“Joint Exit Wave Reconstruction and Multimodal Registration of Transmission Electron Microscopy Image Series”

Von der Fakultät für Mathematik, Informatik und Naturwissenschaften der RWTH Aachen University zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften genehmigte Dissertation

vorgelegt von

Master of Science

Christian Doberstein

aus

Düsseldorf

Berichter: Prof. Dr. Benjamin Berkels
Prof. Dr. Christof Melcher

Tag der mündlichen Prüfung: 19.06.2020

Diese Dissertation ist auf den Internetseiten der Universitätsbibliothek verfügbar.

Abstract

Images generated with a transmission electron microscope (TEM) can reveal information up to the scale of individual atoms. However, the information contained in a TEM image is blurred by aberrations and the partial coherence of the electron beam. Furthermore, the images correspond to the squared amplitude of the image plane electron wave and are thus missing valuable information about the phase. Exit wave reconstruction attempts to solve these problems by reconstructing the electron wave at the exit plane of the specimen, the so-called exit wave, from a series of images recorded with varying focus of the objective lens. This introduces the additional problem of aligning the image series, which is crucial for a successful reconstruction of the exit wave. One possible approach to reconstructing the exit wave involves the minimization of a least squares functional, which is implemented by the well-known MIMAP and MAL algorithms. The MIMAP and MAL algorithms solve the registration problem by alternatingly optimizing the exit wave and the registration. In this thesis, a novel objective functional E_σ for the joint optimization of the exit wave and the registration is proposed. The properties of the forward model of TEM image simulation, which is given by a weighted autocorrelation of the exit wave, are investigated on the basis of the weighted cross-correlation and the novel notion of $\star$-separable weights. The most important weight functions (commonly called transmission cross-coefficients, TCCs) for TEM image simulation are analyzed and integrated into the present framework. The results regarding the forward model are then used for the analysis of the inverse problem. It is shown that the data term of E_σ is not coercive for $\star$-separable TCCs, which in particular implies that the MAL functional is not coercive. One of the main results is the existence of minimizers of the objective functional E_σ, which is shown with the direct method. Additionally, it is shown that the objective functional is not convex in general. These results are complemented by a numerical analysis, which includes the discretization of the objective functional and the treatment of several problems regarding the numerical minimization of E_σ. A novel preconditioner for the exit wave is proposed, showing a reduction of the number of iterations for a given residual energy. The least squares sum in the data term of the objective functional is usually calculated by summing the squared differences of the simulated and experimental images over the same domain for each image. A novel method for the dynamic adjustment of these domains based on the current estimate for the registration is proposed, which allows to use the full image data for the reconstruction while at the same time avoiding the need for a continuation of the images. Numerical experiments are presented that evaluate the utility of the preconditioner and compare the alternating optimization approach with the joint optimization of the exit wave and the registration. Finally, a numerical experiment shows the result of reconstructing the exit wave for a real image series.

Acknowledgements

First of all, I would like to thank my advisor, Prof. Dr. Benjamin Berkels, for his continuous support and guidance during my doctoral studies, for sharing his knowledge and always spontaneously taking the time to discuss new ideas or difficult problems and for giving me the freedom to explore all of my ideas. I am particularly thankful to him for questioning and double-checking all of my results, which revealed some subtle mistakes that would have went unnoticed otherwise, and for suggesting the topic of this thesis that is such a perfect match to my interests.

I am grateful to Dr. Chen Huang for providing me with experimental TEM image series that were very helpful in evaluating the results on real input data.

I would also like to thank my colleagues for many helpful and stimulating discussions about mathematics, computer science and science in general; in particular I wish to thank Alexander Zimmerman, Jan Winkelmann and Kai Schüller.

I thank Christian Schmitz, who offered proofreading the first draft of this thesis and provided many helpful comments.

Finally, I would like to thank my family: my parents, who have supported me from the very beginning of my bachelor studies, and my brothers, for their support and for helping me move quite a few times.

Contents

Introduction

Since its invention by Knoll and Ruska in 1932 [44], electron microscopy has proven to be an indispensable tool for the investigation of biological and non-biological specimen at magnifications beyond the limits of conventional visible light microscopy. The first electron microscope was a transmission electron microscope (TEM), in which a beam of accelerated electrons is transmitted through a specimen to generate a strongly magnified image. The underlying reason why electrons are suitable for high resolution microscopy is the fact that, depending on the accelerating voltage of the electrons, their associated de-Broglie wavelength can be several orders of magnitude smaller than the shortest wavelength of visible light. This makes it possible to record images at atomic resolution and investigate defects such as dislocations, planar faults and stacking faults [62].

Ultimately, the way that images are recorded with a transmission electron microscope is comparable to visible light microscopy in many regards. The electrons in a TEM are diverted by electromagnetic lenses that focus the electrons in a manner very similar to the focusing of photons by optical lenses. One of the main differences between electromagnetic lenses and optical lenses is the fact that a rotationally symmetric magnetic field always results in a positive spherical aberration [60]. In other words, there is no direct way to build “concave” electromagnetic lenses.

While information about the specimen can directly be extracted from the recorded TEM images, it is frequently desirable to post process the data in order to obtain additional information about the specimen. In particular, the recording device of conventional TEMs only measures the squared amplitude of the electron wave in the image plane, meaning that the phase of the image plane electron wave is completely missing in the images. However, at least for very thin specimens, the phase contains most of the high-frequency information about the specimen [65]. Additionally, the images are blurred by the aberrations of the objective lens and the partial coherence of the electron beam, which can make a correct interpretation difficult.

As the electron wave at the exit place of the specimen, the so-called exit wave, is free from aberrations of the objective lens, it can be useful to reconstruct the exit wave from a series of images that have been recorded with varying focusing settings of the objective lens. The well-known MIMAP [38, 39] and MAL [10] algorithms implement such an reconstruction of the exit wave with a variational approach by performing a minimization of suitable least squares functionals. As these functionals measure the distance of the experimental TEM images to images that are simulated from a current estimate of the exit wave, the MIMAP and MAL algorithms necessarily are an indirect approach to exit wave reconstruction.

An important aspect of reconstructing the exit wave from a series of images is the fact that the images are in general not aligned very well to each other due to specimen drift and microscope instabilities. Thus, a registration of the image series has to be performed in addition to the reconstruction of the exit wave, which is implemented in the MIMAP and MAL algorithms with an alternating optimization scheme.

In this thesis, the MIMAP and MAL algorithms are generalized to a joint optimization algorithm, where the exit wave and the registration are optimized simultaneously instead of alternatingly. Based on the functionals from the MIMAP and MAL algorithms, a novel objective functional E_σ is defined that depends on both the current estimate for the exit wave and the registration.

The aim of this thesis is twofold. To date, only very little emphasis has been put on the mathematical properties of the MIMAP and MAL algorithms. Therefore, one of the main aims is to give a complete and coherent mathematical description of the variational approach to exit wave reconstruction. This is facilitated by the novel objective functional E_σ as it joins the optimization of the exit wave and the registration of the image series into a single minimization problem. The main results in this regard are the existence of minimizers of E_σ, its non-convexity and the non-coercivity of its data term. These results are based on a novel representation of the forward model using the notion of the weighted cross-correlation and $\star$-separable weights. The second aim of this thesis is the numerical minimization of the objective functional. For this purpose, several pecularities of the discretized inverse problem are pointed out and corresponding solutions proposed, which have been implemented in a C++ program that has been written as a part of this thesis.

This thesis is divided into 5 chapters. Chapter 1 gives a brief overview of TEM image formation and existing methods for exit wave reconstruction as well as the notation and some common definitions that are used throughout this thesis. In Chapter 2, the forward model of simulating a TEM image is analysed in detail. Structurally, the simulation of TEM images from a given exit wave is described by an autocorrelation of the exit wave weighted by the transfer function of the objective lens. For this reason, the properties of the weighted cross-correlation are investigated, which is done independently from the application to the forward model and electron microscopy. Particular emphasis is also put on the mathematical properties of the different transfer functions that are most commonly used for TEM image simulation. In Chapter 3, the coercivity, the convexity and the existence of minimizers of the objective functional are investigated. A part of the mathematical analysis of E_σ has previously been published in [14] and is extended here in several ways. In Chapter 4, the discretization of the objective functional and different aspects of the numerical minimization of the objective functional are discussed. This chapter also contains a brief description of the implementation of the C++ program that has been used for the numerical experiments. Chapter 5 contains three exemplary exit wave reconstructions, including a comparison of an alternating optimization scheme with the joint optimization of the exit wave and the registration as well as an experiment on a real image series.

It is pointed out that most of the results in this thesis are not limited to transmission electron microscopy (also abbreviated as TEM) and more generally applicable to any least squares problem that involves a weighted cross-correlation as the forward model as long as the weight function fulfills the properties given in Chapter 2.

1 Transmission electron microscopy and exit wave reconstruction

In the first part of this chapter, the imaging process in TEM is outlined by dividing it into four steps: (1) the emission of electrons from an electron source, (2) the transmission of the electrons through the specimen, (3) the diversion of the electrons by the objective lens, and (4) the recording of the electrons by a camera.

This simplified view on image formation in TEM is sufficient for the purpose of this thesis, which is the inversion of the steps (3) and (4). The inversion of these two steps is commonly termed "exit wave reconstruction" or "exit wave restoration". The basic idea behind exit wave reconstruction and its limitations are presented in the second part of this chapter.

The general information about TEM image formation presented in this chapter was extracted from the books [18, 41, 35, 16, 63]. In addition to the brief overview given here, these books also provide more detailed information on the individual steps of image formation as well as on any TEM component that is not mentioned here (e.g. the condenser lenses or the vacuum system).

1.1 Image formation in transmission electron microscopy

The fundamental advantage of electron microscopy over visible light microscopy is the fact that the de-Broglie wavelength $\lambda = h/p$ associated to an electron is inversely proportional to the electron's momentum p, where the proportionality constant h is Planck's constant. For typical electron accelerating voltages in TEM this results in a significantly shorter wavelength than visible light and consequently in a potentially much higher magnification according to the diffraction limit. For example, the de-Broglie wavelength of electrons accelerated through a potential of 100 kV is roughly 3.7 pm, which is five orders of magnitude smaller than the wavelength of visible light. However, unlike visible light microscopy, the resolution prescribed by the diffraction limit is not attained or even surpassed in electron microscopy, which is due to lens aberrations and the considerably higher requirements on the mechanical and electrical stability of the microscope.

The process of image formation in TEM can be divided into four steps as illustrated in Figure 1.1. These steps are described briefly in the following, with special emphasis on the parts relevant to exit wave reconstruction:

1. The electrons are emitted from an electron source, which is most commonly a thermionic emitter or a field emission gun (FEG). FEGs are superior to thermionic

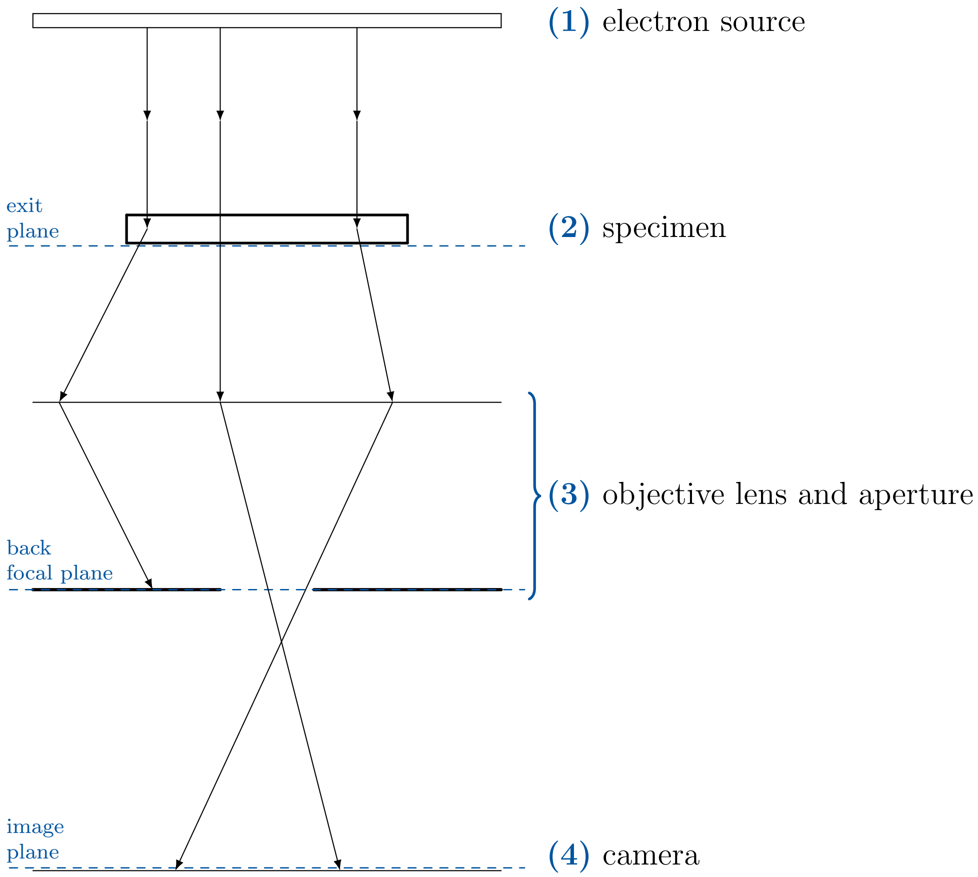

Figure 1.1: Schematic illustration of TEM image formation. The arrows depict three exemplary electron trajectories.

emitters due to their high electron-optical brightness and small energy spread of the emitted electrons, which results in a higher spatial and temporal coherence of the electron beam [35, 56]. The temporal coherence is additionally affected by instabilities in the accelerating voltage and the objective lens currents.

2. Transmission of the electrons through the specimen: because of diffraction inside the specimen, the electrons carry information about the specimen when leaving at its exit plane. This structural information is completely contained within the electron wave at the exit plane of the specimen, which is also called "exit wave" for short. For very thin specimens consisting of only light atoms, a direct interpretation of the exit wave or the TEM images is possible by means of the so-called weak phase object approximation (WPOA). If the specimen is not sufficiently thin or contains heavy atoms, then Bloch wave methods or the multislice algorithm and channelling theory can be utilized for an interpretation of the exit wave or the TEM images [41, 11]. While understanding this part of the image formation process is essential to the analysis of the TEM images and the exit wave, it is not

directly relevant for the reconstruction of the exit wave and therefore not discussed in more detail here.

3. After passing through the specimen, the electrons are focused by the objective lens of the microscope. As the objective lens is not an ideal lens, it introduces several aberrations to the generated images. In terms of the achievable resolution, the most important aberrations are the focus Z, the third-order spherical aberration C_s and the chromatic aberration C_c. The focus is a special aberration as it can be easily adjusted to arbitrary positive or negative values, which can be exploited in several ways. On the one hand, the Scherzer focus [60] yields optimum phase contrast by partially compensating for the positive spherical aberration, whereas the Lichte focus [47] maximizes the amount of information contained in the images at the cost of requiring numerical postprocessing. On the other hand, the objective lens focus can also be varied in order to retrieve a whole series of images with different focus from the same object, which can then be used for the reconstruction of the exit wave as outlined in Section 1.2. The objective aperture is located below the objective lens and acts as a low-pass filter on the electron wave by removing electrons that have been scattered to a high angle depending on the radius of the aperture.

4. In the last stage, the electrons are recorded by a camera. Most commonly, a slow-scan charge-coupled device (CCD) camera is used to record photons emitted from a phosphor screen located in front of the camera that converts an electron into a set of photons. The non-ideality of the electron detection process is described by the modulation transfer function (MTF), which acts as a blurring envelope on the recorded images [49]. Recently, new CMOS (complementary metal oxide semiconductor) cameras have been developed that are capable of detecting electrons directly, which results in an improved MTF and signal-to-noise ratio [27, 48]. However, both image acquisition devices share the fundamental limitation that the recorded image is real-valued and proportional to the squared amplitude of the electron wave in the image plane, lacking any information about the phase of the image plane electron wave. In the case of an ideal detector, the recorded image is a gray-scale pixelated image, whose pixel values are proportional to the number of electrons that arrived at the corresponding area on the image plane. For a real detector like a CCD or a CMOS camera, the resulting image is additionally blurred by the camera's MTF.

An exemplary TEM image of silicon nitride is shown in Figure 1.2.

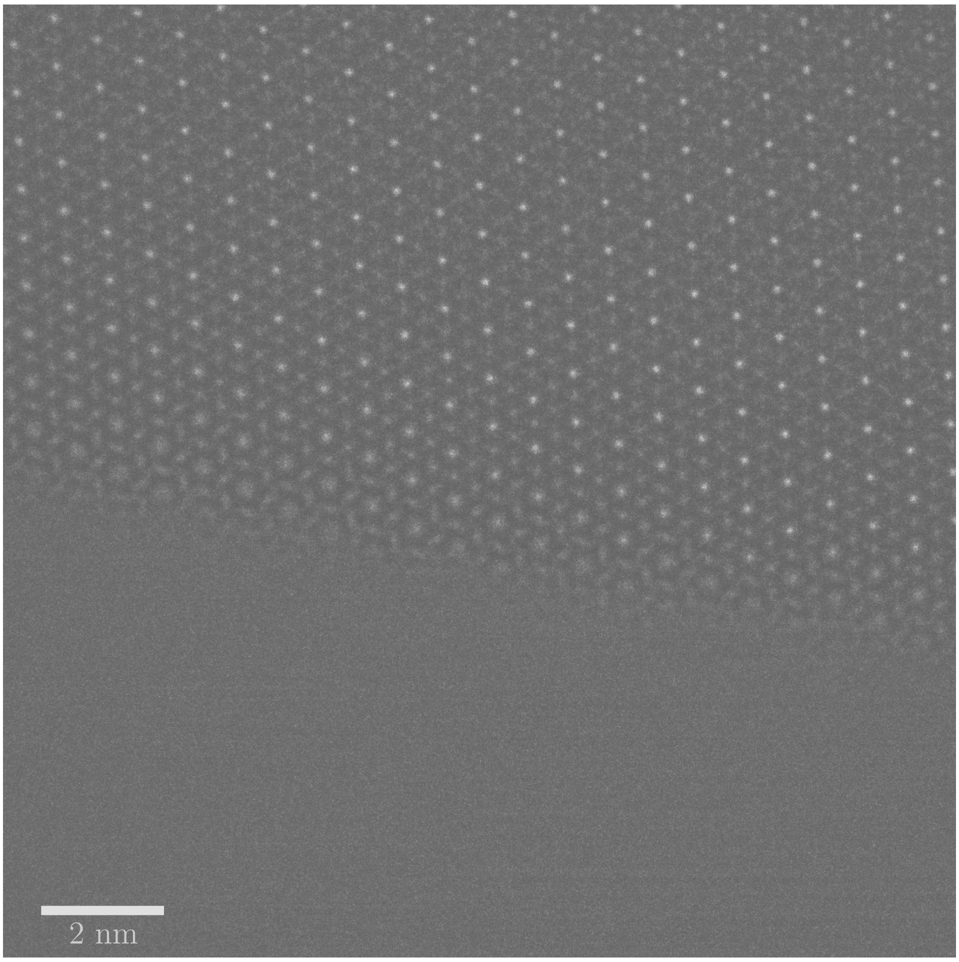

Figure 1.2: Experimental TEM image of Si_3N_4 sampled with 2048×2048 pixels. Image courtesy of Chen Huang from the Oxford Electron Image Analysis Group, University of Oxford.

1.2 Exit wave reconstruction

Although it is possible to analyse a specimen using only a single TEM image, it can be difficult to interpret the image correctly. For example, entire atomic columns may be invisible for certain combinations of aberration coefficients of the objective lens as shown in Figure 1.3. Additionally, it is possible for TEM images to contain features that are caused by the lens aberrations, which may mistakenly be interpreted as a part of the specimen. The natural approach to solve these issues is to reduce the aberrations, which can be done by improving the microscope's hardware components or by postprocessing the images.

Substantial progress has been made in improving the microscope's hardware over the past 30 years. The resolution limit of TEMs was most notably improved by the development of effective aberration correctors such as correctors for the third-order spherical aberration [28] and the chromatic aberration [34]. The use of FEGs in a TEM signif-

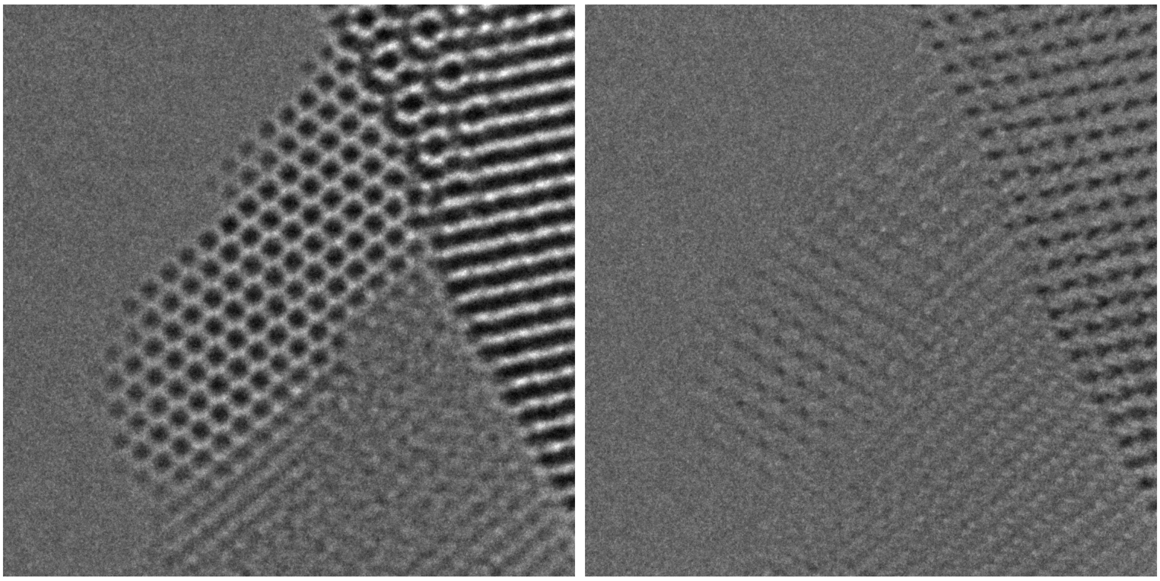

Figure 1.3: Experimental TEM images of CeO_2 depicting the same specimen at different focus values of the objective lens. A comparison of the images shows that a part of the specimen can disappear from the images for an adversely chosen focus value. Images courtesy of Chen Huang from the Oxford Electron Image Analysis Group, University of Oxford.

icantly increased the spatial and temporal coherence and consequently improved the information limit of TEMs equipped with a FEG [56].

Independently of the microscope's hardware components, it is always possible to post process the images numerically. In the following, several algorithms are presented that attempt to reconstruct the exit wave from some kind of image data. As the exit wave is the electron wave at the exit plane of the specimen right before it passes through the objective lens, it is free from aberrations by definition. Besides the removal of the aberrations, exit wave reconstruction has the additional remarkable advantages that the phase of the electron wave is reconstructed and blurring effects due to the partial temporal and spatial coherence are reduced. The existing algorithms for exit wave reconstruction can be divided into two categories: linear and non-linear reconstruction algorithms.

Linear reconstruction algorithms assume that the linear imaging approximation holds, i.e. that only the interferences between the central electron beam and the scattered electrons contribute significantly to the image contrast. The interferences between any two scattered electrons are therefore ignored, which is a valid approximation for very thin specimens. Two widely used linear reconstruction algorithms are the reconstruction with a Wiener filter (see e.g. [50, 36]) and the paraboloid method (PAM) [13], both of which are based on an algorithm originally devised by Schiske [61, 59]. These algorithms reconstruct the exit wave from a series of TEM images depicting the specimen at different focus values, a so-called focus series.

In contrast to the linear reconstruction algorithms, non-linear reconstruction algorithms also take interferences of scattered electrons into account and are therefore more generally applicable to any specimen. A well-known non-linear reconstruction algorithm is the Gerchberg-Saxton algorithm [25], which retrieves the phase of the electron wave from its amplitudes measured in the image plane and the back focal plane. The algorithm by Misell [51] is closely related to the Gerchberg-Saxton algorithm and requires a focus series instead of the electron wave's amplitude in the back focal plane. A comparatively new approach to non-linear exit wave reconstruction consists of solving a partial differential equation called the transport of intensity equation (TIE) for a focus series consisting of two or more images [6, 33].

This thesis is concerned with a further class of non-linear reconstruction algorithms, which may be termed 'variational non-linear reconstruction algorithms'. In these methods, the exit wave is iteratively reconstructed from a series of TEM images by minimizing a certain objective functional. The value of the objective functional depends on the distances of experimental and simulated TEM images, where the simulated images are calculated from a given estimate for the exit wave. In this way, the estimate for the exit wave is improved until a prescribed stopping criterion is fulfilled, in which case it is assumed that the reconstructed exit wave closely resembles the true exit wave of the specimen.

The first of these algorithms was the "multiple input maximum a-posteriori" (MIMAP) algorithm developed by Kirkland [38, 39]. Essentially, the MIMAP algorithm takes a focus image series $\left(\mathcal{I}_j^{\exp}\right)_{j=1,\dots,N}$ as input and minimizes the functional

$$J_\sigma^{\mathrm{MIMAP}}[\Psi] = \sum_{j=1}^{N} \left\| \mathcal{I}_{\Psi,Z_j}^{\mathrm{MIMAP}} - \mathcal{I}_j^{\exp} \right\|_{L^2}^2 + \sigma \left\| \Psi - \Psi_M \right\|_{L^2}^2 , \qquad \sigma > 0,$$

where $\mathcal{I}_{\Psi,Z_j}^{\mathrm{MIMAP}}$ is the forward model that consists of simulating a TEM image from the current estimate for the exit wave Ψ and the focus value Z_j of the j-th input image $\mathcal{I}_j^{\exp}$. Here, Ψ_M is an a-priori estimate of the exit wave, which can be computed from the input images for example. The MIMAP functional is minimized using a gradient-based minimization method. The maximum-likelihood (MAL) algorithm [9, 10, 66] is based on the MIMAP algorithm and differs from the MIMAP algorithm mainly in the way that the simulated images $\mathcal{I}_{\Psi,Z_j}^{\mathrm{MAL}}$ are computed, which results in a large gain in computational efficiency. In the MAL algorithm, the exit wave is reconstructed by minimizing the functional

$$E^{\mathrm{MAL}}[\Psi] = \frac{1}{N} \sum_{j=1}^{N} \left\| \mathcal{I}_{\Psi,Z_j}^{\mathrm{MAL}} - \mathcal{I}_j^{\exp} \right\|_{L^2}^2$$

using a gradient-based minimization method. A complete description of the MIMAP and MAL functionals as well as the minimization algorithms is given in Sections 3.1 and 4.1.

Although a focus series is used as input data in the original MIMAP and MAL algorithms, in principle any other microscope parameter can be varied as long as the real-valued images reveal different parts of the complex-valued exit wave. For example, tilt or combined tilt-focus series are also commonly used in exit wave reconstruction [37, 36, 29] and have their own set of advantages and disadvantages compared to a focus series. Only focus series will be considered in this thesis.

It is important to note an inevitable problem that arises when working with a series of input images. As the individual images of a focus series can not be recorded at exactly the same time, specimen drift and instrumental instabilities cause the images to be slightly shifted with respect to each other. Since the exit wave is reconstructed by combining information from all input images, an accurate registration is crucial for a successful reconstruction. In the MIMAP algorithm, the registration is updated during every iteration by applying a translation to the simulated images $\mathcal{I}_{\Psi,Z_j}^{\mathrm{MIMAP}}$ and additionally minimizing the functional with respect to this translation without taking the update of the exit wave in the same iteration into account. In contrast to this, in the MAL algorithm the images are registered by calculating the cross-correlation of each experimental image with the corresponding simulated image, which is done alternatingly with the minimization of E^{MAL}.

Reconstructing the exit wave from a series of images also has a number of other advantages and disadvantages that have not been mentioned yet. An obvious problem in recording a series of images as opposed to a single image is the radiation damage caused by the electron beam. Since all images in the series are assumed to show the same specimen and atomic structure, care must be taken not to damage the specimen excessively, which is done by using lower accelerating voltages for radiation sensitive specimens. Decreasing the accelerating voltage also decreases the signal-to-noise ratio (SNR), which in turn is partially compensated for by the fact that the reconstructed exit wave is less noisy than the individual images [65, 29].

Independently of the method that is used for aberration correction, the aberration coefficients have to be known to a high precision for a successful correction [68, 29]. However, exit wave reconstruction has the advantage that residual aberrations still present in the reconstructed exit wave can oftentimes be removed or reduced numerically, since the effect of the aberrations on the exit wave is given by a simple multiplication in Fourier space [55, 19, 67]. In contrast to this, there is no such relationship between TEM images and aberration coefficients that can similarly be inverted directly. Furthermore, exit wave reconstruction allows for the correction of aberrations of arbitrary order, which is not possible with hardware correctors.

As mentioned above, the task of exit wave reconstruction involves the inversion of the steps (3) and (4) of TEM image formation in order to retrieve the electron wave at the exit plane of the specimen. It is clear that the third step can only be partially inverted as frequencies removed by the objective aperture are not contained in any image of the series and thus can not be reconstructed. Additionally, for a sufficiently well characterized detector with known MTF, the MTF may be removed from the recorded images numerically with a deconvolution prior to further image processing [19]. In the following, it is therefore assumed that the TEM images are not affected by the MTF of the camera. In this case, the fourth step simply consists of taking the squared amplitude of the complex valued electron wave in the image plane.

1.3 The structure of this thesis

The present thesis is divided into two parts: the theory in Chapters 2 and 3 and the numerics in Chapters 4 and 5. Some of the results presented here have previously been published in the article [14]. In each of these cases a reference to the corresponding result from the article is provided.

It is fairly obvious from its definition in Equation (2.1) that the forward model of simulating a TEM image is structurally identical to an autocorrelation of the exit wave weighted by an appropriate weighting function. Although this is observed in most if not all publications in the field of TEM exit wave reconstruction, the mathematical properties of the weighted cross-correlation have not yet been investigated in the context of exit wave reconstruction. However, the properties of the weighted cross-correlation are essential not only to a complete understanding of the forward model but also to all nontrivial results regarding the inverse problem. Therefore, a multitude of results regarding the weighted cross-correlation is derived in Section 2.1, which are partially based on the corresponding results for the ordinary cross-correlation. In Section 2.2, these results are then applied to the specific weights that are used in TEM, or, in other words, to the forward model of TEM image simulation. It is emphasized that the compact notation for the weighted cross-correlation introduced in Section 2.1 plays a key role in stating these results, as the commonly used, verbose integral notation of the forward model tends to obscure its underlying mathematical structure. Furthermore, the novel notion of a $\star$-separable weight allows one in certain cases to approximate weighted cross-correlations by sums of ordinary cross-correlations, which makes it possible to utilize the convolution theorem and consequently renders many proofs viable in the first place. These results are further generalized to weight functions that are not necessarily $\star$-separable themselves, but can be uniformly approximated by $\star$-separable weights.

One of the main aims of this work is to establish the existence of minimizers for a novel objective functional, which combines the reconstruction of the exit wave and the registration of the focus series into a joint minimization approach. The starting point for the definition of the objective functional are the existing functionals for exit wave reconstruction $J_\sigma^{\mathrm{MIMAP}}$ and J^{MAL} described in Section 3.1. In order to show the existence of minimizers with the direct method of the calculus of variations, it is necessary that the objective functional is coercive and weakly lower semi-continuous. The analysis of the coercivity of the data terms of the MIMAP and MAL functionals in Section 3.2 reveals that the data term is not coercive in general. Consequently, the definition of the novel objective functional E_σ in Section 3.3 includes the nonlinear Tikhonov regularizer that is used in the MIMAP functional. Thus E_σ is coercive and the weak lower semi-continuity is established as a corollary of the results in Chapter 2, which concludes the proof of the existence of minimizers. Finally, it is shown in Section 3.4 that the objective functional is not convex in general, which also applies to the MIMAP and MAL functionals. For a strongly simplified version of the objective functional, this result is further extended to show the non-convexity on any neighborhood of a global minimizer.

The theoretical results are complemented by the practical aspects discussed in Chapter 4, which are important for a successful numerical minimization of the objective functional. At first, a brief summary of the MIMAP and MAL algorithms is given in Section 4.1. In Section 4.2, a discretization of the objective functional E_σ is deduced. This particularly includes a detailed analysis of the non-trivial connection between the discrete Fourier transform and the continuous Fourier transform, which is surprisingly often omitted in the literature. Section 4.3 provides a brief summary of the iterative, gradient-based minimization algorithms that are considered for the minimization of the objective functional. However, there are several issues with a naive minimization of the discretized objective functional that affect both the convergence speed and the correctness of the result. In Section 4.4, three of these problems are identified and appropriate solutions are proposed. On the one hand, scaling of the individual components of a joint minimization problem is frequently necessary to ensure a fast convergence to the correct minimizer, which is described in Section 4.4.2. A problem that is more specifically linked to TEM image simulation is the fact that the contrast in the simulated images wraps around the image borders. Consequently, TEM images should be simulated on an extended domain by adding an additional buffer zone to the discrete exit wave as described in Section 4.4.3. Furthermore, this buffer zone extends the domain of the discrete exit wave, which can be exploited in a remarkable way as shown in Section 4.4.4. In a nutshell, the integration domain of each summand $\left\|\mathcal{I}^{\mathrm{sim}}_{\Psi,Z_j} - \mathcal{I}^{\mathrm{exp}}_{j,t_j}\right\|^2_{L^2}$ of the data term, where $\mathcal{I}^{\mathrm{sim}}_{\Psi,Z_j}$ is the j-th simulated TEM image and $\mathcal{I}^{\mathrm{exp}}_{j,t_j}$ is the j-th experimental TEM image shifted by $t_j \in \mathbb{R}^2$, may be adjusted according to the current estimate for the image shift $t_j \in \mathbb{R}^2$. This completely eliminates errors due to image continuation and maximizes the use of the available information in the input images. Chapter 4 is concluded with a brief overview of the program for exit wave reconstruction that has been written in C++ as part of this thesis and was used to perform the numerical experiments shown here. The sourcecode of the program is available on GitHub [1].

The numerical experiments in Chapter 5 demonstrate the validity of using the objective functional to reconstruct the exit wave, but also reveal an important limitation of the approach in its current form. The first experiment in Section 5.1 shows that the scaling of individual pixels of the discrete exit wave proposed in Section 4.4.2 can decrease the number of necessary iterations of the minimization algorithm. The second experiment in Section 5.2 shows the benefit of a joint approach to exit wave reconstruction and image registration: since a successful exit wave reconstruction is only possible with a good registration of the focus series and vice versa, a joint minimization yields an improved convergence rate as compared to an alternating minimization scheme. Finally, an experimental focus series of a CeO_2 sample is used in Section 5.3 to verify that the objective functional can also be used on real input data. However, the experiment also shows that the quality of the reconstructed exit wave seemingly decreases after a high number of iterations, while the value of the objective functional continues to decrease. This also applies to the MIMAP and MAL algorithms and it is assumed that this is caused by the noise in the experimental images.

1.4 Preliminaries

This section introduces important notational conventions and contains several common definitions that are used throughout this thesis.

Let $d \in \mathbb{N}$. The set $B_r(x)$ with $r \geq 0$ and $x \in \mathbb{R}^d$ generally denotes the d-dimensional open ball of radius r centered around x with respect to the 2-norm, unless a different norm is explicitly specified. The complex conjugate of a number $z = a + \mathrm{i}b \in \mathbb{C}$ is denoted by $z^* = a - \mathrm{i}b$. A function $f : \Omega \to \mathbb{C}$ for $\Omega \subseteq \mathbb{R}^d$ is called Hermitian, if $f^*(-x) = f(x)$ for all $x \in \Omega$, and skew-Hermitian, if $f^*(-x) = -f(x)$ for all $x \in \Omega$. If $f : \Omega \subseteq \mathbb{R}^d \to \mathbb{C}$ is any function, its Hermitian part is given by

$$f^H(x) := \frac{1}{2}\big(f(x) + f^*(-x)\big) \qquad \forall\, x \in \Omega$$

and its skew-Hermitian part is

$$f^{\overline{H}}(x) := \frac{1}{2}\big(f(x) - f^*(-x)\big) \qquad \forall\, x \in \Omega.$$

The space of k times continuously differentiable and complex-valued functions on $\Omega \subseteq \mathbb{R}^d$ is denoted by $C^k(\Omega, \mathbb{C})$ for all $k \in \mathbb{N}_0$. The support of a function $f : \Omega \to \mathbb{C}$ with $\Omega \subseteq \mathbb{R}^d$ is defined as the closed support

$$\operatorname{supp}(f) := \overline{\{x \in \Omega \mid f(x) \neq 0\}} \subseteq \overline{\Omega}$$

and the subspace $C_c^k(\Omega, \mathbb{C}) \subseteq C^k(\Omega, \mathbb{C})$ of compactly supported k times continuously differentiable functions is

$$C_c^k(\Omega, \mathbb{C}) := \{f \in C^k(\Omega, \mathbb{C}) \mid \operatorname{supp}(f) \subseteq \Omega \text{ compact}\}.$$

The scalar product of vectors $x, y \in \mathbb{R}^d$ in finite dimensional real vector spaces is denoted by $x \cdot y = \sum_{j=1}^d x_j y_j$. For infinite dimensional real vector spaces, the only scalar product that is needed in the following is the L^2-scalar product defined as

$$(f, g)_{L^2} := \int_\Omega f^*(x) g(x)\, \mathrm{d}x$$

for all $f, g \in L^2(\Omega, \mathbb{C})$, where $\Omega \subseteq \mathbb{R}^d$. Integrals without explicit integration domains are always understood as integrals over the whole space $\mathbb{R}^d$, possibly by continuing the integrands with zero beyond their domain. This applies in particular to the L^p-norms for $p \in [1, \infty)$, i.e.

$$\|f\|_{L^p} = \|f\|_{L^p(\mathbb{R}^d, \mathbb{C})} = \|f\|_{L^p(\Omega, \mathbb{C})} = \left(\int_\Omega |f(x)|^p \, \mathrm{d}x\right)^{\frac{1}{p}}$$

for all $f \in L^p(\Omega, \mathbb{C})$.

The cross correlation of measurable functions $f, g : \mathbb{R}^d \to \mathbb{C}$ is defined by

$$\big(f \star g\big)(x) := \int f^*(y) g(x+y) \,\mathrm{d}y \qquad \forall\, x \in \mathbb{R}^d$$

and the convolution is defined by

$$\big(f * g\big)(x) := \int f(y) g(x-y) \,\mathrm{d}y \qquad \forall\, x \in \mathbb{R}^d$$

provided that the integrals are well-defined in the sense of Lebesgue-integrability. If $f, g : \Omega \to \mathbb{C}$ are Hermitian functions, then $f \star g = g \star f$. The same equality holds if f and g are both skew-Hermitian. If $f : \Omega \to \mathbb{C}$ is Hermitian and $g : \Omega \to \mathbb{C}$ is skew-Hermitian, then $f \star g = -g \star f$. This is used to derive the following useful identity:

Lemma 1.4.1. *Let $f, g : \Omega \to \mathbb{C}$ with $\Omega \subseteq \mathbb{R}^d$. Then*

$$f \star g + g \star f = 2\big(f^H \star g^H\big) + 2\big(f^{\overline{H}} \star g^{\overline{H}}\big)$$

Proof. Splitting f and g into a sum of their Hermitian and skew-Hermitian parts, it follows that

$$\begin{aligned} f \star g &= \left(f^H + f^{\overline{H}}\right) \star \left(g^H + g^{\overline{H}}\right) = f^H \star g^H + f^H \star g^{\overline{H}} + f^{\overline{H}} \star g^H + f^{\overline{H}} \star g^{\overline{H}} \\ &= g^H \star f^H - g^{\overline{H}} \star f^H - g^H \star f^{\overline{H}} + g^{\overline{H}} \star f^{\overline{H}} = \left(g^H - g^{\overline{H}}\right) \star f^H + \left(g^{\overline{H}} - g^H\right) \star f^{\overline{H}}. \end{aligned}$$

Consequently,

$$\begin{aligned} f \star g + g \star f &= \big(g^H - g^{\overline{H}}\big) \star f^H + \big(g^{\overline{H}} - g^H\big) \star f^{\overline{H}} + g \star f^H + g \star f^{\overline{H}} \\ &= 2\big(g^H \star f^H\big) + 2\big(g^{\overline{H}} \star f^{\overline{H}}\big). \end{aligned}$$

□

The continuous Fourier transform of integrable functions is defined as

$$\mathcal{F} : L^1(\mathbb{R}^d, \mathbb{C}) \to C(\mathbb{R}^d, \mathbb{C}), \quad f \mapsto \left(x \mapsto \mathcal{F}\big(f\big)(x) := \int_{\mathbb{R}^d} f(y) e^{-2\pi \mathrm{i} x \cdot y} \,\mathrm{d}y\right)$$

and extended to an isometry on $L^2(\mathbb{R}^d, \mathbb{C})$ in the usual way. The inverse Fourier transform is similarly defined as

$$\mathcal{F}^{-1} : L^1(\mathbb{R}^d, \mathbb{C}) \to C(\mathbb{R}^d, \mathbb{C}), \quad f \mapsto \left(x \mapsto \mathcal{F}^{-1}\big(f\big)(x) := \int_{\mathbb{R}^d} f(y) e^{2\pi \mathrm{i} x \cdot y} \,\mathrm{d}y\right).$$

One of the most useful properties of the Fourier transform is the convolution theorem, which not only applies to the convolution but also to the cross-correlation. With the above definitions of the Fourier transform and the cross-correlation resp. convolution, we have

1. $\mathcal{F}(f \star g) = \mathcal{F}(f)^* \mathcal{F}(g)$

2. $\mathcal{F}^{-1}(f \star g) = \mathcal{F}^{-1}(f)^* \mathcal{F}^{-1}(g)$
3. $\mathcal{F}(f * g) = \mathcal{F}(f)\mathcal{F}(g)$
4. $\mathcal{F}^{-1}(f * g) = \mathcal{F}^{-1}(f)\mathcal{F}^{-1}(g)$

for all $f, g \in L^1(\mathbb{R}^d, \mathbb{C}) \cup L^2(\mathbb{R}^d, \mathbb{C})$. The first two equations imply

$$\mathcal{F}\big(f \star f\big)(x) = \big|\mathcal{F}\big(f\big)(x)\big|^2 \quad \text{and} \quad \mathcal{F}^{-1}\big(f \star f\big)(x) = \big|\mathcal{F}^{-1}\big(f\big)(x)\big|^2 \qquad \forall\, x \in \mathbb{R}^d$$

for all $f \in L^1(\mathbb{R}^d, \mathbb{C}) \cup L^2(\mathbb{R}^d, \mathbb{C})$. Another useful property of the Fourier transform that is occasionally needed is given in the following lemma.

Lemma 1.4.2. *Let* $f \in L^1(\mathbb{R}^d, \mathbb{C}) \cup L^2(\mathbb{R}^d, \mathbb{C})$. *Then*

- f *is real-valued* $\iff \mathcal{F}(f)$ *is Hermitian*
- f *is imaginary-valued* $\iff \mathcal{F}(f)$ *is skew-Hermitian*

The d-dimensional discrete cross-correlation of $f, g \in \mathbb{C}^{X_1 \times \ldots \times X_d}$ with $X_1, \ldots, X_d \in \mathbb{N}$ is defined as

$$(f \star g)_{x_1,\ldots,x_d} = \sum_{a_1=0}^{X_1} \cdots \sum_{a_d=0}^{X_d} f^*_{a_1,\ldots,a_d} g_{a_1+x_1,\ldots,a_d+x_d} \qquad \forall\, x_i \in \{0, \ldots, X_i - 1\},$$

where g is extended to an element of $\mathbb{C}^{(\mathbb{Z}^d)}$ with an appropriate extension.

An element $g \in \mathbb{C}^{(\mathbb{Z}^d)}$ is called $(N_1, \ldots, N_d)$-periodic if $g_{n_1+a_1N_1,\ldots,n_d+a_dN_d} = g_n$ for all $n, a \in \mathbb{Z}^d$. The d-dimensional discrete Fourier transform of an $(N_1, \ldots, N_d)$-periodic vector $g \in \mathbb{C}^{(\mathbb{Z}^d)}$ is

$$\mathrm{DFT}_N(g)_k := \sum_{n_1=0}^{N_1-1} \cdots \sum_{n_d=0}^{N_d-1} g_{n_1,\ldots,n_d} \exp\left(-2\pi\mathrm{i} \sum_{j=1}^{d} \frac{n_j k_j}{N_j}\right) \qquad \forall\, k \in \mathbb{Z}^d.$$

and the inverse discrete Fourier transform of g is similarly defined by

$$\mathrm{IDFT}_N(g)_k := \sum_{n_1=0}^{N_1-1} \cdots \sum_{n_d=0}^{N_d-1} g_{n_1,\ldots,n_d} \exp\left(2\pi\mathrm{i} \sum_{j=1}^{d} \frac{n_j k_j}{N_j}\right) \qquad \forall\, k \in \mathbb{Z}^d.$$

It is important to note that IDFT_N as defined above is not the inverse of DFT_N, since it is missing the normalization factor $(N_1 \cdots N_d)^{-1}$, i.e.

$$\mathrm{IDFT}_N(\mathrm{DFT}_N(g))_k = N_1 \cdots N_d g_k \qquad \forall\, k \in \mathbb{Z}^d.$$

However, this definition of IDFT_N is convenient here as it is identical to the definition used in the implementation for the numerical experiment.

Only the following version of the discrete convolution theorem is needed here, which is straightforward to verify:

Lemma 1.4.3. *Let* $f, g \in \mathbb{C}^{N \times M}$ *for* $N, M \in \mathbb{N}$. *If* f *and* g *are extended to elements of* $\mathbb{C}^{(\mathbb{Z}^2)}$ *using periodic continuation, then*

$$\mathrm{DFT}(f \star g)_{x,y} = \mathrm{DFT}(f)^*_{x,y}\, \mathrm{DFT}(g)_{x,y}$$

for all $x, y \in \mathbb{Z}$.

2 The forward model: image simulation

In this chapter, a detailed mathematical description of the simulation of TEM images from a given exit wave and microscope parameters is presented. This corresponds to the steps (3) and (4) of TEM image formation as given in Chapter 1 and is one of the core components of the inverse problem of exit wave reconstruction.

Due to the partial spatial coherence of the electron beam, the specimen is illuminated from a small cone of directions instead of just a single direction. Since different illumination angles generate different exit waves, one exit wave is required for each illumination angle in general. However, if the specimen is sufficiently thin with respect to the spatial coherence of the electron source, the spread of the illumination angles can be included in the transfer function of the objective lens instead of the specimen [41]. In this case, only a single exit wave corresponding to illumination parallel to the microscope's optical axis needs to be considered. The relationship between the exit wave in Fourier space $\Psi \in L^2(\mathbb{R}^2, \mathbb{C})$ and a TEM image $\mathcal{I}_{\Psi,Z} \in L^2(\mathbb{R}^2, \mathbb{C})$ in Fourier space is then given by

$$\mathcal{I}_{\Psi,Z}(x) = \int_{\mathbb{R}^2} \Psi^*(y)\Psi(x+y)T_Z(x+y,y)\,\mathrm{d}y \qquad \forall\, x \in \mathbb{R}^2, \tag{2.1}$$

where $Z \in \mathbb{R}$ is the focus value and T_Z is the transmission cross-coefficient (TCC) [41, 7].

The TCC models the phase shifts due to aberrations of the objective lens, the objective aperture and the effects of partial spatial and temporal coherence. Note that only the dependence on the focus is made explicit for the TCC, as all other aberrations and microscope parameters are assumed to stay constant within a focus series. Several versions of the TCC have been developed, each of which put different emphasis on accuracy or computational efficiency. These are discussed in detail in Section 2.2.

Structurally, the formula for TEM image simulation given in Equation (2.1) is an autocorrelation of the exit wave Ψ weighted by the TCC T_Z. Many of the properties of the forward model as well as of the inverse problem given in Chapter 3 can be derived starting from the concept of the weighted cross-correlation introduced in the next section.

2.1 The weighted cross-correlation

The key to understanding the properties of Equation (2.1) is the generalization of the properties of the ordinary cross-correlation to the weighted cross-correlation. Many of the proofs given in this section are a straightforward generalization of well-known proofs for the ordinary cross-correlation (respectively convolution) as given in e.g. [2, 8].

The definition of the weighted cross-correlation is motivated by Equation (2.1) and a direct generalization of the ordinary cross-correlation.

Definition 2.1.1. *Let* $f, g : \mathbb{R}^d \to \mathbb{C}$ *and* $w : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{C}$ *be measurable functions. The weighted cross-correlation* $f \star_w g$ *is defined as*

$$\big(f \star_w g\big)(x) := \int_{\mathbb{R}^d} f^*(y) g(x+y) w(x+y, y) \,\mathrm{d}y$$

for all $x \in \mathbb{R}^d$ *such that the integral is well-defined in the sense of Lebesgue-integrability.*

This definition is extended to functions $f : M_1 \to \mathbb{C}$ and $g : M_2 \to \mathbb{C}$ defined on subsets $M_1, M_2 \subseteq \mathbb{R}^d$ by setting $f(x) := 0$ for all $x \in \mathbb{R}^d \backslash M_1$ and $g(x) := 0$ for all $x \in \mathbb{R}^d \backslash M_2$. It is similarly extended to weight functions $w : M_3 \times M_4 \to \mathbb{C}$ for subsets $M_3, M_4 \subseteq \mathbb{R}^d$ by zero continuation of the weight.

Definition 2.1.1 is kept as general as possible in order to allow for various function spaces and weights. However, in its current generality, the weighted cross-correlation has only very few useful properties. In the following, we therefore consider the weighted cross-correlation for Lebesgue functions f and g. This is reasonable, since the integration in Definition 2.1.1 is carried out over the entire domain of f and g and therefore the value of $\big(f \star_w g\big)(x)$ is well-defined for all $x \in \mathbb{R}^d$ (albeit not necessarily finite).

If Lebesgue functions were also considered for the weight w, then $f \star_w g$ would no longer admit point values as the integration is only carried out over a subset of the weight's domain of measure zero (namely, the diagonals $\{(x+y, y) \mid y \in \mathbb{R}^d\}$). It should be possible to consider Lebesgue functions for the weight as long as the result itself is also treated as a Lebesgue function. However, this approach constitutes an unnecessary complication in the context of TEM image simulation and therefore is not pursued here.

Instead of Lebesgue functions, certain subspaces of bounded functions allow for more useful properties of the weighted cross-correlation in the present context. For subsets $U \subseteq \mathbb{R}^d$, the space of measurable and bounded weight functions is defined as

$$W(U) := \{w : U \times U \to \mathbb{C} \mid w \text{ measurable and bounded}\}.$$

A weight $w \in W(U)$ is called $\star$*-separable*, if there exists an $N \in \mathbb{N}$ and measurable and bounded functions $v_j : U \to \mathbb{C}$ for all $j \in \{1, \dots, N\}$ such that

$$w(x, y) = \sum_{j=1}^{N} v_j(x) v_j^*(y) \qquad \forall\, x, y \in U. \tag{2.2}$$

The subspace of weight functions $w \in W(U)$ that can be uniformly approximated by $\star$-separable weights is defined as

$$W^+(U) := \Big\{ w \in W(U) \;\Big|\; \forall N \in \mathbb{N}\; \exists\, w_N \in W(U) \text{ } \star\text{-separable: } \lim_{N\to\infty} \|w - w_N\|_\infty = 0 \Big\}.$$

The definition of a $\star$-separable weight can be restated as follows: a weight $w \in W(U)$ is $\star$-separable if and only if there exists an $N \in \mathbb{N}$ and measurable and bounded functions

$\tilde{v}_j : U \to \mathbb{C}$ as well as nonnegative coefficients $c_j \geq 0$ for all $j \in \{1, \ldots, N\}$ such that

$$w(x, y) = \sum_{j=1}^{N} c_j \tilde{v}_j(x) \tilde{v}_j^*(y) \qquad \forall\, x, y \in U.$$

This follows from the fact that $c_j \tilde{v}_j(x) \tilde{v}_j^*(y) = v_j(x) v_j^*(y)$ for $v_j := \sqrt{c_j}\tilde{v}_j$. Thus the coefficients c_j must be nonnegative in order for w to be a $\star$-separable weight, which explains the superscript $^+$ in the notation for W^+. A direct application of this "nonnegativity" is given in Lemma 2.1.7 for example, where it is shown that $\mathcal{F}^{-1}(f \star_w f)$ is nonnegative for weights $w \in W^+(U)$ and few additional assumptions on f and U.

In the following, a weight $w \in W(U)$ is occasionally identified with the equivalence class $[w] \in L^\infty(U \times U, \mathbb{C})$ in order to apply Hölder's inequality. Note that both $W(U)$ and $W^+(U)$ contain the constant weight $w \equiv 1$. Therefore, all of the results in this section where $U = \mathbb{R}^d$ may be chosen as the domain are a generalization of the corresponding results for the ordinary cross-correlation.

2.1.1 Upper bounds

Lemmas 2.1.2 and 2.1.3 and Corollary 2.1.4 give basic upper bounds for the weighted cross-correlation of Lebesgue functions. These bounds essentially enable us to identify the weighted cross-correlation of Lebesgue functions again with a function in an appropriate Lebesgue space.

Lemma 2.1.2 ([14, Lemma A.2]). *If $f, g \in L^2(\mathbb{R}^d, \mathbb{C})$ and $w \in W(\mathbb{R}^d)$, then $f \star_w g$ satisfies $(f \star_w g)(x) \in \mathbb{C}$ for all $x \in \mathbb{R}^d$ and $\|f \star_w g\|_\infty \leq \|w\|_\infty \|f\|_{L^2} \|g\|_{L^2}$.*

Proof. For each $x \in \mathbb{R}^d$, define $T_x(y) := x + y$ for all $y \in \mathbb{R}^d$. By definition of the weighted cross-correlation and the Cauchy-Schwarz inequality,

$$\begin{aligned}
\big|\big(f \star_w g\big)(x)\big| &\leq \int |f^*(y) g(x+y) w(x+y, y)| \,\mathrm{d}y \\
&\leq \|w\|_\infty \int |f^*(y) g(x+y)| \,\mathrm{d}y \\
&= \|w\|_\infty \|f^* \cdot (g \circ T_x)\|_{L^1} \\
&\leq \|w\|_\infty \|f^*\|_{L^2} \|g \circ T_x\|_{L^2} \\
&= \|w\|_\infty \|f\|_{L^2} \|g\|_{L^2}
\end{aligned}$$

holds for all $x \in \mathbb{R}^d$. □

As shown above, the (weighted) cross-correlation of square-integrable functions is finite everywhere for $w \in W(\mathbb{R}^d)$. The same is not true for other Lebesgue functions in general, which can be seen with the functions

$$f_\alpha : \mathbb{R} \to \mathbb{R}, \quad x \mapsto \begin{cases} \frac{1}{x^\alpha}, & \text{if } x \in (0,1), \\ 0, & \text{otherwise.} \end{cases} \qquad (\alpha \in \mathbb{R})$$

Then $f_\alpha \in L^1(\mathbb{R},\mathbb{R}) \iff \alpha < 1$ and $f_\alpha \in L^2(\mathbb{R},\mathbb{R}) \iff \alpha < \frac{1}{2}$ and it follows that

$$\big(f_{3/4} \star f_{3/4}\big)(0) = \int_0^1 \frac{1}{y^{3/2}} \,\mathrm{d}y = \infty.$$

Similar to the ordinary cross-correlation, only the L^p norm of the result is bounded in this case, where $p \in [1,\infty)$.

Lemma 2.1.3 ([14, Lemma A.3])**.** *Let* $p \in [1,\infty)$*. If* $f \in L^1(\mathbb{R}^d,\mathbb{C})$*,* $g \in L^p(\mathbb{R}^d,\mathbb{C})$ *and* $w \in W(\mathbb{R}^d)$*, then* $f \star_w g \in L^p(\mathbb{R}^d,\mathbb{C})$ *and* $\|f \star_w g\|_{L^p} \leq \|w\|_\infty \|f\|_{L^1} \|g\|_{L^p}$*.*

Proof. Since the weight is bounded, an upper bound for the L^p-norm of $f \star_w g$ is given by

$$\begin{aligned}
\|f \star_w g\|_{L^p} &= \left(\int \big|\big(f \star_w g\big)(x)\big|^p \,\mathrm{d}x\right)^{\frac{1}{p}} \\
&= \left(\int \left|\int f^*(y)g(x+y)w(x+y,y)\,\mathrm{d}y\right|^p \mathrm{d}x\right)^{\frac{1}{p}} \\
&\leq \left(\int \left(\int |f^*(y)g(x+y)w(x+y,y)| \,\mathrm{d}y\right)^p \mathrm{d}x\right)^{\frac{1}{p}} \\
&\leq \|w\|_\infty \left(\int \left(\int |f^*(y)g(x+y)| \,\mathrm{d}y\right)^p \mathrm{d}x\right)^{\frac{1}{p}}
\end{aligned}$$

By a variant of Minkowski's inequality for integrals [30, Theorem 202],

$$\left(\int \left(\int |f^*(y)g(x+y)| \,\mathrm{d}y\right)^p \mathrm{d}x\right)^{\frac{1}{p}} \leq \int \left(\int |f^*(y)g(x+y)|^p \,\mathrm{d}x\right)^{\frac{1}{p}} \mathrm{d}y,$$

and consequently it follows that

$$\begin{aligned}
\int \left(\int |f^*(y)g(x+y)|^p \,\mathrm{d}x\right)^{\frac{1}{p}} \mathrm{d}y &= \int |f(y)| \left(\int |g(x+y)|^p \,\mathrm{d}x\right)^{\frac{1}{p}} \mathrm{d}y \\
&= \|f\|_{L^1} \|g\|_{L^p} < \infty.
\end{aligned}$$

□

An immediate consequence of Lemma 2.1.3 is that the space $L^1(\mathbb{R}^d,\mathbb{C})$ is closed under weighted cross-correlation. If the support of the weight is bounded, then the same is true for every L^p-space with $p \in [1,\infty)$ as shown next.

Corollary 2.1.4 ([14, Corollary A.4])**.** *If* $p \in [1,\infty)$ *and* $w \in W(U)$ *for a bounded subset* $U \subseteq \mathbb{R}^d$*, then* $f \star_w g \in L^p(\mathbb{R}^d,\mathbb{C})$ *for all* $f, g \in L^p(\mathbb{R}^d,\mathbb{C})$*.*

Proof. Since the support of w is bounded, there is an $r > 0$ such that

$$\operatorname{supp}(w) \subseteq B_r(0) \subseteq \mathbb{R}^d \times \mathbb{R}^d.$$

Let $A := B_r(0) \subseteq \mathbb{R}^d$. Then $w(x,y) = w(x,y)\chi_A(x)\chi_A^*(y)$ for all $x, y \in \mathbb{R}^d$ and it follows that

$$f \star_w g = (f\chi_A) \star_w (g\chi_A).$$

Since $f\chi_A \in L^p(\mathbb{R}^d,\mathbb{C})$ has bounded support, we get $f\chi_A \in L^1(\mathbb{R}^d,\mathbb{C})$ from Hölder's inequality and consequently $f \star_w g \in L^p(\mathbb{R}^d,\mathbb{C})$ by the previous lemma. ☐

As can be seen from the proof of Corollary 2.1.4, one could alternatively require the support of the functions f and g to be bounded instead of requiring U to be bounded.

2.1.2 Symmetry properties

The next lemma lists two symmetry properties of the weighted cross-correlation of square-integrable functions, which have various applications. The first property will mostly be used in the subsequent lemmas, whereas the second property is particularly useful for simplifying expressions with the weighted cross-correlation.

Lemma 2.1.5 ([14, Lemma A.5])**.** *Let $f, g \in L^2(\mathbb{R}^d,\mathbb{C})$ and $w \in W(U)$ for a subset $U \subseteq \mathbb{R}^d$ such that $w^*(x,y) = w(y,x)$ for all $x, y \in U$. Then*

$$\big(f \star_w g\big)(x) = \big(g \star_w f\big)^*(-x) \qquad \forall\, x \in \mathbb{R}^d.$$

Furthermore, if U is bounded, then

$$(h, f \star_w g)_{L^2} = (g \star_w f, h)_{L^2}$$

holds for all Hermitian functions $h \in L^2(\mathbb{R}^d,\mathbb{C})$, that is, for all $h \in L^2(\mathbb{R}^d,\mathbb{C})$ with $h(x) = h^*(-x)$ for all $x \in \mathbb{R}^d$.

Proof. The first statement follows from the definition of the weighted cross-correlation and the symmetry property of the weight with

$$\begin{aligned}
\big(g \star_w f\big)^*(-x) &= \left(\int g^*(y) f(-x+y) w(-x+y, y)\,\mathrm{d}y\right)^* \\
&= \int g(y) f^*(-x+y) w(y, -x+y)\,\mathrm{d}y \\
&= \int g(y+x) f^*(y) w(y+x, y)\,\mathrm{d}y \\
&= \big(f \star_w g\big)(x).
\end{aligned}$$

If U is bounded, then both $f \star_w g$ and $g \star_w f$ are square-integrable by Corollary 2.1.4. Using the first statement, it follows that

$$\begin{aligned}
(h, f \star_w g)_{L^2} &= \int h(x)\big(f \star_w g\big)^*(x)\,\mathrm{d}x = \int h(x)\big(g \star_w f\big)(-x)\,\mathrm{d}x \\
&= \int h(-x)\big(g \star_w f\big)(x)\,\mathrm{d}x = \int h^*(x)\big(g \star_w f\big)(x)\,\mathrm{d}x = (g \star_w f, h)_{L^2}.
\end{aligned}$$ ☐

The following corollary is a direct consequence of the fact that the weighted autocorrelation of square-integrable functions is Hermitian by Lemma 2.1.5.

Corollary 2.1.6 ([14, Lemma A.6])**.** *Let* $f \in L^2(\mathbb{R}^d, \mathbb{C})$ *and* $w \in W(U)$ *for a subset* $U \subseteq \mathbb{R}^d$ *such that* $w^*(x, y) = w(y, x)$ *for all* $x, y \in U$*. Then* $\mathcal{F}^{-1}(f \star_w f)$ *is real-valued.*

Proof. By the first symmetry property of Lemma 2.1.5, we have

$$\big(f \star_w f\big)(x) = \big(f \star_w f\big)^*(-x)$$

for all $x \in \mathbb{R}^d$. This implies that $\mathcal{F}^{-1}(f \star_w f)$ is real-valued by Lemma 1.4.2. □

Note that the proof of the above corollary makes use of the inversion of Friedel's law, which is frequently used in materials science and states that the Fourier transform of a real-valued function is Hermitian.

If one additionally requires the weight to be an element of $W^+(U)$ and U to be bounded, then a stronger result on $\mathcal{F}^{-1}(f \star_w f)$ can be established.

Lemma 2.1.7 ([14, Lemma A.6])**.** *Let* $f \in L^2(\mathbb{R}^d, \mathbb{C})$ *and* $w \in W^+(U)$ *for a bounded subset* $U \subseteq \mathbb{R}^d$ *such that* $w^*(x, y) = w(y, x)$ *for all* $x, y \in U$*. Then* $\mathcal{F}^{-1}(f \star_w f)$ *is real-valued and nonnegative.*

Proof. Since $w \in W^+(U)$, there exists a sequence $(w_N)_{N\in\mathbb{N}} \in W(U)^{\mathbb{N}}$ of $\star$-separable weights with $\lim_{N\to\infty} \|w - w_N\|_\infty = 0$. By definition, it is possible to choose measurable and bounded functions $v_{j,N} : U \to \mathbb{C}$ such that

$$w_N(x, y) = \sum_{j=1}^{N} v_{j,N}(x) v_{j,N}^*(y) \qquad \forall\, x, y \in U \quad \forall\, N \in \mathbb{N}.$$

Using the same reasoning as in the proof of Corollary 2.1.4, we may assume without loss of generality that $\operatorname{supp}(f)$ is bounded. This implies $f \in L^1(\mathbb{R}^d, \mathbb{C})$ and shows that $f \star_{w_N} f \in L^1(\mathbb{R}^d, \mathbb{C})$ for all $N \in \mathbb{N}$ by Lemma 2.1.3. Therefore, $\mathcal{F}^{-1}(f \star_{w_N} f)$ is continuous and the inequality

$$\mathcal{F}^{-1}(f \star_{w_N} f) = \sum_{j=1}^{N} \big|\mathcal{F}^{-1}(f v_{j,N})\big|^2 \geq 0 \tag{2.3}$$

holds for all $N \in \mathbb{N}$ by the linearity of the Fourier transform and the convolution theorem. Since the Fourier transform is unitary, it follows that

$$\begin{aligned}\big\|\mathcal{F}^{-1}(f \star_w f) - \mathcal{F}^{-1}(f \star_{w_N} f)\big\|_{L^2} &= \big\|f \star_{(w - w_N)} f\big\|_{L^2} \\ &\leq \|w - w_N\|_\infty \|f\|_{L^2}^2 \xrightarrow[N\to\infty]{} 0\end{aligned}$$

by Lemma 2.1.2. Thus $\mathcal{F}^{-1}(f \star_{w_N} f) \to \mathcal{F}^{-1}(f \star_w f)$ in L^2 and consequently there exists a subsequence of $\big(\mathcal{F}^{-1}(f \star_{w_N} f)\big)_{N\in\mathbb{N}}$ converging pointwise to $\mathcal{F}^{-1}(f \star_w f)$ almost everywhere. But all elements of $\big(\mathcal{F}^{-1}(f \star_{w_N} f)\big)_{N\in\mathbb{N}}$ are nonnegative by Equation (2.3), which shows that $\mathcal{F}^{-1}(f \star_w f)$ is nonnegative almost everywhere as well. Since $\mathcal{F}^{-1}(f \star_w f)$ is continuous, it follows that $\mathcal{F}^{-1}(f \star_w f) \geq 0$ everywhere. □

2.1.3 Continuity

Similar to the ordinary cross-correlation, the weighted cross-correlation of L^p- and L^q-functions is continuous for suitable weight functions if $\frac{1}{p}+\frac{1}{q}=1$.

