Rank-based optimization techniques for estimation problems in optics

Doelman, Reinier

DOI
10.4233/uuid:9148465a-0855-4cb1-a5da-de8d23dabe81

Publication date
2019

RANK-BASED OPTIMIZATION TECHNIQUES FOR ESTIMATION PROBLEMS IN OPTICS

RANK-BASED OPTIMIZATION TECHNIQUES FOR ESTIMATION PROBLEMS IN OPTICS

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. dr. ir. T.H.J.J. van der Hagen,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op donderdag 5 september 2019 om 15:00 uur

door

Reinier DOELMAN

Bachelor of Laws, Universiteit Leiden, Nederland
Master of Science in Systems and Control, Technische Universiteit Delft, Nederland
geboren te Alphen aan den Rijn, Nederland.

Dit proefschrift is goedgekeurd door de promotor.

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof. dr. ir. M. Verhaegen,	Technische Universiteit Delft, promotor
Onafhankelijke leden:	
Prof. dr. W. M. J. M. Coene	Technische Universiteit Delft
Prof. dr. ir. N. J. Doelman	Universiteit Leiden
Prof. dr. ir. B. Jayawardhana	Rijksuniversiteit Groningen
Prof. dr. ir. J. Schoukens	Vrije Universiteit Brussel
Dr. ir. T. Keviczky	Technische Universiteit Delft
Dr. Y. Shechtman	Technion Israel Institute of Technology

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement No. 339681.

Keywords: Optimization, phase retrieval, identification, blind deconvolution, distributed control, primary mirror control

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

Het spannendste moment van iets weten is vlak voordat je het weet.

Micha Wertheim, 'Voor Je Het Weet'

CONTENTS

ACKNOWLEDGEMENTS

I would like to take this opportunity to thank a number of people that have helped and supported me during the course of my time in Delft as a PhD student.

First and foremost I would like to thank my promotor, Prof. Michel Verhaegen, for giving me the opportunity to do doctoral research, for constructively challenging my ideas, giving feedback and for helping me to develop as a scientist.

My thanks also go out to the colleagues I have worked with during the project, whether that resulted in a joint publication or not. Chengpu, Baptiste, Stojan, Pieter, Thao, Oleg, Elisabeth, Dean, Hai, Hans, Laurens, Paolo and everybody else that worked in the Numerics for Control & Identification group; I am grateful for working with you, your advice and for sharing your ideas and insights.

It was my pleasure to supervise and work with several Master students. Sander, Sjoerd and Wiegert, I am thankful for your hard work, I am proud of the results, and I wish you all the best in your careers.

Furthermore, I would like to thank my co-authors, specifically those that I have not yet named above: Måns Klingspor, Anders Hansson, Johan Löfberg and Renaud Bastaits, for their work and their comments and feedback.

At this point I would also like to acknowledge the people I saw most each day, the ones I shared my office and all kinds of stories and jokes with: Baptiste, Dieky, Sjoerd and Sebastiaan. It was a joy. Kitty, Kiran, Heleen and Marieke, thank you for making things happen when they needed to happen, I appreciate the work you have done.

Of course, there is a long list of people that made the last four years in Delft incredibly fun, whether it was through sports, games of table footbal, coffee breaks and corresponding (sometimes serious, most of the time anything but) discussions and all sorts of other things beside work. Besides people already named before, I'd like to mention here Sachin, Subu, Yasin, Farid, Cees, Niloofar, Nikos, Nico, Jeroen, Joris, Yu, Yiming, Bart, Le, Renshi, Kim, Edwin, and Shuai.

Last but not least a big thank you to my family. Thank you for your immeasurable support.

Reinier Doelman
January 2019, Delft

SUMMARY

Aberrations in optical systems, such as telescopes and microscopes, degrade the quality of the images that can be produced by these systems. For example, an object that is positioned out of focus produces a blurred image on a camera sensor and the turbulent air in the earth's atmosphere reduces the imaging performance of telescopes. In this thesis we only consider wavefront aberrations.

AO can be used to compensate for these wavefront aberrations. The working principle of AO is to quantify by measuring or estimation the wavefront aberration and to dynamically adjust wavefront modulating devices, such as Deformable Mirrors (DMs), to counteract the aberration and thereby improving the optical performance.

The estimation of the wavefront aberration based on images of a point source is called phase retrieval, which is a highly nonlinear estimation problem. The success of the estimation usually depends on the (type of) algorithm, the available information on the aberration that is incorporated in the estimate, and the degree to which the model of the optical system corresponds to reality.

In this thesis we propose a convex optimization-based method for phase retrieval. The method allows for easy inclusion of many types of prior information on the aberration. Furthermore, we develop an efficient implementation of the optimization. The robustness of the approach against measurement noise is investigated and compared with several other state of the art algorithms. Experimental validation shows the algorithm is well able to estimate aberrations in real-life circumstances.

A new type of prior information is introduced to estimate dynamic wavefront aberrations. In literature and in practice, the optical path is split between either a wavefront sensor and a camera, or between multiple cameras in order to reliable estimate an aberration. The inherent problem is that between the sensor and cameras the aberration can differ (Non-Common Path (NCP) errors), and a wrong estimate is used in the compensation by the AO system. We propose a method to estimate the aberration from measurements by a single camera, by assuming that the aberration evolves according to (non-specific) model, i.e. the dynamics are contained in a model-set. At the same time that we estimate the aberration, we also identify the dynamics according to which the aberration evolves over time.

The estimation of the wavefront aberration based on images of an unknown object is called blind deconvolution if both the aberration and object are estimated. Like phase retrieval, this too is a highly nonlinear estimation problem. We propose the first convex-optimization based estimation method for blind deconvolution problems that estimate aberration and object when the images are acquired using coherent illumination. The method allows for the inclusion of many existing types of prior information on the object and/or aberration.

Finally, we analyze controllers for segmented mirrors in large ground-based telescopes. These mirrors consist of many interconnected hexagonal segments. This dis-

tributed nature of the system warrants the investigation into whether the controller that keeps the segments aligned can be designed in such a way that it can be distributed over the segments as well, essentially resulting in a distributed controller where local controllers communicate with each other. What complicates the analysis is that the dynamics across segments are not necessarily decoupled: the wind load can be correlated and the flexibility in the supporting structure of the segments can cause dynamic coupling. We investigate the design of a distributed controller that incorporates these global dynamics. Furthermore, we investigate the performance of the distributed controller and how it relates to the communication and interconnection pattern of the local controllers.

SAMENVATTING

Aberraties in optische systemen, zoals telescopen en microscopen, verslechteren de kwaliteit van de afbeelding die door deze systemen geproduceerd kunnen worden. Bijvoorbeeld, een object dat niet in het brandpunt is gepositioneerd resulteert in een onscherpe afbeelding en de turbulente lucht in de atmosfeer van de aarde beinvloed op een negatieve manier de capaciteit van telescopen om goede afbeelding te maken. In deze dissertatie kijken alleen naar zogenoemde golffrontaberraties.

Adaptieve Optica (Adaptive Optics (AO)) kan gebruikt worden om deze golffrontaberraties te compenseren. Het principe achter AO is dat de golffrontaberratie gekwantificeerd wordt door deze te meten of te schatten en op een dynamische manier golffrontmodulerende instrumenten, zoals vervormbare spiegels (Deformable Mirrors (DMs)), aan te sturen om de aberratie (verstoring) tegen te werken en op deze manier de prestaties van het optische systeem te verbeteren.

Het schatten van de golffrontaberratie op basis van afbeeldingen van een puntbron heet phase retrieval (faseherwinning), en dit is een sterk niet-lineair schattingsprobleem. Het slagen van het schatten wordt doorgaans bepaald door het (type) algoritme, de beschikbare informatie over de aberratie die meegenomen wordt bij het schatten, en de mate waarin het model van het optische systeem overeenkomt met de werkelijkheid.

In deze dissertatie introduceren we een methode gebaseerd op convexe optimalisatie voor het faseherwinningsprobleem. De methode maakt het eenvoudig om een groot aantal verschillende types voorkennis over de aberratie mee te nemen in de schatting. Daarnaast ontwikkelen we een efficiënte implementatie voor de optimalisatie. The robuustheid van de aanpak ten opzichte van meetruis wordt onderzocht en vergeleken met verschillende andere algoritmes die het nieuwste van het nieuwste zijn. Experimentele validatie laat zien dat het algoritme goed in staat is om aberraties te schatten in reële omstandigheden.

We introduceren een nieuw type van voorkennis om een tijd-variërende golffrontaberratie te schatten. Zowel in de literatuur als in de praktijk wordt het optische pad in tweeën gesplitst tussen een golffrontsensor en een camera, of tussen verschillende camera's, om zodoende nauwkeurig de aberratie te schatten. Het inherente probleem is dat tussen de sensor en camera's de aberratie kan verschillen (Non-Common Path (NCP) fouten), en dus een verkeerde schatting wordt gebruikt voor compensatie in het adaptieve optica-systeem. Wij stellen een methode voor om de aberratie te schatten op basis van metingen van een enkele camera, door aan te nemen dat de aberratie over tijd verandert volgens een (niet-specifiek) model, dus dat de dynamica zich in een bepaalde modelset bevinden. Tijdens het schatten van de aberatie schatten we tegelijkertijd de dynamica van het systeem, die beschrijft hoe de aberratie over tijd verandert.

Het schatten van de golffrontaberratie, gebaseerd op afbeeldingen van een onbekend object, heet blinde deconvolutie (engels: blind deconvolution), als zowel de aberratie als het onbekende object worden geschat. Net als faseherwinning is dit ook een sterk

niet-lineair schattingsprobleem. We introduceren de eerste op convexe optimalisatie gebaseerde schattingsmethode voor blinde deconvolutieproblemen die de aberratie en het object schatten wanneer de afbeelding genomen zijn door middel van coherente belichting. De methode staat toe dat vele verschillende bestaande types van voorkennis over het object of aberratie worden meegenomen in de schatting.

Als laatste analyseren we regelaars voor gesegmenteerde spiegels in grote telescopen die zich op aarde bevinden. Deze spiegels bestaan uit vele samengebonden hexagonale segmenten. Dit gedistribueerde aspect van het systeem vraagt om onderzoek naar de vraag of de regelaar die de segmenten richt, op zo een manier kan worden ontworpen, dat deze zelf ook gedistribueerd kan worden over de segmenten, en op deze manier een gedistribueerde regelaar oplevert waarbij lokale regelaars met elkaar communiceren. Wat de analyse lastig maakt is dat de dynamica van de segmenten niet noodzakelijkerwijs ontkoppeld zijn: de wind die tegen de verschillende segmenten blaast kan gecorreleerd zijn en de flexibiliteit van de ondersteunende constructie kan de dynamica van de segmenten koppelen. Wij onderzoeken een ontwerpmethode voor een gedistribueerde regelaar die al deze gekoppelde dynamica meeneemt. Bovendien onderzoeken we de prestaties van de gedistribueerde regelaar en hoe dit gerelateerd is aan de communicatie en het samenschakelingspatroon van de lokale regelaars.

1

INTRODUCTION

1.1. OUTLINE OF THE INTRODUCTION

This introduction will cover some of the essential concepts used in this thesis. In Section 1.2 we discuss optical systems and the pupil function that characterizes them. Furthermore, in this section we introduce the estimation problems in which the pupil function is estimated based on images taken by a camera. Section 1.3 deals with sensors and actuators in Adaptive Optics, through which the performance of an optical system can be improved. Sections 1.4 and 1.5 discuss the existing algorithms relating to the estimation problems introduced in Section 1.2.

1.2. LINEAR SYSTEMS THEORY FOR OPTICAL SYSTEMS AND RELATED INVERSE PROBLEMS

The systems description of Fourier optics in this section follows the line of definitions and assumptions as outlined in [1].

Consider Figure 1.1. Suppose we have an optical system consisting of a collection of lenses and mirrors, with an entrance pupil and an exit pupil, and suppose that the passage of light between these two planes can be adequately described by geometrical optics. This system has an object plane in front (along the direction of the rays of light) of the entrance pupil, and an image plane behind the exit pupil, perpendicular to the optical axis.

Here we consider the case of monochromatic coherent illumination. The complex amplitude distribution of an object in the object plane is denoted by

$$\mathbf{g}_o(p) \in \mathbb{C}, \tag{1.1}$$

where p are the coordinates in the object plane. The complex amplitude distribution of an image in the image plane is denoted by

$$\mathbf{g}_i(u) \in \mathbb{C}, \tag{1.2}$$

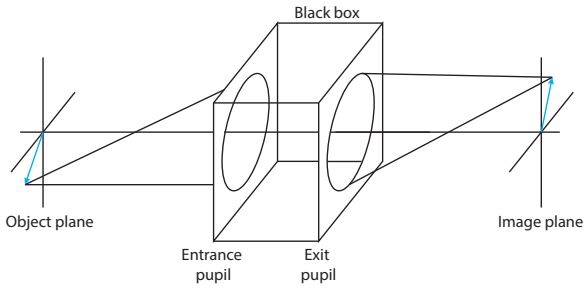


Figure 1.1: A schematic depiction of a generalized optical system including the object plane, entrance and exit pupil, and the image plane. Adapted from [1, Figure 6.1.].

where u are the (two dimensional, reduced[1]) coordinates for the image plane. We assume that the Fraunhofer approximation holds, with the consequence that the amplitude distribution in the image plane is related by a Fourier transform to the amplitude distribution in the exit pupil (or just 'the pupil').

1.2.1. THE POINT SPREAD FUNCTION

In case the amplitude distribution in the object plane is a point source, it can be shown [1, Ch. 4] that a circular aperture of a diffraction-limited system[2] produces a field in the image plane at a distance z that is given by

$$\mathbf{g}_i(r) = \exp\left(jkz\right)\exp\left(j\frac{kr^2}{2z}\right)\frac{\pi w^2}{j\lambda z}\left(2\frac{J_1(kwr/z)}{kwr/z}\right), \tag{1.3}$$

where $k = \frac{2\pi}{\lambda}$ is the wavenumber, r is the radial coordinate, w is the radius of the aperture and J_1 is a Bessel function of the first kind. The intensity distribution of this imaging system is called the Airy pattern,

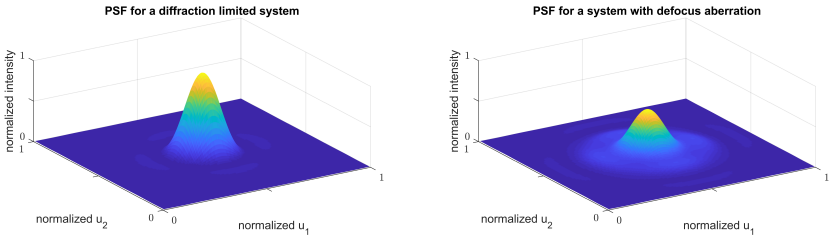
$$\mathbf{i}(r) = \left|\mathbf{g}_i(r)\right|^2 = \left(\frac{\pi w^2}{\lambda z}\right)^2\left(2\frac{J_1(kwr/z)}{kwr/z}\right)^2, \tag{1.4}$$

where $\mathbf{i}$ denotes the intensity in the image plane, see Figure 1.2a.

More generally, for non-diffraction limited systems as well, we call the complex amplitude distribution in the image plane resulting from a point source in the object plane, the amplitude impulse response, denoted $\mathbf{h}$. The intensity $|\mathbf{h}|^2$ is called the intensity impulse response, denoted $\mathbf{s} = |\mathbf{h}|^2$, although this quantity is better known as the Point Spread Function (PSF). As will be discussed in Section 1.2.4, $\mathbf{h}$ and $\mathbf{s}$ are related to the formation of images of amplitude distributions in the object plane that are not point sources. For this reason $\mathbf{h}$ is sometimes called the coherent point spread function, although to prevent confusion with $\mathbf{s}$, we make minimal use of this term in this thesis.

[1] The use of reduced coordinates allows us to disregard magnification and image inversion, and use the same coordinates as in the object plane.

[2] A diffraction-limited system "converts a diverging spherical wave incident on the entrance pupil into a converging spherical wave at the exit pupil." [1, p. 129].



(a) Diffraction limited system.

(b) System with a defocus aberration

Figure 1.2: An example Airy pattern, the point spread function for a diffraction limited optical system on the left, and the PSF for a system with a defocus aberration on the right.

1.2.2. THE GENERALIZED PUPIL FUNCTION

Let $\mathbf{A}(x)$ denote the support function of the aperture,

$$\mathbf{A}(x) = \begin{cases} 1 & \text{if } x \text{ is inside the aperture,} \\ 0 & \text{if } x \text{ is outside the aperture,} \end{cases} \tag{1.5}$$

where x is the position vector in the aperture.

In systems with low numerical aperture, and where the paraxial approximation holds,[3] the aberration (the optical path difference) of the wavefront of light emanating from a point source is related by a scaling to the phase aberration ϕ in the pupil. The complex amplitude in the pupil $\mathbf{A}(x)$ is extended with a complex term to incorporate this phase aberration:

$$\mathscr{P}(x) = \mathbf{A}(x) \exp\left(j\phi(x)\right). \tag{1.6}$$

$\mathscr{P}(x)$ is referred to as the Generalized Pupil Function (GPF).

The relation between the GPF and the amplitude impulse response is a two-dimensional spatial Fourier transform:

$$\mathscr{P}(x) = \mathbf{H}(x) := \mathscr{F}\{\mathbf{h}(u)\}. \tag{1.7}$$

Since $\mathbf{h}$ is the amplitude impulse response, $\mathbf{H}$ is called the amplitude transfer function.

RADIAL BASIS FUNCTIONS

In relation to the GPF, two sets of basis functions are of importance in this thesis. The first set is that of (Gaussian) radial basis functions [2, 3]. A (real-valued) radial basis function G_i is defined as

$$G_i(x) = \mathbf{A}(x) \exp\left(-\mu_i \langle x - x_i, x - x_i \rangle\right), \tag{1.8}$$

where the subscript i is an index for the basis functions, $\mu_i \in \mathbb{R}_+$ determines the spread of the function, and x_i are the two-dimensional coordinates of the center of the basis function. In [2] the approximation of the GPF as the following finite sum has been studied:

$$\mathscr{P}(x) \approx \tilde{\mathscr{P}}(x) = \sum_{i=1}^{n_a} a_i G_i(x), \tag{1.9}$$

[3] This means that several assumptions [1, Section 5.1.2.] relating to small angles hold, which simplify the related mathematics.

Noll index	m	n	name
1	0	0	piston
2	1	1	tip
3	-1	1	tilt
4	0	2	defocus

Table 1.1: Some common names for Zernike polynomials in (1.11) and (1.12) for a few different indices n and m. The names are related to the optical aberration they produce.

with coefficients $a_i \in \mathbb{C}$ and n_a the number of basis functions.

ZERNIKE BASIS FUNCTIONS

The second set of importance are the Zernike polynomials [4]. These basis functions are used to approximate the phase $\phi(x)$ in a circular pupil. Instead of rectangular coordinates x, the functions are usually expressed in polar coordinates (r,θ), where r is the radius and θ the angle. Define the radial polynomial

$$R_n^m(r) = \begin{cases} \sum_{k=0}^{\frac{n-m}{2}} \frac{(-1)^k (n-k)!}{k!(\frac{n+m}{2}-k)!(\frac{n-m}{2}-k)!} r^{n-2k} & m-n \text{ even}, \\ 0 & m-n \text{ odd}. \end{cases} \tag{1.10}$$

The even Zernike polynomials are defined as

$$Z_n^m(r,\theta) = R_n^m \cos(m\theta) \tag{1.11}$$

and the odd ones as

$$Z_n^{-m}(r,\theta) = R_n^m \sin(m\theta), \tag{1.12}$$

where $n \geq m$ and $0 \leq r \leq 1$. It is common to renumber the Zernike polynomials $Z_n^m \to Z_i$ to a sequential index using for example Noll's index [4] or The Optical Society (OSA) standard index [5]. The phase is then approximated as

$$\phi(r,\theta) \approx \tilde{\phi}(r,\theta) = \sum_{i=1}^{n_z} c_i Z_i(r,\theta), \tag{1.13}$$

where $c_i \in \mathbb{R}$ are the Zernike coefficients. A number of Zernike basis functions have conventional names that are related to the aberrations they produce. Some are listed in Table 1.1.

1.2.3. THE PHASE RETRIEVAL PROBLEM

Typically, optical detectors like cameras can only measure the intensity (or amplitude) of the incident light,

$$\mathbf{y}(u) = \mathbf{i}(u) = \left|\mathbf{g}_i(u)\right|^2, \tag{1.14}$$

where $\mathbf{y}$ denotes the measurement – which for now is assumed noise-free.[4] In the optical setting with a point source, the measurement is that of the PSF,

$$\mathbf{y}(u) = \mathbf{s}(u) = |\mathbf{h}(u)|^2 = \left|\mathscr{F}^{-1}\{\mathscr{P}(x)\}\right|^2. \tag{1.15}$$

[4] The symbol $\mathbf{y}$ is used for any measurement; depending on the setting, a hypothetical noise-free one or a noisy one.

The phase retrieval problem is to estimate $\mathscr{P}(x)$ from the measurements $\mathbf{y}(u)$ and other information that is available a priori. For the analysis of the performance of an optical system, sometimes this problem reduces to estimating only the phase aberration in the GPF when the amplitude is known, i.e. estimating $\phi(x) = \angle\mathscr{P}(x)$.

The phase retrieval problem can be vectorized in the following way. Denote with $v = \text{vect}(V)$ the vector obtained from vertically stacking the columns of a matrix V. Assume that the Generalized Pupil Functions has an unknown apodisation and is sampled on an $m \times m$ grid,[5] so that if

$$\mathbf{a} := \text{vect}(\mathscr{P}(x)) \in \mathbb{C}^{m^2}, \tag{1.16}$$

we obtain for the measurement

$$\mathbf{y} = \left|F^{-1}\mathbf{a}\right|^2, \tag{1.17}$$

where $F \in \mathbb{C}^{m^2 \times m^2}$ is the transformation matrix that gives the vectorized two-dimensional Discrete Fourier Transform (DFT) of the vectorized Generalized Pupil Functions, and where $|\cdot|^2$ is meant element-wise. The phase retrieval problem is thus formulated as

$$\begin{aligned} &\text{find} && \mathbf{a} \\ &\text{subject to} && \mathbf{y} = \left|F^{-1}\mathbf{a}\right|^2 \\ &&& \mathbf{a} \in \mathscr{M}, \end{aligned} \tag{1.18}$$

where $\mathscr{M}$ is some set describing the prior information available on $\mathbf{a}$.

Solutions to (1.18) are in general not unique, see [7] for a short discussion. For example, if $\mathbf{a}$ is a feasible solution, then so is $\mathbf{a}\exp\left(j\phi_0\right)$ for any phase shift (piston mode) ϕ_0. In the context of Adaptive Optics (AO, see Section 1.3), this particular ambiguity is of minor concern, since it does not effect the formed image. Only under specific conditions the use of specific algorithms guarantees that - save the trivial ambiguities - the correct solution to (1.18) can be found, see for example [7] for a recent overview and the references in that overview paper. However, in practice good results can be achieved with a variety of algorithms, see also Section 1.4.

Apart from the particular algorithm chosen, a key ingredient for good practical results is the availability and accuracy of prior information [8]. An often used type of prior information is phase diversity, which we will discuss in the following section. In Section 1.2.3 we discuss several other types of prior information.

INCORPORATION OF PHASE DIVERSITY INTO THE PHASE RETRIEVAL PROBLEM

Phase diversity refers to an additive known phase distortion $\phi_D(x)$ in the pupil plane [9, 10]. The GPF that includes the diversity is given by

$$\mathscr{P}_D(x) = \mathbf{A}(x)\exp\left(j\left(\phi(x) + \phi_D(x)\right)\right) = \mathscr{P}(x)\exp\left(j\phi_D(x)\right). \tag{1.19}$$

A measurement of the PSF in the image plane will give a diversity image,

$$\mathbf{y}_D = \left|\mathscr{F}^{-1}\{\mathscr{P}_D\}\right|^2 = \left|\mathscr{F}^{-1}\left\{\mathbf{A}(x)\exp\left(j\left(\phi(x) + \phi_D(x)\right)\right)\right\}\right|^2. \tag{1.20}$$

[5] A modal approximation is also occasionally used [3, 6].

The purpose of using multiple images, each with a different, known phase diversity, is to improve the performance of phase retrieval algorithms, in terms of convergence, ambiguity and robustness to measurement noise. Why it leads to better results, seems to be an open problem. The diversity is typically introduced by means of a deformable mirror (Section 1.3.3) or another phase-modulating device. To introduce a defocus as a phase diversity, it is also possible to move the camera along the optical axis. The micro-lens array in a Shack-Hartmann sensor (Section 1.3.2) that focuses light in the pupil plane onto a camera can also be modeled as a particular phase diversity [11].

The phase diversity can be incorporated into the phase retrieval problem as described in the previous section in the following way. Denote with $W = \mathrm{d}(w)$ the matrix with elements of the vector w on its diagonal. The diversity is expressed in matrix form by

$$\begin{aligned} \mathbf{p}_d &= \mathrm{vect}\left(\exp\left(j\phi_d\right)\right) \in \mathbb{C}^{m^2}, \\ D_d &= \mathrm{d}\left(\mathbf{p}_d\right) \in \mathbb{C}^{m^2 \times m^2}. \end{aligned} \tag{1.21}$$

The diversity image(s) can be expressed now as

$$\mathbf{y}_D = \left|F^{-1} D_d \mathbf{a}\right|^2. \tag{1.22}$$

If n_d is the number of images obtained by applying phase diversities, then the matrices $FD_{d,i}$ for $i = 1, \ldots, n_d$, can be stacked to obtain the matrix U,

$$U = \begin{pmatrix} U_1 \\ \vdots \\ U_{n_d} \end{pmatrix} = \begin{pmatrix} F^{-1} D_{d,1} \\ \vdots \\ F^{-1} D_{d,n_d} \end{pmatrix} \tag{1.23}$$

and the corresponding measurement vectors $\mathbf{y}_{D,i}$ can be stacked in the same manner to obtain the measurement vector $\mathbf{y}$.

The phase retrieval problem - for the case of static aberrations - from images of the PSF can be stated as

$$\begin{aligned} &\text{find} && \mathbf{a} \\ &\text{subject to} && \mathbf{y} = |U\mathbf{a}|^2 \\ & && \mathbf{a} \in \mathscr{M}. \end{aligned} \tag{1.24}$$

OTHER TYPES OF PRIOR INFORMATION USED IN PHASE RETRIEVAL

Magnitude of the GPF The well-known Gerchberg-Saxton (GS) algorithm [12] uses the a priori knowledge of the magnitude of the Generalized Pupil Functions $\mathbf{z}$, i.e. the constraint

$$|\mathbf{a}|^2 = \mathbf{z}. \tag{1.25}$$

Support and positivity constraints Fienup [13] extended the GS algorithm by incorporating support, realness and positivity constraints on $\mathbf{a}$.

