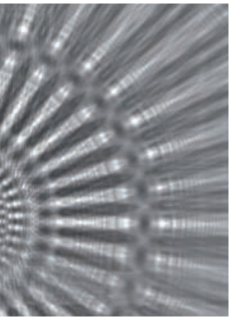
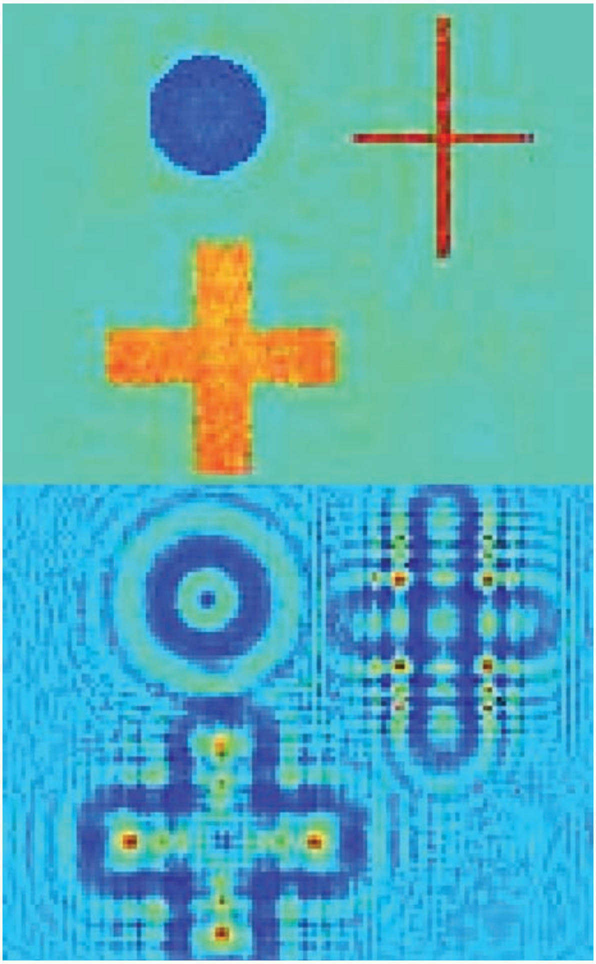
Göttingen Series in
X-ray Physics



Carolin Homann

Phase retrieval problems in x-ray physics

From modeling to efficient algorithms



Universitätsverlag Göttingen

Carolin Homann

Phase retrieval problems in x-ray physics:
From modeling to efficient algorithms

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Published in 2015 by Universitätsverlag Göttingen
as volume 16 in the series „Göttingen series in X-ray physics“

Carolin Homann

Phase retrieval problems in x-ray physics:

From modeling to efficient algorithms

Göttingen series in X-ray physics
Volume 16

Universitätsverlag Göttingen
2015

Bibliographic information published by the Deutsche Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.dnb.de.

Address of the Author
Carolin Homann
e-mail: c.homann@math.uni-goettingen.de

Dissertation
for the award of the degree
"Doctor rerum naturalium"
of the Georg-August-Universität Göttingen
within the doctoral program Mathematical Sciences
of the Georg-August University School of Science (GAUSS)

Thesis Committee:
Prof. Dr. Thorsten Hohage
Prof. Dr. Russell Luke

This work is protected by German Intellectual Property Right Law.
It is also available as an Open Access version through the publisher's homepage and the Göttingen University Catalogue (GUK) at the Göttingen State and University Library (http://www.sub.uni-goettingen.de).
The conditions of the license terms of the online version apply.

Layout: Carolin Homann
Cover: Jutta Pabst
Cover image: Carolin Homann

© 2015 Universitätsverlag Göttingen
http://univerlag.uni-goettingen.de
ISBN: 978-3-86395-210-5
ISSN: 2191-9860

Preface of the series editors

The Göttingen series in x-ray physics is intended as a collection of research monographs in x-ray science, carried out at the Institute for X-ray Physics at the Georg-August-Universität in Göttingen, and in the framework of its related research networks and collaborations.

It covers topics ranging from x-ray microscopy, nano-focusing, wave propagation, image reconstruction, tomography, short x-ray pulses to applications of nanoscale x-ray imaging and biomolecular structure analysis.

In most but not all cases, the contributions are based on Ph.D. dissertations. The individual monographs should be enhanced by putting them in the context of related work, often based on a common long term research strategy, and funded by the same research networks. We hope that the series will also help to enhance the visibility of the research carried out here and help others in the field to advance similar projects.

Prof. Dr. Sarah Köster
Prof. Dr. Tim Salditt
Editors
Göttingen June 2014

Contents

Introduction 1

1 Phase retrieval problems in x-ray physics and inverse medium scattering 5

1.1 Phase retrieval in coherent x-ray imaging 5

1.1.1 Fresnel and projection approximations 5

1.1.2 Fresnel propagator for tempered distributions. 11

1.1.3 Fresnel scaling theorem 15

1.1.4 Phase retrieval in coherent x-ray imaging: An inverse problem in Banach spaces 17

1.2 Phase retrieval in inverse medium scattering 20

2 Validity of the empty beam correction in near field imaging 25

2.1 Motivation 25

2.2 Error estimate 29

2.3 Numerical results 35

3 Tikhonov-type regularization on Banach spaces 41

3.1 Tikhonov regularization and IRNM 41

3.2 Duality theory for linear Tikhonov-type regularization 44

3.2.1 The class of proper, convex and lower semicontinuous functions . 45

3.2.2 Subdifferential and conjugate function 47

3.2.3 Extremal relations and saddle point problem 50

3.3 Duality mappings and Bregman distance 52

4 Generalization of CP to Banach spaces 61

4.1 Chambolle-Pock's first-order primal-dual algorithm 61

4.2 Generalization and convergence results 65

4.3 Generalized resolvents 75

4.3.1 Use of resolvents 75

4.3.2 Duality mapping of Sobolev spaces 81

5 Numerical examples: Solving phase retrieval problems by CP-BS 83

5.1 Preliminaries 83

5.2 Convolution problem 84

5.3 Phase retrieval problem in wavefront reconstructions 89

5.3.1 Numerical approximation of the Fresnel approximation 89

5.3.2 CP-BS as inner solver in the IRNM 90

5.4 Phase retrieval in inverse medium scattering 95

Summary 97

A Appendix: Basics from Fourier theory 99

A.1 Fourier analysis and tempered distributions 99

A.2 Discrete Fourier transform 101

References 107

Acknowledgements 113

Introduction

Within this work, we consider two-dimensional phase retrieval problems that occur in coherent x-ray optics. The fact that x-ray microscopy allows nanoscale resolution makes this imaging technique interesting for visualizing cellular and subcellular structures of biological samples. Often these samples consist of soft tissues with low x-ray absorption which means that the imaginary part of the refractive index corresponding to the object almost vanishes. However, the interaction of x-rays with matter consists not only of absorption: An x-ray wave field passing through a specimen is also modified by phase shifts. So, the detected wave field contains information which allows to reconstruct an image of such a specimen based on the real part of the refractive index. This reconstruction task is complicated by the fact that only the amplitude of the wave field can be measured while the phase information gets lost.

Using the well-established Fraunhofer or Fresnel approximation (see Chapter 1) we obtain a nonlinear operator T that describes the dependence between the unknown sample information ϕ and the data y (the recorded wave field amplitude). Thus, the phase retrieval problems under consideration can be formulated as an operator equation

$$T\phi = y.$$

It describes a nonlinear inverse ill-posed problem. In practice, instead of the exact data y one normally observes only noisy data y^{δ}. These intensity data have a special structure: They are Poisson distributed with the exact data y as mean (cf. [82]). Therefore we call them Poisson data. As analyzed in [47,82], an *iteratively regularized Newton-type method* (IRNM) with Kullback-Leibler-type data fidelity functional is particular suitable in order to solve such kind of problems. This is because the method not only takes into account the special data and noise structure but also exhibits fast convergence. It is an adaption of the well-established iteratively regularized Gauß-Newton method and includes the minimization of a convex Tikhonov-type functional in each iteration step (cf. Figure 1). Thus, the application of a convex optimization method becomes necessary. Due to the Kullback-Leibler data fidelity term, so-called proximal-type algorithms that rely on the so-called resolvent of the data fidelity functional are efficient methods for this purpose. These methods are applicable to a very general class of problems. This gives us great flexibility in the choice of the penalty term which models conditions on the solution. A commonly used representative of this class of algorithms is the *first-order primal-dual algorithm (CP) of Chambolle and Pock* (cf. Section 4.1 or [21]). Especially the fact that, besides using the resolvents of data fidelity and penalty term, the method only involves the evaluation of the derivative of the forward operator T and its adjoint, makes it attractive for our minimization task.

In addition to adapting IRNM in terms of the data to the given phase retrieval problems, we also want to incorporate a priori knowledge about the solution's structure into the reconstruction process. For example, in order to model "blocky" structured solutions, we would like to consider ϕ to be an element of a Sobolev space that is not a Hilbert but a Banach space and choose a penalty term that is based on the corresponding Sobolev space norm. However, the method CP is defined for a Hilbert space setting. Of course,

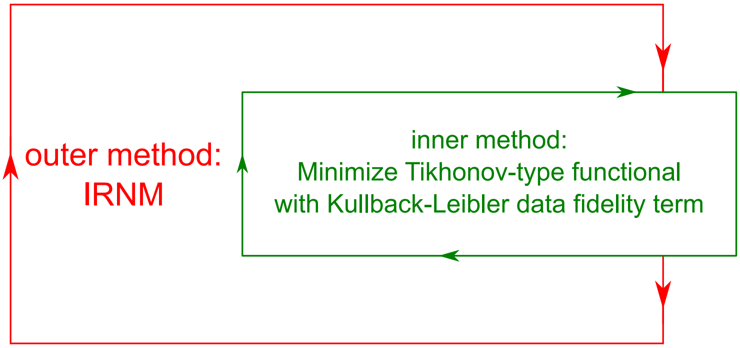

outer method:
IRNM
inner method:
Minimize Tikhonov-type functional
with Kullback-Leibler data fidelity term

Figure 1: Solving phase retrieval problems in x-ray optics, given by an operator equation $T\phi = y$, by the IRNM with Kullback-Leibler data fidelity functional.

the existence of a norm-defining inner product which distinguishes a Hilbert space from a Banach space gives the former space nicer properties, such as the polarization identity, than the latter. On the other hand, the restriction to a Hilbert space setting makes a model less adaptable to certain requirements and situations, as the considered phase retrieval case illustrates. Therefore, we propose a generalization (CP-BS) of CP to a Banach space setting. Under certain conditions, we prove convergence results including also rates of convergence.

This generalization makes the algorithm not only efficiently applicable to the minimization problems we derive from the phase retrieval problems but also allows the solution of a wider class of relevant problems as we will see in Chapter 4. In fact, a problem of the form which is treated by CP-BS arises in many other applications. In order to test the performance of CP-BS , we consider problems that can also efficiently be solved by the original algorithm CP in Hilbert spaces, however a Banach space setting would be more appropriate. For example, it is common to model sparsity constraints with help of L^1-penalization in the Hilbert space $X = L^2$ or $X = L^2(\Omega)$ for some $\Omega \subset \mathbb{R}^2$ (cf. soft thresholding algorithms). Here it would be much more natural to pick $X = L^r$ or $X = L^r(\Omega)$ with $r \approx 1$. This issue will be addressed in numerical examples where we compare CP with a Hilbert space setting to CP-BS with a more problem adapted Banach space setting.

Solving phase retrieval problems in x-ray imaging (under the assumption that the wave field is generated by a perfect point source) is a topic of certain interest in the literature: The commonly used Gerchberg-Saxton-Fienup-type algorithms (cf. [30, 32, 56]) usually disregard the special data structure. On the other hand, they are quite flexible with respect to modeling constraints on the solution ϕ. The fact that many of these methods can also be interpreted as extensions of convex optimization algorithms (cf. [9, 53]) creates a connection to the algorithm CP considered here. By using the IRNM with special data fidelity functional, our approach relies on a well-established regularized technique for solving nonlinear inverse problems. Also other authors suggest (variants of) typical regularization methods from inverse problems for the solution of phase retrieval problems:

See e.g. [29, 40] for nonlinear Tikhonov-type regularization, [27] for a Landweber-type method, and, as mentioned above, [47, 82] for iteratively regularized Newton-type methods. Moreover, [70] which provides a nice overview of regularization methods in Banach spaces lists a phase retrieval problem as application (but again without considering the special data structure). In particular, here it is motivated that for many applications, such as this phase retrieval problem, a Banach space setting is more appropriate than a Hilbert space setting. In general, the application of linear Tikhonov-type regularization or IRNMs to linear respectively nonlinear inverse problems $Tx = y^\delta$ in Banach spaces includes the solution of a convex minimization problem in Banach spaces. Now, in particular if the corresponding Tikhonov-type functional which has to be minimized is nonsmooth the proposed generalization CP-BS is an attractive method for this purpose. We also refer to [14, 55, 62, 78, 80] for other generalizations and extensions of CP.

This work tackles also another aspect of phase retrieval problems in coherent x-ray imaging: The commonly used *empty beam correction* which is also referred to as *product approximation in the detector plane*. It is a preprocessing step with the aim of making the idealized assumption that the coherent x-ray field is generated by a point source applicable. In particular in near field imaging, deviations from this idealized model lead to intensity measurements that are strongly influenced by the empty beam field. The idea behind the empty beam correction is to “factor out” this dependence on the empty beam field. To be more precise, this data correction step consists in dividing the measured data y^δ by the intensities of the empty beam field which one obtains by a further measurement. For a perfect point source this approximation is exact as shown by Giewekemeyer et al., [33, 34]. We study the validity of this simple but also rather crude technique for extended source sizes: We present a quantitative error estimate that allows to determine settings where the empty beam correction is justified, but also shows its limits. In the case that the empty beam correction is not sufficiently accurate, one may use reconstruction methods for the empty beam. See e.g. [36, 64] for recently proposed methods.

This thesis is organized as follows. We introduce and formulate the phase retrieval problems in coherent x-ray imaging in Chapter 1. In order to give a further example for inverse ill-posed problems in Banach spaces, we also consider another class of phase retrieval problems. They occur in inverse medium scattering (see Section 1.2). In Chapter 2 we study the validity of the empty correction. The proposed error estimate is furthermore illustrated and verified by numerical examples. A general motivation of Tikhonov-type regularization and IRNMs is given in Chapter 3 where we additionally specify the convex minimization problems that typically occur in this context (denoted as (**P**)). With regard to these optimization problems (**P**), we also give definitions and results of convex analysis as well as optimization theory which will be required for the generalization of the Chambolle and Pock algorithm CP to Banach spaces. Then we introduce the algorithm CP in Chapter 4 and present a generalization CP-BS to Banach spaces. In particular, we state three different special versions of CP-BS for which we prove convergence results. We close this chapter with considering the generalized resolvents that are included in CP-BS. In Chapter 5 we test the performance of CP-BS by numerical examples. In particular, we use the algorithm as inner solver of the IRNM in order to solve the phase retrieval problems defined in Chapter 1. This thesis closes with a summary of our main results.

1 Phase retrieval problems in x-ray physics and inverse medium scattering

In this section, we introduce two different kinds of phase retrieval problems. The first occurs in imaging by coherent x-ray diffraction and is the underlying application of this work. The second one is a time-harmonic inverse scattering problem for either electromagnetic or acoustic waves. Both problems have in common that by measurements of electric or acoustic wave fields which pass through an object or an inhomogeneous medium of interest we want to reconstruct information on the refractive index of the object. The term 'phase retrieval' refers to the fact that only the amplitude and not the phase of the field can be measured. In both cases we obtain nonlinear inverse problems where a Banach space setting is more appropriate than a Hilbert space setting.

1.1 Phase retrieval in coherent x-ray imaging

The basic setup in this imaging technique is as follows: A sample of interest is illuminated by a coherent x-ray (quasi) point source. For the analysis, the optical axis in direction of the beam is identified with the x_3-axis, where the sample of thickness τ lies in between the *object plane* $\mathbb{P}_0 := \left\{x \in \mathbb{R}^3 \mid x_3 = 0\right\}$ and the plane $\mathbb{P}_{-\tau} := \left\{x := (\mathbf{x}, x_3) \in \mathbb{R}^3 \mid x_3 = -\tau\right\}$ both orthogonal to the beam (see Figure 1.1). The detectors, which are located within a plane $\mathbb{P}_\Gamma$ parallel to object plane $\mathbb{P}_0$ at a distance $\Gamma > 0$, measure only the intensity $|u(\cdot,\Gamma)|^2$ of the electric field $u : \mathbb{R}^3 \to \mathbb{C}$, while the phase information gets lost. From these measurements we want to retrieve information on the refractive index of the sample. This unknown information will be described by a complex function $\phi : D(\phi) \subseteq \mathbb{P}_0 \to \mathbb{C}$. For the aim of reconstructing ϕ we define the Fresnel propagator that approximates the electric field in the *detector plane* $\mathbb{P}_\Gamma := \left\{x = (\mathbf{x}, x_3) \in \mathbb{R}^3 \mid x_3 = \Gamma\right\}$ for given boundary values $u_0 = u(\cdot, 0)$. Based on this mapping we derive a nonlinear operator equation of the form

$$T\phi = |u(\cdot,\Gamma)|^2 ,$$

which describes our problem as a nonlinear inverse ill-posed problem. For a detailed theory of x-ray diffraction imaging we refer the reader to [7,33,60]. The necessary basics of Fourier analysis and distribution theory are given in the appendix.

1.1.1 Fresnel and projection approximations

Let us consider the setting shown in Figure 1.1. As we assume the upper half space $\mathbb{H}^+ := \{x \in \mathbb{R}^3 \mid x_3 > 0\}$ to be vacuum, within this volume the cartesian components of the electric field $u : \mathbb{R}^3 \to \mathbb{C}$ satisfy the *Helmholtz equation*

$$\Delta u + \kappa^2 u = 0 \qquad \text{in } \mathbb{H}^+ := \{x \in \mathbb{R}^3 : x_3 > 0\}, \tag{1.1a}$$

where Δ is the Laplacian defined by $\Delta u = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} + \frac{\partial^2 u}{\partial x_3^2}$ and $\kappa = \frac{2\pi}{\lambda} > 0$ denotes the wavenumber, anti proportional to the wavelength λ. Next, we will derive an operator

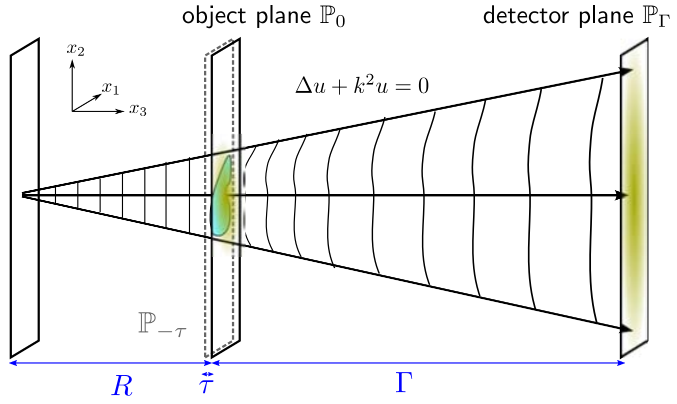


Figure 1.1: Setting of the phase retrieval problem in x-ray diffraction imaging. A spherical-like wave-field generated by a (quasi) point source in the lower half space $\mathbb{H}^- := \left\{x \in \mathbb{R}^3 \mid x_3 < 0\right\}$ traverses a sample of interest at the object plane $\mathbb{P}_0 := \left\{x \in \mathbb{R}^3 \mid x_3 = 0\right\}$. The resulting distribution is indicated by the wavy lines. At a distance $\Gamma > 0$ from the sample, detectors measure the intensity of the diffracted field.

which propagates the field $u_0 := u(\cdot, 0)$ that exits the sample to the detector plane $\mathbb{P}_\Gamma$. This means, in the plane $\mathbb{P}_\Gamma$ we will give an explicit form for the solution of the Helmholtz equation (1.1a) that satisfies the boundary condition

$$u(\mathbf{x}', 0) = u_0(\mathbf{x}') \qquad \mathbf{x}' \in \mathbb{R}^2 . \tag{1.1b}$$

In order to ensure the uniqueness of the solution, we characterize the field as an outgoing wave by assuming it to obey a radiation condition in $\mathbb{H}^+$ as it is defined in [22].

In an arbitrary plane $\mathbb{P}_\Gamma \subseteq \mathbb{H}^+$ we rewrite the (outgoing) solution u of (1.1a) - (1.1b) as $u\,(\mathbf{x}', \Gamma) = \mathcal{F}^{-1}\mathcal{F}\; u|_{P_\Gamma}(\mathbf{x}')$. Then, using the Fourier derivative theorem (cf. Equation (A.1))

$$\left(\frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2}\right)\mathcal{F}^{-1}w\,(\xi') = -\mathcal{F}^{-1}\left(|\xi'|^2 w(\xi')\right) \qquad \xi' \in \mathbb{R}^2,\ w \in L^2(\mathbb{R}^2)$$

and the bijectivity of $\mathcal{F}^{-1} : L^2(\mathbb{R}^2) \to L^2(\mathbb{R}^2)$, we deduce that the Fourier transform of u fulfills the following differential equation:

$$0 = \left(-|\xi'|^2 + \frac{\partial^2}{\partial^2 x_3} + \kappa^2\right)(\mathcal{F}u)\,(\xi', \Gamma) \qquad \forall\,(\xi', \Gamma) \in \mathbb{P}_\Gamma .$$

Thus, $\mathcal{F}u$ is of the form

$$\mathcal{F}u\,(\xi', x_3) = \mathrm{e}^{ix_3\sqrt{\kappa^2-|\xi'|^2}}a(\xi') + \mathrm{e}^{-ix_3\sqrt{\kappa^2-|\xi'|^2}}b(\xi') \qquad (\xi', x_3) \in \mathbb{H}^+,$$

for some complex functions $a, b : \mathbb{H}^+ \to \mathbb{C}$ that are independent of x_3. As the second summand is, in contrast to our assumption, “downgoing” for $|\xi'| < |\kappa|$, we set $b(x) = 0$.

Then the boundary condition $\mathcal{F}u(\cdot,0)=\mathcal{F}u_0$ leads to $\mathcal{F}u\left(\xi',x_3\right)=\mathrm{e}^{ix_3\sqrt{\kappa^2-|\xi'|^2}}\mathcal{F}u_0$. Now it is easy to see that

$$u(\mathbf{x}',x_3)=\mathcal{F}^{-1}e^{ix_3\sqrt{\kappa^2-|\xi'|^2}}\mathcal{F}u_0\,(\mathbf{x}'),\qquad (\mathbf{x}',x_3)\in\mathbb{H}^+$$

is indeed the outgoing solution of (1.1a) - (1.1b). This formula gives an explicit form to the *free space propagation* in vacuum.

We further assume the field to be paraxial such that $|\xi'|\ll\kappa$ for all ξ' for which $|\mathcal{F}u_0(\xi')|$ is not negligible (cf. [60]). Then, we can apply the Taylor approximation

$$\sqrt{\kappa^2-|\xi'|^2}\approx\kappa-\frac{|\xi'|^2}{2k}$$

to the free space propagation which yields its well-known *Fresnel approximation* (also known as *paraxial approximation*):

$$u|_{\mathbb{P}_\Gamma}\approx e^{i\kappa\Gamma}\mathcal{F}^{-1}\left(\exp\left(-i\frac{\Gamma}{2\kappa}|\xi'|^2\right)(\mathcal{F}u_0)\right)=:\mathcal{D}_\Gamma u_0,\quad \Gamma>0. \tag{1.2}$$

Let us rewrite $\xi'\mapsto\exp\left(-i\frac{\Gamma}{2\kappa}|\xi'|^2\right)$ as a *chirp function* χ_f with parameter $f=-\frac{\Gamma}{\kappa}\in\mathbb{R}$:

$$\chi_f:\mathbb{R}^2\to\mathbb{R},\quad \chi_f(\mathbf{x}'):=\mathrm{e}^{\mathrm{i}\frac{f}{2}|\mathbf{x}'|^2}.$$

As we have $\left|\chi_f(\xi')\right|=1$ for any $\xi'\in\mathbb{R}^2$, this function is bounded and the operator $\mathcal{D}_\Gamma$ maps $L^2(\mathbb{R}^2)$ to $L^2(\mathbb{R}^2)$. The following lemma shows that $\mathcal{D}_\Gamma:\phi\mapsto e^{i\kappa\Gamma}\mathcal{F}^{-1}\left(\chi_{-\frac{\Gamma}{\kappa}}\,\mathcal{F}u_0\right)$ defines also a continuous mapping from $S\left(\mathbb{R}^2\right)$ into $S\left(\mathbb{R}^2\right)$.

Lemma 1.1.1. *For any parameter $f\in\mathbb{R}$ the chirp function $\chi_f\in C^\infty$ defines a multiplier on $S\left(\mathbb{R}^2\right)$, i.e. $\chi_f\,\phi=\chi_f\bullet\phi\in S\left(\mathbb{R}^2\right)$ for any $\phi\in S\left(\mathbb{R}^2\right)$. Moreover, the mapping*

$$S\left(\mathbb{R}^2\right)\to S\left(\mathbb{R}^2\right),\quad \phi\mapsto\chi_f\,\phi$$

in continuous.

Proof. First of all, note that for any parameter $f\in\mathbb{R}$ and any multi-index $\gamma\in\mathbb{N}_0^2$ there exist some constants $C_{\alpha,\gamma}>0$ and $m_\gamma>0$ such that

$$\left|D^\gamma\chi_f(\mathbf{x}')\right|\le C_{\alpha,\gamma}\left(1+|\mathbf{x}'|\right)^{m_\gamma}, \tag{1.3}$$

i.e. $D^\gamma\chi_f$ has *polynomial growth.* Let ϕ be in $S\left(\mathbb{R}^2\right)$. Then the definition of $S\left(\mathbb{R}^2\right)$ as well as the inequality $|\mathbf{x}'^\alpha|\le|\mathbf{x}'|^{\alpha_1+\alpha_2}$ yields

$$\begin{aligned}\sup_{\mathbf{x}'\in\mathbb{R}^2}\left|\mathbf{x}'^\alpha D^\gamma\chi_f(\mathbf{x}')\,D^{\beta-\gamma}\phi(\mathbf{x}')\right|&\le C_{\alpha,\gamma}\sup_{\mathbf{x}'\in\mathbb{R}^2}\left|\mathbf{x}'\right|^{\alpha_1+\alpha_2}\left(1+|\mathbf{x}'|\right)^{m_\gamma}\left|D^{\beta-\gamma}\phi(\mathbf{x}')\right|\\&\le C_{\alpha,\gamma}\sum_{j=0}^{m_\gamma}\frac{m_\gamma!}{j!(m_\gamma-j)!}\sup_{\mathbf{x}'\in\mathbb{R}^2}|\mathbf{x}'|^{j+\alpha_1+\alpha_2}\left|D^{\beta-\gamma}\phi(\mathbf{x}')\right|<\infty\end{aligned}$$

for all $\alpha, \beta, \gamma \in \mathbb{N}_0^2$. Here we used that for any even sum $\alpha_1 + \alpha_2 \in \mathbb{N}_0$

$$\sup_{\mathbf{x}' \in \mathbb{R}^2} |\mathbf{x}'|^{\alpha_1+\alpha_2} \left| D^{\beta-\gamma}\phi(\mathbf{x}') \right| = \sup_{\mathbf{x}' \in \mathbb{R}^2} \left| \left(x_1^2 + x_2^2\right)^{\frac{\alpha_1+\alpha_2}{2}} D^{\beta-\gamma}\phi(\mathbf{x}') \right| < \infty$$

holds true. This last fact also shows the equivalent characterization

$$S\left(\mathbb{R}^2\right) = \left\{ \phi \in C^\infty \,\middle|\, \sup_{\mathbf{x}' \in \mathbb{R}^2} |\mathbf{x}'|^{\alpha_1+\alpha_2} |D^\beta \phi| < \infty \text{ for all } \alpha, \beta \in \mathbb{N}_0^2 \right\}.$$

Now the general Leibniz rule

$$D^\beta \chi_f \, \phi = \sum_{\|\gamma\|_{l^1} \le \|\beta\|_{l^1}} \frac{\beta!}{\gamma!(\beta-\gamma)!} D^\gamma \chi_f \, D^{\beta-\gamma}\phi \qquad \text{where } \gamma! := \gamma_1!\gamma_2!$$

implies $\|\chi_f \, \phi\|_{\alpha,\beta} < \infty$ for all $\alpha, \beta \in \mathbb{N}_0^2$. This proves the first assertion. The second one follows analogously: Let $(\phi_n)_{n\in\mathbb{N}}$ be a sequence in $S\left(\mathbb{R}^2\right)$ that converges to some $\phi \in S\left(\mathbb{R}^2\right)$ as $n \to \infty$, i.e. $\|\phi_n - \phi\|_{\alpha,\beta} \to 0$ as $n \to \infty$ for all $\alpha, \beta \in \mathbb{N}_0^2$. Then we can use the same estimates as above to show that $\|\chi_f \, \phi_n - \chi_f \, \phi\|_{\alpha,\beta} \to 0$ as $n \to \infty$ for all $\alpha, \beta \in \mathbb{N}_0^2$ which completes the proof. □

Remark 1.1.2. Note that the last proof not only applies for chirp functions but for any function $\varphi \in C^\infty$ that satisfies Equation (1.3) for all $\gamma \in \mathbb{N}_0^2$. In fact, it is well-known that the class of continuous multipliers on $S\left(\mathbb{R}^2\right)$ consists of all functions $\varphi \in C^\infty$ such that $D^\gamma \varphi$ has polynomial growth for all $\gamma \in \mathbb{N}_0^2$ (see e.g [31, Definition 8.4.1]).

So, the last lemma as well as the continuity of the Fourier transforms $\mathcal{F} : S\left(\mathbb{R}^2\right) \to S\left(\mathbb{R}^2\right)$ and $\mathcal{F}^{-1}$ ensure that $\mathcal{D}_\Gamma : S\left(\mathbb{R}^2\right) \to S\left(\mathbb{R}^2\right)$ is continuous as well . Then by replacing χ_f, with $f = -\frac{\Gamma}{\kappa}$, by its distribution $T_{\chi_f} : \hat{u} \mapsto \int_{\mathbb{R}^2} \chi_f(\mathbf{x}') \, \hat{u}(\mathbf{x}') \, d\mathbf{x}'$ (cf. A.2) we can apply Fouriers convolution formula (A.3) to (1.2). For this purpose the Fourier transform of T_{χ_f}, $f \in \mathbb{R} \backslash \{0\}$ is given by the next lemma, while the well-established case $f = 0$, i.e. $\chi_f = \mathbf{1}$, is studied in Example A.1.2.

Lemma 1.1.3. *For any parameter $f \in \mathbb{R} \backslash \{0\}$ the Fourier transform of T_{χ_f} corresponds to the L^∞-function $\frac{i}{f}\chi_{-\frac{1}{f}}$. Thus, we write*

$$\mathcal{F}\chi_f = \frac{i}{f}\chi_{-\frac{1}{f}}.$$

Proof. Let us consider the derivative $D\,\mathcal{F}T_{\chi_f}(\phi) = T_{\chi_f}(\mathcal{F}D\,\phi)$, $\phi \in S\left(\mathbb{R}^2\right)$. By using (A.1) and substituting $\mathrm{i}\,\xi' \, \chi_f(\xi') = \frac{1}{f} D\,\chi_f(\xi')$ it can be rewritten as

$$D\,\mathcal{F}T_{\chi_f}(\phi) = \mathrm{i}T_{\chi_f}(\xi'\mathcal{F}\phi) = \mathrm{i}\int_{\mathbb{R}^2} \xi' \, \chi_f(\xi') \, (\mathcal{F}\phi)(\xi') \, d\xi' = \frac{1}{f}\int_{\mathbb{R}^2} \left(D\chi_f\right)(\xi') \, (\mathcal{F}\phi)(\xi') \, d\xi'.$$

Now, since $\chi_f(\xi')\mathcal{F}\phi(\xi') \to 0$ as $|\xi'| \to \infty$, partial integration leads to

$$D\,\mathcal{F}T_{\chi_f}(\phi) = -\frac{\mathrm{i}}{f}\int_{\mathbb{R}^2}\chi_f(\xi')\,\xi'(\mathcal{F}\phi)(\xi')\,d\xi' = -\frac{\mathrm{i}}{f}\xi'\mathcal{F}T_{\chi_f}(\phi).$$

Thus, $\mathcal{F}T_{\chi_f}$ is a (weak) solution of the differential equation

$$D\,T + \frac{\mathrm{i}}{f}\xi' T = 0.$$

We rewrite $\mathcal{F}T_{\chi_f}$ as the product of $\frac{\mathrm{i}}{f}\chi_{-\frac{1}{f}} \in C^\infty(\mathbb{R}^2)$ and some tempered distribution $T \in S'(\mathbb{R}^2)$. Then, since $\frac{\mathrm{i}}{f}\chi_{-\frac{1}{f}} \in C^\infty(\mathbb{R}^2)$ solves the classical differential equation

$$D\,\psi(\xi') + \frac{\mathrm{i}}{f}\xi'\psi(\xi') = 0, \quad \psi \in C^1(\mathbb{R}^2),\ \xi' \in \mathbb{R}^2,$$

we conclude (by applying the product rule) that $D\,T$ vanishes on $S(\mathbb{R}^2)$. Hence, $\mathcal{F}T_{\chi_f}$ is of the form T_ψ for $\psi = a\,\frac{\mathrm{i}}{f}\chi_{-\frac{1}{f}}$ and some constant $a \in \mathbb{C}$. It is well-known that

$$\int_{\mathbb{R}} \cos\left(\frac{f}{2}t^2\right)\,dt = \frac{\sqrt{\pi}}{\sqrt{f}}, \quad \int_{\mathbb{R}} \sin\left(\frac{f}{2}t^2\right)\,dt = \frac{\sqrt{\pi}}{\sqrt{f}},$$

such that the boundary condition

$$\psi(0) = \mathcal{F}\chi_f(0) = \frac{1}{2\pi}\int_{\mathbb{R}^2} e^{\mathrm{i}\frac{f}{2}|\mathbf{x}'|^2}\,d\mathbf{x}' = \frac{1}{2\pi}\left(\int_{\mathbb{R}} \cos\left(\frac{f}{2}t^2\right)\,dt + \mathrm{i}\int_{\mathbb{R}} \sin\left(\frac{f}{2}t^2\right)\,dt\right)^2 = \frac{\mathrm{i}}{f}$$

uniquely determines $\psi = \frac{\mathrm{i}}{f}\chi_{-\frac{1}{f}}$. □

Now, applying the formula

$$\mathcal{F}^{-1}(T_{\chi_{-f}}\mathcal{F}u_0) = \frac{1}{2\pi}\left(\mathcal{F}^{-1}T_{\chi_{-f}} * u_0\right) = \frac{-i}{2\pi f}\left(T_{\chi_{\frac{1}{f}}} * u_0\right), \quad u_0 \in S\left(\mathbb{R}^2\right)$$

to (1.2) leads to the *convolution representation* of D_Γ

$$\mathcal{D}_\Gamma\, u_0\,(\mathbf{x}') := \frac{-\mathrm{i}\kappa}{2\pi\Gamma}\,\mathrm{e}^{\mathrm{i}\kappa\Gamma}\int_{\mathbb{R}^2}\chi_{\frac{\kappa}{\Gamma}}(\mathbf{x}'-\mathbf{y}')\,u_0(\mathbf{y}')\,d\mathbf{y}'\,. \tag{1.4}$$

By the extension of this equation to $L^2(\mathbb{R}^2)$ we again obtain an operator mapping from $L^2(\mathbb{R}^2)$ to $L^2(\mathbb{R}^2)$. From expanding the square $|\mathbf{x}' - \mathbf{y}'|^2$ we derive the identity

$$\chi_{\frac{\kappa}{\Gamma}}(\mathbf{x}'-\mathbf{y}') = \chi_{\frac{\kappa}{\Gamma}}(\mathbf{x}')\exp\left(-\mathrm{i}\frac{\kappa}{\Gamma}\mathbf{x}'\cdot\mathbf{y}'\right)\chi_{\frac{\kappa}{\Gamma}}(\mathbf{y}')\,, \tag{1.5}$$

and hence end up with another useful formulation of $\mathcal{D}_\Gamma$:

$$\mathcal{D}_\Gamma\, u_0\,(\mathbf{x}') = \frac{-\mathrm{i}\kappa}{\Gamma}\,\mathrm{e}^{\mathrm{i}\kappa\Gamma}\,\chi_{\frac{\kappa}{\Gamma}}(\mathbf{x}')\,\mathcal{F}\left(\chi_{\frac{\kappa}{\Gamma}}\,u_0\right)\left(\frac{\kappa}{\Gamma}\mathbf{x}'\right). \tag{1.6}$$

In a concrete setting, often the propagation distance $\Gamma > 0$, which controls the oscillations of the chirp functions in (1.2) and (1.6), determines which of the two representations (1.2) and (1.6) is favorable. So, we call Equation (1.2) *near field representation* and (1.6) *far field representation* of the Fresnel approximation. See Section 1.1.2 for more details.

Remark 1.1.4. By its near field representation (1.2) it is easy to see that the Fresnel approximation satisfies the following propagator property

$$\mathcal{D}_R \circ \mathcal{D}_\Gamma = \mathcal{D}_{R+\Gamma}, \quad R, \Gamma > 0.$$

Through defining $\mathcal{D}_0\, u_0 := u_0$, the set of Fresnel propagators $\{\mathcal{D}_\Gamma \mid \Gamma \geq 0\}$ together with this operation becomes a monoid, i.e. a semigroup with an identity element.

Now that we have a representation of the measured intensities $|\mathcal{D}_\Gamma u_0|^2$, given by the boundary values u_0, we specify the unknown sample information $\phi : D(\phi) \subset \mathbb{P}_0 \to \mathbb{C}$ and study its connection to u_0. In the volume between $\mathbb{P}_{-\tau}$ and $\mathbb{P}_0$ the field interacts with matter of which we assume the material properties to vary sufficiently slowly on length scales comparable to the wavelength. Then the Helmholtz equation becomes:

$$\Delta u + n^2\kappa^2 u = 0 \qquad \text{in } \mathbb{M} := \{x \in \mathbb{R}^3 : -\tau < x_3 < 0\}, \tag{1.7}$$

with refraction index $n : \mathbb{M} \to \mathbb{C}$. In vacuum, this refraction index n coincides with 1, cf. (1.1a), and also in each point $x \in \mathbb{M}$ its value $n(x)$ is close to 1, characterizing the rather weak interaction of x-rays with matter. Therefore, we will use the common notation

$$n(x) = 1 - \delta(x) + \mathrm{i}\beta(x), \quad x \in \mathbb{M},$$

where $\delta \ll 1$ and $\beta \ll 1$ are positive real valued functions. By rewriting the field as

$$u(x) = \tilde{u}(x)\, \exp(\mathrm{i}\,\kappa\, x_3)$$

we separate the rapidly oscillating unscattered plane wave component $x_3 \to \exp(\mathrm{i}\,\kappa x_3)$ from the envelope $\tilde{u}$. Then, because of the identity $\frac{\partial^2 u}{\partial x_3^2} = \exp(\mathrm{i}\,\kappa\, x_3)\left(\frac{\partial^2}{\partial x_3^2} + 2\mathrm{i}\kappa\frac{\partial u}{\partial x_3} - \kappa^2\right)\tilde{u}$, we obtain from the Helmholtz equation (1.7) the following equation:

$$0 = \exp(\mathrm{i}\,\kappa\, x_3)\left(\Delta' + \frac{\partial^2}{\partial x_3^2} + 2\mathrm{i}\kappa\frac{\partial}{\partial x_3} + \kappa^2\left(n(x)^2 - 1\right)\right)\tilde{u}(x) \quad x \in \mathbb{M}.$$

Here $\Delta' = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2}$ denotes the Laplacian in two dimensions. By again assuming the interaction of x-rays with matter to be sufficiently weak we can neglect all the second order derivatives which yields the *projection approximation*:

$$0 \approx \exp(\mathrm{i}\,\kappa\, x_3)\left(2\mathrm{i}\frac{\partial}{\partial x_3} + \kappa\left(n(x)^2 - 1\right)\right)\tilde{u}(x). \tag{1.8}$$

Moreover, by using $n(x)^2 - 1 \approx -2\delta(x) + 2\mathrm{i}\beta(x)$ we can approximate $\tilde{u}$ by the solution u_e of the partial differential equation

$$\frac{\partial}{\partial x_3} u_e = -\mathrm{i}\kappa\,(\delta - \mathrm{i}\beta)\, u_e,$$

and hence end up with the boundary condition

$$u_0(\mathbf{x}') \approx u_e(\mathbf{x}', 0) = u_e(\mathbf{x}', -\tau)\ \exp\left(-\mathrm{i}\,\kappa \int_{-\tau}^{0} \delta(x) - \mathrm{i}\beta(x)\, dx_3\right).$$

Now, we specify the unknown information ϕ of interest as the line integral

$$\phi(\mathbf{x}') = \int_{-\tau}^{0} \delta(x) - \mathrm{i}\beta(x)\, dx_3, \qquad x = (\mathbf{x}', x_3)$$

and we define the *object function*

$$O(\mathbf{x}') = \mathrm{e}^{-\mathrm{i}\,\kappa\,\phi(\mathbf{x}')}. \tag{1.9}$$

So, the imaginary part of ϕ, given by $\int_{-\tau}^{0} \beta(x)\, dx_3$, describes the absorption components of the object, while the real part models the phase shifting components. In the following, we mainly focus on the case of pure phase objects, i.e. the amplitude $|O| = \mathrm{e}^{-\kappa \Im(\phi)}$ is assumed to be constant $\mathbf{1}$ or equivalently β vanishes. This is for example a suitable assumption for thin biological samples illuminated by hard x-rays. Introducing the *illumination function* $\iota : \mathbb{R}^3 \to \mathbb{C}$, $\iota(\mathbf{x}', -\tau) = u_e(\mathbf{x}', -\tau)$, which describes the illumination of the sample, the projection approximation reads as

$$u_0(\mathbf{x}') \approx \iota(\mathbf{x}', -\tau)\, O(\mathbf{x}'), \qquad \mathbf{x}' \in \mathbb{R}^2.$$

For simplicity we write

$$u_0 \approx \iota_0\, O, \tag{1.10}$$

which is justified by the assumption that the illumination $\iota_0 = \exp(\mathrm{i}\,\kappa\, x_3)\,\iota(\cdot, -\tau)$ in the object plane $\mathbb{P}_0$ coincides with $\iota(\cdot, -\tau)$ up to a phase factor. This representation of the exit field u_0 as the product of object function O and illumination ι restricted to $\mathbb{P}_0$ requires the sample thickness τ and the wavelength λ to be sufficiently small. See e.g. [33] or [73] for the more general case of *product approximation*.

1.1.2 Fresnel propagator for tempered distributions.

In practice, frequently not only the object function O of a pure phase object (see (1.9)) but also the illumination $\iota_0(\mathbf{x}') := \iota(\mathbf{x}', 0)$ in $\mathbb{P}_0$ has modulus 1 for every $\mathbf{x}' \in \mathbb{R}^2$. Therefore the boundary value $u_0 \approx \iota_0\, O$ in the projection approximation does not belong to the space $L^2(\mathbb{R}^2)$. So, in order to approximate u in the detector plane $\mathbb{P}_\Gamma$ by $D_\Gamma u_0$ for more general u_0, in the following we extend the definition of the Fresnel approximation propagator to the space $S'(\mathbb{R}^2)$ of tempered distributions.

Recall from Lemma 1.1.1 that

$$\mathcal{D}_\Gamma\, \phi = \frac{-\mathrm{i}\kappa}{\Gamma}\, \mathrm{e}^{\mathrm{i}\kappa\Gamma}\, \chi_{\frac{\kappa}{\Gamma}}\, \mathcal{F}\left(\chi_{\frac{\kappa}{\Gamma}}\, u_0\right)\left(\frac{\kappa}{\Gamma}\cdot\right) = \frac{-\mathrm{i}k}{2\pi\Gamma}\, \mathrm{e}^{\mathrm{i}\kappa\Gamma}\, \chi_{\frac{\kappa}{\Gamma}} * \phi \in S\left(\mathbb{R}^2\right) \tag{1.11}$$

for all $\phi \in S\left(\mathbb{R}^2\right)$ and that $\mathcal{D}_\Gamma : S\left(\mathbb{R}^2\right) \to S\left(\mathbb{R}^2\right)$ is continuous. Since the Fresnel propagator $\mathcal{D}_\Gamma$ is similar to the Fourier transform, it seems to be natural to define $\mathcal{D}_\Gamma(T)$ for a tempered distribution $T \in S'\left(\mathbb{R}^2\right)$ in the same way, namely by

$$\mathcal{D}_\Gamma(T)(\phi) := T(\mathcal{D}_\Gamma \phi), \quad \forall \phi \in S\left(\mathbb{R}^2\right). \tag{1.12}$$

The following corollary justifies this definition.

Corollary 1.1.5. *For any $T \in S'\left(\mathbb{R}^2\right)$ the Fresnel propagation $\mathcal{D}_\Gamma T$ defined by* (1.12) *exists as a tempered distribution, i.e. $\mathcal{D}_\Gamma T \in S'\left(\mathbb{R}^2\right)$. Moreover, it is well-defined in the sense that $\mathcal{D}_\Gamma T_\varphi = T_{\mathcal{D}_\Gamma \varphi}$ for all $\varphi \in S\left(\mathbb{R}^2\right)$.*

Proof. As the linearity of $\mathcal{D}_\Gamma T$ is obvious, we only have to show the continuity of $\mathcal{D}_\Gamma T$ in order to prove the first assertion. To this end, we choose a sequence $(\phi_n)_{n\in\mathbb{N}} \subset S\left(\mathbb{R}^2\right)$ with $\phi_n \to 0,\ n \to \infty$. Then we have $\mathcal{D}_\Gamma \phi_n \to 0$ as $n \to \infty$ and also

$$\mathcal{D}_\Gamma T(\phi_n) = T\left(\mathcal{D}_\Gamma \phi_n\right) \to 0, \quad n \to \infty$$

which gives the first assertion. For any $\phi,\ \varphi \in S\left(\mathbb{R}^2\right)$ we have

$$\begin{aligned} T_{\mathcal{D}_\Gamma \varphi}(\phi) &= \frac{-\mathrm{i}\kappa}{2\pi\Gamma}\, \mathrm{e}^{\mathrm{i}\kappa\Gamma} \left\langle \phi, \int_{\mathbb{R}^2} \chi_{\frac{\kappa}{\Gamma}}\left(\cdot - \mathbf{x}'\right) \varphi(\mathbf{x}')\, d\mathbf{x}' \right\rangle_{S(\mathbb{R}^2)} \\ &= \frac{-\mathrm{i}k}{2\pi R}\, \mathrm{e}^{\mathrm{i}\kappa\Gamma} \int_{\mathbb{R}^2} \int_{\mathbb{R}^2} \varphi(\mathbf{x}')\, \chi_{\frac{\kappa}{\Gamma}}\left(\mathbf{x}' - \xi'\right)\ \phi(\xi')\, d\xi'\, d\mathbf{x}' = \left\langle \phi, \mathcal{D}_\Gamma T_\varphi \right\rangle_{S(\mathbb{R}^2)}, \end{aligned}$$

where $\langle \cdot, \cdot \rangle_{S(\mathbb{R}^2)}$ denotes the dual paring $\langle \phi, \varphi \rangle_{S(\mathbb{R}^2)} := \int_{\mathbb{R}^2} \varphi(\mathbf{x}')\, \phi(\mathbf{x}')\, d\mathbf{x}'$. This completes the proof. □

Example 1.1.6. For $\Gamma > 0$ the Fresnel propagation $\mathcal{D}_\Gamma\left(\delta_0\right)$ of the delta distribution $\delta_0 \in S'\left(\mathbb{R}^2\right)$, with

$$\delta_0 : \phi \mapsto \phi(0),$$

is given by the regular distribution $T_f \in S'\left(\mathbb{R}^2\right)$ where

$$f : \mathbb{R}^2 \to \mathbb{C}, \quad \mathbf{x}' \mapsto \frac{-\mathrm{i}k}{2\pi\Gamma} \mathrm{e}^{\mathrm{i}\kappa\Gamma} \chi_{\frac{\kappa}{\Gamma}}(\mathbf{x}'). \tag{1.13}$$

Proof. For $\phi \in S\left(\mathbb{R}^2\right)$ we have

$$\begin{aligned} \left(\mathcal{D}_\Gamma \delta_0\right)(\phi) = \delta_0\left(\mathcal{D}_\Gamma \phi\right) = \left(\mathcal{D}_\Gamma \phi\right)(0) &= \frac{-\mathrm{i}\kappa}{2\pi\Gamma}\, \mathrm{e}^{\mathrm{i}\kappa\Gamma} \int_{\mathbb{R}^2} \chi_{\frac{\kappa}{\Gamma}}\left(\mathbf{x}'\right)\ \phi(\mathbf{x}')\, d\mathbf{x}' \\ &= \frac{-\mathrm{i}\kappa}{2\pi\Gamma}\, \mathrm{e}^{\mathrm{i}\kappa\Gamma} \left\langle \phi, \chi_{\frac{\kappa}{\Gamma}} \right\rangle_{S(\mathbb{R}^2)}. \end{aligned}$$

□

Example 1.1.7. The Fresnel propagation of the constant function $\mathbf{1}$ in the sense of the regular distribution $T_{\mathbf{1}} \in S'\left(\mathbb{R}^2\right)$ is given by

$$\mathcal{D}_\Gamma\left(T_{\mathbf{1}}\right) = \frac{1}{2\pi}\mathrm{e}^{\mathrm{i}k\Gamma} T_{\chi_{-\frac{\Gamma}{\kappa}}} \qquad \Gamma > 0.$$

Proof. From $\mathcal{F}\delta_0 = T_{\mathbf{1}}$ and $\mathcal{F}T_{\chi_{\frac{\kappa}{\Gamma}}} = \frac{\mathrm{i}\Gamma}{k}T_{\chi_{-\frac{\Gamma}{\kappa}}}$ we obtain for $\phi \in S\left(\mathbb{R}^2\right)$:

$$\begin{aligned}
\mathcal{D}_\Gamma\left(T_{\mathbf{1}}\right)(\phi) &= \mathcal{D}_\Gamma\left(\mathcal{F}\delta_0\right)(\phi) = \delta_0\left(\mathcal{D}_\Gamma\left(\mathcal{F}\phi\right)\right) = \frac{-\mathrm{i}\kappa}{2\pi\Gamma}\mathrm{e}^{\mathrm{i}\kappa\Gamma} T_{\chi_{\frac{\kappa}{\Gamma}}}\left(\mathcal{F}\phi\right) \\
&= \frac{-\mathrm{i}\kappa}{2\pi\Gamma}\mathrm{e}^{\mathrm{i}\kappa\Gamma}\mathcal{F}T_{\chi_{\frac{k}{\Gamma}}}(\phi) = \frac{1}{2\pi}\mathrm{e}^{\mathrm{i}\kappa\Gamma}T_{\chi_{-\frac{\Gamma}{\kappa}}}(\phi). \qquad \square
\end{aligned}$$

Now let us consider (tempered) distributions with compact support (cf. Appendix A.1):

$$\mathcal{E}' := \left\{T \in S'\left(\mathbb{R}^2\right) \mid \operatorname{supp} T \text{ is compact}\right\}.$$

Lemma 1.1.8. *Let $T \in \mathcal{E}'$ be a tempered distribution with compact support. Then*

$$g(\xi') = \frac{-i\kappa}{\Gamma}\, e^{i\kappa\Gamma}\, \chi_{\frac{\kappa}{\Gamma}}(\xi')\, \left\langle K_{\mathcal{F}}\left(\frac{\kappa}{\Gamma}\xi',\cdot\right), \chi_{\frac{\kappa}{\Gamma}}T\right\rangle_{S\left(\mathbb{R}^2\right)}$$

with

$$K_{\mathcal{F}} : (\xi', \mathbf{x}') \mapsto \frac{1}{2\pi}e^{-i\xi'\cdot\mathbf{x}'} \qquad \xi', \mathbf{x}' \in \mathbb{R}^2$$

is a $C^\infty(\mathbb{R}^2)$-function that defines a tempered distribution T_g such that $T_g = \mathcal{D}_\Gamma T$. In particular, if $T = T_\omega$ is a regular distribution with $\omega \in L^r(\mathbb{R}^2)$, $r \in [1,\infty]$ compactly supported we obtain

$$g(\xi') = \frac{-i\kappa}{\Gamma}\, e^{i\kappa\Gamma}\, \chi_{\frac{\kappa}{\Gamma}}(\xi')\, \mathcal{F}\left(\chi_{\frac{\kappa}{\Gamma}}\omega\right)\left(\frac{\kappa}{\Gamma}\xi'\right).$$

Equivalently, we have (in the convolution representation)

$$g : \xi' \mapsto \frac{-i\kappa}{2\pi\Gamma}\, e^{i\kappa\Gamma}\, \left\langle \chi_{\frac{\kappa}{\Gamma}}\left(\xi' - \cdot\right), T\right\rangle_{S\left(\mathbb{R}^2\right)} = \frac{-i\kappa}{2\pi\Gamma}\, e^{i\kappa\Gamma}\, \left(\chi_{\frac{\kappa}{\Gamma}} * T\right)(\xi') \quad \xi' \in \mathbb{R}^2.$$

Proof. By [31, Corollary 4.1.2] and $\chi_{\frac{\kappa}{\Gamma}}T \in \mathcal{E}'$ it follows that g is a $C^\infty(\mathbb{R}^2)$-function. Moreover, $T \in S'\left(\mathbb{R}^2\right)$ ensures $\mathcal{D}_\Gamma T \in S'\left(\mathbb{R}^2\right)$. Using [31, Theorem 8.4.1]:

$$\langle \mathcal{F}\phi, E\rangle_{S\left(\mathbb{R}^2\right)} = \mathcal{F}(E)(\phi) = \int_{\mathbb{R}^2} \phi(\xi')\, \langle K_{\mathcal{F}}(\xi',\cdot), E\rangle_{S\left(\mathbb{R}^2\right)}\, d\xi' \quad \phi \in S\left(\mathbb{R}^2\right),\ E \in \mathcal{E}'$$

as well as the multiplier property of $\chi_{\frac{\kappa}{\Gamma}}$ (cf. Lemma 1.1.1) we obtain

$$\begin{aligned}
\mathcal{D}_\Gamma T(\phi) &= \frac{-\mathrm{i}\kappa}{\Gamma}\, \mathrm{e}^{\mathrm{i}\kappa\Gamma} \left\langle \chi_{\frac{\kappa}{\Gamma}}\, \mathcal{F}\left(\chi_{\frac{\kappa}{\Gamma}}\phi\right)\left(\frac{\kappa}{\Gamma}\cdot\right), T\right\rangle_{S\left(\mathbb{R}^2\right)} \\
&= \frac{-\mathrm{i}\kappa}{\Gamma}\, \mathrm{e}^{\mathrm{i}\kappa\Gamma} \int_{\mathbb{R}^2} \phi(\xi')\chi_{\frac{\kappa}{\Gamma}}(\xi')\left\langle K_{\mathcal{F}}\left(\frac{\kappa}{\Gamma}\xi',\cdot\right), \chi_{\frac{\kappa}{\Gamma}}T\right\rangle_{S\left(\mathbb{R}^2\right)} d\xi' = T_g(\phi)
\end{aligned}$$

for all $\phi \in S\left(\mathbb{R}^2\right)$. This proves the first assertion. Expansion (1.5) gives the convolution representation. □

We assume that $O \in L^\infty(\mathbb{R}^2)$ varies only within a rectangle $[-\mathbf{r}_O, \mathbf{r}_O] \subset \mathbb{R}^2$, i.e. there is constant C such that $\tilde{O} := O - C$ is compactly supported in $[-\mathbf{r}_O, \mathbf{r}_O]$. From the linearity of $\mathcal{D}_\Gamma$ and Example 1.1.7 we obtain

$$\mathcal{D}_\Gamma(T_O) = \mathcal{D}_\Gamma(T_{O-C}) + \frac{C}{2\pi} \mathrm{e}^{\mathrm{i}\kappa\Gamma} T_{\chi_{-\frac{\Gamma}{\kappa}}},$$

and thus the Fresnel approximation $\mathcal{D}_\Gamma(T_O)$ in the sense of tempered distributions can be calculated by the Fresnel propagation of the $L^2(\mathbb{R}^2)$-function $\tilde{O} = O - C$:

$$\mathcal{D}_\Gamma(O - C) + C\mathrm{e}^{\mathrm{i}\kappa\Gamma} \chi_{-\frac{\Gamma}{\kappa}}.$$

In order to keep the notation simple, we identify O with $\tilde{O}$ and make the following assumption:

Assumption 1.1.9. *The object function $O \in L^2(\mathbb{R}^2)$ is supported in some rectangle $[-\mathbf{r}_O, \mathbf{r}_O] \subset \mathbb{R}^2$.*

On the other hand we denote by $[-\mathbf{r}_X, \mathbf{r}_X]$, with diameter $\mathbf{r}_X \geq \mathbf{r}_O$, the area in $\mathbb{P}_0$ outside of which the whole field $u_0 \approx \iota_0 O$ (in the projection approximation) is nearly constant. Then, if the dimensionless *Fresnel number*

$$\mathfrak{f} := \frac{\kappa\left(r_{X,1}^2, r_{X,2}^2\right)}{\Gamma} \tag{1.14}$$

is sufficiently small such that $\chi_{\frac{\kappa}{\Gamma}} \approx \mathbf{1}$ on $[-\mathbf{r}_X, \mathbf{r}_X]$, we can neglect this factor in (1.6), which leads to the well-known *Fraunhofer approximation*

$$u(\mathbf{x}', \Gamma) \approx \frac{-\mathrm{i}\kappa}{\Gamma} \mathrm{e}^{\mathrm{i}\kappa\Gamma} \chi_{\frac{\kappa}{\Gamma}}(\mathbf{x}') \, \mathcal{F} u_0\left(\frac{\kappa}{\Gamma}\mathbf{x}'\right). \tag{1.15}$$

Note that $\mathfrak{f} \ll 1$ is in particular satisfied if the propagation distance Γ turns to ∞. For this reason (1.15) is also referred to as *far field diffraction* formula.