Lemma 2.1.8 ([14, Lemma A.8])**.** *Let $U \subseteq \mathbb{R}^d$ be an open set such that the measure of its boundary ∂U is zero and $p, q \in (1,\infty)$ with $\frac{1}{p}+\frac{1}{q}=1$. Let $f \in L^p(U,\mathbb{C})$, $g \in L^q(U,\mathbb{C})$ and $w \in W(U)$. If w is continuous on $U \times U$, then $f \star_w g \in C(\mathbb{R}^d,\mathbb{C})$.*

Proof. Fix an arbitrary $x \in \mathbb{R}^d$ and let $\varepsilon > 0$. We show that $f \star_w g$ is continuous at x. Since $C_c(\mathbb{R}^d,\mathbb{C})$ is dense in $L^q(\mathbb{R}^d,\mathbb{C})$, there is a function $v \in C_c(\mathbb{R}^d,\mathbb{C})$ with $\|g-v\|_{L^q} < \varepsilon$. Define

$$\begin{aligned} G_z &: \mathbb{R}^d \to \mathbb{C}, \quad y \mapsto g(z+y)w(z+y,y),\\ V_z &: \mathbb{R}^d \to \mathbb{C}, \quad y \mapsto v(z+y)w(z+y,y) \end{aligned}$$

for all $z \in \mathbb{R}^d$. By the Hölder and Minkowski inequalities,

$$\begin{aligned} &\big|\big(f \star_w g\big)(x+h) - \big(f \star_w g\big)(x)\big| \\ &\le \int |f^*(y)g(x+h+y)w(x+h+y,y) - f^*(y)g(x+y)w(x+y,y)| \,\mathrm{d}y \\ &\le \|f\|_{L^p} \|G_{x+h} - G_x\|_{L^q} \\ &\le \|f\|_{L^p} \big(\|G_{x+h} - V_{x+h}\|_{L^q} + \|V_{x+h} - V_x\|_{L^q} + \|V_x - G_x\|_{L^q} \big) \end{aligned}$$

holds for all $h \in \mathbb{R}^d$. An upper bound for the first and third summand is given by

$$\|G_z - V_z\|_{L^q} \le \|g-v\|_{L^q} \|w\|_\infty < \varepsilon \|w\|_\infty \qquad \forall\, z \in \mathbb{R}^d.$$

By the continuity of v and the continuity of w on $U \times U$, we get for all $y \in \mathbb{R}^d$ with $y \notin \{y \in U \mid x+y \in \partial U\}$ that

$$\begin{aligned} \lim_{h\to 0} V_{x+h}(y) &= \lim_{h\to 0} v(x+h+y)w(x+h+y,y) \\ &= \begin{cases} v(x+y)w(x+y,y), & \text{if } y \in U \;\;\wedge\;\; x+y \in U,\\ 0, & \text{if } y \in U \;\;\wedge\;\; x+y \in \overline{U}^c,\\ 0, & \text{if } y \in U^c, \end{cases} \\ &= \begin{cases} V_x(y), & y \in U \;\;\wedge\;\; x+y \in U,\\ V_x(y), & y \in U \;\;\wedge\;\; x+y \in \overline{U}^c,\\ V_x(y), & y \in U^c. \end{cases} \end{aligned}$$

This implies $\lim_{h\to 0} V_{x+h}(y) = V_x(y)$ for almost all $y \in \mathbb{R}^d$, since ∂U has zero measure and thus also the set $\{y \in U \mid x+y \in \partial U\}$. Furthermore, since $|V_{x+h}(y)| \le \|w\|_\infty \|v\|_\infty < \infty$ for all $y \in \mathbb{R}^d$ and

$$\operatorname{supp}(V_{x+h}) \subseteq \overline{B_1\big(\operatorname{supp}(v) - x\big)} =: D \qquad \forall\, h \in B_1(0),$$

it follows that $\|w\|_\infty \|v\|_\infty \chi_D$ is an integrable function that dominates the function V_{x+h} for all $h \in B_1(0)$. Now the dominated convergence theorem implies

$$\lim_{h\to 0} \|V_{x+h} - V_x\|_{L^q} = 0$$

and we conclude $\lim_{h\to 0} \left|\big(f \star_w g\big)(x+h) - \big(f \star_w g\big)(x)\right| = 0$. □

In the special case of $p = q = 2$, the continuity of the weighted cross-correlation can be used to derive a characterization of L^2-functions, whose weighted autocorrelation is zero.

Lemma 2.1.9 ([14, Lemma A.9]). *Let $U \subseteq \mathbb{R}^d$ be an open and bounded set such that the measure of its boundary ∂U is zero and choose a continuous weight $w \in W(U)$. If there is a constant $c > 0$ with $w(y,y) \geq c$ for all $y \in U$, then*

$$\|f \star_w f\|_{L^2} = 0 \qquad \Longleftrightarrow \qquad \|f\|_{L^2} = 0$$

holds for all $f \in L^2(U,\mathbb{C})$.

Proof. "$\Leftarrow$": $\|f\|_{L^2} = 0$ implies $f = 0$ almost everywhere and thus $\|f \star_w f\|_{L^2} = 0$.

"$\Rightarrow$": If $\|f \star_w f\|_{L^2} = 0$, then $f \star_w f = 0$ almost everywhere. However, the function $f \star_w f$ is continuous by Lemma 2.1.8, which implies $f \star_w f = 0$. In particular,

$$0 = \big(f \star_w f\big)(0) = \int_U |f(y)|^2 \, w(y,y) \,\mathrm{d}y \geq c \, \|f\|_{L^2}^2 \geq 0$$

and thus $\|f\|_{L^2} = 0$. □

Since the Fourier transform is unitary, Lemma 2.1.9 also shows that

$$\left\|\mathcal{F}^{-1}(f \star_w f)\right\|_{L^2} = 0 \qquad \Longleftrightarrow \qquad \|f\|_{L^2} = 0.$$

Combined with Lemma 2.1.7, this gives a good insight into the possible values of $\mathcal{F}^{-1}(f \star_w f)$, provided that the conditions of both lemmas are fulfilled.

The above lemma can be slightly generalized to continuous weights $w \in W(U)$ with $\lim_{y\to x} w(y,y) = 0$ for some $x \in \partial U$. In this case no global lower bound $c > 0$ exists such that $w(y,y) \geq c$ for all $y \in U$, but it turns out that a "semi-global" lower bound is also sufficient.

In the following lemma, the distance between a point $x \in \mathbb{R}^d$ and a set $M \subseteq \mathbb{R}^d$ is defined as

$$d(x, M) := \inf_{y\in M} \|x - y\|_2 \,.$$

Lemma 2.1.10. *Let $U \subseteq \mathbb{R}^d$ be an open and bounded set such that the measure of its boundary ∂U is zero and define $U_\varepsilon := \{x \in U \mid d(x, \partial U) > \varepsilon\}$. Choose a continuous weight $w \in W(U)$. If for all $\varepsilon > 0$ there is a constant $c_\varepsilon > 0$ with $w(y,y) \geq c_\varepsilon$ for all $y \in U_\varepsilon$, then*

$$\|f \star_w f\|_{L^2} = 0 \qquad \Longleftrightarrow \qquad \|f\|_{L^2} = 0$$

holds for all $f \in L^2(U,\mathbb{C})$.

Proof. Only the second part of the proof of "⇒" is different from Lemma 2.1.9. First, observe that $w(y,y) \geq 0$ for all $y \in U$ by the condition on w. If $f \star_w f = 0$, then

$$\begin{aligned} 0 = \big(f \star_w f\big)(0) &= \int_U |f(y)|^2 \, w(y,y) \, \mathrm{d}y \\ &= \int_{U_\varepsilon} |f(y)|^2 \, w(y,y) \, \mathrm{d}y + \int_{U \setminus U_\varepsilon} |f(y)|^2 \, w(y,y) \, \mathrm{d}y \\ &\geq c_\varepsilon \, \|f\|^2_{L^2(U_\varepsilon)} + 0 \geq 0 \end{aligned}$$

holds for all $\varepsilon > 0$. This implies $\|f\|_{L^2(U_\varepsilon)} = 0$ for all $\varepsilon > 0$ and thus $\|f\|_{L^2(U)} = 0$ by the monotone convergence theorem, since $U_{\varepsilon_1} \subseteq U_{\varepsilon_2}$ for all $\varepsilon_1 > \varepsilon_2 > 0$ and $\bigcup_{\varepsilon > 0} U_\varepsilon = U$. □

2.1.4 Weak lower semi-continuity

In this section, several properties of functionals involving a weighted autocorrelation with a weight $w \in W^+(U)$ are shown as an application of the previous results. Denote by μ_t the modulation by a vector $t \in \mathbb{R}^d$ defined as $\mu_t : \mathbb{R}^d \to \mathbb{C},\ x \mapsto \exp(2\pi \mathrm{i} x \cdot t)$. Based on Lemma 2.1.11, the main result in this section is the weak lower semi-continuity of the functional $(f,t) \mapsto \|f \star_w f - \mu_t c\|^2_{L^2}$ in Proposition 2.1.12. This functional is closely related to the objective functionals for exit wave reconstruction considered in Chapter 3.

Lemma 2.1.11 ([14, Corollary A.7])**.** *Let $U \subseteq \mathbb{R}^d$ be bounded and $w \in W^+(U)$ such that $w^*(x,y) = w(y,x)$ for all $x, y \in U$. Then the functional*

$$F : L^2(U, \mathbb{C}) \to \mathbb{R}, \quad f \mapsto \|f \star_w f\|^2_{L^2}$$

is Fréchet differentiable, continuous and convex.

Proof. The first order Gâteaux differential of F at the position $f \in L^2(U,\mathbb{C})$ in the direction $g \in L^2(U,\mathbb{C})$ can easily be calculated using difference quotients and is given by

$$F'[f] : L^2(U,\mathbb{C}) \to \mathbb{R}, \quad g \mapsto 2\mathrm{Re}\big((f \star_w f, f \star_w g + g \star_w f)_{L^2}\big).$$

By the Cauchy-Schwarz inequality, Lemma 2.1.3 and Lemma 2.1.5, the Gâteaux differential is bounded by

$$\begin{aligned} |F'[f](g)| &\leq 2 \|f \star_w f\|_{L^2} \|f \star_w g + g \star_w f\|_{L^2} \leq 4 \|f \star_w f\|_{L^2} \|f \star_w g\|_{L^2} \\ &\leq 4 \|f \star_w f\|_{L^2} \|w\|_\infty \|f\|_{L^1} \|g\|_{L^2} \end{aligned}$$

for all $g \in L^2(U,\mathbb{C})$, where we have used that $f \star_w f$ is square-integrable by Corollary 2.1.4 and $f \in L^1(U,\mathbb{C})$ since U is bounded. Now note that

$$|F[f+g] - F[f] - F'[f](g)| \leq \|w\|^2_\infty \left(6 \|f\|^2_{L^2} \|g\|^2_{L^2} + 4 \|f\|_{L^2} \|g\|^3_{L^2} + \|g\|^4_{L^2}\right) \quad (2.4)$$

by repeated application of Corollary 2.1.4 and the Cauchy-Schwarz inequality. Summing up, $F'[f]$ is a bounded linear operator that additionally satisfies

$$\begin{aligned} 0 &\leq \lim_{g\to 0} \frac{|F[f+g]-F[f]-F'[f](g)|}{\|g\|_{L^2}} \\ &\overset{(2.4)}{\leq} \lim_{g\to 0} \|w\|_\infty^2 \left(6\,\|f\|_{L^2}^2\,\|g\|_{L^2} + 4\,\|f\|_{L^2}\,\|g\|_{L^2}^2 + \|g\|_{L^2}^3\right) \\ &= 0, \end{aligned}$$

which shows that F is Fréchet differentiable and particularly continuous.

The second order Gâteaux differential of F at a position $f \in L^2(U,\mathbb{C})$ in the direction $g \in L^2(U,\mathbb{C})$ is

$$\begin{aligned} \langle F''[f], g\rangle &= \|f \star_w g + g \star_w f\|_{L^2}^2 + 2\mathrm{Re}\big((f\star_w f, g\star_w g)_{L^2}\big) \\ &= \|f \star_w g + g \star_w f\|_{L^2}^2 + 2\mathrm{Re}\big((\mathcal{F}^{-1}(f\star_w f), \mathcal{F}^{-1}(g\star_w g))_{L^2}\big). \end{aligned}$$

By Lemma 2.1.7, we have $\langle F''[f], g\rangle \geq 0$ for all $f, g \in L^2(U,\mathbb{C})$, which implies that F is convex. □

Proposition 2.1.12 ([14, Proposition 4.1]). *Let $U \subseteq \mathbb{R}^d$ be bounded, $c \in L^2(U,\mathbb{C})$ and $w \in W^+(U)$ with $w^*(x,y) = w(y,x)$ for all $x, y \in U$. Then the functional*

$$F : L^2(U,\mathbb{C}) \times \mathbb{R}^d \to \mathbb{R}, \quad (f,t) \mapsto \|f \star_w f - \mu_t c\|_{L^2}^2$$

is weakly lower semi-continuous.

Proof. Split the functional into three parts as follows

$$\|f\star_w f - \mu_t c\|_{L^2}^2 = \|f\star_w f\|_{L^2}^2 + \|c\|_{L^2}^2 - 2\mathrm{Re}\big((f\star_w f, \mu_t c)_{L^2}\big). \tag{2.5}$$

The leftmost summand, $f \mapsto \|f\star_w f\|_{L^2}^2$, is continuous and convex by Lemma 2.1.11 and thus weakly lower semi-continuous in particular. The second summand is constant and therefore obviously weakly lower semi-continuous.

In order to show that the rightmost summand of Equation (2.5) is weakly lower semi-continuous, we first consider the special case where the weight w can trivially be factorized as $w(x,y) = v(x)v^*(y)$ for a bounded function $v : U \to \mathbb{C}$. In the following, we show that the functional $G[f,t] := \big((fv)\star(fv), \mu_t c\big)_{L^2}$ is weakly continuous on $L^2(U,\mathbb{C}) \times \mathbb{R}^d$ with respect to the norm

$$\|\cdot\| : L^2(U,\mathbb{C}) \times \mathbb{R}^d \to \mathbb{R}_{\geq 0}, \quad (f,t) \mapsto \sqrt{\|f\|_{L^2}^2 + \|t\|_2^2}.$$

Let $(f,t) \in L^2(U,\mathbb{C}) \times \mathbb{R}^d$ and $(f_n,t_n)_{n\in\mathbb{N}} \in (L^2(U,\mathbb{C}) \times \mathbb{R}^d)^{\mathbb{N}}$ be a weakly convergent sequence with $(f_n,t_n) \rightharpoonup (f,t)$. Since

$$\begin{aligned} &\int\int \left|(f_n v)^*(y)(f_n v)(x+y)(\mu_{t_n}c)^*(x)\right|\,\mathrm{d}y\,\mathrm{d}x \\ \leq &\int \|f_n v\|_{L^2}\,\|(f_n v)(x+\cdot)\|_{L^2}\,|c(x)|\,\mathrm{d}x \leq \|f_n\|_{L^2}^2\,\|v\|_\infty^2\,\|c\|_{L^1} < \infty, \end{aligned}$$

we can apply Fubini's theorem to change the integration order so that

$$\begin{aligned}&\big((f_n v)\star(f_n v),\mu_{t_n}c\big)_{L^2}\\ &=\int \big(f_n v\big)^*(y)\int \big(f_n v\big)(x+y)(\mu_{t_n}c)^*(x)\,\mathrm{d}x\,\mathrm{d}y\\ &=\int \big(f_n v\big)^*(y)\int f_n(x)\underbrace{v(x)(\mu_{t_n}c)^*(x-y)}_{=:h_{n,y}(x)}\,\mathrm{d}x\,\mathrm{d}y=(g_n,f_n v)_{L^2},\end{aligned}$$

where $g_n(y) := \int f_n(x)h_{n,y}(x)\,\mathrm{d}x$ for all $y \in \mathbb{R}^d$.

For $h_y(x) := v(x)(\mu_t c)^*(x-y)$, we get

$$|h_{n,y}(x)-h_y(x)| \le 2\,\|v\|_\infty\,|c(x-y)| \qquad \forall\, x,y\in\mathbb{R}^d.$$

Additionally, t_n converges pointwise to t for $n\to\infty$ since weak convergence in the product space $L^2(U,\mathbb{C})\times\mathbb{R}^d$ implies $t_n \rightharpoonup t$ in $\mathbb{R}^d$ and $\mathbb{R}^d$ is finite dimensional. Thus, $h_{n,y}$ converges pointwise to h for all $y\in\mathbb{R}^d$ and consequently $h_{n,y}\to h_y$ in L^2 for all $y\in\mathbb{R}^d$ by the dominated convergence theorem. Since $f_n \rightharpoonup f$, it follows that $g_n(y)\to g(y) := \int f(x)h_y(x)\,\mathrm{d}x$ for all $y\in\mathbb{R}^d$. Thus, $g_n\to g$ is a pointwise converging sequence.

The support of $(g_n)_{n\in\mathbb{N}}$ is uniformly bounded, since the support of f_n and c is contained in $\overline{U}$ for all $n\in\mathbb{N}$. Additionally,

$$|g_n(y)| \le \|f_n\|_{L^2}\,\|v\|_\infty\,\|c\|_{L^2} \le F\,\|v\|_\infty\,\|c\|_{L^2}$$

holds for all $n\in\mathbb{N}$ and $y\in\mathbb{R}^d$, where $F := \sup_{n\in\mathbb{N}}\|f_n\|_{L^2} < \infty$. Applying the dominated convergence theorem once more yields $g_n\to g$ in L^2.

Summing up, $g_n\to g$ in L^2 and $f_n v \rightharpoonup fv$ in L^2 implies $(g_n,f_n v)_{L^2}\to(g,fv)_{L^2}$ and consequently

$$\big((f_n v)\star(f_n v),\mu_{t_n}c\big)_{L^2} = (g_n,f_n v)_{L^2} \to (g,fv)_{L^2} = \big((fv)\star(fv),\mu_t c\big)_{L^2}.$$

Therefore, $G[f,t] = \big((fv)\star(fv),\mu_t c)_{L^2}$ is weakly continuous.

Next, this result is generalized to $H[f,t] := (f\star_w f,\mu_t c)_{L^2}$ using the approximation property of the weight $w\in W^+(U)$ with a sequence of $\star$-separable weights. By the definition of the vector space $W^+(U)$, there exist measurable and bounded functions $v_{j,N}: U\to\mathbb{C}$ such that $\lim_{N\to\infty}\|w-w_N\|_\infty = 0$ for

$$w_N(x,y) = \sum_{j=1}^{N} v_{j,N}(x)v_{j,N}^*(y) \qquad \forall\, x,y\in U.$$

By the previous results,

$$H_N : L^2(U,\mathbb{C})\times\mathbb{R}^d\to\mathbb{C},\quad (f,t)\mapsto (f\star_{w_N} f,\mu_t c)_{L^2} = \sum_{j=1}^{N}\big((fv_{j,N})\star(fv_{j,N}),\mu_t c\big)_{L^2}$$

is weakly continuous for all $N \in \mathbb{N}$.

Let $(f, t) \in L^2(U, \mathbb{C}) \times \mathbb{R}^d$ and $(f_n, t_n)_{n\in\mathbb{N}} \in (L^2(U, \mathbb{C}) \times \mathbb{R}^d)^{\mathbb{N}}$ be a weakly convergent sequence with $(f_n, t_n) \rightharpoonup (f, t)$. Then,

$$\begin{aligned} & |H[f_n, t_n] - H[f, t]| \\ = \; & |H[f_n, t_n] - H_N[f_n, t_n] + H_N[f_n, t_n] - H_N[f, t] + H_N[f, t] - H[f, t]| \\ \leq \; & |H[f_n, t_n] - H_N[f_n, t_n]| + |H_N[f_n, t_n] - H_N[f, t]| + |H_N[f, t] - H[f, t]| \end{aligned}$$

holds for all $N \in \mathbb{N}$. Let $C := \sqrt{\mathrm{Vol}(U)}$ and $F := \sup_{n\in\mathbb{N}} \|f_n\|_{L^2} < \infty$. It follows that $\|g\|_{L^1} \leq C \|g\|_{L^2}$ for all $g \in L^2(U, \mathbb{C})$ and the summands in the above equation can be estimated by

$$\begin{aligned} |H[f_n, t_n] - H_N[f_n, t_n]| &\leq 2 \left\| f_n \star_{(w-w_N)} f_n \right\|_{L^2} \|c\|_{L^2} \\ \text{(Lemma 2.1.3)} &\leq 2 \|f_n\|_{L^1} \|f_n\|_{L^2} \|w - w_N\|_\infty \|c\|_{L^2} \\ &\leq 2CF^2 \|w - w_N\|_\infty \|c\|_{L^2} \end{aligned}$$

and, similarly,

$$|H_N[f, t] - H[f, t]| \leq 2CF^2 \|w - w_N\|_\infty \|c\|_{L^2} .$$

Therefore, combined with $|H_N[f_n, t_n] - H_N[f, t]| \to 0$ for $n \to \infty$, we get

$$\lim_{n\to\infty} |H[f_n, t_n] - H[f, t]| \leq 4CF^2 \|w - w_N\|_\infty \|c\|_{L^2}$$

for all $N \in \mathbb{N}$. Since $\lim_{N\to\infty} \|w - w_N\|_\infty = 0$ by the choice of the sequence $(w_N)_{N\in\mathbb{N}}$, we can conclude $\lim_{n\to\infty} H[f_n, t_n] = H[f, t]$. Thus H is weakly continuous, which implies in particular that $(f, t) \mapsto -2\mathrm{Re}\big((f \star_w f, \mu_t c)_{L^2}\big)$ is weakly lower semi-continuous. Now the claim follows, since finite sums of weakly lower semi-continuous functions are also weakly lower semi-continuous. □

2.2 The transmission cross-coefficient

With the notation introduced in the previous section, the process of TEM image formation can now be expressed in Fourier space as

$$\mathcal{I}_{\Psi,Z} = \Psi \star_{T_Z} \Psi. \tag{2.6}$$

As before, $\Psi \in L^2(\mathbb{R}^2, \mathbb{C})$ is the Fourier space exit wave, $Z \in \mathbb{R}$ is the focus, T_Z is the TCC and $\mathcal{I}_{\Psi,Z}(x)$ is the simulated image in Fourier space. In real space, the simulated image is therefore given by $\mathcal{F}^{-1}\big(\mathcal{I}_{\Psi,Z}\big) = \mathcal{F}^{-1}\big(\Psi \star_{T_Z} \Psi\big)$.

The purpose of this section is to introduce various approximations to the (unknown) correct TCC and to give an overview of their respective properties. By the previous section on the weighted cross-correlation, the following properties are particularly useful for a TCC $T_Z : U \times U \to \mathbb{C}$ with $U \subseteq \mathbb{R}^2$:

I. $T_Z^*(v,w) = T_Z(w,v)$ for all $v,w \in U \subseteq \mathbb{R}^2$. (Lemmas 2.1.5, 2.1.7 and 2.1.11, Corollary 2.1.6, and Proposition 2.1.12.)

II. T_Z is continuous on $U \times U$, where $U \subseteq \mathbb{R}^2$ is an open set with $\operatorname{vol}(\partial U) = 0$. (Lemmas 2.1.8 to 2.1.10.)

III. a) $\exists\, c > 0 :\ T_Z(v,v) \geq c$ for all $v \in U \subseteq \mathbb{R}^2$ (Lemma 2.1.9) *or*
    b) $\forall\, \varepsilon > 0\ \exists\, c_\varepsilon > 0 :\ T_Z(v,v) \geq c_\varepsilon$ for all $v \in U_\varepsilon = \{x \in U \mid d(x,\partial U) > \varepsilon\}$ (Lemma 2.1.10).

IV. $T_Z \in W^+(U)$ for a bounded subset $U \subseteq \mathbb{R}^2$. (Corollary 2.1.4, Lemmas 2.1.7 and 2.1.11, and Proposition 2.1.12.)

Any TCC satisfying these properties is also an element of $W(U)$ due to the fourth property. If a TCC satisfying Properties I–IV is used as the weight in TEM image simulation, then all of the results from Section 2.1 can be applied to Equation (2.6).

Property I has been observed earlier by Ishizuka in [32] for the most general version of the TCC called “full TCC” below. There, it is also shown that a simulated TEM image in Fourier space is Hermitian as a consequence of the TCC’s symmetry. This in turn implies that the simulated real space images are real-valued, as one would expect from a TEM image. The proof given in [32] is identical to the proof of the first part of Lemma 2.1.5.

Throughout the following subsections, the focus value $Z \in \mathbb{R}$ is assumed to be a constant. Along with the description of the TCCs, sufficient conditions on each TCC are given, which ensure that the TCCs satisfy the properties I through IV independently of Z. These results are summarized in the concluding subsection 2.2.7. The summary includes in particular all requirements on the individual components of the TCCs that are necessary for the proofs in Sections 2.2.1 to 2.2.6.

The TCCs are given in decreasing order of accuracy and in increasing order of computational efficiency. For example, image simulation with the full TCC potentially gives the most realistic results, but demands the most computation time and is not practical due to its unknown constituents. The coherent TCC on the other hand gives the least accurate results, but it is straightforward and fast to compute TEM images using this TCC.

2.2.1 The full TCC

If the specimen is sufficiently thin with respect to the spatial coherence of the electron source, then the spread of the illumination angles can be included in the TCC. In this case, the TCC is given by

$$T_Z^{\text{full}} : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{C}, \quad (v,w) \mapsto \int_{\mathbb{R}} \int_{\mathbb{R}^2} s(u) f(\zeta) t_{Z+\zeta}(v+u) t_{Z+\zeta}^*(w+u)\, \mathrm{d}u\, \mathrm{d}\zeta, \tag{2.7}$$

where $Z \in \mathbb{R}$ is the focus [41].

The effects due to the partial spatial and partial temporal coherence are included with the probability density functions $s \in L^1(\mathbb{R}^2, \mathbb{R}_{\geq 0})$ and $f \in L^1(\mathbb{R}, \mathbb{R}_{\geq 0})$, where s is the normalized intensity distribution of the illumination (partial spatial coherence) and f is the normalized focus spread (partial temporal coherence). In practice, s and f are unknown and commonly approximated by Gaussian probability distributions (cf. Section 2.2.3). The function $t_Z : \mathbb{R}^2 \to \mathbb{C}$ is the pupil function, which can be decomposed into the pure phase transfer function p_Z and the aperture function a as

$$t_Z(v) = p_Z(v)a(v) \qquad \forall\, v \in \mathbb{R}^2. \tag{2.8}$$

The pure phase transfer function models the aberrations of the objective lens and is given by

$$p_Z : \mathbb{R}^2 \backslash \{0\} \to \mathbb{C}, \quad v \mapsto \exp\big(-2\pi \mathrm{i} \chi_Z(v)\big). \tag{2.9}$$

The complete wave aberration function χ_Z is

$$\chi_Z : \mathbb{R}^2 \backslash \{0\} \to \mathbb{R}, \quad v \mapsto \frac{1}{\lambda} \sum_{m \in \mathbb{N}} \sum_{\substack{n=0 \\ m+n \text{ even}}}^{m} \frac{1}{m} c_{m,n} \lambda^m \left\| v \right\|_2^m \cos\big(n(\arg(v) - \varphi_{m,n})\big), \tag{2.10}$$

where $(c_{m,n}, \varphi_{m,n}) \in \mathbb{R}^2$ are the coefficients of the objective lens' aberrations and $\lambda \in \mathbb{R}_{>0}$ is the electron wavelength [41, 64, 65]. In this notation, $c_{2,0} = Z$ is the focus and $c_{4,0} = C_s$ is the third-order spherical aberration. For T_Z^{full}, the wave aberration function is the only part of the TCC that ultimately depends on Z. Note that if $n = 0$, then $\varphi_{m,n}$ is arbitrary with regard to the value of the wave aberration function and the scalar value $c_{m,n}$ alone is sufficient to describe the aberration; in this case the aberration is called isotropic and otherwise anisotropic. In practice, only finitely many aberrations are considered and it is assumed from now on that only finitely many aberration coefficients are nonzero. If only focus and third-order spherical aberration are considered and all other aberration coefficients vanish, the wave aberration function reads as

$$\chi_Z(v) = \frac{1}{2} Z \lambda \left\| v \right\|_2^2 + \frac{1}{4} C_s \lambda^3 \left\| v \right\|_2^4 \qquad \forall\, v \in \mathbb{R}^2 \backslash \{0\}. \tag{2.11}$$

The aperture function in Equation (2.8) models the objective aperture, which can be done in one of two ways. Most commonly, the aperture is modeled by an ideal low-pass filter

$$a(v) = \begin{cases} 1, & \text{if } \lambda \left\| v \right\|_2 < \alpha_{\max}, \\ 0, & \text{otherwise}, \end{cases} \qquad \forall\, v \in \mathbb{R}^2, \tag{2.12}$$

where $\alpha_{\max} \in \mathbb{R}_{>0}$ is the maximum semiangle allowed by the objective aperture. Alternatively, the aperture may be modeled with a closed-form approximation to the ideal low-pass filter. An example of a suitable family of approximations is given by

$$a_\delta(v) = \frac{1}{1 + \exp\left(\frac{\lambda^2 \left\| v \right\|_2^2 - \alpha_{\max}^2}{\delta^2}\right)} \qquad \forall\, v \in \mathbb{R}^2, \tag{2.13}$$

where $\delta > 0$ and $\lim_{\delta\to 0} a_\delta(v) = a(v)$ for all $v \notin \partial B_{\alpha_{\max}/\lambda}(0)$ [18]. A smooth aperture function reduces sampling artifacts and may at the same time be more realistic than the ideal low-pass filter function in Equation (2.12) due to partial coherence and the finite thickness of real apertures [42]. Furthermore, a smooth aperture function will prove to be very useful for many theoretical results.

Property IV requires the domain of T_Z^{full} to be contained in $U \times U$, where $U \subseteq \mathbb{R}^2$ is a bounded set, but the support of the full TCC as defined in Equation (2.7) is not bounded in general. For this reason, we additionally require the probability density s and the aperture a to have bounded support. Both of these restrictions pose no limitation on the practical application of the TCC, since both s and a must have bounded support for a realistic microscope. The aperture function given in Equation (2.13) can for example be replaced with

$$\tilde{a}_\delta(v) = \begin{cases} a_\delta(v), & \text{if } \|v\|_2 < \frac{d}{2}, \\ 0, & \text{otherwise} \end{cases} \qquad (d > 0) \tag{2.14}$$

or any smooth approximation thereof with bounded support. For numerical computations, the parameter d may be chosen equal to the length of the diagonal of the TEM images in Fourier space after the discretization.

The analytic form of the wave aberration function's summands as given in Equation (2.10) is impractical for some algebraic manipulations due to the cosine factor. The following lemma utilizes a well-known geometric identity to remove the cosine factor from the summands of χ_Z.

Lemma 2.2.1. *Let $m, n \in \mathbb{N}$ with $n \leq m$ and $c \in \mathbb{R}$. Then*

$$\begin{aligned} \|v\|_2^m \cos\big(n(\arg(v) - c)\big) = \cos(nc)\, \|v\|_2^{m-n} \sum_{j=0}^{\lfloor \frac{n}{2} \rfloor} \binom{n}{2j} (-1)^j v_1^{n-2j} v_2^{2j} \\ + \sin(nc)\, \|v\|_2^{m-n} \sum_{j=0}^{\lfloor \frac{n-1}{2} \rfloor} \binom{n}{2j+1} (-1)^j v_1^{n-(2j+1)} v_2^{2j+1} \end{aligned}$$

holds for all $v \in \mathbb{R}^2 \backslash \{0\}$.

Proof. Using the multiple angle formulas

$$\begin{aligned} \cos(nx) &= \sum_{j=0}^{\lfloor \frac{n}{2} \rfloor} \binom{n}{2j} (-1)^j \cos(x)^{n-2j} \sin(x)^{2j} \qquad \forall\, n \in \mathbb{N}\, \forall\, x \in \mathbb{R}, \\ \sin(nx) &= \sum_{j=0}^{\lfloor \frac{n-1}{2} \rfloor} \binom{n}{2j+1} (-1)^j \cos(x)^{n-(2j+1)} \sin(x)^{2j+1} \qquad \forall\, n \in \mathbb{N}\, \forall\, x \in \mathbb{R} \end{aligned}$$

and the geometric identities

$$\begin{aligned} \|v\|_2 \cos(\arg(v)) &= v_1 \qquad \forall\, v \in \mathbb{R}^2\backslash\{0\}, \\ \|v\|_2 \sin(\arg(v)) &= v_2 \qquad \forall\, v \in \mathbb{R}^2\backslash\{0\}, \end{aligned}$$

the term $\|v\|_2^m \cos\big(n(\arg(v)-c)\big)$ can be expressed as

$$\begin{aligned} &\|v\|_2^m \cos\big(n\arg(v) - nc\big) \\ =\ &\|v\|_2^m \,\big(\cos(n\arg(v))\cos(nc) + \sin(n\arg(v))\sin(nc)\big) \\ =\ &\cos(nc)\,\|v\|_2^{m-n} \sum_{j=0}^{\lfloor \frac{n}{2} \rfloor} \binom{n}{2j} (-1)^j \,\|v\|_2^n \cos\big(\arg(v)\big)^{n-2j} \sin\big(\arg(v)\big)^{2j} \\ &\quad + \sin(nc)\,\|v\|_2^{m-n} \sum_{j=0}^{\lfloor \frac{n-1}{2} \rfloor} \binom{n}{2j+1} (-1)^j \,\|v\|_2^n \cos\big(\arg(v)\big)^{n-(2j+1)} \sin\big(\arg(v)\big)^{2j+1} \\ =\ &\cos(nc)\,\|v\|_2^{m-n} \sum_{j=0}^{\lfloor \frac{n}{2} \rfloor} \binom{n}{2j} (-1)^j v_1^{n-2j} v_2^{2j} \\ &\quad + \sin(nc)\,\|v\|_2^{m-n} \sum_{j=0}^{\lfloor \frac{n-1}{2} \rfloor} \binom{n}{2j+1} (-1)^j v_1^{n-(2j+1)} v_2^{2j+1}. \qquad \square \end{aligned}$$

Remark 2.2.2. The aberration coefficients $(c_{1,1}, \varphi_{1,1})$ for $(m,n) = (1,1)$ represent image shift. The corresponding term of the aberration function can be written as

$$\begin{aligned} &c_{1,1}\,\|v\|_2 \cos(\arg(v) - \varphi_{1,1}) \\ =\ &c_{1,1}\,\|v\|_2 \,\big(\cos(\arg(v))\cos(\varphi_{1,1}) + \sin(\arg(v))\sin(\varphi_{1,1})\big) \\ =\ &c_{1,1}\big(\,\|v\|_2 \cos(\arg(v))\big)\cos(\varphi_{1,1}) + c_{1,1}\big(\,\|v\|_2 \sin(\arg(v))\big)\sin(\varphi_{1,1}) \\ =\ &c_{1,1} v_1 \cos(\varphi_{1,1}) + c_{1,1} v_2 \sin(\varphi_{1,1}) \\ =\ &\underbrace{\begin{pmatrix} c_{1,1}\cos(\varphi_{1,1}) \\ c_{1,1}\sin(\varphi_{1,1}) \end{pmatrix}}_{=:c_\varphi} \cdot v. \end{aligned}$$

This yields a factor $v \mapsto e^{-2\pi \mathrm{i} c_\varphi \cdot v}$ of the pure phase transfer function in Equation (2.9). By Equation (2.7) and the formula for TEM image formation in Equation (2.1), this factor essentially acts like a modulation by c_φ on the exit wave in Fourier space, which in turn is equivalent to a shift by c_φ in real space coordinates.

For the sake of simplicity, we assume in the following that $\operatorname{supp}(s) = \overline{B_{r_s}(0)}$ and $\operatorname{supp}(a) = \overline{B_{r_a}(0)}$ for $r_s, r_a > 0$, where a denotes an aperture function as in Equation (2.12) or Equation (2.14). The circular shape of the aperture function a and the intensity distribution of the illumination s particularly simplifies the proof of Property

III below. Nonetheless, it should be possible to adapt the proofs below to other shapes such as all convex shapes with only minor changes.

Additionally, we assume that for all $\varepsilon > 0$ there are constants $s_\varepsilon, a_\varepsilon > 0$ such that $s(x) \geq s_\varepsilon$ for almost all $x \in B_{r_s-\varepsilon}(0)$ and $a(y) \geq a_\varepsilon$ for all $y \in B_{r_a-\varepsilon}(0)$.

Lemma 2.2.3. *Let* $U := B_{r_s+r_a}(0)$. *Then* $\operatorname{supp}(T_Z^{\text{full}}) \subseteq \overline{U} \times \overline{U}$.

Proof. Let $v \in \mathbb{R}^2 \backslash \overline{U}$ and $w \in \mathbb{R}^2$. For any $u \in \mathbb{R}^2$, the inequality

$$r_s + r_a < \|v\|_2 = \|v + u - u\|_2 \leq \|v + u\|_2 + \|u\|_2 \tag{2.15}$$

implies that $\|v + u\|_2 > r_a$ or $\|u\|_2 > r_s$, since otherwise the inequality in Equation (2.15) does not hold. Since $\operatorname{supp}(s) = \overline{B_{r_s}(0)}$ and $\operatorname{supp}(a) = \overline{B_{r_a}(0)}$, it follows that

$$\begin{aligned} T_Z^{\text{full}}(v, w) &= \int_{\mathbb{R}} \int_{\mathbb{R}^2} s(u) f(\zeta) a(v+u) a(w+u) p_{Z+\zeta}(v+u) p^*_{Z+\zeta}(w+u) \, \mathrm{d}u \, \mathrm{d}\zeta \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}^2} 0 \, \mathrm{d}u \, \mathrm{d}\zeta = 0. \end{aligned}$$

It can be shown similarly that $T_Z^{\text{full}}(v, w) = 0$ for all $(v, w) \in \mathbb{R}^2 \times \mathbb{R}^2 \backslash \overline{U}$. □

Due to this result, the domain of T_Z^{full} is restricted to $U \times U$ in the following, where $U := B_{r_s+r_a}(0)$. The boundary of $U \times U$ is ignored as it is a null set and therefore does not affect the outcome of evaluating $\Psi \star_{T_Z^{\text{full}}} \Psi$.

Property I: The first property, $T_Z^{\text{full}}(v, w) = T_Z^{\text{full}}(w, v)^*$ for all $v, w \in U$, is the easiest property of the full TCC to show. It follows from the fact that the probability densities s and f are real-valued and

$$T_Z^{\text{full}}(w, v)^* = \int_{\mathbb{R}} \int_{\mathbb{R}^2} s(u) f(\zeta) t^*_{Z+\zeta}(w+u) t_{Z+\zeta}(v+u) \, \mathrm{d}u \, \mathrm{d}\zeta = T_Z^{\text{full}}(v, w)$$

for all $v, w \in \mathbb{R}^2$.

Property II: The continuity of the TCC on $U \times U$ can be shown with the dominated convergence theorem. Since $|t_Z(v)| \leq 1$ for all $v \in \mathbb{R}^2$, the integrand of T_Z^{full} is dominated by the function $g \in L^1(\mathbb{R}^2 \times \mathbb{R}, \mathbb{R})$ defined as $g(u, \zeta) := s(u) f(\zeta)$ for all $u \in \mathbb{R}^2$ and all $\zeta \in \mathbb{R}$. Therefore, the continuity follows from the dominated convergence theorem and the fact that the pupil function t_Z is continuous on U.

Property III: For all $v \in \mathbb{R}^2$ the expression $T_Z^{\text{full}}(v, v)$ can be simplified as

$$\begin{aligned} T_Z^{\text{full}}(v, v) &= \int_{\mathbb{R}} \int_{\mathbb{R}^2} s(u) f(\zeta) \underbrace{|t_{Z+\zeta}(v+u)|^2}_{=a(v+u)^2} \, \mathrm{d}u \, \mathrm{d}\zeta = \int_{\mathbb{R}} f(\zeta) \int_{\mathbb{R}^2} s(u) a(v+u)^2 \, \mathrm{d}u \, \mathrm{d}\zeta \\ &= \int_{\mathbb{R}^2} s(u) a(v+u)^2 \, \mathrm{d}u = \big(s \star a^2\big)(v), \end{aligned}$$

since $\int_{\mathbb{R}} f(\zeta)\,\mathrm{d}\zeta = 1$. Since a is bounded and $\operatorname{supp}(a)$ is compact, it follows that $a^2 \in L^2(\mathbb{R}^2, \mathbb{R})$. Therefore, the function $v \mapsto T_Z^{\text{full}}(v, v)$ is continuous on $\mathbb{R}^2$ for every probability density $s \in L^1(\mathbb{R}^2, \mathbb{R}) \cap L^2(\mathbb{R}^2, \mathbb{R})$. However, $T_Z^{\text{full}}(v, v)$ is zero for all $v \notin \overline{U}$ as shown in Lemma 2.2.3. This implies that there is no $c > 0$ with $T_Z^{\text{full}}(v, v) \geq c$ for all $v \in U$. Consequently, Property IIIa does not hold in general for this version of the TCC.

As this property is only needed for Lemma 2.1.9, a possible remedy to this situation is given by its generalization Property IIIb, which is sufficient for the slightly weaker requirements of Lemma 2.1.10. For this generalization a "global" lower bound on all of U is not required. Instead, it is sufficient to find lower bounds $c_\varepsilon > 0$ such that

$$\forall\, \varepsilon > 0 \;\; \exists\, c_\varepsilon > 0 : \quad T_Z^{\text{full}}(v, v) \geq c_\varepsilon \qquad \forall\, v \in U_\varepsilon = \{x \in U \mid d(x, \partial U) > \varepsilon\}.$$

The existence of these lower bounds can be shown as follows. Figure 2.1 illustrates the notation used in the proof.

Let $\varepsilon > 0$ be arbitrary and $v \in U_\varepsilon = B_{r_s + r_a - \varepsilon}(0)$. If $V = B_{r_s}(0) \cap B_{r_a}(-v)$, then

$$T_Z^{\text{full}}(v, v) = \int_{\mathbb{R}^2} s(u) a(v+u)^2 \,\mathrm{d}u = \int_V s(u) a(v+u)^2 \,\mathrm{d}u \geq \int_{V_\delta} s(u) a(v+u)^2 \,\mathrm{d}u$$

for all $\delta > 0$, where $V_\delta = \{x \in V \mid d(x, \partial V) > \delta\}$. Since $\|v\|_2 < r_s + r_a$, the set V is non-empty, which in particular implies that V_δ is non-empty for all sufficiently small $\delta > 0$. Let $s_\delta, a_\delta > 0$ such that $s(x) \geq s_\delta$ for all $x \in B_{r_s - \delta}(0)$ and $a(y) \geq a_\delta$ for all $y \in B_{r_a - \delta}(0)$. A straightforward computation shows $V_\delta = B_{r_s - \delta}(0) \cap B_{r_a - \delta}(-v)$ and it follows that

$$T_Z^{\text{full}}(v, v) \geq \int_{B_{r_s-\delta}(0) \cap B_{r_a-\delta}(-v)} s(u) a(v+u)^2 \,\mathrm{d}u \geq s_\delta a_\delta^2 \operatorname{vol}\bigl(B_{r_s-\delta}(0) \cap B_{r_a-\delta}(-v)\bigr).$$

For any $v \in U_\varepsilon$, the volume $\operatorname{vol}\bigl(B_{r_s-\delta}(0) \cap B_{r_a-\delta}(-v)\bigr)$ is bounded below by the constant $k := \operatorname{vol}\bigl(B_{r_s-\delta}(0) \cap B_{r_a-\delta}(-\tilde{v})\bigr)$, where $\tilde{v} \in \partial U_\varepsilon$. Therefore, if δ is sufficiently small, then a positive lower limit for $T_Z^{\text{full}}(v, v)$ that is independent of the particular choice of $v \in U_\varepsilon$ is

$$T_Z^{\text{full}}(v, v) \geq s_\delta a_\delta^2 k =: c_\varepsilon > 0.$$

Property IV: By the continuity of the TCC on $U \times U$, T_Z^{full} is measurable on $U \times U$ and it is clearly also bounded, since

$$\bigl|T_Z^{\text{full}}(v, w)\bigr| \leq \int_{\mathbb{R}} \int_{\mathbb{R}^2} s(u) f(\zeta) \underbrace{\bigl|t_{Z+\zeta}(v+u) t^*_{Z+\zeta}(w+u)\bigr|}_{\leq 1} \,\mathrm{d}u \,\mathrm{d}\zeta \leq 1$$

for all $v, w \in \mathbb{R}^2$. This shows $T_Z^{\text{full}} \in W(U)$ for the bounded set $U = B_{r_s + r_a}(0) \subseteq \mathbb{R}^2$.

In order to show that the TCC can be uniformly approximated by $\star$-separable weights, the integrand of T_Z^{full} is split into the two parts

$$\begin{aligned} g &: \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}, && (u, \zeta) \mapsto s(u) f(\zeta), \\ q_{v,w} &: \mathbb{R}^2 \times \mathbb{R} \to \mathbb{C}, && (u, \zeta) \mapsto t_{Z+\zeta}(v+u) t^*_{Z+\zeta}(w+u) \end{aligned}$$

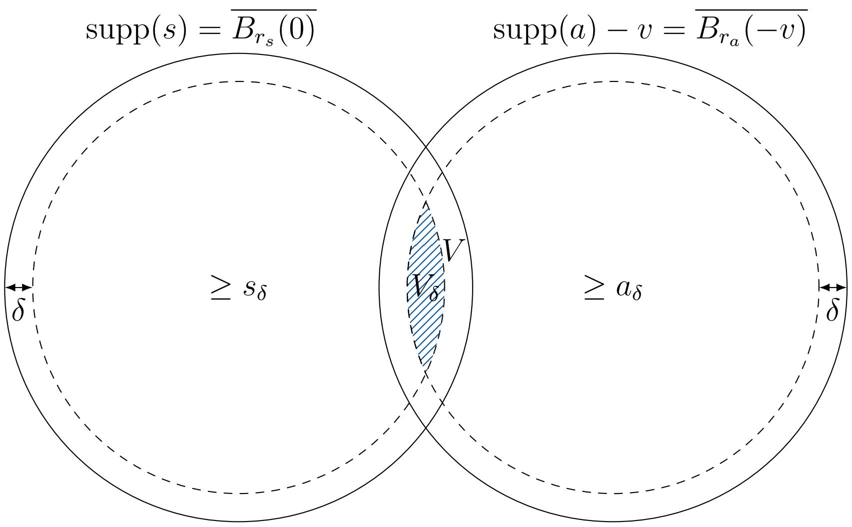

Figure 2.1: Illustration of the proof of Property IIIb.

for all $v, w \in \mathbb{R}^2$. A fundamental part of the proof of the following proposition is the Lipschitz continuity of the integrand of T_Z^{full}, which requires additional assumptions on the probability densities s and f as well as on the aperture function a. In particular, the aperture function a is required to be differentiable, which can be achieved by convolving a with a continuously differentiable mollifier. Here, the supremum norm of a vector-valued function $\alpha : \mathbb{R}^n \to \mathbb{C}^m$ is defined as

$$\|\alpha\|_\infty := \sup_{x \in \mathbb{R}^n} \|\alpha(x)\|_\infty .$$

Proposition 2.2.4 (Factorization property of the TCC)**.** *Assume that the aperture function a as well as the probability density functions s and f are differentiable with*

$$\sup_{v',w' \in \mathbb{R}^2} \|g \nabla q_{v',w'}\|_\infty < \infty \qquad \text{and} \qquad \|\nabla g\|_\infty < \infty. \tag{2.16}$$

Then there exists a sequence $(T_{Z,N})_{N \in \mathbb{N}} \in W(\mathbb{R}^2)^{\mathbb{N}}$ of functions converging uniformly to T_Z^{full} with the following properties:

(i) $T_{Z,N}$ is continuous and bounded on $U \times U$ and zero on $(\mathbb{R}^2 \times \mathbb{R}^2) \backslash (U \times U)$ for all $N \in \mathbb{N}$.

(ii) For all $N \in \mathbb{N}$, $j \in \{1, \ldots, N\}$, there are functions $h_{Z,N,j} : \mathbb{R}^2 \to \mathbb{C}$ such that $h_{Z,N,j}$ is continuous and bounded on U, zero on $\mathbb{R}^2 \backslash U$ and

$$T_{Z,N}(v, w) = \sum_{j=1}^{N} h_{Z,N,j}(v) h^*_{Z,N,j}(w) \qquad \forall\, v, w \in \mathbb{R}^2.$$

Proof. The proof is almost identical to [14, Proposition 2.3]. As the integrand of

$$T_Z^{\mathrm{full}} : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{C}, \quad (v,w) \mapsto \int_{\mathbb{R}} \int_{\mathbb{R}^2} s(u) f(\zeta) t_{Z+\zeta}(v+u) t^*_{Z+\zeta}(w+u) \,\mathrm{d}u \,\mathrm{d}\zeta$$

is continuous in (u,ζ) and already factorized with respect to v and w, we expect the Riemann sums

$$\sum_{\alpha=-M}^{M} \sum_{\beta,\gamma=-M}^{M} \delta_M \delta_M^2 s(u_{\beta,\gamma}) f(\zeta_\alpha) t_{Z+\zeta_\alpha}(v+u_{\beta,\gamma}) t^*_{Z+\zeta_\alpha}(w+u_{\beta,\gamma}),$$

- $M \in \mathbb{N}$,
- $\delta_M = M^{-k} \quad \text{for } k \in \left(\frac{3}{4}, 1\right)$,
- $u_{\beta,\gamma} = (\beta\delta_M, \gamma\delta_M) \in \mathbb{R}^2$,
- $\zeta_\alpha = \alpha\delta_M \in \mathbb{R}$

to be suitable approximations of T_Z^{full} by $\star$-separable weight functions. More precisely, if we define

$$h_{Z,M,\alpha,\beta,\gamma}(v) := \sqrt{\delta_M^3 s(u_{\beta,\gamma}) f(\zeta_\alpha)}\, t_{Z+\zeta_\alpha}(v+u_{\beta,\gamma}) \qquad \forall\, v \in \mathbb{R}^2,$$

then reordering the indices $\alpha,\beta,\gamma \in \{-M,\dots,M\}$ to an index $j \in \{1,\dots,N\}$ with $N := (2M+1)^3$ yields the sought functions $h_{Z,N,j}$. It remains to show that these sums converge uniformly to T_Z^{full} on $U \times U$ as $M \to \infty$, since the functions $h_{Z,N,j}$ are zero on $\mathbb{R}^2 \backslash U = \mathbb{R}^2 \backslash B_{r_s+r_a}(0)$.

Let $\varepsilon > 0$ and fix $v, w \in U$. First, note that the integrand

$$\tau_{v,w} : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{C}, \quad (u,\zeta) \mapsto s(u) f(\zeta) t_{Z+\zeta}(v+u) t^*_{Z+\zeta}(w+u)$$

is Lipschitz continuous. This follows from the fact that the gradient is bounded by

$$\begin{aligned} \|\nabla \tau_{v,w}\|_\infty = \|\nabla(g q_{v,w})\|_\infty &= \|g \nabla q_{v,w} + q_{v,w} \nabla g\|_\infty \le \|g \nabla q_{v,w}\|_\infty + \|\nabla g\|_\infty \\ &\le \sup_{v',w' \in \mathbb{R}^2} \|g \nabla q_{v',w'}\|_\infty + \|\nabla g\|_\infty =: L < \infty. \end{aligned}$$

Let $J_M := [-M\delta_M, (M+1)\delta_M]$ and $I_{x,M} := [x\delta_M, (x+1)\delta_M]$ for all $x \in \mathbb{Z}$ and all $M \in \mathbb{N}$.

Using the triangle inequality twice,

$$\begin{aligned}
&\left|\sum_{\alpha=-M}^{M}\sum_{\beta,\gamma=-M}^{M}\delta_M^3\tau_{v,w}(u_{\beta,\gamma},\zeta_\alpha)-\int_{\mathbb{R}}\int_{\mathbb{R}^2}\tau_{v,w}(u,\zeta)\,\mathrm{d}u\,\mathrm{d}\zeta\right| \\
&\leq \left|\sum_{\alpha=-M}^{M}\sum_{\beta,\gamma=-M}^{M}\delta_M^3\tau_{v,w}(u_{\beta,\gamma},\zeta_\alpha)-\int_{J_M}\int_{J_M\times J_M}\tau_{v,w}(u,\zeta)\,\mathrm{d}u\,\mathrm{d}\zeta\right|+C_M \\
&\leq \sum_{\alpha=-M}^{M}\sum_{\beta,\gamma=-M}^{M}\int_{I_{\alpha,M}}\int_{I_{\beta,M}\times I_{\gamma,M}}\left|\tau_{v,w}(u_{\beta,\gamma},\zeta_\alpha)-\tau_{v,w}(u,\zeta)\right|\,\mathrm{d}u\,\mathrm{d}\zeta+C_M \\
&\leq \sum_{\alpha=-M}^{M}\sum_{\beta,\gamma=-M}^{M}\int_{I_{\alpha,M}}\int_{I_{\beta,M}\times I_{\gamma,M}}3L\underbrace{\left\|(u_{\beta,\gamma},\zeta_\alpha)-(u,\zeta)\right\|_\infty}_{\leq\delta_M}\,\mathrm{d}u\,\mathrm{d}\zeta+C_M \\
&\leq (2M+1)^3\delta_M^3\cdot 3L\delta_M+C_M,
\end{aligned}\tag{2.17}$$

where $C_M=\sup_{v,w\in U}\int_{\mathbb{R}^3\setminus J_M^3}|\tau_{v,w}(x)|\,\mathrm{d}x$. The functions

$$c_M:\overline{U}\times\overline{U}\to\mathbb{R},\quad (v,w)\mapsto\int_{\mathbb{R}^3\setminus J_M^3}|\tau_{v,w}(x)|\,\mathrm{d}x\qquad(M\in\mathbb{N})$$

define a monotone sequence of continuous functions, which converges pointwise to zero on the compact space $\overline{U}\times\overline{U}$. Hence, c_M converges uniformly to zero as $M\to\infty$ by Dini's theorem, which implies $\lim_{M\to\infty}C_M=0$. The first summand in Equation (2.17) also converges to zero for $M\to\infty$, since $\delta_M=M^{-k}$ with $k\in\left(\frac{3}{4},1\right)$. This shows the uniform convergence of the Riemann sums to T_Z^{full} on $U\times U$. □

Intuitively, the requirement $\sup_{v,w\in\mathbb{R}^2}\|g\nabla q_{v,w}\|_\infty<\infty$ ensures that the probability densities s and f decrease at least as fast as the gradient of $q_{v,w}$ grows for all $v,w\in\mathbb{R}^2$. For instance, the TCC T_Z^{Ishizuka} described in Section 2.2.3 approximates the full TCC by using Gaussian probability densities for s and f. In this case, the above condition in Proposition 2.2.4 is clearly fulfilled, since $\nabla q_{v,w}$ grows only polynomially whereas g decreases exponentially. Although using a Gaussian probability density for s violates the requirement on s having bounded support, it turns out that T_Z^{Ishizuka} satisfies the Properties I–IV nonetheless. In other words, the conditions on s given in this section are sufficient for the Properties I–IV to hold, but not necessary.

2.2.2 The large aperture approximation

What is called the "large aperture approximation" here is identical to the full TCC with one key difference: the aperture function is removed from the TCC's integrand by ignoring the dependency of the aperture function on the integration variable u. The resulting TCC is

$$T_Z^{\mathrm{ap}}(v,w):=a(v)a^*(w)\int_{\mathbb{R}}\int_{\mathbb{R}^2}s(u)f(\zeta)p_{Z+\zeta}(v+u)p^*_{Z+\zeta}(w+u)\,\mathrm{d}u\,\mathrm{d}\zeta.\tag{2.18}$$

This is a common simplification that is justified if the diameter of the aperture is sufficiently large compared to the highest frequency of interest [41, 32, 24], which is explained in more detail below. In particular, this can be considered as the first step towards Ishizuka's well established approximation to the full TCC that is described in Section 2.2.3. The large aperture approximation T_Z^{ap} is the version of the TCC that was used in the previously published article [14].

In order to assess the error that is made by the simplification in Equation (2.18) qualitatively, we assume that $\operatorname{supp}(s) = \overline{B_{r_s}(0)}$ and $a = \mathbb{1}_{B_{r_a}(0)}$ for $r_a > r_s > 0$. Then,

$$\begin{aligned} T_Z^{\mathrm{full}}(v,w) &= \int_{\mathbb{R}} \int_{B_{r_s}(0)} s(u) f(\zeta) \underbrace{a(v+u)}_{=1} \underbrace{a(w+u)}_{=1} p_{Z+\zeta}(v+u) p^*_{Z+\zeta}(w+u) \,\mathrm{d}u \,\mathrm{d}\zeta \\ &= T_Z^{\mathrm{ap}}(v,w) \end{aligned}$$

for all $v, w \in B_{r_a - r_s}(0)$ and similarly

$$\begin{aligned} T_Z^{\mathrm{full}}(v,w) &= \int_{\mathbb{R}} \int_{B_{r_s}(0)} s(u) f(\zeta) \underbrace{a(v+u)}_{=0} \underbrace{a(w+u)}_{=0} p_{Z+\zeta}(v+u) p^*_{Z+\zeta}(w+u) \,\mathrm{d}u \,\mathrm{d}\zeta \\ &= T_Z^{\mathrm{ap}}(v,w) \end{aligned}$$

for all $v, w \notin B_{r_a + r_s}(0)$. It follows that $T_Z^{\mathrm{ap}}(v,w) = T_Z^{\mathrm{full}}(v,w)$ holds for all $v, w \in \mathbb{R}^2$ except for possibly some frequencies on the annulus $v, w \in B_{r_a+r_s}(0) \backslash B_{r_a - r_s}(0)$. The contributions to a frequency $x \in \mathbb{R}^2$ of the simulated image are then separated into two parts,

$$\begin{aligned} \mathcal{I}^{\mathrm{sim}}_{\Psi,Z}(x) &= \big(\Psi \star_{T_Z^{\mathrm{full}}} \Psi\big)(x) = \int_{\mathbb{R}^2} \Psi^*(y)\Psi(x+y) T_Z^{\mathrm{full}}(x+y,y)\,\mathrm{d}y \\ &= \int_{M_1(x)} \Psi^*(y)\Psi(x+y) T_Z^{\mathrm{ap}}(x+y,y)\,\mathrm{d}y + \int_{M_2(x)} \Psi^*(y)\Psi(x+y) T_Z^{\mathrm{full}}(x+y,y)\,\mathrm{d}y, \end{aligned} \tag{2.19}$$

where $M_1(x) = \{y \in \mathbb{R}^2 : \|x+y\|_2 \le r_a - r_s \ \wedge\ \|y\|_2 \le r_a - r_s\}$ and $M_2(x) = \mathbb{R}^2 \backslash M_1(x)$. By Lemma 2.2.3, the unbounded set $M_2(x)$ may be replaced by $M_2(x) \cap B_{r_s + r_a}(0)$, since $T_Z^{\mathrm{full}}(x+y,y)$ is zero for all $y \notin B_{r_s+r_a}(0)$ and thus the value of the integral on the right-hand side in Equation (2.19) is not altered. The sets $M_1(x)$ and $M_2(x) \cap B_{r_s+r_a}(0)$ are shown in Figure 2.2. If $\|x\|_2$ is small compared to the aperture radius $r_a \gg r_s$, then the left integral contributes the major part of the value $\mathcal{I}_{\Psi,Z}(x)$, i.e.

$$\begin{aligned} &\int_{M_1(x)} \Psi^*(y)\Psi(x+y) T_Z^{\mathrm{ap}}(x+y,y)\,\mathrm{d}y + \int_{M_2(x)} \Psi^*(y)\Psi(x+y) T_Z^{\mathrm{full}}(x+y,y)\,\mathrm{d}y \\ \approx &\int_{M_1(x)} \Psi^*(y)\Psi(x+y) T_Z^{\mathrm{ap}}(x+y,y)\,\mathrm{d}y + \int_{M_2(x)} \Psi^*(y)\Psi(x+y) T_Z^{\mathrm{ap}}(x+y,y)\,\mathrm{d}y \\ = &\big(\Psi \star_{T_Z^{\mathrm{ap}}} \Psi\big)(x). \end{aligned}$$

Otherwise, if $\|x\|_2$ is not small compared to the aperture radius r_a, then $M_1(x)$ is small compared to $M_2(x)$ and T_Z^{ap} is not necessarily a good approximation to T_Z^{full}.

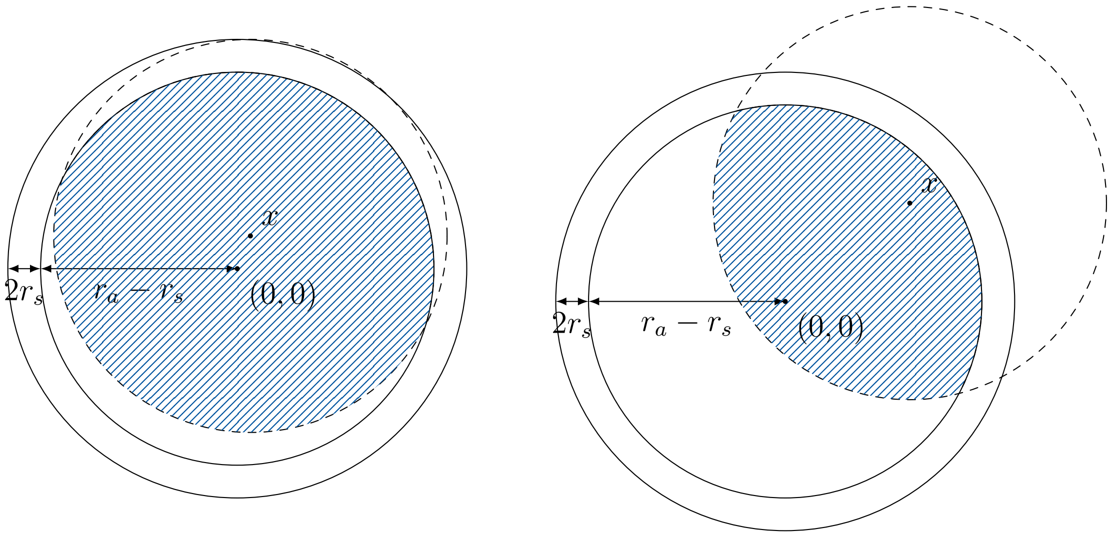

Figure 2.2: The figure on the left shows the sets $M_1(x)$ and $M_2(x) \cap B_{r_s+r_a}(0)$ for $\|x\|_2 \ll r_a$ and the figure on the right shows the same sets for $\|x\|_2 \approx r_a$. The hatched blue region is the set $M_1(x)$ and its complement within the disk of radius $r_a + r_s$ is the set $M_2(x) \cap B_{r_s+r_a}(0)$.

Properties I–IV: The obvious difference between T_Z^{full} and T_Z^{ap} with respect to the Properties I through IV is the domain of the TCC. Since the support of T_Z^{ap} is directly limited by the aperture function, we have $\operatorname{supp}(T_Z^{\text{ap}}) \subseteq \overline{U} \times \overline{U}$ with $U := B_{r_a}(0)$.

The proofs of the first two properties I and II are essentially the same as the proofs for the full TCC given in Section 2.2.1. The proof of the third property III is greatly simplified by the fact that the aperture function is no longer a part of the integrand. It follows that

$$T_Z^{\text{ap}}(v,v) = a(v)^2 \int_{\mathbb{R}} \int_{\mathbb{R}^2} s(u) f(\zeta) \underbrace{|p_{Z+\zeta}(v+u)|^2}_{=1} \,\mathrm{d}u \,\mathrm{d}\zeta = a(v)^2$$

holds for all $v \in \mathbb{R}^2$, which immediately shows that either Property IIIa or its generalization IIIb holds for T_Z^{ap}, depending on the kind of aperture function used.

The fact that T_Z^{ap} is measurable and bounded can also be shown in the same way as for the full TCC and the proof of $T_Z^{\text{ap}} \in W^+(U)$ is very similar to the proof of Proposition 2.2.4. However, since the aperture function is not a part of the integrand of T_Z^{ap}, it is no longer necessary to require a to be differentiable. Using the same notation as in Proposition 2.2.4, the aperture function is left out of the proof and the integrand is instead split into the two parts

$$\begin{aligned} g &: \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}, \quad (u,\zeta) \mapsto s(u) f(\zeta), \\ q_{v,w} &: \mathbb{R}^2 \times \mathbb{R} \to \mathbb{C}, \quad (u,\zeta) \mapsto p_{Z+\zeta}(v+u) p^*_{Z+\zeta}(w+u) \end{aligned}$$

such that $T_Z^{\text{ap}}(v,w) = a(v)a^*(w) \int\int g(u,\zeta) q_{v,w}(u,\zeta) \,\mathrm{d}u \,\mathrm{d}\zeta$. The TCC is then approximated by

$$T^{\mathrm{ap}}_{Z,M}(v,w) = \sum_{\alpha,\beta,\gamma=-M}^{M} h_{Z,M,\alpha,\beta,\gamma}(v)h^*_{Z,M,\alpha,\beta,\gamma}(w) \qquad \forall\, v,w \in \mathbb{R}^2,$$

where

$$h_{Z,M,\alpha,\beta,\gamma}(v) := a(v)\sqrt{\delta_M^3 s(u_{\beta,\gamma})f(\zeta_\alpha)}p_{Z+\zeta_\alpha}(v+u_{\beta,\gamma}) \qquad \forall\, v \in \mathbb{R}^2.$$

A complete proof of Property IV for the large aperture approximation T^{ap}_Z can be found in [14].

2.2.3 Ishizuka's TCC

Both the full TCC and the large aperture approximation described above can not be used in practice without replacing s and f with explicit probability densities. In [32], Ishizuka starts from the full TCC in Section 2.2.1, applies the large aperture approximation from Section 2.2.2 and then substitutes Gaussian probability densities for s and f in order to be able to evaluate the integrals in Equation (2.18) analytically. This TCC is used in the MIMAP algorithm as described in [38]. A derivation of Ishizuka's TCC along the lines of [32] is given in the following.