Smoothness In [14] phase retrieval is applied to the problem of Coherent Diffraction Imaging (CDI), where the measured intensities are the squared amplitudes of the Fourier transform of an image, which constitutes the pupil. The measurement noise affects the estimates of the PSF in the image plane. The Oversampling Smoothness (OSS) algorithm proposed in [14] operates on the assumption that the spatial frequencies as measured in the image plane, vary smoothly over frequency. A spatial filter is used to incorporate this knowledge.

Phase diversity and other linear constraints on the phase aberration In [15] linear equality constraints on the phase are used to decrease the number of parameters to estimate. Phase diversity is a particular example of a linear constraint. To illustrate this, consider the following example. If there are two images,

$$\mathbf{y}_1 = \left|F^{-1}\mathbf{a}_1\right|^2, \quad \mathbf{y}_2 = \left|F^{-1}\mathbf{a}_2\right|^2 \tag{1.26}$$

and the phase in the pupil plane differs by a known amount, i.e. the diversity, we have the optimization problem

$$\begin{aligned} &\text{find} && \mathbf{a}_1, \mathbf{a}_2 \\ &\text{subject to} && \mathbf{y}_1 = \left|F^{-1}\mathbf{a}_1\right|^2, \mathbf{y}_2 = \left|F^{-1}\mathbf{a}_2\right|^2. \\ & && \angle\mathbf{a}_1 - \angle\mathbf{a}_2 = \text{vect}\left(\phi_d\right) \\ & && |\mathbf{a}_1| = |\mathbf{a}_2|, \end{aligned} \tag{1.27}$$

which, by introducing the proper variables, can be recasted into (1.24).

The use of phase diversity to obtain good estimates is a widely used and researched technique [16–25].

Total-Variation Let ∇_i be the matrix computing the discrete spatial gradient at pixel i in the pupil plane. The discrete total variation regularization [26] is expressed through the term

$$\sum_i \|\nabla_i \mathbf{a}_i\|_1 . \tag{1.28}$$

Regularization with this term produces estimates with reduced oscillations in the spatial directions [27–29].

Sparsity The sparsity of a vector $\mathbf{a}$ refers to the case where many of the elements of this vector equal 0. The number of non-zero elements of a vector $\mathbf{a}$ is called the cardinality, denoted with $\|\mathbf{a}\|_0$. This norm is non-convex, and the number of non-zero elements is not always known a prior. Often the 1-norm, $\|\mathbf{a}\|_1$, is used as a convex alternative to induce sparsity in the solution [30]. Relating to phase retrieval, this prior information is used in, among others, [31–38]

Small-phase approximation For small phase aberrations, the highly non-linear measurement model can be approximated with a linear or quadratic model in the phase aberration. This approximation can simplify and speed up the phase retrieval, as demonstrated in for example [39–47].

1

Known temporal model In [48] it is proposed to use the assumption that the temporal evolution of the phase aberration is approximately constant, i.e.

$$\phi_{k+1} = \phi_k, \tag{1.29}$$

where the subscript k denotes the time index, and that therefore the known difference (phase diversity) between two successive aberrations is the correction applied by the Adaptive Optics system.

1.2.4. IMAGE FORMATION

We make a distinction between two types of imaging, coherent imaging and incoherent imaging, because the type of illumination, coherent or incoherent, determines how the recorded image is formed. In both imaging cases the relation between input and output can be described as a (spatial) linear system [1, Section 6.2], although for each illumination case in a different way.

COHERENT ILLUMINATION

In case of coherent illumination, the complex amplitude distribution in the object plane $\mathbf{g}_o$ is related to the complex amplitude distribution in the image plane $\mathbf{g}_i$ by a convolution with the amplitude impulse response $\mathbf{h}$:

$$\mathbf{g}_i = \mathbf{g}_o \star \mathbf{h}, \tag{1.30}$$

where $\star$ denotes the convolution. Taking the (spatial, 2 dimensional) Fourier transform on both sides, we obtain the linear relation between the Fourier transform of $\mathbf{g}_o$ and that of $\mathbf{g}_i$ through the amplitude transfer function $\mathbf{H}$,

$$\mathbf{G}_i = \mathbf{G}_o \mathbf{H}, \tag{1.31}$$

where $\mathbf{G}_i = \mathscr{F}\{\mathbf{g}_i\}$ and $\mathbf{G}_o = \mathscr{F}\{\mathbf{g}_o\}$. The image as recorded by the camera in the image plane is

$$\mathbf{i} = |\mathbf{g}_i|^2 = |\mathbf{g}_o \star \mathbf{h}|^2. \tag{1.32}$$

Coherent imaging plays a role in for example Coherent Diffraction Imaging, ptychography, long range horizontal imaging [49], or the imaging of metamaterials [50, 51].

INCOHERENT ILLUMINATION

In case of incoherent illumination, the intensity in the object plane $\mathbf{f} = |\mathbf{g}_o|^2$ is related to the intensity in the image plane $\mathbf{i} = |\mathbf{g}_i|^2$ by a convolution with the Point Spread Function $\mathbf{s} = |\mathbf{h}|^2$:

$$\mathbf{i} = \mathbf{f} \star \mathbf{s}. \tag{1.33}$$

Taking the (spatial, 2 dimensional) Fourier transform on both sides, we obtain the linear relation between the Fourier transform of $\mathbf{f}$ and that of $\mathbf{i}$ through the Optical Transfer Function (OTF) $\mathbf{S}$,

$$\mathbf{I} = \mathbf{F}\mathbf{S}, \tag{1.34}$$

where $\mathbf{I} = \mathscr{F}\{\mathbf{i}\}$, $\mathbf{F} = \mathscr{F}\{\mathbf{f}\}$ and $\mathbf{S} = \mathscr{F}\{\mathbf{s}\}$.

1.2.5. THE BLIND DECONVOLUTION PROBLEM

Where the phase retrieval problem was the estimation of the amplitude impulse response or GPF from intensity measurements of the image of a point source, the (spatial) blind deconvolution problem in the imaging context adds to this the estimation of the object plane complex amplitude or object plane intensity. In an imaging setting, blind deconvolution can be used for post-processing acquired images to increase the image quality of the final estimates of an object. Another use for blind deconvolution is the estimation of phase aberrations that distort the formed image on the camera, with the purpose of compensating for these aberrations. The blind deconvolution problem in the imaging context is a problem that in different contexts and applications is interpreted differently. The first difference is the imaging case that is considered. That is, an image formed with coherent, or with incoherent light.

For the coherent case this gives the problem

$$\begin{aligned} &\text{find} && \mathbf{h},\ \mathbf{g}_o \\ &\text{subject to} && \mathbf{i_c} = \left|\mathbf{g}_o \star \mathbf{h}\right|^2 \\ & && \mathbf{g}_o \in \mathcal{M}_{\mathbf{g}_o} \\ & && \mathbf{h} \in \mathcal{M}_{\mathbf{h}}, \end{aligned} \tag{1.35}$$

where $\mathcal{M}_{\mathbf{g}_o}$ and $\mathcal{M}_{\mathbf{h}}$ are sets describing the prior information on the variables.

With incoherent illumination the blind deconvolution problem is

$$\begin{aligned} &\text{find} && \mathbf{s},\ \mathbf{f} \\ &\text{subject to} && \mathbf{i_i} = \mathbf{f} \star \mathbf{s} \\ & && \mathbf{s} \geq 0,\ \mathbf{f} \geq 0 \\ & && \mathbf{f} \in \mathcal{M}_{\mathbf{f}} \\ & && \mathbf{s} \in \mathcal{M}_{\mathbf{s}}, \end{aligned} \tag{1.36}$$

and this is the deconvolution problem that is most often encountered in literature, for the reason that most real-world images are taken in this setting.

Even though in (1.36) the Point Spread Function $\mathbf{s}$ can be abstracted to any blurring function, in astronomy-related literature, the (incoherent) imaging case is often discussed in the context of phase aberrations. The blind deconvolution problem is then [11, 52, 53]

$$\begin{aligned} &\text{find} && \phi,\ \mathbf{h},\ \mathbf{f} \\ &\text{subject to} && \mathbf{i_i} = \mathbf{f} \star |\mathbf{h}|^2 \\ & && \mathbf{f} \geq 0 \\ & && \mathbf{h} = \mathscr{F}^{-1}\left\{\mathbf{A}\exp\left(j\phi\right)\right\} \\ & && \mathbf{f} \in \mathcal{M}_{\mathbf{f}} \\ & && \phi \in \mathcal{M}_{\phi}. \end{aligned} \tag{1.37}$$

The object $\mathbf{f} = \left|\mathbf{g}_o\right|^2 \in \mathbb{R}_+^{m\times m}$ is directly estimated (not $\mathbf{g}_o$), being real and positive. The PSF is constrained to represent a phase aberration. This formulation directly extends to the blind deconvolution problem that includes phase diversity images as in Section 1.2.3,

1

but even though the objects **f** are identical in the images, the relation between the different Point Spread Functions is highly nonlinear.

PRIOR INFORMATION USED IN BLIND DECONVOLUTION

Many of the types of prior information used in the phase retrieval literature, see Section 1.2.3, have been applied to the blind deconvolution problem as well. See also [54] for a short overview of regularization applied to blind deconvolution.

Phase diversity If we view the literature from the point of imaging with possible phase aberrations according to (1.37), then what is not always taken into account is the fact that if the blurring is caused by a phase aberration, and we have multiple images of the same object with a different phase diversity (multi-frame blind deconvolution), that there is additional information on the relation between the several intensity impulse response functions $\mathbf{s}_i$. In [55] problem (1.37) this issue is addressed using an GS-like iterative transform algorithm. In [21, 52, 53, 56, 57] the phase aberration is estimated using diversity images in a Bayesian setting and with a gradient descent algorithm, or from a least squared error perspective [58]. [59] uses a parameter search.

Smoothness Prior information on the shape (smoothness) of **s**, based on astronomical data, is used in [60]. [61] assumes piece-wise smoothness.

Sparsity and Total-Variation In order to recover sharp edges in an object, Total-Variation regularization [62, 63] can be applied, potentially combined with sparsity [63].

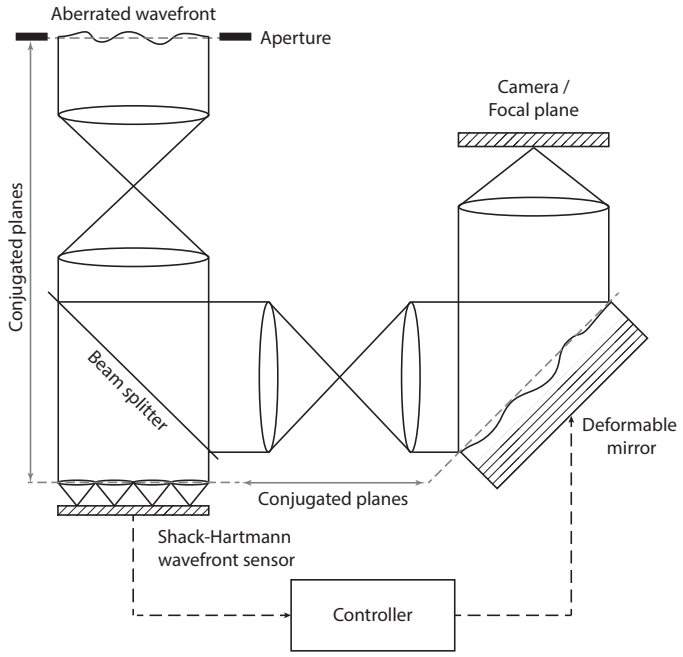
Linear constraints [64] considers the case in astronomy where multiple objects are imaged, with known (linear) relations between the phase (multi-frame blind deconvolution). For the use of translation diversity [24, 25], an object is placed at the aperture, and their relative position is changed to obtain images for parts of the object. To combine these images, a blind deconvolution problem can be formulated where the overlapping parts of the to be estimated object are used as (linear) constraints in the optimization.

1.3. ADAPTIVE OPTICS

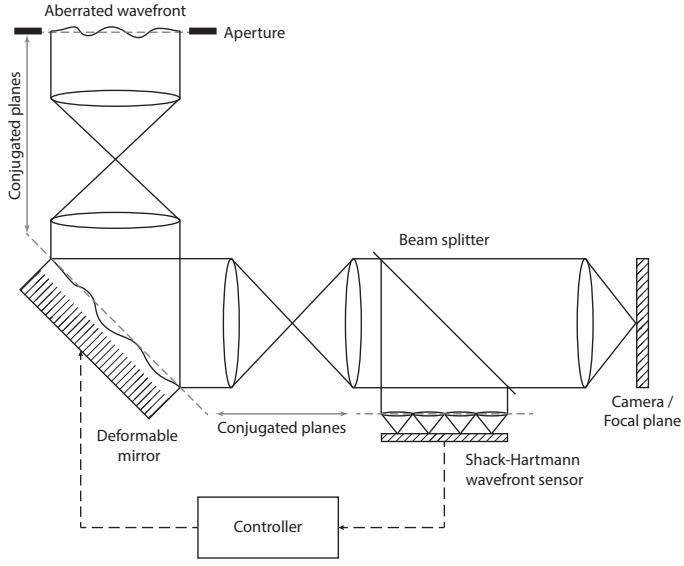
Given an aberrated phase ϕ^{AB} in the pupil plane, Adaptive Optics (AO) can be used to compensate for this aberration and improve the imaging performance of the system [65–67]. Typically use is made of sensors, that give a measurement (or estimate) of the phase, and wavefront-modulating devices such as deformable mirrors, that influence the aberrated wavefront. We discuss the typical AO setup, sensors, actuators and the typical AO control loop.

1.3.1. THE AO SETUP

In Figure 1.3 two AO setups are displayed. The first one, Figure 1.3a, shows an open-loop setup. The wavefront sensor measures the aberrated phase directly, and the deformable mirror compensates the measured aberration. The second setup, Figure 1.3b is a closed-loop AO setup. The wavefront sensor measures the difference between the aberrated



(a) Open loop.



(b) Closed loop.

Figure 1.3: Two different AO setups.

phase and the compensated phase. The actuator commands to the mirror are updated to compensate for this residual on top of the previous aberration correction, thereby

1

regulating the residual to zero.

1.3.2. SENSORS

PUPIL PLANE SENSORS

Pupil plane wavefront sensors attempt to measure the phase of the complex amplitude distribution in a plane optically conjugated to the pupil plane, see Figure 1.3. One such instrument is the Shack-Hartmann (SH) wavefront sensor [68]. The SH wavefront sensor consists of a grid of lenslets, that each focus a part of the aperture onto a Charge-Coupled Device (CCD) or Complementary Metal Oxide Semiconductor (CMOS) camera. If the phase aberration as seen by the SH sensor ϕ^{SH} is zero, each section of the camera's pixels corresponding to a lenslet will have a focused spot in its center, see Figure 1.4. However, if $\phi^{SH} \neq 0$, then the location of the geometric mean of pixel values in the sub-aperture is indicative of the first order spatial gradient of the phase. If we have n_s^2 sub-apertures arranged in a square grid, then we have the relation

$$\mathbf{y}^{SH} = G\phi^{SH}, \tag{1.38}$$

where $\mathbf{y}^{SH} \in \mathbb{R}^{2n_s^2}$ are the measured spatial derivatives (here assumed noise-free) and G is the measurement matrix, whose size depends on the geometry of the sensor. The phase aberration ϕ^{SH} as seen by the Shack-Hartmann sensor can be estimated from the measured spatial derivatives or from the measured intensity distributions. However, some phase aberrations cannot be measured (those in the null space of G) [68–71].

The linear relation between phase and measurement (1.38) enables efficient and numerically stable modelling methods for the phase aberration temporal dynamics like subspace identification [72, 73], or modelling with Vector Auto-Regressive (VAR) models.

The phase aberration as seen by the Shack-Hartmann sensor is on a different optical path than the phase aberration that reaches the camera, see Figure 1.5. The differences between these two phase aberration are called non-common path errors (NCP errors). These NCP errors are encountered in for example astronomy [18–20] and ophthalmic imaging [74–76].

FOCAL PLANE SENSORS

To overcome the control problem arising from non-common path errors, the measurements from the focal plane camera can be used to generate a control signal. Where in the case of a point source the relation between measurements and wavefront error are linear for a Shack-Hartmann sensor (1.38), the relation between the measurements in the focal plane and the wavefront error are highly non-linear, see (1.15) and (1.20).[6]

In [18, 19] this technique is used for calibration and removal of static aberrations in the optical system of the Very Large Telescope (VLT). In [77] the technique is used for coronagraphic imaging. [53] uses focal plane sensors to estimate aberration and object jointly. A comparison in [6] shows experimental results for different phase retrieval techniques that estimate the phase aberration from the focal plane camera.

[6] The comparison is not entirely on an equal footing here: the SH sensor produces an estimate of the aberration that is defined by a resolution determined by the number of lenslets, whereas a focal plane sensor obtains measurements determined by the number of pixels.

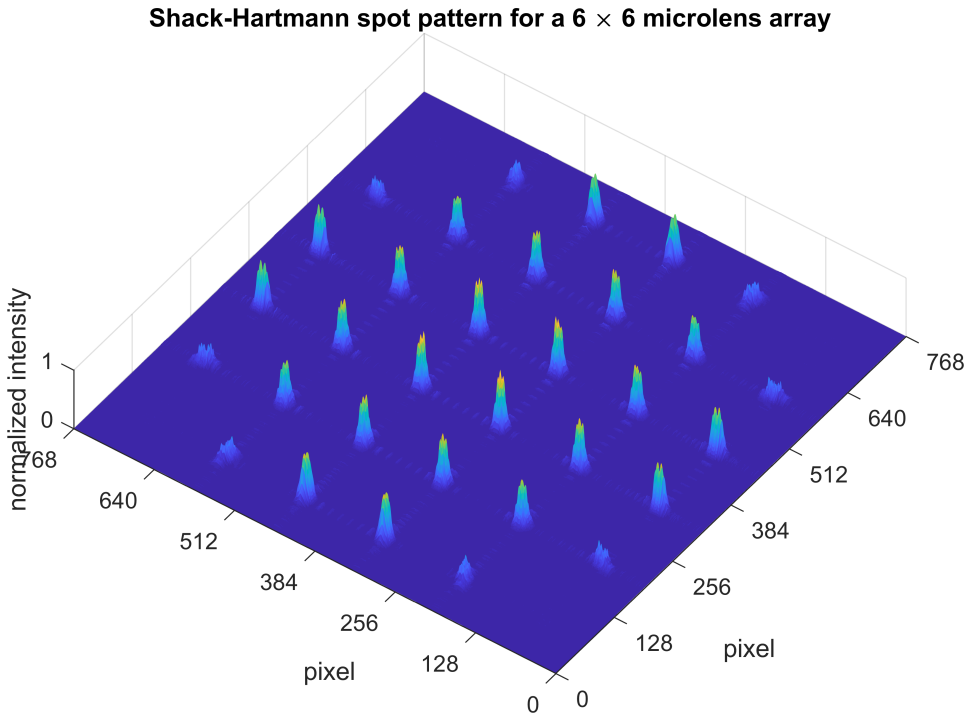


Figure 1.4: The intensity pattern ('spots') in the detector plane of a Shack-Hartmann wavefront sensor generated by a flat wavefront.

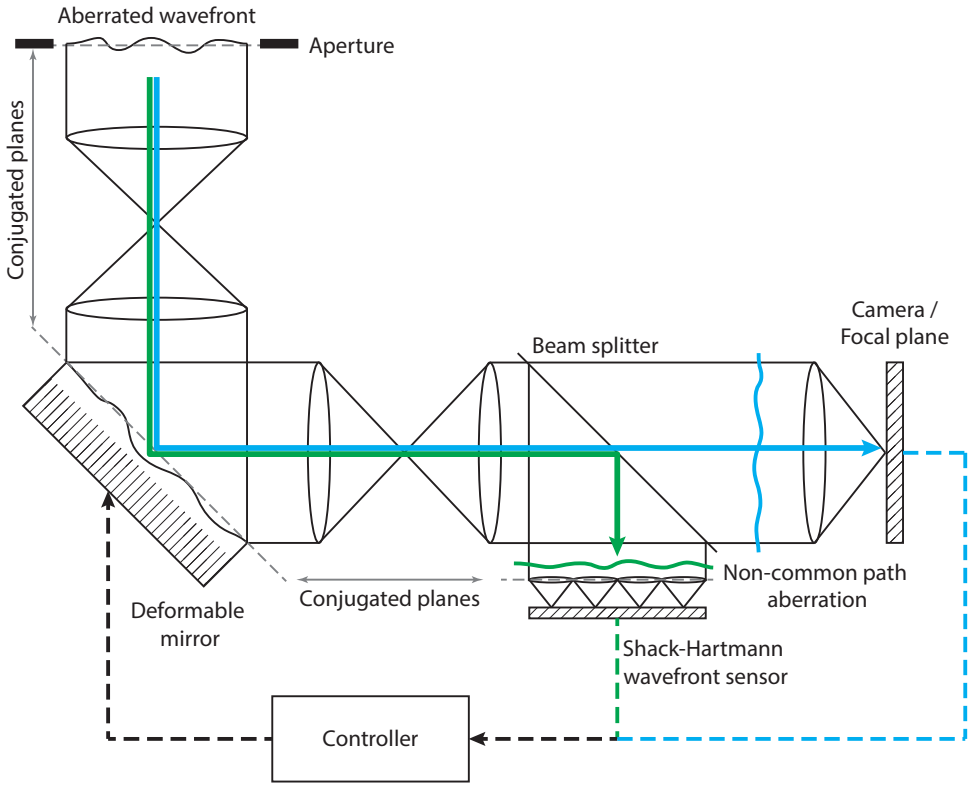


Figure 1.5: The classical AO setup follows the aberration correction along the green path. The additional aberration marked in green is corrected by the mirror, but not seen by the focal plane camera. The additional aberration in blue is seen by the camera, but not corrected by the mirror. A controller that uses the focal plane camera and the blue path to compute the correction by the mirror does not suffer from this problem.

These methods have in common that they employ the use of phase diversity to obtain the correct estimate. An important difference between these methods is that [53] employs two cameras (on different optical paths) to generate the diversity image, whereas [6, 18, 19, 77] use the deformable mirror in the AO system to introduce the diversity and obtain the diversity images sequentially. The first method essentially gives the same non-common path issue as with the SH sensor. The second method needs to assume that the aberration is static.

Using a single focal plane sensor with the purpose of compensating dynamic aberrations is a topic that is seemingly not well researched, both for identification and (optimal) control purposes.

1.3.3. ACTUATORS

In this thesis we only consider deformable mirrors as wavefront modulating devices. We treat segmented mirrors of large astronomical telescopes and (small) deformable mirrors for optical bench setups in separate sections below, due to their size difference and related control issues, even though they conceptually can be put under the same umbrella.

DEFORMABLE MIRRORS

Deformable mirrors modulate the phase in the pupil plane through the introduced phase ϕ^{DM}. In membrane mirrors a reflective membrane changes its shape, and therefore the introduced phase, under the influence of actuator signals $u \in \mathbb{R}^{n_u}$, where n_u is the number of actuators of the mirror. This relationship is typically modeled using the mirror's influence matrix $H \in \mathbb{R}^{n^2 \times n_u}$,[7] where we assume that $\phi^{DM} \in \mathbb{R}^{n^2}$ is sampled on a grid of size $n \times n$,

$$\phi^{DM} = Hu. \tag{1.39}$$

The matrix H can be obtained experimentally by 'poking' the actuators. That is, by actuating the system using $u = \mathbf{e}_i^{n_u}$ for $i = 1, \ldots, n_u$, where $\mathbf{e}_i^N$ is a unit vector of length N that equals 1 on index i and 0 on the other indices. The measured phases then constitute the columns of H. Another way to characterize the mirror is through its effect on the measurements of the Shack-Hartmann sensor in a closed-loop setup,

$$\mathbf{y}^{SH} = \tilde{H}u. \tag{1.40}$$

where $\tilde{H} \in \mathbb{R}^{n_s^2 \times n_u}$.

SEGMENTED MIRRORS

The use of large mirrors in astronomical telescopes brings with it the problem that fabrication of these large mirrors with the required precision is either currently not possible or prohibitively expensive. The solution for this problem is to construct mirrors that consist of segments. For example the Keck telescope, the Thirty Meter Telescope (TMT) and the European Extremely Large Telescope (E-ELT) have or will have segmented primary mirrors. All three consist of grids of hexagonal mirror segments, each of which has

[7] We assume the sampling time of the system is such, that the mirror can be regarded to behave in a static manner and not in a dynamic one.

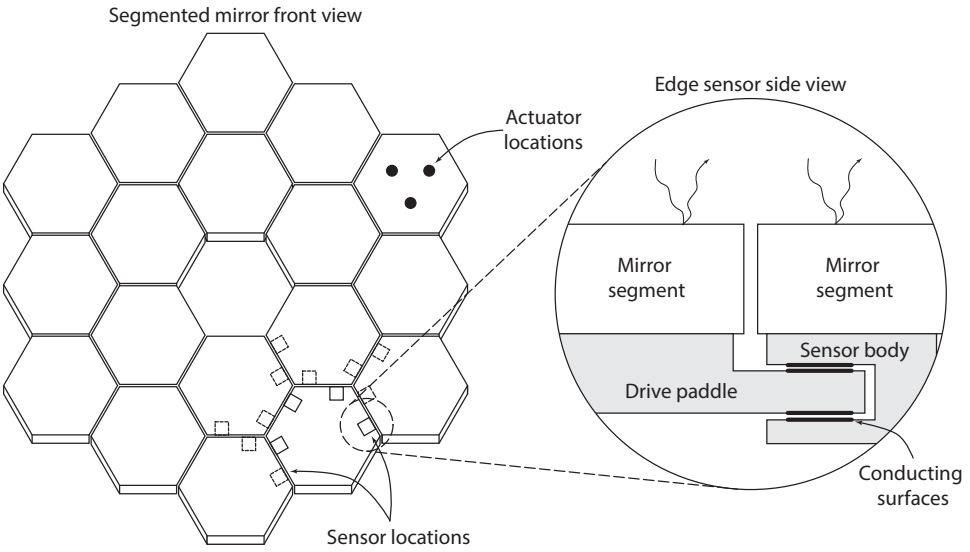


Figure 1.6: An illustration of a two-ring segmented mirror with 18 hexagonal elements, capacitance-based edge sensors measuring relative displacement and actuators on the back of the segments controlling piston, tip and tilt. Adapted from [79, 80].

edge sensors to measure the relative displacement with respect to neighbouring segments and three actuators to control the segment's position. The segmented mirror can be considered to introduce a phase ϕ^{SM} into the system that is linearly related to the (overall) mirror shape [78].