Now we are able to formulate the forward operator T for the introduced problem of retrieving the unknown object information ϕ from the (possibly noisy) intensities measurements y^δ in the detector plane $\mathbb{P}_\Gamma$. y^δ is given by the exact data $y = |u(\cdot, \Gamma)|^2$ possibly perturbed by some noise. Under the use of the projection approximation (1.8) and the Fresnel approximation (see Equation (1.2) or Equation (1.6), respectively) the operator reads as

$$T_{\mathrm{Fresnel}}(\phi)(\mathbf{x}', \Gamma) = \left|\mathcal{D}_\Gamma\left(\iota_0 O(\phi)\right)(\mathbf{x}')\right|^2, \quad (\mathbf{x}', \Gamma) \in \mathbb{P}_\Gamma,$$

while in the case of the Fraunhofer approximation (1.15) we obtain

$$T_{\mathrm{Frau}}(\phi)(\mathbf{x}', \Gamma) = \left|\frac{\kappa}{\Gamma}\right|^2 \left|\mathcal{F}\left(\iota_0 O(\phi)\right)\left(\frac{\kappa}{\Gamma}\mathbf{x}'\right)\right|^2 \quad (\mathbf{x}', \Gamma) \in \mathbb{P}_\Gamma. \tag{1.16}$$

Here the dependence given by Equation (1.9) of the object function O on ϕ is indicated by writing O as a function of ϕ. The following Fresnel scaling theorem provides another useful formulation for each of the two operators.

1.1.3 Fresnel scaling theorem

Let us assume that the sample is illuminated by an ideal point source which is located at $(0, 0, -R) \in \mathbb{H}^-$. Then, because of Example 1.1.6, we state the empty beam field ι_0 in the object plane $\mathbb{P}_0$ to be given by (1.13) for $\Gamma = R$. This Fresnel approximation is justified by the fact that for $|\mathbf{x}'| := \sqrt{x_1^2 + x_2^2} \ll R$ the Taylor approximation $\sqrt{|\mathbf{x}'|^2 + R^2} \approx R + \frac{|\mathbf{x}'|^2}{2R}$ classifies ι_0 as an approximation of the spherical wave scattered from $(0, 0, -R)$:

$$(\mathbf{x}', x_3) \mapsto \frac{\mathrm{e}^{\mathrm{i}\kappa\sqrt{|\mathbf{x}'|^2+(x_3+R)^2}}}{\sqrt{|\mathbf{x}'|^2 + (x_3 + R)^2}}, \tag{1.17}$$

restricted to $\mathbb{P}_0$. Note that the field given by (1.17) is an outgoing solution to the Helmholtz equation (1.1a) in the volume $\left\{x \in \mathbb{R}^3 \mid -R < x_3 < 0\right\}$. Then, assuming the product approximation (1.10) to be valid, u_0 reads as

$$u_0(\mathbf{x}') = \frac{-\mathrm{i}\kappa}{2\pi R}\mathrm{e}^{\mathrm{i}\kappa R}\chi_{\frac{\kappa}{R}}(\mathbf{x}')\,O(\mathbf{x}') \quad \mathbf{x}' \in \mathbb{R}^2.$$

This motivates us to factor out the spherical part of u_0 also in more general cases of illumination caused by a point-like source. It means we write u_0 in the form

$$u_0(\mathbf{x}') = \chi_{\frac{\kappa}{R}}(\mathbf{x}')\,P(\mathbf{x}')\,O(\mathbf{x}'), \quad |\mathbf{x}'| \ll R \tag{1.18}$$

with (presumably) nearly plane envelope P. Inserting this ansatz into the far field representation (1.6) yields

$$\mathcal{D}_\Gamma\, u_0\,(\mathbf{x}') = \frac{-\mathrm{i}k}{\Gamma}\,\mathrm{e}^{\mathrm{i}k\Gamma}\,\chi_{\frac{k}{\Gamma}}(\mathbf{x}')\,\mathcal{F}\left(\chi_{\frac{\kappa M}{\Gamma}}\,P\,O\right)\left(\frac{k}{\Gamma}\mathbf{x}'\right) \tag{1.19}$$

where $M > 0$ denotes the *geometrical magnification*

$$M = \frac{R+\Gamma}{R} \quad \text{with} \quad \frac{M}{\Gamma} = \frac{1}{R} + \frac{1}{\Gamma}.$$

Moreover, introducing effective coordinates

$$\mathbf{x}'_{\mathit{eff}} := \frac{\mathbf{x}'}{M} \quad \text{and} \quad \Gamma_{\mathit{eff}} := \frac{\Gamma}{M},$$

we obtain

$$\begin{aligned}\mathcal{D}_\Gamma\, u_0\,(\mathbf{x}') &= \frac{1}{M}\mathrm{e}^{\mathrm{i}\,\kappa\,\Gamma(1-\frac{1}{M})}\,\chi_{\frac{\kappa}{R+\Gamma}}(\mathbf{x}')\frac{-\mathrm{i}k}{\Gamma_{\mathit{eff}}}\,\mathrm{e}^{\mathrm{i}k\Gamma_{\mathit{eff}}}\,\chi_{\frac{k}{\Gamma_{\mathit{eff}}}}\left(\mathbf{x}'_{\mathit{eff}}\right)\,\mathcal{F}\left(\chi_{\frac{\kappa}{\Gamma_{\mathit{eff}}}}\,P\,O\right)\left(\frac{k}{\Gamma_{\mathit{eff}}}\mathbf{x}'_{\mathit{eff}}\right)\\ &= \frac{1}{M}\mathrm{e}^{-\mathrm{i}\,\kappa\,\Gamma(1-\frac{1}{M})}\,\chi_{\frac{\kappa}{R+\Gamma}}(\mathbf{x}')\,\mathcal{D}_{\Gamma_{\mathit{eff}}}\,P\,O\,\left(\mathbf{x}'_{\mathit{eff}}\right)\end{aligned} \tag{1.20}$$

which gives another formulation for the intensities $T_{\mathrm{Fresnel},P\,O}(\phi)(\cdot, \Gamma) = |\mathcal{D}_\Gamma\, u_0|^2$:

$$|\mathcal{D}_\Gamma\, u_0\,(\mathbf{x}')|^2 = \frac{1}{M^2}\frac{\kappa^2}{\Gamma_{\mathit{eff}}^2}\left|\mathcal{F}\left(\chi_{\frac{k}{\Gamma_{\mathit{eff}}}}\,P\,O\right)\left(\frac{k}{\Gamma_{\mathit{eff}}}\mathbf{x}'_{\mathit{eff}}\right)\right|^2. \tag{1.21}$$

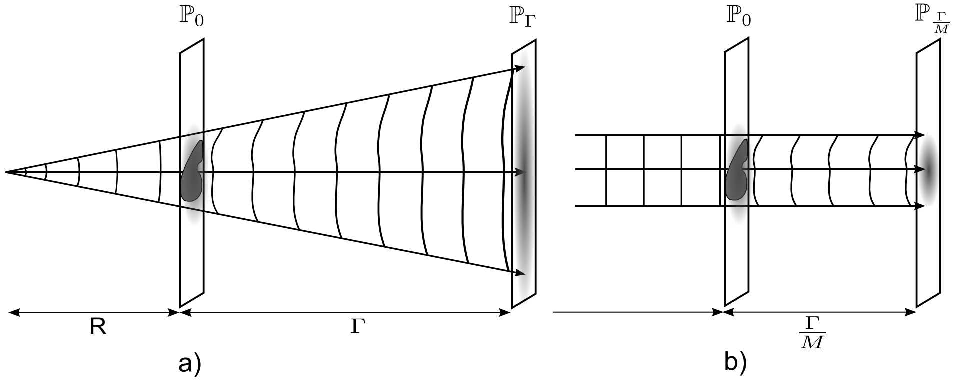


Figure 1.2: Illustration of the Fresnel scaling theorem: a) (quasi) point source illumination. b) plane wave illumination in an effective geometry.

Using again Equation (1.19), we end up with the *Fresnel scaling theorem*:

$$\frac{\kappa^2}{\Gamma^2}\left|\mathcal{F}\left((\chi_{\frac{k}{\Gamma}+\frac{k}{R}}\,PO)\right)\left(\frac{k}{\Gamma}\mathbf{x}'\right)\right|^2 = |\mathcal{D}_\Gamma\, u_0\,(\mathbf{x}')|^2 = \frac{1}{M^2}\frac{\kappa^2}{\Gamma_{\mathit{eff}}^2}\left|\mathcal{F}\left(\chi_{\frac{k}{\Gamma_{\mathit{eff}}}}\,PO\right)\left(\frac{k}{\Gamma_{\mathit{eff}}}\mathbf{x}'_{\mathit{eff}}\right)\right|^2, \quad (1.22)$$

for all $\mathbf{x}' \in \mathbb{R}^2$. The right hand side of this equation is also the limit of the left hand side when R turns to ∞ such that $M \to 1$ and $\frac{\kappa}{R} \to 0$. So, the Fresnel scaling theorem relates the intensity measured in $\mathbb{P}_\Gamma$ for spherical like illumination to the intensity in an effective geometry $M^{-1}\mathbb{P}_{\Gamma_{\mathit{eff}}}$ for illumination by a plane wave where ι_0 is constant. Figure 1.2 gives an illustration.

Accordingly, we can apply the Fresnel scaling theorem (1.22) to the near field representation (1.2) which yields another useful representation of the intensity measurements:

$$|\mathcal{D}_\Gamma\, u_0\,(\mathbf{x}')|^2 = \frac{1}{M^2}\left|\mathcal{F}^{-1}\left(\chi_{-\frac{\Gamma}{Mk}} \bullet \mathcal{F}PO\right)\left(\frac{\mathbf{x}'}{M}\right)\right|^2 \qquad \mathbf{x}' \in \mathbb{R}^2, \quad (1.23)$$

where $\bullet$ is the pointwise multiplication. With respect to the numerical approximation, the crucial difference between the two formulas (1.21) and (1.23) (and likewise between (1.6) and (1.2)) lies in the inverse parameters $\frac{\kappa M}{\Gamma}$ and $-\frac{\Gamma}{\kappa M}$ of the corresponding chirp functions. In order to keep their oscillations small, one usually uses representation (1.21) if both the Fresnel number (1.14) as well as the magnification M are sufficiently small:

$$M\mathfrak{f} = \frac{M\kappa\left(r_{X,1}^2, r_{X,2}^2\right)}{\Gamma} \leq 1$$

and (1.23) otherwise. In the limit case of $M\mathfrak{f} \approx 0$, however, the operator (1.16) based on the Fraunhofer approximation is favorable.

Remark 1.1.10. All the derived representations for the forward operators T_{Fresnel} and T_{Frau} depend on the empty beam field ι_0, which we implicitly presupposed to be given by a spherical or equivalently (in the effective geometry) a plane wave. However, especially in the near field regime where usually formula (1.23) is used this assumption might not be fulfilled. In Chapter 2, we study the validity of a common approach to overcome this problem by 'factoring out' ι_0.

1.1.4 Phase retrieval in coherent X-ray imaging: An inverse problem in Banach spaces

As motivated above, we formulate the introduced phase retrieval problem by an operator equation

$$T\phi = y^\delta, \tag{1.24}$$

where $y^\delta \in \mathbb{P}_\Gamma$ are the (possibly noisy) intensity measurements in the detector plane $\mathbb{P}_\Gamma$ from which we want to retrieve the unknown information $\phi \in X \subset \mathbb{P}_0$ of the object function O. The forward operator $T : X \to Y$ maps ϕ to $\chi_{\frac{k}{R}} O$ composed with the modulus square of either the Fraunhofer (1.2) or the Fresnel (1.15) approximation. See Equations (1.21), (1.23), and (1.16) for the different representations of T_{Fresnel} and T_{Frau} . Here we focus on the case of pure phase objects, i.e. $\phi : \mathbb{P}_0 \to \mathbb{R}$, and consider O as a function of ϕ: $O(\phi) = \mathrm{e}^{\mathrm{i}\kappa\phi}$. Obviously, T is not injektive with respect to a phase shift of 2π. To prevent convergence of iterative algorithms to local minima a condition like $\phi(\xi) \in [-\pi, \pi]$ is typically needed. Moreover, in order to ensure uniqueness and simplify the task of reconstruction, one usually uses the a priori information that ϕ has a compact support $\Omega \subset \mathbb{P}_0$, cf. [51], [57]. However, even if injectivity of T is achieved, we are dealing with a non-linear ill-posed inverse problem (cf. [12], [57]).

In the literature, especially projection based methods are used in order to solve phase retrieval problems. In this context, the first approach was proposed by Gerchberg and Saxton [32], followed by Fienup's error-reduction (ER) and hybrid input-output (HIO) algorithms [30]. These methods form the basis for many developed iterative projection algorithms with this application. They are often referred to as Gerchberg-Saxton-Fienup-type algorithms. We commend [56] for an overview. Since it was shown that ER coincides with a nonconvex alternating projection algorithm and HIO is an instance of the well-known Douglas-Rachford algorithm using (generalized) projections onto nonconvex sets, these methods can be seen as the adaption of classical convex optimization methods to the nonconvex feasibility problem of finding a solution ϕ in the intersection of a convex set A of a priori constraints and the nonconvex set $M := \left\{\phi \mid |T\phi| = y^\delta a.e.\right\}$, see [9,53]. Note that due to noise the intersection $A \cap M$ might be empty. As the Gerchberg-Saxton-Fienup-type algorithms use the projection with respect to the L^2-norm, the problem considered there can be rewritten as

$$\text{find } \phi = \underset{\phi \in A}{\operatorname{argmin}} \|T\phi - y^\delta\|_{L^2}. \tag{1.25}$$

It is a main benefit that these methods are applicable for a very general set A of a priori constraints. In addition, they can be easily adapted to changed conditions. Although in practice these algorithms show sufficiently nice convergence behavior, there is still a lack of mathematical foundation for this nonconvex application. See e.g. [38, 39] for current research in this direction. A second drawback is that the problem formulation (1.25) does not take the specific noise model into account: As usual in photonic imaging, we face a problem with Poisson data, i.e. the data is Poisson distributed with the exact data $y^\dagger$ as mean. In our setup (see Figure 1.1), we have a finite (here even) number $\mathbf{N} := (N_1, N_2) \in \mathbb{N}^2$ of equal sized detectors that are located in the detector plane $\mathbb{P}_\Gamma$. Each of these detectors measures the number

$$y^\delta_{\mathbf{i}} \geq 0, \qquad \mathbf{i} \in \Delta_{\mathbf{N}} = \left\{\frac{N_1}{2}, \cdots, \frac{N_1}{2} - 1\right\} \times \left\{\frac{N_2}{2}, \cdots, \frac{N_2}{2} - 1\right\}$$

of photons reaching the pixel area during the exposure time. Then $y^\delta = \left(y^\delta_{\mathbf{i}}\right)$ is an event derived from a vector of $N_1\, N_2$ independent Poisson distributed random variables. Thus, it is natural (cf. [82] sec. 2.1) to minimize the negative log-likelihood functional which is given (up to an additive constant independent of y) by

$$\phi \mapsto kl_{\mathbf{N}}\left(y^\delta; T\phi\right), \quad kl_{\mathbf{N}}(y^\delta; y) := \begin{cases} \sum_{\mathbf{i}\in\Delta_{\mathbf{N}}} y_{\mathbf{i}} - y^\delta_{\mathbf{i}} \ln(y_{\mathbf{i}}), & \text{if } y|_{y^\delta>0} > 0,\ y|_{y^\delta\geq 0} \geq 0 \\ \infty, & \text{otherwise,} \end{cases} \tag{1.26}$$

or equally in the limit case $\mathbf{N} \to \infty \times \infty$ its non-discrete generalization

$$\mathbb{KL}(y^\delta; y) := \begin{cases} \int_{\mathbb{R}^2} y - y^\delta \ln y \, d\mathbf{x}', & \text{if } y|_{y^\delta>0} > 0,\ y|_{y^\delta\geq 0} \geq 0 \\ \infty, & \text{otherwise,} \end{cases}$$

instead of the L^2-norm (corresponding to l^2-norm in the discrete setting) in (1.25). Here, we set $0 \ln 0 := 0$ and by the notation we indicate that $\mathbb{KL}$ is (up to a constant) the *Kullback-Leibler divergence*.

Regularization of ill-posed problems with Poisson data, also with an example in phase retrieval, was studied in detail in a series of works [47,82,83] by Hohage and Werner. So, motivated by the theory developed there, we solve our problem by a more noise adapted method, the iteratively regularized Newton-type method (IRNM, see Section 3.1):

$$\phi_{n+1} = \operatorname*{argmin}_{\phi\in X} S(y^\delta; T(\phi_n) + T'[\phi_n](\phi - \phi_n)) + \alpha_n R(\phi), \tag{1.27}$$

with a *Kullback-Leibler like data fidelity functional*

$$S(y^\delta; y) = \mathbb{KL}(y^\delta + \epsilon; y + \epsilon), \quad \text{with shift parameter } \epsilon \geq 0, \tag{1.28}$$

an appropriate penalty term R, and regularization parameters $\alpha_n > 0$. Here $T'[\phi_n]$ denotes the *Fréchet derivative* in $\phi_n \in X$, i.e. the linear, bounded operator mapping from X to Y with

$$\lim_{\varphi\in X, \|\varphi\|_X\to 0} \frac{1}{\|\varphi\|_X} \|T(\phi_n + \varphi) - T(\phi_n) - T'[\phi_n](\varphi)\|_Y = 0.$$

Its specific formulation will be present in the following. Compared to problem (1.25), here the a priori constraints are modeled by the preimage space X of $T'[\phi_n]$ and the penalty term $R : X \to \mathbb{R} \cup \{\infty\}$. However, also for this method there still a gap in the convergence theory with respect to phase retrieval problems: The tangential cone condition in [47, 82, 83] has not been verified for our operator T. Moreover, in this work we remove parameter-dependence of the data fidelity functional (1.28) by focusing on the natural case $\epsilon = 0$, although the convergence results were proven for $\epsilon > 0$.

On the other hand, Newton-type methods (e.g. the IRNM) and also Landweber-type iterations that address the reconstruction of phase information from modulus measurements of its Fourier transform with a Banach or Hilbert norm monomial as data fidelity functional $S(y^\delta; y) = \|y - y^\delta\|_Y^s$, $s > 1$, have been suggested in several works, such as [7, 13, 27, 29, 44, 70]. While most of these approaches are based on a Hilbert space setting, choosing X (and possibly also Y) as a Banach space seems to be more appropriate: Since we are especially interested in the reconstruction of blocky structured biological samples (see e.g. Figure 5.5) given on a rectangle $\Omega = (-\mathbf{r}_X, \mathbf{r}_X) \in \mathbb{P}_0$, the Sobolev norm

$$\|\phi\|_{W^{1,r}(\Omega)} = \left(\|\phi\|^r_{L^r(\Omega)} + \|\nabla\phi\|^r_{L^r(\Omega)}\right)^{\frac{1}{r}}, \quad r \geq 1$$

with $r \approx 1$ is particularly appropriate as penalty term, and hence the corresponding Banach space $W^{1,r}(\Omega) := \left\{\phi : \overline{\Omega} \to \mathbb{R} \mid \phi, \nabla\phi \in L^r(\Omega)\right\}$ is the natural choice for X. Furthermore, for an operator similar to $T_{\text{Frau}}(\phi) := T_{\text{Frau},O(\phi)}$, in [70, Section 1.2] it was motivated to consider $X = L^r(\mathbb{R}^2)$ and $Y = L^{r^*}(\mathbb{R}^2)$, where $r^* := \frac{r}{r-1}$ is the conjugate exponent of $r \in (1, 2]$. The Fréchet derivative $T'_{\text{Frau}}[\phi] : X \to Y$ is given by

$$T'_{\text{Frau}}[\phi](h)\,(\mathbf{x}', \Gamma) = \frac{2\,\kappa^2}{\Gamma^2}\Re\left(\overline{\mathcal{F}O(\phi)\left(\frac{\kappa}{\Gamma}\mathbf{x}'\right)}\,\mathcal{F}\left(O'[\phi](h)\right)\left(\frac{\kappa}{\Gamma}\mathbf{x}'\right)\right), \quad \mathbf{x}' \in \mathbb{P}_\Gamma$$

$$O(\phi) = \mathrm{e}^{\mathrm{i}\,\kappa\,\phi}, \quad O'[\phi](h) = \mathrm{i}\,\kappa\,h\,\mathrm{e}^{\mathrm{i}\,\kappa\,\phi}$$

In fact, due to the Hausdorff-Young inequality (A.1) the operator $T'_{\text{Frau}}[\phi]$ is bounded by

$$\left\|T'_{\text{Frau}}[\phi](h)\right\|_{L^{r^*}(\mathbb{R}^2)} \leq C\|h\|_{L^r(\mathbb{R}^2)}$$

for some constant $C > 0$. Analogous, for $T_{\text{Fresnel}}(\phi) := T_{\text{Fresnel},O(\phi)}$, we obtain from (1.22)

$$T'_{\text{Fresnel}}[\phi](h)\,(\mathbf{x}', \Gamma) = \frac{2\,\kappa^2}{\Gamma^2}\Re\left(\overline{\mathcal{F}\left(\chi_{\frac{\kappa M}{\Gamma}}O(\phi)\right)\left(\frac{\kappa}{\Gamma}\mathbf{x}'\right)}\,\mathcal{F}\left(\chi_{\frac{\kappa M}{\Gamma}}\,O'[\phi](h)\right)\left(\frac{\kappa}{\Gamma}\mathbf{x}'\right)\right), \quad \mathbf{x}' \in \mathbb{P}_\Gamma$$

and thus the same estimate holds. So, we can assume that a penalty term R which is given by a Banach space norm $\|\cdot\|_X$ for $X \in \left\{L^r(\Omega), W^{1,r}(\Omega)\right\}$ is a good choice to model the 'structure' of ϕ_n. Note that by defining X as a function space on Ω, we incorporate also the support constraint $\operatorname{supp}\phi_n \subseteq \Omega$.

Because of the pixel-wise measurements we focus our attention on the discrete data fidelity functional $kl_{\mathbf{N}}$: The weighted least square approximation (which we obtain form the second order Taylor expansion of $y \mapsto kl_{\mathbf{N}}\left(y^\delta; y\right) + \sum_{\mathbf{j}\in\Delta_{\mathbf{N}}}\left(y^\delta_{\mathbf{j}} - y^\delta_{\mathbf{j}}\ln y^\delta_{\mathbf{j}}\right)$ at y^δ)

$$kl_{\mathbf{N}}\left(y^\delta; y\right) + \sum_{\mathbf{j}\in\Delta_{\mathbf{N}}}\left(y^\delta_{\mathbf{j}} - y^\delta_{\mathbf{j}}\ln y^\delta_{\mathbf{j}}\right) \approx \frac{1}{2}\left\|\left(\frac{y^\delta_{\mathbf{j}} - y_{\mathbf{j}}}{\sqrt{y_{\mathbf{j}}}}\right)_{\mathbf{j}\in\Delta_{\mathbf{N}}}\right\|^2_{l^2}, \quad y|_{y^\delta>0} > 0,\ y|_{y^\delta\geq 0} \geq 0$$

motivates us to choose either $L^2(\mathbb{R}^2)$ or the weighted Hilbert space $L^2_{W_n}(\mathbb{R}^2)$ with positive weight $W_n = (T(\phi_n) + \epsilon)^{-1}$, $\epsilon > 0$, and norm given by $\|y\|_{L^2_{W_n}(\mathbb{R}^2)} = \|W_n^{\frac{1}{2}} y\|_{L^2(\mathbb{R}^2)}$ as image space Y_n of $T'[\phi_n]$. We can assume that there is no intensity outside the finite (open) detector domain $D \subset \mathbb{R}^2$. Therefore, we restrict the operator T to this domain D and set $Y_n = L^2(D)$ or $Y_n = L^2_{W_n}(D)$. Moreover, [57, Theorem 1] ensures that a compactly supported solution ϕ is uniquely determined by its data $T(\phi)|_D$ restricted to D. In summary, in the n-th iteration step of the IRNM, we consider $T'[\phi_n] : X \to Y_n$ as a linear mapping from $X = L^r(\Omega)$ with $r \in [1, 2)$ to $Y_n \in \left\{L^2_{W_n}(D), L^2(D)\right\}$, or by using the embedding $W^{1,r}(\Omega) \hookrightarrow L^{\frac{2}{2-r}}(\Omega) \subseteq L^r(\Omega)$, we choose $W^{1,r}(\Omega)$ as the preimage space X. Accordingly, the IRNM reads as

$$\phi_{n+1} = \operatorname*{argmin}_{\phi \in X} \mathbb{KL}\left(y^\delta \,;\, T(\phi_n) + T'[\phi_n](\phi - \phi_n)\right) + \frac{\alpha_n}{2}\|\phi\|_X^2, \tag{1.29}$$

where

- X is a Banach space,
- Y is a Banach space, (usually finite-dimensional)
- $T'[\phi_n] : X \to Y$ is a linear operator,
- and $(\alpha_n)_{n\in\mathbb{N}} \geq 0$ is an appropriate sequence of regularization parameters (cf. Section 3.1).

Now that we have derived a promising method, the question arises of how the inner minimization problem (1.29) in each iteration step of IRNM can be solved. Since kl and $\frac{1}{2}\|\cdot\|_X^2$ are convex functions, it is a convex optimization problem which excludes local minima of the Tikhonov-type functional $\phi \mapsto kl\left(y^\delta \,;\, T(\phi_n) + T'[\phi_n](\phi - \phi_n)\right) + \frac{\alpha_n}{2}\|\phi\|_X^2$ (cf. Section 3.2). So, here we find, to some extent, a link to the Gerchberg-Saxton-Fienup-type algorithms based on convex optimization methods.

In order to present a good alternative to these methods, we want use an algorithm which is similar flexible with respect to modeling priori constraints. Due to this requirements so-called proximal-type algorithms are particular suitable for solving the minimization problem (1.29). One finds a wide class of first-order proximal algorithms in the literature for solving the convex problem (1.29), e.g. FISTA [11], ADMM [18], proximal splitting algorithms [25]. We also commend [15, 61] for an overview. However, these methods usually assume X and Y_n to be Hilbert spaces. Therefore, in Chapter 4, we will generalize one of the most common representatives of this group from a Hilbert space to a Banach space setting.

1.2 Phase retrieval in inverse medium scattering

Now we come to another phase retrieval problem where a Banach space setting is more appropriate than a Hilbert space setting: A nonlinear inverse medium scattering problem with 'sparse' contrast as studied in [52]. As opposed to the phase retrieval problems introduced in the previous sections, we consider a more precise model without the projection and Fresnel approximations. Moreover, we aim to reconstruct the whole refractive index n, not just line integrals of n. We also refer to [24, 41] for a detailed introduction to this problem. The problem is given in $m = 2, 3$-dimensions. The considered model describes time-harmonic acoustic waves scattered by an inhomogeneous medium with constant density. For $m = 2$ it describes the scattering of transverse-magnetic (TM) polarized electromagnetic waves scattered by cylindrical, penetrable, isotropic, non-magnetic structures. It is also an approximate model for time-harmonic electromagnetic scattering in $m = 3$ space dimensions for weak and slowly varying inhomogenieties, which is typically the case for x-ray frequencies. Similar the previous sections, the problem consists in retrieving information on the refractive index of some unknown medium from measurements of scattered fields. The setting is as follows: A scattering object of interest is located in a ball $B_\rho = \{x \in \mathbb{R}^m \mid \|x\|_{l^2} \leq \rho\}$ with radius $\rho > 0$ to which we sent successively time-harmonic waves fields $u^i_{d\in D}$ from different directions $\{d \in D\}$ each solving the Helmholtz equation

$$\Delta u^i + \kappa^2\, u^i = 0 \tag{1.30}$$

in m-dimensions. Here one may think of plane waves $u^i_d(x) = \exp(-\mathrm{i}\,\kappa\, x{\cdot}d)$ propagating in direction $-d \in \mathbb{S}^{m-1} := \left\{d \in \mathbb{R}^2 \mid \|d\|_{l^2} = 1\right\}$. As in Section 1.1, κ denotes the wave number and $n : \mathbb{R}^m \to \mathbb{C}$ is the refractive index coinciding with 1 outside of the object's support, in particular on $\mathbb{R}^m \setminus B_\rho$. By $n^2 = 1 + a$ we rewrite n in terms of the contrast $a : \mathbb{R}^m \to \mathbb{C}$ which we assume to have a "sparse" support within B_ρ and nonnegative imaginary part $\Im(a)$. Thus, the Banach space $X = L^r(B_\rho)$ with Lebesgue index $r > 1$ smaller than 2 would be a problem adapted choice for the solution space. Considering an incident field u^i, the resulting total field u obeys the Helmholtz equation

$$\Delta u(x) + \kappa^2\, n^2(x)\, u(x) = 0, \qquad x \in \mathbb{R}^m, \tag{1.31}$$

and the scattered field $u^s = u - u^i$ satisfies the *Sommerfeld's radiation condition*

$$\lim_{r=\|x\|_{l^2}\to\infty} r^{\frac{m-1}{2}} \left(\frac{\partial u^s}{\partial r} - \mathrm{i}k u^s\right) = 0, \qquad \text{uniformly in all directions } \frac{x}{\|x\|_{l^2}} \in \mathbb{R}^m. \tag{1.32}$$

In practice, usually not the scattered field u^s but only the amplitude of its far field pattern u^∞ which is given via the asymptotic

$$u^s(x) = \frac{\exp(\mathrm{i}\,\kappa\, r)}{r^{\frac{m-1}{2}}} \left(u^\infty\left(\frac{x}{r}\right) + O\left(\frac{1}{r}\right)\right), \qquad r := \|x\|_{l^2} \to \infty$$

can be measured. In order to compensate this lack of information, one uses incident fields u^i_d in the from of plane waves $u^i_d(x) = \exp(-\mathrm{i}\,\kappa\, x \cdot d)$ from (almost) all directions $d \in \mathbb{S}^{m-1}$. So, for fixed u^i_d and $d \in \mathbb{S}^{m-1}$ the forward operator $T_{u^i_d}$ maps a to $|u^\infty_d|^2$. We also introduce the operator $F_{u^i_d}$ mapping a to the whole far field u^∞_d which is more common in the literature.

Under appropriate conditions one can show that the field $u = u^s + u^i$, caused by some arbitrary entire solution u^i of (1.30), solves the scattering problem (1.31)-(1.30) if and only if u is a solution to the *Lippmann-Schwinger equation*

$$u(x) + \kappa^2 V(au)(x) = u(x) + \kappa^2 \int_{B_\rho} \Phi(x,z)\, a(z)\, u(z)\, dz = u^i(x), \qquad x \in \mathbb{R}^m, \tag{1.33}$$

or equivalently $u^s = u - u^i$ is a solution to

$$u^s(x) - \kappa^2 V(au^s)(x) = \kappa^2 V(au^i)(x), \qquad x \in \mathbb{R}^m, \tag{1.34}$$

where

$$V(\varphi)(x) := \int_{\mathbb{R}^2} \Phi(x,z)\, \varphi(z)\, dz$$

denotes the volume potential and

$$\Phi(x,z) := \begin{cases} \frac{\mathrm{i}}{4} H_0^{(1)} (\kappa \, \|x-z\|_{l^2}) & m=2 \\ \frac{1}{4\pi} \frac{\exp(\kappa \, \|x-z\|_{l^2})}{\|x-z\|_{l^2}} & m=3 \end{cases}$$

is the fundamental solution of the Helmholtz equation. In the considered case $a \in L^r(B_\rho)$, with $r > 1$ and $\Im(a) \geq 0$, Lechleiter et al proved in [52] that for any incident field $u^i \in L^s(B_\rho)$, $s > \frac{r}{r-1}$, which satisfies (1.30), the Lippmann-Schwinger equations (1.33) and (1.34) are uniquely solvable in $L^s(B_\rho)$. Moreover, let $v \in L^s(B_\rho)$ be the solution of (1.34). Then $u^s := \kappa^2 V(a(v + u^i))$ belongs to

$$W^{2,\frac{rs}{r+s}}_{\mathrm{loc}}(\mathbb{R}^m) := \left\{ v : \mathbb{R}^2 \to \mathbb{C} \,|\, v \in W^{2,\frac{rs}{r+s}}(B_R) \text{ for all } R > 0 \right\} \hookrightarrow L^s_{\mathrm{loc}}(\mathbb{R}^m)$$

and it is a solution to the Helmholtz equation $\Delta u^s + \kappa^2 n^2 u^s = -\kappa^2\, a\, u^i$ in $L^s_{\mathrm{loc}}(\mathbb{R}^m)$ subject to Sommerfeld's radiation condition. Multiplying both sides of Equation (1.33) by a yields the following operator equation (cf. [41, Eq. (5)])

$$a\, u = (I + \kappa^2 a\, V)^{-1} (a\, u^i).$$

We assume $a\, u \in L^2(B_\rho)$ (which is assured e.g. if $r \leq 2$) and introduce the operator

$$E : L^2(B_\rho) \to L^2(\mathbb{S}^1), \quad E(v)(\vartheta) = \int_{B_\rho} \exp(-\mathrm{i}\, \kappa \vartheta \cdot z)\, v(z)\, dz,$$

which defines the far field pattern u^∞ corresponding to $a\, u$ via $u^\infty = -\kappa^2 \gamma_m E(a\, u)$ for $\gamma_2 = \frac{\exp(\frac{\mathrm{i}\pi}{4})}{\sqrt{8\pi\kappa}}$ and $\gamma_3 = \frac{1}{4\pi}$. Then the nonlinear operator F_{u^i} reads as

$$F_{u^i}(a) = -\kappa^2 \gamma_m E\left((I + \kappa^2 a\, V)^{-1} (a\, u^i) \right)$$

and also gives $T_{u^i}(a) = |F_{u^i}(a)|$. Moreover, it can be shown ([41, Proposition 2.1]) that

$$F_{u^i} : D(F) := \left\{ a \in L^\infty(B_\rho) \,|\, (I + \kappa^2\, aV)^{-1} \text{ is boundedly invertible} \right\} \to L^2(\mathbb{S}^{m-1})$$

is Fréchet differentiable for fixed $u^i \in L^2(\mathbb{R}^2)$ with

$$F'_{u^i}[a](h) = -\kappa^2\, \gamma_m E\left((I + \kappa^2 a\, V)^{-1}(h\, u)\right).$$

Consequently, we have $T'_{u^i}[a](h)(\vartheta) = 2\Re\left(\overline{F_{u^i}(a)(\vartheta)} F'_{u^i}[a](h)(\vartheta)\right)$ for any $a, h \in D(F)$ and $\vartheta \in \mathbb{S}^{m-1}$. Note that both the operator T and its Fréchet derivative require the solution of a Lippmann-Schwinger equation for which Vainikko ([79]) proposed a fast solution method.

Also in this case the IRNM

$$a_{n+1} = \operatorname*{argmin}_{a \in X} S\left(\left(u_d^{\infty,\delta}\right)_{d\in D}; \left(T_{u_d^i}(a_n) + T'_{u_d^i}[a_n](a - a_n)\right)_{d\in\mathbb{S}^{m-1}}\right) + \alpha_n\, R(a), \tag{1.35}$$

is an attractive method for solving the ill-posed operator equation $\left(T_{u_d^i}(a) = \left|u_d^{\infty,\delta}\right|^2\right)_{d\in D}$. Here $D \subseteq \mathbb{S}^{m-1}$ denotes the set of incident wave directions and $\left|u_d^{\infty,\delta}\right|^2$ are the given possibly noisy intensity measurements caused by some plane wave u_d^i, $d \in D$. We also refer the reader to [41,42,46] for applications of (modified) iteratively regularized Newton-type methods to this problem (with non-sparse solutions). In order to promote sparse solutions, $R(a) = \|a\|_{L^1}$ or $R(a) = \frac{1}{2}\|a\|_X^2$ are appropriate choices for the penalty term. So, we have to minimize a quite general Tikhonov-type functional in each step of the IRNM. Also for this purpose proximal-type algorithms are particular suitable. Reconstructing sparse solutions with the help L^1-penalization is a commonly used approach and will be considered in more detail for a "simpler" problem.

2 Validity of the empty beam correction in near field imaging

This chapter addresses the problem that all representations of a forward operator for the introduced phase retrieval problems involve the illumination ι_0 in the object plane. More precisely, due to the assumed projection approximation $u_0 \approx \iota_0 \, O(\phi)$ the empty beam field ι_0 is related to the object function O in such a way that for the aim of retrieving the unknown object information ϕ from the recorded intensities $|u(\cdot, \Gamma)|^2 \approx |\mathcal{D}_\Gamma u_0|^2$ in the detector plane $\mathbb{P}_\Gamma$ the knowledge of ι_0 is essential. However, in particular in a near field setting already small imperfections in the waveguide system often cause strong deviations from the idealized assumption ι_0 to be the spherical wave (1.17) or equivalently the plane wave $\iota(\cdot, x_3) = \exp(\mathrm{i}\,\kappa x_3)$ in the effective geometry. See e.g. [6] for the negative influence of mirror figure errors in a Kirkpatrick-Baez (KB) mirror system. Further experimental illustrations are given in Figures 2.1 and 2.2. So, in a general near field setting (indicated by a sufficiently large product $M\mathfrak{f}$) the function ι_0 especially depends on the concrete experimental setup such that it often does not comply with a generalized formula. A common approach [33,34] to deal with this problem is to take a further intensity measurement of the empty beam field $I_{\iota_0} \approx |\mathcal{D}_\Gamma \iota_0|^2$ and then to approximate the intensity of the object function O (in the effective geometry) by the quotient of the two measurements

$$\frac{|u(\cdot, \Gamma)|^2}{I_{\iota_0}}.$$

So, the intensity data of the product $u_0 \approx \iota_0 \, O(\phi)$ is here approximated by the product of the intensities of each factor ι_0 and $O(\phi)$. However, because the Fresnel approximation is based on Fourier transforms, one expects the product approximation (1.10) in the object plane to result in a convolution in the detector plane. For this reason, we also call this empty beam correction *product approximation in the detector plane*. This approach is motivated by the fact that it actually holds if ι is given by an ideal point source (see [33, 34]) and has the huge advantage that the difficult modeling or reconstruction of ι_0 is avoided.

The following sections build up on the work that has been published by us in [36,48]. We study the validity of the product approximation in the detector plane in case of extended source sizes by providing a rigorous error estimate which also identifies the relevant experimental parameters. Moreover, in Section 2.3 we verify our conditions by numerical simulations. The physical experiments in this chapter were conducted by A.-L. Robisch, J. Hagemann, and T. Salditt with the help of M. Bartels, M. Krenkel, and C. Olendrowitz, all from the Institute for X-ray Physics, Göttingen.

2.1 Motivation

Let us start with two experimental results illustrating the influence of the concrete experimental setup on the validity of this product approximation in the detector plane.

In the first example, shown in Figure 2.1, the distortions of the inhomogeneous illumination $\iota : \mathbb{R}^3 \to \mathbb{C}$ which are caused by imperfections within the focusing system still occur after the division of the measured intensities in presence of the object by the intensities of the empty beam field. Here the sample is given by a *C. elegans* nematode, prepared by high-pressure freezing and epon-embedding, placed on a 1 mm thick glass [59]. The experiment was carried out at the Göttingen Instrument for Nano Imaging with coherent X-rays (GINIX), which is installed at the undulator beamline P10 of the PETRA III storage ring of the Deutsches Elektronen SYnchrotron (DESY). For previous cone-beam reconstructions of the same object we refer to [49, 69]. The worm is partly illuminated (near the end) by a monochromatic 7.9 keV beam (bandwidth 0.01%). A Kirkpatrick-Baez (KB) mirror system focuses the beam horizontally and vertically to a source size of 422×185 nm (FWHM, horz x vert) with a distance $R = 190$ mm to the object plane. The holograms, shown in Fig. 2.1 (a) and (b), were taken at a distance $R + \Gamma = 5.4$ m by a scintillator-microscope-based (Optique Peter) detector equipped with a 20 µm LuAG:Ce scintillator (Crytur) using a PCO2000 CCD (2048 by 2048 pixels) in combination with a 4-fold magnifying objective (Olympus) at accumulation time of 2 s. So, we have a magnification of $M = 28.4$. The camera's physical pixel size of 7.4 µm leads to pixel sizes of 1.85 µm in the detector plane and 65 nm in the object plane. Comparing the recorded intensity measurements (see Fig. 2.1 (a) and (b)) for the object and the empty beam to the result (Fig. 2.1 (c)) of the division of both, we see that the strong artifacts caused by both mirror imperfections and dirt on different vacuum windows (KB chamber, flight tube) not only occur in the holograms (a) and (b) but still clearly appear after the empty beam correction in the area of the worm. Although in the less interesting region which carries no object information the artifacts seem to 'divide out', the described approximation is not applicable in this example.

On the other hand, [59] provides examples for the empty beam correction performing well. Besides cleaner windows and optimization of parameters like the photon energy or the sample distance R the main reason for the successive application there is probably the use of a clean-up pinhole which compactifies the source size.

The influence of the source size is also studied by the second experimental example that is shown in Figure 2.2. Here, in order to further reduce the source size generated by the KB mirror system above, a lithographic bonded silicon channel [34, 35] with a length of 1mm and a cross-section size of 90×70 nm (horz x vert) is used. This in fact leads to a source size smaller than the samples finest feature size of 50 nm where a Siemens star test pattern (model ATN/XRESO-50HC, NTT-AT) consisting of a 200-nm-thick tantalum layer serves as sample. It was positioned at $R = 19.8$ mm behind the waveguide exit. Compared to the previous example we now obtain for the empty beam a hologram, shown in Figure 2.2 (a), which comes very close to that of an ideal point source illumination. Also the intensity data in presence of the sample and the result of the empty beam correction (see Figure 2.2 (b), (c)) look much cleaner so that in this case the product approximation in the detector plane seems to be suitable. Note that due to the small source size all the structures of the sample clearly occur in the holograms 2.2 (b) and (c). On the other hand, by causing distortion through inserting a wavefront modifyer which consists of 2 µm thick vertical stripes fabricated in Tungsten a few mm behind the waveguide, we again obtain strongly

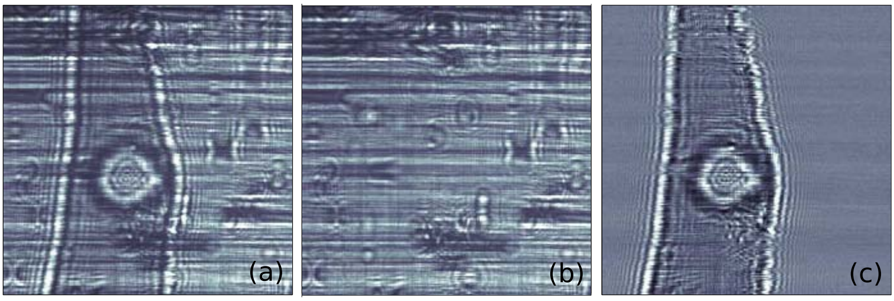


Figure 2.1: Illustration of the described empty beam correction in a near field regime: The intensity data of the pure object function (given by a part of a C. elegans, a transparent nematode) is approximated by the quotient of the intensity data in presence of the worm (a) and the intensity measurements of the empty beam (b). It can be seen that the intensity data in (a) is highly distorted due to strong artifacts in the empty beam (b) caused by mirror imperfections and other spurious elements such as dirt on vacuum windows. However, strong artifacts still appear in the region with object information even after the division by the empty beam. So, the product approximation in the detector plane is not satisfactory in this example. The square image size corresponds to 133 µm in the sample plane.

aberrated holograms not only for the empty beam and the object but also after the empty beam division (see Figure 2.2 (d), (e), (f)).

Before studying the empty beam correction's validity for extended source sizes, let us formulate this product approximation in the detector plane in mathematical terms and give, in analogy to [33, 34], its motivation for ideal point source/plane wave illumination. More generally, we show that the approximation is exact in this case not only for the measured intensities but also for the propagated fields. Recall from Section 1.1.3 that if the illumination ι is obtained by a perfect point source in $(0, 0, -R) \in \mathbb{P}_{-R}$ we can assume ι_0 to be given as:

$$\iota_0(\mathbf{x}') = \mathcal{D}_R\, \delta_0(\mathbf{x}') = \frac{-\mathrm{i}\,\kappa}{2\pi\, R} \mathrm{e}^{\mathrm{i}\kappa R} \chi_{\frac{\kappa}{R}}(\mathbf{x}'), \qquad \mathbf{x}' \in \mathbb{R}^2$$

and analogously in the detector plane $\mathbb{P}_\Gamma$ we have (cf. Remark 1.1.4):

$$\iota(\mathbf{x}', \Gamma) = \mathcal{D}_\Gamma\, \iota_0(\mathbf{x}') = \mathcal{D}_{R+\Gamma}\, \delta_0(\mathbf{x}') = \frac{-\mathrm{i}\,\kappa}{2\pi\,(R+\Gamma)} \mathrm{e}^{\mathrm{i}\kappa(R+\Gamma)} \chi_{\frac{\kappa}{R+\Gamma}}(\mathbf{x}'), \quad \mathbf{x}' \in \mathbb{R}^2.$$

Using the projection approximation (1.10) as well as $\chi_{\frac{\kappa}{M\Gamma}}(\mathbf{x}') = \chi_{\frac{M\kappa}{\Gamma}}\left(\frac{\mathbf{x}'}{M}\right)$ for $\mathbf{x}' \in \mathbb{R}^2$ we end up with an exact product approximation in terms of the Fresnel approximation:

$$\begin{aligned}
\mathcal{D}_\Gamma\, u_0\,(\mathbf{x}') &= \left(\frac{-\mathrm{i}\kappa}{2\pi}\right)^2 \frac{1}{R\,\Gamma} \mathrm{e}^{\mathrm{i}\kappa(R+\Gamma)} \left(\mathcal{F}\,\chi_{\frac{M\kappa}{\Gamma}} \mathcal{O}\right)\left(\frac{\kappa}{\Gamma}\mathbf{x}'\right) \\
&= \left(\frac{-\mathrm{i}\kappa}{2\pi\,(R+\Gamma)} \mathrm{e}^{i\kappa(R+\Gamma)} \chi_{\frac{\kappa}{R+\Gamma}}(\mathbf{x}')\right)\left(\frac{-\mathrm{i}\kappa\, M}{2\pi\,\Gamma} \chi_{\frac{\kappa}{M\Gamma}}(\mathbf{r}')\, \mathcal{F}\left(\chi_{\frac{\kappa M}{\Gamma}}\, \mathcal{O}\right)\left(\frac{\kappa}{\Gamma}\mathbf{x}'\right)\right) \qquad (2.1) \\
&= \mathrm{e}^{\mathrm{i}\kappa\frac{\Gamma}{M}}\, \mathcal{D}_\Gamma\, \iota_0\,(\mathbf{x}')\;\, \mathcal{D}_{\frac{\Gamma}{M}} \mathcal{O}\left(\frac{\mathbf{x}'}{M}\right) \qquad\quad \mathbf{x}' \in \mathbb{R}^2.
\end{aligned}$$

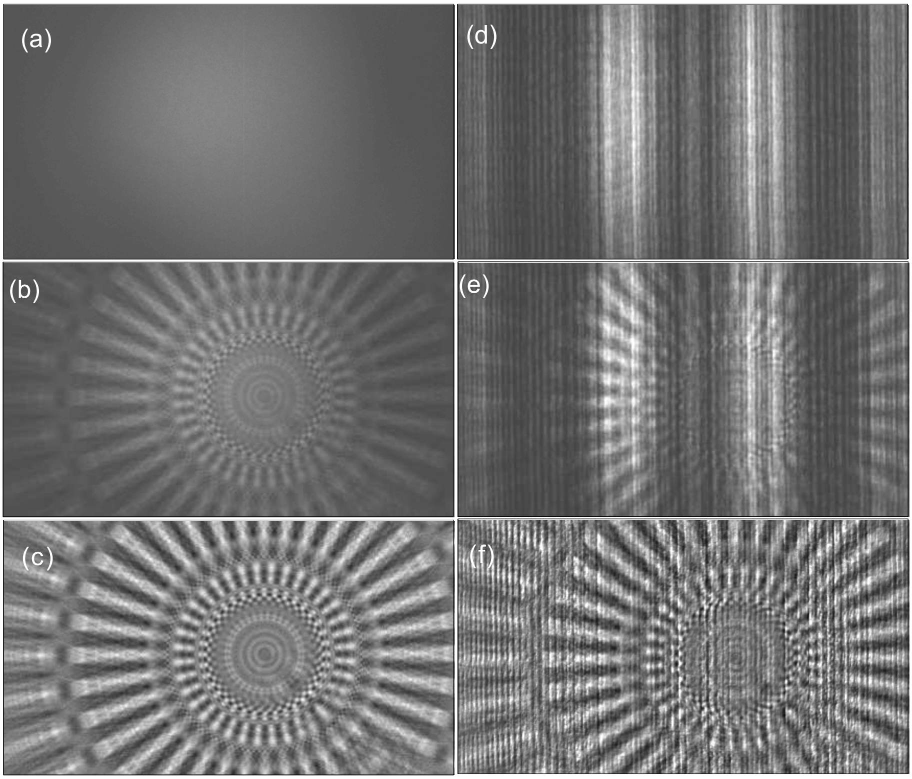


Figure 2.2: Illustration of the described empty beam correction with a fine structured Siemens star as test pattern. While in the left column the holograms are given for a nearly ideal spherical or parallel beam (depending on the geometry concerned) in the right column the beam is deliberately scrambled by a wavefront modifyer in form of W stripes positioned a few mm behind the waveguide. From top to bottom the images show the empty beam hologram ((a), (d)), intensity data of the sample ((b),(e)), and the result of empty beam correction ((c),(f)), respectively. One pixel corresponds to 25.5 nm in the object plane.

Note that $\mathcal{D}_{\frac{\Gamma}{M}} O\left(\frac{\mathbf{x}'}{M}\right) = \mathcal{D}_{\Gamma_{\mathit{eff}}} O\left(\mathbf{x}'_{\mathit{eff}}\right)$ describes the propagated object function O in the effective geometry $(\mathbf{x}'_{\mathit{eff}}, \Gamma_{\mathit{eff}}) \in \frac{1}{M}\mathbb{P}_{\Gamma_{\mathit{eff}}}$ as introduced in Section 1.1.3. By taking the modulus square of (2.1) we obtain the validity of the product approximation for the detected intensities which is investigated here:

$$\left|\mathcal{D}_{\Gamma}\, u_0(\mathbf{x}')\right|^2 = \left|\mathcal{D}_{\Gamma}\, \iota_0(\mathbf{x}')\right|^2 \left|\mathcal{D}_{\frac{\Gamma}{M}}\, O\left(\frac{\mathbf{x}'}{M}\right)\right|^2 \qquad \mathbf{x}' \in \mathbb{R}^2. \tag{2.2}$$

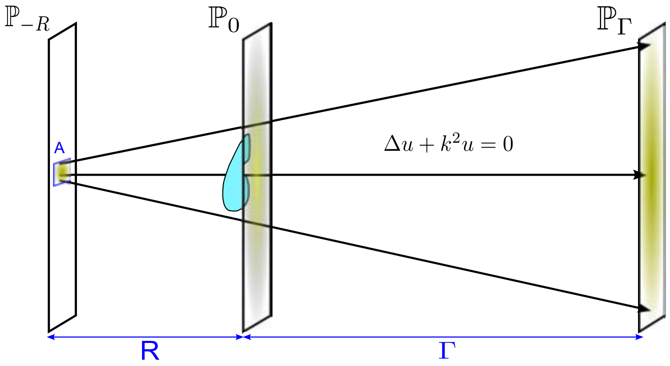


Figure 2.3: Illustration of the assumed setup with an extended x-ray source size A.

2.2 Error estimate

Now our aim is to gain deeper insight on the validity of (2.2) by considering more general illumination functions ι_0. As in particular the source size seems to be a critical parameter, we study this prediction by the following assumption:

Assumption 2.2.1. *Let ι_0 be generated by the extended source $A = [-a, a]^2 \times \{-R\} \in \mathbb{P}_{-R}$ with diameter $a \geq 0$ and $R > 0$ in the form that it is the Fresnel approximation of a* source function $\omega_A : \mathbb{P}_{-R} \to \mathbb{C} \in L^1(\mathbb{R}^2)$ *supported in A (in the sense of Lemma* (1.1.8) *where we identify the regular distribution $T_{\omega_A} \in \mathcal{E}'$ with ω_A):*

$$\iota_0(\mathbf{x}') = \mathcal{D}_R\left(\omega_A\right)(\mathbf{x}') = \frac{-ik}{R}\, e^{i\kappa R}\, \chi_{\frac{\kappa}{R}}\left(\mathbf{x}'\right)\, \left(\mathcal{F}\left(\chi_{\frac{\kappa}{R}}\omega_A\right)\right)\left(\frac{\kappa}{R}\mathbf{x}'\right), \quad \mathbf{x}' \in \mathbb{R}^2. \tag{2.3}$$

So, ι_0 is a C^∞-function and due to [31, Lemma 8.4.1] it is also a multiplier on $S'\left(\mathbb{R}^2\right)$. Since there are delta sequences $(\delta_n)_{n\in\mathbb{N}} \subset L^1(\mathbb{R}^2)$ converging to δ_0 in $\mathcal{E}'$ as $n \to \infty$, the case of point source illumination $\iota_0 = \mathcal{D}_R\,\delta_0$ studied above is a boundary value of this assumption. Figure 2.3 sketches the setup. Furthermore, we define a dimensionless parameter depending on the object function O. For this purpose, we state in analogy to Assumption 1.1.9 that for a known constant $C \geq 0$ the function $\tilde{O} = O - C \in L^2([-\mathbf{r}_O, \mathbf{r}_O])$ is compactly supported. Then the Fourier transform $\mathcal{F}\tilde{O}$ is integrable and $\mathcal{F}\tilde{O}(\mathbf{x}')$ tends to 0 as $|\mathbf{x}'| \to \infty$, where $\tilde{O}$'s degree of smoothness determines the rate of decay. Note that due to

$$\frac{\mathcal{D}_\Gamma(\iota_0\, O)}{\mathcal{D}_\Gamma\, \iota_0} = \frac{\mathcal{D}_\Gamma(\iota_0\, \tilde{O})}{\mathcal{D}_\Gamma\, \iota_0} - C$$

studying the empty beam correction with respect to $\tilde{O}$ is as interesting as with respect to O. It might be even more practical since the reconstruction of $\tilde{O}$ from $\left|\mathcal{D}_\Gamma\, \tilde{O}\right|^2$, which also gives O, could be performed without applying $\mathcal{D}$ to a distribution but to a $L^2(\mathbb{R}^2)$-function.