Derivation of Ishizuka's TCC: Neglecting terms of higher than first order in u, the wave aberration function $\chi_{Z+\zeta}(v+u)$ is Taylor-expanded at (Z,v) and approximated by

$$\chi_{Z+\zeta}(v+u) \approx C_1^{Z,v} + u\cdot C_2^{Z,v} + \zeta C_3^{Z,v} + \zeta u \cdot C_4^{Z,v} \tag{2.20}$$

for suitable constants $C_1^{Z,v}, C_3^{Z,v} \in \mathbb{R}$ and $C_2^{Z,v}, C_4^{Z,v} \in \mathbb{R}^2$ independent of u and ζ. Plugging this into the pure phase transfer function results in

$$\begin{aligned} T^{\mathrm{ap}}_Z(v,w) &= a(v)a(w)\int_{\mathbb{R}}\int_{\mathbb{R}^2} s(u)f(\zeta)\exp\big(-2\pi\mathrm{i}(\chi_{Z+\zeta}(v+u)-\chi_{Z+\zeta}(w+u))\big)\,\mathrm{d}u\,\mathrm{d}\zeta \\ &\approx a(v)a(w)\int_{\mathbb{R}}\int_{\mathbb{R}^2} s(u)f(\zeta)\exp\big(-2\pi\mathrm{i}(D_1^{Z,v,w}+u\cdot D_2^{Z,v,w}+\zeta D_3^{Z,v,w}+\zeta u\cdot D_4^{Z,v,w})\big)\,\mathrm{d}u\,\mathrm{d}\zeta \\ &= a(v)a(w)e^{-2\pi\mathrm{i}D_1^{Z,v,w}}\int_{\mathbb{R}} f(\zeta)e^{-2\pi\mathrm{i}\zeta D_3^{Z,v,w}}\int_{\mathbb{R}^2} s(u)e^{-2\pi\mathrm{i}u\cdot(D_2^{Z,v,w}+\zeta D_4^{Z,v,w})}\,\mathrm{d}u\,\mathrm{d}\zeta \end{aligned}$$

for all $v,w \in B_{r_a}(0)$, where $D_i^{Z,v,w} = C_i^{Z,v} - C_i^{Z,w}$ for $i=1,\dots,4$. For the sake of readability, we drop the dependency of $D_i^{Z,v,w}$ on Z, v and w and write D_i for short. Choosing normalized one- and two-dimensional Gaussian functions for s and f, i.e.

$$s(u) = \frac{1}{\pi q_0^2}\exp\left(-\frac{\|u\|_2^2}{q_0^2}\right) \qquad \forall\, u \in \mathbb{R}^2 \qquad (q_0 > 0), \tag{2.21}$$

$$f(\zeta) = \frac{1}{\sqrt{\pi\Delta^2}}\exp\left(-\frac{\zeta^2}{\Delta^2}\right) \qquad \forall\, \zeta \in \mathbb{R} \qquad (\Delta > 0), \tag{2.22}$$

the above integrals can be evaluated analytically: since $e^{-C^2} = \int_{\mathbb{R}} \frac{1}{\sqrt{\pi}} e^{-2\mathrm{i}Cx} e^{-x^2}\,\mathrm{d}x$ holds for all $C \in \mathbb{R}$ (see Lemma 2.2.6), it follows that

$$\int_{\mathbb{R}^2} s(u) e^{-2\pi\mathrm{i}u\cdot(D_2+\zeta D_4)}\,\mathrm{d}u\,\mathrm{d}\zeta = \frac{1}{\pi q_0^2}\prod_{j=1}^{2}\int_{\mathbb{R}} \exp\left(-\frac{u_j^2}{q_0^2}\right) e^{-2\pi\mathrm{i}u_j(D_2+\zeta D_4)_j}\,\mathrm{d}u_j$$

$$= \prod_{j=1}^{2}\int_{\mathbb{R}} \frac{1}{\sqrt{\pi}} e^{-u_j^2} e^{-2\pi\mathrm{i}q_0 u_j(D_2+\zeta D_4)_j}\,\mathrm{d}u_j = \prod_{j=1}^{2} e^{-\pi^2 q_0^2 (D_2+\zeta D_4)_j^2} = e^{-\pi^2 q_0^2 \|D_2+\zeta D_4\|_2^2}.$$

Continuing, we get

$$\int_{\mathbb{R}} f(\zeta) e^{-2\pi\mathrm{i}\zeta D_3} e^{-\pi^2 q_0^2\|D_2+\zeta D_4\|_2^2}\,\mathrm{d}\zeta$$

$$= e^{-\pi^2 q_0^2\|D_2\|_2^2} \int_{\mathbb{R}} f(\zeta) e^{-2\pi\mathrm{i}\zeta D_3} e^{-\pi^2 q_0^2\left(2\zeta D_2\cdot D_4 + \zeta^2\|D_4\|_2^2\right)}\,\mathrm{d}\zeta$$

$$= e^{-\pi^2 q_0^2\|D_2\|_2^2} \frac{1}{\sqrt{\pi}\Delta} \int_{\mathbb{R}} e^{-2\pi\mathrm{i}\zeta D_3} \exp\left(-\left(\frac{1}{\Delta^2} + \pi^2 q_0^2 \|D_4\|_2^2\right)\zeta^2 - \left(2\pi^2 q_0^2 D_2\cdot D_4\right)\zeta\right)\mathrm{d}\zeta.$$

Let $K := \sqrt{\frac{1}{\Delta^2} + \pi^2 q_0^2 \|D_4\|_2^2}$ and substitute $\zeta \to \frac{1}{K}\zeta$. Then the above is equal to

$$e^{-\pi^2 q_0^2\|D_2\|_2^2} \frac{1}{\sqrt{\pi}\Delta}\frac{1}{K} \int_{\mathbb{R}} e^{-2\pi\mathrm{i}\frac{1}{K}\zeta D_3} \exp\left(-\left(\zeta^2 + \frac{2\pi^2 q_0^2 D_2\cdot D_4}{K}\zeta\right)\right)\mathrm{d}\zeta$$

and completing the square with $\widetilde{K} := \frac{\pi^2 q_0^2 D_2\cdot D_4}{K}$ yields

$$e^{-\pi^2 q_0^2\|D_2\|_2^2} \frac{1}{\sqrt{\pi}\Delta}\frac{1}{K} \int_{\mathbb{R}} e^{-2\pi\mathrm{i}\frac{D_3}{K}\zeta} \exp\left(-\left(\zeta^2 + 2\widetilde{K}\zeta + \widetilde{K}^2 - \widetilde{K}^2\right)\right)\mathrm{d}\zeta$$

$$= e^{-\pi^2 q_0^2\|D_2\|_2^2} \frac{1}{\sqrt{\pi}\Delta}\frac{1}{K} e^{\widetilde{K}^2} \int_{\mathbb{R}} e^{-2\pi\mathrm{i}\frac{D_3}{K}\zeta} \exp\left(-\left(\zeta + \widetilde{K}\right)^2\right)\mathrm{d}\zeta$$

$$= e^{-\pi^2 q_0^2\|D_2\|_2^2} \frac{1}{\sqrt{\pi}\Delta}\frac{1}{K} e^{\widetilde{K}^2} e^{2\pi\mathrm{i}\frac{D_3}{K}\widetilde{K}} \int_{\mathbb{R}} e^{-2\pi\mathrm{i}\frac{D_3}{K}\zeta} e^{-\zeta^2}\,\mathrm{d}\zeta$$

$$= e^{-\pi^2 q_0^2\|D_2\|_2^2} \frac{1}{\Delta}\frac{1}{K} e^{\widetilde{K}^2} e^{2\pi\mathrm{i}\frac{D_3}{K}\widetilde{K}} \exp\left(-\left(\frac{\pi D_3}{K}\right)^2\right)$$

$$= \underbrace{e^{-\pi^2 q_0^2\|D_2\|_2^2}}_{=:E_s(v,w)} \underbrace{\frac{1}{\sqrt{u}} e^{\frac{1}{u}\pi^4 q_0^4 \Delta^2 (D_2\cdot D_4)^2}}_{=:E_x(v,w)} \underbrace{e^{\frac{2}{u}\mathrm{i}\Delta^2\pi^3 q_0^2 D_3 (D_2\cdot D_4)}}_{=:P(v,w)} \underbrace{e^{-\frac{1}{u}\pi^2\Delta^2 D_3^2}}_{=:E_f(v,w)}, \tag{2.23}$$

where $u := (\Delta K)^2 = 1 + \Delta^2\pi^2 q_0^2 \|D_4\|_2^2$. If a high resolution TEM is used to investigate a weakly scattering object, then E_x, P and u may be approximated by unity [32]. The constants $C_i^{Z,v}$ from the Taylor expansion in Equation (2.20) are

$$\begin{aligned} C_1^{Z,v} &= \chi_Z(v), \\ C_2^{Z,v} &= \nabla\chi_Z(v), \\ C_3^{Z,v} &= \frac{1}{2}\lambda\, \|v\|_2^2 \quad \text{and} \\ C_4^{Z,v} &= \lambda v \end{aligned}$$

for all $v \in \mathbb{R}^2$.

Summing up, the resulting approximation of the full TCC is

$$\begin{aligned} T_Z^{\text{Ishizuka}}(v,w) &:= a(v)a^*(w)p_Z(v)p_Z^*(w)E_s^Z(v,w)E_f(v,w) \\ &= t_Z(v)t_Z^*(w)E_s^Z(v,w)E_f(v,w) \qquad\qquad \forall\, v,w \in \mathbb{R}^2 \end{aligned} \tag{2.24}$$

with

$$E_s^Z(v,w) = \exp\left(-\left(\frac{\pi\alpha}{\lambda}\right)^2 \|\nabla\chi_Z(v) - \nabla\chi_Z(w)\|_2^2\right) \in [0,1], \tag{2.25}$$

$$E_f(v,w) = \exp\left(-\frac{1}{4}(\pi\Delta\lambda)^2 \left(\|v\|_2^2 - \|w\|_2^2\right)^2\right) \in [0,1] \tag{2.26}$$

for all $v, w \in \mathbb{R}^2$. Here, $\alpha := q_0\lambda$ is approximately equal to the half angle of beam convergence and Δ is also called the focal spread parameter.

Both E_s^Z and E_f act as damping functions on the TCC since their range is $[0,1]$. The first function, E_s^Z, approximates the effects due to the partial spatial coherence and is also referred to as the spatial coherence envelope. The function E_f on the other hand approximates the effects due to the partial temporal coherence and is also called the temporal coherence envelope. The gradients in the spatial coherence envelope can easily be calculated using Lemma 2.2.1. If the wave aberration function from Equation (2.11) is used, then

$$\nabla\chi_Z(v) = \begin{pmatrix} Z\lambda v_1 + C_s\lambda^3 v_1^3 + C_s\lambda^3 v_1 v_2^2 \\ Z\lambda v_2 + C_s\lambda^3 v_2^3 + C_s\lambda^3 v_1^2 v_2 \end{pmatrix} \qquad \forall\, v \in \mathbb{R}^2. \tag{2.27}$$

Remark 2.2.5. In [35, p. 16] it is noted that although using a Gaussian probability density for the temporal coherence is not strictly correct, it is still a reasonable choice as the "width" of the probability density is more important than its precise shape. Here, the width of a probability density $f : \mathbb{R} \to \mathbb{R}_{\geq 0}$ stands for the width of the corresponding confidence interval at a given confidence level, which is derived from the temporal coherence of the microscope. As seen above, using a Gaussian probability density has the advantage that the calculations can be performed analytically. Nonetheless, the theory in Sections 2.2.1 and 2.2.2 is general enough to encompass other probability densities in the full TCC or the large aperture approximation as well.

Let $r_a > 0$ such that $\operatorname{supp}(a) = \overline{B_{r_a}(0)}$. Similar to the large aperture approximation in the previous section, the support of T_Z^{Ishizuka} is given by $\operatorname{supp}\left(T_Z^{\text{Ishizuka}}\right) = \overline{U} \times \overline{U}$ with $U := B_{r_a}(0)$. Furthermore, we continue to assume that for all $\varepsilon > 0$ there is a constant $a_\varepsilon > 0$ such that $a(x) \geq a_\varepsilon$ for all $x \in B_{r_a - \varepsilon}(0)$.

Properties I–III: The first property holds, since

$$T_Z^{\text{Ishizuka}}(w,v)^* = a^*(w)a(v)p_Z^*(w)p_Z(v)\underbrace{E_s^Z(w,v)}_{=E_s^Z(v,w)}\underbrace{E_f(w,v)}_{=E_f(v,w)} = T_Z^{\text{Ishizuka}}(v,w)$$

is valid for all $v, w \in \mathbb{R}^2$. Property II, i.e. the continuity of T_Z^{Ishizuka} on $U \times U$, follows directly from the definition in Equation (2.24). The equation

$$T_Z^{\text{Ishizuka}}(v,v) = a(v)^2\underbrace{|p_Z(v)|^2}_{=1}\underbrace{E_s^Z(v,v)}_{=1}\underbrace{E_f(v,v)}_{=1} = a(v)^2$$

shows that Property IIIa or Property IIIb holds depending on the particular kind of aperture function that is used.

Property IV: The TCC is clearly measurable and bounded with $T_Z^{\text{Ishizuka}}(v,w) \leq 1$ for all $v, w \in \mathbb{R}^2$, which implies $T_Z^{\text{Ishizuka}} \in W(U)$. One way to show that the approximation property holds for T_Z^{Ishizuka} involves, in a certain sense, reversing the analytical integration in the derivation of the TCC and then approximating the integral representation with Riemann sums as before. The reversal of the analytical integration can be achieved by means of the following lemma.

Lemma 2.2.6. *For every $C \in \mathbb{R}$,*

$$\exp(-C^2) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} e^{-2\mathrm{i}Cx} e^{-x^2}\,\mathrm{d}x.$$

Proof. The integral expression on the right-hand side of the equation can be rewritten as the one-dimensional Fourier transform

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} e^{-2\mathrm{i}Cx} e^{-x^2}\,\mathrm{d}x = \int_{-\infty}^{\infty} \sqrt{\pi} e^{-\pi^2 y^2} e^{-2\pi\mathrm{i}Cy}\,\mathrm{d}y = \sqrt{\pi}\big(\mathcal{F}f\big)(C)$$

of the Gaussian function $f(y) := e^{-\pi^2 y^2}$. The Fourier transform of f is (see e.g. [17])

$$\big(\mathcal{F}f\big)(w) = \sqrt{\frac{\pi}{\pi^2}} e^{-\frac{(\pi w)^2}{\pi^2}} = \frac{1}{\sqrt{\pi}} e^{-w^2} \qquad \forall\, w \in \mathbb{R}$$

and it follows that

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} e^{-2\mathrm{i}Cx} e^{-x^2}\,\mathrm{d}x = \sqrt{\pi}\big(\mathcal{F}f\big)(C) = e^{-C^2}. \qquad \square$$

The spatial and temporal coherence envelopes can thus be expressed as

$$\begin{aligned} E_s^Z(v,w) &= \exp\left(-\left(\frac{\pi\alpha}{\lambda}\right)^2 \|\nabla\chi_Z(v) - \nabla\chi_Z(w)\|_2^2\right) \\ &= \prod_{k=1}^{2} \exp\left(-\left(\frac{\pi\alpha}{\lambda}(\partial_k\chi_Z(v) - \partial_k\chi_Z(w))\right)^2\right) \\ &= \prod_{k=1}^{2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} \exp\Big(-2\mathrm{i}\underbrace{\frac{\pi\alpha}{\lambda}\big(\partial_k\chi_Z(v) - \partial_k\chi_Z(w)\big)}_{=:C_s^{Z,k}(v,w)} x\Big) \exp\left(-x^2\right)\,\mathrm{d}x \end{aligned} \tag{2.28}$$

and

$$\begin{aligned} E_f(v,w) &= \exp\left(-\left(\frac{\pi\Delta\lambda}{2}\big(\,\|v\|_2^2 - \|w\|_2^2\,\big)\right)^2\right) \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} \exp\Big(-2\mathrm{i}\underbrace{\frac{\pi\Delta\lambda}{2}\big(\,\|v\|_2^2 - \|w\|_2^2\,\big)}_{=:C_f(v,w)}\,x\Big)\exp(-x^2)\,\mathrm{d}x \end{aligned} \tag{2.29}$$

for all $v, w \in \mathbb{R}^2$.

The major advantage of expressing the envelope functions in this way is the fact that $C_s^{Z,k}$ respectively C_f is not squared anymore. This results in an integrand that can be factorized trivially with respect to v and w in contrast to the non-integral representations of E_s^Z and E_f in Equations (2.25) and (2.26), which can not be factorized directly.

The proof of the uniform convergence of Riemann sums approximating the three integrals in Equations (2.28) and (2.29) proceeds essentially in the same way as the proof of Proposition 2.2.4. First, the Lipschitz continuity of the integrands is established, which is then used to compute a uniform upper bound on the difference of the Riemann sums and the integrals. However, unlike the proof of Proposition 2.2.4, the uniform convergence is shown for each of the three integrals individually at first.

Lemma 2.2.7. *Let $C : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ be a bounded function and*

$$f_{v,w}(x) := \frac{1}{\sqrt{\pi}} \exp\big(-2\mathrm{i}C(v,w)x\big)\exp(-x^2) \qquad \forall\, x \in \mathbb{R}\ \ \forall\, v,w \in \mathbb{R}^2.$$

There exists a Lipschitz constant $L > 0$ independent of $v, w \in \mathbb{R}^2$ with

$$|f_{v,w}(x) - f_{v,w}(y)| \le L\,|x-y| \qquad \forall\, x,y \in \mathbb{R}$$

for all $v, w \in \mathbb{R}^2$.

Proof. If the first exponential function is split into a cosine and a sine term, then

$$\begin{aligned} &|f_{v,w}(x) - f_{v,w}(y)| \\ &= \left|\frac{1}{\sqrt{\pi}}\exp\big(-2\mathrm{i}C(v,w)x\big)\exp(-x^2) - \frac{1}{\sqrt{\pi}}\exp\big(-2\mathrm{i}C(v,w)y\big)\exp(-y^2)\right| \\ &= \frac{1}{\sqrt{\pi}}\bigg|\Big(\cos\big(2C(v,w)x\big) - \mathrm{i}\sin\big(2C(v,w)x\big)\Big)e^{-x^2} \\ &\qquad - \Big(\cos\big(2C(v,w)y\big) - \mathrm{i}\sin\big(2C(v,w)y\big)\Big)e^{-y^2}\bigg| \\ &\le \frac{1}{\sqrt{\pi}}\left|\cos\big(2C(v,w)x\big)e^{-x^2} - \cos\big(2C(v,w)y\big)e^{-y^2}\right| \\ &\qquad + \frac{1}{\sqrt{\pi}}\left|\sin\big(2C(v,w)x\big)e^{-x^2} - \sin\big(2C(v,w)y\big)e^{-y^2}\right| \end{aligned}$$

for all $v, w \in \mathbb{R}^2$ and all $x, y \in \mathbb{R}$. Since C is bounded, there are constants $L_{\cos}$ and $L_{\sin}$ independent of v, w such that the above sum is bounded by

$$|f_{v,w}(x) - f_{v,w}(y)| \leq \frac{1}{\sqrt{\pi}} L_{\cos} |x-y| + \frac{1}{\sqrt{\pi}} L_{\sin} |x-y| = \frac{1}{\sqrt{\pi}} (L_{\cos} + L_{\sin}) |x-y|$$

and therefore $L := \frac{1}{\sqrt{\pi}}(L_{\cos} + L_{\sin})$ is the desired Lipschitz constant. More explicitly, since

$$\begin{aligned}\left|\frac{d}{dx}\cos(2C(v,w)x)e^{-x^2}\right| &= \left|-2C(v,w)\sin(2C(v,w)x)e^{-x^2} - 2x\cos(2C(v,w)x)e^{-x^2}\right| \\ &\leq 2\left\|C\right\|_\infty + \left|2xe^{-x^2}\right| \leq 2\left\|C\right\|_\infty + 1\end{aligned}$$

for all $x \in \mathbb{R}$ and all $v, w \in \mathbb{R}^2$, we may set $L_{\cos} := 2\left\|C\right\|_\infty + 1$ and similarly for $L_{\sin}$. □

Lemma 2.2.7 can not be applied to the functions $C_s^{Z,j}$ and C_f directly, since they are not bounded. However, this problem can be easily circumvented by replacing $C_s^{Z,j}$ and C_f with

$$\begin{aligned}\widetilde{C}_s^{Z,j} &: \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}, \quad (v,w) \mapsto C_s^{Z,j}(v,w)\mathbb{1}_U(v)\mathbb{1}_U(w) \\ \widetilde{C}_f &: \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}, \quad (v,w) \mapsto C_f(v,w)\mathbb{1}_U(v)\mathbb{1}_U(w)\end{aligned}$$

In doing so, the values of the TCC remain unchanged since $T_Z^{\text{Ishizuka}}(v,w) = 0$ for all $(v,w) \in \mathbb{R}^2 \backslash U \times U$, but the result is that $\widetilde{C}_s^{Z,j}$ and $\widetilde{C}_f$ are bounded functions.

Proposition 2.2.8 (Factorization property)**.** *Let C and $f_{v,w}$ as in Lemma 2.2.7 and*

$$T(v,w) := \int_{-\infty}^{\infty} f_{v,w}(x)\,\mathrm{d}x \qquad \forall\, v, w \in \mathbb{R}^2.$$

If C can be written as $C(v,w) = c(v) - c(w)$ for a suitable function $c : \mathbb{R}^2 \to \mathbb{R}$, then there exists a sequence $(T_N)_{N\in\mathbb{N}}$ of functions converging uniformly to T with the following property:

- *For all $N \in \mathbb{N}$, $j \in \{1, \ldots, N\}$, there are bounded functions $h_{N,j} : \mathbb{R}^2 \to \mathbb{C}$ such that*

$$T_N(v,w) = \sum_{j=1}^{N} h_{N,j}(v) h_{N,j}^*(w) \qquad \forall\, v, w \in \mathbb{R}^2.$$

Proof. Consider the Riemann sums

$$\begin{aligned}\sum_{j=-M}^{M} \delta_M f_{v,w}(j\delta_M) &= \sum_{j=-M}^{M} \frac{\delta_M}{\sqrt{\pi}} \exp\big(-2\mathrm{i}C(v,w)j\delta_M\big)\exp\big(-(j\delta_M)^2\big) \\ &= \sum_{j=-M}^{M} \frac{\delta_M \exp\big(-(j\delta_M)^2\big)}{\sqrt{\pi}} \exp\big(-2\mathrm{i}c(v)j\delta_M\big)\exp\big(-2\mathrm{i}c(w)j\delta_M\big)^*\end{aligned}$$

for $M \in \mathbb{N}$ and $\delta_M = M^{-k}$ with $k \in \big(\frac{1}{2}, 1\big)$. If we define

$$h_{M,j} : \mathbb{R}^2 \to \mathbb{C}, \quad v \mapsto \sqrt{\frac{\delta_M \exp\big(-(j\delta_M)^2\big)}{\sqrt{\pi}}} \exp\big(-2\mathrm{i}c(v)j\delta_M\big)$$

for all $M \in \mathbb{N}$ and all $j \in \{-M, \dots, M\}$, then the functions

$$T_M : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{C}, \quad (v, w) \mapsto \sum_{j=-M}^{M} h_{M,j}(v) h_{M,j}^*(w)$$

are approximations to T of the prescribed structure. It remains to show that the sequence $(T_M)_{M\in\mathbb{N}}$ converges uniformly to T.

Let $v, w \in \mathbb{R}^2$. By Lemma 2.2.7 there exists a Lipschitz constant $L > 0$ of $f_{v,w}$ that is independent of v and w. Using the triangle inequality, it follows that

$$\begin{aligned}
&\left| \sum_{j=-M}^{M} \delta_M f_{v,w}(j\delta_M) - \int_{-\infty}^{\infty} f_{v,w}(x)\,\mathrm{d}x \right| \\
&\le \left| \sum_{j=-M}^{M} \delta_M f_{v,w}(j\delta_M) - \int_{-M\delta_M}^{(M+1)\delta_M} f_{v,w}(x)\,\mathrm{d}x \right| + D_M(v,w) \\
&= \left| \sum_{j=-M}^{M} \int_{j\delta_M}^{(j+1)\delta_M} f_{v,w}(j\delta_M)\,\mathrm{d}x - \sum_{j=-M}^{M} \int_{j\delta_M}^{(j+1)\delta_M} f_{v,w}(x)\,\mathrm{d}x \right| + D_M(v,w) \\
&\le \sum_{j=-M}^{M} \int_{j\delta_M}^{(j+1)\delta_M} \left| f_{v,w}(j\delta_M) - f_{v,w}(x) \right| \,\mathrm{d}x + D_M(v,w) \\
&\le (2M+1)\delta_M L \delta_M + D_M = (2M+1)L\delta_M^2 + D_M(v,w)
\end{aligned}$$

where

$$D_M(v,w) := \int_{-\infty}^{-M\delta_M} \left| f_{v,w}(x) \right| \,\mathrm{d}x + \int_{(M+1)\delta_M}^{\infty} \left| f_{v,w}(x) \right| \,\mathrm{d}x \le \frac{2}{\sqrt{\pi}} \int_{M\delta_M}^{\infty} \exp(-x^2)\,\mathrm{d}x.$$

Recall that $\delta_M = M^{-k}$ with $k \in \big(\frac{1}{2}, 1\big)$. On the one hand, $(2M+1)L\delta_M^2$ converges to zero for $M \to \infty$, since $k > \frac{1}{2}$. On the other hand, $M\delta_M = M^{1-k} \to \infty$ for $M \to \infty$, since $k < 1$, which implies that $D_M \to 0$ uniformly for $M \to \infty$. □

Putting it all together, the spatial and temporal coherence envelopes are approximated by

$$E_s^Z(v,w) = \lim_{M\to\infty} \prod_{k=1}^{2} T_{Z,M}^{\text{spatial},k}(v,w) = \lim_{M\to\infty} \prod_{k=1}^{2} \sum_{j=-M}^{M} h_{Z,M,j}^{\text{spatial},k}(v) h_{Z,M,j}^{\text{spatial},k}(w)^*,$$

$$E_f(v,w) = \lim_{M\to\infty} T_M^{\text{temporal}}(v,w) = \lim_{M\to\infty} \sum_{j=-M}^{M} h_{M,j}^{\text{temporal}}(v) h_{M,j}^{\text{temporal}}(w)^*$$

for all $v, w \in \mathbb{R}^2$, where

$$h_{Z,M,j}^{\text{spatial},k}(v) := \gamma_{M,j} \exp\left(-2\mathrm{i}\left(\frac{\pi\alpha}{\lambda}\partial_k\chi_Z(v)\right) j\delta_M\right),$$
$$h_{M,j}^{\text{temporal}}(v) := \gamma_{M,j} \exp\left(-2\mathrm{i}\left(\frac{\pi\Delta\lambda}{2}\|v\|_2^2\right) j\delta_M\right)$$

with $\gamma_{M,j} := \sqrt{\delta_M \exp(-(j\delta_M)^2)/\sqrt{\pi}}$ and $\delta_M = M^{-k'}$ for $k' \in \left(\frac{1}{2}, 1\right)$. By Proposition 2.2.8, each of the three sequences $(T_{Z,M}^{\text{spatial},1})_{M\in\mathbb{N}}$, $(T_{Z,M}^{\text{spatial},2})_{M\in\mathbb{N}}$ and $(T_{M}^{\text{temporal}})_{M\in\mathbb{N}}$ converges uniformly to the respective integral function that it approximates. From Equations (2.28) and (2.29) it can be seen that the limit functions are bounded. Therefore, the sequence of the products

$$\left(T_{Z,M}^{\text{spatial},1} T_{Z,M}^{\text{spatial},2} T_{M}^{\text{temporal}}\right)_{M\in\mathbb{N}}$$

converges uniformly to the product of their respective limit functions, which is $E_s^Z E_f$. In conclusion,

$$T_Z^{\text{Ishizuka}}(v, w) = \lim_{M\to\infty} \sum_{j,k,l=-M}^{M} h_{Z,M,j,k,l}(v) h_{Z,M,j,k,l}^*(w) \qquad \forall\, v, w \in \mathbb{R}^2 \tag{2.30}$$

with uniform convergence, where

$$h_{Z,M,j,k,l}(v) := t_Z(v) h_{Z,M,j}^{\text{spatial},1}(v) h_{Z,M,k}^{\text{spatial},2}(v) h_{M,l}^{\text{temporal}}(v) \qquad \forall\, v \in \mathbb{R}^2$$

for all $M \in \mathbb{N}$ and all $j, k, l \in \{-M, \ldots, M\}$.

2.2.4 The focal integration approximation

As mentioned in Section 1.2, the MAL algorithm is based on the MIMAP algorithm and mainly differs from the MIMAP algorithm in the way the simulated TEM images are computed. The main contribution of the MAL algorithm is the use of the so-called focal integration approximation instead of using Ishizuka's TCC directly.

The focal integration approximation can be derived from Equation (2.30) by fixing a finite $M \in \mathbb{N}$ instead of taking the limit on the right-hand side. Furthermore, in the focal integration approximation only the temporal coherence envelope is approximated in this way, whereas the spatial coherence envelope is replaced by $E_s^Z(v, w) \approx E_s^Z(v, 0) E_s^Z(w, 0)$. While this yields an approximation of the temporal coherence envelope E_f with Riemann sums corresponding to the integral

$$E_f(v, w) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} \exp\Big(-2\mathrm{i}\frac{\pi\Delta\lambda}{2}\big(\|v\|_2^2 - \|w\|_2^2\big)x\Big) \exp(-x^2)\,\mathrm{d}x \qquad \forall\, v, w \in \mathbb{R}^2$$

as given in Equation (2.29), in the original MAL article [10] a slightly different integral representation of E_f given by substituting $x \to \frac{1}{\sqrt{2}\Delta}x$, i.e.

$$\begin{aligned} E_f(v,w) &= \int_{-\infty}^{\infty} f_{\widetilde{\Delta}}(x) \exp\left(-\pi \mathrm{i} \lambda x \left(\|v\|_2^2 - \|w\|_2^2\right)\right) \,\mathrm{d}x \qquad \forall\, v,w \in \mathbb{R}^2, \\ f_{\widetilde{\Delta}}(x) &= \frac{1}{\sqrt{2\pi}\widetilde{\Delta}} \exp\left(-\frac{x^2}{2\widetilde{\Delta}^2}\right) \qquad \forall\, x \in \mathbb{R}, \\ \widetilde{\Delta} &= \frac{1}{\sqrt{2}}\Delta, \end{aligned} \tag{2.31}$$

is approximated with Riemann sums. The resulting TCC is

$$T_Z^{\mathrm{MAL}}(v,w) := t_Z(v) t_Z^*(w) E_s^Z(v,0) E_s^Z(w,0) \sum_{j=-M}^{M} \delta f_{\widetilde{\Delta}}(j\delta) \exp\left(-\pi \mathrm{i} \lambda j \delta \left(\|v\|_2^2 - \|w\|_2^2 \right)\right) \tag{2.32}$$

for all $v, w \in \mathbb{R}^2$.

In practice, this can be a very good approximation to Ishizuka's TCC. The approximation of the spatial coherence envelope by $E_s^Z(v,w) \approx E_s^Z(v,0)E_s^Z(w,0)$ is justified for highly coherent field emission guns [10] and the approximation of the temporal coherence envelope can clearly be arbitrarily good, depending on the magnitude of $M \in \mathbb{N}$. However, the focal integration approximation has the conceptual disadvantage that two additional parameters, $M \in \mathbb{N}$ and $\delta > 0$, need to be adjusted appropriately in order to get reasonable results. Note that the parameters M and δ_M in Section 2.2.3 are not a part of T_Z^{Ishizuka} but merely intermediate variables in the proof of IV. The authors in [10] suggest to choose $M \in \{2,3\}$ depending on the desired level of accuracy and $\delta \approx \Delta$.

In the following, the same conditions as in Sections 2.2.2 and 2.2.3 are imposed on the aperture function. As before, the support of T_Z^{MAL} satisfies $\mathrm{supp}(T_Z^{\mathrm{MAL}}) \subseteq \overline{U} \times \overline{U}$ for $U := B_{r_a}(0)$. Furthermore, it is assumed that the wave aberration function contains only isotropic aberrations, that is, the coefficients of the anisotropic aberrations are assumed to be zero.

Properties I and II: Both the symmetry property $T_Z^{\mathrm{MAL}}(v,w)^* = T_Z^{\mathrm{MAL}}(w,v)$ for all $v, w \in \mathbb{R}^2$ as well as the fact that T_Z^{MAL} is continuous on $U \times U$ can directly be verified with the definition in Equation (2.32).

Property III: For all $v \in \mathbb{R}^2$, we have

$$T_Z^{\mathrm{MAL}}(v,v) = a(v)^2 E_s^Z(v,0)^2 \sum_{j=-M}^{M} \delta f_{\widetilde{\Delta}}(j\delta).$$

The temporal coherence factor $\sum_{j=-M}^{M} \delta f_{\widetilde{\Delta}}(j\delta)$ is equal to a positive constant $c_1 > 0$ that is independent of v. The spatial coherence factor is

$$E_s^Z(v,0)^2 = \exp\left(-\left(\frac{\pi\alpha}{\lambda}\right) \|\nabla\chi_Z(v) - \nabla\chi_Z(0)\|_2^2\right)^2 = \exp\left(-\left(\frac{\pi\alpha}{\lambda}\right) \|\nabla\chi_Z(v)\|_2^2\right)^2,$$

where $\nabla\chi_Z(0) = 0$ follows from Equation (2.10). Since χ_Z is continuously differentiable, $v \mapsto \|\nabla\chi_Z(v)\|_2^2$ is a continuous function and there exists a constant $\xi < \infty$ such that

$$\max_{v\in\overline{U}} \|\nabla\chi_Z(v)\|_2^2 = \xi.$$

This implies $E_s^Z(v,0)^2 \geq \exp\big(-\big(\frac{\pi\alpha}{\lambda}\big)\xi\big)^2 =: c_2 > 0$ for all $v \in \overline{U}$ and consequently

$$T_Z^{\mathrm{MAL}}(v,v) \geq c_1 c_2 a(v)^2 \qquad \forall\, v \in U.$$

This shows that Property IIIa or IIIb holds depending on the particular choice of the aperture function.

Property IV: The measurability and boundedness of T_Z^{MAL} follows immediately from Equation (2.32). Since E_s^Z is real-valued and $f_{\tilde{\Delta}}$ is nonnegative, the TCC T_Z^{MAL} can be written as

$$T_Z^{\mathrm{MAL}}(v,w) = \sum_{j=-M}^{M} h_{Z,j}(v) h_{Z,j}^*(w) \qquad \forall\, v, w \in \mathbb{R}^2$$

with

$$h_{Z,j}(v) := t_Z(v) E_s^Z(v,0) \sqrt{\delta f_{\tilde{\Delta}}(j\delta)} \exp\big(-\pi \mathrm{i} \lambda j \delta \, \|v\|_2^2\big) \qquad \forall\, v \in \mathbb{R}^2$$

for all $j \in \{-M, \dots, M\}$. Thus, T_Z^{MAL} is itself a $\star$-separable weight and can therefore trivially be approximated with a uniformly convergent sequence of $\star$-separable weights.

Remark 2.2.9. Using T_Z^{MAL} instead of T_Z^{Ishizuka} has the advantage that the Fourier transform may be used to accelerate the image simulation greatly. By Equation (2.1), the relationship between a real space TEM image at focus Z and the corresponding exit wave Ψ is

$$\begin{aligned}
\mathcal{F}^{-1}(\mathcal{I}_{\Psi,Z}) &= \mathcal{F}^{-1}(\Psi \star_{T_Z^{\mathrm{full}}} \Psi) \approx \mathcal{F}^{-1}(\Psi \star_{T_Z^{\mathrm{MAL}}} \Psi) = \sum_{j=-M}^{M} \mathcal{F}^{-1}\big((\Psi h_{Z,j}) \star (\Psi h_{Z,j})\big) \\
&= \sum_{j=-M}^{M} \big|\mathcal{F}^{-1}(\Psi h_{Z,j})\big|^2
\end{aligned}$$

and the rightmost side can be evaluated efficiently using the fast Fourier transform.

2.2.5 The quasi-coherent TCC

All of the previously considered TCCs can be associated with a non-linear imaging model. In a non-linear imaging model, all interferences are taken into account, including in particular the interferences between two frequencies $v, w \in \mathbb{R}^2$ with $v \neq 0$ and $w \neq 0$. In a linear imaging model on the other hand, only those interferences between two frequencies $v, w \in \mathbb{R}^2$ are considered where either $v = 0$ or $w = 0$. This is equivalent to saying that the TCC corresponding to a linear imaging model is of the form $h_Z(v)h_Z^*(w)$

for a suitable function $h_Z : \mathbb{R}^2 \to \mathbb{C}$. In a linear imaging model, a TEM image is therefore given by

$$\mathcal{F}^{-1}(\Psi \star_{T_Z^{\text{linear}}} \Psi) = \mathcal{F}^{-1}\big((\Psi h_Z) \star (\Psi h_Z)\big) = \big|\mathcal{F}^{-1}(\Psi h_Z)\big|^2 = \big|\mathcal{F}^{-1}(\Psi) * \mathcal{F}^{-1}(h_Z)\big|^2, \tag{2.33}$$

where $T_Z^{\text{linear}}(v, w) = h_Z(v)h_Z^*(w)$ for all $v, w \in \mathbb{R}^2$. This shows that in real space coordinates the image plane electron wave $\mathcal{F}^{-1}(\Psi) * \mathcal{F}^{-1}(h_Z)$ is simply a convolution of the exit plane electron wave $\mathcal{F}^{-1}(\Psi)$ with the point spread function $\mathcal{F}^{-1}(h_Z)$.

In the quasi-coherent approximation, the coherence envelopes are approximated by

$$\begin{aligned} E_s^Z(v, w) &\approx E_s^Z(v, 0)E_s^Z(w, 0) \qquad \forall\, v, w \in \mathbb{R}^2, \\ E_f(v, w) &\approx E_f(v, 0)E_f(w, 0) \qquad \forall\, v, w \in \mathbb{R}^2 \end{aligned}$$

and the corresponding TCC is given by

$$T_Z^{\text{q-coh}}(v, w) := t_Z(v)t_Z^*(w)E_s^Z(v, 0)E_s^Z(w, 0)E_f(v, 0)E_f(w, 0) \qquad \forall\, v, w \in \mathbb{R}^2.$$

This TCC is clearly of the form $T_Z^{\text{q-coh}}(v, w) = h_Z(v)h_Z^*(w)$ with

$$h_Z(v) := t_Z(v)E_s^Z(v, 0)E_f(v, 0).$$

Properties I–IV: If the aperture function satisfies the same conditions as before and $U = B_{r_a}(0)$, then all of the properties can be shown just like in the previous section for the focal integration approximation.

2.2.6 The coherent TCC

For a perfectly coherent TEM, the coherence parameters α and Δ for the spatial and temporal coherence are equal to zero. By the definition of the coherence envelopes in Equations (2.25) and (2.26), this yields $E_s^Z(v, w) = 1$ and $E_f(v, w) = 1$ for all $v, w \in \mathbb{R}^2$. Hence, the TCC for coherent imaging is

$$T_Z^{\text{coh}}(v, w) := t_Z(v)t_Z^*(w) = a(v)a^*(w)p_Z(v)p_Z^*(w) \qquad \forall\, v, w \in \mathbb{R}^2.$$

Alternatively, this TCC can be retrieved from the full TCC in Section 2.2.1 directly. Consider sequences $(s_n)_{n\in\mathbb{N}} \in L^1(\mathbb{R}^2, \mathbb{R})^{\mathbb{N}}$ and $(f_n)_{n\in\mathbb{N}} \in L^1(\mathbb{R}, \mathbb{R})^{\mathbb{N}}$ of probability density functions with

$$\lim_{n\to\infty} \int_{B_\varepsilon(0)} s_n(u)\,\mathrm{d}u = 1 \qquad \text{and} \qquad \lim_{n\to\infty} \int_{(-\varepsilon,\varepsilon)} f_n(\zeta)\,\mathrm{d}\zeta = 1$$

for all $\varepsilon > 0$. Then

$$\lim_{n\to\infty} \lim_{m\to\infty} \int_{\mathbb{R}} \int_{\mathbb{R}^2} s_n(u)f_m(\zeta)t_{Z+\zeta}(v + u)t_{Z+\zeta}^*(w + u)\,\mathrm{d}u\,\mathrm{d}\zeta = t_Z(v)t_Z^*(w) = T_Z^{\text{coh}}(v, w)$$

holds for all $v, w \in \mathbb{R}^2$ such that the aperture function a is continuous in v and w. This follows from applying the auxiliary lemma 2.2.10 twice.

Lemma 2.2.10. *Let $d \in \mathbb{N}$ and $(g_n)_{n\in\mathbb{N}} \in L^1(\mathbb{R}^d, \mathbb{R}_{\geq 0})$ be a sequence of nonnegative integrable functions with $\int_{\mathbb{R}^d} g_n(x)\,\mathrm{d}x = 1$ for all $n \in \mathbb{N}$ and*

$$\lim_{n\to\infty} \int_{B_\varepsilon(0)} g_n(x)\,\mathrm{d}x = 1$$

for all $\varepsilon > 0$. If $f : \mathbb{R}^d \to \mathbb{C}$ is bounded and continuous in 0, then

$$\lim_{n\to\infty} \int_{\mathbb{R}^d} g_n(x) f(x)\,\mathrm{d}x = f(0).$$

Proof. For all $\varepsilon > 0$ we have

$$\begin{aligned}
&\left|\int_{\mathbb{R}^d} g_n(x) f(x)\,\mathrm{d}x - f(0)\right| \\
&= \left|\int_{\mathbb{R}^d} g_n(x)\big(f(x) - f(0)\big)\,\mathrm{d}x\right| \\
&\leq \int_{B_\varepsilon(0)} g_n(x)\,|f(x) - f(0)|\,\mathrm{d}x + \int_{\mathbb{R}^d\setminus B_\varepsilon(0)} g_n(x)\,|f(x) - f(0)|\,\mathrm{d}x \\
&\leq \sup_{x\in B_\varepsilon(0)} |f(x) - f(0)| \int_{B_\varepsilon(0)} g_n(x)\,\mathrm{d}x + 2\,\|f\|_\infty \int_{\mathbb{R}^d\setminus B_\varepsilon(0)} g_n(x)\,\mathrm{d}x.
\end{aligned}$$

Taking the limit $n \to \infty$, it follows that

$$\limsup_{n\to\infty} \left|\int_{\mathbb{R}^d} g_n(x) f(x)\,\mathrm{d}x - f(0)\right| \leq \sup_{x\in B_\varepsilon(0)} |f(x) - f(0)|\,.$$

Since this inequality holds for all $\varepsilon > 0$ and f is continuous in 0, the claim follows. □

Properties I–IV: Assuming that the aperture function satisfies the same conditions as in the previous sections, we set $U := B_{r_a}(0)$ and all of the properties follow directly from the definition of T_Z^{coh}.

2.2.7 Summary

The notation and results from Sections 2.2.1 to 2.2.6 are summarized below.

Transmission cross-coefficients: In the following tables, the definition of the TCCs is given along with sufficient conditions on their constituents such that all of the properties I–IV hold.

2.2.1: The full TCC

$$T_Z^{\text{full}}(v,w) = \int_{\mathbb{R}} \int_{\mathbb{R}^2} s(u) f(\zeta) p_{Z+\zeta}(v+u) a(v+u) p^*_{Z+\zeta}(w+u) a^*(w+u) \,\mathrm{d}u \,\mathrm{d}\zeta$$

$U = B_{r_s+r_a}(0)$

• $s : \mathbb{R}^2 \to \mathbb{R}$	differentiable probability density with bounded support, $\operatorname{supp}(s) = \overline{B_{r_s}(0)}$, and $\forall\, \varepsilon > 0 \;\exists\, s_\varepsilon > 0 :\; s(x) \geq s_\varepsilon \;\forall\, x \in B_{r_s-\varepsilon}(0)$.
• $f : \mathbb{R} \to \mathbb{R}$	differentiable probability density.
• $p_Z : \mathbb{R}^2 \to \mathbb{C}$	pure phase transfer function as defined in Equation (2.9).
• $a : \mathbb{R}^2 \to \mathbb{R}$	bounded, differentiable function with $0 \leq \lvert a(x)\rvert \leq 1 \;\forall\, x \in \mathbb{R}$, bounded support, $\operatorname{supp}(a) = \overline{B_{r_a}(0)}$, and $\forall\, \varepsilon > 0 \;\exists\, a_\varepsilon > 0 :\; a(x) \geq a_\varepsilon \;\forall\, x \in B_{r_a-\varepsilon}(0)$.

Additionally, the functions need to satisfy Equation (2.16) from Proposition 2.2.4.

2.2.2: The large aperture approximation

$$T_Z^{\text{ap}}(v,w) = a(v) a^*(w) \int_{\mathbb{R}} \int_{\mathbb{R}^2} s(u) f(\zeta) p_{Z+\zeta}(v+u) p^*_{Z+\zeta}(w+u) \,\mathrm{d}u \,\mathrm{d}\zeta$$

$U = B_{r_a}(0)$

• $s : \mathbb{R}^2 \to \mathbb{R}$	differentiable probability density with bounded support, $\operatorname{supp}(s) = \overline{B_{r_s}(0)}$, and $\forall\, \varepsilon > 0 \;\exists\, s_\varepsilon > 0 :\; s(x) \geq s_\varepsilon \;\forall\, x \in B_{r_s-\varepsilon}(0)$.
• $f : \mathbb{R} \to \mathbb{R}$	differentiable probability density.
• $p_Z : \mathbb{R}^2 \to \mathbb{C}$	pure phase transfer function as defined in Equation (2.9).
• $a : \mathbb{R}^2 \to \mathbb{R}$	bounded function with $0 \leq \lvert a(x)\rvert \leq 1 \;\forall\, x \in \mathbb{R}$, bounded support, $\operatorname{supp}(a) = \overline{B_{r_a}(0)}$, and $\forall\, \varepsilon > 0 \;\exists\, a_\varepsilon > 0 :\; a(x) \geq a_\varepsilon \;\forall\, x \in B_{r_a-\varepsilon}(0)$.

Additionally, the functions need to satisfy Equation (2.16) from Proposition 2.2.4 with g and $q_{v,w}$ as given in Section 2.2.2.

2.2.3: Ishizuka's TCC

$$T_Z^{\text{Ishizuka}}(v,w) = a(v) a^*(w) p_Z(v) p^*_Z(w) E_s^Z(v,w) E_f(v,w)$$

$U = B_{r_a}(0)$

• $p_Z : \mathbb{R}^2 \to \mathbb{C}$	pure phase transfer function as defined in Equation (2.9).
• $a : \mathbb{R}^2 \to \mathbb{R}$	bounded function with $0 \leq \lvert a(x)\rvert \leq 1 \;\forall\, x \in \mathbb{R}$, bounded support, $\operatorname{supp}(a) = \overline{B_{r_a}(0)}$, and $\forall\, \varepsilon > 0 \;\exists\, a_\varepsilon > 0 :\; a(x) \geq a_\varepsilon \;\forall\, x \in B_{r_a-\varepsilon}(0)$.
• $E_s^Z : \mathbb{R}^2 \times \mathbb{R}^2 \to [0,1]$	spatial coherence envelope as defined in Equation (2.25).
• $E_f : \mathbb{R}^2 \times \mathbb{R}^2 \to [0,1]$	temporal coherence envelope as defined in Equation (2.26).

2.2.4: The focal integration approximation

$$T_Z^{\mathrm{MAL}}(v,w) = t_Z(v)t_Z^*(w)E_s^Z(v,0)E_s^Z(w,0)\sum_{j=-M}^{M}\delta f_{\widetilde{\Delta}}(j\delta)\exp\big(-\pi \mathrm{i}\lambda j\delta\big(\|v\|_2^2-\|w\|_2^2\big)\big)$$

$U = B_{r_a}(0)$

• $t_Z : \mathbb{R}^2 \to \mathbb{C}$	pupil function $t_Z(v) = p_Z(v)a(v)$ for all $v \in \mathbb{R}^2$.
• $p_Z : \mathbb{R}^2 \to \mathbb{C}$	pure phase transfer function as defined in Equation (2.9).
• $a : \mathbb{R}^2 \to \mathbb{R}$	bounded function with $0 \leq \lvert a(x)\rvert \leq 1\ \forall\, x \in \mathbb{R}$, bounded support, $\mathrm{supp}(a) = \overline{B_{r_a}(0)}$, and $\forall\, \varepsilon > 0\ \exists\, a_\varepsilon > 0 :\ a(x) \geq a_\varepsilon\ \forall\, x \in B_{r_a-\varepsilon}(0)$.
• $E_s^Z : \mathbb{R}^2 \times \mathbb{R}^2 \to [0,1]$	spatial coherence envelope as defined in Equation (2.25).
• $f_{\widetilde{\Delta}} : \mathbb{R} \to \mathbb{R}$	Gaussian function as defined in Equation (2.31).

2.2.5: The quasi-coherent TCC

$$T_Z^{\text{q-coh}}(v,w) = a(v)a^*(w)p_Z(v)p_Z^*(w)E_s^Z(v,0)E_s^Z(w,0)E_f(v,0)E_f(w,0)$$

$U = B_{r_a}(0)$

• $p_Z : \mathbb{R}^2 \to \mathbb{C}$	pure phase transfer function as defined in Equation (2.9).
• $a : \mathbb{R}^2 \to \mathbb{R}$	bounded function with $0 \leq \lvert a(x)\rvert \leq 1\ \forall\, x \in \mathbb{R}$, bounded support, $\mathrm{supp}(a) = \overline{B_{r_a}(0)}$, and $\forall\, \varepsilon > 0\ \exists\, a_\varepsilon > 0 :\ a(x) \geq a_\varepsilon\ \forall\, x \in B_{r_a-\varepsilon}(0)$.
• $E_s^Z : \mathbb{R}^2 \times \mathbb{R}^2 \to [0,1]$	spatial coherence envelope as defined in Equation (2.25).
• $E_f : \mathbb{R}^2 \times \mathbb{R}^2 \to [0,1]$	temporal coherence envelope as defined in Equation (2.26).

2.2.6: The coherent TCC

$$T_Z^{\mathrm{coh}}(v,w) = a(v)a^*(w)p_Z(v)p_Z^*(w)$$

$U = B_{r_a}(0)$

• $p_Z : \mathbb{R}^2 \to \mathbb{C}$	pure phase transfer function as defined in Equation (2.9).
• $a : \mathbb{R}^2 \to \mathbb{R}$	bounded function with $0 \leq \lvert a(x)\rvert \leq 1\ \forall\, x \in \mathbb{R}$, bounded support, $\mathrm{supp}(a) = \overline{B_{r_a}(0)}$, and $\forall\, \varepsilon > 0\ \exists\, a_\varepsilon > 0 :\ a(x) \geq a_\varepsilon\ \forall\, x \in B_{r_a-\varepsilon}(0)$.

Auxiliary functions: The following functions are a part of the various TCCs.

- Pure phase transfer function (Equation (2.9)):

$$p_Z(v) = \exp\big(-2\pi \mathrm{i}\chi_Z(v)\big) \qquad \forall\, v \in \mathbb{R}^2 \quad (Z \in \mathbb{R}).$$

- Wave aberration function (Equation (2.10)):

$$\chi_Z(v) = \frac{1}{\lambda}\sum_{m\in\mathbb{N}}\sum_{\substack{n=0\\ m+n \text{ even}}}^{m} \frac{1}{m} c_{m,n}\lambda^m \|v\|_2^m \cos\big(n(\arg(v) - \varphi_{m,n})\big) \qquad \forall\, v \in \mathbb{R}^2$$

  ($Z \in \mathbb{R},$ only finitely many coefficients $c_{m,n} \neq 0$).

- Aperture function (Equations (2.12) to (2.14)):

$$a(v) = \begin{cases} 1, & \text{if } \lambda\|v\|_2 < \alpha_{\max}, \\ 0, & \text{otherwise} \end{cases} \qquad \forall\, v \in \mathbb{R}^2 \quad \text{or}$$

$$a_\delta(v) = \begin{cases} \left(1 + \exp\left(\frac{\lambda^2\|v\|_2^2 - \alpha_{\max}^2}{\delta^2}\right)\right)^{-1}, & \text{if } \|v\|_2 < \frac{d}{2}, \\ 0, & \text{otherwise} \end{cases} \qquad \forall\, v \in \mathbb{R}^2 \quad (d, \delta > 0)$$

  (or a smooth approximation to the above functions with bounded support).

- Spatial coherence envelope (Equation (2.25)):

$$E_s^Z(v,w) = \exp\left(-\left(\frac{\pi\alpha}{\lambda}\right)^2 \|\nabla\chi_Z(v) - \nabla\chi_Z(w)\|_2^2\right) \qquad \forall\, v, w \in \mathbb{R}^2.$$

- Temporal coherence envelope (Equation (2.26)):

$$E_f(v,w) = \exp\left(-\frac{1}{4}(\pi\Delta\lambda)^2 \left(\|v\|_2^2 - \|w\|_2^2\right)^2\right) \qquad \forall\, v, w \in \mathbb{R}^2.$$

- Gaussian in the focal integration approximation (Equation (2.31)):

$$f_{\widetilde{\Delta}}(x) = \frac{1}{\sqrt{2\pi}\widetilde{\Delta}} \exp\left(-\frac{x^2}{2\widetilde{\Delta}^2}\right) \qquad \forall\, x \in \mathbb{R}.$$

Notation: The following tables give a brief overview of the notation used in the previous sections. The same notation will be kept in the subsequent chapters.

General variables

λ	Electron wavelength
$\alpha_{\max}$	Maximum semiangle allowed by the objective aperture

Aberrations

$(c_{m,n}, \varphi_{m,n})$	Aberration coefficients ($m \in \mathbb{N}, n \in \{0,\dots,m\} : m+n$ even)
Z	focus ($Z = c_{2,0}$)
C_s	third-order spherical aberration ($C_s = c_{4,0}$)

Coherence parameters	
q_0	"Width" of the Gaussian function used for the intensity distribution of the illumination s (see Equation (2.21))
α	Approximately equal to the half angle of beam convergence (spatial coherence, $\alpha = q_0\lambda$)
Δ	Focal spread parameter (temporal coherence, see Equation (2.22))
$\widetilde{\Delta}$	$\widetilde{\Delta} = \frac{1}{\sqrt{2}}\Delta$

2.3 Properties of the forward model

The results presented so far allow us to formulate and proof several properties of TEM images that are simulated with any of the previously introduced TCCs. In this section, the function $T_Z : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{C}$ stands for any of the TCCs from Section 2.2.

A basic, but important, property of the simulated images $\Psi \star_{T_Z} \Psi$ is the fact that they are square-integrable for any exit wave $\Psi \in L^2(\mathbb{R}^2, \mathbb{C})$, which follows from Corollary 2.1.4. Additionally, the support of the simulated images is bounded in Fourier space, which is a consequence of the bandwidth limit imposed by the objective aperture. If $r > 0$ is chosen such that $\operatorname{supp}(T_Z) \subseteq \overline{U} \times \overline{U}$ with $U := B_r(0)$, then it can be seen from

$$\big(\Psi \star_{T_Z} \Psi\big)(x) = \int_{\mathbb{R}^2} \Psi^*(y)\Psi(x+y)T_Z(x+y,y)\,\mathrm{d}y \qquad \forall\, x \in \mathbb{R}^2$$

that $\operatorname{supp}(\Psi \star_{T_Z} \Psi) \subseteq B_{2r}(0)$. It should be noted that this implies that the support of real space TEM images is in fact unbounded, although the recorded images naturally only have a limited size. This error will be treated as a discretization error in algorithms for exit wave reconstruction (see Section 4.2).

Lemma 2.3.1. *The simulated real space images $\mathcal{F}^{-1}(\Psi \star_{T_Z} \Psi)$ are smooth functions for all $\Psi \in L^2(\mathbb{R}^2, \mathbb{C})$.*

Proof. By Lemma 2.1.2 and since $\operatorname{supp}(\Psi \star_{T_Z} \Psi) \subseteq B_{2r}(0)$, the simulated Fourier space images $\Psi \star_{T_Z} \Psi$ are bounded with bounded support. For all $n \in \mathbb{N}^2$, there is a multivariate polynomial function $p_n : \mathbb{R}^2 \to \mathbb{C}$ such that

$$\left|\big(\Psi \star_{T_Z} \Psi\big)(y)\frac{\partial^n}{\partial x^n}e^{2\pi \mathrm{i} x\cdot y}\right| = \left|\big(\Psi \star_{T_Z} \Psi\big)(y)p_n(y)\right| =: g_n(y) \qquad \forall\, x, y \in \mathbb{R}^2$$

and $g_n \in L^1(\mathbb{R}^2, \mathbb{C})$ is an integrable function since $\operatorname{supp}(\Psi \star_{T_Z} \Psi) \subseteq B_{2r}(0)$ and $\Psi \star_{T_Z} \Psi$ is integrable. Thus, by the dominated convergence theorem, the partial derivatives of the real space images exist and are given by

$$\frac{\partial^n}{\partial x^n}\mathcal{F}^{-1}\big(\Psi \star_{T_Z} \Psi\big)(x) = \int_{\mathbb{R}^2} \big(\Psi \star_{T_Z} \Psi\big)(y)\frac{\partial^n}{\partial x^n}e^{2\pi \mathrm{i} x\cdot y}\,\mathrm{d}y \qquad \forall\, x \in \mathbb{R}^2$$

for all $n \in \mathbb{N}^2$. □

Since images recorded with a TEM correspond to the squared amplitude of the electron wave, they are not just real-valued but also nonnegative. It can be shown that this also pertains to simulated TEM images as long as the TCC satisfies the factorization property IV. This is an immediate consequence of Lemma 2.1.7 and is formulated here as a corollary for completeness:

Corollary 2.3.2. *The simulated real space images* $\mathcal{F}^{-1}(\Psi \star_{T_Z} \Psi)$ *are real-valued and nonnegative for all* $\Psi \in L^2(\mathbb{R}^2, \mathbb{C})$.

Another intuitive, yet mathematically not obvious property of the forward model is the fact that if the exit wave is translated in real space coordinates, then the simulated images are translated the same way. In order to show this, the translation and the modulation by a vector $z \in \mathbb{R}^2$ are denoted by

$$\begin{aligned} \phi_z &: \mathbb{R}^2 \to \mathbb{R}^2, \quad x \mapsto x + z, \\ \mu_z &: \mathbb{R}^2 \to \mathbb{C}, \quad x \mapsto \exp(2\pi \mathrm{i} x \cdot z) \end{aligned}$$

in the following lemma.

Lemma 2.3.3. *If* $\psi \in L^2(\mathbb{R}^2, \mathbb{C})$ *and* $z \in \mathbb{R}^2$, *then*

$$\mathcal{F}^{-1}\big(\mathcal{F}(\psi \circ \phi_z) \star_{T_Z} \mathcal{F}(\psi \circ \phi_z)\big) = \mathcal{F}^{-1}(\Psi \star_{T_Z} \Psi) \circ \phi_z,$$

where $\Psi = \mathcal{F}(\psi) \in L^2(\mathbb{R}^2, \mathbb{C})$.

Proof. A universal property of the Fourier transform is that it converts a modulation μ_z to a translation ϕ_z and vice versa. More precisely, the equations

$$\mathcal{F}(f \circ \phi_z) = \mathcal{F}(f)\mu_z \qquad \text{and} \qquad \mathcal{F}^{-1}(f\mu_z) = \mathcal{F}^{-1}(f) \circ \phi_z \tag{2.34}$$

hold for all $f \in L^2(\mathbb{R}^2, \mathbb{C})$. This implies

$$\begin{aligned} &\big(\mathcal{F}(\psi \circ \phi_z) \star_{T_Z} \mathcal{F}(\psi \circ \phi_z)\big)(x) \\ &= \big((\Psi\mu_z) \star_{T_Z} (\Psi\mu_z)\big)(x) \\ &= \int_{\mathbb{R}^2} \Psi^*(y)\Psi(x+y)\mu_z^*(y)\mu_z(x+y)T_Z(x+y, y)\,\mathrm{d}y \\ &= \mu_z(x) \int_{\mathbb{R}^2} \Psi^*(y)\Psi(x+y)T_Z(x+y, y)\,\mathrm{d}y \\ &= \mu_z(x)\big(\Psi \star_{T_Z} \Psi\big)(x) \end{aligned}$$

for all $x \in \mathbb{R}^2$. The claim follows by taking the inverse Fourier transform on the left-hand side and the right-hand side of the above chain of equality and applying Equation (2.34) once more. ☐

Lemma 2.3.3 is stated in real space coordinates using translations in order to provide an intuitive view of the result. However, on occasion the equivalent identities

$$(\Psi\mu_z) \star_{T_Z} (\Psi\mu_z) = (\Psi \star_{T_Z} \Psi)\mu_z \quad \text{and} \tag{2.35}$$

$$\mathcal{F}^{-1}\big((\Psi\mu_z) \star_{T_Z} (\Psi\mu_z)\big) = \mathcal{F}^{-1}(\Psi \star_{T_Z} \Psi) \circ \phi_z \tag{2.36}$$

are more useful for algebraic manipulations than applying Lemma 2.3.3 directly.

3 The inverse problem: exit wave reconstruction

In this chapter, three closely related variational approaches to exit wave reconstruction are presented. These approaches are all based on the minimization of a given objective functional, whose mathematical properties are investigated in this chapter. The algorithms and numerical methods that are used to find a minimizer computationally are discussed in Chapter 4.

First, a brief overview of the MIMAP and MAL functionals is given in Section 3.1. One of the main differences of the functionals is the fact that the MIMAP functionals include a generalized Tikhonov regularizer for the exit wave, whereas the MAL functional does not contain a regularizer for the exit wave. For this reason, the importance of the regularizer with regard to the coercivity of the functional is investigated in Section 3.2, where it is shown that the MAL functional is not coercive. Then, a novel objective functional E_σ for joint exit wave reconstruction and image registration is introduced in Section 3.3. This functional extends the MIMAP and MAL functionals by incorporating the registration more directly into the functional and thus allows for a simultaneous optimization of the exit wave and the registration. A part of the properties of E_σ given here have previously been published in [14]. The main result in this section is the existence of minimizers of E_σ. Finally, the convexity of the objective functional is investigated in Section 3.4.

In a nutshell, the problem of reconstructing the exit wave from a series of TEM images taken at different focus values can be phrased as follows:

Given: A series of $N \in \mathbb{N}$ images $g_1, \dots, g_N \in L^2(\mathbb{R}^2, \mathbb{R}_{\geq 0})$ and the corresponding focus values $Z_1, \dots, Z_N \in \mathbb{R}$.

Task: Find a function $\Psi \in L^2(\mathbb{R}^2, \mathbb{C})$ such that

$$g_j \approx \mathcal{F}^{-1}(\Psi \star_{T^{\mathrm{full}}_{Z_j}} \Psi) \qquad \forall\, j \in \{1, \dots, N\}. \tag{3.1}$$

As stated in Section 1.2, only focus image series will be considered here, although theoretically any other microscope parameter that is involved in the forward model can be varied as well. If, for example, the accelerating voltage or the spherical aberration were varied in the experimental image series, then the TCC would depend on the electron wavelength λ or the spherical aberration C_s instead of the focus Z. This does not affect the theory in Chapter 2 in any way, since all of the instrument's parameters are assumed to be constant for the simulation of a single image. Thus, the results presented here continue to hold if other parameters than the focus are varied within the image series.

3.1 Existing variational approaches for exit wave reconstruction

In both the MIMAP and the MAL algorithm, the exit wave is reconstructed by minimizing a functional of the kind

$$J_\sigma[\Psi] = \sum_{j=1}^{N} \left\|\Psi \star_{T_{Z_j}} \Psi - \mathcal{F}(g_j)\right\|_{L^2}^2 + \sigma \left\|\Psi - \Psi_M\right\|_{L^2}^2 \qquad \forall\, \Psi \in L^2(\mathbb{R}^2, \mathbb{C}), \tag{3.2}$$

where $\sigma \geq 0$ and T_{Z_j} is the approximation to $T_{Z_j}^{\text{full}}$ that is used in the respective algorithm. As before, $g_1, \dots, g_N \in L^2(\mathbb{R}^2, \mathbb{R}_{\geq 0})$ are experimental TEM images in real space, which have been acquired with the focus values $Z_1, \dots, Z_N \in \mathbb{R}$. The parameter $\Psi_M \in L^2(\mathbb{R}^2, \mathbb{C})$ is an a-priori estimate of the exit wave.

Thus, the functional in Equation (3.2) consists of a data term measuring the squared L^2 distance of the simulated and experimental TEM images in Fourier space and a generalized Tikhonov regularizer for the exit wave. Note that the data term is well-defined, since $\Psi \star_{T_{Z_j}} \Psi$ is square-integrable by Corollary 2.1.4.

3.1.1 The MIMAP functionals

In the original MIMAP algorithm as described in [38], the MIMAP functionals are defined as

$$J_\sigma^{\text{MIMAP},m}[\Psi] = \sum_{j=1}^{N} \left\|\Psi \star_{T_j^{\text{MIMAP},m}} \Psi - \mathcal{F}(g_j)\right\|_{L^2}^2 + \sigma \left\|\Psi - \Psi_M\right\|_{L^2}^2$$

for $m \in \{1, 2\}$, where σ is inversely proportional to the signal-to-noise ratio. Two different kinds of TCCs are considered for the image simulation: an extension of the quasi-coherent TCC and an extension of Ishizuka's TCC.

The original MIMAP functionals also include a term for a background constant, which is omitted here. This term was intended to account for an artificial constant that is added to the images if they are recorded using photographic film plates and copied subsequently, which was common at the time of writing of [38] in 1984.

The extension of the quasi-coherent TCC: The MIMAP algorithm as described in [38] makes use of a slight generalization of the quasi-coherent TCC described in Section 2.2.5, which includes two additional effects. On the one hand, the term u in the derivation of Ishizuka's TCC in Equation (2.23) is not approximated by unity. On the other hand, the registration and a Debye-Waller temperature coefficient are included in the TCC, both of which are, strictly speaking, not a part of the microscope's transfer function.

Explicitly, the quasi-coherent TCC used in the MIMAP algorithm is

$$T_j^{\mathrm{MIMAP},1}(v,w) := h_j(v)h_j^*(w) \qquad \forall\, v,w \in \mathbb{R}^2$$

with

$$h_j(v) = \exp\left(\mathrm{i}\widetilde{\chi}_{Z_j}(v)\right)\frac{1}{\sqrt{1+\varepsilon_0\left\|v\right\|_2^2}}\exp\left(-\frac{B\left\|v\right\|_2^2}{4}\right)\exp(2\pi\mathrm{i}v\cdot t_j) \qquad \forall\, v \in \mathbb{R}^2$$

for all $j \in \{1,\dots,N\}$.

