[n] &= a_j \star \overline{h}\,\overline{g}[2n],\\
d^2_{j+1}[n] &= a_j \star \overline{g}\,\overline{h}[2n],\\
d^3_{j+1}[n] &= a_j \star \overline{g}\,\overline{g}[2n].
\end{aligned} \tag{3.15}$$

The fast wavelet transform using filter banks can be generalized to the multivariate case to compute the approximation and wavelet coefficients. For details we refer the reader to [80, Chapter 7.7.3]. It should be clear that all the ideas from the one-dimensional case can be carried over to the two-dimensional setting if we use separable wavelets.

Remark 3.1.16. The separable construction has several drawbacks. Since we only compute horizontal, vertical and diagonal directions, the tensor-product structure prefers those directions. Singularities along other curves are thus not captured sufficiently. The separable two-dimensional wavelet transform is inherently isotropic and therefore not aware of different directions. ○

The next section introduces shearlets, one approach to circumvent these drawbacks while maintaining most of the advantages that the wavelet transform has.

3.2. Shearlets

Before we describe the construction of compactly supported shearlets, we want to briefly review the development of anisotropic wavelet-like function systems. The introduction of wavelets to the field of signal processing had an enormous impact on applications as well as on the development of the theoretical aspects. Wavelets were soon recognized to provide sparse representations for most one-dimensional signals [79, 80], a fact that has been extensively exploited for various applications such as denoising of signals, reconstruction or extraction of special features. It was further possible to establish a connection between the regularity of functions in Besov spaces with the decay of wavelet coefficients of these functions [32].

However, a naïve tensor-product approach for multivariate approximation using wavelets for images is not optimal anymore in the sense of N-term approximations. This is due to the fact that tensor-product wavelets are intrinsically isotropic and cannot efficiently represent structures along curves. This circumstance led to the development of various function systems (often summarized under the term $*$-lets) that are based on wavelets and try to overcome isotropy in order to achieve better approximation rates, namely brushlets [83], ridgelets [16], contourlets [33], curvelets [17], and shearlets [61, 45, 78] – just to mention a few. A different approach is used by the *Easy Path Wavelet Transform (EPWT)*, see [95]. Here, the idea is to construct a (smooth) path through the pixels of the discrete image and to apply a one-dimensional wavelet transform along this path. Hence, the EPWT is an adaptive transform in contrast to the other transforms mentioned above. Its approximation properties are well understood and its usefulness for image processing is well established, see [49, 97, 96].

3.2.1. Construction of Shearlets

This section introduces shearlets using compactly supported generators as proposed in [69]. We briefly introduce the shearlet transform and review some of its properties as well es its numerical implementation. For an extensive treatment of the latter, we refer to [60].

An important ingredient for the construction of shearlets are the scaling, shearing and sampling matrices.

Definition 3.2.1 (Scaling, shearing, and sampling matrices). *For $j \in \mathbb{Z}$ and $k \in \mathbb{Z}$ we define the anisotropic scaling matrices*

$$A_{2^j} := \begin{pmatrix} 2^j & 0 \\ 0 & 2^{\lfloor j/2 \rfloor} \end{pmatrix}, \qquad \widetilde{A}_{2^j} := \begin{pmatrix} 2^{\lfloor j/2 \rfloor} & 0 \\ 0 & 2^j \end{pmatrix}.$$

For $k \in \mathbb{Z}$ we define the shearing matrices

$$S_k := \begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix}, \qquad S_k^T := \begin{pmatrix} 1 & 0 \\ k & 1 \end{pmatrix}.$$

For a sampling vector $c = (c_1, c_2)$ with sampling parameters $c_1, c_2 > 0$ we define the sampling matrices

$$M_c := \begin{pmatrix} c_1 & 0 \\ 0 & c_2 \end{pmatrix}, \qquad \widetilde{M}_c := \begin{pmatrix} c_2 & 0 \\ 0 & c_1 \end{pmatrix}.$$

Shearlets are constructed for two different *cones*, cf. Figure 3.3. The following definition introduces these cones. Note that this is not a general definition for cones but only a very special case that we will be working with here.

Definition 3.2.2 (Vertical, horizontal and truncated cones). *The horizontal and vertical cones are defined as*

$$\mathcal{K}_{\mathrm{h}} := \left\{ \xi = (\xi_1, \xi_2) \in \mathbb{R}^2 \mid |{}^{\xi_2}/_{\xi_1}| \leq 1 \right\},$$
$$\mathcal{K}_{\mathrm{v}} := \left\{ \xi = (\xi_1, \xi_2) \in \mathbb{R}^2 \mid |{}^{\xi_1}/_{\xi_2}| \leq 1 \right\}.$$

Using the definition of the centered rectangle

$$\mathcal{R} := \left\{\xi = (\xi_1, \xi_2) \in \mathbb{R}^2 \mid \|\xi\|_\infty < 1\right\},$$

we define the truncated horizontal cone *and the* truncated vertical cone *by*

$$\begin{aligned}
\mathcal{K}_{\mathrm{h}}^{\mathrm{t}} &:= \mathcal{K}_{\mathrm{h}} \ \setminus \ \mathcal{R} \\
&= \left\{\xi = (\xi_1, \xi_2) \in \mathbb{R}^2 \mid \ \xi_1 \geq 1, \ |\xi_2/\xi_1| \leq 1\right\} \cup \left\{\xi = (\xi_1, \xi_2) \in \mathbb{R}^2 \mid \xi_1 \leq -1, |\xi_2/\xi_1| \leq 1\right\}, \\
\mathcal{K}_{\mathrm{v}}^{\mathrm{t}} &:= \mathcal{K}_{\mathrm{h}} \ \setminus \ \mathcal{R} \\
&= \left\{\xi = (\xi_1, \xi_2) \in \mathbb{R}^2 \mid \ \xi_2 \geq 1, \ |\xi_1/\xi_2| \leq 1\right\} \cup \left\{\xi = (\xi_1, \xi_2) \in \mathbb{R}^2 \mid \xi_2 \leq -1, |\xi_1/\xi_2| \leq 1\right\}.
\end{aligned}$$

These cones are highlighted by bold lines in Figure 3.3. The shearlet system is a union of a low-pass element and two individual generating shearlets which are defined according to the horizontal respectively the vertical cone.

Definition 3.2.3 (Definition 2.1 from [60]). *Let* $\phi, \psi, \tilde{\psi} \in L^2(\mathbb{R}^2)$*, and* $c = (c_1, c_2) \in \mathbb{R}_+^2$*. We define the set of low-pass elements for* $\xi \in \mathcal{R}$ *by*

$$\Phi\left(\phi; c_1\right) := \left\{\phi_m = \phi\left(\cdot - c_1 m\right) : \ m \in \mathbb{Z}^2\right\},$$

the set of shearlet elements on the horizontal cone for $\xi \in \mathcal{K}_{\mathrm{h}}^{\mathrm{t}}$ *by*

$$\Psi\left(\psi; c\right) := \left\{\psi_{j,k,m} = 2^{\frac{3}{4}j}\psi\left(S_k A_{2^j} \cdot - M_c m\right) : \ j \geq 0, \ |k| \leq \lceil 2^{j/2} \rceil, \ m \in \mathbb{Z}^2\right\}$$

and the set of shearlet elements on the vertical cone for $\xi \in \mathcal{K}_{\mathrm{h}}^{\mathrm{t}}$ *by*

$$\widetilde{\Psi}\left(\tilde{\psi}; c\right) := \left\{\widetilde{\psi}_{j,k,m} = 2^{\frac{3}{4}j}\widetilde{\psi}\left(S_k^T \widetilde{A}_{2^j} \cdot - \widetilde{M}_c m\right) : \ j \geq 0, \ |k| \leq \lceil 2^{j/2} \rceil, \ m \in \mathbb{Z}^2\right\}.$$

The complete shearlet system *is then given by*

$$SH\left(\phi, \psi, \widetilde{\psi}; c\right) := \Phi\left(\phi; c_1\right) \cup \Psi\left(\psi; c\right) \cup \widetilde{\Psi}\left(\widetilde{\psi}; c\right).$$

The associated shearlet transform *maps a function* $f \in L^2(\mathbb{R}^2)$ *onto a series of shearlet coefficients*

$$c_{j,k,m}(f) := \left(\left\langle f, \phi_m \right\rangle, \left\langle f, \psi_{j,k,m} \right\rangle, \left\langle f, \widetilde{\psi}_{j,k,m} \right\rangle\right)$$

with the standard L^2*-scalar product. Elements from the shearlet system are called* shearlets.

More precisely, this construction is referred to as *cone-adapted regular discrete shearlet system*, cf. [70]. The reason for this can be seen in Figure 3.3 where the bold lines constitute the cones for which the shearlets are constructed individually. The question that arises from the definition is which functions $\phi, \psi, \widetilde{\psi}$ are suitable in order to achieve the desired properties and a fast numerical implementation. The *classical shearlets* are generated by separable, band-limited functions

$$\widehat{\psi}(\xi_1, \xi_2) = \widehat{\psi_1}(\xi_1)\,\widehat{\psi_2}\left(\frac{\xi_2}{\xi_1}\right), \quad \widehat{\widetilde{\psi}}(\xi_1, \xi_2) = \widehat{\widetilde{\psi}}_1\left(\frac{\xi_1}{\xi_2}\right)\widehat{\widetilde{\psi}}_2(\xi_2)$$

where ψ_1 and $\widetilde{\psi_1}$ are orthogonal wavelets and ψ_2 and $\widetilde{\psi_2}$ are bump functions. This construction directly yields a Parseval frame due to the tiling of the frequency plane, see Figure 3.3. Since we want to approximate objects that have compact support in spatial domain, it is natural to use compactly supported shearlets. The difficulty when designing shearlets with compact support (and hence functions that are *not* band-limited in the frequency plane) is to achieve a similar frequency tiling as in Figure 3.3.

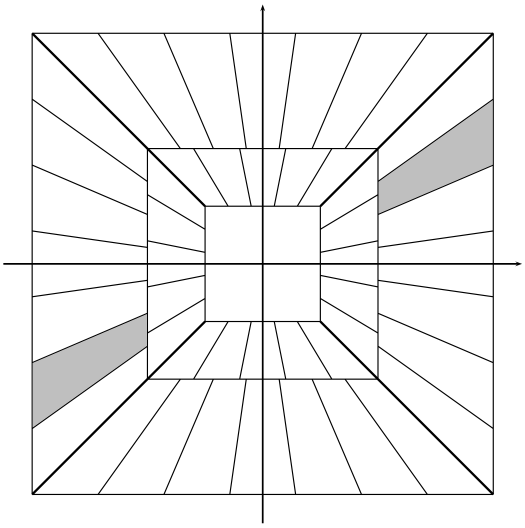


Figure 3.3.: Induced tiling of the frequency plane $\widehat{\mathbb{R}}^2$ for classical band-limited, cone-adapted shearlets

The first construction is due to Lim [69] where the shearlets form a frame with reasonable frame bounds A and B. Here, separable generators are used and a fast numerical implementation is derived. In [70], a construction based on non-separable compactly supported generators was presented together with a numerical implementation.

Remark 3.2.4. It is an open problem whether or not it is possible to construct compactly supported shearlets that form a Parseval frame for $L^2(\mathbb{R}^2)$, cf. [55]. ○

3.2.2. Compactly Supported Shearlets

In this section we want to briefly recall the characterization and assumptions for which shearlet frames with compactly supported generators exist.

Theorem 3.2.5 (Theorem II.3 from [70]). *Let* $\phi, \psi \in L^2(\mathbb{R}^2)$ *be compactly supported such that their Fourier transforms satisfy the decay conditions*

$$\hat{\phi}\left(\xi_1, \xi_2\right) \leq C_1 \min\left\{1, |\xi_1|^{-\gamma}\right\} \min\left\{1, |\xi_2|^{-\gamma}\right\} \tag{3.16}$$

and

$$\left|\hat{\psi}\left(\xi_1, \xi_2\right)\right| \leq C_2 \min\left\{1, |\xi_1|^{\alpha}\right\} \min\left\{|\xi_1|^{-\gamma}\right\} \min\left\{|\xi_2|^{-\gamma}\right\} \tag{3.17}$$

with positive constants $C_1, C_2 < \infty$ *and* $\alpha > \gamma > 3$. *Define* $\tilde{\psi}\left(x_1, x_2\right) = \psi\left(x_2, x_1\right)$ *and assume further that there exists a constant* $A > 0$ *such that*

$$\left|\hat{\phi}\left(\xi\right)\right|^2 + \sum_{j \geq 0} \sum_{|k| \leq 2^{\lceil j/2 \rceil}} \left|\hat{\psi}\left(S_k^T A_{2^{-j}} \xi\right)\right|^2 + \left|\hat{\tilde{\psi}}\left(S_k \tilde{A}_{2^{-j}} \xi\right)\right|^2 > A. \tag{3.18}$$

Then there exists a sampling vector $c = (c_1, c_2) \in \mathbb{R}_+^2$ *such that the generated shearlet system* $SH\left(\phi, \psi, \tilde{\psi}; c\right)$ *is a frame for* $L^2(\mathbb{R}^2)$.

Theorem 3.2.5 indicates the requirements on the wavelets chosen as generators. Using compactly supported wavelets one has to ensure smoothness and vanishing moments of ψ to fulfill (3.16) and (3.17). The boundedness from below away from zero in (3.18) resembles the conditions for the Fourier transform of the scaling function ϕ in the one-dimensional wavelet case. While this condition is less strict it also does not yield a Riesz basis. However, it still ensures a sufficient covering of the frequency plane and thus one obtains a frame for $L^2(\mathbb{R}^2)$ as stated by the theorem.

3.2.3. Discrete Shearlets

This section outlines the general idea of the implementation of a discrete shearlet transform. There exist several MATLAB toolboxes for band-limited as well as compactly supported shearlets, see [48, 60, 69, 70]. Here, we will focus on non-separable shearlets with compactly supported generators as discussed in [60] and follow the exposition therein. First, we introduce separable, compactly supported shearlets.

A key aspect in deriving a fast algorithm is the discretization of the shear operation S_k onto a grid $\mathbb{Z}^2$. We consider the horizontal cone and start with a two-dimensional tensor-product scaling function and wavelet, i.e.,

$$\phi(x) = \phi_1 \otimes \phi_1(x), \qquad \psi(x) = \psi_1 \otimes \phi_1(x) \tag{3.19}$$

where $x = (x_1, x_2) \in \mathbb{R}^2$. These generators are called *separable shearlet generators*. The construction for the vertical cone follows analogously while interchanging the coordinates x_1 and x_2 in the construction. The assumptions of Theorem 3.2.5 can be fulfilled by choosing a suitable scaling function and wavelet. Prominent examples that can be chosen (and are implemented in the ShearLab toolbox) include Coiflets, Daubechies wavelets, Symmlets, and their corresponding scaling functions.

While the abstract definition of the transform assumes a continuous function and calculates an infinite sequence of coefficients, the discrete transform will have to work on a finite number of samples with a finite number of scales and shearing directions. One thus assumes that the coefficients on some fine scale are given and then uses a procedure similar to Figure 3.2 to compute the coefficients on different scales. However, in practice, analogously to the wavelet transform, those coefficients are not directly available and an approximation has to be made at some point. Furthermore, in order to be able to apply a similar filter bank scheme as for wavelets, one needs to take care of the shearing operator. In order to do that, we first prove a fact about the commutation of the scaling and shearing matrix and consider its implications for the inner products $\left\langle f, \psi_{j,k,m} \right\rangle$.

Lemma 3.2.6. *With the notation as in Definition 3.2.1 we have*

$$S_k A_{2^j} = A_{2^j} S_{2^{-\lfloor j/2 \rfloor} k} \tag{3.20}$$

and hence

$$\psi_{j,k,m}(\cdot) = \psi_{j,0,m}(S_{2^{-\lfloor j/2\rfloor}k}\cdot). \tag{3.21}$$

Therefore, we can calculate the shearlet coefficients $\left\langle f, \psi_{j,k,m}\right\rangle$ *by*

$$\left\langle f, \psi_{j,k,m}\right\rangle = \left\langle f\left(S_{-2^{-\lfloor j/2\rfloor}k}\cdot\right), \psi_{j,0,m}\right\rangle. \tag{3.22}$$

Proof. For the first statement, we calculate

$$\begin{aligned}
S_k A_{2^j} &= \begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 2^j & 0 \\ 0 & 2^{\lfloor j/2\rfloor} \end{pmatrix} = \begin{pmatrix} 2^j & k2^{\lfloor j/2\rfloor} \\ 0 & 2^{\lfloor j/2\rfloor} \end{pmatrix} \\
A_{2^j} S_{2^{-\lfloor j/2\rfloor}k} &= \begin{pmatrix} 2^j & 0 \\ 0 & 2^{\lfloor j/2\rfloor} \end{pmatrix}\begin{pmatrix} 1 & k2^{-\lfloor j/2\rfloor} \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 2^j & k2^{\lfloor j/2\rfloor} \\ 0 & 2^{\lfloor j/2\rfloor} \end{pmatrix}.
\end{aligned}$$

For the second statement, we use the first statement and obtain

$$\begin{aligned}
\psi_{j,0,m}\left(S_{2^{-\lfloor j/2\rfloor}k}\cdot\right) &= 2^{\frac{j+\lfloor j/2\rfloor}{4}}\psi\left(S_0 A_{2^j} S_{2^{-\lfloor j/2\rfloor}k}\cdot - M_c m\right) \\
&\overset{S_0=\mathrm{Id}}{=} 2^{\frac{j+\lfloor j/2\rfloor}{4}}\psi\left(A_{2^j} S_{2^{-\lfloor j/2\rfloor}k}\cdot - M_c m\right) \\
&\overset{(3.20)}{=} 2^{\frac{j+\lfloor j/2\rfloor}{4}}\psi\left(S_k A_{2^j}\cdot - M_c m\right) \\
&= \psi_{j,k,m}(\cdot).
\end{aligned}$$

For the third statement, we write down the inner product, substitute with $y = S_{2^{\lfloor j/2\rfloor}k}x$, and obtain

$$\begin{aligned}
\left\langle f, \psi_{j,k,m}\right\rangle &= \int_{\mathbb{R}^2} f(x)\psi_{j,k,m}(x)\,\mathrm{d}x \\
&= \int_{\mathbb{R}^2} f(x)\,\psi_{j,0,m}\left(S_{2^{-\lfloor j/2\rfloor}k}\cdot\right)\,\mathrm{d}x \\
&= \int_{\mathbb{R}^2} \left|\det S_{2^{-\lfloor j/2\rfloor}k}\right| f\left(S_{-2^{-\lfloor j/2\rfloor}k}x\right)\psi_{j,0,m}(x)\,\mathrm{d}x \\
&= \left\langle f\left(S_{2^{-\lfloor j/2\rfloor}k}\cdot\right), \psi_{j,0,m}(\cdot)\right\rangle
\end{aligned}$$

as claimed, since $\left|\det S_{2^{-\lfloor j/2\rfloor}k}\right| = 1$. □

Remark 3.2.7. Calculating the shearlet coefficients hence is possible by calculating a discrete tensor-product wavelet transform with an anisotropic scaling matrix of the sheared data. The next steps are then to develop a faithful discretization of the shear operator. We further describe how to compute shearlet coefficients for non-separable generators. ○

If we discretize the shear operator $S_{2^{-\lfloor j/2 \rfloor}k}$, we need to make sure that it is well defined. In this current form, the shear operator $S_{2^{-\lfloor j/2 \rfloor}k}$ does not preserve the grid $\mathbb{Z}^2$, i.e.,

$$S_{2^{-\lfloor j/2 \rfloor}k}\left(\mathbb{Z}^2\right) \neq \mathbb{Z}^2.$$

More specifically, we need to refine the grid along the horizontal axis x_1, since

$$\begin{aligned} 2^{-\lfloor j/2 \rfloor}\mathbb{Z} \times \mathbb{Z} &= \begin{pmatrix} 2^{-\lfloor j/2 \rfloor} & 0 \\ 0 & 1 \end{pmatrix}\left(\mathbb{Z}^2\right) = \begin{pmatrix} 2^{-\lfloor j/2 \rfloor} & 0 \\ 0 & 1 \end{pmatrix}\left(S_k\mathbb{Z}^2\right) \\ &= S_{2^{-\lfloor j/2 \rfloor}k}\begin{pmatrix} 2^{-\lfloor j/2 \rfloor} & 0 \\ 0 & 1 \end{pmatrix}\left(\mathbb{Z}^2\right) = S_{2^{-\lfloor j/2 \rfloor}k}\left(2^{-\lfloor j/2 \rfloor}\mathbb{Z} \times \mathbb{Z}\right) \end{aligned}$$

where we used

$$\begin{pmatrix} 2^{-\lfloor j/2 \rfloor} & 0 \\ 0 & 1 \end{pmatrix} S_k = \begin{pmatrix} 2^{-\lfloor j/2 \rfloor} & 2^{-\lfloor j/2 \rfloor}k \\ 0 & 1 \end{pmatrix} = S_{2^{-\lfloor j/2 \rfloor}k}\begin{pmatrix} 2^{-\lfloor j/2 \rfloor} & 0 \\ 0 & 1 \end{pmatrix}$$

and that $\mathbb{Z}^2$ is invariant under the action of S_k. This leads to a grid $2^{-\lfloor j/2 \rfloor}\mathbb{Z} \times \mathbb{Z}$ for each decomposition level. However, it is still necessary to compute the sheared sampling values $f\left(S_{2^{-j/2}k}\cdot\right)$ from the given input, cf. [70]. One possibility is to apply an interpolation of the sampling values to obtain values on the refined grid, calculate the sheared sampling value and perform a downsampling afterwards. Therefore, we assume that the function $f \in L^2(\mathbb{R}^2)$ can be written as

$$f(x) = \sum_{n \in \mathbb{Z}^2} f_J[n] 2^J \phi\left(2^J x_1 - n_1, 2^J x_2 - n_2\right) \tag{3.23}$$

with ϕ as defined in (3.19). Let $\uparrow^{2^{\lfloor j/2 \rfloor}}$ denote the upsampling by a factor of $2^{\lfloor j/2 \rfloor}$ and $\downarrow_{2^{\lfloor j/2 \rfloor}}$ the downsampling by the same factor along the x_1-axis. Denote by $\star_1$ the convolution along the x_1-axis. One obtains interpolated sampling values $\tilde{f_J}$ on a finer grid $2^{-\lfloor j/2 \rfloor}\mathbb{Z} \times \mathbb{Z}$ by

$$\tilde{f_J} := \left((f_J)_{\uparrow 2^{\lfloor j/2 \rfloor}} \star_1 h_{\lfloor j/2 \rfloor}\right).$$

Note that $\tilde{f_J}$ is unchanged in x_2 direction and interpolated along the x_1-axis yielding $2^{-\lfloor j/2 \rfloor}$ as many sampling values. This is necessary to apply the shear operator. The application of $S_{2^{-\lfloor j/2 \rfloor}k}$ to discrete data f_J is hence achieved by computing the discretized shear operator

$$S^{\mathrm{d}}_{2^{-\lfloor j/2 \rfloor}k} f_J := \left((\tilde{f_J}[S_k \cdot]) \star_1 \overline{h}_{\lfloor j/2 \rfloor} \right)_{\downarrow_{2^{\lfloor j/2 \rfloor}}} \tag{3.24}$$

where, as above, we used the convention $\overline{h}[n] = h[-n]$. This result is central for the computation of the shearlet coefficients as we will discuss in the next section.

Non-Separable Shearlet Generator

For separable shearlets, we have considered generators of the form

$$\psi^{\mathrm{sep}}(x) := \psi_1 \otimes \phi_1(x) \tag{3.25}$$

where ψ_1 is a wavelet and ϕ_1 a scaling function. It is discussed in [60] that non-separable generators can achieve better numerical results. Non-separable generators are constructed in Fourier domain by

$$\widehat{\psi}(\xi) = P(\xi_1/2, \xi_2)\widehat{\psi^{\mathrm{sep}}}(\xi) \tag{3.26}$$

where P is the Fourier series of a two-dimensional fan filter.[2] Following [60], it is possible to construct compactly supported wavelets ψ_1, scaling functions ϕ_1, and finite two-dimensional fan filters $\{p[n]\}_{n \in \mathbb{Z}}$ such that the requirements of Theorem 3.2.5 are met, and one obtains a shearlet frame with reasonable frame bounds. We denote the Fourier coefficients of the trigonometric polynomials

$$\hat{h}_j(\xi_1) = \prod_{k=0}^{j-1} \hat{h}(2^k \xi_1), \quad \hat{g}_j(\xi_1) = \hat{g}\left(\frac{2^j \xi_1}{2}\right) \hat{h}_{j-1}(\xi_1) \tag{3.27}$$

by $\left\{h_j[n]\right\}_{n \in \mathbb{Z}}$ and $\left\{g_j[n]\right\}_{n \in \mathbb{Z}}$ respectively where $\hat{h}_0 \equiv 1$.

In the last section we have seen how the shear operator can be suitably discretized.

---

[2] Fan filters have been studied originally in the electrical engineering community, cf. [1]. The term *fan filter* describes those filters that have a wedge-like support in frequency domain. The fan filter is in this case responsible for directional sensitivity of the shearlet generators. For more details and examples, we refer to [1, 3, 24].

Now, we turn our attention to the computation of the shearlet coefficients $\left\langle f, \psi_{j,k,m} \right\rangle$ itself. We have seen in Lemma 3.2.6 that

$$\psi_{j,k,m}(\cdot) = \psi_{j,0,m}(S_{2^{-\lfloor j/2 \rfloor}k} \cdot) \qquad \text{and} \qquad \left\langle f, \psi_{j,k,m} \right\rangle = \left\langle f(S_{-2^{-\lfloor j/2 \rfloor}k} \cdot), \psi_{j,0,m} \right\rangle .$$

Recall that $\psi_{j,k,m}$ was defined as

$$\psi_{j,k,m}(x) = 2^{\frac{3}{4}j} \psi(S_k A_{2^j} \cdot - M_c m)$$

which for $k = 0$ and $M_c = \mathrm{Id}$ simplifies to

$$\psi_{j,0,m}(x) = 2^{\frac{3}{4}j} \psi(A_{2^j} \cdot - m). \tag{3.28}$$

Computing the Fourier transform of (3.28) yields

$$\widehat{\psi}_{j,0,m}(\xi) = 2^{-\frac{3}{4}j} \mathrm{e}^{-2\pi \mathrm{i} \left\langle m, A_{2^j}^{-1} \xi \right\rangle} \widehat{\psi}(A_{2^j}^{-1} \xi). \tag{3.29}$$

Using the definition of $\widehat{\psi}$ in (3.26) with ψ^{sep} as in (3.25) and plugging it into (3.29) yields

$$\widehat{\psi}_{j,0,m}(\xi) = 2^{-\frac{3}{4}j-1} \mathrm{e}^{-2\pi \mathrm{i} \left\langle m, A_{2^j}^{-1} \xi \right\rangle} P(A_{2^j}^{-1} Q^{-1} \xi) \hat{g}(2^{-j-1} \xi_1) \hat{h}(2^{-\lfloor j/2 \rfloor -1} \xi_2) \hat{\phi}(A_{2^j}^{-1} 2^{-1} \xi)$$

with $Q = \mathrm{diag}\,(2, 1)$. This can be shown to be equal to

$$\hat{\psi}_{j,0,m}(\xi) = 2^{-J} \mathrm{e}^{-2\pi \mathrm{i} \left\langle m, A_{2^j}^{-1} \xi \right\rangle} P(A_{2^j}^{-1} Q^{-1} \xi) \hat{g}_{J-j} \otimes \hat{h}_{J - \lfloor j/2 \rfloor}(2^{-J} \xi) \hat{\phi}(2^{-J} \xi) \tag{3.30}$$

where we refer to [60, Section 3.2] for details. Since we assumed that f is of the form (3.23) one obtains for its Fourier transform

$$\hat{f}(\xi) := 2^{-J} \hat{f}_J(2^{-J} \xi) \hat{\phi}(2^{-J} \xi). \tag{3.31}$$

One can show by using (3.31) and (3.30) that the shearlet coefficients are of the form

$$\left\langle f, \psi_{j,0,m} \right\rangle \quad = 2^{-2J} \int_{\mathbb{R}^2} \hat{f}_J(2^{-J} \xi) \left| \hat{\phi}(2^{-J} \xi) \right|^2 \mathrm{e}^{2\pi \mathrm{i} \left\langle m, A_{2^j}^{-1} \xi \right\rangle} P^*(A_{2^j}^{-1} Q^{-1} \xi) \hat{W}_j^*(2^{-J} \xi) \hat{\phi}(2^{-J} \xi) \, \mathrm{d}\xi$$

where the filter W_j is defined by $W_j = g_{J-j} \otimes h_{J-\lfloor j/2 \rfloor}$. Using a rescaling $\eta = 2^{-J}\eta$ and

$$\sum_{n \in \mathbb{Z}^2} \left|\hat{\phi}(\xi + n)\right|^2 = 1$$

one can then compute that

$$\left\langle f, \psi_{j,0,m} \right\rangle = \int_{[0,1]^2} \hat{f}_J(\eta) \mathrm{e}^{2\pi \mathrm{i} \left\langle m, A_{2^j}^{-1} 2^J \eta \right\rangle} P^*(2^J A_{2^j}^{-1} Q^{-1} \eta) \hat{W}^*(\eta) \, \mathrm{d}\eta.$$

It is shown in [60, Section 3.2.1] that one can compute the shearlet coefficients using discrete convolutions

$$\left\langle f, \psi_{j,0,m} \right\rangle = \left(f_J \star \left(\overline{p_j \star W_j}\right)\right) \left[A_{2^j}^{-1} 2^J m\right] \tag{3.32}$$

where $\left\{p_j[n]\right\}_{n \in \mathbb{Z}}$ are the Fourier coefficients of $P\left(2^{J-j-1}\xi_1, 2^{J-j/2}\xi_2\right)$. If we choose $p_j \equiv 1$ and omit the anisotropic scaling matrix A_{2^j}, (3.32) simplifies to the two-dimensional wavelet formula, see (3.15). Furthermore, denote the *digital shearlet filters*

$$\psi^{\mathrm{d}}_{j,k} = S^{\mathrm{d}}_{k/2^{j/2}} \left(p_j \star W_j\right) \tag{3.33}$$

with the discretized shearing operator as described above. Similarly as before, the corresponding shearlets on the other cone can be obtained by a changing the order of variables and will be denoted by $\tilde{\psi}^{\mathrm{d}}_{j,k}$.

For a more detailed treatment of the construction of compactly supported generator functions we refer to [56], for more information on non-separable generators we refer to [70] and to [60] for a detailed discussion on the numerical discretization. Suitable choices for the aforementioned low-pass, high-pass and fan filters will be discussed in the next section together with more details on the numerical implementation in ShearLab.

Inverse Transform

Since the shearlet system is not a basis, the dual frame elements (here denoted *dual shearlets*) have to be known in order to compute an inverse transform using the same scheme as for the forward transform. In general, this is often not the case and the pseudo-inverse is computed to perform an inverse transform.

However, in certain circumstances it is possible to give precise formulas for the dual shearlets and thus derive an inverse algorithm. This is described in [60] and studied in more detail in [59]. Here, we focus merely on giving the formulas that are important to understand the idea of the inverse algorithm.

Choosing a separable low-pass filter

$$\hat{\phi}^{\mathrm{d}}(\xi) = \hat{h}_J(\xi_1) \cdot \hat{h}_J(\xi_2)$$

with h_J as defined in (3.27) one defines the *dual frame weights*

$$\hat{\Psi}^{\mathrm{d}}(\xi) := \left|\hat{\phi}^{\mathrm{d}}(\xi)\right|^2 + \sum_{j=0}^{J-1} \sum_{|k|\leq 2^{\lfloor j/2 \rfloor}} \left(\left|\hat{\psi}^{\mathrm{d}}_{j,k}(\xi)\right|^2 + \left|\hat{\tilde{\psi}}^{\mathrm{d}}_{j,k}(\xi)\right|^2 \right)$$

where $\hat{\psi}^{\mathrm{d}}_{j,k}$ is the Fourier series of the digital shearlet filter $\psi^{\mathrm{d}}_{j,k}$ as defined in (3.33). The *dual shearlets* can then be defined by

$$\varphi^{\mathrm{d}}(\xi) = \frac{\hat{\phi}^{\mathrm{d}}(\xi)}{\hat{\Psi}^{\mathrm{d}}(\xi)}, \quad \gamma^{\mathrm{d}}_{j,k}(\xi) = \frac{\hat{\psi}^{\mathrm{d}}_{j,k}(\xi)}{\hat{\Psi}^{\mathrm{d}}(\xi)}, \quad \tilde{\gamma}^{\mathrm{d}}_{j,k}(\xi) = \frac{\hat{\tilde{\psi}}^{\mathrm{d}}_{j,k}(\xi)}{\hat{\Psi}^{\mathrm{d}}(\xi)}.$$

It is shown in [60] that these definitions yield a reconstruction formula that can be implemented similarly as the forward transform using point-wise multiplications in Fourier domain to realize the convolutions.

Remark 3.2.8. Note that although the low-pass filter is separable, the resulting generating functions are not. This separability is only valid for the low-pass filter which itself does not represent any directional features and therefore separability is not important. ○

3.2.4. Numerical Implementation

This section explains the numerical implementation of the shearlet transform in more detail. We will illustrate the behavior using a simple example. We start with an image $x \in \mathbb{R}^{d_1 \times d_2}$, perform the transform to obtain the shearlet coefficients, and perform an inverse transform. For each of these steps we will introduce the functions involved using the shearlet toolbox *ShearLab3D* in version 1.1.[3]

---

[3] Although the name might suggest that this toolbox only computes the shearlet transform of three-dimensional data, it offers both, a two-dimensional and a three-dimensional transform.

Forward Transform

The forward transform can be separated into two parts. First, a shearlet system has to be computed depending on the size of the image and the number of decomposition levels which imply the number of shearings.

In this step one has to chose a suitable quadrature mirror filter. We have seen that the shearlets are constructed on cones. Since shearlets on the border of two neighboring cones are very similar, the toolbox offers to compute only one of the two shearlets to save computation time and memory. The simplest version[4] to create a shearlet system is by calling the function

```
shearletSystem = SLgetShearletSystem2D(useGPU, rows, cols, nScales)
```

where `rows` and `cols` determine the size of the image and `scales` the number of scales. The boolean input `useGPU` specifies if the shearlet system will be stored on a GPU. In order to illustrate the method, we show the returned structure in Table 3.1. The shearlets are computed using convolutions which are calculated as a pointwise multiplication in Fourier domain. Therefore, the shearlet variable is complex-valued although they are real-valued indeed (up to machine precision). The size denotes the size of the shearlets and the shearLevels are determined by the number of decomposition levels, i.e., the number of shearings on the corresponding scale. The variable `full` denotes if a full system is computed or if the shearlets on the border of the cone are only computed once as mentioned above. The variable `nShearlets` stores the number of shearlets and `shearletIdxs` the indices of the shearlets in the format [cone, scale, shearing]. In `dualFrameWeights` the sums of the absolute squared shearlets are stored which is neccessary for the inverse transform. Furthermore, `RMS` stores the root mean squares of all shearlets for normalization purposes.

The storage on a GPU is indicated by `useGPU` while `isComplex` is always zero, since real-valued shearlets are used. Once a shearlet system is computed for a given size, number of scales (and further optional parameters), the forward transform is computed using

```
coeffs = SLsheardec2D(x,shearletSystem)
```

---

[4] In this case, a couple of default options are used. For example, a default quadrature mirror filter is chosen and a default number of shearing levels. For a detailed outline of these options we refer to the comments in the Matlab file *SLgetShearletSystem2D.m* inside the ShearLab3D toolbox.