Assumption and Definition 2.2.2. *Under the assumption that there is a constant* $C \geq 0$ *such that* $\tilde{O} := O - C \in L^2(\mathbb{R}^2)$ *is compactly supported in a rectangle* $[-\mathbf{r}_O, \mathbf{r}_O]$ *we define for any spatial frequency* $\rho > 0$ *the finite parameter*

$$\epsilon_O\,(\rho) := \frac{\int_{\|\xi'\|_{l^1} > \rho} |(\mathcal{F}\tilde{O})(\xi')|\, d\xi'}{\int_{\mathbb{R}^2} |(\mathcal{F}\tilde{O})(\xi')|\, d\xi'}. \tag{2.4}$$

Although the compactly supported function $\tilde{O}$ can not be band-limited, we can assume $\mathcal{F}\tilde{O}$ to vanish almost outside a rectangle $\left\{\xi' \in \mathbb{R}^2 \mid \|\xi'\|_{l^1} \leq \rho_0\right\}$ with some $\rho_0 > 0$. This implies $\epsilon_O\,(\rho_0) \approx 0$. Noting that a rotation by 45 degrees turns $\left\{\xi' \in \mathbb{R}^2 \mid \|\xi'\|_{l^1} \leq \rho_0\right\}$ into $\left[-\frac{\rho_0}{\sqrt{2}}, \frac{\rho_0}{\sqrt{2}}\right]^2$, we obtain from the Appendix A.2 that ρ_0 is related via $\frac{\rho_0}{\sqrt{2}} = \pi \frac{\mathbf{N}}{2\mathbf{r}_O}$ to the grid $\left\{\frac{2\mathbf{r}_O}{\mathbf{N}}\mathbf{j} \mid \mathbf{j} \in \Delta_\mathbf{N}\right\}$ which discretizes the support $[-\mathbf{r}_O, \mathbf{r}_O]$. As the value $p_X = \frac{2\mathbf{r}_O}{\mathbf{N}}$ defines the corresponding pixel size we conclude that the parameter ρ_0 is reciprocal to the smallest relevant feature size of the object.

With these definitions our main result with respect to empty beam correction reads as follows:

Theorem 2.2.3. *Under the assumptions* (2.2.1) *and* (2.2.2) *the dimensionless parameters* ϵ_O *given by Equation (2.4) and*

$$\epsilon_A(\rho) := \frac{\Gamma}{R\,M}\,a\,\rho = \frac{\Gamma}{R+\Gamma}\,a\,\rho, \quad \rho > 0$$

determine the validity of the product approximation in the detector plane $\mathbb{P}_\Gamma$

$$\mathcal{D}_\Gamma\left(\iota_0\,\tilde{O}\right)(\xi') \approx e^{-i\kappa\frac{\Gamma}{M}}\,\mathcal{D}_\Gamma\,\iota_0\,(\xi')\,\mathcal{D}_{\frac{\Gamma}{M}}\tilde{O}\left(\frac{1}{M}\xi'\right), \quad \xi' \in \mathbb{R}^2 \tag{2.5}$$

by the error estimate:

$$\frac{\left|\mathcal{D}_\Gamma\left(\iota_0\,\tilde{O}\right)(\xi') - e^{-i\kappa\frac{\Gamma}{M}}\,\mathcal{D}_\Gamma\iota_0(\xi')\,\mathcal{D}_{\frac{\Gamma}{M}}\tilde{O}\left(\frac{1}{M}\xi'\right)\right|}{\frac{\kappa}{4\pi^2(R+\Gamma)}\left(\int_A |\omega_A(\mathbf{r}')|\,d\mathbf{r}'\right)\left(\int_{\mathbb{R}^2}|(\mathcal{F}\tilde{O})(\mathbf{y}')|\,d\mathbf{y}'\right)} \leq \inf_{\rho>0}\left(\epsilon_A\,(\rho) + 2\,\epsilon_O\,(\rho)\right). \tag{2.6}$$

Proof. First of all, let us expand the Fresnel approximation of the empty beam ι_0 as well as of the product $\iota_0\,\tilde{O} \approx u_0$ in the detector plane $\mathbb{P}_\Gamma$ where ι_0 is given by Equation (2.3). By applying Fouriers convolution theorem (A.4) (for distributions) as well as

$$\mathcal{F}\,T\,(c\mathbf{x}') = \frac{1}{c^2}\,\mathcal{F}\,T\left(\frac{\mathbf{x}'}{c}\right), \quad \mathbf{x}' \in \mathbb{R}^2, T \in S'\left(\mathbb{R}^2\right), c \in \mathbb{R},$$

the propagated illumination function reads:

$$\begin{aligned}\mathcal{D}_\Gamma(\iota_0)(\xi') &= \frac{-\kappa^2}{R\Gamma}\,\mathrm{e}^{\mathrm{i}\kappa(R+\Gamma)}\,\chi_{\frac{\kappa}{\Gamma}}(\xi')\,\mathcal{F}\Big(\chi_{\frac{\kappa}{R}+\frac{\kappa}{\Gamma}}\,\Big(\mathcal{F}\big(\chi_{\frac{\kappa}{R}}\omega_A\big)\Big(\frac{\kappa}{R}\cdot\Big)\Big)\Big)\Big(\frac{\kappa}{\Gamma}\xi'\Big)\\ &= \frac{-k}{\Gamma}\,\mathrm{e}^{\mathrm{i}\kappa(R+\Gamma)}\,\chi_{\frac{\kappa}{\Gamma}}(\xi')\Big(\mathcal{F}\big(\chi_{\frac{\kappa}{R}+\frac{\kappa}{\Gamma}}\big) * \chi_{\frac{\kappa}{R}}\omega_A\Big(-\frac{R}{\kappa}\cdot\Big)\Big)\Big)\Big(\frac{\kappa}{\Gamma}\xi'\Big) \quad \xi' \in \mathbb{R}^2.\end{aligned} \tag{2.7}$$

With the help of Lemma 1.1.3 we obtain for the Fourier transform of the chirp function $\chi_{\frac{\kappa}{R}+\frac{\kappa}{\Gamma}} = \chi_{\frac{\kappa}{\Gamma}M}$ an expansion analog to (1.5):

$$\mathcal{F}\chi_{\frac{\kappa}{\Gamma}+\frac{\kappa}{R}}\left(\frac{\kappa}{\Gamma}\xi' - \mathbf{z}' - \mathbf{y}'\right) = \frac{-\mathrm{i}\,k\,M}{\Gamma}\mathcal{F}\chi_{\frac{\kappa}{\Gamma}+\frac{\kappa}{R}}\left(\frac{\kappa}{\Gamma}\xi' - \mathbf{z}'\right)\,\mathrm{e}^{\mathrm{i}\frac{\Gamma}{\kappa M}\mathbf{y}'\cdot\left(\frac{\kappa}{\Gamma}\xi'-\mathbf{z}'\right)}\,\mathcal{F}\chi_{\frac{\kappa}{\Gamma}+\frac{\kappa}{R}}(\mathbf{y}')$$

for all $\xi', \mathbf{y}', \mathbf{z}' \in \mathbb{R}^2$. Substituting this in the Fresnel propagation of $\iota_0\,\tilde{O}$ we end up with

$$\begin{aligned}
&\mathcal{D}_\Gamma\left(\iota_0\,\tilde{O}\right)(\xi') = \frac{-\kappa^2}{R\,\Gamma}\,\mathrm{e}^{\mathrm{i}\kappa(R+\Gamma)}\,\chi_{\frac{\kappa}{\Gamma}}(\xi')\,\mathcal{F}\left(\tilde{O}\,\chi_{\frac{\kappa}{\Gamma}+\frac{\kappa}{R}}\,\left(\mathcal{F}\left(\chi_{\frac{\kappa}{R}}\omega_A\right)\left(\frac{\kappa}{R}\cdot\right)\right)\right)\left(\frac{\kappa}{\Gamma}\xi'\right)\\
&= \frac{-R}{4\pi^2\,\Gamma}\,\mathrm{e}^{\mathrm{i}\kappa(R+\Gamma)}\,\chi_{\frac{\kappa}{\Gamma}}(\xi')\,\left(\mathcal{F}\left(\tilde{O}\right) * \mathcal{F}\left(\chi_{\frac{\kappa}{\Gamma}+\frac{\kappa}{R}}\right) * \left(\chi_{\frac{\kappa}{R}}\omega_A\right)\left(-\frac{R}{\kappa}\cdot\right)\right)\left(\frac{\kappa}{\Gamma}\xi'\right)\\
&= \frac{-R}{4\pi^2\,\Gamma}\,\mathrm{e}^{\mathrm{i}\kappa(R+\Gamma)}\,\chi_{\frac{\kappa}{\Gamma}}(\xi')\,\int_{\mathbb{R}^2}\int_{\mathbb{R}^2}\mathcal{F}\tilde{O}(\mathbf{y}')\,\mathcal{F}\chi_{\frac{\kappa}{\Gamma}M}\left(\frac{\kappa}{\Gamma}\xi' - \mathbf{z}' - \mathbf{y}'\right)\left(\chi_{\frac{\kappa}{R}}\omega_A\right)\left(-\frac{R}{\kappa}\mathbf{z}'\right)\,d\mathbf{y}'\,d\mathbf{z}'\\
&= \frac{\mathrm{i}\,\kappa\,R\,M}{4\pi^2\,\Gamma^2}\,\mathrm{e}^{\mathrm{i}\kappa(R+\Gamma)}\,\chi_{\frac{\kappa}{\Gamma}}(\xi') \qquad (2.8)\\
&\int_{\mathbb{R}^2}\int_{\mathbb{R}^2} k(\mathbf{y}', \xi', \mathbf{z}')\,\mathcal{F}\tilde{O}(\mathbf{y}')\,\mathcal{F}\chi_{\frac{\kappa}{\Gamma}M}(\mathbf{y}')\,\mathcal{F}\chi_{\frac{\kappa}{\Gamma}M}\left(\frac{\kappa}{\Gamma}\xi' - \mathbf{z}'\right)\,\left(\chi_{\frac{\kappa}{R}}\omega_A\right)\left(-\frac{R}{\kappa}\mathbf{z}'\right)\,d\mathbf{y}'\,d\mathbf{z}' \qquad (2.9)
\end{aligned}$$

where

$$k(\xi', \mathbf{y}', \mathbf{z}') := \mathrm{e}^{\mathrm{i}\frac{\Gamma}{\kappa M}\mathbf{y}'\cdot\left(\frac{\kappa}{\Gamma}\xi'-\mathbf{z}'\right)}.$$

This step is justified by equation (A.4), [31, Remark p. 103] and Lemma 1.1.8. Now we observe that the two variables $\mathbf{y}'$ and $\mathbf{z}'$ are only coupled by the first factor in line (2.9) that is given by the function k. Therefore, we consider the case of $\epsilon_A(\rho) = \frac{\Gamma a \rho}{R M} \ll 1$ and $\epsilon_O(\rho) \ll 1$ which yields the approximation

$$\mathrm{e}^{\mathrm{i}\frac{\Gamma}{\kappa M}\mathbf{y}'\cdot\left(\frac{\kappa}{\Gamma}\xi'-\mathbf{z}'\right)} \approx \mathrm{e}^{\mathrm{i}\frac{1}{M}\mathbf{y}'\cdot\xi'}, \quad \text{for } \frac{R}{\kappa}\mathbf{z}' \in A = [-a, a]^2,\ \|\mathbf{y}'\|_{l^1} = |y_1| + |y_2| \le \rho. \qquad (2.10)$$

In fact, this uncoupling together with (2.7) and the representation

$$\mathcal{D}_{\frac{\Gamma}{M}}\tilde{O}\left(\frac{1}{M}\xi'\right) = \frac{-\mathrm{i}\,R\,M}{2\pi\,\Gamma}\left(\tilde{O} * \chi_{\frac{\kappa}{\Gamma}M}\right)\left(\frac{1}{M}\xi'\right) = \frac{-\mathrm{i}\,R\,M}{2\pi\,\Gamma}\mathcal{F}^{-1}\left(\mathcal{F}\tilde{O}\,\mathcal{F}\chi_{\frac{\kappa}{\Gamma}M}\right)\left(\frac{1}{M}\xi'\right) \qquad (2.11)$$

lead to a product approximation in $\mathbb{P}_\Gamma$:

$$\begin{aligned}
&\mathcal{D}_\Gamma\left(\iota_0\,\tilde{O}\right)(\xi')\\
&\approx \frac{\mathrm{i}\,\kappa\,R\,M}{2\pi\,\Gamma^2}\,\mathrm{e}^{\mathrm{i}\kappa(R+\Gamma)}\,\chi_{\frac{\kappa}{\Gamma}}(\xi')\,\mathcal{F}^{-1}\left(\mathcal{F}\tilde{O}\,\mathcal{F}\chi_{\frac{\kappa}{\Gamma}M}\right)\left(\frac{1}{M}\xi'\right)\,\left(\mathcal{F}\left(\chi_{\frac{\kappa}{\Gamma}M}\right) * \chi_{\frac{\kappa}{R}}\omega_A\left(-\frac{R}{\kappa}\cdot\right)\right)\left(\frac{\kappa}{\Gamma}\xi'\right)\\
&= \mathrm{e}^{-\mathrm{i}\kappa\frac{\Gamma}{M}}\,\mathcal{D}_{\frac{\Gamma}{M}}\tilde{O}\left(\frac{1}{M}\xi'\right)\,\mathcal{D}_\Gamma\left(\iota_0\right)(\xi')\,.
\end{aligned}$$

Now we take this approach as a guideline to estimate the pointwise error

$$\epsilon(\xi') := \left|\mathcal{D}_\Gamma\left(\iota_0\,\tilde{O}\right)(\xi') - \mathrm{e}^{-\mathrm{i}\kappa\frac{\Gamma}{M}}\,\mathcal{D}_{\frac{\Gamma}{M}}\tilde{O}\left(\frac{1}{M}\xi'\right)\,\mathcal{D}_\Gamma\iota_0(\xi')\right|.$$

Inserting (2.7), (2.9), and (2.11) into this pointwise error and using the the identities $\left|\mathcal{F}\chi_{\frac{\kappa}{\Gamma}M}\right| \equiv \frac{\Gamma}{\kappa M}$ and $\int_{(\kappa/R)A} \left|\omega_A\left(-\frac{R}{\kappa}\mathbf{z}'\right)\right| d\mathbf{z}' = (\kappa/R)^2 \int_A |\omega_A(\mathbf{r}')|\, d\mathbf{r}'$ we obtain:

$$\begin{aligned}
\epsilon(\xi') \le \frac{\kappa R M}{4\pi^2\,\Gamma^2} \quad & \int_{\mathbb{R}^2} \left|\mathcal{F}\chi_{\frac{\kappa}{\Gamma}M}\left(\frac{\kappa}{\Gamma}\xi' - \mathbf{z}'\right)\right| \left|\left(\chi_{\frac{\kappa}{R}}\omega_A\right)\left(-\frac{R}{\kappa}\mathbf{z}'\right)\right| \\
& \int_{\mathbb{R}^2} \left|\mathrm{e}^{\mathrm{i}\frac{1}{M}\mathbf{y}'\cdot\xi'}\right| \left|\mathrm{e}^{-\mathrm{i}\frac{\Gamma}{\kappa M}\mathbf{y}'\cdot\mathbf{z}'} - 1\right| \left|\mathcal{F}\tilde{O}(\mathbf{y}')\right| \left|\mathcal{F}\chi_{\frac{\kappa}{\Gamma}M}(\mathbf{y}')\right| d\mathbf{y}'\, d\mathbf{z}' \qquad (2.12) \\
\le \frac{\kappa}{4\pi^2 M R} \quad & \left(\int_A \left|\omega_A(\mathbf{r}')\right| d\mathbf{r}'\right) \sup_{\mathbf{z}'\in(\kappa/R)A} \left(\int_{\mathbb{R}^2} \left|\mathrm{e}^{-\mathrm{i}\frac{\Gamma}{\kappa M}\mathbf{y}'\cdot\mathbf{z}'} - 1\right| \left|\mathcal{F}\tilde{O}(\mathbf{y}')\right| d\mathbf{y}'\right).
\end{aligned}$$

Moreover, for any $\rho > 0$ we have

$$\begin{aligned}
\sup_{\|\mathbf{y}'\|_{l^1}\le\rho,\, \mathbf{z}'\in(\kappa/R)A} \left|\mathrm{e}^{-\mathrm{i}\frac{\Gamma}{\kappa M}\mathbf{y}'\cdot\mathbf{z}'} - 1\right| &= \sup_{\|\mathbf{y}'\|_{l^1}\le\rho,\, \mathbf{r}'\in A} \left|\mathrm{e}^{-\mathrm{i}\frac{\Gamma}{R+\Gamma}\mathbf{y}'\cdot\mathbf{r}'} - 1\right| \le \sup_{|t|\le\epsilon_A(\rho)} \left|\mathrm{e}^{it} - 1\right| \\
&= \sup_{|t|\le\epsilon_A(\rho)} \left|\int_0^t i\mathrm{e}^{is}\,\mathrm{d}s\right| \le \sup_{|t|\le\epsilon_A(\rho)} \int_0^t \left|i\mathrm{e}^{is}\right| \mathrm{d}s = \epsilon_A(\rho)
\end{aligned}$$

and thus the second integral on the right hand side of Equation (2.12) can be bounded as follows

$$\begin{aligned}
\sup_{\mathbf{z}'\in(\kappa/R)A} \int_{\mathbb{R}^2} \left|\mathrm{e}^{-\mathrm{i}\frac{\Gamma}{\kappa M}\mathbf{y}'\cdot\mathbf{z}'} - 1\right| |\mathcal{F}\tilde{O}(\mathbf{y}')|\, d\mathbf{y}' &\le \epsilon_A(\rho) \int_{\mathbb{R}^2} |\mathcal{F}\tilde{O}(\mathbf{y}')|\, d\mathbf{y}' + 2\int_{\|\mathbf{y}'\|_{l^1}>\rho} |\mathcal{F}\tilde{O}(\mathbf{y}')|\, d\mathbf{y}' \\
&\le (\epsilon_A(\rho) + 2\epsilon_O(\rho)) \int_{\mathbb{R}^2} |\mathcal{F}\tilde{O}(\mathbf{y}')|\, d\mathbf{y}'.
\end{aligned}$$

Inserting the last estimate in (2.12) completes the proof. □

First of all, note that by the inequalities

$$\begin{aligned}
|\mathcal{D}_\Gamma\, \iota_0(\mathbf{y}')| = \left|\mathcal{D}_{R+\Gamma}\, \omega_A(\mathbf{y}')\right| &= \frac{\kappa}{2\pi(R+\Gamma)} \left|\int_{\mathbb{R}^2} \chi_{\frac{\kappa}{\Gamma}}(\mathbf{y}' - \xi')\, \omega_A(\xi')\, d\xi'\right| \\
&\le \frac{\kappa}{2\pi(R+\Gamma)} \int_A \left|\omega_A(\xi')\right| d\xi', \quad \mathbf{y}' \in \mathbb{R}^2 \qquad (2.13) \\
\left|\mathcal{D}_{\frac{\Gamma}{M}}\tilde{O}\left(\tfrac{1}{M}\mathbf{z}'\right)\right| \le \frac{1}{2\pi}\int_{\mathbb{R}^2} \left|\chi_{-\frac{\Gamma}{\kappa M}}(\xi')\, \mathcal{F}\tilde{O}(\xi')\right| d\xi' &= \frac{1}{2\pi}\int_{\mathbb{R}^2} \left|\mathcal{F}\tilde{O}(\xi')\right| d\xi', \quad \mathbf{z}' \in \mathbb{R}^2
\end{aligned}$$

we obtain a natural connection between the left hand side of our error estimate (2.6) and the relative error

$$\delta(\xi') := \frac{\left|\mathcal{D}_\Gamma\left(\iota_0\,\tilde{O}\right)(\xi') - \mathrm{e}^{-\mathrm{i}\kappa\frac{\Gamma}{M}}\, \mathcal{D}_\Gamma\, \iota_0(\xi')\, \mathcal{D}_{\frac{\Gamma}{M}}\tilde{O}\left(\frac{1}{M}\xi'\right)\right|}{\sup_{\mathbf{y}'\in\mathbb{R}^2} |\mathcal{D}_\Gamma\iota_0(\mathbf{y}')| \sup_{\mathbf{z}'\in\mathbb{R}^2} \left|\mathcal{D}_{\frac{\Gamma}{M}}\tilde{O}\left(\frac{1}{M}\mathbf{z}'\right)\right|}, \qquad \xi' \in \mathbb{R}^2. \qquad (2.14)$$

In particular, if the values a and ρ_0 or the parameters $f = \frac{\kappa}{R+\Gamma}$ and $f = \frac{\Gamma}{\kappa M}$ of the corresponding chirp function are sufficiently small these inequalities provide a good estimate of the denominator

$$d = \frac{\kappa}{4\pi^2(R+\Gamma)} \left(\int_A |\omega_A(\mathbf{r}')|\, d\mathbf{r}'\right) \left(\int_{\mathbb{R}^2} |(\mathcal{F}\tilde{O})(\mathbf{y}')|\, d\mathbf{y}'\right)$$

in Equation (2.6). So, in these cases the error (2.6) can really be seen as a relative error.

Now let us discuss the right hand side of (2.6). While obviously the source size a should be small to ensure validity of the product approximation (2.5) in the detector plane, stating a condition with respect to ρ is more complicated since the parameter $\epsilon_A(\rho)$ is increasing in ρ while $\epsilon_O(\rho)$ is decreasing. Under the assumption from above that there is a critical size ρ_0 of relevant Fourier components in $\tilde{O}$ such that $\epsilon_O(\rho_0) \approx 0$ we obtain for $\frac{\Gamma}{R+\Gamma} < 1$ the simple error bound:

$$\frac{\left|\mathcal{D}_\Gamma\left(\iota_0\,\tilde{O}\right)(\xi') - \mathrm{e}^{-\mathrm{i}\,\kappa\frac{\Gamma}{M}}\ \mathcal{D}_\Gamma\,\iota_0(\xi')\ \mathcal{D}_{\frac{\Gamma}{M}}\tilde{O}\left(\frac{1}{M}\xi'\right)\right|}{\frac{\kappa}{4\pi^2(R+\Gamma)}\left(\int_A|\omega_A(\mathbf{r}')|\,d\mathbf{r}'\right)\left(\int_{\mathbb{R}^2}|(\mathcal{F}\tilde{O})(\mathbf{y}')|\,d\mathbf{y}'\right)} \le \rho_0\,a, \qquad \xi' \in \mathbb{R}^2.$$

This implies the condition ρ_0 to be small. More precisely, in order to ensure the approximation's validity, *the source size a of the illumination should be much smaller than the size* $\Delta x_O := \frac{\sqrt{2}\pi}{\rho_0}$ *of the smallest relevant feature in the object.*

Next we state an error estimation for the measured intensities. Setting

$$r := \inf_{\rho>0}\left(\epsilon_A(\rho) + 2\,\epsilon_O(\rho)\right),$$

we derive from inequalities (2.6) and (2.13) with the help of

$$\left||a|^2 - |b|^2\right| \le |a-b|\;|a+b| \le |a-b|\;\;(2|b| + |a-b|) \qquad \text{for any } a, b \in \mathbb{C}$$

the following inequality:

$$\begin{aligned}\left|\left|\mathcal{D}_\Gamma\left(\iota_0\,\tilde{O}\right)(\xi')\right|^2 - \left|\mathcal{D}_{\frac{\Gamma}{M}}\tilde{O}\left(\frac{1}{M}\xi'\right)\right|^2\left|\mathcal{D}_\Gamma\,\iota_0(\xi')\right|^2\right| &\le d\,r\left(2\left|\mathcal{D}_{\frac{\Gamma}{M}}\tilde{O}\left(\frac{1}{M}\xi'\right)\right|\left|\mathcal{D}_\Gamma\,\iota_0(\xi')\right| + d\,r\right)\\ &\le d^2\left(2r + r^2\right).\end{aligned}$$

Moreover, this estimation can be simplified in the relevant case of $r \le 1$ by using $r^2 \le r$.

Corollary 2.2.4. *In addition to the assumptions of Theorem 2.2.3, suppose that the right hand side of* (2.6) *given by* $\inf_{\rho>0}\left(\epsilon_A(\rho) + 2\,\epsilon_O(\rho)\right)$ *is less or equal to* 1. *Then the error of the product approximation with respect to the measured intensities is bounded for all* $\xi' \in \mathbb{R}^2$ *by*

$$\frac{\left|\left|\mathcal{D}_\Gamma\left(\iota_0\,\tilde{O}\right)(\xi')\right|^2 - \left|\mathcal{D}_{\frac{\Gamma}{M}}\tilde{O}\left(\frac{1}{M}\xi'\right)\right|^2\left|\mathcal{D}_\Gamma\,\iota_0(\xi')\right|^2\right|}{\frac{\kappa^2}{16\pi^4(R+\Gamma)^2}\left(\int_A|\omega_A(\mathbf{r}')|\,d\mathbf{r}'\right)^2\left(\int_{\mathbb{R}^2}|(\mathcal{F}\tilde{O})(\mathbf{y}')|\,d\mathbf{y}'\right)^2} \le 3\inf_{\rho>0}\left(\epsilon_A(\rho) + 2\,\epsilon_O(\rho)\right). \qquad (2.15)$$

For numerical reasons we also like to state Theorem 2.2.3 in the effective geometry with coordinates (cf. Section 1.1.3)

$$\mathbf{x}'_{\mathit{eff}} := \frac{\mathbf{x}'}{M}, \qquad \Gamma_{\mathit{eff}} := \frac{\Gamma}{M}.$$

For this purpose, we rewrite the propagated empty beam field $\iota_0 = \mathcal{D}_\Gamma \omega_A$ given by (2.3) in the form of Equation (1.18):

$$\iota_0 = \chi_{\frac{\kappa}{R}}\, P, \quad \text{with} \quad P(\mathbf{x}') = \frac{-\mathrm{i}\kappa}{R} \mathrm{e}^{ikR}\, \mathcal{F}\left(\chi_{\frac{\kappa}{R}}\, \omega_A\right)\left(\frac{\kappa}{R}\mathbf{x}'\right), \quad \mathbf{x}' \in \mathbb{R}^2. \tag{2.16}$$

Due to Lemma 1.1.8 the envelope P is a $C^\infty(\mathbb{R}^2)$-function with $T_P \in S'\left(\mathbb{R}^2\right)$. Then conclude from

$$\mathcal{F}P(\xi') = \frac{-\mathrm{i}R}{\kappa} \mathrm{e}^{ikR}\, \chi_{\frac{\kappa}{R}}\, \omega_A\left(-\frac{R}{\kappa}\xi'\right), \quad \xi' \in \mathbb{R}^2 \tag{2.17}$$

that the Fourier transform of P is compactly supported in $[-b, b]^2$ with $b := \frac{\kappa a}{R}$.

Corollary 2.2.5. *In addition to Assumption 2.2.2, suppose that the illumination* $\iota_0 = \chi_{\frac{\kappa}{R}} P$ *in the object plane* $\mathbb{P}_0$ *is given by a band limited probe field* $P \in S'\left(\mathbb{R}^2\right)$ *with bandwidth* $b = \frac{\kappa a}{R}$ *and* $\mathcal{F}P \in L^1([-b, b]^2)$. *Then we have for any* $\mathbf{x}'_{\mathit{eff}} \in \frac{1}{M}\mathbb{R}^2$ *the following error estimate*

$$\frac{\left|\mathcal{D}_{\Gamma_{\mathit{eff}}}(P\tilde{O})(\mathbf{x}'_{\mathit{eff}}) - e^{-i\kappa\Gamma_{\mathit{eff}}}\mathcal{D}_{\Gamma_{\mathit{eff}}}\tilde{O}(\mathbf{x}'_{\mathit{eff}})\, \mathcal{D}_{\Gamma_{\mathit{eff}}}P(\mathbf{x}'_{\mathit{eff}})\right|}{\frac{1}{4\pi^2}\left(\int_{\mathbb{R}^2} |\mathcal{F}P(\mathbf{r}')|\, d\mathbf{r}'\right)\left(\int_{\mathbb{R}^2} |\mathcal{F}\tilde{O}(\mathbf{y}')|\, d\mathbf{y}'\right)} \leq \inf_{\rho>0}\left(\rho b \frac{\Gamma_{\mathit{eff}}}{\kappa} + 2\,\epsilon_O(\rho)\right). \tag{2.18}$$

Proof. Obviously there is an one-to-one correspondence between the probe field P given by (2.16) with $\mathcal{F}T_P \in \mathcal{E}'$ and the source function ω_A which we consider as the regular distribution $T_{\omega_A} \in \mathcal{E}'$. With the additional assumption $\omega_A, \mathcal{F}P \in L^1(\mathbb{R}^2)$, Equation (2.17) yields

$$\int_{\mathbb{R}^2} |\omega_A(\mathbf{r}')|\, d\mathbf{r}' = \frac{R}{\kappa}\int_{\mathbb{R}^2} |\mathcal{F}P(\mathbf{r}')|\, d\mathbf{r}'.$$

Now, the assertion immediately follows by inserting $a = \frac{Rb}{\kappa}$ as well as the representation (1.20) of the Fresnel propagations for $\mathcal{D}_\Gamma\, \iota_0$ and $\mathcal{D}_{z_2}(\iota_0\tilde{O})$ into Equation (2.6). □

Let us interpret the area $\left\{\xi' \in \mathbb{R}^2 \mid \|\xi'\|_{l^1} \leq \rho_0\right\}$ with diameter $\rho_0 = \frac{\sqrt{2}\pi}{\Delta x_O}$ given by the finest relevant feature size Δx_O of the object as the support of $\mathcal{F}\tilde{O}$. Then a remarkable fact of the error estimate (2.18) in near field imaging is that the roles of P and $\tilde{O}$ are completely symmetric within this inequality where one support can be transformed into the other by using a 45 degree rotation. So, analogously, we define the finest relevant feature size Δx_P of the probe P which is related via (the Nyquist sampling rate):

$$\Delta x_P = \frac{\pi}{b} = \frac{\pi R}{\kappa a}$$

to the support size a. Then we see that up to a factor $\sqrt{2}$ also the corresponding support sizes a and ρ are interchangeable in Corollary 2.2.5. Usually the parameters Δx_O and Δx_P coincide with the "correlation lengths“, i.e. the typical length scales over which P and $\tilde{O}$ vary.

2.3 Numerical results

Based on the theory developed in the previous section, we now evaluate the results by numerical examples. From Theorem 2.2.3 we obtain the parameters ρ_0 and a to be critical for the validity of the product approximation (2.5) so that their influence is of special interest for us. As we are considering a near field regime where it is numerically favorable to represent the Fresnel propagator in an effective geometry, we in particular focus on the error estimate given by Corollary (2.2.5). In this parallel beam case the source size a finds its expression in the bandwidth $b = \frac{\kappa a}{R}$ of the probe field P. Moreover, the parameters b and ρ_0 are enclosed in the corresponding correlation lengths Δx_P and Δx_O . In order to be independent from the sharpness of the inequalities (2.13), we study the relative error δ given by Equation (2.14) instead of the right hand side of (2.18). Note from the proof of Corollary (2.2.5) that with respect to an effective geometry δ reads:

$$\delta(\mathbf{r}') = \frac{\left|\mathcal{D}_{\Gamma_{eff}}(P\tilde{O})(\mathbf{r}'_{eff}) - e^{-i\kappa\Gamma_{eff}}\mathcal{D}_{\Gamma_{eff}}\tilde{O}(\mathbf{r}'_{eff})\,\mathcal{D}_{\Gamma_{eff}}P(\mathbf{r}'_{eff})\right|}{\left\|\mathcal{D}_{\Gamma_{eff}}P\right\|_{\infty}\,\left\|\mathcal{D}_{\Gamma_{eff}}\tilde{O}\right\|_{\infty}}.$$

Introducing the 'effective' Fresnel number $\mathfrak{f}_{eff} := \frac{\kappa\left(\mathbf{r}^2_{X,1},\mathbf{r}^2_{X,2}\right)}{\Gamma_{eff}}$ and assuming $\epsilon(\rho_0) = 0$, Equation (2.6) yields

$$\delta(\mathbf{r}') \leq \frac{\sqrt{2}}{4\,\|\mathfrak{f}_{eff}\|_{l^1}}\frac{\mathbf{r}_X}{\Delta x_P}\cdot\frac{\mathbf{r}_X}{\Delta x_O}\frac{\left(\int_{\mathbb{R}^2}|\mathcal{F}P(\mathbf{k}')|\,d\mathbf{k}'\right)\left(\int_{\mathbb{R}^2}|\mathcal{F}\tilde{O}(\mathbf{k}')|\,d\mathbf{k}'\right)}{\left\|\mathcal{D}_{\Gamma_{eff}}P\right\|_{\infty}\,\left\|\mathcal{D}_{\Gamma_{eff}}\tilde{O}\right\|_{\infty}}. \tag{2.19}$$

In order to create probe fields P which differ only in their bandwidth b, we filter a 'fine structured' complex function $\tilde{P}$ by sinc filters with different cutoff frequencies b:

$$P = \mathcal{F}^{-1}\left(\mathbf{1}_{[-b,b]^2} \bullet \mathcal{F}\tilde{P}\right),$$

where $\mathbf{1}_{[-b,b]^2}(\mathbf{x}')$ is 1 if $\mathbf{x}' \in [-b,b]^2$ and vanishes otherwise. The amplitude of $\tilde{P}$ is chosen to be Dürer's Melencolia I, see Figure 2.4 (b), while an image of a mandrill, depicted in Figure 2.4 (a), serves as the phase. The resolution of the images is 512×512 pixels and they are padded with ones (for the amplitude image) and zeros (for the phase image) to obtain a resolution of 2048×2048 pixels in time and Fourier space, respectively. Then we vary b in the range between 1 and 1024 pixels in the Fourier domain. We assume that the distances are $R = 6$ mm and $\Gamma = 519$ mm, the wavenumber is $\kappa = 86\,\mathrm{nm}^{-1}$, and the pixel size is 2.2µm yielding an effective pixel size of 25 nm. Therefore, the effective source size $a = \frac{Rb}{\kappa}$ corresponding to b is in the interval $[9\,\mathrm{nm}, 8748\,\mathrm{nm}]$. As sample we choose a pure phase object where the phase is a grating in $[-0.3, 0]$ with structure size $\Delta x_O = 8, 16, 32, 128\,\mathrm{px}$ generated by sine functions (see Figure 2.4 c) for $\Delta x_O = 32\,\mathrm{px}$). According to our theory, in this example one clearly obtains aberrations in the approximation $\left|\mathcal{D}_{\Gamma_{eff}}P\right|^{-2}\left|\mathcal{D}_{\Gamma_{eff}}PO\right|^2$ (see Figure 2.4 f)) compared to the true intensities $\left|\mathcal{D}_{\Gamma_{eff}}O\right|^2$, shown in Figure 2.4 g). The corresponding error $|\mathcal{D}_{z_{eff}}O(\mathbf{r}'_{eff})|^2 - |\mathcal{D}_{z_{eff}}(P\cdot O)(\mathbf{r}'_{eff})|^2\,|\mathcal{D}_{z_{eff}}(P)(\mathbf{r}'_{eff})|^{-2}$ is depicted in Figure 2.4 i). By comparison with the studied error δ (Figure 2.4 h)) we see that both terms have the same behavior

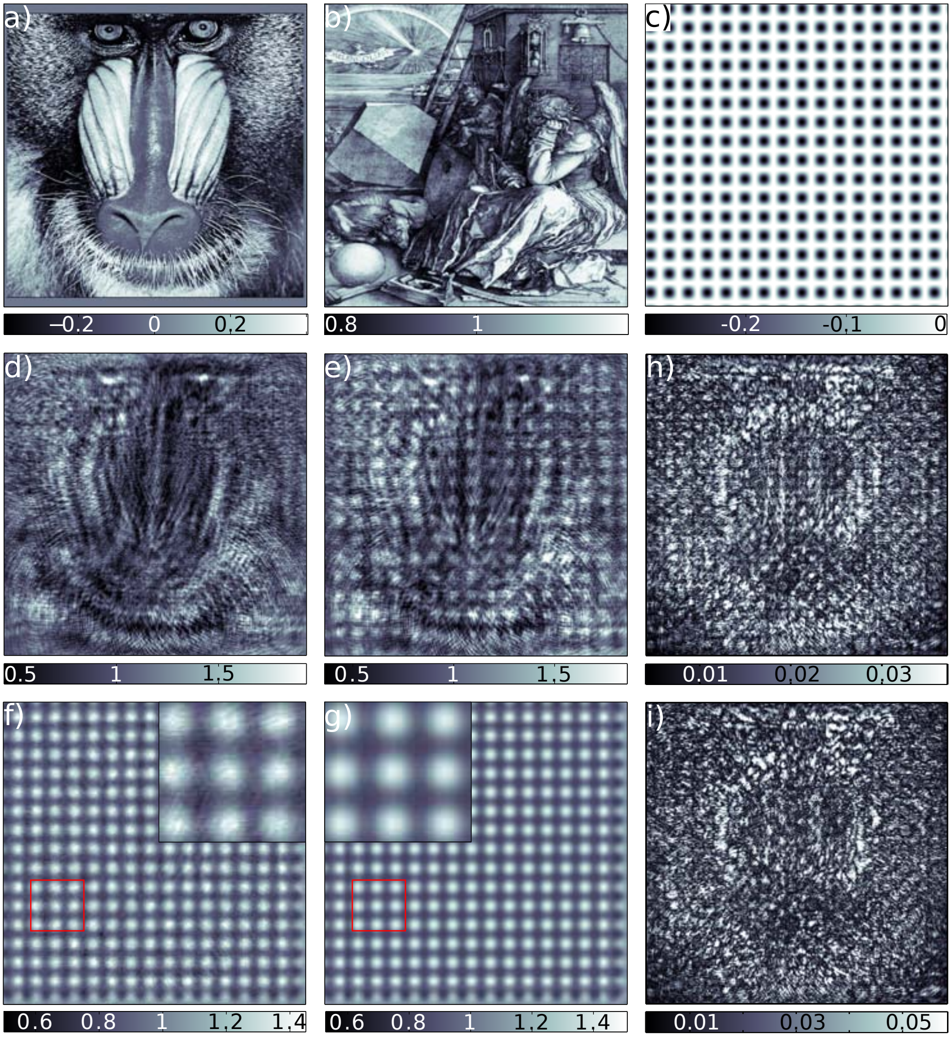


Figure 2.4: Overview of the simulation setup and data generation. The resolution is 512×512 pixels. a) + b): Images of a mandrill and Dürer's Melencolia I which serve as phase (in $[-0.4, 0.4]$ rad) and amplitude (in $[0.8, 1.2]$) for the complex (unfiltered) probe field P. c): Phases of the object function where the feature size constant Δx_O is 32 pixels. d): Intensities $|\mathcal{D}_{\Gamma_{eff}} P|^2$ where the empty beam field is defined by a) and b). e): Intensities $|\mathcal{D}_{\Gamma_{eff}} P\,O|^2$. f): Result of the empty beam correction, i.e. e) divided by d). g): Intensities $|\mathcal{D}_{\Gamma_{eff}} O|^2$ which are supposed to be approximated by the empty beam correction. h): δ-error matrix as defined by Equation (2.14). i): error matrix $\big|(\mathrm{g}) - (\mathrm{f})\big|$. For better visual comparison of f) and g) the insets show a zoom into the region indicated by a red square.

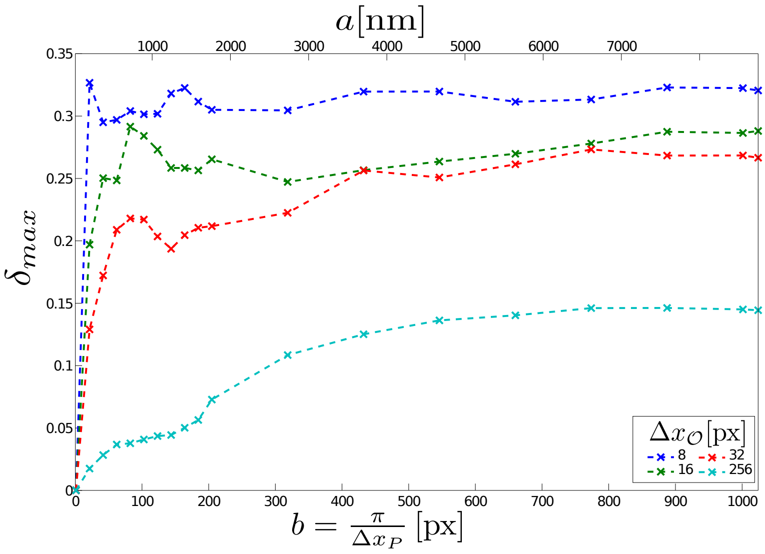


Figure 2.5: Error $\delta_{\max} = \max \delta$ under the setup of Figure 2.4 for different correlation lengths Δx_O and bandwidths b given in units of pixels. In addition to b the x-label on top indicates the corresponding effective source size a in nanometer.

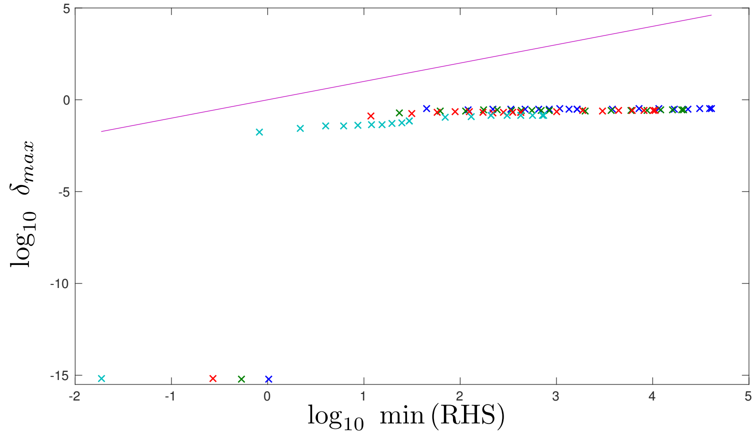
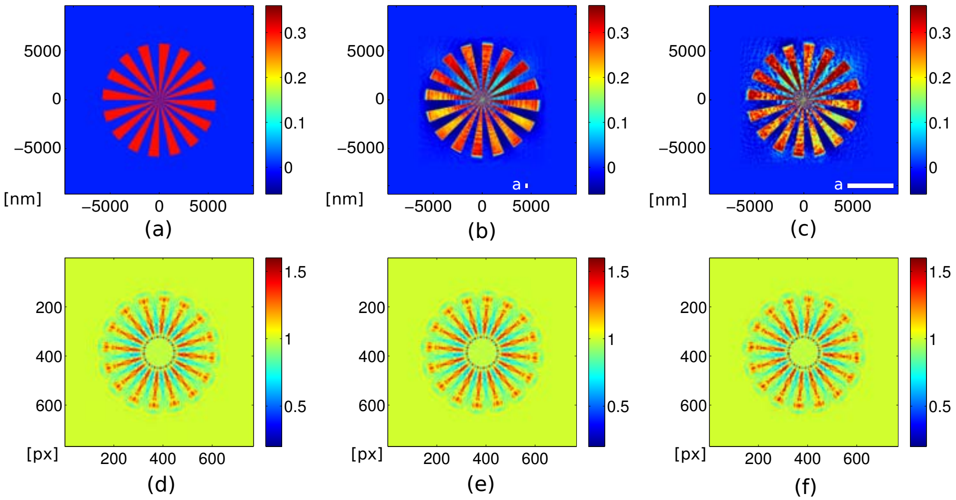


Figure 2.6: Illustration of the accuracy of the error estimate (2.19). $log_{10}\,\delta_{max}$ is plotted versus log_{10} of the minimum of the right hand side (RHS) of (2.19). The same color coding as in Figure 2.5 is used. The purple solid curve is the identity.



(a) (b) (c) (d) (e) (f)

Figure 2.7: Influence of the empty beam correction on the reconstruction. Under the experimental setup described above the (real valued) phase ϕ of the object function $O(\phi) = \exp(\mathrm{i}\kappa\phi)$ is the Siemens star shown in (a). The illumination is given by a probe field P with different source sizes a which is generated by filtering the 'Mandrill-Dürer-Probe', from Figure 2.4. (d) are the intensities $|\mathcal{D}_{\Gamma_{eff}} O|^2$ corresponding to plane wave illumination ($a \approx 0$). (e) and (f) show the empty beam corrected data for $a = 183\,\mathrm{nm}$ and $a = 4667\,\mathrm{nm}$, respectively. (b) and (c) are the reconstructions from this data (e) and (f), performed by the IRNM with $\alpha_n = 0.01\,0.5^{n-1}$ after 12 iterations.

and yield values in the same range. Note that as the object functions O of sine gratings are sufficiently well-behaved we here considered O instead of $\tilde{O}$.

Based on the estimate (2.19), in Figures 2.5 and 2.6 we study the maximum error

$$\delta_{max} := \max_{\xi' \in M^{-1}\mathbb{P}_{\Gamma_{eff}}} \delta(\xi')$$

(for the shifted object function $\tilde{O}$) with respect to the finest relevant feature sizes Δx_P and Δx_O. As expected, the error increases when either Δx_P or Δx_O is decreased. So, the smaller the structures in the object, the more stringent is the requirement to keep the correlation length Δx_P of the wavefront aberrations large or equivalently the source size a small. This also means that for quite extended source sizes a it can not be assured that small structures in the object will be transferred correctly after application of the empty beam division. Note that the case of a plane wave illumination where $a \to 0$ is well captured. In Figure 2.6 we see that the right hand side of (2.19) provides a sufficiently good upper bound of δ where we can assume that the accuracy of (2.19) relies in particular on the sharpness of the inequalities (2.13). The outliers where δ_{max} vanishes belong to the case of plane wave illumination and thus verify the proven exactness of the product approximation in the detector plane for this case.

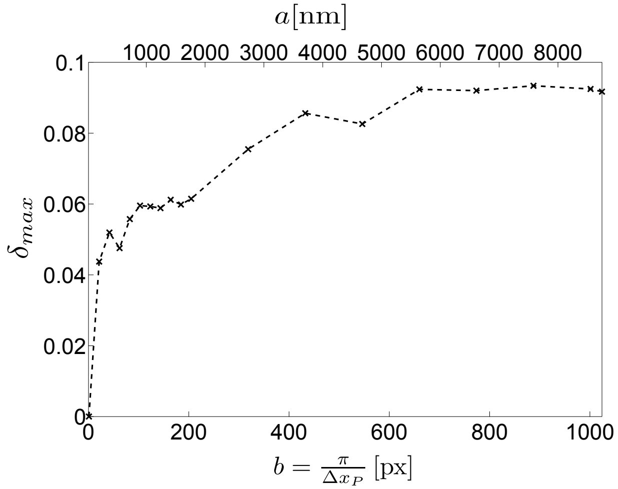


Figure 2.8: Error $\delta_{\max} = \max \delta$ as the function of the bandwidth b for the object function O (instead of $\tilde{O}$) of a Siemens star test pattern (cf. Figure 2.7 (a)). The setup is described in Figure 2.7. In addition to b, the x-label on top indicates the corresponding effective source size a in nanometer.

Figure 2.7 illustrates the influence of the error δ_{max} which is given as a function of the source size a on the reconstruction of a phase shifting Siemens star shown in (a). Although the results of the empty beam correction in (e) and (f) (obtained for $a = 183\,\text{nm}$ and $a = 4667\,\text{nm}$, respectively) look quite similar to the exact intensities $\left|\mathcal{D}_{\Gamma_{eff}} O\right|^2$ shown in (d), the reconstructions differ considerably. Note that by the IRNM of the form

$$x_{n+1} = \operatorname*{argmin}_{\phi\in X} \frac{1}{2}\left\| \frac{|\mathcal{D}_{\Gamma_{eff}} P O|^2}{|\mathcal{D}_{\Gamma_{eff}} P|^2} - T_{\text{Fresnel}}(\phi_n) - T'_{\text{Fresnel}}[\phi_n](\phi-\phi_n)\right\|_Y^2 + \frac{\alpha_n}{2}\|\phi\|_X^2,$$

with $\alpha_n := 0.01\, 2^{-n+1}$, $X = l^{1.5}$, and $Y = l^2$ we already used a method that takes noisy data into account. The inner minimization problem is solved by the generalized Chambolle-Pock algorithm CP-BS that will be introduced in Section 4.2. We applied the version CP-BS 1 described in Theorem 4.2.1, with $\tau = 6$ and $\sigma = 0.96\, \|T'_{\text{Fresnel}}[\phi_n]\|^{-2}\, \tau^{-1}$. In both cases ($a = 183\,\text{nm}$ and $a = 4667\,\text{nm}$) we obtained after 12 IRNM iterations approximations which are closest to the true solution. So, this example underlines for δ_{max} being large the need of including further information on the illumination function in the reconstruction process, e.g. the approximated source size. In this context, promising algorithms for the simultaneous reconstruction of the empty beam ι_0 and the object function by taking intensity measurements at several distances have been proposed recently [36,64]. Figures 2.7 and 2.8 also illustrate that in order to predict the reconstruction's quality the error value δ_{max} has to be considered in relation to the hologram's contrast. Since the image information is contained in the deviation of the pixel values (in the normalized hologram) from unity amplitude, for the given weak object with an averaged contrast of 5% an error δ of a similar, small level significantly deteriorates the reconstruction result.

3 Tikhonov-type regularization on Banach spaces

In the following section, we motivate iteratively regularized Newton-type methods (IRNM) for solving nonlinear inverse ill-posed problems $Tx = y^\delta$ such as the phase retrieval problem introduced in the previous chapter. Then, in order to develop a method for the inner convex optimization problems of the IRNM which consist of finding the minimizer of a general linear Tikhonov-type functional

$$x \mapsto S(y^\delta; Ax) + \alpha R(x)$$

with a linear operator A and not necessary differentiable data misfit $S(y^\delta; \cdot)$ and/or penalty term R. Sections 3.2 and 3.3 provide the necessary definitions and results.

3.1 Tikhonov regularization and IRNM

First of all, let us consider a linear ill-posed problem given by the operator equation

$$Tx = y,$$

where $T : D(T) \subset X \to Y$ is a linear operator between the Banach spaces X and Y. Moreover, as in applications the exact right-hand side y is usually not available, but only noisy data y^δ, we assume that our given data $y^\delta \in Y$ might be disturbed with a noise level $\delta > 0$ such that the misfit between y and y^δ, which is measured by some data fidelity functional $\tilde{S}(y^\delta; \cdot) : Y \to [0, \infty]$, is bounded by $\delta > 0$:

$$\tilde{S}(y^\delta; y) \leq \delta.$$

However, due to the ill-posedness, the minimization problem

$$\operatorname*{argmin}_{x \in X} \tilde{S}(y^\delta; Tx)$$

neither needs to be solvable nor stable. To overcome these drawbacks, we use a Tikhonov-type regularization of the general form

$$x_\alpha = \operatorname*{argmin}_{x \in X} S(y^\delta; Tx) + \alpha R(x), \tag{3.1}$$

where $\alpha > 0$ is the regularization parameter. The penalty term $R : X \to [0, +\infty)$ as well as the data fidelity functional $S(y^\delta; \cdot) : Y \to [0, +\infty)$, most likely picked as $S = \tilde{S}$, are chosen problem adapted and belong to the class $\Gamma(X)$ and $\Gamma(Y)$ of proper, convex, and lower semicontinuous (l.s.c.) functions (see definitions in the next section). This approach was introduced by Tikhonov ([75, 76]) for a Hilbert space setting with special interest in S and R given by Hilbert space norms: $S(y^\delta; y) = \|y - y^\delta\|_X^2$ and $R(x) = \|x - x_0\|_Y^2$. However, there are a lot of applications where such a quadratic Tikhonov functional does not adequately reflect the properties of the problem. That is for example the case if the solution is known to have a blocky structure or is sparse, which is typically modeled by choosing the total variation, a non-Hilbertian-Sobolev norm, or the L^1-norm as penalty term R, respectively. For the same reasons, a general Banach space setting is preferable to a Hilbert space setting. General Tikhonov-type regularization has been studied intensely during the last decade. For instance, see [70] and the references therein for the case of general penalties and norm monomials as data fidelity functionals, [82] for Kullback-Leibler like data fidelity functionals, and [63] for the even more general theory of a vector space setting. Under certain assumptions, including, inter alia, the existence of an exact solution $x \in D(T)$ which is uniquely determined by the exact right hand side $y = T(x)$, the following properties of the Tikhonov-type regularization (3.1) can usually be shown (cf. e.g. [63]):

- The method (3.1) is *well-defined*, i.e. for any $\alpha > 0$ and $y^\delta \in Y$ there exists a minimizer $x_\alpha \in X$.
- The method (3.1) is *stable* in the sense that for any arbitrary $\alpha > 0$ the minimizer x_α depends continuously on the data y^δ.
- The method (3.1) *converges* in the sense that together with an appropriate parameter choice $\alpha(\delta, y^\delta)$, depending on the noise level δ and/or the given data y^δ, or any monotonically decreasing sequence of noise levels $\delta_{k\in\mathbb{N}} \searrow 0$ and any (noisy) data sequence $y^\delta_{k\in\mathbb{N}}$, with $S(y^\delta_k; y) \leq \delta_k$, the corresponding sequence $x_{\alpha,k}$ converges to the exact solution x, as $k \to \infty$. Typically, a sufficient condition on the parameter choice rule (independent of y^δ) is:

  $$\alpha(\delta, y^\delta) \to 0 \quad \text{and} \quad \frac{\delta}{\alpha(\delta, y^\delta)} \to 0, \qquad \text{as } \delta \to \infty.$$

  This rule reflects the understanding that for less noise on the given data y^δ also less regularization of the problem is required.

Here, 'convergence' does not necessarily mean strong convergence in the corresponding Banach space, but it can also denote weak convergence of subsequences.