The leftmost factor $\exp\left(\mathrm{i}\widetilde{\chi}_{Z_j}(v)\right)$ is closely related to the function h_{Z_j} of the quasi-coherent TCC $T_{Z_j}^{\text{q-coh}}$ defined in Section 2.2.5, which is shown in the following. The function $\widetilde{\chi}_{Z_j}$ comprises the wave aberration function and the coherence envelopes and is defined as

$$\begin{aligned}
\operatorname{Re}\left(\widetilde{\chi}_{Z_j}(v)\right) &= \frac{\pi\lambda\left\|v\right\|_2^2}{1+\varepsilon_0\left\|v\right\|_2^2}\left(-\frac{1}{2}C_s\left(1-\varepsilon_0\left\|v\right\|_2^2\right)\lambda^2\left\|v\right\|_2^2 - Z\right) && \forall\, v \in \mathbb{R}^2,\\
\operatorname{Im}\left(\widetilde{\chi}_{Z_j}(v)\right) &= \frac{\pi^2\alpha^2\left\|v\right\|_2^2\left(-C_s\lambda^2\left\|v\right\|_2^2 - Z\right)^2}{1+\varepsilon_0\left\|v\right\|_2^2} + \frac{\frac{1}{4}\pi^2\lambda^2\Delta^2\left\|v\right\|_2^4}{1+\varepsilon_0\left\|v\right\|_2^2} && \forall\, v \in \mathbb{R}^2,
\end{aligned}$$

where $\varepsilon_0 = \pi^2\alpha^2\Delta^2$ [24, 70, 38]. By Equations (2.11) and (2.27), the real and imaginary parts of $\widetilde{\chi}_{Z_j}$ can be written as

$$\operatorname{Re}\left(\widetilde{\chi}_{Z_j}(v)\right) = \frac{-2\pi\chi_{Z_j}(v)}{1+\varepsilon_0\left\|v\right\|_2^2} + \frac{\frac{1}{2}C_s\varepsilon_0\pi\lambda^3\left\|v\right\|_2^6}{1+\varepsilon_0\left\|v\right\|_2^2} \qquad \forall\, v \in \mathbb{R}^2, \tag{3.3}$$

$$\operatorname{Im}\left(\widetilde{\chi}_{Z_j}(v)\right) = \frac{\left(\frac{\pi\alpha}{\lambda}\right)^2\left\|\nabla\chi_{Z_j}(v)\right\|_2^2}{1+\varepsilon_0\left\|v\right\|_2^2} + \frac{\frac{1}{4}(\pi\Delta\lambda)^2\left\|v\right\|_2^4}{1+\varepsilon_0\left\|v\right\|_2^2} \qquad \forall\, v \in \mathbb{R}^2, \tag{3.4}$$

assuming that only the focus and the spherical aberration are considered for the wave aberration function. Note that the numerators of the two fractions in Equation (3.4) are identical to the negated exponents of the spatial and temporal coherence envelopes in Equations (2.25) and (2.26) with $w = 0$. If $u(v) := 1 + \Delta^2\pi^2\lambda^2q_0^2\left\|v\right\|_2^2 = 1+\varepsilon_0\left\|v\right\|_2^2$ is approximated by 1 as in Equation (2.23) in the derivation of Ishizuka's TCC and the second term in Equation (3.3) is neglected, then the function $\widetilde{\chi}_{Z_j}$ reduces to

$$\widetilde{\chi}_{Z_j}(v) \approx -2\pi\chi_{Z_j}(v) + \mathrm{i}\left(\left(\frac{\pi\alpha}{\lambda}\right)^2\left\|\nabla\chi_{Z_j}(v)\right\|_2^2 + \frac{1}{4}(\pi\Delta\lambda)^2\left\|v\right\|_2^4\right) \qquad \forall\, v \in \mathbb{R}^2$$

and consequently

$$\exp\left(\mathrm{i}\widetilde{\chi}_{Z_j}(v)\right) \approx p_{Z_j}(v)E_s^{Z_j}(v,0)E_f(v,0) \qquad \forall\, v \in \mathbb{R}^2.$$

The constant $B \in \mathbb{R}$ is a Debye-Waller temperature coefficient that models thermal vibrations of the specimen during image acquisition. Strictly speaking, the corresponding

Debye-Waller factor $\exp\big(-B\,\|v\|_2^2/4\big)$ is not related to the microscope's transfer function. Nevertheless, it is added to the TCC with the intention of reducing the blurring of the exit wave caused by thermal vibrations [38, p. 154].

Finally, the registration of the TEM images is also incorporated into the TCC by means of the modulation $\mu_{t_j}(v) := \exp(2\pi \mathrm{i} v \cdot t_j)$ with $t_j \in \mathbb{R}^2$. Since the Fourier transform converts modulations to translations and vice versa, it follows from Equation (2.33) and

$$\begin{aligned}\mathcal{F}^{-1}(\Psi \star_{T_j^{\mathrm{MIMAP},1}} \Psi) = \big|\mathcal{F}^{-1}(\Psi h_j)\big|^2 = \Big|\mathcal{F}^{-1}\big(\Psi \tilde{h}_j \mu_{t_j}\big)\Big|^2 = \Big|\mathcal{F}^{-1}\big(\Psi \tilde{h}_j\big) \circ \phi_{t_j}\Big|^2 \\ = \Big|\mathcal{F}^{-1}\big(\Psi \tilde{h}_j\big)\Big|^2 \circ \phi_{t_j}\end{aligned}$$

that the modulation by μ_{t_j} indeed translates the simulated images in real space. Here, $\phi_y : \mathbb{R}^2 \to \mathbb{R}^2, x \mapsto x + y$ is the translation by $y \in \mathbb{R}^2$ and $\tilde{h}_j := h_j/\mu_{t_j}$.

Although the possibility of an objective aperture is mentioned in [38, p. 163], no aperture term is explicitly included in the function h_j. If an aperture function is added to the definition of h_j, then the properties I–IV also hold for $T_j^{\mathrm{MIMAP},1}$. Adding an aperture function poses no limitation on the practical application of the MIMAP functional, since the aperture function may be chosen equal to unity on the entire (finite) domain of the experimental TEM images.

The extension of Ishizuka's TCC: Another, more generally valid form of the TCC that is considered in [38] for the MIMAP functional is an extension of Ishizuka's TCC. It is defined as

$$T_j^{\mathrm{MIMAP},2}(v, w) := \widetilde{T}_{Z_j}(v, w) k_j(v) k_j^*(w) \qquad \forall\, v, w \in \mathbb{R}^2$$

for all $j \in \{1, \dots, N\}$, where $\widetilde{T}_{Z_j}$ denotes the last line of Equation (2.23) with $Z = Z_j$ and

$$k_j(v) := \exp\left(-\frac{B\,\|v\|_2^2}{4}\right) \exp(2\pi \mathrm{i} v \cdot t_j) \qquad \forall\, v \in \mathbb{R}^2.$$

Thus, just like the extension of the quasi-coherent TCC $T_j^{\mathrm{MIMAP},1}$, the extension of Ishizuka's TCC does not replace u by 1 and includes a Debye-Waller factor as well as the registration. The statement on the Debye-Waller factor carries over from $T_j^{\mathrm{MIMAP},1}$ to $T_j^{\mathrm{MIMAP},2}$. If an aperture function is added to k_j, then the support of $T_j^{\mathrm{MIMAP},2}$ is bounded. The modulation again acts as a translation on the simulated images in real space by Lemma 2.3.3.

If an aperture function is added, then the TCC also satisfies the properties I–III along with $T_j^{\mathrm{MIMAP},2} \in W(U)$ for an appropriate bounded set $U \subseteq \mathbb{R}^2$. However, it is not clear if the TCC also satisfies the factorization property IV, which can likely be checked with another elongated proof along the lines of Proposition 2.2.8. The factorization property of $T_j^{\mathrm{MIMAP},2}$ is not investigated in more detail here, since this TCC can be considered as an intermediate stage between the large aperture approximation and Ishizuka's TCC, both of which satisfy the factorization property.

3.1.2 The MAL functional

In the MAL algorithm, the functional

$$J^{\mathrm{MAL}}[\Psi] = \frac{1}{N} \sum_{j=1}^{N} \left\| \Psi \star_{T_{Z_j}^{\mathrm{MAL}}} \Psi - \mathcal{F}(g_j \circ \phi_{t_j}) \right\|_{L^2}^2$$

is minimized in order to reconstruct the exit wave from a focus image series [10]. The MAL algorithm therefore does not utilize a regularizer for the exit wave and it is argued that the regularizer in the MIMAP functionals obstructs convergence instead of improving it [10, p. 119]. In Section 2.2.4, it has been shown that the TCC $T_{Z_j}^{\mathrm{MAL}}$ satisfies all of the properties I–IV.

The translations ϕ_{t_j} that are applied to the experimental input images g_j for all $j = 1, \dots, N$ are not included explicitly in the MAL functional in [10]. Rather, it is stated that the registration of the image series is updated after every minimization step of the functional, which implies that the experimental input images g_j are replaced with the shifted images $g_j \circ \phi_{t_j}$.

3.2 Coercivity

The coercivity of the functionals is a crucial property if one wants to show the existence of minimizers with the direct method. Here, a function $f : V \to W$ between two normed vector spaces $(V, \|\cdot\|_V)$ and $(W, \|\cdot\|_W)$ is called coercive if and only if

$$\lim_{k \to \infty} \|f(v_k)\|_W = \infty$$

holds for all sequences $(v_k)_{k \in \mathbb{N}} \in V^{\mathbb{N}}$ with $\lim_{k\to\infty} \|v_k\|_V = \infty$.

The MIMAP functionals are clearly coercive because of the regularizer $\sigma \|\Psi - \Psi_M\|_{L^2}^2$. This is a beneficial, but likely unintentional, side-effect of the regularizer, as its original purpose was to improve the reliability of a minimizer in the case of a low signal-to-noise ratio. However, since the regularizer was removed from the functional in the MAL algorithm, it is not clear if the MAL functional is coercive as well. It is therefore natural to ask if the regularizer is necessary for the coercivity of the functional.

3.2.1 A special case: low-pass filter approximation

A partial answer to this question is given below in multiple steps, which show that the functional

$$L^2(U, \mathbb{C}) \to \mathbb{R}, \qquad \Psi \mapsto \|(\Psi p_Z) \star (\Psi p_Z) - G\|_{L^2}^2 \tag{3.5}$$

is not coercive for any $G \in L^2(\mathbb{R}^2, \mathbb{C})$ and any measurable subset $U \subseteq \mathbb{R}^2$ with non-empty interior. This functional is identical to the MAL functional in the case of perfectly coherent illumination and only a single experimental input image.

The intuition behind the non-coercivity of Equation (3.5) is given by considering the functional $J[f] = \|f \star f\|_{L^2}$ with $f \in L^2(\mathbb{R}^d, \mathbb{C})$, $d \in \mathbb{N}$.

Lemma 3.2.1. *The functional*

$$J : L^2(\mathbb{R}^d, \mathbb{C}) \to \mathbb{R}, \quad f \mapsto \|f \star f\|_{L^2}$$

is not coercive for all $d \in \mathbb{N}$.

Proof. Define the functions

$$g_\delta : \mathbb{R}^d \to \mathbb{R}, \quad x \mapsto \begin{cases} \prod_{i=1}^d \frac{1}{\sqrt{x_i}}, & \text{if } 1 \leq x_i \leq \delta \ \forall i \in \{1, \ldots, d\}, \\ 0, & \text{otherwise} \end{cases}$$

for all $\delta > 1$. Then

$$\|g_\delta\|_{L^2} = \sqrt{\int_{[1,\delta]^d} \prod_{i=1}^d \frac{1}{x_i} \, \mathrm{d}x} = \sqrt{\left(\int_1^\delta \frac{1}{y} \, \mathrm{d}y \right)^d} = \sqrt{\ln(\delta)^d} \xrightarrow[\delta \to \infty]{} \infty$$

and

$$\|g_\delta^2\|_{L^2} = \sqrt{\int_{[1,\delta]^d} \prod_{i=1}^d \frac{1}{x_i^2} \, \mathrm{d}x} = \sqrt{\left(\int_1^\delta \frac{1}{y^2} \, \mathrm{d}y \right)^d} = \sqrt{\left(1 - \frac{1}{\delta} \right)^d} \xrightarrow[\delta \to \infty]{} 1 < \infty.$$

If we set $f_\delta := \mathcal{F}(g_\delta)$ for all $\delta > 1$, then $f_\delta \in L^2(\mathbb{R}^d, \mathbb{C})$ and it follows that

$$J[f_\delta] = \|f_\delta \star f_\delta\|_{L^2} = \left\| \left| \mathcal{F}^{-1} f_\delta \right|^2 \right\|_{L^2} = \|g_\delta^2\|_{L^2} \xrightarrow[\delta \to \infty]{} 1,$$

but $\|f_\delta\|_{L^2} = \|g_\delta\|_{L^2} \xrightarrow[\delta \to \infty]{} \infty$, since the Fourier transform is unitary. Therefore J is not coercive. ☐

In order to show that the functional in Equation (3.5) is not coercive, the above observation is generalized to

$$L^2(U, \mathbb{C}) \to \mathbb{R}, \qquad f \mapsto \|f \star f\|_{L^2} \tag{3.6}$$

for measurable sets $U \subseteq \mathbb{R}^d$ with non-empty interior. This implies that the functional in Equation (3.5) is not coercive, since $p_Z(v) \neq 0$ for all $v \in \mathbb{R}^2$ and

$$\|(\Psi p_Z) \star (\Psi p_Z) - G\|_{L^2}^2 \leq 2 \, \|(\Psi p_Z) \star (\Psi p_Z)\|_{L^2}^2 + 2 \, \|G\|_{L^2}^2$$

for all $\Psi \in L^2(A, \mathbb{C})$ and all $G \in L^2(\mathbb{R}^2, \mathbb{C})$.

Since the support of g_δ in Lemma 3.2.1 is bounded for all $\delta > 1$, the support of its Fourier transform $f_\delta = \mathcal{F}(g_\delta)$ is necessarily unbounded. Therefore, it is not directly possible to apply the proof of Lemma 3.2.1 to a functional with the domain $L^2(U, \mathbb{C})$ for measurable sets $U \neq \mathbb{R}^d$. In order to construct a suitable sequence of functions with bounded support, the functions f_δ are multiplied with the characteristic function

$\mathbb{1}_{[-\frac{1}{2},\frac{1}{2}]}$. It is well known that $\mathcal{F}(\operatorname{sinc}) = \mathbb{1}_{[-\frac{1}{2},\frac{1}{2}]}$, where sinc is the normalized cardinal sine function defined as

$$\operatorname{sinc}(x) = \begin{cases} \frac{\sin(\pi x)}{\pi x}, & \text{if } x \neq 0, \\ 1, & \text{if } x = 0. \end{cases}$$

By the convolution theorem, the multiplication of f_δ with the characteristic function $\mathbb{1}_{[-\frac{1}{2},\frac{1}{2}]}$ corresponds to a convolution of g_δ with sinc, since

$$\mathcal{F}(g_\delta * \operatorname{sinc}) = \mathcal{F}(g_\delta)\mathcal{F}(\operatorname{sinc}) = f_\delta \mathbb{1}_{[-\frac{1}{2},\frac{1}{2}]}.$$

In other words, convolving g_δ with sinc corresponds to applying an ideal low-pass filter with cutoff frequency $\frac{1}{2}$ to g_δ.

It remains to show that $g_\delta * \operatorname{sinc}$ is sufficiently "close" to g_δ in the sense of

$$\lim_{\delta\to\infty} \|g_\delta * \operatorname{sinc}\|_{L^2} = \infty \qquad \text{and} \qquad \lim_{\delta\to\infty} \left\|(g_\delta * \operatorname{sinc})^2\right\|_{L^2} < \infty.$$

The central tool to this end is Lemma 3.2.2, which provides a method to estimate how well a scalar function $g : \mathbb{R} \to \mathbb{R}$ is approximated by $g * k$ for an appropriate function $k : \mathbb{R} \to \mathbb{R}$.

Interval arithmetic is utilized in order to compute bounds for the real-valued functions in Lemma 3.2.2. The sum and the difference of two intervals $I_1 := [a, b]$ and $I_2 := [c, d]$ are defined as

$$\begin{aligned} I_1 + I_2 &:= \{x + y \mid x \in I_1,\ y \in I_2\} = [a + c, b + d], \\ I_1 - I_2 &:= \{x - y \mid x \in I_1,\ y \in I_2\} = [a - d, b - c] \end{aligned}$$

for all $a, b, c, d \in \mathbb{R}$ with $a \leq b$ and $c \leq d$. Furthermore, a scalar $e \in \mathbb{R}$ is identified with the interval $[e, e]$.

Lemma 3.2.2 ([14, Lemma B.1])**.** *Let $g : \mathbb{R} \to \mathbb{R}$ be monotonic, bounded and positive, $k \in L^1_{\mathrm{loc}}(\mathbb{R}, \mathbb{R})$ and $N > 0$. If there are constants $\alpha_N, \beta_N > 0$ such that*

- $\left|\int_I k(y)\,\mathrm{d}y\right| \leq \alpha_N$ *for all intervals $I \subseteq [-N, N]$,*
- $\left|\int_I k(y)\,\mathrm{d}y\right| \leq \beta_N$ *for all bounded intervals $I \subseteq \mathbb{R}\backslash[-N, N]$,*

then

$$\big(g * k\big)(x) \in \delta_N g(x - N) + |g(x + N) - g(x - N)|\,[-\alpha_N, \alpha_N] + 4\,\|g\|_\infty\,[-\beta_N, \beta_N],$$

where $\delta_N = \int_{[-N,N]} k(y)\,\mathrm{d}y$, holds for all $x \in \mathbb{R}$.

Proof. Let $x \in \mathbb{R}$ and split the integral into three parts,

$$\big(g * k\big)(x) = \int_{-N}^{N} g(x - y)k(y)\,\mathrm{d}y + \int_{N}^{\infty} g(x - y)k(y)\,\mathrm{d}y + \int_{-\infty}^{-N} g(x - y)k(y)\,\mathrm{d}y.$$

By the second mean value theorem of integral calculus [31], there exists a $B \in [-N, N]$ with

$$\begin{aligned}\int_{-N}^{N} g(x-y)k(y)\,\mathrm{d}y &= g(x+N)\int_{-N}^{B} k(y)\,\mathrm{d}y + g(x-N)\int_{B}^{N} k(y)\,\mathrm{d}y \\ &= g(x+N)\int_{-N}^{B} k(y)\,\mathrm{d}y + g(x-N)\left(\delta_N - \int_{-N}^{B} k(y)\,\mathrm{d}y\right) \\ &= \big(g(x+N) - g(x-N)\big)\int_{-N}^{B} k(y)\,\mathrm{d}y + \delta_N g(x-N) \\ &\in |g(x+N) - g(x-N)|\,[-\alpha_N, \alpha_N] + \delta_N g(x-N).\end{aligned}$$

Bounds for the other two integrals can also be found using the second mean value theorem: for every $M > N$ there exists a $Q_M \in [N, M]$ with

$$\begin{aligned}\int_{N}^{\infty} g(x-y)k(y)\,\mathrm{d}y &= \lim_{M\to\infty}\int_{N}^{M} g(x-y)k(y)\,\mathrm{d}y \\ &= \lim_{M\to\infty} g(x-N)\int_{N}^{Q_M} k(y)\,\mathrm{d}y + g(x-M)\int_{Q_M}^{M} k(y)\,\mathrm{d}y \\ &\in 2\,\|g\|_\infty\,[-\beta_N, \beta_N]\end{aligned}$$

and similarly for the third integral. ☐

As mentioned above, the normalized cardinal sine function $k := \operatorname{sinc}$ is of particular interest as the convolution kernel. If $I = [c, d] \subset \mathbb{R}$ is a bounded interval, then the integral $\int_I k(y)\,\mathrm{d}y$ can be expressed in terms of the sine integral $\mathrm{Si}(x) = \int_0^x \sin(y)/y\,\mathrm{d}y$ by means of

$$\int_I k(y)\,\mathrm{d}y = \frac{1}{\pi}\int_{\pi c}^{\pi d}\frac{\sin(y)}{y}\,\mathrm{d}y = \frac{1}{\pi}\big(\mathrm{Si}(\pi d) - \mathrm{Si}(\pi c)\big).$$

By rewriting the integral in this way, suitable values for α_N and β_N as well as a bound on δ_N can be determined from the properties of Si.

It is well known that $|\mathrm{Si}(x)| \leq \mathrm{Si}(\pi) \leq 2$ for all $x \in \mathbb{R}$. Moreover, we have $\left|\mathrm{Si}(x) - \frac{\pi}{2}\right| \leq \frac{1}{x}$ for all $x > 0$, which can be shown as follows [12]: the equation

$$\frac{\pi}{2} - \mathrm{Si}(x) = \int_x^\infty \frac{\sin(t)}{t}\,\mathrm{d}t = \int_0^\infty \frac{\sin(x+t)}{x+t}\,\mathrm{d}t = \int_0^\infty \frac{\sin(x)\cos(t) + \cos(x)\sin(t)}{x+t}\,\mathrm{d}t$$

holds for all $x > 0$, since $\int_0^\infty \frac{\sin(t)}{t}\,\mathrm{d}t = \frac{\pi}{2}$. Similar to Parseval's theorem for the Fourier transform, a useful property of the Laplace transform is the fact that the Laplace transform $\mathcal{L}$ can be applied to the integrand as $\int_0^\infty f(x)g(x)\,\mathrm{d}x = \int_0^\infty \mathcal{L}(f)(y)\mathcal{L}^{-1}(g)(y)\,\mathrm{d}y$. Applying the Laplace transform to the numerator and the inverse Laplace transform to $(x+t)^{-1}$ yields [46]

$$\begin{aligned}\int_0^\infty \frac{\sin(x)\cos(t) + \cos(x)\sin(t)}{x+t}\,\mathrm{d}t &= \int_0^\infty \left(\sin(x)\frac{s}{s^2+1} + \cos(x)\frac{1}{s^2+1}\right)e^{-sx}\,\mathrm{d}s \\ &= \int_0^\infty \frac{(\sin(x), \cos(x)) \cdot (s, 1)}{s^2+1}e^{-sx}\,\mathrm{d}s\end{aligned}$$

and, using the Cauchy-Schwarz inequality, it follows that

$$\begin{aligned}\left|\mathrm{Si}(x)-\frac{\pi}{2}\right| &\le \int_0^\infty \frac{\|(\sin(x),\cos(x)\|_2\,\|(s,1)\|_2}{s^2+1}e^{-sx}\,\mathrm{d}s\\ &= \int_0^\infty \frac{1}{\sqrt{s^2+1}}e^{-sx}\,\mathrm{d}s < \int_0^\infty e^{-sx}\,\mathrm{d}s = \frac{1}{x}.\end{aligned}$$

A direct consequence of this inequality is that $\left|\mathrm{Si}(x)+\frac{\pi}{2}\right| \le \frac{1}{-x}$ holds for all $x<0$ because of the point symmetry of the sine integral.

Now fix an arbitrary $N>0$. If $I=[c,d]\subseteq[-N,N]$, then

$$\left|\int_I k(y)\,\mathrm{d}y\right| = \frac{1}{\pi}\left|\mathrm{Si}(\pi d)-\mathrm{Si}(\pi c)\right| \le \frac{1}{\pi}(2+2) = \frac{4}{\pi} =: \alpha_N,$$

On the other hand, if we consider an interval $I=[c,d]\subseteq(N,\infty)$, then

$$\begin{aligned}\left|\int_I k(y)\,\mathrm{d}y\right| &= \frac{1}{\pi}\left|\mathrm{Si}(\pi d)-\mathrm{Si}(\pi c)\right| = \frac{1}{\pi}\left|\Big(\mathrm{Si}(\pi d)-\frac{\pi}{2}\Big)-\Big(\mathrm{Si}(\pi c)-\frac{\pi}{2}\Big)\right|\\ &\le \frac{1}{\pi}\left(\frac{1}{\pi d}+\frac{1}{\pi c}\right) \le \frac{2}{\pi^2 N} =: \beta_N\end{aligned}$$

and similarly $\left|\int_I k(y)\,\mathrm{d}y\right| \le \frac{2}{\pi^2 N} = \beta_N$ for $I=[c,d]\subseteq(-\infty,-N)$. The value of δ_N is bounded by

$$|\delta_N-1| = \left|\int_{[-N,N]} k(y)\,\mathrm{d}y - 1\right| = \left|\frac{2}{\pi}\,\mathrm{Si}(\pi N)-1\right| \le \frac{2}{\pi}\frac{1}{\pi N} = \frac{2}{\pi^2 N}.$$

Lemma 3.2.3 ([14, Lemma B.2]). *The functional* $f\mapsto \|f\star f\|_{L^2}$ *with* $f\in L^2(I,\mathbb{C})$ *is not coercive for any proper interval* $I\subseteq\mathbb{R}$.

Proof. For $\delta>1$ let $g_\delta:\mathbb{R}\to\mathbb{R}$ with $g_\delta(x)=\frac{1}{\sqrt{x}}$ for all $x\in[1,\delta]$ and $g_\delta(x)=0$ for all $x\in\mathbb{R}\backslash[1,\delta]$. The function g_δ is neither monotonic nor positive but it can be written as the difference of two monotonic, bounded and positive functions $g_{\delta,1}$ and $g_{\delta,2}$, e.g.

$$g_{\delta,1}(x)=\begin{cases}1, & \text{if } x<1,\\ 2, & \text{if } x\ge 1\end{cases} \qquad \text{and} \qquad g_{\delta,2}(x)=\begin{cases}1, & \text{if } x<1,\\ 2-\frac{1}{\sqrt{x}}, & \text{if } 1\le x\le\delta,\\ 2, & \text{if } x>\delta.\end{cases}$$

Applying Lemma 3.2.2 to $g_{\delta,j}$ for $j\in\{1,2\}$ and using the values for α_N and β_N derived above as well as the bound on δ_N yields bounds $\big(g_{\delta,j}*k\big)(x)\in[L_{\delta,j,N}(x),U_{\delta,j,N}(x)]$ for all $x\in\mathbb{R}$ and all $N>0$, where

$$\begin{aligned}L_{\delta,j,N}(x) &:= g_{\delta,j}(x-N)\left(1-\frac{2}{\pi^2 N}\right) - |g_{\delta,j}(x+N)-g_{\delta,j}(x-N)|\,\frac{4}{\pi} - \frac{16}{\pi^2 N},\\ U_{\delta,j,N}(x) &:= g_{\delta,j}(x-N)\left(1+\frac{2}{\pi^2 N}\right) + |g_{\delta,j}(x+N)-g_{\delta,j}(x-N)|\,\frac{4}{\pi} + \frac{16}{\pi^2 N}.\end{aligned}$$

Therefore, $g_\delta * k$ is bounded by

$$\begin{aligned}\big(g_\delta * k\big)(x) &= \big(g_{\delta,1} * k\big)(x) - \big(g_{\delta,2} * k\big)(x)\\ &\in [L_{\delta,1,N}(x) - U_{\delta,2,N}(x), U_{\delta,1,N}(x) - L_{\delta,2,N}(x)]\end{aligned}$$

for all $x \in \mathbb{R}$ and all $N > 0$. In particular, choosing $x = 2N$ yields the lower bounds

$$\big(g_\delta * k\big)(2N) \geq \begin{cases} \left(1 - \frac{12-4\sqrt{3}}{3\pi}\right) \frac{1}{\sqrt{N}} - \frac{40}{\pi^2}\frac{1}{N} + \frac{2}{\pi^2}\frac{1}{N\sqrt{N}}, & \text{if } N \in \left[1, \frac{\delta}{3}\right], \\ \left(1 - \frac{4}{\pi}\right) \frac{1}{\sqrt{N}} - \frac{40}{\pi^2}\frac{1}{N} + \frac{2}{\pi^2}\frac{1}{N\sqrt{N}}, & \text{if } N \in \left(\frac{\delta}{3}, \delta\right], \\ -\frac{40}{\pi^2 N}, & \text{if } N \in (\delta, \infty) \end{cases}$$

and the upper bounds

$$\big(g_\delta * k\big)(2N) \leq \begin{cases} \left(1 + \frac{12-4\sqrt{3}}{3\pi}\right) \frac{1}{\sqrt{N}} + \frac{40}{\pi^2}\frac{1}{N} - \frac{2}{\pi^2}\frac{1}{N\sqrt{N}}, & \text{if } N \in \left[1, \frac{\delta}{3}\right], \\ \left(1 + \frac{4}{\pi}\right) \frac{1}{\sqrt{N}} + \frac{40}{\pi^2}\frac{1}{N} - \frac{2}{\pi^2}\frac{1}{N\sqrt{N}}, & \text{if } N \in \left(\frac{\delta}{3}, \delta\right], \\ \frac{40}{\pi^2 N}, & \text{if } N \in (\delta, \infty) \end{cases}$$

for all $N > 1$. A bound on $\big(g_\delta * k\big)(x)$ for negative values of x can be found similarly by choosing $x = -N$, which yields $\big(g_\delta * k\big)(-N) \in \left[\frac{-36}{\pi^2 N}, \frac{36}{\pi^2 N}\right]$ for all $N > 0$. Furthermore, $g_\delta * k$ is uniformly bounded by

$$\left\|g_\delta * k\right\|_\infty \leq \left\|g_\delta\right\|_{L^4} \left\|k\right\|_{L^{4/3}} \leq \left\|g_\infty\right\|_{L^4} \left\|k\right\|_{L^{4/3}} < \infty$$

for all $\delta > 1$ by Hölder's inequality. Using these bounds and the fact that the coefficient $1 - \frac{12-4\sqrt{3}}{3\pi}$ of $\frac{1}{\sqrt{N}}$ in the lower bound of $\big(g_\delta * k\big)(2N)$ is positive, it follows that

- $\exists\, c_1, x_{\min} > 0 :\ \big(g_\delta * k\big)(x) \geq c_1 \frac{1}{\sqrt{x}}$ for all $x \in \left[x_{\min}, \frac{\delta}{3}\right]$ and all $\delta > 1$.
- $\exists\, c_2 > 0 :\ \big(g_\delta * k\big)(x) \leq c_2 \frac{1}{\sqrt{x}}$ for all $|x| \geq 1$ and all $\delta > 1$. Furthermore, we have $\big(g_\delta * k\big)(x) \leq \left\|g_\infty\right\|_{L^4} \left\|k\right\|_{L^{4/3}}$ for all $|x| < 1$ and all $\delta > 1$.

This implies

$$\lim_{\delta\to\infty} \left\|g_\delta * k\right\|_{L^2} = \infty \qquad \text{and} \qquad \lim_{\delta\to\infty} \left\|(g_\delta * k)^2\right\|_{L^2} < \infty.$$

Without loss of generality, we assume that $I = \left[-\frac{1}{2}, \frac{1}{2}\right]$. Let $f_\delta := (\mathcal{F} g_\delta)\mathbb{1}_{[-\frac{1}{2},\frac{1}{2}]}$ for all $\delta > 1$. Then $f_\delta \in L^2\left(\left[-\frac{1}{2}, \frac{1}{2}\right], \mathbb{C}\right)$ and

$$\begin{aligned}\lim_{\delta\to\infty} \left\|f_\delta\right\|_{L^2} &= \lim_{\delta\to\infty} \left\|\mathcal{F}^{-1} f_\delta\right\|_{L^2} = \lim_{\delta\to\infty} \left\|g_\delta * k\right\|_{L^2} = \infty,\\ \lim_{\delta\to\infty} \left\|f_\delta \star f_\delta\right\|_{L^2} &= \lim_{\delta\to\infty} \left\|\left(\mathcal{F}^{-1} f_\delta\right)^2\right\|_{L^2} = \lim_{\delta\to\infty} \left\|(g_\delta * k)^2\right\|_{L^2} < \infty.\end{aligned}$$

□

Next, the result for the one-dimensional case in Lemma 3.2.3 is generalized to arbitrary subsets $U \subseteq \mathbb{R}^d$ with non-empty interior, which completes the proof of the non-coercivity of the functional in Equation (3.6).

Corollary 3.2.4 ([14, Lemma B.3]). *The functional $f \mapsto \|f \star f\|_{L^2}$ with $f \in L^2(U,\mathbb{C})$ is not coercive for any measurable set $U \subseteq \mathbb{R}^d$ with non-empty interior.*

Proof. Since the interior of U is non-empty, there are $a, b \in \mathbb{R}$ with $a < b$ such that $[a,b]^d \subseteq U$. By Lemma 3.2.3 there is a sequence $(f_n)_{n\in\mathbb{N}}$ of functions $f_n \in L^2([a,b],\mathbb{C})$ with $\lim_{n\to\infty} \|f_n\|_{L^2} = \infty$ and $\lim_{n\to\infty} \|f_n \star f_n\|_{L^2} = 1$. Consider the sequence $(h_n)_{n\in\mathbb{N}}$ given by

$$h_n : U \to \mathbb{C}, \qquad x \mapsto \begin{cases} f_n(x_1), & \text{if } x \in [a,b]^d, \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$\|h_n\|_{L^2}^2 = \int_{[a,b]^d} |f_n(x_1)|^2 \,\mathrm{d}x = (b-a)^{d-1} \|f_n\|_{L^2}^2 \xrightarrow[n\to\infty]{} \infty$$

and

$$\begin{aligned} \left|\big(h_n \star h_n\big)(x)\right| &= \left|\int_{\mathbb{R}^d} h_n(y) h_n(x+y)\,\mathrm{d}y\right| = \left|\int_{[a,b]^d \cap ([a,b]^d - x)} f_n(y_1) f_n(y_1+x_1)\,\mathrm{d}y\right| \\ &\le \begin{cases} (b-a)^{d-1} \left|\big(f_n \star f_n\big)(x_1)\right|, & \text{if } x \in B^\infty_{b-a}(0), \\ 0, & \text{otherwise} \end{cases} \end{aligned}$$

for all $x \in \mathbb{R}^d$, where $B^\infty_{b-a}(0) = \big\{x \in \mathbb{R}^d : \|x\|_\infty < b-a\big\}$. This implies

$$\begin{aligned} \|h_n \star h_n\|_{L^2}^2 &\le (b-a)^{2d-2} \int_{B^\infty_{b-a}(0)} \left|\big(f_n \star f_n\big)(x_1)\right|^2 \,\mathrm{d}x \\ &\le (b-a)^{2d-2}(b-a)^{d-1} \int_{[-(b-a),b-a]} \left|\big(f_n \star f_n\big)(x_1)\right|^2 \,\mathrm{d}x \\ &= (b-a)^{3d-3} \|f_n \star f_n\|_{L^2}^2 \xrightarrow[n\to\infty]{} (b-a)^{3d-3} < \infty. \end{aligned}$$

□

3.2.2 The general case

Let $U \subseteq \mathbb{R}^2$ be measurable with non-empty interior. The negative result from the previous section suggests that the MAL functional is not coercive even in the general case of partially coherent illumination and multiple input images. As a generalization of the results from Section 3.2.1, it is shown that the functionals

$$L^2(U,\mathbb{C}) \to \mathbb{R}, \qquad \Psi \mapsto \sum_{j=1}^N \left\|\Psi \star_{T_{Z_j}} \Psi - G_j\right\|_{L^2}^2 \tag{3.7}$$

are not coercive regardless of the choice of $N \in \mathbb{N}$, $G_j \in L^2(\mathbb{R}^2,\mathbb{C})$ and $Z_j \in \mathbb{R}$, given that $T_{Z_j} \in W(U)$ is a $\star$-separable TCC. Since the TCC T_Z^{MAL} is $\star$-separable, it follows that the MAL functional is not coercive.

First, note that it is sufficient to consider the functional

$$J : L^2(U,\mathbb{C}) \to \mathbb{R}, \qquad \Psi \mapsto \sum_{j=1}^N \left\|\Psi \star_{T_{Z_j}} \Psi\right\|_{L^2}^2 \tag{3.8}$$

instead of Equation (3.7), since

$$\sum_{j=1}^{N} \left\| \Psi \star_{T_{Z_j}} \Psi - G_j \right\|_{L^2}^2 \leq 2 \sum_{j=1}^{N} \left\| \Psi \star_{T_{Z_j}} \Psi \right\|_{L^2}^2 + 2 \sum_{j=1}^{N} \| G_j \|_{L^2}^2$$

holds for all $\Psi \in L^2(U, \mathbb{C})$. Thus, if $(\Psi_n)_{n\in\mathbb{N}} \in L^2(U, \mathbb{C})^{\mathbb{N}}$ is a sequence with $\|\Psi_n\|_{L^2} \to \infty$ for $n \to \infty$ and $\liminf_{n\to\infty} J[\Psi_n] < \infty$, then $(\Psi_n)_{n\in\mathbb{N}}$ also satisfies

$$\liminf_{n\to\infty} \sum_{j=1}^{N} \left\| \Psi_n \star_{T_{Z_j}} \Psi_n - G_j \right\|_{L^2}^2 < \infty.$$

Proposition 3.2.5. *Let $N \in \mathbb{N}$ and $Z_j \in \mathbb{R}$ for all $j \in \{1, \ldots, N\}$. Furthermore, let $U \subseteq \mathbb{R}^2$ be a measurable set with non-empty interior and $T_Z \in W(U)$ a $\star$-separable TCC with*

$$T_Z(v, w) = \sum_{k=1}^{K} h_{Z,k}(v) h_{Z,k}^*(w) \qquad \forall\, v, w \in U \tag{3.9}$$

for a suitable $K \in \mathbb{N}$ and suitable measurable and bounded functions $h_{Z,k} : \mathbb{R}^2 \to \mathbb{C}$. If $\mathcal{F}^{-1}(h_{Z_j,k}) \in L^1(\mathbb{R}^2, \mathbb{C})$ for all $j \in \{1, \ldots, N\}$ and all $k \in \{1, \ldots, K\}$, then the functional

$$J : L^2(U, \mathbb{C}) \to \mathbb{R}, \qquad \Psi \mapsto \sum_{j=1}^{N} \left\| \Psi \star_{T_{Z_j}} \Psi \right\|_{L^2}^2$$

is not coercive.

Proof. Let $\Psi \in L^2(U, \mathbb{C})$ and $\psi := \mathcal{F}^{-1}(\Psi)$. By the decomposition of the TCC in Equation (3.9), the convolution theorem, and the fact that the Fourier transform is unitary, we have

$$J[\Psi] = \sum_{j=1}^{N} \left\| \sum_{k=1}^{K} (\Psi h_{Z_j,k}) \star (\Psi h_{Z_j,k}) \right\|_{L^2}^2 = \sum_{j=1}^{N} \left\| \sum_{k=1}^{K} \left| \mathcal{F}^{-1}(\Psi h_{Z_j,k}) \right|^2 \right\|_{L^2}^2 .$$

Rewriting the squared L^2-norm as an inner product and applying the Cauchy-Schwarz inequality, it follows that

$$\begin{aligned} J[\Psi] &= \sum_{j=1}^{N} \sum_{k,l=1}^{K} \left(\left| \mathcal{F}^{-1}(\Psi h_{Z_j,k}) \right|^2, \left| \mathcal{F}^{-1}(\Psi h_{Z_j,l}) \right|^2 \right)_{L^2} \\ &\leq \sum_{j=1}^{N} \sum_{k,l=1}^{K} \left\| \left| \mathcal{F}^{-1}(\Psi h_{Z_j,k}) \right|^2 \right\|_{L^2} \left\| \left| \mathcal{F}^{-1}(\Psi h_{Z_j,l}) \right|^2 \right\|_{L^2} . \end{aligned}$$

Applying the convolution theorem once more as well as the “counterpart” of Lemma 2.1.3 for convolutions, we get

$$\begin{aligned}
J[\Psi] &\leq \sum_{j=1}^{N} \sum_{k,l=1}^{K} \left\| \left| \psi * \mathcal{F}^{-1}(h_{Z_j,k}) \right|^2 \right\|_{L^2} \left\| \left| \psi * \mathcal{F}^{-1}(h_{Z_j,l}) \right|^2 \right\|_{L^2} \\
&= \sum_{j=1}^{N} \sum_{k,l=1}^{K} \left\| \psi * \mathcal{F}^{-1}(h_{Z_j,k}) \right\|_{L^4}^2 \left\| \psi * \mathcal{F}^{-1}(h_{Z_j,l}) \right\|_{L^4}^2 \\
&\leq \sum_{j=1}^{N} \sum_{k,l=1}^{K} \|\psi\|_{L^4}^2 \left\| \mathcal{F}^{-1}(h_{Z_j,k}) \right\|_{L^1}^2 \|\psi\|_{L^4}^2 \left\| \mathcal{F}^{-1}(h_{Z_j,l}) \right\|_{L^1}^2 \\
&= \left\| |\psi|^2 \right\|_{L^2}^2 \sum_{j=1}^{N} \sum_{k,l=1}^{K} \left\| \mathcal{F}^{-1}(h_{Z_j,k}) \right\|_{L^1}^2 \left\| \mathcal{F}^{-1}(h_{Z_j,l}) \right\|_{L^1}^2 ,
\end{aligned}$$

assuming that $\psi = \mathcal{F}^{-1}(\Psi) \in L^4(\mathbb{R}^2, \mathbb{C})$ for now. Since $\mathcal{F}^{-1}(h_{Z_j,k})$ is integrable for all j, k by assumption, there exists a constant $C > 0$ with

$$\sum_{j=1}^{N} \sum_{k,l=1}^{K} \left\| \mathcal{F}^{-1}(h_{Z_j,k}) \right\|_{L^1}^2 \left\| \mathcal{F}^{-1}(h_{Z_j,l}) \right\|_{L^1}^2 = C < \infty,$$

which shows that $J[\Psi] \leq C \left\| |\psi|^2 \right\|_{L^2}^2$.

By Corollary 3.2.4, there exists a sequence $(\Psi_n)_{n\in\mathbb{N}} \in L^2(U, \mathbb{C})^{\mathbb{N}}$ with $\|\Psi_n\|_{L^2} \to \infty$ as $n \to \infty$ and $\lim_{n\to\infty} \|\Psi_n \star \Psi_n\|_{L^2} < \infty$. If we define $\psi_n := \mathcal{F}^{-1}(\Psi_n)$ for all $n \in \mathbb{N}$, then

$$\lim_{n\to\infty} \left\| |\psi_n|^2 \right\|_{L^2} = \lim_{n\to\infty} \left\| \mathcal{F}^{-1}(\Psi_n \star \Psi_n) \right\|_{L^2} = \lim_{n\to\infty} \|\Psi_n \star \Psi_n\|_{L^2} < \infty.$$

The same line of reasoning shows $\left\| |\psi_n|^2 \right\|_{L^2} = \|\Psi_n \star \Psi_n\|_{L^2} < \infty$ and consequently $\psi_n \in L^4(\mathbb{R}^2, \mathbb{C})$ for all $n \in \mathbb{N}$. Finally, it follows that

$$\liminf_{n\to\infty} J[\Psi_n] \leq \lim_{n\to\infty} C \left\| |\psi_n|^2 \right\|_{L^2}^2 < \infty,$$

although $\lim_{n\to\infty} \|\Psi_n\|_{L^2} = \infty$. Thus J is not coercive. □

The assumption on $\mathcal{F}^{-1}(h_{Z_j,k})$ being an integrable function for all $j \in \{1, \dots, N\}$ and all $k \in \{1, \dots, K\}$ is for instance satisfied if $h_{Z_j,k} \in C_c^4(\mathbb{R}^2, \mathbb{C})$. If the aperture function is four times continuously differentiable, then this is true for all of the $\star$-separable TCCs in Sections 2.2.4 to 2.2.6. An explicit upper bound for the L^1-norm of $\mathcal{F}^{-1}(h_{Z_j,k})$ can then be derived as follows:

Since $h_{Z_j,k} \in C_c^4(\mathbb{R}^2, \mathbb{C})$, we have

$$\left(1 + x_1^4 + x_2^4\right) \mathcal{F}^{-1}\left(h_{Z_j,k}\right)(x) = \mathcal{F}^{-1} \left(\left(1 + \frac{\partial^4}{\partial x_1^4} + \frac{\partial^4}{\partial x_2^4} \right) h_{Z_j,k} \right)(x)$$

for all $x \in \mathbb{R}^2$. This implies

$$\left|\mathcal{F}^{-1}\big(h_{Z_j,k}\big)(x)\right| = \frac{\mathcal{F}^{-1}\left(\left(1 + \frac{\partial^4}{\partial x_1^4} + \frac{\partial^4}{\partial x_2^4}\right) h_{Z_j,k}\right)(x)}{1 + x_1^4 + x_2^4} \leq \frac{\left\|\left(1 + \frac{\partial^4}{\partial x_1^4} + \frac{\partial^4}{\partial x_2^4}\right) h_{Z_j,k}\right\|_{L^1}}{1 + x_1^4 + x_2^4}$$

for all $x \in \mathbb{R}^2$ since the Fourier transform of an integrable function f is uniformly bounded by the L^1 norm of f. Therefore it follows that

$$\left\|\mathcal{F}^{-1}(h_{Z_j,k})\right\|_{L^1} \leq \left\|\left(1 + \frac{\partial^4}{\partial x_1^4} + \frac{\partial^4}{\partial x_2^4}\right) h_{Z_j,k}\right\|_{L^1} \int_{\mathbb{R}^2} \frac{1}{1 + x_1^4 + x_2^4} \,\mathrm{d}x < \infty.$$

Unfortunately, there is no direct way to generalize the proof of Proposition 3.2.5 to TCCs $T_Z \in W^+(U)$ that are not $\star$-separable. In order to see this, let $T_Z \in W^+(U)$ be one of the TCCs from Sections 2.2.1 to 2.2.3 and $(T_{Z,K})_{K\in\mathbb{N}}$ the corresponding sequence of $\star$-separable functions converging uniformly to T_Z with

$$T_{Z,K}(v,w) = \sum_{k=1}^{K} h_{Z,K,k}(v) h^*_{Z,K,k}(w) \qquad \forall\, v, w \in U$$

for all $K \in \mathbb{N}$, where $h_{Z,K,k} : \mathbb{R}^2 \to \mathbb{C}$ are the functions given in the respective section. The straightforward generalization of Proposition 3.2.5 would require the existence of a constant $C' < \infty$ such that

$$\left\|\left(1 + \frac{\partial^4}{\partial x_1^4} + \frac{\partial^4}{\partial x_2^4}\right) h_{Z_j,K,k}\right\|_{L^1} \leq C' K^{-\frac{1}{2}} \qquad \forall\, k \in \{1, \dots, K\}$$

holds for all $K \in \mathbb{N}$, which then implied that there is a constant $C < \infty$ with

$$\sum_{j=1}^{N} \sum_{k,l=1}^{K} \left\|\mathcal{F}^{-1}(h_{Z_j,K,k})\right\|_{L^1}^2 \left\|\mathcal{F}^{-1}(h_{Z_j,K,l})\right\|_{L^1}^2 \leq C < \infty$$

for all $K \in \mathbb{N}$. However, since the partial derivatives of the pure phase transfer function are unbounded, it seems unlikely that such a constant C' exists.

3.3 Joint exit wave reconstruction and image registration

Due to the close connection of exit wave reconstruction and the registration of the image series, it is natural to consider solving both problems simultaneously instead of alternatingly. In this section, a novel objective functional for exit wave reconstruction is proposed, which combines these two problems into a single functional. This way, it is no longer necessary to switch between the reconstruction and the registration during every iteration, which allows us to investigate the mathematical properties of the algorithm as a whole more easily.

It should be noted that most of the parts necessary for this joint approach are already apparent in the MIMAP functionals. However, in the MIMAP algorithm the minimization of the functional is still performed alternatingly with respect to the exit wave and the translations.

In the entire section, T stands for any of the TCCs from Section 2.2. The notation from the previous chapters is adopted in this section. Furthermore, A denotes the interior of the support of the aperture function, $B_{r_a}(0)$, if T is any of the TCCs from Sections 2.2.2 to 2.2.6 and $A := B_{r_s+r_a}(0)$ if T is the full TCC from Section 2.2.1.

Definition 3.3.1. *Let $N \in \mathbb{N}$ and $g_1, \ldots, g_N \in L^1(\mathbb{R}^2, \mathbb{R}_{\geq 0}) \cap L^2(\mathbb{R}^2, \mathbb{R}_{\geq 0})$ be a series of real space TEM images with the corresponding focus values $Z_1, \ldots, Z_N \in \mathbb{R}$ such that $\operatorname{supp}(\mathcal{F}(g_j)) \subseteq \overline{2A}$ for all $j \in \{1, \ldots, N\}$. The objective functional for joint reconstruction and registration is defined as*

$$E_\sigma : L^2(A, \mathbb{C}) \times (\mathbb{R}^2)^N \to \mathbb{R}, \quad (\Psi, t) \mapsto \frac{1}{N} \sum_{j=1}^{N} \left\| \Psi \star_{T_{Z_j}} \Psi - \mathcal{F}(g_j \circ \phi_{t_j}) \right\|_{L^2}^2 + \sigma \left\| \Psi - \Psi_M \right\|_{L^2}^2$$

for $\sigma \geq 0$ and $\Psi_M \in L^2(A, \mathbb{C})$.

The functional is well-defined, since the simulated Fourier space images $\Psi \star_{T_{Z_j}} \Psi$ are square-integrable by Corollary 2.1.4 and the shifted experimental images $\mathcal{F}(g_j \circ \phi_{t_j})$ in Fourier space are clearly also square-integrable for all $j \in \{1, \ldots, N\}$. The domain of the exit wave is restricted to the support of the aperture function, since all frequencies outside of A are filtered out due to

$$\Psi \star_{T_{Z_j}} \Psi = (\Psi \mathbb{1}_A) \star_{T_{Z_j}} (\Psi \mathbb{1}_A)$$

and thus do not contribute to the simulated images.

The restriction $\operatorname{supp}(\mathcal{F}(g_j)) \subseteq \overline{2A}$ on the frequencies of the experimental images does not affect the reconstruction, since $\operatorname{supp}(\Psi \star_{T_Z} \Psi) \subseteq \overline{2A}$ for all $\Psi \in L^2(A, \mathbb{C}), Z \in \mathbb{R}$ and $\operatorname{supp}(\mathcal{F}(g_j \circ \phi_y)) = \operatorname{supp}(\mathcal{F}(g_j))$ for all $y \in \mathbb{R}^2$, which is a consequence of the fact that translations are converted to modulations with the Fourier transform. Thus only image plane frequencies v with $|v| < 2r_a$, where r_a is the aperture radius, can be reconstructed with the objective functional. Hence a low-pass filter of radius $2r_a$ may be applied to the experimental input images $g_1, \ldots, g_N$ without changing E_σ in any way.

There are several equivalent ways of expressing the data term of the objective functional. Instead of comparing the simulated images with the experimental images in Fourier space, the data term can also be transformed to real space, since the Fourier transform is unitary. This results in

$$E_\sigma[\Psi, t] = \frac{1}{N} \sum_{j=1}^{N} \left\| \mathcal{F}^{-1}(\Psi \star_{T_{Z_j}} \Psi) - g_j \circ \phi_{t_j} \right\|_{L^2}^2 + \sigma \left\| \Psi - \Psi_M \right\|_{L^2}^2 . \tag{3.10}$$

As the Fourier transform converts translations to modulations and the L^2-norm is invariant of translations or modulations, further ways to express the functional E_σ are

$$E_\sigma[\Psi, t] = \frac{1}{N}\sum_{j=1}^{N} \left\|\Psi \star_{T_{Z_j}} \Psi - \mu_{t_j}\mathcal{F}(g_j)\right\|_{L^2}^2 + \sigma\left\|\Psi - \Psi_M\right\|_{L^2}^2 \tag{3.11}$$

$$= \frac{1}{N}\sum_{j=1}^{N} \left\|\left(\Psi \star_{T_{Z_j}} \Psi\right)\mu_{-t_j} - \mathcal{F}(g_j)\right\|_{L^2}^2 + \sigma\left\|\Psi - \Psi_M\right\|_{L^2}^2 . \tag{3.12}$$

The version given in Equation (3.11) is used for the numerical minimization of the objective functional in Chapter 5. By Section 3.1.1 and Equation (2.35), the functional in Equation (3.12) is essentially identical to the MIMAP functionals.

Remark 3.3.2. Although the objective functional in Definition 3.3.1 is defined such that its domain contains translations for all N input images, only $N-1$ translations can be determined by minimizing the functional if $\sigma = 0$ or $\Psi_M = 0$. The reason for this is that no reference frame for the translations is fixed, which can be seen mathematically as follows. If $t' = (t_1+\tau, \dots, t_N+\tau) \in (\mathbb{R}^2)^N$ with $\tau \in \mathbb{R}^2$, then

$$\begin{aligned}
E_\sigma[\Psi, t'] &= \frac{1}{N}\sum_{j=1}^{N} \left\|\Psi \star_{T_{Z_j}} \Psi - \mu_{t_j+\tau}\mathcal{F}(g_j)\right\|_{L^2}^2 + \sigma\left\|\Psi\right\|_{L^2}^2 \\
&= \frac{1}{N}\sum_{j=1}^{N} \left\|(\Psi \star_{T_{Z_j}} \Psi)\mu_{-\tau} - \mu_{t_j}\mathcal{F}(g_j)\right\|_{L^2}^2 + \sigma\left\|\Psi\right\|_{L^2}^2 \\
&= \frac{1}{N}\sum_{j=1}^{N} \left\|(\Psi\mu_{-\tau}) \star_{T_{Z_j}} (\Psi\mu_{-\tau}) - \mu_{t_j}\mathcal{F}(g_j)\right\|_{L^2}^2 + \sigma\left\|\Psi\right\|_{L^2}^2 \\
&= E_\sigma[\Psi\mu_{-\tau}, t]
\end{aligned}$$

for all $\Psi \in L^2(A, \mathbb{C})$ by Equation (2.35). This implies that $E_\sigma[\Psi\mu_\tau, t'] = E_\sigma[\Psi, t]$, which shows that the functional's value is unaffected by a "global" modulation (respectively translation) of the exit wave and all images.

However, the reference frame for the translation can be prescribed arbitrarily. For the practical minimization of the objective functional this problem is therefore easily circumvented by keeping e.g. $t_1 = 0 \in \mathbb{R}^2$ fixed, which means that the domain of E_σ is effectively reduced to $L^2(A, \mathbb{C}) \times (\mathbb{R}^2)^{N-1}$.

Alternatively, if $\sigma \neq 0$ and $\Psi_M \neq 0$, then the a-priori estimate Ψ_M can act as some kind of "soft" reference frame. This is meant in the sense that the exit wave Ψ may still deviate slightly from the reference frame prescribed by Ψ_M without changing the value of the functional very much. Note that this only works if Ψ_M contains some structural information about the exit wave. If Ψ_M merely corresponds to the mean value $m \in \mathbb{C}$ of the real space exit wave, then

$$\left\|\mathcal{F}^{-1}(\Psi) - m\right\|_{L^2}^2 = \left\|\mathcal{F}^{-1}(\Psi)\circ\phi_{-\tau} - m\right\|_{L^2}^2 = \left\|\mathcal{F}^{-1}(\Psi\mu_{-\tau}) - m\right\|_{L^2}^2$$

for all $\tau \in \mathbb{R}^2$ and the functional's value is again unaffected by a global modulation of the exit wave and all images.

For the remainder of this section, the coefficient $\sigma \geq 0$ of the regularizer is fixed.

3.3.1 Gâteaux differentials

Let $t \in (\mathbb{R}^2)^N$. The differentials with respect to the exit wave can be deduced from the fact that the function $\Psi \mapsto E_\sigma[\Psi, t]$ restricted to any one-dimensional affine subspace of its domain $L^2(A, \mathbb{C})$ is actually a polynomial of degree 4. This can easily be verified using the properties of the weighted cross-correlation from Section 2.1:

Lemma 3.3.3 (Extension of [14, Lemma 3.2]). *Let $t \in (\mathbb{R}^2)^N$. For all $\Psi, \Phi \in L^2(A, \mathbb{C})$ with $\Phi \neq 0$ there are coefficients $\left(C^j_{\Psi,\Phi,t,\sigma}\right)_{j=0,\dots,4} \in \mathbb{R}^5$ such that*

$$E_\sigma[\Psi + \tau\Phi, t] = \sum_{j=0}^{4} C^j_{\Psi,\Phi,t,\sigma} \tau^j$$

for all $\tau \in \mathbb{R}$. Furthermore, $C^4_{\Psi,\Phi,t,\sigma} > 0$.

Proof. Denote the modulated input images by $G_j := \mu_{t_j} \mathcal{F}(g_j)$, the difference of the simulated and experimental images in Fourier space by $D_{\Psi,j} := \Psi \star_{T_{Z_j}} \Psi - G_j$ and the regularizer by $R_\Psi := \|\Psi - \Psi_M\|^2_{L^2}$. Then

$$\begin{aligned} E_\sigma[\Psi + \tau\Phi, t] &= \frac{1}{N} \sum_{j=1}^{N} \left\| (\Psi + \tau\Phi) \star_{T_{Z_j}} (\Psi + \tau\Phi) - G_j \right\|^2_{L^2} + \sigma R_{\Psi+\tau\Phi} \\ &= \frac{1}{N} \sum_{j=1}^{N} \left\| D_{\Psi,j} + \tau(\Psi \star_{T_{Z_j}} \Phi + \Phi \star_{T_{Z_j}} \Psi) + \tau^2(\Phi \star_{T_{Z_j}} \Phi) \right\|^2_{L^2} + \sigma R_{\Psi+\tau\Phi} \end{aligned}$$

for all $\tau \in \mathbb{R}$. Expanding and collecting coefficients with the same power of τ yields $E_\sigma[\Psi + \tau\Phi, t] = \sum_{j=0}^{4} C^j_{\Psi,\Phi,t,\sigma} \tau^j$ with

$$\begin{aligned} C^0_{\Psi,\Phi,t,\sigma} &= \frac{1}{N} \sum_{j=1}^{N} \|D_{\Psi,j}\|^2_{L^2} + \sigma R_\Psi = E_\sigma[\Psi, t], \\ C^1_{\Psi,\Phi,t,\sigma} &= \frac{2}{N} \sum_{j=1}^{N} \mathrm{Re}\Big(\big(D_{\Psi,j}, \Psi \star_{T_{Z_j}} \Phi + \Phi \star_{T_{Z_j}} \Psi\big)_{L^2} \Big) + 2\sigma \mathrm{Re}\Big((\Psi - \Psi_M, \Phi)_{L^2} \Big), \\ C^2_{\Psi,\Phi,t,\sigma} &= \frac{1}{N} \sum_{j=1}^{N} \left\| \Psi \star_{T_{Z_j}} \Phi + \Phi \star_{T_{Z_j}} \Psi \right\|^2_{L^2} + 2\mathrm{Re}\Big(\big(D_{\Psi,j}, \Phi \star_{T_{Z_j}} \Phi\big)_{L^2} \Big) + \sigma \|\Phi\|^2_{L^2}, \\ C^3_{\Psi,\Phi,t,\sigma} &= \frac{2}{N} \sum_{j=1}^{N} \mathrm{Re}\Big(\big(\Psi \star_{T_{Z_j}} \Phi + \Phi \star_{T_{Z_j}} \Psi, \Phi \star_{T_{Z_j}} \Phi\big)_{L^2} \Big), \\ C^4_{\Psi,\Phi,t,\sigma} &= \frac{1}{N} \sum_{j=1}^{N} \left\| \Phi \star_{T_{Z_j}} \Phi \right\|^2_{L^2}. \end{aligned}$$

Since the real space images g_j are real-valued, it follows that $G_j \in L^2(\mathbb{R}^2, \mathbb{C})$ are Hermitian functions for all $j \in \{1, \dots, N\}$. Using the symmetry properties from Lemma 2.1.5 and the fact that $D_{\Psi,j}$ is Hermitian, the coefficients of the odd powers of τ can therefore be simplified to

$$\begin{aligned}
C^1_{\Psi,\Phi,t,\sigma} &= \frac{4}{N}\sum_{j=1}^{N} \mathrm{Re}\Big(\big(D_{\Psi,j}, \Psi \star_{T_{Z_j}} \Phi\big)_{L^2}\Big) + 2\sigma \mathrm{Re}\Big((\Psi - \Psi_M, \Phi)_{L^2}\Big),\\
C^3_{\Psi,\Phi,t,\sigma} &= \frac{4}{N}\sum_{j=1}^{N} \mathrm{Re}\Big(\big(\Psi \star_{T_{Z_j}} \Phi, \Phi \star_{T_{Z_j}} \Phi\big)_{L^2}\Big).
\end{aligned}$$

The coefficient $C^4_{\Psi,\Phi,t,\sigma}$ of τ^4 is positive by Lemma 2.1.9 respectively Lemma 2.1.10. □

The Gâteaux differentials of E_σ with respect to the exit wave Ψ can now simply be read off the polynomials' coefficients.

Corollary 3.3.4 (Extension of [14, Corollary 3.3]). *Let $t \in (\mathbb{R}^2)^N$ and $\Psi, \Phi \in L^2(A, \mathbb{C})$ with $\Phi \neq 0$. The Gâteaux differentials of E_σ at (Ψ, t) in the direction $(\Phi, 0)$ are*

$$\big\langle \partial_\Psi E^{(n)}_\sigma[\Psi, t], \Phi\big\rangle := \frac{d}{d\varepsilon}\big\langle \partial_\Psi E^{(n-1)}_\sigma[\Psi + \varepsilon\Phi, t], \Phi\big\rangle\Big|_{\varepsilon=0} = \begin{cases} n! C^n_{\Psi,\Phi,t,\sigma}, & n \in \{1, \dots, 4\},\\ 0, & n \geq 5.\end{cases}$$

As one would expect, the function $t \mapsto E_\sigma[\Psi, t]$ for a fixed $\Psi \in L^2(\mathbb{R}^2, \mathbb{C})$ does not possess a similar structure that can be utilized for the computation of the differentials. However, for the derivative based minimization methods that are used to minimize E_σ in Chapter 4, only the first order Gâteaux differential is needed, which can be calculated more directly without using difference quotients:

Lemma 3.3.5 ([14, Lemma 3.4]). *Let $\Psi \in L^2(A, \mathbb{C})$ and $t, \tilde{t} \in (\mathbb{R}^2)^N$ with $\tilde{t} \neq 0$. The first order Gâteaux differential of E_σ at (Ψ, t) in the direction $(0, \tilde{t})$ is*

$$\big\langle \partial_t E_\sigma[\Psi, t], \tilde{t}\big\rangle = -\frac{2}{N}\sum_{j=1}^{N} \mathrm{Re}\left(\big(\Psi \star_{T_{Z_j}} \Psi - \mu_{t_j}\mathcal{F}(g_j), \nu_{\tilde{t}_j}\mu_{t_j}\mathcal{F}(g_j)\big)_{L^2}\right),$$

where $\nu_{\tilde{t}_j}(x) = 2\pi \mathrm{i} x \cdot \tilde{t}_j$ for all $x \in \mathbb{R}^2$.

Proof. By the definition of E_σ, we have

$$\big\langle \partial_t E_\sigma[\Psi, t], \tilde{t}\big\rangle = \frac{1}{N}\sum_{j=1}^{N} \frac{d}{d\varepsilon}\int_{\mathbb{R}^2} \Big|\underbrace{(\Psi \star_{T_{Z_j}} \Psi)(x) - \mu_{t_j + \varepsilon\tilde{t}_j}(x)\mathcal{F}(g_j)(x)}_{=:R_{j,\varepsilon}(x)}\Big|^2 \,\mathrm{d}x\Big|_{\varepsilon=0}.$$

The partial derivative of the integrand is given by

$$\frac{d}{d\varepsilon}\left|R_{j,\varepsilon}(x)\right|^2 = \frac{d}{d\varepsilon}\left(R_{j,\varepsilon}(x)R^*_{j,\varepsilon}(x)\right) = 2\mathrm{Re}\left(\left(\frac{d}{d\varepsilon}R_{j,\varepsilon}(x)\right)R^*_{j,\varepsilon}(x)\right),$$

where

$$\frac{d}{d\varepsilon} R_{j,\varepsilon}(x) = -\nu_{\tilde{t}_j}(x)\mu_{t_j+\varepsilon\tilde{t}_j}(x)\mathcal{F}(g_j)(x).$$

This derivative is dominated by

$$\left|\frac{d}{d\varepsilon}\left|R_{j,\varepsilon}(x)\right|^2\right| \leq 2\left|\nu_{\tilde{t}_j}(x)\right| \left|\mathcal{F}(g_j)(x)\right| \left(\left|\left(\Psi \star_{T_{Z_j}} \Psi\right)(x)\right| + \left|\mathcal{F}(g_j)(x)\right|\right)$$

for all $\varepsilon > 0$, which is an integrable function by the Cauchy-Schwarz inequality, Corollary 2.1.4 and Definition 3.3.1. By the dominated convergence theorem, the integrable upper bound that is independent of ε allows us to interchange integration and differentiation, which yields

$$\begin{aligned}\langle \partial_t E[\Psi, t], \tilde{t}\rangle &= \frac{1}{N}\sum_{j=1}^{N}\int_{\mathbb{R}^2}\frac{d}{d\varepsilon}\left|R_{j,\varepsilon}(x)\right|^2\Big|_{\varepsilon=0}\,\mathrm{d}x \\ &= -\frac{2}{N}\sum_{j=1}^{N}\mathrm{Re}\left(\left(\Psi \star_{T_{Z_j}} \Psi - \mu_{t_j}\mathcal{F}(g_j), \nu_{\tilde{t}_j}\mu_{t_j}\mathcal{F}(g_j)\right)_{L^2}\right). \qquad \square\end{aligned}$$

3.3.2 Existence of minimizers

By the direct method of the calculus of variations, minimizers of the objective functional E_σ exist if E_σ is coercive, weakly lower semi-continuous and its domain is a reflexive Banach space. Additionally, instead of considering the full domain $L^2(A,\mathbb{C}) \times \mathbb{R}^N$ of E_σ, it is also sufficient to restrict the admissible set to a weakly closed subset as is done below. Most of the work in showing the coercivity and the weak lower semi-continuity has already been carried out in earlier sections. The purpose of this section is to combine the individual results in order to yield a complete proof of the existence of minimizers.

In order to ensure that E_σ is coercive on its entire domain, the admissible set is restricted to $L^2(A,\mathbb{C}) \times \overline{B_r(0)}^N$ for an $r > 0$. From a mathematical point of view, the additional parameter r must be considered as a workaround to ensure the coercivity of the functional in the translations. Unlike usual rigid registration problems, it is not easily possible to define an $r > 0$ that can serve as an upper bound based on the diameter of the image domain. Since the support of the TCC is bounded, the domain of the simulated images in real space is necessarily unbounded. Consequently, it is not clear whether there exists a minimizing sequence of E_σ with bounded translations for all choices of input images $g_1, \dots, g_N$. However, in practice the restriction on the translations is justified by the fact that the sample is moving within a bounded domain during image acquisition. Therefore, this is a purely mathematical limitation that is irrelevant to the practical application.