If a static relation is assumed between the actuator displacements u_a and the edge sensor readings $\mathbf{y}^e$, the relation between the two [81, 82] is

$$\mathbf{y}^e = Ju_a, \tag{1.41}$$

where J is the influence matrix.

The segments are mounted on a supporting structure. If the supporting structure causes the segments to dynamically interact (and (1.41) does not hold), we can describe the system (disregarding disturbances) using a dynamic model [82],

$$\begin{aligned} M\ddot{x} + V\dot{x} + Kx &= Bu, \\ \mathbf{y}^e &= G_e x, \end{aligned} \tag{1.42}$$

where M is the mass matrix, V the damping matrix and K contains the spring constants. B is the input matrix, x the system state and u the vector of actuator displacements. The matrix G_e describes the relation between system states and measurements.

1.3.4. THE CONTROL LOOP

This description of the control loop follows [83, 84] and is a description of the temporal modelling of the system as displayed Figure 1.3b.

The Shack-Hartmann measurement that is available to the controller at time k, denoted as $\mathbf{y}_k^{SH}$, is dependent on the phase aberration at the sensor at time $k-1$,

$$\mathbf{y}_k^{SH} = G\phi_{k-1}^{SH} + w_k, \tag{1.43}$$

where G is the measurement matrix, $w_k \sim \mathcal{N}(0, \Sigma_w)$ is the Gaussian measurement noise with zero mean and covariance Σ_w at time k and $\phi^{SH} \in \mathbb{R}^{n^2}$ is the (vectorized) phase aberration at the sensor. The phase introduced by the deformable mirror at time k depends on the input signals as sent to the mirror at time $k-1$,

$$\phi_k^{DM} = Hu_{k-1}. \tag{1.44}$$

This means that the estimate $\hat{\mathbf{y}}_{k|k-1}$ (i.e. the estimate for time k, dependent on information available on time $k-1$) is given by

$$\hat{\mathbf{y}}_{k|k-1}^{SH} = G\hat{\phi}_{k|k-1}^{AB} - GHu_{k-2}. \tag{1.45}$$

MODELING OF AN AO SYSTEM WITH A DYNAMIC PHASE ABERRATION

For simplicity, assume we have a dynamic model of the aberration available in the form of a Vector Auto-Regressive model of order 2 (VAR(2)):

$$\phi_{k+1}^{AB} = A_1\phi_k^{AB} + A_2\phi_{k-1}^{AB} + v_k, \tag{1.46}$$

where $v_k \sim \mathcal{N}(0, \Sigma_v)$ is white Gaussian noise driving the time evolution of the aberration and $A_1,\ A_2 \in \mathbb{R}^{n^2 \times n^2}$ are coefficient matrices.[8] Define the state vector

$$x_k = \begin{pmatrix} \phi_k^{AB\,T} & \phi_{k-1}^{AB\,T} & u_{k-1}^T & u_{k-2}^T \end{pmatrix}^T. \tag{1.47}$$

In closed loop we have the relation

$$\phi_k^{SH} = \phi_k^{AB} - \phi_k^{DM}, \tag{1.48}$$

and so the state-space system becomes [83, 84]

$$\begin{aligned} x_{k+1} &= Ax_k + B_u u_k + B_v v_k \\ \mathbf{y}_k^{SH} &= Cx_k + D_w w_k \end{aligned} \tag{1.49}$$

where

$$A = \begin{pmatrix} A_1 & A_2 & 0 & 0 \\ I & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & I & 0 \end{pmatrix},\ B_u = \begin{pmatrix} 0 \\ 0 \\ H \\ 0 \end{pmatrix},\ B_v = \begin{pmatrix} I \\ 0 \\ 0 \\ 0 \end{pmatrix},\ C = G\begin{pmatrix} 0 & I & 0 & -H \end{pmatrix},\ D_w = I. \tag{1.50}$$

Other models for the time evolution dynamics of the phase aberration are for example (see [84]) a simple random walk process (integrator), $\phi_{k+1}^{AB} = \phi_k^{AB} + v_k$, identified models [72, 85–88], or models based on first-principles modeling [83, 84].

[8] Alternatively, the time evolution of a (small) limited set of Zernike coefficients (Section 1.2.2) can be used as a temporal model, to reduce the size of the coefficient matrices A_1 and A_2

1

Using Linear Quadratic Gaussian (LQG) control, the phase can be optimally compensated by the mirror. If $\hat{\phi}^{AB}_{k+1|k}$ is the optimal estimate of the phase aberration at the next time step, then the LQG control input is

$$u_k = H^+ \hat{\phi}^{AB}_{k+1|k}, \tag{1.51}$$

where H^+ is the pseudo-inverse of H [84].

ADVANTAGES AND DISADVANTAGES OF THE CLASSICAL AO SETUP

The use of a classical AO setup has several advantages:

- The measurement of the phase derivatives with a SH sensor and the construction of the phase estimate can be done quickly.
- The determination of the appropriate actuator commands is, in the simplest case, a matrix-vector multiplication [89]. In the dynamic case, as demonstrated in the previous subsection, optimal control can be applied to compensate the aberration.
- The use of a Shack-Hartmann sensor with a large grid of subapertures can give an accurate reconstruction of the phase [71].

There are also some inherent downsides to the classical setup:

- The light that enters the system is divided by the beam splitter over two optical paths. If the light source is weak, this will negatively affect the signal-to-noise ratio at the camera.
- The Shack-Hartmann sensor is used to estimate the phase aberration as it reaches the SH sensor, but this might be different from the aberration that reaches the camera. This is encountered in for example astronomy [19, 20, 22, 23, 58, 90] and opthalmology [76].

As described in Section 1.3.2, the use of focal plane sensors, extracting the aberration from the focal plane camera itself, can mitigate these problem at the cost of an increased computational complexity. For an analysis of the computational problems and existing solutions, we review in the next two section phase retrieval algorithms (extracting the phase aberration from an image of a point source) and blind deconvolution algorithms (extracting the phase aberration from an image of an extended object).

1.4. AN OVERVIEW OF PHASE RETRIEVAL ALGORITHMS

Due to the importance of the phase retrieval problem, there exist a number of review articles that give an overview of the different algorithms that have been developed to solve the phase retrieval problem [7, 8, 35, 91, 92]. We treat here two main categories. The first category is that of iterative transform (or iterative projection) methods. These were the first to be developed and are widely used in practice. The second category is that of convex optimization based algorithms, because they are relevant to the methods proposed in this thesis. After we set out these two categories, we briefly touch on other approaches.

1.4.1. ITERATIVE, FOURIER TRANSFORM-BASED METHODS

Many of the most popular methods for general phase retrieval fall into the category of iterative transform methods, whose most well known members are the Gerchberg-Saxton (GS) algorithm [12] and Fienup's Hybrid Input-Output (HIO) algorithm [13, 93].

In [8] a very clear overview is given to show the difference between many of the popular algorithms in this class. Therefore we will follow the presentation in [8] to briefly set out the main algorithms. To simplify the presentation we assume there is one image and no diversity. Furthermore, $\mathbf{a}$ and $\mathbf{h}$ are vectorized. $\mathbf{a}(x)$ is defined on a support set $\mathscr{D}$. The corresponding problem is

$$\begin{aligned} &\text{find} && \mathbf{a} \\ &\text{subject to} && \mathbf{y} = \left|F^{-1}\mathbf{a}\right|^2. \end{aligned} \tag{1.52}$$

The algorithms make use of projection and reflection operations, which is why this class of algorithms is also often denoted as that of alternating projection algorithms.

Let P_C denote the projection operator of a signal onto a set C. The reflector R_C is defined as

$$R_C = 2P_C - I, \tag{1.53}$$

where I is the identity map,

$$x' = I(x) = x. \tag{1.54}$$

The magnitude constraint projection $P_{\mathbf{y}}$ is the projection

$$P_{\mathbf{y}}(\mathbf{a}) = F\mathbf{h}, \text{ where } \mathbf{h} = \begin{cases} \sqrt{\mathbf{y}}\frac{F^{-1}\mathbf{a}}{|F^{-1}\mathbf{a}|} & \text{if } F^{-1}\mathbf{a} \neq 0, \\ \sqrt{\mathbf{y}} & \text{otherwise.} \end{cases} \tag{1.55}$$

The support constraint projection $P_{\mathscr{D}}$ is the projection

$$P_{\mathscr{D}}(\mathbf{a}(x)) = \begin{cases} \mathbf{a}(x) & \text{if } x \in \mathscr{D}, \\ 0 & \text{otherwise.} \end{cases} \tag{1.56}$$

If $\mathbf{a}(x)$ is assumed real and nonnegative, and nonnegativity constraints are also incorporated, the projection is

$$P_+(\mathbf{a}(x)) = \begin{cases} \max(0, \mathbf{a}(x)) & \text{if } x \in \mathscr{D}, \\ 0 & \text{otherwise.} \end{cases} \tag{1.57}$$

Fienup's HIO algorithm [13] consists of the updates

$$\mathbf{a}_{k+1} = \begin{cases} P_{\mathbf{y}}(\mathbf{a}_k(x)) & \text{if } x \in \mathscr{D} \text{ and } P_{\mathbf{y}}(\mathbf{a}_k(x)) \geq 0 \\ \mathbf{a}_k(x) - \beta_k P_{\mathbf{y}}(\mathbf{a}_k(x)) & \text{otherwise.} \end{cases} \tag{1.58}$$

Here β_k is a tuning parameter. In [94] it is shown that with only support constraints, (1.58) is equivalent to

$$\mathbf{a}_{k+1} = \frac{1}{2}\left(R_{\mathscr{D}}(R_{\mathbf{y}} + (\beta_k - 1)P_{\mathbf{y}}) + I + (1 - \beta_k)P_{\mathbf{y}}\right)(\mathbf{a}_k), \tag{1.59}$$

The Difference Map algorithm [95] consists of the update

$$\begin{aligned} \mathbf{a}_{k+1} &= \left(I + \beta\left(P_+P_1 - P_{\mathbf{y}}P_2\right)\right)(\mathbf{a}_k), \\ P_1 &= (1+\gamma_1)P_{\mathbf{y}} - \gamma_1 I \\ P_2 &= (1+\gamma_2)P_+ - \gamma_2 I \end{aligned} \tag{1.60}$$

The Relaxed Averaged Alternating Reflections (RAAR) algorithm [96] consists of the update

$$\mathbf{a}_{k+1} = \left(\frac{1}{2}\beta(R_+R_{\mathbf{y}} + I) + (1-\beta)P_{\mathbf{y}}\right)(\mathbf{a}_k) \tag{1.61}$$

The Gerchberg-Saxton algorithm [12] considers a slightly different case relating to (1.52), because it assumes the presence of the additional constraint

$$\mathbf{z} = |\mathbf{a}|^2, \tag{1.62}$$

but not any support or nonnegativity constraints. If $P_{\mathbf{z}}$ is the projection

$$P_{\mathbf{z}}(\mathbf{a}) = \begin{cases} \sqrt{\mathbf{z}}\frac{\mathbf{a}}{|\mathbf{a}|} & \text{if } \mathbf{a} \neq 0, \\ \sqrt{\mathbf{z}} & \text{otherwise.} \end{cases}, \tag{1.63}$$

then the GS algorithm is the update sequence

$$\mathbf{a}_{k+1} = (P_{\mathbf{z}}P_{\mathbf{y}})(\mathbf{a}_k). \tag{1.64}$$

For iterative transform methods a global convergence result does not exist, only local convergence results [97]. Advantages are that the Fourier transforms can be efficiently carried out using the Fast Fourier Transform (FFT) algorithm, and projections such as (1.55) can be applied with element-wise operations.

1.4.2. CONVEX OPTIMIZATION-BASED RETRIEVAL METHODS

A different category of phase retrieval algorithms are based on convex relaxations of (1.18). These relaxation-based methods have enjoyed great attention in the last few years; the overview articles [7, 98] compare a number of approaches.

Notice that (1.18), when described in row-by-row fashion, reads

$$\mathbf{y}_{[i]} = \left|U_{[i,:]}\mathbf{a}\right|^2, \; i = 1, \ldots, n_y. \tag{1.65}$$

The square brackets indicate the element index and the colon indicates all elements along that dimension. (1.65) can be reformulated as follows:

$$\begin{aligned} \mathbf{y}_{[i]} &= \left|U_{[i,:]}\mathbf{a}\right|^2 \\ &= (U_{[i,:]}\mathbf{a})^H(U_{[i,:]}\mathbf{a}) \\ &= \text{trace}\left((\mathbf{a}^H U_{[i,:]}^H)(U_{[i,:]}\mathbf{a})\right) \\ &= \text{trace}\left(U_{[i,:]}^H U_{[i,:]}\mathbf{a}\mathbf{a}^H\right) \\ &= \text{trace}\left(\mathbf{U}_i\mathbf{A}\right), \end{aligned} \tag{1.66}$$

where $\cdot^H$ is the Hermitian transpose, $\mathbf{U}_i = U^H_{[i,:]} U_{[i,:]}$, $\mathbf{A} = \mathbf{a}\mathbf{a}^H \in \mathbb{C}^{n_a \times n_a}$, and n_a are the number of coefficients in $\mathbf{a}$. The principle of defining the symmetric positive definite rank-1 matrix $\mathbf{A}$ and viewing the quadratic measurements as linear measurements in a higher dimension, is called 'lifting' and is the basis of the well-known PhaseLift algorithm [99].

The exact reformulation of problem (1.18) is then

$$\begin{aligned} &\text{find} && \mathbf{A} \\ &\text{subject to} && \mathbf{y}_{[i]} = \text{trace}\left(\mathbf{U}_i \mathbf{A}\right),\ i = 1, \ldots, n_y, \\ & && \text{rank}\left(\mathbf{A}\right) = 1 \\ & && \mathbf{A} \succeq 0 \end{aligned} \tag{1.67}$$

where n_y is the number of measurements

The PhaseLift algorithm consists of two steps. The first step is to solve the semidefinite relaxation of (1.67) to obtain $\mathbf{A}'$:

$$\begin{aligned} \mathbf{A}' \in \arg\min_{\mathbf{A}} \quad & \text{trace}\left(\mathbf{A}\right) \\ \text{subject to} \quad & \mathbf{y}_{[i]} = \text{trace}\left(\mathbf{U}_i \mathbf{A}\right),\ i = 1, \ldots, n_y, \\ & \mathbf{A} \succeq 0. \end{aligned} \tag{1.68}$$

The second step is to take the resulting optimal $\mathbf{A}'$ and find the vector $\mathbf{a}'$ by solving

$$\mathbf{a}' \in \arg\min_{\mathbf{a}} \left\| \mathbf{a}\mathbf{a}^H - \mathbf{A}' \right\|_F^2, \tag{1.69}$$

which can be computed with an eigendecomposition or singular value decomposition (SVD) of $\mathbf{A}'$.

The PhaseCut algorithm [100, 101] is based on a different convex relaxation of (1.18) and itself based on the approach in [102]. The starting point is to split the amplitude and phase components $v \in \mathbb{C}^{n_y}$, and reformulate the optimization problem to

$$\begin{aligned} \min_{\mathbf{a}, v} \quad & \left\| U\mathbf{a} - \text{d}\left(\sqrt{\mathbf{y}}\right) v \right\|_2^2 \\ \text{subject to} \quad & |v| = \mathbf{1}_{n_y} \end{aligned} \tag{1.70}$$

Given an optimal v, the optimal $\mathbf{a}$ can be found by solving the ordinary least squares problem, $\mathbf{a} = U^\dagger \, \text{d}\left(\sqrt{\mathbf{y}}\right) v$. Substituting this into (1.70), one obtains

$$\begin{aligned} \min_{\mathbf{a}, v} \quad & \left\| (UU^\dagger - I) \, \text{d}\left(\sqrt{\mathbf{y}}\right) v \right\|_2^2 \\ \text{subject to} \quad & |v| = \mathbf{1}_{n_y} \end{aligned} \tag{1.71}$$

The objective function is written into a quadratic form to obtain

$$v^H N v, \tag{1.72}$$

where $N = ((UU^\dagger - I)\,\mathrm{d}\left(\sqrt{\mathbf{y}}\right))^H((UU^\dagger - I)\,\mathrm{d}\left(\sqrt{\mathbf{y}}\right))$. Similar to PhaseLift, define $V = vv^H \in \mathbb{C}^{n_y \times n_y}$. The optimization problem then becomes

$$\begin{aligned} \min_V \quad & \operatorname{trace}(VN) \\ \text{subject to} \quad & \widehat{\mathrm{d}}(V) = \mathbf{1}_{n_y}, \\ & V \succeq 0, \\ & \operatorname{rank}(V) = 1, \end{aligned} \tag{1.73}$$

where $\widehat{\mathrm{d}}(V)$ are the values on the diagonal of V.

The convex relaxation of (1.18) for PhaseCut is the problem

$$\begin{aligned} \min_V \quad & \operatorname{trace}(VN) \\ \text{subject to} \quad & \widehat{\mathrm{d}}(V) = \mathbf{1}_{n_y}, \\ & V \succeq 0 \end{aligned} \tag{1.74}$$

The optimal v is recovered from V in a similar manner to PhaseLift.

The PhaseLift and PhaseCut algorithms share the trait that a vector in a quadratic expression is 'lifted' to a full matrix. The advantage of doing this, is that it gets rid of a product of decision variables and is therefore a crucial step towards the convex relaxation of the phase retrieval problem. The trade-off is that the lifting returns a rank constraint. This rank constraint is then dropped and traded-in for a low-rank inducing convex objective function. There are also a number of disadvantages to this lifting approach. The first is quadratic increase in the number of decision variables, which affects storage requirements [103] and algorithm runtime (the problem becomes a Semidefinite Programming problem (SDP), [104]). The second is the difficulty it brings to including prior information on the (original) decision variable into the problem, since this decision variable itself has been substituted.

In case of PhaseLift, to avoid the increase in decision variables, it is proposed in [99] to keep the matrix $\mathbf{A}$ factored into $\mathbf{A} = \mathbf{a}\mathbf{a}^H$ or a higher rank factorization, and iteratively update the factors, see also [103]. In this way, the algorithm is somewhat connected to the GS algorithm. Fienup showed in [13] how the iterative projection algorithm is related to a basic gradient descent scheme. In [105] it is shown how the factored form of PhaseLift is connected to a regularized gradient descent scheme.

The recently proposed PhaseMax convex relaxation [106–108] avoids this lifting and changes the nonconvex equality constraint into a convex inequality constraint,

$$\begin{aligned} \max_{\mathbf{a}} \quad & \mathbf{a}_{\text{init}}^T \mathbf{a} \\ \text{subject to} \quad & \sqrt{\mathbf{y}} \geq |U\mathbf{a}|, \end{aligned} \tag{1.75}$$

where $\mathbf{a}_{\text{init}}$ is an initial guess and the inequality holds element-wise. The PhaseLamp algorithm [109] is an iterative application of PhaseMax.

An important aspect of the convex methods listed here is the presence of the results in literature on recovery guarantees. That is, the circumstances under which the method recovers the correct solution with high probability. The common assumption in these

proofs is that the rows of the matrix U are random Gaussian vectors.[9] Since this assumption does not necessarily hold for the imaging case we study, we do not treat this here. Instead we refer to several articles where these algorithm properties are developed and compared [98–100, 111–114].

1.4.3. OTHER PHASE RETRIEVAL METHODS

A third common approach is to formulate a cost (or loss) function, and to minimize this cost function with a gradient-descent-based method. In phase retrieval the cost functions are usually highly nonlinear, and convergence to a global minimum usually cannot be guaranteed. With careful initialization, and under specific circumstances, some convergence can be guaranteed, as described in literature on Wirtinger flow algorithms [111, 115–118]. Given the cost function

$$f(\mathbf{a}) = \left\| \mathbf{y} - |U\mathbf{a}|^2 \right\|_2^2, \tag{1.76}$$

these algorithms make use of the Wirtinger derivative of the cost function [111, 115] with respect to the complex-valued vector $\mathbf{a}$,

$$\frac{\partial f(\mathbf{a})}{\partial \mathbf{a}} = 4 \sum_{i=1}^{n_y} (\mathbf{y}_i - \left| U_{[i,:]}\mathbf{a} \right|^2)(U_{[i,:]}^H U_{[i,:]})\mathbf{a}. \tag{1.77}$$

This derivative can be expressed in matrix-vector notation as

$$\frac{\partial f(\mathbf{a})}{\partial \mathbf{a}} = 4U^H \Psi_{n_y} \left(U\mathbf{a} \otimes I_{n_y} \right) \left(|U\mathbf{a}|^2 - y \right) \tag{1.78}$$

where Ψ_{n_y} is the selection matrix [119] such that for a matrix X of size $n_y \times n_y$ it holds that

$$\Psi_{n_y} \operatorname{vect}(X) = \widehat{\mathrm{d}}(X), \tag{1.79}$$

the vector of values on the diagonal of X, and $\otimes$ denotes the Kronecker product.

A different approach is to directly optimize the actuator input values or the aberrated phase based on an image-based metric [75, 120–122]. This technique is common in sensorless Adaptive Optics.

Recently, also neural network-based approaches have been proposed for phase retrieval [123–125].

1.5. AN OVERVIEW OF BLIND DECONVOLUTION ALGORITHMS

We make two distinctions when it comes to blind deconvolution algorithms. The first is whether the algorithm applies to the coherent or incoherent case in Section 1.2.5. The second is whether the algorithm is based on convex optimization or on non-convex optimization.

[9] In recent work the results are slightly extended [110].

1.5.1. NON-CONVEX OPTIMIZATION FOR THE INCOHERENT CASE

In the incoherent case, we have measurements formed according to

$$\mathbf{i_i} = \mathbf{s} \star \mathbf{f}, \text{ where } \mathbf{s} = |\mathbf{h}|^2, \ \mathbf{f} = \left|\mathbf{g}_o\right|^2, \tag{1.80}$$

where $\mathbf{f}$ is the object and $\mathbf{s}$ the squared amplitude of the amplitude impulse response. The Fourier transform of this expression reads

$$\mathbf{I_i} = \mathbf{SF}. \tag{1.81}$$

If $\mathbf{s}$ is known, Richardson [126] and Lucy [127] independently proposed the algorithm that is referred to as the Richardson-Lucy algorithm to deconvolve the object from the measurements with the following iteration:

$$\mathbf{f}_{k+1}(u) = \left(\frac{\mathbf{i_i}(u)}{\mathbf{f}_k(u) \star \mathbf{s}(u)} \star \mathbf{s}(-u) \right) \mathbf{f}_k = P_{\mathbf{s}}(\mathbf{f}_k(u)) := P_{\mathbf{s}}(\mathbf{f}_k). \tag{1.82}$$

If not $\mathbf{s}$ but $\mathbf{f}$ is known, the expression is similar and denoted $P_{\mathbf{f}}(\mathbf{s}_k)$. If $\mathbf{s}$ is unknown, so in the case of blind deconvolution, the iterations can simply be alternated [128],

$$\begin{aligned} \mathbf{f}_{k+1} &= P^m_{\mathbf{s}_k}(\mathbf{f}_k) = \underbrace{P_{\mathbf{s}_k} \cdots P_{\mathbf{s}_k}}_{m \text{ times}}(\mathbf{f}_k) \\ \mathbf{s}_{k+1} &= P^m_{\mathbf{f}_{k+1}}(\mathbf{s}_k) \end{aligned} \tag{1.83}$$

where m denotes the number of repeated iterations. The algorithms enjoys popularity for both deconvolution and blind deconvolution, see for example [129–132].

Other methods adapts the Gerchberg-Saxton/Fienup algorithm to the blind deconvolution problem [133–135], by iteratively applying the image plain constraints that $\mathbf{s}$ and $\mathbf{f}$ are real and positive functions, and the pupil constraint (1.80).

The blind deconvolution problem is also often addressed in a (post-processing) Bayesian framework, maximizing the Maximum Likelihood (ML) or the Maximum a posteriori (MAP) probability density. See [136] for an overview of Bayesian methods. Applications include microscopy [137] and astronomy [21, 60].

1.5.2. CONVEX OPTIMIZATION FOR THE INCOHERENT CASE

In the last few years a new method has been developed for blind deconvolution in the incoherent case, based on a convex relaxation of the problem [138, 139]. This convex relaxation approach is close in nature to the PhaseLift algorithm [99], because it relies on 'lifting' of decision variables.

In the discrete convolution case, notice that the measurement $\mathbf{i_i}$ is a summation of products of elements in $\mathbf{s}$ and $\mathbf{f}$. Let $\bar{\mathbf{i}}_\mathbf{i}, \bar{\mathbf{s}}$ and $\bar{\mathbf{f}}$ denote the vectorized variables. Then there exists a selection matrix U such that

$$\mathbf{i_i} = \mathbf{s} \star \mathbf{f} \quad \Leftrightarrow \quad \bar{\mathbf{i}}_\mathbf{i} = U \operatorname{vect}\left(\bar{\mathbf{s}}\bar{\mathbf{f}}^T\right) \tag{1.84}$$

Now let $U_{[i,:]}$ denote the i'th row of U and let $\mathbf{A} = \bar{\mathbf{s}}\bar{\mathbf{f}}^T$. The i'th row (1.84) now reads

$$\mathbf{i}_{\mathbf{i},[i]} = U_{[i,:]} \operatorname{vect}(\mathbf{A}) = \operatorname{trace}(\mathbf{U}_i \mathbf{A}), \tag{1.85}$$

where vect$\left(\mathbf{U}_i^T\right) = U_{[i,:]}^T$. It means that the blind deconvolution problem for the incoherent case can equivalently be written as

$$\begin{aligned} &\text{find} && \mathbf{A} \\ &\text{subject to} && \mathbf{i}_{\mathbf{i},[i]} = \text{trace}\left(\mathbf{U}_i\mathbf{A}\right),\ i = 1,\ldots,n_y \\ & && \text{rank}\left(\mathbf{A}\right) = 1. \end{aligned} \tag{1.86}$$

To obtain the convex relaxation of this problem, the rank constraint is dropped and the objective function becomes the nuclear norm of $\mathbf{A}$,

$$\begin{aligned} &\min_{\mathbf{A}} && \|\mathbf{A}\|_* \\ &\text{subject to} && \mathbf{i}_{\mathbf{i},[i,:]} = \text{trace}\left(\mathbf{U}_i\mathbf{A}\right),\ i = 1,\ldots,n_y. \end{aligned} \tag{1.87}$$

The nuclear norm of a matrix is defined as the sum of its singular values,

$$\|X\|_* = \sum_i \sigma_i(X). \tag{1.88}$$

The minimization of the nuclear norm of a matrix promotes low-rank solutions [140].

Based on this formulation, research has been done on how to incorporate prior information like sparsity [36, 37, 141], the demixing of sums of convolutions (idem), conditions that give recovery guarantees [138, 139, 142, 143], and techniques to improve the computational complexity [144].