Field	Value	Min	Max
shearlets	⟨512 × 512 × 49 complex double⟩	–	–
size	[512 512]	512	512
shearLevels	[1 1 2 2]	1	2
full	0	0	0
nShearlets	49	49	49
shearletIdxs	⟨49 × 3 double ⟩	-4	4
dualFrameWeights	⟨512 × 512 double ⟩	0.0669	1.000
RMS	⟨1 × 49 double ⟩	0.0231	0.1049
useGPU	0	0	0
isComplex	0	0	0

Table 3.1.: An example shearlet system structure computed using ShearLab3D 1.1 and MATLAB 2013a

where x denotes the input image. The coefficients on each scale are computed by point-wise multiplications of the shearlet system with the image x in Fourier domain. The variable `coeffs` is of the size 512 × 512 × 49 (in this example) or more generally rows × cols × nScales.

Inverse Transform

For the inverse transform, the ShearLab3D toolbox uses the same shearlet system as for the forward transform where each coefficient is weighted by the dual frame weights saved in `shearletSystem.dualFrameWeights` in order to obtain the dual shearlet frame. The calculation is therefore performed analogously to the forward transform by point-wise multiplications in Fourier domain. In this example, we retrieve the original image x using the command

```
x = SLshearrec2D(coeffs,shearletSystem);
```

where the shearlet system can be pre-computed and used for all images of the same size assuming the same number of scales and shear levels are used. As described in the last section, the dual shearlets can be obtained by rescaling with the dual frame weights Ψ^{d}. This makes the inverse transform fast and stable and the generated shearlet system universally applicable for forward and inverse transform.

Computing the Shearlet System

Since the forward and inverse transform reduce to a point-wise multiplication of the data with the shearlet system in Fourier domain, the more interesting aspect of the

Index	0	1	2	3	4	5	6	7	8
Value	0.010493	-0.026348	-0.051777	0.27635	0.58257	0.27635	-0.051777	-0.026348	0.010493

Table 3.2.: Filter coefficients for the maximally flat symmetric 9-tap filter

transform is how the shearlet system can be computed. We touched this topic briefly in the last section. The Matlab function of interest here is `SLgetShearletSystem2D`. This method computes the necessary filters for a given set of options such as image size, number of scales, number of shearings. We have seen in the last section that we basically need to specify a low-pass filter h and a fan filter P. The low-pass filter then induces a high-pass filter via (3.27). While, in principle, every filter that fulfills the assumptions of Theorem 3.2.5 is possible, the choice of the filter influences the numerical performance and the frame bounds of the generated shearlet system. In ShearLab 3D, the default choice for the low-pass filter is a so called *maximally flat symmetric 9-tap low-pass filter*. Here, *maximally flat* refers to filters that realize a maximum number of vanishing moments for a given length such as Daubechies wavelets mentioned in Remark 3.1.12. The coefficients for this filter are given in Table 3.2.

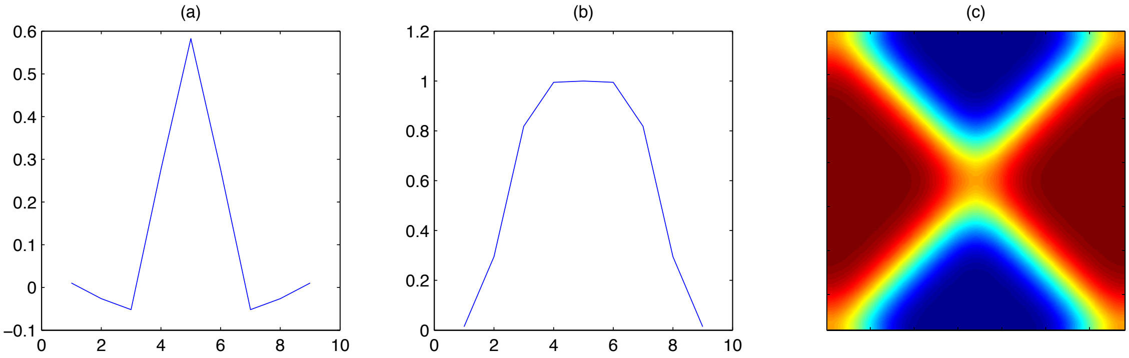


Figure 3.4.: Illustration of filters used for the discrete shearlet transform: (a) maximally flat, symmetric 9-tap low-pass filter, (b) magnitude of its Fourier transform, (c) magnitude of the Fourier transform of the maximally flat 2D fan filter.

Remark 3.2.9. Since there are no symmetric, compactly supported orthogonal wavelets, this filter only approximately fulfills orthogonality, see [60, Section 5.1]. As mentioned in [60], these filters are very similar to Cohen-Daubechies-Feauveau (CDF) 9/7 wavelets but trade some higher regularity and vanishing moments for the maximal flatness.

For the fan filter P, also a maximally flat filter is used that in this case is a so called *2D fan filter*. The coefficients can be computed using

```
modulate2(dfilters('dmaxflat4','d')./sqrt(2),'c');
```

where both functions are part of the Nonsubsampled Contourlet Toolbox [24] but also available in ShearLab 3D. Figure 3.4 illustrates the chosen low-pass filter, the magnitude of its Fourier transform as well as the magnitude of the Fourier transform of the fan filter. The latter illustrates the directionality of the filter. For further details on the implementation we refer to [60] and to the Matlab code itself which is freely available on `http://shearlet.org`.

4. Iterative Algorithms for Phase Retrieval

In the last chapters we have introduced the mathematical model that describes the imaging process and we presented the problem of phase retrieval. We further discussed shearlets, a representation system that has good approximation properties for a class of images and thus can be employed as an *a priori* condition.

In this chapter we consider algorithms that aim to solve the phase retrieval problem. Most of these algorithms can be grouped together as instances of generic fixed-point iterations. We study these algorithms and their development in order to understand the traditional perception of this problem. As briefly mentioned before, most often the algorithms can be interpreted to solve a *feasibility problem*, i.e., one seeks a solution as a point in the intersection of two (or more) sets. From the point of modeling, we distinguish between *measurement sets* and *constraint sets* although they are formally treated the same way. The measurement set describes the measured data in the transformed domain, in this case the intensities in Fresnel domain. The solution $x \in \mathbb{E}$, where $\mathbb{E} = C^{d_1 \times d_2}$ or $\mathbb{E} = C^d$ with $d = d_1 \cdot d_2$, should therefore satisfy an equation of the type

$$|Ux|_\circ = m \tag{4.1}$$

where $|\cdot|_\circ$ denotes the point-wise modulus, $U : \mathbb{E} \to \mathbb{E}$ is a unitary mapping, e.g. $U = \mathcal{D}_\tau[\cdot]$, and $m \in \mathbb{R}_+^d$ describes the data.[1] The set of *feasible solutions* can thus be written as

$$M = \{x \in \mathbb{E} \mid |Ux|_\circ = m\} . \tag{4.2}$$

The set M contains infinitely many elements x and a suitable solution can only be

---

[1] Of course, if $\mathbb{E} = \mathbb{C}^{d_1 \times d_2}$ then $m \in \mathbb{R}_+^{d_1 \times d_2}$ accordingly. Note that here $\mathcal{D}_\tau$ denotes a discrete version of the aforementioned Fresnel transform.

distinguished when posing additional constraints on the solution x.

For example, we enforce physical constraints of the object such as compact support or real-valuedness. A potential *constraint set* describing a real-valued object with compact support would be

$$C = \left\{ x \in \mathbb{R}^d \mid \operatorname{supp} x \subset D \right\}$$

for some *a priori* known $D \subset \{1, \dots, d\}$. This approach results in the feasibility problem

$$\text{find } x \in M \cap C. \tag{4.3}$$

In other words, the solution of the problem should be a function (or vector) with a prescribed support that simultaneously fulfills the measurements defined by (4.2) or (4.1), respectively. Later on in this chapter, we will introduce further possible constraints in order to achieve meaningful solutions of the phase retrieval problem. Such problems as described by (4.3) can be solved using projection algorithms.

These iterative phase retrieval algorithms make use of the fact that the projections onto the individual sets is known and fast to compute. However, since M given by (4.2) is not convex, the difficulty of these methods lies in the rigorous mathematical convergence analysis. The non-convexity leads to several local minima of the corresponding minimization problem and to a multi-valued projection operator. A common technique to circumvent or mitigate this difficulty of non-convex optimization problems are convex relaxations. However, these are not applicable in this case.

4.1. Projection Algorithms and Feasibility Formulation

The term *projection algorithm* describes iterative algorithms of the form

$$x^{(k+1)} = T_{A,B} x^{(k)} \tag{4.4}$$

where $T_{A,B}$ is a mapping that consists of (linear) combinations of projections onto sets $A, B \subset \mathbb{E}$. A prime example of this is the *method of alternating projections*

$$x^{(k+1)} = P_B P_A x^{(k)} \tag{4.5}$$

where P_A and P_B denote the projections onto the sets A and B. Typically, one assumes that $A \cap B \neq \emptyset$ but in certain situations some algorithms even converge to local best approximation points if $A \cap B = \emptyset$. Generally, the *projection* (or *best approximation*) Π_A onto a set A in a Euclidean space $\mathbb{E}$ is a multi-valued mapping $\Pi_A : \mathbb{E} \rightrightarrows \mathbb{E}$ defined by

$$\Pi_A(x) := \operatorname*{argmin}_{y \in A} \left\| x - y \right\|_{\mathbb{E}} . \tag{4.6}$$

A *projector onto A* (or *projection operator*) will then be denoted by $P_A \in \Pi_A$, cf. [5, Definition 3.7]. In most cases that we consider, the projector is single-valued and thus $\Pi_A = \{P_A\}$. In this case, the set A is called *Chebyshev set*. Indeed, every closed and convex set is a Chebyshev set. In finite dimensions, even the converse is true while this is an open question in infinite dimensions, cf. [7].

Remark 4.1.1. We can rewrite the minimization problem in (4.6) as follows. Denote by

$$\iota_A(x) := \begin{cases} 0, & \text{if } x \in A \\ +\infty, & \text{if } x \notin A \end{cases} \tag{4.7}$$

the *indicator functional of a set A*. Then, (4.6) is equivalent to

$$\Pi_A(x) = \operatorname*{argmin}_{y \in \mathbb{E}} \iota_A(y) + \left\| x - y \right\|_{\mathbb{E}} . \tag{4.8}$$

Since ι_A is convex if and only if A is convex and ι_A is lower semi-continuous if and only if A is closed, see [5, Section 1.9], it is immediately obvious why the properties of the set A are important for the existence and uniqueness of the projection operator. ○

Remark 4.1.2. In the phase retrieval problem, the forward operator $|Ux|_\circ$ is non-linear and the set described by (4.2) is not convex. The uniqueness of the projection operator is therefore not guaranteed. However, we will see that at least locally the projection is single-valued which corresponds to a notion of regularity called *prox-regular*. Indeed, the set M described in (4.2) is prox-regular, see [50, Theorem 9.6]. ○

In the optics community, several variants of the algorithm (4.4) are considered with special choices for A and B. We can only mention a selection and we discuss briefly the convergence results in Remark 4.1.3. For a more detailed discussion on further aspects of convergence for convex and non-convex sets, we refer to [50] and [6]. The latter provides an overview of convex algorithms, their non-convex counterparts, and a rich collection of literature.

In 1972, Gerchberg and Saxton proposed an algorithm that essentially is an alternating projection algorithm onto the sets

$$\begin{aligned} M_1 &= \left\{ x \in \mathbb{C}^d \mid |\hat{x}| = m_1 \right\}, \\ M_2 &= \left\{ x \in \mathbb{C}^d \mid |x| = m_2 \right\} \end{aligned} \tag{4.9}$$

for measurements $m_1, m_2 \in \mathbb{R}_+^d$. It uses an initial random guess $x^{(0)} = m_2 \odot \varphi_2^{(0)}$ where $\varphi_2^{(0)} = \mathrm{e}^{\mathrm{i}\zeta}$ and $\zeta = (\zeta[j])_{j=1}^d$ is uniformly distributed on $[0, 2\pi)$, see [39]. Here, in slight abuse of our problem formulation, two measurement sets are given. However, from the mathematical viewpoint, there is no difference between measurement sets and constraint sets.[2] The original version of this algorithm was stated in terms of a flowchart, see Figure 4.1. In this flowchart, the phase functions are computed as

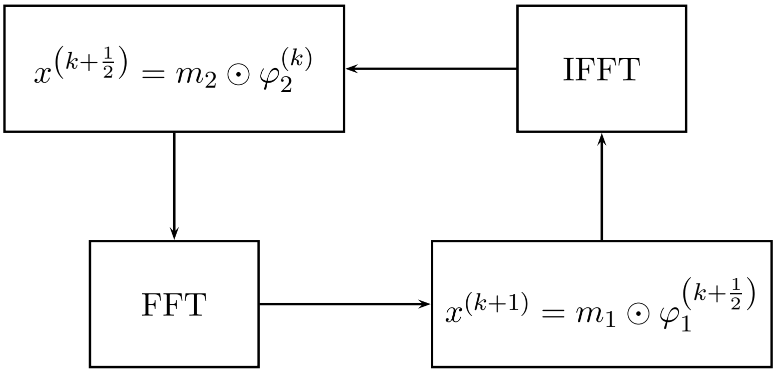

$x^{(k+\frac{1}{2})} = m_2 \odot \varphi_2^{(k)}$
IFFT
FFT
$x^{(k+1)} = m_1 \odot \varphi_1^{(k+\frac{1}{2})}$

Figure 4.1.: Schematic flowchart describing the Gerchberg-Saxton algorithm, based on the original figure from [39].

$$\varphi_1^{(k+1/2)}[j] = \begin{cases} \exp\left(\mathrm{i}\frac{\hat{x}^{(k+1/2)}[j]}{\left|\hat{x}^{(k+1/2)}[j]\right|}\right), & \left|\hat{x}^{(k+1/2)}[j]\right| \neq 0 \\ \exp(\mathrm{i}\theta), & \left|\hat{x}^{(k+1/2)}[j]\right| = 0 \end{cases} \qquad \varphi_2^{(k)}[j] = \begin{cases} \exp\left(\mathrm{i}\frac{x^{(k)}[j]}{\left|x^{(k)}[j]\right|}\right), & \left|x^{(k)}[j]\right| \neq 0 \\ \exp(\mathrm{i}\theta), & \left|x^{(k)}[j]\right| = 0, \end{cases}$$

for any $\theta \in [0, 2\pi)$, i.e., the new iterates are the alternating projections onto the sets

---

[2] This is true as long as the properties of the sets are not determined by their practical meaning. In this thesis, we are concerned with non-convex measurement sets where the constraints sets are typically convex. Hence, in this setting, the problem that the Gerchberg-Saxton algorithm solves is actually harder since both sets are non-convex.

M_1 and M_2 defined in (4.9). The projections onto these sets are given by

$$
\begin{aligned}
P_{M_1} x^{(k+1/2)} &= F^{-1} \begin{cases} m_1[j] \frac{F x^{(k+1/2)}[j]}{|F x^{(k+1/2)}[j]|}, & \text{if } F x^{(k+1/2)}[j] \neq 0 \\ m_1[j] \exp(\mathrm{i}\theta), & \text{if } F x^{(k+1/2)}[j] = 0 \end{cases} \\
P_{M_2} x^{(k)} &= \begin{cases} m_2[j] \frac{x^{(k)}[j]}{|x^{(k)}[j]|}, & \text{if } x^{(k)}[j] \neq 0 \\ m_2[j] \exp(\mathrm{i}\theta), & \text{if } x^{(k)}[j] = 0, \end{cases}
\end{aligned}
$$

where F denotes the discrete Fourier transform, see [50, Theorem 9.6] for a proof.

Although the non-convex nature of the problem was not yet understood, the authors of [39] noticed that the algorithm often stagnates at stationary points away from the true solution.

Remark 4.1.3 (Convergence Results for Alternating Projections). Following the remarks in [50], we want to briefly comment on the development of the convergence theory of the method of alternating projections. As mentioned, the first result dates back to von Neumann who established convergence for the case where A and B are subspaces, see [106]. For closed, convex sets $A_1, \dots, A_n$ with $\cap_{i=1}^n A_i \neq \emptyset$, Gubin et al. proved in [44] that the *cylindric projection* $P_{A_1} \cdots P_{A_n}$ converges to a point in the intersection. A proof for linear rates using the regularity of the intersection of the sets is given in [4]. Surveys on the alternating projections method in the convex setting can be found in [30, 31].

The first results for non-convex sets were published in [65, 64] followed by further quantifications of regularity in [10, 11, 12]. For a comprehensive treatment of the non-convex case, we refer to [50]. ○

In [37], Fienup proposed the error reduction algorithm which also is a variant of the method of alternating projections, cf. [63], with M as in (4.2) and the constraint set

$$S = \left\{ x \in \mathbb{R}^d \mid \operatorname{supp}(x) \subset D \right\} \tag{4.10}$$

for some set $D \subset \{j \in \mathbb{N} \mid 1 \leq j \leq d\}$. Here, the measurement set M is still non-convex but the constraint set S, where the constraint is on the support of the iterates in object domain, is an affine subspace. The projection onto S is given by

$$P_S x^{(k)}[j] = \begin{cases} x^{(k)}[j], & j \in D \\ 0, & \text{otherwise.} \end{cases}$$

Originally, the error-reduction algorithm was proposed in the form

$$x^{(k+1)}[j] = \begin{cases} P_M x^{(k)}[j], & j \in D \\ 0, & \text{otherwise} \end{cases}$$

but can be rewritten in the form $x^{(k+1)} = P_S P_M x^{(k)}$ with M as in (4.2) and S as in (4.10). Furthermore, using this notation, Fienup proposed the *hybrid input-output algorithm* (HIO)

$$x^{(k+1)}[j] = \begin{cases} P_M\, x^{(k)}[j], & j \in D \wedge x^{(k)}[j] \geq 0 \\ x^{(k)}[j] - \beta_k P_M x^{(k)}[j], & \text{otherwise.} \end{cases}$$

This notation is widely used in the optics community. However, in most cases it was shown that these algorithms can be written as fixed point iterations.

In [6, 7] the equivalence between HIO for $\beta_k \equiv 1$ and the *Douglas-Rachford algorithm*[3]

$$x^{(k+1)} = \frac{1}{2}\left(R_S R_M + \mathrm{Id}\right) x^{(k)} \tag{4.11}$$

has been shown where S as in (4.10) and M as defined in (4.2). Here, R_M is the *reflection over the set M* defined as $R_M := 2P_M - \mathrm{Id}$ and R_S defined likewise.

A more general result on the equivalence of HIO and a relaxed Douglas-Rachford can be found in [8], see also [75] for more details. The authors of [8] also proposed the *hybrid projection reflection* (HPR) algorithm which is given by

$$x^{(k+1)} = \frac{1}{2}\left(R_{S_+}\left(R_M + (\beta_k - 1)\, P_M\right) + \mathrm{Id} + (1 - \beta_k)\, P_M\right) x^{(k)}$$

where

$$S_+ = \left\{x \in \mathbb{R}^d_+ \mid \operatorname{supp} x \subset D\right\} \tag{4.12}$$

and with the projection

$$P_{S_+} x^{(k)}[j] = \begin{cases} \max\left\{\operatorname{Re}\left(x^{(k)}[j]\right), 0\right\}, & j \in D \\ 0, & \text{otherwise.} \end{cases} \tag{4.13}$$

---

[3] In the more general formulation, the projections onto sets are replaced by proximity operators, see [71]. It was originally developed for the numerical solution of heat conduction problems, see [34].

Besides all these, there are a lot more variants used in practice, e.g., the difference map where Elser showed a correspondence with HIO in some settings in [36]. A discussion of convergence properties in the convex case is given in [9]. For more details and newer results in the non-convex setting, we refer to [50]. The Douglas-Rachford algorithm is the building block for the *relaxed averaged alternating reflections* (RAAR) algorithm proposed by Luke in [75]. Written as a fixed point iteration, the algorithm to find $x \in A \cap B$ for two known sets A and B is given by

$$x^{(k+1)} = \frac{\beta_k}{2} \left(R_A R_B + \mathrm{Id} \right) x^{(k)} + (1 - \beta_k) \, P_B x^{(k)} \tag{4.14}$$

where $\beta_k \in (0, 1)$, i.e., it is a convex combination of the Douglas-Rachford algorithm and the projection onto P_B. In practice, one often choses $B = M$ and $A = S$ or $A = S_+$. This algorithm can be seen as a relaxation of (4.11). Its properties are discussed in detail in [75, 76].

Remark 4.1.4 (Convergence Results for RAAR). The convergence behavior of (4.14) in the convex case, i.e., when A and B are closed and convex, is well understood. We will discuss it in Section 4.3 which will lead to a better understanding of the relaxation parameter β_k which is of special interest for inconsistent feasibility problems, i.e., when $A \cap B = \emptyset$. Inconsistent problems often occur in practice due to noisy measurements or vague a priori information. ○

4.2. New *a priori* Conditions Based on Sparsity

In this section we introduce new *a priori* conditions based on sparsity of the solution in a suitable dictionary for the phase retrieval problem. There are several possibilities how to incorporate sparsity (or compressibility) of the vector of frame coefficients characterizing the solution in a dictionary. We attempt to incorporate this sparsity constraint into the feasibility formulation. We show that this approach has its difficulties and does not lead to wanted properties such as convex constraint sets. However, this problem can be solved using the observation that the threshold operator used[4] is a proximity operator in the transformed domain. This procedure leads to a better understanding of the behavior of the algorithm and motivates the construction of

---

[4] We will define the threshold operator later as the solution of a minimization problem. Intuitively, a threshold operator sets small, negligible coefficients to zero and maintains only the most significant coefficients.

other threshold operators.

When we replace the projection onto a set by the thresholding operation of frame coefficients, it is not obvious how to identify this operation with a feasibility problem. Indeed, it turns out that it is hard to find a set that represents the constraints in a meaningful way that is also convex at the same time.

Naively, given a sparsifying linear transform $T : \mathbb{R}^d \to \mathbb{R}^n$ one may pose the constraint set for a given $n_0 < n$ by

$$S_0 := \left\{ x \in \mathbb{R}^d \mid \|Tx\|_0 \leq n_0 \right\}$$

where $\|\cdot\|_0$ denotes the semi-norm that counts the number of non-zero entries, i.e. for $y \in \mathbb{R}^n$ define

$$\left\| y \right\|_0 := \# \{ j \in \{1, \dots, n\} \mid y[j] \neq 0 \} .$$

Remark 4.2.1. Although the set S_0 is non-convex, it can be used in practice. Consider the problem to find $x \in U \cap S_0$ where U is an affine subspace and $T = \mathrm{Id}$. In this case, [52] showed global linear convergence to a point $x \in U \cap S_0$ for the method of alternating projections and local linear convergence for the Douglas-Rachford algorithm.

However, if M is non-convex as it is in the phase retrieval problem and one considers the feasibility problem to find $x \in M \cap S_0$, these results no longer hold. We will therefore consider alternative approaches. ○

A typical approach which is based on the widespread convex relaxation of the ℓ_0-semi-norm is obtained by replacing it with the ℓ_1-norm. This relaxation yields the constraint set

$$S_1 := \left\{ x \in \mathbb{R}^d \mid \|Tx\|_1 \leq n_1 \right\} .$$

Observe however, that this constraint is qualitatively different from S_0 and does not meet our requirements. The simplest idea to enforce the sparsity condition $x \in S_0$ is to apply a so-called hard-threshold operator that only retains the n_0 frame coefficients with largest modulus. However, in image proccessing the hard-threshold operator is often replaced by the soft-threshold operator which is given by

$$\left(S_\gamma x\right)[j] := \begin{cases} x[j] - \gamma, & x[j] > \gamma, \\ x[j] + \gamma, & x[j] < -\gamma, \\ 0, & \left|x[j]\right| \leq \gamma. \end{cases} \tag{4.15}$$

with a suitable threshold parameter $\gamma > 0$. Thresholding operations have been shown to provide decent reconstruction techniques in image processing, cf. e.g. [26, 73, 35, 85, 19, 14]. Further threshold operations will be investigated at the end of this chapter as well as in the numerical evaluation of the algorithm. In [72], we studied the discrete shearlet transform as sparsifying transform T and proposed to use

$$S_{T,\gamma} := \left\{x \in \mathbb{R}^d \mid \exists h \in \mathbb{R}^d \,:\, \|Th\|_2 \leq c_m \wedge Tx = S_\gamma Th, \right\}$$

where c_m depends on the the threshold-parameter γ. This set is indeed a suitable choice to model the application of the soft-threshold operator. However, it is also a non-convex set.

Lemma 4.2.2 (Lemma 3.2 from [72]). *Given a linear frame transform $T : \mathbb{R}^d \to \mathbb{R}^n$ with $n \geq d$, measurements $m \in \mathbb{R}^d_+$, and $c_m > \gamma\sqrt{2}$, the set*

$$S_{T,\gamma} := \left\{x \in \mathbb{R}^d \mid \exists h \in \mathbb{R}^d \,:\, \|Th\|_2 \leq c_m \wedge Tx = S_\gamma Th, \right\},$$

where S_γ as in (4.15)*, is not convex.*

Proof. Consider $h_1, h_2 \in \mathbb{R}^n$ such that

$$\begin{aligned} Th_1 &= (c_m, 0, \dots, 0)^T \\ Th_2 &= \left(\sqrt{c_m^2 - (\gamma+\varepsilon)^2}, \varepsilon + \gamma, 0, \dots, 0\right)^T, \qquad \varepsilon > 0. \end{aligned}$$

and hence $\left\|Th_j\right\|_2 \leq c_m$ for $j = 1, 2$. Identify $x_1, x_2 \in S_{T,\gamma}$ by

$$\begin{aligned} Tx_1 &= S_\gamma Th_1 = (c_m - \gamma, 0, \dots, 0)^T \\ Tx_2 &= S_\gamma Th_2 = \left(\sqrt{c_m^2 - (\gamma+\varepsilon)^2} - \gamma, \varepsilon, 0, \dots, 0\right). \end{aligned}$$

Here, we need that $\sqrt{c_m^2 - (\gamma + \varepsilon)^2} > \gamma$ in order to achieve this representation which implies $c_m > \gamma\sqrt{2}$ and hence is fulfilled by assumption. We now prove that $S_{T,\gamma}$ is not convex by contradiction. We have $x_1, x_2 \in S_{T,\gamma}$ but we will show that the convex combination $1/2\,(x_1 + x_2) \notin S_{T,\gamma}$. Therefore, consider

$$\frac{1}{2}(Tx_1 + Tx_2) = \left(\frac{1}{2}(c_m - \gamma) + \frac{1}{2}\left[\sqrt{c_m^2 - (\gamma + \varepsilon)^2} - \gamma\right], \frac{\varepsilon}{2}, 0, \dots, 0\right).$$

Then given that $1/2\left(c_m + \sqrt{c_m - (\gamma + \varepsilon)^2}\right) > \gamma$, which can be ensured with $c_m > 5\gamma/4$ and is fulfilled by assumption, we have that

$$Th = \left(\frac{1}{2}\left(c_m + \sqrt{c_m^2 - (\gamma + \varepsilon)^2}\right), \frac{\varepsilon}{2} + \gamma, 0, \dots, 0\right)$$

is the vector with minimal norm such that $1/2\,(Tx_1 + Tx_2) = S_\gamma Th$. For sufficiently small $\varepsilon > 0$ we obtain

$$\|Th\|_2^2 = \left(\frac{1}{2}\left[c_m + \sqrt{c_m^2 - (\gamma + \varepsilon)^2}\right]\right)^2 + \left(\frac{\varepsilon}{2} + \gamma\right)^2 > c_m^2$$

since for $\varepsilon \to 0$ we observe that

$$\frac{1}{4}\left[c_m + \sqrt{c_m^2 - \gamma^2}\right]^2 + \gamma^2 > c_m^2$$

which then yields $1/2(x_1 + x_2) \notin S_{T,\gamma}$. □

Remark 4.2.3. Note that the non-convexity of the set is a consequence of the non-linearity of the soft-threshold operator and still holds true for tight frames. A new viewpoint to the problem will let us circumvent such difficulties and leads to a framework that connects threshold operations and so-called proximity operators which are generalizations of projection operators. Proximity operators have similar properties as projection operators which makes them suitable for our application. ○

We have seen in Remark 4.1.1 that the projection of a point $x \in \mathbb{E}$ onto a closed, convex set A is given by

$$P_A(x) = \operatorname*{argmin}_{y \in A} \frac{1}{2}\left\|x - y\right\|_{\mathbb{E}} = \operatorname*{argmin}_{y \in \mathbb{E}} \iota_A(y) + \frac{1}{2}\left\|x - y\right\|_{\mathbb{E}}$$

with ι_A as defined in (4.7). This motivates the following generalization.

Definition 4.2.4 (Proximity Operator, Definition 12.23 from [5]). *Let* $f \in \Gamma_0(\mathbb{E})$ *and* $x \in \mathbb{E}$. *The mapping*

$$\operatorname{prox}_f(x) = \operatorname*{argmin}_{y \in \mathbb{E}} f(y) + \frac{1}{2} \left\| x - y \right\|_{\mathbb{E}}^2$$

is the proximity operator *(or* proximal mapping*) of* f *at the point* x. *We further define for* $\gamma > 0$ *the* proximity operator *(or* proximal mapping*) of* γf *at the point* x *by*

$$\operatorname{prox}_{\gamma f}(x) = \operatorname*{argmin}_{y \in \mathbb{E}} f(y) + \frac{1}{2\gamma} \left\| x - y \right\|_{\mathbb{E}}^2 .$$

Using this definition, we immediately see that the projection onto a closed, convex set is the proximity operator of the indicator functional of this set, i.e., $P_A(x) = \operatorname{prox}_{\iota_A}(x)$. If $f = \gamma \left\| \cdot \right\|_1$, this penalty is less strict than the above model, but also promotes sparsity while being convex. Furthermore, $\operatorname{prox}_{\gamma \|\cdot\|_1}$ can be explicitly derived as a point-wise operation which makes it suitable for a fast implementation. The next proposition shows that the soft-threshold operator is the proximity operator with respect to $f = \gamma \left\| \cdot \right\|_1$. This result is widely known in the literature, see, e.g., [26].

Proposition 4.2.5. *The solution of*

$$\operatorname{prox}_{\gamma \|\cdot\|_1}(x) = \operatorname*{argmin}_{y \in \mathbb{R}^d} \left\| y \right\|_1 + \frac{1}{2\gamma} \left\| x - y \right\|_2^2 \tag{4.16}$$

is given component-wise by

$$\left(S_\gamma x \right)[j] := \operatorname{prox}_{\gamma \|\cdot\|_1}(x)[j] = \begin{cases} x[j] - \gamma, & x[j] > \gamma, \\ x[j] + \gamma, & x[j] < -\gamma, \\ 0, & \left| x[j] \right| \leq \gamma. \end{cases} \tag{4.17}$$

Proof. In (4.16) we have an unconstrained optimization problem where the objective is smooth away from zero and continuous everywhere. Furthermore, it is convex and decouples into one-dimensional subproblems, therefore we can consider the one-dimensional problem and search for zeros of the derivative. For $y[j] = 0$ we consider the subdifferential. Therefore we want to solve for all $j = 1, \dots, d$ the first order optimality condition

$$0 \in \partial\Big(\gamma \left|y[j]\right| + \frac{1}{2}\left(x[j] - y[j]\right)^2\Big).$$

Since $|\cdot|$ and $(x[j] - \cdot)^2$ are in $\Gamma_0(\mathbb{R}^d)$ and $\operatorname{dom}(x[j] - \cdot)^2 = \mathbb{R}$, we can apply the subdifferential sum rule from Fact B.1.10 which yields

$$0 \in \gamma \frac{y[j]}{\left|y[j]\right|} + x[j] - y[j]. \tag{4.18}$$

with $\frac{y[j]}{|y[j]|} = [-1, 1]$ if $y[j] = 0$ since the subdifferential of $\left|y[j]\right|$ is given by

$$\partial\left|y[j]\right| = \begin{cases} 1, & y[j] > 0, \\ -1, & y[j] < 0, \\ [-1, 1], & y[j] = 0. \end{cases}$$

Plugging this into (4.18) we obtain (4.17). □

Remark 4.2.6. It is of course possible to use other penalty functions than the ℓ_1-norm. As long as they are lower semi-continuous, proper and convex, they will always lead to unique proximity operators. However, it is not guaranteed that the solutions will be easy to compute which is the case for the ℓ_1-norm. We will discuss such proximity operators that are simple threshold operators in Section 4.6. ○

Introducing soft-threshold operators for frames.

We have seen in Proposition 4.2.5 how to compute the proximity operator with respect to the ℓ_1-norm. We are interested in sparsity constraints in the transformed domain and consider the proximity operator

$$\operatorname{prox}_{\gamma\|T\cdot\|_1}(x) = \operatorname*{argmin}_{y \in \mathbb{R}^d} \left\|Ty\right\| + \frac{1}{2\gamma}\left\|x - y\right\|_2^2 \tag{4.19}$$

with a linear frame transform $T : \mathbb{R}^d \to \mathbb{R}^n$. In general, the solution $\operatorname{prox}_{\gamma\|T\cdot\|_1}$ of (4.19) will not be the same as $T^{-1}\operatorname{prox}_{\gamma\|\cdot\|_1} T$. However, this will be the case when T is a unitary mapping, cf. [35].

Remark 4.2.7. There are two approaches to introduce the algorithm. On the one hand, we want to solve the minimization problem

$$\min_{u \in \mathbb{R}^n} \|u\|_1 + \frac{1}{2\gamma} \|v - u\|_2^2$$

where $u, v \in \mathbb{R}^n$ are vectors containing frame coefficients which yields the proximity operator in the transformed domain

$$\operatorname{prox}_{\|\gamma \cdot\|_1}(x) = \operatorname*{argmin}_{u \in \mathbb{R}^n} \|u\|_1 + \frac{1}{2\gamma} \|v - u\|_2^2 .$$

This mapping enforces sparse or compressible solutions in the transformed domain. In order to use this in the algorithm, one would then define the mapping

$$T^{-1} \operatorname{prox}_{\gamma\|\cdot\|_1}(Tx) = \left\{ y \in \mathbb{R}^d \mid Ty = \operatorname{prox}_{\gamma\|\cdot\|_1}(Tx) \right\}. \tag{4.20}$$

Depending on the properties of T, this mapping is indeed different from $\operatorname{prox}_{\|Tx\|_1}(Tx)$.

On the other hand, the theoretical analysis of the operator defined in (4.19) may seem easier at first sight. In the following, we will compare those two operations and analyze the mapping defined in (4.20). The numerical implementation will use the mapping $T^{-1} \operatorname{prox}_{\|\cdot\|_1, \gamma}(Tx)$. ○

Lemma 4.2.8 (Section II from [35]). *Let $T : \mathbb{R}^d \to \mathbb{R}^d$ be a unitary, linear mapping. Then*

$$\operatorname{prox}_{\gamma\|T\cdot\|_1} = T^{-1} \circ \operatorname{prox}_{\gamma\|\cdot\|_1} \circ T. \tag{4.21}$$

Proof. First, we define

$$\Gamma(y) := \left\|Ty\right\|_1 + \frac{1}{2\gamma} \left\|x - y\right\|_2^2 .$$

Substituting $u = Tx, v = Ty$ we have

$$\Gamma(v) = \|v\|_1 + \frac{1}{2\gamma} \left\|T^{-1}(u - v)\right\|_2^2 .$$

Since T was assumed to be unitary, this yields

$$\Gamma(v) = \|v\|_1 + \frac{1}{2\gamma} \|u - v\|_2^2 .$$

The corresponding proximity operator hence is

$$\operatorname*{argmin}_{v \in \mathbb{R}^n} \|v\|_1 + \frac{1}{2\gamma} \|u - v\|_2^2 = \operatorname{prox}_{\gamma\|\cdot\|_1} \circ T.$$

Applying the inverse T^{-1} to the left yields the claim. □

This result can be generalized to linear mappings $T : \mathbb{R}^d \to \mathbb{R}^n$ with weaker assumptions and for general penalty functions $f \in \Gamma_0(\mathbb{R}^d)$. The following is a finite dimensional simplification of [22, Proposition 11].