As shown in Section 3.2, the objective functional is not coercive with respect to the exit wave if $\sigma = 0$ and the TCC is $\star$-separable (with the regularity assumptions on $h_{Z_j,k}$ in Proposition 3.2.5), which likely also applies to TCCs that are not $\star$-separable. In the case $\sigma > 0$ the objective functional is clearly coercive with respect to the exit wave.

The weak lower semi-continuity of the data term of E_σ follows from Proposition 2.1.12, Equation (3.11) and the fact that finite sums of weakly lower semi-continuous functions are also weakly lower semi-continuous. Since norms are weakly lower semi-continuous on their respective spaces, the regularizer of E_σ is weakly lower semi-continuous as well. Overall, E_σ is a weakly lower semi-continuous functional for all $\sigma \geq 0$.

Since $L^2(A,\mathbb{C}) \times (\mathbb{R}^2)^N$ is a reflexive Banach space and $L^2(A,\mathbb{C}) \times \overline{B_r(0)}^N$, as a closed and convex subset, is also weakly sequentially closed, the existence of minimizers of E_σ now follows with the direct method:

Theorem 3.3.6 ([14, Theorem 4.2])**.** *If $\sigma > 0$, then there exist $\Psi_* \in L^2(A,\mathbb{C})$ and $t_* \in \overline{B_r(0)}^N$ with*

$$E_\sigma[\Psi_*, t_*] = \inf_{\Psi \in L^2(A,\mathbb{C}),\, t \in \overline{B_r(0)}^N} E_\sigma[\Psi, t].$$

3.3.3 Optimization of the focus parameters

The objective functional can be extended to

$$E_\sigma^{\text{ext}} : L^2(A,\mathbb{C}) \times (\mathbb{R}^2)^N \times \mathbb{R}^N \to \mathbb{R}, \quad (\Psi, t, Z) \mapsto \frac{1}{N} \sum_{j=1}^N \left\| \Psi \star_{T_{Z_j}} \Psi - \mathcal{F}(g \circ \phi_{t_j}) \right\|_{L^2}^2 + \sigma \left\| \Psi - \Psi_M \right\|_{L^2}^2$$

in order to optimize for the focus parameters as well. The first order Gâteaux differential of E_σ^{ext} with respect to the focus is computed by interchanging differentiation and integration twice, which allows us to express the differential in terms of the derivative of the TCC. To this end, we exclude the full TCC and the large aperture approximation in the following and therefore restrict the TCC T_Z to one of the TCCs from Sections 2.2.3 to 2.2.6.

Lemma 3.3.7. *Let $\Psi \in L^2(A,\mathbb{C})$, $t \in (\mathbb{R}^2)^N$ and $Z, \tilde{Z} \in \mathbb{R}^N$ with $\tilde{Z} \neq 0$. The first order Gâteaux differential of E_σ^{ext} at (Ψ, t, Z) in the direction $(0, 0, \tilde{Z})$ is*

$$\left\langle \partial_Z E_\sigma^{ext}[\Psi, t, Z], \tilde{Z} \right\rangle = \frac{2}{N} \sum_{j=1}^N \operatorname{Re}\Big(\big(\Psi \star_{T'_{Z_j,\tilde{Z}_j}} \Psi, \Psi \star_{T_{Z_j}} \Psi - G_j \big)_{L^2} \Big),$$

where $T'_{Z_j,\tilde{Z}_j}(v,w) = \frac{d}{d\varepsilon} T_{Z_j+\varepsilon\tilde{Z}_j}(v,w)\Big|_{\varepsilon=0}$.

Proof. Let $G_j := \mu_{t_j} \mathcal{F}(g_j)$. By definition, the Gâteaux differential is

$$\left\langle \partial_Z E_\sigma^{\text{ext}}[\Psi, t, Z], \tilde{Z} \right\rangle = \frac{d}{d\varepsilon} \frac{1}{N} \sum_{j=1}^N \int_{\mathbb{R}^2} \left| \big(\Psi \star_{T_{Z_j+\varepsilon\tilde{Z}_j}} \Psi \big)(x) - G_j(x) \right|^2 \mathrm{d}x \Bigg|_{\varepsilon=0}. \tag{3.13}$$

The proof proceeds backwards and begins with interchanging integration and differentiation for the inner integral

$$\frac{d}{d\varepsilon} \big(\Psi \star_{T_{Z_j+\varepsilon\tilde{Z}_j}} \Psi \big)(x) \Big|_{\varepsilon=0} = \frac{d}{d\varepsilon} \int_{\mathbb{R}^2} \Psi(y) \Psi^*(x+y) T_{Z_j+\varepsilon\tilde{Z}_j}(x+y, y) \, \mathrm{d}y \Bigg|_{\varepsilon=0}. \tag{3.14}$$

Since T_Z is continuously differentiable with respect to Z, it is straightforward to verify with a case-by-case analysis for the different TCCs that for all bounded intervals $I \subseteq \mathbb{R}$ there exists a constant $C > 0$ such that

$$\left| \frac{d}{d\varepsilon} T_{Z+\varepsilon}(x,y)\Big|_{\varepsilon=0} \right| \leq C \qquad \forall\, Z \in I \;\forall x, y \in A.$$

This follows from the fact that the derivative $\frac{d}{d\varepsilon} T_{Z+\varepsilon}(x,y)$ is continuous in Z, x and y on $I \times A^2$. (The derivatives of T_Z^{Ishizuka} and T_Z^{MAL} with respect to Z are calculated in Section 4.4.1.) In particular, if we choose $I = [Z_j - 1, Z_j + 1]$, then there is a constant $C > 0$ such that the difference quotients satisfy

$$\left| \frac{T_{Z_j+\varepsilon}(x,y) - T_{Z_j}(x,y)}{\varepsilon} \right| \leq C \qquad \forall\, \varepsilon \in [-1,1] \;\forall\, x, y \in A \tag{3.15}$$

by the mean value theorem. Due to Equation (3.15) and the fact that $\Psi \in L^2(A, \mathbb{C})$, the function $y \mapsto \Psi(y)\Psi^*(x+y)C$ is an integrable function dominating the sequence of difference quotients that converges to the focus derivative of the integrand in Equation (3.14). By the dominated convergence theorem, differentiation and integration can be interchanged, yielding

$$\frac{d}{d\varepsilon} \left(\Psi \star_{T_{Z_j+\varepsilon\tilde{Z}_j}} \Psi\right)(x)\Big|_{\varepsilon=0} = \int_{\mathbb{R}^2} \Psi(y)\Psi^*(x+y) \frac{d}{d\varepsilon}\, T_{Z_j+\varepsilon\tilde{Z}_j}(x+y, y)\Big|_{\varepsilon=0} \,\mathrm{d}y. \tag{3.16}$$

Continuing with the outer integral, we have to find an integrable bound to the sequence of difference quotients that converge pointwise almost everywhere to the derivative

$$\begin{aligned} &\frac{d}{d\varepsilon} \left| \left(\Psi \star_{T_{Z_j+\varepsilon\tilde{Z}_j}} \Psi\right)(x) - G_j(x) \right|^2 \Bigg|_{\varepsilon=0} \\ &= 2\mathrm{Re}\left(\frac{d}{d\varepsilon} \left(\Psi \star_{T_{Z_j+\varepsilon\tilde{Z}_j}} \Psi\right)(x)\Big|_{\varepsilon=0} \left((\Psi \star_{T_{Z_j}} \Psi)(x) - G_j(x)\right)^* \right) \end{aligned} \tag{3.17}$$

By Equation (3.15) and the fact that $\mathrm{supp}(T_Z) \subseteq \overline{A} \times \overline{A}$, there is a constant $C > 0$ such that

$$\begin{aligned} &\left| \frac{\left(\Psi \star_{T_{Z_j+\varepsilon\tilde{Z}_j}} \Psi\right)(x) - \left(\Psi \star_{T_{Z_j}} \Psi\right)(x)}{\varepsilon} \right| \\ &= \left| \int_A \Psi(y)\Psi^*(x+y) \frac{T_{Z_j+\varepsilon\tilde{Z}_j}(x+y,y) - T_{Z_j}(x+y,y)}{\varepsilon} \,\mathrm{d}y \right| \\ &\leq \int_A \left|\Psi(y)\Psi^*(x+y)\right| C \,\mathrm{d}y \leq \|\Psi\|_{L^2}^2\, C \end{aligned}$$

for all $\varepsilon \in [-1,1]$ and all $x \in A$. Consequently, $x \mapsto 2\,\|\Psi\|_{L^2}^2\, C \left|(\Psi \star_{T_{Z_j}} \Psi)(x) - G_j(x)\right|$ is an integrable function dominating the sequence of difference quotients that converge to Equation (3.17). Applying the dominated convergence theorem once more, it follows that

$$\begin{aligned}
\left\langle \partial_Z E_\sigma^{\text{ext}}[\Psi,t,Z],\tilde{Z}\right\rangle &= \frac{1}{N}\sum_{j=1}^N \int_{\mathbb{R}^2} \frac{d}{d\varepsilon} \left|\left(\Psi \star_{T_{Z_j+\varepsilon\tilde{Z}_j}} \Psi\right)(x) - G_j(x)\right|^2\Bigg|_{\varepsilon=0} \,\mathrm{d}x \\
&= \frac{1}{N}\sum_{j=1}^N \int_{\mathbb{R}^2} 2\mathrm{Re}\left(\frac{d}{d\varepsilon}\left(\Psi \star_{T_{Z_j+\varepsilon\tilde{Z}_j}} \Psi\right)(x)\Big|_{\varepsilon=0} \cdot \left(\left(\Psi \star_{T_{Z_j}} \Psi\right)(x) - G_j(x)\right)^*\right) \mathrm{d}x \\
&\overset{(3.16)}{=} \frac{2}{N}\sum_{j=1}^N \mathrm{Re}\Big(\left(\Psi \star_{T'_{Z_j,\tilde{Z}_j}} \Psi, \Psi \star_{T_{Z_j}} \Psi - G_j\right)_{L^2}\Big). \qquad \square
\end{aligned}$$

Remark 3.3.8. Similar to the translations, only the relative focus between two successive images can be determined in this way, which can easily be seen with the coherent TCC. Let $Z' = (Z_1+\zeta,\dots,Z_N+\zeta) \in (\mathbb{R}^2)^N$ for $\zeta \in \mathbb{R}$ and $Z \in (\mathbb{R}^2)^N$. Then, using the definition of the wave aberration function in Equation (2.10), we get

$$T^{\text{coh}}_{Z'_j}(v,w) = T^{\text{coh}}_{Z_j+\zeta}(v,w) = t_{Z_j}(v)t^*_{Z_j}(w)q_\zeta(v)q^*_\zeta(w) \qquad \forall\, v,w \in \mathbb{R}^2$$

with $q_\zeta(v) = \exp\left(-\pi\mathrm{i}\zeta\lambda\left\|v\right\|_2^2\right)$. If $\Psi_M = 0$, it follows that

$$\begin{aligned}
E_\sigma^{\text{ext}}[\Psi,t,Z'] &= \frac{1}{N}\sum_{j=1}^N \left\|\Psi \star_{T^{\text{coh}}_{Z_j+\zeta}} \Psi - \mathcal{F}(g\circ\phi_{t_j})\right\|_{L^2}^2 + \sigma\left\|\Psi\right\|_{L^2}^2 \\
&= \frac{1}{N}\sum_{j=1}^N \left\|(\Psi q_\zeta) \star_{T^{\text{coh}}_{Z_j}} (\Psi q_\zeta) - \mathcal{F}(g\circ\phi_{t_j})\right\|_{L^2}^2 + \sigma\left\|\Psi\right\|_{L^2}^2 \qquad (3.18)\\
&= E_\sigma^{\text{ext}}[\Psi q_\zeta,t,Z]
\end{aligned}$$

for all $\Psi \in L^2(A,\mathbb{C})$ and all $t \in (\mathbb{R}^2)^N$. This implies $E_\sigma^{\text{ext}}\left[\Psi\frac{1}{q_\zeta},t,Z'\right] = E_\sigma^{\text{ext}}[\Psi,t,Z]$ and shows that the value of the functional E_σ^{ext} is invariant under simultaneously changing the “global” base focus value by ζ and substituting $\Psi \to \Psi\frac{1}{q_\zeta}$ for the exit wave. It is not obvious if a similar relationship also holds for other TCCs. Kirkland [38, p. 168] mentions that the impossibility of determining the base focus value was verified in a numerical experiment, but without providing any details on his approach.

There is however one key difference to the translations: while the absolute value of the translations can be defined with respect to an arbitrary reference frame, the absolute value of the focus values can not be arbitrarily prescribed and is very important for a successful exit wave reconstruction. If the correct base focus value is not known, then the reconstructed electron wave is *not* the exit wave since the result is multiplied by q_ζ as shown in Equation (3.18) and the exit wave is free from aberrations by definition. The base focus value must therefore be determined to a sufficiently high accuracy in a different way (for example by comparison with a multislice simulation if at least one image contains a crystal of known structure; see e.g. [43]).

3.4 Convexity

The convexity of the objective functional is an important property with regard to the uniqueness of minimizers and the numerical minimization of the functional. It is obvious that the data term of the objective functional E_σ is not convex with respect to the translations t for arbitrary input image series $(g_j)_{j=1,\dots,N}$. However, it is not clear if the data term is convex with respect to the exit wave, i.e. if the functional $\Psi \mapsto E_0[\Psi, t]$ is convex for a fixed $t \in (\mathbb{R}^2)^N$.

In Section 3.4.1, it is shown that the data term of the objective functional is not convex except in a trivial case. Furthermore, in Section 3.4.2 the local convexity in a neighborhood of a minimizer is investigated, where it is shown that a functional related to the objective functional is not convex in any neighborhood of a global minimizer. Although this result is obtained only for a strongly simplified version of the objective functional by disregarding all aberrations and coherence effects as well as the regularizer, it can be understood as an indication of how badly ill-posed the non-regularized inverse problem is.

3.4.1 Global convexity

First, consider the trivial case where all input images are zero, i.e. $g_j = 0$ for all $j \in \{1, \dots, N\}$. In this case, the functional $\Psi \mapsto E_0[\Psi, t]$ is convex by Lemma 2.1.11 and $\Psi = 0$ is the unique global minimizer by Lemma 2.1.9 respectively Lemma 2.1.10. For this reason, we restrict ourselves to the case $g_j \neq 0$ for at least one $j \in \{1, \dots, N\}$ for the remainder of this section.

In order to show that the data term of the objective functional is not convex with respect to the exit wave, it is sufficient to show its non-convexity on a one-dimensional linear subspace of $L^2(A, \mathbb{C})$. The objective functional on one-dimensional linear subspaces of $L^2(A, \mathbb{C})$ can be characterized as follows.

Lemma 3.4.1. *Let* $t \in (\mathbb{R}^2)^N$ *and* $\Phi \in L^2(A, \mathbb{C})$ *with* $\Phi \neq 0$. *Then*

$$E_0[\alpha\Phi, t] = C^4_{0,\Phi,t,0}\alpha^4 + C^2_{0,\Phi,t,0}\alpha^2 + C^0_{0,\Phi,t,0}$$

for all $\alpha \in \mathbb{R}$, *where* $C^4_{0,\Phi,t,0} > 0$ *and*

$$C^2_{0,\Phi,t,0} = -\frac{2}{N}\sum_{j=1}^{N} \operatorname{Re}\left(\left(g_j \circ \phi_{t_j}, \mathcal{F}^{-1}\left(\Phi \star_{T_{Z_j}} \Phi\right)\right)_{L^2}\right).$$

Proof. By Lemma 3.3.3, the data term of E_σ can be written as $E_0[\alpha\Phi, t] = \sum_{j=0}^{4} C^j_{0,\Phi,t,0}\alpha^j$ for all $\alpha \in \mathbb{R}$ with $C^4_{0,\Phi,t,0} > 0$. Since $\Psi = 0$ and $\sigma = 0$, the coefficients of the odd powers of α are zero and the coefficient of α^2 simplifies to

$$\begin{aligned} C^2_{0,\Phi,t,0} &= -\frac{2}{N}\sum_{j=1}^{N}\operatorname{Re}\left(\left(\mu_{t_j}\mathcal{F}(g_j), \Phi \star_{T_{Z_j}} \Phi\right)_{L^2}\right) \\ &= -\frac{2}{N}\sum_{j=1}^{N}\operatorname{Re}\left(\left(g_j \circ \phi_{t_j}, \mathcal{F}^{-1}\big(\Phi \star_{T_{Z_j}} \Phi\big)\right)_{L^2}\right). \qquad \square \end{aligned}$$

Let $t \in (\mathbb{R}^2)^N$. In order to show that the functional $\Psi \mapsto E_0[\Psi, t]$ is not convex, it is therefore sufficient to find a direction $\Phi \in L^2(A, \mathbb{C})$ with $C^2_{0,\Phi,t,0} < 0$. This follows from the fact that a biquadratic polynomial $f(x) = ax^4 + bx^2 + c$ with $a \geq 0$ is convex if and only if $b \geq 0$.

Since the input images g_j are nonnegative and $\mathcal{F}^{-1}\big(\Phi \star_{T_{Z_j}} \Phi\big)$ is nonnegative as well for all $\Phi \in L^2(A, \mathbb{C})$ by Lemma 2.1.7, it follows that $C^2_{0,\Phi,t,0} \leq 0$ for all $\Phi \in L^2(A, \mathbb{C})$. In order to establish the non-convexity of $\Psi \mapsto E_0[\Psi, t]$, the challenge therefore lies in finding a direction $\Phi \in L^2(A, \mathbb{C})$ such that the coefficient $C^2_{0,\Phi,t,0}$ is strictly negative.

Assume for now that we formally extend the weighted cross-correlation and the Fourier transform to the space of tempered distributions. Then, we can choose $\Phi := \delta_0$ and get that $\mathcal{F}^{-1}(\Phi \star_{T_{Z_j}} \Phi) = T_{Z_j}(0, 0) > 0$ is a constant and positive function by Property III, which corresponds to the simulated TEM image of the plane electron wave (i.e. merely the electron beam without any interaction with a specimen). Then, formally,

$$C^2_{0,\Phi,t,0} = -\frac{2}{N}\sum_{j=1}^{N} T_{Z_j}(0,0)\left(g_j \circ \phi_{t_j}, 1\right)_{L^2} = -\frac{2}{N}\sum_{j=1}^{N} T_{Z_j}(0,0)\left\|g_j\right\|_{L^1} < 0,$$

since $g_j \geq 0$ for all $j \in \{1, \ldots, N\}$ and $\|g_j\|_{L^1} > 0$ for at least one image. The same conclusion can be reached without resorting to tempered distributions by considering a mollifier sequence for Φ instead, which is done in Proposition 3.4.2.

Proposition 3.4.2. *Let $t \in (\mathbb{R}^2)^N$. If $g_j \neq 0$ for at least one $j \in \{1, \ldots, N\}$, then $\Psi \mapsto E_0[\Psi, t]$ is not convex.*

Proof. By Lemma 3.4.1, it is sufficient to find a direction $\Phi \in L^2(A, \mathbb{C})$ with $C^2_{0,\Phi,t,0} < 0$. Denote the d-dimensional ball with respect to the supremum norm by

$$B^{\infty,d}_r(c) := \left\{x \in \mathbb{R}^d : \|x - c\|_\infty < r\right\} \qquad \forall\, r \in \mathbb{R}, c \in \mathbb{R}^d, d \in \mathbb{N}$$

and choose any sequence of directions $(\Phi_n)_{n\in\mathbb{N}} \in L^2(A, \mathbb{C})^{\mathbb{N}}$ such that $\Phi_n \in L^1(A, \mathbb{R}_{\geq 0})$ with $\int_A \Phi_n(x)\,\mathrm{d}x = 1$ for all $n \in \mathbb{N}$ and

$$\lim_{n\to\infty}\int_{B^{\infty,2}_\delta(0)\cap A} \Phi_n(x)\,\mathrm{d}x = 1 \qquad \forall\, \delta > 0.$$

The integrand of $(g_j \circ \phi_{t_j}, \mathcal{F}^{-1}(\Phi_n \star_{T_{Z_j}} \Phi_n))_{L^2}$ is dominated by the integrable function $g_j \circ \phi_{t_j}\left\|T_{Z_j}\right\|_\infty$, since

$$\left|\mathcal{F}^{-1}\big(\Phi_n \star_{T_{Z_j}} \Phi_n\big)(x)\right| \leq \left\|\Phi_n \star_{T_{Z_j}} \Phi_n\right\|_{L^1} \leq \|\Phi_n\|^2_{L^1}\left\|T_{Z_j}\right\|_\infty = \left\|T_{Z_j}\right\|_\infty$$

for all $x \in \mathbb{R}^2$ by the definition of the inverse Fourier transform and Lemma 2.1.3. Assume that $\lim_{n\to\infty} \mathcal{F}^{-1}\big(\Phi_n \star_{T_{Z_j}} \Phi_n\big)(x) = T_{Z_j}(0,0) > 0$ holds for all $x \in \mathbb{R}^2$. By the dominated convergence theorem,

$$\lim_{n\to\infty} C^2_{0,\Phi_n,t,0} = -\frac{2}{N}\sum_{j=1}^{N} T_{Z_j}(0,0)\,\big\|g_j \circ \phi_{t_j}\big\|_{L^1} = -\frac{2}{N}\sum_{j=1}^{N} T_{Z_j}(0,0)\,\|g_j\|_{L^1} < 0,$$

since $g_j \geq 0$ for all $j \in \{1,\dots,N\}$ and $\|g_j\|_{L^1} > 0$ for at least one image. Consequently, there exists an $n \in \mathbb{N}$ such that $C^2_{0,\Phi_n,t,0} < 0$.

It remains to show that $\lim_{n\to\infty} \mathcal{F}^{-1}\big(\Phi_n \star_{T_{Z_j}} \Phi_n\big)(x) = T_{Z_j}(0,0)$ is indeed true for all $x \in \mathbb{R}^2$. For this purpose, we define the auxiliary functions

$$\Theta_n : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}_{\geq 0}, \quad (y,z) \mapsto \begin{cases} \Phi_n(z)\Phi_n(y+z), & \text{if } z \in A \text{ and } y+z \in A, \\ 0, & \text{otherwise} \end{cases}$$

for all $n \in \mathbb{N}$ and

$$f_{x,j} : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{C}, \quad (y,z) \mapsto T_{Z_j}(y+z,z)e^{2\pi \mathrm{i} y\cdot x}$$

for all $x \in \mathbb{R}^2$ and all $j \in \{1,\dots,N\}$. Since

$$\begin{aligned} &\int_{\mathbb{R}^2}\int_{\mathbb{R}^2} \big|\Phi_n^*(z)\Phi_n(y+z)T_{Z_j}(y+z,z)e^{2\pi\mathrm{i}y\cdot x}\big| \,\mathrm{d}y\,\mathrm{d}z \\ \leq\ &\big\|T_{Z_j}\big\|_\infty \int_{\mathbb{R}^2} |\Phi_n^*(z)| \int_{\mathbb{R}^2} |\Phi_n(y+z)| \,\mathrm{d}y\,\mathrm{d}z \\ =\ &\big\|T_{Z_j}\big\|_\infty < \infty, \end{aligned}$$

we may apply Fubini's theorem and get

$$\begin{aligned} \mathcal{F}^{-1}\big(\Phi_n \star_{T_{Z_j}} \Phi_n\big)(x) &= \int_{\mathbb{R}^2}\int_{\mathbb{R}^2} \Phi_n^*(z)\Phi_n(y+z)T_{Z_j}(y+z,z)\,\mathrm{d}z\, e^{2\pi\mathrm{i}y\cdot x}\,\mathrm{d}y \\ &= \int_{\mathbb{R}^2\times\mathbb{R}^2} \Theta_n(y,z)f_{x,j}(y,z)\,\mathrm{d}\begin{pmatrix} y \\ z \end{pmatrix} \end{aligned}$$

for all $x \in \mathbb{R}^2$, $n \in \mathbb{N}$ and $j \in \{1,\dots,N\}$. The function $f_{x,j}$ is bounded and continuous in 0 for all $x \in \mathbb{R}^2$, $j \in \{1,\dots,N\}$ since T_{Z_j} is continuous in 0 and the sequence $(\Theta_n)_{n\in\mathbb{N}}$ satisfies $\Theta_n \in L^1(\mathbb{R}^4,\mathbb{R}_{\geq 0})$ with $\int_{\mathbb{R}^4} \Theta_n(x)\,\mathrm{d}x = 1$ for all $n \in \mathbb{N}$. Furthermore,

$$\lim_{n\to\infty} \int_{B_\delta^{\infty,4}(0)} \Theta_n(x)\,\mathrm{d}x \leq \lim_{n\to\infty} \int_{\mathbb{R}^4} \Theta_n(x) = 1$$

and

$$\begin{aligned} \lim_{n\to\infty} \int_{B_\delta^{\infty,4}(0)} \Theta_n(x)\,\mathrm{d}x &= \lim_{n\to\infty} \int_{B_\delta^{\infty,2}(0)} \Phi_n(z) \int_{B_\delta^{\infty,2}(0)} \Phi_n(y+z)\,\mathrm{d}y\,\mathrm{d}z \\ &\geq \lim_{n\to\infty} \int_{B_{\delta/2}^{\infty,2}(0)} \Phi_n(z) \int_{B_\delta^{\infty,2}(0)+z} \Phi_n(y)\,\mathrm{d}y\,\mathrm{d}z \\ &\geq \lim_{n\to\infty} \int_{B_{\delta/2}^{\infty,2}(0)} \Phi_n(z) \int_{B_{\delta/2}^{\infty,2}(0)} \Phi_n(y)\,\mathrm{d}y\,\mathrm{d}z = 1 \end{aligned}$$

for all $\delta > 0$ imply that $\lim_{n\to\infty} \int_{B_\delta^{\infty,4}(0)} \Theta_n(x)\,\mathrm{d}x = 1$ for all $\delta > 0$. By Lemma 2.2.10 and the fact that all norms in $\mathbb{R}^4$ are equivalent, it follows that

$$\begin{aligned}\lim_{n\to\infty} \mathcal{F}^{-1}\big(\Phi_n \star_{T_{Z_j}} \Phi_n\big)(x) &= \lim_{n\to\infty} \int_{\mathbb{R}^2\times\mathbb{R}^2} \Theta_n(y,z) f_{x,j}(y,z)\,\mathrm{d}\begin{pmatrix} y \\ z\end{pmatrix} \\ &= f_{x,j}(0,0) = T_{Z_j}(0,0)\end{aligned}$$

for all $x \in \mathbb{R}^2$. $\square$

This concludes the proof of the non-convexity of the functional $\Psi \mapsto E_0[\Psi,t]$. If the coefficient σ of the regularizer is positive, then $\Psi \mapsto E_\sigma[\Psi,t]$ may or may not be convex depending on the size of σ. Similar to Lemma 3.4.1,

$$E_\sigma[\alpha\Phi,t] = C^4_{0,\Phi,t,\sigma}\alpha^4 + C^2_{0,\Phi,t,\sigma}\alpha^2 + C^0_{0,\Phi,t,\sigma}$$

for all $\alpha \in \mathbb{R}$ and all $\sigma \geq 0$, where $C^4_{0,\Phi,t,\sigma} > 0$ and

$$C^2_{0,\Phi,t,\sigma} = -\frac{2}{N}\sum_{j=1}^N \operatorname{Re}\left(\left(g_j \circ \phi_{t_j}, \mathcal{F}^{-1}\big(\Phi \star_{T_{Z_j}} \Phi\big)\right)_{L^2}\right) + \sigma \left\|\Phi\right\|^2_{L^2}.$$

As shown in the proof of Proposition 3.4.2, there exists a direction $\Phi \in L^2(A,\mathbb{C})$ with

$$c := -\frac{2}{N}\sum_{j=1}^N \operatorname{Re}\left(\left(g_j \circ \phi_{t_j}, \mathcal{F}^{-1}\big(\Phi \star_{T_{Z_j}} \Phi\big)\right)_{L^2}\right) < 0$$

and it follows that $C^2_{0,\Phi,t,\sigma} = c + \sigma\left\|\Phi\right\|^2_{L^2} < 0$ for all $\sigma < -c\left\|\Phi\right\|^{-2}_{L^2}$. Therefore, the functional $\Psi \mapsto E_\sigma[\Psi,t]$ is not convex for all $\sigma \in \big[0, -c\left\|\Phi\right\|^{-2}_{L^2}\big)$.

On the other hand, if the second order Gâteaux differential $\big\langle \partial_\Psi E^{(2)}_\sigma[\Psi,t],\Phi\big\rangle$ is non-negative for all $\Psi,\Phi \in L^2(A,\mathbb{C})$ with $\Phi \neq 0$, then the functional $\Psi \mapsto E_\sigma[\Psi,t]$ is convex. First, note that

$$\left|\left(g_j \circ \phi_{t_j}, \mathcal{F}^{-1}\big(\Phi \star_{T_{Z_j}} \Phi\big)\right)_{L^2}\right| \leq \left\|g_j\right\|_{L^2}\left\|\Phi \star_{T_{Z_j}} \Phi\right\|_{L^2} \leq \left\|g_j\right\|_{L^2}\left\|\Phi\right\|^2_{L^2}\left\|T_{Z_j}\right\|_\infty \tag{3.19}$$

holds by the Cauchy-Schwarz inequality and Lemma 2.1.2. Denote the difference of the simulated images and the experimental images by $D_{\Psi,j} := \Psi \star_{T_{Z_j}} \Psi - \mathcal{F}(g_j \circ \phi_{t_j})$ for all $\Psi \in L^2(A,\mathbb{C})$ and all $j \in \{1,\ldots,N\}$. By Lemma 3.3.3 and Corollary 3.3.4, the second order Gâteaux differential of E_σ with respect to the exit wave is given by $\big\langle \partial_\Psi E^{(2)}_\sigma[\Psi,t],\Phi\big\rangle = 2!C^2_{\Psi,\Phi,t,\sigma}$ with

$$\begin{aligned}C^2_{\Psi,\Phi,t,\sigma} &= \frac{1}{N}\sum_{j=1}^N \left\|\Psi \star_{T_{Z_j}} \Phi + \Phi \star_{T_{Z_j}} \Psi\right\|^2_{L^2} + 2\operatorname{Re}\Big(\big(D_{\Psi,j}, \Phi \star_{T_{Z_j}} \Phi\big)_{L^2}\Big) + \sigma\left\|\Phi\right\|^2_{L^2} \\ &\geq \frac{2}{N}\sum_{j=1}^N \operatorname{Re}\left(\big(\Psi \star_{T_{Z_j}} \Psi - \mathcal{F}(g_j \circ \phi_{t_j}), \Phi \star_{T_{Z_j}} \Phi\big)_{L^2}\right) + \sigma\left\|\Phi\right\|^2_{L^2} \\ &\geq -\frac{2}{N}\sum_{j=1}^N \operatorname{Re}\left(\left(g_j \circ \phi_{t_j}, \mathcal{F}^{-1}\big(\Phi \star_{T_{Z_j}} \Phi\big)\right)_{L^2}\right) + \sigma\left\|\Phi\right\|^2_{L^2},\end{aligned}$$

since $(\Psi \star_{T_{Z_j}} \Psi, \Phi \star_{T_{Z_j}} \Phi)_{L^2} = \big(\mathcal{F}^{-1}(\Psi \star_{T_{Z_j}} \Psi), \mathcal{F}^{-1}(\Phi \star_{T_{Z_j}} \Phi)\big)_{L^2} \geq 0$ by Lemma 2.1.7. Using Equation (3.19), it follows that

$$\begin{aligned} C^2_{\Psi,\Phi,t,\sigma} &\geq -\frac{2}{N}\sum_{j=1}^{N} \|g_j\|_{L^2} \|\Phi\|^2_{L^2} \left\|T_{Z_j}\right\|_\infty + \sigma \|\Phi\|^2_{L^2} \\ &= \left(-\frac{2}{N}\sum_{j=1}^{N} \|g_j\|_{L^2} \left\|T_{Z_j}\right\|_\infty + \sigma\right) \|\Phi\|^2_{L^2} \end{aligned}$$

for all $\Psi, \Phi \in L^2(A, \mathbb{C})$. This shows that the functional $\Psi \mapsto E_\sigma[\Psi, t]$ is convex for all $\sigma \geq \frac{2}{N}\sum_{j=1}^{N} \|g_j\|_{L^2} \left\|T_{Z_j}\right\|_\infty$. It is interesting to note that $(\Psi_*, t_*) := (0,0)$ is the unique global minimizer of E_σ in this case (unique up to the choice of t_*, which is arbitrary and irrelevant to the value of the objective functional if $\Psi_* = 0$): this follows from the fact that

$$E_\sigma[\Psi, t] = E_\sigma[0 + 1 \cdot \Psi, t] = \underbrace{C^4_{0,\Psi,t,\sigma}}_{>0} \cdot 1^4 + \underbrace{C^2_{0,\Psi,t,\sigma}}_{\geq 0} \cdot 1^2 + \underbrace{C^0_{0,\Psi,t,\sigma}}_{=E_\sigma[0,t]} > E_\sigma[0,t] = E_\sigma[0,0]$$

holds for all $\Psi \in L^2(A, \mathbb{C})\backslash\{0\}$ and all $t \in (\mathbb{R}^2)^N$ by Lemma 3.4.1.

3.4.2 Local convexity in the neighborhood of a global minimizer

In the ideal situation of an aberration-free and fully coherent TEM, image simulation reduces to a plain autocorrelation of the exit wave. The corresponding functional is then

$$J : L^2(A, \mathbb{C}) \to \mathbb{R}, \quad \Psi \mapsto \|\Psi \star \Psi - \mathcal{F}g\|^2_{L^2}$$

for a non-empty and bounded subset $A \subseteq \mathbb{R}^2$ and an image $g \in L^2(A, \mathbb{R}_{\geq 0})$. Since A is bounded, it follows as in Corollary 2.1.4 that $\Psi \star \Psi \in L^2(\mathbb{R}^2, \mathbb{C})$ and thus J is well-defined.

The proof of the following proposition makes use of the clamping operator defined as $[\cdot]_a^b : \mathbb{R} \to [a, b],\ x \mapsto [x]_a^b := \max(\min(x, b), a)$ for all $a, b \in \mathbb{R}$ with $a \leq b$. A useful property of the clamping operator is

$$x \left[\frac{1}{x}\right]_{-d}^{d} \leq 1 \qquad \forall\, x \in \mathbb{R}\backslash\{0\} \tag{3.20}$$

for all $d \geq 0$, which can be shown with a simple case analysis: if $x \in \mathbb{R}\backslash\{0\}$ with $|x| < \frac{1}{d}$, then $\left|\frac{1}{x}\right| > d$ and consequently

$$x \left[\frac{1}{x}\right]_{-d}^{d} = x\operatorname{sign}(x) d = |x|\, d < \frac{1}{d} d = 1.$$

Otherwise, if $x \in \mathbb{R}\backslash\{0\}$ with $|x| \geq \frac{1}{d}$, then $\left|\frac{1}{x}\right| \leq d$ and

$$x \left[\frac{1}{x}\right]_{-d}^{d} = x\frac{1}{x} = 1.$$

Proposition 3.4.3. *Let* $\emptyset \neq A \subseteq \mathbb{R}^2$ *be a bounded subset and* $g \in L^2(A, \mathbb{R}_{\geq 0})\backslash\{0\}$. *If there exists a* $\Psi_* \in L^2(A, \mathbb{C})$ *with* $\mathcal{F}g = \Psi_* \star \Psi_*$, *then the functional*

$$J : L^2(A, \mathbb{C}) \to \mathbb{R}, \quad \Psi \mapsto \|\Psi \star \Psi - \mathcal{F}g\|_{L^2}^2$$

is not convex on any neighborhood of a global minimizer.

Proof. Fix a global minimizer $\Psi_* \in L^2(A, \mathbb{C})$. Since g is nonzero and $\Psi_* \star \Psi_* = \mathcal{F}g$, it follows that $\Psi_* \neq 0$. Similar to Corollary 3.3.4, the second order Gâteaux differential of J is given by

$$\langle J''[\Psi], \Phi\rangle = 2\,\|\Psi \star \Phi + \Phi \star \Psi\|_{L^2}^2 + 4\mathrm{Re}\left((\Psi \star \Psi - \mathcal{F}g, \Phi \star \Phi)_{L^2}\right)$$

for all $\Psi, \Phi \in L^2(A, \mathbb{C})$. We consider the line between Ψ_* and the origin and show that for every $\varepsilon \in (0, 2)$ there exists a direction $\Phi \in L^2(A, \mathbb{C})\backslash\{0\}$ such that

$$\langle J''[(1 - \varepsilon)\Psi_*], \Phi\rangle < 0.$$

Since $\Psi_* \star \Psi_* = \mathcal{F}g$, the second order Gâteaux differential at $(1-\varepsilon)\Psi_*$ can be expressed as

$$\begin{aligned}
&\langle J''[(1 - \varepsilon)\Psi_*], \Phi\rangle \\
&= 2(1-\varepsilon)^2\,\|\Psi_* \star \Phi + \Phi \star \Psi_*\|_{L^2}^2 + 4\mathrm{Re}\left(\left((1-\varepsilon)^2(\Psi_* \star \Psi_*) - \Psi_* \star \Psi_*, \Phi \star \Phi\right)_{L^2}\right) \\
&= 2(1-\varepsilon)^2\,\|\Psi_* \star \Phi + \Phi \star \Psi_*\|_{L^2}^2 + 4(\varepsilon^2 - 2\varepsilon)\mathrm{Re}\left((\Psi_* \star \Psi_*, \Phi \star \Phi)_{L^2}\right)
\end{aligned} \tag{3.21}$$

for all $\Phi \in L^2(A, \mathbb{C})$. Let $\psi_* := \mathcal{F}^{-1}(\Psi_*)$ and consider the functions $\phi_{d,r} \in L^2(\mathbb{R}^2, \mathbb{C})\backslash\{0\}$ for all $d, r > 0$ given by

$$\mathrm{Re}\big(\phi_{d,r}\big)(x) = \begin{cases} 0, & \text{if } x \notin B_r(0), \\ 0, & \text{if } x \in B_r(0) \text{ and } \mathrm{Im}\big(\widetilde{\psi}_*(x)\big) = 0, \\ 1, & \text{if } x \in B_r(0) \text{ and } \mathrm{Re}\big(\widetilde{\psi}_*(x)\big) = 0 \text{ and } \mathrm{Im}\big(\widetilde{\psi}_*(x)\big) \neq 0, \\ \left[\frac{1}{\mathrm{Re}\left(\widetilde{\psi}_*(x)\right)}\right]_{-d}^{d}, & \text{otherwise,} \end{cases}$$

$$\mathrm{Im}\big(\phi_{d,r}\big)(x) = \begin{cases} 0, & \text{if } x \notin B_r(0), \\ 0, & \text{if } x \in B_r(0) \text{ and } \mathrm{Re}\big(\widetilde{\psi}_*(x)\big) = 0, \\ 1, & \text{if } x \in B_r(0) \text{ and } \mathrm{Im}\big(\widetilde{\psi}_*(x)\big) = 0 \text{ and } \mathrm{Re}\big(\widetilde{\psi}_*(x)\big) \neq 0, \\ \left[\frac{-1}{\mathrm{Im}\left(\widetilde{\psi}_*(x)\right)}\right]_{-d}^{d}, & \text{otherwise,} \end{cases}$$

for all $x \in \mathbb{R}^2$, where the function $\widetilde{\psi}_* : \mathbb{R}^2 \to \mathbb{C}$ is an arbitrary, but fixed representative of $\psi_* \in L^2(\mathbb{R}^2, \mathbb{C})$. If we define $\Phi_{d,r} := \mathcal{F}(\phi_{d,r})$ for all $d, r > 0$, then $\Phi_{d,r} \in L^2(\mathbb{R}^2, \mathbb{C})\backslash\{0\}$ since $\phi_{d,r}$ is square integrable and nonzero for all $d, r > 0$. Then $\mathrm{Re}(\psi_*) = \mathcal{F}^{-1}\big(\Psi^H\big)$ and $i\mathrm{Im}(\psi_*) = \mathcal{F}^{-1}\big(\Psi^{\overline{H}}\big)$ since Ψ^H is Hermitian and $\Psi^{\overline{H}}$ is skew-Hermitian and similarly for $\phi_{d,r}$.

Using Lemma 1.4.1 and the fact that the Fourier transform is unitary, the first summand in Equation (3.21) is equal to

$$\begin{aligned}
&2(1-\varepsilon)^2 \left\|\Psi_* \star \Phi + \Phi \star \Psi_*\right\|_{L^2}^2 \\
&= 2(1-\varepsilon)^2 \left\|2\left(\Phi^H \star \Psi_*^H\right) + 2\left(\Phi^{\overline{H}} \star \Psi_*^{\overline{H}}\right)\right\|_{L^2}^2 \\
&= 8(1-\varepsilon)^2 \left\|\mathcal{F}^{-1}\left(\Phi^H\right)^* \mathcal{F}^{-1}\left(\Psi_*^H\right) + \mathcal{F}^{-1}\left(\Phi^{\overline{H}}\right)^* \mathcal{F}^{-1}\left(\Psi_*^{\overline{H}}\right)\right\|_{L^2}^2 \\
&= 8(1-\varepsilon)^2 \left\|\mathrm{Re}(\phi)\mathrm{Re}(\psi_*) + \mathrm{Im}(\phi)\mathrm{Im}(\psi_*)\right\|_{L^2}^2
\end{aligned}$$

for all $\Phi \in L^2(A, \mathbb{C})$ with $\phi := \mathcal{F}^{-1}(\Phi)$. Substituting $\Phi_{d,r}$ for Φ, it follows that

$$\left\|\Psi_* \star \Phi_{d,r} + \Phi_{d,r} \star \Psi_*\right\|_{L^2}^2 = 4\left\|\mathrm{Re}(\phi_{d,r})\mathrm{Re}(\psi_*) + \mathrm{Im}(\phi_{d,r})\mathrm{Im}(\psi_*)\right\|_{L^2}^2 \underset{d\to\infty}{\longrightarrow} 0 \tag{3.22}$$

for all $r > 0$, which is a consequence of the definition of $\phi_{d,r}$ and can be shown as follows. Let $\theta_{d,r} := \mathrm{Re}(\phi_{d,r})\mathrm{Re}(\psi_*) + \mathrm{Im}(\phi_{d,r})\mathrm{Im}(\psi_*)$. By the definition of $\phi_{d,r}$ and Equation (3.20), the functions $\theta_{d,r}$ are essentially bounded by 2 for all $d, r > 0$. Furthermore, $\theta_{d,r}$ converges pointwise almost everywhere to 0 as $d \to \infty$ and $\mathrm{supp}(\theta_{d,r}) \subseteq \overline{B_r(0)}$ for all $r > 0$ so that the convergence in Equation (3.22) follows from the dominated convergence theorem. Note that the prescribed value of 1 for $\mathrm{Re}(\phi_{d,r})$ and $\mathrm{Im}(\phi_{d,r})$ in the third case is irrelevant to this limit, but important for the properties below.

Continuing with the second summand in Equation (3.21), substituting $\Phi_{d,r}$ for Φ and taking the inverse Fourier transform yields

$$4(\varepsilon^2 - 2\varepsilon)\mathrm{Re}\left((\Psi_* \star \Psi_*, \Phi_{d,r} \star \Phi_{d,r})_{L^2}\right) = 4(\varepsilon^2 - 2\varepsilon)\mathrm{Re}\left(\left(|\psi_*|^2, |\phi_{d,r}|^2\right)_{L^2}\right).$$

Because of the three properties

1. $|\phi_{d,r}(x)| \geq |\phi_{d',r}(x)|$ for all $x \in \mathbb{R}^2$ and all $d \geq d'$,
2. $\phi_{d,r}(x) \neq 0$ for almost all $x \in B_r(0)$ with $\psi_*(x) \neq 0$ and
3. $\psi_* \neq 0$,

it follows that there is an $r_* > 0$ such that

$$\mathrm{Re}\left(\left(|\psi_*|^2, |\phi_{d,r_*}|^2\right)_{L^2}\right) \geq \mathrm{Re}\left(\left(|\psi_*|^2, |\phi_{1,r_*}|^2\right)_{L^2}\right) =: c > 0$$

for all $d \geq 1$.

Summing up,

$$\limsup_{d\to\infty} \langle J''[(1-\varepsilon)\Psi_*], \Phi_{d,r_*}\rangle \leq 2(1-\varepsilon)^2 \cdot 0 + 4(\varepsilon^2 - 2\varepsilon)c < 0$$

for all $\varepsilon \in (0, 2)$. This implies that for every $\varepsilon \in (0, 2)$ there exists a $d \geq 1$ with $\langle J''[(1-\varepsilon)\Psi_*], \Phi_{d,r_*}\rangle < 0$. □

4 Numerical minimization of the objective functional

This chapter addresses numerous issues that are relevant to a numerical minimization of the objective functional.

In Section 4.1, the MIMAP and MAL algorithms that are used for the minimization of the corresponding MIMAP and MAL functionals are reviewed briefly. Although the methods used for the registration of the focus series in the MIMAP and MAL algorithms appear to be very different at first glance, it is pointed out that both methods actually yield the same result.

In Section 4.2, the discretization of TEM images and the exit wave is formalized and the precise relationship between the discrete Fourier transform and the continuous Fourier transform is investigated in detail. This forms the basis for the subsequent discretization of the objective functional given in the remainder of this section.

In Section 4.3, the minimization algorithms that are considered for the minimization of the discretized objective functional are briefly summarized. This includes the steepest descent, the nonlinear conjugate gradient method, Quasi-Newton type methods and the Gauss-Newton method.

In Section 4.4, several issues with the naive minimization of the objective functional are addressed, which are both important for the convergence speed of the minimization and the correctness of the result. In particular, an efficient formula for the computation of the partial derivatives of the discrete objective functional is derived if the TCC is $\star$-separable. Furthermore, three extensions of the discrete objective functional are discussed: a simple preconditioner in the form of several scaling factors (Section 4.4.2), extending the size of the reconstructed exit wave to include a buffer zone (Section 4.4.3) and individual integration domains for each input image (Section 4.4.4).

Finally, Section 4.5 provides a brief overview of the implementation of the C++ program that has been used to perform the exit wave reconstructions shown in this thesis.

4.1 The MIMAP and MAL algorithms

For the sake of convenience, the definition of the MIMAP and MAL functionals from Section 3.1 is repeated here. The MIMAP functionals are defined as

$$J_\sigma^{\mathrm{MIMAP},m}[\Psi] = \sum_{j=1}^{N} \left\| \Psi \star_{T_j^{\mathrm{MIMAP},m}} \Psi - \mathcal{F}(g_j) \right\|_{L^2}^2 + \sigma \left\| \Psi - \Psi_M \right\|_{L^2}^2$$

for $m \in \{1, 2\}$ and the MAL functional is

$$J^{\mathrm{MAL}}[\Psi] = \frac{1}{N} \sum_{j=1}^{N} \left\| \Psi \star_{T_{Z_j}^{\mathrm{MAL}}} \Psi - \mathcal{F}(g_j \circ \phi_{t_j}) \right\|_{L^2}^2 .$$

The TCCs of the MIMAP functionals are given in Section 3.1.1 and T_Z^{MAL} is defined in Section 2.2.4.

4.1.1 The MIMAP algorithm

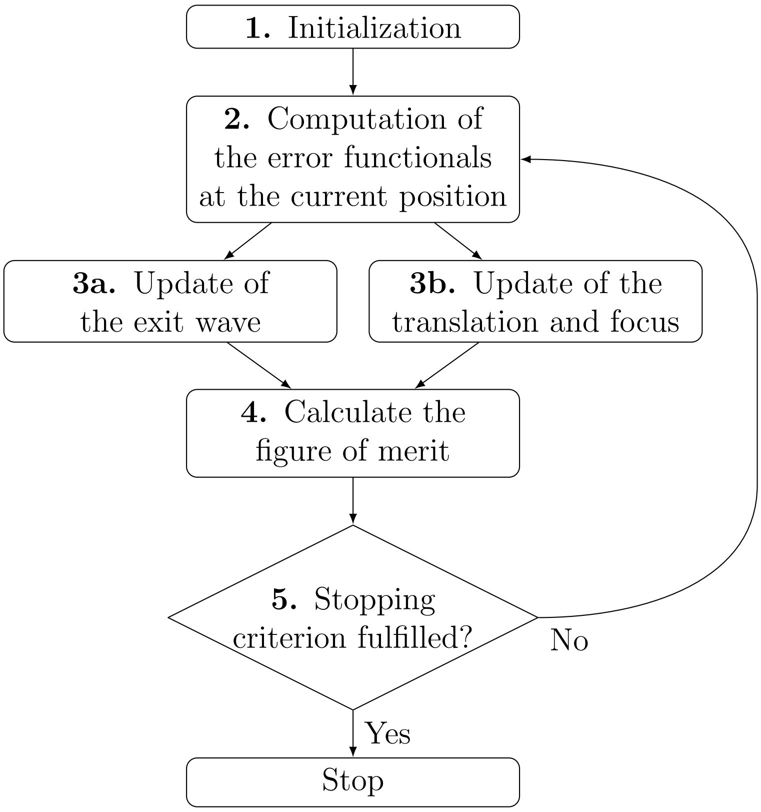

Figure 4.1: Flow diagram of the MIMAP algorithm

The MIMAP algorithm consists of the six steps illustrated in Figure 4.1. In the original algorithm as described in [38], the exit wave is iteratively reconstructed along with an optimization of both the image shifts and the focus values. Although the reconstruction of the exit wave is performed in parallel with the optimization of the image shifts and focus values, these steps are still performed independently from one another. From a theoretical point of view, this approach does not guarantee a decrease of the energy in every iteration, although in practice an update of the exit wave is unlikely to change the descent direction for the translations significantly and vice versa.

The minimization of the MIMAP functionals is very similar to a steepest descent with manual step size control and works as follows:

1. Initialization: In the first step, the initial guess for the exit wave, the translations and the focus values is computed.

The a-priori estimate Ψ_M from the regularizer and the initial guess for the exit wave $\Psi^{(0)}$ are initialized to the same value. Kirkland suggests using either the result from any fast linear reconstruction method or $\Psi_M := \Psi^{(0)} := \mathcal{F}(\tilde{g})a$, where a is the aperture function and

$$\tilde{g}(x) = \sqrt{g_k(x)} + \mathrm{i}\left(\sqrt{g_k(x)} - \sqrt{\|g_k\|_\infty}\right) \tag{4.1}$$

for a $k \in \{1, \dots, N\}$ such that g_k is the “best” focus image (i.e. such that Z_k is closest to the Scherzer focus). The derivation of Equation (4.1) is based on the weak phase object approximation.

The translations are initialized to zero and the focus values are initialized to the experimentally measured values.

2. Computation of the error functionals: In the article [38], the minimization of the MIMAP functionals is rephrased as a fixed-point problem by expressing the updates as

$$\begin{aligned}\Psi^{(k+1)} &= \Psi^{(k)} + \gamma_1\left(\Psi^{(k)}\right)\widetilde{E}_1^m\left[\Psi^{(k)}\right],\\ \left(t^{(k+1)}, Z^{(k+1)}\right) &= \left(t^{(k)}, Z^{(k)}\right) + \gamma_2\widetilde{E}_2^m\left[t^{(k)}, Z^{(k)}\right],\end{aligned} \qquad (k \in \mathbb{N}_0),$$

where $\widetilde{E}_1^m, \widetilde{E}_2^m$ for $m \in \{1, 2\}$ are suitable error functionals and $\gamma_1^m : L^2(\mathbb{R}^2, \mathbb{C}) \to \mathbb{R}$, $\gamma_2 \in \mathbb{R}$ are the corresponding step sizes. The error functionals satisfy $\widetilde{E}_1^m[\Psi] \to 0$ and $\widetilde{E}_2^m[t, Z] \to 0$ as (Ψ, t, Z) tends to a local minimizer of $J_\sigma^{\mathrm{MIMAP},m}$.

Note that the derivation of the error functionals for $m = 1$ and $m = 2$ is done very differently in [38], although the methods are essentially identical. The error functionals of $J_\sigma^{\mathrm{MIMAP},1}$ are derived from the continuous real space functional, whereas the error functionals of $J_\sigma^{\mathrm{MIMAP},2}$ are derived from the discretized Fourier space functional. In order to maintain a consistent approach, here the error functionals are all given with regard to the continuous Fourier space versions of the functionals.

The error functional $\widetilde{E}_1^m$ for the exit wave is simply the Gâteaux differential

$$\widetilde{E}_1^m[\Psi] = \partial_\Psi J_\sigma^{\mathrm{MIMAP},m}[\Psi] \in L^2(\mathbb{R}^2, \mathbb{C})',$$

which can be identified with an element of $L^2(\mathbb{R}^2, \mathbb{C})$ via the canonical isomorphism $L^2(\mathbb{R}^2, \mathbb{C})' \cong L^2(\mathbb{R}^2, \mathbb{C})$. The step size is chosen such that the error functional is on the same scale as the current estimate for the exit wave, resulting in

$$\gamma_1^m(\Psi) = -\frac{1}{4}\left(\sum_{j=1}^N \left\|\Psi_{T_{Z_j}^{\mathrm{MIMAP},m}}\Psi\right\|_{L^2}^2\right)^{-1} \qquad \forall\, \Psi \in L^2(\mathbb{R}^2, \mathbb{C}).$$

The error functional $\widetilde{E}_2^m$ is evaluated by calculating the derivatives of $J_\sigma^{\mathrm{MIMAP},m}$ with respect to the translations in both directions and the focus values symbolically. For every $j \in \{1, \ldots, N\}$, all three of these derivatives have the form $\int f_j(x) g_{j,k}(x) \,\mathrm{d}x$ with $k \in \{1, 2, 3\}$ and suitable functions $f_j, g_{j,k} : \mathbb{R}^2 \to \mathbb{R}$, where f_j depends on Z_j and t_j. It follows that all derivatives vanish if $f_j = 0$ for all $j \in \{1, \ldots, N\}$ and thus the error functional $\widetilde{E}_2^m$ is calculated by finding appropriate values for Z_j and t_j such that $f_j = 0$ for all $j \in \{1, \ldots, N\}$. The constant $\gamma_2 := \frac{1}{2}$ is chosen as the step size. More details can be found in [38, sections 5.3 and 6.3].

3a/3b. Update of the arguments: This step simply consists of adding the values of the error functionals calculated in step 2 to the previous estimate and setting $\left(\Psi^{(k+1)}, t^{(k+1)}, Z^{(k+1)}\right)$ as the current estimate.

4. Calculate the figure of merit: Treat $J_\sigma^{\mathrm{MIMAP},m}$ as a function of the exit wave, the translations and the focus values. If $\left(\Psi^{(k)}, t^{(k)}, Z^{(k)}\right)$ is the current estimate, then

$$\varepsilon^{(k)} := \frac{J_0^{\mathrm{MIMAP},m}\left[\Psi^{(k)}, t^{(k)}, Z^{(k)}\right]}{\sum_{j=1}^{N} \|g_j\|_{L^2}^2}$$

is the figure of merit for the current iteration.

5. Stopping criterion: The minimization is stopped if $\varepsilon^{(k)} \approx \varepsilon^{(k-1)}$ within a given accuracy, that is, if the figure of merit is not sufficiently decreasing anymore.

Remark 4.1.1. The MIMAP algorithm was later improved in [40] to perform the minimization with the nonlinear conjugate gradient method using the Polak-Ribière formula (see. Equation (4.11) in Section 4.3) for the descent direction. Furthermore, an extension of the algorithm is suggested in [40] to correct for rotational misalignment of the image series as well. This can be implemented in a straightforward way by minimizing the MIMAP functionals with respect to the rotation of each image after every few iterations.

4.1.2 The MAL algorithm

The MAL algorithm as described in [10] consists of the six steps illustrated in Figure 4.2. As opposed to the MIMAP algorithm, no optimization of the focus values and rotational misalignment is incorporated in the algorithm. Another slight difference is the fact that the exit wave and the translations are optimized alternatingly instead of in parallel. The MAL algorithm is implemented in the software package TrueImage [45] and an application of the MAL algorithm can be found in [5] for example.

The algorithm works as follows:

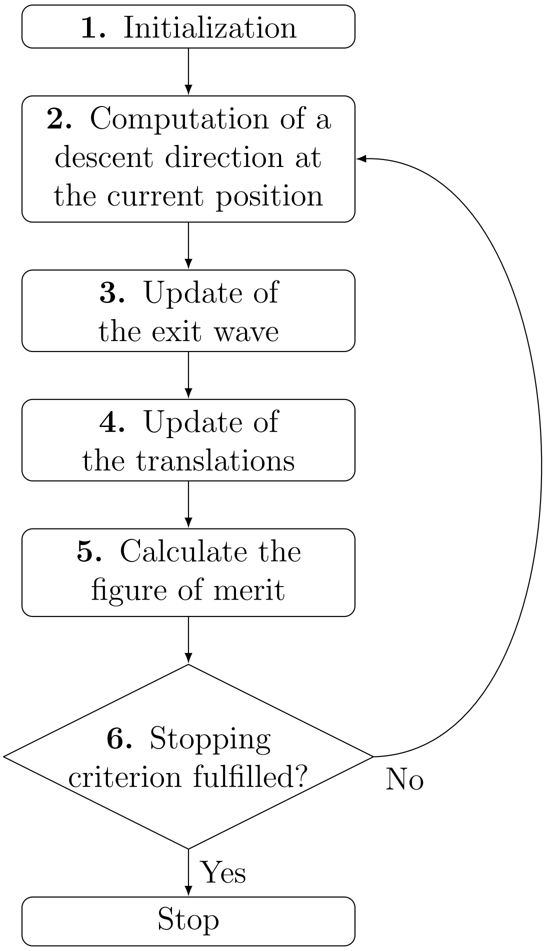

Figure 4.2: Flow diagram of the MAL algorithm

1. Initialization: At first, an initial guess for the exit wave and the translations is computed.

Two different kinds of initial estimates are suggested for the exit wave. On the one hand, the reconstructed exit wave after one iteration of the paraboloid method (PAM) [13] can be used as an initial guess. On the other hand, a constant exit wave corresponding to the square root of the mean image intensity may also be used. The latter is particularly useful for the verification of the validity of this approach for exit wave reconstruction, as in this case the initial guess is essentially unbiased and does not contain any a-priori information except for the mean image intensity.

The image series is pre-aligned by finding the peak position of the cross-correlation of every pair of successive images in the series.

2. Computation of a descent direction: In [10, p. 115], the derivatives of the MAL functional are calculated in the "directions" $\Phi^*(v)$ and $\Phi(-v)$ for all frequencies $v \in \mathbb{R}^2$.

This results in

$$\frac{2}{N}\sum_{j=1}^{N}\int_{\mathbb{R}^2} D_{\Psi,j}(v-w)T_{Z_j}(w,v)\Psi^*(w)\,\mathrm{d}w \tag{4.2}$$

as the derivative at the position Ψ in the direction $\Phi^*(v)$ and

$$\frac{2}{N}\sum_{j=1}^{N}\int_{\mathbb{R}^2} D_{\Psi,j}(v-w)T_{Z_j}(-v,-w)\Psi^*(-w)\,\mathrm{d}w \tag{4.3}$$

as the derivative at the position Ψ in the direction $\Phi(-v)$ [10, Eqs. (19a) and (19b)], where $D_{\Psi,j} = \Psi \star_{T_{Z_j}^{\mathrm{MAL}}} \Psi - G_j$ with $G_j = \mu_{t_j}\mathcal{F}(g_j)$. A comparison with Lemma 3.3.3 shows that

$$\begin{aligned}(4.2) &= C^1_{\Psi,\delta_v,t,0} + \mathrm{i}C^1_{\Psi,\mathrm{i}\delta_v,t,0} \quad\text{and}\\ (4.3) &= C^1_{\Psi,\delta_{-v},t,0} + \mathrm{i}C^1_{\Psi,\mathrm{i}\delta_{-v},t,0}.\end{aligned}$$

Either both (4.2) and (4.3) are combined in order to calculate a descent direction $d(\Psi)$ (cf. [10, section 4.1]) or $d(\Psi) := (4.2)$ is used as the descent direction.

3. Update of the exit wave: Based on the current descent direction and the search direction of the previous iteration $s\big(\Psi^{(k-1)}\big)$, the search direction $s\big(\Psi^{(k)}\big)$ for the current iteration is calculated using the nonlinear conjugate gradient method with the Polak-Ribière formula (see Equation (4.11) in Section 4.3).

The step size γ is calculated as follows. The function $p(\gamma) := J^{\mathrm{MAL}}\big[\Psi^{(k)} + \gamma s\big(\Psi^{(k)}\big)\big]$ is a polynomial of degree 4 in γ, which is approximated with the quadratic polynomial q defined by $p(0)$, $p\big(\frac{1}{2}\big)$ and $p(1)$. The minimizer of q is then chosen as the step size γ, assuming that q is convex.

Consequently, the exit wave is updated by $\Psi^{(k+1)} := \Psi^{(k)} + \gamma s\big(\Psi^{(k)}\big)$ and k is incremented by 1.

4. Update of the translations: The alignment of the images is updated during the reconstruction by finding the maximizer of

$$t_j \mapsto \big(g_j \star f_{\Psi^{(k)},j}\big)(t_j) = \int_{\mathbb{R}^2} f_{\Psi^{(k)},j}(y+t_j)g_j(y)\,\mathrm{d}y,$$

which then yields the new shift of the j-th input image g_j in pixels. Here, the function $f_{\Psi^{(k)},j} = \mathcal{F}^{-1}\big(\Psi^{(k)} \star_{T_{Z_j}^{\mathrm{MAL}}} \Psi^{(k)}\big)$ is the j-th simulated image in real space.

5. Calculate the figure of merit: No specific figure of merit is defined for the MAL algorithm. However, in the examples in [10, 66] essentially the same figure of merit as in the MIMAP algorithm is used, i.e.

$$\varepsilon^{(k)} := J^{\mathrm{MAL}}\left[\Psi^{(k)}\right],$$

where it is assumed that the image intensities are normalized.

6. Stopping criterion: The same stopping criterion as in the MIMAP algorithm may be used in the MAL algorithm as well. If $\varepsilon^{(k)} \approx \varepsilon^{(k-1)}$ within a prescribed tolerance, then the minimization is stopped and the reconstructed exit wave is returned.

4.1.3 Comparison of the registration methods

The registration of the image series is handled differently in the MAL and the MIMAP algorithms. In the MIMAP algorithm, the alignment of the experimental images is improved during every iteration by considering $J_\sigma^{\mathrm{MIMAP},m}$ as a function of the translations t_j and performing one minimization step of the functional with respect to the translations. In the MAL algorithm on the other hand, the registration is improved by calculating the peak position of the cross-correlation of each experimental image with the corresponding simulated image.

Although these methods follow entirely different approaches, they are in fact equivalent from the point of view of continuous functions (i.e. before considering the discretization). Let $m \in \{1,2\}$ and $\widetilde{T}_j^{\mathrm{MIMAP},m}(v,w) := T_j^{\mathrm{MIMAP},m}(v,w)/(\mu_{t_j}(v)\mu_{t_j}^*(w))$. Then

$$\begin{aligned} J_\sigma^{\mathrm{MIMAP},m}[\Psi] &= \sum_{j=1}^{N} \left\| \Psi \star_{T_j^{\mathrm{MIMAP},m}} \Psi - \mathcal{F}(g_j) \right\|_{L^2}^2 + \sigma \left\| \Psi - \Psi_M \right\|_{L^2}^2 \\ &= \sum_{j=1}^{N} \left\| \big(\Psi \star_{\widetilde{T}_j^{\mathrm{MIMAP},m}} \Psi\big)\mu_{t_j} - \mathcal{F}(g_j) \right\|_{L^2}^2 + \sigma \left\| \Psi - \Psi_M \right\|_{L^2}^2 \\ &= \sum_{j=1}^{N} \Big\| \underbrace{\mathcal{F}^{-1}(\Psi \star_{\widetilde{T}_j^{\mathrm{MIMAP},m}} \Psi)}_{=: f_{\Psi,j,m}} \circ \phi_{t_j} - g_j \Big\|_{L^2}^2 + \sigma \left\| \Psi - \Psi_M \right\|_{L^2}^2 \end{aligned}$$

by Equation (2.35) and since the Fourier transform is unitary. Therefore, minimizing the MIMAP functionals

$$J_\sigma^{\mathrm{MIMAP},m}[\Psi] = \sum_{j=1}^{N} \left\| f_{\Psi,j,m} \right\|_{L^2}^2 + \left\| g_j \right\|_{L^2}^2 - 2\big(f_{\Psi,j,m} \circ \phi_{t_j}, g_j\big)_{L^2} + \sigma \left\| \Psi - \Psi_M \right\|_{L^2}^2$$

with respect to t_j is equivalent to maximizing

$$\big(f_{\Psi,j,m} \circ \phi_{t_j}, g_j\big)_{L^2} = \int_{\mathbb{R}^2} f_{\Psi,j,m}(y + t_j) g_j(y)\,\mathrm{d}y = \big(g_j \star f_{\Psi,j,m}\big)(t_j)$$

for all $j \in \{1,\dots,N\}$. The latter is precisely the method used in the MAL algorithm for updating the registration (except for the different TCC).

Remark 4.1.2. The plain cross-correlation registration method has been improved by various authors (see e.g. [69, 58]) in order to make the registration more stable for highly periodic images or to correct for rotational misalignment. The phase correlation function (PCF, [50, 36]) can additionally be used to determine the misalignment along the "z-axis", i.e. the focus difference between two images.

4.2 Discretization

The theory regarding the forward model and the inverse problem in Chapters 2 and 3 is formulated in a continuous setting, since this is convenient for the mathematical analysis of the objective functional. However, for a numerical reconstruction of the exit wave the objective functional needs to be discretized. The details of the discretization are described in this section, which includes the discretization of the images and the relationship between the continuous Fourier transform and the discrete Fourier transform.