1.5.3. NON-CONVEX OPTIMIZATION FOR THE COHERENT CASE

The blind deconvolution problem for the coherent case is expressed as

$$\begin{aligned} &\text{find} && \mathbf{h},\ \mathbf{g}_i,\ \mathbf{g}_o \\ &\text{subject to} && \mathbf{y_c} = \left|\mathbf{g}_i\right|^2 \\ & && \mathbf{g}_i = \mathbf{h} \star \mathbf{g}_o \end{aligned} \tag{1.89}$$

The equivalent problem in the Fourier domain is

$$\begin{aligned} &\text{find} && \mathbf{H},\ \mathbf{G}_i,\ \mathbf{G}_o \\ &\text{subject to} && \mathbf{Y_c} = \mathbf{G}_i \bigstar \mathbf{G}_i \\ & && \mathbf{G}_i = \mathbf{H}\mathbf{G}_o, \end{aligned} \tag{1.90}$$

where $\bigstar$ denotes the autocorrelation. In [24, 25] the estimation of wavefront errors in CDI is analyzed and the authors choose to minimize a weighted least squared error metric with a gradient descent scheme. [145] compares the performance of several gradient descent schemes (first and second order schemes on amplitude and intensity based error metrics) showing superior robustness to noise for amplitude based metrics. Refinement of a guessed object and wavefront aberration in a Maximum Likelihood context can be found in [146]. One important aspect in these refinement schemes is the initial guess, which [147] suggests could be provided using Machine Learning.

In [148, 149] the extended Ptychographical Iterative Engine (ePIE) is proposed, an iterative transform algorithm for ptychography. The algorithm makes use of multiple images with shifted pupils, but for the sake of simplicity, we do not discuss that here.

Let $\mathbf{G}_i = \mathbf{H}\mathbf{G}_o$, and define the functions

$$P_{\mathbf{y}}(\mathbf{G}_i) = F\mathbf{g}_i, \text{ where } \mathbf{g}_i = \begin{cases} \sqrt{\mathbf{y}}\frac{F^{-1}\mathbf{g}_i}{|F^{-1}\mathbf{g}_i|} & \text{if } F^{-1}\mathbf{g}_i \neq 0, \\ \sqrt{\mathbf{y}} & \text{otherwise.} \end{cases} \tag{1.91}$$

$$\begin{aligned} P_{\mathbf{H}}(\mathbf{G}_i) &= \frac{\mathbf{H}^H}{|\mathbf{H}|^2_{\max}}\mathbf{G}_i \\ P_{\mathbf{G}_o}(\mathbf{G}_i) &= \frac{\mathbf{G}_o^H}{|\mathbf{G}_o|^2_{\max}}\mathbf{G}_i \end{aligned} \tag{1.92}$$

The ePIE algorithm consists of the alternating updates of $\mathbf{H}$ and $\mathbf{G}_o$:

$$\begin{aligned} \mathbf{G}_{i,k+1} &= \mathbf{H}_k\mathbf{G}_{o,k} \\ \mathbf{H}_{k+1} &= \mathbf{H}_k + \alpha P_{\mathbf{G}_{o,k}}((P_{\mathbf{y_c}} - I)(\mathbf{G}_{i,k+1})) \\ &= (I + \alpha P_{\mathbf{G}_{o,k}}(P_{\mathbf{y_c}} - I)\mathbf{G}_{o,k})\mathbf{H}_k \\ \mathbf{G}_{o,k+1} &= \mathbf{G}_{o,k} + \beta P_{\mathbf{H}_k}((P_{\mathbf{y_c}} - I)(\mathbf{G}_{i,k+1})), \\ &= (I + \beta P_{\mathbf{H}_k}((P_{\mathbf{y_c}} - I)\mathbf{H}_k)\mathbf{G}_{o,k} \end{aligned} \tag{1.93}$$

for some choice of parameters α and β.

1.6. MOTIVATION AND OUTLINE OF THIS THESIS

1.6.1. MOTIVATION

The motivation for the research in this thesis comes from the gaps we identify in the current literature on phase retrieval, blind deconvolution and their relation to the system and control setting for Adaptive Optics. Convex methods have shown their success in solving the phase retrieval problem and the incoherent blind deconvolution problem, but have some limitations.

1. Not all forms of prior information are easily included due to the 'lifting' these methods typically employ.
2. The lifting of these variables creates optimization problems of a high computational complexity.
3. Among the convex methods, no convex optimization algorithm seems to exist for coherent blind deconvolution.
4. Among convex methods for either case of illumination, phase diversity is not employed as prior information in the deconvolution problem.
5. Generally, the time evolution of the phase aberration in phase retrieval is underutilized as a form of prior information.

The aim of the research in this thesis is to develop convex optimization-based methods that are able to address these issues, investigate their performance and convergence properties and where possible, verify their performance on experimental measurement data.

1.6.2. OUTLINE

This thesis is organized as follows.

COPR Chapter 2 we present a novel convex optimization-based method for the phase retrieval problem. The convex relaxation does not require 'lifting' and is affine the decision variables. This enables the easy inclusion of prior information. Since the relaxation is parameterized, we also suggest an iterative algorithm with acronym COPR (Convex Optimization-based Phase Retrieval) with some results on convergence properties for specific cases. We present two different approaches to reduce the computational burden. First, we use Gaussian Radial Basis Functions (GRBFs) to parameterize the pupil function. Secondly, we provide a tailored ADMM algorithm to solve the relaxed problem in an efficient way. The Alternating Direction Method of Multipliers (ADMM) algorithm requires per iteration only parallelized Singular Value Decompositions (SVDs) of 2×2 matrices, and a matrix vector multiplication.

This chapter is based on "Reinier Doelman, Nguyen H. Thao and Michel Verhaegen, Solving large-scale general phase retrieval problems via a sequence of convex relaxations. Journal of the Optical Society of America A 35(8), pp. 1410–1419. OSA (2018)"

A phase-evolution-dynamics-set prior In Chapter 3 we propose a convex relaxation-based method for a problem with two different interpretations. The first interpretation of the problem is a phase retrieval problem for a time series of images, where the phase changes dynamically over time. The novel idea is to regularize the phase retrieval problem with prior information that the phase change according to a dynamic model (a VAR model of specified order), but the model itself is unknown.[10] The novel constraint is a bilinear constraint, and the proposed method easily extends to other prior information that can be formulated as bilinear constraints. A final novel aspect of our method is that – in the context of phase retrieval – most images in a sequence can be taken in-focus, and phase diversity does not seem to be necessary at all time instances in our numerical simulation for successful retrieval results, thereby solving the non-common-path error problem.

The second interpretation of the problem is that of blind Wiener system identification with squared output measurements. The proposed solution for these system identification problems is to solve a convex relaxation of the problem, if necessary in an iterative manner. Compared to existing literature on blind Wiener system identification, which use Bayesian methods, our method does not seem to require a careful initialization.

This chapter is based on "Reinier Doelman, Måns Klingspor, Anders Hansson, Johan Löfberg and Michel Verhaegen, Identification of the dynamics of time-varying phase

[10] If the model were known, the constraint would be a linear constraint, see Section 1.2.3 and 1.2.5.

aberrations from time histories of the point-spread function. In preparation, 2019"

Convex optimization based blind deconvolution for images taken with coherent light Chapter 4 proposes a convex relaxation of the blind deconvolution problem for the imaging setting with coherent illumination. The application of convex relaxation for this setting seems to be novel. The method enjoys easy extension to the incoherent case, and incorporation of many different types of prior information. Compared to other convex relaxation-based algorithms for blind deconvolution, our method is novel with regard to keeping the pupil function parameterized (or amplitude transfer function), and not lifting its squared amplitude, the intensity impulse response function. This allows the application of phase diversity to the blind deconvolution problem, which is novel for convex-optimization based blind deconvolution in any setting. Similar to the COPR algorithm, the computation of the solution to the convex relaxation can be split into completely parallelized SVDs and the solutions to ordinary least squares problems.

This chapter is based on "Reinier Doelman and Michel Verhaegen, Convex optimization-based blind deconvolution for images taken with coherent illumination. Submitted to Journal of the Optical Society of America A, 2018."

Distributed wind-load compensation for segmented mirror telescopes The large, segmented primary mirrors in ground-based telescopes suffer from dynamic disturbances due to for example the wind loading and the vibrations introduced by other equipment. The difficulty in rejecting these disturbances is partly due to the dynamic coupling the segments have, introduced by the flexibility in the support structure on which the segments are mounted, and the spatial correlation in the wind loading. From a system engineering point-of-view it would be desirable to have a distributed controller, where each segment has its own controller, and the controllers communicate with each other. In Chapter 5 we ask the question how the controllers could best be interconnected to achieve the required performance of the overall system, a question that current literature on control for segmented mirrors has not answered yet.

This chapter is based on "Reinier Doelman, Sander Dominicus, Renaud Bastaits and Michel Verhaegen, Systematically structured $\mathscr{H}_2$ optimal control for truss-supported segmented mirrors. IEEE Transactions on Control Systems Technology 99, pp. 1–8. IEEE 2018."

REFERENCES

[1] J. W. Goodman, Introduction to Fourier optics. Roberts and Company Publishers, 2005.

[2] A. Martinez-Finkelshtein, D. Ramos-Lopez, and D. Iskander, "Computation of 2D Fourier transforms and diffraction integrals using Gaussian radial basis functions," Applied and Computational Harmonic Analysis, vol. 43, no. 3, pp. 424–448, 2017.

[3] P. J. Piscaer, A. Gupta, O. Soloviev, and M. Verhaegen, "Modal-based phase retrieval

using Gaussian radial basis functions," J. Opt. Soc. Am. A, vol. 35, pp. 1233–1242, Jul 2018.

[4] R. J. Noll, "Zernike polynomials and atmospheric turbulence," JOSA, vol. 66, no. 3, pp. 207–211, 1976.

[5] L. N. Thibos, R. A. Applegate, J. T. Schwiegerling, and R. Webb, "Standards for reporting the optical aberrations of eyes," Journal of refractive surgery, vol. 18, no. 5, pp. S652–S660, 2002.

[6] J. Antonello and M. Verhaegen, "Modal-based phase retrieval for adaptive optics," JOSA A, vol. 32, no. 6, pp. 1160–1170, 2015.

[7] Y. Shechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev, "Phase retrieval with application to optical imaging: a contemporary overview," IEEE signal processing magazine, vol. 32, no. 3, pp. 87–109, 2015.

[8] D. R. Luke, J. V. Burke, and R. G. Lyon, "Optical wavefront reconstruction: Theory and numerical methods," SIAM review, vol. 44, no. 2, pp. 169–224, 2002.

[9] R. A. Gonsalves, "Phase retrieval and diversity in adaptive optics," Optical Engineering, vol. 21, no. 5, p. 215829, 1982.

[10] L. M. Mugnier, A. Blanc, and J. Idier, "Phase diversity: a technique for wave-front sensing and for diffraction-limited imaging," Advances in Imaging and Electron Physics, vol. 141, pp. 1–76, 2006.

[11] R. G. Paxman and J. H. Seldin, "Fine-resolution astronomical imaging with phase-diverse speckle," in Digital Image Recovery and Synthesis II, vol. 2029, pp. 287–299, International Society for Optics and Photonics, 1993.

[12] R. W. Gerchberg, "A practical algorithm for the determination of phase from image and diffraction plane pictures," Optik, vol. 35, pp. 237–246, 1972.

[13] J. R. Fienup, "Phase retrieval algorithms: a comparison," Applied optics, vol. 21, no. 15, pp. 2758–2769, 1982.

[14] J. A. Rodriguez, R. Xu, C.-C. Chen, Y. Zou, and J. Miao, "Oversampling smoothness: an effective algorithm for phase retrieval of noisy diffraction intensities," Journal of applied crystallography, vol. 46, no. 2, pp. 312–318, 2013.

[15] M. G. Lofdahl, "Multi-frame blind deconvolution with linear equality constraints," in Image reconstruction from incomplete data II, vol. 4792, pp. 146–156, International Society for Optics and Photonics, 2002.

[16] D. Acton, D. Soltau, and W. Schmidt, "Full-field wavefront measurements with phase diversity.," Astronomy and Astrophysics, vol. 309, pp. 661–672, 1996.

[17] J. Fang and D. Savransky, "Amplitude and phase retrieval with simultaneous diversity estimation using expectation maximization," JOSA A, vol. 35, no. 2, pp. 293–300, 2018.

[18] M. Hartung, A. Blanc, T. Fusco, F. Lacombe, L. Mugnier, G. Rousset, and R. Lenzen, "Calibration of NAOS and CONICA static aberrations-experimental results," Astronomy & Astrophysics, vol. 399, no. 1, pp. 385–394, 2003.

[19] A. Blanc, T. Fusco, M. Hartung, L. Mugnier, and G. Rousset, "Calibration of NAOS and CONICA static aberrations-application of the phase diversity technique," Astronomy & Astrophysics, vol. 399, no. 1, pp. 373–383, 2003.

[20] M. A. van Dam, D. Le Mignant, and B. A. Macintosh, "Performance of the Keck Observatory adaptive-optics system," Applied Optics, vol. 43, no. 29, pp. 5458–5467, 2004.

[21] M. G. Löfdahl and G. Scharmer, "Wavefront sensing and image restoration from focused and defocused solar images.," Astronomy and Astrophysics Supplement Series, vol. 107, pp. 243–264, 1994.

[22] J.-F. Sauvage, T. Fusco, G. Rousset, and C. Petit, "Calibration and precompensation of noncommon path aberrations for extreme adaptive optics," JOSA A, vol. 24, no. 8, pp. 2334–2346, 2007.

[23] T. Fusco, C. Petit, G. Rousset, J. F. Sauvage, A. Blanc, J. M. Conan, and J. L. Beuzit, "Optimization of the pre-compensation of non-common path aberrations for adaptive optics systems," in Adaptive Optics: Methods, Analysis and Applications, p. AWB2, Optical Society of America, 2005.

[24] G. R. Brady, M. Guizar-Sicairos, and J. R. Fienup, "Optical wavefront measurement using phase retrieval with transverse translation diversity," Optics express, vol. 17, no. 2, pp. 624–639, 2009.

[25] M. Guizar-Sicairos and J. R. Fienup, "Measurement of coherent X-ray focused beams by phase retrieval with transverse translation diversity," Optics express, vol. 17, no. 4, pp. 2670–2685, 2009.

[26] N. Dey, L. Blanc-Feraud, C. Zimmer, P. Roux, Z. Kam, J.-C. Olivo-Marin, and J. Zerubia, "Richardson–Lucy algorithm with Total Variation regularization for 3D confocal microscope deconvolution," Microscopy research and technique, vol. 69, no. 4, pp. 260–266, 2006.

[27] Y. Wang, J. Yang, W. Yin, and Y. Zhang, "A new alternating minimization algorithm for total variation image reconstruction," SIAM Journal on Imaging Sciences, vol. 1, no. 3, pp. 248–272, 2008.

[28] H. Chang, Y. Lou, Y. Duan, and S. Marchesini, "Total Variation–based phase retrieval for Poisson noise removal," SIAM Journal on Imaging Sciences, vol. 11, no. 1, pp. 24–55, 2018.

[29] H. Jiang, C. Song, C.-C. Chen, R. Xu, K. S. Raines, B. P. Fahimian, C.-H. Lu, T.-K. Lee, A. Nakashima, J. Urano, et al., "Quantitative 3D imaging of whole, unstained cells by using X-ray diffraction microscopy," Proceedings of the National Academy of Sciences, vol. 107, no. 25, pp. 11234–11239, 2010.

[30] R. Tibshirani, "Regression shrinkage and selection via the Lasso," Journal of the Royal Statistical Society. Series B (Methodological), pp. 267–288, 1996.

[31] Y. Shechtman, A. Beck, and Y. C. Eldar, "GESPAR: Efficient phase retrieval of sparse signals," IEEE transactions on signal processing, vol. 62, no. 4, pp. 928–938, 2014.

[32] H. T. Nguyen, D. R. Luke, O. Soloviev, and M. Verhaegen, "SOPR for sparse phase retrieval," arXiv preprint arXiv:1804.01878, 2018.

[33] H. Ohlsson, A. Y. Yang, R. Dong, and S. S. Sastry, "Compressive phase retrieval from squared output measurements via semidefinite programming," arXiv preprint arXiv:1111.6323, 2011.

[34] H. Ohlsson and Y. C. Eldar, "On conditions for uniqueness in sparse phase retrieval," in Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, pp. 1841–1845, IEEE, 2014.

[35] P. Netrapalli, P. Jain, and S. Sanghavi, "Phase retrieval using alternating minimization," in Advances in Neural Information Processing Systems, pp. 2796–2804, 2013.

[36] A. Flinth, "Sparse blind deconvolution and demixing through $\ell_{1,2}$-minimization," Advances in Computational Mathematics, vol. 44, no. 1, pp. 1–21, 2018.

[37] P. Jung, F. Krahmer, and D. Stöger, "Blind demixing and deconvolution at near-optimal rate," preparation, 2018.

[38] N. Kazemi and M. D. Sacchi, "Sparse multichannel blind deconvolution," Geophysics, vol. 79, no. 5, pp. V143–V152, 2014.

[39] W. J. Wild, "Linear phase retrieval for wave-front sensing," Optics letters, vol. 23, no. 8, pp. 573–575, 1998.

[40] C. U. Keller, V. Korkiakoski, N. Doelman, R. Fraanje, R. Andrei, and M. Verhaegen, "Extremely fast focal-plane wavefront sensing for extreme adaptive optics," in Adaptive Optics Systems III, vol. 8447, p. 844721, International Society for Optics and Photonics, 2012.

[41] C. Smith, R. Marinica, and M. Verhaegen, "Real-time wavefront reconstruction from intensity measurements," in Proceedings of the 3rd AO4ELT Conference: Adaptive Optics for Extremely Large Telescopes, Florence, Italy, 26-31 May 2013, Arcetri Astrophysical Observatory, 2013.

[42] C. S. Smith, R. Marinica, J. Antonello, J. Arnold, and M. Verhaegen, "Focal-plane wavefront estimation and control using the extended Kalman filter," in Optomechatronic Technologies (ISOT), 2012 International Symposium on, pp. 1–6, IEEE, 2012.

[43] C. Smith, R. Marinică, A. Den Dekker, M. Verhaegen, V. Korkiakoski, C. Keller, and N. Doelman, "Iterative linear focal-plane wavefront correction," JOSA A, vol. 30, no. 10, pp. 2002–2011, 2013.

[44] R. Marinica, C. S. Smith, and M. Verhaegen, "State feedback control with quadratic output for wavefront correction in adaptive optics," in Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on, pp. 3475–3480, IEEE, 2013.

[45] S. Meimon, T. Fusco, and L. M. Mugnier, "Lift: a focal-plane wavefront sensor for real-time low-order sensing on faint sources," Optics letters, vol. 35, no. 18, pp. 3036–3038, 2010.

[46] M. Wilby, C. Keller, J.-F. Sauvage, T. Fusco, D. Mouillet, J.-L. Beuzit, and K. Dohlen, "A 'fast and furious' solution to the low-wind effect for SPHERE at the VLT," in Adaptive Optics Systems V, vol. 9909, p. 99096C, International Society for Optics and Photonics, 2016.

[47] V. Korkiakoski, C. U. Keller, N. Doelman, M. Kenworthy, G. Otten, and M. Verhaegen, "Fast & furious focal-plane wavefront sensing," Applied optics, vol. 53, no. 20, pp. 4565–4579, 2014.

[48] R. A. Gonsalves, "Adaptive optics by sequential diversity imaging," in European Southern Observatory Conference and Workshop Proceedings, vol. 58, p. 121, 2002.

[49] A. MacDonald, Blind deconvolution of anisoplanatic images collected by a partially coherent imaging system. PhD thesis, Air Force Institute of Technology, Wright-Patterson Air Force Base Ohio, 2004.

[50] D. P. Calle, Image formation with plasmonic nanostructures. PhD thesis, Universitat de València, 2015.

[51] D. Pastor, T. Stefaniuk, P. Wróbel, C. J. Zapata-Rodríguez, and R. Kotyński, "Determination of the point spread function of layered metamaterials assisted with the blind deconvolution algorithm," Optical and Quantum Electronics, vol. 47, no. 1, pp. 17–26, 2015.

[52] R. Paxman and J. Fienup, "Optical misalignment sensing and image reconstruction using phase diversity," JOSA A, vol. 5, no. 6, pp. 914–923, 1988.

[53] R. G. Paxman, T. J. Schulz, and J. R. Fienup, "Joint estimation of object and aberrations by using phase diversity," JOSA A, vol. 9, no. 7, pp. 1072–1085, 1992.

[54] E. Thiébaut, "Optimization issues in blind deconvolution algorithms," in Astronomical Data Analysis II, vol. 4847, pp. 174–184, International Society for Optics and Photonics, 2002.

[55] N. Baba, H. Tomita, and N. Miura, "Iterative reconstruction method in phase-diversity imaging," Applied optics, vol. 33, no. 20, pp. 4428–4433, 1994.

[56] T. J. Schulz, "Multiframe blind deconvolution of astronomical images," JOSA A, vol. 10, no. 5, pp. 1064–1073, 1993.

[57] C. R. Vogel, T. F. Chan, and R. J. Plemmons, "Fast algorithms for phase-diversity-based blind deconvolution," in Adaptive Optical System Technologies, vol. 3353, pp. 994–1006, International Society for Optics and Photonics, 1998.

[58] L. Mugnier, J.-F. Sauvage, T. Fusco, A. Cornia, and S. Dandy, "On-line long-exposure phase diversity: a powerful tool for sensing quasi-static aberrations of extreme adaptive optics imaging systems.," Optics Express, vol. 16, no. 22, pp. 18406–18416, 2008.

[59] R. A. Gonsalves and R. Chidlaw, "Wavefront sensing by phase retrieval," in Applications of Digital Image Processing III, vol. 207, pp. 32–40, International Society for Optics and Photonics, 1979.

[60] R. Mourya, L. Denis, J.-M. Becker, and É. Thiébaut, "A blind deblurring and image decomposition approach for astronomical image restoration," in European Conference on Signal Processing 2015, 2015.

[61] Y.-L. You and M. Kaveh, "A regularization approach to joint blur identification and image restoration," IEEE Transactions on Image Processing, vol. 5, no. 3, pp. 416–428, 1996.

[62] T. F. Chan and C.-K. Wong, "Total Variation blind deconvolution," IEEE transactions on Image Processing, vol. 7, no. 3, pp. 370–375, 1998.

[63] X. Gong, B. Lai, and Z. Xiang, "A l_0 sparse analysis prior for blind Poissonian image deconvolution," Optics express, vol. 22, no. 4, pp. 3860–3865, 2014.

[64] M. Van Noort, L. R. Van Der Voort, and M. G. Löfdahl, "Solar image restoration by use of multi-frame blind de-convolution with multiple objects and phase diversity," Solar Physics, vol. 228, no. 1-2, pp. 191–215, 2005.

[65] R. Tyson, Principles of adaptive optics. CRC press, 2010.

[66] F. Roddier, Adaptive optics in astronomy. Cambridge university press, 1999.

[67] R. Davies and M. Kasper, "Adaptive optics for astronomy," Annual Review of Astronomy and Astrophysics, vol. 50, pp. 305–351, 2012.

[68] B. C. Platt and R. Shack, "History and principles of Shack–Hartmann wavefront sensing," Journal of refractive surgery, vol. 17, no. 5, pp. S573–S577, 2001.

[69] C. C. de Visser and M. Verhaegen, "Wavefront reconstruction in adaptive optics systems using nonlinear multivariate splines," JOSA A, vol. 30, no. 1, pp. 82–95, 2013.

[70] E. Brunner, C. C. de Visser, and M. Verhaegen, "Nonlinear spline wavefront reconstruction from Shack–Hartmann intensity measurements through small aberration approximations," JOSA A, vol. 34, no. 9, pp. 1535–1549, 2017.

[71] E. Brunner, Spline-based wavefront reconstruction for Shack-Hartmann measurements. PhD thesis, Delft University of Technology, 2018.

[72] K. Hinnen, M. Verhaegen, and N. Doelman, "A data-driven $\mathscr{H}_2$-optimal control approach for adaptive optics," IEEE Transactions on Control Systems Technology, vol. 16, no. 3, pp. 381–395, 2008.

[73] M. Verhaegen and V. Verdult, Filtering and system identification: a least squares approach. Cambridge university press, 2007.

[74] A. Roorda, F. Romero-Borja, W. J. Donnelly III, H. Queener, T. J. Hebert, and M. C. Campbell, "Adaptive optics scanning laser ophthalmoscopy," Optics express, vol. 10, no. 9, pp. 405–412, 2002.

[75] H. Hofer, N. Sredar, H. Queener, C. Li, and J. Porter, "Wavefront sensorless adaptive optics ophthalmoscopy in the human eye," Optics express, vol. 19, no. 15, pp. 14160–14171, 2011.

[76] Y. N. Sulai and A. Dubra, "Non-common path aberration correction in an adaptive optics scanning ophthalmoscope," Biomedical optics express, vol. 5, no. 9, pp. 3059–3073, 2014.

[77] J.-F. Sauvage, L. Mugnier, B. Paul, and R. Villecroze, "Coronagraphic phase diversity: a simple focal plane sensor for high-contrast imaging," Optics Letters, vol. 37, no. 23, pp. 4808–4810, 2012.

[78] G. Chanan, D. G. MacMartin, J. Nelson, and T. Mast, "Control and alignment of segmented-mirror telescopes: matrices, modes, and error propagation," Applied Optics, vol. 43, no. 6, pp. 1223–1232, 2004.

[79] R. C. Jared, A. Arthur, S. Andreae, A. Biocca, R. W. Cohen, J. M. Fuertes, J. Franck, G. Gabor, J. Llacer, T. S. Mast, et al., "WM Keck Telescope segmented primary mirror active control system," in Advanced Technology Optical Telescopes IV, vol. 1236, pp. 996–1009, International Society for Optics and Photonics, 1990.

[80] R. H. Minor, A. Arthur, G. Gabor, H. G. Jackson, R. C. Jared, T. S. Mast, and B. A. Schaefer, "Displacement sensors for the primary mirror of the WM Keck telescope," in Advanced Technology Optical Telescopes IV, vol. 1236, pp. 1009–1018, International Society for Optics and Photonics, 1990.

[81] M. Dimmler, T. Erm, B. Bauvir, B. Sedghi, H. Bonnet, M. Müller, and A. Wallander, "E-ELT primary mirror control system," in Ground-based and Airborne Telescopes II, vol. 7012, p. 70121O, International Society for Optics and Photonics, 2008.

[82] R. Bastaits, Extremely large segmented mirrors: dynamics, control and scale effects. PhD thesis, Université libre de Bruxelles, 2010.

[83] B. Le Roux, J.-M. Conan, C. Kulcsár, H.-F. Raynaud, L. M. Mugnier, and T. Fusco, "Optimal control law for classical and multiconjugate adaptive optics," JOSA A, vol. 21, no. 7, pp. 1261–1276, 2004.