Proposition 4.2.9 (Proposition 11 from [22]). *Let $f \in \Gamma_0(\mathbb{R}^n)$, $T : \mathbb{R}^d \to \mathbb{R}^n$ linear such that $TT^* = \nu \operatorname{Id}$ for some $\nu \in (0, \infty)$. Then $f \circ T \in \Gamma_0(\mathbb{R}^d)$ and*

$$\operatorname{prox}_{f \circ T} = \operatorname{Id} + \nu^{-1} T^* \circ (\operatorname{prox}_{\nu f} - \operatorname{Id}) \circ T. \tag{4.22}$$

Proof. For a proof we refer to [22]. □

Remark 4.2.10. Proposition 4.2.9 is indeed a generalization of Lemma 4.2.8. Suppose that T is unitary, then $T^* = T^{-1}$ and $TT^* = \operatorname{Id}$, hence $\nu = 1$. Therefore, (4.22) reduces to (4.21). Unfortunately, even for tight frames, we only have $T^*T = \nu\operatorname{Id}$, but not $TT^* = \nu\operatorname{Id}$, since, in general, $n > d$.

We now introduce a new algorithm based on soft-thresholding of frame coefficients. This algorithm uses the proximity operator of the ℓ_1-norm in the transformed domain.

Definition 4.2.11 (Exact RAAR-(T, γ, ℓ_1)). *Let $T : \mathbb{R}^d \to \mathbb{R}^n$ be a linear frame transform, i.e., T maps a vector $x \in \mathbb{R}^d$ onto its frame coefficients $y = Tx \in \mathbb{R}^n$ for $n \geq d$. Furthermore, let $\beta_k > 0$ for all $k \in \mathbb{N}$. Consider the measurement set*

$$M = \left\{ x \in \mathbb{C}^d \mid |Ux|_\circ = m \right\}$$

with a unitary transform $U : \mathbb{C}^d \to \mathbb{C}^d$ *and denote by*

$$\operatorname{prox}_{\gamma\|T\cdot\|_1}(x) := \operatorname*{argmin}_{y\in\mathbb{R}^d} \left\|Ty\right\|_1 + \frac{1}{2\gamma}\left\|x-y\right\|_2^2. \tag{4.23}$$

Then we define the exact RAAR-(T,γ,ℓ_1) iteration *for a given* $x^{(0)} \in \mathbb{R}^d$ *by*

$$x^{(k+1)} = \frac{\beta_k}{2}\left(R_{\gamma\|T\cdot\|_1}R_M + \mathrm{Id}\right)x^{(k)} + (1-\beta_k)\,P_M x^{(k)} \tag{4.24}$$

where $R_M := 2P_M - \mathrm{Id}$ *and*

$$R_{\gamma\|T\cdot\|_1} := 2\operatorname{prox}_{\gamma\|T\cdot\|_1} - \mathrm{Id}.$$

We have seen before that the proximity operator defined by (4.23) does not necessarily coincide with $T^{-1}\operatorname{prox}_{\gamma\|\cdot\|_1} T = T^{-1}S_\gamma T$. Indeed, this is only the case if T is unitary. Although one could obtain the proximity operator by solving a simple convex minimization problem, it may be computationally expensive since the transform T and its inverse transform T^{-1} have to be computed in each minimization step. Therefore, we propose an inexact version of this algorithm.

Definition 4.2.12 (Inexact RAAR-(T,γ,ℓ_1)). *Let* $T : \mathbb{R}^d \to \mathbb{R}^n$ *be a linear frame transform, i.e.,* T *maps a vector* $x \in \mathbb{R}^d$ *onto its frame coefficients* $y = Tx \in \mathbb{R}^n$ *for* $n \geq d$*. Denote by* $T^\dagger$ *the pseudo-inverse of* T*. Furthermore, let* $\beta_k > 0$ *for all* $k \in \mathbb{N}$*. Consider the measurement set*

$$M = \left\{x \in \mathbb{C}^d \mid |Ux|_\circ = m\right\}$$

with a unitary transform $U : \mathbb{C}^d \to \mathbb{C}^d$ *and denote by*

$$P_{S_{T,\gamma}} := T^\dagger S_\gamma T$$

an approximation to the proximity operator $\operatorname{prox}_{\gamma\|T\cdot\|_1}(x)$*. Then we define the* inexact RAAR-(T,γ,ℓ_1) iteration *for a given* $x^{(0)} \in \mathbb{R}^d$ *by*

$$x^{(k+1)} = \frac{\beta_k}{2}\left(R_{S_{T,\gamma}}R_M + \mathrm{Id}\right)x^{(k)} + (1-\beta_k)\,P_M x^{(k)} \tag{4.25}$$

where $R_M := 2P_M - \mathrm{Id}$ *and*

$$R_{S_{T,\gamma}} := 2P_{S_{T,\gamma}} - \mathrm{Id} = 2T^\dagger S_\gamma T - \mathrm{Id}.$$

Remark 4.2.13. Since M is non-convex and as it is not obvious if $P_{S_{T,\gamma}}$ is a proximity operator, it is not immediately clear if convergence can be expected at all. Therefore, we first compare the algorithm to a convex analogue of the Douglas-Rachford algorithm. It will turn out that for a special instance, a prominent convergence statement will hold. Furthermore, we will investigate the fully discrete non-convex instance of this iteration and show that the iterates are bounded which will yield Cesaro-convergence in this setting. ○

4.3. Convergence Results for Exact RAAR-(T, γ, ℓ_1) in the Convex Setting

In this section we study the convergence behavior of different instances of the proposed algorithm in the convex setting. While these results do not carry over to the non-convex setting, they are first and foremost interesting on their own since the algorithm can be used for convex problems, too. Furthermore, these results do gain some insight on the relaxation parameter β_k and the fixed-points of the iteration. They will also indicate what type of convergence to expect. For example, in the convex setting, it is not the sequence $x^{(k)}$ itself that will converge to the solution of the problem but the shadow sequence $y^{(k)} = \mathrm{prox}_{\gamma g}(x^{(k)})$. Moreover, we will see that the original RAAR iteration converges to *nearest points* for inconsistent feasibility problems and that the parameter β_k controls the initial behavior of the algorithm as well as the location of the fixed-points with respect to the measurement set M. These results, while important on their own, provide a meaningful heuristic for the non-convex setting.

4.3.1. Convergence of RAAR-($\mathrm{Id}, \gamma, \ell_1$) for $\beta_k \equiv 1$

Consider the minimization of the sum of two functions $f, g \in \Gamma_0(\mathbb{R}^d)$:

$$\min\{f(x) + g(x)\}. \tag{4.26}$$

We first cite a convergence result for an algorithm which uses proximity operators to solve (4.26). Afterwards, we will show that this algorithm coincides with the Douglas-Rachford iteration for special choices of f and g,

$$x^{(k+1)} = \frac{1}{2}(R_B R_A + \mathrm{Id})x^{(k)} \tag{4.27}$$

which we introduced in (4.11) for the phase retrieval problem.

Definition 4.3.1 (Zeros of Multivalued Mappings, Equation 1.8 in [5]). *Let* $H : \mathbb{E} \rightrightarrows \mathbb{E}$ *be a (multivalued) mapping. We define the* zeros of H *by*

$$\operatorname{zer} H = H^{-1}0 = \{x \in \mathbb{E} \mid 0 \in Hx\}. \tag{4.28}$$

We will apply this definition to subdifferentials of convex functions (which may be multivalued mappings) as a necessary optimality criterion. For a proof of the following proposition, we refer to [5]. Before we state the next result, we need to introduce the notion of *uniform convexity*.

Definition 4.3.2 (Definition 10.5 from [5]). *Let* $f : \mathbb{E} \to \mathbb{R}_\infty$ *be proper. Define the increasing* modulus function $\phi : \mathbb{R}_+ \to [0, \infty]$ *that only vanishes at* 0*. If for all* $x, y \in \operatorname{dom} f$ *and* $\alpha \in (0, 1)$ *it holds that*

$$f(\alpha x + (1-\alpha)y) + \alpha(1-\alpha)\phi(\left\|x-y\right\|_{\mathbb{E}}) \le \alpha f(x) + (1-\alpha)f(y) \tag{4.29}$$

then f *is* uniformly convex with modulus ϕ*. If* (4.29) *holds for all* $x, y \in C$ *where* $C \subset \operatorname{dom} f$ *is nonempty,* f *is* uniformly convex on C*. If* (4.29) *holds for all* $x, y \in \operatorname{dom} f$ *and* $\alpha \in (0, 1)$ *with* $\phi = \beta/2\,|\cdot|$ *with* $\beta > 0$*,* f *is* strongly convex.

We now state the convergence result for an algorithm that solves (4.26). Since we are interested in the finite-dimensional setting, we state the finite dimensional version.

Proposition 4.3.3 (Corollary 27.4 from [5]). *Let* $f, g \in \Gamma_0(\mathbb{E})$ *such that*

$$\operatorname{zer}(\partial f + \partial g) \neq \emptyset.$$

Further, let $(\lambda^{(k)})_{k\in\mathbb{N}} \subset [0,2]$ *such that*

$$\sum_{k\in\mathbb{N}} \lambda^{(k)}(2-\lambda^{(k)}) = +\infty,$$

let $\gamma > 0$ *and* $x^{(0)} \in \mathbb{E}$. *For all* $k \in \mathbb{N}$ *set*

$$\begin{aligned} y^{(k)} &= \operatorname{prox}_{\gamma g}\left(x^{(k)}\right), \\ z^{(k)} &= \operatorname{prox}_{\gamma f}\left(2y^{(k)} - x^{(k)}\right), \\ x^{(k+1)} &= x^{(k)} + \lambda^{(k)}\left(z^{(k)} - y^{(k)}\right). \end{aligned} \tag{4.30}$$

Then there exists $x \in \mathbb{E}$ *such that the following hold:*

1. $\operatorname{prox}_{\gamma g}(x) \in \operatorname{argmin}(f+g)$.
2. $(y^{(k)} - z^{(k)})_{k\in\mathbb{N}}$ *converges to* 0.
3. $(x^{(k)})_{k\in\mathbb{N}}$ *converges to* x.
4. $(y^{(k)})_{k\in\mathbb{N}}$ *and* $(z^{(k)})_{k\in\mathbb{N}}$ *converge to* $\operatorname{prox}_{\gamma g}(x)$.
5. *If one of the following holds additionally:*
   a) f *is uniformly convex on every nonempty bounded subset of* $\operatorname{dom}(\partial f)$.
   b) g *is uniformly convex on every nonempty bounded subset of* $\operatorname{dom}(\partial g)$.

   then $\operatorname{prox}_{\gamma g}(x)$ *which is unique minimizer of* $f+g$.

Remark 4.3.4. Note that although the sequence $(x^{(k)})_{n\in\mathbb{N}}$ converges to x, we are interested in the *shadow sequence* $(y^{(k)})_{k\in\mathbb{N}} = \left(\operatorname{prox}_{\gamma g}(x^{(k)})\right)_{k\in\mathbb{N}}$ which converges to the unique minimizer of $f+g$ in the case that 5a) or 5b) holds.

Lemma 4.3.5. *For closed, convex sets* $A, B \subset \mathbb{R}^d$ *such that* $A \cap B \neq \emptyset$, $f = \iota_A$, $g = \iota_B$, *and* $\gamma = 1$, (4.30) *is equivalent to*

$$x^{(k+1)} = \frac{1}{2}(R_A R_B + \mathrm{Id})x^{(k)} \tag{4.31}$$

and hence solves the feasibility problem to find $x \in A \cap B$.

Proof. For $f = \iota_A, g = \iota_B$ and $\gamma = 1$ we obtain

$$\begin{aligned}\operatorname{prox}_{\gamma f} &= \operatorname{prox}_{\iota_A} = P_A\\ \operatorname{prox}_{\gamma g} &= \operatorname{prox}_{\iota_B} = P_B\end{aligned}$$

which leads to

$$\begin{aligned}y^{(k)} &= P_B\left(x^{(k)}\right)\\ z^{(k)} &= P_A\left(2y^{(k)} - x^{(k)}\right)\\ x^{(k+1)} &= x^{(k)} + \lambda^{(k)}\left(z^{(k)} - y^{(k)}\right).\end{aligned}$$

Rewriting this with $\lambda^{(k)} \equiv 1$ and using $R_B = 2P_B - \mathrm{Id}$ yields

$$x^{(k+1)} = (P_A R_B - P_B)x^{(k)} + x^{(k)}.$$

On the other hand, (4.31) can be rewritten as

$$\begin{aligned}x^{(k+1)} &= \frac{1}{2}\left(R_A R_B + I\right)x^{(k)}\\ &= \frac{1}{2}R_A R_B x^{(k)} + \frac{1}{2}x^{(k)}\\ &= \frac{1}{2}\left(2P_A - \mathrm{Id}\right)\left(2P_B - \mathrm{Id}\right)x^{(k)} + \frac{1}{2}x^{(k)}\\ &= \frac{1}{2}\left(4P_A P_B - 2P_B + \mathrm{Id} - 2P_A\right)x^{(k)} + \frac{1}{2}x^{(k)}\\ &= \left(2P_A P_B - P_B + \frac{1}{2}\mathrm{Id} - P_A\right)x^{(k)} + \frac{1}{2}x^{(k)}\\ &= \left(2P_A P_B - P_B - P_A + \mathrm{Id}\right)x^{(k)}\\ &= \left(P_A\left(2P_B - \mathrm{Id}\right) - P_B + \mathrm{Id}\right)x^{(k)}\\ &= \left(P_A R_B - P_B\right)x^{(k)} + x^{(k)}\end{aligned}$$

which establishes the claim. □

Using this relation, we can now use the convergence result from Proposition 4.3.3 and apply it to a convex instance of our proposed algorithm (4.25). Beforehand, we will need some results in order to verify the assumptions from Proposition 4.3.3. Mainly, we will need to verify $\operatorname{zer}(\partial f + \partial g) \neq \emptyset$. For proofs of the following results, we refer the interested reader to [5].

Theorem 4.3.6 (Proposition 27.2 from [5]). *Let $f, g \in \Gamma_0(\mathbb{E})$ such that one of the following holds:*

1. $\operatorname{argmin}(f+g) \neq \emptyset$ *and* $0 \in \operatorname{sri}(\operatorname{dom} f - \operatorname{dom} g)$
2. $\operatorname{argmin}(f+g) \subset \operatorname{argmin} f \cap \operatorname{argmin} g \neq \emptyset$
3. $f = \iota_A$ *and* $g = \iota_B$ *where* $A, B \subset \mathcal{H}$ *are closed and convex with* $A \cap B \neq \emptyset$.

Then $\operatorname{argmin}(f+g) = \operatorname{zer}(\partial f + \partial g) \neq \emptyset$.

Definition 4.3.7 (Cones). *Let $C \subset \mathbb{E}$, $\lambda \in \mathbb{R}$ then we denote the set $\lambda C = \{\lambda x \mid x \in C\}$. The set C is a* cone *if for all $\lambda > 0$ it holds that $C = \lambda C$. Furthermore,* $\operatorname{cone} C$ *denotes the smallest cone that contains C.*

Remark 4.3.8 (Strong relative interior, Definition 6.9 in [5]). The notation $\operatorname{sri} C$ for subset $C \subset \mathbb{E}$ denotes the *strong relative interior of* C, a weaker notion of interiority defined by

$$\operatorname{sri} C = \left\{ x \in C \mid \operatorname{cone}(C - x) = \overline{\operatorname{span}(C - x)} \right\}.$$

By [5, Example 6.10] it holds that $\operatorname{int} C \subset \operatorname{sri} C \subset C$. Hence, it will be sufficient to show that $0 \in \operatorname{int}(\operatorname{dom} f - \operatorname{dom} g)$ since it implies that $0 \in \operatorname{sri}(\operatorname{dom} f - \operatorname{dom} g)$. ○

Theorem 4.3.9 (Corollary 11.15 from [5]). *Let $f, g \in \Gamma_0(\mathbb{E})$. Suppose* $\operatorname{dom} f \cap \operatorname{dom} g \neq \emptyset$ *such that f is coercive and g is bounded from below.*

Then $f + g$ is coercive and it has a minimizer in $\mathbb{E}$. If f or g is strictly convex, then $f + g$ has exactly one minimizer over $\mathbb{E}$.

Consider the exact RAAR-(T, γ, ℓ_1) iteration from Definition 4.2.12 which reads

$$x^{(k+1)} = \frac{\beta_k}{2}\left(R_{\gamma\|T\cdot\|_1} R_M + \operatorname{Id}\right) x^{(k)} + (1 - \beta_k)\, P_M x^{(k)}$$

where $R_M := 2P_M - \mathrm{Id}$ and

$$R_{\gamma\|T\cdot\|_1} := 2\,\mathrm{prox}_{\gamma\|T\cdot\|_1} - \mathrm{Id}.$$

Suppose now that M is a closed, convex subset of $\mathbb{E}$ and $T = \mathrm{Id}$. In this case with $\beta_k \equiv 1$, the iteration simplifies to

$$x^{(k+1)} = \frac{1}{2}\left(R_{\gamma\|\cdot\|_1} R_M + \mathrm{Id}\right) x^{(k)}. \tag{4.32}$$

Proposition 4.3.10. *Let M be a closed, convex, nonempty subset of $\mathbb{E}$ and $T = \mathrm{Id}$. Then the prerequisites of Proposition 4.3.3 are fulfilled for the iteration (4.32). Hence, the consequences 1) – 5) of Proposition 4.3.3 hold true, especially there is a $x \in \mathbb{E}$ such that the sequence $x^{(k)}$ converges to x and $\mathrm{prox}_{\gamma\|\cdot\|_1}(x)$ is a minimizer of the optimization problem $\min f + g$ with $f = \gamma\|\cdot\|_1$ and $g = \iota_M$.*

Proof. We need to prove that $f, g \in \Gamma_0(\mathbb{E})$ such that

$$\mathrm{zer}\,(\partial f + \partial g) \neq \emptyset. \tag{4.33}$$

We have $f = \gamma\|\cdot\|_1$ and $g = \iota_M$. While it is obvious that $f \in \Gamma_0(\mathbb{E})$, g is lower semicontinuous since M is closed, it is convex since M is convex and proper since $M \neq \emptyset$. Hence, $g \in \Gamma_0(\mathbb{E})$.

For (4.33) we use Theorem 4.3.6. Hence, we have to check that $\mathrm{argmin}\,(f + g) \neq \emptyset$ and $0 \in \mathrm{sri}\,(\mathrm{dom}\, f - \mathrm{dom}\, g)$. For the first condition apply Theorem 4.3.9. Since $\mathrm{dom}\, g = M$ and $\mathrm{dom}\, f = \mathbb{E}$ we have $\mathrm{dom}\, f \cap \mathrm{dom}\, g = M \neq \emptyset$. Furthermore, g is coercive and f is bounded from below (by zero). Therefore, Theorem 4.3.9 implies that $\mathrm{argmin}\,(f + g) \neq \emptyset$.

Last to check is that $0 \in \mathrm{sri}\,(\mathrm{dom}\, f - \mathrm{dom}\, g)$. Since $\mathrm{dom}\, f = \mathbb{E}$ and $\mathrm{dom}\, g = M$ with $M \subset \mathbb{E}$, we have $\mathrm{dom}\, f - \mathrm{dom}\, g = \mathbb{E}$ and therefore $0 \in \mathrm{int}\,(\mathrm{dom}\, f - \mathrm{dom}\, g)$. Hence, by Theorem 4.3.6 we have $\mathrm{zer}\,(\partial f + \partial g) = \mathrm{argmin}\, f + g \neq \emptyset$. □

4.3.2. Fixed-Points of RAAR-$(\mathrm{Id}, 1, \iota_A)$

In this section we cite a result from [75] on the convergence behavior of the original RAAR algorithm in the convex case and the influence of the relaxation parameter β_k. We see that the algorithm converges to *nearest points* for inconsistent (convex) feasibility problems and the results, describing the set of fixed-points, provide insight on the influence of the relaxation parameter. Further results for convex and prox-regular sets are given in [76].

Definition 4.3.11 (Nearest Points, Gap Vector from [75]). *Let $A, B \subset \mathbb{E}$ be closed and convex sets. We denote by $E \subset A$ the* points of A nearest to B, *i.e.,*

$$E := \operatorname*{argmin}_{x \in A} \operatorname{dist}_B(x) \quad \textit{where} \quad \operatorname{dist}_B(x) := \inf_{y \in B} \left\| x - y \right\|_{\mathcal{H}},$$

and likewise by $F \subset B$ the points of B nearest to A. *We define the* gap vector *by*

$$g := P_{\overline{(B-A)}}(0).$$

The following theorem characterizes the fixed-points of the RAAR-sequence (4.14) if both sets A and B are closed and convex.

Theorem 4.3.12 (Theorem 2.2 from [75]). *For $0 < \beta < 1$ the fixed-points of*

$$\mathcal{R}_\beta := \frac{\beta}{2}(R_A R_B + \mathrm{Id}) + (1 - \beta) P_B$$

are given by

$$\operatorname{Fix} \mathcal{R}_\beta = F - \frac{\beta}{1 - \beta} g.$$

Furthermore, for all $u \in \operatorname{Fix}(\mathcal{R}_\beta)$ it holds that:

1. $u = P_B u - \frac{\beta}{1-\beta} g$
2. $P_B u - P_A R_B u = g$
3. $P_B u \in F$ *and* $P_A P_B u \in E$.

Theorem 4.3.12 shows that the RAAR algorithm has fixed points weather or not A and B are disjoint. The fixed-points are those inside the set B nearest to A shifted by the scaled gap vector $\beta/(1-\beta)\, g$ where $g = P_{\mathrm{cl}(B-A)}(0)$. Therefore, the parameter β in the

RAAR algorithm controls the location of the fixed-points and hence the convergence behavior. Although this result only holds in the convex setting, it provides an intuition on the influence of β. Recall that B typically takes the role of the measurement set and the gap vector points away from B in the direction of A. Therefore, the parameter β controls how close the fixed-points of the iteration lie to B. For a more detailed discussion we refer to [75].

4.4. A Proximity Operator for Tight Frames

Before we analyze the inexact algorithm, we present a result on the firmly non-expansiveness of the involved operator. More precisely, we will prove that for tight frames T the operator

$$P_{S_{T,\gamma}} = T^* \operatorname{prox}_{\gamma\|\cdot\|_1} T$$

is a proximity operator with respect to a proper, lower semi-continuous, convex function. This immediately implies the firmly non-expansiveness by Proposition B.1.8. Recall that for tight frames we have $T^\dagger = T^*$ and $T^*T = \nu\mathrm{Id}$ where ν denotes the frame constant, i.e. the upper frame bound which is identical to the lower frame bound for tight frames.

Proposition 4.4.1 (Proposition 4.a from [88]). *Let* $f \in \Gamma_0(\mathbb{R}^d)$. *Then*

$$y \in \partial_f(x) \Longleftrightarrow x = \operatorname{prox}_f(x + y).$$

We will use this result to show that $P_{S_{T,\gamma}}$ is a proximity operator, i.e., we will need to prove that there exists a $f \in \Gamma_0(\mathbb{R}^d)$ such that $y \in \partial_f(x)$ whenever $x = \nu T^* S_\gamma T(x + y)$.

Definition 4.4.2 (Graph of set-valued mappings, [5]). *Let* $H : \mathbb{E} \rightrightarrows \mathbb{E}$. *The* graph of *H is defined by*

$$\operatorname{gra} H := \{(x, u) \in \mathbb{E} \times \mathbb{E} \mid u \in H(x)\}.$$

Definition 4.4.3 (Definition from Section 2 in [98]). *Let $H : \mathbb{E} \rightrightarrows \mathbb{E}$ and $n \geq 2$ with $n \in \mathbb{N}$. Then H is* n-cyclically monotone *if for all $(x_1, \ldots, x_{n+1}) \in \mathbb{E}^{n+1}$ and $(u_1, \ldots, u_n) \in \mathbb{E}^n$ it holds that*

$$\left.\begin{array}{c} (x_1, u_1) \in \operatorname{gra} H \\ \vdots \\ (x_n, u_n) \in \operatorname{gra} H \\ x_{n+1} = x_1 \end{array}\right\} \Longrightarrow \sum_{j=1}^{n} \left\langle x_{j+1} - x_j, u_j \right\rangle \leq 0. \tag{4.34}$$

If (4.34) holds for all $n \geq 2$ then H is cyclically monotone*. If H is cyclically monotone and* $\operatorname{gra} H$ *cannot be enlarged without violating this property, H is called* maximally cyclically monotone*.*

Theorem 4.4.4 (Rockafellar, [98]. Theorem 22.14 from [5]). *Let $H : \mathbb{E} \rightrightarrows \mathbb{E}$. Then H is maximally cyclically monotone if and only if there exists $f \in \Gamma_0(\mathbb{E})$ such that $H = \partial f$.*

Theorem 4.4.5 (Minty, [86]. Theorem 21.1 from [5]). *Let $H : \mathbb{E} \rightrightarrows \mathbb{E}$ be monotone. Then H is maximally monotone if and only if* $\operatorname{ran}(\mathrm{Id} + H) = \mathbb{E}$.

We now prove the main result of this section using the aforementioned results.

Theorem 4.4.6. *Let $T \in \mathbb{R}^{n \times d}$ (where $n \geq d$) be a tight frame, i.e., $T^*T = \nu \mathrm{Id}$. Then the operator defined by*

$$P_{S_{T,\gamma}} := T^* S_\gamma T \tag{4.35}$$

is the proximity operator of a proper, lower semi-continuous and convex function.

Proof. We define the mapping $H : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ by

$$y \in H(x) :\Leftrightarrow x = \nu T^* S_\gamma T(x + y). \tag{4.36}$$

Hence, by Proposition 4.4.1, the mapping $P_{S_{T,\gamma}}$ is a proximity operator if $H = \partial f$ for some $f \in \Gamma_0(\mathbb{R}^d)$. We thus need to show that H is maximally cyclically monotone which will yield the claim by using Theorem 4.4.4.

1. Note that we can rewrite (4.36) as follows:

$$\begin{aligned} y \in H(x) &\Longleftrightarrow x = \nu T^* S_\gamma T(x+y) \\ &\Longleftrightarrow \nu T^* T x = \nu T^* S_\gamma T(x+y) \\ &\Longleftrightarrow \exists\, u \in \ker T^* : u + Tx = S_\gamma T(x+y) \end{aligned} \tag{4.37}$$

where we used $T^*T = \nu\mathrm{Id}$. Recall the definition of the soft-threshold operator

$$(S_\gamma x)[j] = \begin{cases} x[j] - \gamma, & x[j] > \gamma \\ x[j] + \gamma, & x[j] < -\gamma \\ 0, & x[j] \in (-\gamma, \gamma). \end{cases} \tag{4.38}$$

We denote $x' := x + y$ or $x = x' - y$ respectively and rewrite (4.37) as

$$y \in H(x) \Leftrightarrow u + T(x' - y) = S_\gamma T x' \tag{4.39}$$

for some $u \in \ker T^*$. Using (4.38) we can write

$$S_\gamma T x' = T x' - t \tag{4.40}$$

where $t = (t[j])_{j=1}^n$ with

$$t[j] := \begin{cases} \gamma, & (Tx')[j] \geq \gamma \\ -\gamma, & (Tx')[j] \leq -\gamma \\ (Tx')[j], & (Tx')[j] \in (-\gamma, \gamma). \end{cases}$$

Moreover, using (4.39) this yields

$$u + Tx' - Ty = Tx' - t,$$

i.e.,

$$t = Ty - u \tag{4.41}$$

or $Ty = t + u$ and thus

$$y = \nu T^* T y = \nu T^*(t+u) = \nu T^* t. \tag{4.42}$$

2. For the next step of the proof, let $(x_1, y_1), (x_2, y_2) \in \operatorname{gra} H$, i.e. $y_1 \in H(x_1), y_2 \in H(x_2)$. Then there are $t_1, t_2 \in \mathbb{R}^n$ and $u_1, u_2 \in \ker T^*$ such that by (4.40)

   $$\begin{aligned} S_\gamma T(x_1 + y_1) &= T(x_1 + y_1) - t_1, \\ S_\gamma T(x_2 + y_2) &= T(x_2 + y_2) - t_2, \end{aligned}$$

   with $t_1 = Ty_1 - u_1, t_2 = Ty_2 - u_2$. Now, let $x_1' := x_1 + y_1$ and $x_2' := x_2 + y_2$. We will show that

   $$\left\langle S_\gamma T(x_1 + y_1), t_2 - t_1 \right\rangle \le 0. \tag{4.43}$$

   First, observe that for the j-th component of $t_2 - t_1$ we have

   $$(t_2 - t_1)[j] = \left\{ \begin{array}{ll} 2\gamma, & (Tx_2')[j] \ge \gamma \wedge (Tx_1')[j] \le -\gamma, \\ \gamma - (Tx_2')[j], & (Tx_2')[j] \ge \gamma \wedge (Tx_1')[j] \in (-\gamma, \gamma), \\ 0, & (Tx_2')[j] \ge \gamma \wedge (Tx_1')[j] \ge \gamma, \\ \hline \gamma + (Tx_2')[j], & (Tx_2')[j] \in (-\gamma, \gamma) \wedge (Tx_1')[j] \le -\gamma, \\ [Tx_2'][j] - [Tx_1'[j], & (Tx_2')[j] \in (-\gamma, \gamma) \wedge (Tx_1')[j] \in (-\gamma, \gamma), \\ -\gamma + (Tx_2')[j], & (Tx_2')[j] \in (-\gamma, \gamma) \wedge (Tx_1')[j] \ge \gamma, \\ \hline 0, & (Tx_2')[j] \le -\gamma \wedge (Tx_1')[j] \le -\gamma, \\ -\gamma - (Tx_1')[j], & (Tx_2')[j] \le -\gamma \wedge (Tx_1')[j] \in (-\gamma, \gamma), \\ -2\gamma, & (Tx_2')[j] \le -\gamma \wedge (Tx_1')[j] \ge \gamma. \end{array} \right.$$

   Therefore, if $(T(x_1 + y_1))[j] = (Tx_1')[j] \ge \gamma$ we have $(S_\gamma T(x_1'))[j] \ge 0$ and

   $$(S_\gamma Tx_1')[j] \cdot (t_2 - t_1)[j] = \underbrace{(S_\gamma Tx_1')[j]}_{\ge 0} \cdot \underbrace{(t_2[j] - \gamma)}_{\le 0} \le 0.$$

Similarly, for $(Tx_1')[j] \leq -\gamma$ we have $(S_\gamma Tx_1')[j] \leq 0$ and

$$(S_\gamma Tx_1')[j] \cdot (t_2 - t_1)[j] = \underbrace{(S_\gamma Tx_1')[j]}_{\leq 0}\,\underbrace{(t_2[j] + \gamma)}_{\geq 0} \leq 0.$$

Finally, for $(Tx_1')[j] \in (-\gamma, \gamma)$ we have $(S_\gamma Tx_1')[j] = 0$ and therefore

$$(S_\gamma Tx_1')[j] \cdot (t_2 - t_1)[j] = 0.$$

To conclude, it holds that

$$\left\langle S_\gamma T(x + y), t_2 - t_1 \right\rangle = \sum_{j=1}^{n} (S_\gamma Tx_1')[j] \cdot (t_2 - t_1)[j] \leq 0.$$

3. We now use this fact to prove that H is maximally cyclically monotone. Therefore let $n \in \mathbb{N}$ with $n \geq 2$ be arbitrary and choose $(x_i, y_i) \in \operatorname{gra} H$ for $i = 1, \ldots, n$ and define $x_{n+1} := x_1$. Moreover, denote $t_i := T(x_i + y_i) - S_\gamma T(x_i + y_i)$. Then by using (4.42) we obtain

$$\begin{aligned}\sum_{i=1}^{n} \langle x_{i+1} - x_i, y_i \rangle &\overset{(4.42)}{=} \sum_{i=1}^{n} \langle x_{i+1} - x_i, \nu T^* t_i \rangle \\ &= \nu \sum_{i=1}^{n} \langle Tx_{i+1} - Tx_i, t_i \rangle .\end{aligned}$$

Let $u_i := Ty_i - t_i$. Then $u_i \in \ker T^*$, and we have

$$\begin{aligned}\sum_{i=1}^{n} \langle x_{i+1} - x_i, y_i \rangle &= \nu \sum_{i=1}^{n} \langle (Tx_{i+1} + u_{i+1}) - (Tx_i + u_i) - u_{i+1} + u_i, t_i \rangle \\ &\overset{(4.39)}{=} \nu \sum_{i=1}^{n} \left(\left\langle S_\gamma T(x_{i+1} + y_{i+1}) - S_\gamma T(x_i + y_i), t_i \right\rangle + \langle u_i - u_{i+1}, t_i \rangle \right).\end{aligned}$$

Using (4.43), we obtain the estimate

$$\nu \sum_{i=1}^{n} \left\langle S_\gamma T(x_{i+1} + y_{i+1}) - S_\gamma T(x_i + y_i), t_i \right\rangle = \nu \sum_{i=1}^{n} \left\langle S_\gamma T(x_{i+1} + y_{i+1}), t_i - t_{i+1} \right\rangle \leq 0$$

for the first sum where $t_{n+1} = t_1$. For the second sum we obtain

$$\begin{aligned}
\nu \sum_{i=1}^{n} \langle u_i - u_{i+1}, t_i \rangle &\overset{(4.41)}{=} \nu \sum_{i=1}^{n} \langle u_i - u_{i+1}, Ty_i - u_i \rangle \\
&= \nu \sum_{i=1}^{n} \langle u_i - u_{i+1}, -u_i \rangle + \nu \sum_{i=1}^{n} \langle u_i - u_{i+1}, Ty_i \rangle \\
&= \nu \sum_{i=1}^{n} \langle u_i, -u_i \rangle + \nu \sum_{i=1}^{n} \langle u_{i+1}, u_i \rangle + \nu \sum_{i=1}^{n} \Big\langle \underbrace{T^*(u_i - u_{i+1})}_{=0}, y_i \Big\rangle \\
&= -\nu \sum_{i=1}^{n} \|u_i\|^2 + \nu \sum_{i=1}^{n} \langle u_{i+1}, u_i \rangle \leq 0
\end{aligned}$$

since $\langle u_{i+1}, u_i \rangle \leq {}^1\!/\!{}_2(\|u_i\|^2 + \|u_{i+1}\|^2)$. This establishes

$$\sum_{i=1}^{n} \langle x_{i+1} - x_i, y_i \rangle \leq 0$$

for all $n \in \mathbb{N}, n \geq 2$ where $(x_i, y_i) \in \operatorname{gra} H$ and $x_{n+1} = x_1$. Therefore, by definition, H is cyclically monotone.

4. Finally, we show that H is maximally cyclically monotone. Using Theorem 4.4.5 we need to show that H is monotone and that $\operatorname{ran}(\mathrm{Id} + H) = \mathbb{R}^d$ since we already established that H is cyclically monotone. In particular, H is monotone, since it is cyclically 2-monotone. For arbitrary $z \in \mathbb{R}^d$ choose

$$x := \nu T^* S_\gamma T z \in \mathbb{R}^d.$$

By (4.36) with $z = x + (z - x)$ we have

$$z - x \in H(x),$$

i.e., $z \in H(x) + x$. Therefore, $\operatorname{ran}(\mathrm{Id} + H) = \mathbb{R}^d$ with H monotone. Hence, H is maximally cyclically monotone.

By Theorem 4.4.4 there exists a function $f \in \Gamma_0(\mathbb{R}^d)$ such that $H = \partial f$. Hence, by Proposition 4.4.1 and with (4.36), this yields the claim. □

Remark 4.4.7. Note that although the proximity operator $P_{S_{T,\gamma}}$ used in the inexact RAAR-(T, γ, ℓ_1) is not the same as $\operatorname{prox}_{\gamma\|T\cdot\|_1}$ for general tight frames, the two coincide for orthogonal bases, i.e., when T is a unitary mapping. If it holds that $TT^* = \nu \mathrm{Id}$, Proposition 4.2.9 shows that the corresponding proximity operator can be explicitly calculated. It is an open problem if this is also true if one has $T^*T = \nu \mathrm{Id}$ which is the case here.