Definition 4.2.1. *Let $d \in \mathbb{N}$, $X \in \mathbb{N}^d$ and $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$. A function*

$$\hat{f} : \{0, \dots, X_1 - 1\} \times \dots \times \{0, \dots, X_d - 1\} \subseteq \mathbb{Z}^d \to \mathbb{K}$$

is called a discrete d-dimensional image. It is identified with an element of $\mathbb{K}^{X_1 \times \dots \times X_d}$ by means of the canonical isomorphism

$$\big(\{0, \dots, X_1 - 1\} \times \dots \times \{0, \dots, X_d - 1\} \to \mathbb{K}\big) \xrightarrow{\cong} \mathbb{K}^{X_1 \times \dots \times X_d}.$$

Let $d \in \mathbb{N}$ and $X \in \mathbb{N}^d$. Define $[a, b] := [a_1, b_1] \times \dots \times [a_d, b_d] \subseteq \mathbb{R}^d$ for all $a, b \in \mathbb{R}^d$. A locally integrable function $f \in L^1_{\text{loc}}(\mathbb{R}^d, \mathbb{C})$ can be discretized on a bounded rectangular domain $[a, b] \subseteq \mathbb{R}^d$ with $a, b \in \mathbb{R}^d$ by setting

$$\hat{f}(x) := \frac{1}{\mathrm{vol}(S_x)} \int_{S_x} f(y) \, \mathrm{d}y \qquad \forall\, x \in \{0, \dots, X_1 - 1\} \times \dots \times \{0, \dots, X_d - 1\}, \tag{4.4}$$

where the sets S_x are defined as $S_x := [p_x, p_{x+1}] \subseteq \mathbb{R}^d$ with $1 = (1, \dots, 1) \in \mathbb{Z}^d$ and $p_x \in \mathbb{R}^d$ given by

$$(p_x)_j := a_j + x_j \frac{b_j - a_j}{X_j} \qquad \forall\, j \in \{1, \dots, d\}.$$

Here, x refers to a corner of S_x instead of its center as this corresponds to the discretization of the objective functional described later.

The following auxiliary lemma is essential to making the connection between the continuous Fourier transform and the discrete Fourier transform explicit.

Lemma 4.2.2. *Let $a, b \in \mathbb{R}$ with $a < b$ and $M \subseteq \mathbb{R}$ bounded. Let $f : [a, b] \to \mathbb{C}$ be Riemann-integrable and $g : [a, b] \times M \to \mathbb{C}$ such that*

- *$y \mapsto g(y, x)$ is Riemann-integrable for all $x \in M$,*
- *there exists a constant $C > 0$ with*

$$|g(y, x) - g(y, x')| \leq C\, |x - x'| \qquad \forall\, y \in [a, b]\ \forall x, x' \in M.$$

Then the functions $h_N : M \to \mathbb{C}$ defined as the Riemann sums

$$h_N(x) := \frac{b - a}{N} \sum_{n=0}^{N-1} f(y_n) g(y_n, x) \qquad \forall\, x \in M$$

with $y_n := a + \frac{b-a}{N} n$ *for all* $n \in \{0, \dots, N-1\}$ *converge uniformly to*

$$h : M \to \mathbb{C}, \quad x \mapsto \int_a^b f(y) g(y,x) \, \mathrm{d}y$$

as $N \to \infty$.

Proof. Let $\varepsilon > 0$. Since f is Riemann-integrable, $|f|$ is Riemann integrable as well and

$$\lim_{N \to \infty} \frac{b-a}{N} \sum_{n=0}^{N-1} |f(y_n)| = \int_a^b |f(y)| \, \mathrm{d}y. \tag{4.5}$$

Therefore, there exists a constant $C' > 0$ with $\frac{b-a}{N} \sum_{n=0}^{N-1} |f(y_n)| \leq \int_a^b |f(y)| \, \mathrm{d}y + C'$ for all $N \in \mathbb{N}$ because the limit in Equation (4.5) exists. Since M is bounded, there exists a finite subset $S \subseteq M$ with

$$\forall\, x \in M \; \exists\, \tilde{x} \in S : \quad |x - \tilde{x}| < \frac{\varepsilon}{3C \left(\int_a^b |f(y)| \, \mathrm{d}y + C' \right)}. \tag{4.6}$$

Now choose $N \in \mathbb{N}$ sufficiently large such that $|h(\tilde{x}) - h_N(\tilde{x})| < \frac{\varepsilon}{3}$ for all $\tilde{x} \in S$. This is possible because $y \mapsto f(y) g(y,x)$ is Riemann-integrable for all $x \in M$ as a product of Riemann-integrable functions and S only has finitely many elements. Let $x \in M$ be arbitrary and choose a $\tilde{x} \in S$ that satisfies Equation (4.6). It follows from the above that

$$\begin{aligned} |h(x) - h_N(x)| &= |h(x) - h(\tilde{x}) + h(\tilde{x}) - h_N(\tilde{x}) + h_N(\tilde{x}) - h_N(x)| \\ &\leq |h(x) - h(\tilde{x})| + |h(\tilde{x}) - h_N(\tilde{x})| + |h_N(\tilde{x}) - h_N(x)| \\ &\leq |h(x) - h(\tilde{x})| + \frac{\varepsilon}{3} + |h_N(\tilde{x}) - h_N(x)| \end{aligned}$$

with

$$|h(x) - h(\tilde{x})| \leq \int_a^b |f(y)| \, |g(y,x) - g(y,\tilde{x})| \, \mathrm{d}y \leq C \, |x - \tilde{x}| \int_a^b |f(y)| \, \mathrm{d}y < \frac{\varepsilon}{3}$$

and

$$\begin{aligned} |h_N(\tilde{x}) - h_N(x)| &\leq \frac{b-a}{N} \sum_{n=0}^{N-1} |f(y_n)| \, |g(y_n, \tilde{x}) - g(y_n, x)| \\ &\leq C \, |x - \tilde{x}| \left(\int_a^b |f(y)| \, \mathrm{d}y + C' \right) < \frac{\varepsilon}{3}. \end{aligned}$$

Putting it all together, $|h(x) - h_N(x)| < \varepsilon$ for all $x \in M$, which concludes the proof of the uniform convergence. □

The main result of Lemma 4.2.2 is the fact that the convergence is uniform, as the pointwise convergence of $h_N \to h$ follows immediately from the Riemann-integrability of the integrand $y \mapsto f(y)g(y,x)$ for all $x \in M$. It is also pointed out that Lemma 4.2.2 imposes almost no restrictions on the choice of f, which will be beneficial in the following.

Recall that in Section 1.4, the continuous Fourier transform of an integrable function $f \in L^1(\mathbb{R},\mathbb{C})$ is defined as

$$\mathcal{F}\big(f\big)(x) = \int_{\mathbb{R}} f(y)\exp(-2\pi \mathrm{i} xy)\,\mathrm{d}y \qquad \forall\, x \in \mathbb{R}$$

and the discrete Fourier transform of an N-periodic vector $g \in \mathbb{C}^{\mathbb{Z}}$ is defined as

$$\mathrm{DFT}_N(g)_k = \sum_{n=0}^{N-1} g_n \exp\left(-2\pi\mathrm{i}\frac{nk}{N}\right) \qquad \forall\, k \in \mathbb{Z}.$$

In the following, it is useful to consider a slightly different definition for the discrete Fourier transform. Let $N \in 2\mathbb{N}$ and define the “shifted” discrete Fourier transform $\mathrm{DFTS}_N(g)$ of an N-periodic vector $g \in \mathbb{C}^{\mathbb{Z}}$ as

$$\mathrm{DFTS}_N(g)_k := \sum_{n=-\frac{N}{2}}^{\frac{N}{2}-1} g_n \exp\left(-2\pi\mathrm{i}\frac{nk}{N}\right) \qquad \forall\, k \in \mathbb{Z}.$$

This variant of the DFT naturally occurs as a Riemann-sum approximation of the one-dimensional continuous Fourier transform in Corollary 4.2.3. It is related to the discrete Fourier transform by

$$\begin{aligned}
\mathrm{DFTS}_N(g)_k &= \sum_{n=-\frac{N}{2}}^{-1} g_n \exp\left(-2\pi\mathrm{i}\frac{nk}{N}\right) + \sum_{n=0}^{\frac{N}{2}-1} g_n \exp\left(-2\pi\mathrm{i}\frac{nk}{N}\right) \\
&= \sum_{n=N-\frac{N}{2}}^{N-1} g_{n-N} \exp\left(-2\pi\mathrm{i}\frac{(n-N)k}{N}\right) + \sum_{n=0}^{\frac{N}{2}-1} g_n \exp\left(-2\pi\mathrm{i}\frac{nk}{N}\right) \\
&= \sum_{n=N-\frac{N}{2}}^{N-1} g_n \exp\left(-2\pi\mathrm{i}\frac{nk}{N}\right) + \sum_{n=0}^{\frac{N}{2}-1} g_n \exp\left(-2\pi\mathrm{i}\frac{nk}{N}\right) \\
&= \mathrm{DFT}_N(g)_k
\end{aligned} \tag{4.7}$$

for all $k \in \mathbb{Z}$.

Combining Equation (4.7) with the following corollary, the discrete Fourier transform can be interpreted as a discretization of the continuous Fourier transform.

Corollary 4.2.3. *Let $f \in L^1(\mathbb{R},\mathbb{C})$ be Riemann-integrable on $[-L, L]$ for all $L > 0$. For all $B > 0$ and all $\varepsilon > 0$ there are constants $N \in 2\mathbb{N}$ and $L > 0$ with*

$$\left|\mathcal{F}(f)\left(\frac{k}{2L}\right) - \frac{2L}{N}\,\mathrm{DFTS}_N(g)_k\right| < \varepsilon$$

for all $k \in \mathbb{Z}$ *with* $|k| \leq 2LB$*, where* $g \in \mathbb{C}^{\mathbb{Z}}$ *is the* N*-periodic vector given by* $g_n := f(y_n)$ *for all* $n \in \{-\frac{N}{2}, \dots, \frac{N}{2} - 1\}$ *with* $y_n = \frac{2L}{N} n$*. If* $\operatorname{supp}(f) \subseteq [-a, a]$ *for an* $a > 0$*, then it is possible to choose* $L := a$ *independently of* B *and* ε*.*

Proof. Let $\varepsilon > 0$. Since f is integrable, there exists an $L > 0$ with $\|f\|_{L^1(\mathbb{R}\setminus[-L,L])} < \frac{\varepsilon}{2}$, which implies

$$\begin{aligned} \big|\mathcal{F}\big(f\big)(x) - \mathcal{F}\big(f\mathbb{1}_{[-L,L]}\big)(x)\big| &= \left|\int_{\mathbb{R}\setminus[-L,L]} f(y) e^{-2\pi \mathrm{i} xy} \,\mathrm{d}y\right| \\ &\leq \int_{\mathbb{R}\setminus[-L,L]} |f(y)| \underbrace{\left|e^{-2\pi \mathrm{i} xy}\right|}_{=1} \mathrm{d}y < \frac{\varepsilon}{2} \end{aligned}$$

for all $x \in \mathbb{R}$. By assumption, f is Riemann-integrable and if we define the function $g : [-L, L] \times [-B, B] \to \mathbb{C}$ by $g(y, x) := e^{-2\pi \mathrm{i} x \cdot y}$, then $y \mapsto g(y, x)$ is clearly also Riemann-integrable on $[-L, L]$ for all $x \in [-B, B]$. Furthermore, the existence of the constant $C > 0$ from Lemma 4.2.2 follows from the fact that g is continuously differentiable and $[-L, L] \times [-B, B]$ is compact. Therefore, there exists an $N \in 2\mathbb{N}$ such that

$$\left|\mathcal{F}\big(f\mathbb{1}_{[-L,L]}\big)(x) - \frac{2L}{N} \sum_{n=-\frac{N}{2}}^{\frac{N}{2}-1} f(y_n) \exp(-2\pi \mathrm{i} y_n x)\right| < \frac{\varepsilon}{2}$$

for all $x \in [-B, B]$, where $y_n := \frac{2L}{N} n$. In particular, for $x = x_k := \frac{1}{2L} k$ with $k \in \mathbb{Z}$ and $|k| \leq 2LB$ it follows that

$$\begin{aligned} \left|\mathcal{F}\big(f\big)(x_k) - \frac{2L}{N} \operatorname{DFTS}_N(g)_k\right| &= \left|\mathcal{F}\big(f\big)(x_k) - \frac{2L}{N} \sum_{n=-\frac{N}{2}}^{\frac{N}{2}-1} f(y_n) \exp(-2\pi \mathrm{i} y_n x_k)\right| \\ &\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon, \end{aligned}$$

since $x_k \in [-B, B]$. If there is an $a > 0$ such that $\operatorname{supp}(f) \subseteq [-a, a]$, then choosing $L := a$ yields $\big|\mathcal{F}\big(f\big)(x) - \mathcal{F}\big(f\mathbb{1}_{[-L,L]}\big)(x)\big| = 0$ for all $x \in \mathbb{R}$. ☐

Both Lemma 4.2.2 and Corollary 4.2.3 can be generalized to higher dimensional functions by using the multidimensional Riemann integral (see e.g. [15] for the definition and relevant properties of the multidimensional Riemann integral).

Let $\mathbb{R}^d_{>0} := \{x \in \mathbb{R}^d \mid x_j > 0 \;\forall j\}$ and $N \in (2\mathbb{N})^d$. Define the d-dimensional DFTS of an $(N_1, \dots, N_d)$-periodic element $g \in \mathbb{C}^{(\mathbb{Z}^d)}$ as

$$\operatorname{DFTS}_N(g)_k := \sum_{n_1=-\frac{N_1}{2}}^{\frac{N_1}{2}-1} \cdots \sum_{n_d=-\frac{N_d}{2}}^{\frac{N_d}{2}-1} g_{n_1,\dots,n_d} \exp\left(-2\pi \mathrm{i} \sum_{j=1}^{d} \frac{n_j k_j}{N_j}\right) \qquad \forall\, k \in \mathbb{Z}^d.$$

Here, an element $g \in \mathbb{C}^{(\mathbb{Z}^d)}$ is called $(N_1, \dots, N_d)$-periodic if $g_{n_1+a_1N_1,\dots,n_d+a_dN_d} = g_n$ for all $n, a \in \mathbb{Z}^d$. Similar to the one-dimensional case, the d-dimensional DFTS is equal to the d-dimensional DFT.

The following generalizations of Lemma 4.2.2 and Corollary 4.2.3 hold with identical proofs:

Lemma 4.2.4. *Let* $d, d' \in \mathbb{N}$, $a, b \in \mathbb{R}^d$ *with* $a_j < b_j$ *for all* $j \in \{1, \ldots, d\}$ *and* $M \subseteq \mathbb{R}^{d'}$ *bounded. Let* $f : [a, b] \to \mathbb{C}$ *be Riemann-integrable and* $g : [a, b] \times M \to \mathbb{C}$ *such that*

- $y \mapsto g(y, x)$ *is Riemann-integrable for all* $x \in M$,
- *there exists a constant* $C > 0$ *with*

$$|g(y, x) - g(y, x')| \le C \left\|x - x'\right\|_2 \qquad \forall\, y \in [a, b]\ \forall x, x' \in M.$$

Then the functions $h_N : M \to \mathbb{C}$ *defined as the Riemann sums*

$$h_N(x) := \frac{\operatorname{vol}([a, b])}{N^d} \sum_{n_1,\ldots,n_d=0}^{N-1} f(y_{n_1,\ldots,n_d}) g\big(y_{n_1,\ldots,n_d}, x\big) \qquad \forall\, x \in M$$

with $y_{n_1,\ldots,n_d} := (a_1 + \frac{b_1-a_1}{N} n_1, \ldots, a_d + \frac{b_d-a_d}{N} n_d)$ *converge uniformly to*

$$h : M \to \mathbb{C}, \quad x \mapsto \int_{[a,b]} f(y) g(y, x)\, \mathrm{d}y$$

as $N \to \infty$.

Corollary 4.2.5. *Let* $d \in \mathbb{N}$ *and* $f \in L^1(\mathbb{R}^d, \mathbb{C})$ *be Riemann-integrable on* $[-L, L]$ *for all* $L \in \mathbb{R}^d_{>0}$. *For all* $B \in \mathbb{R}^d_{>0}$ *and all* $\varepsilon > 0$ *there are constants* $N \in 2\mathbb{N}$ *and* $L \in \mathbb{R}^d_{>0}$ *such that*

$$\left| \mathcal{F}(f) \left(\frac{k_1}{2L_1}, \ldots, \frac{k_d}{2L_d} \right) - \frac{\operatorname{vol}([-L, L])}{N^d} \operatorname{DFTS}_N(g)_{k_1,\ldots,k_d} \right| < \varepsilon$$

for all $k \in \mathbb{Z}^d$ *with* $|k_j| \le 2L_j B_j$ *for all* $j = 1, \ldots, d$, *where* $g \in \mathbb{C}^{(\mathbb{Z}^d)}$ *is* $(N_1, \ldots, N_d)$-*periodic and defined by*

$$g_n := f(y_n) \qquad \forall\, n \in \left\{ -\frac{N_1}{2}, \ldots, \frac{N_1}{2} - 1 \right\} \times \ldots \times \left\{ -\frac{N_d}{2}, \ldots, \frac{N_d}{2} - 1 \right\} \subseteq \mathbb{Z}^d$$

with $y_n = \big(\frac{2L_1}{N_1} n_1, \ldots, \frac{2L_d}{N_d} n_d\big)$. *If* $\operatorname{supp}(f) \subseteq [-a, a]$ *for an* $a \in \mathbb{R}^d_{>0}$, *then it is possible to choose* $L = a$ *independently of* B *and* ε.

As the sign in the exponential terms of the Fourier transforms is irrelevant for the proofs of Lemma 4.2.2 and Corollary 4.2.3, it is straightforward to adapt the proofs to the inverse Fourier transform as well. If the inverse DFTS of an $(N_1, \ldots, N_d)$-periodic element $g \in \mathbb{C}^{(\mathbb{Z}^d)}$ is defined as

$$\operatorname{IDFTS}_N(g)_k := \sum_{n_1=-\frac{N_1}{2}}^{\frac{N_1}{2}-1} \cdots \sum_{n_d=-\frac{N_d}{2}}^{\frac{N_d}{2}-1} g_{n_1,\ldots,n_d} \exp\left(2\pi \mathrm{i} \sum_{j=1}^{d} \frac{n_j k_j}{N_j} \right) \qquad \forall\, k \in \mathbb{Z}^d$$

then we get the same approximation property as in Corollary 4.2.5 for the inverse transforms if the Fourier transform is replaced with the inverse Fourier transform and the DFTS is replaced with the IDFTS.

For the discretization of the objective functional, the exit wave and the TEM images are discretized as piecewise constant functions on a rectilinear grid of size $X \times Y$ for $X, Y \in 2\mathbb{N}$ with grid widths $h_x = \frac{L_X}{X}$ respectively $h_y = \frac{L_Y}{Y}$ for $L_X, L_Y > 0$. It is convenient to neglect the microscope's magnification by referring to the TEM images in the same coordinate space as the exit wave. The physical dimensions of the truncated and discretized exit wave and TEM images are then $L_X \times L_Y$ in real space respectively $\big(X/L_X\big) \times \big(Y/L_Y\big)$ in Fourier space. The origin is placed at the center of this subsection so that the Fourier space domains are restricted to $Q := \big[-\frac{X}{2L_X}, \frac{X}{2L_X}\big] \times \big[-\frac{Y}{2L_Y}, \frac{Y}{2L_Y}\big]$. Explicitly, the uniform subdivision of Q into $X \cdot Y$ rectangles is given by the subsets $\big\{S_{(x,y)} := [p_{(x,y)}, p_{(x+1,y+1)}] \subseteq Q \;\big|\; (x,y) \in \{0, \dots, X-1\} \times \{0, \dots, Y-1\}\big\}$, where

$$p_{(x,y)} := \left(\frac{x - \frac{X}{2}}{L_X}, \frac{y - \frac{Y}{2}}{L_Y}\right) \in \mathbb{R}^2.$$

Both the exit wave and the TEM images can be approximated by using the mean value formula in Equation (4.4). Overall, this discretization of the exit wave and the TEM images involves two separate kinds of errors: the approximation by piecewise constant functions within the grid cells and the restriction of the domain to Q.

In addition to the exit wave and the TEM images, the TCCs $T_Z : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{C}$ for $Z \in \mathbb{R}$ and the modulations $\mu_t : \mathbb{R}^2 \to \mathbb{C}$ for $t \in \mathbb{R}^2$ also need to be discretized. The TCCs are restricted to the domain $Q \times Q$ and discretized as a piecewise constant function on a rectilinear grid of size $(X \times Y)^2$, whereas the modulations are restricted to Q and discretized as a piecewise constant function on a rectilinear grid of size $X \times Y$. In the following, the discrete TCCs $\widehat{T}_Z \in \mathbb{C}^{(X \times Y)^2}$ and the discrete modulations $\widehat{\mu}_t \in \mathbb{C}^{X \times Y}$ are understood to be defined in terms of point evaluations of T_Z and μ_t at the top left corner of the respective pixels.

Furthermore, the weighted cross-correlation is approximated by the discrete weighted cross-correlation. For $\hat{f}, \hat{g} \in \mathbb{C}^{X \times Y}$ and $\widehat{w} \in \mathbb{C}^{(X \times Y)^2}$ the discrete weighted cross-correlation $\hat{f} \star_{\widehat{w}} \hat{g} \in \mathbb{C}^{X \times Y}$ is defined by

$$\big(\hat{f} \star_{\widehat{w}} \hat{g}\big)_{x,y} := \sum_{a=0}^{X-1} \sum_{b=0}^{Y-1} \hat{f}^*_{a,b} \hat{g}_{a+x,b+y} \widehat{w}_{(a+x,b+y),(a,b)}$$

for all $x \in \{0, \dots, X-1\}$ and $y \in \{0, \dots, Y-1\}$, where $\hat{g}$ and $\widehat{w}$ are extended to elements of $\mathbb{C}^{\mathbb{Z} \times \mathbb{Z}}$ respectively $\mathbb{C}^{(\mathbb{Z} \times \mathbb{Z})^2}$ with an appropriate continuation, e.g. zero continuation or periodic continuation. This definition does not contain the scaling by the grid widths $\frac{L_X}{X}$ and $\frac{L_Y}{Y}$ and thus is not a Riemann-sum approximation of the continuous weighted cross-correlation.

Recall that the objective functional for exit wave reconstruction is defined as

$$E_\sigma[\Psi,t] = \frac{1}{N}\sum_{j=1}^{N} \left\|\Psi \star_{T_{Z_j}} \Psi - \mu_{t_j}\mathcal{F}(g_j)\right\|_{L^2}^2 + \sigma\left\|\Psi - \Psi_M\right\|_{L^2}^2$$

for all $\Psi \in L^2(A,\mathbb{C})$ and all $t \in (\mathbb{R}^2)^N$. Denote by $\widehat{\Psi}_M \in \mathbb{C}^{X\times Y}$ a discretization of Ψ_M on the same grid and domain as the discretization of the exit wave and by $\widehat{g}_j$ for all $j \in \{1,\dots,N\}$ the discretization of the TEM images. A discretization of the objective functional E_σ is given by

$$\begin{aligned}
&\widehat{E}_\sigma\big[\widehat{\Psi},t\big] := \\
&\frac{1}{N}\sum_{j=1}^{N}\frac{L_X}{X}\frac{L_Y}{Y}\sum_{x=0}^{X-1}\sum_{y=0}^{Y-1}\left|\frac{L_X}{X}\frac{L_Y}{Y}\left(\widehat{\Psi}\star_{\widehat{T}_{Z_j}}\widehat{\Psi}\right)_{x,y} - (\widehat{\mu}_{t_j})_{x,y}\frac{L_X}{X}\frac{L_Y}{Y}\operatorname{DFT}\left(\widehat{g}_j\right)_{x,y}\right|^2 \\
&+\sigma\frac{L_X}{X}\frac{L_Y}{Y}\sum_{x=0}^{X-1}\sum_{y=0}^{Y-1}\left|\widehat{\Psi}_{x,y} - \left(\widehat{\Psi}_M\right)_{x,y}\right|^2 \\
=\;&\frac{L_X^3L_Y^3}{NX^3Y^3}\sum_{j=1}^{N}\sum_{x=0}^{X-1}\sum_{y=0}^{Y-1}\left|\left(\widehat{\Psi}\star_{\widehat{T}_{Z_j}}\widehat{\Psi}\right)_{x,y} - (\widehat{\mu}_{t_j})_{x,y}\operatorname{DFT}\left(\widehat{g}_j\right)_{x,y}\right|^2 \\
&+\sigma\frac{L_XL_Y}{XY}\sum_{x=0}^{X-1}\sum_{y=0}^{Y-1}\left|\widehat{\Psi}_{x,y} - \left(\widehat{\Psi}_M\right)_{x,y}\right|^2
\end{aligned}\tag{4.8}$$

for all $\widehat{\Psi} \in \mathbb{C}^{X\times Y}$ and $t \in (\mathbb{R}^2)^N$. The use of the discrete Fourier transform as an approximation to the continuous Fourier transform is justified by Corollary 4.2.5. For the calculation of the discrete weighted autocorrelation, zero continuation of the exit wave is reasonable, since Ψ is the exit wave in Fourier space and the low frequencies are located near $\left(\frac{X}{2},\frac{Y}{2}\right)$. If the images g_j are approximated by discrete images using Equation (4.4), then the resulting discrete images $\widehat{g}_j$ are in general not equal to the discrete images $g \in \mathbb{C}^{(\mathbb{Z}^d)}$ given in Corollary 4.2.5 (which, in any case, can only be calculated for images with point values). However, if the continuous images $g_j \in L^1(\mathbb{R}^2,\mathbb{R}_{\geq 0}) \cap L^2(\mathbb{R}^2,\mathbb{R}_{\geq 0})$ do have point values, then it is reasonable to assume that both approximations are very similar for sufficiently large X, Y. The extended objective functional E_σ^{ext} is equally discretized.

The discretization of the forward model in Equation (4.8) is useful if the TCC is not $\star$-separable and the weighted autocorrelation has to be calculated by evaluating the discrete weighted autocorrelation explicitly. If the discrete Fourier transforms of the input images are precomputed, then the complexity of evaluating $\widehat{E}_\sigma$ is $\mathcal{O}\big(N(XY)^2\big)$.

If the TCC is $\star$-separable, there exist $M \in \mathbb{N}$ and bounded and measurable functions $v_{k,j} : \mathbb{R}^2 \to \mathbb{C}$ with

$$T_{Z_j}(x,y) = \sum_{k=0}^{M} v_{k,j}(x)v_{k,j}^*(y) \qquad \forall\, x,y \in \mathbb{R}^2.$$

Let $\widehat{v}_{k,j} \in \mathbb{C}^{X \times Y}$ be discretizations of $v_{k,j}$ as piecewise constant functions on a grid of size $X \times Y$ on the domain Q. If the functions $v_{k,j}$ are approximated with $\widehat{v}_{k,j}$ by point evaluation at the top left corner of each pixel just like the TCC, then the analog relationship also holds for the discretizations, i.e.

$$\left(\widehat{T}_{Z_j}\right)_{a,b} = \sum_{k=0}^{M} \left(\widehat{v}_{k,j}\right)_a \left(\widehat{v}_{k,j}\right)_b^*$$

for all $a, b \in \{0, \ldots, X-1\} \times \{0, \ldots, Y-1\}$. It follows that

$$\widehat{\Psi} \star_{\widehat{T}_{Z_j}} \widehat{\Psi} = \sum_{k=0}^{M} \left(\widehat{\Psi}\widehat{v}_{k,j}\right) \star \left(\widehat{\Psi}\widehat{v}_{k,j}\right) = \sum_{k=0}^{M} \frac{1}{XY} \operatorname{IDFT}\left(\left|\operatorname{DFT}\left(\widehat{\Psi}\widehat{v}_{k,j}\right)\right|^2\right) \tag{4.9}$$

by Lemma 1.4.3 and the fact that $\frac{1}{XY}\operatorname{IDFT}(\operatorname{DFT}(f)) = f$ for all $f \in \mathbb{C}^{X \times Y}$. Here, the product of elements of $\mathbb{C}^{X \times Y}$ on the right-hand side of Equation (4.9) is defined by pointwise multiplication. The complexity of calculating the right-hand side is given by $\mathcal{O}\big(2MXY\log(XY)\big)$ if the fast Fourier transform is used. In this way, the overall complexity of evaluating the objective functional is reduced to $\mathcal{O}\big(2NMXY\log(XY)\big)$.

Remark 4.2.6. Lemma 1.4.3 requires that the discrete exit wave $\widehat{\Psi} \in \mathbb{C}^{X \times Y}$ is extended to an element of $\mathbb{C}^{\mathbb{Z} \times \mathbb{Z}}$ using periodic continuation. As opposed to the zero continuation, periodic continuation of the exit wave is not sensible in general, but must be considered as a necessary evil in order to be able to utilize the fast Fourier transform. In Section 4.4.3, an approach is presented to alleviate the error caused by the periodic continuation of the exit wave.

4.3 Minimization algorithms

There are several classes of algorithms to choose from for the numerical minimization of the objective functional.

First of all, derivative-free optimization algorithms such as simulated annealing or genetic algorithms can be used. An advantage of these algorithms is that they attempt to approximate a global minimum and do not get stuck in local minima as easily as local gradient-based methods. However, the drawback is that many evaluations of the objective functional are necessary for a good approximation of a global minimizer. Since the objective functional E_σ is continuously differentiable and the computational complexity of calculating the derivatives is on the same order as the evaluation of the functional (see Section 4.4.1), it is reasonable to use the additional information provided by the derivatives.

Available derivative-based optimization algorithms include the gradient descent and the nonlinear conjugate gradient method. These are the methods that are used in the MIMAP and MAL algorithms. Another possibility is using a quasi-Newton algorithm, which is based on Newton's method but does not require the computation and inversion of the Hessian. The same applies to the Gauss-Newton algorithm, which is designed to find minimizers of nonlinear least squares functionals such as the objective functional for exit wave reconstruction.

In principle, methods using higher order derivatives such as Newton's method could also be used to approximate a minimizer. An advantage of Newton's method would be the local quadratic convergence in the neighborhood of a minimizer. Nevertheless, it is the calculation of the second order derivatives of the functional E_σ that makes Newton's method unsuitable for a numerical minimization of the functional. Not only is the Hessian matrix very expensive to calculate, but its calculation should also be avoided for the translations due to the presence of noise in the experimental input data [52].

A very brief summary of the optimization algorithms that are considered for the minimization of the objective functional is given in the following. All of these algorithms return a critical point of the corresponding function, which is not necessarily a local minimizer but can also be a saddle point. More details can be found in [53].

Gradient descent The gradient descent algorithm consists of a series of line searches along the direction of the negative gradient (steepest descent direction), which are repeated until a given convergence criterion is fulfilled. More precisely, the gradient descent consists of the following steps:

```
Data: A function $f \in C^1(\Omega, \mathbb{R})$ with $\Omega \subseteq \mathbb{R}^n$ and an initial guess $x_0 \in \Omega$.
Result: A critical point $x_*$ of $f$.

Set $i := 0$
Repeat: until convergence:
  1. Calculate $d_i := -\nabla f(x_i)$
  2. Choose a step size $\tau_i > 0$ for the line search in the direction $d_i$
  3. Set $x_{i+1} := x_i + \tau_i d_i$
  4. Update $i \leftarrow i + 1$
Set $x_* := x_i$
```

The gradient descent is based on the fact that the function's value locally decreases the fastest in the direction of the negative gradient. The step size τ_i can be calculated using Armijo's rule [3] for example.

Nonlinear conjugate gradient Although the gradient descent always proceeds in the locally optimal descent direction, it is in general a very inefficient method for the approximation of a minimizer. This is due to the fact that within a given number of iterations many of the search directions are oftentimes parallel or near-parallel. Instead of choosing a locally optimal descent direction, the nonlinear conjugate gradient algorithm tries to find search directions that are more favorable globally. This is done by taking the history of search directions into account in order to choose directions that are approximately orthogonal to each other.

```
Data: A function $f \in C^1(\Omega, \mathbb{R})$ with $\Omega \subseteq \Omega$ and an initial guess $x_0 \in \mathbb{R}^n$.
Result: A critical point $x_*$ of $f$.

Set $i := 0$, $\beta_0 := 0$ and $d_{-1} := 0$
Repeat: until convergence:
  1. Calculate $d_i := -\nabla f(x_i) + \beta_i d_{i-1}$
  2. Choose a step size $\tau_i > 0$ for the line search in the direction $d_i$
  3. Set $x_{i+1} := x_i + \tau_i d_i$
  4. Calculate $\beta_{i+1}$
  5. Update $i \leftarrow i + 1$
Set $x_* := x_i$
```

Implementations of the nonlinear conjugate gradient method mainly differ in the calculation of the weighting factor β_{i+1} for the previous search direction. The nonlinear conjugate gradient method was originally devised by Fletcher and Reeves in [21] as an extension of the linear conjugate gradient method, where they proposed to choose the parameter β_{i+1} as

$$\beta_{i+1} := \frac{\|\nabla f(x_{i+1})\|_2^2}{\|\nabla f(x_i)\|_2^2}. \tag{4.10}$$

A widespread alternative is the Polak-Ribière formula [57]

$$\beta_{i+1} := \frac{\|\nabla f(x_{i+1})\|_2^2 - \nabla f(x_{i+1})^T \nabla f(x_i)}{\|\nabla f(x_i)\|_2^2}. \tag{4.11}$$

Both of these formulas are heuristics for the nonlinear optimization and can equally well be used as the weighting factor for the previous search direction.

The step size τ_i can again be calculated using Armijo's rule just like in the gradient descent algorithm. The first iteration of the conjugate gradient method is then equal to the steepest descent, since $d_{-1} = 0$.

Quasi-Newton Quasi-Newton methods provide an alternative to Newton's method if the calculation of second order derivatives is not feasible. The key idea is to approximate the Hessian matrix by regular updates that are performed based on the change in the gradients between two successive iterations. The basic structure of quasi-Newton algorithms is as follows:

```
Data: A function $f \in C^1(\Omega, \mathbb{R})$ with $\Omega \subseteq \mathbb{R}^n$ and an initial guess $x_0 \in \Omega$.
Result: A critical point $x_*$ of $f$.

Set $i := 0$ and $B_0 := I \in \mathbb{R}^{n \times n}$
Repeat: until convergence:
  1. Calculate $d_i := -B_i^{-1} \nabla f(x_i)$
  2. Choose a step size $\tau_i > 0$ for the line search in the direction $d_i$
  3. Set $x_{i+1} := x_i + \tau_i d_i$
  4. Calculate $B_{i+1}$
  5. Update $i \leftarrow i + 1$
Set $x_* := x_i$
```

Variants of the quasi-Newton method differ in the calculation of the approximation of the Hessian B_i and its inverse B_i^{-1}. A popular method is the BFGS algorithm that was proposed independently by Broyden, Fletcher, Goldfarb and Shanno in 1970. In the BFGS algorithm, the approximation to the Hessian is updated by

$$B_{i+1} := B_i + \frac{y_i y_i^T}{y_i^T \Delta x_i} - \frac{B_i \Delta x_i (B_i \Delta x_i)^T}{\Delta x_i^T B_i \Delta x_i},$$

where $y_i := \nabla f(x_{i+1}) - \nabla f(x_i)$ and $\Delta x_i := x_{i+1} - x_i$ [20]. An advantage over the gradient descent is that additional information about the Hessian is used in the calculation of the search direction. For this reason, a Quasi-Newton algorithm is closer to Newton's method than a gradient descent, which results in an improved, superlinear convergence speed. Also note that since the first approximation of the Hessian matrix $B_0 = I$ is the identity matrix, the first iteration of the quasi-Newton method is equal to one iteration of the gradient descent.

Gauss-Newton In contrast to the algorithms introduced so far, the Gauss-Newton algorithm is specifically designed to minimize functionals that are given as a sum of squares. Utilizing this additional structure makes the minimization algorithm more stable and efficient [53]. Similar to Newton's method, the solution of the nonlinear problem is approximated by iteratively solving a sequence of linearized problems.

```
Data: A function $F \in C^1(\Omega, \mathbb{R}^m)$ with $\Omega \subseteq \mathbb{R}^n$ and an initial guess $x_0 \in \Omega$.
Result: A critical point $x_*$ of $E : \Omega \to \mathbb{R},\ x \mapsto \|F(x)\|_2^2$.

Set $i := 0$
Repeat: until convergence:
  1. Solve $\big(DF\big)(x_i)^T \big(DF\big)(x_i) d_i = -\big(DF\big)(x_i)^T F(x_i)$ for $d_i$
  2. Choose a step size $\tau_i > 0$ for the line search in the direction $d_i$
  3. Set $x_{i+1} := x_i + \tau_i d_i$
  5. Update $i \leftarrow i + 1$
Set $x_* := x_i$
```

The normal equations in step 1 originate from a linear approximation of F using the Taylor expansion $F(x) \approx F(x_i) + \big(DF\big)(x_i) \cdot (x - x_i)$, which replaces the original nonlinear minimization problem by

$$\Big\|F(x_i) + \big(DF\big)(x_i) d_i\Big\|_2^2 \ \to\ \min.$$

If the Jacobian matrix is small, then the normal equations can be solved using standard techniques from numerical analysis. Otherwise, it is possible to resort to iterative methods that solve the normal equations approximately with a given precision [26, 22].

For the discretization of the objective functional $\widehat{E}_\sigma$ from Equation (4.8), the vector valued function $F : \mathbb{R}^{2XY+2N} \to \mathbb{R}^{2NXY+2XY}$ with $\widehat{E}_\sigma\big[\widehat{\Psi}, t\big] = \Big\|F\big[\widehat{\Psi}, t\big]\Big\|_2^2$ is defined such that the entries of $F\big[\widehat{\Psi},t\big] \in \mathbb{R}^{2NXY+2XY}$ are

- $\sqrt{\frac{L_X^3 L_Y^3}{NX^3Y^3}}\mathrm{Re}\left(\big(\widehat{\Psi} \star_{\widehat{T}_{Z_j}} \widehat{\Psi}\big)_{x,y} - (\widehat{\mu}_{t_j})_{x,y}\,\mathrm{DFT}\left(\widehat{g}_j\right)_{x,y}\right)$
- $\sqrt{\frac{L_X^3 L_Y^3}{NX^3Y^3}}\mathrm{Im}\left(\big(\widehat{\Psi} \star_{\widehat{T}_{Z_j}} \widehat{\Psi}\big)_{x,y} - (\widehat{\mu}_{t_j})_{x,y}\,\mathrm{DFT}\left(\widehat{g}_j\right)_{x,y}\right)$

for $j \in \{1,\dots,N\}$, $x \in \{0,\dots,X-1\}$ and $y \in \{0,\dots,Y-1\}$ as well as

- $\sqrt{\sigma\frac{L_X L_Y}{XY}}\mathrm{Re}\left(\widehat{\Psi}_{x,y} - \big(\widehat{\Psi}_M\big)_{x,y}\right)$
- $\sqrt{\sigma\frac{L_X L_Y}{XY}}\mathrm{Im}\left(\widehat{\Psi}_{x,y} - \big(\widehat{\Psi}_M\big)_{x,y}\right)$

for $x \in \{0,\dots,X-1\}$ and $y \in \{0,\dots,Y-1\}$. The Jacobian DF is a dense matrix of the size $(2NXY+2XY)(2XY+2N)$. Even for a moderate number of images and pixels in each image, its size is prohibitively large for storing the Jacobian using conventional hardware. Hence, the usual approaches for solving the normal equations in step 1 by calculating the Cholesky factorization of $\big(DF\big)(x_i)^T\big(DF\big)(x_i)$, the QR factorization of $\big(DF\big)(x_i)$ or a singular value decomposition of $\big(DF\big)(x_i)$ are not viable. Instead, it is possible to use the iterative LSMR algorithm [22] to solve the normal equations approximately. The advantage of the LSMR algorithm is that, instead of working with the Jacobian directly, it only needs to be able to compute matrix-vector products with the Jacobian and its transpose.

4.4 Practical aspects of the minimization

In this section, several results are collected that are of only little relevance to the theory, but very useful or even necessary for a successful numerical minimization of the objective functional. On the one hand, the computation of the derivatives and the scaling of the objective functional's arguments discussed in Sections 4.4.1 and 4.4.2 mainly improve the convergence speed of the minimization and reduce the computation time. On the other hand, the approaches that deal with the wrap around error and image continuation in Sections 4.4.3 and 4.4.4 are necessary for the correctness of the reconstructed exit wave.

4.4.1 Computation of the partial derivatives

The efficient numerical computation of the partial derivatives of $\widehat{E}_\sigma$ involves a few subtleties that are discussed in this section. It turns out that if the TCC is $\star$-separable, then the complete gradient can be calculated in linearithmic time $\mathcal{O}\big(XY\log(XY)\big)$ in the number of pixels of the discrete exit wave $\widehat{\Psi} \in \mathbb{C}^{X\times Y}$. If the TCC is not $\star$-separable, then the complexity is quadratic in the number of pixels, i.e. $\mathcal{O}\big((XY)^2\big)$.

The partial derivatives are calculated separately for the exit wave, the translations and the focus values and then combined to yield the complete gradient of the discretized objective functional at a given position. Although the derivatives are deduced from a discretization of the Gâteaux differentials of E_σ from Section 3.3, they are in fact identical to the partial derivatives of the discrete functional $\widehat{E}_\sigma$.

Exit wave: By Lemma 3.3.3 and Corollary 3.3.4, the first order Gâteaux differential of E_σ at a position $(\Psi, t) \in L^2(A, \mathbb{C}) \times (\mathbb{R}^2)^N$ in the direction $(\Phi, 0) \in L^2(A, \mathbb{C}) \times (\mathbb{R}^2)^N$ is

$$\langle \partial_\Psi E_\sigma[\Psi, t], \Phi \rangle = \frac{4}{N} \sum_{j=1}^{N} \mathrm{Re}\left(\left(D_{\Psi,j}, \Psi \star_{T_{Z_j}} \Phi \right)_{L^2} \right) + 2\sigma \mathrm{Re}\left((\Psi - \Psi_M, \Phi)_{L^2} \right),$$

where $D_{\Psi,j} = \Psi \star_{T_{Z_j}} \Psi - \mu_{t_j} \mathcal{F}(g_j)$ does not depend on Φ. Using the same notation as in Section 4.2, a discretization of $\langle \partial_\Psi E_\sigma[\Psi, t], \Phi \rangle$ is given by

$$\begin{aligned}
&\frac{4}{N} \sum_{j=1}^{N} \mathrm{Re}\left(\frac{L_X}{X} \frac{L_Y}{Y} \sum_{x=0}^{X-1} \sum_{y=0}^{Y-1} \frac{L_X}{X} \frac{L_Y}{Y} \big(\widehat{D}_{\Psi,j}\big)^*_{x,y} \frac{L_X}{X} \frac{L_Y}{Y} \big(\widehat{\Psi} \star_{\widehat{T}_{Z_j}} \widehat{\Phi}\big)_{x,y} \right) \\
&\quad + 2\sigma \mathrm{Re}\left(\frac{L_X}{X} \frac{L_Y}{Y} \sum_{x=0}^{X-1} \sum_{y=0}^{Y-1} \big(\widehat{\Psi} - \widehat{\Psi}_M\big)^*_{x,y} \widehat{\Phi}_{x,y} \right) \\
&= \frac{4 L_X^3 L_Y^3}{N X^3 Y^3} \sum_{j=1}^{N} \mathrm{Re}\left(\sum_{x=0}^{X-1} \sum_{y=0}^{Y-1} \big(\widehat{D}_{\Psi,j}\big)^*_{x,y} \big(\widehat{\Psi} \star_{\widehat{T}_{Z_j}} \widehat{\Phi}\big)_{x,y} \right) \\
&\quad + 2\sigma \frac{L_X L_Y}{XY} \mathrm{Re}\left(\sum_{x=0}^{X-1} \sum_{y=0}^{Y-1} \big(\widehat{\Psi} - \widehat{\Psi}_M\big)^*_{x,y} \widehat{\Phi}_{x,y} \right)
\end{aligned} \tag{4.12}$$

for $\widehat{\Psi}, \widehat{\Psi}_M, \widehat{\Phi} \in \mathbb{C}^{X \times Y}$, where $\widehat{D}_{\Psi,j} = \widehat{\Psi} \star_{\widehat{T}_{Z_j}} \widehat{\Psi} - \widehat{\mu}_{t_j} \,\mathrm{DFT}(\widehat{g}_j)$.

Let $a \in \{0, \ldots, X-1\}$, $b \in \{0, \ldots, Y-1\}$ and define $\widetilde{\Phi} \in \mathbb{C}^{X \times Y}$ by $\widetilde{\Phi}_{a,b} := 1$ and $\widetilde{\Phi}_{a',b'} := 0$ for all $(a', b') \neq (a, b)$. The partial derivative of $\widehat{E}_\sigma$ with respect to $\mathrm{Re}\big(\widehat{\Psi}_{a,b}\big)$ can be computed by substituting $\widehat{\Phi} = \widetilde{\Phi}$ in Equation (4.12), which yields

$$\begin{aligned}
\frac{\partial}{\partial \mathrm{Re}\big(\widehat{\Psi}_{a,b}\big)} \widehat{E}_\sigma\big[\widehat{\Psi}, t\big] = &\frac{4 L_X^3 L_Y^3}{N X^3 Y^3} \sum_{j=1}^{N} \mathrm{Re}\left(\sum_{x=0}^{X-1} \sum_{y=0}^{Y-1} \big(\widehat{D}_{\Psi,j}\big)^*_{x,y} \widehat{\Psi}^*_{a-x,b-y} \big(\widehat{T}_{Z_j}\big)_{(a,b),(a-x,b-y)} \right) \\
&+ 2\sigma \frac{L_X L_Y}{XY} \mathrm{Re}\left(\big(\widehat{\Psi} - \widehat{\Psi}_M\big)^*_{a,b} \right).
\end{aligned} \tag{4.13}$$

In this expression, the discretized exit wave and TCC are implicitly extended to elements of $\mathbb{C}^{\mathbb{Z} \times \mathbb{Z}}$ respectively $\mathbb{C}^{(\mathbb{Z} \times \mathbb{Z})^2}$ using the same continuation as the one that is used for the computation of the functional's value $\widehat{E}_\sigma\big[\widehat{\Psi}, t\big]$ (cf. Section 4.2). A similar expression can be derived if $\widetilde{\Phi}_{a,b} = \mathrm{i}$ instead of $\widetilde{\Phi}_{a,b} = 1$, which yields the partial derivative of $\widehat{E}_\sigma$ with respect to $\mathrm{Im}\big(\widehat{\Psi}_{a,b}\big)$. In general, the complexity of evaluating Equation (4.13) is $\mathcal{O}(NXY)$, assuming that $\widehat{D}_{\Psi,j}$ has been precomputed for all $j \in \{1, \ldots, N\}$. Since the partial derivatives have to be calculated for all $2XY$ possible directions, the overall complexity is $\mathcal{O}\big(2N(XY)^2\big)$.

However, if the TCC is $\star$-separable, then the first summand in Equation (4.13) can be rearranged to allow for a faster computation of the partial derivatives by utilizing the fast Fourier transform. Let $M \in \mathbb{N}$ and $\widehat{v}_{1,j}, \dots, \widehat{v}_{M,j} \in \mathbb{C}^{X \times Y}$ with

$$\big(\widehat{T}_{Z_j}\big)_{x',y'} = \sum_{k=1}^{M} \big(\widehat{v}_{k,j}\big)_{x'} \big(\widehat{v}_{k,j}\big)^*_{y'}$$

for all $x', y' \in \{0, \dots, X-1\} \times \{0, \dots, Y-1\}$. Then

$$\begin{aligned}
&\sum_{x=0}^{X-1}\sum_{y=0}^{Y-1} \big(\widehat{D}_{\Psi,j}\big)^*_{x,y} \widehat{\Psi}^*_{a-x,b-y} \big(\widehat{T}_{Z_j}\big)_{(a,b),(a-x,b-y)} \\
&= \sum_{k=1}^{M} \big(\widehat{v}_{k,j}\big)_{a,b} \sum_{x=0}^{X-1}\sum_{y=0}^{Y-1} \big(\widehat{D}_{\Psi,j}\big)^*_{x,y} \big(\widehat{\Psi}\widehat{v}_{k,j}\big)^*_{a-x,b-y} \\
&= \sum_{k=1}^{M} \big(\widehat{v}_{k,j}\big)_{a,b} \big(\widehat{D}_{\Psi,j} * (\widehat{\Psi}\widehat{v}_{k,j})\big)^*_{a,b},
\end{aligned}$$

where $*$ denotes the discrete convolution. Using the discrete convolution theorem, it follows that

$$\begin{aligned}
\frac{\partial \widehat{E}_\sigma[\widehat{\Psi}, t]}{\partial \mathrm{Re}\big(\widehat{\Psi}_{a,b}\big)} &= \frac{4L_X^3 L_Y^3}{NX^3Y^3} \sum_{j=1}^{N}\sum_{k=1}^{M} \mathrm{Re}\left(\big(\widehat{v}_{k,j}\big)_{a,b} \frac{1}{XY} \,\mathrm{IDFT}\left(\mathrm{DFT}\big(\widehat{D}_{\Psi,j}\big)\,\mathrm{DFT}\big(\widehat{\Psi}\widehat{v}_{k,j}\big)\right)^*_{a,b} \right) \\
&\quad + 2\sigma \frac{L_X L_Y}{XY} \mathrm{Re}\left(\big(\widehat{\Psi} - \widehat{\Psi}_M\big)^*_{a,b} \right).
\end{aligned}$$

This reduces the overall complexity to $\mathcal{O}\big(2NMXY \log(XY)\big)$.

Translation: The computation of the derivatives of the objective functional with respect to the translations is straightforward. Let $j \in \{1, \dots, N\}$, $k \in \{1, 2\}$ and $\tilde{t} \in (\mathbb{R}^2)^N$ with

$$\tilde{t}_{j',k'} = \begin{cases} 1, & \text{if } j' = j \text{ and } k' = k, \\ 0, & \text{otherwise} \end{cases}$$

for all $j' \in \{1, \dots, N\}$ and all $k' \in \{1, 2\}$. By Lemma 3.3.5, the Gâteaux differential of E_σ at a point $(\Psi, t) \in L^2(A, \mathbb{C}) \times (\mathbb{R}^2)^N$ in the direction $(0, \tilde{t})$ is

$$\langle \partial_t E_\sigma[\Psi, t], \tilde{t} \rangle = -\frac{2}{N} \mathrm{Re}\left(\big(\Psi \star_{T_{Z_j}} \Psi - \mu_{t_j}\mathcal{F}(g_j), \nu_k \mu_{t_j} \mathcal{F}(g_j)\big)_{L^2} \right),$$

where $\nu_k(x) = 2\pi \mathrm{i} x_k$ for all $x \in \mathbb{R}^2$. The product $\nu_k \mu_{t_j}$ is discretized just like μ_{t_j} as a piecewise constant function on a grid of size $X \times Y$ on the domain Q using point evaluations of the continuous function $\nu_k \mu_{t_j}$ at the top left corner of the respective pixels.

If the discretization of $\nu_k\mu_{t_j}$ is denoted by $\widehat{\nu_k\mu_{t_j}} \in \mathbb{C}^{X\times Y}$, then a discretization of the Gâteaux differential is

$$\begin{aligned}
&-\frac{2}{N}\mathrm{Re}\left(\frac{L_XL_Y}{XY}\sum_{x=0}^{X-1}\sum_{y=0}^{Y-1}\frac{L_XL_Y}{XY}\big(\widehat{D}_{\Psi,j}\big)^*_{x,y}\big(\widehat{\nu_k\mu_{t_j}}\big)_{x,y}\frac{L_XL_Y}{XY}\,\mathrm{DFT}\left(\widehat{g}_j\right)_{x,y}\right)\\
&= -\frac{2L_X^3L_Y^3}{NX^3Y^3}\mathrm{Re}\left(\sum_{x=0}^{X-1}\sum_{y=0}^{Y-1}\big(\widehat{D}_{\Psi,j}\big)^*_{x,y}\big(\widehat{\nu_k\mu_{t_j}}\big)_{x,y}\,\mathrm{DFT}\left(\widehat{g}_j\right)_{x,y}\right)\\
&= \frac{\partial}{\partial t_{j,k}}\widehat{E}_\sigma\big[\widehat{\Psi},t\big]
\end{aligned}\tag{4.14}$$

with $\widehat{D}_{\Psi,j} = \widehat{\Psi} \star_{\widehat{T}_{Z_j}} \widehat{\Psi} - \widehat{\mu}_{t_j}\,\mathrm{DFT}(\widehat{g}_j)$ as before.

It is evident that the overall complexity of calculating the partial derivatives with respect to the translations is limited by the computation of $D_{\Psi,j}$. The complexity of this computation is either $\mathcal{O}\big((XY)^2\big)$ or $\mathcal{O}\big(MXY\log(XY)\big)$ for an $M\in\mathbb{N}$, depending on whether the TCC is $\star$-separable or not. Since the derivatives need to be calculated for all $j\in\{1,\dots,N\}$ and all $k\in\{1,2\}$, this yields the overall complexity $\mathcal{O}\big(2N(XY)^2\big)$ respectively $\mathcal{O}\big(2NMXY\log(XY)\big)$. However, if the values of $D_{\Psi,j}$ are cached from the calculation of the exit wave derivatives and the Fourier transforms of the input images are precomputed, then the effective complexity of evaluating Equation (4.14) once is $\mathcal{O}(XY)$.

Focus: Let $j\in\{1,\dots,N\}$ and $\tilde{Z}\in\mathbb{R}^N$ with $\tilde{Z}_j = 1$ and $\tilde{Z}_k = 0$ for all $k\neq j$. By Lemma 3.3.7, the Gâteaux differential of E_σ^{ext} at a point $(\Psi,t,Z)\in L^2(A,\mathbb{C})\times(\mathbb{R}^2)^N\times\mathbb{R}^N$ in the direction $(0,0,\tilde{Z})$ is

$$\left\langle \partial_Z E_\sigma^{\mathrm{ext}}[\Psi,t,Z],\tilde{Z}\right\rangle = \frac{2}{N}\mathrm{Re}\left(\left(\Psi\star_{T'_{Z_j,1}}\Psi,\Psi\star_{T_{Z_j}}\Psi-G_j\right)_{L^2}\right),$$

where $T'_{Z_j,1}(v,w) = \frac{d}{d\varepsilon}T_{Z_j+\varepsilon}(v,w)\big|_{\varepsilon=0}$ for all $v,w\in\mathbb{R}^2$. If $\widehat{T}'_{Z_j}\in\mathbb{C}^{(X\times Y)^2}$ is the discretization of $T'_{Z_j,1}$ as a piecewise constant function on a grid of size $X\times Y$ using point evaluations, then the discretization of the Gâteaux differential satisfies

$$\begin{aligned}
&\frac{2}{N}\mathrm{Re}\left(\frac{L_XL_Y}{XY}\sum_{x=0}^{X-1}\sum_{y=0}^{Y-1}\frac{L_XL_Y}{XY}\big(\widehat{D}_{\Psi,j}\big)_{x,y}\frac{L_XL_Y}{XY}\big(\widehat{\Psi}\star_{\widehat{T}'_{Z_j}}\widehat{\Psi}\big)^*_{x,y}\right)\\
&= \frac{2L_X^3L_Y^3}{NX^3Y^3}\mathrm{Re}\left(\sum_{x=0}^{X-1}\sum_{y=0}^{Y-1}\big(\widehat{D}_{\Psi,j}\big)_{x,y}\big(\widehat{\Psi}\star_{\widehat{T}'_{Z_j}}\widehat{\Psi}\big)^*_{x,y}\right)\\
&= \frac{\partial}{\partial Z_j}\widehat{E}_\sigma^{\mathrm{ext}}\big[\widehat{\Psi},t,Z\big].
\end{aligned}$$

For the remainder of this section, the computation of $\widehat{\Psi}\star_{\widehat{T}'_{Z_j}}\widehat{\Psi}$ is discussed for the TCCs T_Z^{Ishizuka} and T_Z^{MAL}.

If $T_Z = T_Z^{\text{Ishizuka}}$ and only the focus and the spherical aberration are considered in the wave aberration function, then the derivative of the TCC is

$$\begin{aligned}\frac{d}{d\varepsilon} T_{Z+\varepsilon}(v,w)\bigg|_{\varepsilon=0} &= \Big(-\pi \mathrm{i}\lambda\left(\|v\|_2^2 - \|w\|_2^2\right) \\ &-2\left(\frac{\pi\alpha}{\lambda}\right)^2 \Big(Z\lambda^2\|v-w\|_2^2 + \lambda^4 C_s\left(\|v\|_2^4 + \|w\|_2^4 - (v\cdot w)\left(\|v\|_2^2 + \|w\|_2^2\right)\right)\Big)\Big) T_Z(v,w)\end{aligned}$$

for all $Z \in \mathbb{R}$ and all $v, w \in \mathbb{R}^2$ using Equation (2.11). Here, the derivative of the product of the phase transfer functions is calculated using

$$p_{Z+\varepsilon}(v)p_{Z+\varepsilon}^*(w) = p_Z(v)p_Z^*(w)\exp\left(-\pi \mathrm{i}\varepsilon\lambda\left(\|v\|^2 - \|w\|^2\right)\right) \qquad \forall\, v, w \in \mathbb{R}^2. \tag{4.15}$$

For the derivative of the spatial coherence envelope, the identity

$$\begin{aligned}&\frac{d}{d\varepsilon} E_s^{Z+\varepsilon}(v,w)\bigg|_{\varepsilon=0} \\ &= -2\left(\frac{\pi\alpha}{\lambda}\right)^2 \Big(Z\lambda^2\|v-w\|_2^2 + C_s\lambda^4\left(\|v\|_2^4 + \|w\|_2^4 - (v\cdot w)\left(\|v\|_2^2 + \|w\|_2^2\right)\right)\Big) E_s^Z(v,w),\end{aligned} \tag{4.16}$$

follows with a straightforward, but tedious computation.

Just like the TCC T_Z^{Ishizuka} itself, the derivative $T'_{Z,1}$ is not $\star$-separable and therefore does not allow for a fast computation of $\widehat{\Psi}\star_{\widehat{T}'_Z}\widehat{\Psi}$ using the FFT. Therefore, the complexity of calculating the partial derivatives with respect to all focus values is $\mathcal{O}\big(N(XY)^2\big)$ for Ishizuka's TCC.

If $T_Z = T_Z^{\text{MAL}}$, then T_Z can be written as $T_Z(v,w) = \sum_{k=-M}^{M} c_k \tilde{h}_{Z,k}(v)\tilde{h}_{Z,k}^*(w)$ for all $v, w \in \mathbb{R}^2$ and all $Z \in \mathbb{R}$, where

$$\tilde{h}_{Z,k}(v) = t_Z(v)E_s^Z(v,0)\exp\left(-\pi \mathrm{i}\lambda k\delta\|v\|_2^2\right)$$

and $c_k > 0$ are appropriate constants (cf. Section 2.2.4). Using Equations (4.15) and (4.16), it follows with the product rule that

$$\begin{aligned}\frac{d}{d\varepsilon}\tilde{h}_{Z+\varepsilon,k}(v)\tilde{h}_{Z+\varepsilon,k}^*(w)\bigg|_{\varepsilon=0} &= -\pi \mathrm{i}\lambda\left(\|v\|_2^2 - \|w\|_2^2\right)\tilde{h}_{Z,k}(v)\tilde{h}_{Z,k}^*(w) \\ &\quad - 2\left(\frac{\pi\alpha}{\lambda}\right)^2\left(Z\lambda^2\|v\|_2^2 + C_s\lambda^4\|v\|_2^4\right)\tilde{h}_{Z,k}(v)\tilde{h}_{Z,k}^*(w) \\ &\quad - 2\left(\frac{\pi\alpha}{\lambda}\right)^2\left(Z\lambda^2\|w\|_2^2 + C_s\lambda^4\|w\|_2^4\right)\tilde{h}_{Z,k}(v)\tilde{h}_{Z,k}^*(w) \\ &= \left(-\pi \mathrm{i}\lambda\|v\|_2^2 - 2\left(\frac{\pi\alpha}{\lambda}\right)^2\left(Z\lambda^2\|v\|_2^2 + C_s\lambda^4\|v\|_2^4\right)\right)\tilde{h}_{Z,k}(v)\tilde{h}_{Z,k}^*(w) \\ &\quad + \Bigg(\underbrace{-\pi \mathrm{i}\lambda\|w\|_2^2 - 2\left(\frac{\pi\alpha}{\lambda}\right)^2\left(Z\lambda^2\|w\|_2^2 + C_s\lambda^4\|w\|_2^4\right)}_{=:\alpha_Z(w)}\Bigg)^*\tilde{h}_{Z,k}(v)\tilde{h}_{Z,k}^*(w).\end{aligned}$$

This implies

$$T'_{Z,1}(v,w) = \sum_{k=-M}^{M} c_k \left(\alpha_Z(v)\tilde{h}_{Z,k}(v)\tilde{h}^*_{Z,k}(w) + \tilde{h}_{Z,k}(v)\alpha^*_Z(w)\tilde{h}^*_{Z,k}(w) \right)$$

for all $v, w \in \mathbb{R}^2$. Unlike T_Z, it is no longer clear if $T'_{Z,1}$ is $\star$-separable. However, the structure of $T'_{Z,1}$ still allows for a fast computation of $\Psi \star_{T'_{Z,1}} \Psi$ using the fast Fourier transform, which is shown next.

If $\widehat{\alpha}_Z, \widehat{h}_{Z,k}, \widehat{\alpha_Z h_{Z,k}} \in \mathbb{C}^{X \times Y}$ denote the discretizations of $\alpha_Z, \tilde{h}_{Z,k}, \alpha_Z \tilde{h}_{Z,k}$ as piecewise constant functions on a grid of size $X \times Y$ using point evaluations, then

$$\big(\widehat{T}'_{Z_j}\big)_{a,b} = \sum_{k=-M}^{M} c_k \left(\big(\widehat{\alpha_Z h_{Z,k}}\big)_a \big(\widehat{h}_{Z,k}\big)^*_b + \big(\widehat{h}_{Z,k}\big)_a \big(\widehat{\alpha_Z h_{Z,k}}\big)^*_b \right)$$

for all $a, b \in \{0, \ldots, X-1\} \times \{0, \ldots, Y-1\}$. Consequently, $\widehat{\Psi} \star_{\widehat{T}'_{Z_j}} \widehat{\Psi}$ can be computed as

$$\begin{aligned}
\widehat{\Psi} \star_{\widehat{T}'_{Z_j}} \widehat{\Psi} &= \sum_{k=-M}^{M} c_k \left(\big(\widehat{\Psi}\widehat{\alpha_Z h_{Z,k}}\big) \star \big(\widehat{\Psi}\widehat{h}_{Z,k}\big) + \big(\widehat{\Psi}\widehat{h}_{Z,k}\big) \star \big(\widehat{\Psi}\widehat{\alpha_Z h_{Z,k}}\big) \right) \\
&= \sum_{k=-M}^{M} \frac{c_k}{XY} \operatorname{IDFT}\left(\operatorname{DFT}\left(\big(\widehat{\Psi}\widehat{\alpha_Z h_{Z,k}}\big) \star \big(\widehat{\Psi}\widehat{h}_{Z,k}\big) + \big(\widehat{\Psi}\widehat{h}_{Z,k}\big) \star \big(\widehat{\Psi}\widehat{\alpha_Z h_{Z,k}}\big) \right) \right) \\
&= \sum_{k=-M}^{M} \frac{2c_k}{XY} \operatorname{IDFT}\left(\operatorname{Re}\left(\operatorname{DFT}\big(\widehat{\Psi}\widehat{\alpha_Z h_{Z,k}}\big)^* \operatorname{DFT}\big(\widehat{\Psi}\widehat{h}_{Z,k}\big) \right) \right). \qquad (4.17)
\end{aligned}$$

Evaluating Equation (4.17) requires $3(2M+1)$ discrete Fourier transforms. If the fast Fourier transform is used, the overall complexity of calculating the partial derivatives with respect to all focus values is thus $\mathcal{O}\big(3(2M+1)NXY\log(XY)\big)$.

4.4.2 Scaling

The objective functional $\widehat{E}^{\text{ext}}_\sigma$ joins three very different optimization parameters: the exit wave, the image shifts and the focus values. Let $\big(\widehat{\Psi}, \widehat{t}, \widehat{Z}\big) \in \mathbb{C}^{X \times Y} \times (\mathbb{R}^2)^N \times \mathbb{R}^N$ be the current position during a minimization of $\widehat{E}^{\text{ext}}_\sigma$ and $\big(\widehat{\Psi}_s, \widehat{t}_s, \widehat{Z}_s\big) \in \mathbb{C}^{X \times Y} \times (\mathbb{R}^2)^N \times \mathbb{R}^N$ be the current search direction. Because of the different nature of the optimization parameters, the optimal step sizes for the individual parameters $\widehat{\Psi}$, $\widehat{t}$ and $\widehat{Z}$ may differ by orders of magnitude. More explicitly, denote by $\tau_1, \tau_2, \tau_3 > 0$ the optimal step sizes such that

$$\begin{aligned}
\widehat{E}^{\text{ext}}_\sigma\big[\widehat{\Psi} + \tau_1\widehat{\Psi}_s, \widehat{t}, \widehat{Z}\big] &= \min_{\tau>0} \widehat{E}^{\text{ext}}_\sigma\big[\widehat{\Psi} + \tau\widehat{\Psi}_s, \widehat{t}, \widehat{Z}\big], \\
\widehat{E}^{\text{ext}}_\sigma\big[\widehat{\Psi}, \widehat{t} + \tau_2\widehat{t}_s, \widehat{Z}\big] &= \min_{\tau>0} \widehat{E}^{\text{ext}}_\sigma\big[\widehat{\Psi}, \widehat{t} + \tau\widehat{t}_s, \widehat{Z}\big] \qquad \text{and} \\
\widehat{E}^{\text{ext}}_\sigma\big[\widehat{\Psi}, \widehat{t}, \widehat{Z} + \tau_3\widehat{Z}_s\big] &= \min_{\tau>0} \widehat{E}^{\text{ext}}_\sigma\big[\widehat{\Psi}, \widehat{t}, \widehat{Z} + \tau\widehat{Z}_s\big].
\end{aligned}$$

If τ_1, τ_2 and τ_3 are not on the same order of magnitude, then the convergence speed of the iterative minimization can decrease severely. It is also important to notice that the algorithm may converge to different local minima depending on the choice of τ_1, τ_2 and τ_3.

Such a functional is said to be poorly scaled, which is a problem that can be resolved by considering the scaled functional

$$\widehat{E}_\sigma^{\mathrm{ext,scaled}}\big[\widehat{\Psi}, \widehat{t}, \widehat{Z}\big] := \widehat{E}_\sigma^{\mathrm{ext}}\big[c_1\widehat{\Psi}, c_2\widehat{t}, c_3\widehat{Z}\big]$$

instead of the original functional for appropriately chosen $c_1, c_2, c_3 > 0$. The scaling does not affect the value of the objective functional, provided that the argument is scaled with the inverse coefficients $c_1^{-1}, c_2^{-1}, c_3^{-1}$. Thus, the computed local minimizer of $\widehat{E}_\sigma^{\mathrm{ext,scaled}}$ has to be scaled with c_1, c_2, c_3 in order to get the desired local minimum of $\widehat{E}_\sigma^{\mathrm{ext}}$. However, the partial derivatives of the objective functional are scaled by the coefficients c_1, c_2, c_3 according to the chain rule, which makes it possible to adjust τ_1, τ_2 and τ_3 to be roughly on the same order of magnitude. Replacing the objective functional $\widehat{E}_\sigma^{\mathrm{ext}}$ with $\widehat{E}_\sigma^{\mathrm{ext,scaled}}$ can be considered as a simple form of preconditioning.