Example 3.1.1. As a simple illustration, let us consider a linear ill-posed problem $Tx = y$ with convolution operator

$$T(x) : [-1, 1] \to \mathbb{R}, \quad T(x)(t) := \int_{-\frac{1}{2}}^{\frac{1}{2}} x(s)\, k(t-s)\, ds, \quad k(t) := \exp(-5|t|) \qquad (3.2)$$

and sparse solution $x : [-1/2, 1/2] \to \mathbb{R}$ (see Figure 3.1). This sparsity constraint is incorporated by setting $R(x) = \|x\|_{L^1([-0.5,0.5])}$ in (3.1). Moreover, as instead of the exact data y, only data y^δ distrubed by 18 % normal distributed noise is given, we choose $S(y^\delta; y) = \frac{1}{2}\|y^\delta - y\|^2_{L^2([-1,1])}$ as data fidelity functional. So, accordingly, $X = L^r([-1/2, 1/2])$, with $r \in (1, 2]$ and $Y = L^2([-1, 1])$ seems to be an appropriate setting. Figure 3.1 b) shows the minimizer x_α of the Tikhonov-type functional

$$x \mapsto \frac{1}{2}\|y^\delta - y\|^2_{L^2([-1,1])} + \alpha\|x\|_{L^1([-\frac{1}{2}, \frac{1}{2}])}$$

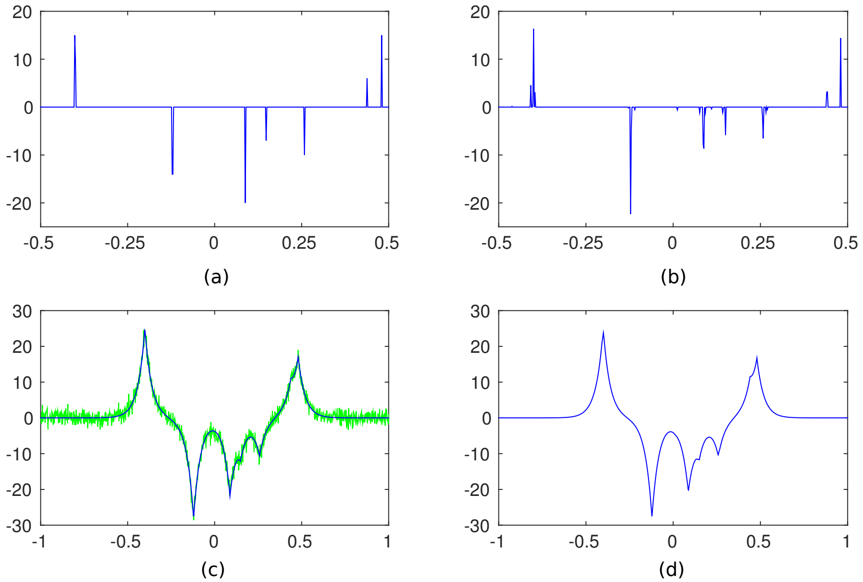


Figure 3.1: Sparse convolution problem: (a) exact solution, (b) reconstruction x_α, (c) exact (blue) and given (green) data, (d) reconstructed data.

for $\alpha = 5$. Although we obtain the same x_α for all preimage spaces $X = L^r([-1/2, 1/2])$, with $r \in (1, 2]$, intuitively r close to 1 seems to be preferable. In fact, we will illustrate that a problem adapted choice for X and Y will have positive effects on the performance of our method for solving convex optimization problems of the form (3.1).

Now let us consider nonlinear inverse problems $Tx = y$ with a nonlinear Fréchet differentiable operator $T : D(T) \subset X \to Y$. As e.g. also studied in [70, 82], for given (noisy) data y^δ the corresponding solution can similarly be approximated by the iteratively regularized Newton-type method (IRNM)

$$x_{n+1} = \operatorname*{argmin}_{x \in X} S(y^\delta; T(x_n) + T'[x_n](x - x_n)) + \alpha_n R(x) \tag{3.3}$$

for some initial guess $x_0 \in X$ and positive regularization parameters $(\alpha_n)_{n\in\mathbb{N}}$ which now also depend on the iteration step n. The idea behind this method is to apply a Tikhonov-type regularization not to the nonlinear operator equation $Tx = y$, since this often leads to a Tikhonov-type functional with local minima, but iteratively to the linearized versions

$$T(x_n) + T'[x_n](x - x_n) \approx y^\delta, \quad n \in \mathbb{N}.$$

Based on the works [47, 82, 83], our interest in the IRNM is to use it for solving the phase retrieval problems introduced in Section 1.1. Under appropriate conditions, one typically obtains the following properties of the IRNM equipped with a parameter choice rule for $\alpha_n(\delta, y^\delta)$ (cf. [82]):

- The IRNM (3.3) is *well-defined*, i.e. for any iteration step $n \in \mathbb{N}$ and any $y^\delta \in Y$ there exists a minimizer x_{n+1}.
- The method (3.3) *converges for exact data* in the sense that if $y^\delta = y$ there either exists a finite index $N \in \mathbb{N}$ such that x_N coincides with the exact solution x, or the sequence $(x_n)_{n\in\mathbb{N}}$ converges to x, as n turns to ∞.
- The method (3.3) *converges for noisy data* in the sense that together with an appropriate rule for a stopping index $N(\delta, y^\delta) \in \mathbb{N}$ for any monotonically decreasing sequence of noise levels $\delta_{k\in\mathbb{N}} \searrow 0$ and (noisy) data $y^\delta_{k\in\mathbb{N}}$, with $S(y^\delta_k; y) \le \delta_k$ the corresponding sequence $x_{N(\delta_k, y^\delta_k)}$ converges to the exact solution x, as $k \to \infty$.

Typically, the parameter choice rule for $\alpha_n(\delta, y^\delta)$ has to satisfy the following condition:

$$\alpha_0 \le 1, \qquad 1 \le \frac{\alpha_n(\delta, y^\delta)}{\alpha_{n+1}(\delta, y^\delta)} \le C \quad \text{for } n \in \mathbb{N}, \qquad \alpha_n(\delta, y^\delta) \searrow 0 \quad \text{as } n \to \infty,$$

for some constant $C \ge 1$. Moreover, with the help of source conditions or variational inequalities, also convergence rates have been proven. But, as already mentioned in Section 1.1, in order to apply the IRNM with the non-quadratic data fidelity term (1.28) and very general penalty term we also need a quite general method for solving the inner minimization problem. To this end, we rewrite this optimization problem in the common, basic form

$$\bar{x} = \operatorname{argmin}_{x\in X} \left(g(Tx) + f(x)\right), \qquad (\mathbf{P})$$

where $g = S(y^\delta; \cdot)$ and $f = \alpha R$. Together with standard assumptions given by the IRNM and its application to phase retrieval problems we consider a problem which frequently arises in many other contexts as well. For example, common approaches in image deblurring (e.g. the ROF model [68]) or sparse signal restoration (e.g. the LASSO problem [74]) can be interpreted as Tikhonov-type regularizations of the form (3.1) (covered by (**P**)). In order to formulate the corresponding assumptions and to present a method for solving (**P**), we will need some basic definitions and results from convex analysis.

3.2 Duality theory for linear Tikhonov-type regularization

The following basics from convex analysis on Banach spaces can be found e.g. in [5, 86]. As above, we assume X, Y, and Z to be reflexive and real Banach spaces with corresponding norms given by $\|\cdot\|_X$, $\|\cdot\|_Y$, and $\|\cdot\|_Z$, respectively. Z^*, equipped with the norm

$$\|z^*\|_{Z^*} = \sup\{|z^*(z)| = |\langle z, z^*\rangle_Z| \mid z \in Z,\ \|z\|_Z = 1\},$$

denotes the corresponding topological dual space of such a Banach space Z, paired by $\langle\cdot,\cdot\rangle_Z : Z \times Z^* \to \mathbb{R}$. Moreover, we introduce the set $\overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$ of real numbers extended by $+\infty$ and $-\infty$.

3.2.1 The class of proper, convex and lower semicontinuous functions

First of all, let us define some key properties of data fidelity functionals $g = S(y^\delta; \cdot) : Y \to \overline{\mathbb{R}}$ and penalty terms $f = \alpha R : X \to \overline{\mathbb{R}}$.

Definition 3.2.1 ([5] Def. 1.1, Prop. 1.2)**.** A function $h : A \to \overline{\mathbb{R}}$ given on a convex subset A of Z is called *convex* if for any $z, u \in A$ and any $\lambda \in [0, 1]$ the following inequality holds:

$$h(\lambda z + (1 - \lambda)u) \leq \lambda h(z) + (1 - \lambda)h(u).$$

If this inequality strictly holds for all $z \neq u$ we call h *strictly convex*. Equivalently, convexity of h is given if and only if its *epigraph* $\operatorname{epi} h := \{(z, \alpha) \in A \times \mathbb{R} \,|\, h(z) \leq \alpha\}$ is a convex subset of $A \times \mathbb{R}$.

With respect to Tikhonov-type functionals it makes no sense to consider functions f, g taking the value $-\infty$. Therefore we introduce the following property:

Definition 3.2.2. A convex function $h : Z \to \overline{\mathbb{R}}$ is called *proper* if for any point $z \in Z$ we have $h(z) > -\infty$ and there is at least one point $z \in Z$ where $h(z)$ is different from $+\infty$.

Assuming f and g to be proper and convex and at least one of the functions $x \mapsto g(Tx)$ or f to be even strictly convex ensures the uniqueness of the minimizer of the then also strictly convex function $F : x \mapsto (g(Tx) + f(x))$:

Theorem 3.2.3. *Let both $f : X \to \overline{\mathbb{R}}$ and $g : Y \to \overline{\mathbb{R}}$ be proper and convex and $T : X \to Y$ a linear, bounded operator. Then every local minimum of the convex function $F : x \mapsto (g(Tx) + f(x))$ is also a global minimum of F. Moreover, the set of minimizers*

$$\{\bar{x} \in X \,|\, \bar{x} = argmin_{x \in X} (g(Tx) + f(x))\}$$

is convex. If in addition f is strictly convex or g is strictly convex and T is injective, then F has at most one global minimum.

Proof. It is easy to see that the sum F of the two convex functions f and $g(T\cdot)$ is convex as well. Moreover, F is strictly convex if one of the conditions of the last assertion is satisfied. With this knowledge the assertions are given by Propositions 2.5.6 and 2.5.8. in [86]. □

Moreover, we want to allow f and g to be discontinuous in a certain manner.

Theorem and Definition 3.2.4. *A function $h : Z \to \overline{\mathbb{R}}$ is called* lower semicontinuous (l.s.c) *if one of the following equivalent conditions holds true:*

- *For any $z \in Z$ and any sequence $(z_n)_{n \in \mathbb{N}} \subseteq Z$ converging (with respect to $\|\cdot\|_Z$) to z as $n \to \infty$, we have*
$$\liminf_{n \to \infty} h(Z_n) = h(z).$$

- *For any $\alpha \in \mathbb{R}$ the level set $L_\alpha := \{z \in Z \,|\, h(z) \leq \alpha\}$ is closed.*
- *The epigraph epi f is closed in $Z \times \mathbb{R}$.*

Proof. See e.g. Definition 1.2. and Proposition 1.3 in Chap. 2 of [5]. □

Now that we have defined all the properties that f and g should satisfy, we introduce for any reflexive Banach space Z the set of proper, convex and lower semicontinuous functions

$$\Gamma(Z) := \left\{h : Z \to \overline{\mathbb{R}} \,|\, h \text{ is proper, convex and l.s.c} \right\}$$

and assume $f \in \Gamma(X)$ and $g \in \Gamma(Y)$. It is easy to check that then also the sum

$$x \to g(Tx) + f(x)$$

belongs to the class $\Gamma(X)$.

Example 3.2.5. The following typical choices for f and g are proper, convex and l.s.c:

(i) For any Banach space Z and any exponent $r \geq 1$ the norm monomial $\|\cdot\|_Z^r : Z \to \overline{\mathbb{R}}$ belongs to $\Gamma(Z)$.

(ii) The indicator function

$$\chi_C : Z \to \mathbb{R}, \quad \chi_C(z) = \begin{cases} 0 & z \in C \\ +\infty & \text{otherwise.} \end{cases} \tag{3.4}$$

of a set $C \subset X$ belongs to $\Gamma(Z)$ if and only if C is nonempty, convex and closed.

(iii) For any $\mathbf{N} \in \mathbb{N}^2$ and any $y^\delta \in Y = l^{r^*}(\triangle_{\mathbf{N}})$ with $y^\delta_{\mathbf{i}} \geq 0$ for any index $\mathbf{i} \in \mathbf{N}$ and $r^* > 1$ the function $kl_{\mathbf{N}}\left(y^\delta; \cdot\right) : Y \to \overline{\mathbb{R}}$ given by Equation (1.26) belongs to $\Gamma(Y)$.

Proof. The first assertion directly follows from the triangle inequality and the convexity and continuity of $z \mapsto |z|^r$. For the second assertion we consider the epigraph epiχ_C, which is convex and closed if and only if C has the same properties. $kl_{\mathbf{N}}\left(y^\delta; \cdot\right)$ is convex since it is a linear combination of the convex functions $a \mapsto a$ for $a \geq 0$ and $a \mapsto -\ln a$ for $a > 0$. Similarly, it is lower semicontinuous as any summand

$$h_{\mathbf{i}} : y_{\mathbf{i}} \mapsto \begin{cases} y_{\mathbf{i}} - y^\delta_{\mathbf{i}} \ln(y_{\mathbf{i}}), & \text{if } y^\delta_{\mathbf{i}} > 0 \text{ and } y_{\mathbf{i}} > 0, \\ y_{\mathbf{i}} & \text{if } y^\delta_{\mathbf{i}} = 0 \text{ and } y_{\mathbf{i}} \geq 0 \\ \infty & \text{otherwise,} \end{cases}$$

is. Finally, because of $-\infty < h_{\mathbf{i}}(y^\delta_{\mathbf{i}}) \leq h_{\mathbf{i}}(y_{\mathbf{i}})$ for all $y_{\mathbf{i}}$ in the case of $y^\delta_{\mathbf{i}} > 0$ and $h_{\mathbf{i}}(0) = 0 \leq h_{\mathbf{i}}(y_{\mathbf{i}})$ for all $y_{\mathbf{i}}$ in the case $y^\delta_{\mathbf{i}} = 0$ the function $kl_{\mathbf{N}}\left(y^\delta; \cdot\right)$ is proper. □

3.2.2 Subdifferential and conjugate function

Next, we would like to characterize the solutions to the problem (**P**). For this purpose we introduce the subdifferential which generalizes the derivative:

Definition 3.2.6. For a proper and convex function $h : Z \to \overline{\mathbb{R}}$ on a Banach space Z the (possibly set-valued) subdifferential $\partial h : Z \rightrightarrows Z^*$ of h is given by

$$\partial h(z) := \{z^* \in Z^* \mid \langle u - z, z^* \rangle_Z \leq h(u) - h(z), \text{ for all } u \in Z\}.$$

This definition directly yields the following necessary and sufficient condition for solutions of convex minimization problems:

Theorem 3.2.7. *(Theorem 2.5.7 [86]) Let $h : Z \to \overline{\mathbb{R}}$ be a proper and convex function. Then $\bar{x} \in Z$ with $h(\bar{x}) < \infty$ is a minimizer of h if and only if $0 \in \partial h(\bar{x})$.*

Proof. $h(\bar{x}) \leq h(u)$ for all $u \in Z$ is equivalent to $0 \leq h(u) - h(\bar{x})$ for all $u \in Z$ which holds true if and only if $0 \in \partial h$. □

One can show that a proper and convex function $h : Z \to \overline{\mathbb{R}}$ is subdifferentiable at a point $z \in Z$ if it is finite and continuous in z . Moreover, if h is continuous at a point $z \in Z$ and $\partial h(z)$ is single valued, then also the Gâteaux derivative $h'(z)$ exists in z with $\partial h(z) = h'(z)$. On the other hand, if h is Gâteaux differentiable at a point $z \in Z$, the derivative $h'(z)$ coincides with the subdifferential at z. Thus, the subdifferential can be seen as a generalized Gâteaux derivative.

Example 3.2.8. Let us consider the non-differentiable function $h : \mathbb{R} \to \overline{\mathbb{R}}$, $h(x) = |x|$ which obviously is in $\Gamma(\mathbb{R})$. We have $\partial h(x) = \operatorname{sign}(x)$ for $\mathbb{R} \setminus \{0\}$ and $\partial h(0) = [-1, 1]$. So, we see that in the origin the subdifferential corresponds to all lines passing $h(0) = 0$ which lie completely under the graph of h, c.f. Figure 3.2. This statement applies accordingly to all other points $z \in \mathbb{R} \setminus \{0\}$ where the subdifferential coincides with the Gâteaux derivative and hence, is given by the tangent to h at z. Theorem 3.2.7 determines $z = 0$ as the unique minimizer of h.

Proposition 3.2.9. *The subdifferential of the norm monomial $h_{2,Z}(z) = \frac{1}{2}\|z\|_Z^2$ is given by*

$$\partial h_{2,Z}(z) = \left\{z^* \in Z^* \mid \langle z, z^* \rangle_Z = \|z\|_Z^2 = \|z^*\|_{Z^*}^2 \right\}.$$

Proof. [cf. [5, p. 102]] Suppose that $z^* \in Z^*$ satisfies $\langle z, z^* \rangle_Z = \|z\|_Z^2 = \|z^*\|_{Z^*}^2$. Then for any $u \in Z$ we have

$$\langle u - z, z^* \rangle_Z \leq \|u\|_Z \|z\|_Z - \|z\|_Z^2 = \frac{-1}{2} (\|u\|_Z - \|z\|_Z)^2 + \frac{1}{2}\|u\|_Z^2 - \frac{1}{2}\|z\|_Z^2 \leq h_{2,Z}(u) - h_{2,Z}(z)$$

and hence $z^* \in \partial h_{2,Z}(z)$. Conversely, assume that $z^* \in \partial h_{2,Z}$ such that the inequality $\langle u - z, z^* \rangle_Z \leq \frac{1}{2}\|u\|_Z^2 - \frac{1}{2}\|z\|_Z^2$ holds true for any $u \in Z$. Substituting $u = z + \lambda v$ for $\lambda > 0$ and $v \in Z$ yields

$$\lambda \langle v, z^* \rangle_Z \leq \frac{1}{2} \left((\|z\|_Z + \lambda \|v\|_Z)^2 - \|z\|_Z^2 \right) \leq \lambda \|z\|_Z \|v\|_Z + \frac{\lambda^2}{2}\|v\|_Z^2.$$

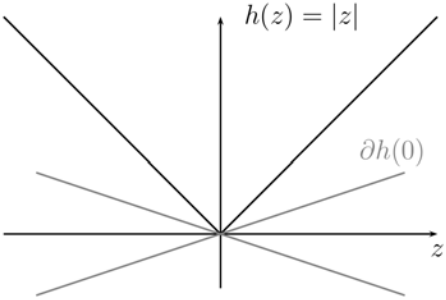


Figure 3.2: Illustration of the subdifferential

Then dividing by λ and taking the limit for $\lambda \to 0$ leads to $|\langle v, z^* \rangle_Z| \leq \|z\|_Z \|v\|_Z$ for any $v \in Z$. On the other hand, for $u = (1-\lambda)z$ with $\lambda \in (0,1)$ we obtain

$$-\lambda \langle z, z^* \rangle_Z \leq \frac{1}{2}\left((1-\lambda)^2 \|z\|_Z^2 - \|z\|_Z^2\right) \leq -\frac{\lambda^2}{2}\|z\|_Z^2,$$

and thus $\langle z, z^* \rangle_Z \geq \|z\|_Z^2$. This proves the assertion. □

So, if Z is a Hilbert space, the subdifferential $\partial h_{2,Z}$ is the identity operator (cf. Theorem 3.3.6). This fact is in particular useful for minimizing problems of the form

$$\operatorname*{argmin}_{z \in Z} \tau h(z) + \frac{1}{2}\|z - u\|_Z^2, \tag{3.5}$$

on a Hilbert space Z where $h \in \Gamma(Z)$, $\tau > 0$ and $u \in Z$: Under the assumption that h is finite and continuous at at least one point in Z we can apply the sum rule for the subdifferential. Then because of 3.2.7, a solution $\bar{x} \in Z$ to (3.5), which is unique due to Theorem 3.2.3, satisfies

$$u \in (\tau \partial h + I)(\bar{x}).$$

It is well-known ([65]) that the operator $(\tau\partial h + I)$ is bijective for any $h \in \Gamma(Z)$ and any $\tau > 0$. Therefore its inverse $(\tau\partial h + I)^{-1} : Z^* = Z \to Z$, the *resolvent of h*, is well-defined and single-valued. This well-established operator $(\tau\partial h + I)^{-1}$ is also referred to as *proximity operator*, abbreviated $prox_{\tau h}$, (see e.g. [26] and the references therein) and can be interpreted as the generalization of the orthogonal projection

$$\operatorname*{argmin}_{z \in C} \|z - u\|_Z^2 = \operatorname*{argmin}_{z \in Z} \iota_C(z) + \frac{1}{2}\|z - u\|_Z^2 = (\partial \iota_C + I)^{-1}(u)$$

onto convex sets $C \subset Z$. It is a major benefit that for a lot of interesting penalty terms and data fidelity functionals the corresponding resolvent has a closed or at least “sufficiently simple” form as the following example illustrates:

Example 3.2.10. Let us consider the l^1-norm $h(z) = \sum_{i=1}^N |z_i|$ on the N-dimensional Hilbert space $l^2(\{1,\dots N\})$. Due to Example 3.2.8 the subdifferential reads as

$$(\partial h(z))_i = \begin{cases} \operatorname{sign}(z_i) & z_i \neq 0 \\ [-1,1] & z_i = 0, \end{cases}$$

where $i \in \{1,\dots N\}$. Therefore the resolvent for $\tau > 0$ is given by the *shrinkage operation*

$$\left((\tau\partial h + I)^{-1}(z^*)\right)_i = \max\{|z_i^*| - \tau, 0\} \operatorname{sign}(z_i^*) \quad i \in \{1,\dots N\}.$$

As a second example we determine the resolvent of $kl_{\mathbf{N}}(y^\delta;\cdot) : l^2(\triangle_{\mathbf{N}}) \to \overline{\mathbb{R}}$ given by Equation (1.26) for some $y^\delta \in l^2(\triangle_{\mathbf{N}})$ with nonnegative entries and some $\mathbf{N} \in \mathbb{N}^2$. Computing the subdifferential at a point y in the effective domain $\left\{y \in l^2(\triangle_{\mathbf{N}}) \mid kl_{\mathbf{N}}(y) < \infty\right\}$:

$$\left(\partial\, kl_{\mathbf{N}}(y^\delta; y)\right)_{\mathbf{i}} = 1 - \frac{y^\delta_{\mathbf{i}}}{y_{\mathbf{i}}} \quad \text{for } y^\delta_{\mathbf{i}} > 0, \quad \text{and} \quad \left(\partial\, kl_{\mathbf{N}}(y^\delta; y)\right)_{\mathbf{i}} = 1 \quad \text{for } y^\delta_{\mathbf{i}} = 0,$$

we see that at y the resolvent $\bar{p} := \left(\tau\,\partial\, kl_{\mathbf{N}}(y^\delta;\cdot) + I\right)^{-1}(y)$ is the nonnegative solution of

$$(\tau - y_{\mathbf{i}})\,\bar{p}_{\mathbf{i}} - \tau\, y^\delta_{\mathbf{i}} + \bar{p}^2_{\mathbf{i}} = 0, \qquad \mathbf{i} \in \triangle_{\mathbf{N}},$$

and hence

$$\bar{p}_{\mathbf{i}} = \left(\left(\tau\,\partial\, kl_{\mathbf{N}}(y^\delta;\cdot) + I\right)^{-1}(y)\right)_{\mathbf{i}} = \frac{y_{\mathbf{i}} - \tau + \sqrt{(y_{\mathbf{i}} - \tau)^2 + 4\,\tau\, y^\delta_{\mathbf{i}}}}{2} \qquad \mathbf{i} \in \triangle_{\mathbf{N}}.$$

The obvious way of generalizing this concept of resolvents to Banach spaces is to replace the Hilbert norm and space in (3.5) by a Banach norm $\|\cdot\|_Z$ and the corresponding Banach space Z. Also the literature refers to this as resolvent on Banach spaces (cf. [23, p. 168], [10]). However, we will proceed in a different way: On a Hilbert space Z we can apply the polarization identity (see (3.14)) to Equation (3.5) yielding

$$(\tau\partial h + I)^{-1}(u) = \operatorname*{argmin}_{z \in Z} \tau h(z) + \frac{1}{2}\|z\|_Z^2 + \langle z, u\rangle_Z\,. \tag{3.6}$$

Accordingly, on a Banach space Z we consider the minimization problem

$$\operatorname*{argmin}_{z \in Z} \tau\, h(z) + \frac{1}{2}\|z\|_Z^2 + \langle z, u\rangle_Z \tag{3.7}$$

where $u \in Z^*$.

But before we introduce this generalization in detail, let us come back to our functional $F : x \longrightarrow (g(Tx) + f(x)) \in \Gamma(X)$ which we want to minimize. Assuming that there is a point $u \in X$ where F is finite and continuous we can apply the sum rule as well as the chain rule for subdifferentials to F:

$$\partial F(x) = T^*\partial g(Tx) + \partial f(x), \qquad x \in X.$$

Here T^* denotes the adjoint of T. So, in opposite to problem (3.5), for F it might be difficult to compute the minimizer $\bar{x}$ by the optimality condition $0 \in \partial F(\bar{x})$ as it depends on both the subdifferentials of f and g linked with T and T^*. To avoid this, it is a common approach to utilize the relation between (**P**) and its so called Fenchel dual problem, which will be defined in the following.

Theorem and Definition 3.2.11. *For* $h : Z \to \overline{\mathbb{R}}$ *the function*

$$h^* : Z^* \to \overline{\mathbb{R}}, \qquad h^*(z^*) := \sup_{z \in Z} \langle z, z^* \rangle_Z - h(z),$$

is called conjugate or Fenchel conjugate *of h. It has the following properties:*

(i) h^* is convex and lower semicontinuous in Z^*. Moreover, if $h \in \Gamma(Z)$, then it follows that $h^* \in \Gamma(Z^*)$.

(ii) Let h be proper, then $h^{**} = h$ holds true if and only if $h \in \Gamma(Z)$.

(iii) If h is proper and convex, ∂h^* coincides with the inverse of ∂h, i.e.:

$$z \in \partial h(z^*) \Leftrightarrow z^* \in \partial h^*(z).$$

Proof. Writing the epigraph $\operatorname{epi} h^* = \cap_{z \in Z} \operatorname{epi} l_z$ as the intersection of the continuous, linear functions $l_z : z^* \mapsto \langle z, z^* \rangle_Z - h(z)$ we see that it is closed and convex, which prooves the first part of (i). For the second part of (i) as well as assertions (ii) and (iii) see Corollary 1.4, Theorem 1.4 and Proposition 2.1 in Chap. 2 of [5], respectively. □

Example 3.2.12. For $r \in (1, \infty)$ consider the function $h_{r,Z}(z) = \frac{1}{r}\|z\|_Z^r \in \Gamma(Z)$. With the help of Theorem 3.2.7 we obtain that $\|z^*\|_{Z^*}^{\frac{1}{r-1}}$ minimizes the function $\lambda \mapsto \frac{1}{r}\lambda^r - \lambda\|z^*\|_{Z^*}$ given on $[0, \infty)$ and thus the conjugate of $h_{r,Z}$ reads as

$$h_{r,Z}^*(z^*) = \sup_{z \in Z} \left(\langle z, z^* \rangle_Z - \frac{1}{r}\|z\|_Z^r \right) = \sup_{z \in Z} \left(\|z\|_Z \|z^*\|_{Z^*} - \frac{1}{r}\|z\|_Z^r \right) = \frac{1}{r^*}\|z^*\|_{Z^*}^{r^*} = h_{r^*,Z^*}(z^*)$$

where $r^* := \frac{r}{r-1}$ is the conjugate exponent of r.

3.2.3 Extremal relations and saddle point problem

Now let us introduce the perturbation function

$$H : X \times Y \to \overline{\mathbb{R}}, \quad (x, y) \mapsto g(Tx - y) + f(x)$$

which embeds our problem (**P**) given by $\operatorname{argmin}_{x \in X} H(x, 0)$ into a family of problems:

$$\operatorname*{argmin}_{x \in X} H(x, y) \qquad y \in Y.$$

Obviously, we have $H \in \Gamma(X \times Y)$. Calculating the conjugate of H

$$\begin{aligned} H^*(x^*, p) &= \sup_{x\in X, y\in Y} \{\langle x, x^*\rangle_X + \langle y, p\rangle_Y - g(Tx - y) - f(x)\} \\ &= \sup_{x\in X, w\in Y} \{\langle x, x^*\rangle_X + \langle w, (-p)\rangle_Y - g(w) - f(x) + \langle Tx, p\rangle_Y\} \\ &= \sup_{x\in X} \{\langle x, x^* + T^*p\rangle_X - f(x)\} + g^*(-p) = f^*(x^* + T^*p) + g^*(-p), \end{aligned}$$

we call the maximization problem

$$\operatorname*{argmax}_{p\in Y^*} -H^*(0, p) = \operatorname*{argmax}_{p\in Y^*} (-f^*(T^*p) - g^*(-p)) \quad (\mathbf{D}), \tag{3.8}$$

the *dual problem* of the *primal problem* (**P**) with respect to H. First of all, note that, due to

$$H^*(0, p) \geq \langle 0, p\rangle_Y - H(x, 0) = -(g(Tx) + f(x)) \quad \text{for all } p \in Y^*, x \in X,$$

the extremal value $m_{(\mathbf{P})} := \inf_{x\in X} g(Tx) + f(x)$ in the primal problem (**P**) is always greater or equal to the extremal value $M_{(\mathbf{D})} := \sup_{p\in Y^*} -f^*(T^*p) - g^*(-p)$ in the dual problem (**D**). This relation $M_{(\mathbf{D})} \leq m_{(\mathbf{P})}$ is referred to as *weak duality*. In the case of *strong duality* $m_{(\mathbf{P})} = M_{(\mathbf{D})}$ one obtains the following crucial relationship between solutions to (**P**) and (**D**):

Theorem 3.2.13. *Let there are at least one point* $x_0 \in X$ *and one point* $p_0 \in Y^*$ *where the function values* $f(x_0)$, $g(Tx_0)$, $f^*(T^*p_0), g^*(-p_0)$ *are finite. We have:*

- *If in addition* g *is continuous at* Tx_0, *then the dual problem* (**D**) *has a solution and strong duality* $m_{(\mathbf{P})} = M_{(\mathbf{D})}$ *holds.*
- *If in addition* f^* *is continuous at* T^*p_0, *then the primal problem* (**P**) *has a solution and strong duality* $m_{(\mathbf{P})} = M_{(\mathbf{D})}$ *holds.*
- $\bar{x} \in X$ *is a solution to the primal problem* (**P**) *and* $\bar{p} \in Y^*$ *a solution to the dual problem* (**D**) *and strong duality* $g(T\bar{x}) + f(\bar{x}) = m_{(\mathbf{P})} = M_{(\mathbf{D})} = -f^*(T^*\bar{p}) - g^*(-\bar{p})$ *holds if and only if*
$$T^*\bar{p} \in \partial f(\bar{x}), \qquad -\bar{p} \in \partial g(T\bar{x}). \tag{3.9}$$

Proof. Theorem 2.2., Lemma 2.2. and Theorem 2.4. in Chap. 3 of [5] □

Moreover, solutions to (**P**) and (**D**) build the saddle point of the *Lagrange function* (also referred to as Lagrangian):

$$L(x, p) = -(H(x, \cdot))^*(p) = -\sup_{y\in Y} \{\langle y, p\rangle_Y - g(Tx - y) - f(x)\} = f(x) - \langle T\,x, p\rangle_Y - g^*(-p)$$

associated to (**P**) and (**D**) via

$$\sup_{p\in Y^*} L(x, p) = g(Tx) + f(x), \quad \sup_{x\in X} L(x, p) = -f^*(T^*p) - g^*(-p).$$

Theorem 3.2.14. *Under the assumptions of Theorem 3.2.13, the following assertions are equivalent:*

- $\bar{x} \in X$ *is a solution to the primal problem* (**P**) *and* $\bar{p} \in Y^*$ *a solution to the dual problem* (**D**) *and strong duality* $g(T\bar{x}) + f(\bar{x}) = m_{(\mathbf{P})} = M_{(\mathbf{D})} = -f^*(T^*\bar{p}) - g^*(-\bar{p})$ *holds.*
- $(\bar{x}, \bar{p}) \in X \times Y^*$ *is a* saddle point *of* L*, i.e.*

$$L(x, \bar{p}) \leq L(\bar{x}, \bar{p}) \leq L(\bar{x}, p), \quad \textit{for all } x \in X,\ p \in Y^*. \tag{3.10}$$

If one of these assertions holds we obtain the minmax equality

$$m_{(\mathbf{P})} = M_{(\mathbf{D})} = \min_{x\in X} \max_{p\in Y^*} L(x, p) = \max_{p\in Y^*} \min_{x\in X} L(x, p).$$

Proof. see [5, Theorem 2.7 in Chap. 3]. □

More precisely, L has a saddle point $(\bar{x}, \bar{p})$ if and only if L satisfies the minmax equality. Therefore the saddle point problem of finding a pair $(\bar{x}, \bar{p}) \in X \times Y^*$ obeying (3.10) can be formulated as

$$\operatorname*{argmin}_{x\in X} \operatorname*{argmax}_{p\in Y^*} - \langle Tx, p\rangle_Y + f(x) - g^*(-p) \quad (\mathbf{S}).$$

Characterizing a solution $\bar{x}$ to (**P**) by the extremal relations (3.9) which are because of Assertion 3.2.11 (iii) equivalent to

$$T^*\bar{p} \in \partial f(\bar{x}), \qquad T\bar{x} \in \partial g^*(-\bar{p}) \tag{3.11}$$

has the advantage that $\bar{x}$ is connected to the dual solution $\bar{p}$ via rather simple equations. This fact is also used by the primal-dual algorithm (CP) of Chambolle and Pock, studied in Chapter 4, which tries to solve the primal and the dual problem simultaneously. In order to generalize this method from Hilbert and Banach spaces X and Y^*, where X and X^* as well as Y and Y^* not necessarily coincide, we will need some concepts of the Banach spaces and the relations to their duals.

3.3 Duality mappings and Bregman distance

The following definitions and results can be found in Chapter I and II of [23]. First we make some assumptions on the geometry of the Banach spaces X and Y.

Definition 3.3.1. Consider a real Banach space Z. The function $\delta_Z : [0, 2] \to [0, 1]$,

$$\delta_Z(\epsilon) := \inf \left\{ 1 - \left\| \frac{1}{2} (z + u) \right\|_Z \;\middle|\; z, u \in Z,\ \|z\|_Z = \|u\|_Z = 1, \|z - u\|_Z \geq \epsilon \right\}$$

is called the *modulus of convexity* and $\rho_Z : [0,\infty) \to \mathbb{R}$,

$$\rho_Z(\tau) := \frac{1}{2} \sup \{ \|z+u\|_Z - \|z-u\|_Z - 2 \mid z,u \in Z,\ \|z\|_Z = 1, \|u\|_Z \leq \tau \}$$

is the *modulus of smoothness*. Then Z is said to be

- *strictly convex* if for any $z, u \in Z$ with $z \neq u$ and $\|z\|_Z = \|u\|_Z = 1$ we have $\|z+u\|_Z < 2$.
- *uniformly convex* if $\delta_Z(\epsilon) > 0$ for any $\epsilon \in (0,2]$.
- *$r-$convex* with $r \in (1,\infty)$ (also referred to as *the modulus of convexity is of power type r*) if there exists a constant $C > 0$ such that
$$\delta_Z(\epsilon) \geq C\,\epsilon^r, \quad \epsilon \in [0,2].$$
- *smooth* if for any $z \in Z \setminus \{0\}$ there exists a unique $z^* \in Z^*$ with $\|z^*\|_{Z^*} = 1$ such that $\langle z, z^* \rangle_Z = \|z\|_Z$.
- *uniformly smooth* if $\frac{\rho_Z(\tau)}{\tau} \to 0$ as $\tau \to 0$.
- *$r-$smooth* with $r > 1$ (also referred to as *the modulus of smoothness is of power type r*) if there exists a constant $G > 0$ such that
$$\rho_X(\tau) \leq G\,\tau^r, \quad \tau \in [0,\infty).$$

Note that $r-$convexity with $r \in (1,\infty)$ yields uniform convexity, which in turn implies strict convexity. Conversely, on a finite dimensional Banach space Z the continuous function $(z,u) \mapsto 1 - \left\|\frac{1}{2}(x+u)\right\|_Z$ attains its minimum on the compact set $\mathbb{S}_Z \times \mathbb{S}_Z$ of unit spheres $\mathbb{S}_Z := \{z \in Z \mid \|z\|_Z = 1\}$ so that Z is strictly convex if and only if it is uniformly convex.

Example 3.3.2. (i) If Z is a Hilbert space the parallelogram law

$$\|z+u\|_Z^2 + \|z-u\|_Z^2 = 2(\|z\|_Z^2 + \|u\|_Z^2),$$

yields for any $\epsilon \in [0,2]$

$$\delta_Z(\epsilon) = \inf \left\{ 1 - \sqrt{\frac{1}{2}\left(\|z\|_Z^2 + \|u\|_Z^2\right) - \frac{1}{4}\|z-u\|_Z^2} \;\middle|\; \|x\|_Z = \|u\|_Z = 1, \|x-u\|_Z \geq \epsilon \right\}$$
$$= 1 - \sqrt{1 - \frac{\epsilon^2}{4}} \geq \frac{\epsilon^2}{8},$$

and hence Z is 2-convex.

(ii) The reflexive Banach spaces l^r, L^r and $W^{1,r}$ with $r \in (1, \infty)$ are $\max\{r, 2\}$-convex and $\min\{r, 2\}$-smooth, cf. [85]. One can easily check that the same holds for the weighted sequence space l^r_W with positive weight $W = (w_j)_{j \in \mathbb{N}}$ and norm given by

$$\|x\|_{l^r_W} := \left(\sum_j w_j \, |x_j|^r \right)^{\frac{1}{r}} = \left\| \left(w_j^{\frac{1}{r}} x_j \right)_j \right\|_{l^r} .$$

Observe that in the last example no Banach space has a modulus of convexity of power type q smaller than 2. In fact, there is no Banach space at all with this property as the following theorem states:

Theorem 3.3.3. *The Banach space Z has a modulus of convexity of power type $r \in (1, \infty)$ if and only if its dual space Z^* has a modulus of smoothness of power type r^* where $r^* \in (1, \infty)$ denotes the* conjugate exponent *satisfying*

$$\frac{1}{r} + \frac{1}{r^*} = 1.$$

Moreover, if Z is r-convex, then it is reflexive and r is greater or equal to 2.

Proof. The first assertion follows by the Lindenstrauss duality theory, see [54, Sec. 1.e]. For the second see [23, Corollary 2.15 in Chapter II] and [85, p. 193]. □

In particular in the 'symmetric case' $r = r^* = 2$ an r-convex Banach space Z takes characteristics similar to Hilbert space properties which will play an important role for the generalization of CP to Banach spaces. So, in the following, we assume both X and the dual space Y^* to be (reflexive,) smooth and 2-convex Banach spaces. Because of the last theorem and [23, Theorem 1.3 in Chap. II], here the second statement is equivalent to the condition that Y is a reflexive, strictly convex and 2-smooth Banach space. Note that by Example 3.3.2 (ii), these conditions are consistent with preferable choices of the preimage and image space X and Y for the phase retrieval problems introduced in Chapter 1. So, based on the assumptions for the phase retrieval problem, we are now able to specify the inner minimization problem in the IRNM 3.3, rewritten as (**P**), which we consider within this work:

Problem 3.3.4. *For a linear, bounded operator $T : X \to Y$ given on spaces X and Y let us consider the convex optimization problem*

$$\bar{x} = argmin_{x \in X} \left(g(Tx) + f(x) \right) \quad (\mathbf{P}), \tag{3.12}$$

where:

- *X is a reflexive, smooth and 2-convex Banach space,*
- *Y is a reflexive, strictly convex and 2-smooth Banach space,*
- *$f \in \Gamma(X)$, $g \in \Gamma(Y)$,*

- *there exists at least one point $x_0 \in X$ and one point $p_0 \in Y^*$ such that the function values $f(x_0)$, $g(Tx_0)$, $f^*(T^*p_0), g^*(-p_0)$ are finite and g is continuous at Tx_0 and f^* is continuous at T^*p_0.*

Now let us introduce the so-called duality mappings $J_{q,Z}$ which give useful relations between a Banach space Z and its dual Z^*:

Definition 3.3.5. Let Z be a Banach space. For $q \in (1, \infty)$ the (set-valued) mapping

$$J_{q,Z} : Z \rightrightarrows Z^*, \quad J_{q,Z}(x) := \left\{ z^* \in Z^* \mid \langle z, z^* \rangle_Z = \|z\|_Z \, \|z^*\|_{Z^*} \, , \|z^*\|_{Z^*} = \|z\|_Z^{q-1} \right\}$$

is called *duality mapping* with respect to the weight function $t \mapsto t^{q-1}$.

In the case $q = 2$, $J_{2,Z}$ is also referred to as *normalized duality mapping* and we use the short notation $J_Z = J_{2,Z}$. From Proposition 3.2.9 we already know that the normalized duality mapping J_Z is given by the subdifferential of the norm monomial $z \mapsto \frac{1}{2}\|z\|_Z^2$. For arbitrary $q \in (1, \infty)$ it follows from the theorem of Asplund ([3]) that $J_{q,Z}$ coincides with the subdifferential $\partial h_{q,Z}$ of $h_{q,Z}(z) := \frac{1}{q}\|z\|_Z^q$. In addition to this characterization, by the next theorem we summarize some important properties of the duality mappings:

Theorem 3.3.6. *For any (real) Banach space Z and $q \in (1, \infty)$ we have:*

(i) If Z is smooth, then $J_{q,Z}$ is single-valued.

(ii) Z is reflexive if and only if $J_{q,Z}$ is surjective.

(iii) If Z is strictly convex, then $J_{q,Z}$ is strictly monotone and hence injective.

(iv) If Z is 2-convex and smooth, then $J_{q,Z}$ is bijective with inverse $J_{q^*,Z^*} : Z^* \to Z^{**} = Z$ where $q^* \in (1, \infty)$ denotes the conjugate exponent of q, i.e. $\frac{1}{q} + \frac{1}{q^*} = 1$.

(v) If Z is a Hilbert space, then the normalized duality mapping J_Z is the identity. Moreover, Z is a Hilbert space if and only if J_Z is linear.

(vi) $J_{q,Z}(z) = \partial h_{q,Z}$, where $h_{q,Z}(z) := \frac{1}{q}\|z\|_Z^q$.

(vii) If Z is q-convex, as a consequence of the Xu-Roach inequalities *there exists a constant $C_Z > 0$ such that the following inequality holds for all $u, z \in Z$:*

$$\frac{1}{q}\|z-u\|_Z^q \geq \frac{1}{q}\|z\|_Z^q - \left\langle u, J_{q,Z}(z) \right\rangle_Z + \frac{C_Z}{q}\|u\|_Z^q \tag{3.13}$$

Proof. Assertions (i) – (iii) are given by Corollary 4.5 in Chap. I, Theorem 3.4 in Chap. II and Corollary 1.9 in Chap. II of [23], respectively. Assertion (iv) follows from the previous ones together with Corollary 3.5 in Chap. II of [23]. For (v) see Proposition 4.8 in Chap. I of [23]. (vi) is a special case of the Theorem of Asplund [3] and (vii) is a consequence of the Xu-Roach inequalities ([85]). □

So, under the regularity assumptions of Problem 3.3.4 on the Banach spaces X and Y the corresponding duality mappings are quite well-behaved as well, but in general nonlinear. Due to $(iv) - (v)$ we can interpret $J_{q,Z}$ as the generalization of the identity mapping to this Banach space setting. From this point of view inequality (3.13) generalizes the polarization identity

$$\frac{1}{2}\|z-u\|_Z^2 = \frac{1}{2}\|z\|_Z^2 - \langle z,u\rangle_Z + \frac{1}{2}\|u\|_Z^2, \quad x,u \in Z \tag{3.14}$$

which characterizes a Hilbert space Z. The following example illustrates the duality mapping's nonlinearity at one side and their strong dependence on norm monomials on the other. As remarked in [70], by the theorem of Asplund 3.3.6 (vi) we can expect the numerical computation of a duality mapping at some point to be of a comparable complexity as the evaluation of the corresponding norm to a certain power.

Example 3.3.7. For the reflexive, smooth and strictly convex Banach space $Z = l^r$ with $r \in (1,\infty)$ the dual space with respect to standard scalar product is $Z^* = l^{r^*}$. From the last theorem we derive the duality mapping to be given as:

$$J_{q,l^r} : l^r \to l^{r^*}, \qquad J_{q,l^r}(x) = \frac{1}{\|x\|_{l^r}^{r-q}} |x|^{r-1}\,\mathrm{sign}(x), \qquad \text{for } q \in (1,\infty)$$

which has to be understood componentwise.

Considering again the symmetric case $q = q^* = 2$ and a reflexive, smooth, strictly convex Banach space Z, then for $C_Z = 1$ the right hand side of (3.13) becomes the so-called *Bregman distance*

$$\mathcal{B}_Z(u,z) := \frac{1}{2}\|u\|_Z^2 - \frac{1}{2}\|z\|_Z^2 - \langle u-z, J_Z(z)\rangle_Z = \frac{1}{2}\|u\|_Z^2 + \frac{1}{2}\|z\|_Z^2 - \langle u, J_Z(z)\rangle_Z\,,$$

with gauge function $h_{2,Z} : z \mapsto \frac{1}{2}\|z\|_Z^2$ (see Figure 3.3 for an illustration). So, with respect to the polarization identity, this Bregman distance behaves more like a Hilbert space norm squared than the functional $\Phi_Z(u,z) \mapsto \frac{1}{2}\|u-z\|_Z^2$. Obviously, if Z is a Hilbert space $\mathcal{B}_Z$ coincides with Φ_Z. Since $h_{2,Z} : z \mapsto \frac{1}{2}\|z\|_Z^2$ is strictly convex and we have $J_Z = \partial h_{2,Z}$, $\mathcal{B}_Z$ has also the following metric property in common with Φ_Z:

$$\mathcal{B}_Z(u,z) \geq 0, \quad \text{where equality holds if and only if } u = z. \tag{3.15}$$

Note that $\mathcal{B}_Z$ is not a metric as symmetry is not fulfilled. Nevertheless, a kind of symmetry with respect to the duals holds true:

$$\mathcal{B}_{Z^*}\left(J_Z(v), J_Z(x)\right) = \mathcal{B}_Z(x,v)\,. \tag{3.16}$$

Moreover, the Bregman distance associated with $h_{2,Z}$ satisfies the *three-point identity*:

$$\begin{aligned} \mathcal{B}_Z(u,x) + \mathcal{B}_Z(v,u) &= \frac{1}{2}\|v\|_Z^2 - \frac{1}{2}\|x\|_Z^2 - \langle u-x, J_Z(x)\rangle_Z - \langle v-u, J_Z(u)\rangle_Z \\ &= \mathcal{B}_Z(v,x) + \langle v-u, J_Z(x) - J_Z(u)\rangle_Z\,, \quad x,u,v \in Z. \end{aligned} \tag{3.17}$$

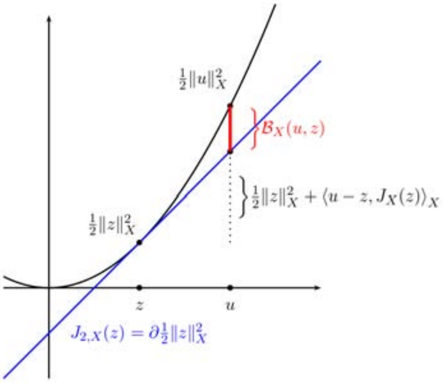

$\frac{1}{2}\|u\|_X^2$
$\mathcal{B}_X(u,z)$
$\frac{1}{2}\|z\|_X^2 + \langle u-z, J_X(z)\rangle_X$
$\frac{1}{2}\|z\|_X^2$
z
u
$J_{2,X}(z) = \partial\frac{1}{2}\|z\|_X^2$

Figure 3.3: Illustration of the Bregman distance with the gauge function $h_{2,Z} : z \mapsto \frac{1}{2}\|z\|_Z^2$

Although we will mainly focus on $\mathcal{B}_Z$, we also would like to introduce the more general class of Bregman distances associated with $h_{q,Z}$ where we again restrict our considerations on the case of single-valued, bijective duality mappings:

Definition 3.3.8. Let Z be a reflexive, smooth and strictly convex Banach space. For a parameter $q \in (1, \infty)$ the function

$$\mathcal{B}_{q,Z}(u,z) := \frac{1}{q}\|u\|_Z^q - \frac{1}{q}\|z\|_Z^q - \left\langle u - z, J_{q,Z}(z)\right\rangle_Z = \frac{1}{q}\|u\|_Z^q + \frac{1}{q^*}\|z\|_Z^q - \langle u, J_Z(z)\rangle_Z$$

is called *Bregman distance* with gauge function $h_{q,Z} : z \mapsto \frac{1}{q}\|z\|_Z^q$.

It is easy to check that for any $q \in (1, \infty)$ the Bregman distance $\mathcal{B}_{q,Z}$ satisfies the properties (3.15), (3.16), (3.17) as well.

Proposition 3.3.9. *Let Z be a Banach space which is smooth and convex of power type $q \in (1, \infty)$. Then there exists a constant $C_Z > 0$ such that inequality* (3.13) *holds and we have:*

$$\mathcal{B}_{q,Z}(u,z) \geq \frac{C_Z}{q}\|u - z\|_Z^q, \qquad z, u \in Z \tag{3.18}$$

Proof. In order to show that the constant C_Z in (3.18) is the same as in (3.13) we repeat the proof given in [16]: The assertion directly follows by substituting (3.13) for $u = z - v$ and arbitrary $v, z \in Z$ into the Bregman distance:

$$\mathcal{B}_{q,Z}(v,z) = \frac{1}{q}\|z - (z - v)\|_Z^q - \frac{1}{q}\|z\|_Z^q - \left\langle v - z, J_{q,Z}(z)\right\rangle_Z \geq \frac{C_Z}{q}\|z - v\|_Z^q.$$

□

Example 3.3.10. The connection of the last proposition to inequality (3.13) helps to find optimal constants C_Z. For example, in [84] it is shown that for the 2-convex Banach space $Z = l^r$ with $r \in (1, 2]$ the inequality (3.13) holds for any constant $C_Z < r - 1$.

Note that due to Theorem 3.3.3 the inequality (3.18) is only stated for $q \geq 2$. The assumption of Problem 3.3.4 that X and Y^* are reflexive, smooth and 2-convex Banach spaces provides that there exist positive constants C_X and C_{Y^*}, such that the inequalities:

$$\mathcal{B}_X(x,u) \geq \frac{C_X}{2}\|x-u\|_X^2\,, \qquad \text{and} \qquad \mathcal{B}_{Y^*}(y^*,p) \geq \frac{C_{Y^*}}{2}\|y^*-p\|_{Y^*}^2, \tag{3.19}$$

hold for all $x, u \in X$ and all $y^*, p \in Y^*$. Consequently, we obtain the following inequality which plays a key role in our convergence analysis of the generalized Chambolle-Pock algorithm:

Proposition 3.3.11. *Under the assumptions of Problem 3.3.4 we have for any $x, u \in X$, any $y^*, p \in Y^*$ and any positive constant α:*

$$\begin{aligned} &\left|\langle T(x-u), p-y^*\rangle_Y\right| \\ &\quad \leq \|T\|\left(\frac{\alpha\,\min\left\{\mathcal{B}_X(x,u), \mathcal{B}_X(u,x)\right\}}{C_X} + \frac{\min\left\{\mathcal{B}_{Y^*}(p,y^*), \mathcal{B}_{Y^*}(y^*,p)\right\}}{\alpha\, C_{Y^*}}\right), \end{aligned} \tag{3.20}$$

where $\|T\| = \max\left\{\|Tx\|_Y \mid x \in X, \|x\|_X = 1\right\}$ *denotes the operator norm.*

Proof. Applying Cauchy-Schwarz's inequality as well as the special case of Young's inequality:

$$ab \leq \frac{\alpha\, a^2}{2} + \frac{b^2}{2\,\alpha}, \qquad a, b \geq 0 \tag{3.21}$$

with $a := \|x_k - x_{k-1}\|_X$, and $b := \|p_{k+1} - p_k\|_{Y^*} \in \mathbb{R}$ leads to

$$\left|\langle T(x-u), p-y^*\rangle_Y\right| \leq \|T\|\,\|x-u\|_X\,\|p-y^*\|_{Y^*} \leq \|T\|\left(\alpha\,\frac{\|x-u\|_X^2}{2} + \frac{\|p-y^*\|_{Y^*}^2}{2\,\alpha}\right).$$

Now the inequalities (3.19) give the assertion. □

The last proposition is also the reason for restricting Problem 3.3.4 to 2-convex Banach spaces X and Y^* which of course already covers a lot of interesting problems of the form (**P**). We observe that the corresponding proof only works with the use of Young's inequality

$$ab \leq \frac{\alpha\, a^q}{q} + \frac{b^{q^*}}{q^*\,\alpha^{q^*-1}}, \qquad a, b \geq 0, \alpha > 0$$

for the symmetric case $q = 2 = q^*$. In the general case $q \in (1,\infty)\backslash\{2\}$ either q or the conjugate exponent q^* must be smaller than 2 such that inequality (3.18) can not be applied. However, here the question arises if under more general conditions on the moduli of convexity of X and Y^* (which are still assumed to be reflexive, smooth and strictly convex) the following more general form of inequality (3.20) holds true for some $q \in (1,\infty)$, some $r > 0$ and some positive constants C_X, C_{Y^*}:

$$\begin{aligned} &\left|\langle T(x-u), p-y^*\rangle_Y\right| \\ &\leq \|T\|^r\left(\frac{\min\left\{\mathcal{B}_{q,X}(x,u), \mathcal{B}_{q,X}(u,x)\right\}}{C_X} + \frac{\min\left\{\mathcal{B}_{q^*,Y^*}(p,y^*), \mathcal{B}_{q^*,Y^*}(y^*,p)\right\}}{C_{Y^*}}\right), \end{aligned} \tag{3.22}$$

for all $x, u \in X$, $y^*, p \in Y^*$. Also in the case of X and Y^* being 2-convex, it would be helpful if (3.22) is satiesfied for some $q \in (1, \infty) \backslash \{2\}$. Although making additional assumptions on the operator T would be also conceivable, we study the validity of (3.22) for a special operator T, namely the identity (and thus $X = Y$), since $T = I$ should be sufficiently well-behaved in any case. The three-point identity (3.17) shows that, if $\left\langle v - w, p - J_{q,Y}(w) \right\rangle_Y$ is nonnegative for some points $v, w \in Y$ and $p \in Y^*$, the following inequality holds:

$$\left| \left\langle v - w, p - J_{q,Y}(w) \right\rangle_Y \right| \leq \mathcal{B}_{q,Y}(v, w) + \mathcal{B}_{q^*,Y^*}(p, w).$$

So, in this special case the validity of symmetry properties of the form

$$\mathcal{B}_{q,Y}(v, w) \leq C^{\text{sym}}_{q,Y} \mathcal{B}_{q,Y}(w, v) \quad \text{and} \quad \mathcal{B}_{q^*,Y^*}(p, y^*) \leq C^{\text{sym}}_{q^*,Y^*} \mathcal{B}_{q^*,Y^*}(y^*, p), \tag{3.23}$$

for some positive constants $C_1^{\text{sym}}, C_2^{\text{sym}} > 0$ and $v, w \in Y, p, y^* \in Y^*$ would give the assertion. However, it is not clear if there exists a constant $q \in (1, \infty)$ and a non-Hilbertian Banach space Y such that this condition is fulfilled by $\mathcal{B}_{q,Y}$ and $\mathcal{B}_{q^*,Y^*}$, respectively. Let us assume, for a moment, that for some Y and some q the Bregman distances satisfy (3.23) as well as the following triangle property for some $C^{\text{tria}}_{q,Y}$, $C^{\text{tria}}_{q^*,Y^*} > 0$:

$$\mathcal{B}_{q,Y}(v, y) - \mathcal{B}_{q,Y}(v, u) \leq C_1^{\text{tria}} \mathcal{B}_{q,Y}(u, y), \quad \mathcal{B}_{q,Y}(v, y) - \mathcal{B}_{q,Y}(u, y) \leq C_2^{\text{tria}} \mathcal{B}_{q,Y}(v, u),$$

on suitable subsets $U \subseteq Y$ and $P \subseteq Y^*$. Then the well-known four-point identity (see e.g. [8])

$$\begin{aligned} \langle v - w, p - y^* \rangle_Y = {} & \mathcal{B}_{q,Y}\left(v, J_{q^*,Y^*}(y^*)\right) - \mathcal{B}_{q,Y}\left(v, J_{q^*,Y^*}(p)\right) \\ & + \mathcal{B}_{q,Y}\left(w, J_{q^*,Y^*}(p)\right) - \mathcal{B}_{q,Y}\left(w, J_{q^*,Y^*}(y^*)\right) \quad v, w \in Y,\ p, y^* \in Y^*, \end{aligned}$$

yields

$$\begin{aligned} & \left(\frac{1}{2} + \frac{1}{2}\right) |\langle v - w, p - y^* \rangle_Y| \\ & \quad \leq \frac{C_1^{\text{tria}}}{2} \mathcal{B}_{q,Y}\left(J_{q^*,Y^*}(p), J_{q^*,Y^*}(y^*)\right) + \frac{C_1^{\text{tria}}}{2} \mathcal{B}_{q,Y}\left(J_{q^*,Y^*}(y^*), J_{q^*,Y^*}(p)\right) \\ & \qquad + \frac{C_2^{\text{tria}}}{2} \mathcal{B}_{q,Y}(v, w) + \frac{C_2^{\text{tria}}}{2} \mathcal{B}_{q,Y}(w, v) \\ & \quad = \frac{C_1^{\text{tria}}}{2} \mathcal{B}_{q^*,Y^*}(y^*, p) + \frac{C_1^{\text{tria}}}{2} \mathcal{B}_{q^*,Y^*}(p, y^*) + \frac{C_2^{\text{tria}}}{2} \mathcal{B}_{q,Y}(v, w) + \frac{C_2^{\text{tria}}}{2} \mathcal{B}_{q,Y}(w, v) \\ & \quad \leq \frac{C_1^{\text{tria}}\left(C^{\text{sym}}_{q^*,Y^*} + 1\right)}{2} \min\left\{\mathcal{B}_{q^*,Y^*}(y^*, p), \mathcal{B}_{q^*,Y^*}(p, y^*)\right\} \\ & \qquad + \frac{C_2^{\text{tria}}(C^{\text{sym}}_{q,Y} + 1)}{2} \min\left\{\mathcal{B}_{q,Y}(v, w), \mathcal{B}_{q,Y}(w, v)\right\}. \end{aligned}$$

for all $v, w \in U, p, y^* \in P$. This fuels the hope that there exist at least suitable subsets $U \subseteq X$ and $P \subseteq Y^*$ on which (3.22) is satisfied for some $q \neq 2$ or non-2-convex Banach spaces X and Y^*. Unfortunately, we did not succeed in proving this conjecture.