4.5. Boundedness and Cesàro-Convergence of RAAR-(T, γ, ℓ_1)

While we have studied convex analogues of the proposed algorithm in the last sections, we now analyze the convergence behavior in the non-convex setting. The following results are partially published in [72]. Here, we only analyze the inexact version RAAR-(T, γ, ℓ_1) from Definition 4.2.12 which will be used in practice. However, similar results can be obtained for the exact version as well since in the exact version one could use the fact that the mapping $\mathrm{prox}_{\gamma\|T\cdot\|_1}$ is firmly non-expansive.

Using a similar property we are able to prove the boundedness of the sequence and a result on the fixed-points of the inexact RAAR-(T, γ, ℓ_1) algorithm. For frames, the upper bound on the sequence will thus contain both the lower and the upper frame bound in the estimate. However, using the same technique as for the lower bound of the sequence, we can still establish an estimate that is independent of the upper frame bound. For tight frames, the lower and upper frame bound coincide and we obtain sharper estimates. All occurring norms in the following proofs will be Euclidean norms if not otherwise noted.

Lemma 4.5.1. *Let $T : \mathbb{R}^d \to \mathbb{R}^n$ be the analysis operator of a frame, i.e., there are constants $c_2 \geq c_1 > 0$ such that*

$$c_1 \|x\| \leq \|Tx\| \leq c_2 \|x\| \qquad \forall\, x \in \mathbb{R}^d \tag{4.44}$$

then for all $x \in \mathbb{R}^d$ it holds that

$$\left\|R_{S_{T,\gamma}} x\right\| \leq \frac{c_2}{c_1} \|x\| .$$

Proof. By definition we have

$$
\begin{aligned}
\left\|R_{S_{T,\gamma}}x\right\| = \left\|2P_{S_{T,\gamma}}x - x\right\| = \left\|2T^{\dagger}S_{\gamma}Tx - x\right\| &\overset{(4.44)}{\leq} \frac{1}{c_1}\left\|2S_{\gamma}Tx - Tx\right\| \\
&\leq \frac{1}{c_1}\left(\sum_{j=1}^{n}\left|2S_{\gamma}(Tx)[j] - (Tx)[j]\right|^2\right)^{\frac{1}{2}} . \qquad (4.45)
\end{aligned}
$$

For the summands we have for $(Tx)[j] > \gamma$ that

$$
\left|2S_{\gamma}(Tx)[j] - (Tx)[j]\right| = \left|2[(Tx)[j] - \gamma] - (Tx)[j]\right| = \left|(Tx)[j] - 2\gamma\right| \leq \left|(Tx)[j]\right| .
$$

Similarly, we obtain for $(Tx)[j] < -\gamma$ that

$$
\left|2S_{\gamma}(Tx)[j] - (Tx)[j]\right| = \left|2[(Tx)[j] + \gamma] - (Tx)[j]\right| = \left|(Tx)[j] + 2\gamma\right| \leq \left|(Tx)[j]\right| .
$$

For $\left|(Tx)[j]\right| \leq \gamma$ we have $S_{\gamma}(Tx)[j] = 0$ and therefore

$$
\left|2S_{\gamma}(Tx)[j] - (Tx)[j]\right| = \left|0 - (Tx)[j]\right| = \left|(Tx)[j]\right|
$$

and hence, $\left|2S_{\gamma}(Tx)[j] - (Tx)[j]\right| \leq \left|(Tx)[j]\right|$. Plugging this into (4.45) yields

$$
\left\|R_{S_{T,\gamma}}x\right\| \leq \frac{1}{c_1}\left(\sum_{j=1}^{n}\left|(Tx)[j]\right|^2\right)^{\frac{1}{2}} = \frac{1}{c_1}\|Tx\| \overset{(4.44)}{\leq} \frac{c_2}{c_1}\|x\| .
$$

□

Using this result we can prove the boundedness the inexact RAAR-(T, γ, ℓ_1) algorithm in the discrete, finite dimensional setting. Using that estimate we will be able to prove convergence of the Cesàro-sequence generated by the iterates of that algorithm.

Theorem 4.5.2. *Suppose $T : \mathbb{R}^d \to \mathbb{R}^n$ is analysis operator of a frame with frame bounds $c_2 \geq c_1 > 0$ and denote the measurements by $m \in \mathbb{R}^d_+$. Then for all $x^{(k)}$, $k \in \mathbb{N}$ with $x^{(0)} \in M$, the sequence generated by*

$$
x^{(k+1)} = \frac{\beta_k}{2}\left(R_{S_{T,\gamma}}R_M + \mathrm{Id}\right)x^{(k)} + (1 - \beta_k)P_M x^{(k)} \qquad (4.46)
$$

is bounded by

$$\max\left\{0, \|m\| - 3\beta_k \gamma \frac{\sqrt{n}}{c_1}\right\} \leq \left\|x^{(k)}\right\| \leq \frac{\beta_k}{c_1}\left(c_2 \|m\| + \gamma \sqrt{n}\right) + (1-\beta_k)\|m\|. \tag{4.47}$$

Proof. By definition of the soft-threshold operator we have for a frame, i.e. when (4.44) holds, the bound

$$\left\|x - P_{S_{T,\gamma}} x\right\| = \frac{1}{c_1}\left\|Tx - S_\gamma Tx\right\| \leq \frac{\gamma\sqrt{n}}{c_1}, \tag{4.48}$$

since every component of the vector Tx is at most changed by γ. This means, for every component we have

$$\left|Tx[j] - (S_\gamma Tx)[j]\right|^2 \leq \gamma^2.$$

The definition of the 2-norm yields (4.48). By the triangle inequality it follows that

$$\begin{aligned}\left\|x^{(k+1)}\right\| &= \left\|\frac{\beta_k}{2}\left(R_{S_{T,\gamma}} R_M + \mathrm{Id}\right)x^{(k)} + (1-\beta_k) P_M x^{(k)}\right\| \\ &\leq \frac{\beta_k}{2}\left\|\left(R_{S_{T,\gamma}} R_M + \mathrm{Id}\right)x^{(k)}\right\| + (1-\beta_k)\left\|P_M x^{(k)}\right\|.\end{aligned}$$

By assumption we have $\|x\| = \|m\|$ for all $x \in M$, therefore $\left\|P_M x^{(k)}\right\| = \|m\|$, and hence

$$\left\|x^{(k+1)}\right\| \leq \frac{\beta_k}{2}\left\|\left(R_{S_{T,\gamma}} R_M + \mathrm{Id}\right)x^{(k)}\right\| + (1-\beta_k)\|m\|.$$

We can rewrite the first term in the following way

$$\begin{aligned}(R_{S_{T,\gamma}} R_M + \mathrm{Id})x^{(k)} &= \left(R_{S_{T,\gamma}}(R_M + \mathrm{Id}) + \mathrm{Id} - R_{S_{T,\gamma}}\right)x^{(k)} \\ &= R_{S_{T,\gamma}}\left(R_M + \mathrm{Id}\right)x^{(k)} + \left(\mathrm{Id} - R_{S_{T,\gamma}}\right)x^{(k)}.\end{aligned}$$

Plugging this in and using the triangle inequality we obtain

$$\left\|x^{(k+1)}\right\| \leq \frac{\beta_k}{2}\left\|R_{S_{T,\gamma}}\left(R_M + \mathrm{Id}\right)x^{(k)}\right\| + \left\|\left(\mathrm{Id} - R_{S_{T,\gamma}}\right)x^{(k)}\right\| + (1-\beta_k)\|m\|. \tag{4.49}$$

Furthermore, note that

$$R_{S_{T,\gamma}}\left(R_M + \mathrm{Id}\right) = R_{S_{T,\gamma}}\left(2P_M - \mathrm{Id} + \mathrm{Id}\right) = 2R_{S_{T,\gamma}}P_M$$

and for the second term on the right hand side in (4.49) we have by (4.48)

$$\left\|\left(\mathrm{Id} - R_{S_{T,\gamma}}\right)x^{(k)}\right\| = \left\|\left(I - (2P_{S_{T,\gamma}} - I\right)x^{(k)}\right\| = 2\left\|x^{(k)} - P_{S_{T,\gamma}}x^{(k)}\right\| \leq 2\frac{\gamma\sqrt{n}}{c_1}.$$

This yields the bound

$$\left\|x^{(k+1)}\right\| \leq \beta_k\left\|R_{S_{T,\gamma}}P_M x^{(k)}\right\| + \beta_k\frac{\gamma\sqrt{n}}{c_1} + (1-\beta_k)\left\|m\right\|. \tag{4.50}$$

Using Lemma 4.5.1 we can estimate

$$\left\|R_{S_{T,\gamma}}P_M x^{(k)}\right\| \leq \frac{c_2}{c_1}\left\|P_M x^{(k)}\right\| = \frac{c_2}{c_1}\left\|m\right\|$$

and obtain

$$\begin{aligned}\left\|x^{(k+1)}\right\| &\leq \beta_k\frac{c_2}{c_1}\left\|m\right\| + \beta_k\frac{\gamma\sqrt{n}}{c_1} + (1-\beta_k)\left\|m\right\| \\ &\leq \frac{\beta_k}{c_1}\left(c_2\left\|m\right\| + \gamma\sqrt{n}\right) + (1-\beta_k)\left\|m\right\|.\end{aligned}$$

The next step is to prove the lower bound. Elementary calculations yield

$$\begin{aligned}\left\|x^{(k+1)}\right\| &= \left\|\frac{\beta_k}{2}\left(R_{S_{T,\gamma}}R_M + \mathrm{Id}\right)x^{(k)} + (1-\beta_k)P_M x^{(k)}\right\| \\ &= \left\|\frac{\beta_k}{2}\left(\left(2P_{S_{T,\gamma}} - \mathrm{Id}\right)(2P_M - \mathrm{Id}) + \mathrm{Id}\right)x^{(k)} + (1-\beta_k)P_M x^{(k)}\right\| \\ &= \left\|\frac{\beta_k}{2}\left(4P_{S_{T,\gamma}}P_M x^{(k)} - 2P_{S_{T,\gamma}}x^{(k)} - 2P_M x^{(k)} + 2x^{(k)}\right) + (1-\beta_k)P_M x^{(k)}\right\| \\ &= \left\|\beta_k\left(2P_{S_{T,\gamma}}P_M x^{(k)} - P_{S_{T,\gamma}}x^{(k)} - P_M x^{(k)}\right) + x^{(k)} + (1-\beta_k)P_M x^{(k)}\right\| \\ &= \left\|P_M x^{(k)} + 2\beta_k P_{S_{T,\gamma}}P_M x^{(k)} - \beta_k P_{S_{T,\gamma}}x^{(k)} - 2\beta_k P_M x^{(k)} + \beta_k x^{(k)}\right\| \\ &= \left\|P_M x^{(k)} + 2\beta_k\left(P_{S_{T,\gamma}}P_M - P_M\right)x^{(k)} + \beta_k\left(\mathrm{Id} - P_{S_{T,\gamma}}\right)x^{(k)}\right\| \\ &= \left\|\left(\mathrm{Id} + 2\beta_k\left(P_{S_{T,\gamma}} - \mathrm{Id}\right)\right)P_M x^{(k)} + \beta_k\left(\mathrm{Id} - P_{S_{T,\gamma}}\right)x^{(k)}\right\|.\end{aligned}$$

Using the inverse triangle inequality of the form

$$\|a+b\| \geq \|a\| - \|b\|$$

we obtain by (4.48) the lower bound

$$\begin{aligned} \left\|x^{(k+1)}\right\| &\geq \left\|\left(\mathrm{Id} + 2\beta_k \left(P_{S_{T,\gamma}} - \mathrm{Id}\right)\right) P_M x^{(k)}\right\| - \beta_k \left\|\left(\mathrm{Id} - P_{S_{T,\gamma}}\right) x^{(k)}\right\| \\ &\geq \left\|P_M x^{(k)}\right\| - 2\beta_k \frac{\gamma\sqrt{n}}{c_1} - \beta_k \frac{\gamma\sqrt{n}}{c_1} = \|m\| - 3\beta_k \frac{\gamma\sqrt{n}}{c_1}. \end{aligned}$$

Since $\|\cdot\|$ must be non-negative, we have

$$\left\|x^{(k+1)}\right\| \geq \max\left\{0,\ \|m\| - 3\beta_k \frac{\gamma\sqrt{n}}{c_1}\right\}.$$

Finally, we prove that there is no fixed-point in M. Consider $x^{(k)} \in M$, i.e., $P_M x^{(k)} = x^{(k)}$. Then we have

$$\begin{aligned} x^{(k+1)} - x^{(k)} &= \frac{\beta_k}{2}\left(R_{S_{T,\gamma}} R_M + \mathrm{Id}\right) x^{(k)} + (1-\beta_k) P_M x^{(k)} - x^{(k)} \\ &= \frac{\beta_k}{2}\left(R_{S_{T,\gamma}} (2P_M - \mathrm{Id}) x^{(k)} + x^{(k)}\right) + (1-\beta_k)(P_M - \mathrm{Id}) x^{(k)} \\ &= \frac{\beta_k}{2}\left(R_{S_{T,\gamma}} P_M - \mathrm{Id}\right) x^{(k)} + \left(\frac{\beta_k}{2} R_{S_{T,\gamma}} + (1-\beta_k)\,\mathrm{Id}\right) \underbrace{(P_M - \mathrm{Id})\, x^{(k)}}_{=0,\ \text{if } x^{(k)} \in M}. \end{aligned}$$

If there was a fixed-point $x \in M$ then

$$x^{(k+1)} - x^{(k)} = \frac{\beta_k}{2}\left(R_{S_{T,\gamma}} P_M - \mathrm{Id}\right) x^{(k)} = \beta_k \left(P_{S_{T,\gamma}} - \mathrm{Id}\right) x^{(k)}$$

must vanish. But for $\left\|x^{(k)}\right\| > 0$ and $\beta_k \geq \varepsilon > 0$ this is not possible. Therefore, no fixed-point $x \in M$ exists. □

Remark 4.5.3. Using the same technique for the upper bound as we used for the lower bound, we can achieve an estimate that is independent of the upper frame bound, i.e.,

$$\max\left\{0, \|m\| - 3\beta_k \gamma \frac{\sqrt{n}}{c_1}\right\} \leq \left\|x^{(k)}\right\| \leq \|m\| + 3\beta_k \gamma \frac{\sqrt{n}}{c_1}.$$

○

Corollary 4.5.4 (Theorem 3.1 from [72]). *If T is analysis operator of a tight frame, i.e. $c_2 = c_1$, the sequence generated by the inexact RAAR-(T, γ, ℓ_1) is bounded by*

$$\max\left\{0, \|m\| - 3\beta_k \gamma \frac{\sqrt{n}}{c_1}\right\} \leq \left\|x^{(k)}\right\| \leq \|m\| + \beta_k \frac{\gamma\sqrt{n}}{c_1}. \tag{4.51}$$

Remark 4.5.5. Since the sequence generated by the algorithm is bounded, see Lemma 4.5.4, this guarantees the existence of accumulation points, i.e., there exists a converging subsequence. Moreover, we this implies that the Cesàro sequence $z^{(k)} := {}^1\!/\!_k \sum_{j=1}^{k} x^{(k)}$ converges. ○

Remark 4.5.6. Corollary 4.5.4 provides an intuition on how to choose the thresholding parameter γ. If we choose an adaptive γ_k with $\gamma_k \to 0$ as $k \to \infty$, then $P_{S_{T,\gamma_k}}$ becomes the identity (modulo a frame constant). In the case of a Parseval frame with frame constant $\nu = 1$ this would imply that

$$x^{(k+1)} - x^{(k)} = \beta_k \left(P_{S_{T,\gamma_k}} - \mathrm{Id}\right) x^{(k)} \to 0$$

if $P_{S_{T,\gamma_k}} \to \mathrm{Id}$ which is the case for $\gamma_k \to 0$, since

$$(S_\gamma x)[j] = \begin{cases} x[j] - \gamma_k, & x[j] > \gamma_k, \\ x[j] + \gamma_k, & x[j] < -\gamma_k, \\ 0, & \text{otherwise} \end{cases} \xrightarrow{\gamma_k \to 0} \begin{cases} x[j], & x[j] > 0, \\ x[j], & x[j] < 0, \\ 0, & x[j] = 0 \end{cases} = \mathrm{Id}.$$

Corollary 4.5.7. *The Cesàro sequence $z^{(k)} = {}^1\!/\!_k \sum_{j=1}^{k} x^{(j)}$ where $x^{(j)}$ are the iterates obtained by (4.46) with $x^{(0)} \in M$ converges.*

Proof. For $\beta_k \in (0, 1)$ bounded by $\beta_{\max} := \max_{k \geq 1} \beta_k$ and use the estimate (4.47) to obtain

$$\left\|x^{(k+1)}\right\| \leq \frac{c_2}{c_1} \|m\| + \frac{\beta_{\max}\gamma\sqrt{n}}{c_1}$$

where we used that ${}^{c_2}/_{c_1} \geq 1$. We can now estimate

$$\begin{aligned}
\left\|z^{(k+1)} - z^{(k)}\right\| &= \left\|\frac{1}{k+1}\sum_{j=1}^{k+1} x^{(j)} - \frac{1}{k}\sum_{j=1}^{k} x^{(j)}\right\| \\
&= \left\|\frac{1}{k+1}x^{(k+1)} + \left(\frac{1}{k+1} - \frac{1}{k}\right)\sum_{j=1}^{k} x^{(j)}\right\| \\
&\leq \frac{1}{k+1}\left\|x^{(k+1)}\right\| + \frac{1}{k(k+1)}\left\|\sum_{j=1}^{k} x^{(j)}\right\| \\
&\leq \frac{1}{k+1}\left(\frac{c_2}{c_1}\|m\| + \beta_{k+1}\frac{\gamma\sqrt{n}}{c_1}\right) + \frac{1}{k+1}\left(\frac{c_2}{c_1}\|m\| + \beta_{\max}\frac{\gamma\sqrt{n}}{c_1}\right) \\
&\leq \frac{2}{k+1}\left(\frac{c_2}{c_1}\|m\| + \beta_{\max}\frac{\gamma\sqrt{n}}{c_1}\right)
\end{aligned}$$

and hence, $\left\|z^{(k+1)} - z^{(k)}\right\| \to 0$ for $k \to \infty$. □

4.6. Generalizations to other Threshold Functions

We have seen that general proximity operators of proper, lower semi-continuous, convex functions can be used within the exact RAAR-(T, γ, ℓ_1) algorithm. From a theoretical perspective, all these proximity operators have the necessary properties, i.e., are firmly non-expansive. However, for a fast and stable numerical implementation, it is crucial that the minimization problem

$$\operatorname{prox}_{\gamma f}(x) = \operatorname*{argmin}_{y \in \mathbb{R}^d} f(y) + \frac{1}{2\gamma}\left\|x - y\right\|^2$$

can be solved efficiently. Ideally, the solution is explicitly given as a point-wise operation. The following result from [19] shows that a wide class of such *shrinkage mappings* are indeed proximity operators. The proof of this result is given in the appendix of [19].

Theorem 4.6.1 (Theorem 1 from [19]). *Let $v_\gamma : \mathbb{R}_+ \to \mathbb{R}_+$ be continuous with*

$$v_\gamma(x) = \begin{cases} 0, & x \leq \gamma \\ \leq x, & x \in [\gamma, \infty) \end{cases}$$

such that v_γ is strictly increasing on $[\gamma, \infty)$. Furthermore, define

$$V(x)[j] := v_\gamma(\left|x[j]\right|)\,\mathrm{sign}\,(x[j]).$$

Then

$$V_\gamma(x) = \mathrm{prox}_{\gamma G}(x) \qquad \textit{where} \quad G(x) = \sum_{j=1}^{d} g(x[j])$$

with g even, non-decreasing and continuous on $[0, \infty)$, differentiable on $(0, \infty)$ and non-differentiable at 0 with $\partial g(0) = [-1, 1]$. Moreover, if $x - v_\gamma(x)$ is non-increasing on $[\gamma, \infty)$, then g is concave on $[0, \infty)$ and G satisfies the triangle inequality.

Remark 4.6.2. The soft-threshold operator S_γ can be rewritten as

$$(S_\gamma x)[j] = \max\left\{\left|x[j]\right| - \gamma, 0\right\}\mathrm{sign}\,(x[j]),$$

i.e., $v_\gamma(\left|x[j]\right|) = \max\left\{\left|x[j]\right| - \gamma, 0\right\}$. We know that $G(x) = \sum_{j=1}^{d} g(x[j]) = \sum_{j=1}^{d}\left|x[j]\right|$ in this case. Note that Theorem 4.6.1 is not constructive in the sense that given a shrinkage mapping v_γ one could reconstruct the penalty function G. However, one does know by the result that the shrinkage mapping V is the proximity operator of a sufficiently well behaved function. Furthermore, in some cases it is possible to explicitly derive the penalty function. ○

Example 4.6.3. Introducing a parameter p, a class of shrinkage mappings is given by

$$v_\gamma^p(x) = \max\left\{x - \gamma^{2-p}\,x^{p-1}, 0\right\}$$

for which the soft-thresholding is given by choosing $p = 1$ and hard-thresholding can be obtained as the limiting case $p \to \infty$. In [19], a new shrinkage mapping is introduced, called *smooth-hard shrinkage* defined for $x \in \mathbb{R}_+$ by

$$v_\gamma^{\mathrm{SH}}(x) = \begin{cases} x\exp\left(-\frac{\alpha}{(\mathrm{e}^{x-\gamma}-1)^2}\right), & x \geq \gamma \\ 0, & \text{else} \end{cases} \tag{4.52}$$

with parameter $\alpha > 0$. The motivation of smooth-hard shrinkage is the conjecture that the discontinuity of the hard-threshold operator leads to inferior reconstruction results. ○

Based on the observation that constructions following Theorem 4.6.1 lead to proximity operators, we propose the following algorithm.

Definition 4.6.4 (Inexact RAAR-(T, γ, V)). *Let $T : \mathbb{R}^d \to \mathbb{R}^n$ be a linear frame transform, i.e., T maps a vector $x \in \mathbb{R}^d$ onto its frame coefficients $y = Tx \in \mathbb{R}^n$ for $n \geq d$. Denote by $T^\dagger$ the pseudo-inverse of T. Furthermore, let $\beta_k > 0$ for all $k \in \mathbb{N}$. Consider the measurement set*

$$M = \left\{x \in \mathbb{C}^d \mid |Ux|_\circ = m\right\}$$

with a unitary transform $U : \mathbb{C}^d \to \mathbb{C}^d$ and denote by

$$P_{T,V_\gamma} := T^\dagger V_\gamma T$$

with

$$V(x)[j] := v_\gamma(\left|x[j]\right|)\,\mathrm{sign}\,(x[j])$$

such that v_γ fulfills the prerequisites of Theorem 4.6.1. Then we define the RAAR-(T, γ, V) iteration *for a given $x^{(0)} \in \mathbb{R}^d$ by*

$$x^{(k+1)} = \frac{\beta_k}{2}\left(R_{S_{T,V_\gamma}} R_M + \mathrm{Id}\right) x^{(k)} + (1 - \beta_k)\, P_M x^{(k)} \tag{4.53}$$

where $R_M := 2P_M - \mathrm{Id}$ and

$$R_{S_{T,V_\gamma}} := 2P_{S_{T,V_\gamma}} - \mathrm{Id} = 2T^\dagger V_\gamma T - \mathrm{Id}.$$

In Chapter 5 we will compare the performance of the algorithm given by (4.53) for soft-thresholding and smooth-hard shrinkage. The main advantage of such shrinkage mappings is that they offer a flexibility of the type of penalty. Furthermore, they all lead to simple point-wise operations which make them a favorable choice of proximity operator.

4.7. Numerical Implementation

This section deals with the numerical implementation of the projection onto the non-convex measurement set. For details on the shearlet transform we refer to Chapter 3.2. We are first and foremost concerned with the numerical implementation of the projection operator P_M. The thresholding operator is easily implemented as a fast point-wise operation on a vector $x \in \mathbb{R}^d$. The projection onto the set

$$M = \left\{ x \in \mathbb{R}^d \mid |Ux|_\circ = m \right\}$$

is given by $P_M x = U^{-1} y$ with

$$y[j] = \begin{cases} m[j] \frac{Ux[j]}{|Ux[j]|}, & \text{if } |Ux|\,[j] \neq 0 \\ m[j] \exp(\mathrm{i}\varphi), & \text{if } |Ux|\,[j] = 0 \end{cases}$$

for some $\varphi \in [0, 2\pi)$ where one typically choses $\varphi = 0$. Trying to numerically implement this projection, one is faced with the situation that $|Ux|_\circ$ is in the range of the machine precision, but not identically zero. To circumvent this situation, we follow [77] in using a different operator which is more stable and that is based on a smooth perturbation of the modulus function $|Ux|_\circ$.

As discussed in [77], a suitable smooth perturbation of the modulus function for $u \in \mathbb{R}$ is given by

$$\kappa_\varepsilon(u) = \frac{u^2}{\sqrt{u^2 + \varepsilon^2}}. \tag{4.54}$$

This function converges to $|\cdot|$ uniformly in ε with bounded gradient, as we will see now.

Lemma 4.7.1 (Section 5.2 from [77]). *Let $\kappa_\varepsilon(u)$ as in (4.54) and $\varepsilon > 0$. Then the following properties hold:*

i) $\kappa_\varepsilon(0) = 0$

ii) $\operatorname{supp}(\kappa_\varepsilon(u)) = \operatorname{supp}(u)$

iii) $\left| |u| - \kappa_\varepsilon(u) \right| \leq \varepsilon$*, i.e.,* $\kappa_\varepsilon(u) \to |u|$ *uniformly in* ε*.*

iv) $\left\| \kappa_\varepsilon'(u) \right\|_\infty \leq \frac{4}{3}\sqrt{\frac{3}{2}}$

Proof. The first property is obvious, ii) follows directly from the definition in (4.54). For iii) consider

$$\begin{aligned}\left|\frac{u^2}{(u^2+\varepsilon^2)^{1/2}}-|u|\right| &= \left|\frac{u^2-|u|\,(u^2+\varepsilon^2)^{1/2}}{(u^2+\varepsilon^2)^{1/2}}\right| \\ &\leq \left|\frac{u^2-|u|\,(u^2+\varepsilon^2)^{1/2}}{|u|}\right| \\ &= \left||u|-\left(u^2+\varepsilon^2\right)^{1/2}\right| = \left(u^2+\varepsilon^2\right)^{1/2}-|u| \\ &\leq |u|+\varepsilon-|u| = \varepsilon.\end{aligned}$$

iv) Since κ_ε is convex, we will determine the bound on the gradient by calculating the extremal points of the derivative. The derivative is given by

$$\frac{\mathrm{d}\kappa_\varepsilon(u)}{\mathrm{d}u} = \frac{u^3+2\varepsilon^2 u}{(u^2+\varepsilon^2)^{3/2}}$$

for every fixed $\varepsilon > 0$, thus yielding the second derivative

$$\frac{\mathrm{d}^2\kappa_\varepsilon(u)}{\mathrm{d}u^2} = \frac{2\varepsilon^4-\varepsilon^2u^2}{(u^2+\varepsilon^2)^{5/2}} \overset{!}{=} 0. \tag{4.55}$$

Equation (4.55) has the roots

$$u_{1,2} = \pm\sqrt{2}\varepsilon,$$

therefore the gradient of κ_ε is bounded by

$$\kappa_\varepsilon\left(\pm\sqrt{2}\varepsilon\right) = \pm\frac{4}{3}\sqrt{\frac{3}{2}}.$$

□

5. Numerical Evaluation

In this chapter we numerically evaluate the performance of the proposed algorithm based on simulated measurements. In the next sections, we introduce some more details on the numerical implementation and explain how the performance of the algorithms is evaluated. Furthermore, we will discuss how the simulated data is obtained.

The computational complexity of the shearlet transform in the algorithm can be reduced drastically by pre-processing. Since the generated shearlet system does only depend on the image size, the number of scales and the filters that are used, we can pre-compute the shearlet system and re-use the same system for every iteration in the algorithm.

Finally, we will evaluate the performance of the proposed algorithm for different types of objects and compare them with different well-known, a priori constraints. Additionally, we will combine the newly proposed sparsity constraint with the so-called range constraint. We will see that this leads to drastically improved performance especially for data with noise.

5.1. Discrete Shearlet Transform and Soft-Thresholding

Although the framework developed in Chapter 4 is general enough to allow for a lot of different transforms, we will focus on the discrete shearlet transform discussed in Chapter 3. The proposed soft-thresholding operator

$$P_{S_{T,\gamma}} = T^* S_\gamma T$$

is implemented as follows. Given a discrete image $x \in \mathbb{R}^{d_1 \times d_2}$, compute the discrete shearlet transform $Tx \in \mathbb{R}^{d_1 \times d_2 \times s}$ where s denotes the redundancy of the transform.

For each component of $y := Tx$ we compute the soft-thresholding point-wise[1] by

$$S_\gamma y = \operatorname{sign}(y).*\max(\operatorname{abs}(y) - \gamma, 0).$$

Similarly, the smooth-hard shrinkage $V_{\gamma,\alpha} y$ can be computed by

$$\begin{aligned} v &= y.*(\operatorname{abs}(y) \geq \gamma) \\ V_{\gamma,\alpha}(y) &= \operatorname{sign}(v).*\operatorname{abs}(v).*\exp\Big((-\alpha./(\exp(\operatorname{abs}(v) - \gamma) - 1).^2\Big). \end{aligned}$$

Figure 5.1 illustrates the different thresholding operations and the influence of the smoothing parameter α for the smooth-hard shrinkage. For vanishing α, the smooth-hard-shrinkage approximates the hard-thresholding depicted in the left plot.

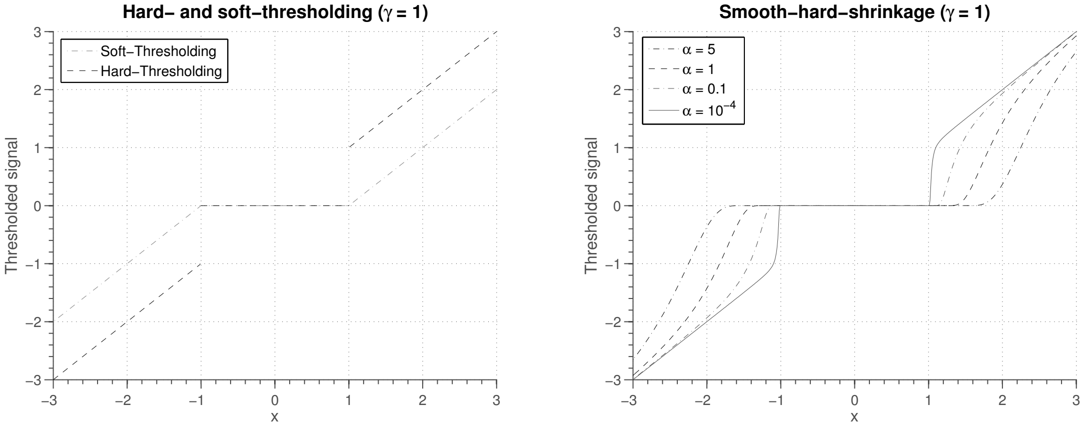


Figure 5.1.: Illustration of different thresholding operations and the influence of the smoothing parameter α occurring in smooth-hard shrinkage.

We obtain the final result $P_{S_{T,\gamma}}$ using the inverse discrete shearlet transform. The number of scales used for the discrete shearlet will depend on the size of the image and hence, the computational complexity will increase for larger images. However, as mentioned above, the computation of the shearlet system can be done in pre-processing. The filters that we will use for the numerical evaluation are those mentioned in Section 3.2.4. We will furthermore consider the operator

$$P^+_{S_{T,\gamma}} = P_{\mathrm{R}} P_{S_{T,\gamma}} \tag{5.1}$$

where P_{R} denotes the projection onto the range constraint set.

---

[1] Note that we used the Matlab notation for point-wise multiplication here. The functions sign, max, and abs act point-wise and hence return a vector for vectorial inputs.

5.2. Range Constraint

The set R describing the range constraint depends on the type of the object. For phase objects, the range constraint R_p describes objects with negative phase, amplitude objects have range constraints R_a where the amplitude of the object is smaller or equal one. For mixed objects, both constraints apply, i.e., the object has an amplitude smaller or equal one and a negative phase described by the set R_m. To conclude, these sets are defined as

$$\begin{aligned} R_\mathrm{p} &:= \left\{x \in \mathbb{C}^d \mid \varphi(x) \leq 0\right\}, \\ R_\mathrm{a} &:= \left\{x \in \mathbb{C}^d \mid |x|_\circ \leq 1\right\}, \\ R_\mathrm{m} &:= \left\{x \in \mathbb{C}^d \mid \varphi(x) \leq 0 \wedge |x|_\circ \leq 1\right\}, \end{aligned}$$

where $\varphi(x)$ denotes the point-wise phase of x and $|x|_\circ$ the point-wise amplitude of x. The projections onto these sets are given by

$$\begin{aligned} P_{R_\mathrm{p}} &= |x|_\circ \odot \exp\left(\mathrm{i} \cdot \min\left\{\varphi(x), 0\right\}\right), \\ P_{R_\mathrm{a}} &= \min\left\{|x|_\circ, 1\right\} \odot \exp\left(\mathrm{i}\varphi(x)\right), \\ P_{R_\mathrm{m}} &= \min\left\{|x|_\circ, 1\right\} \odot \exp\left(\mathrm{i} \cdot \min\left\{\varphi(x), 0\right\}\right). \end{aligned}$$

5.3. Simulating the Object Transmission Function

This section describes the simulation of the object transmission function. This part is crucial for the application of the algorithm to experimental data as the object transmission function is part of the mathematical model describing the experimental process. It will further give rise to a splitting approach that will be necessary in order to obtain meaningful results.

In the description of the simulation details we will follow the presentation in [92]. Recall from Chapter 2 that the exit wave field can be approximated by

$$U_\omega(x_1, x_2, 0) \approx \mathrm{e}^{\mathrm{i}k\tau} P(x_1, x_2) O(x_1, x_2).$$

with *illumination function*

$$P(x_1, x_2) := U_\omega(x_1, x_2, -\tau)$$

and *object transmission function*

$$O(x_1, x_2) := \exp\left(-k \int_{-\tau}^{0} \beta(x_1, x_2, z)\,\mathrm{d}z - \mathrm{i}k \int_{-\tau}^{0} \delta(x_1, x_2, z)\,\mathrm{d}z\right).$$

For monochromatic plane wave illumination one has $P(x_1, x_2) \equiv 1$ for all x_1, x_2 and hence we obtain

$$U_\omega(x_1, x_2, 0) \approx \mathrm{e}^{\mathrm{i}k\tau} O(x_1, x_2).$$

Hence, the important step is the simulation of the object transmission function $O(x_1, x_2)$. For homogeneous objects that are assumed here, $\beta(x_1, x_2, z)$ and $\delta(x_1, x_2, z)$ are constant in z-direction and hence

$$O(x_1, x_2) = \exp\left(-k\beta\Delta z(x_1, x_2) - \mathrm{i}k\delta\Delta z(x_1, x_2)\right)$$

where $\Delta z(x_1, x_2)$ describes the lateral thickness profile and β, δ are *constants*, cf. [92, Chapter 6]. Given a maximal thickness τ and defining the *relative lateral thickness profile* $T(x_1, x_2) := {}^{\Delta z(x_1, x_2)}/_{\tau}$ we can write

$$O(x_1, x_2) = \exp\left[-k\tau\beta T(x_1, x_2) - \mathrm{i}k\tau\delta T(x_1, x_2)\right].$$

In simulations, $T(x_1, x_2)$ will be a discretized gray scale image on which we will enforce the constraints in object domain such as support, positivity or sparsity. Depending on β and δ, we will either call the objects of interest *amplitude objects* (if δ is negligible), *phase objects* (if β is negligible) or *mixed objects* if neither of them can be neglected.

5.4. Details on the Numerical Implementation

The relaxation parameter β_k for the RAAR algorithm is chosen as

$$\beta_k = \exp((-k/\beta_{\text{switch}})^3)\beta_0 + (1 - \exp((-k/\beta_{\text{switch}})^3) * \beta_{\text{max}} \tag{5.2}$$

with $\beta_0 = 0.99$, $\beta_{\text{switch}} = 20$, and $\beta_{\text{max}} = 0.55$ according to [75]. In some cases with noise, we will choose $\beta_k \equiv \beta_{\text{max}}$ which leads to better reconstruction results.