[84] C. Kulcsár, H.-F. Raynaud, J.-M. Conan, C. Correia, and C. Petit, "Control design and turbulent phase models in adaptive optics: A state-space interpretation," in Adaptive Optics: Methods, Analysis and Applications, p. AOWB1, Optical Society of America, 2009.

[85] K. Hinnen, M. Verhaegen, and N. Doelman, "Exploiting the spatiotemporal correlation in adaptive optics using data-driven $\mathscr{H}_2$-optimal control," JOSA A, vol. 24, no. 6, pp. 1714–1725, 2007.

[86] B. Sinquin and M. Verhaegen, "QUARKS: Identification of large-scale Kronecker vector-autoregressive models," IEEE Transactions on Automatic Control, 2018.

[87] B. Sinquin and M. Verhaegen, "Tensor-based predictive control for extremely large-scale single conjugate adaptive optics," JOSA A, vol. 35, no. 9, pp. 1612–1626, 2018.

[88] B. Sinquin and M. Verhaegen, "K4SID: Large-scale subspace identification with kronecker modeling," IEEE Transactions on Automatic Control, 2018.

[89] C. Correia, H.-F. Raynaud, C. Kulcsár, and J.-M. Conan, "On the optimal reconstruction and control of adaptive optical systems with mirror dynamics," JOSA A, vol. 27, no. 2, pp. 333–349, 2010.

[90] D. Ren, B. Dong, Y. Zhu, and D. J. Christian, "Correction of non–common-path error for extreme adaptive optics," Publications of the Astronomical Society of the Pacific, vol. 124, no. 913, p. 247, 2012.

[91] H. H. Bauschke, P. L. Combettes, and D. R. Luke, "Phase retrieval, error reduction algorithm, and Fienup variants: a view from convex optimization," JOSA A, vol. 19, no. 7, pp. 1334–1345, 2002.

[92] V. Katkovnik, "Phase retrieval from noisy data based on sparse approximation of object phase and amplitude," arXiv preprint arXiv:1709.01071, 2017.

[93] J. R. Fienup, "Reconstruction of an object from the modulus of its Fourier transform," Optics letters, vol. 3, no. 1, pp. 27–29, 1978.

[94] H. H. Bauschke, P. L. Combettes, and D. R. Luke, "Hybrid projection–reflection method for phase retrieval," JOSA A, vol. 20, no. 6, pp. 1025–1034, 2003.

[95] V. Elser, "Solution of the crystallographic phase problem by iterated projections," Acta Crystallographica Section A: Foundations of Crystallography, vol. 59, no. 3, pp. 201–209, 2003.

[96] D. R. Luke, "Relaxed averaged alternating reflections for diffraction imaging," Inverse problems, vol. 21, no. 1, p. 37, 2004.

[97] D. Russell Luke, N. H. Thao, and M. K. Tam, "Quantitative convergence analysis of iterated expansive, set-valued mappings," Mathematics of Operations Research, vol. 43, no. 4, pp. 1143–1176, 2018.

[98] K. Jaganathan, Y. C. Eldar, and B. Hassibi, "Phase retrieval: An overview of recent developments," arXiv preprint arXiv:1510.07713, 2015.

[99] E. J. Candes, T. Strohmer, and V. Voroninski, "PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming," Communications on Pure and Applied Mathematics, vol. 66, no. 8, pp. 1241–1274, 2013.

[100] I. Waldspurger, A. d'Aspremont, and S. Mallat, "Phase recovery, MaxCut and complex semidefinite programming," Mathematical Programming, vol. 149, no. 1-2, pp. 47–81, 2015.

[101] F. Fogel, I. Waldspurger, and A. d'Aspremont, "Phase retrieval for imaging problems," Mathematical programming computation, vol. 8, no. 3, pp. 311–335, 2016.

[102] M. X. Goemans and D. P. Williamson, "Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming," Journal of the ACM (JACM), vol. 42, no. 6, pp. 1115–1145, 1995.

[103] A. Yurtsever, M. Udell, J. A. Tropp, and V. Cevher, "Sketchy decisions: Convex low-rank matrix optimization with optimal storage," arXiv preprint arXiv:1702.06838, 2017.

[104] L. Vandenberghe, V. R. Balakrishnan, R. Wallin, A. Hansson, and T. Roh, "Interior-point algorithms for semidefinite programming problems derived from the KYP lemma," in Positive polynomials in control, pp. 195–238, Springer, 2005.

[105] L.-H. Yeh, "Analysis and comparison of Fourier ptychographic phase retrieval algorithms," tech. rep., Technical Report No. UCB/EECS-2016-86 (University of California, 2016), 2016.

[106] T. Goldstein and C. Studer, "PhaseMax: Convex phase retrieval via basis pursuit," IEEE Transactions on Information Theory, 2018.

[107] O. Dhifallah and Y. M. Lu, "Fundamental limits of PhaseMax for phase retrieval: A replica analysis," in Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2017 IEEE 7th International Workshop on, pp. 1–5, IEEE, 2017.

[108] P. Hand and V. Voroninski, "Corruption robust phase retrieval via linear programming," arXiv preprint arXiv:1612.03547, 2016.

[109] O. Dhifallah, C. Thrampoulidis, and Y. M. Lu, "Phase retrieval via linear programming: Fundamental limits and algorithmic improvements," in Communication, Control, and Computing (Allerton), 2017 55th Annual Allerton Conference on, pp. 1071–1077, IEEE, 2017.

[110] D. Gross, F. Krahmer, and R. Kueng, "A partial derandomization of PhaseLift using spherical designs," Journal of Fourier Analysis and Applications, vol. 21, no. 2, pp. 229–266, 2015.

[111] E. J. Candes, X. Li, and M. Soltanolkotabi, "Phase retrieval via Wirtinger flow: Theory and algorithms," IEEE Transactions on Information Theory, vol. 61, no. 4, pp. 1985–2007, 2015.

[112] E. J. Candès and X. Li, "Solving quadratic equations via PhaseLift when there are about as many equations as unknowns," Foundations of Computational Mathematics, vol. 14, no. 5, pp. 1017–1026, 2014.

[113] I. Waldspurger, "Phase retrieval with random Gaussian sensing vectors by alternating projections," IEEE Transactions on Information Theory, 2018.

[114] A. S. Bandeira, J. Cahill, D. G. Mixon, and A. A. Nelson, "Saving phase: Injectivity and stability for phase retrieval," Applied and Computational Harmonic Analysis, vol. 37, no. 1, pp. 106–125, 2014.

[115] E. J. Candes, X. Li, and M. Soltanolkotabi, "Phase retrieval from coded diffraction patterns," Applied and Computational Harmonic Analysis, vol. 39, no. 2, pp. 277–299, 2015.

[116] Y. Chen and E. Candes, "Solving random quadratic systems of equations is nearly as easy as solving linear systems," in Advances in Neural Information Processing Systems, pp. 739–747, 2015.

[117] G. Wang, G. B. Giannakis, and Y. C. Eldar, "Solving systems of random quadratic equations via truncated amplitude flow," IEEE Transactions on Information Theory, vol. 64, no. 2, pp. 773–794, 2018.

[118] G. Wang, G. Giannakis, Y. Saad, and J. Chen, "Solving most systems of random quadratic equations," in Advances in Neural Information Processing Systems, pp. 1867–1877, 2017.

[119] J. R. Magnus, "Linear structures," Griffin's statistical monographs and courses, no. 42, 1988.

[120] J. Zhang, N. Pégard, J. Zhong, H. Adesnik, and L. Waller, "3D computer-generated holography by non-convex optimization," Optica, vol. 4, no. 10, pp. 1306–1313, 2017.

[121] F. A. South, Y.-Z. Liu, A. J. Bower, Y. Xu, P. S. Carney, and S. A. Boppart, "Wavefront measurement using computational adaptive optics," JOSA A, vol. 35, no. 3, pp. 466–473, 2018.

[122] H. R. Verstraete, S. Wahls, J. Kalkman, and M. Verhaegen, "Model-based sensorless wavefront aberration correction in optical coherence tomography," Optics letters, vol. 40, no. 24, pp. 5722–5725, 2015.

[123] Y. Rivenson, Y. Zhang, H. Günaydın, D. Teng, and A. Ozcan, "Phase recovery and holographic image reconstruction using deep learning in neural networks," Light: Science & Applications, vol. 7, no. 2, p. 17141, 2018.

[124] S. Jiang, K. Guo, J. Liao, and G. Zheng, "Solving Fourier ptychographic imaging problems via neural network modeling and tensorflow," arXiv preprint arXiv:1803.03434, 2018.

[125] C. Metzler, P. Schniter, A. Veeraraghavan, et al., "prDeep: Robust phase retrieval with a flexible deep network," in International Conference on Machine Learning, pp. 3498–3507, 2018.

[126] W. H. Richardson, "Bayesian-based iterative method of image restoration," JOSA, vol. 62, no. 1, pp. 55–59, 1972.

[127] L. B. Lucy, "An iterative technique for the rectification of observed distributions," The astronomical journal, vol. 79, p. 745, 1974.

[128] D. Fish, A. Brinicombe, E. Pike, and J. Walker, "Blind deconvolution by means of the Richardson–Lucy algorithm," JOSA A, vol. 12, no. 1, pp. 58–65, 1995.

[129] B. Kim and T. Naemura, "Blind deconvolution of 3D fluorescence microscopy using depth-variant asymmetric PSF," Microscopy research and technique, vol. 79, no. 6, pp. 480–494, 2016.

[130] B. Kim and T. Naemura, "Blind depth-variant deconvolution of 3D data in wide-field fluorescence microscopy," Scientific reports, vol. 5, p. 9894, 2015.

[131] G. Molodij, S. Keil, T. Roudier, N. Meunier, and S. Rondi, "A method for single image restoration based on the principal ergodic," JOSA A, vol. 27, no. 11, pp. 2459–2467, 2010.

[132] F. Tsumuraya, N. Miura, and N. Baba, "Iterative blind deconvolution method using Lucy's algorithm," Astronomy and Astrophysics, vol. 282, pp. 699–708, 1994.

[133] G. Ayers and J. C. Dainty, "Iterative blind deconvolution method and its applications," Optics letters, vol. 13, no. 7, pp. 547–549, 1988.

[134] D. Wilding, O. Soloviev, P. Pozzi, G. Vdovin, and M. Verhaegen, "Blind multi-frame deconvolution by tangential iterative projections (TIP)," Optics Express, vol. 25, no. 26, pp. 32305–32322, 2017.

[135] D. Wilding, P. Pozzi, O. Soloviev, G. Vdovin, and M. Verhaegen, "Pupil mask diversity for image correction in microscopy," Optics Express, vol. 26, no. 12, pp. 14832–14841, 2018.

[136] P. Campisi and K. Egiazarian, Blind image deconvolution: theory and applications. CRC press, 2016.

[137] J. Markham and J.-A. Conchello, "Parametric blind deconvolution: a robust method for the simultaneous estimation of image and blur," JOSA A, vol. 16, no. 10, pp. 2377–2391, 1999.

[138] A. Ahmed, B. Recht, and J. Romberg, "Blind deconvolution using convex programming," IEEE Transactions on Information Theory, vol. 60, no. 3, pp. 1711–1732, 2014.

[139] A. Ahmed, A. Cosse, and L. Demanet, "A convex approach to blind deconvolution with diverse inputs," in Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2015 IEEE 6th International Workshop on, pp. 5–8, IEEE, 2015.

[140] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," SIAM review, vol. 52, no. 3, pp. 471–501, 2010.

[141] T. Strohmer and K. Wei, "Painless breakups—efficient demixing of low rank matrices," Journal of Fourier Analysis and Applications, pp. 1–31, 2017.

[142] D. Stöger, P. Jung, F. Krahmer, et al., "Blind deconvolution and compressed sensing," in Compressed Sensing Theory and its Applications to Radar, Sonar and Remote Sensing (CoSeRa), 2016 4th International Workshop on, pp. 24–27, IEEE, 2016.

[143] S. Ling and T. Strohmer, "Simultaneous blind deconvolution and blind demixing via convex programming," in Signals, Systems and Computers, 2016 50th Asilomar Conference on, pp. 1223–1227, IEEE, 2016.

[144] S. Ling and T. Strohmer, "Regularized gradient descent: a non-convex recipe for fast joint blind deconvolution and demixing," Information and Inference: A Journal of the IMA, 2017.

[145] L.-H. Yeh, J. Dong, J. Zhong, L. Tian, M. Chen, G. Tang, M. Soltanolkotabi, and L. Waller, "Experimental robustness of Fourier ptychography phase retrieval algorithms," Optics express, vol. 23, no. 26, pp. 33214–33240, 2015.

[146] P. Thibault and M. Guizar-Sicairos, "Maximum-likelihood refinement for coherent diffractive imaging," New Journal of Physics, vol. 14, no. 6, p. 063004, 2012.

[147] S. W. Paine and J. R. Fienup, "Machine learning for improved image-based wavefront sensing," Optics letters, vol. 43, no. 6, pp. 1235–1238, 2018.

[148] A. M. Maiden and J. M. Rodenburg, "An improved ptychographical phase retrieval algorithm for diffractive imaging," Ultramicroscopy, vol. 109, no. 10, pp. 1256–1262, 2009.

[149] A. M. Maiden, M. J. Humphry, F. Zhang, and J. M. Rodenburg, "Superresolution imaging via ptychography," JOSA A, vol. 28, no. 4, pp. 604–612, 2011.

2

SOLVING LARGE-SCALE GENERAL PHASE RETRIEVAL PROBLEMS VIA A SEQUENCE OF CONVEX RELAXATIONS

We present a convex relaxation-based algorithm for large-scale general phase retrieval problems. General phase retrieval problems include i.a. the estimation of the phase of the optical field in the pupil plane based on intensity measurements of a point source recorded in the image (focal) plane. The non-convex problem of finding the complex field that generates the correct intensity is reformulated into a rank constraint problem. The nuclear norm is used to obtain the convex relaxation of the phase retrieval problem. A new iterative method, indicated as Convex Optimization-based Phase Retrieval (COPR), is presented, with each iteration consisting of solving a convex problem. In the noise-free case and for a class of phase retrieval problems the solutions of the minimization problems converge linearly or faster towards a correct solution. Since the solutions to nuclear norm minimization problems can be computed using semidefinite programming, and this tends to be an expensive optimization in terms of scalability, we provide a fast Alternating Direction Method of Multipliers (ADMM) algorithm that exploits the problem structure. The performance of the COPR algorithm is demonstrated in a realistic numerical simulation study, demonstrating its improvements in reliability and speed with respect to state-of-the-art methods. Furthermore, COPR is tested on experimental measurements.

Parts of this chapter have been published in Reinier Doelman, Nguyen H. Thao, and Michel Verhaegen, "Solving large-scale general phase retrieval problems via a sequence of convex relaxations," J. Opt. Soc. Am. A 35, 1410-1419 (2018), [1].

2.1. INTRODUCTION

Recovery of a signal from several measured intensity patterns, also known as the phase retrieval problem, is of great interest in optics and imaging. Recently it was shown in [2] that the problem of estimating the wavefront aberration from measurements of the Point Spread Functions can be formulated as a phase retrieval problem.

In this paper, we consider the general phase retrieval problem [3]:

$$\text{find} \quad \mathbf{a} \in \mathbb{C}^{n_a} \text{ such that } \mathbf{y}_i = |\mathbf{u}_i^H \mathbf{a}|^2 \quad \text{for } i = 1, \ldots, n_y,$$

where $\mathbf{y}_i \in \mathbb{R}_+$ and $\mathbf{u}_i \in \mathbb{C}^{n_a}$ are known and $(\cdot)^H$ denotes the Hermitian transpose of a vector (matrix). For brevity the following compact notation will be used in this paper to denote this general noise-free phase retrieval problem:

$$\text{find} \quad \mathbf{a} \in \mathbb{C}^{n_a} \text{ such that } \mathbf{y} = |U\mathbf{a}|^2, \tag{2.1}$$

where $\mathbf{y} \in \mathbb{R}_+^{n_y}$ are the measurements and $U \in \mathbb{C}^{n_y \times n_a}$ is the propagation matrix. With noise on the measurements y_i, we consider the following related optimization problem:

$$\min_{\mathbf{a} \in \mathbb{C}^{n_a}} \quad \left\| \mathbf{y} - |U\mathbf{a}|^2 \right\|, \tag{2.2}$$

where $\|\cdot\|$ denotes a vector norm of interest.

The sparse variant of the phase retrieval problem corresponds to the case that the unknown parameter $\mathbf{a}$ is a sparse vector. A special case of this problem is when the measurements are the magnitude of the Fourier transform of multiples of $\mathbf{a}$ with certain phase diversity patterns. A number of algorithms utilizing the Fourier transform have been proposed for solving this class of phase retrieval problems [4–6].

The fundamental nature of (2.1) has given rise to a wide variety of solution methods that have been developed for specific variants of this problem since the observation of Sayre in 1952 that phase information of a scattered wave may be recovered from the recorded intensity patterns at and between Bragg peaks of a diffracted wave [7]. Direct methods [8] usually use insights about the crystallographic structure and randomization to search for the missing phase information. The requirement of such a-priori structural information and the expensive computational complexity often limit the application of these methods in practice.

A second class of methods first devised by Gerchberg and Saxton [9] and Fienup [4] can be described as variants of the method of alternating projections on certain sets defined by the constraints. For an overview of these methods and latter refinements we refer the reader to [5, 10].

In [11] (2.1) is relaxed to a convex optimization problem. The inclusion of the sparsity constraint in the same framework of convex relaxations has been considered in [12]. However, as reported in [6] the combination of matrix lifting and Semidefite Programming (SDP) makes this method not suitable for large-scale problems. To deal with large-scale problems, the authors of [6] have proposed an iterative solution method, called GESPAR, which appears to yield promising recovery of very sparse signals. However, this method consists of a heuristic search for the support of $\mathbf{a}$ in combination with a variant of Gauss-Newton method, whose computational complexity is often expensive. These algorithmic features are potential drawbacks of GESPAR.

In this paper, we propose a sequence of convex relaxations for the phase retrieval problem in (2.1). Contrary to existing convex relaxation schemes such as those proposed in [11, 12], matrix lifting is not required in our strategy. The obtained convex problems are affine in the unknown parameter vector **a**. Contrary to [13], our strategy does not require the tuning of regularization parameters when the measurements are corrupted by noise. We then present an algorithm based on the Alternating Direction Method of Multipliers (ADMM) that can solve the resulting optimization problems effectively. This potentially addresses the restriction of current SDP-based methods to only relatively small-scale problems.

In Section 2.2 we formulate the estimation problem of our interest for both zonal and modal forms. In Section 2.3 we propose an algorithm for solving this problem. The algorithm is based on iteratively minimizing a nuclear norm. The nuclear norm of a matrix is the sum of its singular values. Its benefit in optimization is that it is used as a convex relaxation to the rank function [14]. The convexity enables direct use of standard software libraries for solving convex optimization problems. However, since it is a computationally heavy minimization problem, we suggest an ADMM-based algorithm based on [15] in Section 2.4 that exploits the problem structure and is therefore more efficient in practical cases. This ADMM algorithm features two minimization problems whose solutions can be computed exactly and with complexity $\mathcal{O}\left(n_y n_a\right)$, where n_y is the number of measurements and n_a is the number of unknown variables. To find these solutions either a least-squares problem has to be solved or the Singular Value Decompositionss of 2×2 matrices have to be computed. Analytic solutions for the ADMM algorithm update steps will be presented in Subsections 2.4.1 and 2.4.2. The convergence behaviour of the algorithm proposed in Section 2.3 is analyzed in Section 2.5. Compared to the other sections, the mathematical analysis in this section is more involved, which is often the case for convergence analyses. In Section 2.6 we describe and discuss the results of a number of numerical experiments that demonstrate the promising performances of our algorithms. In Section 2.7 we test COPR on experimental measurements. We end with concluding remarks in Section 2.8.

2.2. WAVEFRONT ESTIMATION FROM INTENSITY MEASUREMENTS

The problem of phase retrieval from the Point Spread Function images can be approached from 2 directions. We take the opportunity to present them in a unified way. We first describe the problem in zonal form, and then in modal form. The modal form approach used in this paper seems less popular than the zonal form one.

2.2.1. PROBLEM FORMULATION IN ZONAL FORM

In [2] it was shown that reconstructing the wavefront from Charge-Coupled Device (CCD) recorded images of a point source may also be formulated as a phase retrieval problem. These recorded images are called Point Spread Functions (PSFs). As such approaches avoid the requirement of extra hardware to sense the wavefront, such as a Shack-Hartmann wavefront sensor, the problem is relevant and summarized here.

The PSF is derived from the magnitude of the Fourier transform of the Generalized Pupil Function (GPF). For an aberrated optical system the GPF is defined as the complex

valued function [16]:

$$\mathscr{P}(\rho,\theta) = \mathbf{A}(\rho,\theta)e^{j\phi(\rho,\theta)}, \tag{2.3}$$

where ρ (radius) and θ (angle) specify the normalized polar coordinates in the exit pupil plane of the optical system. In (2.3), $\mathbf{A}(\rho,\theta)$ is the amplitude apodisation function and $\phi(\rho,\theta)$ is the phase aberration function.

The aim of the wavefront reconstruction problem is to estimate $\phi(\rho,\theta)$. Once this phase aberration of an optical system has been estimated, it can be corrected by using phase modulating devices such as deformable mirrors.

In order to estimate $\phi(\rho,\theta)$, a known phase diversity pattern $\phi_d(\rho,\theta)$ can be introduced (e.g., by using a deformable mirror) to transform the GPF in a controlled manner into the aberrated GPF:

$$\mathscr{P}_d(\rho,\theta) = \mathbf{A}(\rho,\theta)e^{j\phi(\rho,\theta)}e^{j\phi_d(\rho,\theta)}. \tag{2.4}$$

The noise-free intensity pattern of $\mathscr{P}_d(\rho,\theta)$ measured at the image plane is denoted

$$\mathbf{y}_d = \left|\mathscr{F}^{-1}\left\{\mathbf{A}(\rho,\theta)e^{j\phi(\rho,\theta)}e^{j\phi_d(\rho,\theta)}\right\}\right|^2. \tag{2.5}$$

If we sample the function $\mathscr{P}_d(\rho,\theta)$ at points corresponding to a square grid of size $m\times m$ on the pupil plane, then $\mathbf{A}(\rho,\theta)$, $\phi_d(\rho,\theta)$ and $\phi(\rho,\theta)$ are square matrices of that size.

Let us define vect$(\cdot)$ the vectorization operator such that vect(Z) yields the vector obtained by stacking the columns of matrix Z into a column vector. The inverse operator vect$^{-1}(\cdot)$, which maps a column vector of size m^2 to a square matrix of size $m\times m$, is also well defined. Let in particular the matrix Z and the vector $\mathbf{a}$ be defined as:

$$Z = \mathbf{A}(\rho,\theta)e^{j\phi(\rho,\theta)} \in \mathbb{C}^{m\times m}, \quad \mathbf{a} = \text{vect}(Z) \in \mathbb{C}^{m^2}.$$

With the definition of the vector $\bar{\mathbf{h}}_d$:

$$\bar{\mathbf{h}}_d = \text{vect}\left(e^{j\phi_d(\rho,\theta)}\right) \in \mathbb{C}^{m^2},$$

and with $D_d = \text{d}\left(\bar{\mathbf{h}}_d\right) \in \mathbb{C}^{m^2\times m^2}$ the diagonal matrix with diagonal entries taken from the vector $\bar{\mathbf{h}}_d$, we can write the noise-free intensity measurements in (2.5) as

$$\mathbf{y}_d = \left|\mathscr{F}^{-1}\left\{e^{j\phi_d(\rho,\theta)}Z\right\}\right|^2 = \left|\mathscr{F}^{-1}\left\{\text{vect}^{-1}(D_d\mathbf{a})\right\}\right|^2.$$

As the Fourier transform is a linear operator, we can write our noise-free intensity measurements in the form:

$$\mathbf{y}_d = |U_d\mathbf{a}|^2, \tag{2.6}$$

where in this case U_d is a unitary matrix.

By stacking the vectors $\mathbf{y}_d$ and the matrices U_d, obtained from the n_d images with n_d different phase diversities, correspondingly into the vector $\mathbf{y}$ and the matrix U (of size $n_d m^2 \times m^2$), the problem of finding $\mathbf{a}$ from noise-free intensity measurements can be formulated as in (2.1) and that from noisy measurements can be formulated as in (2.2) for $n_a = m^2$ and $n_y = n_d m^2$.

It is worth noting that the dimension of the unknown **a** with m in the range of a couple of hundreds turns this problem into a non-convex large-scale optimization problem. For such a problem the implementation of PhaseLift [13] using standard semidefinite programming, using libraries like MOSEK [17], will not be tractable because of the large matrix dimensions of the unknown quantity. If we assume that the computational complexity of semidefinite programming with matrix constraints of size $n \times n$ increases with $\mathcal{O}\left(n^6\right)$ [18], then a naive implementation of the PhaseLift method applied to (2.2) involving a single image has worst-case computational complexity of $\mathcal{O}\left(m^{12}\right)$.

2.2.2. PROBLEM FORMULATION IN MODAL FORM

In general, only approximate solutions can be expected for a phase retrieval problem. In the modal form of the phase retrieval problem, also considered in [2] for Extended Nijboer-Zernike (ENZ) basis functions, the GPF is assumed to be well approximated by a weighted sum of basis functions. We make use of real-valued radial basis functions [19] with complex coefficients to approximate the GPF. These are studied in the scope of wavefront estimation in [20] and an illustration of these basis function on a 4×4 grid in the pupil plane is given in Figure 2.1.

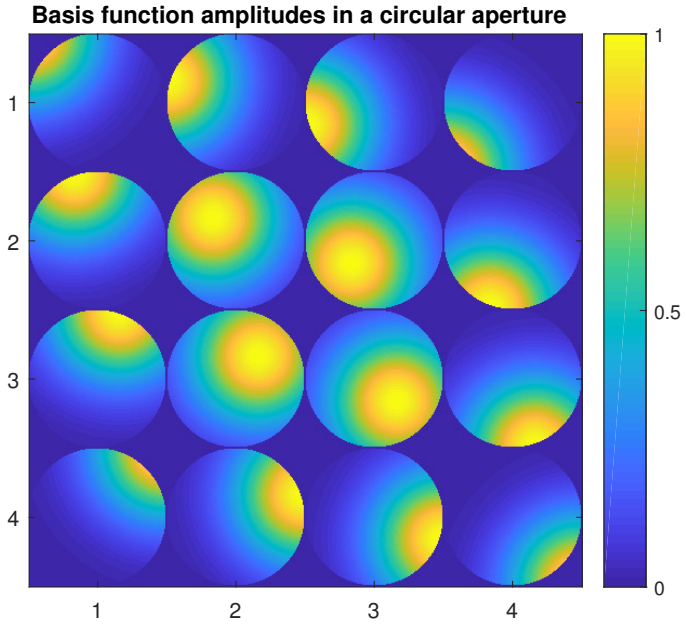


Figure 2.1: 16 radial basis functions with centers in a 4×4 grid, with circular aperture support.