Let us come back to the concept of resolvents on Hilbert spaces which serves for a whole class of convex optimization methods for solving problems of the form (**P**) as a basic tool (cf. [25], [11]). Also the algorithm of Chambolle and Pock (CP, [21]) relies on the efficient evaluation of resolvents. As mentioned above, for its generalization to Banach spaces we will need a generalization of (3.6) as well. For this purpose we consider the canonical generalization (3.7) of the right hand side of (3.6). Note that under the use of the Bregman distance B_Z it reads as:

$$\operatorname*{argmin}_{z\in Z} \tau\, h(z) + B_Z(z, J_{Z^*}(u)),$$

where $h \in \Gamma(Z)$, $\tau > 0$ and $u \in Z^*$. Obviously, B_Z is proper and convex with respect to the first argument such that Theorem 3.2.3 ensures the uniqueness of a solution to (3.6). Assuming that there is a point in Z where $h \in \Gamma(Z)$ is finite and continuous, Theorem 3.2.7 together with the sum rule and Theorem 3.3.6 (*vi*) characterizes the solution $\bar{z}$ by

$$0 \in (\tau\partial h + J_Z)(\bar{z}) - u.$$

Since Rockafellar proved ([65], Proposition 1) that the operator $(\tau\partial h + J_Z)^{-1} : Z^* \to Z$ is well-defined and single-valued on any reflexive Banach space Z, our generalized resolvent is given by

$$(\tau\,\partial h + J_Z)^{-1}(u) = \operatorname*{argmin}_{z\in Z} \tau\, h(z) + B_Z(z, J_{Z^*}(u)). \tag{3.24}$$

In particular, under the assumptions of Problem 3.3.4 due to Theorem 3.2.11, the generalized resolvents of f and g :

$$(\tau\,\partial f + J_X)^{-1}(x^*), \quad (\sigma\,\partial g^* + J_{Y^*})^{-1}(y) \tag{3.25}$$

are well-defined and single valued for any $x^* \in X^*$, $y \in Y$ and any parameters $\tau, \sigma > 0$. In Section 4.3 we will study these operators in more detail.

4 Generalization of the Chambolle-Pock's algorithm to Banach spaces

Based on our preliminary considerations of concepts in convex analysis and of duality theory for linear Tikhonov-type functionals

$$x \mapsto S(y^\delta; Tx) + \alpha R(x),$$

in the following section we now introduce the first-order primal-dual algorithm of Chambolle and Pock (CP , [21]) for solving such minimization Problems 3.3.4 on Hilbert spaces X, Y^*. In Section 4.2 we then present a generalization (CP-BS) of CP for the given Banach space setting of Problem 3.3.4 and prove corresponding convergence results. The generalized resolvents (3.25) on which the algorithm CP-BS relies on are the topic of Section 4.3. The following sections are an extended version of the work that we published in [45].

4.1 Chambolle-Pock's first-order primal-dual algorithm

Recall from Theorem 3.2.13 that the saddle point problem (**S**) associated with Problem 3.3.4 is solvable and a solution pair $(\bar{x}, -\bar{p}) \in X \times Y^*$ is characterized by (3.11):

$$T\bar{x} \in \partial g^*(\bar{p}), \qquad -T^*\bar{p} \in \partial f(\bar{x}). \tag{4.1}$$

In order to motivate CP on Hilbert spaces X and Y we rewrite these extremal relations for some parameters $\tau, \sigma > 0$ as:

$$\begin{aligned}
(\bar{p} + \sigma T\bar{x}) - \bar{p} &\in \sigma\partial g^*(\bar{p}) & \Leftrightarrow \qquad & \bar{p} = (\sigma\partial g^* + I)^{-1}(\bar{p} + \sigma T\bar{x}) \\
(\bar{x} - \tau T^*\bar{p}) - \bar{x} &\in \tau\partial f(\bar{x}) & \Leftrightarrow \qquad & \bar{x} = (\tau\partial f + I)^{-1}(\bar{x} - \tau T^*\bar{p}).
\end{aligned}$$

So, we have derived a relation between a solution $\bar{x}$ to the problem (**P**) and the solution $-\bar{p}$ to its dual problem (**D**) that base on the resolvents of f and g^*. Under the reasonable assumption that evaluating these resolvents has a comparable complexity as the evaluation of the operator T, the algorithm CP combines these both equations together with an over-relaxation step $\hat{x}_{k+1} := x_{k+1} + \theta_k(x_{k+1} - x_k)$ for $\theta_k \in [0, 1]$ to the following iterative scheme:

Algorithm 1 (CP, [21]). *For* $(\tau_k, \sigma_k)_{k\in\mathbb{N}} \subseteq (0, \infty)$, $(\theta)_{k\in\mathbb{N}} \subseteq [0, 1], (x_0, p_0) \in X \times Y^*$, $\hat{x}_0 := x_0$, *set:*

$$p_{k+1} := (\sigma_k \partial g^* + I)^{-1}(p_k + \sigma_k T\hat{x}_k) \tag{4.2}$$

$$x_{k+1} := (\tau_k \partial f + I)^{-1}(x_k - \tau_k T^* p_{k+1}) \tag{4.3}$$

$$\hat{x}_{k+1} := x_{k+1} + \theta_k(x_{k+1} - x_k) \tag{4.4}$$

As pointed out in [21], line (4.4) serves as an approximation for the desirable implicit choice $\hat{x}_k = x_{k+1}$. Obviously, the algorithm is designed to solve the corresponding saddle point problem (**S**), i.e. to find a solution $\bar{x} \in X$ to the primal problem (**P**) and the corresponding solution $-\bar{p}$ to the dual problem (**D**) simultaneously. Therefore on a bounded subset $B_1 \times B_2 \subset X \times Y$ the objective value can be expressed by the *partial primal-dual gap*:

$$\begin{aligned}\mathcal{G}_{B_1\times B_2}(x,p) &:= \max_{-p'\in B_2} L(x,-p') - \min_{x'\in B_1} L(x,-p)\\ &= \max_{-p'\in B_2} \left(\langle Tx,p'\rangle_Y - g^*(p') + f(x)\right) - \min_{x'\in B_1} \left(\langle Tx',p\rangle_Y - g^*(p) + f(x')\right).\end{aligned}$$

Observe that if there is a saddle point $(\bar{x},-\bar{p})$ in $B_1 \times B_2 \subset X \times Y$ the functional

$$\mathcal{G}_{B_1\times B_2}(x,p) \geq L(x,-\bar{p}) - L(\bar{x},-p)$$

is non-negative for all $(x,p) \in X \times Y$ and vanishes at $(\bar{x},-\bar{p})$. Conversely, one can show that if $\mathcal{G}_{B_1\times B_2}$ vanishes at an interior point $(\bar{x},-\bar{p})$ of $B_1 \times B_2 \subset X \times Y$, then the pair $(\bar{x},-\bar{p})$ solves the saddle point problem (**S**). Moreover, we introduce for $(x,y^*) \in X \times Y^*$ the misfit functional

$$\triangle_k(x,y^*) := \frac{\|y^*-p_k\|^2}{\sigma_k} + \frac{\|x-x_k\|^2}{\tau_k}.$$

With these definitions Chambolle and Pock stated a parameter choice and proved the following convergence result:

Theorem 4.1.1 (Theorem 1 [21])**.** *Suppose that assumptions of Problem 3.3.4 hold true. For some σ,τ with $\sqrt{\sigma\tau}\|T\| < 1$ we choose constant parameters $\sigma_k = \sigma$, $\tau_k = \tau$ and $\theta_k = 1$ in Algorithm 1. Then for the resulting version, denoted as CP 1, the following assertions hold true:*

- *The sequence $(x_k,p_k)_{k\in\mathbb{N}}$ remains bounded in the form that for any solution $(\bar{x},-\bar{p})$ of the saddle point problem* (**S**) *in $X\times Y$ there exists a constant $C < \left(1-\|T\|^2\,\sigma\,\tau\right)^{-1}$ such that for any $N \in \mathbb{N}$:*
$$\triangle_N(\bar{x},\bar{p}) \leq C\,\triangle_0(\bar{x},\bar{p}).$$
- *We define the sequence $\left(x^N,p^N\right)_{N\in\mathbb{N}}$ of mean values $x^N := \frac{1}{N}\sum_{k=1}^N x_k \in X$ and $y^N := \frac{1}{N}\sum_{k=1}^N y_k \in Y$. For any $N \in \mathbb{N}$ and any bounded set $B_1 \times B_2 \subset X \times Y$ the restricted primal-dual gap $\mathcal{G}_{B_1\times B_2}\left(x^N,p^N\right)$ is bounded by*
$$D(B_1,B_2) := \frac{1}{N}\sup_{(x,y^*)\in B_1\times B_2} \triangle_0\left(y^*,x\right).$$
Moreover, for every weak cluster point $(\tilde{x},\tilde{p})$ of the sequence $\left(x^N,p^N\right)_{N\in\mathbb{N}}$, $(\tilde{x},-\tilde{p})$ solves the saddle point problem (**S**)*.*
- *If we further assume the Hilbert spaces X and Y to be finite dimensional, then there exists a solution $(\bar{x},-\bar{p})$ to the saddle point problem* (**S**) *such that the sequence (x_k,p_k) converges strongly to $(\bar{x},\bar{p})$.*

He and Yuan ([37]) recognized that this version CP 1 can be also interpreted as an proximal point algorithm (PPA) for finding $(\bar{x}, \bar{p}) \in X \times Y$ such that $0 \in K(\bar{x}, \bar{p})$, where

$$K : X \times Y \to X \times Y, \quad K(x, p) = \begin{pmatrix} \partial f(x) + T^* p \\ \partial g^*(p) - Tx. \end{pmatrix} \tag{4.5}$$

Note that $0 \in K(\bar{x}, \bar{p})$ corresponds to the optimality condition (4.1). In general for a maximal monotone operator $K : Z \to Z$ on some Hilbert space Z, i.e. K is monotone and its graph is not properly contained in the graph of another monotone operator, a proximal point algorithm defines the $k+1$-iterate by

$$0 \in K(u_{k+1}) + \frac{1}{r_k}(u_{k+1} - u_k),$$

or equivalently by the variational inequality

$$\left\langle v - u_{k+1}, K(u_{k+1}) + \frac{1}{r_k}(u_{k+1} - u_k) \right\rangle_Z \geq 0, \quad \text{for all } v \in Z,$$

where $r_k, k \in \mathbb{N}$ are positive parameters and $\frac{1}{r_k}(u_{k+1} - u_k)$ is referred to as proximal term (cf. [67]). Setting $u_k = (x_k, p_k)^t$ the algorithm CP can be rewritten in PPA-form with (generalized) linear proximal term $M(u_{k+1} - u_k)$ as follows ([37, 62]):

$$\left\langle \begin{pmatrix} x - x_{k+1} \\ p - p_{k+1} \end{pmatrix}, K(x_{k+1}, p_{k+1}) + M \begin{pmatrix} x_{k+1} - x_k \\ p_{k+1} - p_k \end{pmatrix} \right\rangle_{X \times Y} \geq 0, \quad M := \begin{pmatrix} \frac{1}{\tau} I & -T^* \\ -\theta T & \frac{1}{\sigma} I \end{pmatrix}$$

for all $(x, p) \in X \times Y$. Since the theory of proximal point algorithms covers only the case of symmetric, positive definite matrices M, θ needs to be 1 in order for CP to be a PPA.

In [21] Chambolle and Pock also give parameter choice rules for which they could prove convergence rates. For this purpose an additional assumption on f or / and g^* is required:

Definition 4.1.2. A function $h : Z \to \overline{\mathbb{R}}$ on a Hilbert space Z is called *strongly convex with modulus* $\gamma > 0$ if it satisfies the following inequality :

$$h(\lambda z + (1 - \lambda) u) + \lambda (1 - \lambda) \gamma \|z - u\|_Z^2 \leq \lambda h(z) + (1 - \lambda) h(u) \quad z, u \in Z, \ \lambda \in [0, 1]. \tag{4.6}$$

If this inequality holds true for $\lambda = \frac{1}{2}$ we call h *strongly midconvex with modulus* $\gamma > 0$, while h is said to be *midconvex* if it obeys (4.6) for $\lambda = \frac{1}{2}$ and $\gamma = 0$.

Obviously, strong convexity implies strong midconvexity as well as convexity. Moreover, one can show (see [58]) that h is strongly (mid)convex with modulus $\gamma > 0$ if and only if the function $h - \gamma \| \cdot \|_Z^2$ is (mid)convex. Therefore the Hilbert norm monomial $h_{2,Z}(x) = \frac{1}{2} \| \cdot \|_Z^2$ is strongly (mid)convex with modulus $\gamma = \frac{1}{2}$. In [21] the following consequence of f being strongly midconvex with modulus $\gamma > 0$ is used (see [21, Eq. (35)]):

Corollary 4.1.3. *On a Hilbert space Z, let $h : Z \to \overline{\mathbb{R}}$ be a proper, convex and strongly midconvex function with modulus $\gamma > 0$. Then for any $u, z \in Z$ and any $z^* \in \partial h(z)$ we have*

$$h(u) - h(z) \geq \langle u - z, z^* \rangle_Z + \frac{\gamma}{2} \|z - u\|_Z^2. \tag{4.7}$$

Moreover, the function $\Phi_{2,Z}(Z) = \frac{1}{2}\|x - z_0\|_Z^2$ *satisfies this inequality for any* $\gamma \in (0, 1)$ *and any shift vector* $z_0 \in Z$.

Proof. The first assertion follows directly from the definitions of strong midconvexity and of the subdifferential:

$$h(u) - h(z) \geq 2\,h\left(\frac{z+u}{2}\right) - 2h(z) + \frac{\gamma}{2}\|z-u\|_Z^2 \geq \langle u - z, z^* \rangle_Z + \frac{\gamma}{2}\|z-u\|_Z^2,$$

for any $u, x \in Z$ and $x^* \in \partial h(x)$. Rewriting the polarization identity (3.14) as

$$\Phi_{2,Z}(u) - \Phi_{2,Z}(z) = \langle u - z, z - z_0 \rangle_Z + \frac{1}{2}\|z-u\|_Z^2, \quad u, z - z_0 = \partial\Phi_{2,Z}(z) \in Z$$

completes the proof. □

Stating a parameter rule in dependence of the modulus of midconvexity of f and/or g^*, Chambolle and Pock reached the following accelerated versions of CP :

Theorem 4.1.4 (Theorem 2 and 3 in [21]). *Under the assumptions of Problem 3.3.4, suppose that X and Y are Hilbert spaces and* $(\bar{x}, -\bar{p})$ *is a solution to* (**S**).

(i) *Assuming that* f *satisfies condition* (4.7) *for some* $\gamma > 0$, *we obtain a second version (CP* 2*) of algorithm 1 by choosing the parameters* $(\sigma_k, \tau_k)_{k\in\mathbb{N}}$, $(\theta_k)_{k\in\mathbb{N}}$ *as follows:*

- $\sigma_0\tau_0\|T\|^2 = 1$
- $\theta_k := (1 + 2\,\gamma\,\tau_k)^{-\frac{1}{2}}, \quad \tau_{k+1} := \theta_k\,\tau_k, \quad \sigma_{k+1} := \theta_k^{-1}\,\sigma_k.$

Then the sequence $(x_k, p_k)_{k\in\mathbb{N}}$ *we receive from CP* 2 *satisfies the following error bound: For any* $\epsilon > 0$ *there exists a* $N_0 \in \mathbb{N}$ *such that*

$$\|\bar{x} - x_N\|_Z^2 \leq \frac{1+\epsilon}{N^2}\left(\frac{\|\bar{x} - x_0\|_Z^2}{\gamma^2\,\tau_0^2} + \frac{\|\bar{p} - p_0\|_Z^2}{\gamma^2\,\sigma_0\,\tau_0}\right),$$

for all $N \geq N_0$.

(ii) *We assume that both* f *and* g^* *satisfy condition* (4.7) *with some* $\gamma > 0$ *and* $\delta > 0$, *respectively. Then we obtain a third version (CP* 3*) of algorithm 1 by the constant parameter choice* $\sigma_k = \sigma, \tau_k = \tau, \theta_k = \theta, k \in \mathbb{N}$ *with:*

- $\mu \leq \frac{2\sqrt{\gamma\,\delta}}{\|T\|}$
- $\sigma = \frac{2\mu}{\delta}, \quad \tau = \frac{2\mu}{\gamma}, \quad \theta \in \left[\frac{1}{1+\mu}, 1\right].$

The sequence $(x_k, p_k)_{k\in\mathbb{N}}$ *which we receive from CP* 3 *satisfies the following error bound:*

$$(1-\omega)\,\delta\,\|\bar{p} - p_N\|_Z^2 + \gamma\,\|\bar{x} - x_N\|_Z^2 \leq \omega^N\left(\delta\,\|\bar{p} - p_0\|_Z^2 + \gamma\,\|\bar{x} - x_0\|_Z^2\right),$$

where $\omega = (1+\theta)(2+\mu)^{-1} < 1$.

Equivalently, the assertion (*i*) holds true if we assume g^* to satisfy condition (4.7) instead of f and interchange the roles of τ_k and σ_k. So if f or g^* is strongly midconvex we obtain by CP 2 convergence in $O\left(k^{-1}\right)$, while in the case where both functions are strongly midconvex CP 3 convergences in $O\left(\omega^{\frac{1}{k}}\right)$ to $(\bar{x}, \bar{p})$.

Before we come to the generalization of CP to Banach spaces we also like to refer to other generalizations of this method: In addition to the above-mentioned preconditioned versions, there exist extended variants for solving monotone inclusion problems ([14,78]) and also to the case of nonlinear operators T ([80]). Recently, Lorenz and Pock ([55]) proposed a quite general forward-backward algorithm for monotone inclusion problems with CP as a special case.

4.2 Generalization and convergence results

Now, under the assumptions of Problem 3.3.4 let us consider the extremal relations (4.1) for a solution pair $(\bar{x}, -\bar{p})$ of (**S**) in a Banach spaces setting. The idea of the generalization is to rewrite these conditions in an analogous way as above with the help of positive parameters σ, τ and the duality mappings $J_X : X \to X^*, J_Y : Y \to Y^*$:

$$\begin{aligned}
(J_{Y^*}(\bar{p}) + \sigma T\bar{x}) - J_{Y^*}(\bar{p}) \in \sigma\partial g^*(\bar{p}) \quad &\Leftrightarrow \quad \bar{p} = (\sigma\partial g^* + J_{Y^*})^{-1}(J_{Y^*}(\bar{p}) + \sigma T\bar{x}) \\
(J_X(\bar{x}) - \tau T^*\bar{p}) - J_X(\bar{x}) \in \tau\partial f(\bar{x}) \quad &\Leftrightarrow \quad \bar{x} = (\tau\partial f + J_X)^{-1}(J_X(\bar{x}) - \tau T^*\bar{p}).
\end{aligned} \tag{4.8}$$

Consequently, we propose the following iterative method:

Algorithm 2 (CP-BS). *For* $(\tau_k, \sigma_k)_{k\in\mathbb{N}} \subseteq (0, \infty)$, $(\theta)_{k\in\mathbb{N}} \subseteq [0, 1], (x_0, p_0) \in X \times Y^*$, $\hat{x}_0 := x_0$, *set:*

$$p_{k+1} := (\sigma_k\,\partial g^* + J_{Y^*})^{-1}(J_{Y^*}(p_k) + \sigma_k\,T\hat{x}_k) \tag{4.9}$$

$$x_{k+1} := (\tau_k\,\partial f + J_X)^{-1}(J_X(x_k) - \tau_k\,T^*p_{k+1}) \tag{4.10}$$

$$\hat{x}_{k+1} := x_{k+1} + \theta_k\,(x_{k+1} - x_k) \tag{4.11}$$

First of all, note that, because of the duality mapping's nonlinearity, this algorithm is in general nonlinear. As in Theorem 4.1.1 we also like to bound the distance of one element of the sequence $(x_k, p_k)_{k\in\mathbb{N}}$ to the solution $(\bar{x}, -\bar{p})$. For this purpose we generalize the measure $\triangle_k$ to the given general Banach space case by

$$\triangle_k(x, y^*) := \frac{\mathcal{B}_{Y^*}(y^*, p_k)}{\sigma_k} + \frac{\mathcal{B}_X(x, x_k)}{\tau_k} \qquad (x, y^*) \in X \times Y^*.$$

For bounded subsets $B_1 \times B_2 \in X \times Y^*$ the partial primal-dual gap $\mathcal{G}_{B_1\times B_2}$ generalizes in a straightforward way

$$\mathcal{G}_{B_1\times B_2}(x, p) := \max_{-p'\in B_2} L(x, -p') - \min_{x'\in B_1} L(x, -p) \quad (x, p) \in X \times Y^*,$$

where L denotes the Lagrange function associated with (**S**). With these definitions we can prove the following extended version of Theorem 4.1.1:

Theorem 4.2.1. *Suppose that the assumptions of Problem 3.3.4 hold true. We choose positive parameters* σ, τ *such that* $\sqrt{\sigma\tau}\|T\| < \min\{C_X, C_{Y^*}\}$ *where* C_X, C_{Y^*} *are given by* (3.18) *and set* $\sigma_k = \sigma$, $\tau_k = \tau$ *and* $\theta_k = 1$ *for all* $k \in \mathbb{N}$ *in Algorithm 2. Then for the resulting version of CP-BS, denoted as CP-BS* 1*, the following assertions hold true:*

(i) The sequence $(x_k, p_k)_{k\in\mathbb{N}}$ *remains bounded in the form that for any solution* $(\bar{x}, -\bar{p})$ *to the saddle point problem* (**S**) *in there exists a constant*

$$C < \left(1 - \frac{\|T\|^2\,\sigma\,\tau}{C_X\,C_{Y^*}}\right)^{-1},$$

such that for any $N \in \mathbb{N}$

$$\triangle_N(\bar{x}, \bar{p}) \le C\,\triangle_0(\bar{x}, \bar{p}). \tag{4.12}$$

(ii) We define the sequence $\left(x^N, p^N\right)_{N\in\mathbb{N}}$ *of mean values* $x^N := \frac{1}{N}\sum_{k=1}^{N} x_k \in X$ *and* $y^N := \frac{1}{N}\sum_{k=1}^{N} y_k \in Y^*$*. For any* $N \in \mathbb{N}$ *and any bounded set* $B_1 \times B_2 \subset X \times Y^*$ *the restricted primal-dual gap* $\mathcal{G}_{B_1\times B_2}\left(x^N, p^N\right)$ *is bounded by*

$$D(B_1, B_2) := \frac{1}{N} \sup_{(x,y^*)\in B_1\times B_2} \triangle_0\,(y^*, x)\,.$$

Moreover, for every weak cluster point $(\tilde{x}, \tilde{p})$ *of the sequence* $\left(x^N, p^N\right)_{N\in\mathbb{N}}$, $(\tilde{x}, -\tilde{p})$ *solves the saddle point problem* (**S**).

(iii) If we further assume the Banach spaces X and Y to be finite dimensional, then there exists a solution $(\bar{x}, -\bar{p})$ *to the saddle point problem* (**S**) *such that the sequence* (x_k, p_k) *converges strongly to* $(\bar{x}, \bar{p})$*.*

Proof. Using the property (3.17) of the Bregman distance the misfit functional $\triangle_k(x, y^*)$ for some $(x, y^*) \in X \times Y^*$ can be expanded as

$$\begin{aligned}
\triangle_k(x, y^*) =& \frac{\mathcal{B}_Y\left(J_{Y^*}(p_k), J_{Y^*}(y^*)\right)}{\sigma} + \frac{\mathcal{B}_{X^*}\left(J_X(x_k), J_X(x)\right)}{\tau} \\
=& \frac{\mathcal{B}_Y\left(J_{Y^*}(p_{k+1}), J_{Y^*}(y^*)\right)}{\sigma} - \left\langle \frac{J_{Y^*}(p_k) - J_{Y^*}(p_{k+1})}{\sigma}, y^* - p_{k+1} \right\rangle_Y \\
&+ \frac{\mathcal{B}_Y\left(J_{Y^*}(p_k), J_{Y^*}(p_{k+1})\right)}{\sigma} + \frac{\mathcal{B}_{X^*}\left(J_X(x_{k+1}), J_X(x)\right)}{\tau} \\
&- \left\langle x - x_{k+1}, \frac{J_X(x_k) - J_X(x_{k+1})}{\tau} \right\rangle_X + \frac{\mathcal{B}_{X^*}\left(J_X(x_k), J_X(x_{k+1})\right)}{\tau}.
\end{aligned}$$

The iteration formulas (4.9) and (4.10) imply that

$$\frac{1}{\sigma}\left(J_{Y^*}(p_k) - J_{Y^*}(p_{k+1})\right) + T\hat{x}_k \in \partial g^*(p_{k+1}) \quad \text{and} \quad \frac{1}{\tau}\left(J_X(x_k) - J_X(x_{k+1})\right) - T^* p_{k+1} \in f(x_{k+1}).$$

So, by the definition of the subdifferential we obtain:

$$g^*(y^*) - g^*(p_{k+1}) \geq \left\langle \frac{J_{Y^*}(p_k) - J_{Y^*}(p_{k+1})}{\sigma}, y^* - p_{k+1} \right\rangle_Y + \langle T\hat{x}_k, y^* - p_{k+1}\rangle_Y \tag{4.13}$$

$$f(x) - f(x_{k+1}) \geq \left\langle x - x_{k+1}, \frac{J_X(x_k) - J_X(x_{k+1})}{\tau} \right\rangle_X - \langle T(x - x_{k+1}), p_{k+1}\rangle_X . \tag{4.14}$$

Plugging this into the expansion of $\triangle_k(x, y^*)$ it follows from Equation (3.16) that

$$\begin{aligned}
\triangle_k(x,y^*) \geq\, & g^*(p_{k+1}) - g^*(y^*) - \langle T\hat{x}_k, p_{k+1} - y^*\rangle_Y \\
& + f(x_{k+1}) - f(x) - \langle T(x_{k+1} - x), -p_{k+1}\rangle_Y \\
& + \frac{\mathcal{B}_{Y^*}(y^*, p_{k+1})}{\sigma} + \frac{\mathcal{B}_{Y^*}(p_{k+1}, p_k)}{\sigma} + \frac{\mathcal{B}_X(x, x_{k+1})}{\tau} + \frac{\mathcal{B}_X(x_{k+1}, x_k)}{\tau} \\
& + \langle Tx_{k+1}, p_{k+1} - y^*\rangle_Y - \langle T(x_{k+1} - x), p_{k+1}\rangle_Y \\
& + \langle Tx_{k+1}, y^*\rangle_Y - \langle Tx, p_{k+1}\rangle_Y \\
=\, & [< Tx_{k+1}, y^* >_Y - g^*(y^*) + f(x_{k+1})] - [\langle Tx, p_{k+1}\rangle_Y - g^*(p_{k+1}) + f(x)] \\
& + \triangle_{k+1}(x, y^*) + \frac{\mathcal{B}_{Y^*}(p_{k+1}, p_k)}{\sigma} + \frac{\mathcal{B}_X(x_{k+1}, x_k)}{\tau} \\
& + \langle T(x_{k+1} - \hat{x}_k), p_{k+1} - y^*\rangle_Y .
\end{aligned} \tag{4.15}$$

In order to estimate the last summand of this inequality, we insert (4.11) with $\theta_k = 1$ and apply Proposition 3.3.11 with $\alpha := \left(\frac{\sigma}{\tau}\right)^{\frac{1}{2}} > 0$ yielding

$$\begin{aligned}
& \langle T((x_{k+1} - x_k) - (x_k - x_{k-1})), p_{k+1} - y^*\rangle_Y \\
& = \langle T(x_{k+1} - x_k), p_{k+1} - y^*\rangle_Y - \langle T(x_k - x_{k-1}), p_k - y^*\rangle_Y - \langle T(x_k - x_{k-1}), p_{k+1} - p_k\rangle_Y \\
& \geq \langle T(x_{k+1} - x_k), p_{k+1} - y^*\rangle_Y - \langle T(x_k - x_{k-1}), p_k - y^*\rangle_Y \\
& - \frac{\|T\|\,\sigma^{\frac{1}{2}}\,\tau^{\frac{1}{2}}}{C_X} \frac{\mathcal{B}_X(x_k, x_{k-1})}{\tau} - \frac{\|T\|\,\sigma^{\frac{1}{2}}\,\tau^{\frac{1}{2}}}{C_{Y^*}} \frac{\mathcal{B}_{Y^*}(p_{k+1}, p_k)}{\sigma}.
\end{aligned}$$

Thus, we conclude that

$$\begin{aligned}
\triangle_k(x,y^*) \geq\, & [\langle Tx_{k+1}, y^*\rangle_Y - g^*(y^*) + f(x_{k+1})] - [\langle Tx, p_{k+1}\rangle_Y - g^*(p_{k+1}) + f(x)] \\
& + \triangle_{k+1}(x, y^*) + \left(1 - \frac{\|T\|\,\sigma^{\frac{1}{2}}\,\tau^{\frac{1}{2}}}{C_{Y^*}}\right) \frac{\mathcal{B}_{Y^*}(p_{k+1}, p_k)}{\sigma} \\
& - \frac{\|T\|\,\sigma^{\frac{1}{2}}\,\tau^{\frac{1}{2}}}{C_X} \frac{\mathcal{B}_X(x_k, x_{k-1})}{\tau} + \frac{\mathcal{B}_X(x_{k+1}, x_k)}{\tau} \\
& + \langle T(x_{k+1} - x_k), p_{k+1} - y^*\rangle_Y - \langle T(x_k - x_{k-1}), p_k - y^*\rangle_Y .
\end{aligned} \tag{4.16}$$

Summing from $k = 0$ to $N - 1$ leads to

$$\begin{aligned}
&\triangle_0(x,y^*) + \left|\langle T\left(x_N - x_{N-1}\right), p_N - y^*\rangle_Y\right| \\
&\qquad \geq \sum_{k=0}^{N} \left[\langle Tx_{k+1}, y^*\rangle_Y - g^*(y^*) + f(x_{k+1})\right] - \left[\langle Tx, p_{k+1}\rangle_Y - g^*(p_{k+1}) + f(x)\right] \\
&\qquad + \triangle_N(x,y^*) + \left(1 - \frac{\|T\|\,\sigma^{\frac{1}{2}}\,\tau^{\frac{1}{2}}}{C_{Y^*}}\right)\sum_{k=1}^{N} \frac{\mathcal{B}_{Y^*}\left(p_k, p_{k-1}\right)}{\sigma} + \frac{\mathcal{B}_X\left(x_N, x_{N-1}\right)}{\tau} \\
&\qquad + \left(1 - \frac{\|T\|\,\sigma^{\frac{1}{2}}\,\tau^{\frac{1}{2}}}{C_X}\right)\sum_{k=1}^{N-1} \frac{\mathcal{B}_X\left(x_k, x_{k-1}\right)}{\tau}.
\end{aligned}$$

Now we apply again Proposition 3.3.11 with $\alpha = \frac{C_x}{\|T\|\,\tau}$:

$$\left|\langle T\left(x_N - x_{N-1}\right), p_N - y^*\rangle_Y\right| \leq \frac{\mathcal{B}_X\left(x_N, x_{N-1}\right)}{\tau} + \frac{\|T\|^2\,\sigma\,\tau}{C_X\,C_{Y^*}}\frac{\mathcal{B}_{Y^*}\left(y^*, p_N\right)}{\sigma}, \tag{4.17}$$

so that we deduce

$$\begin{aligned}
\triangle_0(x,y^*) &\geq \sum_{k=0}^{N} \left[\langle Tx_{k+1}, y^*\rangle_Y - g^*(y^*) + f(x_{k+1})\right] - \left[\langle Tx, p_{k+1}\rangle_Y - g^*(p_{k+1}) + f(x)\right] \\
&\quad + \frac{\mathcal{B}_X\left(x, x_N\right)}{\tau} + \left(1 - \frac{\|T\|\,\sigma^{\frac{1}{2}}\,\tau^{\frac{1}{2}}}{C_X}\right)\sum_{k=1}^{N-1} \frac{\mathcal{B}_X\left(x_k, x_{k-1}\right)}{\tau} \\
&\quad + \left(1 - \frac{\|T\|^2\,\sigma\,\tau}{C_X\,C_{Y^*}}\right)\frac{\mathcal{B}_{Y^*}\left(y^*, p_N\right)}{\sigma} + \left(1 - \frac{\|T\|\,\sigma^{\frac{1}{2}}\,\tau^{\frac{1}{2}}}{C_{Y^*}}\right)\sum_{k=1}^{N} \frac{\mathcal{B}_{Y^*}\left(p_k, p_{k-1}\right)}{\sigma}.
\end{aligned} \tag{4.18}$$

Here, because of the choice $\sigma^{\frac{1}{2}}\,\tau^{\frac{1}{2}} < \frac{\min\{C_X, C_{Y^*}\}}{\|T\|}$, we obtain only positive coefficients. For a solution $(\bar{x}, -\bar{p})$ to the saddle point problem (**S**) we set $(x, y^*) = (\bar{x}, \bar{p})$. Then, due to the extremal relations (4.1), every summand in the first line of (4.18) is non negative as well:

$$\begin{aligned}
&\left[-\langle Tx_{k+1}, -\bar{p}\rangle_Y - g^*(y^*) + f(x_{k+1})\right] - \left[\langle T\bar{x}, p_{k+1}\rangle_Y - g^*(p_{k+1}) + f(\bar{x})\right] \\
&= f(x_{k+1}) - f(\bar{x}) - \langle x_{k+1} - \bar{x}, -T^*\bar{p}\rangle_X + g^*(p_{k+1}) - g^*(\bar{p}) - \langle T\bar{x}, p_{k+1} - \bar{p}\rangle_Y \geq 0.
\end{aligned} \tag{4.19}$$

This proves assertion (i). The second assertion follows directly along the lines of the corresponding proof for Theorem 4.1.1 in [21], p. 124: Because of (4.18) and (4.19) we have

$$\begin{aligned}
&\mathcal{G}_{B_1\times B_2}\left(x^N, p^N\right) = \sup_{(x,y^*)\in B_1\times B_2} \left[\langle Tx_N, y^*\rangle_Y - g^*(y^*) + f(x_N)\right] - \left[\langle Tx, p_N\rangle_Y - g^*(p_N) + f(x)\right] \\
&\leq \frac{1}{N} \sup_{(x,y^*)\in B_1\times B_2} \sum_{k=0}^{N} \left[\langle Tx_{k+1}, y^*\rangle_Y - g^*(y^*) + f(x_{k+1})\right] - \left[\langle Tx, p_{k+1}\rangle_Y - g^*(p_{k+1}) + f(x)\right] \\
&\leq \frac{1}{N} \sup_{(x,y^*)\in B_1\times B_2} \triangle_0(x,y^*).
\end{aligned}$$

Moreover, observe that the sequence $\left(x^N, p^N\right)_{N\in\mathbb{N}}$ is bounded. So, it has at least one cluster point $(\tilde{x}, \tilde{p}) \in X \times Y^*$ and because of the last inequality at this point the global primal-dual gap is non-positive:

$$\mathcal{G}(\tilde{x}, \tilde{p}) \leq [< T\tilde{x}, y^* >_Y - g^*(y^*) + f(\tilde{x})] - [\langle Tx, \tilde{p}\rangle_Y - g^*(\tilde{p}) + f(x)] \leq 0.$$

We conclude that $(\tilde{x}, -\tilde{y})$ must be a solution to the saddle point problem (**S**) which completes the proof of (ii). Also for the last assertion which requires the assumption that X and Y are finite dimensional, we apply the same arguments as in [21], p. 124, to (4.18): In order to prove strong convergence of the bounded sequence $(x_k, p_k)_{k\in\mathbb{N}}$ let us consider a convergent subsequence $(x_{l(k)}, p_{l(k)})_{k\in\mathbb{N}}$ with limit $(\tilde{x}, \tilde{p})$. Due to inequality (4.18) we have

$$\lim_{k\to\infty} \mathcal{B}_X(x_k, x_{k-1}) = \lim_{k\to\infty} \mathcal{B}_{Y^*}(p_k, p_{k-1}) = 0,$$

such that $(\tilde{x}, \tilde{p})$ must be the limit of $(x_{l(k)-1}, p_{l(k)-1})_{k\in\mathbb{N}}$ as well. Consequently, $(\tilde{x}, \tilde{p})$ obeys Equation (4.8) and thus $(\tilde{x}, -\tilde{p})$ is a solution to the saddle point problem (**S**). Now it remains to show that the whole sequence convergences to this point. For this purpose we substitute $(x, y^*) = (\tilde{x}, \tilde{p})$ in (4.16) and sum again from $k = l(k)$ to some $N-1$ greater or equal to $l(k)$:

$$\begin{aligned} \triangle_{l(k)}(\tilde{x}, \tilde{p}) \geq \triangle_N(\tilde{x}, \tilde{p}) &+ \left(1 - \frac{\|T\|\,\sigma^{\frac{1}{2}}\,\tau^{\frac{1}{2}}}{C_{Y^*}}\right) \sum_{k=l(k)+1}^{N} \frac{\mathcal{B}_{Y^*}(p_k, p_{k-1})}{\sigma} + \frac{\mathcal{B}_X(x_N, x_{N-1})}{\tau} \\ &- \frac{\mathcal{B}_X(x_{l(k)}, x_{l(k)-1})}{\tau} + \left(1 - \frac{\|T\|\,\sigma^{\frac{1}{2}}\,\tau^{\frac{1}{2}}}{C_X}\right) \sum_{k=l(k)}^{N-1} \frac{\mathcal{B}_X(x_k, x_{k-1})}{\tau} \\ &+ \langle T(x_N - x_{N-1}), p_N - \tilde{p}\rangle_Y - \langle T(x_{l(k)} - x_{l(k)-1}), p_{l(k)} - \tilde{p}\rangle_Y, \end{aligned}$$

where we used $[< Tx_{k+1}, \tilde{p} >_Y - g^*(\tilde{p}) + f(x_{k+1})] - [\langle T\tilde{x}, p_{k+1}\rangle_Y - g^*(p_{k+1}) + f(\tilde{x})] \geq 0$ for all $k \in \mathbb{N}$. By taking the limit $l(k), N \to \infty$ immediately the convergence of the sequence $(x_k, p_k)_{k\in\mathbb{N}}$ with respect to the Bregman divergence and hence the strong convergence $(x_k, p_k) \to (\tilde{x}, \tilde{p})$ for $k \to \infty$ follows. □

Remark 4.2.2. This generalization CP-BS 1 covers also the preconditioned version of CP 1 introduced in [62]: There X and Y are of the form $X = \Upsilon^{\frac{1}{2}} H_X$ equipped with $\|x\|_X = \|\Upsilon^{-\frac{1}{2}} x\|_{H_X}$ and $Y = \Sigma^{-\frac{1}{2}} H_Y$ equipped with $\|y\|_Y = \|\Sigma^{\frac{1}{2}} x\|_{H_Y}$, where H_X, H_Y are Hilbert spaces and Υ, Σ symmetric, positive definite matrices. Considering the dual spaces $X^* = \Upsilon^{-\frac{1}{2}} H_X$ and $Y^* = \Sigma^{\frac{1}{2}} H_Y$ with respect to the scalar product on the corresponding Hilbert spaces, the duality mappings read as

$$J_X(x) = \Upsilon^{-1} x, \qquad J_{Y^*}(y) = \Sigma^{-1} y.$$

Due to their linearity line (4.9) and (4.10) take the form of update rule (4) in [62]:

$$\begin{aligned} p_{k+1} &= (\sigma_k \Sigma\, \partial g^* + I)^{-1} (p_k + \sigma_k\, \Sigma T \hat{x}_k), \\ x_{k+1} &= (\tau_k \Upsilon\, \partial f + I)^{-1} (x_k - \tau_k\, \Upsilon T^* p_{k+1}). \end{aligned} \tag{4.20}$$

The convergence proof of the extended algorithm (4.20) (proposed in [62]) relies on the idea ([37]) to reformulate the update rule as the proximal point algorithm:

$$\left\langle \begin{pmatrix} x - x_{k+1} \\ p - p_{k+1} \end{pmatrix}, K(x_{k+1}, p_{k+1}) + M \begin{pmatrix} x_{k+1} - x_k \\ p_{k+1} - p_k \end{pmatrix} \right\rangle_{X \times Y^*} \geq 0, \quad M := \begin{pmatrix} \Upsilon^{-1} & -T^* \\ -\theta T & \Sigma^{-1} \end{pmatrix}$$

for all $(x, p) \in X \times Y^*$ where $K : X \times Y^* \to X^* \times Y$ is given by Equation (4.5) and the matrix M is symmetric and positive definite for $\theta = 1$ ([62, Lemma 1]). We also like to rewrite our method CP-BS in this form: Setting $p_1 = p_0$ and hence starting with the update of x_k in line (4.10) it follows that

$$\begin{aligned} \begin{pmatrix} 0 \\ 0 \end{pmatrix} &= \begin{pmatrix} \partial f(x_{k+1}) + T^* p_{k+1} - T^* p_{k+1} + T^* p_k + \frac{1}{\tau_k} (J_X(x_{k+1}) - J_X(x_k)) \\ \partial g^*(p_{k+1}) - T x_{k+1} - \theta_k T x_{k+1} + \theta_k T x_k + \frac{1}{\sigma_k} (J_{Y^*}(p_{k+1}) - J_{Y^*}(p_k)) \end{pmatrix} \\ &= K(x_{k+1}, p_{k+1}) + \begin{pmatrix} \frac{1}{\tau_k} (J_X(x_{k+1}) - J_X(x_k)) \\ \frac{1}{\sigma_k} (J_{Y^*}(p_{k+1}) - J_{Y^*}(p_k)) \end{pmatrix} + \begin{pmatrix} T^*(p_k - p_{k+1}) \\ \theta_k T(x_k - x_{k+1}) \end{pmatrix}. \end{aligned}$$

So, our algorithm 2 takes the structure of a PPA with a nonlinear proximal term in some way. Although there exist generalizations of PPAs with nonlinear proximal terms (cf. e.g. [4]) to the best of the author's knowledge non of these is of this special form (with a Banach space setting). On the other hand, one finds generalizations of PPA to Banach spaces that include a similar update rule for $u_{n+1} = (x_{n+1}, p_{k+1})$ as one step in their iteration (see [20, 50]). The update u_{n+1} is then used to define two sets $H_k, W_k \subseteq X \times Y^*$ such that $K^{-1}(0) \subseteq H_k \cap W_k$. There the projection (with respect to the Bregman distance) of the initial guess $u_0 = (x_0, p_0)$ to the intersection $H_k \cap W_k$ gives the new iterate. These additional projection steps are required in order to achieve strong convergence (cf. [71]). Compared to the proposed method CP-BS the generalized PPAs in [20, 50] have the advantage that less restrictive conditions on the Banach spaces X, Y are needed to assure strong convergence but the disadvantage that the necessary computation of the projection with respect to the Bregman distance might be too complicated in practice.

Next we want to also generalize the accelerated forms CP 2, CP 3 of the CP -algorithm which base on the assumption that f and/or g^* satisfy condition (4.7). For this purpose we extended this property to Banach spaces:

Definition 4.2.3. On a Banach space Z, let $h : Z \to \overline{\mathbb{R}}$ be a proper, convex function. We say that h satisfy a *Bregman midconvex property with modulus* $\gamma > 0$ if for any $u, z \in Z$ and any $z^* \in \partial h(z)$ the following inequality holds true:

$$h(u) - h(z) \geq \langle u - z, z^* \rangle_Z + \gamma\, \mathcal{B}_Z(u, z). \tag{4.21}$$

Now, we can formulate a convergence result that uses the assumption that f satisfies this Bregman midconvex property. The case that (4.21) holds true for g^* instead of f follows analogously.

Theorem 4.2.4. *Under the assumptions of Problem 3.3.4 suppose that f satisfies the Bregman midconvex property* (4.21) *with modulus $\gamma > 0$. Moreover, let $(\bar{x}, -\bar{p}) \in X \times Y^*$ be a solution to* (**S**)*. We obtain a generalized version CP-BS 2 of CP 2 by choosing the parameters $(\sigma_k, \tau_k)_{k \in \mathbb{N}}$, $(\theta_k)_{k \in \mathbb{N}}$ in algorithm 2 as follows:*

- $\sqrt{\sigma_0}\,\sqrt{\tau_0}\,\|T\| \le \min\{C_X, C_{Y^*}\}$, *(where $\tau_0, \sigma_0 > 0$)*
- $\theta_k := (1+\gamma\,\tau_k)^{-\frac{1}{2}}$, $\qquad \tau_{k+1} := \theta_k\,\tau_k$, $\qquad \sigma_{k+1} := \theta_k^{-1}\,\sigma_k$.

Then the sequence $(x_k, p_k)_{k\in\mathbb{N}}$ we receive from CP-BS 2 satisfies the following error bound: For any $\epsilon > 0$ there exists an $N_0 \in \mathbb{N}$ such that

$$\mathcal{B}_X(\bar{x}, x_N) \le \frac{4+4\,\epsilon}{N^2}\left(\frac{\mathcal{B}_X(\bar{x}, x_0)}{\gamma^2\,\tau_0^2} + \frac{\mathcal{B}_{Y^*}(\bar{p}, p_0)}{\gamma^2\,\sigma_0\,\tau_0}\right), \tag{4.22}$$

for all $N \ge N_0$ and hence $(x_k)_{k\in\mathbb{N}}$ convergences strongly to the solution $\bar{x}$ of (**P**).

Proof. We go back to the estimate (4.15) and set $(x, y^*) := (\bar{x}, \bar{p})$. For $u = \bar{x}, z = x_{k+1}$ and $x^* := {}^1\!/_{\tau_k}\,(J_X(x_k) - J_X(x_{k+1})) - T^*\,p_{k+1} \in \partial f(x_{k+1})$ the Bregman midconvex property (4.21) of f gives:

$$\begin{aligned} f(\bar{x}) - f(x_{k+1}) \ge &\left\langle \bar{x} - x_{k+1}, \frac{J_X(x_k) - J_X(x_{k+1})}{\tau_k}\right\rangle_X - \langle T(\bar{x} - x_{k+1}), p_{k+1}\rangle_X \\ &+ \gamma\,\mathcal{B}_X(\bar{x}, x_{k+1}). \end{aligned} \tag{4.23}$$

Thus, replacing (4.14) by (4.23) and applying (4.19) to the right hand side of (4.15) leads to the inequality:

$$\begin{aligned} \triangle_k(\bar{x}, \bar{p}) \ge &\left(\gamma + \frac{1}{\tau_k}\right)\mathcal{B}_X(\bar{x}, x_{k+1}) + \frac{\mathcal{B}_{Y^*}(\bar{p}, p_{k+1})}{\sigma_k} + \frac{\mathcal{B}_{Y^*}(p_{k+1}, p_k)}{\sigma_k} \\ &+ \frac{\mathcal{B}_X(x_{k+1}, x_k)}{\tau_k} + \langle T(x_{k+1} - \hat{x}_k), p_{k+1} - \bar{p}\rangle_Y. \end{aligned} \tag{4.24}$$

Now we use Proposition 3.3.11 with $\alpha = \frac{C_X}{\|T\|\,\sigma_k\,\tau_k}$

$$-\theta_{k-1}\,\langle T(x_k - x_{k-1}), p_{k+1} - p_k\rangle_Y \ge -\frac{\mathcal{B}_{Y^*}(p_{k+1}, p_k)}{\sigma_k} - \frac{\theta_{k-1}^2\,\|T\|^2\,\tau_{k-1}\,\sigma_k}{C_X\,C_{Y^*}}\frac{\mathcal{B}_X(x_k, x_{k-1})}{\tau_{k-1}},$$

insert the over-relaxation step (4.11) such that we end up with

$$\begin{aligned} \triangle_k(\bar{x}, \bar{p}) \ge &(1+\gamma\,\tau_k)\,\frac{\tau_{k+1}}{\tau_k}\frac{\mathcal{B}_X(\bar{x}, x_{k+1})}{\tau_{k+1}} + \frac{\sigma_{k+1}}{\sigma_k}\frac{\mathcal{B}_{Y^*}(\bar{p}, p_{k+1})}{\sigma_{k+1}} \\ &+ \frac{\mathcal{B}_X(x_{k+1}, x_k)}{\tau_k} - \frac{\theta_{k-1}^2\,\|T\|^2\,\tau_{k-1}\,\sigma_k}{C_X\,C_{Y^*}}\frac{\mathcal{B}_X(x_k, x_{k-1})}{\tau_{k-1}} \\ &+ \langle T(x_{k+1} - x_k), p_{k+1} - \bar{p}\rangle_Y - \theta_{k-1}\,\langle T(x_k - x_{k-1}), p_k - \bar{p}\rangle_Y. \end{aligned}$$

The choice of the parameters ensures that

$$(1+\gamma\,\tau_k)\,\frac{\tau_{k+1}}{\tau_k} = \theta_k^{-1} \ge 1, \quad \frac{\sigma_{k+1}}{\sigma_k} = \theta_k^{-1} \ge 1, \quad \text{and} \quad \frac{\tau_k}{\tau_{k+1}} = \theta_k^{-1} \ge 1 \quad \text{for all } k \in \mathbb{N}.$$

Moreover, because of $\min(C_X, C_{Y^*}) \geq \|T\|\, \tau_0^{\frac{1}{2}} \sigma_0^{\frac{1}{2}} = \|T\|\, \tau_k^{\frac{1}{2}} \sigma_k^{\frac{1}{2}}$ for all $k \in \mathbb{N}$ we have

$$\frac{1}{\tau_k} \frac{\theta_{k-1}^2 \|T\|^2 \tau_{k-1} \sigma_k}{C_X C_{Y^*}} = \frac{1}{\tau_{k-1}} \frac{\|T\|^2 \tau_k \sigma_k}{C_X C_{Y^*}} \leq \frac{1}{\tau_{k-1}}, \qquad k \in \mathbb{N}.$$

Therefore,

$$\begin{aligned}\frac{\triangle_k(\bar{x},\bar{p})}{\tau_k} \geq & \frac{\triangle_{k+1}(\bar{x},\bar{p})}{\tau_{k+1}} + \frac{\mathcal{B}_X(x_{k+1},x_k)}{\tau_k^2} - \frac{\mathcal{B}_X(x_k,x_{k-1})}{\tau_{k-1}^2} \\ & + \frac{1}{\tau_k} \langle T(x_{k+1}-x_k), p_{k+1}-\bar{p}\rangle_Y - \frac{1}{\tau_{k-1}} \langle T(x_k - x_{k-1}), p_k - \bar{p}\rangle_Y\end{aligned}$$

holds true. Now, summing these inequalities from $k = 0$ to $N-1$ for some $N > 0$ with $x_{-1} := x_0$ and applying (4.17) with $\tau = \tau_{N-1}$ yields

$$\begin{aligned}\frac{\triangle_0(\bar{x},\bar{p})}{\tau_0} &\geq \frac{\triangle_N(\bar{x},\bar{p})}{\tau_N} + \frac{\mathcal{B}_X(x_N,x_{N-1})}{\tau_{N-1}^2} + \frac{1}{\tau_{N-1}} \langle T(x_N - x_{N-1}), p_N - \bar{p}\rangle_Y \\ &\geq \frac{\triangle_N(\bar{x},\bar{p})}{\tau_N} - \frac{\|T\|^2}{C_X C_{Y^*}} \mathcal{B}_{Y^*}(\bar{p},p_N).\end{aligned}$$

By multiplying by τ_N^2 and using the identity $\tau_N \sigma_N = \tau_0 \sigma_0$ we obtain the following error bound:

$$\frac{\tau_N^2}{\tau_0 \sigma_0}\left(1 - \frac{\|T\|^2}{C_X C_{Y^*}} \tau_0 \sigma_0\right) \mathcal{B}_{Y^*}(\bar{p},p_N) + \mathcal{B}_X(\bar{x},x_N) \leq \tau_N^2 \left(\frac{\mathcal{B}_{Y^*}(\bar{p},p_0)}{\sigma_0 \tau_0} + \frac{\mathcal{B}_X(\bar{x},x_0)}{\tau_0^2}\right).$$

Substituting γ by $\frac{\gamma}{2}$ in Lemma 1-2 and Corollary 1 in [21] shows that for any $\epsilon > 0$ there exists a $N_0 \in \mathbb{N}$ (depending on ϵ and $\gamma\,\tau_0$) with $\tau_N^2 \leq 4(1+\epsilon)(N\gamma)^{-2}$ for all $N \geq N_0$. This completes the proof. □

Note that compared to the error estimate in Theorem 4.1.4 (*ii*) the error bound for the generalized version CP-BS 2 is 4-times larger. That is due to the fact that in the Hilbert space case the positive term (4.19) was estimated with the help of the extremal relation $-T^*\bar{p} \in \partial f(\bar{x})$ and property (4.21) as:

$$[-\langle T x_{k+1}, -\bar{p}\rangle_Y - g^*(y^*) + f(x_{k+1})] - [\langle T\bar{x}, p_{k+1}\rangle_Y - g^*(p_{k+1}) + f(\bar{x})] \geq \frac{\gamma}{2}\|x_{k+1} - \bar{x}\|_X^2.$$

Then this inequality gives the larger coefficient $2\gamma + \frac{1}{\tau_k}$ of $\mathcal{B}_X(\bar{x}, x_{k+1}) = \frac{1}{2}\|x_{k+1} - \bar{x}\|_X^2$ in estimate (4.24), which cause the smaller error bound. Although we obtain a similar inequality

$$[-\langle T x_{k+1}, -\bar{p}\rangle_Y - g^*(y^*) + f(x_{k+1})] - [\langle T\bar{x}, p_{k+1}\rangle_Y - g^*(p_{k+1}) + f(\bar{x})] \geq \gamma\, \mathcal{B}_X(x_{k+1}, \bar{x})$$

in the Banach space setting, its application would just cause an additional summand at the right hand side of (4.24) as the Bregman distance is not symmetric.