Splitting Approach

Since the shearlet transform and the shrinkage operators are only defined for real-valued objects, we will use a suitable splitting. Since, for physical reasons, it is reasonable to assume that the measured object is sparse in amplitude and or phase, we will use a corresponding decomposition of the complex-valued wave function and apply the proximity operator to the amplitude and phase component individually. Therefore, we will use the operator

$$R_{(\gamma_1,\gamma_2)\|\cdot\|_1} = 2\left(T^{-1}\mathrm{prox}_{\gamma_1\|\cdot\|_1} T\,|\cdot|_\circ \cdot \exp\left[\mathrm{i}\,T^{-1}\mathrm{prox}_{\gamma_2\|\cdot\|_1} T\varphi(\cdot)\right]\right) - \mathrm{Id}. \tag{5.3}$$

instead of $R_{S_{T,\gamma}}$ and $P_{(\gamma_1,\gamma_2)\|\cdot\|_1}$ respectively. As we will see in the numerical examples later, this approach is justified since the assumption of cartoon-like images is fulfilled in phase and amplitude individually.

5.4.1. Error Measures

We have seen in the last chapter that for some algorithms it may not be the immediate iterates $x^{(k)}$ but the shadow sequences that are of interest. It is therefore crucial when comparing, e.g., the error decay, to monitor the right variable. While in the case of the method of alternating projections these are exactly the iterates itself, for the Douglas-Rachford algorithm and the RAAR-variants, we will monitor the shadow sequence $\mathrm{prox}_{\gamma g}(x^{(k)})$. Furthermore, the question arises which measure we will use in order to judge the reconstructions. In image processing, the *peak-signal-to-noise-ratio* (PSNR)

$$\mathrm{PSNR}\left(x^{(k)}\right) := 10\log_{10}\left(\frac{\max_x^2}{\left\|x^{(k)} - x\right\|^2}\right)$$

is a widely used error measure where $\max_x$ denotes the maximal possible entry of x. Using normalization, this can be set to 1 if the solution x only exhibits values in $[0, 1]$. It is also quite common to simply use the Euclidean distance of the sequence to the original solution as an error measure. Both of these methods require that the solution to the phase retrieval problem is known. This will be the case for simulated data but not so for real data. In [8], the error measure

$$E\left(x^{(k)}\right) := \frac{\left\|P_A(P_B(x^{(k)}) - P_B(x^{(k)})\right\|^2}{\left\|P_B(x^{(k)})\right\|^2} \tag{5.4}$$

is suggested which, in the case for two sets A and B, measures the squared distance from $P_B(x^{(k)})$ to A. This measure is motivated by the practical intuition that one is more interested in the object-constraint rather than the data which may be error-prone. This measure can be used for experimental data when the true solution x is not known.

In some figures we will only plot the decay of the error measure up to a certain number of iterations which may differ from the number of iterations of the algorithm. The purpose is to highlight the most important part of the behavior of the error. We only do this for cases where the error measure is almost constant after this certain number of iterations.

5.5. Numerical Results for the Reconstruction of Simulated Data

In order to be able to reliably monitor the error decay of the algorithm according to the measure suggested above, we will be using simulated data. Therefore we will consider the cases with exact data, i.e., without noise, as well as data that is corrupted by simulated Poisson noise of different intensities. We will use the algorithm for different types of objects (amplitude, pure phase and mixed objects) in order to estimate the applicability for different types of experimental data. In the following, we denote by $x \in \mathbb{R}^{256\times 256}$ the image of a cell depicted in Figure 5.6, originally published in [40].

5.5.1. Exact Data

As discussed in the previous section, the monitored sequence is in this case $P_{S_{T,\gamma}}x^{(k)}$ and not $x^{(k)}$ itself. The root-mean-squared error is given by

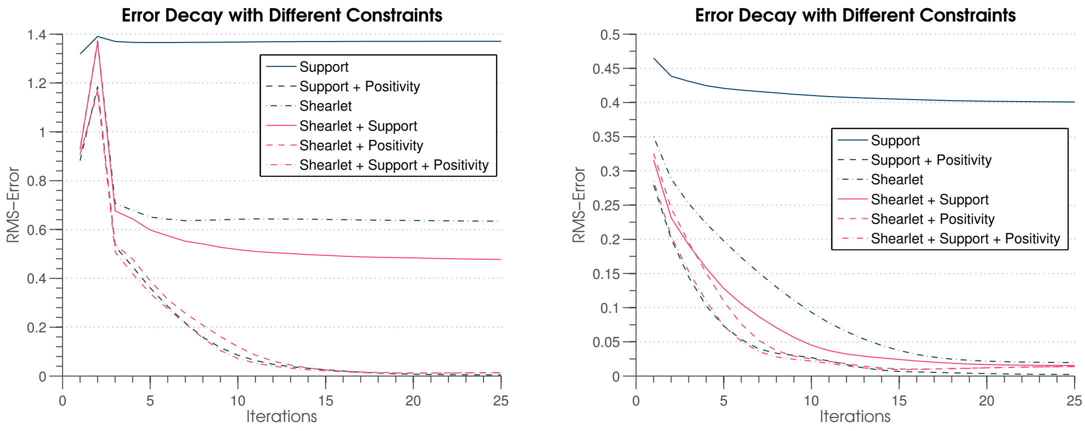
$$E_{\mathrm{RMS}}(x^{(k)}) = \frac{\left\|x - P_{S_{T,\gamma}}x^{(k)}\right\|_F}{\sqrt{d}}$$

where $d = d_1 \cdot d_2$ is the number of pixels and x denotes the true solution. Here, since $x \in \mathbb{R}^{d_1 \times d_2}$, we use the Frobenius norm which is defined by

$$\|x\|_F = \sqrt{\sum_{j=1}^{d_1} \sum_{k=1}^{d_2} \left|x[j,k]\right|^2}.$$

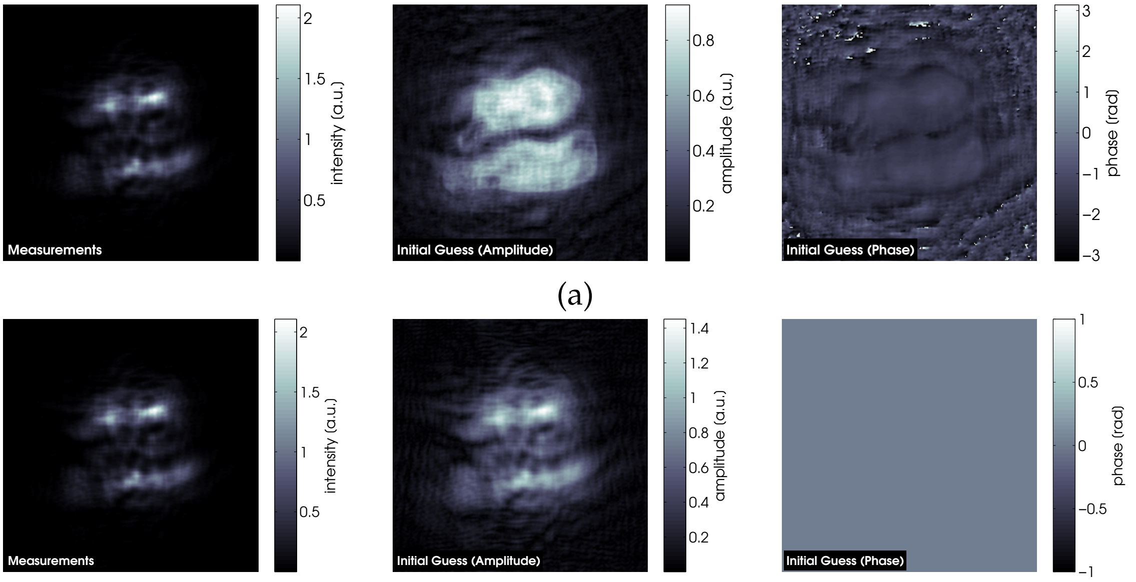
Real-Valued Objects

First, we consider the case of exact data for a real-valued object. This means, the measurements m are described by $m := |D_\tau x|_\circ$ where D_τ denotes the discretized Fresnel transform. The experimental parameters are $\lambda = 0.1\text{nm}$, $z = 100\text{mm}$. The considered object is of size 256 × 256 pixel. The soft-threshold parameter was set to $\gamma_k = {}^{\gamma_0}/_k$ with $\gamma_0 = 0.5$. This resembles the numerical experiments published in [72]. For real-valued objects, depending on the Fresnel number, the measurements m may be closer to the solution than the simple back-projection $\mathcal{D}_\tau^{-1} m$ without any phase information. We therefore also examine the behavior for the starting guess $x^{(0)} = m$. The plotted root-mean-squared errors in Figure 5.2(a) coincide with the visual impressions from Figure 5.4. While the simple support constraint does not yield a meaningful reconstruction, all other constraints do. However, the constructions that incorporate positivity outperform the shearlet constraint and even the combined shearlet and support constraint by far. In this setting, the positivity constraint is the dominating constraint and yields almost perfect reconstructions. However,



(a) (b)

Figure 5.2.: Decay of the root-mean-squared error of a real-valued object for different constraints on exact data with starting guess (a) $x^{(0)} = \mathcal{D}_\tau^{-1} m$ and (b) $x^{(0)} = m$



(a)

(b)

Figure 5.3.: Two different initial guesses for a real-valued object with exact data where in (a) $x^{(0)} = \mathcal{D}_\tau^{-1} m$ and in (b) $x^{(0)} = m$

using the different starting guess $x^{(0)} = m$ shows different results for some cases. In Figure 5.2(b) we observe a better decay in the error for both shearlet and combined shearlet and support constraint. The faster error decay coincides with visually better reconstructions for those cases depicted in Figure 5.5. This raises the conjecture that the performance of the algorithm is critically influenced by the starting guess. This conjecture is justified by the observation that the optimization problem which the algorithm solves is non-convex. However, this problem is independent of the choice of constraints and will therefore not be studied in more detail.

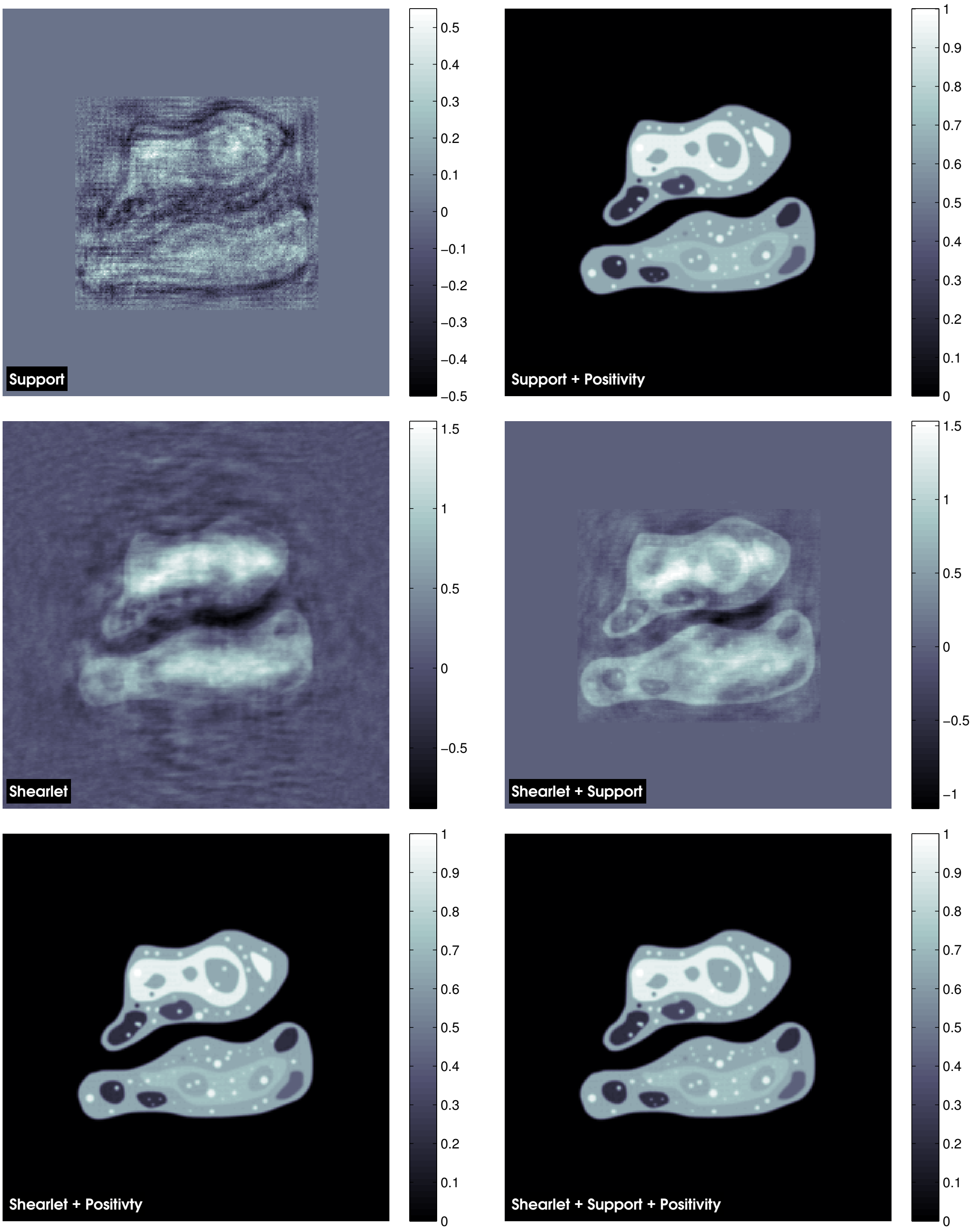


Figure 5.4.: Reconstructions of a real-valued object from exact measurements for different constraints with starting guess $x^{(0)} = \mathcal{D}_\tau^{-1} m$.

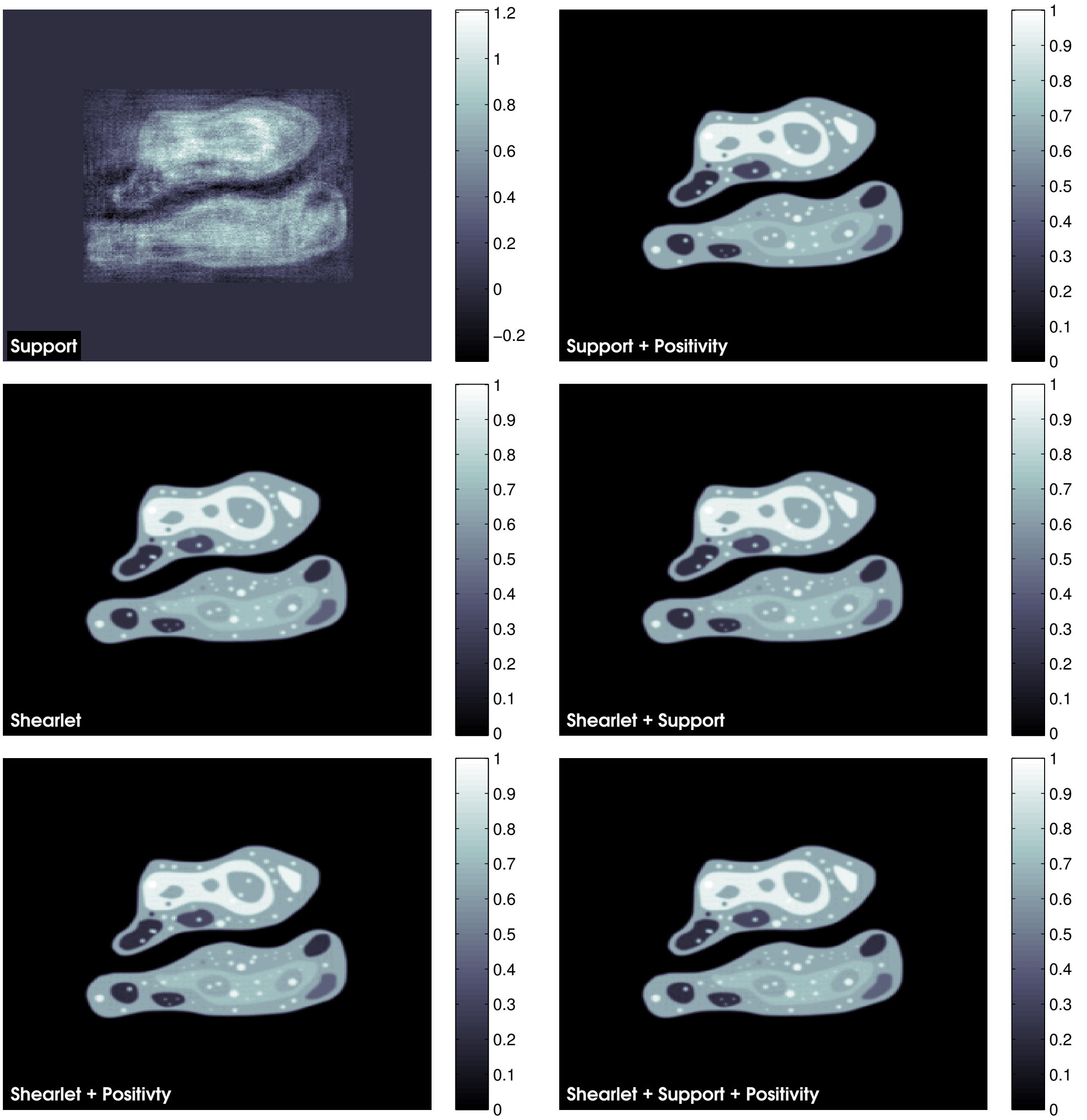


Figure 5.5.: Reconstructions of a real-valued object from exact measurements for different constraints with starting guess $x^{(0)} = m$.

Amplitude Objects

In order to provide simulations for a physically more sound scenario, we consider amplitude objects in the following section. Using the formulation of the object transmission function above, this corresponds to objects where δ is negligible. Numerically, we set $\delta = 0$ for these objects. The full set of parameters used for the numerical simulation and reconstruction is given in Table 5.1. Although $\delta = 0$ and hence the imaginary part of the complex exponential vanishes, due to noise or reconstruction artifacts, the phase may not be equal to one everywhere. Therefore, we will also depict the phase of the reconstruction in order to indicate the artifacts. However, if the type of object is known, this could potentially be incorporated by additional constraints. The measurements and initial guesses for the amplitude and phase of the object are depicted in Figure 5.7.

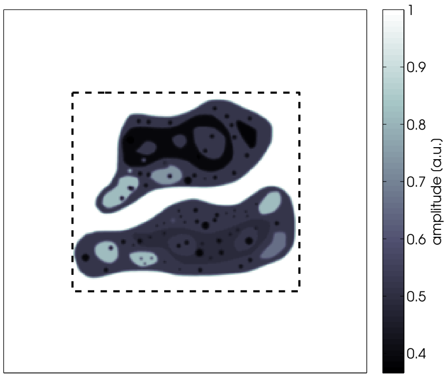


Figure 5.6.: Amplitude object with estimated support (dashed line) used for numerical simulation of the object transmission function

As can be seen in the initial guess for the amplitude in Figure 5.7, the measurements already resemble the initial object. Hence, the error decay for all four constraints, which is depicted in Figure 5.8, resembles the excellent reconstructions seen in Figure 5.9.[2] The differences that can be seen are the artifacts in the reconstructed phases. However, note that these oscillations that can be seen are at most of the order 10^{-4} and hence in this setting negligible.

---

[2] Note that in order to make the plot more readable, the error decay is only shown for the first 100 iterations. Later on, the errors for all constraints become almost stationary.

Parameter		Value
Decrement of real part of refractive index	δ	0
Imaginary part of refractive index	β	$8 \cdot 10^{-6}$
Maximal lateral thickness	τ	$20 \cdot 10^{-6}\mathrm{m}$
Wavelength	λ	10^{-10}
Phase shift distribution	P	0
Amplitude distribution	A	τx
Wave-function in object plane	U	$\exp(-k\beta A)$
Number of iterations	N	500
RAAR relaxation parameter	β_k	cf. (5.2)
Threshold parameter	γ_k	$\gamma_k = {}^{\gamma_0}/_k$ with $\gamma_0 = 0.007$

Table 5.1.: Parameters for the numerical simulation of the object transmission function for amplitude objects with exact data

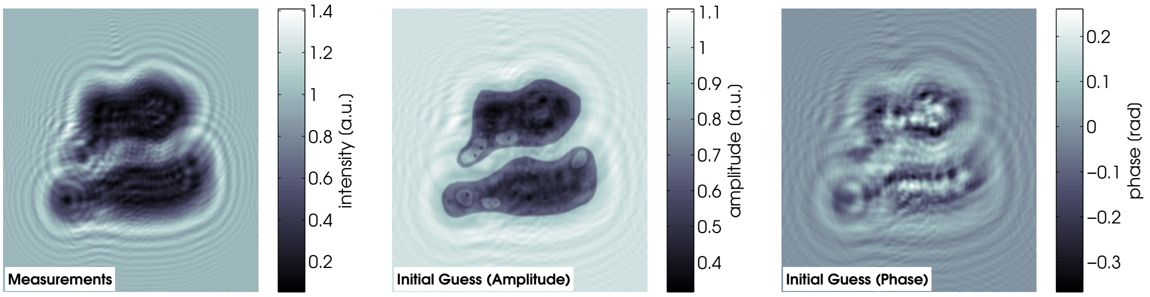
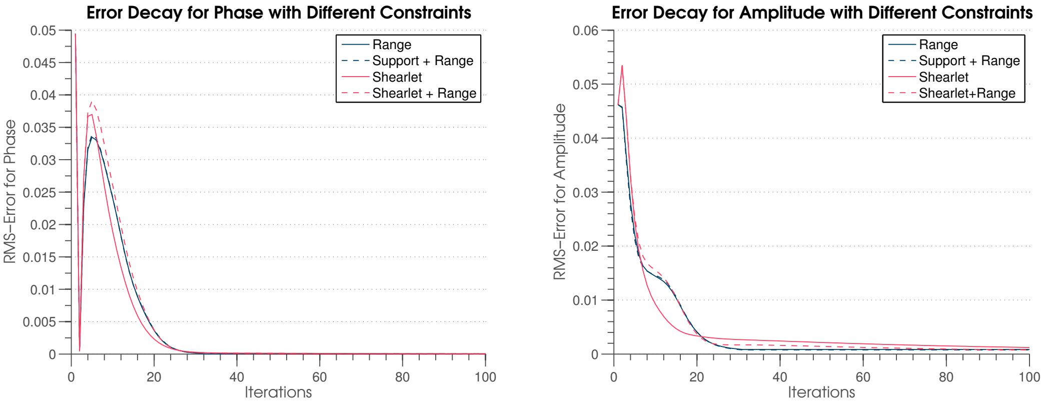


Figure 5.7.: Exact measurements of an amplitude object and initial guesses for phase and amplitude



(a) (b)

Figure 5.8.: Decay of the root-mean-squared error of an amplitude object for different constraints on exact data for (a) phase and (b) amplitude

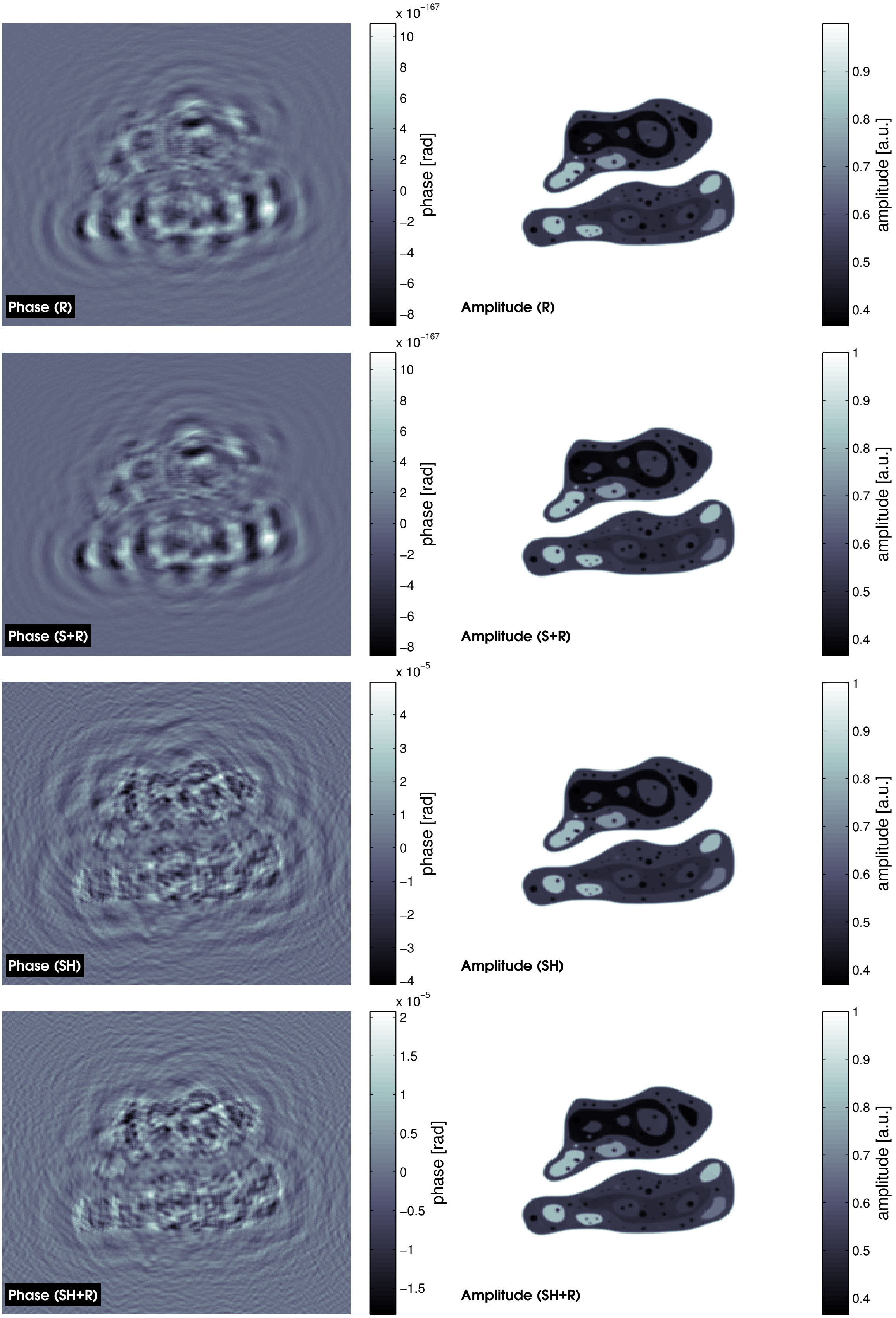


Figure 5.9.: Reconstructions of an amplitude object from exact measurements for different constraints

Phase Objects

Next, we want to study the performance of the different constraints for phase objects. The parameters for this setup are depicted in Table 5.2. The dashed line around the object marks the estimated support which is used for the support and range constraint. Since the real part of the refractive index for this type of object is zero everywhere, the amplitude of the object is one in each pixel element. For the reconstruction, we depict both the phase and amplitude of the reconstruction. Especially for Poisson data where the object cannot be recovered exactly, it is to be expected that the amplitude will not be exactly one everywhere. The measurements and initial guesses for phase and amplitude are shown in 5.11. Qualitatively, the initial guess for the phase is already close to the solution. Hence, the error decay in Figure 5.12 behaves expectedly well for all constraints. The bump for the error decay of the amplitude is due to the fact that the initial guess is not optimal in the sense that one knows from the type of object that the amplitude has to be constantly one. However, despite the sub-optimal choice of starting guesses all methods deliver decent reconstructions. The shearlet constraints alone however lacks the proper scaling of the entries since no range constraint is applied here. Nonetheless, qualitatively, the object is reconstructed as well as for the combined shearlet plus range constraint, see Figure 5.13.

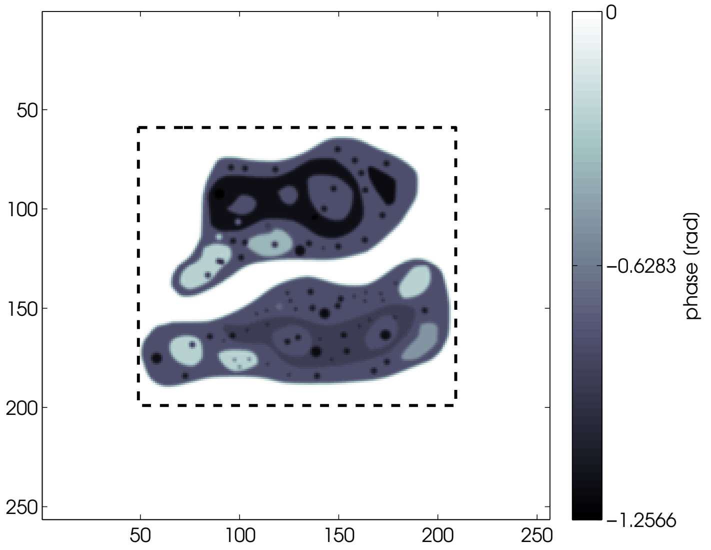


Figure 5.10.: Phase object with estimated support (dashed line) used for numerical simulation of the object transmission function

Parameter		Value
Decrement of real part of refractive index	δ	$1.6 \cdot 10^{-6}$
Imaginary part of refractive index	β	0
Maximal lateral thickness	τ	$20 \cdot 10^{-6}$m
Wavelength	λ	10^{-10}
Phase shift distribution	P	τx
Amplitude distribution	A	0
Wave-function in object plane	U	$\exp(-ik\delta P)$
Number of iterations	N	500
RAAR relaxation parameter	β_k	cf. (5.2)
Threshold parameter	γ_k	$\gamma_k = \gamma_0/k$ with $\gamma_0 = 0.07$

Table 5.2.: Parameters for numerical simulation of phase objects with exact data

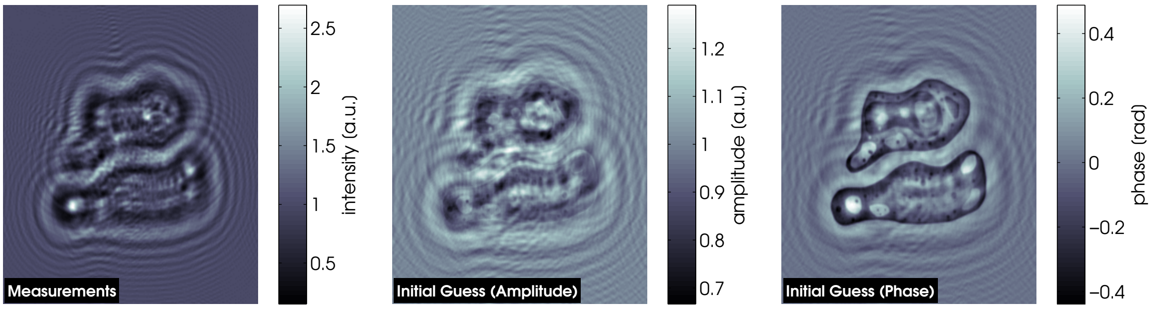
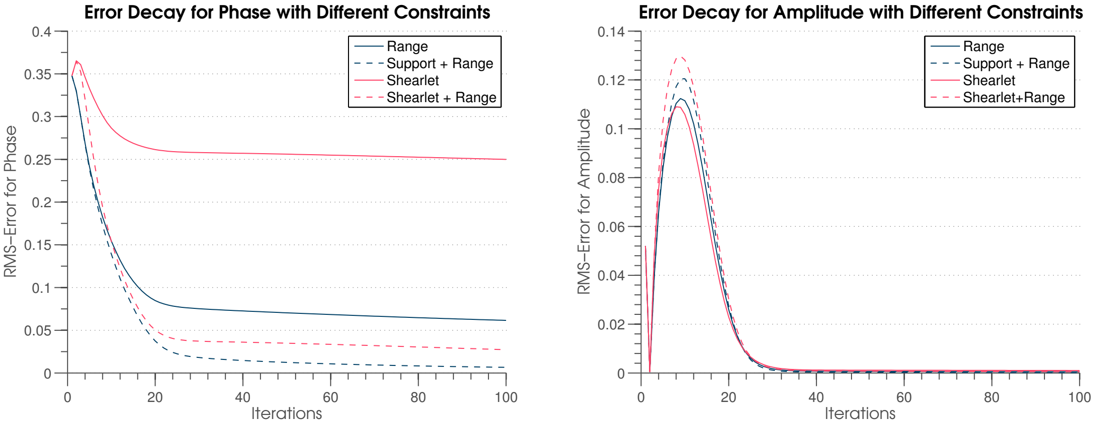


Figure 5.11.: Exact measurements of the phase object and initial guesses for phase and amplitude



(a) (b)

Figure 5.12.: Decay of the root-mean-squared error of a phase object for different constraints on exact data for (a) phase and (b) amplitude

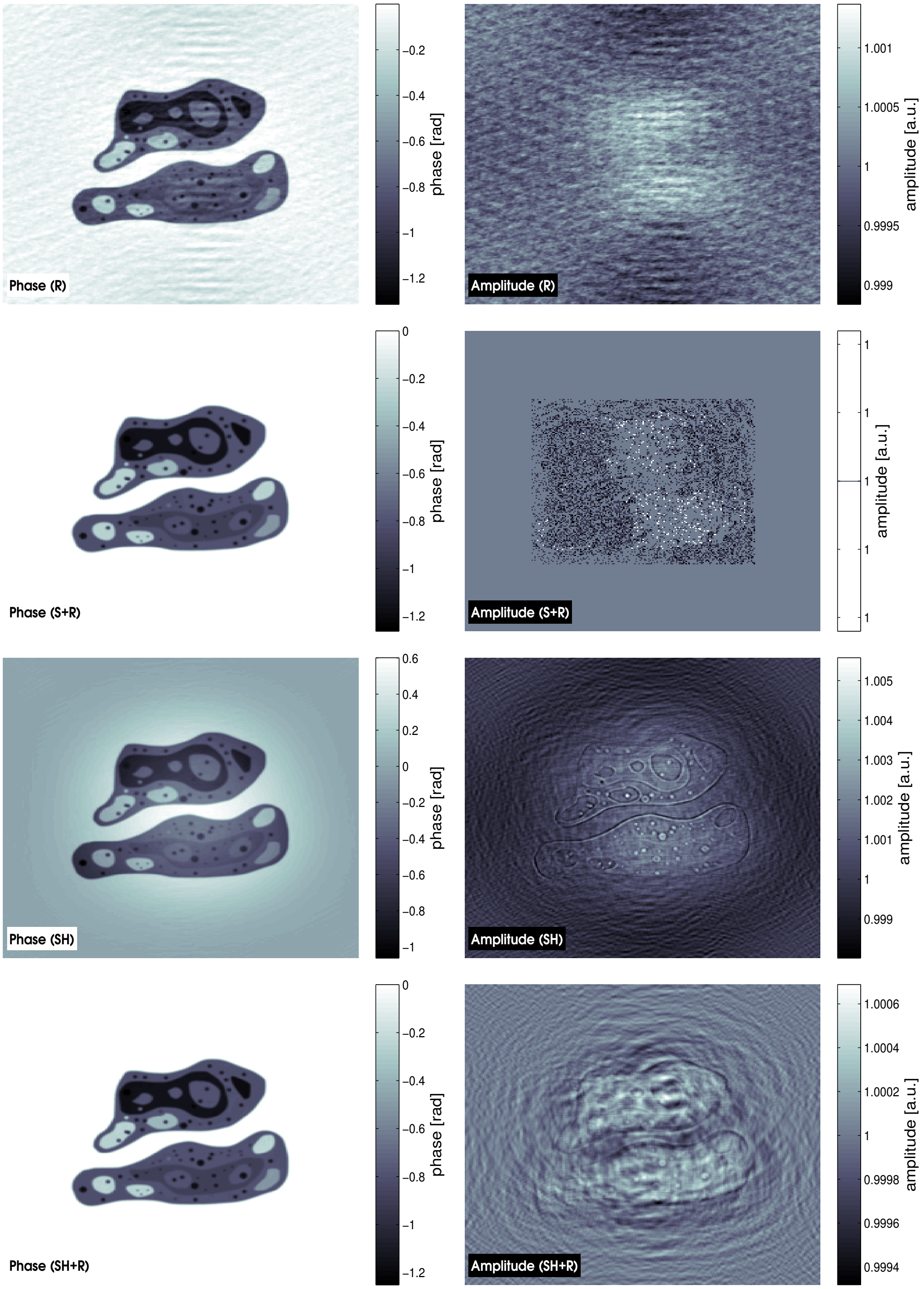


Figure 5.13.: Reconstructions of a phase object from exact measurements for different constraints

Mixed Objects

For mixed objects, we will use the splitting approach described in (5.3). The assumption here is that both the amplitude and the phase can be sparsely represented using shearlets. The operator $P_{S_{T,\gamma}}$ will therefore be applied for the phase and amplitude separately. The mixed object for phase and amplitude is depicted in Figure 5.14, the parameters for the simulation as well as for the algorithm are given in Table 5.3.