Switching from the polar coordinates (ρ, θ) to the Cartesian coordinates (x, y) in the

pupil plane, let us consider the radial basis functions and the approximate GPF given by

$$\begin{aligned} G_i(x,y) &= \chi(x,y)e^{-\lambda_i((x-x_i)^2+(y-y_i)^2)}, \\ \mathscr{P}(x,y) &\approx \widetilde{\mathscr{P}}(x,y,\mathbf{a}) = \sum_{i=1}^{n_a} a_i G_i(x,y), \end{aligned} \tag{2.7}$$

where (x_i, y_i) are the centers of basis functions $G_i(x,y)$, $a_i \in \mathbb{C}$, $\lambda_i \in \mathbb{R}_+$ determines the spread of that function, $\chi(x,y)$ denotes the support of the aperture, and $\mathbf{a}$ is the coefficient parameter vector to be estimated. The parameters λ_i are usually taken equal for all basis functions and for their tuning we refer to [20].

The aberrated GPF corresponding to the introduction of phase diversity ϕ_d is

$$\widetilde{\mathscr{P}}_d(x,y,\mathbf{a},\phi_d) = \sum_{i=1}^{n_a} a_i G_i(x,y) e^{j\phi_d(x,y)}. \tag{2.8}$$

The normalized complex PSF (amplitude impulse response) is the 2-dimensional inverse Fourier transform of the GPF [21, 22]. The aberrated complex PSF corresponding to the aberrated GPF in (2.8) is given as

$$\mathbf{h}_d(u,v) = \sum_{i=1}^{n_a} a_i \mathscr{F}^{-1}\left\{G_i(x,y)e^{j\phi_d(x,y)}\right\} = \sum_{i=1}^{n_a} a_i U_{d,i}(u,v), \tag{2.9}$$

where (u,v) are the Cartesian coordinates in the image plane of the optical system.

We now drop the dependency on the coordinates and vectorize expression (2.9) for all n_d diversities that have been applied to obtain the following compact form of a single matrix-vector multiplication,

$$\bar{\mathbf{h}} = U\mathbf{a}. \tag{2.10}$$

The vector $\bar{\mathbf{h}}$ is the obtained vectorization and combination over all the aberrated complex PSFs, and the matrix U is the vectorized and concatenated version of the functions $U_{d,i}$ sampled on a grid of size $m \times m$.

Let the intensity of the PSFs be recorded on the corresponding grid of pixels of size $m \times m$, and let the vectorization of this intensity pattern for different phase diversities be concatenated into the vector $\mathbf{y}$. We can again formulate the problem of finding $\mathbf{a}$ from noise-free intensity measurements as in (2.1) and from noisy measurements as in (2.2) for $n_y = m^2 n_d$.

It is worth noting that the dimension of $\mathbf{a}$ is not dependent on the size of the sample grid (the size of the problem). This is the fundamental advantage of the modal form formulation over the zonal form one, for which the size of $\mathbf{a}$ directly depends on the size of the problem, i.e. $n_a = m^2$.

In this paper two steps are combined to deal with the large-scale nature of optimization (2.2):

1. The unknown pupil function $\mathscr{P}(\rho,\theta)$ can be represented as a linear combination of a number of basis functions. In [2] use has been made of the ENZ basis functions, while in [20] use is made of radial basis functions instead of ENZ ones. The radial basis functions are used here as [20] demonstrated their advantages over the ENZ type.

2. A new strategy is proposed for solving optimization (2.1) via a sequence of convex optimization problems. Each of the subproblems can be solved effectively by an iterative ADMM algorithm that exploits the problem structure.

In the following we assume that the problem is normalized such that all entries of $\mathbf{y}$ have values between 0 and 1.

2.3. THE COPR ALGORITHM

Equation 2.1 is equivalent to a rank constraint. Define the matrix-valued function

$$L(A,B,C,X,Y) = \begin{pmatrix} C + AY + XB + XY & A + X \\ B + Y & I \end{pmatrix}, \tag{2.11}$$

where I is the identity matrix of appropriate size. Let $\mathbf{b} \in \mathbb{C}^{n_a}$ be a coefficient vector. For notational convenience, we will denote

$$\begin{aligned} &M(U,\mathbf{a},\mathbf{b},\mathbf{y}) := \\ &L\left(\mathrm{d}\left(\mathbf{a}^H U^H\right), \mathrm{d}\left(U\mathbf{a}\right), \mathrm{d}\left(\mathbf{y}\right), \mathrm{d}\left(\mathbf{b}^H U^H\right), \mathrm{d}\left(U\mathbf{b}\right)\right). \end{aligned}$$

Our proposed algorithm in this paper relies on the following fundamental result.

Lemma 2.3.1 ([23])**.** For any $\mathbf{b} \in \mathbb{C}^{n_a}$, the constraint $\mathbf{y} = |U\mathbf{a}|^2$ is equivalent to the constraint

$$\operatorname{rank}\left(M(U,\mathbf{a},\mathbf{b},\mathbf{y})\right) = n_y.$$

For addressing problem (2.2), Lemma 2.3.1 suggests a consideration of the following approximate problem, for a user-selected parameter vector $\mathbf{b}$,

$$\min_{\mathbf{a}\in\mathbb{C}^{n_a}} \operatorname{rank}\left(M(U,\mathbf{a},\mathbf{b},\mathbf{y})\right). \tag{2.12}$$

Since (2.12) is a non-convex problem and to anticipate the presence of measurement noise, we propose to solve the following convex optimization problem:

$$\min_{\mathbf{a}\in\mathbb{C}^{n_a}} f(\mathbf{a}) := \left\| M(U,\mathbf{a},\mathbf{b},\mathbf{y}) \right\|_*, \tag{2.13}$$

where $\|\cdot\|_*$ denotes the nuclear norm of a matrix, the sum of its singular values [14, 24]. The motivation to choose M (and L) in the structure of (2.11), is that it is affine in the unknown $\mathbf{a}$. By relaxing the rank constraint into (2.13) we obtain a convex relaxation without 'lifting' (substituting) the variables as is the case with PhaseLift. One advantage is that the solution for $\mathbf{a}$ can be easily influenced if we have prior knowledge. For example, in the case that prior knowledge on the problem indicates that $\mathbf{a}$ is a sparse vector, the objective function in (2.13) can easily be extended with an ℓ_1-regularization to stimulate sparse solutions, since the vector $\mathbf{a}$ appears affinely in $M(U,\mathbf{a},\mathbf{b},\mathbf{y})$:

$$\min_{\mathbf{a}\in\mathbb{C}^{n_a}} f(\mathbf{a}) + \lambda \|\mathbf{a}\|_1, \tag{2.14}$$

for some regularization parameter λ.

Note that for $\mathbf{b} = -\mathbf{a}$,

$$\left\| M(U, \mathbf{a}, -\mathbf{a}, \mathbf{y}) \right\|_* = \left\| \mathbf{y} - |U\mathbf{a}|^2 \right\|_1 + n_y. \tag{2.15}$$

Since the result of optimization 2.13 might not produce a desired solution sufficiently fitting the measurements, we propose the iterative Convex Optimization-based Phase Retrieval (COPR) algorithm, outlined in Algorithm 1.

Algorithm 1 Convex Optimization-based Phase Retrieval (COPR)

1: **procedure** COPR($\mathbf{b}, \tau$) ▷ Some guess for $\mathbf{b}$
2: **while** $\left\| \mathbf{y} - |U\mathbf{a}|^2 \right\|_1 > \tau$ **do** ▷ Termination criterion
3: $\mathbf{a}_+ \in \arg\min_{\mathbf{a}} \left\| M(U, \mathbf{a}, \mathbf{b}, \mathbf{y}) \right\|_*$
4: $\mathbf{b}_+ \leftarrow -\mathbf{a}_+$
5: **end while**
6: **end procedure**

The nuclear norm is a convex function, and standard software like YALMIP [25] or CVX [26] can be used to concisely implement Algorithm 1. However, the nuclear norm minimization in Algorithm 1 is the main computational burden for an implementation. Usual implementations of the nuclear norm involve semidefinite constraints, and require a semidefinite optimization solver. If we assume that their computational complexity increases with $\mathcal{O}\left(n^6\right)$ [18] with constraint on matrices of size $n \times n$, then minimizing the nuclear norm of the matrix $M(U, \mathbf{a}, \mathbf{b}, \mathbf{y})$ of size $2n_y \times 2n_y$ is computationally infeasible even for relatively small-scale problems. Therefore, we propose a tailored ADMM algorithm of which the computational complexity of the iterations scales $\mathcal{O}\left(n_y n_a\right)$, and requires the inverse of a matrix of size $2n_a \times 2n_a$ for every iteration of Algorithm 1.

2.4. EFFICIENT COMPUTATION OF THE SOLUTION TO (2.13)

The minimization problem (2.13) can be reformulated as:

$$\min_{X, \mathbf{a}} \|X\|_* \quad \text{subject to} \quad X = M(U, \mathbf{a}, \mathbf{b}, \mathbf{y}). \tag{2.16}$$

Applying the ADMM optimization technique [15, 27] to the constraint optimization problem (2.16), we obtain the steps in Algorithm 2.

The advantage of using this ADMM formulation is that both of the update steps (2.17) and (2.18) have solutions that can be computed analytically. The efficient computation of the solutions are described in the following two subsections.

2.4.1. EFFICIENT COMPUTATION OF THE SOLUTION TO (2.17)

Upon inspection of (2.17), we see that this is a complex-valued standard least squares problem since $M(U, \mathbf{a}, \mathbf{b}, \mathbf{y})$ is parameterized affinely in $\mathbf{a}$. Let $\mathscr{R}(\cdot)$ and $\mathscr{I}(\cdot)$ respectively denote the real and the imaginary parts of a complex object. Let the subscripts $(\cdot)_1$, $(\cdot)_2$ and $(\cdot)_3$ respectively denote the top-left, top-right and bottom-left submatrices according to (2.11). Define

$$\mathbf{Z} = \mathbf{X} + \frac{1}{\rho}\mathbf{Y}, \quad X = \mathrm{d}\left(b^H U^H\right).$$

Algorithm 2 An ADMM algorithm for solving (2.16) based on [15]

procedure NN-ADMM($\mathbf{b}, \mathbf{y}, \rho, \tau$)
  $\mathbf{a} \leftarrow -\mathbf{b}$
  $\mathbf{X} \leftarrow M(U, \mathbf{a}, \mathbf{b}, \mathbf{y})$
  $\mathbf{Y} \leftarrow 0$
  while $\left| \left\| M(U, \mathbf{a}_+, \mathbf{b}, \mathbf{y}) \right\|_* - \left\| M(U, \mathbf{a}, \mathbf{b}, \mathbf{y}) \right\|_* \right| > \tau$ **do**
    $\mathbf{a}_+ \in$

$$\arg\min_{\mathbf{a}} \left\| \mathbf{X} - M(U, \mathbf{a}, \mathbf{b}, \mathbf{y}) + \frac{1}{\rho}\mathbf{Y} \right\|_F^2 \tag{2.17}$$

    $\mathbf{X}_+ \in$

$$\arg\min_{\mathbf{X}} \|\mathbf{X}\|_* + \frac{\rho}{2} \left\| \mathbf{X} - M(U, \mathbf{a}_+, \mathbf{b}, \mathbf{y}) + \frac{1}{\rho}\mathbf{Y} \right\|_F^2 \tag{2.18}$$

    $\mathbf{Y}_+ \leftarrow \mathbf{Y} + \rho\left(\mathbf{X}_+ - M(U, \mathbf{a}_+, \mathbf{b}, \mathbf{y})\right)$
    update ρ according to the rules in [27]
  end while
end procedure

In the sequel, let $\widehat{\mathrm{d}}(P)$ denote the vector with the diagonal entries of a square matrix P.

Reordering the elements in (2.17), separating the real and the imaginary parts, removing all matrix elements in the argument of the Frobenius norm that do not depend on $\mathbf{a}$, and vectorizing the result, gives the following least squares problem:

$$\min_{\mathbf{x}} \|\mathbf{u}_{ADMM} - \mathbf{u}_{COPR} - AB\mathbf{x}\|_2^2. \tag{2.19}$$

The variables $\mathbf{u}_{ADMM}$, $\mathbf{u}_{COPR}$, A, B and $\mathbf{x}$ are given by

$$\begin{aligned}
\mathbf{u}_{ADMM} &= \begin{pmatrix} \widehat{\mathrm{d}}(\mathscr{R}(\mathbf{Z}_1)) \\ \widehat{\mathrm{d}}(\mathscr{R}(\mathbf{Z}_2)) \\ \widehat{\mathrm{d}}(\mathscr{R}(\mathbf{Z}_3)) \\ \widehat{\mathrm{d}}(\mathscr{I}(\mathbf{Z}_2)) \\ \widehat{\mathrm{d}}(\mathscr{I}(\mathbf{Z}_3)) \end{pmatrix}, \qquad \mathbf{u}_{COPR} = \begin{pmatrix} \mathbf{y} + \widehat{\mathrm{d}}\left(|X|^2\right) \\ \widehat{\mathrm{d}}(\mathscr{R}(X)) \\ \widehat{\mathrm{d}}(\mathscr{R}(X)) \\ \widehat{\mathrm{d}}(\mathscr{I}(X)) \\ -\widehat{\mathrm{d}}(\mathscr{I}(X)) \end{pmatrix}, \\
A &= \begin{pmatrix} 2\mathscr{R}(X) & 2\mathscr{I}(X) \\ I & 0 \\ I & 0 \\ 0 & I \\ 0 & -I \end{pmatrix}, \qquad B = \begin{pmatrix} \mathscr{R}(U) & -\mathscr{I}(U) \\ -\mathscr{I}(U) & -\mathscr{R}(U) \end{pmatrix},
\end{aligned} \tag{2.20}$$

and $\mathbf{x} = \begin{pmatrix} \mathscr{R}(\mathbf{a})^T & \mathscr{I}(\mathbf{a})^T \end{pmatrix}^T$. This means that the optimal solution to (2.19) is given by

$$\mathbf{x}^* = (B^T A^T AB)^{-1} B^T A^T (\mathbf{u}_{ADMM} - \mathbf{u}_{COPR}).$$

During the ADMM iterations only $\mathbf{u}_{ADMM}$ changes. The inverse $(B^T A^T AB)^{-1}$ has to be computed once for every iteration of Algorithm 1 (i.e. it remains constant throughout the ADMM iterations). Since the complexity of computing an inverse is $\mathscr{O}\left(n^3\right)$ for matrices

of size $n \times n$, the computational complexity of this inverse process scales cubically with the number of basis functions.

Once this inverse matrix is obtained, the optimal solution to the least squares problem in (2.19) can be computed by a simple matrix-vector multiplication, whose complexity scales with $\mathcal{O}(n_y n_a)$.

Another approach that avoids the dense matrix-matrix multiplication in the computation of the inverse, is to use thin QR factorizations. Let

$$A = Q_A R_A, \quad B = Q_B R_B, \quad \text{and} \quad Q_C R_C = R_A Q_B. \tag{2.21}$$

B is dense and tall, but the (thin) QR factorization of B has to be computed only once. The QR factorization of the sparse A matrix is itself sparse (A is by permutation block diagonal with blocks of size 5 by 2). $R_A Q_B$ is dense and tall, but of smaller size than AB. The optimal solution can be computed using back substitution or using the pseudo-inverse of the triangular matrix $R_C R_B$, where both these matrices are of size $2n_a \times 2n_a$. The solution is given by

$$\mathbf{x}^* = (R_C R_B)^\dagger Q_C^T Q_A^T (\mathbf{u}_{ADMM} - \mathbf{u}_{COPR}) \tag{2.22}$$

Note that in the case that the objective term includes regularization as in (2.14), the optimization (2.19) should be modified appropriately to include the additive regularization term $\lambda \|\mathbf{a}\|_1$.

2.4.2. EFFICIENT COMPUTATION OF THE SOLUTION TO (2.18)

The optimization in (2.18) is of the form

$$\arg\min_X \|X\|_* + \lambda \|X - C\|_F^2. \tag{2.23}$$

The solution can be computed using singular value soft-thresholding, see [15, Theorem 2.1].

Let $C = U_C \Sigma_C V_C^T$ be the Singular Value Decomposition of $C \in \mathbb{C}^{2n_y \times 2n_a}$.

Lemma 2.4.1 ([28])**.** The solution $\mathbf{X}$ to (2.23) has singular vectors U_C and V_C.

Proof. Let $X = U_X \Sigma_X V_X^T$ be a Singular Value Decomposition of X. Then

$$\|X\|_* + \lambda \|X - C\|_F^2 = \operatorname{trace}(\Sigma_X) + \lambda (\langle X, X\rangle + \langle C, C\rangle - 2\langle X, C\rangle).$$

Using Von Neumann's trace inequality we get

$$\begin{aligned} &\min_X (\operatorname{trace}(\Sigma_X) + \lambda (\langle X, X\rangle + \langle C, C\rangle - 2\langle X, C\rangle)) \\ \geq &\min_X (\operatorname{trace}(\Sigma_X) + \lambda (\langle X, X\rangle + \langle C, C\rangle - 2\operatorname{trace}(\Sigma_X \Sigma_C))) \end{aligned}$$

with equality holds true when C and X are simultaneously unitarily diagonalizable. The optimal solution $\mathbf{X}$ to (2.23) therefore has the same singular vectors as C, i.e. $U_{\mathbf{X}} = U_C$, $V_{\mathbf{X}} = V_C$. □

Denote the singular values of C in descending order as $\sigma_{C,1},\ldots,\sigma_{C,2n_y}$, and those of X similarly. Thanks to Lemma 2.4.1, (2.23) can be simplified to

$$\arg\min_{\sigma_{X,i}} \sum_{i=1}^{2n_y} \left(\sigma_{X,i} + \lambda\left(\sigma_{X,i} - \sigma_{C,i}\right)^2\right). \tag{2.24}$$

This problem is completely decoupled in $\sigma_{X,i}$ and the optimal solution to (2.24) is computed with

$$\sigma_{\mathbf{X},i} = \max\left(0, \sigma_{C,i} - \frac{1}{2\lambda}\right), \quad i = 1,\ldots,2n_y.$$

By row and column permutations, the matrix C is block-diagonal with blocks of size 2×2. The Singular Value Decomposition (SVD) of this permuted matrix therefore involves block-diagonal matrices U_C, Σ_C and V_C and these blocks can be obtained separately and in parallel. Since the blocks are of size 2×2, the SVD can be obtained analytically.

This shows that a valid SVD can be computed very efficiently, in $\mathcal{O}(1)$. That is, in theory, in a computation time independent of the number of pixels in the image, the number of images taken or of the number of basis functions.

2.5. CONVERGENCE ANALYSIS OF ALGORITHM 1

Algorithm 1 can be reformulated as a Picard iteration $\mathbf{a}_{k+1} \in T(\mathbf{a}_k)$, where the fixed point operator $T : \mathbb{C}^{n_a} \to \mathbb{C}^{n_a}$ is given by

$$T(\mathbf{a}) = \arg\min_{\mathbf{x}\in\mathbb{C}^{n_a}} \left\| M(U,\mathbf{x},-\mathbf{a},\mathbf{y}) \right\|_*. \tag{2.25}$$

Our subsequent analysis will show that the set of fixed points, Fix T, the set of $\mathbf{a}$'s for which $\mathbf{a} = T(\mathbf{a})$, of T is in general nonconvex and as a result, iterations generated by T can not be Fejér monotone [29, Definition 5.1 of] with respect to Fix T. That is, each new iterate is not guaranteed to be closer to all fixed points in Fix T. Therefore, the widely known convergence theory based on the properties of Fejér monotone operators and averaging operators is not applicable to the operator T given at (2.25).

In this section, we make an attempt to prove convergence of Algorithm 1, which has been observed from our numerical experiments, via a relatively new developed convergence theory based on the theory of pointwise almost averaging operators [30]. It is worth mentioning that we are not aware of any other analysis schemes addressing convergence of Picard iterations generated by general nonaveraging fixed point operators. Our discussion consists of two stages. Based on the convergence theory developed in [30], we first formulate a convergence criterion for Algorithm 1 (Proposition 2.5.1) under rather abstract assumptions on the operator T. Due to the highly complicated structure of the nuclear norm of a general complex matrix, we are unable to verify these mathematical conditions for general matrices U. However, we will verify that they are well satisfied in the case that U is a unitary matrix (Theorem 2.5.2). From the latter result, we heuristically hope that Algorithm 1 still enjoys the convergence result when the matrix U is close to being unitary in a certain sense. In Section 2.6 we demonstrate that convergence is obtained in practice for the imaging case.

It is a common prerequisite for analyzing local convergence of a fixed point algorithm that the set of solutions to the original problem is nonempty. That is, there exists $\mathbf{a} \in \mathbb{C}^{n_a}$ such that $\mathbf{y} = |U\mathbf{a}|^2$. Before stating the convergence result, we need to verify that the fixed point set of T is nonempty.

Lemma 2.5.1. The fixed point operator T defined at (2.25) holds

$$\{\mathbf{a} \,|\, \mathbf{y} = |U\mathbf{a}|^2\} \subseteq \operatorname{Fix} T := \{\mathbf{a} \in \mathbb{C}^{n_a} \,|\, \mathbf{a} \in T(\mathbf{a})\}.$$

Proof. See Appendix A.1. □

The next proposition provides an abstract convergence result for Algorithm 1. Fix T is supposed to be closed. In the sequel, the metric projection associated with a set Ω is denoted P_Ω,

$$P_\Omega(x) := \{\omega \in \Omega \,|\, \|x - \omega\| = \operatorname{dist}(x, \Omega)\}, \quad \forall x.$$

Proposition 2.5.1. [30, simplified version of Theorem 2.2 of] Let $S \subset \operatorname{Fix} T$ be closed with $T(\mathbf{a}^*) \subset \operatorname{Fix} T$ for all $\mathbf{a}^* \in S$ and let W be a neighborhood of S. Suppose that T satisfies the following conditions.

(i) T is pointwise averaging at every point of S with constant $\alpha \in (0,1)$ on W. That is, for all $\mathbf{a} \in W$, $\mathbf{a}_+ \in T(\mathbf{a})$, $\mathbf{a}^* \in P_S(\mathbf{a})$ and $\mathbf{a}_+^* \in T(\mathbf{a}^*)$,

$$\|\mathbf{a}_+ - \mathbf{a}_+^*\|^2 \leq \|\mathbf{a} - \mathbf{a}^*\|^2 - \frac{1-\alpha}{\alpha} \|(\mathbf{a}_+ - \mathbf{a}) - (\mathbf{a}_+^* - \mathbf{a}^*)\|^2. \tag{2.26}$$

(ii) The set-valued mapping $\psi := T - \mathrm{Id}$ is metrically subregular on W for 0 with constant $\gamma > 0$, where Id is the Identity mapping. That is,

$$\gamma \operatorname{dist}(\mathbf{a}, \psi^{-1}(0)) \leq \operatorname{dist}(0, \psi(\mathbf{a})), \quad \forall \mathbf{a} \in W. \tag{2.27}$$

(iii) It holds $\operatorname{dist}(\mathbf{a}, S) \leq \operatorname{dist}(\mathbf{a}, \operatorname{Fix} T)$ for all $\mathbf{a} \in W$.

Then all Picard iterations $\mathbf{a}_{k+1} \in T(\mathbf{a}_k)$ starting in W satisfy $\operatorname{dist}(\mathbf{a}_k, S) \to 0$ as $k \to \infty$ at least linearly.

The pointwise property instead of the standard averaged property is imposed in (i) of Propositon 2.5.1 allows us to deal with the intrinsic nonconvexity of the fixed point set Fix T. The metric subregularity assumption imposed in (ii) technically ensures adequate progression of the iterates relative to the distance from the current iterate to the fixed point set. This is not only a technical assumption, but also a necessary condition for local linear convergence of a fixed point algorithm, Theorem 3.12 of [31]. Condition (iii) is, on one hand, a technical assumption and becomes redundant when $S = \operatorname{Fix} T$. On the other hand, the set S allows one to exclude from the analysis possible inhomogeneous fixed points of T, at which the algorithm often exposes weird convergence behavior [30, see Example 2.1 of].

The size of neighborhood W appearing in Proposition 2.5.1 indicates the robustness of the algorithm in terms of erroneous input (the distance from the starting point to a nearest solution).

We now apply the abstract result of Proposition 2.5.1 to the following special, but important case.

Theorem 2.5.2. Let $U \in \mathbb{C}^{n_a \times n_a}$ be unitary and $\mathbf{a}^* \in \mathbb{C}^{n_a}$ be such that $|U\mathbf{a}^*|^2 = \mathbf{y}$. Then every Picard iteration generated by Algorithm 1 $\mathbf{a}_{k+1} \in T(\mathbf{a}_k)$ starting sufficiently close to $\mathbf{a}^*$ converges linearly to a point $\tilde{\mathbf{a}} \in \mathrm{Fix}\ T$ satisfying $|U\tilde{\mathbf{a}}|^2 = \mathbf{y}$.

Proof. See Appendix A.2. □

2.6. NUMERICAL EXPERIMENTS

Four important numerical aspects of the COPR algorithm, including convergence, flexibility, complexity, and robustness, are tested on relevant problems. First we discuss convergence and the number of iterations of COPR and the ADMM algorithm. Second, we demonstrate the flexibility of the convex relaxation by comparing the COPR algorithm with an added ℓ_1-regularization to the PhaseLift method [13] and to the Compressive Sensing Phase Retrieval (CPRL) method in [12] on an under-determined sparse estimation problem. Then we compare the practically observed computational complexity of COPR and a naive implementation of PhaseLift [13]. Finally, we investigate the robustness of COPR relative to noise in a Monte-Carlo simulation for 25 and 100 basis functions. We compare four algorithms: COPR, PhaseLift [13], a basic alternating projections method (Section 4.3 in [13]) and an averaged projections method based on [32]. We note that the latter method fundamentally employs the Fourier transform at every iteration and hence is, in generally, not applicable for phase retrieval in the modal form.

2.6.1. CONVERGENCE

The while-loops in Algorithms 1 and 2 can be run for a fixed number of iterations. Figure 2.2 shows four such combinations for a typical problem with 5 images of size 256 × 256 and 64 basis functions. All cases are identically initialized with coefficients that best approximate a flat wavefront. As can be seen from the figure and the line with a square marker, only one COPR iteration is necessary here, as the ADMM algorithm slowly converges towards 0. However, stopping the ADMM algorithm after a limited number of iterations and having more than one COPR iteration can have a clear benefit, since faster convergence is achieved this way.