Finally, we generalize CP 3 to Banach spaces where we assume that also g^* satisfies the Bregman midconvex property (4.21) with some modulus $\delta > 0$. This gives a method with linear convergence:

Theorem 4.2.5. *Under the assumptions of Problem 3.3.4, we assume both f and g^* to satisfy the Bregman midconvex property* (4.21) *with modulus $\gamma > 0$ and $\delta > 0$, respectively. Moreover, let $(\bar{x}, -\bar{p}) \in X \times Y^*$ denote a solution to* (**S**)*. Choosing the parameters $\sigma_k = \sigma, \tau_k = \tau$, and $\theta_k = \theta$ in Algorithm 2 constant for any $k \in \mathbb{N}$ such that*

- $\mu \leq \frac{\sqrt{\gamma\delta}\min\{C_X, C_{Y^*}\}}{\|T\|}$
- $\sigma = \frac{\mu}{\delta}, \quad \tau = \frac{\mu}{\gamma}$
- $\theta \in \left[\frac{1}{1+\mu}, 1\right]$

we obtain a generalized version CP-BS 3 *of CP* 3*. Then the sequence $(x_k, p_k)_{k\in\mathbb{N}}$ we receive from CP-BS* 3 *satisfies the following error bound:*

$$(1-\omega)\,\delta\,\mathcal{B}_{Y^*}(\bar{p}, p_N) + \gamma\,\mathcal{B}_X(\bar{x}, x_N) \leq \omega^N\left(\delta\,\mathcal{B}_{Y^*}(\bar{p}, p_0) + \gamma\,\mathcal{B}_X(\bar{x}, x_0)\right) \quad N \in \mathbb{N}, \tag{4.25}$$

with $\omega < 1$ is given by (4.28)*. Hence, the sequence $(x_k)_{k\in\mathbb{N}}$ converges (with respect to $\|\cdot\|_X$) in $\mathcal{O}\left(\omega^{\frac{k}{2}}\right)$ to the solution $\bar{x}$ of* (**P**)*.*

Proof. In analogy to the proof of Theorem 4.2.4, we obtain from the property (4.21) of f and g^* a sharper estimate for (4.15) where we set $(x, y^*) = (\bar{x}, \bar{y})$. For this purpose, we replace (4.14) by (4.23) and (4.13) by

$$g^*(\bar{p}) - g^*(p_{k+1}) \geq \left\langle \frac{J_{Y^*}(p_k) - J_{Y^*}(p_{k+1})}{\sigma}, \bar{p} - p_{k+1} \right\rangle_Y + \langle T\hat{x}_k, \bar{p} - p_{k+1}\rangle_Y + \delta\,\mathcal{B}_{Y^*}(\bar{p}, p_{k+1}).$$

This, together with (4.19), leads to

$$\begin{aligned} \triangle_k(\bar{x}, \bar{p}) \geq &\left(\delta + \frac{1}{\sigma}\right)\mathcal{B}_{Y^*}(\bar{p}, p_{k+1}) + \left(\gamma + \frac{1}{\tau}\right)\mathcal{B}_X(\bar{x}, x_{k+1}) + \frac{\mathcal{B}_{Y^*}(p_{k+1}, p_k)}{\sigma} \\ &+ \frac{\mathcal{B}_X(x_{k+1}, x_k)}{\tau} + \langle T(x_{k+1} - \hat{x}_k), p_{k+1} - \bar{p}\rangle_Y\,. \end{aligned} \tag{4.26}$$

Using (4.11) as well as (3.20) with some constant $\alpha > 0$ we can estimate the last term in the following way:

$$\begin{aligned} \langle T(x_{k+1} - \hat{x}_k), p_{k+1} - \bar{p}\rangle_Y &= \langle T(x_{k+1} - x_k), p_{k+1} - \bar{p}\rangle_Y - \omega\,\langle T(x_k - x_{k-1}), p_k - \bar{p}\rangle_Y \\ &\quad - \omega\,\langle T(x_k - x_{k-1}), p_{k+1} - p_k\rangle_Y \\ &\quad - (\theta - \omega)\,\langle T(x_k - x_{k-1}), p_{k+1} - \bar{p}\rangle_Y \\ &\geq \langle T(x_{k+1} - x_k), p_{k+1} - \bar{p}\rangle_Y - \omega\,\langle T(x_k - x_{k-1}), p_k - \bar{p}\rangle_Y \\ &\quad - \omega\,\|T\|\frac{\mathcal{B}_{Y^*}(p_{k+1}, p_k)}{C_{Y^*}\,\alpha} - \theta\,\|T\|\,\alpha\frac{\mathcal{B}_X(x_k, x_{k-1})}{C_X} \\ &\quad - (\theta - \omega)\,\|T\|\frac{\mathcal{B}_{Y^*}(\bar{p}, p_{k+1})}{C_{Y^*}\,\alpha}, \end{aligned}$$

for any $\omega \in [(1+\mu)^{-1}, \theta]$. Now we set $\alpha = \omega\left(\frac{\gamma}{\delta}\right)^{\frac{1}{2}}$ such that $\frac{\|T\|\mu\,\omega}{C_{Y^*}\,\alpha} \leq \delta = \frac{\mu}{\sigma}$ and $\frac{\mu\|T\|\alpha}{C_X} \leq \omega\,\gamma$ and multiply inequality (4.26) by μ:

$$\begin{aligned} \mu \bigtriangleup_k(\bar{x},\bar{p}) \geq & \left(1+\mu-\frac{1}{\omega}\right)\mu \bigtriangleup_{k+1}(\bar{x},\bar{p}) + \frac{\mu}{\omega}\bigtriangleup_{k+1}(\bar{x},\bar{p}) - \frac{(\theta-\omega)\,\delta}{\omega}\mathcal{B}_{Y^*}(\bar{p},p_{k+1}) \\ & + \mu\,\langle T(x_{k+1}-x_k), p_{k+1}-\bar{p}\rangle_Y - \mu\,\omega\,\langle T(x_k - x_{k-1}), p_k - \bar{p}\rangle_Y \\ & + \gamma\,\mathcal{B}_X(x_{k+1},x_k) - \theta\,\omega\,\gamma\,\mathcal{B}_X(x_k,x_{k-1}). \end{aligned} \tag{4.27}$$

As in Theorem 4.1.4 we choose

$$\omega := \frac{1+\theta}{2+\mu} \geq \frac{1+\theta}{2+\frac{\sqrt{\gamma\,\delta}\min\{C_X,C_{Y^*}\}}{\|T\|}}, \tag{4.28}$$

in order to ensure that

$$\left(1+\mu-\frac{1}{\omega}\right)\mu\bigtriangleup_{k+1}(\bar{x},\bar{p}) - \frac{(\theta-\omega)\,\delta}{\omega}\mathcal{B}_{Y^*}(\bar{p},p_{k+1}) \geq 0.$$

Thus, multiplying (4.27) with ω^{-k} and summing from $k = 0$ to $N-1$ for some $N > 0$ (where we set $x_{-1} = x_0$) leads to

$$\begin{aligned} \mu\bigtriangleup_0(\bar{x},\bar{p}) \geq \omega^{-N}\mu\bigtriangleup_N(\bar{x},\bar{p}) + \omega^{-N+1}\,\gamma\,\mathcal{B}_X(x_N,x_{N-1}) \\ + \omega^{-N+1}\,\mu\,\langle T(x_N - x_{N-1}), p_N - \bar{p}\rangle_Y\,. \end{aligned}$$

Finally, by using Proposition 3.3.11 with $\alpha = (\gamma/\delta)^{\frac{1}{2}}$ as well as $\|T\|\,\mu\,\alpha \leq \gamma\min\{C_X, C_{Y^*}\}$ and $\|T\|\mu/\alpha \leq \delta\,\min\{C_X, C_{Y^*}\}$, we obtain the inequality:

$$\mu\bigtriangleup_0(\bar{x},\bar{p}) \geq \omega^{-N}\mu\bigtriangleup_N(\bar{x},\bar{p}) - \omega^{-N+1}\,\delta\,\mathcal{B}_{Y^*}(\bar{p},p_N)\,.$$

Now the assertion follows from relation (3.18). □

Note that letting $\sigma_k,\ \tau_k \to \infty$ in lines (4.9) and (4.10) the algorithm CP-BS becomes an iteration of the extremal relations (4.1), while setting $\sigma_k = \tau_k = 0$ eliminates the dependence on (4.1). Consequently, the parameters $\sigma_k,\ \tau_k,\ k \in \mathbb{N}$ should be chosen as large as possible. This advice was also confirmed by our numerical experiments, where we relaxed the upper bound of the product $\tau_k\,\sigma_k$ given by the theory in the following way:

Remark 4.2.6. Due to the application of Proposition 3.3.11 in the proofs of Theorems 4.2.1, 4.2.4 and 4.2.5 the proposed parameter choice rules of the versions CP-BS 1-3 depend on the constants C_X and C_{Y^*} given by (3.18). Obviously, if X and Y are Hilbert spaces we have $C_X = C_{Y^*} = 1$. Also in the Banach space case for $u = 0$ the inequality

$$\mathcal{B}_X(x,u) = \frac{1}{2}\|x\|_X^2 \geq \frac{C_X}{2}\|x-u\|_X^2 \quad x \in X$$

is sharp for $C_X = 1$. Accordingly, for a specific application we might only require estimate (3.20) on bounded domains where the constants C_X and C_{Y^*} are not optimal. In fact, numerical experiments indicate that we obtain faster convergence of CP-BS 1-3 if we relax the parameter choice of σ and τ by replacing the minimum $\min\{C_X\,,\, C_{Y^*}\}$ by a value $C \in [C_X\,C_{Y^*}, 1)$ close to 1. However, the condition $C < 1$, or equivalently $\tau_0\,\sigma_0\,\|T\|^2 < 1$, seems to be required in order to obtain convergence to the solution $(\bar{x}, -\bar{p})$.

4.3 Generalized resolvents

4.3.1 Use of resolvents

In this section we discuss the special generalization of the resolvent in our algorithm with the focus on its evaluation complexity. Recall from Section 3.3 that on a reflexive Banach space Z the generalized resolvent $(\tau\partial h + J_Z)^{-1} : Z^* \to Z$ of $h \in \Gamma(Z)$ and $\tau > 0$ is well-defined and single-valued. Moreover, if we assume that there is at least one point in Z where h is finite and continuous, then Equation (3.24):

$$(\tau\partial h + J_Z)^{-1}(u) = \operatorname*{argmin}_{z\in Z} \tau\, h(z) + \mathcal{B}_Z(z, J_{Z^*}(u))$$

holds true. Setting $F = J_Z$ the resolvent $(\partial h + J_Z)^{-1}$ is obviously closely related but not identical to the F-resolvents $(A + F)^{-1}F$ of maximal monotone operators A as used in the generalized PPA of [50] and studied in [10]. Rewriting the corresponding resolvent $(\tau\partial h + J_Z)^{-1}J_Z : Z \to Z$ as a minimization problem (under the assumption that $h \in \Gamma(Z)$ is finite and continuous at some point):

$$(\tau\partial h + J_Z)^{-1}J_Z(u) = \operatorname*{argmin}_{z\in Z} \tau\, h(z) + \mathcal{B}_Z(z, u)$$

we see that it is another generalization with respect to the Bregman distance.

On a Hilbert space Y Moreau's decomposition (e.g. [66, Theorem 31.5])

$$(\sigma\partial g^* + I)^{-1}(y) = y - \sigma\,(\partial g + \sigma I)^{-1}(y)\,, \quad y \in Y,$$

defines the resolvent of g^* by the resolvent of g. The following generalization gives us the opportunity to also calculate the generalized resolvent $(\sigma_k\, g^* + J_{Y^*})^{-1}$ in line (4.9) without knowledge of g^*. Moreover, it connects our generalization of the resolvent to the generalization

$$y \mapsto \bar{y} = \operatorname*{argmin}_{w\in Y} \sigma\, g(w) + \frac{1}{2}\|w - y\|_Y^2$$

which is common in the literature (cf. Section 3.2.2). By Theorem (3.2.7) as well as [23, Theorem 3.4] which states the surjectivity of $(\sigma\, J_Y \circ \partial g + I) : Y \to Y$ we obtain for any $g \in \Gamma(Y)$ that is continuous at one point, any $\sigma > 0$ and any $y \in Y$ the characterization

$$\operatorname*{argmin}_{w\in Y} \sigma\, g(w) + \frac{1}{2}\|w - y\|_Y^2 = \bar{y} = (\sigma\, J_{Y^*} \circ \partial g + I)^{-1}(y).$$

Lemma 4.3.1. *Suppose that Y is a reflexive Banach space and $g \in \Gamma(Y)$ is finite and continuous at at least one point. Then for any $\sigma > 0$ the following identity holds true:*

$$(\sigma\partial g^* + J_{Y^*})^{-1}(y) = J_Y\left(y - \sigma\,(J_{Y^*} \circ \partial g + \sigma I)^{-1}(y)\right), \quad y \in Y. \tag{4.29}$$

Proof. For $y \in Y$ let $\bar{p} \in Y^*$ be a solution to the minimization problem

$$\begin{aligned}\bar{p} &= \operatorname*{argmin}_{p\in Y^*} (\sigma g^*(p) - \langle y, p\rangle_Y + h_{2,Y^*}(p)) \\ &= \operatorname*{argmax}_{p\in Y^*} \left(-g^*(p) + \left\langle \frac{y}{\sigma}, p\right\rangle_Y - \frac{1}{\sigma} h_{2,Y^*}(p)\right),\end{aligned} \tag{4.30}$$

where $h_{2,Y^*}(p) := \frac{1}{2}\|p\|^2_{Y^*}$. Then $\bar{p}$ can be rewritten as $\bar{p} = (\partial\sigma g^* + J_{Y^*})^{-1}(y)$, cf. Equation (3.24). From Example 3.2.12 we conclude $\left(\frac{1}{\sigma}h_{2,Y^*}\right)^*(y) = \sigma\, h_{2,Y}(y)$. Therefore $\bar{y} \in Y$ given by

$$\bar{y} = (J_{Y^*} \circ \partial g + \sigma\, I)^{-1}(y) = \operatorname{argmin}_{z\in Y}\left(\sigma h_{2,Y}\left(z - \frac{y}{\sigma}\right) + g(z)\right),$$

is the solution $\bar{y} \in Y$ to the Fenchel dual problem corresponding to the primal problem (4.30) (cf. problems (3.12), (3.8)). Now the extremal relations (3.9) imply that $-\bar{p} \in \sigma\,\partial h_{2,Y}\left(\bar{y} - \frac{y}{\sigma}\right) = \sigma J_Y\left(\bar{y} - \frac{y}{\sigma}\right)$. Thus, we end up with

$$-(\partial\sigma g^* + J_{Y^*})^{-1}(y) = -\bar{p} = \sigma J_Y\left((J_{Y^*} \circ \partial g + \sigma I)^{-1}(y) - \frac{y}{\sigma}\right).$$

□

For the application of CP-BS to the inner minimization problems of the IRNM (3.3), where $g(y) = S(y^\delta; y + T(x_n) - T'[x_n](x_n))$, we are interested in the generalized resolvent $(\sigma\partial\tilde{g}^* + J_{Y^*})^{-1}$ of shifted functions $\tilde{g} = g(\cdot + y_0) \in \Gamma(Y)$ with $y_0 \in Y$. Due to

$$(J_{Y^*} \circ \partial\tilde{g} + \sigma I)^{-1}(y) = (J_{Y^*} \circ \partial g + \sigma I)^{-1}(y + \sigma y_0) - y_0, \quad \forall y \in Y$$

the last corollary gives the relation

$$\begin{aligned}(\sigma\partial\tilde{g}^* + J_{Y^*})^{-1}(y) &= J_Y\left(y + \sigma y_0 - \sigma\,(J_{Y^*} \circ \partial g + \sigma I)^{-1}(y + \sigma\, y_0)\right) \\ &= (\sigma\partial g^* + J_{Y^*})^{-1}(y + \sigma\, y_0), \qquad y \in Y.\end{aligned} \tag{4.31}$$

Calculating the generalized resolvent $(\tau\partial f + J_X)^{-1}$ of shifted penalty terms $f = \alpha\, R(x - x_0)$ with $x_0 \in X$ becomes more complicated as we will discuss later on.

Next, we want to study some standard examples for f and g in (**P**). First of all, under assumption of Problem 3.12 let us consider the case that f and g are given by the squares of the corresponding Banach space norms:

$$f(x) := \frac{1}{2}\|x\|^2_X, \quad g(y) = \frac{1}{2}\|y - y_0\|^2_Y, \tag{4.32}$$

where $y_0 \in Y$ is again a shift vector. As already mentioned, in inverse Problems it is quite natural to minimize a (shifted) Banach norm monomial on the corresponding Banach space. If X and Y are Hilbert spaces the resolvents corresponding to (4.32) reduce to scalar multiplications

$$(\tau\partial f + I)^{-1}(x^*) = \frac{1}{1+\tau}x^*, \quad (\sigma\partial g^* + I)^{-1}(y) = y - \frac{\sigma}{1+\sigma}\left(y + \frac{1}{\sigma}y_0\right) = \frac{y - \sigma y_0}{1+\sigma}$$

for any $\tau, \sigma > 0$ and any $x^* \in X$, $y \in Y$. Here we used Moreau's decomposition in order to calculate the resolvent of g^*. And also property (4.7) is satisfied by $f : x \mapsto \frac{1}{2}\|x\|^2_X$ and $g^* : y^* \mapsto \frac{1}{2}\left(\|y^*\|^2_{Y^*} + \langle y_0, y^*\rangle_Y\right)$ (cf. Example 3.2.12 and Corollary 4.1.3). So let us check if also in our Banach space setting the choice (4.32) is predestined for the application of the generalized algorithm CP-BS : Applying the Theorem of Asplund as well as the generalization of Moreau's decomposition (4.29), we deduce that the generalized resolvents of f and g^* are given by corresponding duality mappings:

$$(\tau\partial f + J_X)^{-1}(x^*) = (\tau J_X + J_X)^{-1}(x^*) = \frac{1}{\tau+1} J_{X^*}(x^*) \tag{4.33}$$

$$(\sigma\partial g^* + J_{Y^*})^{-1}(y) = J_Y\left(y - \sigma\left(J_Y\, J_{Y^*}(\cdot - y_0) + \sigma I\right)^{-1}(y)\right) = J_Y\left(\frac{y - \sigma y_0}{\sigma + 1}\right). \tag{4.34}$$

This, for instance, guarantees a closed form of the operators $(\tau\partial f + J_X)^{-1} : X^* \to X$ and $(\sigma\partial g^* + J_{Y^*})^{-1} : Y \to Y^*$ with g, f defined by Equation (4.32) for any Banach space $X, Y \in \left\{ l^r_W \mid r \in (1,\infty), W \text{ positive weight} \right\}$. However, if we consider f and g to be given as $f(x) := \frac{1}{2}\|x\|^2_Z$ and $g(y) = \frac{1}{2}\|y - y_0\|^2_Z$ for a Banach norm $\|\cdot\|_Z$ in the original Hilbert space setting of CP , where $X \neq Z$ and $Y \neq Z$, typically a whole system of nonlinear equations has to be solved in order to calculate the resolvents $(\tau J_X + I)^{-1}$ and $(\sigma(J_{Y^*} + y_0) + I)^{-1}$. So the generalization CP-BS is efficiently applicable to a wider class of functions f and g than CP . Moreover, for any $u, x \in X$, $x^* \in \partial f(x) = J_X(x)$ and any $y^*, p \in Y^*$, $y = \partial J_{Y^*}(p)$ we have

$$f(u) - f(x) - \langle u - x, x^*\rangle_X = \mathcal{B}_X(u, x), \quad g^*(p) - g^*(y^*) - \langle y, p - y^*\rangle_Y = \mathcal{B}_Y(y^*, p).$$

Thus, we conclude:

Corollary 4.3.2. *f and g^*, defined by* (4.32)*, satisfy the Bregman midconvex property* (4.21) *for any modulus $\gamma \in (0, 1]$.*

So the case (4.32) not only provides sufficiently simple generalized resolvents but also allows the application of the accelerated versions CP-BS 2 and CP-BS 3. More general, the following Corollary shows that if the evaluation of the duality mappings J_X and J_Y is sufficiently cheap, then for any arbitrary exponent $r \in (1, \infty)$ the generalized resolvents of

$$f(x) := \frac{1}{r}\|x\|^r_X, \qquad g(y) := \frac{1}{r}\|y - y_0\|^r_Y, \quad \text{where } y_0 \in Y \tag{4.35}$$

become rather simple:

Corollary 4.3.3. *For $\sigma, \tau > 0$ and f, g given by* (4.35)*, we have*

$$(\tau\partial f + J_X)^{-1}(x^*) = \frac{1}{\tau\alpha^{r-2} + 1} J_{X^*}(x^*), \quad x^* \in X \tag{4.36}$$

$$(\sigma\partial g^* + J_{Y^*})^{-1}(y) = \frac{1}{\beta^{r-2} + \sigma} J_Y(y - \sigma y_0), \quad y \in Y, \tag{4.37}$$

where $\alpha \geq 0$ is the maximal solution of $\tau\alpha^{r-1} + \alpha = \|x^*\|_{X^*}$ and $\beta \geq 0$ the maximal solution of $\beta^{r-1} + \sigma\beta = \|y - \sigma y_0\|_Y$.

Proof. We set $x := (\tau\partial f + J_X)^{-1}(x^*)$. Then the identity $\partial f(x) = J_{r,X}(x) = \|x\|_X^{r-2}\, J_X(x)$ implies that $\left(\tau\,\|x\|_X^{r-2} + 1\right) J_X(x) = x^*$. Applying the norm to both sides of this equation we conclude $\alpha = \|x\|_X = \|J_X(x)\|_{X^*} \geq 0$. Then inserting α gives the first assertion. In order to prove Equation (4.37) we set $\tilde{y} := (J_{Y^*} \circ \partial g + \sigma I)^{-1}(y)$. Due to the identity $\partial g(x) = J_{r,X}(x) = \|x\|_X^{r-2}\, J_X(x)$ we have

$$y - \sigma y_0 = J_{Y^*} \circ \partial g(\tilde{y}) + \sigma(\tilde{y} - y_0) = \left(\|\tilde{y} - y_0\|_Y^{r-2} + \sigma\right)(\tilde{y} - y_0)\,.$$

Now the application of the norm $\|\cdot\|_Y$ yields $\beta = \|\tilde{y} - y_0\|_Y$ and $\tilde{y} = \frac{y + \beta^{r-2} y_0}{\beta^{r-2} + \sigma}$. By Lemma 4.3.1 the assertion follows:

$$(\sigma\partial g^* + J_{Y^*})^{-1}(y) = J_Y(y - \sigma\tilde{y}) = J_Y\left(\frac{y - \sigma\, y_0}{\beta^{r-2} + \sigma}\right).$$

□

As mentioned in Section 3.3 there exist closed or sufficiently simple resolvents also for other interesting, not necessary differentiable functions g^* and f. The next example shows that under the assumption that the evaluation of J_X (or J_Z) is sufficiently cheap, this also applies for the generalized resolvents.

Example 4.3.4. (i) Suppose that X is an N-dimensional reflexive Banach space. Then, by analogy to Example 3.2.10, we obtain that the generalized resolvent of the norm $f(x) = \|x\|_{l^1}$ and $\tau > 0$ is given by

$$(\tau\partial f + J_X)_i^{-1}(x^*) = J_{X^*}\left(\max\{|x_i^*| - \tau, 0\}\ \mathrm{sign}(x_i^*)\right) \qquad x^* \in X^*,\ i \in \{1, \ldots, N\}\,.$$

So here the generalized resolvent differs from the resolvent just by the duality mapping.

(ii) For any $N \in \mathbb{N}$ and any positive weight $W = (w_i)_{i=1,\ldots,N}$ the weighted sequence space $Y = l^2_W(\{1, \ldots, N\})$ is a Hilbert space. The corresponding the inner product is given by $\langle y, y^* \rangle = \sum w_i\, y_i\, y_i^*$. But with respect to the inner product $\left\langle y_i, y_i^* \right\rangle = \sum y_i\, y_i^*$ the space $Y = l^2_W(\{1, \ldots, N\})$ has the interpretation of a Banach space with $l^2_{W^{-1}}$ as its dual. In the sense of this second view let us determine the generalized resolvent $(\sigma\partial g^* + J_{Y^*})^{-1}$ of $g = kl_{\mathbf{N}}(y^\delta; \cdot) : l^2_W(\triangle_{\mathbf{N}}) \to \overline{\mathbb{R}}$, where the definitions are identical to those in Example 3.2.10. In analogy to Example 3.2.10 we obtain by $J_{l^2_W}(y) = W y$ the identity

$$\left(\left(J_{l^2_{W^{-1}}} \circ \partial kl(y^\delta; \cdot) + \sigma I\right)^{-1}(y)\right)_{\mathbf{i}} = \frac{y_{\mathbf{i}} - w_{\mathbf{i}}^{-1}}{2\,\sigma} + \frac{\sqrt{\left(y_{\mathbf{i}} - w_{\mathbf{i}}^{-1}\right)^2 + 4\,\sigma\, w_{\mathbf{i}}^{-1}\, y_{\mathbf{i}}^\delta}}{2\sigma}$$

for any $y \in l^2_W(\triangle_{\mathbf{N}})$ and any $\mathbf{i} \in \triangle_{\mathbf{N}}$. Hence, we obtain the generalized resolvent of $kl(y^\delta; \cdot)^*$ by Lemma 4.3.1:

$$\left(\left(\sigma\partial kl(y^\delta; \cdot)^* + J_{l^2_{W^{-1}}}\right)^{-1}(y)\right)_{\mathbf{i}} = \frac{w_{\mathbf{i}}\, y_{\mathbf{i}} + 1}{2} + \frac{w_{\mathbf{i}}\,\sqrt{\left(y_{\mathbf{i}} - w_{\mathbf{i}}^{-1}\right)^2 + 4\,\sigma\, w_{\mathbf{i}}^{-1}\, y_{\mathbf{i}}^\delta}}{2}.$$

(iii) On a reflexive Banach space X we consider the indicator function $\chi_C \in \Gamma(X)$, given by 3.4, of a convex set $C \subset X$ including an open subset. Then we obtain for any $x^* \in X^*$ and any positive τ

$$(\tau\,\partial\iota_C + J_X)^{-1}\,(x^*) = \operatorname*{argmin}_{z\in X}\left(\tau\chi_C(z) - \langle z, x^*\rangle_X + \frac{1}{2}\|z\|_X^2\right)$$
$$= \operatorname*{argmin}_{z\in C}\left(\frac{1}{2}\,\|J_{X^*}(x^*)\|_X^2 - \langle z, J_X\,(J_{X^*}(x^*))\rangle_X + \frac{1}{2}\|z\|_X^2\right) = \pi_C(x^*)$$

where $\pi_C : X^* \to C$, $\pi_C(x^*) := \operatorname*{argmin}_{z\in C}\mathcal{B}_X(z, J_{X^*}(x^*))$ denotes the generalized projection introduced by Alber [2].

In inverse problems one frequently also wants to minimize Tikhonov-type functionals (3.1) which incorporates a shift vector $x_0 \neq 0$ in the penalty term R:

$$x_\alpha = \operatorname*{argmin}_{x\in X} S\,(y^\delta; Tx) + \alpha\, R(x - x_0).$$

This approach, for example, allows to take an initial guess $x_0 \neq 0$ for the solution into account. For a Hilbert space X a closed form of the resolvent $(\tau\,\partial\alpha R + I)^{-1}$ directly provides a closed form of the resolvent of the shifted penalty term $f := \alpha\, R(\cdot - x_0)$:

$$(\tau\,\partial f + I)^{-1}\,(x^*) = (\tau\,\partial\alpha\, R + I)^{-1}\,(x^* - x_0) + x_0, \quad x^* \in X^* = X.$$

However, that is not the case in the more general Banach space setting where the duality mappings J_X are nonlinear. One way to overcome this drawback, is to consider the equivalent minimization problem

$$x_\alpha = \operatorname*{argmin}_{u\in X} \tilde{S}\,(y^\delta; Tu) + \alpha\, R(u).$$

with shifted data fidelity functional $\tilde{S}(y^\delta; y) = S\,(y^\delta; y + Tx_0)$. Then the generalized resolvent $\left(\sigma\,\tilde{S}^*(y^\delta;\cdot) + J_{Y^*}\right)^{-1}$ can be evaluated with the help of Equation (4.31). Another possibility to deal with this problem is to redefine the Banach space X. Although this approach is more complicated, we want to study it in detail because it characterizes the preimage space X as a powerful tool for modeling conditions on the solution. So, instead of X let us consider the "shifted" space $Z = \{[x] = x + x_0\,|x \in X\} = X + x_0$, with addition $\oplus : Z^2 \to Z$ and scalar multiplication $\odot : \mathbb{C}\times Z \to Z$ given by

$$[x_1]\oplus[x_2] := [x_1 + x_2] = (x_1 + x_2) + x_0, \quad \lambda\odot[x] := [\lambda\, x] = \lambda x + x_0.$$

Equipped with the norm $\|[x]\|_Z := \|x\|_X$ it becomes a Banach space. The dual space with respect to the scalar product

$$\langle [x], [x^*]_*\rangle_Z = \langle x, x^*\rangle_X$$

is $Z^* = X^*/\,\{J_X(x_0)\} = \{[x]_* = x^* + J_X(x_0)\,|x^* \in X^*\}$ where the corresponding norm reads as

$$\|[x^*]_*\|_{Z^*} = \sup_{\|[x]\|_Z\le 1}\,|\,\langle [x], [x^*]_*\rangle_Z\,| = \sup_{\|x\|_X\le 1}\,|\,\langle x, x^*\rangle_X\,| = \|x^*\|_{X^*}.$$

Obviously, the map $\pi : X \to Z,\ x \mapsto x + x_0 = [x]$ is an isometric isomorphism with adjoint $\pi^* : Z^* \to X^*$, $\pi^*([x^*]_*) = x^*$ and inverse $\pi^{-1}([x]) = x$. Therefore we have

$$J_Z([x]) = (\pi^*)^{-1} \circ J_X \circ \pi^{-1}([x]) = J_X(x) + J_X(x_0), \quad [x] = x + x_0 \in Z$$

as well as

$$J_{Z^*}([x^*]_*) = \pi \circ J_{X^*} \circ \pi^*([x^*]_*) = J_{X^*}(x^*) + x_0 \quad [x^*]_* = x^* + J_X(x_0) \in Z^*.$$

Suppose that X satisfies the assumptions of Problem 3.12, i.e. X is reflexive, 2-convex and smooth. Then one can easily check that this also applies for Z. By the next corollary we translate the update rule of Algorithm 2 given on the space Z into terms of the original space X:

Corollary 4.3.5. *Under the assumptions of Problem 3.12, consider the space $Z = X + x_0$ introduced above. Let $f_Z \in \Gamma(Z)$, $g_Z \in \Gamma(Y)$ denote the functions given by $f_Z([x]) = f(x)$ and $g_Z(y) = g(y + T(x_0))$, respectively. By the definition $T_Z([x]) := T(\pi^{-1}([x])) = T(x)$ for all $[x] \in Z$ the linear operator $T : X \to Y$ determines a linear operator T_Z on Z. Applying algorithm CP-BS to the problem*

$$[\bar{x}] = \operatorname*{argmin}_{[x]\in Z} g_Z(T_Z[x]) + f_Z([x]) = \operatorname*{argmin}_{[x]\in Z} g(T[x]) + f([x] - x_0).$$

the update rule becomes:

$$\begin{aligned}
p_{k+1} &= (\sigma_k\, \partial g_Z^* + J_{Y^*})^{-1} \left(J_{Y^*}(p_k) + \sigma_k\, T_Z\, [\hat{x}_{k+1}]\right) \\
&= (\sigma_k\, \partial g^* + J_{Y^*})^{-1} \left(J_{Y^*}(p_k) + \sigma_k\, T\, [\hat{x}_{k+1}]\right) \\
[x_{k+1}] &= (\tau \odot \partial f \oplus J_Z)^{-1} \left(J_Z([x_k]) \ominus \tau \odot T_Z^* p_{k+1}\right) \\
&= (\tau\, \partial f + J_X)^{-1} \left(J_Z([x_k]) - \tau\, T^* p_{k+1} - J_X(x_0)\right) + x_0 \\
[\hat{x}_{k+1}] &= [x_{k+1}] \oplus \theta_k \odot ([x_{k+1}] \ominus [x_k]) = [x_{k+1}] - \theta_k\, ([x_{k+1}] - [x_k])
\end{aligned}$$

Proof. The linearity of T_Z directly follows from the identity

$$T_Z(\lambda \odot ([u] \oplus [x])) = T_Z([\lambda(u + x)])) = T\,(\lambda(u + z)) \quad [u], [x] \in Z, \lambda \in \mathbb{R}.$$

Rewriting

$$J_{Y^*}(p_k) + \sigma_k\, T_Z\, [\hat{x}_{k+1}] = \left(\sigma_k\, \partial g_Z^* + J_{Y^*}\right) p_{k+1} = \left(\sigma_k\, \partial g^* + J_{Y^*}\right) p_{k+1} - \sigma_k T x_0$$

together with $T_Z\, [\hat{x}_{k+1}] + T x_0 = T[\hat{x}_{k+1}]$ yields the update of the dual variable p_k. One can easily check the expression for the update of the over-relaxation step. Thus, it remains to verify the update rule for the primal variable x_k defined in line (4.10). For this purpose we consider the subdifferential of f_Z at $[x] \in Z$:

$$\begin{aligned}
\partial f_Z([x]) &= \{[x^*]_* \in Z^* \mid \langle [u] \ominus [x], [x^*]_* \rangle_Z \le f_Z([u]) - f_Z([x]) = f(u) - f(x), \forall u \in X\} \\
&= \{x^* + J_X(x_0) \in X^* \mid \langle u - x + x_0 - x_0, x^* \rangle_X \le f(u) - f(x), \forall u \in X\} \\
&= \partial f(x) + J_X(x_0) = [\partial f(x)]_*.
\end{aligned}$$

Consequently, the generalized resolvent $[x] := (\tau \odot \partial f_Z \oplus J_Z)^{-1}([x^*]_*)$ at $[x^*]_* \in Z^*$ satisfies

$$[x^*]_* = [\tau\,\partial f(x)]_* \oplus [J_X(x)]_* = [\tau\,\partial f(x) + J_X(x)]_*,$$

and hence it can be rewritten as

$$(\tau \odot \partial f_Z \oplus J_Z)^{-1}([x^*]_*) = (\tau\,\partial f + J_X)^{-1}([x^*]_* - J_X(x_0)) + x_0.$$

Now the assertion follows from

$$\begin{aligned} J_Z([x_k]) \ominus \tau \odot T_Z^* p_{k+1} &= [J_X(x_k)]_* \ominus \tau \odot [T^* p_{k+1}]_* = [J_Z(x_k) - \tau\, T^* p_{k+1}]_* \\ &= J_Z([x_k]) - \tau\, T^* p_{k+1}. \end{aligned} \qquad \square$$

So, this reformulation of CP-BS allows to incorporate a shift vector $x_0 \neq 0$ without increasing the algorithm's complexity compared to $x_0 = 0$. Note that if $T : X \to Y$ is the Fréchet derivative of an operator $F : X \to Y$ at some point $u \in X$ (cf. IRNM), the linear operator $T_Z : Z \to Y$ is the Fréchet derivative of $F_Z : Z \to Y, F_Z(z) := F(z)$ at $[u - x_0] = u$.

4.3.2 Duality mapping of Sobolev spaces

For the aim of solving phase retrieval problems by the IRNM, we also consider X to be given by the discretization of the Sobolev spaces $W^{1,r}(\Omega) := \left\{\phi : \overline{\Omega} \to \mathbb{R} \mid \phi, \nabla\phi \in L^r\right\}$, $r \in (1, \infty)$, on an open interval $\Omega = (-\mathbf{r}_X, \mathbf{r}_X) \subset \mathbb{R}^2$. Recall from Example 3.3.2 that, associated with the norm

$$\|\phi\|_{W^{1,r}(\Omega)} = \left(\|\phi\|^r_{L^r(\Omega)} + \|\nabla\phi\|^r_{L^r(\Omega)}\right)^{\frac{1}{r}},$$

$W^{1,r}(\Omega)$ is a separable, reflexive, $\max\{r, 2\}$-convex, and $\min\{r, 2\}$-smooth Banach space. One opportunity to obtain a corresponding duality mapping is to use the well-known subspace $W_0^{1,r}(\Omega) := \left\{\phi : \overline{\Omega} \to \mathbb{R} \mid \phi, \nabla\phi \in L^r,\ \phi|_{\partial\Omega} = 0\right\}$ with homogeneous boundary condition. By equipping $W_0^{1,r}(\Omega)$ with the equivalent norm $\|\nabla\phi\|_{L^r(\Omega)}$, the normalized duality mapping $J_{W_0^{1,r}(\Omega)} : W_0^{1,r}(\Omega) \to W^{-1,r^*}(\Omega)$ with respect to the $L^2(\Omega)$-scalar product reads as (see [28])

$$\left\langle J_{W_0^{1,r}(\Omega)}(\phi), \varphi\right\rangle_{W_0^{1,r}(\Omega)} = \|\phi\|^{2-r}_{W_0^{1,r}} \left\langle J_{r,W_0^{1,r}(\Omega)}(\phi), \varphi\right\rangle_{W_0^{1,r}(\Omega)} = -\|\phi\|^{2-r}_{W_0^{1,r}} \int_\Omega \left(|\nabla\phi|^{r-2}\nabla\phi\right)\nabla\varphi.$$

In order to avoid the considerable effort of evaluating this duality mapping, we use the characterization $W^{s,r}(\Omega) = H^{s,r}(\Omega), s \in \mathbb{N}, r \in (1, \infty)$ by fractional Sobolev spaces $H^{s,r}(\Omega)$ (also known as Liouville spaces or Bessel potential spaces (see e.g. [1, pp. 252], and [77, pp. 208]), which we consider on $\Omega = (-\mathbf{r}_X, \mathbf{r}_X)$ with periodic boundary conditions: To define these spaces we introduce the Bessel potential operators $\Lambda_s := (I - \Delta)^s$ by

$$\Lambda_s\,\phi\,(\mathbf{x}') := \sum_{\mathbf{k}\in\mathbb{Z}^2} \left(1 + \left|\frac{\pi\mathbf{k}}{\mathbf{r}_X}\right|^2\right)^{s/2} c_{\mathbf{k}}\left(\phi_{2\mathbf{r}_X}\right)\,\exp\left(\pi\mathrm{i}\frac{\mathbf{k}}{\mathbf{r}_X}\cdot\mathbf{x}'\right), \qquad s \in \mathbb{R},$$

a-priori for $\phi \in C^\infty(\Omega)$ where $c_{\mathbf{k}}\left(\phi_{\mathbf{r}_X}\right) := (r_{X,1}\, r_{X,2})^{-2} \int_\Omega \exp(\frac{-\pi \mathrm{i}\, \mathbf{k}}{\mathbf{r}_X} \cdot \mathbf{x}')\, \phi(\mathbf{x}')\, d\mathbf{x}'$ denote the Fourier coefficients of the $2\mathbf{r}_X$-periodization $\phi_{2\mathbf{r}_X}$ (see Equations (A.5), (A.6)). Note that $\Lambda_0 = I$ and $\Lambda_s \Lambda_t = \Lambda_{s+t}$ for all $s, t \in \mathbb{R}$. For $s \geq 0$ and $r \in (1, \infty)$ the operators Λ_{-s} have continuous extensions to $L^r(\Omega)$, and so the Sobolev spaces

$$H^{s,r}(\Omega) := \Lambda_{-s}\, L^r(\Omega) \qquad \text{with norms} \qquad \|\phi\|_{H^{s,r}(\Omega)} := \|\Lambda_s\, \phi\|_{L^r(\Omega)}$$

are well defined. Actually, this definition also makes sense for $s < 0$, and (with respect to the $L^2(\Omega)$-inner product) we have the duality relation

$$(H^{s,r}(\Omega))^* = H^{-s,r^*}(\Omega)$$

for $1/r + 1/r^* = 1$ (see e.g. [72, §13.6]). The normalized duality mapping

$$J_{H^{s,r}(\Omega)} : H^{s,r}(\Omega) \to H^{-s,r^*}(\Omega)$$

has a sufficiently simple form

$$J_{H^{s,r}(\Omega)} = \Lambda_{-s}\, J_{L^r}\, \Lambda_s.$$

$H^{s,r}(\Omega)$ is a separable, reflexive, $\max\{r, 2\}$-convex and $\min\{r, 2\}$-smooth Banach space (cf. [1]). In order to define the discrete counterpart $h^{s,r}(\triangle(\mathbf{r}_X))$ of $H^{s,r}(\Omega)$ we discretize the rectangle $\overline{\Omega} = [-\mathbf{r}_X, \mathbf{r}_X] \subset \mathbb{P}_0$ by the grid (cf. Appendix A.2)

$$\Delta(\mathbf{r}_X) := \left\{ \mathbf{r}_X \bullet \frac{2\mathbf{j}}{\mathbf{N}} := \left(r_{X,1} \frac{2\, j_1}{N_1}, r_{X,2} \frac{2\, j_2}{N_2} \right) \middle|\ \mathbf{j} \in \Delta_{\mathbf{N}} \right\}, \tag{4.38}$$

with

$$\Delta_{\mathbf{N}} := \left\{ -\frac{N_1}{2}, -\frac{N_1}{2} + 1, \ldots, \frac{N_1}{2} - 1 \right\} \times \left\{ -\frac{N_2}{2}, -\frac{N_2}{2} + 1, \ldots, \frac{N_2}{2} - 1 \right\}$$

and a sufficiently large number of sample points $\mathbf{N} := (N_1, N_2) \in \mathbb{N}^2$. Here $\bullet$ again denotes the pointwise multiplication. Then sampling $\phi \in H^{s,r}(\Omega)$, $r \in (1, \infty)$, $s \in \mathbb{R}$ on this grid we obtain $\underline{\phi}_{\mathbf{r}_X} \in h^{s,r}(\triangle(\mathbf{r}_X))$ and Λ_s becomes

$$\underline{\Lambda_s}\, \underline{\phi}_{\mathbf{r}_X} = \mathcal{F}_{\mathbf{N}}^{-1} \left[\left(1 + \left|\xi_{\mathbf{j}}\right|^2\right)^{\frac{s}{2}}_{\mathbf{j} \in \triangle_{\mathbf{N}}} \bullet \mathcal{F}_{\mathbf{N}} \left(\underline{\phi}_{\mathbf{r}_X} \right) \right], \qquad \xi_{\mathbf{j}} := \frac{\mathbf{r}_\xi \bullet 2\mathbf{j}}{\mathbf{N}} \in \triangle(\mathbf{r}_\xi)$$

where $\mathbf{r}_\xi = \frac{\pi}{2} \left(\frac{N_2}{r_{X,1}}, \frac{N_2}{r_{X,2}} \right)$. Accordingly we define

$$J_{h^{s,r}(\triangle(\mathbf{r}_X))} = \underline{\Lambda_{-s}}\, J_{l^r(\triangle(\mathbf{r}_X))}\, \underline{\Lambda_s},$$

as well as $\|\underline{\phi}_{\mathbf{r}_X}\|_{h^{s,r}(\triangle(\mathbf{r}_X))} = \|\underline{\Lambda_s}\, \underline{\phi}_{\mathbf{r}_X}\|_{l^r(\triangle(\mathbf{r}_X))}$, for any $\underline{\phi}_{\mathbf{r}_X} \in h^{s,r}(\triangle(\mathbf{r}_X))$, $r \in (1, \infty)$, $s \in \mathbb{R}$. So, the duality mapping $J_{h^{s,r}(\triangle(\mathbf{r}_X))}$ and the forward operators T_{Fresnel}, T_{Frau} associated with the phase retrieval problems in x-ray imaging, are of comparable complexity.

5 Numerical examples: Solving phase retrieval problems by CP-BS

In this section, we test the performance of the proposed algorithm 2 (CP-BS). As motivated above, we apply our method to the regularization functionals (3.1) and (3.3) with the aim of solving linear or nonlinear inverse problems $Tx = y$ in Banach spaces. For this purpose the linear convolution problem introduced by Example 3.1.1 poses a good test problem since Tikhonov-type regularization (3.1), unlike the IRNM (3.3), involves no outer method. The reconstruction of the phase information ϕ in phase retrieval problems occurring in coherent x-ray imaging (cf. Section 1.1) will be the topic of Section 5.3. Most of these examples have been already described in [45]. As a last example, we consider in Section 5.4 a medium scattering problem as defined in Section 1.2. But, first let us summarize some properties and results which are required for the implementation.

5.1 Preliminaries

In most examples, the preimage and image space of the operator $T : X \to Y$ will be (finite dimensional) weighted sequence spaces

$$l^r_W(I) := \left\{ x = (x_i)_{i\in I} \in \mathbb{R}^{\#I} \;\middle|\; \|x\|_{l^r_W} = \left(\sum_{i\in I} w_i |x_i|^r \right)^{\frac{1}{r}} < \infty \right\},$$

with $r \in (1, \infty)$, finite index set I, and positive weight $W = (w_i)_{i\in I}$ (cf. Example 3.3.2 (ii)). In these cases the operator norm which is required for the parameter choice of CP-BS 1 - CP-BS 3 can be efficiently calculated by the power method of Boyd [17]. To be more precise, the iterative method computes the operator norm $\|A\|$ of nonnegative matrices $A : l^r(I_X) \to l^s(I_Y)$ defined on unweighted finite dimensional sequence spaces $l^r(I_X), l^s(I_Y)$, $r,\ s \in (1, \infty)$ and the relative norm $\|Au_0\|_{l^s(I_Y)}$, where

$$u_0 = \operatorname*{argmax}_{u\in l^r(I_X)} \left\{ \frac{\|Au\|_{l^s(I_Y)}}{\|u\|_{l^s(I_Y)}} \mid \|u\|_{l^r(I_X)} = 1 \right\}$$

if A has negative entries as well. So, in order to apply this method to $X = l^r_W(I_X)$ and $Y = l^s_V(I_Y)$, with weights $W,\ V$ different from $\mathbf{1}$, we redefine the operator $T : X \to Y$ as a mapping on the unweighted spaces $A : l^r(I_X) \to l^s(I_Y)$, $A := V^{\frac{1}{s}} \bullet T \bullet W^{-\frac{1}{r}}$ such that

$$\|A\| = \max_{u\in l^r(I_X)} \left\{ \|Au\|_{l^s(I_Y)} \mid \|u\|_{l^r(I_X)} = 1 \right\} = \max_{x\in l^r_W(I_X)} \left\{ \|Tx\|_{l^s_V(I_Y)} \mid \|x\|_{l^r_W(I_X)} = 1 \right\}.$$

For the aim of modeling blocky structured solution, we will also choose discrete Sobolev spaces $h^{s,r}(\triangle(\mathbf{r}_X)), s \in \mathbb{R}, r \in (1, \infty)$ given on some grid $\triangle(\mathbf{r}_X)$ as preimage spaces X (see Section 4.3.2). Again, the operator norm $\|T\|$ of a nonnegative linear mapping $T : h^{s,r}(\triangle(\mathbf{r}_X)) \to l^s_V(I_Y)$ can be computed by the power method proposed in [17]: Setting $A = V^{\frac{1}{s}} \bullet T\, \Lambda_{-s} : l^r(\triangle(\mathbf{r}_X)) \to l^s_V(I_Y)$, we have

$$\|A\| = \max_{\phi\in h^{s,r}(\triangle(\mathbf{r}_X))} \left\{ \|T\phi\|_{l^s_V(I_Y)} \mid \|\Lambda_s\, \phi\|_{l^r(\triangle(\mathbf{r}_X))} = 1 \right\} = \|T\|.$$

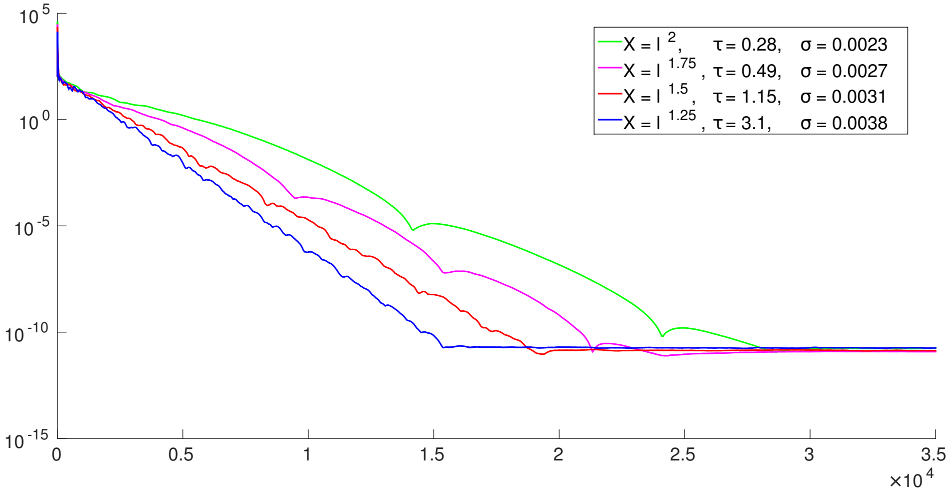


Figure 5.1: Convergence plot for CP-BS 1 applied to the sparse convolution problem of Figure 3.1. For different choices of X the error $\|x_k - x_\alpha\|_{l^1}$ between an element of the sequence $(x_k)_{k\in\mathbb{N}}$ and the true minimizer x_α of (3.1) is shown as function of the iteration step k. The parameters τ and σ are optimally chosen (according to Remark 4.2.6).

Based on this preliminary work, we are now able to study the performance of CP-BS 1-3 applied to different Tikhonov-type functionals. For all versions of the algorithm we relax the parameter choice of σ and τ according to Remark 4.2.6.

5.2 Convolution problem

Let us start the numerical examination of the generalized algorithm CP-BS with the convolution problem introduced in Example 3.1.1. In order to test the intuition that a problem adapted choice of X and Y will have positive effects on the algorithm's performance, we consider again the special sparse convolution problem of Figure 3.1 and apply CP-BS 1 to the corresponding Tikhonov-type regularization

$$\bar{x} = \bar{x}_\alpha = \operatorname*{argmin}_{x\in X} \frac{1}{2}\|y^\delta - Tx\|^2_{l^2(I_Y)} + 5\|x\|_{l^1(I_Y)},$$

with different - more or less problem adapted - preimage spaces $X = l^r(I_X)$ where $r = 2, 1.75, 1.5, 1.25$. Here $I_X = \left\{-\frac{1}{2}, -\frac{1}{2} + \frac{1}{N-1}, ..., \frac{1}{2}\right\}$ denotes the discretization of $\left[\frac{1}{2}, \frac{1}{2}\right]$ with $N = 501$ sample points and $I_Y = \left\{-1, -1 + \frac{1}{N-1}, ..., 1\right\}$ the one of $[-1, 1]$. Then the operator $T : X \to Y$ becomes the discrete convolution which can be rewritten in terms of discrete Fourier transforms as

$$T(x)(j) = \left(\mathcal{F}^{-1}_{2N-1}\left(\hat{k} \bullet \mathcal{F}_{2N-1}\tilde{x}\right)\right)_j, \qquad j \in I_Y,\ x \in X,$$

σ	0.007	(τ)	0.0023	(τ)	0.00075	(τ)
$X = l^2(I_X)$	76476	(0.0915)	22368	(0.279)	39418	(0.854)
$X = l^{1.25}(I_X)$	26271	(1.7)	16575	(4.8)	38710	(14.66)

Table 5.1: According to Figure 5.1, the table compares the performance of CP 1 with $X = l^2(I_X)$ and CP-BS 1 with $X = l^{1.25}(I_X)$ for different choices of σ. It shows the necessary number of iterations until the error $\|x_\alpha - x_k\|_{l^1}$ is less than 10^{-5}, averaged over 100 experiments.

where $\tilde{x} = \{\tilde{x}_i \mid \tilde{x}_i := x(i) \text{ for } i \in I_x,\ \tilde{x}_i := 0 \text{ for } i \in I_Y \backslash I_X\}$ is the vector $x = (x_i)_{i \in I_X}$ padded with $N-1$ zeros and $\hat{k}$ is discrete Fourier transform $\mathcal{F}_{2N-1}\tilde{k}$ of the zero-padded kernel k. The generalized resolvents of $g(y) = \frac{1}{2}\|y^\delta - y\|^2_{l^2(I_Y)}$ and $f(x) = 5\|x\|_{l^1(I_Y)}$ are given by Equation (4.34) and Example 4.3.4. Their closed forms ensure that the algorithm CP-BS is efficiently applicable. Inspired by the optimality condition $T\bar{x} \in \partial g^*(\bar{p})$, where $g^*(p) = \frac{1}{2}\|p\|^2_{l^2(I_Y)} + \left\langle y^\delta, p \right\rangle_{l^2}$, we choose

$$T x_0 \in \partial g^*(p_0) = p_0 + y^\delta \Leftrightarrow p_0 = T x_0 - y^\delta \tag{5.1}$$

and $x_0 = 0$ as an initial guess.

Figure 5.1 and Table 5.1 show that a problem adapted choice of X, which in this case means $X = l^r(I_X)$ with $r \approx 1$, actually accelerates the convergence of CP-BS: The original algorithm CP 1 applied to the Hilbert space $l^2(I_X)$ needs at least ≈ 29581 iterations to converge, while for its generalization CP-BS 1 applied to $X = l^{1.25}(I_X)$ only ≈ 15534 iterations are necessary (cf. Figure 5.1). Moreover, we note that the optimal choice of σ, which also defines τ via $\tau = C\sigma^{-1}\|T\|^{-2} \approx \sigma^{-1}\|T\|^{-2}$ for some $C \lesssim 1$, is not so different for the various Banach spaces. This makes sense in so far that, in the right hand side of Equation (4.12), the parameter σ weights the initial distance $\mathcal{B}_{Y^*}(\bar{p}, p_0)$ in the dual variable p, which is the same for any X, while τ weights $\mathcal{B}_X(\bar{x}, x_0)$, where we have $\mathcal{B}_{l^{1.25}}(\bar{x}, x_0) \approx 3.6\,\mathcal{B}_{l^2}(\bar{x}, x_0)$ (in this particular setting). Accordingly, we found that the optimal parameter choice also depends (to a smaller extent) on the concrete given data y^δ. However, it seems that the optimal parameter choice is not only determined by the term $\triangle_0(\bar{x}, \bar{p})$ together with the (relaxed) condition $\sqrt{\sigma}\,\sqrt{\tau}\,\|T\| \lesssim 1$: Although, for $X = l^{1.25}$, the $\mathcal{B}_{Y^*}(\bar{p}, p_0) = 25418$ is ≈ 11 times greater than $\mathcal{B}_X(\bar{x}, x_0) = 2273$, the quotient $\frac{0.0038}{3.1} \approx 0.0012$ of the corresponding optimal parameters σ and τ is much less than 1. Also, $\frac{\mathcal{B}_{Y^*}(\bar{p}, p_0)}{\|T\|^2 \mathcal{B}_X(\bar{x}, x_0)} \approx 0.14$ does not give the right relation $\frac{\sigma}{\tau}$. In practice, of course, an optimal parameter choice is normally not known. However, Table 5.1 illustrates that also for any other (reasonable) choice of σ the version CP-BS 1 with $X = l^{1.25}(I_X)$ is preferable to CP 1 with $X = l^2(I_X)$. Here we chose $\tau \in (\sigma^{-1}\|T\|^{-2} - 2^{-6}, \sigma^{-1}\|T\|^{-2})$ for the Hilbert space case $X = l^2(I_X)$ and $\tau \in \left[\sigma^{-1}\|T\|^{-2}C_1, \sigma^{-1}\|T\|^{-2}C_2\right]$ with $C_1 = 0.89$ and $C_2 = 0.96 \in [0.25, 1]$ for the Banach space case $X = l^{1.25}(I_X)$ (cf. Remark 4.2.6).