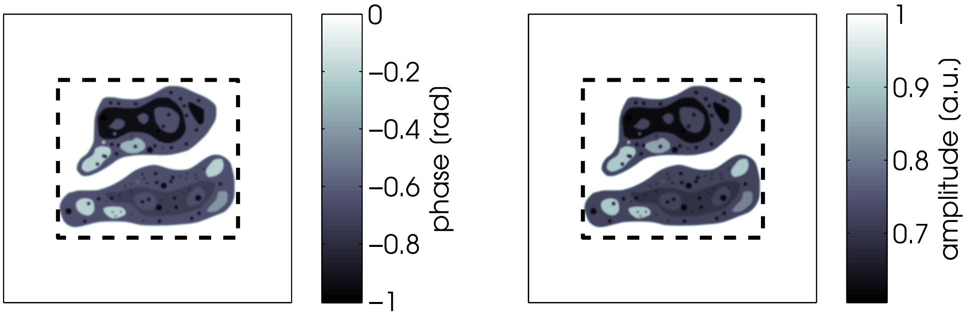


Figure 5.14.: Mixed object with estimated support (dashed line) used for numerical simulation of the object transmission function

Parameter		Value
Decrement of real part of refractive index	δ	$1.6 \cdot 10^{-6}$
Imaginary part of refractive index	β	$8 \cdot 10^{-7}$
Maximal lateral thickness	τ	10^{-6}
Wavelength	λ	10^{-10}
Phase shift distribution	P	τx
Amplitude distribution	A	τx
Wave-function in object plane	U	$\exp(-ik\delta P - k\beta A)$
Number of iterations	N	500
RAAR relaxation parameter	β_k	see (5.2)
Threshold parameter	γ_k	$\gamma_k = \gamma_0/k$ with $\gamma_0 = 0.01$

Table 5.3.: Parameters for the numerical simulation of the object transmission function for mixed objects with exact data

We used the same threshold parameter γ_k both for the phase and for the amplitude part of the splitting. As it turns out, mixed objects seem to be the most complicated objects to recover. For the case with exact data, all methods perform worse than for pure phase or pure amplitude objects. Nonetheless, the error decay behaves qualitatively similar to the other cases, see Figure 5.16. While the shearlet plus range constraint performs similar for the phase of the mixed object as the range and support plus range constraint, both shearlet constraints outperform the other constraints for the amplitude. This behavior can be visually tracked in Figure 5.17 where the reconstructions of the amplitude using shearlet constraints (with or without range constraint) have less oscillatory artifacts than the two other methods that do not use shearlet soft-thresholding. As before, the reconstruction using the shearlet constraint alone provides qualitatively good results but suffers from incomplete range information as can be seen in the color bar and also from the dark background in the reconstructions. Thus, combining shearlet and range constraints leads to improved reconstructions throughout all test cases. We will now investigate the behavior of

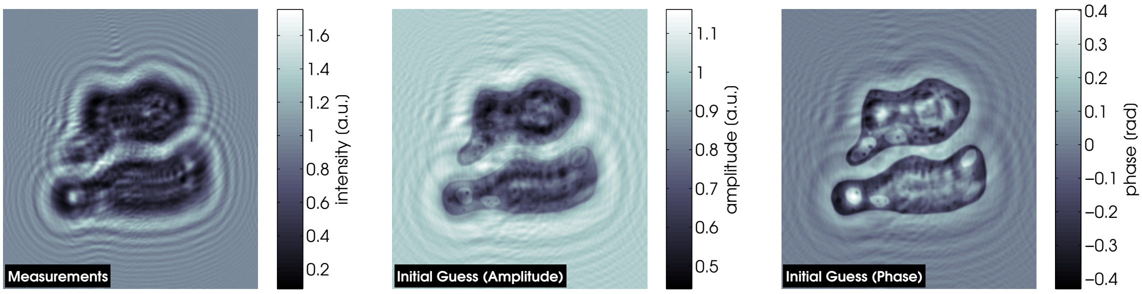
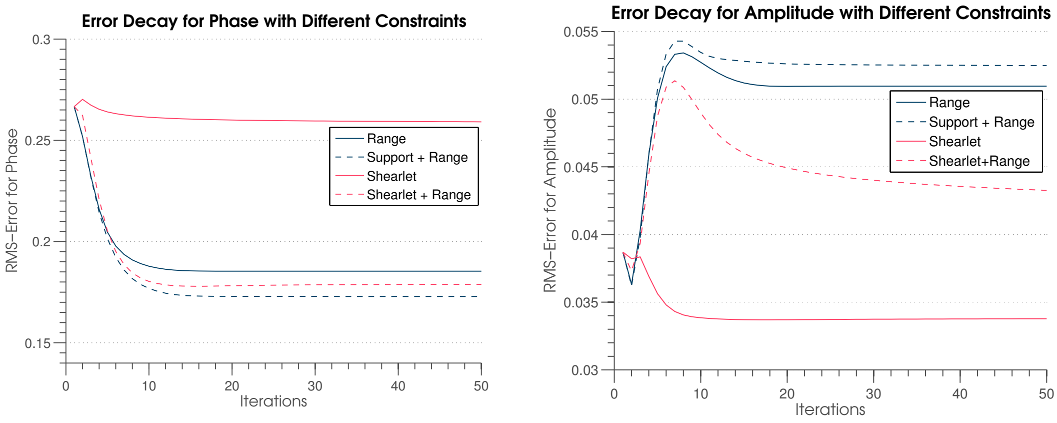


Figure 5.15.: Exact measurements of a mixed object and initial guesses for phase and amplitude



(a) (b)

Figure 5.16.: Decay of the root-mean-squared error of a mixed object for different constraints on exact data for (a) phase and (b) amplitude

the different methods for the same objects but with data that is corrupted by Poisson noise.

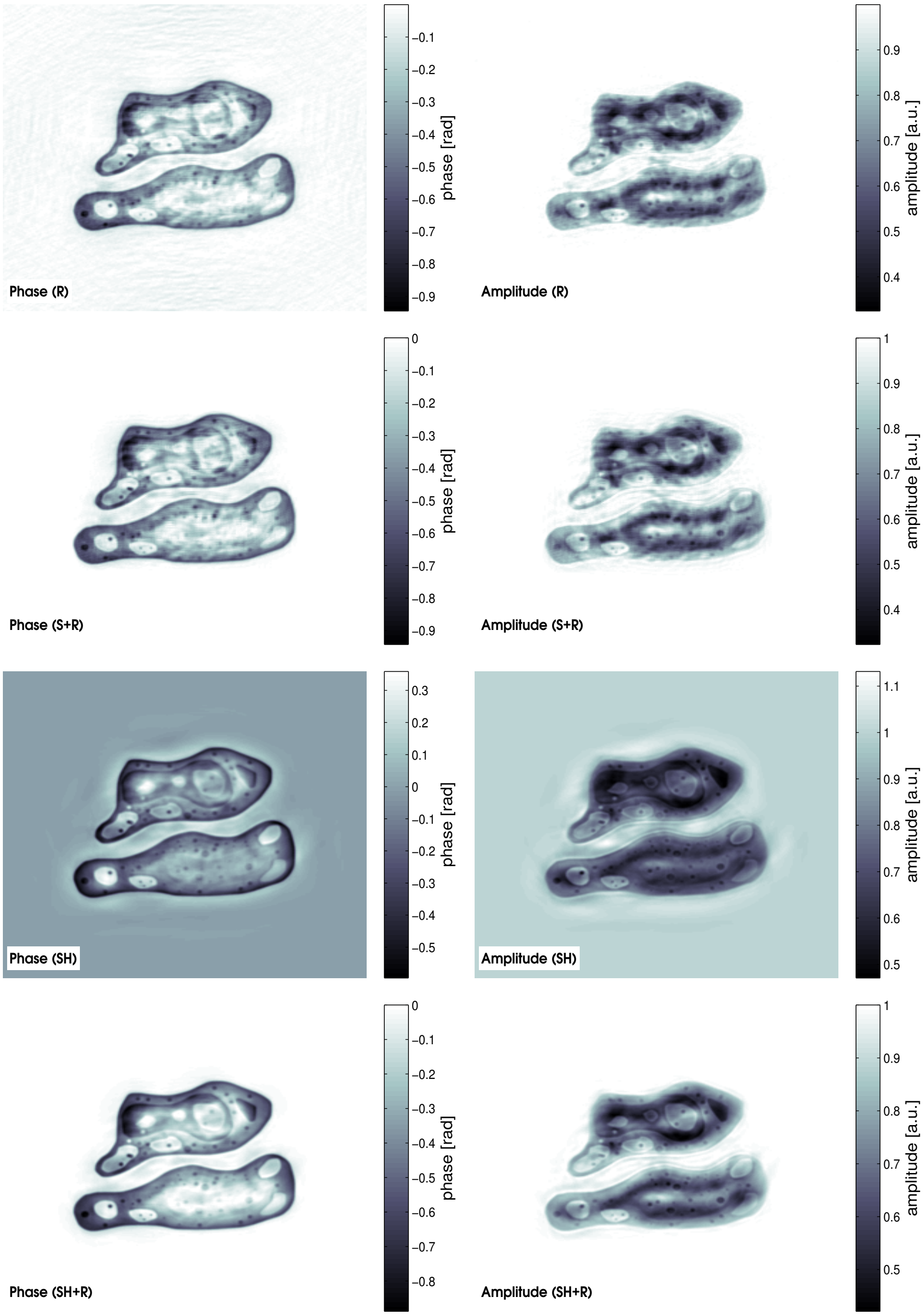


Figure 5.17.: Reconstructions of a mixed object from exact measurements for different constraints

5.5.2. Poisson Data

For the evaluation of the method we use a constant noise level which can be interpreted as an expected number of 50 photons per pixel. Although we may have different noise for each type of object, every constraint will be evaluated (for each object) with the exact same data. The measurements and initial guesses will be depicted similarly as for the case of exact data. It turned out that contrary to the case with exact data, the best results can be achieved to use a constant relaxation parameter $\beta_k \equiv \beta$ as well as a constant thresholding parameter $\gamma_k \equiv \gamma$. The choice of the parameters will be shown in the corresponding tables. The relaxation parameter is $\beta = 0.55$ in all cases. This aligns with the observations from [92] where a similar behavior was found.

Amplitude Object

For the amplitude object, we use the same physical parameters as before but perturb the data with Poisson noise to simulate the physical measurement process more realistically.

Parameter		Value
Decrement of real part of refractive index	δ	0
Imaginary part of refractive index	β	$8 \cdot 10^{-6}$
Maximal lateral thickness	τ	$20 \cdot 10^{-6}$m
Wavelength	λ	10^{-10}
Phase shift distribution	P	0
Amplitude distribution	A	τx
Wave-function in object plane	U	$\exp(-k\beta A)$
Number of iterations	N	500
RAAR relaxation parameter	β_k	$\beta_k \equiv 0.55$
Threshold parameter	γ_k	$\gamma_k \equiv 0.005$

Table 5.4.: Parameters for the numerical simulation of the object transmission function for amplitude objects with Poisson data

The corrupted measurements and the corresponding initial guesses are shown in Figure 5.18. The error decay for both phase and amplitude is shown in 5.19. Similar to exact data, the error in the phase decays for all methods very fast during the first few iterations. However, simply using a range constraint on the amplitude leads to a diverging error, cf. 5.19(b). The support plus range constraint stabilizes on a level

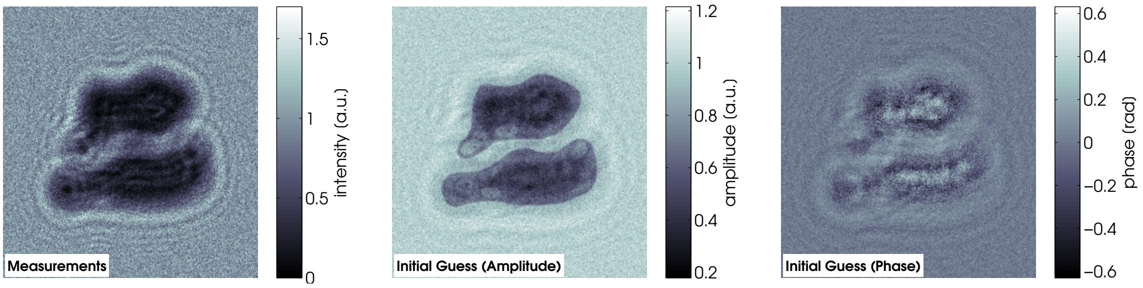
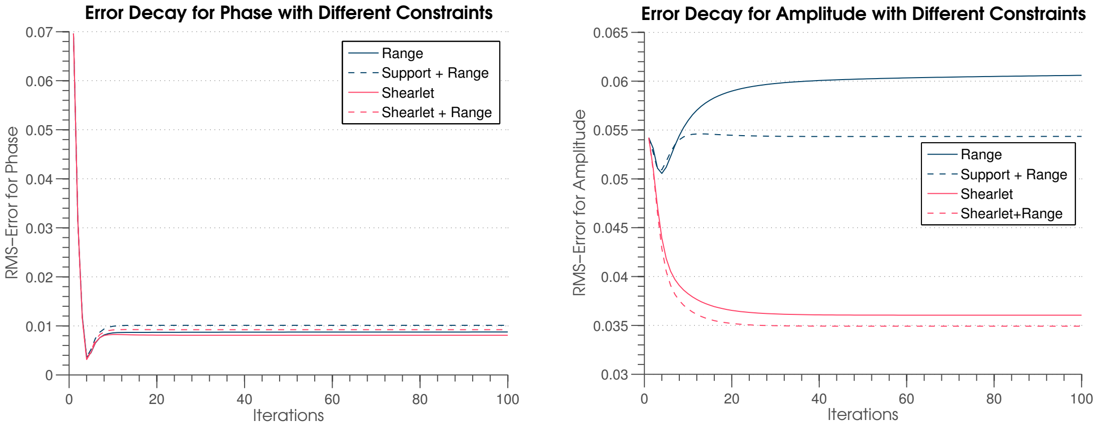


Figure 5.18.: Poisson measurements of the amplitude object and initial guesses for phase and amplitude

which is very close to the error simply using the initial guess. Both shearlet constraints outperform the other two constraints as can be seen in Figure 5.20. The reconstructions using the shearlet constraints lead to an improved reconstruction quality and removal of most of the noise.



(a) (b)

Figure 5.19.: Decay of the root-mean-squared error of an amplitude object for different constraints on Poisson data for (a) phase and (b) amplitude

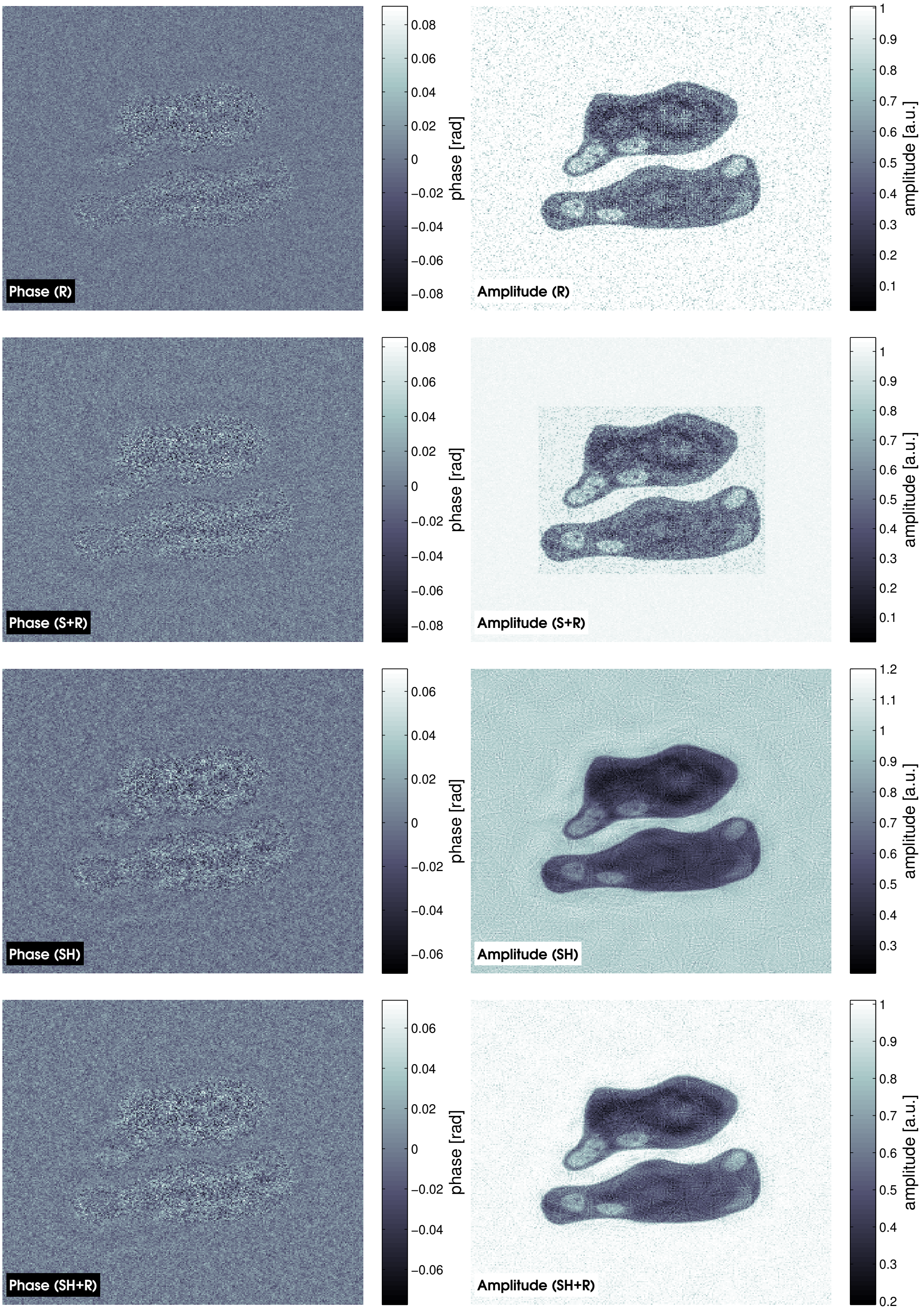


Figure 5.20.: Reconstructions of an amplitude object from Poisson measurements for different constraints

Phase Object

In this section we compare the four different constraints on the measurement of a simulated pure phase object that is corrupted by Poisson noise. The parameters of the setup are given in Table 5.5. The measurements as well as the initial guesses for phase and amplitude are given in Figure 5.21.

Parameter		Value
Decrement of real part of refractive index	δ	$1.6 \cdot 10^{-6}$
Imaginary part of refractive index	β	0
Maximal lateral thickness	τ	$20 \cdot 10^{-6}\mathrm{m}$
Wavelength	λ	10^{-10}
Phase shift distribution	P	τx
Amplitude distribution	A	0
Wave-function in object plane	U	$\exp(-ik\delta P)$
Number of iterations	N	500
RAAR relaxation parameter	β_k	$\beta_k \equiv 0.55$
Threshold parameter	γ_k	$\gamma_k \equiv 0.007$

Table 5.5.: Parameters for the numerical simulation of the object transmission function for phase objects with Poisson noise

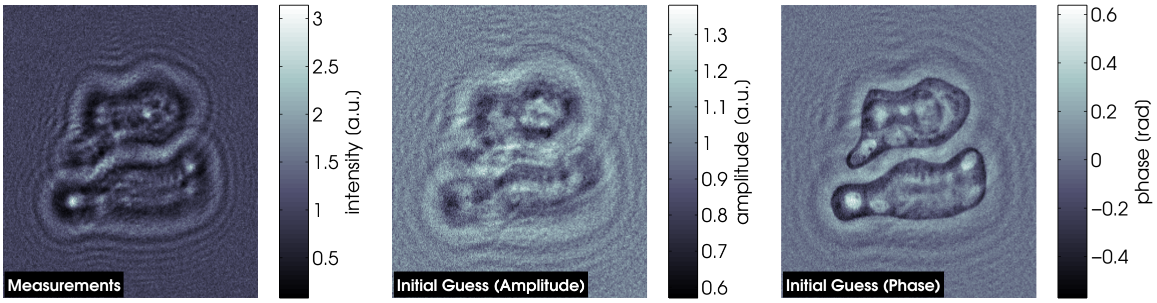
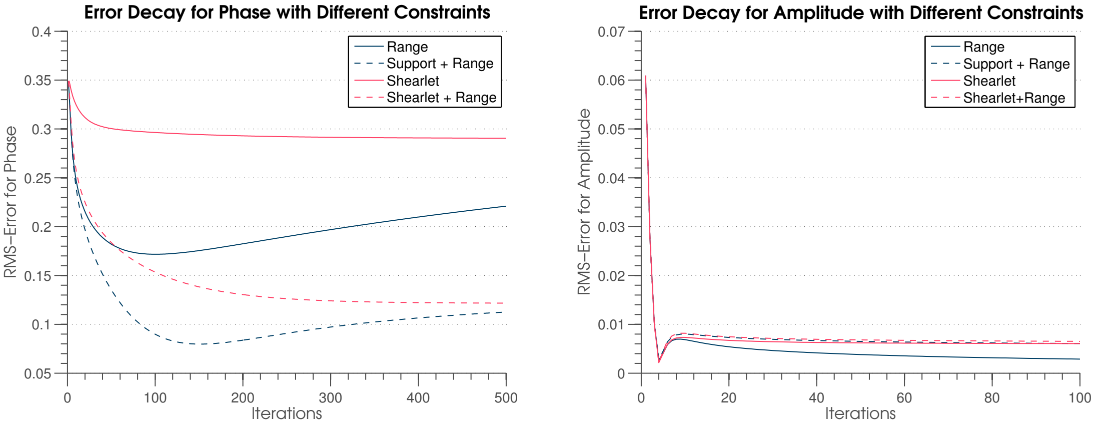


Figure 5.21.: Poisson measurements of the phase object and initial guesses for phase and amplitude

In Figure 5.22 the decay of the root-mean-squared error is depicted. While the error for the amplitude decays for all constraints almost instantly, cf. Figure 5.22(b), the behavior of the error for the phase is different. While the error for the range constraint first decays, it diverges after about 100 iterations, reflected by the poor reconstruction results, see Figure 5.23. The three other constraints decay comparibly fast where the shearlet and range constraint becomes stationary on a similar level as the support and



(a) (b)

Figure 5.22.: Decay of the root-mean-squared error of a phase object for different constraints on Poisson measurements for (a) phase and (b) amplitude

range constraint does. Although the reconstruction results for the shearlet constraint are qualitatively similar as the shearlet and range constraint, due to the lack of information on the range of the values, the quantitive result is worse than the combined approach, see the color bar in Figure 5.23. The shearlet plus range constraint leads to a visually improved reconstruction result compared to the support and range constraint. The slightly better error decay of the latter can be explained by the fact that due to the support constraint, the reconstruction is identical to the original object on all parts outside of the support box where the shearlet and range constraint does have some small artifacts which sum up overall to a larger error.

Note that the reconstructed amplitude for all four different constraints predominantly consists of noise with some artifacts of the original object (shearlet and shearlet plus range constraint) or the box constraint in the case of the support and range constraint. For the reconstructed phases, the noise is clearly visible for the range and the support plus range constraint. For both shearlet constraint types, the smoothing effect of the shearlet soft-thresholding leads to reconstructions with less noise.

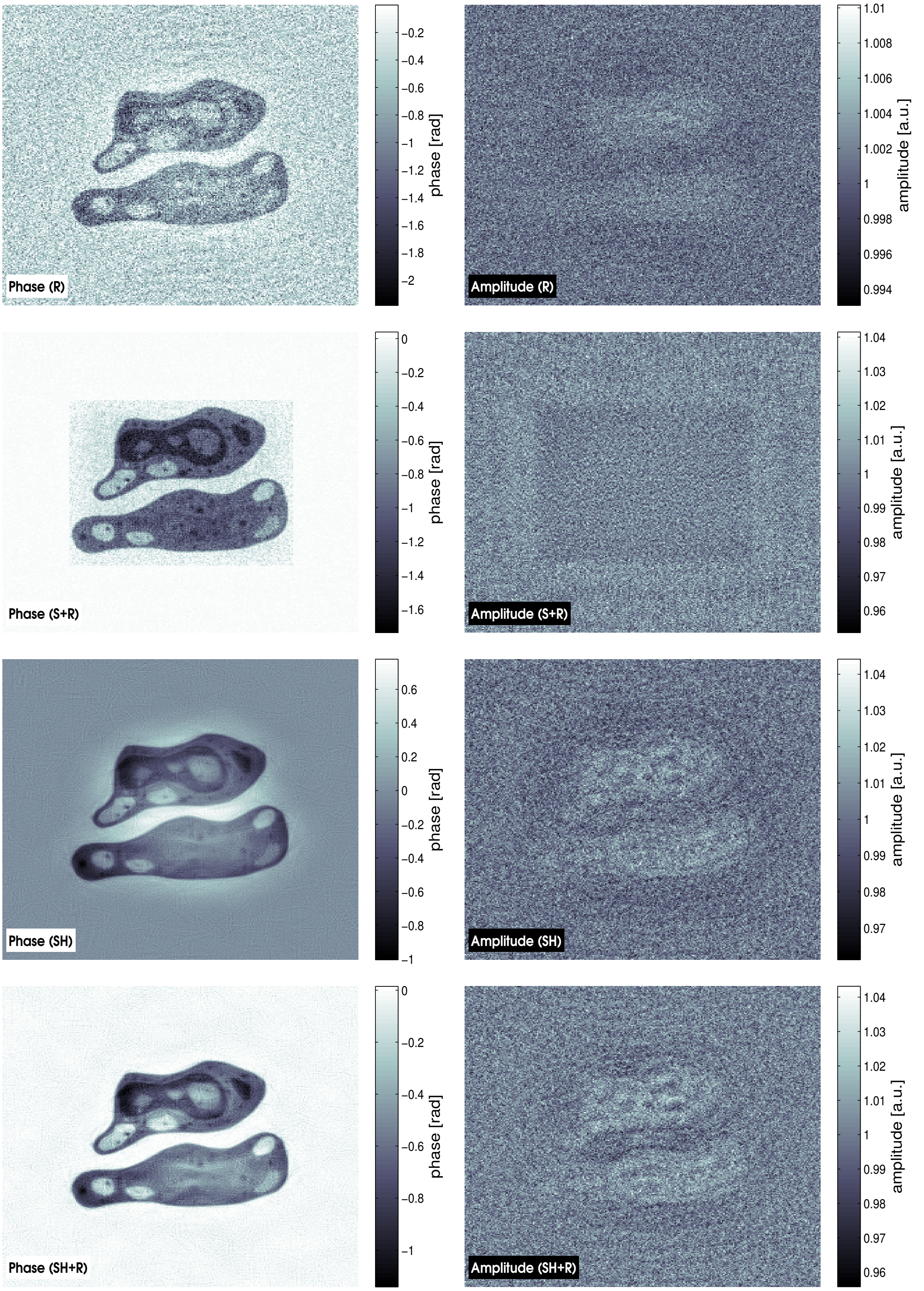


Figure 5.23.: Reconstructions of a phase object from Poisson measurements for different constraints

Mixed Objects

We now compare the different constraints on a mixed object using Poisson corrupted measurements. The parameters used for the reconstruction are given in Table 5.6 and the measurements with the initial guesses are shown in 5.24.

Parameter		Value
Decrement of real part of refractive index	δ	$1.6 \cdot 10^{-6}$
Imaginary part of refractive index	β	$8 \cdot 10^{-7}$
Maximal lateral thickness	τ	10^{-6}
Wavelength	λ	10^{-10}
Phase shift distribution	P	τx
Amplitude distribution	A	τx
Wave-function in object plane	U	$\exp\left(-\mathrm{i}k\delta P - k\beta A\right)$
Number of iterations	N	500
RAAR relaxation parameter	β_k	$\beta_k \equiv 0.55$
Threshold parameter	γ_k	$\gamma_k \equiv 0.005$

Table 5.6.: Parameters for the numerical simulation of the object transmission function for mixed objects with Poisson noise

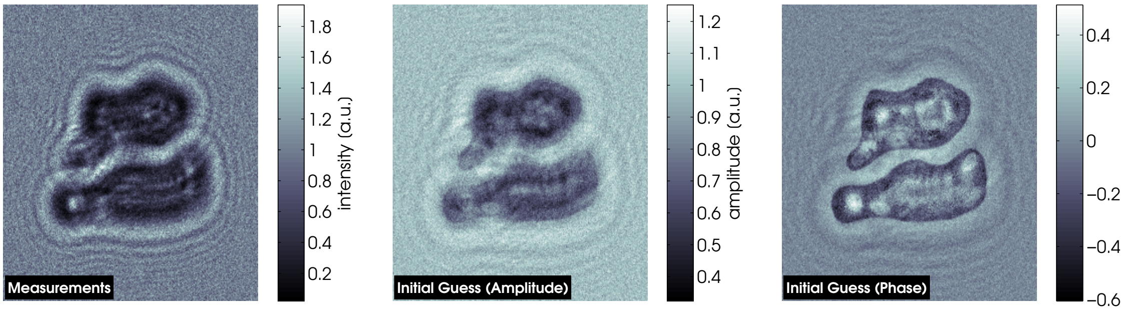
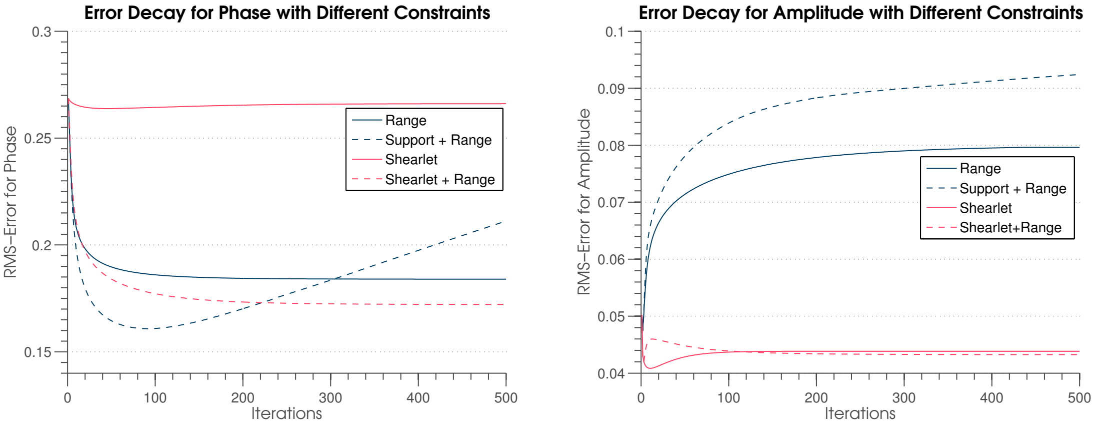


Figure 5.24.: Poisson measurements of a mixed object and initial guesses for phase and amplitude

Comparable to the case of a mixed object with exact data, the reconstructions compared to pure phase or pure amplitude objects are worse for all types of constraints. The error for the support plus range constraint diverges for both phase and amplitude, see Figure 5.25. The best decay is achieved by both shearlet constraints for the amplitude, for the phase the best methods are the range constraint and the shearlet plus range constraint. The reconstructions are shown in Figure 5.26. The quality of the reconstructions matches the decay of the error shown in Figure 5.25. Despite none



(a) (b)

Figure 5.25.: Decay of the root-mean-squared error of a mixed object for different constraints on Poisson measurements for (a) phase and (b) amplitude

of the methods delivering convincing results in this case, the shearlet as well as the shearlet plus range constraint deliver the best of these reconstructions.

This behavior could have been expected since the reconstruction of a mixed object was already inferior with exact data for all constraints compared to pure phase or pure amplitude objects.

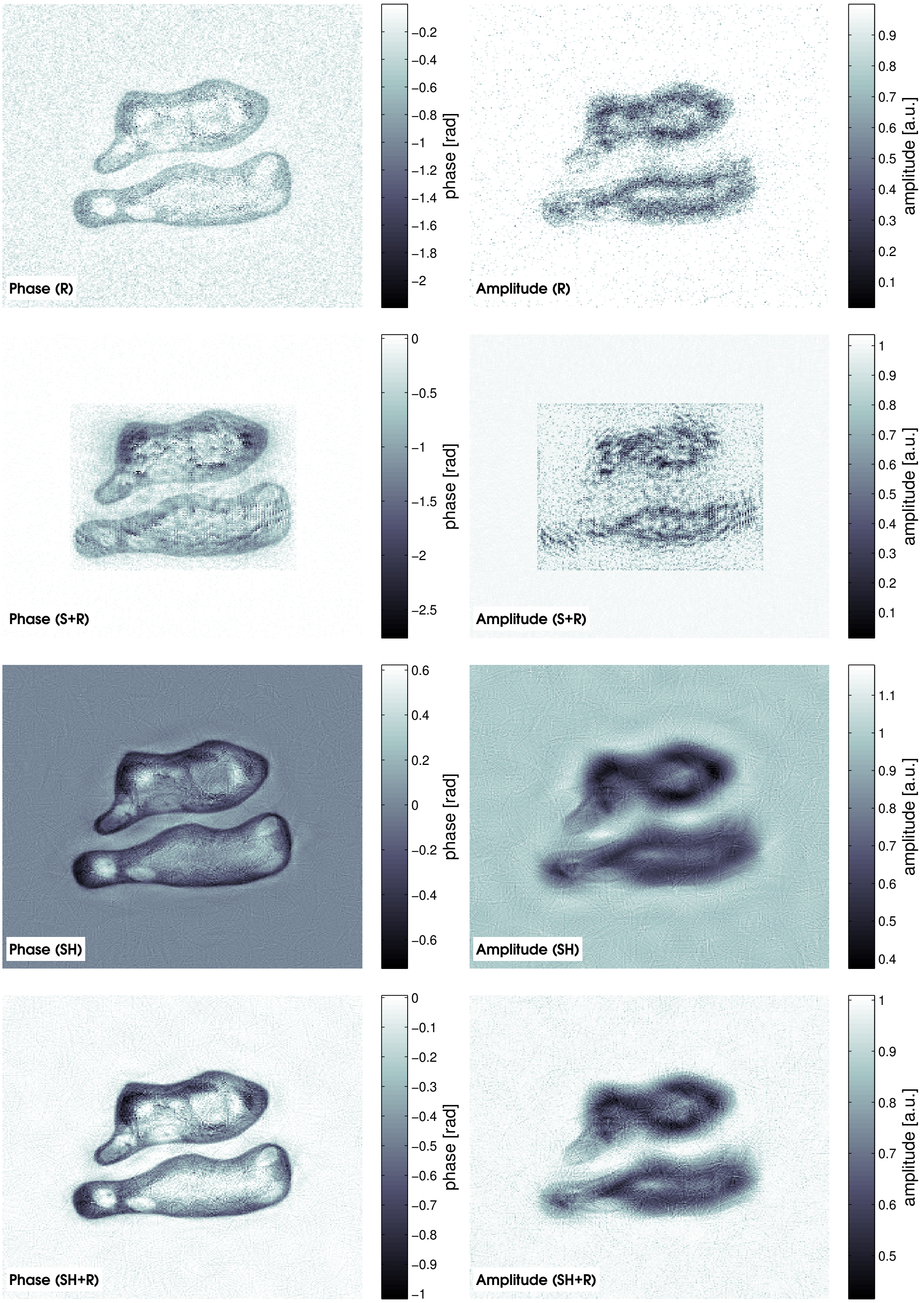


Figure 5.26.: Reconstructions of a mixed object from Poisson measurements for different constraints

5.6. Numerical Results using Smooth-Hard Shrinkage

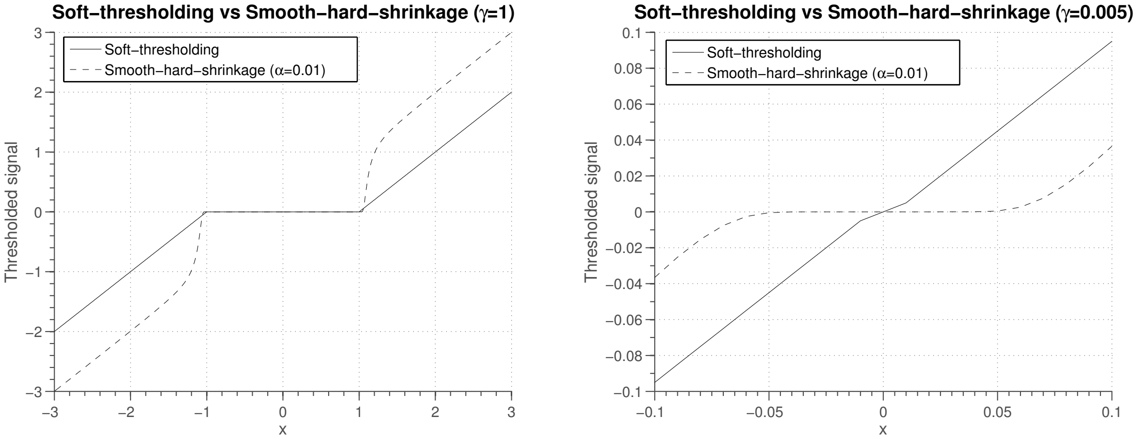
In this section, we investigate the performance of smooth-hard shrinkage of shearlet coefficients. More precisely, we will use the RAAR-(T, γ, V) algorithm defined in Definition 4.6.4 where

$$V(x)[j] := v_\gamma(\left|x[j]\right|) \operatorname{sign}(x[j])$$

with the smooth-hard shrinkage defined in (4.52) which is given for parameters $\alpha, \gamma > 0$ by

$$v_\gamma^{\mathrm{SH}}(x) = \begin{cases} x \exp\left(-\frac{\alpha}{(\mathrm{e}^{x-\gamma}-1)^2}\right), & x \geq \gamma \\ 0, & \text{else.} \end{cases}$$

The corresponding operator is then defined by $P_{S_{T,\gamma,\alpha,V}} = T^\dagger V T$.