2.6.2. APPLICATION OF COPR TO COMPRESSIVE SENSING PROBLEMS

The first problem is to estimate 16 coefficients from 8 measurements, where the optimal vector is known to be sparse.

We generate a sparse coefficient vector $\mathbf{a}$ with two randomly generated non-zero complex elements. We generate two images ($n_d = 2,\ m = 128$) by applying two different amounts of defocus with Zernike coefficients $-\frac{\pi}{8}$ and $\frac{\pi}{8}$, respectively. From each image we use the center 2×2 pixels, resulting in a total of $n_y = 8$ measurements.

The applied algorithms are the COPR algorithm, the COPR algorithm with an additional ℓ_1-regularization, the PhaseLift algorithm [13] and the CPRL algorithm of [12]. The results are displayed in Figure 2.3. As can be seen from the figure, COPR and PhaseLift fail to retrieve the correct solution. The CPRL method and the regularized COPR algorithm compute the correct solution.

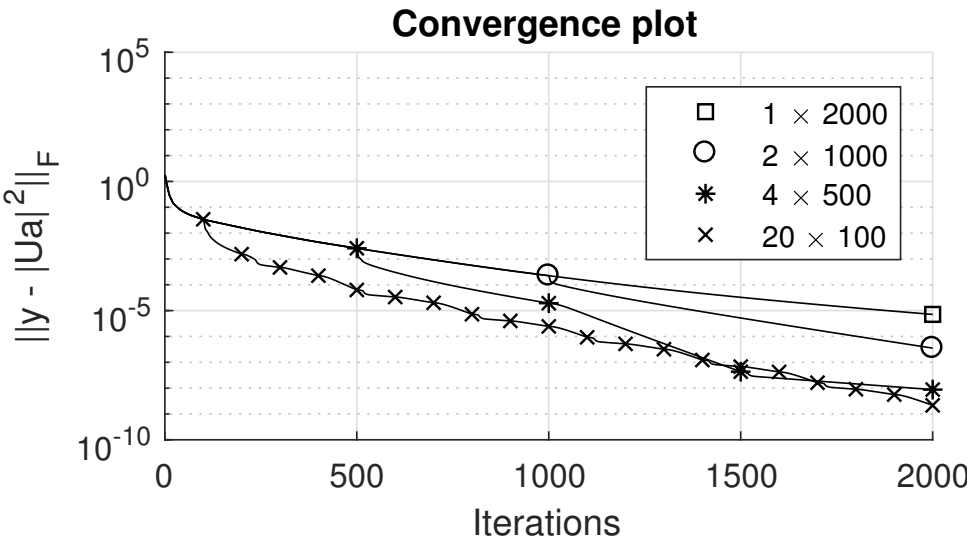


Figure 2.2: Convergence plot for 4 different combinations of COPR iterations and ADMM iterations. Denoted in the legend are first the number of COPR iterations, and then the number of ADMM iterations used to solve (2.16) in each COPR iteration. Markers denote each new COPR iteration.

2.6.3. COMPUTATIONAL COMPLEXITY

The second problem demonstrates the trends of the required computation time when the number of estimated coefficients increases. The underlying estimation problem consists of 7 images with different amounts of defocus applied as phase diversity, where each image is of size 64 by 64 pixels. A subset of 20 by 20 pixels of each image is used in the estimation. We compare the COPR algorithm to the PhaseLift algorithm, which is implemented according to optimization problem (2.5) in [13]. For PhaseLift, the reported time is the time it takes the MOSEK solver [17] to solve the optimization problem. This does not include the time taken by YALMIP [25] to convert the problem as given to the solver-specific form. For COPR, the initial guesses for the coefficients are drawn randomly from a Gaussian distribution, the number of iterations is set beforehand according to convergence to the correct solution, and the total time is recorded. The implementation of COPR does not exploit the parallelism referred to in Section 2.4.2. By convergence we mean that the estimated vector $\hat{\mathbf{a}}$ satisfies the tolerance criterion:

$$\min_{c\in\mathbb{C},\,|c|=1} \left\| c\hat{\mathbf{a}} - \mathbf{a}^* \right\|_2^2 \leq 10^{-5}, \tag{2.28}$$

where $\mathbf{a}^*$ is the exact solution.

The minimization over the parameter c ensures that the (unobservable) piston mode in the phase is canceled.[1] The computational complexity of PhaseLift is, as implemented, approximately $\mathcal{O}\left(n^4\right)$. The MOSEK solver ran into numerical issues for more than 25 estimated parameters. The COPR algorithm's computational complexity is approximately

[1] Let $\begin{pmatrix}\hat{\mathbf{a}} & \mathbf{a}^*\end{pmatrix} = QR$ be the QR decomposition. Then $\angle c^* = \angle \frac{R_{12}}{R_{11}}$.

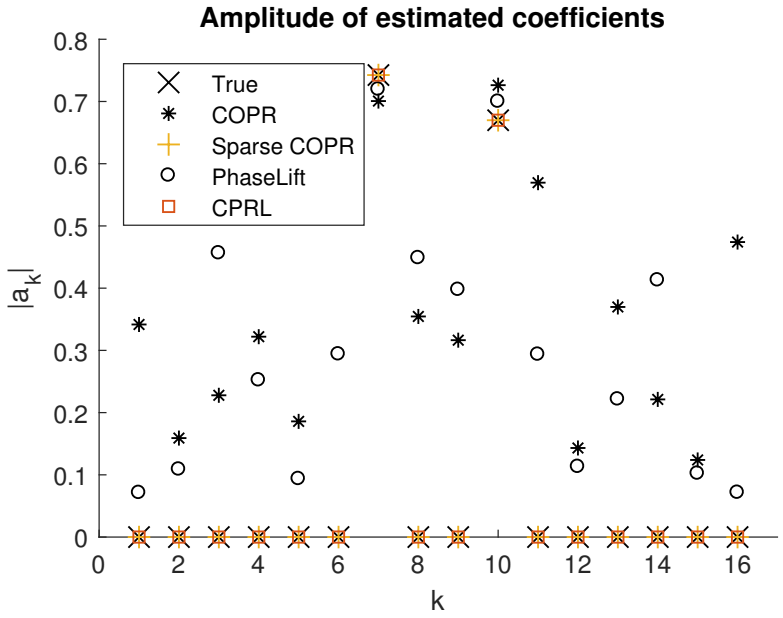


Figure 2.3: The absolute values of 16 estimated coefficients according to 4 different algorithms.

$\mathcal{O}(n)$. The better complexity is offset by a longer computation time for very small problems.

2.6.4. ROBUSTNESS TO NOISE

When estimating an unknown phase aberration, it is more logical to evaluate the performance of the algorithm on its ability to estimate the phase, and not the coefficients of basis functions.

We assume the phase is randomly generated with a deformable mirror. Let $H \in \mathbb{R}^{m^2 \times n_u}$ be the mirror's influence matrix and $\mathbf{u} \in \mathbb{R}^{n_u}$ be the input to the mirror's actuators, such that

$$\phi_{DM} = H\mathbf{u}. \tag{2.29}$$

The input values u_i are drawn from the uniform distribution between 0 and 1. The mirror has $n_u = 44$ actuators and the images have sides $m = 128$. The aperture radius is 0.4.

Five different defocus diversities are applied with Zernike coefficients uniformly spaced between $-\frac{\pi}{2}$ and $\frac{\pi}{2}$. Gaussian noise is added to the obtained images such that

$$\mathbf{y} = \max(0, \left|\mathcal{F}^{-1}\left\{\mathcal{P}_d(\rho,\theta)\right\}\right|^2 + \varepsilon),\ \varepsilon \in N(0, \sigma I). \tag{2.30}$$

and σ is the noise variance. No denoising methods were applied. The Signal-to-Noise

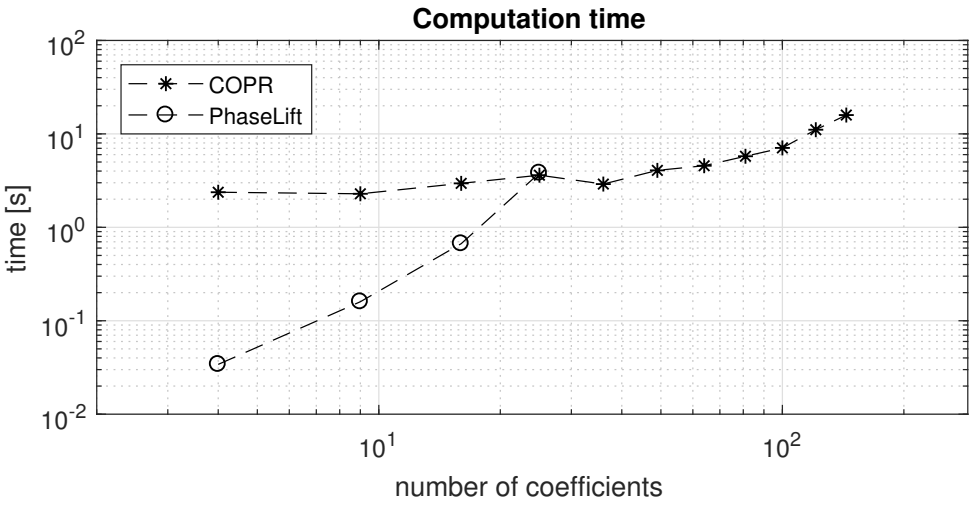


Figure 2.4: A computation time comparison between PhaseLift and COPR for different numbers of coefficients.

Ratio (SNR) is computed according to

$$10\log_{10}\frac{\left\|\mathbf{y}-\left|\mathscr{F}^{-1}\left\{\mathscr{P}_d(\rho,\theta)\right\}\right|^2\right\|_2^2}{\left\|\left|\mathscr{F}^{-1}\left\{\mathscr{P}_d(\rho,\theta)\right\}\right|^2\right\|_2^2}. \tag{2.31}$$

The phase is estimated from $\mathbf{y}$ using four different algorithms. The first is the COPR algorithm. The second is the averaged projections (AvP) algorithm [32]. The AvP algorithm is an extension of the well-known Gerchberg-Saxton algorithm [33] for solving problems with multiple images and is in the same class of algorithms as the Hybrid-Input-Output algorithm and the Difference Map [34, 35]. This makes this algorithm relevant for comparison. The third is the alternating projections (AlP) method ([13], Section 4.3), and the fourth algorithm is the PhaseLift method [13].

The COPR, PhaseLift and the AlP method are applied to estimate the phase using 25 basis functions, where the initial guesses for the coefficients are those coefficients that best approximate a flat wavefront. The AvP method is not based on the use of basis functions but on projection and the Fourier transform.

We make use of the Strehl ratio as a measure of optical quality. The Strehl ratio S is the ratio of the maximum intensity of the aberrated PSF and that of the unaberrated one and can be approximated with the expression of Mahajan:

$$S \approx e^{-\delta^2},$$

where $\delta = \left\|\phi_{DM} - \hat{\phi}\right\|_2$ and the mean residual phase has been removed [36].

For every noise level, 100 different phases were generated with the deformable mirror model (2.29). The results are presented in Figure 2.5. The resulting Strehl-ratio's are plotted with a trend line connecting the 50% quantiles. Figure 2.6 gives a qualitative comparison of the estimates for a single case.

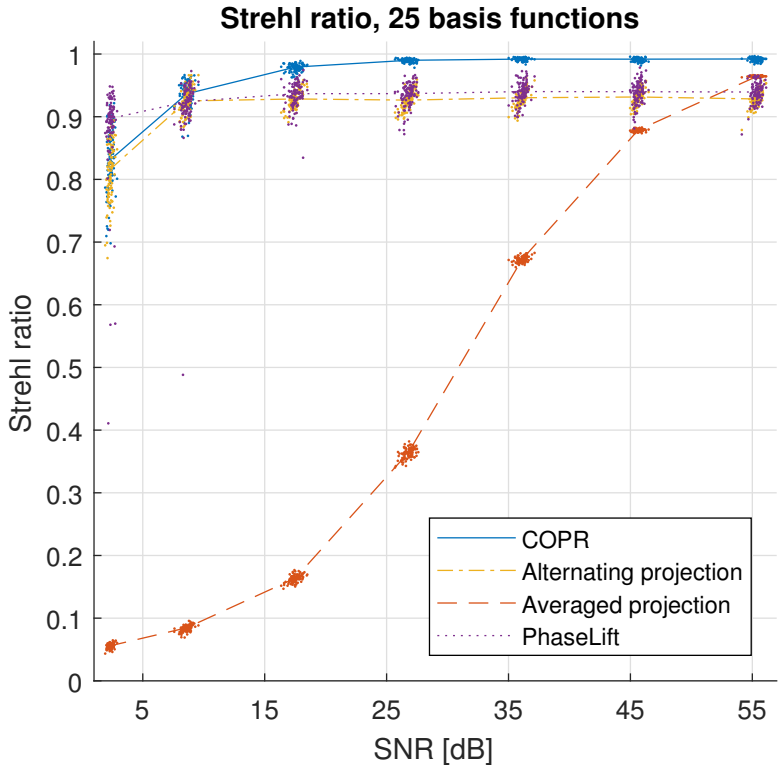


Figure 2.5: The Strehl ratio of the estimated phase aberration as a function of SNR.

In the case of PhaseLift, the tuning parameter that trades off measurement fit and the rank of the 'lifted' matrix is tuned once and applied to all problems. This has the effect that the reported performance is not as high as it could be with optimal tuning for individual problems. This points to another advantage of COPR: the absence of tuning parameters aside from the choice of basis functions.

The two figures show that COPR is robust to noise and gives accurate phase estimates for a wide range of noise levels.

2.7. EXPERIMENTAL VALIDATION

In this section we validate COPR on experimentally obtained data. The optical setup and the conditions of the data collection have been described in [37]. We thank Oleg Soloviev and Thao Nguyen for providing the data and the estimates of the DRAP algorithm as presented in [37].

The measured Point Spread Functions are displayed on the top row of Figure 2.7, where each PSF is an image with a size of 256×256 pixels. From each image 10000 pixels were used in the optimization ($n_y = 5 \cdot 10^4$). The number of basis functions employed is 225, where $\lambda = 60$. The COPR algorithm was run for 20 iterations, producing the fit that can be seen in Figure 2.8. The result of each COPR iteration was computed using

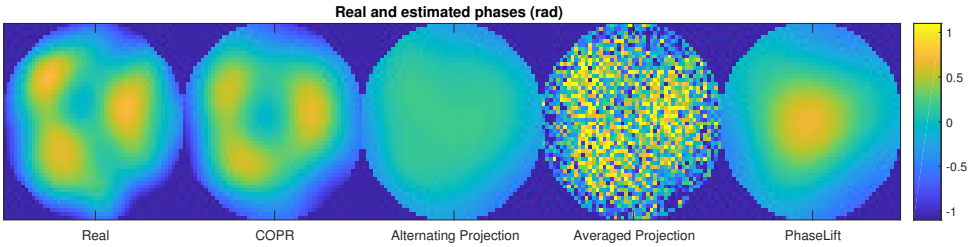


Figure 2.6: Example PSF and phase estimates of the COPR, Alternating Projection [13], Averaged Projection [32] and PhaseLift [13] algorithms for 3 PSF measurements with an SNR of approximately 36dB.

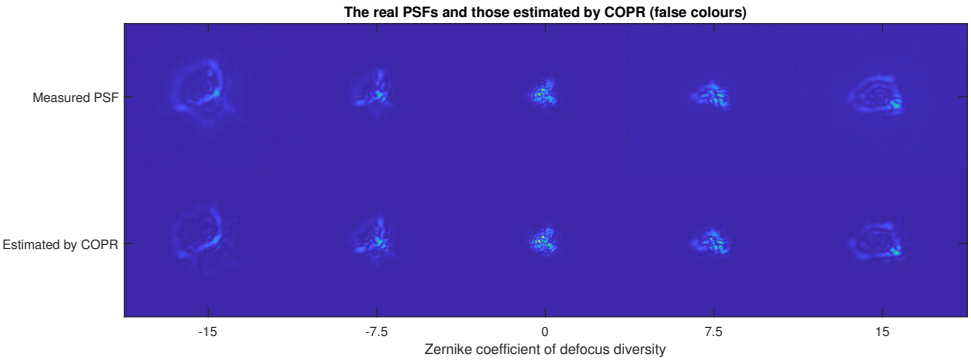


Figure 2.7: The measured Point Spread Functions (top) and the respective estimates (bottom). False colours are used according to $x \mapsto x^{0.9}$ for better visibility.

an ADMM optimization with 80 iterations. A breakdown of the computation time of the algorithm is recorded in Table 2.1. The implementation of COPR, that did not exploit parallelization opportunities, took approximately 10 minutes to compute the results reported here. As can be seen from the table, most time in an ADMM iteration (~90%) was spent updating the variable X, for which a large number of Singular Value Decompositionss were computed in a serial manner even though parallelization is an option.

The resulting coefficients produces estimated PSFs as shown in Figure 2.7 on the bottom row. As can be seen from the figure, the measured PSFs and estimated PSFs closely agree.

The estimated phase in the pupil plane is given in Figure 2.9, together with the measured phase, and the phase as estimated according to the DRAP method in [37]. As can be seen from the figure, the estimated phase by COPR resembles the phase as measured by the Shack-Hartmann wavefront sensor and the phase as estimated by the DRAP algorithm.

2.8. CONCLUDING REMARKS

The convex relaxations in solving the phase retrieval problem as proposed in (2.13) have the advantage over current convex relaxation methods, such as PhaseLift, that our strat-

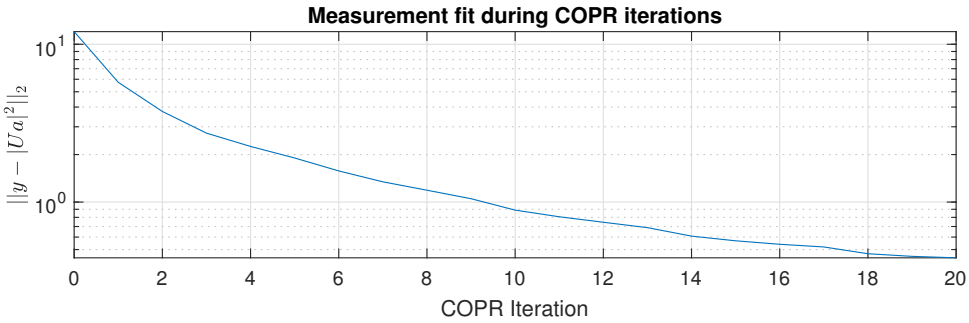


Figure 2.8: The estimates of the coefficients by the COPR algorithm iteratively produce a better fit with respect to the measurements.

Iteration	Update			overhead	total
	a	**X**	**Y**		
ADMM iteration	0.03	0.27	0.01	0.02	0.33
Nuclear norm	2.02	21.46	0.67	5.82	29.97
COPR total	40.42	429.28	13.39	116.21	599.30

Table 2.1: Approximate computation times for the COPR algorithm in seconds. The algorithm executed 20 nuclear norm minimizations and 80 ADMM iterations for every nuclear norm minimization. The time reported in the column with 'overhead' is the difference between the total time and the sum of the update times. It comprises for example computation time for pseudo-inverses and QR factorizations, that are required for the least squares problem that is the update of **a**. The total computation time is approximately 10 minutes (599 seconds).

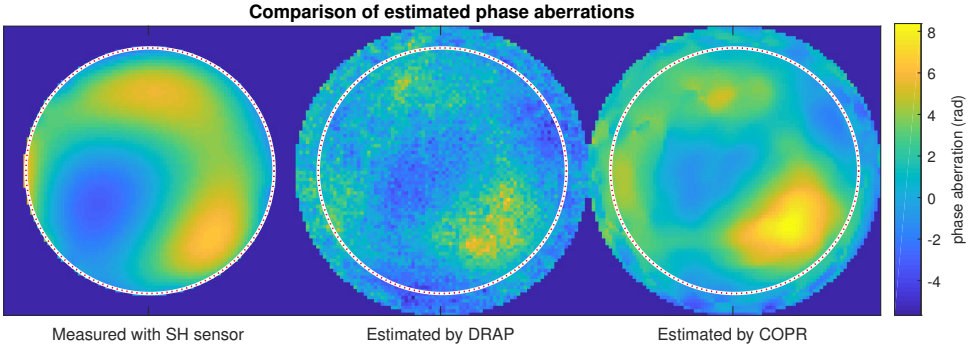


Figure 2.9: The measured phase (left), the estimate of the phase using the DRAP algorithm and the estimate of the phase using the COPR algorithm. The measured phase has been flipped and rotated by $\pi/2$ radians. The (relative) size of the aperture, the flipping and the rotation have been manually tuned.

egy is affine in the coefficients that are to be estimated. This allows for easy extension of the proposed method to phase retrieval problems that incorporate prior knowledge on the coefficients by regularization of the objective function. One such successful extension is the regularization with the ℓ_1-norm to find sparse solutions, as demonstrated in Figure 2.3.

In Section 2.4 an ADMM algorithm was proposed for efficient computation of the solution to (2.13). The result is that for the COPR algorithm a better computational complexity is observed compared to PhaseLift, see Figure 2.4. COPR is also able to solve phase estimation problems with larger numbers of parameters.

The required computations are favourable both in computation time and accuracy (they have simple analytic solutions) and in worst-case scaling behaviour $\mathcal{O}\left(n_y n_a\right)$ for every ADMM iteration, where n_y is the number of pixels and n_a is the number of basis functions.

We discussed convergence properties of the COPR algorithm in Section 2.5 and showed that for selected problems this convergence is linear or faster.

Finally, COPR has been shown to be robust against measurement noise, and outperform the two projection-based methods whose naive forms are often sensitive to noise as expected.

We are aware that in practice the performance of projection methods can be substantially better than what we have observed in this study provided that appropriate denoising techniques are also applied. Keeping aside from the matter of using denoising techniques, we have chosen to compare the algorithms in their very definition forms.

2.9. FUNDING INFORMATION

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement No. 339681.

REFERENCES

[1] R. Doelman, N. H. Thao, and M. Verhaegen, "Solving large-scale general phase retrieval problems via a sequence of convex relaxations," J. Opt. Soc. Am. A, vol. 35, pp. 1410–1419, Aug 2018.

[2] J. Antonello and M. Verhaegen, "Modal-based phase retrieval for adaptive optics," JOSA A, vol. 32, no. 6, pp. 1160–1170, 2015.

[3] Y. Shechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev, "Phase retrieval with application to optical imaging," IEEE Signal Processing Magazine, vol. May, pp. 87–109, 2015.

[4] J. R. Fienup, "Phase retrieval algorithms: a comparison," Applied optics, vol. 21, no. 15, pp. 2758–2769, 1982.

[5] D. R. Luke, J. V. Burke, and R. G. Lyon, "Optical wavefront reconstruction: theory and numerical methods," SIAM Rev., vol. 44, no. 2, pp. 169–224, 2002.

[6] Y. Shechtman, A. Beck, and Y. C. Eldar, "GESPAR: Efficient phase retrieval of sparse signals," IEEE transactions on signal processing, vol. 62, no. 4, pp. 928–938, 2014.

[7] D. Sayre, "Some implications of a theorem due to Shannon," Acta Crystallography [Online], vol. 5, no. 6, p. 843, 1952.

[8] H. Hauptman, "The direct methods of X-ray crystallography," Science, vol. 233, no. 4760, pp. 178–183, 1986.

[9] R. Gerchberg and W. Saxton, "A practical algorithm for the determination of phase from image and diffraction plane pictures," Optik, vol. 35, pp. 237–246, 1972.

[10] H. Bauschke, P. Combetters, and D. Luke, "Phase retrieval, error reduction algorithm and Fienup variants: a view from convex optimization," JOSA A, vol. 19, no. 7, pp. 1334–1345, 2002.

[11] E. J. Candes, X. Li, and M. Soltanolkotabi, "Phase retrieval via Wirtinger flow: Theory and algorithms," IEEE Transactions on Information Theory, vol. 61, no. 4, pp. 1985–2007, 2015.

[12] H. Ohlsson, A. Y. Yang, R. Dong, and S. S. Sastry, "Compressive phase retrieval from squared output measurements via semidefinite programming," arXiv preprint arXiv:1111.6323, 2011.

[13] E. J. Candes, T. Strohmer, and V. Voroninski, "Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming," Communications on Pure and Applied Mathematics, vol. 66, no. 8, pp. 1241–1274, 2013.

[14] M. Fazel, H. Hindi, and S. P. Boyd, "A rank minimization heuristic with application to minimum order system approximation," in American Control Conference, 2001. Proceedings of the 2001, vol. 6, pp. 4734–4739, IEEE, 2001.

[15] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," SIAM Journal on Optimization, vol. 20, no. 4, pp. 1956–1982, 2010.

[16] J. Goodman, Introduction to Fourier optics. McGraw-hill, 2008.

[17] MOSEK ApS, The MOSEK optimization toolbox for MATLAB manual. Version 7.1 (Revision 28)., 2015.

[18] L. Vandenberghe, V. R. Balakrishnan, R. Wallin, A. Hansson, and T. Roh, "Interior-point algorithms for semidefinite programming problems derived from the KYP lemma," Positive polynomials in control, pp. 579–579, 2005.

[19] A. Martinez-Finkelshtein, D. Ramos-Lopez, and D. Iskander, "Computation of 2D Fourier transforms and diffraction integrals using Gaussian radial basis functions," Applied and Computational Harmonic Analysis, 2016.

[20] P. J. Piscaer, A. Gupta, O. Soloviev, and M. Verhaegen, "Modal-based phase retrieval using Gaussian radial basis functions," In preparation, 2018.

[21] A. J. Janssen, "Extended Nijboer–Zernike approach for the computation of optical point-spread functions," JOSA A, vol. 19, no. 5, pp. 849–857, 2002.

[22] J. Braat, P. Dirksen, and A. J. Janssen, "Assessment of an extended Nijboer–Zernike approach for the computation of optical point-spread functions," JOSA A, vol. 19, no. 5, pp. 858–870, 2002.

[23] R. Doelman and M. Verhaegen, "Sequential convex relaxation for convex optimization with bilinear matrix equalities," in Proceedings of the European Control Conference, 2016.

[24] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," SIAM review, vol. 52, no. 3, pp. 471–501, 2010.

[25] J. Löfberg, "YALMIP : A toolbox for modeling and optimization in MATLAB," in In Proceedings of the CACSD Conference, (Taipei, Taiwan), 2004.

[26] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1." http://cvxr.com/cvx, Mar. 2014.

[27] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," Foundations and Trends in Machine Learning, vol. 3, no. 1, pp. 1–122, 2011.

[28] V. Larsson and C. Olsson, "Convex low rank approximation," International Journal of Computer Vision, vol. 120, no. 2, pp. 194–214, 2016.

[29] H. H. Bauschke and P. L. Combettes, Convex Analysis and Monotone Operator Theory in Hilbert Spaces. CMS Books Math./Ouvrages Math. SMC, New York: Springer, 2011.