By considering the generalized resolvent $(\tau\partial f + J_X)^{-1}$ of $f = \|\cdot\|_{l^1}$, Figure 5.2 presents an explanation of this acceleration: Rewriting (cf. Example 4.3.4 (*i*))

$$(\tau\partial f + J_X)^{-1} = J_{X^*} \circ (\tau\partial f + I)^{-1},$$

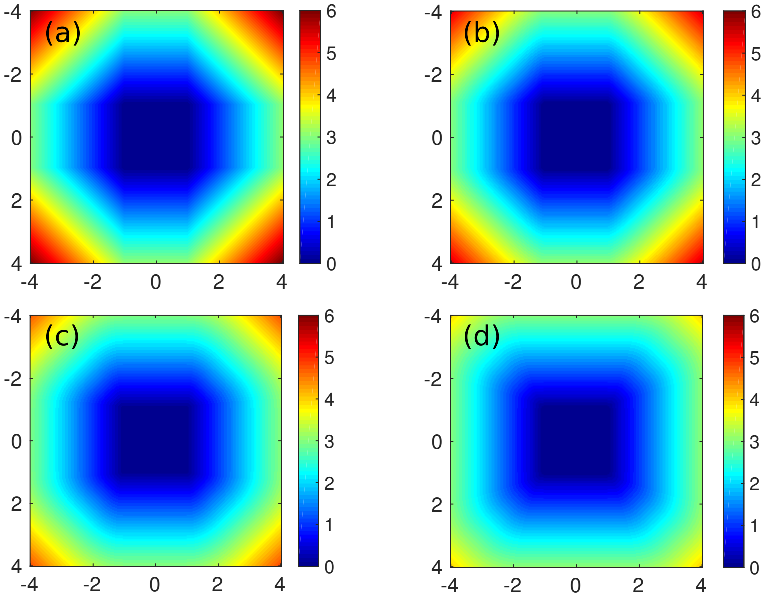


Figure 5.2: Influence of the different Banach spaces $X = l^r$, $r = 2, 1.75, 1.5, 1.25$ (Figure (a)-(d)), onto the generalized resolvent $(\tau\partial f + J_X)^{-1}$ of $f = \|\cdot\|_{l^1}$ and $\tau = 1$. For any point $x^* \in X^*$ in the rectangle $[-4, 4]^2$ the figures show the l^1-norm of the generalized resolvent at this point, i.e. $\|(\tau\partial f + J_X)^{-1}(x^*)\|_{l^1}$.

we see that this generalization of the resolvent $(\tau\partial f + I)^{-1} : l^2(I_X) \to l^2(I_X)$ leads to a further adaption of the resolvent's image to the corresponding Banach space X. So, the generalized resolvent on $X = l^r(I_X)$ with $r \approx 1$ implements the sparsity constraint better than the l^2-resolvent does. In general, the formulation (see Equation (3.24))

$$(\tau\partial f + J_X)^{-1}(u) = \operatorname*{argmin}_{x \in X} \tau f + \mathcal{B}_X(x, J_{X^*}(u)), \qquad u \in X^*,$$

implies that, by using a Bregman distance $\mathcal{B}_X$ that reflects the problem properties best, step (4.10) also gives the most likely iterate $x_{k+1} = (\tau\partial f + J_X)^{-1}(J_X(x_k) - \tau_k T^* p_{k+1})$. Likewise, we can expect that adjusting the image space Y improves the convergence rate of CP-BS .

Next, we want to reconstruct the piecewise differentiable function $\bar{x}$, shown in Figure 5.3 (a), from noisy data y^δ which are again given by the exact data $T\bar{x}$ (Figure 5.3 (b)) to which 5% normal distributed noise was been added. For this purpose Tikhonov-type regularization of the form

$$x_\alpha = \operatorname*{argmin}_{x \in X} \frac{1}{2}\|Tx - y^\delta\|_Y^2 + \frac{\alpha}{2}\|x\|_X^2, \tag{5.2}$$

with $X = l^{1.5}(I_X)$-penalty term, $\alpha = 1$, and $Y = l^2(I_X)$, seems to be an appropriate choice. Since $f(y) = \frac{1}{2}\|Tx - y^\delta\|_Y^2$ as well as $g(x) = \frac{\alpha}{2}\|x\|_X^2$ satisfy the Bregman midconvex

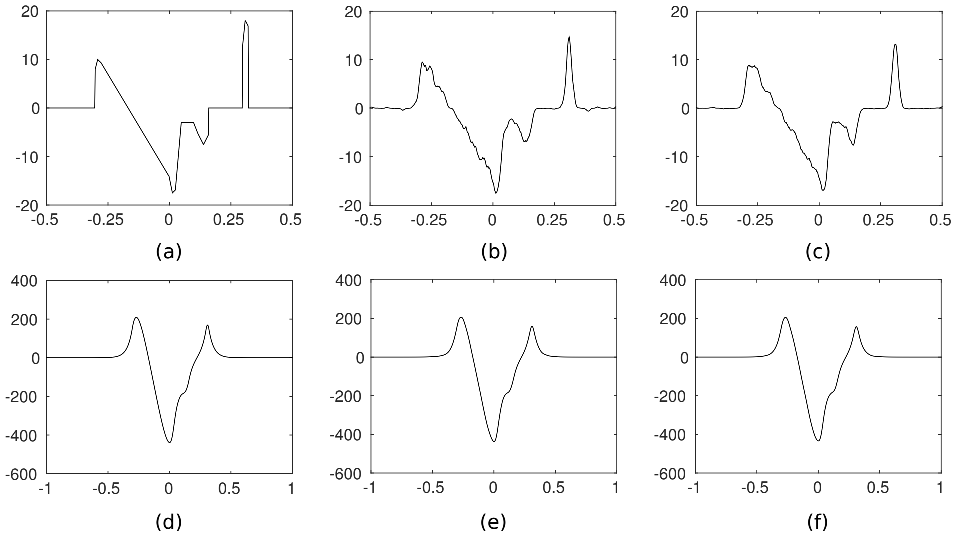


Figure 5.3: Convolution problem with exact solution given by (a) and exact data (d). For data perturbed by 5% normal distributed noise, (b) and (c) show the minimizer of the Tikhonov-type functionals 5.2 and 5.3, respectively. (e) and (f) are the reconstructed data corresponding to (b) and (c).

property (4.21), all introduced versions CP-BS 1 - 3 of our Algorithm 2 can be used for the computation of x_α. As above, we set $x_0 := 0$ and $p_0 := Tx_0 - y^\delta = -y^\delta$. Tuning the corresponding parameters in an optimal way, we obtain superimposable convergence curves for all three variants which almost coincide with the valid blue curve in Figure 5.4. Experimentally, we found the following parameter choice to be optimal:

- CP-BS 1: $\tau \approx 0.06$ and $\sigma = \frac{C}{\tau \|T\|^2} \approx 0.06$, where $C = 0.96$
- CP-BS 2: $\tau_0 \approx 0.06$ and $\sigma_0 = \frac{C}{\tau_0 \|T\|^2} \approx 0.06$, where $C = 0.96$, and $\gamma \approx 0.025$
- CP-BS 3: $\mu = \frac{C\sqrt{\gamma\delta}}{\|T\|}$, where $C = 0.98$, $\gamma = \delta \in (0, 1]$ and $\theta \in \left[\frac{1}{1+\mu}, 1\right]$

By "≈" we indicate that small deviations (in the range of ± 0.01) from τ, τ_0, (changing σ and σ_0 as well), and γ give almost the same result. So, in comparison to CP-BS 1 and CP-BS 2, here the use of variant CP-BS 3 does not accelerate the convergence as one might expect from the theory but provides for a whole set of parameters the optimal error decay. This last aspect of CP-BS 3 is quite obvious as the multiplication of γ and δ with the same constant does not change the values of τ and σ in CP-BS 3 and for a sufficiently large operator norm $\|T\|$ (≈ 16.32 in this example) also the influence on θ is rather small. Moreover, the parameters of CP-BS 1 and CP-BS 3 can be chosen identically. Letting $\delta \to 0$, the parameter choice rule of CP-BS 2 becomes identically to the one of CP-BS 1. Therefore, in the following we focus on CP-BS 2 with a reasonable large modulus γ or δ,

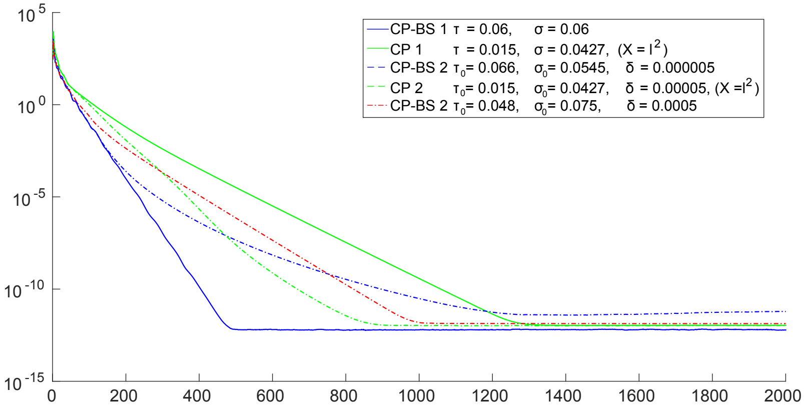


Figure 5.4: Convergence plot for the algorithms CP-BS 1 and CP-BS 2 (interchanging the roles of τ and δ) with $X = l^r(I_X)$, $r := 1.5$ and the algorithms CP 1 and CP 2 with $X = l^2(I_X)$ applied to the minimization Problem 5.3. The figure shows the error $\|x_k - x_\alpha\|_{l^r}$ between an element of the corresponding sequence $(x_k)_{k\in\mathbb{N}}$ and the true minimizer x_α of (3.1) per iteration step k. The parameters τ, σ, and δ (reasonable large) are optimally chosen.

respectively. In comparison to the first sparse example, where more than 15000 iterations were necessary in order to reach the solution, here we only need 558 iterations. This is probably due to the more regular penalty term.

On the other hand, also the Tikhonov-type regularization

$$x_\alpha = \operatorname*{argmin}_{x\in X} \frac{1}{2}\|Tx - y^\delta\|_Y^2 + \frac{\alpha}{r}\|x\|_{l^{1.5}(I_X)}^r, \tag{5.3}$$

with $r = 1.5$ and $\alpha = 22.5$ appears to be appropriate for this reconstruction task (cf. Figure 5.3). For $X = l^{1.5}(I_X)$ Equation (4.36) assures that the generalized resolvent of the corresponding penalty function $f(x) = \frac{\alpha}{r}\|x\|_{l^{1.5}(I_X)}^r$ is sufficiently simple as well. So, versions CP-BS 1 and CP-BS 2 (now for g^* satisfying the Bregman midconvex property with some modulus δ) can be applied for computing the minimizer x_α. Moreover, in this special, case also the resolvent $(\tau\,\partial f + I)^{-1}$ on the Hilbert space $X = l^2(I_X)$ can be efficiently evaluated: For any point $u \in X^* = l^2(I_X)$ the image $x := (\tau\,\partial f + I)^{-1}(u)$ solves

$$u - x = \tau\, J_{r,l^r(I_X)}(x) = \tau\,|x|^{\frac{1}{2}}\operatorname{sign}(x) = \tau\,|x|^{\frac{1}{2}}\operatorname{sign}(u) \quad \Leftrightarrow \quad 0 = |x|^2 - \tau^2\,|x| - 2|u||x| + |u|^2,$$

and hence for any index $i \in I_X$ there are only 2 candidates for x_i:

$$x_i = \operatorname{sign}(x_i)\,|x_i| \in \left\{\operatorname{sign}(u_i)\,\left(2\,|u_i| + \tau^2 \pm \tau\,\sqrt{4|u_i| + \tau^2}\right)/2\right\}.$$

This gives us another opportunity to compare the original versions CP 1 and CP 2 with an appropriate Hilbert space setting $X = l^2(I_X)$ and $Y = l^2(I_Y)$ to their generalizations CP-BS 1 and CP-BS 2 combined with the more natural choice $X = l^{1.5}(I_X)$. Figure 5.4

shows that CP-BS 1 with an optimal parameter choice (solid blue curve) converges clearly faster than all the other versions (under the condition to choose δ too small). Although the improvement by the generalization is not that pronounced in the case of CP-BS 2, which requires even more iterations than CP 2 to converge, this example again demonstrates the positive effect of a problem adapted choice of the space X onto the algorithm's performance.

5.3 Phase retrieval problem in wavefront reconstructions

Now, let us consider phase retrieval problems in coherent x-ray imaging, as described in Section 1.1, which were our main motivation for generalizing the Chambolle-Pock algorithm to Banach spaces. For this purpose we need numerical approximations of the forward operators T_{Fresnel} and T_{Frau} which are given by the Fresnel and Fraunhofer approximations, respectively (cf. Equations (1.21), (1.23), and (1.16)). Then the discretization of the Fréchet derivatives $T'_{\text{Fresnel}}[\phi], T'_{\text{Frau}}[\phi]$ and their adjoints follows analogously. In particular the L^2-adjoint of $T'_{\text{Fresnel}}[\phi]$ given in Section 1.1.4 reads

$$(T'_{\text{Fresnel}}[\phi])^*(g)\,(\mathbf{x}') = \frac{2\,\kappa^2}{\Gamma^2}\,\Re\left(\overline{\chi_{\frac{\kappa M}{\Gamma}}(\mathbf{x}')\,\mathrm{i}\,\kappa\,\mathrm{e}^{\mathrm{i}\,\kappa\,\phi(\mathbf{x}')}}\;\mathcal{F}^{-1}\left(\mathcal{F}\left(\chi_{\frac{\kappa M}{\Gamma}}O(\phi)\right)\left(\frac{\kappa}{\Gamma}\cdot\right)\,g\right)(\mathbf{x}')\right),$$

for all $g \in X^*$, $\mathbf{x}' \in \mathbb{R}^2$.

5.3.1 Numerical approximation of the Fresnel approximation

The numerical implementation of the Fresnel and Fraunhofer approximation basically relies on the discrete Fourier transform (A.11), providing a good approximation of its continuous counterpart, as studied in Appendix A.2. Making use of Assumption 1.1.9 (as well as the projection approximation) that implies

$$\operatorname{supp} u_0 \subset [-\mathbf{r}_X, \mathbf{r}_X] := \{\mathbf{x}' = (x_1, x_2) \in \mathbb{R}^2 \mid |x_1| \le r_{X,1}, |x_2| \le r_{X,2}\},$$

we again discretize the corresponding rectangle $[-\mathbf{r}_X, \mathbf{r}_X] \subset \mathbb{P}_0$ by the grid

$$\Delta(\mathbf{r}_X) := \left\{\mathbf{r}_X \bullet \frac{2\mathbf{j}}{\mathbf{N}} := \left(r_{X,1}\frac{2\,j_1}{N_1}, r_{X,2}\frac{2\,j_2}{N_2}\right) \,\middle|\, \mathbf{j} \in \Delta_{\mathbf{N}}\right\},$$

defined by Equation (4.38) with sample points $\mathbf{N} := (N_1, N_2) \in \mathbb{N}^2$. In order to use the well-established fast implementations of $\mathcal{F}_{\mathbf{N}}$ and $\mathcal{F}_{\mathbf{N}}^{-1}$, which have a complexity of $O(N_1\,N_2 \log_2 N_1\,N_2)$, it is favorable to choose N_j, $j = 1, 2$, as a power of two. Then, introducing the notation $\underline{u_0}_{\mathbf{r}_X,\mathbf{j}} = u_0\left(\frac{2\mathbf{r}_X \bullet \mathbf{j}}{\mathbf{N}}\right)$, $\mathbf{j} \in \Delta_{\mathbf{N}}$, for the sampled function u_0 with $\operatorname{supp} u_0 \subseteq [-\mathbf{r}_X, \mathbf{r}_X]$, the discrete Fourier transform $\frac{2 r_{X,1}\, r_{X,2}}{\pi\,\sqrt{N_1, N2}}\mathcal{F}_{\mathbf{N}}\underline{u_0}_{\mathbf{r}_X}$ approximates $\mathcal{F}\,u_0$ on the grid

$$\Delta(\mathbf{r}_\xi) = \left\{-r_{\xi,1}, -r_{\xi,1} + \frac{2\,r_{\xi,1}}{N_1}, \dots, r_{\xi,1} - \frac{2\,r_{\xi,1}}{N_1}\right\} \times \left\{-r_{\xi,2}, r_{\xi,2} + \frac{2\,r_{\xi,2}}{N_2}, \dots, r_{\xi,2} - \frac{2\,r_{\xi,2}}{N_2}\right\},$$

with $\mathbf{r}_\xi := \frac{\pi}{2}\left(\frac{N_1}{r_{X,1}}, \frac{N_2}{r_{X,2}}\right)$ by Equation (A.13): $\frac{2r_{X,1}\, r_{X,2}}{\pi\,\sqrt{N_1 N_2}}\left(\mathcal{F}_{\mathbf{N}}\underline{u_0}_{\mathbf{r}_X}\right)_{\mathbf{j}} \approx \underline{\mathcal{F}u_0}_{\mathbf{r}_\xi,\mathbf{j}}$. Accordingly, for the Fourier transform $\widehat{u_0} := \mathcal{F}u_0$, we obtain from (A.14)

$$\frac{2r_{\xi,1}\, r_{\xi,2}}{\pi\,\sqrt{N_1 N_2}}\left(\mathcal{F}_{\mathbf{N}}^{-1}\underline{\widehat{u_0}}_{\mathbf{r}_\xi}\right)_{\mathbf{j}} \approx \underline{\mathcal{F}^{-1}\widehat{u_0}}_{\mathbf{r}_X,\mathbf{j}}.$$

Note from Appendix A.2 that $\Delta(\mathbf{r}_\xi)$ is sampled with the Nyquist sampling rate of $\frac{\pi}{\mathbf{r}_X}$. Since $\mathcal{F}u_0(\xi) \approx 0$ at any point $\xi \in \mathbb{R}^2\backslash[-\mathbf{r}_\xi, \mathbf{r}_\xi]$, we can assume u_0 to have a bandwidth smaller or equal to $\mathbf{r}_\xi$. Then the discrete far field representation of the Fresnel approximation is given by:

$$\begin{aligned}\mathcal{D}_\Gamma u_0\left(\frac{\pi\,\Gamma\,\mathbf{j}}{\mathbf{r}_X}\right) &= \mathcal{D}_\Gamma u_0\left(\frac{\Gamma}{\kappa}\frac{2\mathbf{r}_\xi \bullet \mathbf{j}}{\mathbf{N}}\right)\\ &\approx \frac{-\mathrm{i}\,\kappa\,2\,r_{x,1}r_{x,2}}{\pi\,\Gamma\,\sqrt{N_1 N_2}}\mathrm{e}^{\mathrm{i}\kappa\Gamma}\,\chi_{\frac{\kappa}{\Gamma}}\left(\frac{\Gamma}{\kappa}\frac{2\mathbf{r}_\xi \bullet \mathbf{j}}{\mathbf{N}}\right)\left(\mathcal{F}_{\mathbf{N}}\left(\underline{\chi}_{\frac{\kappa}{\Gamma},\mathbf{r}_X} \bullet \underline{u_0}_{\mathbf{r}_X}\right)\right)_{\mathbf{j}}, \qquad \mathbf{j} \in \Delta_{\mathbf{N}},\end{aligned} \tag{5.4}$$

while the discrete near field representation reads as:

$$\mathcal{D}_\Gamma u_0\left(\frac{2\mathbf{r}_X \bullet \mathbf{j}}{\mathbf{N}}\right) \approx \mathrm{e}^{\mathrm{i}\,\kappa\,\Gamma}\left(\mathcal{F}_{\mathbf{N}}^{-1}\left(\underline{\chi}_{-\frac{\Gamma}{\kappa},r_\xi} \bullet \mathcal{F}_{\mathbf{N}}\underline{u_0}_{\mathbf{r}_X}\right)\right)_{\mathbf{j}}, \qquad \mathbf{j} \in \Delta_{\mathbf{N}}\,.$$

Here we set

$$\underline{\chi}_{\frac{\kappa}{\Gamma},\mathbf{r}_X,\mathbf{j}} = \mathrm{e}^{\frac{\mathrm{i}}{2}\left(\mathfrak{f}_1\left|\frac{2\,j_1}{N_1}\right|^2 + \mathfrak{f}_2\left|\frac{2\,j_2}{N_2}\right|^2\right)} \quad \text{and} \quad \underline{\chi}_{-\frac{\Gamma}{\kappa},r_\xi,\mathbf{j}} = \mathrm{e}^{\frac{-\mathrm{i}}{2}\left(\mathfrak{f}_1^{-1}\frac{\pi^2\,N_1^2}{4}\left|\frac{2\,j_1}{N_1}\right|^2 + \mathfrak{f}_2^{-1}\frac{\pi^2\,N_2^2}{4}\left|\frac{2\,j_2}{N_2}\right|^2\right)},$$

which explicates the influence of the Fresnel number $\mathfrak{f}$ to the chirp functions oscillations. On the other hand, we discretize the near field formula (1.23) given the effective geometry by:

$$|\mathcal{D}_\Gamma u_0|^2\left(\frac{M\,2\,\mathbf{r}_X \bullet \mathbf{j}}{\mathbf{N}}\right) \approx \frac{1}{M^2}\left|\mathcal{F}_{\mathbf{N}}^{-1}\left(\underline{\chi}_{-\frac{\Gamma}{M\,\kappa},r_\xi} \bullet \mathcal{F}_{\mathbf{N}}\underline{u_0}_{\mathbf{r}_X}\right)\right|_{\mathbf{j}}^2, \qquad \mathbf{j} \in \Delta_{\mathbf{N}}\,. \tag{5.5}$$

The numerical approximation of the Fraunhofer diffraction formula (1.15) follows directly from (5.4) by neglecting the chirp function $\underline{\chi}_{\frac{\kappa}{\Gamma},\mathbf{r}_X}$.

5.3.2 CP-BS as inner solver in the IRNM

Recall that our aim consists in applying the IRNM

$$\bar{\phi}_{n+1} = \underset{\phi \in X}{\operatorname{argmin}}\; kl\left(y^\delta;\; T(\bar{\phi}_n) + T'[\bar{\phi}_n](\phi - \bar{\phi}_n)\right) + \frac{\alpha_n}{2}\|\phi\|_X^2 \tag{5.6}$$

not only with the noise adapted data misfit term kl given by Equation (1.26) but also on an appropriate Banach space X corresponding to the penalty term $\phi \mapsto R(\phi) = \frac{1}{2}\|\phi\|_X^2$. Here, we simplified the notation by writing ϕ instead of $\underline{\phi}_{\mathbf{r}_X}$ for the discretized phase information given on the grid $\triangle(\mathbf{r}_X)$. Moreover, we introduced an overbar in order to distinguish the iterates $\bar{\phi}_{n+1}$ of the outer IRNM-solver from those, denoted as ϕ_k, of the inner CP-BS -solvers (given by the primal variable). As stressed in Section 4.3, it is due to the generalization to Banach spaces that the algorithm CP becomes efficiently applicable for this Banach norm penalty R with $X \in \left\{ l^r(\triangle(\mathbf{r}_X)), h^{1,r}(\triangle(\mathbf{r}_X)) \,|r \in (1,2) \right\}$. To be more precise, the generalized resolvent of R is up to a constant factor the duality mapping (cf. Equation (4.33)) which ensures a closed form in these cases but also implies greater computational cost for the application of CP-BS in case of $X = h^{1,r}(\triangle(\mathbf{r}_X))$ than for $X = l^r(\triangle(\mathbf{r}_X))$. Example 4.3.4 together with Equation (4.31) gives the generalized resolvent of $y \mapsto kl\left(y^\delta;\ T(y + \bar{\phi}_n) - T'[\bar{\phi}_n](\bar{\phi}_n)\right)$ on the weighted Hilbert space $Y_n = l^2_{W_n}(\triangle_{\mathbf{N}})$ (treated as a Banach space if $W_n \neq \mathbf{1}$) with $W_n \in \left\{ (T(\bar{\phi}_n) + \epsilon)^{-1}, \mathbf{1} \right\}$ (cf. Section 1.1.4). As initial guesses $\phi_0 = x_0$ and p_0 for the inner solver CP-BS we take the previous iterates $\phi_0 = \bar{\phi}_n$ and $p_0 = \bar{p}_n$, starting with the second iteration $n + 1 = 2$.

Figure 5.5 illustrates the influence of the choice of

$$X \in \left\{ l^2(\triangle(\mathbf{r}_X)), l^{1.5}(\triangle(\mathbf{r}_X)), h^{1,1.25}(\triangle(\mathbf{r}_X)) \right\}$$

with the associated penalty term to the reconstruction in a far field regime. Here, the Fresnel number $\mathfrak{f}$ is set to zero, and hence the forward operator $T = T_{\mathrm{Frau}} : X \to Y := l^2(\triangle_{\mathbf{N}})$ is based on the Fraunhofer approximation. Moreover, in order to obtain intensity data in a realistic range, T_{Frau} is multiplied by the factor such that total intensity $\|y^\delta\|_{l^1}$ is 10^6. For the reconstruction, we also multiply the regularization parameters α_k, $k \in \mathbb{N}$ by 10^6, which corresponds to a normalization of the data. As exact solution $\bar{\phi} \in X$, we take an example from [34]: The piecewise constant phase $\bar{\phi}$ (Figure 5.5 (a)) of the simulated object function $O(\bar{\phi}) = \mathrm{e}^{\mathrm{i}\bar{\phi}}$ belongs to two unstained biological cells. In practice, the region $\overline{\Omega} = [-\mathbf{r}_X, \mathbf{r}_X]$ including the support of the sample can be marked by putting a further constant phase shifting object into the beam. This is also simulated here by adding the characteristic function χ_Ω of a 174×188 pixel rectangle $\overline{\Omega}$ to the true phase $\bar{\phi}$ (256 × 256 pixels). In order to incorporate this additional information into the IRNM, we redefine the penalty term as $R(\phi) = \frac{1}{2}\|\phi - \mathbf{1}\|^2_X$, or equivalently consider the space $X + \mathbf{1}$, given on $\triangle(\mathbf{r}_X)$, instead of $X \in \left\{ l^2(\triangle(\mathbf{r}_X)), l^{1.5}(\triangle(\mathbf{r}_X)), h^{1,1.25}(\triangle(\mathbf{r}_X)) \right\}$ (cf. Corollary 4.3.5). Accordingly, in the first iteration, $\bar{\phi}_0 = \mathbf{1}$ serves as the initial guess for the primal variable x of CP-BS which also defines p_0 via Equation (5.1). For experimental chosen regularization parameters α_k, the IRNM was stopped when the $l^{1.5}$-error $\varrho(n + 1) := \|\phi_{n+1} - \bar{\phi}\|_{l^{1.5}}$ between the reconstruction ϕ_n and the true solution $\bar{\phi}$ reached a minimum. In Figure 5.5, this error $\varrho(N)$ for the selected reconstruction ϕ_N ((d)-(f)) is $\varrho(N) \approx 142$ in (d), $\varrho(N) \approx 119$ in (e), and $\varrho(N) \approx 117$ in (f). We also see that choosing an $l^{1.5}$- space and penalty term ((e), (h)) gives a slightly better reconstruction than in the l^2-setting ((d), (g)). A bit more obvious is the improvement in (f), where a Sobolev norm space and penalty is used. However, in this ill-posed setting, it remains questionable whether the rather small benefit of choosing $X = h^{1,1.25}(\triangle(\mathbf{r}_X))/\{\mathbf{1}\}$ justifies the larger effort of evaluating the corresponding generalized resolvent of R. A compromise would be to start the IRNM with $X = l^{1.5}(\triangle(\mathbf{r}_X))/\{\mathbf{1}\}$ and to set $X = h^{1,1.25}(\triangle(\mathbf{r}_X))/\{\mathbf{1}\}$ only for the last iterations, with a possible adaption of the regularization parameter choice rule.

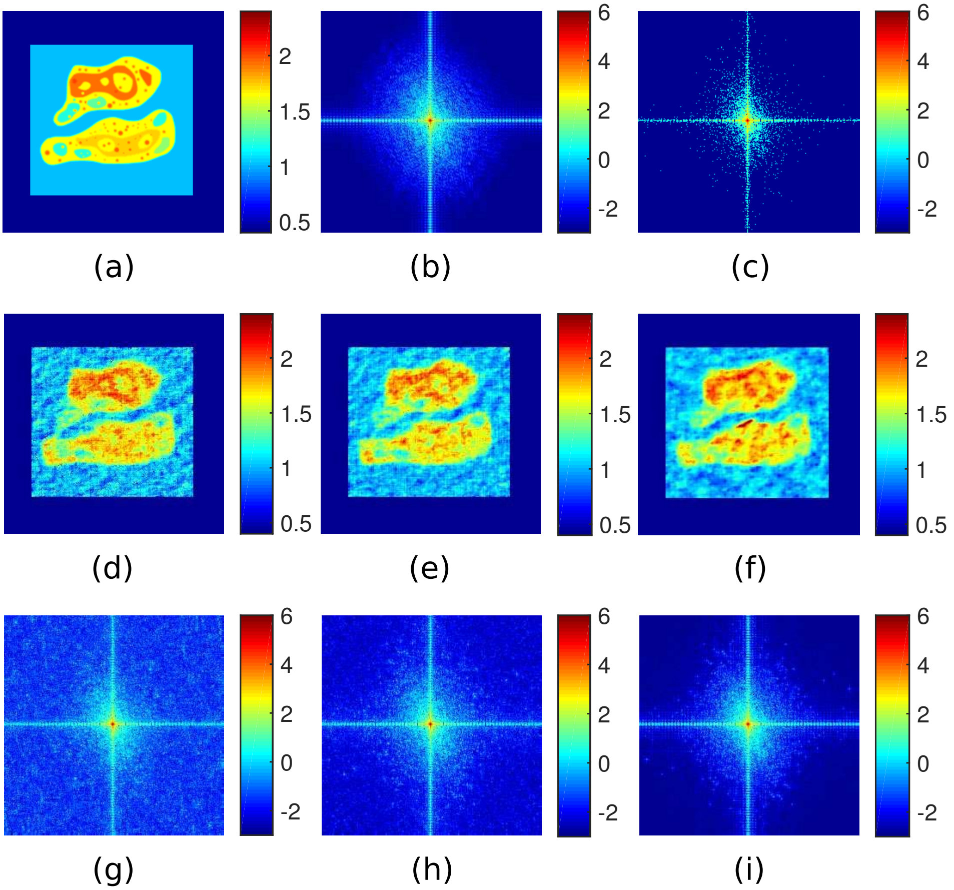
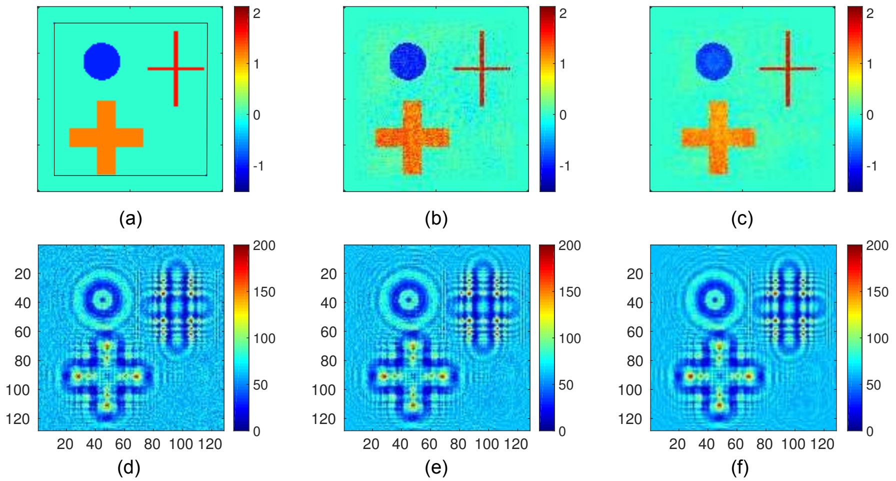


Figure 5.5: Reconstruction of a pure phase shifting cell test pattern (a) (taken from [34]) from far field data (c), where $M\mathfrak{f} = 0$, for different choices of X. In order to stress the region Ω, its indicator function χ_Ω is added to the phase. We have: (a) sum of true phase and χ_Ω , (b) log_{10} of the exact data, and (c) log_{10} of the given noisy data = (b) + Poisson noise. The reconstructions are performed by the IRNM (1.29) where the support information is incorporated in the Banach space X by considering the space $U + \mathbf{1}$ of an appropriate Banach space U given on Ω (cf. Corollary 4.3.5). Under the following setting, the reconstructed phase maps are given in the middle row with corresponding data below : (d), (g) reconstruction for $U = l^2(\triangle(\mathbf{r}_X))$ with $\alpha_n = 5^{-5}\,(2/3)^{n-1}$, after 9 iterations. (e), (h) reconstruction for $U = l^{1.5}(\triangle(\mathbf{r}_X))$ with $\alpha_n = 10^{-6}\,(2/3)^{n-1}$ after 8 iterations. (f), (i) reconstruction for $U = h^{1,1.25}(\triangle(\mathbf{r}_X))$ (discretized) with $\alpha_n = 5^{-7}\,(1/2)^{n-1}$ after 15 iterations. The inner minimization problem of the IRNM is solved by the algorithm CP-BS 1 given by Theorem 4.2.1 with $\tau = 0.01, 0.05, 0.4$ (from left to right) and $\sigma = \|T'(\phi)\|^{-2}\tau^{-1}0.96$.



(a) (b) (c)

(d) (e) (f)

Figure 5.6: Reconstructions $\bar{\phi}_N$ by the IRNM where ϱ attains its minimum (in both cases): (a) True phase, where the frame shows the given support $\bar{\Omega} = [-\mathbf{r}_X, \mathbf{r}_X]$ (b) reconstruction ϕ_{15} for $X = l^{1.5}(\triangle(\mathbf{r}_X))$ and $\alpha_n = 20\,2^{-n+1}$, (c) reconstruction ϕ_{37} for $X = h^{1,1.1}\,(\triangle(\mathbf{r}_X))$ and $\alpha_n = 0.25\,2^{-n+1}$, (d)-(e) corresponding data. As initial guess for the first iteration we used $\bar{\phi}_0 = 0$ as well as Equation (5.1) defining p_0. The inner minimization problems were solved by the CP-BS 1 with $\tau = 0.2$ in (b),(e) and $\tau = 24$ in (c),(f) choosing $\sigma = 0.96\,\|T'[\bar{\phi}_n]\|_X^{-2}\tau^{-1}$ for the $n+1$-th iteration step. Surprisingly, in (c), the IRNM seems to convergence, i.e. the reconstruction ϕ_N and also the error $\varrho(N)$ barely changes after 20 iterations.

As a second example, we consider a less ill-posed phase retrieval problem in a near field regime. (a) and (d) of Figure 5.6 show the true solution $\bar{\phi}$ and the given data y^δ. Accordingly, the operator $T = T_{\mathrm{Fresnel}}$ is formulated in the parallel beam geometry, i.e. we use the representation (1.23) discretized by Equation (5.5). With $\mathbf{N} = (N_1, N_2) = (128, 128)$ pixels in the detector and the object plane, the geometrical magnification M is $2\,N_1$ and the Fresnel number $\mathfrak{f}$ is $\mathfrak{f}_1 := \mathfrak{f}_2 = \frac{2\,\pi\, r_{X,2}^2}{\lambda \Gamma} = 2\,\pi > 1$, yielding rather strong oscillations of $\underline{\chi}_{-\frac{\Gamma}{M\kappa},\mathbf{r}_\xi}$. The given support $[-\mathbf{r}_X, \mathbf{r}_X]$ is 106 × 106 pixels. Applying the IRNM (5.6) with $X \in \left\{l^{1.5}(\triangle(\mathbf{r}_X)), h^{1,1.1}\,(\triangle(\mathbf{r}_X))\right\}$, we obtain an almost perfect reconstruction (Figure 5.6 (c)) in case of the Sobolev preimage space and penalty term, while in the $l^{1.5}$-case some aberrations occur. Here, again we selected the approximation $\bar{\phi}_N$ with (alomost) the smallest error ϱ which corresponds to an optimal stopping rule. For $\epsilon = 0.1$, the image space Y_n of $T_{\mathrm{Fresnel}}[\bar{\phi}_n]$ is defined as $Y_n = l^2_{W_n}(\triangle_{\mathbf{N}})$. Figure 5.7 (cyan curve) illustrates that, compared to $Y_n = l^2(\triangle_{\mathbf{N}})$, this more adapted choice accelerates the convergence of the inner solver CP-BS . This observation is in accordance with the results in [62] where the special version of CP-BS 1 for Hilbert spaces, preconditioned by a matrix, was introduced (cf. Remark 4.2.2). However, in the far field regime of the first example, weighting $l^2(\triangle_{\mathbf{N}})$ by $W_n = (T(\bar{\phi}_n) + \epsilon)^{-1}$ or $(\max(T(\bar{\phi}_n), \epsilon))^{-1}$ for some $\epsilon > 0$ leads to instabilities.

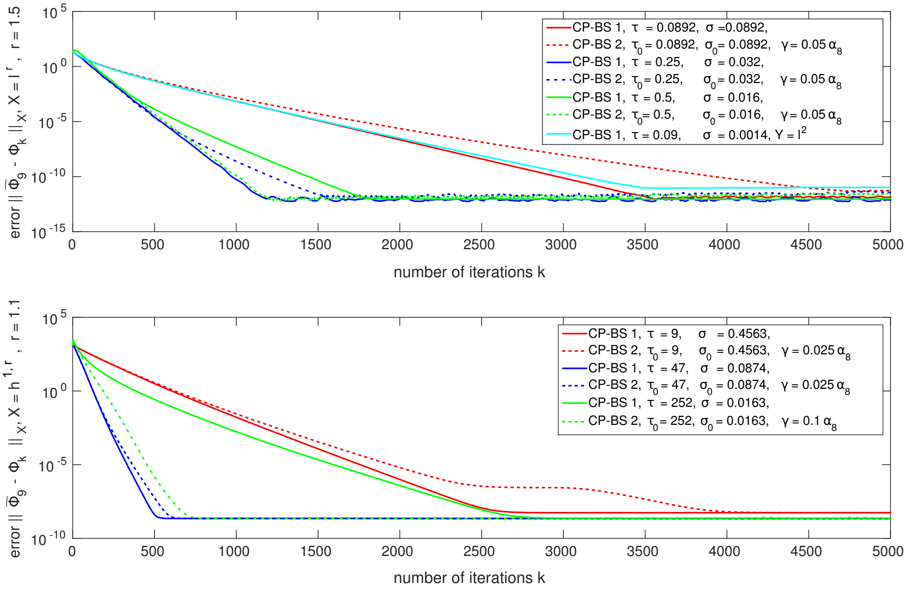
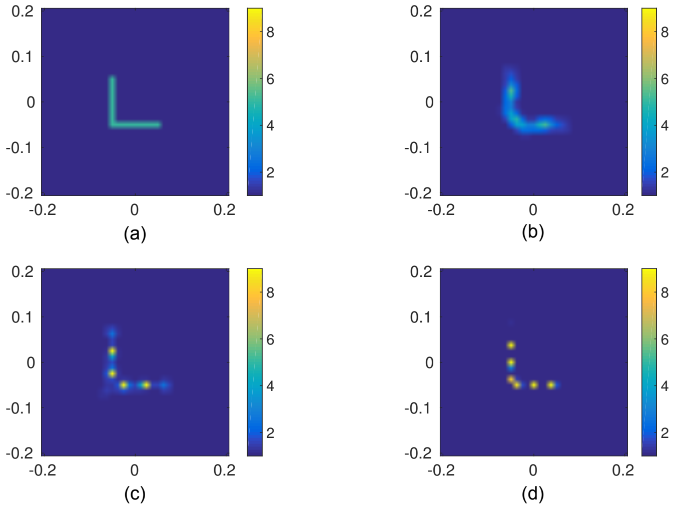


Figure 5.7: IRNM (5.6) applied to the phase retrieval problem as shown in Figure 5.6 for $X = l^{1.5}(\triangle_{\mathbf{N}})$ (top) and $X = h^{1,1.1}(\triangle_{\mathbf{N}})$ (bottom). For fixed $n = 8$, the figures show the error $\|\bar{\phi}_9 - \phi_k\|_X$ of an element of the sequence $(\phi_k)_{k\in\mathbb{N}}$ which we obtain from the algorithms CP-BS 1 and CP-BS 2 and the true minimizer $\bar{\phi}_8$ of (5.6) per iteration step k. For a given τ (or τ_0), we set $\sigma = 0.96\,\|T'[\bar{\phi}_n]\|^{-2}\tau^{-1}$ (σ_0 analogously). According to the theory, we observed that choosing the value $\frac{\gamma}{\alpha_8}$ “sufficiently” small and $\tau_0 = \tau$, $\sigma_0 = \sigma$ both versions CP-BS 1 and CP-BS 2 give the same error curves. The solid blue curves (and also the cyan curve where $Y = l^2(\triangle_{\mathbf{N}})$) present the optimal parameter choices for CP-BS 1, which coincides with best error decay we obtained.

In the convergence plots in Figure 5.7 which are given for the $n + 1 = 9$-th iteration step of the IRNM for both choices of X, we also compare the versions CP-BS 1 and CP-BS 2. In both cases (top and bottom), we have again found no parameter choice for CP-BS 2 such that the resulting error $\|\bar{\phi}_N - \phi_k\|_X$ decays faster than the solid blue curves, representing the optimal choices of σ and τ for CP-BS 1. We also observed that, if we further decrease $\tau = \tau_0$ and hence increase $\sigma = \sigma_0$, CP-BS 1 achieves still a better result than CP-BS 2 does with a “reasonably large modulus” γ (cf. red curves). If we multiply the modulus γ that corresponds to the dashed blue curve in the case $X = l^{1.5}(\triangle_{\mathbf{N}})$ by a factor $\frac{1}{2}$ such that $\gamma = 0.025\alpha_8$, we again obtain the optimal error decay. This illustrates our impression that CP-BS 2 reacts quite sensitively to changes of γ. On the other hand, for $\tau = \tau_0$ sufficiently large (see the green curves) and suitably chosen γ, CP-BS 2 convergences much faster than CP-BS 1 does. So, for the considered problem the version CP-BS 1 seems to be more favorable if $\tau = \tau_0$ is not much lager than $100\,\sigma$ in the $l^{1.5}$-case and $\tau \leq 2000\,\sigma$ in the Sobolev-case. If we pick $\tau = \tau_0 \gg \sigma$, the method CP-BS 2 with some γ not too small becomes the better choice. This is in accordance



(a) (b) (c) (d)

Figure 5.8: IRNM applied to the medium scattering problem: (a) Refractive index n^2 given by the true contrast $\bar{a}$ via $n^2 = 1 + \bar{a}$. (b) Reconstruction results for $X = l^{1.5}$, $R(a) = \|a\|_X^2$, $\alpha_n = 8\,2^{-n+1}$, and inner solver CP-BS 3 with $\delta = \gamma = 0.5$ and $\mu = 0.98\frac{\sqrt{\delta\gamma}}{\|T'[a_n]\|}$. (c) Reconstruction results for $X = l^{1.25}$, $R(a) = \|a\|_X^2$, and $\alpha_n = 10\,2^{-n+1}$. Up to iteration $n = 11$ we chose same μ as in (b) for inner solver CP-BS 3 with $\delta = \gamma = 0.5$. For $n = 12, 13$ we took $\mu = 0.6\frac{\sqrt{\delta\gamma}}{\|T'[a_n]\|}$. (d) Reconstruction results for $X = l^{1.5}$, $R(a) = \|a\|_{l^1}$, $\alpha_0 = 10\,\frac{2}{3}^{n-1}$, and inner solver CP-BS 1 with $\tau = 8$. The discrepancy principal (5.7) (with $\nu = 2$ chose the reconstruction a_N with index $N = 12$ in (b), $N = 14$ in (c), and $N = 15$ in (d).

with the strategy suggested in [21], which requires a good approximation of the distance $\mathcal{B}_X(\bar{x}, x_0)$. Translating the advice in our Banach space setting, it states that, in order keep the right hand side of Equation (4.22) small, one should choose $\gamma\tau_0 \gg \mathcal{B}_X(\bar{x}, x_0)$. In Figure 5.7 we have: $\mathcal{B}_X(\bar{\phi}_9, \phi_0) = \mathcal{B}_X(\bar{\phi}_9, \bar{\phi}_8) \approx 242$ for (b), (e) and $\mathcal{B}_X(\bar{\phi}_9, \phi_0) \approx 8\,10^5$ for (c), (f).

5.4 Phase retrieval in inverse medium scattering

We conclude our numerical investigation of the algorithm CP-BS by a nonlinear inverse medium scattering problem with 'sparse' contrast, as introduced in Section 1.2. The two-dimensional example shown in Figure 5.8 basically adapts the setting of Figure 1 in [52], while there the given data consist of point measurements of u^s on some surface: The true real-valued contrast $\bar{a}$, with $\bar{a}(\mathbf{x}') = 4$ for $\mathbf{x}' \in \operatorname{supp}(\bar{a})$, (shown in (a)) has a sparse support which is loacted in a ball B_ρ of radius $\rho = 0.2$. For $j = 1, \ldots 16$ equidistantly sampled directions $d_j = \exp{(2\pi\,\mathrm{i}\,j/16)} \in \mathbb{S}^1$ and plane waves $u^i_j(\mathbf{x}') = \exp(-\mathrm{i}\,\kappa\,\mathbf{x}' \cdot d_j)$ we calculate the far field patterns $u_j^\infty(\vartheta_k)$ at $\vartheta_k := \exp\left(2\pi\,\mathrm{i}\,\frac{k}{32}\right)$, $k = 1, \ldots, 32$. For the discretization and efficient evaluation of the corresponding forward operators F_{d_j}, $j = 1, \ldots 16$, the Fréchet derivatives, and their adjoints we refer to [41] (while do not use the two-grid method). Note that this implementation relies on the fast solution method for the Lippmann-Schwinger equation suggested by Vainikko ([79]). The ball $B_{2\rho}$ on which the (4ρ-periodizations of the) functions Φ, $\bar{a}, a$, $a\,u$ and $a\,u^i$ are considered is sampled by the grid $\triangle(2\rho)$ with $\mathbf{N} = (64, 64)$ grid points. The given data $y^\delta = \left(u^{\infty,\delta}_{d_j}(\vartheta_k)\right)_{k,j}$ is the vector of far field patterns $u_j^\infty(\vartheta_k)$, $k = 1, \dots, 32$, $j = 1, \dots 16$ to which 8% normally distributed noise has been added. Thus, we apply the IRNM (1.35) with l^2-data misfit functional $S(y; y^\delta) = \frac{1}{2}\|y - y^\delta\|^2_{l^2}$ and corresponding image space $Y = l^2$. In order to model the sparseness of $\operatorname{supp} a$, the penalty term R is chosen to be either $R = \|\cdot\|_{l^1}$ (cf. the sparse convolution problem 3.1) or $R = \frac{1}{2}\|\cdot\|^2_X$ given on a Banach space $X = l^r$ with $r = 1.25, 1.5$. We force the IRNM to generate a real-valued result by setting the imaginary part of the reconstruction a_{n+1} to zero after any iteration step n. As stopping criterion we use the discrepancy principal, i.e. we stop the iteration for the first index $n = N$ where

$$S(y^\delta; (T_{u^i_j}(a_{n+1}))_{j=1,\dots,J}) \leq \nu\, 0.08, \tag{5.7}$$

with $\nu = 2 > 1$ and take the corresponding iterate a_N, $N = n + 1$ as the result. In iteration step $n = 12$ of the IRNM corresponding to Figure 5.8 (c), the inner minimization problem defines an example where CP-BS 3 does not converge for the relaxed parameter choice $\mu = 0.98\frac{\sqrt{\delta\gamma}}{\|T'[a_n]\|}$ (cf. Remark 4.2.6). This might be also due to a wrong operator norm $\|T'[a_n]\|$ that is given by the power method of Boyd, since here it is not clear whether the matrix corresponding to $T'[a_n]$ is nonnegative. However, if we replace $\min\{C_{l^{1.25}}, C_{l^2}\} = 0.25$ by 0.6 in the parameter choice rule defined by Theorem 4.2.5 the resulting sequence $x_{k\in\mathbb{N}}$ convergences. Figure 5.8 shows that the l^1-penalization gives a slightly too sparse reconstruction (see (c)), while the choice $R = \frac{1}{2}\|\cdot\|^2_X$ with $X = l^{1.5}$ promotes a too expanded support. So, $R = \frac{1}{2}\|\cdot\|^2_X$ with $X = l^{1.25}$ turns out to be favorable. Here we see the advantage of CP-BS 's generality which easily allows an adaption of the constraints (defined by R and X with respect to the solution properties and by S and Y with respect to the data properties).

Summary

Within this work, we studied two different issues with regard to phase retrieval problems in coherent x-ray imaging: The validity of the *empty beam correction* as well as an efficient, *problem adapted reconstruction method.*

Our main result with respect to the empty beam correction is a parameter-dependent error estimate (see Chapter 1). Moreover, we translated this result to the effective geometry which is commonly used in near field imaging. The presented analysis not only explains this data correction step in mathematical terms, and thus provides a deeper understanding, but also allows us to formulate conditions on a setup that justifies this *product approximation in the detector plane*. In particular, our estimate shows that the error in the empty beam correction is mainly influenced by two parameters: The source size (or equivalently the wavefront smoothness) and the characteristic length scale in the sample. This prediction was also verified by numerical simulations. Although usually not all quantities of the proposed error bound are precisely known in practice, our analysis can be a helpful tool to design and evaluate physical experiments. On the other hand, it underlies the need for efficient reconstruction methods not only for the object function but also for the illumination function in order to achieve high resolution in case of an extended source.

A further main result of this thesis is the generalization (denoted as CP-BS) of the Chambolle-Pock algorithm CP from a Hilbert space to a Banach space setting. For this purpose, we translated some concepts defined on Hilbert spaces to the considered Banach space setting (cf. Section 3.2 and Chapter 4). In particular, the proposed algorithm CP-BS includes a generalization of the resolvents as well as of a midconvex property. Moreover, under certain conditions we showed strong convergence of the algorithm CP-BS , and under additional regularity assumptions we also proved rates of convergence that are similar to those of CP . This means, the version CP-BS 2 converges in $O\left(\frac{1}{k}\right)$ and CP-BS 3 converges even linearly.

Then the generalized algorithm CP-BS allowed us the efficient application of a really problem adapted method to phase retrieval problems in coherent x-ray imaging: The IRNM with Kullback-Leibler data misfit functional $\mathbb{KL}$ and non-Hilbertian norm penalty term (such as a L^r- or $H^{1,r}$-penalty term with $r \in (1, 2)$) given on Banach spaces. Such a combination of IRNM or Tikhonov-type regularization with CP-BS as inner solver defines an attractive method for solving nonlinear respectively linear inverse problems in Banach spaces, as we also illustrated by numerical examples.

There are two main advantages of the generalization CP-BS over CP : First of all, the method becomes (efficiently) applicable to further interesting problems (**P**), such as problems with penalty functions $f(x) = \frac{1}{2}\|x\|_X^2$ that base on Banach space norms. Secondly, in numerical simulations we obtained significantly faster convergence by adapting the Banach spaces to the problem properties. This performance improvement is for example relevant for the crucial task of reconstructing sparse solutions.

There are some questions that could not be conclusively settled here. So, in numerical examples, we compared the different versions of CP-BS to each other and showed that a good parameter choice yields faster convergence. However, providing precise parameter choice rules is still a topic of interest in order to make this otherwise very simple method more user-friendly. Another open question is the validity of inequality (3.22). Proving this conjecture for $q \neq 2$ or a more general Banach space setting than we did in Proposition 3.3.11 would allow a further generalization of CP-BS .

A Appendix: Basics from Fourier theory

In this chapter we summarize some basics from Fourier analysis and distribution theory and study the discrete Fourier transform as the numerical approximation of its continuous counterpart. The definition and results of the following section can be found in the books [31] and [81]. Section A.2 is based on [19] and the lecture notes on "Inverse Problems II" by Thorsten Hohage [43].