The scaling of the individual elements of $\widehat{\Psi} \in \mathbb{C}^{X\times Y}$, $\widehat{t} \in (\mathbb{R}^2)^N$ and $\widehat{Z} \in \mathbb{R}^N$ is less critical for the convergence speed of the minimization than the above scaling of the components as a whole. While it is expected that a non-uniform scaling of the elements of $\widehat{t}$ and $\widehat{Z}$ is not useful, scaling the individual exit wave pixels non-uniformly can indeed be beneficial for the convergence speed. This is indicated by Figure 4.3, which shows that the high frequencies of the exit wave are reconstructed very slowly compared to the low frequencies. Denote by $P \in \mathbb{C}^{X\times Y}$ a matrix that realizes a scaling of the individual exit wave pixels if the discrete functional $\widehat{E}_\sigma^{\mathrm{ext,scaled}}$ is replaced with

$$\widehat{E}_\sigma^{\mathrm{ext}}\big[c_1P\widehat{\Psi}, c_2\widehat{t}, c_3\widehat{Z}\big].$$

Here, the product $P\widehat{\Psi}$ is to be understood as a componentwise product of elements of $(\mathbb{R}^2)^{X\times Y}$, i.e. the real and imaginary parts are considered separately. By the chain rule, the partial derivatives with respect to $\mathrm{Re}\big(\widehat{\Psi}_{x,y}\big)$ are then given by

$$\frac{\partial}{\partial \mathrm{Re}\big(\widehat{\Psi}_{x,y}\big)}\left(\widehat{E}_\sigma^{\mathrm{ext}}\big[c_1P\widehat{\Psi}, c_2\widehat{t}, c_3\widehat{Z}\big]\right) = c_1\mathrm{Re}(P_{x,y})\left(\frac{\partial \widehat{E}_\sigma^{\mathrm{ext}}}{\partial \mathrm{Re}\big(\widehat{\Psi}_{x,y}\big)}\right)\big[c_1P\widehat{\Psi}, c_2\widehat{t}, c_3\widehat{Z}\big]$$

for all $(x, y) \in \{0, \ldots, X-1\} \times \{0, \ldots, Y-1\}$ and similarly for $\mathrm{Im}\big(\widehat{\Psi}_{x,y}\big)$. The matrix P is called scale mask in the following.

The circular shape of the power spectra in Figure 4.3 motivates the construction of an isotropic scale mask $P \in (\mathbb{R}^2)^{X\times Y} \cong \mathbb{C}^{X\times Y}$. Here, isotropic refers to a scale mask that is rotationally symmetric around its center. Recall that the domain of the discrete exit wave $\widehat{\Psi} \in \mathbb{C}^{X\times Y}$ is $Q = \big[-\frac{X}{2L_X}, \frac{X}{2L_X}\big] \times \big[-\frac{Y}{2L_Y}, \frac{Y}{2L_Y}\big]$, which covers frequencies with magnitudes in the range $[0, f_{\max}]$ with $f_{\max} := \sqrt{(X/(2L_X))^2 + (Y/(2L_Y))^2}$. In order to construct an isotropic scale mask, it is sufficient to prescribe a value of the scale mask

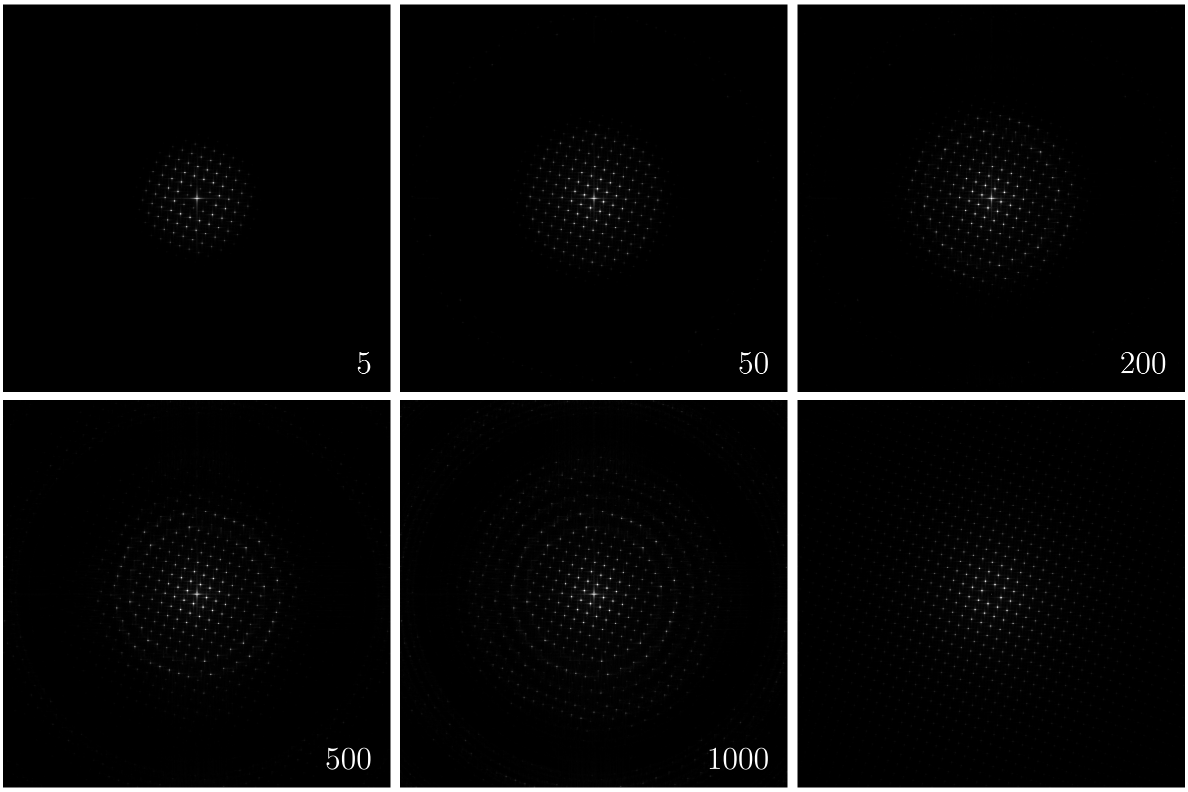

Figure 4.3: The power spectra of the reconstructed exit wave after 5, 50, 200, 500 and 1000 iterations of the conjugate gradient method for a hypothetical specimen consisting of a standard lattice of uranium atoms. The power spectrum of the correct exit wave, which was used for the simulation of the focus series, is displayed in the bottom right image. No scaling of the individual exit wave pixels was performed during the reconstruction. It can clearly be seen that the high frequencies are only reconstructed after a high number of iterations, whereas the low frequencies are reconstructed much faster. The complete experiment is presented in Section 5.1.

for each magnitude in $[0, f_{\max}]$. Since a pixel $(x, y) \in \{0, \ldots, X-1\} \times \{0, \ldots, Y-1\}$ of the discrete exit wave corresponds to the frequencies $S_{(x,y)} = [p_{(x,y)}, p_{(x+1,y+1)}] \subseteq Q$ with

$$p_{(x,y)} = \left(\frac{x - \frac{X}{2}}{L_X}, \frac{y - \frac{Y}{2}}{L_Y} \right) \in \mathbb{R}^2,$$

the pixels on the diagonal $D := \left\{ \left(\mathrm{rd}\left(d\frac{X}{2}\right), \mathrm{rd}\left(d\frac{Y}{2}\right) \right) \,\middle|\, d \in [0,1] \right\} \subseteq \mathbb{Z}^2$ provide a suitable sampling of the magnitudes $[0, f_{\max}]$, where $\mathrm{rd} : \mathbb{R} \to \mathbb{Z}$ rounds a real number to the nearest integer. If $X = Y$, then D is equal to $\{(d, d) \mid d = 0, \ldots, X/2\}$.

The values of $P_{x,y}$ for $(x, y) \in D$ are determined as follows. Define $\widehat{\Phi}_{x,y} \in \mathbb{C}^{X \times Y}$ for

$(x, y) \in D$ by

$$\left(\widehat{\Phi}_{x,y}\right)_{x',y'} := \begin{cases} (\widehat{\Psi}_s)_{x,y}, & \text{if } (x', y') = (x, y), \\ 0, & \text{otherwise} \end{cases}$$

for all $(x', y') \in \{0, \dots, X-1\} \times \{0, \dots, Y-1\}$. For each $(x, y) \in D$, exponents $e^{\text{re}}_{x,y}, e^{\text{im}}_{x,y} \in \mathbb{Z}$ are computed such that $\tau^{\text{re}}_{x,y} := 2^{e^{\text{re}}_{x,y}}$ and $\tau^{\text{im}}_{x,y} := 2^{e^{\text{im}}_{x,y}}$ are the optimal power-of-two step sizes in the sense that

$$\begin{aligned} &\widehat{E}_\sigma\big[\widehat{\Psi} + \tau^{\text{re}}_{x,y}\text{Re}\big(\widehat{\Phi}_{(x,y)}\big), t\big] \quad \text{and} \\ &\widehat{E}_\sigma\big[\widehat{\Psi} + \mathrm{i}\tau^{\text{im}}_{x,y}\text{Im}\big(\widehat{\Phi}_{(x,y)}\big), t\big] \end{aligned}$$

are a minimal among all exponents in a given range, i.e. $e^{\text{re}}_{x,y}, e^{\text{im}}_{x,y} \in [-10, \dots, 10]$ for example. The exponents are placed into vectors $(e^{\text{re}}_{0,0}, \dots, e^{\text{re}}_{X/2,Y/2}) \in \mathbb{Z}^{|D|}$ respectively $(e^{\text{im}}_{0,0}, \dots, e^{\text{im}}_{X/2,Y/2}) \in \mathbb{Z}^{|D|}$ and smoothed with the mean value filter $(\frac{1}{5}, \dots, \frac{1}{5}) \in \mathbb{Z}^5$ in order to remove outliers. This yields the smoothed exponents $(\tilde{e}^{\text{re}}_{0,0}, \dots, \tilde{e}^{\text{re}}_{X/2,Y/2}) \in \mathbb{Z}^{|D|}$ and $(\tilde{e}^{\text{im}}_{0,0}, \dots, \tilde{e}^{\text{im}}_{X/2,Y/2}) \in \mathbb{Z}^{|D|}$. Finally, the entries $P_{x,y}$ of the scale mask for $(x, y) \in D$ are defined as

$$\begin{aligned} \text{Re}(P_{x,y}) &:= 2^{\tilde{e}^{\text{re}}_{x,y}/6} \quad \text{and} \\ \text{Im}(P_{x,y}) &:= 2^{\tilde{e}^{\text{im}}_{x,y}/6}. \end{aligned}$$

In this way, the partial derivatives of the objective functional are scaled accordingly by $2^{\tilde{e}^{\text{re}}_{x,y}/6}$ or $2^{\tilde{e}^{\text{im}}_{x,y}/6}$, which helps in leveling out the step sizes for the individual pixels. An additional scaling of the exponents is necessary to prevent an overly strong amplification of those frequencies for which a large exponent $\tilde{e}^{\text{re}}_{x,y}$ or $\tilde{e}^{\text{im}}_{x,y}$ has been computed. The factor $\frac{1}{6}$ has been found to perform well in practice as it reduces the overall range of the scaling factors sufficiently and at the same time it is small enough so that it does not reduce all scaling factors to unity.

In Section 5.1, it is shown that such a scale mask for the exit wave can indeed decrease the number of iterations that are necessary for a reconstruction of the high frequencies. However, the computation of the scale mask is very costly and therefore the scale mask in its current form is of little practical use as a preconditioner. Instead, it can be understood as a proof of concept and may be used as a starting point for the derivation of other preconditioners.

4.4.3 Contrast dislocation and the wrap around error

The information corresponding to a single atom in a TEM image may be spread within a relatively large area as compared to the actual size of the atom. This effect is caused by the aberrations of the objective lens and termed contrast dislocation. The left image in Figure 4.4 depicts contrast dislocation caused by the focus and spherical aberration for a single gold atom. Although contrast dislocation makes the direct interpretation of TEM images difficult, it is mostly irrelevant to exit wave reconstruction. This is indeed one of the main reasons and advantages of using exit wave reconstruction in the first place: if the aberrations are sufficiently well known, then the simulated TEM images are subject to the same kind of contrast dislocation as the experimental images and a reconstruction of the exit wave makes the direct analysis of TEM images superfluous.

However, there is one important exception where contrast dislocation is problematic for a successful exit wave reconstruction. Because of the discretization, contrast dislocation can affect the results near the images' edges if the specimen is not perfectly periodic with respect to the discrete image domains. On the one hand, the experimental images only contain incomplete information about atoms near the images' edges, since a part of the information is located outside of the image domain due to contrast dislocation. On the other hand, contrast dislocation in conjunction with the periodicity of the discrete Fourier transform causes the wrap around error in simulated TEM images as follows.

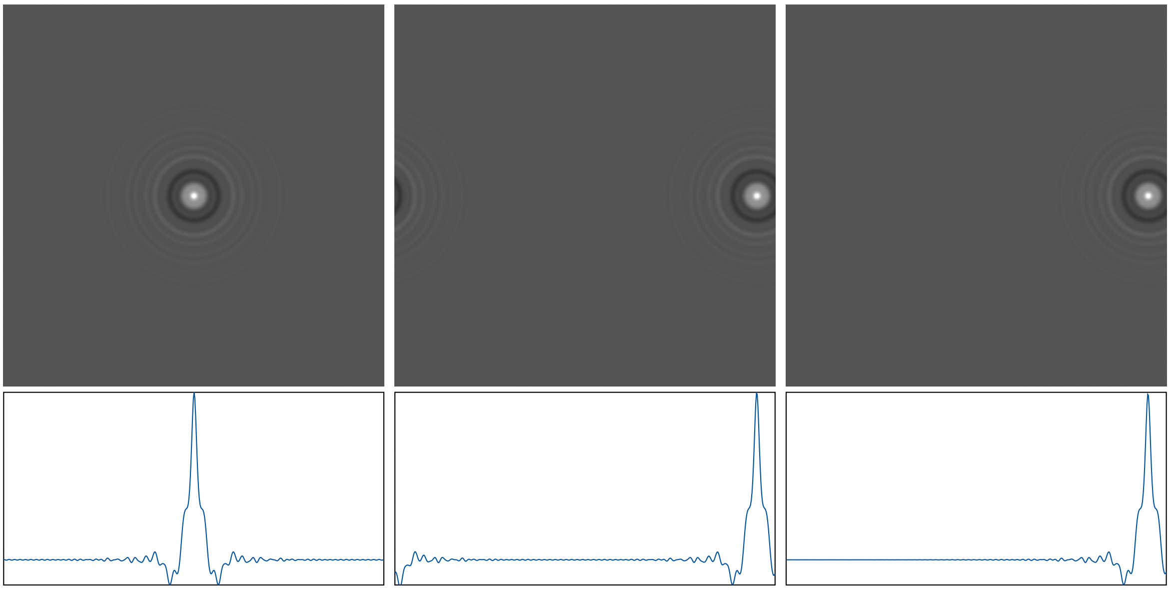

Figure 4.4: Simulated TEM images of a single gold atom sampled at 512×512 pixels and depicting an area of 1.6×1.6 nm^2, showing contrast dislocation in the form of the additional "rings" surrounding the atom. The left and middle images were simulated without an additional buffer zone around the images. The right image was initially simulated on a grid of size 1024×1024 (with a buffer zone with of 256 pixel) and then cropped to the central section of size 512×512. The plots in the bottom row display the pixel values of the images at the cross-section $(x, 256)$ with $x = 0, \ldots, 511$, which range from 0.846 (black) to 1.644 (white). The microscope parameters for all images are: focus $Z = 4.5$ nm, spherical aberration $C_s = -700$ nm, accelerating voltage $U = 200$ kV (electron wavelength $\lambda \approx 0.00250793$ nm), objective aperture semiangle $\alpha_{\max} = 125$ mrad, coherence parameters $\alpha = 1$ mrad and $\Delta = 2.8$ nm. The images were simulated using T_Z^{MAL} with $M = 4$ and $\delta = 1.25$ nm.

The discrete Fourier transform of a vector $g \in \mathbb{C}^N$ for $N \in \mathbb{N}$ is necessarily N-periodic, since

$$\begin{aligned}\mathrm{DFT}_N(g)_{k+N} &= \sum_{n=0}^{N-1} g_n \exp\left(-2\pi\mathrm{i}\frac{n(k+N)}{N}\right)\\ &= \sum_{n=0}^{N-1} g_n \exp\left(-2\pi\mathrm{i}\frac{nk}{N}\right)\underbrace{\exp(-2\pi\mathrm{i}n)}_{=1}\\ &= \mathrm{DFT}_N(g)_k\end{aligned}$$

holds for all $k \in \mathbb{Z}$. The same applies to the inverse discrete Fourier transform. Because of this periodicity, opposite edges of the TEM images are effectively adjacent to each other in numerical computations. Therefore, the simulated atoms wrap around the images' edges as illustrated in the middle image of Figure 4.4 (cf. [41, ch. 4.4]).

The experimental TEM images do not contain the wrap around error and consequently this error needs to be reduced for a meaningful exit wave reconstruction. If the specimen is almost perfectly periodic with respect to the discrete images' domains, then the wrap around error is small and can be ignored. However, in the general case a buffer zone should be added around the real space images to account for the wrap around error [54, 41, 66]. Otherwise, a numerical minimization of the objective functional would attempt to adapt the exit wave in order to compensate for the wrap around error, which then yields a better match of the simulated and experimental images with a worse estimate of the true exit wave.

If a buffer zone of width $W \in \mathbb{N}_0$ is used, then the discrete exit wave $\widehat{\Psi}$ is an element of $\mathbb{C}^{(X+2W)\times(Y+2W)}$ and the discrete input images $\widehat{g}_j$ need to be extended to $\mathbb{R}^{(X+2W)\times(Y+2W)}$ using e.g. constant continuation with the mean value or periodic continuation. The data term of the discrete objective functional from Equation (4.8) is then transformed to real space (in order to allow for a real space buffer zone) and adapted to

$$\frac{1}{N}\sum_{j=1}^{N}\frac{1}{XY}\sum_{x=0}^{X-1}\sum_{y=0}^{Y-1}\left|\mathrm{IDFT}\left(\widehat{\Psi}\star_{\widehat{T}_{Z_j}}\widehat{\Psi}-\widehat{\mu}_{t_j}\,\mathrm{DFT}\left(\widehat{g}_j\right)\right)_{x+W,y+W}\right|^2, \tag{4.18}$$

where a scaling factor of $L_X^3 L_Y^3/(X^2Y^2)$ has been omitted and $\widehat{T}_{Z_j}$ respectively $\widehat{\mu}_{t_j}$ are similarly extended to elements of $\mathbb{C}^{((X+2W)\times(Y+2W))^2}$ respectively $\mathbb{C}^{(X+2W)\times(Y+2W)}$. Note that this data term ignores the values of the simulated and input images in the buffer zone.

The importance of using a buffer zone for non-periodic specimens is confirmed with a simple experiment, where the exit wave is reconstructed from a focus series consisting of the 4 images shown in Figure 4.5. Using these images as the input data, two reconstructions have been performed with the nonlinear conjugate gradient method. The first reconstruction does not use a buffer zone and the second reconstruction uses a buffer zone with a width of 256 pixels. Therefore, the first reconstruction yields an exit wave of size 512 × 512 pixels, whereas the second reconstruction computes an exit wave of size 1024 × 1024 pixels. The TCC T_Z^{MAL} with $M = 4$ and $\delta = 1.25$ nm was used for the forward model in both cases.

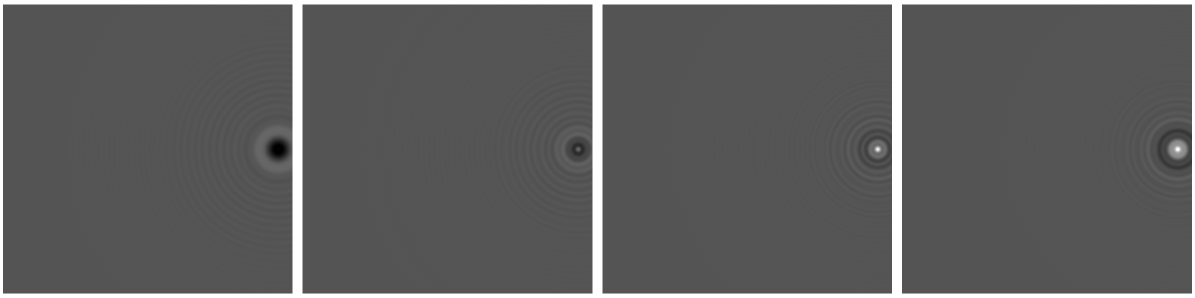

Figure 4.5: Simulated focus series used as input data to the exit wave reconstruction. The images are sampled at 512×512 pixels and depict an area of 1.6×1.6 nm^2. The focus values are -3 nm, -0.5 nm, 2 nm and 4.5 nm (from left to right) and the other microscope parameters are identical to the ones in Figure 4.4. All images have been simulated with a buffer zone width of 256 pixels using T_Z^{MAL} with $M = 4$ and $\delta = 1.25$ nm. Note that the rightmost image is identical to the rightmost image in Figure 4.4.

As the initial guess for the exit wave a constant exit wave in real space corresponding to the mean image intensity was used. Neither the translations nor the focus values were optimized and instead kept fixed at the correct values throughout the minimization. Note that there is no specimen movement within the focus series in this example. The coefficient σ of the regularizer was set to zero.

Figure 4.6 shows the energy of the modified objective functional given in Equation (4.18) for the first 100 iterations for both reconstructions. It can clearly be seen that using a buffer zone results in a significantly lower residual or, equivalently, a significantly better match of the input images and the simulated images. Note that in both cases a perfect reconstruction is not possible. This is because of the first problem caused by contrast dislocation in exit wave reconstruction mentioned above. As the gold atom is located near the image's edge, the input images are missing the information about the atom that is located on the right side of the right image edge.

Finally, the amplitude and phase of the reconstructed exit waves in Figures 4.7 and 4.8 show that using a buffer zone not only results in a lower energy, but also a better reconstruction of the exit wave.

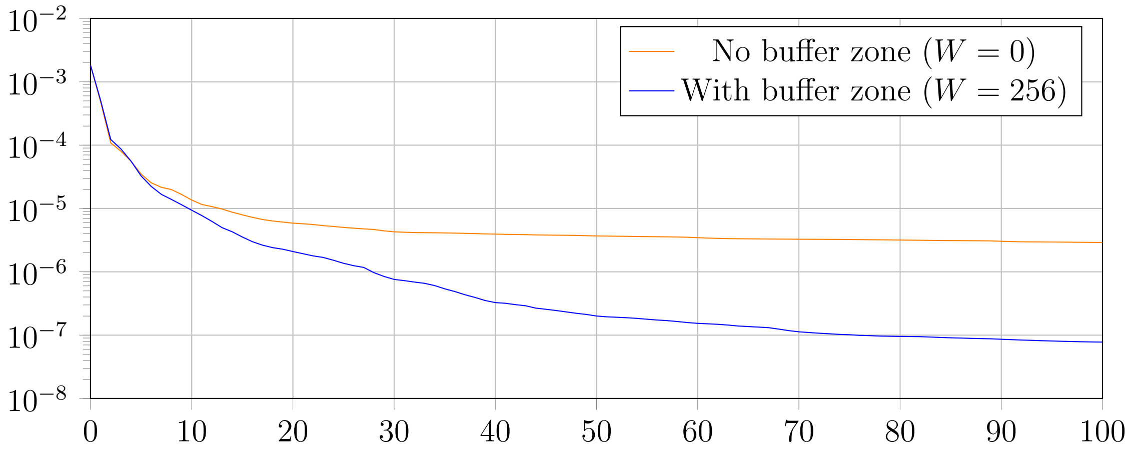

Figure 4.6: The value of the modified objective functional in Equation (4.18) for the first 100 iterations of both reconstructions.

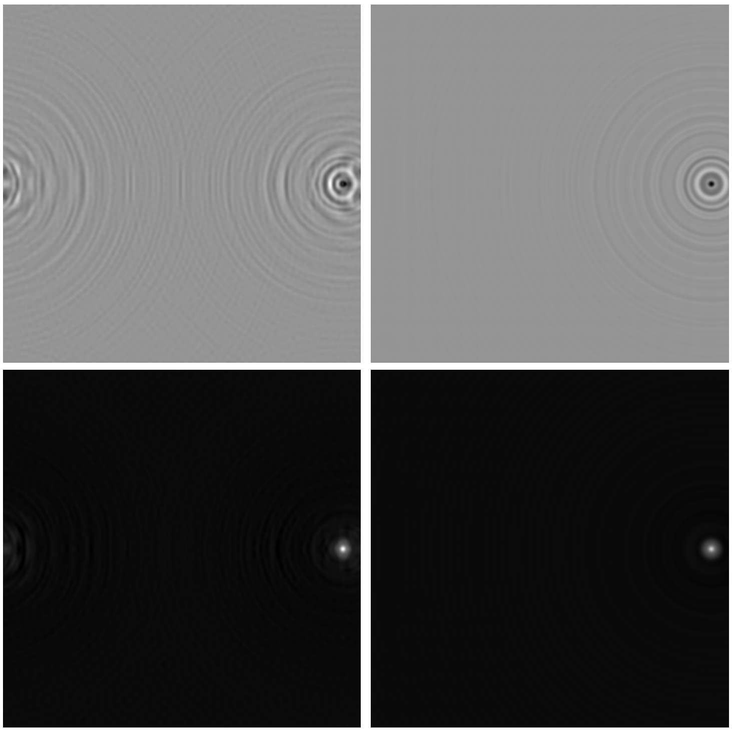

Figure 4.7: The amplitudes (top) and phases (bottom) of the reconstructed exit waves after 100 iterations. The images on the left show the result of the reconstruction without a buffer zone ($W = 0$). The images on the right show the result of the reconstruction using a buffer zone (of width $W = 256$), cropped to the central section of size 512×512.

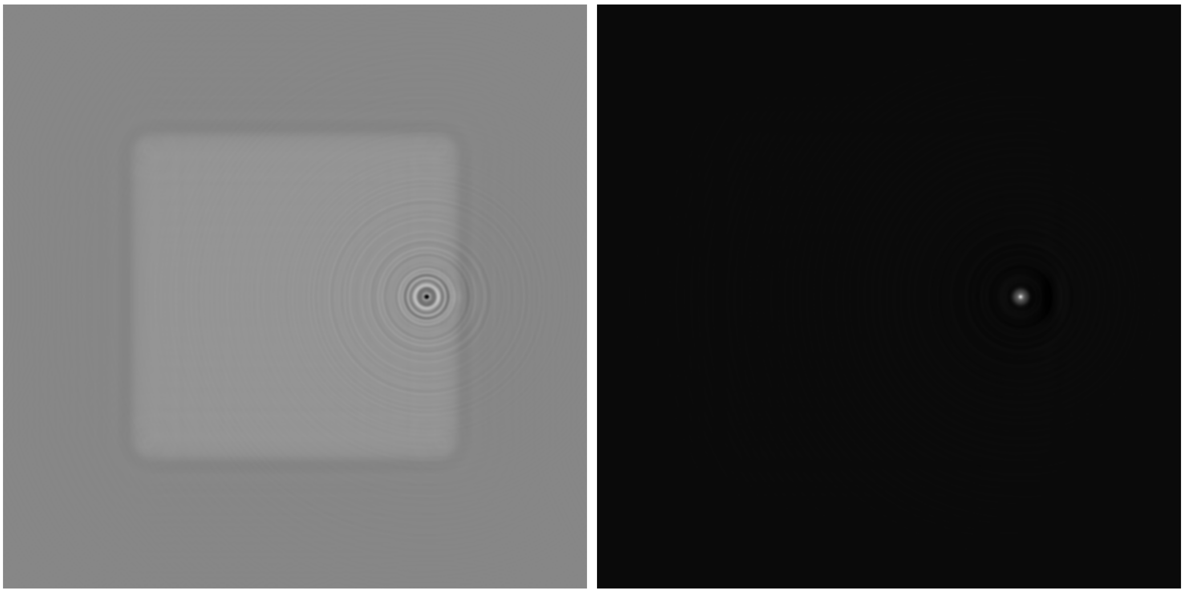

Figure 4.8: The full amplitude (left) and phase (right) of the reconstructed exit wave of size 1024 × 1024 pixels after 100 iterations, including the buffer zone of width $W = 256$. The central subsection of size 512 × 512 pixels is identical to the images in the right column of Figure 4.7.

4.4.4 Dynamic integration domains

Unlike the example in the previous section, the images in a real focus series are in general not aligned very well to each other and the specimen drift can be quite significant. Here, we follow the approach from Remark 3.3.2 and choose the first image $\widehat{g}_1$ as the reference by keeping $t_1 = 0$ fixed throughout the minimization. In this case, the reconstructed exit wave is aligned to $\widehat{g}_1$. This means that for all the other images $\widehat{g}_2, \ldots, \widehat{g}_N$ in the series, only these subsections of $\widehat{g}_j \circ \widehat{\phi_{t_j}}$ contribute to the reconstructed exit wave which are contained in the domain of $\widehat{g}_1$. Worse yet, if the shifted images are calculated with $\widehat{\mu}_{t_j} \operatorname{DFT}\left(\widehat{g}_j\right)$ in Fourier space as it is done in the present work, then the image shift is implicitly performed with periodic continuation, which causes artifacts in the reconstructed exit wave.

However, if the exit wave reconstruction is performed with a buffer zone that is sufficiently large to encompass the shifted images for all translations t_j occurring during the reconstruction, then these problems can be circumvented completely by adjusting the integration domains in the objective functional to the domains of the shifted images. On top of that, using this approach, the area of the reconstructed exit wave is larger than the domain of any of the individual images in the focus series. This is achieved by adjusting the integration domains in Equation (4.18), which yields the modified version of the data term

$$\frac{1}{N} \sum_{j=1}^{N} \frac{1}{X_j Y_j} \sum_{x=0}^{X_j-1} \sum_{y=0}^{Y_j-1} \left| \operatorname{IDFT}\left(\widehat{\Psi} \star_{\widehat{T}_{Z_j}} \widehat{\Psi} - \widehat{\mu}_{t_j} \operatorname{DFT}\left(\widehat{g}_j\right)\right)_{x+\tilde{x}_j+W, y+\tilde{y}_j+W} \right|^2 \tag{4.19}$$

with $(X_j, Y_j) \in \mathbb{N}^2$ and $\tilde{x}_j, \tilde{y}_j \in \mathbb{Z}$ such that $(\tilde{x}_j + W, \tilde{y}_j + W) \in \mathbb{N}_0^2$ for all $j \in \{1, \dots, N\}$, where $W \in \mathbb{N}_0$ is the width of the buffer zone. The integration domain of the j-th real space residual $R_j := \operatorname{IDFT}\left(\widehat{\Psi} \star_{\widehat{T}_{Z_j}} \widehat{\Psi} - \widehat{\mu}_{t_j} \operatorname{DFT}\left(\widehat{g}_j\right)\right)$ is thus the rectangle

$$I_j := \{(x + \tilde{x}_j + W, y + \tilde{y}_j + W) \mid x = 0, \dots, X_j - 1 \text{ and } y = 0, \dots, Y_j - 1\}.$$

In other words, $(\tilde{x}_j + W, \tilde{y}_j + W) \in \mathbb{N}_0^2$ is the top left position in image coordinates on the screen and $(X_j, Y_j) \in \mathbb{N}^2$ is the size of the integration domain. If the integration domains I_j are carefully chosen such that

$$(p_j)_1 - (t_j)_1 \frac{X}{L_X} \in [W, W + X] \qquad \wedge \qquad (p_j)_2 - (t_j)_2 \frac{Y}{L_Y} \in [W, W + Y] \tag{4.20}$$

for all $p_j \in I_j$ and all translations $t \in (\mathbb{R}^2)^N$ considered during the reconstruction, then a continuation of the discrete input images beyond their domain is no longer necessary as only values within the domain of the real space images are accessed in Equation (4.19). The translations that are considered during the entire reconstruction are necessarily unknown a priori. However, initially a large range of possible translations may be considered based on an upper bound for the translations, which is then successively refined during the minimization as outlined in the following.

The condition in Equation (4.20) requires that X_j and Y_j are initially set to a value much smaller than X and Y because of the typically very large errors in the translations t_j at the beginning of the minimization. However, the translations converge quickly and the magnitudes of the changes in t from one iteration to the next decrease as a consequence. This in turn can be used to increase the sizes (X_j, Y_j) and adjust the positions $(\tilde{x}_j, \tilde{y}_j)$ of the integration domains as described in the following paragraph.

Denote by $t^{(k)} \in (\mathbb{R}^2)^N$ the estimate for the translations in the k-th iteration and fix an $n \in \mathbb{N}$. Let $j \in \{2, \dots, N\}$, $l \in \{1, 2\}$ and define the maximum absolute value of the changes of the translation $(t_j)_l$ during the previous n iterations as

$$\left(\bar{t}_j^{(k)}\right)_l := \max\left\{\left|\left(t_j^{(k-k')}\right)_l - \left(t_j^{(k-k'-1)}\right)_l\right| \;:\; k' = 0, \dots, n-1\right\}$$

for $k \geq n$. The position and size of the integration domain I_j can be adjusted based on the assumption that

$$\left(t_j^{(k')}\right)_l \in \left[\left(t_j^{(k)}\right)_l - \delta\left(\bar{t}_j^{(k)}\right)_l, \left(t_j^{(k)}\right)_l + \delta\left(\bar{t}_j^{(k)}\right)_l\right] \qquad \forall\, k' \geq k$$

for an appropriate $\delta \geq 1$, which states that the final value for the translation $(t_j)_l$ has been computed within an accuracy of $2\delta\left(\bar{t}_j^{(k)}\right)_l$. In practice, setting $\delta = 20$ turned out to work well. If the integration domains are adjusted every n iterations, the sizes of the integration domains (X_j, Y_j) eventually converge to $(X - 1, Y - 1)$ and the positions $(\tilde{x}_j, \tilde{y}_j)$ converge to

$$\tilde{x}_j = \lim_{k \to \infty} \left\lceil \left(t_j^{(\mathrm{k})}\right)_1 \frac{X}{L_X} \right\rceil \qquad \text{and} \qquad \tilde{y}_j = \lim_{k \to \infty} \left\lceil \left(t_j^{(\mathrm{k})}\right)_2 \frac{Y}{L_Y} \right\rceil.$$

For $j \geq 2$, the size $(X-1, Y-1)$ is the maximum possible size for the integration domains I_j while ensuring that the condition in Equation (4.20) holds, since $(t_j)_1 X/L_X$ and $(t_j)_2 Y/L_Y$ are not integers in general. It is pointed out that the sizes and positions of the integration domains are integers and as such converge within a finite number of iterations in practice. Therefore, the integration domains stay constant after a finite number of iterations and the results from Chapter 3 carry over to the modified objective functional in Equation (4.19), even if the integration domains are adjusted during the minimization with the method described above.

The advantage of adjusting the position and size of the integration domains during the exit wave reconstruction is illustrated by an example reconstruction. An excerpt of the simulated focus series that was used as the input data is shown in Figure 4.9. The reconstruction of the exit wave and the registration of the image series was performed simultaneously using the nonlinear conjugate gradient method and the scaling factors for the exit wave and the translations were set to $c_1 = 10^2$ and $c_2 = 10^3$ (cf. Section 4.4.2). No scaling of the individual exit wave pixels was performed and the coefficient σ of the regularizer was set to zero. The integration domains were updated every 10 iterations based on the changes in the translations in the previous 10 iterations. The same TCC and buffer zone width was used for the reconstruction as for the simulation of the focus series, i.e. T_Z^{MAL} with $M = 4$, $\delta = 1$ nm and a buffer zone of width $W = 256$ pixels. As the initial guess for the exit wave a constant exit wave in real space corresponding to the mean image intensity was used. The initial guess for the translations was calculated from a cross-correlation of successive images in the focus series. The focus parameters were not optimized and kept constant and equal to the correct values throughout the minimization.

Figure 4.10 shows the energy for the first 1000 iterations of the nonlinear conjugate gradient method, where it can be seen that the energy increases after $10, 20, \ldots$ iterations. Because the integration domains are updated every 10 iterations, their sizes (X_j, Y_j) and positions $(\tilde{x}_j + W, \tilde{y}_j + W)$ may change. As a consequence, the simulated images and the input images are compared at new pixel positions $(x + \tilde{x}_j + W, y + \tilde{y}_j + W)$ in Equation (4.19). This temporarily increases the value of the objective functional until a good reconstruction at these new pixels has been achieved.

The plots in Figure 4.11 show the updates of the integration domains and the translation values during the reconstruction. Initially, the integration domain positions and sizes are set as follows:

Image	1	2	3	…	22	23	24
$\tilde{x}_j$	0	11	21	…	211	221	231
$\tilde{y}_j$	0	7	13	…	137	143	149
X_j	512	501	491	…	301	291	281
Y_j	512	501	491	…	301	291	281

The integration domain sizes (X_j, Y_j) decrease as j increases, since the initial guess for the translations is calculated from the relative shifts of successive images, which in turn are estimated with the cross-correlation. The error in the relative shifts is accumulated in the absolute shifts with respect to the first image $\widehat{g}_1$, i.e. the translations

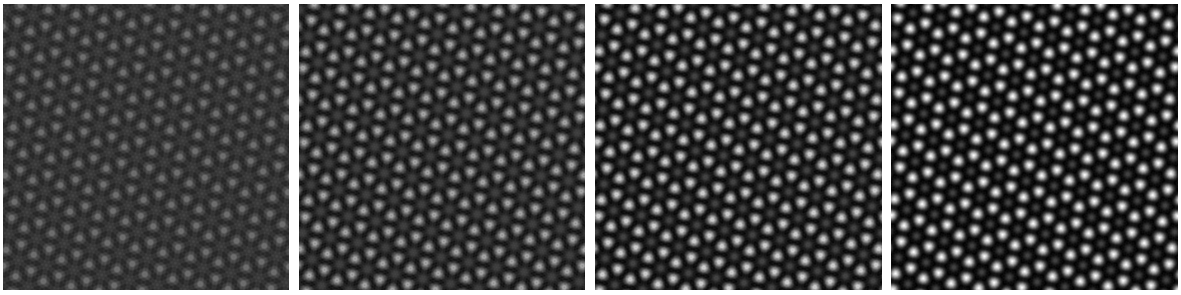

Figure 4.9: From left to right: images 13–16 of the focus series used for the reconstruction. The series consists of 24 images sampled at 512×512 pixels that each depict an area of 2.5×2.5 nm^2. The specimen drift between successive images in the series is 0.025 nm in x direction and 0.008 nm in y direction. The images were simulated from a simulated exit wave of a specimen that consists of carbon atoms arranged in a honeycomb structure. (The exit wave is calculated from a linear superposition of the projected potentials of the carbon atoms and thus not a realistic exit wave for graphene.) The focus value of the j-th image in the series is $Z_j = -10 + (j-1) \cdot 0.9$ nm for all $j \in \{1, \ldots, 24\}$ and the other microscope parameters are: spherical aberration $C_s = -70$ nm, accelerating voltage $U = 300$ kV (electron wavelength $\lambda \approx 0.00196875$ nm), objective aperture semiangle $\alpha_{\max} = 125$ mrad, coherence parameters $\alpha = 0.4$ mrad and $\Delta = 2.9$ nm. The images were simulated using T_Z^{MAL} with $M = 4$, $\delta = 1$ nm and a buffer zone of width $W = 256$.

$t_j^{(0)}$. Consequently, the accuracy of the initial guess $t_j^{(0)}$ generally decreases as j increases. For the initial guess, it was simply assumed that the translation of the j-th image is accurate within $10j$ pixels in both directions for all $j \in \{2, \ldots, 24\}$. This is certainly an overly careful error bound on the initial guess for the translations. However, since the error bounds are adjusted every 10 iterations, it is insignificant that the first 10 iterations are calculated on very small integration domains in this way.

After the first 10 iterations, the integration domains are updated to

Image	1	2	3	...	22	23	24
$\tilde{x}_j$	0	8	12	...	108	115	122
$\tilde{y}_j$	0	5	4	...	35	40	46
X_j	512	505	507	...	507	504	500
Y_j	512	502	507	...	507	500	492

and after another 40 iterations, the integration domains are

Image	1	2	3	...	22	23	24
$\tilde{x}_j$	0	6	11	...	108	113	119
$\tilde{y}_j$	0	3	4	...	35	37	39
X_j	512	510	510	...	510	511	510
Y_j	512	509	510	...	510	510	509

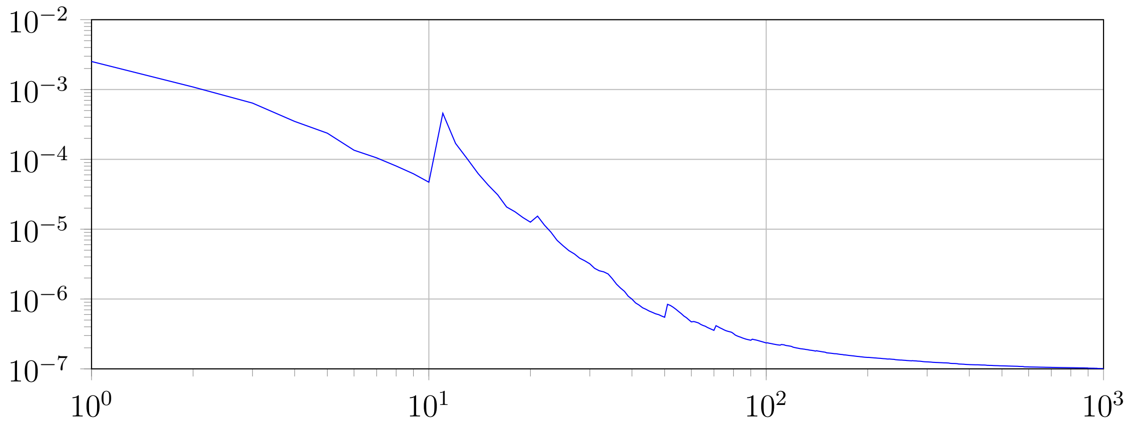

Figure 4.10: The value of the objective functional in Equation (4.19) for the first 1000 iterations of the minimization algorithm. The energy is not monotonically decreasing, since the integration domains are updated every 10 iterations and the objective functional is modified if the position or size of any integration domain changes.

After 720 iterations, the estimate for the translations is sufficiently good so that the automatic update of the integration domains comes to an end. At this time, the integration domain sizes are $(X_j, Y_j) = (511, 511)$ for $j = 2, \ldots, 24$ and $(X_1, Y_1) = (512, 512)$.

Figures 4.12 and 4.13 show the distance of the estimates for the exit wave and the translations to the correct values. Even with a buffer zone and dynamic integration domains, a perfect reconstruction of the exit wave near the image borders is not possible due to contrast dislocation (cf. Section 4.4.3). For this reason, the distance of the estimate for the exit wave to the correct exit wave is also shown for subsections of the domain $\{1, \ldots, 1024\} \times \{1, \ldots, 1024\}$ that are smaller than the central subsection of size 512×512. During the first 30 iterations, the translations are only slightly improved. The reason for this is likely that a good reconstruction of the exit wave is required in order to improve the registration with a comparison of simulated and experimental images, as is done by a minimization of the objective functional. After 70 iterations, the error in the estimated translations stays constantly below 0.1 pixel.

Besides the advantage that a continuation of the input images is no longer necessary with dynamic integration domains, each image in the focus series is used almost completely during the reconstruction. More precisely, a subsection of the size $(X-1) \times (Y-1)$ is utilized instead of a potentially much smaller subsection if the same integration domain is used for all images. Consequently, the area of the reconstructed exit wave is larger than the area of any image from the focus series, which can be seen in Figure 4.14.

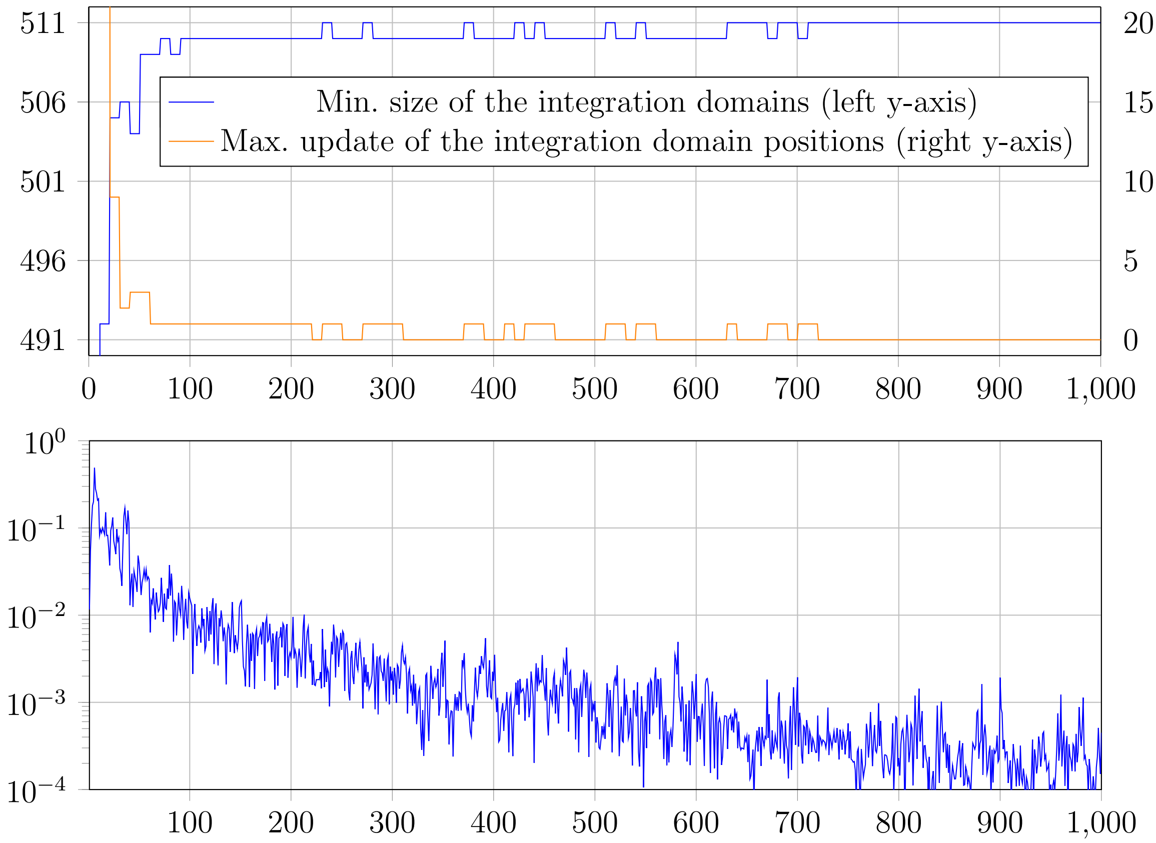

Figure 4.11: Upper plot: denote by $\left(X_j^{(k)}, Y_j^{(k)}\right)$ the sizes and by $\left(\tilde{x}_j^{(k)}, \tilde{y}_j^{(k)}\right)$ the positions of the integration domains after the iteration k. The blue graph shows the minimum of $X_j^{(k)}$ and $Y_j^{(k)}$ for all $j \in \{1, \dots, 24\}$, i.e.

$$\min\left(\left\{X_j^{(k)} \mid j = 1, \dots, 24\right\} \cup \left\{Y_j^{(k)} \mid j = 1, \dots, 24\right\}\right)$$

for all $k \in \{0, \dots, 1000\}$. The orange graph shows the maximum absolute value of the changes in the integration domain positions, i.e.

$$\max\left(\left\{\left|\tilde{x}_j^{(k)} - \tilde{x}_j^{(k-10)}\right| : j = 1, \dots, 24\right\} \cup \left\{\left|\tilde{y}_j^{(k)} - \tilde{y}_j^{(k-10)}\right| : j = 1, \dots, 24\right\}\right)$$

for all $k \in \{10, 20, 30, \dots, 1000\}$. Lower plot: the changes of the translations in pixel in the supremum norm, i.e. $\max\left(M_1^{(k)} \cup M_2^{(k)}\right)$ with

$$M_1^{(k)} := \left\{\frac{X}{L_X}\left|\left(t_j^{(k)}\right)_1 - \left(t_j^{(k-1)}\right)_1\right| : j = 1, \dots, 24\right\},$$
$$M_2^{(k)} := \left\{\frac{Y}{L_Y}\left|\left(t_j^{(k)}\right)_2 - \left(t_j^{(k-1)}\right)_2\right| : j = 1, \dots, 24\right\}$$

for $k \in \{1, \dots, 1000\}$.

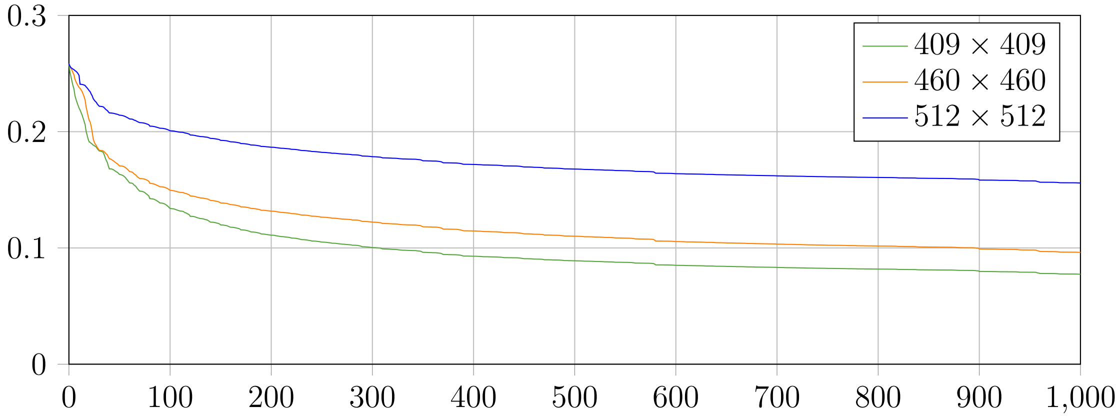

Figure 4.12: Denote by $\psi^{(k)} := \mathrm{IDFT}\left(\widehat{\Psi}^{(k)}\right) \in \mathbb{C}^{1024\times 1024}$ the estimate for the exit wave in real space after the k-th iteration and by $\widetilde{\psi} \in \mathbb{C}^{1024\times 1024}$ the real space exit wave that was used for the simulation of the focus series. The plots show the distance of the current estimate to $\widetilde{\psi}$ in the supremum norm for the central subsections of size 409×409, 460×460 and 512×512 pixels, i.e. $\left\|\left(\psi^{(k)} - \widetilde{\psi}\right)\mathbb{1}_M\right\|_\infty$, where M is the corresponding central subsection of the domain $\{1,\ldots,1024\} \times \{1,\ldots,1024\}$.

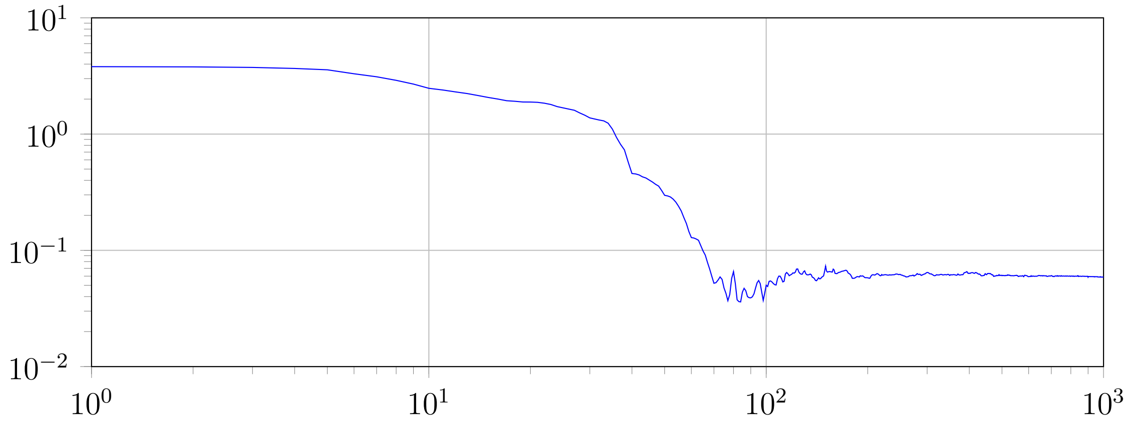

Figure 4.13: The distance of the estimate for the translations $t^{(k)} \in (\mathbb{R}^2)^N$ to the correct translation values $t^* \in (\mathbb{R}^2)^N$ in pixels with respect to the supremum norm for all iterations $k \in \{1,\ldots,1000\}$.

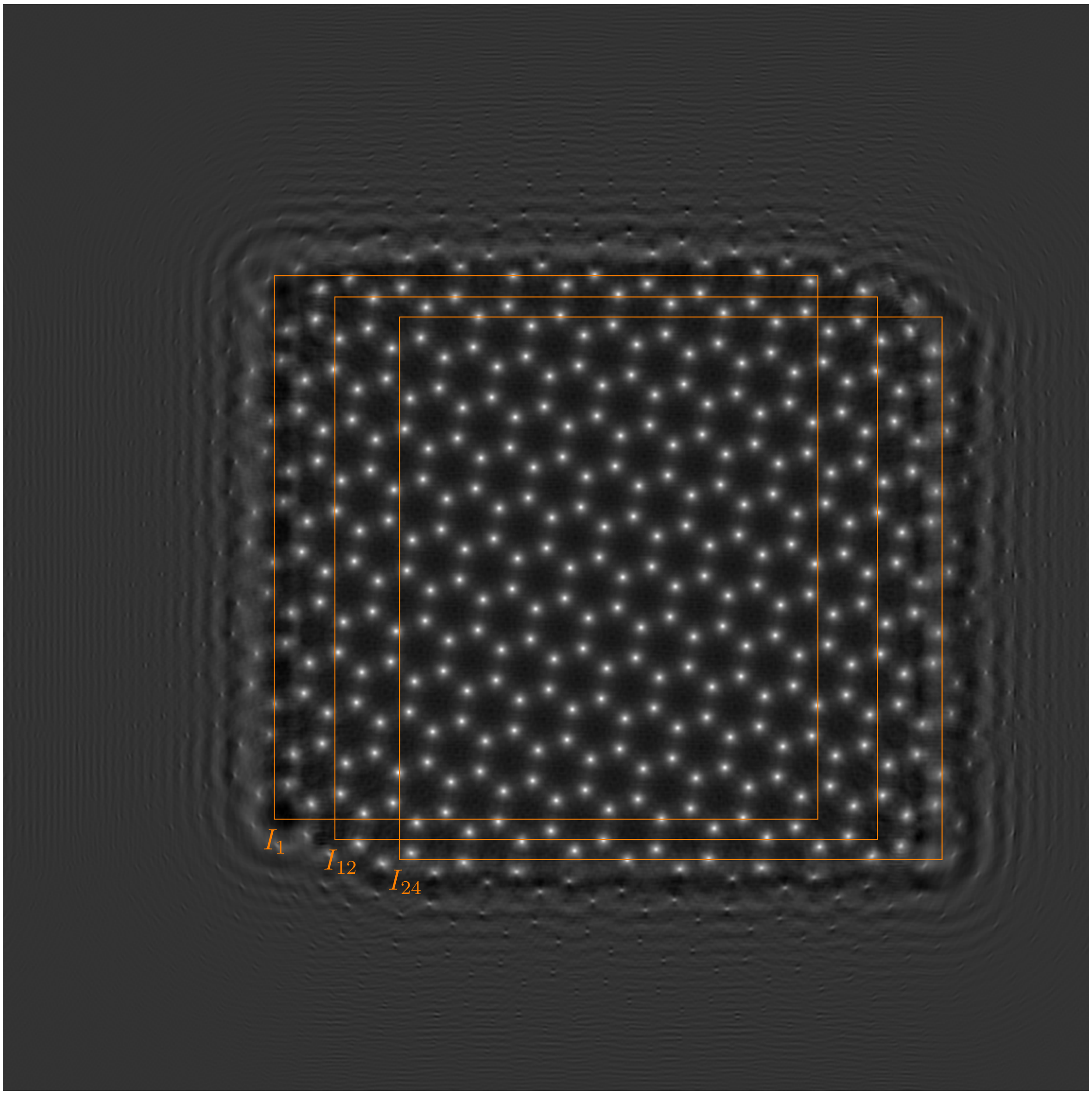

Figure 4.14: The phase of the reconstructed exit wave after 1000 iterations with the integration domains for $\widehat{g}_1$, $\widehat{g}_{12}$ and $\widehat{g}_{24}$ drawn on top. Outside of the intersection $\bigcap_{j=1}^{24} I_j = I_1 \cap I_{24}$, the reconstructed exit wave is influenced only by a part of the focus series.

4.5 Implementation

The numerical minimization of the objective functional has been implemented in a C++ program written within the framework of the Quocmesh library [23]. The source code is available on GitHub [1], along with all parameter files that are necessary to reproduce the experiments in Chapter 4 and Sections 5.1 and 5.2.

Two different kinds of TCCs can be used in the program: the focal integration approximation T_Z^{MAL} and Ishizuka's TCC T_Z^{Ishizuka}. Since Ishizuka's TCC is not $\star$-separable, the simulation of TEM images with T_Z^{Ishizuka} is quite expensive. However, the simulation of a TEM image $\widehat{\Psi} \star_{\widehat{T}_Z} \widehat{\Psi}$ for $\widehat{\Psi} \in \mathbb{C}^{X \times Y}$ can trivially be parallelized, since the individual pixel values

$$\big(\widehat{\Psi} \star_{\widehat{T}_Z} \widehat{\Psi}\big)_{x,y} = \sum_{a=0}^{X-1} \sum_{b=0}^{Y-1} \widehat{\Psi}^*_{a,b} \widehat{\Psi}_{a+x,b+y} \big(\widehat{T}_Z\big)_{(a+x,b+y),(a,b)}$$

for $x \in \{0, \ldots, X-1\}, y \in \{0, \ldots, Y-1\}$ are completely independent from each other. This high degree of parallelism is ideally suited for a computation on GPUs instead of CPUs if a non-$\star$-separable TCC is used. It was observed experimentally on our machine that the computation of TEM images with T_Z^{Ishizuka} on a GPU is roughly 50 times faster than the same computation on a multi-core processor. For this reason, the computations in the program are automatically performed on the (potentially multiple) available GPUs if T_Z^{Ishizuka} is used for image simulation. This is done by invoking suitable OpenCL kernels for the image simulation and derivative computations.

In an earlier version of the program the Gauss-Newton algorithm was implemented with both the LSMR and the QR method. However, as mentioned in Section 4.3, computing the QR factorization of the Jacobian in the Gauss-Newton algorithm is not viable even for relatively small image sizes. If the normal equations from the Gauss-Newton algorithm were solved approximately using the iterative LSMR algorithm instead, then the decrease in the objective functional's value in each iteration was comparable to the Quasi-Newton BFGS algorithm and the nonlinear CG method. However, a single iteration of the Gauss-Newton algorithm with LSMR took roughly 3 times as much computation time than a single iteration of the Quasi-Newton BFGS algorithm or the nonlinear CG method. For these reasons, the Gauss-Newton algorithm is not implemented in the current version of the program.

The most important files of the C++ program are shown in Figure 4.15. A brief overview of the purpose and contents of each of these files is given in the following:

Reconstruction.cpp: Starting point for the exit wave reconstruction. Performs the initialization and contains the main minimization loop.

Preconditioner.h: Implements the circular scale mask P for the exit wave as described in Section 4.4.2.

Functional.h: Contains the base code for the evaluation of the discrete objective functional with the data term given by Equation (4.19).

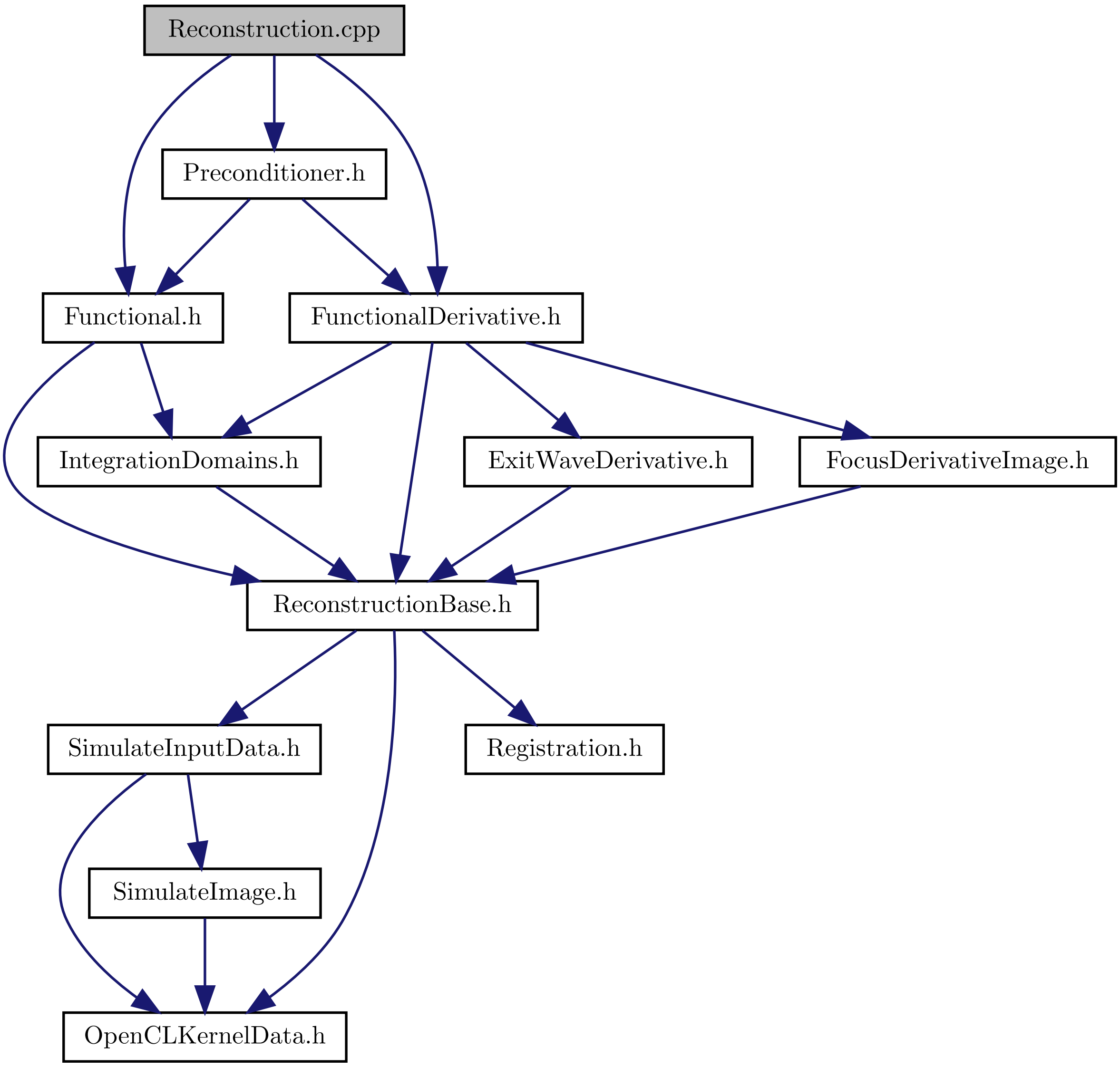

Figure 4.15: Reduced include graph of the program showing only the most important project files. Technical files that contain I/O routines and the implementation of certain auxiliary classes are omitted, as well as all Quocmesh and C++ STL includes. Generated using doxygen and graphviz.

FunctionalDerivative.h: Contains the base code for the computation of the gradient of the discrete objective functional with respect to the exit wave and the focus values as well as the full code for the computation of the derivatives with respect to the translations.

IntegrationDomains.h: Realizes the update of the integration domains on the basis of the changes in the translations in the previous iterations as described in Section 4.4.4.

ExitWaveDerivative.h: Implements the computation of the partial derivatives of the objective functional with respect to the exit wave pixels for the TCCs T_Z^{MAL} and T_Z^{Ishizuka}. The partial derivatives are computed according to the formulas given in Section 4.4.1. The effects of scaling (Section 4.4.2) and different integration domains (Section 4.4.4) on the derivatives are accounted for in FunctionalDerivative.h.

FocusDerivativeImage.h: Implements the computation of the partial derivatives of the objective functional with respect to the focus values for the TCCs T_Z^{MAL} and T_Z^{Ishizuka} using the formulas from Section 4.4.1. Similarly to the computation of the partial derivatives with respect to the exit wave pixels, the effects of scaling (Section 4.4.2) and different integration domains (Section 4.4.4) on the derivatives are already accounted for in FunctionalDerivative.h.

ReconstructionBase.h: Contains two fundamental data structures which are the basis for the implementation of the objective functional: (1) the `Arguments` class that contains the current estimate for the exit wave, the translations and the focus values and (2) the `InputData` class that stores the focus series and all constant microscope parameters such as the accelerating voltage and aberration coefficients other than the focus.

SimulateInputData.h: Implements functions for the simulation of exit waves corresponding to a superposition of single atoms or a lattice of point charges. Additionally, this file includes a function for the simulation of a focus series from a given exit wave, which may then be used as input data to the minimization algorithm.

Registration.h: Implements the registration of two images by calculating their cross-correlation with the FFT. This is used to calculate an initial guess for the image shifts.

SimulateImage.h: Implements the simulation of TEM images for the TCCs T_Z^{MAL} and T_Z^{Ishizuka}. In the case T_Z^{MAL}, the images are simulated using Equation (4.9), whereas for T_Z^{Ishizuka} the images are simulated using the definition of the discrete weighted autocorrelation directly.

OpenCLKernelData.h: Contains the interface to perform computations on the GPU with T_Z^{Ishizuka}. This includes the simulation of TEM images and the computation of the partial derivatives of the objective functional with respect to the exit wave pixels and the focus values.

5 Experiments

This chapter contains three numerical experiments that are used to analyse the behavior of the objective function from different point of views.