(a) (b)

Figure 5.27.: Soft-thresholding compared to smooth-hard shrinkage for different parameters

In Figure 5.27(a), the difference between the soft-thresholding used in the previous section and the smooth-hard shrinkage used in this section, is depicted for $\gamma = 1$ and the selected parameter $\alpha = 0.01$. In Figure 5.27(b), the same functions are depicted but with $\gamma = 0.005$ and $\alpha = 0.01$ which are more realistic parameters for the numerical reconstructions. Here, the functions look rather different. The region where the signal is thresholded is much broader compared to the soft-thresholding and the transition is a lot smoother.

Phase Objects and Smooth-Hard Thresholding with Poisson Data

The parameters are, except for the newly introduced smoothing parameter α, the same as for the numerical experiment of a phase object with Poisson data, cf. Table 5.5. For completeness, Table 5.7 lists all the parameters necessary for this simulation. The initial guesses are therefore the same as in the last section for a phase object, although the noise on the measurements may be slightly different.

Parameter		Value
Decrement of real part of refractive index	δ	$1.6 \cdot 10^{-6}$
Imaginary part of refractive index	β	0
Maximal lateral thickness	τ	$20 \cdot 10^{-6}\text{m}$
Wavelength	λ	10^{-10}
Phase shift distribution	P	τx
Amplitude distribution	A	0
Wave-function in object plane	U	$\exp(-\mathrm{i}k\delta P)$
Number of iterations	N	2000
RAAR relaxation parameter	β_k	$\beta_k \equiv 0.55$
Threshold parameter	γ_k	$\gamma_k \equiv 0.007$
Smoothing parameter	α	$\alpha = 0.01$

Table 5.7.: Parameters for the numerical simulation of the object transmission function for phase objects with Poisson noise using smooth-hard thresholding

The measurements and initial guesses of this experiment are depicted in Figure 5.28. However, since the decay of the error for the shearlet constraint with smooth-hard thresholding indicated a further decay, we ran the algorithm for $N = 2000$ instead of $N = 500$ iterations compared to the previous experiment where the error became stationary after fewer iterations. The error decay using the smooth-hard shrinkage compares favorably to the soft-thresholding. For both methods, with or without combined range constraint, the smooth-hard shrinkage performs better than simple soft-thresholding. Interestingly, the smooth-hard shrinkage without range constraint outperforms the combined approach which includes the range constraint. However, visually, cf. Figure 5.30, the combined approach performs better. This can especially be seen comparing the color bar for all reconstruction results. This is even more visible in the reconstructions depicted in Figure 5.30. In contrast to the soft-thresholding of shearlet coefficients, the smooth-hard thresholding of shearlet coefficients even leads

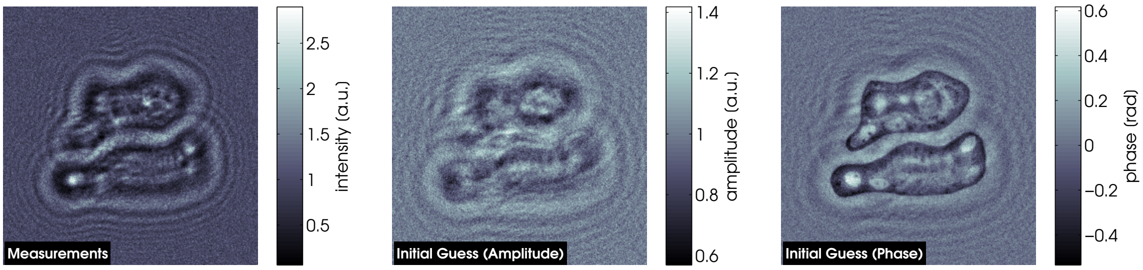
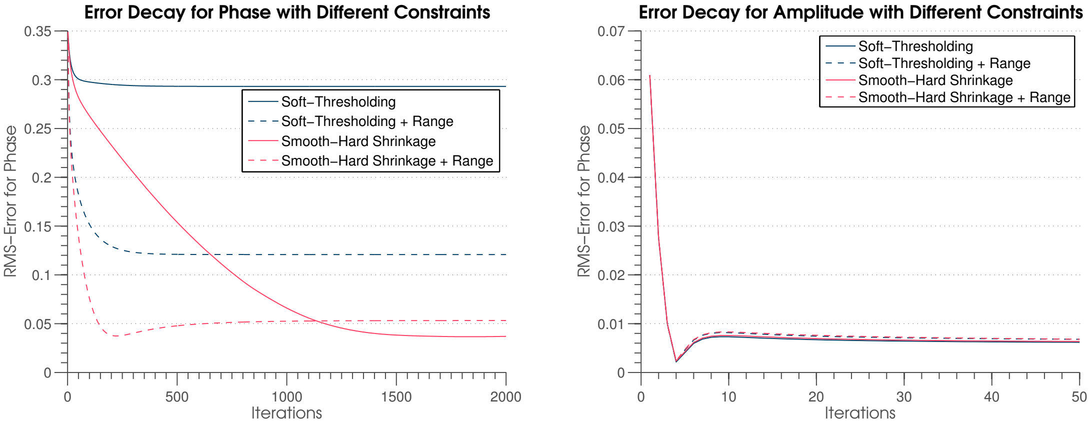


Figure 5.28.: Poisson measurements of a phase object and initial guesses for phase and amplitude



(a) (b)

Figure 5.29.: Decay of the root-mean-squared error of a phase object for different constraints comparing soft-thresholding and smooth-hard shrinkage of shearlet coefficients on Poisson measurements for (a) phase and (b) amplitude

to a quantitative recovery, i.e., the range of the object is approximately recovered without using any a priori information. This means, in this setting, an additional range constraint is not necessary. To conclude, smooth-hard shrinkage seems to outperform the method using soft-thresholding. However, introducing yet another parameter (the smoothing parameter α, see (4.52)) is one more degree of freedom which may crucially influence the performance of the reconstructions. Despite that, especially in situations where the others may fail or deliver unsatisfying results, this methods provides an alternative.

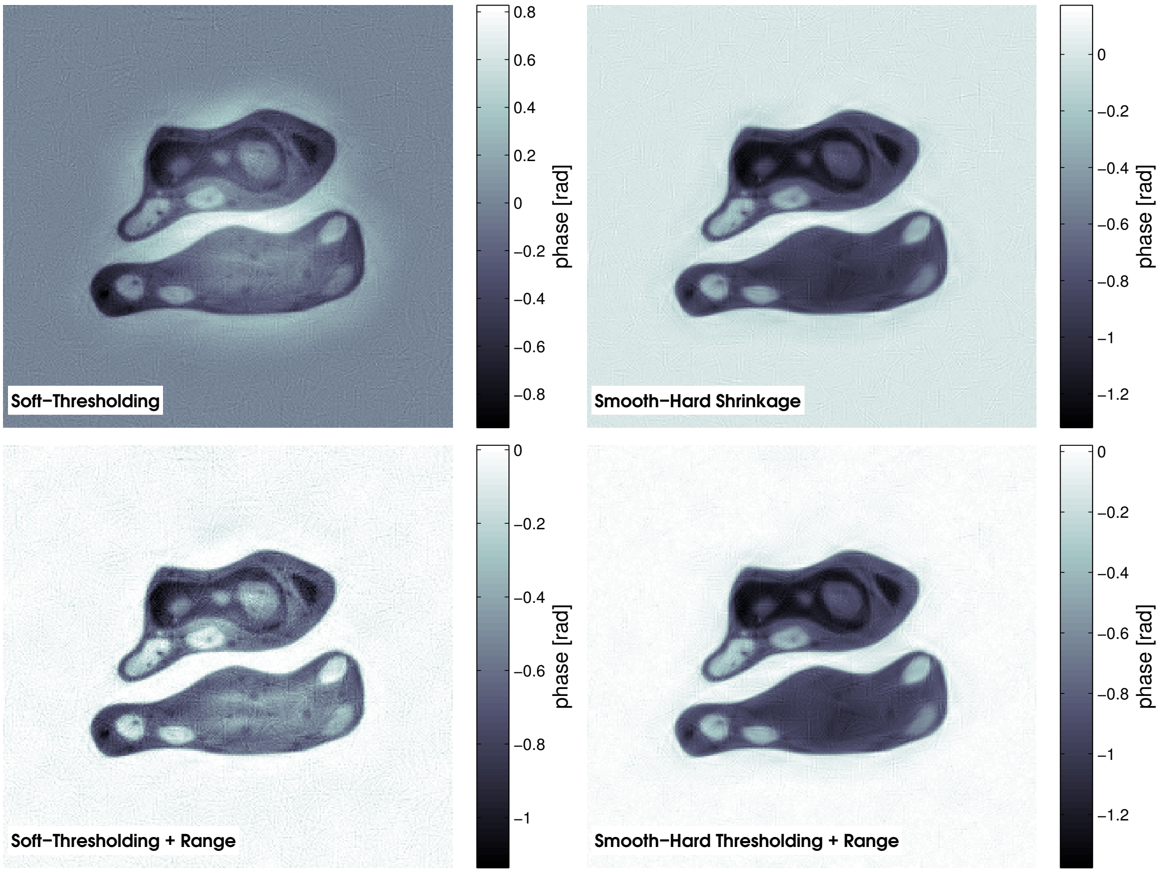


Figure 5.30.: Reconstructions of a phase object from Poisson measurements for different constraints comparing the reconstruction quality using soft-thresholding or smooth-hard shrinkage of shearlet coefficient

5.7. Applicability to Experimental Data

The applicability of the proposed method on experimental data was proven in collaboration with the group of Prof. Dr. Tim Salditt from the Institute for X-ray Physics, Göttingen. The results are promising and the proposed combined method outperforms existing methods. Similar as in our numerical simulations, the shearlet constraint alone does not provide quantitative information but still qualitatively good reconstructions. Combined with the range constraint, one obtains qualitatively and quantitatively superior reconstructions to the other methods. The results are published in [93], a detailed analysis of the method is further given in [92].

6. Conclusion

In this thesis we studied the applicability of sparsity constraints for the discrete, two-dimensional phase retrieval problem. The investigation focused on sparsity constraints which were previously applied in image processing applications such as denoising, deblurring and inpainting. Since fast and stable numerical algorithms are important, we proposed to use compactly supported shearlets that are both theoretically well understood and field-tested in several imaging applications. Therefore, after introducing the mathematical model in Chapter 2, we studied the construction of compactly supported shearlets and their discrete numerical realization in Chapter 3.

The phase retrieval problem can be modeled as a non-convex feasibility problem. We showed in Chapter 4 how to incorporate such sparsity constraints into existing methods. We chose the relaxed averaged alternating reflections algorithm as basis for our method. Using proximity operators which generalize projection operators, we were able to draw several connections to other existing methods. For $\beta_k \equiv 1$, the exact version of our algorithms corresponds to an instance of the Douglas-Rachford algorithm. We used results from [5] to prove the convergence of that algorithm in the convex setting. The investigation of the convex case is concluded by an illustration of the fixed-points of the original method as discussed in [75].

For tight frames, we showed that the operator used in the inexact version of the algorithm, namely $P_{S_{T,\gamma}} = T^* S_\gamma T$, is the proximity operator of a proper, convex, lower semi-continuous function. This result justifies the widespread usage of such mappings in algorithms. Using an estimate for the reflector $R_{S_{T,\gamma}} = 2P_{S_{T,\gamma}} - \mathrm{Id}$ we further proved the boundedness and convergence of the Cesàro sequence of the proposed algorithm for general frames. However, we also showed that for a starting guess $x^{(0)} \in M$ there are no fixed-points of the iteration that lie in M. This poses the question how to chose the soft-thresholding parameter γ. In the absence of noise using $\gamma_k \xrightarrow{k\to\infty} 0$ we obtain reconstructions that are visually not distinguishable from the true solution. Nonetheless, in the case of Poisson noise, a fixed-point in M may not even be wanted since the set M may not contain the true solution. It is therefore not surprising that a constant (or at least non-vanishing) soft-thresholding parameter leads to better results. We close the section with a remark on generalizations of the method to other threshold functions. It can be shown that under some mild assumptions, those threshold functions are again proximity operators of proper, convex, and lower semicontinuous functions at least for the exact version of our algorithm.

We concluded our investigations with the numerical evaluation of our method. Using simulations with exact and Poisson data we showed the usefulness of the proposed method. Depending on the situation, our method is able to outperform other state-of-the-art methods. Furthermore, it is possible to combine the sparsity constraint with other constraints and achieve even better results. Finally, the applicability was examined, in cooperation with the group of Prof. Dr. Tim Salditt, using experimental data. Again, our findings show that the method performs comparably to other methods and achieve significantly better reconstructions if combined with existing constraints.

To conclude, using sparsity constraint for the phase retrieval problem is a promising approach and there are several problems where this approach may be used in the future. Since modern imaging modalities allow for three-dimensional data acquisition, the next step is to apply these methods to three-dimensional phase retrieval data. It is furthermore possible to use these constraints together with Gauss-Newton methods which also provide very good reconstructions results. Using sparsity constraints may improve these methods even further.

Open problems encompass the convergence and convergence rates of the proposed method. The crucial question here is the behavior of the operator $P_{S_{T,\gamma}}$ in the neighborhood of the fixed-points of the iteration. This study will have to take the properties of the transform T into account and hence will lead to assumptions on the mapping T.

Furthermore, the proposed method may be used with different transforms other than the discrete shearlet transform. Using dictionaries, one may be able to construct frames which are better adapted for certain situations and thus have better approximation properties.

A central aspect of the method is the choice of parameters. Introducing the soft-thresholding adds another parameter (next to the relaxation parameter of the original RAAR algorithm) which has to be chosen suitably as optimal reconstructions depend on optimal parameters. Future research will have to take this problem into account and should therefore focus on the development of strategies for the selection of the soft-threshold parameters in an automated fashion.

Finally, the proposed method is also applicable for convex problems. If the projection onto the constraint set is known and fast to compute, the proposed method will provide an efficient algorithm incorporating soft-thresholding of frame coefficients. The results for the convex case presented in Chapter 4 indicate that convergence theory for the convex case covers our method in some parts already.

A. The Mathematical Model

Fundamental Solution to the Helmholtz Equation

Lemma A.1.1. *The function* $G_0 : \mathbb{R}^3 \to \mathbb{C}$ *with (fixed) parameter* $p \in \operatorname{int}(\Omega)$

$$G_0(x;p) = \frac{e^{ik\left|x-p\right|}}{\left|x-p\right|}$$

is a fundamental solution for the Helmholtz equation, i.e.,

$$\left(\Delta + k^2\right) G_0(x;p) = 4\pi\delta(x-p). \tag{A.1}$$

Proof. First, we consider the case $x \neq p$. Recall that the Laplacian in spherical coordinates applied to a function $G_0 : \mathbb{R}^3 \to \mathbb{C}$ is given by

$$\Delta G_0\left(r, \vartheta, \varphi\right) = \underbrace{\frac{1}{r}\left(\frac{\partial^2}{\partial r^2} r G_0(r,\vartheta,\varphi)\right)}_{=:\Delta_r G_0(r,\vartheta,\varphi)} + \underbrace{\frac{1}{r^2 \sin\vartheta}\frac{\partial}{\partial\vartheta}\left(\sin\vartheta \frac{\partial G_0(r,\vartheta,\varphi)}{\partial\vartheta}\right) + \frac{1}{r^2\sin^2\vartheta}\frac{\partial^2 G_0(r,\vartheta,\varphi)}{\partial\varphi^2}}_{=\Delta_{\vartheta,\varphi}G_0(r,\vartheta,\varphi)}.$$

The function $G_0(x;p)$ is radially symmetric, i.e., it only depends on the distance $\left|x-p\right|$. Therefore, $\Delta_{\vartheta,\varphi} G_0\left(r,\vartheta,\varphi\right) = 0$. We denote $r := \left|x-p\right|$ and for $r > 0$ we therefore have

$$\begin{aligned}\Delta G_0(x;p) + k^2 G_0(x;p) &= \Delta_r G_0(x;p) + k^2 G_0(x;p)\\ &= \frac{1}{r}\frac{\partial^2}{\partial r^2} r \frac{e^{ikr}}{r} + k^2 \frac{e^{ikr}}{r} = \frac{1}{r}\left(\frac{\partial^2}{\partial r^2} + k^2\right) e^{ikr}\end{aligned}$$

where

$$\frac{\partial^2 e^{ikr}}{\partial r^2} = -k^2 e^{ikr}$$

and hence for all $x \in \Omega \setminus \mathbb{B}_\varepsilon(p)$,

$$\left(\Delta + k^2\right) G_0(x;p) = 0.$$

For $x = p$ we consider an arbitrarily small volume $\mathbb{B}_\varepsilon(p)$ around $x = p$ and study the limit $\varepsilon \searrow 0$. For the right hand side of (A.1) this yields

$$\int_{\mathbb{B}_\varepsilon(p)} -4\pi\delta(x-p)\,\mathrm{d}x = -4\pi$$

which is independent of ε. For the left hand side we have

$$\int_{\mathbb{B}_\varepsilon(p)} \left(\Delta + k^2\right) G_0(x;p)\,\mathrm{d}x = \int_{\mathbb{B}_\varepsilon(p)} \operatorname{div}\left(\nabla G_0(x;p)\right)\mathrm{d}x + k^2 \int_{\mathbb{B}_\varepsilon(p)} G_0(x;p)\,\mathrm{d}x.$$

Writing the second integral of the right-hand-side in spherical coordinates and using that $G_0(\cdot;p)$ is rotationally invariant, we obtain

$$k^2 \int_{\mathbb{B}_\varepsilon(p)} G_0(x;p)\,\mathrm{d}x = 4\pi k^2 \int_0^\varepsilon r^2 G_0(x;p)\,\mathrm{d}r = 4\pi k^2 \int_0^\varepsilon r\,\mathrm{e}^{\mathrm{i}kr}\,\mathrm{d}r \quad \to \quad 0$$

as $\varepsilon \searrow 0$. For the first integral we apply the divergence theorem by Gauss[1] and obtain

$$\begin{aligned}\int_{\mathbb{B}_\varepsilon(p)} \operatorname{div}\left(\nabla G_0(x;p)\right)\,\mathrm{d}x &= \int_{\partial\mathbb{B}_\varepsilon(p)} \nabla G_0(x;p)\cdot n\,\mathrm{d}\sigma \\ &= \int_{\partial\mathbb{B}_\varepsilon(p)} \left(\frac{\mathrm{i}k\mathrm{e}^{\mathrm{i}kr}}{r} - \frac{\mathrm{e}^{\mathrm{i}kr}}{r^2}\right)\mathrm{d}\sigma \\ &= 4\pi\varepsilon^2\left(\frac{\mathrm{i}k\mathrm{e}^{\mathrm{i}k\varepsilon}}{\varepsilon} - \frac{\mathrm{e}^{\mathrm{i}k\varepsilon}}{\varepsilon^2}\right) \quad \to \quad -4\pi\end{aligned}$$

since $4\pi\mathrm{i}k\varepsilon\,\mathrm{e}^{\mathrm{i}k\varepsilon} \to 0$ and $4\pi\,\mathrm{e}^{\mathrm{i}k\varepsilon} \to 4\pi$ for $\varepsilon \searrow 0$. This concludes the proof. □

---

[1] Here, n denotes the outward facing normal vector on $\mathbb{B}_\varepsilon(p)$ and hence, the derivative in normal direction is simply the derivative with respect to r.

Properties of Poisson Data

Solution of the Associated Differential Equation

An extensive description of variation of constants is given in [107] and [53]. We recall a brief version of it.

Fact A.1.2 (Variation of Constants). *Consider a general inhomogeneous differential equation*

$$y'(x) = A(x)y(x) + b(x) \tag{A.2}$$

with antiderivative $F(x) = \int_{x_0}^{x} A(t)\,\mathrm{d}t$. *Then all solutions of the homogeneous equation* $y' = Ay$ *are given by*

$$\left\{y(x) = C\mathrm{e}^{F(x)} \mid C \in \mathbb{R}\right\}.$$

The ansatz for the inhomogeneity consists in letting $C = C(x)$, *i.e.,*

$$y(x) = C(x)\mathrm{e}^{F(x)} \tag{A.3}$$

or

$$y'(x) = C(x)A(x)\mathrm{e}^{F(x)} + C'(x)\mathrm{e}^{F(x)} = A(x)y(x) + C'(x)\mathrm{e}^{F(x)}.$$

Hence, y in (A.3) *solves the differential equation* (A.2) *if and only if*

$$C'(x) = b(x)\mathrm{e}^{-F(x)},$$

i.e.,

$$C(x) = \int_{x_0}^{x} b(t)\mathrm{e}^{-F(t)}\,\mathrm{d}t + C$$

for some $C \in \mathbb{R}$. *Therefore, the set of all solutions to the inhomogeneous equation is given by*

$$\left\{y(x) = \mathrm{e}^{F(x)}\left[\int_{x_0}^{x} b(t)\mathrm{e}^{-F(t)}\,\mathrm{d}t + C\right] \mid C \in \mathbb{R}\right\}.$$

Usually, one would use this to obtain a solution of the differential equation and using an induction argument for $K-1 \to K$. Since the solution is known in this case and the proof is rather technical, we simply differentiate the solution and check that it fulfills the differential equation (A.4).

Fact A.1.3. *The solution for the differential equation*

$$\frac{\mathrm{d}\mathbb{P}(K,t,t+\tau)}{\mathrm{d}\tau} = \lambda(t+\tau)\left[\mathbb{P}(K-1,t,t+\tau) - \mathbb{P}(K,t,t+\tau)\right] \tag{A.4}$$

with initial value $P(0,t,t) = 1$ *is given by*

$$\mathbb{P}(K,t,t+\tau) = \frac{\left[\int_t^{t+\tau} \lambda(\xi)\,\mathrm{d}\xi\right]^K}{K!} \exp\left[-\int_t^{t+\tau} \lambda(\xi)\,\mathrm{d}\xi\right].$$

Proof. Since the solution is known, we only need to differentiate the solution in order to check that it fulfills (A.4). We use the product and chain rule to obtain

$$\begin{aligned}\frac{\mathrm{d}\mathbb{P}(K,t,t+\tau)}{\mathrm{d}\tau} &= \underbrace{\frac{K\cdot\left[\int_t^{t+\tau} \lambda(\xi)\,\mathrm{d}\xi\right]^{K-1}}{K!} \exp\left[-\int_t^{t+\tau} \lambda(\xi)\,\mathrm{d}\xi\right]}_{=\mathbb{P}(K-1,t,t+\tau)} \lambda(t+\tau) \\ &\quad + \underbrace{\frac{\left[\int_t^{t+\tau} \lambda(\xi)\,\mathrm{d}\xi\right]^{K}}{K!} \exp\left[-\int_t^{t+\tau} \lambda(\xi)\,\mathrm{d}\xi\right]}_{=\mathbb{P}(K,t,t+\tau)} (-\lambda(t+\tau)) \\ &= \lambda(t+\tau)\left[\mathbb{P}(K-1,t,t+\tau) - \mathbb{P}(K,t,t+\tau)\right]\end{aligned}$$

where we used ${}^{K}/_{K!} = {}^{K}/_{K\cdot(K-1)!} = {}^{1}/_{(K-1)!}$. ☐

B. Fundamental Results from Convex Analysis

In this chapter we introduce fundamental results from convex analysis in order to prove a result regarding the asymptotic behaviour of the modified RAAR iteration. Therefore we carefully derive the most important result of this section that the proximal mapping of the ℓ_1-norm is firmly non-expansive. This chapter is based on [5].

Firmly non-expansiveness of Proximal Mappings

Lemma B.1.1 (Lemma 2.12 from [5]). *For a real Hilbert space* $\mathcal{H}$ *and* $x,\ y \in \mathcal{H}$ *the following hold:*

(i) $\langle x, y\rangle \leq 0 \ \Leftrightarrow\ \forall \alpha \in \mathbb{R}_+ : \|x\| \leq \left\|x - \alpha y\right\| \ \Leftrightarrow\ \forall \alpha \in [0, 1] : \|x\| \leq \left\|x - \alpha y\right\|$

(ii) $x \perp y \ \Leftrightarrow\ \forall \alpha \in \mathbb{R} : \|x\| \leq \left\|x - \alpha y\right\| \ \Leftrightarrow\ \forall \alpha \in [-1, 1] : \|x\| \leq \left\|x - \alpha y\right\|$

Proof.

(i) Observe that for all $\alpha \in \mathbb{R}$ we have

$$\begin{aligned}
\left\|x - \alpha y\right\|^2 - \|x\|^2 &= \langle x - \alpha y, x - \alpha y\rangle - \langle x, x\rangle \\
&= \langle x - \alpha y, x\rangle - \langle x - \alpha y, \alpha y\rangle - \langle x, x\rangle \\
&= \langle x, x\rangle - \langle \alpha y, x\rangle - \alpha \langle x - \alpha y, y\rangle - \langle x, x\rangle \\
&= -\alpha \langle y, x\rangle - \alpha \left(\langle x, y\rangle - \langle \alpha y, y\rangle\right) \\
&= -\alpha \langle x, y\rangle - \alpha \langle x, y\rangle + \alpha^2 \left\|y\right\|^2 \\
&= \alpha \Big(\alpha \left\|y\right\|^2 - 2 \langle x, y\rangle\Big),
\end{aligned}$$

i.e.,

$$\left\|x-\alpha y\right\|^2-\|x\|^2=\alpha\left(\alpha\left\|y\right\|^2-2\left\langle x,y\right\rangle\right) \tag{B.1}$$

and

$$\|x\|^2=\left\|x-\alpha y\right\|^2-\alpha^2\left\|y\right\|^2+2\alpha\left\langle x,y\right\rangle.$$

For $\alpha\in\mathbb{R}$ we have $\alpha^2\left\|y\right\|^2\geq 0$ and can estimate

$$\|x\|^2\leq\left\|x-\alpha y\right\|^2+2\alpha\left\langle x,y\right\rangle.$$

By assumption we have $\langle x,y\rangle\leq 0$. Hence, we have for all $\alpha\in\mathbb{R}_+$ the estimate

$$\|x\|\leq\left\|x-\alpha y\right\|,$$

and especially for all $\alpha\in[0,1]$. Conversely, suppose that $\|x\|^2\leq\left\|x-\alpha y\right\|^2$ holds for all $\alpha\in[0,1]$. Using (B.1) we have

$$\underbrace{\left\|x-\alpha y\right\|^2-\|x\|^2}_{\geq 0}=\alpha\left(\alpha\left\|y\right\|^2-2\left\langle x,y\right\rangle\right)$$

and therefore

$$\langle x,y\rangle\leq\frac{\alpha\left\|y\right\|^2}{2}.$$

For $\alpha\searrow 0$ the claim follows.

(ii) Note that $x\perp y\ \Leftrightarrow\ (\langle x,y\rangle\leq 0\ \wedge\ \langle x,-y\rangle\leq 0)$ and use (i). □

Lemma B.1.2 (Lemma 2.13 from [5]). *Let* $(x_i)_{i\in I},(u_i)_{i\in I}\subset\mathcal{H}$ *be finite families in a real Hilbert space* $\mathcal{H}$, $(\alpha_i)_{i\in I}\subset\mathbb{R}$ *a finite family in* $\mathbb{R}$ *such that* $\sum_{i\in I}\alpha_i=1$ *and* $0\in I$. *Then the following holds:*

(i) $\left\langle\sum_{i\in I}\alpha_i x_i,\sum_{j\in I}\alpha_j u_j\right\rangle+\frac{1}{2}\sum_{i\in I}\sum_{j\in I}\alpha_i\alpha_j\left\langle x_i-x_j,u_i-u_j\right\rangle=\sum_{i\in I}\alpha_i\left\langle x_i,u_i\right\rangle$,

(ii) $\left\|\sum_{i\in I}\alpha_i x_i\right\|^2+\frac{1}{2}\sum_{i\in I}\sum_{j\in I}\alpha_i\alpha_j\left\|x_i-x_j\right\|^2=\sum_{i\in I}\alpha_i\|x_i\|^2$.

Proof. First observe that

$$\begin{aligned}\left\langle\sum_{i\in I}\alpha_i x_i, \sum_{j\in I}\alpha_j u_j\right\rangle &= \left\langle\alpha_0 x_0, \sum_{j\in I}\alpha_j u_j\right\rangle + \left\langle\sum_{i\in I\setminus\{0\}}\alpha_i x_i, \sum_{j\in I}\alpha_j u_j\right\rangle\\ &= \sum_{i\in I}\left\langle\alpha_i x_i, \sum_{j\in I}\alpha_j u_j\right\rangle = \sum_{i\in I}\alpha_i\left\langle x_i, \sum_{j\in I}\alpha_j u_j\right\rangle\\ &= \sum_{i\in I}\sum_{j\in I}\alpha_i\left\langle x_i, \alpha_j u_j\right\rangle = \sum_{i\in I}\sum_{j\in I}\alpha_i\alpha_j\left\langle x_i, u_j\right\rangle\end{aligned}$$

Therefore we have

$$2\left\langle\sum_{i\in I}\alpha_i x_i, \sum_{j\in I}\alpha_j u_j\right\rangle = \sum_{i\in I}\sum_{j\in I}\alpha_i\alpha_j\left(\left\langle x_i, u_j\right\rangle + \left\langle x_j, u_i\right\rangle\right). \tag{B.2}$$

Expanding the sum of scalar product yields

$$\begin{aligned}\left\langle x_i, u_j\right\rangle + \left\langle x_j, u_i\right\rangle &= \left\langle x_i + x_j - x_j, u_j + u_i - u_i\right\rangle + \left\langle x_j, u_i\right\rangle\\ &= \left\langle x_i - x_j, u_j + u_i - u_i\right\rangle + \left\langle x_j, u_j + u_i - u_i\right\rangle + \left\langle x_j, u_i\right\rangle\\ &= \left\langle x_i - x_j, -u_i + u_j\right\rangle + \left\langle x_i - x_j, u_i\right\rangle + \left\langle x_j, u_j + u_i - u_i\right\rangle + \left\langle x_j, u_i\right\rangle\\ &= -\left\langle x_i - x_j, u_i - u_j\right\rangle + \langle x_i, u_i\rangle - \left\langle x_j, u_i\right\rangle + \left\langle x_j, u_i\right\rangle + \left\langle x_j, u_j\right\rangle\\ &= \langle x_i, u_i\rangle + \left\langle x_j, u_j\right\rangle - \left\langle x_i - x_j, u_i - u_j\right\rangle.\end{aligned}$$

Plugging this into (B.2) yields by $\sum_{i\in I}\alpha_i = 1$

$$\begin{aligned}2\left\langle\sum_{i\in I}\alpha_i x_i, \sum_{j\in I}\alpha_j u_j\right\rangle &= \sum_{i\in I}\sum_{j\in I}\alpha_i\alpha_j\left(\langle x_i, u_i\rangle + \left\langle x_j, u_j\right\rangle - \left\langle x_i - x_j, u_i - u_j\right\rangle\right)\\ &= 2\sum_{i\in I}\alpha_i\langle x_i, u_i\rangle - \sum_{i\in I}\sum_{j\in I}\alpha_i\alpha_j\left\langle x_i - x_j, u_i - u_j\right\rangle\end{aligned}$$

and which yields the first equation of the Lemma. For the second equation consider the case where $(u_i)_{i\in I} := (x_i)_{i\in I}$. Then we have

$$\begin{aligned}\left\langle\sum_{i\in I}\alpha_i x_i, \sum_{i\in I}\alpha_i x_i\right\rangle &= \left\|\sum_{i\in I}\alpha_i x_i\right\|^2,\\ \sum_{i\in I}\alpha_i\langle x_i, x_i\rangle &= \sum_{i\in I}\alpha_i\|x_i\|^2,\end{aligned}$$

$$\sum_{i\in I}\sum_{j\in I}\alpha_i\alpha_j\left\langle x_i - x_j, x_i - x_j\right\rangle = \sum_{i\in I}\sum_{j\in I}\alpha_i\alpha_j\left\|x_i - x_j\right\|^2$$

and therefore the second equation in the Lemma holds. □

Corollary B.1.3 (Corollary 2.14 from [5]). *Let $x, y \in \mathcal{H}, \alpha \in \mathbb{R}$. Then the Lemma above implies*

$$\left\|\alpha x + (1-\alpha)y\right\|^2 + \alpha(1-\alpha)\left\|x - y\right\|^2 = \alpha\left\|x\right\|^2 + (1-\alpha)\left\|y\right\|^2.$$

As a next step, we define contraction properties of mappings which will be important for convergence analysis. The following theorem will characterize equivalences between different properties of the operator T and the reflector $2T - \mathrm{Id}$.