A.1 Fourier analysis and tempered distributions

Let

$$\mathcal{F} : L^2(\mathbb{R}^2) \to L^2(\mathbb{R}^2), \quad (\mathcal{F}\varphi)(\xi') := \frac{1}{2\pi}\int_{\mathbb{R}^2} e^{-\mathrm{i}\xi'\cdot\mathbf{y}'}\varphi(\mathbf{y}')\,d\mathbf{y}'$$

denote the two-dimensional (continuous, bijective) Fourier transform with its (continuous) inverse

$$\mathcal{F}^{-1} : L^2(\mathbb{R}^2) \to L^2(\mathbb{R}^2), \quad (\mathcal{F}\varphi)(\mathbf{x}') := \frac{1}{2\pi}\int_{\mathbb{R}^2} e^{\mathrm{i}\mathbf{x}'\cdot\mathbf{y}'}\varphi(\mathbf{y}')\,d\mathbf{y}'.$$

Here $\mathbf{x}'\cdot\mathbf{y}'$ is the standard scalar product $x_1\,y_1 + x_2\,y_2$ of the two vectors $\mathbf{x}' = (x_1, x_2) \in \mathbb{R}^2$ and $\mathbf{y}' = (y_1, y_2) \in \mathbb{R}^2$. By the following theorem we summarize some properties of the Fourier transform which are repeatedly used in this work:

Theorem A.1.1. *The Fourier transform has the following properties:*

(i) For any $\varphi \in L^2(\mathbb{R}^2)$ and any $c \in \mathbb{R}\setminus\{0\}$ we have:

$$\mathcal{F}(\varphi(c\cdot))(\xi') = \frac{1}{|c|^2}\mathcal{F}\varphi\left(\frac{\xi'}{c}\right), \quad \xi' \in \mathbb{R}^2.$$

(ii) For any $\varphi \in L^2(\mathbb{R}^2)$ we have

$$\mathcal{F}\mathcal{F}\varphi(\mathbf{x}') = \varphi(-\mathbf{x}') \quad \mathbf{x}' \in \mathbb{R}^2.$$

(iii) On the space $L^1(\mathbb{R}^2)$ the Fourier transform $\mathcal{F} : L^1(\mathbb{R}^2) \to C_0(\mathbb{R}^2)$ exists as a continuous, linear, and bound operator with

$$\|\mathcal{F}\varphi\|_{L^\infty} \le \frac{1}{2\pi}\|\varphi\|_{L^1(\mathbb{R}^2)} \quad \varphi \in L^1(\mathbb{R}^2).$$

(iv) For any $\varphi \in L^1(\mathbb{R}^2) \cap L^2(\mathbb{R}^2)$ and $\psi \in L^1(\mathbb{R}^2)$ the convolution theorem *holds true:*

$$\mathcal{F}(\varphi * \psi) = 2\pi\,(\mathcal{F}\varphi \bullet \mathcal{F}\psi).$$

(v) For any $\varphi \in L^2(\mathbb{R}^2)$ and $\psi \in L^2(\mathbb{R}^2)$ the convolution theorem *holds true:*

$$\mathcal{F}(\varphi * \psi) = 2\pi\,(\mathcal{F}\varphi \bullet \mathcal{F}\psi).$$

(vi) For any $\varphi \in L^2(\mathbb{R}^2)$ and $\psi \in L^2(\mathbb{R}^2)$ the Plancherel formula *holds true:*

$$\langle \mathcal{F}\varphi, \mathcal{F}\psi \rangle_{L^2(\mathbb{R}^2)} = \langle \varphi, \psi \rangle_{L^2(\mathbb{R}^2)} \,.$$

Proof. The scaling property (i) is easy to check. *(ii)* follows directly from the identity $\mathcal{F}^{-1}\varphi(-\xi') = \mathcal{F}\varphi(\xi')$ for all $\varphi \in L^2(\mathbb{R}^2)$ and all $\xi' \in \mathbb{R}^2$ by applying $\mathcal{F}$. For the convolution theorems (iv) and (v) we refer to [31, Theorem 8.1.3] and [81, VII.5.10] . The assertions *(iii)* and (vi) can be found e.g. in [81, Satz V.2.2 and Equation (V.16)]. □

Now let us consider the *Schwartz space*

$$S\left(\mathbb{R}^2\right) := \left\{ \phi \in C^\infty(\mathbb{R}^2) \;\middle|\; \sup_{\mathbf{x}' \in \mathbb{R}^2} |\mathbf{x}'^\alpha D^\beta \phi(\mathbf{x}')| \leq \infty \text{ for all } \alpha, \beta \in \mathbb{N}_0^2 \right\}$$

where $D^\beta \phi = \frac{\partial^{\beta_1 + \beta_2}\phi}{\partial x_1^{\beta_1}\,\partial x_2^{\beta_2}}$. We equip $S\left(\mathbb{R}^2\right)$ with its usual topology generated by the seminorms

$$\|\phi\|_{\alpha,\beta} := \sup_{\mathbf{x}' \in \mathbb{R}^2} |\mathbf{x}'^\alpha D^\beta \phi(\mathbf{x}')| \qquad \alpha, \beta \in \mathbb{N}_0^2$$

and define its dual space $S'(\mathbb{R}^2)$, i.e. the space of tempered distributions. Since $S\left(\mathbb{R}^2\right)$ is dense in $L^r(\mathbb{R}^2)$ for any $r \in [1, \infty)$, and the introduced Fourier transform $\mathcal{F}$ can be considered as the unique continuous extension of $\mathcal{F} : S\left(\mathbb{R}^2\right) \to S\left(\mathbb{R}^2\right)$, the last theorem also applies for $\varphi, \psi \in S\left(\mathbb{R}^2\right)$. For $r \in [1, 2]$ and $r^* = \frac{r}{r-1}$ assertion *(iii)* can be generalized to the *Hausdorff-Young inequality* (cf. e.g. [81, Section V.2]):

$$\|\mathcal{F}\varphi\|_{L^{r^*}(\mathbb{R}^2)} \leq \frac{1}{(2\pi)^{2(1/r - 1/r^*)}} \|\varphi\|_{L^r(\mathbb{R}^2)} \quad \varphi \in S\left(\mathbb{R}^2\right).$$

Moreover, for $\alpha \in \mathbb{N}_0^2$, $\varphi \in S\left(\mathbb{R}^2\right)$ and $\psi(\mathbf{x}') := \mathbf{x}'^\alpha \varphi(\mathbf{x}')$, we have (e.g. [81, Satz V.2.10]):

$$D^\alpha(\mathcal{F}\varphi) = (-\mathrm{i})^{|\alpha|}\mathcal{F}\psi, \quad \text{and} \quad \mathcal{F}(D^\alpha\varphi)(\xi') = \mathrm{i}^{|\alpha|}\,\xi'^\alpha(\mathcal{F}\varphi)(\xi'). \tag{A.1}$$

The Fourier transform of a tempered distribution $T \in S'\left(\mathbb{R}^2\right)$ is given by

$$(\mathcal{F}T)(\varphi) = T(\mathcal{F}\varphi), \quad \varphi \in S\left(\mathbb{R}^2\right),$$

which defines an isomorphism $\mathcal{F} : S'\left(\mathbb{R}^2\right) \to S'\left(\mathbb{R}^2\right)$.

Example A.1.2. For $a \in \mathbb{R}^2$ let us introduce the delta-distribution $\delta_a \in S'\left(\mathbb{R}^2\right)$ given by $\delta_a(\phi) := \phi(a)$. Then we have

$$\mathcal{F}\delta_a(\phi) = \delta_a(\mathcal{F}\phi) = \mathcal{F}\phi(a) = \frac{1}{2\pi}\int_{\mathbb{R}^2} \mathrm{e}^{-\mathrm{i}a\cdot\mathbf{x}'}\phi(\mathbf{x}')\,d\mathbf{x}' \,.$$

In particular, the Fourier transform of $\phi \mapsto 2\pi\,\phi(0)$ is the constant "function" $T_\mathbf{1} \in S'\left(\mathbb{R}^2\right)$ mapping $\phi \in S\left(\mathbb{R}^2\right)$ to $\int_{\mathbb{R}^2} \mathbf{1}(\mathbf{x}')\,\phi(\mathbf{x}')\,d\mathbf{x}'$.

Given a function $\varphi \in L^r(\mathbb{R}^2)$ with $r \in [1,\infty]$ we denote by $T_\varphi \in S'\left(\mathbb{R}^2\right)$ the regular distribution

$$T_\varphi(\phi) = \int_{\mathbb{R}^2} \varphi(x)\,\phi(x)\,dx \quad \phi \in S\left(\mathbb{R}^2\right). \tag{A.2}$$

Then the convolution with $\phi \in S\left(\mathbb{R}^2\right)$ is defined as $T_\varphi * \phi = T_{\varphi*\phi}$. With these definitions Fourier's convolution theorem generalizes to

$$\mathcal{F}\left(T_\varphi * \psi\right) = 2\pi\left(\mathcal{F}T_\varphi \bullet \mathcal{F}\psi\right) \tag{A.3}$$

for all $\varphi \in L^r,\ r \in [1,\infty]$ and $\psi \in S\left(\mathbb{R}^2\right)$ (cf. [19]). More generally, let us consider the space

$$\mathcal{E}' := \left\{T \in S'\left(\mathbb{R}^2\right) \mid \operatorname{supp} T \text{ is compact}\right\},$$

of (tempered) distributions with compact support where the support of a distribution $T \in S'\left(\mathbb{R}^2\right)$ is defined as (cf. [31, Definition 1.4.1])

$$\operatorname{supp} T := \mathbb{R}^2 \setminus \bigcup \left\{U \subseteq \mathbb{R}^2 \mid T(\phi) = 0 \text{ for all } \phi \in C^\infty(\mathbb{R}^2) \text{ with } \operatorname{supp}\phi \subset U\right\}.$$

$\mathcal{E}'$ can also be identified with the dual of $C^\infty(\mathbb{R}^2)$ (cf. [31, p. 35, 97]). For any $E \in \mathcal{E}'$ we have ([31, Theorem 8.4.1]) $\mathcal{F}E = T_f$ where $f(\xi') = \langle K_{\mathcal{F}}(\xi',\cdot), E\rangle_{S(\mathbb{R}^2)} \in C^\infty(\mathbb{R}^2)$ and $K_{\mathcal{F}}(\xi',\mathbf{x}') := \frac{1}{2\pi}e^{-\mathrm{i}\xi'\cdot\mathbf{x}'}$. The convolution of $E \in \mathcal{E}'$ and $T \in S'\left(\mathbb{R}^2\right)$ is a tempered distribution given by

$$E * T(\phi) = \left\langle \langle \phi(\mathbf{x}'+\mathbf{y}'), E(\mathbf{x}')\rangle_{S(\mathbb{R}^2)}, T(\mathbf{y}')\right\rangle_{S(\mathbb{R}^2)} = \left\langle \langle \phi(\mathbf{x}'+\mathbf{y}'), T(\mathbf{x}')\rangle_{S(\mathbb{R}^2)}, E(\mathbf{y}')\right\rangle_{S(\mathbb{R}^2)}$$

for all $\phi \in S\left(\mathbb{R}^2\right)$. Moreover, Fourier's convolution theorem holds true ([31, Theorem 8.4.2]):

$$\mathcal{F}(E * T) = 2\pi\left(\mathcal{F}E \bullet \mathcal{F}T\right) \in S'\left(\mathbb{R}^2\right). \tag{A.4}$$

A.2 Discrete Fourier transform

The numerical implementation of the continuous Fourier transform is usually performed by the discrete Fourier transform. In order to study this relationship and to estimate the error of the approximation, we introduce for a given function $\varphi \in L^2(\mathbb{R}^2)$ and a period $\mathbf{T} = (T_1, T_2) > 0$ its $\mathbf{T}$-*periodization*:

$$\varphi_{\mathbf{T}}(\mathbf{x}') := \sum_{\mathbf{l}\in\mathbb{Z}^2} \varphi(\mathbf{x}' + \mathbf{T}\bullet\mathbf{l}) \quad \mathbf{x}' \in \mathbb{R}^2. \tag{A.5}$$

Note that for $\varphi \in L^2([-\mathbf{a},\mathbf{a}])$ supported in a rectangle $[-\mathbf{a},\mathbf{a}] \subset \mathbb{R}^2$ the $2\mathbf{a}$-periodization $\varphi_{2\mathbf{a}} \in L^2(\mathbb{R})$ is given by

$$\varphi_{2\mathbf{a}}(\mathbf{x}' + 2\mathbf{a}\,\mathbf{l}) = \varphi(\mathbf{x}'), \quad \mathbf{x}' \in [-\mathbf{a},\mathbf{a}],\ \mathbf{l} \in \mathbb{Z}^2, \mathbf{a}\,\mathbf{l} = \mathbf{a}\bullet\mathbf{l}$$

and hence extends φ $2\mathbf{a}$-periodically to $\mathbb{R}^2$.

Lemma A.2.1. *Let us assume that for a function* $\varphi \in L^2(\mathbb{R}^2)$ *the sum in* (A.5) *is uniformly absolute convergent. Then we have*

$$c_{\mathbf{k}}(\varphi_{\mathbf{T}}) := \frac{1}{T_1\,T_2}\int_{[-\mathrm{T}/2,\mathrm{T}/2]} \varphi_{\mathbf{T}}(\mathbf{x}')\, e^{-2\pi i \frac{\mathbf{k}}{\mathbf{T}}\cdot\mathbf{x}'}\, d\mathbf{x}' = \frac{2\pi}{T_1\,T_2}\mathcal{F}\varphi\left(\frac{2\pi\mathbf{k}}{\mathbf{T}}\right) \tag{A.6}$$

where $\frac{\mathbf{k}}{\mathbf{T}} = \left(\frac{k_1}{T_1}, \frac{k_2}{T_2}\right)$*. If* $\varphi \in S\left(\mathbb{R}^2\right)$*, the Fourier series of the* $\mathbf{T}$*-periodic function* $\varphi_{\mathbf{T}}$

$$\varphi_{\mathbf{T}}(\mathbf{x}') = \sum_{\mathbf{k}\in\mathbb{Z}^2} c_{\mathbf{k}}(\varphi_{\mathbf{T}})\, e^{2\pi i \frac{\mathbf{k}}{\mathbf{T}}\cdot\mathbf{x}'}, \quad \mathbf{x}' \in \mathbb{R}^2 \tag{A.7}$$

convergences absolute and uniformly and hence, for $\mathbf{x}' = 0$ *we end up with the* Poisson summation formula

$$\sum_{\mathbf{l}\in\mathbb{Z}^2} \varphi(T\mathbf{l}) = \frac{2\pi}{T_1 T_2}\sum_{\mathbf{l}\in\mathbb{Z}^2} \mathcal{F}\varphi\left(\frac{2\pi\mathbf{l}}{\mathbf{T}}\right).$$

Proof. Because of the uniform convergence, we can interchange the order of summation and integration which leads to the first assertion

$$\begin{aligned} \frac{2\pi}{T_1 T_2}\mathcal{F}\varphi\left(\frac{2\pi\mathbf{k}}{\mathbf{T}}\right) &= \frac{1}{T_1 T_2}\int_{\mathbb{R}^2} \varphi(\mathbf{x}')\, \mathrm{e}^{-\frac{\mathrm{i}2\pi\mathbf{k}\cdot\mathbf{x}'}{\mathbf{T}}}\, d\mathbf{x}' \\ &= \frac{1}{T_1 T_2}\sum_{\mathbf{l}\in\mathbb{Z}^2}\int_{[-\frac{\mathbf{T}}{2},\frac{\mathbf{T}}{2}]^2} \varphi(\mathbf{x}' + \mathbf{T}\bullet\mathbf{l})\, \mathrm{e}^{-\frac{\mathrm{i}2\pi\mathbf{k}\cdot\mathbf{x}'}{\mathbf{T}}}\, d\mathbf{x}' \\ &= \frac{1}{T_1 T_2}\int_{[-\frac{\mathbf{T}}{2},\frac{\mathbf{T}}{2}]^2} \varphi_{\mathbf{T}}(\mathbf{x}')\, \mathrm{e}^{-\frac{\mathrm{i}2\pi\mathbf{k}\cdot\mathbf{x}'}{\mathbf{T}}}\, d\mathbf{x}' = c_{\mathbf{k}}(\varphi_{\mathbf{T}}). \end{aligned}$$

Then, as $\mathcal{F}\varphi \in S\left(\mathbb{R}^2\right)$, there is a constant $C > 0$ such that:

$$\sum_{\mathbf{k}\in\mathbb{Z}^2} |c_{\mathbf{k}}(\varphi_{\mathbf{T}})| = \frac{2\pi}{T_1 T_2}\sum_{\mathbf{k}\in\mathbb{Z}^2}\left|\mathcal{F}\varphi\left(\frac{2\pi\mathbf{k}}{\mathbf{T}}\right)\right| \le \frac{2\pi}{T_1 T_2}\sum_{\mathbf{k}\in\mathbb{Z}^2}\frac{C}{\left(1 + 2\pi\left|\frac{\mathbf{k}}{\mathbf{T}}\right|\right)^4} < \infty.$$

So, the Fourier coefficients $c_{\mathbf{k}}(\varphi_{\mathbf{T}})$ are in l^1. This implies that the Fourier series of $\varphi_{\mathbf{T}}$ converges absolutely and uniformly to a continuous function $\tilde{\varphi}$. $\tilde{\varphi}$ has the same Fourier coefficients as $\varphi_{\mathbf{T}}$. Thus, we have $\tilde{\varphi} = \varphi_{\mathbf{T}}$ almost everywhere. As $\varphi_{\mathbf{T}}(\mathbf{x}') := \sum_{\mathbf{l}\in\mathbb{Z}^2}\varphi(\mathbf{x}' + T\mathbf{l})$ is a series of continuous functions, this series converges absolutely and uniformly on each ball in $\mathbb{R}^2$ and therefore $\varphi_{\mathbf{T}}$ is continuous as well. This proves the identity $\varphi_{\mathbf{T}}(\mathbf{x}') = \tilde{\varphi}(\mathbf{x}')$ for all $\mathbf{x}' \in \mathbb{R}$. □

Corollary A.2.2. *Let* $\varphi \in L^2(\mathbb{R}^2)$ *be a band limited function with bandwidth* $\mathbf{b} \in \mathbb{R}^2_+$*, i.e. supp* $\mathcal{F}\varphi \in [-\mathbf{b}, \mathbf{b}]$*. Moreover, we assume the Fourier transform* $\mathcal{F}\varphi \in L^2(\mathbb{R}^2)$ *to be piecewise differentiable. Then* $\mathcal{F}\varphi$ *can be represented by its Fourier series*

$$\mathcal{F}\varphi(\xi') = \frac{\pi}{2\,b_1\,b_2}\sum_{\mathbf{j}\in\mathbb{Z}^2}\varphi\left(\frac{\mathbf{j}\,\pi}{\mathbf{b}}\right)e^{-i\pi\frac{\mathbf{j}\cdot\xi'}{\mathbf{b}}}, \tag{A.8}$$

at each point $\xi' \in [-\mathbf{b}, \mathbf{b}]$, *where* $\mathcal{F}\varphi$ *is continuous. Vice versa a piecewise differentiable function* $\varphi \in L^2(\mathbb{R}^2)$ *with compact support in a rectangle* $[-\mathbf{a}, \mathbf{a}] \neq \emptyset$ *is given by*

$$\varphi(\mathbf{x}') = \frac{\pi}{2\, a_1\, a_2} \sum_{\mathbf{k}\in\mathbb{Z}^2} \mathcal{F}\varphi\left(\frac{\mathbf{k}\,\pi}{\mathbf{a}}\right) e^{i\,\pi \frac{\mathbf{k}\cdot\mathbf{x}'}{\mathbf{a}}} \tag{A.9}$$

at each point $\mathbf{x}' \in [-\mathbf{a}, \mathbf{a}]$, *where* φ *is continuous.*

Proof. [cf. [19, p. 213]] For $\mathcal{F}\varphi \in L^2(\mathbb{R}^2)$ supported in $[-\mathbf{b}, \mathbf{b}]$ consider the $2\mathbf{b}$-periodization $(\mathcal{F}\varphi)_{2\mathbf{b}}\,(\xi') = \sum_{\mathbf{l}\in\mathbb{Z}^2} \mathcal{F}\varphi(\xi' + 2\mathbf{b} \bullet \mathbf{l})$. Due to Dirichlet, we obtain from the piecewise differentiability of $\mathcal{F}\varphi$ that, at a point $\xi' \in \mathbb{R}^2$ where $(\mathcal{F}\varphi)_{2\mathbf{b}}$ is continuous, the function value $(\mathcal{F}\varphi)_{2\mathbf{b}}\,(\xi')$ is equal to its Fourier series

$$\sum_{\mathbf{j}\in\mathbb{Z}^2} \hat{c}_{\mathbf{j}}((\mathcal{F}\varphi)_{2\mathbf{b}})\, e^{-\mathrm{i}\,\pi \frac{\mathbf{j}\cdot\xi'}{\mathbf{b}}},$$

with coefficients

$$\hat{c}_{\mathbf{j}}((\mathcal{F}\varphi)_{2\mathbf{b}}) := \frac{\pi}{2\, b_1\, b_2} \int_{[-\mathbf{b},\mathbf{b}]} (\mathcal{F}\varphi(\mathbf{x}'))_{2\mathbf{b}}\; e^{\frac{\mathrm{i}\,\pi\,\mathbf{j}\cdot\mathbf{x}'}{\mathbf{b}}}\; d\mathbf{x}' = \frac{\pi}{2\, b_1\, b_2} \varphi\left(\frac{\pi\,\mathbf{j}}{\mathbf{b}}\right).$$

Now restricting $(\mathcal{F}\varphi)_{2\mathbf{b}}$ to $[-\mathbf{b}, \mathbf{b}]$, shows (A.8). The second assertion follows analogously where the Fourier coefficients of the Fourier series of $(\varphi)_{2\mathbf{a}}$ are given in Lemma A.2.1. □

Equation (A.9) shows that the discrete set of function values $\left\{\mathcal{F}\varphi\left(\mathbf{x}'_{\mathbf{k}}\right), \mid \mathbf{x}'_{\mathbf{k}} = \frac{\mathbf{k}\pi}{\mathbf{a}}, \mathbf{k} \in \mathbb{Z}^2\right\}$ completely determines a function $\varphi \in L^2(\mathbb{R}^2)$ that is compact supported in a rectangle $[-\mathbf{a}, \mathbf{a}]$. So, in order to reconstruct φ from its Fourier transform $\mathcal{F}\varphi$ uniformly sampled on a grid $\{\xi\mathbf{k} \mid \mathbf{k} \in \mathbb{Z}\}$, the sample rate ξ should be smaller or equal to the *Nyquist sampling rate* $\frac{\pi}{\mathbf{a}}$. Choosing $\mathbf{N} = (N_1, N_2) \in \mathbb{N}^2$ we sample φ on a grid

$$\left\{\frac{2\mathbf{a}}{\mathbf{N}}\mathbf{j} : \mathbf{j} \in \Delta_{\mathbf{N}}\right\} = \left\{-a_1, -a_1 + \frac{2\,a_1}{N_1}, \ldots, a_1 - \frac{2\,a_1}{N_1}\right\} \times \left\{-a_2, a_2 + \frac{2\,a_2}{N_2}, \ldots, a_2 - \frac{2\,a_2}{N_2}\right\}$$

where

$$\Delta_{\mathbf{N}} := \left\{-\frac{N_1}{2}, -\frac{N_1}{2} + 1, \ldots, \frac{N_1}{2} - 1\right\} \times \left\{-\frac{N_2}{2}, -\frac{N_2}{2} + 1, \ldots, \frac{N_2}{2} - 1\right\},$$

and set $\underline{\varphi}_{\mathbf{a}} := \left(\varphi\left(\frac{2\,a_1\,\mathbf{j}_1}{\mathbf{N}_1}, \frac{2\,a_2\,\mathbf{j}_2}{\mathbf{N}_2}\right)\right)_{\mathbf{j}\in\Delta_{\mathbf{N}}}$. Then for $\mathbf{b} = \frac{\pi\mathbf{N}}{2\mathbf{a}} = \frac{\pi}{2}\left(\frac{N_1}{a_1}, \frac{N_2}{a_2}\right)$ (and $\xi = \frac{\pi}{a}$) we obtain from Equation (A.9) that:

$$\underline{\varphi}_{\mathbf{a},\mathbf{j}} := \varphi\left(\frac{2\,\mathbf{a}\,\mathbf{j}}{\mathbf{N}}\right) = \frac{\pi}{2\,a_1 a_2} \sum_{\mathbf{k}\in\mathbb{Z}^2} \mathcal{F}\varphi\left(\frac{\pi\,\mathbf{k}}{\mathbf{a}}\right) e^{2\pi\,\mathrm{i}\,\frac{\mathbf{k}\cdot\mathbf{j}}{\mathbf{N}}} = \frac{2\,b_1 b_2}{\pi\,N_1 N_2} \sum_{\mathbf{k}\in\mathbb{Z}^2} \mathcal{F}\varphi\left(\frac{2\mathbf{b}\,\mathbf{k}}{\mathbf{N}}\right) e^{2\pi\,\mathrm{i}\,\frac{\mathbf{k}\cdot\mathbf{j}}{\mathbf{N}}}.$$

If $\varphi, \ldots, \varphi^{(m)} \in C$ and $\varphi^{(m)}$ is piecewise differentiable, the Fourier coefficients

$$\underline{\mathcal{F}\varphi}_{\mathbf{b},\mathbf{k}} := \mathcal{F}\varphi\left(\frac{2\mathbf{b}\,\mathbf{k}}{\mathbf{N}}\right) = \mathcal{F}\varphi\left(\frac{\mathbf{k}\,\pi}{a}\right)$$

satisfy the asymptotic $\underline{\mathcal{F}\varphi}_{\mathbf{b},\mathbf{k}} \in \mathcal{O}\left(\left(\frac{2}{\mathbf{k}}\right)^{m+1}\right)$. So, for a sufficiently large $\mathbf{N} = (N_1, N_2) \in \mathbb{N}^2$, we can assume that $\mathcal{F}\varphi$ vanishes outside of the square $[-\mathbf{b}, \mathbf{b}]$. This leads to the approximation:

$$\underline{\varphi}_{\mathbf{a},\mathbf{j}} = \varphi\left(\frac{2\,\mathbf{a}\,\mathbf{j}}{\mathbf{N}}\right) \approx \frac{2\,b_1 b_2}{\pi\, N_1 N_2} \sum_{\mathbf{k}\in\Delta_{\mathbf{N}}} \mathcal{F}\varphi\left(\frac{2\mathbf{b}\,\mathbf{k}}{\mathbf{N}}\right) e^{2\pi\,\mathrm{i}\,\frac{\mathbf{k}\cdot\mathbf{j}}{\mathbf{N}}} = \frac{2\,b_1 b_2}{\pi\,\sqrt{N_1 N_2}} \left(\mathcal{F}_{\mathbf{N}}^{-1} \underline{\mathcal{F}\varphi}_{\mathbf{b}}\right)_{\mathbf{j}}. \tag{A.10}$$

Here $\mathcal{F}_{\mathbf{N}} : \mathbb{C}^{N_1} \times \mathbb{C}^{N_2} \to \mathbb{C}^{N_1} \times \mathbb{C}^{N_2}$ is the *discrete Fourier transform* and $\mathcal{F}_{\mathbf{N}}^{-1}$ is its inverse which are given by

$$(\mathcal{F}_{\mathbf{N}}\underline{\psi})_{\mathbf{k}} := \frac{1}{\sqrt{N_1 N_2}} \sum_{\mathbf{j}\in\Delta_{\mathbf{N}}} \exp\left(-2\pi i\, \frac{\mathbf{j}}{\mathbf{N}} \cdot \mathbf{k}\right) \underline{\psi}_{\mathbf{j}}, \qquad \mathbf{k} \in \Delta_{\mathbf{N}}\,, \tag{A.11}$$

$$(\mathcal{F}_{\mathbf{N}}^{-1}\underline{\psi})_{\mathbf{j}} := \frac{1}{\sqrt{N_1 N_2}} \sum_{\mathbf{k}\in\Delta_{\mathbf{N}}} \exp\left(2\pi i\, \mathbf{k} \cdot \frac{\mathbf{j}}{\mathbf{N}}\right) \underline{\psi}_{\mathbf{k}}, \qquad \mathbf{j} \in \Delta_{\mathbf{N}}\,. \tag{A.12}$$

Applying $\mathcal{F}_{\mathbf{N}}$ to (A.10), we see that this approximation corresponds to

$$(\mathcal{F}\varphi)\left(\frac{2\,\mathbf{b}\,\mathbf{j}}{\mathbf{N}}\right) \approx \frac{2\,a_1 a_2}{\pi\,\sqrt{N_1 N_2}} \left(\mathcal{F}_{\mathbf{N}} \underline{\varphi}_{\mathbf{a}}\right)_{\mathbf{j}}. \tag{A.13}$$

Analogously, for a function $\varphi \in L^2(\mathbb{R}^2)$ with bandwidth $\mathbf{b} \in \mathbb{R}^2$ we conclude that

$$(\mathcal{F}^{-1}\varphi)\left(\frac{2\,\mathbf{b}\,\mathbf{j}}{\mathbf{N}}\right) \approx \frac{2\,a_1 a_2}{\pi\,\sqrt{N_1 N_2}} \left(\mathcal{F}_{\mathbf{N}}^{-1} \underline{\varphi}_{\mathbf{a}}\right)_{\mathbf{j}}. \tag{A.14}$$

In the following we will give an error estimate for the last approximations which are crucial for numerical use of the Fourier transform.

Proposition A.2.3. *Let $\varphi \in C(\mathbb{R}^2)$ be a $\mathbf{T}$-periodic function with absolutely convergent Fourier series:*

$$\varphi(\mathbf{x}') = \sum_{\mathbf{k}\in\mathbb{Z}^2} c_{\mathbf{k}}(\varphi_{\mathbf{T}})\, e^{\frac{2\pi i}{\mathbf{T}}\mathbf{k}\cdot\mathbf{x}'}, \quad \textit{where} \quad c_{\mathbf{k}}(\varphi_{\mathbf{T}}) := \frac{1}{T_1 T_2} \int_{-T_2/2}^{T_2/2} \int_{-T_1/2}^{T_1/2} \varphi_{\mathbf{T}}(\mathbf{x}')\, e^{-\frac{2\pi i}{\mathbf{T}}\mathbf{k}\cdot\mathbf{x}'}\, d\mathbf{x}'.$$

Then for $\underline{\varphi}_{\mathbf{T}/2,\mathbf{j}} := \varphi\left(\frac{\mathbf{T}\mathbf{j}}{\mathbf{N}}\right)$, $\mathbf{j} \in \Delta_{\mathbf{N}}$ we have

$$\frac{1}{\sqrt{N_1 N_2}} \left(\mathcal{F}_{\mathbf{N}} \underline{\varphi}_{\mathbf{T}/2}\right)_{\mathbf{k}} = \sum_{\mathbf{l}\in\mathbb{Z}^2} c_{\mathbf{k}+\mathbf{N}\mathbf{l}}\,(\varphi)\,. \tag{A.15}$$

Proof. By inserting the Fourier series of φ at some point $\mathbf{x}' = \frac{\mathbf{T}}{\mathbf{N}}\mathbf{l}$ with $\mathbf{l} \in \Delta_{\mathbf{N}}$, the right hand side of (A.15) reads as:

$$\frac{1}{\sqrt{N_1 N_2}} \left(\mathcal{F}_{\mathbf{N}} \underline{\varphi}_{\mathbf{T}/2}\right)_{\mathbf{k}} = \frac{1}{N_1 N_2} \sum_{\mathbf{j}\in\Delta_{\mathbf{N}}} \varphi\left(\frac{\mathbf{T}\mathbf{k}}{\mathbf{N}}\right) \mathrm{e}^{\left(-2\pi i\, \frac{\mathbf{j}\cdot\mathbf{k}}{\mathbf{N}}\right)} = \frac{1}{N_1 N_2} \sum_{\mathbf{l}\in\Delta_{\mathbf{N}}} c_{\mathbf{l}}(\varphi_{\mathbf{T}}) \sum_{\mathbf{j}\in\Delta_{\mathbf{N}}} \mathrm{e}^{\left(2\pi i\, \frac{\mathbf{j}\cdot(\mathbf{l}-\mathbf{k})}{\mathbf{N}}\right)}.$$

Thus, showing that the series $\sum_{\mathbf{j}\in\Delta_{\mathbf{N}}} \mathrm{e}^{\left(2\pi i\,\frac{\mathbf{j}\cdot(\mathbf{l}-\mathbf{k})}{\mathbf{N}}\right)}$ is N_1N_2 for $(\mathbf{l}-\mathbf{k}) \bmod \mathbf{N} = 0$ and vanishes everywhere else, completes the proof. For this purpose we recall for $N \in \mathbb{N}$ the identity

$$\sum_{j=0}^{N-1} \exp\left(2\pi \mathrm{i}\,\frac{j\,m}{N}\right) = \begin{cases} N & m \bmod N = 0 \\ 0 & \text{else.} \end{cases}$$

This together with $\exp\left(2\pi \mathrm{i}\,\frac{-j\,m}{N}\right) = \exp\left(2\pi \mathrm{i}\,\frac{(N-j)\,m}{N}\right)$ yields:

$$\begin{aligned} \sum_{\mathbf{j}\in\Delta_{\mathbf{N}}} \mathrm{e}^{\left(2\pi i\,\frac{\mathbf{j}\cdot(\mathbf{l}-\mathbf{k})}{\mathbf{N}}\right)} &= \sum_{\mathbf{j}_2=0}^{N_2-1}\sum_{\mathbf{j}_1=0}^{N_1-1} \mathrm{e}^{\left(2\pi i\,\frac{\mathbf{j}_1\cdot(\mathbf{l}_1-\mathbf{k}_1)}{N_1}\right)}\mathrm{e}^{\left(2\pi i\,\frac{\mathbf{j}_2\cdot(\mathbf{l}_2-\mathbf{k}_2)}{N_2}\right)} \\ &= \begin{cases} N_1N_2 & (\mathbf{l}_1-\mathbf{k}_1) \bmod N_1 = (\mathbf{l}_2-\mathbf{k}_2) \bmod N_2 = 0 \\ 0 & \text{else} \end{cases}. \qquad \square \end{aligned}$$

Note that for a function $\varphi \in L^2(\mathbb{R})$ with $\operatorname{supp}\varphi \subset \left[-\frac{\mathbf{T}}{2}, \frac{\mathbf{T}}{2}\right]$ and uniform absolute convergent $\frac{2\pi\mathbf{N}}{\mathbf{T}}$-periodization $(\mathcal{F}\varphi)_{\frac{2\pi\mathbf{N}}{\mathbf{T}}}$ of the Fourier transform we also obtain the identity (A.15) from Lemma A.2.1:

$$\begin{aligned} \sum_{l\in\mathbb{Z}^2} c_{\mathbf{k}+\mathbf{N}\mathbf{l}}\left(\varphi_{\mathbf{T}}\right) &= \frac{2\pi}{T_1T_2}\sum_{\mathbf{l}\in\mathbb{Z}^2}\mathcal{F}\varphi\left(\frac{2\pi}{\mathbf{T}}\left(\mathbf{k}+\mathbf{N}\mathbf{l}\right)\right) = \frac{2\pi}{T_1T_2}\left(\mathcal{F}\varphi\right)_{\frac{2\pi\mathbf{N}}{\mathbf{T}}}\left(\frac{2\pi}{\mathbf{T}}\mathbf{k}\right) \\ &= \frac{1}{N_1N_2}\sum_{\mathbf{j}\in\Delta_{\mathbf{N}}}\varphi\left(\frac{\mathbf{T}\mathbf{k}}{\mathbf{N}}\right)\exp\left(-2\pi i\,\frac{\mathbf{j}}{\mathbf{N}}\cdot\mathbf{k}\right) = \frac{1}{\sqrt{N_1N_2}}\left(\mathcal{F}_{\mathbf{N}}\underline{\varphi}_{\mathbf{T}/2}\right)_{\mathbf{k}}. \end{aligned}$$

Proposition A.2.4. *Let us assume that* $\varphi \in C(\mathbb{R}^2)$ *is a* $\mathbf{T}$*-periodic function with absolutely convergent Fourier series and weak derivatives* $\varphi^{(1)}, \dots \varphi^{(m)} \in L^1([-\mathbf{T}/2, \mathbf{T}/2])$ *up to order* $m \in \mathbb{N}$, $m > d$. *Then the error for the approximation of the continuous Fourier transform by the discrete one is bounded by a measure of smoothness*

$$\mu_m(\varphi) := \max\left\{\left\|\frac{\partial^m\varphi}{\partial x_1^m}\right\|_{L^1(\mathbb{R}^2)}, \left\|\frac{\partial^m\varphi}{\partial x_2^m}\right\|_{L^1(\mathbb{R}^2)}\right\}$$

in such a way that

$$\left|\frac{1}{\sqrt{N_1N_2}}\left(\mathcal{F}_{\mathbf{N}}\underline{\varphi}_{\mathbf{T}/2}\right)_{\mathbf{k}} - c_{\mathbf{k}}\left(\varphi\right)\right| \le \frac{C}{\|\mathbf{N}\|_{l^2}^m}\mu_m(\varphi), \qquad \mathbf{k} \in \Delta_{\mathbf{N}} \tag{A.16}$$

where $C > 0$ *is a constant and* $\mathbf{N} \in \mathbb{N}^2$ *is the number of sample points.*

Proof. The identity

$$c_{\mathbf{k}}\left(\frac{\partial^m\varphi}{\partial^m x_j}\right) = \left(\frac{2\pi\mathrm{i}}{T_j}k_j\right)^m c_{\mathbf{k}}\left(\varphi\right), \qquad \mathbf{k} \in \Delta_{\mathbf{N}},\ j = 1, 2$$

together with $|c_{\mathbf{k}}(\varphi)| \leq \|\varphi\|_{L^1(\mathbb{R}^2)}$ and Proposition (A.2.3) yields:

$$\left|\frac{1}{\sqrt{N_1 N_2}}\left(\mathcal{F}_{\mathbf{N}}\underline{\varphi}_{\mathbf{T}/2}\right)_{\mathbf{k}} - c_{\mathbf{k}}(\varphi)\right| \leq \sum_{\mathbf{l}\in\mathbb{Z}^2\setminus\{0\}} |c_{\mathbf{k}+\mathbf{N}\mathbf{l}}(\varphi)| \leq \sum_{\mathbf{l}\in\mathbb{Z}^2\setminus\{0\}} \left(\frac{\|\mathbf{T}\|_{l^\infty}}{2\pi\ \|\mathbf{k}+\mathbf{N}\mathbf{l}\|_{l^\infty}}\right)^m \mu_m(\varphi).$$

Now the assertion follows by

$$\begin{aligned}
\sum_{l\in\mathbb{Z}^2\setminus\{0\}} \left(\frac{\|\mathbf{T}\|_{l^\infty}}{2\pi\ \|\mathbf{k}+\mathbf{N}\mathbf{l}\|_{l^\infty}}\right)^m &\leq \sqrt{2}\left(\frac{\|\mathbf{T}\|_{l^\infty}}{2\pi}\right)^m \sum_{l\in\mathbb{Z}^2\setminus\{0\}} \|\mathbf{k}+\mathbf{N}\mathbf{l}\|_{l^2}^{-m} \\
&\leq \sqrt{2}\left(\frac{\|\mathbf{T}\|_{l^\infty}}{2\pi\,\|\mathbf{N}\|_{l^2}}\right)^m \sum_{l\in\mathbb{Z}^2\setminus\{0\}} \left\|\frac{\mathbf{k}}{\mathbf{N}}+\mathbf{l}\right\|_{l^2}^{-m} \\
&\leq \left(\frac{C\ \|\mathbf{T}\|_{l^\infty}}{2\pi\|\mathbf{N}\|_{l^2}}\right)^m \int_{\|\mathbf{x}'\|_{L^2} > \frac{\|\mathbf{T}\|_{l^\infty}}{2}} (x_1\ x_2)^{-m}\, d\mathbf{x}' \\
&= \left(\frac{C\ \|\mathbf{T}\|_{l^\infty}}{2\pi\|\mathbf{N}\|_{l^2}}\right)^m \int_{\frac{\|\mathbf{T}\|_{l^\infty}}{2}}^{\infty} (r)^{1-m}\ dr \\
&= \left(\frac{C}{\pi\|\mathbf{N}\|_{l^2}}\right)^m \frac{\|\mathbf{T}\|_{l^\infty}^2}{4(m-2)}.
\end{aligned}$$

□

References

[1] R. Adams and J. Fournier. *Sobolev Spaces*. Pure and Applied Mathematics. Elsevier Science, 2003.

[2] Y. I. Alber. Generalized projection operators in Banach spaces: properties and applications. *Functional Differential Equations, Proceedings of the Israel Seminar in Ariel*, 1:1–21, 1994.

[3] E. Asplund. Positivity of duality mappings. *Bulletin of the American Mathematical Society*, 73(2):200–203, 1967.

[4] A. Auslender, M. Teboulle, and S. Ben-Tiba. A Logarithmic-Quadratic Proximal Method for Variational Inequalities. In *Computational Optimization*, pages 31–40. Springer US, 1999.

[5] V. Barbu and T. Precupanu. *Convexity and Optimization in Banach Spaces*. Mathematics and Its Applications (East European Series). Bucureşti: D. Reidel Publishing Company, 1986.

[6] R. Barrett, R. Baker, P. Cloetens, Y. Dabin, C. Morawe, H. Suhonen, R. Tucoulou, A. Vivo, and L. Zhang. Dynamically-figured mirror system for high-energy nanofocusing at the ESRF. *SPIE*, pages 813904–813912, 2011.

[7] M. Bartels. *Cone-beam x-ray phase contrast tomography of biological samples*. PhD thesis, Universität Göttingen, 2013.

[8] H. Bauschke, R. Burachik, P. Combettes, V. Elser, D. Luke, and H. Wolkowicz. *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*. Springer Optimization and Its Applications. Springer, 2011.

[9] H. H. Bauschke, P. L. Combettes, and D. R. Luke. Phase retrieval, error reduction algorithm, and Fienup variants: a view from convex optimization. *J. Opt. Soc. Amer. A*, 19(7):1334–1345, 2002.

[10] H. H. Bauschke, X. Wang, and L. Yao. General resolvents for monotone operators: characterization and extension, Biomedical Mathematics: Promising Directions. In *in Imaging, Therapy Planning and Inverse Problems, Medical Physics Publishing*, pages 57–74, 2010.

[11] A. Beck and M. Teboulle. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[12] B. Blaschke, H. Engl, W. Grever, and M. Klibanov. An application of Tikhonov regularization to phase retrieval. *Nonlinear World*, 3(4):771–786, 1996.

[13] B. Blaschke-Kaltenbacher and H. Engl. Regularization methods for nonlinear ill-posed problems with applications to phase reconstruction. In *Inverse Problems in Medical Imaging and Nondestructive Testing*, pages 17–35. Springer, Wien-New York, 1997. (Oberwolfach Workshop, 1996).

[14] R. I. Boţ, E. R. Csetnek, and A. Heinrich. On the convergence rate improvement of a primal-dual splitting algorithm for solving monotone inclusion problems. *arXiv:1303.2875*, 2013.

[15] R. I. Boţ, E. R. Csetnek, and C. Hendrich. Recent developments on primal-dual splitting methods with applications to convex minimization. In *Mathematics Without Boundaries, Surveys in Interdisciplinary Research*, page 40. Springer, 2014.

[16] T. Bonesky, K. S. Kazimierski, P. Maass, F. Schöpfer, and T. Schuster. Minimization of Tikhonov Functionals in Banach spaces. *Abstract and Applied Analysis*, 2008.

[17] D. W. Boyd. The Power Method for l^p Norms. *Linear Algebra and its Applications*, 9:95–101, 1974.

[18] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

[19] R. Brigola. *Fourieranalysis, Distributionen und Anwendungen*. Vieweg-Lehrbuch angewandte Mathematik. Vieweg, 1997.

[20] L. Ceng, G. Mastroeni, and J. Yao. An Inexact Proximal-Type Method for the Generalized Variational Inequality in Banach Spaces. *Journal of Inequalities and Applications*, 2007(078124), 2007.

[21] A. Chambolle and T. Pock. A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging. *J. Math. Imaging Vis.*, 40(1):120–145, 2011.

[22] M. Cheney and D. Isaacson. Inverse problems for a perturbed dissipative half-space. *Inverse Problems*, (11):856–88, 1995.

[23] I. Cioranescu. *Geometry of Banach Spaces, Duality Mappings and Nonlinear Problems*. Mathematics and Its Applications. Springer, 1990.

[24] D. Colton and R. Kress. *Inverse acoustic and electromagnetic scattering theory*, volume 93. Springer, 2012.

[25] P. L. Combettes and J.-C. Pesquet. Proximal Splitting Methods in Signal Processing. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 185–212. Springer, 2011.

[26] W. V. Combettes, P.L. Signal recovery by proximal forward-backward splitting. *Multiscale Model. Simul.*, 4:1168–1200, 2005.

[27] V. Davidoiu, B. Sixou, M. Langer, and F. Peyrin. Non-linear iterative phase retrieval based on Frechet derivative. *Opt. Express*, 19(23):22809–22819, 2011.

[28] J. P. Dinca, G. and J. Mawhin. Variational and topological methods for Dirichlet problems with p-Laplacian. *Portugaliae Mathematica. Nova Série*, 58(3):339–378, 2001.

[29] D. C. Dobson. Phase reconstruction via nonlinear least-squares. *Inverse Problems*, 8(4):541, 1992.

[30] J. R. Fienup. Reconstruction of an object from the modulus of its Fourier transform. *Opt. Lett.*, 3(1):27–29, 1978.

[31] F. G. Friedlander. *Introduction to the theory of distributions*. Cambridge University Press, 1982.

[32] R. W. Gerchberg and W. O. Saxton. A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik*, 35:237–250, 1972.

[33] K. Giewekemeyer. *A study on new approaches in coherent x-ray microscopy of biological specimens*. PhD thesis, Universität Göttingen, 2011.

[34] K. Giewekemeyer, S. 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.

[35] J. Haber. Optimization of the Wafer Bondig Process for the Fabrication of Lithographic X-Ray Waveguides. Master's thesis, Universität Göttingen, 2013.

[36] J. Hagemann, A.-L. Robisch, D. R. Luke, C. Homann, T. Hohage, P. Cloetens, H. Suhonen, and T. Salditt. Reconstruction of wave front and object for inline holography from a set of detection planes. *Opt. Express*, 22(10):11552–11569, 2014.

[37] B. He and X. Yuan. Convergence Analysis of Primal-Dual Algorithms for a Saddle-Point Problem: From Contraction Perspective. *SIAM Journal on Imaging Sciences*, 5(1):119–149, 2012.

[38] R. Hesse. *Fixed Point Algorithms for Nonconvex Feasibility with Applications*. PhD thesis, Universität Göttingen, 2014.

[39] R. Hesse and D. Luke. Nonconvex Notions of Regularity and Convergence of Fundamental Algorithms for Feasibility Problems. *SIAM Journal on Optimization*, 23(4):2397–2419, 2013.

[40] B. Hofmann, B. Kaltenbacher, C. Pöschl, and O. Scherzer. A convergence rates result for Tikhonov regularization in Banach spaces with non-smooth operators. *Inverse Problems*, 23(3):987, 2007.

[41] T. Hohage. On the numerical solution of a three-dimensional inverse medium scattering problem. *Inverse Problems*, 17(6):1743, 2001.

[42] T. Hohage. Fast numerical solution of the electromagnetic medium scattering problem and applications to the inverse problem. *Journal of Computational Physics*, 214(1):224 – 238, 2006.

[43] T. Hohage. Inverse problems II. University Lecture, summer term 2012.

[44] T. Hohage, K. Giewekemeyer, and T. Salditt. Iterative reconstruction of a refractive index from x-ray or neutron reflectivity measurements. *Physical Review E*, 77(5):051604, 2008.

[45] T. Hohage and C. Homann. A Generalization of the Chambolle-Pock Algorithm to Banach Spaces with Applications to Inverse Problems. *arXiv:1412.0126*, 2014.

[46] T. Hohage and S. Langer. Acceleration techniques for regularized Newton methods applied to electromagnetic inverse medium scattering problems. *Inverse Problems*, 26(7):074011, 2010.

[47] T. Hohage and F. Werner. Iteratively regularized Newton-type methods with general data misfit functionals and applications to Poisson data. *Numerische Mathematik*, 123(4):745–779, 2013.

[48] C. Homann, T. Hohage, J. Hagemann, A.-L. Robisch, and T. Salditt. On the validity of the empty beam correction in near field imaging. *Physical Review A*. to appear.

[49] S. Kalbfleisch, H. Neubauer, S. P. Krüger, M. Bartels, M. Osterhoff, D. D. Mai, K. Giewekemeyer, B. Hartmann, M. Sprung, and T. Salditt. The Göttingen Holography Endstation of Beamline P10 at PETRA III/DESY. *AIP Conference Proceedings*, 1365(1):96–99, 2011.

[50] S. Kamimura and W. Takahashi. Strong Convergence of a Proximal-Type Algorithm in a Banach Space. *SIAM J. on Optimization*, 13(3):938–945, 2002.

[51] M. V. Klibanov. On the recovery of a 2-D function from the modulus of its Fourier transform. *J. Math. Anal. Appl.*, 323(2):818–843, 2006.

[52] A. Lechleiter, K. S. Kazimierski, and M. Karamehmedović. Tikhonov regularization in l^p applied to inverse medium scattering. *Inverse Problems*, 29(7):075003, 2013.

[53] A. Levi and H. Stark. Image restoration by the method of generalized projections with application to restoration from magnitude. *J. Opt. Soc. Am. A*.

[54] J. Lindenstrauss and L. Tzafriri. *Classical Banach Spaces II: Function Spaces*. Ergebnisse der Mathematik und ihrer Grenzgebiete. Springer, 1979.

[55] D. A. Lorenz and T. Pock. An accelerated forward-backward algorithm for monotone inclusions. *CoRR*, abs/1403.3522, 2014.

[56] S. Marchesini. A unified evaluation of iterative projection algorithms for phase retrieval. *Review of Scientific Instruments*, 78(1):011301, 2007.

[57] S. Maretzke. A uniqueness result for propagation-based phase contrast imaging from a single measurement. *arXiv:1409.4794*, 2014.

[58] K. Nikodem and Z. Pales. Characterizations of inner product spaces by strongly convex functions. *Banach Journal of Mathematical Analysis*, 5(1):83–87, 2011.

[59] C. Olendrowitz, M. Bartels, M. Krenkel, A. Beerlink, R. Mokso, M. Sprung, and T. Salditt. Phase-contrast x-ray imaging and tomography of the nematode Caenorhabditis elegans. *Phys. Med. Biol.*, 57(16):5309, 2012.

[60] D. M. Paganin. *Coherent X-ray Optics*. Oxford Series on Synchrotron Radiation. New York: Oxford University Press, 2006.

[61] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):123–231, 2013.

[62] T. Pock and A. Chambolle. Diagonal Preconditioning for First Order Primal-dual Algorithms in Convex Optimization. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, pages 1762–1769, 2011.

[63] C. Pöschle. *Tikhonov Regularization with General Residual Term*. PhD thesis, Universität Insbruck, 2008.

[64] A.-L. Robisch and T. Salditt. Phase retrieval for object and probe using a series of defocus near-field images. *Opt. Express*, 21(20):23345–23357, 2013.

[65] R. Rockafellar. On the maximal monotonicity of subdifferential mappings. *Pacific J. Math.*, 33(1):209–216, 1970.

[66] R. Rockafellar. *Convex Analysis*. Princeton University Press, 1997.

[67] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

[68] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear Total Variation Based Noise Removal Algorithms. *Phys. D*, 60(1-4):259–268, Nov. 1992.

[69] T. Salditt, S. Kalbfleisch, M. Osterhoff, S. P. Krüger, M. Bartels, K. Giewekemeyer, H. Neubauer, and M. Sprung. Partially coherent nano-focused x-ray radiation characterized by Talbot interferometry. *Opt. Express*, 19(10):9656–9675, 2011.

[70] T. Schuster, B. Kaltenbacher, B. Hofmann, and K. Kazimierski. *Regularization Methods in Banach Spaces*, volume 10 of *Radon Series on Computational and Applied Mathematics*. De Gruyter, 2012.

[71] M. Solodov and B. Svaiter. Forcing strong convergence of proximal point iterations in a Hilbert space. *Mathematical Programming*, 87(1):189–202, 2000.

[72] M. Taylor. *Partial Differential Equations: Nonlinear Equations*, volume 3. Springer, New York, 1996.

[73] P. Thibault, M. Dierolf, A. Menzel, O. Bunk, C. David, and F. Pfeiffer. High-Resolution Scanning X-ray Diffraction Microscopy. *Science*, 321(5887):379–382, 2008.

[74] R. Tibshirani. Regression Shrinkage and Selection Via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

[75] A. N. Tikhonov. Regularization of incorrectly posed problems. *Soviet Math. Doklady*, 4:1624–1627, 1963.

[76] A. N. Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Doklady*, 4:1035–1038, 1963.

[77] H. Triebel. *Theory of function spaces*, volume 78 of *Monographs in mathematics*. Basel: Birkhäuser Verlag, 1983.

[78] B. C. Vũ. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Computational Mathmatics*, 38(3):667–681, 2013.

[79] G. Vainikko. Fast Solvers of the Lippmann-Schwinger Equation. In *Direct and Inverse Problems of Mathematical Physics*, volume 5 of *Int. Soc. for Anal., Appl. and Comput.*, pages 423–440. 2000.

[80] T. Valkonen. A primal-dual hybrid gradient method for nonlinear operators with applications to MRI. *Inverse Problems*, 30(5):055012, 2014.

[81] D. Werner. *Funktionalanalysis*. Springer-Lehrbuch. Springer, 2011.

[82] F. Werner. *Inverse problems with Poisson data: Tikhonov-type regularization and iteratively regularized Newton methods*. PhD thesis, University of Göttingen, 2012.

[83] F. Werner and T. Hohage. Convergence rates in expectation for Tikhonov-type regularization of inverse problems with Poisson data. *Inverse Problems*, 28:104004, 2012.

[84] Z.-B. Xu. Characteristic inequalities of l^p spaces and their applications. *Acta Math. Sinica*, 32(2):209–218, 1989.

[85] Z.-B. Xu and G. Roach. Characteristic inequalities of uniformly convex and uniformly smooth Banach spaces. *Journal of Mathematical Analysis and Applications*, 157(1):189–210, 1991.

[86] C. Zălinescu. *Convex analysis in general vector spaces*. River Edge, NJ : World Scientific, 2002.

Acknowledgements

Firstly, I would like to thank my supervisor Prof. Dr. Thorsten Hohage for his continuous guidance, encouragement, and motivation. He introduced me to the very interesting fields of phase retrieval problems and convex optimization. I am grateful for our fruitful discussions and the invaluable suggestions he offered.

I would also like to express my thanks to Prof. Dr. Russell Luke not only for supporting me as my second supervisor but also for reviewing this thesis as second referee.

Very special thanks go to Prof. Dr. Tim Salditt and his working group. I am particular thankful to Prof. Dr. Tim Salditt, Johannes Hagemann, and Anna-Lena Robisch for fruitful discussions concerning the empty beam correction and for giving the mathematical theory a physical meaning. The great collaboration between our working groups has been a nice source of motivation and inspiration.

I further would like to thank my former teacher Reinhard Loges for having inspired my deep interest in mathematics and for his continued support of this interest.

More thanks go to my colleagues from the Institute for Numerical and Applied Mathematics, in particular the members of my working group, for providing a great working atmosphere. Especially I would like to express my thanks to Helen Schomburg, Dr. Robert Hesse, and Dr. Frank Werner for some good advice, helpful discussions, and encouraging words.

I am grateful to Verena, Judith, and Carsten for proofreading parts of this thesis.

Besonderer Dank gilt meinen Eltern und meiner Schwester, auf deren Unterstützung und Verständnis ich mich stets verlassen konnte und die immer ein offenes Ohr für mich haben.

My very special thanks go to Carsten for all his invaluable support and understanding. He shows it to me in every circumstance.

Finally, I would like to thank the DFG Collaborative Research Center SFB 755 “Nanoscale Photonic Imaging” and the Graduiertenkolleg 1023 “Identifikation in mathematischen Modellen: Synergie stochastischer und numerischer Methoden” for providing the funding of my research.

Curriculum Vitae

Carolin Homann, M.Sc.

Address	Institute for Numerical and Applied Mathematics University of Göttingen Lotzestraße 16-18 37083 Göttingen / Germany
Office:	+49/551/394510
Email:	c.homann@math.uni-goettingen.de

Personal Details

Gender	Female
Date of birth	14th of August, 1986
Place of birth	Peine, Germany
Citizenship	German

Education

Since 10/2011	Ph.D. student of mathematics at the University of Göttingen (Germany). Supervisor: Professor Dr. Thorsten Hohage.
10/2006-09/2011	Student of mathematics at the University of Göttingen (Germany). 09/2009 Bachelor of Science, Advisor: Professor Dr. Rainer Kress. 09/2011 Master of Science, Advisor: Professor Dr. Rainer Kress.
09/1999–06/2006	Secondary school ’Ratsgymnasium Peine’ in Peine (Germany). 06/2006 Allgemeine Hochschulreife.
08/1997–07/1999	Comprehensive school ’Orientierungsstufe Edemissen’ in Edemissen (Germany).
08/1993–07/1997	Primary school ’Grundschule Abbensen’ in Edemissen-Abbensen (Germany).

Research Experience

Since 10/2011	Research assistant at the Institute for Numerical and Applied Mathematics, University of Göttingen (Germany) in the Collaborative Research Centre 755 Nanoscale Photonic Imaging of the German Research Foundation.

In phase retrieval problems that occur in imaging by coherent x-ray diffraction, one tries to reconstruct information about a sample of interest from possibly noisy intensity measurements of the wave field traversing the sample. The mathematical formulation of these problems bases on some assumptions. Usually one of them is that the x-ray wave field is generated by a point source. In order to address this very idealized assumption, it is common to perform a data preprocessing step, the so-called empty beam correction. Within this work, we study the validity of this approach by presenting a quantitative error estimate. Moreover, in order to solve these phase retrieval problems, we want to incorporate a priori knowledge about the structure of the noise and the solution into the reconstruction process. For this reason, the application of a problem adapted iteratively regularized Newton-type method becomes particularly attractive. This method includes the solution of a convex minimization problem in each iteration step. We present a method for solving general optimization problems of this form. Our method is a generalization of a commonly used algorithm which makes it efficiently applicable to a wide class of problems. We also proof convergence results and show the performance of our method by numerical examples.

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

ISBN 978-3-86395-210-5
ISSN 2191-9860

Universitätsverlag Göttingen