In the first experiment in Section 5.1, the use of a scale mask to scale individual exit wave pixels is evaluated (cf. Section 4.4.2). This is done by a comparison of three minimization strategies that employ different scale masks, which shows that a slightly lower energy can indeed be reached if a scale mask is used and updated regularly. In Section 5.2, the simultaneous optimization of the exit wave and the translations is compared with an alternating minimization scheme. It is shown that the overall performance with regard to reconstructing the exit wave is comparable, although the joint approach needs significantly fewer iterations to optimize the translations. In the concluding experiment in Section 5.3, the objective functional is used to reconstruct the exit wave for a real image series. It is shown that the estimates for the exit wave are structurally identical to exit waves that have been reconstructed with other methods. Furthermore, it is pointed out that the noise contained in the input images is reproduced to some extent in the simulated images at the expense of the quality of the exit wave.

5.1 Improved reconstruction of high frequencies with scaling

In this section, the value of preconditioning the exit wave is illustrated by an example, where the scaling of the individual exit wave pixels as described in Section 4.4.2 is used as the preconditioner. For this purpose, we adapt the notation from Section 4.4.2 and denote the scale mask for the exit wave at the k-th iteration by $P^{(k)} \in (\mathbb{R}^2)^{X \times Y}$.

The results of three different minimization strategies are compared:

1. No preconditioning: $P^{(k)}_{x,y} = (1,1)$ for all $x \in \{0, \dots, X-1\}$, $y \in \{0, \dots, Y-1\}$ and all iterations $k \in \mathbb{N}$.
2. Calculate the scale mask once before the first iteration and keep it constant for the entire minimization: $P^{(k)} = P^{(0)}$ for all $k \in \mathbb{N}$.
3. Calculate the scale mask before the first iteration and recalculate it every 30 iterations based on the current estimate.

An excerpt of the simulated focus series that was used as the input data for all three minimization strategies is shown in Figure 5.1. Only the exit wave was optimized using the nonlinear conjugate gradient method, whereas the translations and focus values were kept constant and equal to the correct values throughout the minimization. As the initial guess for the exit wave, a plane wave corresponding to the mean image intensity of the input images was used. In addition to the scaling of the exit wave pixels, the exit wave as a whole was also scaled by $c_1 = 10^2$ in order to keep the step size during the minimization close to 1, which is advantageous for the underlying implementation of the nonlinear conjugate gradient method. An exit wave buffer zone of size $W = 256$ pixels was used during the minimization. Clearly, the integration domains do not need to be adjusted as the translations stay constant. Therefore, the integration domains were immediately placed at the optimal positions and set to the maximum size of 511×511 pixels for the images $\widehat{g}_2, \dots, \widehat{g}_{12}$ respectively 512×512 pixels for the reference image $\widehat{g}_1$. All three reconstructions as well as the computation of the scale masks $P^{(k)}$ were performed using T_Z^{MAL} with the parameters $M = 4$ and $\delta = 1.5$ nm. The coefficient σ of the generalized Tikhonov regularizer was set to zero.

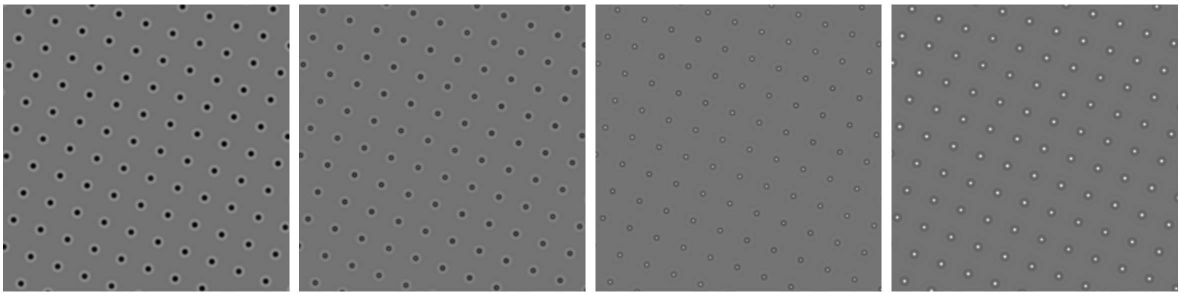

Figure 5.1: From left to right: images 6–9 of the focus series used for the exit wave scaling experiment. The series consists of 12 images in total sampled at 512×512 pixels that each depict an area of 6.4×6.4 nm^2. The specimen drift between successive images in the series is roughly 0.037 nm in x direction and 0.018 nm in y direction. The images were simulated from a simulated exit wave of an artificial specimen that consists of a lattice of uranium atoms. The focus value of the j-th image in the series is $Z_j = (-10 + (j-1) \cdot 1.5)$ nm for all $j \in \{1, \dots, 12\}$ and the other microscope parameters are: spherical aberration $C_s = -70$ nm, accelerating voltage $U = 300$ kV (electron wavelength $\lambda \approx 0.00196875$ nm), objective aperture semiangle $\alpha_{\max} = 125$ mrad, coherence parameters $\alpha = 0.4$ mrad and $\Delta = 3.8$ nm. The images were simulated using T_Z^{Ishizuka} and a buffer zone of width $W = 512$.

Figure 5.2 shows the energy during the first 1000 iterations for all three minimization strategies. First, it can be seen that the approach of using $P^{(0)}$ throughout the entire minimization is not useful. While this approach temporarily yields a lower energy than using no preconditioner for the iterations 25–110, the scale mask $P^{(0)}$ is not useful for the iterations $k \geq 110$. This is expected to be caused by the fact that the scale mask $P^{(0)}$ is calculated based on the initial guess for the exit wave, which is equal to a constant wave in real space and thus significantly different from all further estimates of the exit wave $\Psi^{(k)}$ at the iterations $k \geq 1$. If the scale mask is updated every 30 iterations,

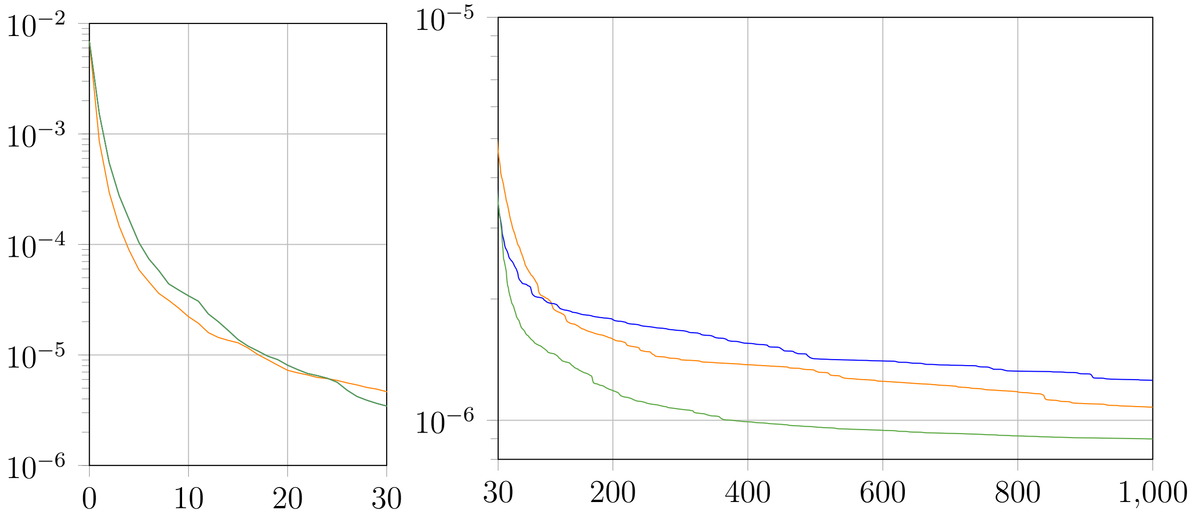

Figure 5.2: The energy for the first 1000 iterations of the three different minimization strategies: no scaling of the individual exit wave pixels (1, orange), initial calculation of the scale mask (2, blue), initial calculation of the scale mask and updating the scale mask every 30 iterations (3, green). For the second and third minimization strategy, the energy is identical during the first 30 iterations.

then the minimization indeed yields a lower energy after 1000 iterations compared to the other two approaches. The images of the scale masks shown in Figure 5.3 confirm the observations from the energy: the scale mask $P^{(0)}$ differs greatly from the other scale masks since it is computed from the initial guess for the exit wave. For this reason, $P^{(0)}$ is only useful in the beginning of the minimization.

Figures 5.4 and 5.5 show the power spectra as well as the amplitude and phase of the estimated exit waves after 1000 iterations. Since the scale masks are close to 1 near the low frequencies, the reconstruction of the low frequencies is almost identical for all three minimizations strategies. Instead, the scale masks accelerate the reconstruction of high frequencies, which results in a lower energy and consequently a slightly better match of the input images and the images simulated from the estimate for the exit wave by the definition of the objective functional.

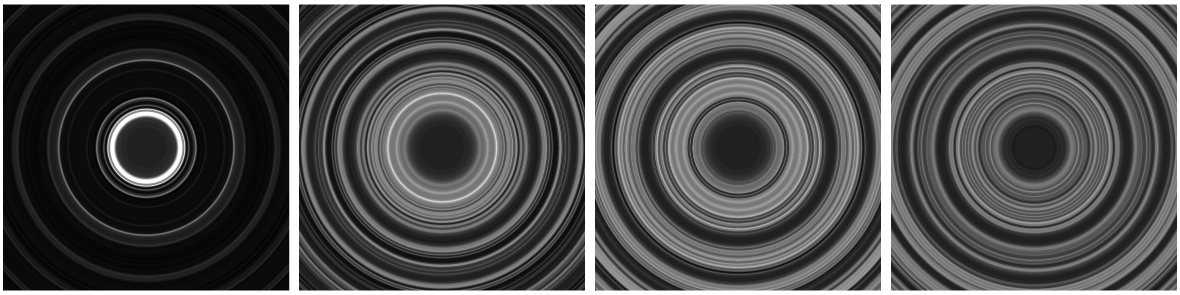

Figure 5.3: The real part of the scale masks $P_1^{(0)}$, $P_1^{(30)}$, $P_1^{(180)}$, $P_1^{(990)}$ (from left to right) during the third minimization strategy, where the scale masks are updated every 30 iterations. The scale masks range from 0.56 (black) to 10.08 (white). The imaginary parts of the scale masks are very similar to the real parts.

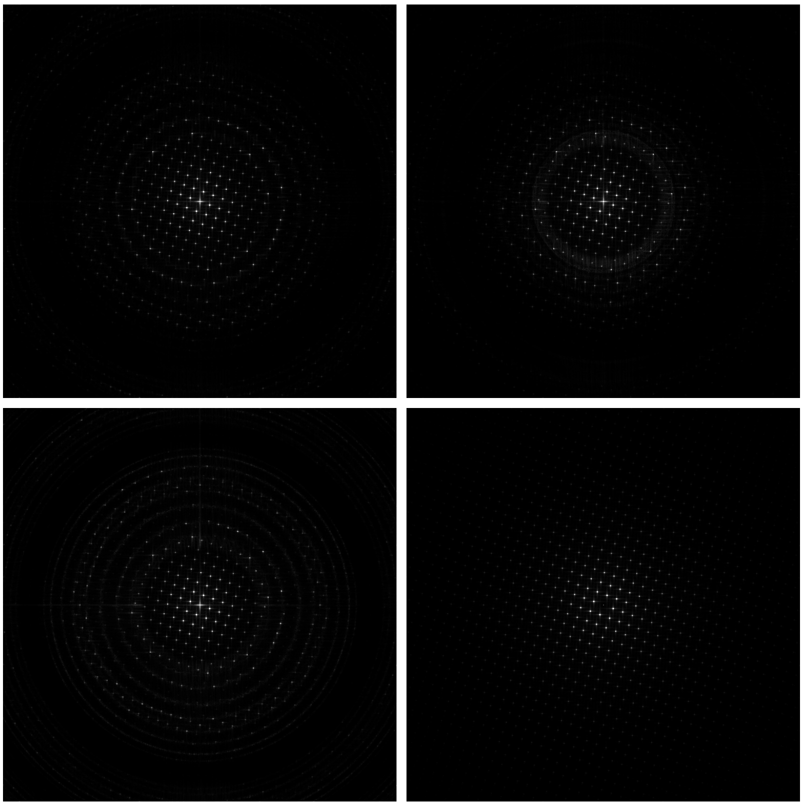

Figure 5.4: The power spectra of the reconstructed exit waves after 1000 iterations and the correct exit wave: no scaling of the individual exit wave pixels (top left), initial calculation of the scale mask (top right), initial calculation of the scale mask and recalculating the scale mask every 30 iterations (bottom left), correct exit wave (bottom right). The circular ring pattern in the power spectra are an artifact of the TCC T_Z^{MAL} and do not occur if the TCC T_Z^{Ishizuka} is used for the forward model.

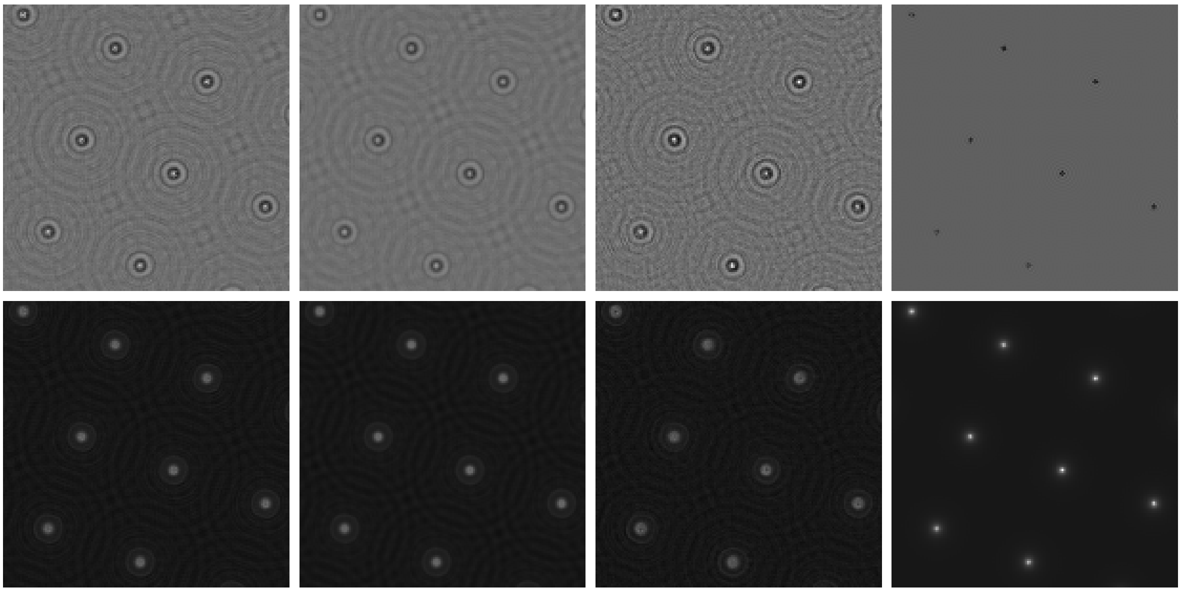

Figure 5.5: Subsections of the amplitudes (top row) and phases (bottom row) of the reconstructed exit waves (first three columns) and the correct exit wave (rightmost column). Scaling method for the exit wave pixels during the minimization (from left to right): no scaling of the individual exit wave pixels, initial calculation of the scale mask, initial calculation of the scale mask and recalculating the scale mask every 30 iterations. Note that a perfect reconstruction of the exit wave is not possible because (1) different forward models were used for the simulation of the input images and the simulation during the reconstruction and (2) because of contrast dislocation (see Section 4.4.3). Using a different forward model for the input data simulation and the reconstruction is meaningful as in this way the experiments are closer to the application of the objective functional to real input data.

5.2 Alternating vs. joint minimization

The MIMAP and MAL algorithms both use some kind of alternating minimization scheme for the reconstruction of the exit wave and the optimization of the translation parameters. Since a successful exit wave reconstruction crucially depends on a good registration of the focus series, it is interesting to investigate whether a joint minimization algorithm performs equally well or even better than an alternating minimization.

In this section, two experiments are described that were performed on exactly the same input data. The only difference between the two experiments is that the first experiment performed the exit wave reconstruction and image registration simultaneously, whereas in the second experiment alternatingly 5 iterations were performed for the optimization of the exit wave and 5 iterations for the optimization of the translations.

The only difference between the two experiments is that the minimization in the first experiment was performed by a joint optimization of the exit wave and the translations, whereas the minimization in the second experiment alternatingly performed 5 iterations for the optimization of the exit wave and 5 iterations for the optimization of the translations. The optimization of the exit wave and the translations was performed for 5 successive iterations at a time instead of just 1 iteration because of the conjugate gradient method. If the optimization of the exit wave and the translations was alternated after every iteration, then the conjugate gradient method would reduce to a the steepest descent as the previous descent directions can not be used to adapt the following direction.

An excerpt of the simulated focus series that was used as the input data to the minimizations is shown in Figure 5.6. The exit wave was initially set to a plane wave corresponding to the mean image intensity and the translations were initialized based on a cross-correlation registration of successive images in the series. No scaling of individual exit wave pixels was used, but the scaling factors for the exit wave and the translations were set to $c_1 = c_2 = 10^2$ in order to keep the step size during the minimization near 1. A buffer zone of size $W = 256$ was used during the minimization and the integration domains were adjusted every 10 iterations with the method described in Section 4.4.4. The minimizations were performed using the nonlinear conjugate gradient method and T_Z^{MAL} with the parameters $M = 3$, $\delta = 3.5$ nm as the TCC for image simulation.

In total, 1000 iterations have been performed for both minimizations. Figure 5.7 shows the energy of the objective functional during the experiments. It can be seen that the alternating minimization scheme takes considerably more iterations to achieve the same reduction of the energy as the joint minimization scheme. The energy after 500 iterations of the joint minimization is roughly $2.52 \cdot 10^{-7}$ and the energy after 1000 iterations of the alternating minimization is approximately $7.44 \cdot 10^{-7}$. Although in both cases the exit wave has been optimized for a total of 500 iterations, the energy is by a factor of 3 larger for the alternating minimization strategy as compared to the joint minimization strategy. Note that the number of iterations given for the alternating minimization strategy is equal to the sum of the iterations that have been performed for the optimization of the exit wave and the optimization of the translations. Although the exit wave is only optimized for the half of the iterations, the computational cost of

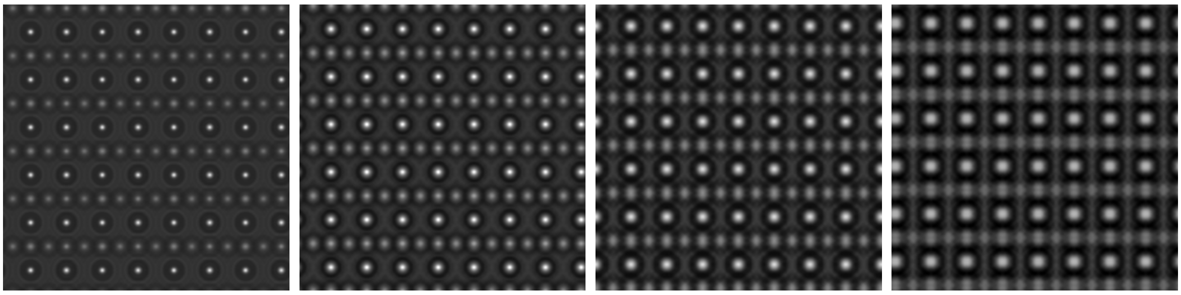

Figure 5.6: From left to right: images 8–11 of the focus series used for the comparison of alternating and joint minimization. The series consists of 16 images sampled at 512×512 pixels that each depict an area of approximately 2.93×2.93 nm^2. The specimen drift between successive images in the series is -0.04 nm in x direction and 0.03 nm in y direction. The exit wave that was used for the simulated of the focus series corresponds to a cubic $SrTiO_3$ lattice that was simulated with the multislice method [11] using the Dr. Probe software [4]. The focus value of the j-th image in the series is given by $Z_j = (-20 + (j-1) \cdot 3)$ nm for all $j \in \{1, \ldots, 16\}$ and the other microscope parameters are: spherical aberration $C_s = -700$ nm, accelerating voltage $U = 300$ kV (electron wavelength $\lambda \approx 0.00196875$ nm), objective aperture semiangle $\alpha_{\max} = 125$ mrad, coherence parameters $\alpha = 0.4$ mrad, $\Delta = 3.8$ nm. The images were simulated using T_Z^{MAL} with the parameters $M = 3$, $\delta = 3.5$ nm and a buffer zone of width $W = 256$.

performing a single iteration is almost identical for both experiments, i.e. 1000 iterations of the joint minimization approach take roughly the same time as 1000 iterations of the alternating minimization approach (c.f. Section 4.4.1). For this reason the number of iterations of the alternating minimization approach is not rescaled for the plots in this section.

The difference between the two approaches is also visible in Figure 5.8, which shows the accuracy of the registration during the experiments. There, it can be seen that the optimization of the translations takes significantly more iterations for the alternating minimization approach as compared to the joint minimization approach.

Nevertheless, although the estimates for the translations are worse for the alternating minimization approach, the reconstruction of the exit wave is surprisingly almost equally good in both reconstructions. This can be seen in Figure 5.9, which shows that the correct exit wave is equally well approximated by both minimization strategies and that there is only very little visible difference between the amplitudes and phases of the reconstructed exit waves. The joint minimization strategy only seems to be able to reconstruct the exit wave slightly better near the border of the integration domains.

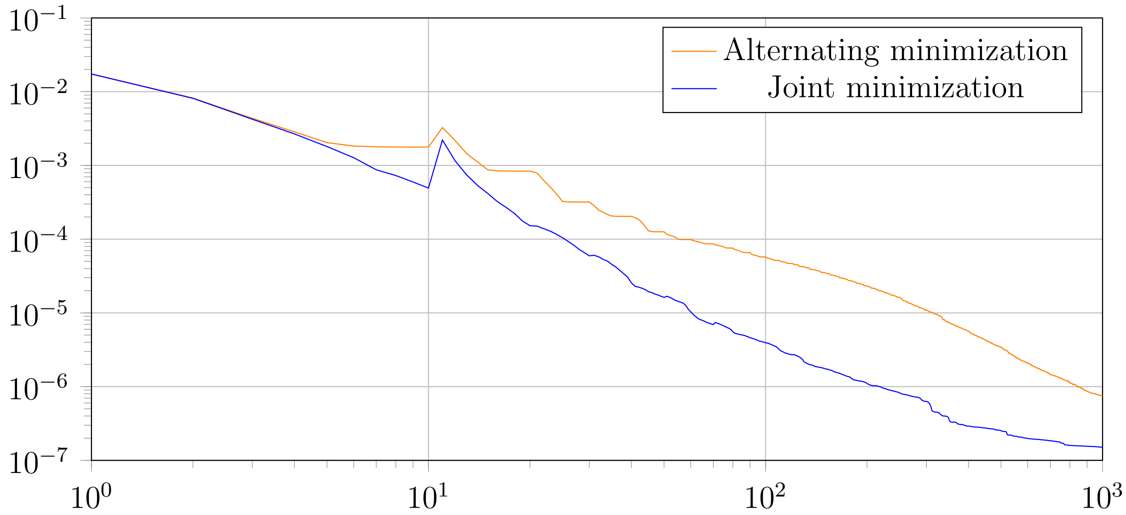

Figure 5.7: The value of the objective functional for the first 1000 iterations of both the alternating minimization approach and the joint minimization approach.

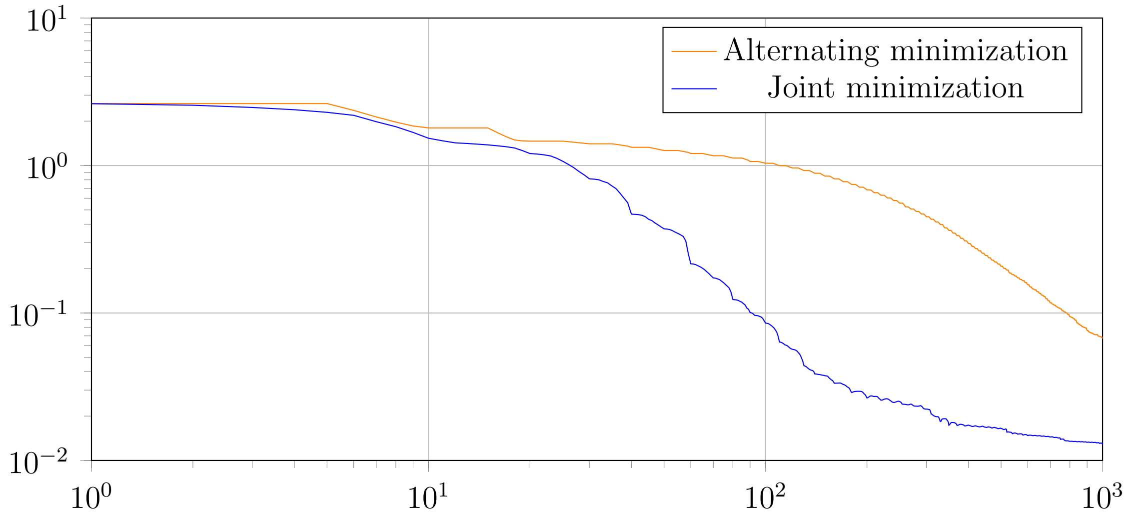

Figure 5.8: The distance of the estimate for the translations $t^{(k)} \in (\mathbb{R}^2)^N$ to the correct translation values $t^* \in (\mathbb{R}^2)^N$ in pixels with respect to the supremum norm for all iterations $k \in \{1, \ldots, 1000\}$.

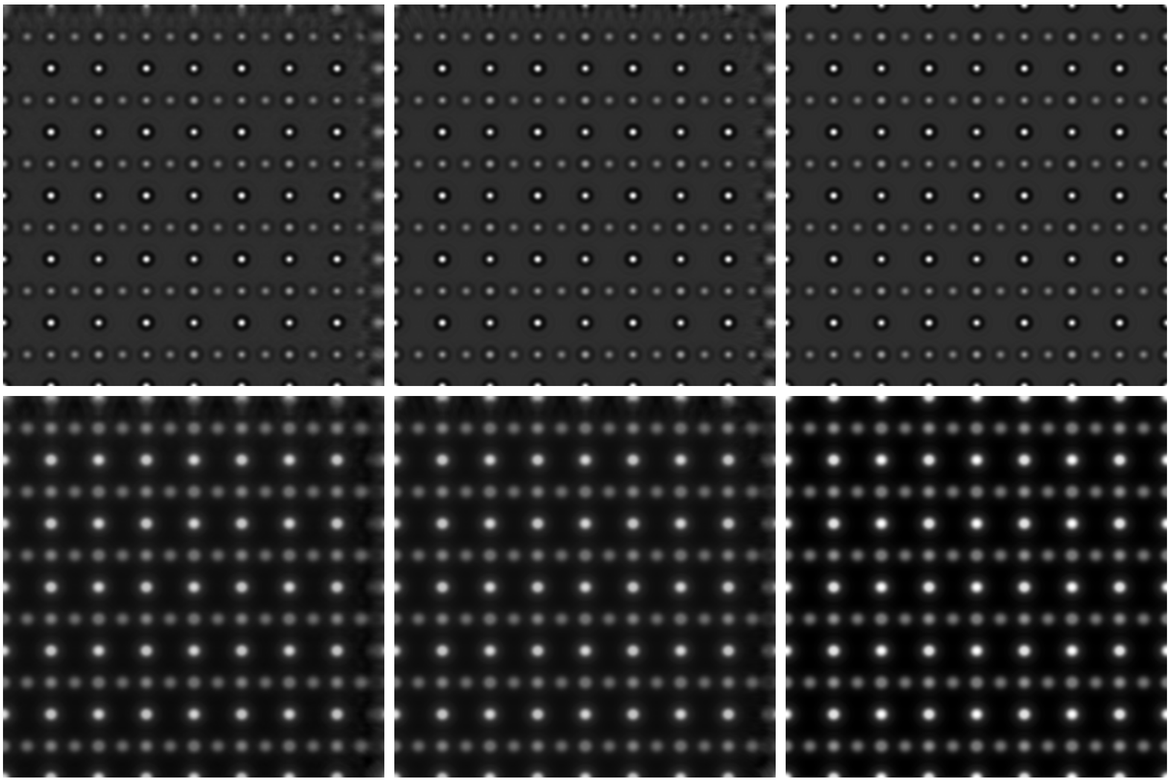

Figure 5.9: The amplitudes (top row) and phases (bottom row) of the reconstructed exit wave after 1000 iterations and the correct exit wave. The left column shows the exit wave estimate from the alternating minimization approach and the middle column shows the result of the joint minimization approach. The right column shows the amplitude and phase of the exit wave that has been used for the simulation of the input data. The value range of the phases of the reconstructed exit waves is -0.24 (white) to 1.24 (black). This range differs from the range of the correct phase $[0.08, 1.46]$ mainly by a constant offset of approximately 0.26, which was subtracted from the correct phase for the comparison. This is due to the fact that the value of the objective functional is invariant to a global phase shift, which implies that the correct exit wave can only be determined up to a constant that can arbitrarily be added to the phase. The buffer zone has been removed from the reconstructed exit waves for the comparison with the correct exit wave.

5.3 Exit wave reconstruction with a real focus series

In this section, the exit wave is reconstructed for an experimental TEM image series of CeO_2. The result is compared to the exit waves from two other reconstructions that have been performed on the same input data. The results of the other two reconstructions have kindly been provided by Chen Huang from the Oxford Electron Image Analysis Group at the University of Oxford.

Figure 5.10 shows an excerpt of the image series that was used for the reconstruction. The objective functional was minimized using the nonlinear conjugate gradient method and the TCC T_Z^{MAL} with the parameters $M = 3$ and $\delta = 1.8$ nm was used for the simulation of TEM images. Since no objective aperture was employed during the recording of the image series, the radius of the aperture function r_a was set to a value that exceeds the magnitude of the highest frequencies that are contained in the discrete images. For the practical minimization of the objective functional, this is equivalent to using no aperture function at all. A constant electron wave corresponding to the mean image intensity in real space was used as the initial guess for the exit wave. The translations were initialized based on the relative shifts between successive images that have been calculated with the cross-correlation. The focus values were kept constant for the entire minimization. As in Section 5.2, no scaling of the individual exit wave pixels was performed. However, the exit wave and the translations were similarly scaled by $c_1 = c_2 = 10^2$ in order to keep the stepsize during the minimization close to 1. A buffer zone of size $W = 128$ pixels was used during the minimization and the integration domains were adjusted every 10 iterations with the method described in Section 4.4.4. No regularization of the exit wave was used, i.e. the coefficient σ of the regularizer was set to zero.

It can be seen in Figure 5.11 that the energy is reduced the most during the first 20 iterations. This agrees with the observations in [10, 66], where it is reported that convergence of the MAL algorithm is reached in as little as 10 to 15 iterations. However, the energy still decreases after the first 20 iterations, albeit more slowly, until it stays more or less constant after roughly 200 iterations.

The initial guess for the translations is very close to the true image shifts so that the changes in the translations shown in Figure 5.12 are small from the very beginning of the minimization. Correspondingly, the integration domains after the first 10 iterations are already very close to their maximum size of 767 × 767 pixels:

Image	1	2	3	4	...	18	19	20	21
X_j	768	763	763	765	...	764	763	760	761
Y_j	768	764	766	766	...	763	766	765	766

After 100 iterations, the integration domain sizes are

Image	1	2	3	4	...	18	19	20	21
X_j	768	767	766	767	...	767	766	766	766
Y_j	768	767	766	766	...	767	767	767	767

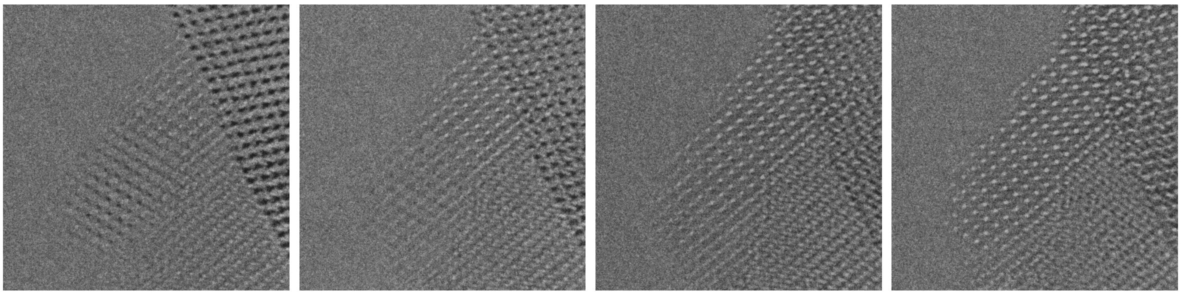

Figure 5.10: Images 10–13 of the experimental focus series of $\mathrm{CeO_2}$ that was used as input data to the objective functional. In total, the series consists of 21 images sampled at 768 × 768 pixels that each depict an area of approximately 6.96×6.96 nm^2. The focus value of the j-th image in the series is given by $Z_j = (-20 + (j-1) \cdot 2)$ nm for all $j \in \{1, \dots, 21\}$ and the other microscope parameters are: spherical aberration $C_s = -200$ nm, accelerating voltage $U = 300$ kV (electron wavelength $\lambda \approx 0.00197$ nm), coherence parameters $\alpha = 0.1$ mrad and $\Delta = 2.8$ nm. No objective aperture was used during the image acquisition. Image series courtesy of Chen Huang from the Oxford Electron Image Analysis Group, University of Oxford.

A visual examination of the registration shows that at this point, the simulated images are very well aligned to the experimental images.

Figures 5.13 and 5.14 show the amplitudes and phases of the estimated exit waves after 5, 10, 15 and 20 iterations. The comparison with the reference exit waves shows that the structure is reconstructed very well, although the contrast is lower for the exit wave reconstructions that have been performed with the objective functional E_σ.

After the first 20 iterations, the quality of the exit wave estimates deteriorates as is shown in Figure 5.15. However, at the same time the energy of the objective functional continues to decrease, which implies that the approximation of the experimental images by the simulated images improves. This is illustrated in Figure 5.16, which shows that the simulated images partially reproduce the noise in the experimental input images and therefore yield a better match and a lower energy. Experiments with other values of the parameters M and δ of the TCC T_Z^{MAL} and the TCC T_Z^{Ishizuka} resulted in the same outcome. Therefore, it is assumed that the deterioration of the exit wave is directly caused by the noise in the input images and unrelated to the particular TCC that is used in the forward model. Additionally, this effect has not been observed in any of the experiments that have been performed on simulated image series without noise and the use of the generalized Tikhonov regularizer (i.e. $\sigma > 0$) does not seem to reduce this effect. The images shown here imply that the minimization of the objective functional is not meaningful for noisy input data if no additionally regularization of the exit wave is used that prevents this effect. This problem likely has not been noticed or neglected before, as it is generally reported that convergence of the MIMAP or MAL algorithms is reached after a low number of iterations [10, 66, 39].

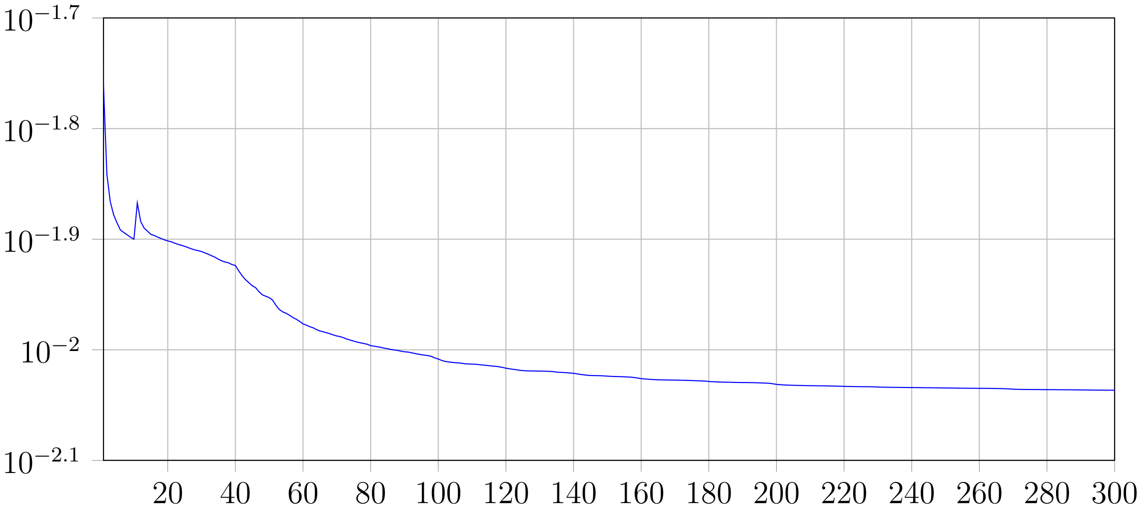

Figure 5.11: The value of the objective functional during the first 300 iterations.

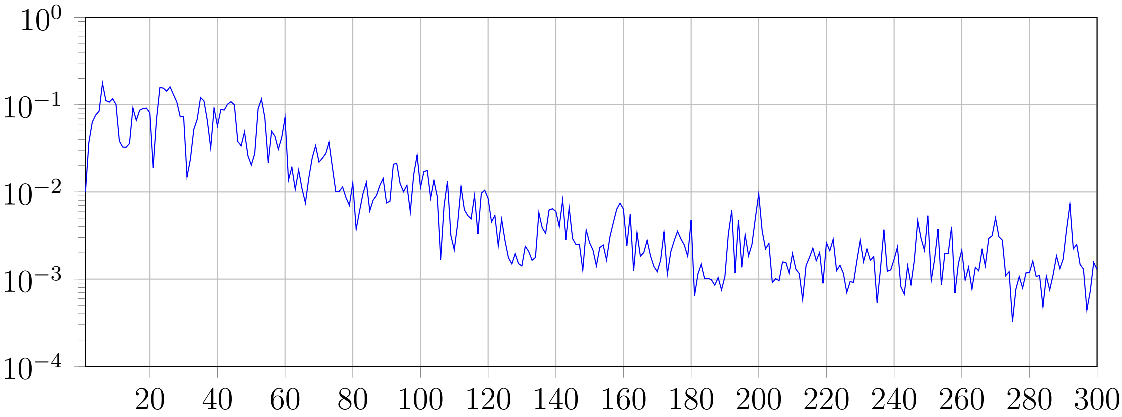

Figure 5.12: The changes in the translations with respect to the supremum norm in pixels, i.e. $\max\left(M_1^{(k)} \cup M_2^{(k)}\right)$ with

$$M_1^{(k)} := \left\{ \frac{X}{L_X} \left| \left(t_j^{(k)}\right)_1 - \left(t_j^{(k-1)}\right)_1 \right| : j = 1, \dots, 24 \right\},$$

$$M_2^{(k)} := \left\{ \frac{Y}{L_Y} \left| \left(t_j^{(k)}\right)_2 - \left(t_j^{(k-1)}\right)_2 \right| : j = 1, \dots, 24 \right\}$$

for $k \in \{1, \dots, 300\}$, where $t_j^{(k)}$ denotes the estimate for the translations after the k-th iteration.

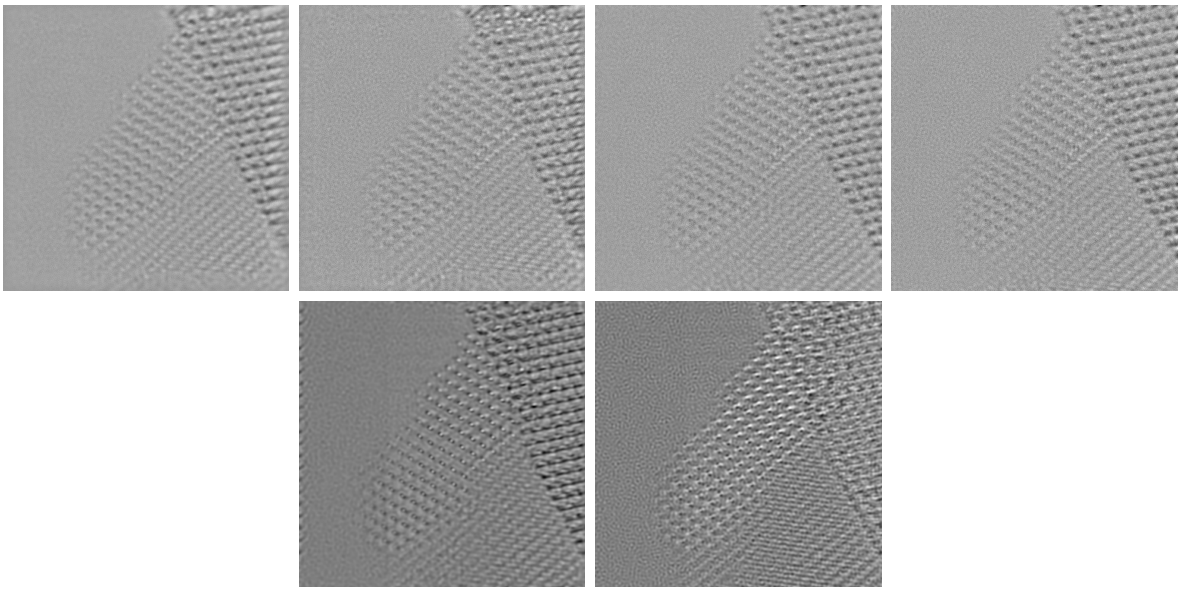

Figure 5.13: Top row (from left to right): The amplitudes of the estimate for the exit wave after 5, 10, 15 and 20 iterations. The bottom row shows the amplitude of an exit wave that has been reconstructed with an iterative method similar to the MAL algorithm (left image) and a linear Wiener-filter method (right image). The contrast of the amplitudes of the reconstructed exit wave has been rescaled independently of the reference images.

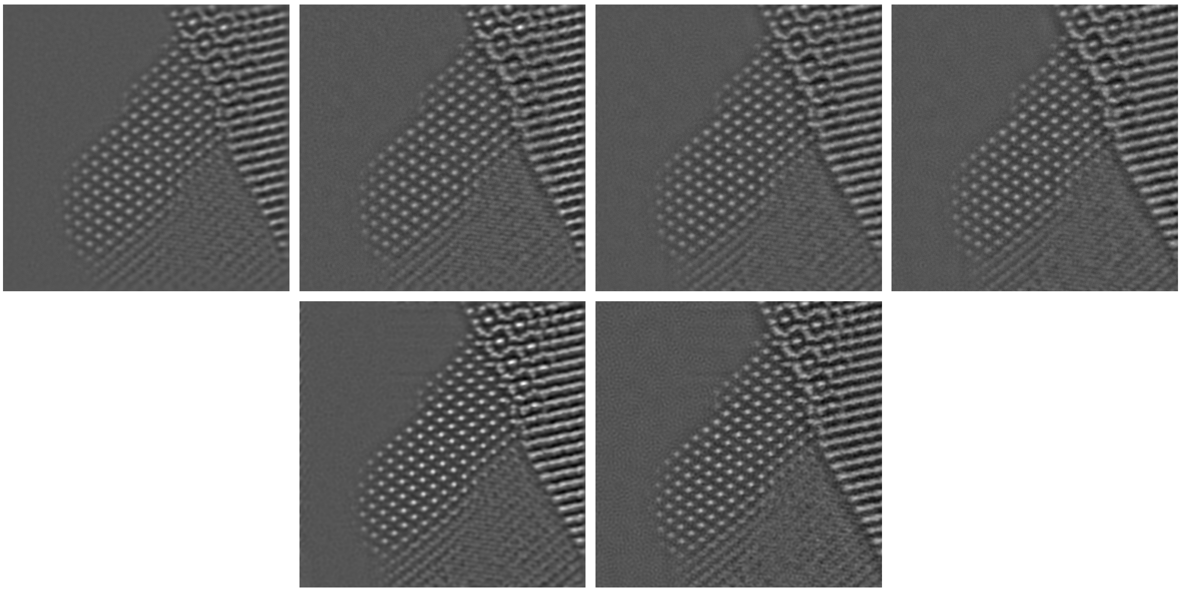

Figure 5.14: Top row (from left to right): The phases of the estimate for the exit wave after 5, 10, 15 and 20 iterations. The bottom row shows the phase of an exit wave that has been reconstructed with an iterative method similar to the MAL algorithm (left image) and a linear Wiener-filter method (right image). The contrast of the phases of the reconstructed exit wave has been rescaled independently of the reference images.

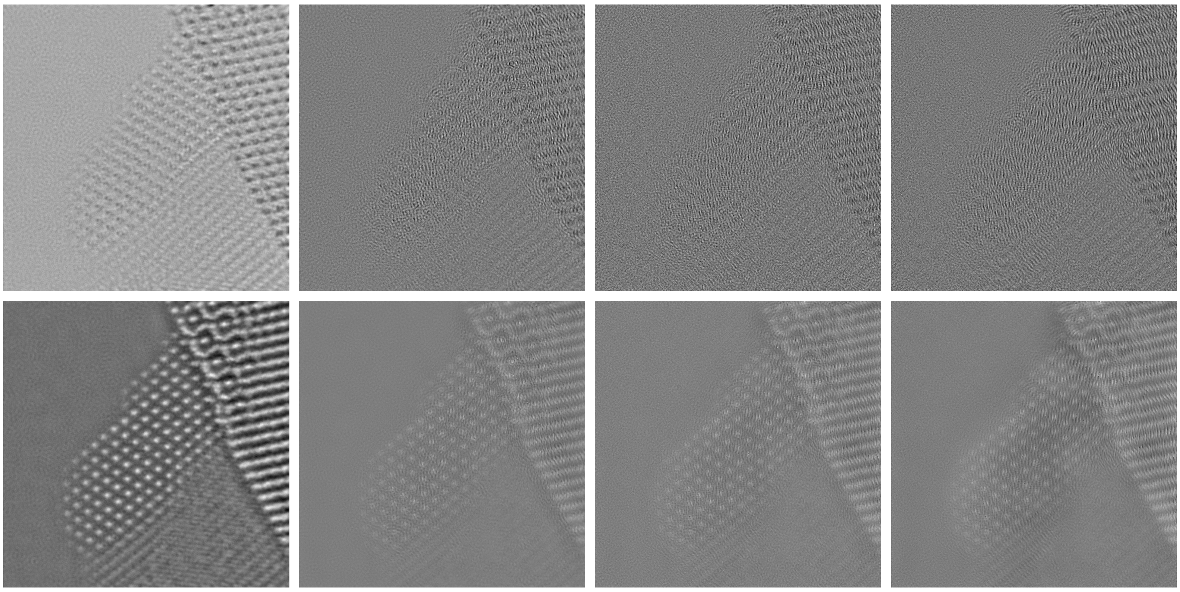

Figure 5.15: The amplitudes (top row) and phases (bottom row) of the estimates for the exit wave after 20, 50, 100 and 200 iterations (from left to right). It can be seen that the quality of the estimates gradually decreases as the number of iterations increases. The contrast has been enhanced for each image individually in order to show the effect most clearly.

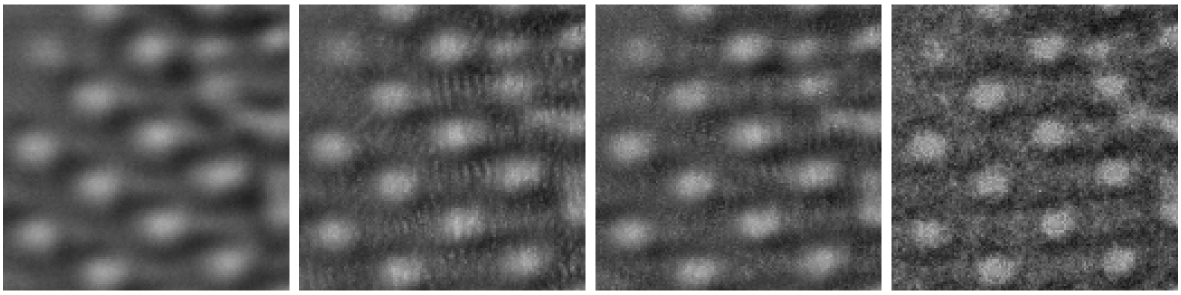

Figure 5.16: A subsection of the simulated images for the focus value Z_{15} from the current estimate of the exit wave after 20, 50 and 200 iterations (from left to right). More explicitly, the images depict a subsection of the size 128×128 pixels of IDFT $\left(\widehat{\Psi}^{(k)} \star_{\widehat{T}^{\mathrm{MAL}}_{Z_{15}}} \widehat{\Psi}^{(k)}\right)$ for $k \in \{20, 50, 200\}$. The rightmost image shows the corresponding subsection of the experimental image $\widehat{g}_{15}$.

Conclusions

In this thesis, a novel approach for joint exit wave reconstruction and image registration has been proposed by introducing the objective functional E_σ.

The analysis of the objective functional was facilitated by the results regarding the forward model. In particular, the concepts of the weighted cross-correlation and $\star$-separable weights were introduced and served as the central tool for the investigation of the properties of simulated TEM images. The notion of the weighted cross-correlation allowed to carry over several results from the ordinary cross-correlation with suitable conditions on the weight function. Most importantly, in several instances it was shown that the convolution theorem can be applied in a meaningful way to the weighted cross-correlation if the weight can be uniformly approximated by $\star$-separable weights. This formed the basis for the proof of the weak lower semi-continuity of the objective functional and was also used to show the useful property that the simulated TEM images are real-valued and nonnegative. In order to apply the results to the forward model of TEM image simulation, the most widely used TCCs were analyzed and it was shown that each of these TCCs satisfies the properties I–IV under suitable conditions on their constituent functions.

One of the main results of this thesis was to establish the existence of minimizers for the objective functional, which was done with the direct method. It was shown that the data term of the objective functional is not coercive with respect to the exit wave for $\star$-separable weight functions, which suggested the introduction of a nonlinear Tikhonov regularizer. As a first step towards the investigation of the uniqueness of minimizers, it was shown that the objective functional is not convex except in the trivial case where all input images are identical to zero. This result was extended for a simplified version of the functional that corresponds to a perfectly coherent and aberration-free TEM, where it was shown that this functional is not convex on any neighborhood of a global minimizer.

The analysis of the forward model and the inverse problem was complemented by a numerical analysis of the objective functional. A discretization of the objective functional was derived with particular emphasis on the precise connection between the continuous Fourier transform and the discrete Fourier transform. Several practical aspects of the numerical minimization were discussed and a novel preconditioner for the exit wave was proposed. The dynamic adjustment of the integration domains was introduced, which enables the use of the full input data for the reconstruction while removing the errors caused by image continuation at the same time.

Finally, several experiments have been described that evaluate the use of the objective functional from different perspectives. As a proof of concept, it was shown that a preconditioner for the exit wave can improve the reconstruction of high frequencies and decrease the number of necessary iterations. The novel joint minimization approach was compared with the usual alternating minimization approach, which showed that the joint minimization strategy performed slightly better. The experiment on noisy input data revealed that the noise is partially reproduced in the simulated images after a high number of iterations, which reduces the usefulness of the estimates for the exit wave. Therefore, the minimization of the data term of the objective functional for noisy input data is not useful without an additional regularization of the exit wave that prevents this effect, which is not achieved with the nonlinear Tikhonov regularizer.

Bibliography

[1] Source code of the C++ program for exit wave reconstruction. `https://github.com/CDoberstein/ExitWaveReconstruction`.

[2] H. W. Alt. *Linear Functional Analysis*. Springer, 2016. `https://doi.org/10.1007/978-1-4471-7280-2`.

[3] L. Armijo. Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics*, 16(1):1 – 3, 1966. `https://doi.org/10.2140/pjm.1966.16.1`.

[4] J. Barthel. Dr. probe: A software for high-resolution stem image simulation. *Ultramicroscopy*, 193:1 – 11, 2018. `https://doi.org/10.1016/j.ultramic.2018.06.003`.

[5] J. Barthel, T. E. Weirich, G. Cox, H. Hibst, and A. Thust. Structure of $Cs_{0.5}[Nb_{2.5}W_{2.5}O_{14}]$ analysed by focal-series reconstruction and crystallographic image processing. *Acta Materialia*, 58:3764 – 3772, 2010. `https://doi.org/10.1016/j.actamat.2010.03.016`.

[6] M. Beleggia, M. A. Schofield, V. V. Volkov, and Y. Zhu. On the transport of intensity technique for phase retrieval. *Ultramicroscopy*, 102:37 – 49, 2004. `https://doi.org/10.1016/j.ultramic.2004.08.004`.

[7] M. Born and E. Wolf. *Principles of optics*. Cambridge University Press, 7th edition, 2009.

[8] K. Bredies and D. Lorenz. *Mathematische Bildverarbeitung*. Vieweg+Teubner, 2011. `https://doi.org/10.1007/978-3-8348-9814-2`.

[9] W. Coene, G. Janssen, M. Op de Beeck, and D. Van Dyck. Phase retrieval through focus variation for ultra-resolution in field-emission transmission electron microscopy. *Physical Review Letters*, 69:3743 – 3746, 1992. `https://doi.org/10.1103/PhysRevLett.69.3743`.

[10] W. Coene, A. Thust, M. Op de Beeck, and D. Van Dyck. Maximum-Likelihood Method for Focus-Variation Image Reconstruction in High Resolution Transmission Electron Microscopy. *Ultramicroscopy*, 64:109 – 135, 1996. `https://doi.org/10.1016/0304-3991(96)00010-1`.

[11] J. M. Cowley and A. F. Moodie. The Scattering of Electrons by Atoms and Crystals. I. A New Theoretical Approach. *Acta Cryst.*, 10:609 – 619, 1957. `https://doi.org/10.1107/S0365110X57002194`.

[12] J. D'Aurizio. Upper bound for the sine integral. Mathematics Stack Exchange. `https://math.stackexchange.com/q/2956604` (version: 2018-10-15).

[13] M. Op de Beeck, D. Van Dyck, and W. Coene. Wave function reconstruction in HRTEM: the parabola method. *Ultramicroscopy*, 64:167 – 183, 1996. `https://doi.org/10.1016/0304-3991(96)00058-7`.

[14] C. Doberstein and B. Berkels. A least-squares functional for joint exit wave reconstruction and image registration. *Inverse Problems*, 35(5), 2019. `https://doi.org/10.1088/1361-6420/ab0b04`.

[15] J. J. Duistermaat and J. A. C. Kolk. *Multidimensional Real Analysis II: Integration.* Cambridge University Press, 2004.

[16] R. F. Egerton. *Physical Principles of Electron Microscopy.* Springer, 2005. `https://doi.org/10.1007/b136495`.

[17] A. Erdélyi, editor. *Tables of Integral Transforms*, volume I. McGraw-Hill Book Company, New York, 1954.

[18] R. Erni. *Aberration-Corrected Imaging in Transmission Electron Microscopy.* Imperial College Press, 2nd edition, 2015. `https://doi.org/10.1142/p960`.

[19] R. Erni, M. D. Rossell, and P. N. H. Nakashima. Optimization of exit-plane waves restored from HRTEM through-focal series. *Ultramicroscopy*, 110:151 – 161, 2010. `https://doi.org/10.1016/j.ultramic.2009.10.015`.

[20] R. Fletcher. A New Approach to Variable Metric Algorithms. *Computer Journal*, 13:317 – 322, 1970. `https://doi.org/10.1093/comjnl/13.3.317`.

[21] R. Fletcher and C. M. Reeves. Function minimization by conjugate gradients. *The Computer Journal*, 7(2):149 – 154, 1964. `https://doi.org/10.1093/comjnl/7.2.149`.

[22] D. C.-L. Fong and M. Saunders. LSMR: An Iterative Algorithm for Sparse Least-Squares Problems. *SIAM Journal on Scientific Computing*, 33:2950 – 2971, 2011. `https://doi.org/10.1137/10079687X`.

[23] AG Rumpf, Institute for Numerical Simulation, University of Bonn. Quocmesh software library. `https://archive.ins.uni-bonn.de/numod.ins.uni-bonn.de/software/quocmesh/index9aff.html`.

[24] J. Frank. The Envelope of Electron Microscopic Transfer Functions for Partially Coherent Illumination. *Optik*, 38:519 – 536, 1973.

[25] R. W. Gerchberg and W. O. Saxton. A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik*, 35(2):237 – 246, 1972.

[26] S. Gratton, A. S. Lawless, and N. K. Nichols. Approximate Gauss-Newton methods for nonlinear least squares problems. *SIAM Journal on Optimization*, 18(1):106 – 132, 2007. `https://doi.org/10.1137/050624935`.

[27] N. Guerrini, R. Turchetta, G. Van Hoften, R. Henderson, G. McMullan, and A. R. Faruqi. A high frame rate, 16 million pixels, radiation hard CMOS sensor. *Journal of Instrumentation*, 6(3), 2011. `https://doi.org/10.1088/1748-0221/6/03/C03003`.

[28] M. Haider, H. Rose, S. Uhlemann, E. Schwan, B. Kabius, and K. Urban. A spherical-aberration-corrected 200 kV transmission electron microscope. *Ultramicroscopy*, 75:53 – 60, 1998. `https://doi.org/10.1016/S0304-3991(98)00048-5`.

[29] S. Haigh. *Super Resolution Tilt Series Exit Wave Restoration from Aberration Corrected Images*. PhD thesis, University of Oxford, 2007.

[30] G. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities*. Cambridge University Press, 2nd edition, 1988.

[31] E. W. Hobson. On the Second Mean-Value Theorem of the Integral Calculus. *Proceedings of the London Mathematical Society*, s2-7:14 – 23, 1909. `https://doi.org/10.1112/plms/s2-7.1.14`.

[32] K. Ishizuka. Contrast Transfer of Crystal Images in TEM. *Ultramicroscopy*, 5:55 – 65, 1980. `https://doi.org/10.1016/0304-3991(80)90011-X`.

[33] K. Ishizuka and B. Allman. Phase measurement of atomic resolution image using transport of intensity equation. *Journal of Electron Microscopy*, 54:191 – 197, 2005. `https://doi.org/10.1093/jmicro/dfi024`.

[34] B. Kabius, P. Hartel, M. Haider, H. Müller, S. Uhlemann, U. Loebau, J. Zach, and H. Rose. First application of Cc-corrected imaging for high-resolution and energy-filtered TEM. *Journal of Electron Microscopy*, 58:147 – 155, 2009. `https://doi.org/10.1093/jmicro/dfp021`.

[35] A. I. Kirkland, S. L. Y. Chang, and J. L. Hutchison. Atomic resolution transmission electron microscopy. In *Science of Microscopy*, chapter 1. Springer, New York, NY, 2007. `https://doi.org/10.1007/978-0-387-49762-4_1`.

[36] A. I. Kirkland and R. R. Meyer. “Indirect” High-Resolution Transmission Electron Microscopy: Aberration Measurement and Wavefunction Reconstruction. *Microscopy and Microanalysis*, 10:401 – 413, 2004. `https://doi.org/10.1017/S1431927604040437`.

[37] A. I. Kirkland, W. O. Saxton, K.-L. Chau, K. Tsuno, and M. Kawasaki. Super-resolution by aperture synthesis: tilt series reconstruction in CTEM. *Ultramicroscopy*, 57:355–374, 1995. `https://doi.org/10.1016/0304-3991(94)00191-O`.

[38] E. J. Kirkland. Improved High Resolution Image Processing of Bright Field Electron Micrographs: I. Theory. *Ultramicroscopy*, 15:151 – 172, 1984. `https://doi.org/10.1016/0304-3991(84)90037-8`.

[39] E. J. Kirkland. Improved High Resolution Image Processing of Bright Field Electron Micrographs: II. Experiment. *Ultramicroscopy*, 17:87 – 103, 1985. `https://doi.org/10.1016/0304-3991(85)90002-6`.

[40] E. J. Kirkland. High Resolution Image Processing of Electron Micrographs. *Scanning Microscopy Supplement*, 2:139 – 147, 1988.

[41] E. J. Kirkland. *Advanced Computing in Electron Microscopy*. Springer, 2nd edition, 2010. `https://doi.org/10.1007/978-1-4419-6533-2`.

[42] E. J. Kirkland, R. F. Loane, and J. Silcox. Simulation of annular dark field STEM images using a modified multislice method. *Ultramicroscopy*, 23:77 – 96, 1987. `https://doi.org/10.1016/0304-3991(87)90229-4`.

[43] E. J. Kirkland and B. M. Siegel. Nonlinear high resolution image processing of conventional transmission electron micrographs: II. Experiment. *Ultramicroscopy*, 9:65 – 74, 1982. `https://doi.org/10.1016/0304-3991(82)90229-7`.

[44] M. Knoll and E. Ruska. Das Elektronenmikroskop. *E. Z. Physik*, 78:318 – 339, 1932. `https://doi.org/10.1007/BF01342199`.

[45] C. Kübel and A. Thust. TrueImage: A software package for focal-series reconstruction in HRTEM. In T. E. Weirich, J. L. Lábár, and X. Zou, editors, *Electron Crystallography*, pages 373 – 392. Springer Netherlands, 2006. `https://doi.org/10.1007/1-4020-3920-4_23`.

[46] P. K. F. Kuhfittig. *Introduction to the Laplace Transform*. Springer, Boston, MA, 1978. `https://doi.org/10.1007/978-1-4899-2201-4`.

[47] H. Lichte. Optimum focus for taking electron holograms. *Ultramicroscopy*, 38:13 – 22, 1991. `https://doi.org/10.1016/0304-3991(91)90105-F`.

[48] G. McMullan, A. R. Faruqi, and R. Henderson. Direct electron detectors. *Methods in Enzymology*, 579:1 – 17, 2016. `https://doi.org/10.1016/bs.mie.2016.05.056`.

[49] R. R. Meyer and A. I. Kirkland. Characterisation of the Signal and Noise Transfer of CCD Cameras for Electron Detection. *Microscopy Research and Technique*, 49:269 – 280, 2000. `https://doi.org/10.1002/(SICI)1097-0029(20000501)49:3%3C269::AID-JEMT5%3E3.0.CO;2-B`.

[50] R. R. Meyer, A. I. Kirkland, and W. O. Saxton. A new method for the determination of the wave aberration function for high resolution TEM: 1. Measurement of the symmetric aberrations. *Ultramicroscopy*, 92:89 – 109, 2002. `https://doi.org/10.1016/S0304-3991(02)00071-2`.

[51] D. L. Misell. An examination of an iterative method for the solution of the phase problem in optics and electron optics: I. Test calculations. *Journal of Physics D: Applied Physics*, 6:2200 – 2216, 1973. `https://doi.org/10.1088/0022-3727/6/18/305`.

[52] J. Modersitzki. *Numerical Methods for Image Registration.* Oxford University Press, 2004. `https://doi.org/10.1093/acprof:oso/9780198528418.001.0001`.

[53] J. Nocedal and S. J. Wright. *Numerical Optimization.* Springer, 2nd edition, 2006. `https://doi.org/10.1007/978-0-387-40065-5`.

[54] C. Ophus and T. Ewalds. Guidelines for quantitative reconstruction of complex exit waves in HRTEM. *Ultramicroscopy*, 113:88 – 95, 2012. `https://doi.org/10.1016/j.ultramic.2011.10.016`.

[55] C. Ophus, H. I. Rasool, M. Linck, A. Zettl, and J. Ciston. Automatic software correction of residual aberrations in reconstructed HRTEM exit waves of crystalline samples. *Advanced Structural and Chemical Imaging*, 2:15 – 24, 2016. `https://doi.org/10.1186/s40679-016-0030-1`.

[56] M. Otten and W. Coene. High-resolution imaging on a field emission TEM. *Ultramicroscopy*, 48:77 – 91, 1993. `https://doi.org/10.1016/0304-3991(93)90173-U`.

[57] E. Polak and G. Ribiere. Note sur la convergence de méthodes de directions conjuguées. *ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique*, 3(R1):35 – 43, 1969.

[58] B. S. Reddy and B. N. Chatterij. An FFT-Based Technique for Translation, Rotation, and Scale-Invariant Image Registration. *IEEE Transactions on Image Processing*, 5(8):1266–1271, 1996. `https://doi.org/10.1109/83.506761`.

[59] W. O. Saxton. What is the focus variation method? Is it new? Is it direct? *Ultramicroscopy*, 55:171 – 181, 1994. `https://doi.org/10.1016/0304-3991(94)90168-6`.

[60] O. Scherzer. The Theoretical Resolution Limit of the Electron Microscope. *Journal of Applied Physics*, 20:20 – 29, 1949. `https://doi.org/10.1063/1.1698233`.

[61] P. Schiske. Image reconstruction by means of focus series. *Journal of Microscopy*, 207:154, 2002. (Reprint of the original german article from 1968.) `https://doi.org/10.1046/j.1365-2818.2002.01042.x`.

[62] D. J. Smith. The realization of atomic resolution with the electron microscope. *Rep. Prog. Phys.*, 60:1513 – 1580, 1997. `https://doi.org/10.1088/0034-4885/60/12/002`.

[63] J. C. H. Spence. *High-Resolution Electron Microscopy.* Oxford University Press, 4th edition, 2013. `https://doi.org/10.1093/acprof:oso/9780199668632.001.0001`.

[64] A. Thust. Focal-Series Reconstruction in HRTEM: Fundamentals and Applications. In *Probing the Nanoworld.* Forschungszentrum Jülich, 2007.

[65] A. Thust. Focal-Series Reconstruction. In *Transmission Electron Microscopy*, chapter 9. Springer, 2016. `https://doi.org/10.1007/978-3-319-26651-0`.

[66] A. Thust, W. Coene, M. Op de Beeck, and D. Van Dyck. Focal-series reconstruction in HRTEM: simulation studies on non-periodic objects. *Ultramicroscopy*, 64:211 – 230, 1996. `https://doi.org/10.1016/0304-3991(96)00011-3`.

[67] A. Thust, M. H. F. Overwijk, W. M. J. Coene, and M. Lentzen. Numerical correction of lens aberrations in phase-retrieval HRTEM. *Ultramicroscopy*, 64:249 – 264, 1996. `https://doi.org/10.1016/0304-3991(96)00022-8`.

[68] S. Uhlemann and M. Haider. Residual wave aberrations in the first spherical aberration corrected transmission electron microscope. *Ultramicroscopy*, 72:109 – 119, 1998. `https://doi.org/10.1016/S0304-3991(97)00102-2`.

[69] M. van Heel, M. Schatz, and E. Orlova. Correlation functions revisited. *Ultramicroscopy*, 46:307 – 316, 1992. `https://doi.org/10.1016/0304-3991(92)90021-B`.

[70] R. H. Wade and J. Frank. Electron microscope transfer functions for partially coherent axial illumination and chromatic defocus spread. *Optik*, 49:81 – 92, 1977.