Definition B.1.4 (Definition 4.1 from [5]). *Let $D \subset \mathcal{H}$ be nonempty and $T : D \to \mathcal{H}$. Then T is*

(i) firmly non-expansive *if*

$$\forall x, y \in D: \qquad \left\|Tx - Ty\right\|^2 + \left\|(\mathrm{Id} - T)\,x - (\mathrm{Id} - T)\,y\right\|^2 \leq \left\|x - y\right\|^2, \tag{B.3}$$

(ii) non-expansive *if it is Lipschitz continuous with constant 1, i.e.,*

$$\forall x, y \in D: \qquad \left\|Tx - Ty\right\| \leq \left\|x - y\right\|, \tag{B.4}$$

(iii) quasi-non-expansive *if*

$$\forall x \in D, \forall y \in \operatorname{Fix} T: \qquad \left\|Tx - y\right\| \leq \left\|x - y\right\|, \tag{B.5}$$

(iv) strictly quasi-non-expansive *if*

$$\forall x \in D \setminus \operatorname{Fix} T, y \in \operatorname{Fix} T: \qquad \left\|Tx - y\right\| < \left\|x - y\right\|. \tag{B.6}$$

It is immediate to see that firmly non-expansiveness implies non-expansiveness and furthermore quasi-non-expansiveness is implied by both of them.

Proposition B.1.5 (Proposition 4.2 from [5]). *Let* $D \subset \mathcal{H}$ *be nonempty and* $T : D \to \mathcal{H}$*. Then the following assertions are equivalent:*

(i) T *is firmly non-expansive.*

(ii) $\mathrm{Id} - T$ *is firmly non-expansive.*

(iii) $2T - \mathrm{Id}$ *is non-expansive.*

(iv) $\forall x,\ y \in D:\ \left\|Tx - Ty\right\|^2 \leq \langle x - y, Tx - Ty\rangle$.

(v) $\forall x,\ y \in D:\ 0 \leq \langle Tx - Ty, (\mathrm{Id} - T)\,x - (\mathrm{Id} - T)\,y\rangle$.

(vi) $\forall x,\ y \in D,\ \forall \alpha \in [0,1]:\ \left\|Tx - Ty\right\| \leq \left\|\alpha(x - y) + (1 - \alpha)(Tx - Ty)\right\|$.

Proof. (i) ⇔ (ii). Follows immediately with $S := \mathrm{Id} - T$, hence $\mathrm{Id} - S = T$ which implies S is firmly non-expansive if and only if T is firmly non-expansive.

(i) ⇔ (iii). Fix $x, y \in D$, define $R := 2T - \mathrm{Id}$ and set

$$\begin{aligned}
\mu &:= \left\|Tx - Ty\right\|^2 + \left\|(\mathrm{Id} - T)\,x - (\mathrm{Id} - T)\,y\right\|^2 - \left\|x - y\right\|^2, \\
\nu &:= \left\|Tx - Ty\right\|^2 + \left\|x - y\right\|^2.
\end{aligned}$$

Furthermore, we have

$$\begin{aligned}
\left\|Rx - Ry\right\|^2 &= \left\|(2T - \mathrm{Id})x - (2T - \mathrm{Id})y\right\|^2 \\
&= \left\|2Tx - x - 2Ty + y\right\|^2 \\
&= \left\|2(Tx - Ty) + (1 - 2)(x - y)\right\|^2
\end{aligned}$$

Corollary B.1.3 with $\alpha = 2$ states that

$$\|2u + (1 - 2)v\|^2 - 2\,\|u - v\|^2 = 2\,\|u\|^2 + (1 - 2)\,\|v\|^2 .$$

Setting $u := Tx - Ty, v := x - y$ we have

$$\begin{aligned}
\left\|Rx - Ry\right\|^2 &= \left\|2(Tx - Ty) + (1 - 2)(x - y)\right\|^2 \\
&= 2\left\|Tx - Ty\right\|^2 - \left\|x - y\right\|^2 + 2\left\|Tx - Ty - (x - y)\right\|^2
\end{aligned}$$

$$\begin{aligned}
&= 2\left\|Tx - Ty\right\|^2 - \left\|x - y\right\|^2 + 2\left\|Tx - x - Ty + y\right\|^2 \\
&= 2\left\|Tx - Ty\right\|^2 - \left\|x - y\right\|^2 + 2\left\|(\mathrm{Id} - T)x + Ty - y\right\|^2 \\
&= 2\left\|Tx - Ty\right\|^2 - \left\|x - y\right\|^2 + 2\left\|(\mathrm{Id} - T)x - (\mathrm{Id} - T)y\right\|^2
\end{aligned}$$

Therefore we obtain

$$\left\|Rx - Ry\right\|^2 = 2\left\|Tx - Ty\right\|^2 - \left\|x - y\right\|^2 + 2\left\|(\mathrm{Id} - T)x - (\mathrm{Id} - T)y\right\|^2 .$$

Substracting $\left\|x - y\right\|^2$ on both sides yields

$$\left\|Rx - Ry\right\|^2 - \left\|x - y\right\|^2 = 2\left(\left\|Tx - Ty\right\|^2 + \left\|(\mathrm{Id} - T)x - (\mathrm{Id} - T)y\right\|^2 - \left\|x - y\right\|^2\right),$$

i.e., $\nu = 2\mu$. This implies that $\nu \leq 0$ if and only if $\mu \leq 0$ if and only if T is firmly non-expansive.

(i) ⇔ (iv). Observe that

$$\begin{aligned}
\left\|(\mathrm{Id} - T)x - (\mathrm{Id} - T)y\right\|^2 &= \left\|x - Tx - y + Ty\right\|^2 = \left\|Ty - Tx + x - y\right\|^2 \\
&= \left\|(Tx - Ty) - (x - y)\right\|^2 \\
&= \left\|Tx - Ty\right\|^2 + \left\|x - y\right\|^2 - 2\left\langle x - y, Tx - Ty\right\rangle .
\end{aligned}$$

Using the definition of firmly non-expansiveness yields that T is firmly non-expansive if and only if

$$2\left\|Tx - Ty\right\|^2 + \left\|x - y\right\|^2 - 2\left\langle x - y, Tx - Ty\right\rangle \leq \left\|x - y\right\|^2,$$

i.e.,

$$\left\|Tx - Ty\right\|^2 - \left\langle x - y, Tx - Ty\right\rangle \leq 0$$

and hence, $\left\|Tx - Ty\right\|^2 \leq \left\langle x - y, Tx - Ty\right\rangle$.

(iv) ⇔ (v). Starting with (iv) we have

$$\begin{aligned}
\left\|Tx - Ty\right\|^2 \leq \left\langle x - y, Tx - Ty\right\rangle \quad &\Leftrightarrow \quad \left\langle Tx - Ty, Tx - Ty\right\rangle \leq \left\langle x - y, Tx - Ty\right\rangle \\
&\Leftrightarrow \quad 0 \leq \left\langle x - y - Tx + Ty, Tx - Ty\right\rangle \\
&\Leftrightarrow \quad 0 \leq \left\langle (\mathrm{Id} - T)x - (\mathrm{Id} - T)y, Tx - Ty\right\rangle
\end{aligned}$$

$$\Leftrightarrow \quad 0 \leq \langle Tx - Ty, (\mathrm{Id} - T)x - (\mathrm{Id} - T)y \rangle$$

and hence (iv).

(v) ⇔ (vi) Using Lemma B.1.1 we have

$$\langle u, v \rangle \leq 0 \quad \Leftrightarrow \quad \forall \alpha \in [0,1] : \|u\| \leq \|u - \alpha v\|.$$

Multiplying (v) with −1 gives

$$\Big\langle \underbrace{Tx - Ty}_{=:u}, \underbrace{(T - \mathrm{Id})x - (T - \mathrm{Id})y}_{=:v} \Big\rangle \leq 0.$$

Using Lemma B.1.1 implies

$$\begin{aligned} \|u\| \leq \|u - \alpha v\| \quad &\Leftrightarrow \quad \left\|Tx - Ty\right\| \leq \left\|Tx - Ty - \alpha\left((T - \mathrm{Id})x - (T - \mathrm{Id})y\right)\right\| \\ &\Leftrightarrow \quad \left\|Tx - Ty\right\| \leq \left\|Tx - Ty - \alpha Tx + \alpha x + \alpha Ty - \alpha y\right\| \\ &\Leftrightarrow \quad \left\|Tx - Ty\right\| \leq \left\|\alpha(x - y) + (1 - \alpha)(Tx - Ty)\right\|. \end{aligned}$$

This completes the proof. □

Definition B.1.6 (Definition 12.23 from [5]). *Let $f \in \Gamma_0(\mathcal{H})$ and $x \in \mathcal{H}$. Then* $\mathrm{prox}_f(x)$ *is the unique point that satisfies*

$$f(x) = \min_{y \in \mathcal{H}} \left\{ f(y) + \frac{1}{2}\|x - y\|^2 \right\} = f\left(\mathrm{prox}_f(x)\right) + \frac{1}{2}\left\|x - \mathrm{prox}_f(x)\right\|^2.$$

The operator $\mathrm{prox}_f : \mathcal{H} \to \mathcal{H}$ *is the* proximity operator *or* proximal mapping *of f.*

The next proposition characterizes the proximal point of f.

Proposition B.1.7 (Proposition 12.26 from [5]). *Let $f \in \Gamma_0(\mathcal{H})$ and $x, p \in \mathcal{H}$. Then*

$$p = \mathrm{prox}_f(x) \quad \Leftrightarrow \quad \forall y \in \mathcal{H} : \langle y - p, x - p \rangle + f(p) \leq f(y).$$

Proof. First, suppose that $p = \operatorname{prox}_f(x)$ and set for any arbitrary $y \in \mathcal{H}$ and all $\alpha \in (0,1) : p_\alpha := \alpha y + (1-\alpha)p$. By definition of the proximity operator we have

$$f(p) + \frac{1}{2}\left\|x - p\right\|^2 \leq f\left(p_\alpha\right) + \frac{1}{2}\left\|x - p_\alpha\right\|^2$$

which is equivalent to

$$f(p) \leq f\left(\alpha y + (1-\alpha)p\right) + \frac{1}{2}\left\|x - p_\alpha\right\|^2 - \frac{1}{2}\left\|x - p\right\|^2.$$

Using the convexity of f we obtain

$$f(p) \leq \alpha f(y) + (1-\alpha)f(p) + \frac{1}{2}\left\langle x - p_\alpha, x - p_\alpha\right\rangle - \frac{1}{2}\left\langle x - p, x - p\right\rangle.$$

Expanding the first inner product gives

$$\begin{aligned}\left\langle x - p_\alpha, x - p_\alpha\right\rangle &= \left\langle x - \alpha y - p + \alpha p, x - p\right\rangle + \left\langle x - \alpha y - p + \alpha p, \alpha p - \alpha y\right\rangle \\ &= \left\langle x - p, x - p\right\rangle + \alpha\left\langle p - y, x - p\right\rangle + \alpha\left\langle x - \alpha y - p + \alpha p, p - y\right\rangle \\ &= \left\langle x - p, x - p\right\rangle - \alpha\left\langle x - p, y - p\right\rangle + \alpha\left\langle x - p, p - y\right\rangle + \alpha^2\left\langle y - p, y - p\right\rangle \\ &= \left\langle x - p, x - p\right\rangle - 2\alpha\left\langle y - p, x - p\right\rangle + \alpha^2\left\|y - p\right\|^2.\end{aligned}$$

Therefore, we obtain

$$f(p) \leq \alpha f(y) + (1-\alpha)f(p) - \alpha\left\langle y - p, x - p\right\rangle + \frac{\alpha^2}{2}\left\|y - p\right\|^2,$$

i.e.,

$$0 \leq \alpha f(y) - \alpha f(p) - \alpha\left\langle y - p, x - p\right\rangle + \frac{\alpha^2}{2}\left\|y - p\right\|^2$$

or after dividing by α

$$0 \leq f(y) - f(p) - \left\langle y - p, x - p\right\rangle + \frac{\alpha}{2}\left\|y - p\right\|^2.$$

For $\alpha \searrow 0$ we obtain the desired result.

We are now able to state the main result of this section.

Proposition B.1.8 (Proposition 12.27 from [5]). *For $f \in \Gamma_0(\mathcal{H})$ both* $\operatorname{prox}_f(x)$ *and* $\operatorname{Id} - \operatorname{prox}_f(x)$ *are firmly non-expansive.*

Proof. Let $x, y \in \mathcal{H}$ and set $p := \operatorname{prox}_f(x)$ and $q := \operatorname{prox}_f(y)$. By Proposition B.1.7 we have

$$\begin{aligned} \langle q-p, x-p\rangle + f(p) &\le f(q) \\ \langle p-q, y-q\rangle + f(q) &\le f(p) \end{aligned}$$

Since $p, q \in \operatorname{dom} f$, we can add up the two inequalities yielding

$$\langle q-p, x-p\rangle + \langle p-q, y-p\rangle + f(p) + f(q) \le f(p) + f(q).$$

Substracting $f(p) + f(q)$ and multiplying by -1 yields

$$0 \le -\langle q-p, x-p\rangle - \langle p-q, y-p\rangle \quad \Leftrightarrow \quad 0 \le \langle p-q, x-p\rangle - \langle p-q, y-p\rangle$$

and hence

$$0 \le \langle p-q, (x-p) - (y-q)\rangle .$$

But this is exactly (v) in Proposition B.1.5 and hence, prox_f and $\operatorname{Id} - \operatorname{prox}_f$ are firmly non-expansive.

Corollary B.1.9. *The operator* $\operatorname{prox}_{\ell_1}(x)$ *is firmly non-expansive.*

Proof. Since the ℓ_1-norm is continuous and convex it is in $\Gamma_0(\mathcal{H})$ and hence its proximal mapping is firmly non-expansive.

Subdifferentials of Convex Functions

We cite a fact from [5], the subdifferential sum rule. This result will be used in the proof of Proposition 4.2.5.

Fact B.1.10 (Theorem 16.37 from [5]). *Let* $f, g \in \Gamma_0(\mathcal{H})$ *and*

$$0 \in \operatorname{sri}(\operatorname{dom} f - \operatorname{dom} g),$$

then

$$\partial(f+g) = \partial f + \partial g.$$

Bibliography

[1] R. Ansari. Efficient IIR and FIR fan filters. *IEEE Transactions on Circuits and Systems*, 34(8):941–945, 1987.

[2] L. Baghaei, A. Rad, B. Dai, P. Pianetta, R. F. Pease, and J. Miao. Iterative phase recovery using wavelet domain constraints. *Journal of Vacuum Science and Technology B*, 27, 2009.

[3] R. H. Bamberger and M. J. T. Smith. A filter bank for the directional decomposition of images: Theory and design. *IEEE Transactions on Signal Processing*, 40(4):882–893, 1992.

[4] H. H. Bauschke and J. M. Borwein. On the convergence of von Neumann's alternating projection algorithm for two sets. *Set-Valued Analysis*, 1(2):185–212, 1993.

[5] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2011.

[6] H. H. Bauschke, P. L. Combettes, and D. R. Luke. On the structure of some phase retrieval algorithms. In *International Conference on Image Processing*, volume 2. IEEE, 2002.

[7] H. H. Bauschke, P. L. Combettes, and D. R. Luke. Phase retrieval, error reduction algorithm, and Fienup variants: A view from convex optimization. *Journal of the Optical Society of America A*, 19(7):1334–1345, 2002.

[8] H. H. Bauschke, P. L. Combettes, and D. R. Luke. Hybrid projection–reflection method for phase retrieval. *Journal of the Optical Society of America A*, 20(6):1025–1034, 2003.

[9] H. H. Bauschke, P. L. Combettes, and D. R. Luke. Finding best approximation pairs relative to two closed convex sets in Hilbert spaces. *Journal of Approximation Theory*, 127(2):178–192, 2004.

[10] H. H. Bauschke, D. R. Luke, H. M. Phan, and X. Wang. Restricted normal cones and the method of alternating projections: applications. *Set-Valued and Variational Analysis*, 21(3):475–501, 2013.

[11] H. H. Bauschke, D. R. Luke, H. M. Phan, and X. Wang. Restricted normal cones and the method of alternating projections: theory. *Set-Valued and Variational Analysis*, 21(3):431–473, 2013.

[12] H. H. Bauschke, D. R. Luke, H. M. Phan, and X. Wang. Restricted normal cones and sparsity optimization with affine constraints. *Foundations of Computational Mathematics*, 14(1):63–83, 2014.

[13] M. Born and E. Wolf. *Principles of Optics*. Cambridge University Press, Cambridge, England, 7th (expanded) edition, 1999.

[14] K. Bredies, D. A. Lorenz, and S. Reiterer. Minimization of non-smooth, non-convex functionals by iterative thresholding. *Journal of Optimization Theory and Applications*, 165(1):78–112, 2015.

[15] A. Calderón. Intermediate spaces and interpolation. *Studia Mathematica*, 1(Special Series):31–34, 1963.

[16] E. J. Candès. *Ridgelets: Theory and Applications*. PhD thesis, Department of Statistics, Stanford University, 1998.

[17] E. J. Candès and D. L. Donoho. New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities. *Communications on Pure and Applied Mathematics*, 57(2):219–266, 2004.

[18] E. J. Candès, Y. C. Eldar, T. Strohmer, and V. Voroninski. Phase retrieval via matrix completion. *SIAM Journal on Imaging Sciences*, 6(1):199–225, 2013.

[19] R. Chartrand. Shrinkage mappings and their induced penalty functions. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1026–1029. IEEE, 2014.

[20] H. Cheng, H. Liu, Q. Zhang, and S. Wei. Phase retrieval using the transport-of-intensity equation. In *Fifth International Conference on Image and Graphics*, pages 417–421. IEEE, 2009.

[21] O. Christensen. Six (seven) problems in frame theory. In *New Perspectives on Approximation and Sampling Theory*, pages 337–358. Springer, 2014.

[22] P. L. Combettes and J.-C. Pesquet. A Douglas–Rachford splitting approach to nonsmooth convex variational signal recovery. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):564–574, 2007.

[23] G. Cook. The Singular Mind of Terry Tao: A prodigy grows up to become one of the greatest mathematicians in the world. *New York Times Magazine*, `http://www.nytimes.com/2015/07/26/magazine/the-singular-mind-of-terry-tao.html?_r=1`, 24 July 2015.

[24] A. L. Da Cunha, J. Zhou, and M. N. Do. The nonsubsampled contourlet transform: theory, design, and applications. *IEEE Transactions on Image Processing*, 15(10):3089–3101, 2006.

[25] I. Daubechies. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, Philadelphia, 1992.

[26] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57(11):1413–1457, 2004.

[27] V. Davidoiu, B. Sixou, M. Langer, and F. Peyrin. Non-linear iterative phase retrieval based on Frechet derivative and projection operators. In *9th IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 106–109. IEEE, 2012.

[28] V. Davidoiu, B. Sixou, M. Langer, and F. Peyrin. Nonlinear phase retrieval using projection operator and iterative wavelet thresholding. *IEEE Signal Processing Letters*, 19(9):579–582, 2012.

[29] V. Davidoiu, B. Sixou, M. Langer, and F. Peyrin. Nonlinear approaches for the single-distance phase retrieval problem involving regularizations with sparsity constraints. *Applied Optics*, 52(17):3977–3986, 2013.

[30] F. Deutsch. The method of alternating orthogonal projections. In *Approximation Theory, Spline Functions and Applications*, pages 105–121. Springer, 1992.

[31] F. Deutsch. *Best Approximation in Inner Product Spaces*. Springer Science & Business Media, 2012.

[32] R. A. DeVore, G. C. Kyriazis, and P. Wang. Multiscale characterizations of Besov spaces on bounded domains. *Journal of Approximation Theory*, 93:273–292, 1998.

[33] M. N. Do and M. Vetterli. The contourlet transform: an efficient directional multiresolution image representation. *IEEE Transactions on Image Processing*, 14(12):2091–2106, 2005.

[34] J. Douglas and H. H. Rachford. On the numerical solution of heat conduction problems in two and three space variables. *Transactions of the American Mathematical Society*, pages 421–439, 1956.

[35] M. Elad. Why simple shrinkage is still relevant for redundant representations? *IEEE Transactions on Information Theory*, 52(12):5559–5569, 2006.

[36] V. Elser. Phase retrieval by iterated projections. *Journal of the Optical Society of America A*, 20(1):40–55, 2003.

[37] J. R. Fienup. Reconstruction of an object from the modulus of its Fourier transform. *Optics Letters*, 3(1):27–29, 1978.

[38] J. R. Fienup. Phase retrieval algorithms: A comparison. *Applied Optics*, 21(15):2758–2769, 1982.

[39] R W. Gerchberg and W. O. Saxton. A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik*, 35:237–246, 1972.

[40] K. Giewekemeyer, S.P. Krüger, S. Kalbfleisch, M. Bartels, C. Beta, and T. Salditt. X-ray propagation microscopy of biological cells using waveguides as a quasipoint source. *Physical Review A*, 83(2):023804, 2011.

[41] J. W. Goodman. *Statistical Optics*. New York, Wiley-Interscience, 1985.

[42] J. W. Goodman. *Introduction to Fourier Optics*. Roberts and Company Publishers, Englewood, Colorado, 3rd edition, 2005.

[43] A. Grossmann and J. Morlet. Decomposition of Hardy functions into square integrable wavelets of constant shape. *SIAM Journal on Mathematical Analysis*, 15(4):723–736, 1984.

[44] L. G. Gubin, B. T. Polyak, and E. V. Raik. The method of projections for finding the common point of convex sets. *USSR Computational Mathematics and Mathematical Physics*, 7(6):1–24, 1967.

[45] K. Guo, G. Kutyniok, and D. Labate. Sparse multidimensional representations using anisotropic dilation and shear operators. In *Wavelets and Splines*, pages 189–201, Athens, GA, 2006.

[46] T. E. Gureyev and K. A. Nugent. Phase retrieval with the transport-of-intensity equation. II. Orthogonal series solution for nonuniform illumination. *Journal of the Optical Society of America A*, 13(8):1670–1682, 1996.

[47] T. E. Gureyev, A. Roberts, and K. A. Nugent. Phase retrieval with the transport-of-intensity equation: matrix solution with use of Zernike polynomials. *Journal of the Optical Society of America A*, 12(9):1932–1941, 1995.

[48] S. Häuser and G. Steidl. Fast finite shearlet transform: A tutorial. *arXiv preprint. arXiv: 1202.1773*, 2014.

[49] D. Heinen and G. Plonka. Wavelet shrinkage on paths for denoising of scattered data. *Results in Mathematics*, 62(3-4):337–354, 2012.

[50] R. Hesse. *Fixed Point Algorithms for Nonconvex Feasibility with Applications*. PhD thesis, Georg-August-Universität Göttingen, 2014.

[51] R. Hesse and D. R. Luke. Nonconvex notions of regularity and convergence of fundamental algorithms for feasibility problems. *SIAM Journal on Optimization*, 23(4):2397–2419, 2013.

[52] R. Hesse, D. R. Luke, and P. Neumann. Alternating projections and Douglas–Rachford for sparse affine feasibility. *IEEE Transactions on Signal Processing*, 62(18):4868–4881, 2014.

[53] H. Heuser. *Gewöhnliche Differentialgleichungen*. B. G. Teubner, Stuttgart, 1989.

[54] T. Hohage and F. Werner. Iteratively regularized Newton-type methods for general data misfit functionals and applications to Poisson data. *Numerische Mathematik*, 123(4):745–779, 2013.

[55] R. Houska. The nonexistence of shearlet scaling functions. *Applied and Computational Harmonic Analysis*, 32(1):28–44, 2012.

[56] P. Kittipoom, G. Kutyniok, and W.-Q. Lim. Construction of Compactly Supported Shearlet Frames. *Constructive Approximation*, 35(1):21–72, 2012.

[57] M. Krenkel, M. Bartels, and T. Salditt. Transport of intensity phase reconstruction to solve the twin image problem in holographic x-ray imaging. *Optics Express*, 21(2):2220–2235, 2013.

[58] R. Kress. *Linear Integral Equations*. Springer, New York, Heidelberg, Dordrecht, London, 3rd edition, 2014.

[59] G. Kutyniok and W.-Q Lim. Dualizable shearlet frames and sparse approximation. *Constructive Approximation*, doi:10.1007/s00365-016-9330-x, 2016.

[60] G. Kutyniok, W.-Q Lim, and R. Reisenhofer. ShearLab 3D: Faithful digital shearlet transforms based on compactly supported shearlets. *ACM Transactions on Mathematical Software*, 42(1), 2015.

[61] D. Labate, W.-Q. Lim, G. Kutyniok, and G. Weiss. Sparse multidimensional representation using shearlets. In *Wavelets XI, SPIE Proc. 5914*, pages 254–262, Bellingham, WA, 2005.

[62] M. Langer, P. Cloetens, and F. Peyrin. Fourier-wavelet regularization of phase retrieval in x-ray inline phase tomography. *Journal of the Optical Society of America A*, 26(8):1876–1881, 2009.

[63] A. Levi and H. Stark. Image restoration by the method of generalized projections with application to restoration from magnitude. *Journal of the Optical Society of America A*, 1(9):932–943, 1984.

[64] A. S. Lewis, D. R. Luke, and J. Malick. Local linear convergence for alternating and averaged nonconvex projections. *Foundations of Computational Mathematics*, 9(4):485–513, 2009.

[65] A. S. Lewis and J. Malick. Alternating projections on manifolds. *Mathematics of Operations Research*, 33(1):216–234, 2008.

[66] M. Liebling. *On Fresnelets, Interference Fringes, and Digital Holography*. PhD thesis, École Polytechnique Fédérale de Lausanne, 2004.

[67] M. Liebling, T. Blu, and M. Unser. Fresnelets: New Multiresolution Wavelet Bases for Digital Holography. *IEEE Transactions on Image Processing*, 12(1):29–43, 2003.

[68] M. Liebling, T. Blu, and M. Unser. Complex-wave retrieval from a single off-axis hologram. *Journal of the Optical Society of America A*, 21(3):367–377, 2004.

[69] W.-Q Lim. The discrete shearlet transform: A new directional transform and compactly supported shearlet frames. *IEEE Transactions on Image Processing*, 19(5):1166–1180, 2010.

[70] W.-Q Lim. Nonseparable shearlet transform. *IEEE Transactions on Image Processing*, 22(5):2056–2065, 2013.

[71] P.-L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979.

[72] S. Loock and G. Plonka. Phase retrieval for Fresnel measurements using a shearlet sparsity constraint. *Inverse Problems*, 30(5):055005, 2014.

[73] D. A. Lorenz. *Wavelet Shrinkage in Signal and Image Processing: An Investigation of Relations and Equivalences*. PhD thesis, Bremen University, 2004.

[74] D. R. Luke. *Analysis of Optical Wavefront Reconstruction and Deconvolution in Adaptive Optics*. PhD thesis, University of Washington, 2001.

[75] D. R. Luke. Relaxed averaged alternating reflections for diffraction imaging. *Inverse Problems*, 21(1):37–50, 2005.

[76] D. R. Luke. Finding best approximation pairs relative to a convex and prox-regular set in a Hilbert space. *SIAM Journal on Optimization*, 19(2):714–739, 2008.

[77] D. R. Luke, J. V. Burke, and R. G. Lyon. Optical wavefront reconstruction: Theory and numerical methods. *SIAM Review*, 44(2):169–224, 2002.

[78] J. Ma and G. Plonka. The curvelet transform: A review of recent applications. *IEEE Signal Processing Magazine*, 27(2):118–133, 2010.

[79] S. G. Mallat. Multiresolution approximations and wavelet orthonormal bases of $L^2(\mathbb{R})$. *Transactions of the American Mathematical Society*, 315(1):69–87, 1989.

[80] S. G. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, San Diego, 1999.

[81] S. Maretzke. Regularized Newton methods for simultaneous Radon inversion and phase retrieval in phase contrast tomography. *arXiv preprint arXiv:1502.05073*, 2015.

[82] S. Maretzke, M. Bartels, M. Krenkel, T. Salditt, and T. Hohage. Regularized Newton methods for x-ray phase contrast and general imaging problems. *Optics Express*, 24(6):6490–6506, 2016.

[83] F. G. Meyer and R. R. Coifman. Opérateurs de Calderón-Zygmund. *Applied and Computational Harmonic Analysis*, 4:147–187, 1996.

[84] Y. Meyer and D. H. Salinger. *Wavelets and Operators*, volume 1. Cambridge University Press, Cambridge, England, 1992.

[85] C. A. Micchelli, L. Shen, and Y. Xu. Proximity algorithms for image models: Denoising. *Inverse Problems*, 27(4):045009, 2011.

[86] G. J. Minty. Monotone (nonlinear) operators in Hilbert space. *Duke Mathematical Journal*, 29(3):341–346, 1962.

[87] M. L. Moravec, J. K. Romberg, and R. G. Baraniuk. Compressive phase retrieval. In *Optical Engineering + Applications*, pages 670120–1–670120–11, 2007.

[88] J.-J. Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société mathématique de France*, 93:273–299, 1965.

[89] S. Mukherjee and C. S. Seelamantula. Fienup algorithm with sparsity constraints: Application to frequency-domain optical-coherence tomography. *IEEE Transactions on Signal Processing*, 62(18):4659–4672, 2014.

[90] H. Ohlsson, A. Y. Yang, R. Dong, and S. S. Sastry. Compressive phase retrieval from squared output measurements via semidefinite programming. *arXiv preprint. arXiv:1111.6323*, 2011.

[91] D. Paganin. *Coherent X-Ray Optics*. Oxford University Press, Oxford, England, 2006.

[92] A. Pein. Sparsity Constraint for Phase Retrieval in X-Ray Propagation Imaging. Master's thesis, Georg-August-Universität Göttingen, 2015.

[93] A. Pein, S. Loock, G. Plonka, and T. Salditt. Using sparsity information for iterative phase retrieval in x-ray propagation imaging. *Optics Express*, 24(8):8332–8343, 2016.

[94] M. A. Pinsky. *Introduction to Fourier Analysis and Wavelets*. Brooks/Cole, Pacific Grove, CA, 2002.

[95] G. Plonka. The easy path wavelet transform: A new adaptive wavelet transform for sparse representation of two-dimensional data. *Multiscale Modeling & Simulation*, 7(3):1474–1496, 2009.

[96] G. Plonka, A. Iske, and S. Tenorth. Optimal representation of piecewise Hölder smooth bivariate functions by the easy path wavelet transform. *Journal of Approximation Theory*, 176:42–67, 2013.

[97] G. Plonka, S. Tenorth, and A. Iske. Optimally sparse image representation by the easy path wavelet transform. *International Journal of Wavelets, Multiresolution and Information Processing*, 10(01):1250007, 2012.

[98] R. T. Rockafellar. Characterization of the subdifferentials of convex functions. *Pacific Journal of Mathematics*, 17(3):497–510, 1966.

[99] B. Seifert, H. Stolz, M. Donatelli, D. Langemann, and M. Tasche. Multilevel Gauss–Newton methods for phase retrieval problems. *Journal of Physics A: Mathematical and General*, 39(16):4191, 2006.

[100] A. Sommerfeld. Die Greensche Funktion der Schwingungsgleichung. *Jahresberichte der Deutschen Mathematiker Vereinigung*, 21:309–353, 1912.

[101] A. Souvorov, T. Ishikawa, and A. Kuyumchyan. Multiresolution phase retrieval in the Fresnel region by use of wavelet transform. *Journal of the Optical Society of America A*, 23(2):279–287, 2006.

[102] G. Strang. *Calculus*. Wellesley-Cambridge Press, Wellesley, MA, 1991.

[103] W. A. Strauss. *Partial Differential Equations – An Introduction*. John Wiley and Sons, Ltd., New York, 2nd edition, 1992.

[104] M. R. Teague. Deterministic phase retrieval: A Green's function solution. *Journal of the Optical Society of America A*, 73(11):1434–1441, 1983.

[105] D. G. Voelz and M. C. Roggemann. Digital simulation of scalar optical diffraction: Revisiting chirp function sampling criteria and consequences. *Applied Optics*, 48(32):6132–6142, 2009.

[106] J. von Neumann. *Functional Operators: Vol. II*. Princeton University Press, Princeton, NJ, 1950.

[107] W. Walter. *Gewöhnliche Differentialgleichungen*. Springer, Berlin, 6th edition, 1996.

[108] J. Weng, J. Zhong, and C. Hu. Phase reconstruction of digital holography with the peak of the two-dimensional Gabor wavelet transform. *Applied Optics*, 48(18):3308–3316, 2009.

[109] F. Werner. On convergence rates for iteratively regularized Newton-type methods under a Lipschitz-type nonlinearity condition. *Journal of Inverse and Ill-posed Problems*, 23(1):75–84, 2015.

[110] B. Xue and S. Zheng. Phase retrieval using the transport of intensity equation solved by the FMG-CG method. *Optik – International Journal for Light and Electron Optics*, 122(23):2101–2106, 2011.

[111] R. M. Young. *An Introduction to Non-Harmonic Fourier Series*. Academic Press, New York, 1980.

Curriculum Vitae

Dipl.-Math. Stefan Loock

Adress

Postal	Institut für Numerische und Angewandte Mathematik Georg-August Universität Göttingen Lotzestraße 16-18 37083 Göttingen, Germany
Phone	+49 (0)551 – 39 20075
E-Mail	s.loock@math.uni-goettingen.de

Personal Details

Date of birth	October 30, 1986
Place of birth	Stade
Citizenship	German

Academic Education

Since 09/2012	Doctoral student of mathematics Thesis: *Phase Retrieval with Sparsity Constraints* Georg-August-Universität Göttingen Advisor: Prof. Dr. Gerlind Plonka-Hoch
10/2007 – 09/2012	Diploma student of mathematics and physics (minor) Thesis: *Die verallgemeinerte Totalvariation zur Rauschreduktion in Bildern*, Universität Bremen Advisor: Dr. Stefan Schiffler

Internship

09/2010 – 10/2010	Deutsches Zentrum für Luft- und Raumfahrt Institut für Raumfahrtsysteme, Bremen

Alternative Civilian Service

07/2006 – 03/2007	DJH Cuxhaven-Duhnen

Education

1999 – 2006	Secondary school “Gymnasium Warstade”

Research Experience

since 07/2014	Member of "Project C11 – Fresnel wavelets for coherent diffractive imaging" – DFG Collaborative Research Center 755: Nanoscale Photonic Imaging
09/2012 – 06/2014	Associate member of "Project C11 – Fresnel wavelets for coherent diffractive imaging" – DFG Collaborative Research Center 755: Nanoscale Photonic Imaging
09/2012 – 06/2014	Scientific Assistant at Institute for Numerical and Applied Mathematics, Georg-August-Universität Göttingen
2009 – 2012	Student Assistent, Universität Bremen

Talks & Posters

03/2016	Joint Annual Meeting of GAMM and DMV, Braunschweig
02/2016	International Symposium 2016 on Biological Dynamics: From Microscopic to Mesoscopic Scales, Grimma
03/2015	Approximation Methods and Functions Spaces, Hasenwinkel
01/2015	CRC 755 Winter School 2015, Teistungen
09/2014	Mathematical Signal Processing and Phase Retrieval, Göttingen
01/2014	24. Rhein-Ruhr-Workshop, Bestwig
10/2013	Advances in Mathematical Image Processing, Annweiler
09/2013	Autumn School of CRC 755 and CRC 937, Wildbad Kreuth
04/2013	1st International Symposium on Nanoscale Photonic Imaging, Göttingen

Teaching Experience

2015	Teaching Assistant: Applied Mathematics for Teacher Trainees
2014/2015	Teaching Assistant: Variational Analysis III
2014	Teaching Assistant: Numerical Mathematics II
2013/2014	Teaching Assistant: Numerical Mathematics I
2013	Teaching Assistant: Numerical Mathematics II
2012/2013	Teaching Assistant: Mathematical Methods of Image Reconstruction
2012/2013	Teaching Assistant: Compressed Sensing
2011/2012	Tutor: Functional Analysis

Publications

1. S. Loock and G. Plonka. Phase retrieval for Fresnel measurements using a shearlet sparsity constraint. *Inverse Problems*, **30**(5):055005, 2014.
2. A. Pein, S. Loock, G. Plonka, and T. Salditt. Using sparsity information for iterative phase retrieval in x-ray propagation imaging. *Optics Express*, **24**(8):8332–8343, 2016.
3. S. Loock and G. Plonka. Iterative phase retrieval using sparsity constraints. *Proceedings in Applied Mathematics and Mechanics*, accepted, 2016.