[30] D. R. Luke, N. H. Thao, and M. K. Tam, "Quantitative convergence analysis of iterated expansive, set-valued mappings," Math. Oper. Res. to appear.

[31] D. R. Luke, M. Teboulle, and N. H. Thao, "Necessary conditions for linear convergence of iterated expansive, set-valued mappings with application to alternating projections." under peer-review.

[32] D. R. Luke, "Matlab Proxtoolbox," 2018. http://http://num.math.uni-goettingen.de/proxtoolbox/.

[33] R. W. Gerchberg and W. O. Saxton, "A practical algorithm for the determination of phase from image and diffraction plane pictures," Optik, vol. 35, p. 237, 1972.

[34] C.-C. Chen, J. Miao, C. Wang, and T. Lee, "Application of optimization technique to noncrystalline x-ray diffraction microscopy: Guided hybrid input-output method," Physical Review B, vol. 76, no. 6, p. 064113, 2007.

[35] V. Elser, "Solution of the crystallographic phase problem by iterated projections," Acta Crystallographica Section A: Foundations of Crystallography, vol. 59, no. 3, pp. 201–209, 2003.

[36] F. Roddier, Adaptive optics in astronomy. Cambridge university press, 1999.

[37] N. H. Thao, O. Soloviev, and M. Verhaegen, "Convex combination of alternating projection and Douglas-Rachford algorithm for phase retrieval," 08 2018.

3

IDENTIFICATION OF THE DYNAMICS OF TIME-VARYING PHASE ABERRATIONS FROM TIME HISTORIES OF THE POINT-SPREAD FUNCTION

To optimally compensate time-varying phase aberrations with Adaptive Optics (AO), a model of the dynamics of the aberrations is required to predict the phase aberration at the next time step. We model the time-varying behaviour of the phase aberration, expressed in Zernike modes, by assuming that the temporal dynamics of the Zernike coefficients can be described by a vector-autoregressive (Vector Auto-Regressive (VAR)) model. We propose an iterative method based on a convex heuristic for a rank constrained optimization problem, to jointly estimate the parameters of the VAR model and the Zernike coefficients from a time series of measurements of the Point Spread Function (PSF) of the optical system. By assuming the phase aberration is small, the relation between aberration and PSF measurements can be approximated by a quadratic function. As such, our method is a blind identification method for linear dynamics in a stochastic Wiener system with a quadratic nonlinearity at the output and a phase retrieval method that uses a time-evolution-model constraint and a single image at every time step.

This chapter has been published as a journal publication, “Reinier Doelman, Måns Klingspor, Anders Hansson, Johan Löfberg and Michel Verhaegen, Identification of the dynamics of time-varying phase aberrations from time histories of the point-spread function. Journal of the Optical Society of America A, 2019."

3.1. INTRODUCTION

Phase aberrations in optical systems cause blurring in the images taken with these systems. In order to improve the degraded image quality due to aberrations in an optical system, Adaptive Optics (AO) can be used to compensate for these aberrations online, or post-processing techniques can be used. For example, in (most) high performance telescopes these aberrations are wavefront (phase) aberrations induced by turbulence, misalignment, etc. To compensate for the phase aberration, both for online computations as well as for post processing of image data, information on this aberration is required. A classical method to obtain these phase aberrations is using a Shack-Hartmann (SH) wavefront sensor [1], see Figure 3.1. This sensor measures the spatial

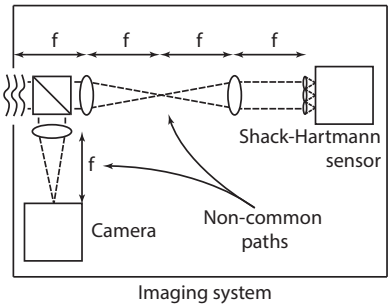


Figure 3.1: The classical optical setup for estimating temporal dynamics of an aberrated wavefront. This setup includes three lenses with focal length f.

derivatives of the wavefront and from these measurements the wavefront aberration itself can be estimated, as for example used in [2, 3] for (quasi-)static phase aberrations. To optimally compensate a dynamic aberration, a prediction of future aberrations has to be made. To predict the phase aberration at a future time step, a model is required which describes the (time) dynamics of the aberration. This model can be obtained from, for example, physical modelling [4, 5], which is not always possible and/or not always accurate. Also, [4] lists a number of different model assumptions. A different way to obtain a model is from identification [6, 7] based on data from the SH wavefront sensor. However, the use of a Shack-Hartmann wavefront sensor does not allow for the identification of (dynamic) non-common path errors; as can be seen in Figure 3.1, the optical paths between the incoming wavefront and the wavefront sensor, respectively the camera, are different. Any additional aberration that occurs in only one of the two paths gives a mismatch between the estimated aberration and the aberration as seen by the camera. This issue is encountered in for example astronomy [8–10] and ophthalmic imaging [11–13].

The phase aberration can be estimated from the camera measurements by using phase retrieval methods [2, 3, 9, 10, 14, 15]. These techniques use two (or more) images, of which one usually has an added phase diversity. The phase diversity is often a defocus [16, 17], since this diversity image can be obtained by simply moving the camera out-of-focus. From the time series of estimated phase aberrations a model can be identified using the same techniques as with the SH wavefront sensor measurements. However, taking multiple images with different phase diversities but the same aberration, might be

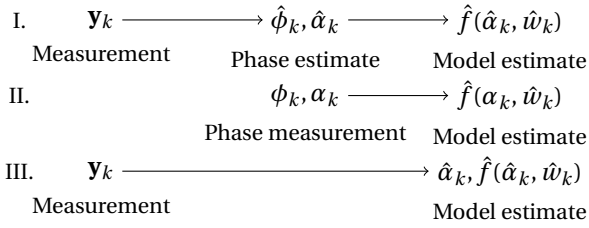


Figure 3.2: An overview of identification methods. $\mathbf{y}$ denotes the measurement, ϕ is the phase aberration, α is the vector of Zernike coefficients, w a noise signal, k is a time index, and $f(\cdot)$ is the model function. I. Phase-retrieval first, then model estimate. II. Model estimate based on phase measurements. III. Our method: model estimation and phase retrieval from PSF measurements.

impossible in a dynamical setting, without introducing non-common paths in the setup. In addition to the drawback of taking multiple images, the approach to first reconstruct at each time instance the phase aberration and subsequently modeling the temporal dynamics of these reconstructed aberrations may suffer from accumulation of the estimation errors in the two–step process. Identification of the phase aberration dynamics based on measurements by the camera directly, has not yet been investigated. The prior knowledge that the phase aberration behaves in a non-specific dynamical manner (i.e. the dynamics are contained in a specific model set), has to our knowledge not previously been incorporated in a phase aberration estimation procedure.

In this chapter we introduce a method to both estimate the phase aberrations and the aberration dynamics directly from intensity measurements of the PSF, when the phase aberration dynamics can be described with a Vector Auto-Regressive (VAR) model, driven by a stochastic input, see Figure 3.2. Hereby we circumvent the problem of non-common paths by requiring multiple images for every time-step and avoid the accumulation of errors that the two-step process has.

To accomplish this, we assume the phase aberrations are small and the PSF can be well-approximated by a second-order Taylor series expansion of the image intensities as a function of the Zernike coefficients of the wavefront aberration. We show how the estimation problem has a convex heuristic from which we can iteratively estimate both the phase aberrations and their temporal dynamics.

Since we estimate both the aberrations and dynamics, one way to view this method is as a system identification method with PSF data. A second way to view this method is as a phase retrieval method with a time-evolution model constraint in the pupil plane. In this sense it also differs from the method in [18, 19], where linear, but known, constraints on the phase aberration are used to reduce the parameter search space; our constraints are bilinear. The available literature applicable to specifically noise driven linear systems with nonlinear outputs, seems to be quite sparse. In [20] an identification method is proposed for blind identification of Wiener systems, but an invertible nonlinearity is assumed. Applicable identification methods for these type of systems are based on a Bayesian approaches [21–24] using Maximum Likelihood (ML) and Expectation Maximization (EM) algorithms to jointly estimate the dynamics and nonlinearity itself. However, since the type of nonlinearity is known, we exploit the fact that our esti-

mation problem has a convex heuristic in the sought-after parameters.

Article overview In Section 3.2 we set out the mathematical notation, the optical conventions and the problem statement. Section 3.3 contains a reformulation of the estimation problem, and introduces its convex heuristic. Furthermore, we propose an iterative algorithm that uses the heuristic to compute an estimated model and phase. In Section 3.4 we conduct a numerical experiment to compare the performance of our method to two straightforward approaches. Section 3.5 contains the conclusion and some suggestions for future research.

3.1.1. NOTATION

In this paper we make use of the vectorization function $\text{vect}(\cdot) : X \to \overline{x}, X \in \mathbb{R}^{m\times n}, \overline{x} \in \mathbb{R}^{mn}$, i.e. a linear transformation of a matrix X into a column vector $\overline{x}$, stacking the columns of X. This transformation is invertible, and we thus also define $\text{vect}^{-1}(\cdot) : \overline{x} \to X$. The nuclear norm of a matrix X is defined as $\|X\|_* = \sum_i \sigma_i(X)$, the sum of the singular values of X. $\|X\|_F$ denotes the Frobenius norm of X. As a performance measure, given a true sequence $\{\alpha_k\}_{k=1}^K$ and an estimated sequence $\{\hat{\alpha}_k\}_{k=1}^K$, we define

$$\text{VAF}(\alpha, \hat{\alpha}) = \max\left(0, \left(1 - \frac{\sum_{k=1}^K \|\alpha_k - \hat{\alpha}_k\|_2^2}{\sum_{k=1}^K \|\alpha_k\|_2^2}\right)\right) \tag{3.1}$$

where VAF is abbreviation for Variance Accounted For. Finally, given matrices $X_1, ..., X_n$ we define the matrix direct sum as $\bigoplus_{n=1,...,N} X_n := \text{blkdiag}(X_1, X_2, \ldots, X_N)$, so e.g.

$$\bigoplus_{n=1,2} X_n := \text{blkdiag}(X_1, X_2) = \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix}. \tag{3.2}$$

3.2. PROBLEM DESCRIPTION

3.2.1. LINEAR AND QUADRATIC APPROXIMATIONS OF THE PSF FOR SMALL PHASE ABERRATIONS

We follow the description of the optical setup in [25], where a linear quadratic controller is designed using quadratic output measurements based on Taylor series expansions of the PSF. The controller is based on an Linear Time-Invariant (LTI) model of the disturbance. We consider the same optical problem and Taylor approximation setting, but focus on the model identification.

The Point Spread Function of an optical system is the inverse Fourier transform of the Generalized Pupil Function (GPF). The GPF is the complex-valued function

$$\mathscr{P}(x, y) = \mathbf{A}(x, y) \exp\big(j\phi(x, y)\big), \tag{3.3}$$

where (x, y) are the Cartesian coordinates in the pupil plane, $\mathbf{A}(x, y)$ is the real-valued aperture apodisation function, $\phi(x, y)$ is the real-valued phase function, and $j^2 = -1$. We assume that $\phi(x, y) = \phi_a(x, y) + \phi_d(x, y)$ where $\phi_a(x, y)$ is the phase aberration function and $\phi_d(x, y)$ is the phase diversity function. We assume that $\phi_a(x, y)$ can be well-

approximated by a weighted sum of Zernike basis functions,

$$\overline{\phi}_a(x,y,\alpha) := \sum_{r=1}^{s} Z_r(x,y)\alpha_r \tag{3.4}$$

where $Z_r(x,y)$ is the r'th basis function, and $\alpha_r \in \mathbb{R}$ are the weights. Similarly, but with different weights, $\overline{\phi}_d(x,y,\beta)$ approximates $\phi_d(x,y)$. Hence

$$\overline{\phi}(x,y,\alpha,\beta) = \overline{\phi}_a(x,y,\alpha) + \overline{\phi}_d(x,y,\beta)$$

and it follows from the definition in (3.4) that

$$\overline{\phi}(x,y,\alpha,\beta) = \overline{\phi}(x,y,\alpha+\beta,0)$$

because of linearity in the weights. Now, we define a grid of points $\tilde{x} \times \tilde{y}$ where $\tilde{x} = \{x_1,\ldots,x_m\}$, $\tilde{y} = \{y_1,\ldots,y_m\}$,. Over this grid, we define

$$\Phi(\alpha,\beta) = \begin{bmatrix} \overline{\phi}(x_1,y_1,\alpha,\beta) & \ldots & \overline{\phi}(x_m,y_1,\alpha,\beta) \\ \vdots & \ldots & \vdots \\ \overline{\phi}(x_1,y_m,\alpha,\beta) & \ldots & \overline{\phi}(x_m,y_m,\alpha,\beta) \end{bmatrix} \tag{3.5}$$

and with this definition we can express the vectorization of $\Phi(\alpha,\beta)$ as a matrix multiplication

$$\operatorname{vect}\big(\Phi(\alpha,\beta)\big) = Z(\alpha+\beta), \tag{3.6}$$

where $Z \in \mathbb{R}^{m^2 \times s}$ for a matrix Z composed of $Z_r(x_k,y_k)$. Similarly we define

$$\Gamma = \begin{bmatrix} \mathbf{A}(x_1,y_1) & \ldots & \mathbf{A}(x_m,y_1) \\ \vdots & \ldots & \vdots \\ \mathbf{A}(x_1,y_m) & \ldots & \mathbf{A}(x_m,y_m) \end{bmatrix} \tag{3.7}$$

The complex field in the imaging plane with incoherent illumination is the inverse Fourier transform of the GPF. Taking intensity measurements with a noise-free camera gives the PSF, the squared amplitude of this field:

$$\mathbf{y}(\alpha,\beta) = \operatorname{vect}\left(\left|\mathscr{F}^{-1}\left\{\Gamma \odot \exp\left(j \operatorname{vect}^{-1}\left(Z(\alpha+\beta)\right)\right)\right\}\right|^2\right), \tag{3.8}$$

where $\mathbf{y}(\alpha,\beta) \in \mathbb{R}_+^{p^2}$ and p^2 is the number of pixels, $\odot$ denotes the Hadamard product, $\exp(\cdot)$ denotes the element-wise exponential function and $|\cdot|^2$ denotes the square of the absolute value of the elements of the matrix.

A linear approximation of the PSF measurements for the i'th pixel is given by a first-order Taylor expansion of a small aberration α about the diversity β [25]:

$$\mathbf{y}_i(\alpha,\beta) = D_{0,i}(\beta) + D_{1,i}(\beta)\alpha + \mathscr{O}\left(\|\alpha\|^2\right), \tag{3.9}$$

where $\mathscr{O}\left(\|\alpha\|^2\right)$ denotes terms of order 2 and higher. The matrices $D_{0,i}$ and $D_{1,i}$ are given by

$$\begin{aligned} D_{0,i}(\beta) &= \mathbf{y}_i(\alpha,\beta)\big|_{\alpha=0} && \in \mathbb{R}, \\ D_{1,i}(\beta) &= \left.\frac{\partial \mathbf{y}_i(\alpha,\beta)}{\partial \alpha}\right|_{\alpha=0} && \in \mathbb{R}^{1\times s}. \end{aligned} \tag{3.10}$$

The first-order approximation has limited approximation power. For larger aberrations a second order Taylor expansion can be used [25],

$$\begin{aligned}\mathbf{y}_i(\alpha,\beta) &= D_{0,i}(\beta) + D_{1,i}(\beta)\alpha \\ &\quad + \frac{1}{2}\alpha^T D_{2,i}(\beta)\alpha + \mathcal{O}\left(\|\alpha\|^3\right),\end{aligned} \tag{3.11}$$

where

$$D_{2,i}(\beta) = \left.\frac{\partial^2 \mathbf{y}_i(\alpha,\beta)}{\partial \alpha^T \partial \alpha}\right|_{\alpha=0} \in \mathbb{R}^{s\times s}. \tag{3.12}$$

Since the quadratic approximation holds for aberrations of larger magnitudes, and an identification method that is designed for this approximation would therefore be valid for a larger number of cases, we continue with the model with a quadratic approximation of the PSF and assume $D_{2,i}$ to be non-zero. We use Zernike polynomials normalized to 1 radian amplitude. To give an indication of the validity of the approximation, consider the Zernike modes with OSA/ANSI-index 3 to 9. Drawing Zernike coefficients from a normal distribution, the quadratic approximation of the PSF is a good approximation with

$$\mathrm{VAF}(y_i(\alpha,\beta) - D_{0,i}(\beta), D_{1,i}(\beta)\alpha + \frac{1}{2}\alpha^T D_{2,i}(\beta)\alpha)) > 0.9, \tag{3.13}$$

where VAF stands for Variance Accounted For (defined in the notation section), for $\|\alpha\|_2 <$ 1.0 to 1.4 with a defocus diversity β ranging between 0 and 0.5. The linear approximation is invalid without defocus and only valid up to $\|\alpha\|_2 < 0.3$ for $\beta = 0.5$. This trend also holds for similar values of β. Aberrations of small magnitudes can for example be encountered in adaptive optics systems operating in closed loop. See also [26] for a discussion.

3.2.2. VAR MODELS AND THE IDENTIFICATION PROBLEM

We assume that the total phase $\Phi(\alpha,\beta)$ is time dependent. In vectorized form this becomes

$$\mathrm{vect}\left(\Phi(\alpha_k,\beta_k)\right) = Z(\alpha_k + \beta_k), \tag{3.14}$$

where we use k as the time index. Similarly, the i'th pixel at time k is denoted with $\mathbf{y}_i(\alpha_k,\beta_k)$.

The assumption on the model structure is that the vector of Zernike coefficients of the phase aberration evolves according to a vector valued autonomous auto-regressive model of order N (VAR(N)):

$$\begin{aligned}\alpha_k &= f(\alpha_{k-1},\ldots,\alpha_{k-N},w_k) \\ &= A_1\alpha_{k-1} + \ldots + A_N\alpha_{k-N} + w_k,\end{aligned} \tag{3.15}$$

where $A_n \in \mathbb{R}^{s\times s}$ are coefficient matrices and $w_k \in \mathbb{R}^s$ is driving, white noise. This is a common dynamic model for, for example, turbulent phase [4, 5, 27, 28].

The system that generates the measurements $\left\{\mathbf{y}_i(\alpha_k,\beta_k)\right\}_{k=1,\ldots,K}^{i=1,\ldots,p^2}$ becomes

$$\begin{aligned}\alpha_k &= f(\alpha_{k-1},\ldots,\alpha_N,w_k) \\ &= A_1\alpha_{k-1} + \ldots + A_N\alpha_{k-N} + w_k, \\ \mathbf{y}_i(\alpha_k,\beta_k) &= D_{0,i}(\beta_k) + D_{1,i}(\beta_k)\alpha_k + \alpha_k^T D_{2,i}(\beta_k)\alpha_k + v_{i,k},\end{aligned} \tag{3.16}$$

where $v_{i,k}$ is a noise signal consisting of measurement noise and the approximation error $\mathcal{O}\left(\|\alpha\|^3\right)$ in the Taylor expansion of $\overline{\phi}_a$.

The identification problem to find $\{\alpha_k\}_{k=1}^K$, $\{A_n\}_{n=1}^N$, $\{w_k\}_{k=1}^K$ and $\left\{v_{i,k}\right\}_{i=1,\dots,p^2}^{k=1,\dots,K}$ in (3.16) is cast into a minimization problem:

$$\begin{aligned}
&\text{minimize} && \sum_{i=1}^{K} \|w_k\|_2^2 + \gamma \sum_{k=1}^{K} \sum_{i=1}^{p^2} \left\|v_{i,k}\right\|_2^2 \\
&\text{over} && A_n, \alpha_k, w_k, v_{i,k} \\
&\text{subject to} && \alpha_k = f(\alpha_{k-1}, \dots, \alpha_N, w_k), \\
& && \mathbf{y}_i(\alpha_k, \beta_k) = D_{0,i}(\beta_k) + D_{1,i}(\beta_k)\alpha_k \\
& && \quad + \alpha_k^T D_{2,i}(\beta_k)\alpha_k + v_{i,k}, \\
&\text{for} && n = 1, \dots, N,\ i = 1, \dots, p^2, \\
& && k = 1, \dots, K,
\end{aligned} \tag{3.17}$$

for a trade-off parameter $\gamma \in \mathbb{R}_+$. This formulation can be seen as a generalization of a standard state reconstruction problem (see for example [29]), where the difference lies in the quadratic term in the output and the unknown parameter values of the model.

In the following section, we reformulate the equality constraints, which are both bilinear, into rank constraints. Subsequently we use a heuristic formulation for the rank constraints and create a convex optimization problem.

3.3. BLIND IDENTIFICATION FROM QUADRATIC MEASUREMENTS

3.3.1. REFORMULATING (3.17) INTO A RANK CONSTRAINED PROBLEM

The time-evolution of the Zernike coefficients can be written as a matrix equation in the following manner.

$$\underbrace{\begin{pmatrix}\alpha_K & \dots & \alpha_{N+1}\end{pmatrix}}_{\mathscr{A}_K} = \underbrace{\begin{pmatrix}A_1 & \dots & A_N\end{pmatrix}}_{A} \underbrace{\begin{pmatrix}\alpha_{K-1} & \alpha_{K-2} & \dots & \alpha_N \\ \alpha_{K-2} & \alpha_{K-3} & \dots & \alpha_{N-1} \\ \vdots & \vdots & \vdots & \vdots \\ \alpha_{K-N} & \dots & \dots & \alpha_1\end{pmatrix}}_{\mathscr{H}} + W \tag{3.18}$$

where $\mathscr{H}$ is a Hankel matrix and $W = \begin{pmatrix}w_K & \cdots & w_{N+1}\end{pmatrix}$. Now, the measurement equations in (3.16) can be rearranged to

$$\mathbf{y}_i(\alpha_k, \beta_k) - D_{0,i}(\beta_k) - D_{1,i}(\beta_k)\alpha_k - v_{i,k} = \alpha_k^T D_{2,i}(\beta_k)\alpha_k. \tag{3.19}$$

Furthermore, let $\hat{W}$ be the estimate of W and

$$\begin{aligned} D_y &= \bigoplus_{i,k} \mathbf{y}_i(\alpha_k,\beta_k) - D_{0,i}(\beta_k) - D_{1,i}(\beta_k)\alpha_k, \\ D_\alpha &= \bigoplus_{i,k} \alpha_k, \\ D_2 &= \bigoplus_{i,k} D_{2,i}(\beta_k), \\ \hat{V} &= \bigoplus_{i,k} v_{i,k}. \end{aligned} \tag{3.20}$$

Now, (3.17) can be rewritten compactly as

$$\begin{aligned} &\text{minimize} && \|\hat{W}\|_F^2 + \gamma \|\hat{V}\|_F^2 \\ &\text{over} && \alpha_k,\ A,\ \hat{W},\ \hat{V} \\ &\text{subject to} && \mathscr{A}_K - \hat{W} = A\mathscr{H} \\ & && D_{\mathbf{y}} - \hat{V} = D_\alpha^T D_2 D_\alpha, \end{aligned} \tag{3.21}$$

The optimization problem in (3.21) is an optimization problem with two bilinear equality constraints. Following [30], we will convert these constraints into equivalent rank constraints using Lemma 3.3.1.

Lemma 3.3.1 ([30])**.** Let the matrix-valued function $L(\cdot)$ be defined as

$$\begin{aligned} &L(A,P,B,C,\mathbf{X},\mathbf{Y}) = \\ &\begin{pmatrix} C + AP\mathbf{Y} + \mathbf{X}PB + \mathbf{X}P\mathbf{Y} & (A+\mathbf{X})P \\ P(B+\mathbf{Y}) & P \end{pmatrix}. \end{aligned} \tag{3.22}$$

For this matrix it holds that

$$\operatorname{rank} L(A,P,B,C,\mathbf{X},\mathbf{Y}) = \operatorname{rank} P \iff APB = C \tag{3.23}$$

for any choice of $\mathbf{X},\mathbf{Y}$ and any non-zero P of appropriate size.

Define now the two matrices M_{VAR} and M_{meas}:

$$\begin{aligned} M_{VAR} &:= L\left(A, I_{Ns}, \mathscr{H}, \mathscr{A}_K - \hat{W}, \mathbf{X}_1, \mathbf{Y}_1\right), \\ M_{meas} &:= L\left(D_\alpha^T, D_2, D_\alpha, D_{\mathbf{y}} - \hat{V}, \mathbf{X}_2, \mathbf{Y}_2\right). \end{aligned} \tag{3.24}$$

Here I_{Ns} is an identity matrix of size $Ns \times Ns$. Applying Lemma (3.3.1) to the two constraints in problem (3.21) gives us

$$\begin{aligned} \operatorname{rank} M_{VAR} &= \operatorname{rank} I_{Ns} = Ns \\ \operatorname{rank} M_{meas} &= \operatorname{rank} D_2 \end{aligned} \tag{3.25}$$

as equivalent constraints. Problem (3.21) can now be formulated as

$$\begin{aligned} &\text{minimize} && \|\hat{W}\|_F^2 + \gamma \|\hat{V}\|_F^2 \\ &\text{over} && \alpha_k,\ A,\ \hat{W},\ \hat{V} \\ &\text{subject to} && \operatorname{rank} M_{VAR} = \operatorname{rank} I_{Ns} = Ns \\ & && \operatorname{rank} M_{meas} = \operatorname{rank} D_2 \end{aligned} \tag{3.26}$$

3.3.2. A CONVEX HEURISTIC FOR (3.26)

Rank constrained problems (or problems with bilinear matrix equalities) are in general Non-deterministic Polynomial-time (NP)-hard to solve [31]. The proposed solution is to solve a convex heuristic for the problem by adding the sum of the nuclear norms of the matrices M_{VAR} and M_{meas} in (3.25) to the objective function and verifying their rank afterwards. The fact that the two matrices are affinely parameterized by the decision variables A, α_k, w_k and $v_{i,k}$, even though problem (3.17) is not, allows the application of the nuclear norm to make the problem convex. The advantage of using the nuclear norm is that standard software like YALMIP [32] or CVX [33] is available to implement the convex optimization problem. Alternatively to employing the nuclear norm, other rank-minimizing heuristics could be applied, like for example the use of the truncated nuclear norm, [34].

We introduce a regularization parameter λ to weigh the nuclear norms, following from the two rank constraints in (3.26), against each other and a parameter ξ to weigh the original objective function with the low rank inducing terms, and obtain the convex problem:

$$\underset{A,\alpha_k,\hat{W},\,\hat{V}}{\text{minimize}} \quad \left\|\hat{W}\right\|_F^2 + \gamma\left\|\hat{V}\right\|_F^2 + \xi\left(\lambda||M_{VAR}||_* + ||M_{meas}||_*\right). \tag{3.27}$$

In this optimization problem $\hat{W}$ appears in the first and third term (see (3.24)) and $\hat{V}$ likewise in the second and third term, and there are three parameters (γ,ξ,λ) to tune.

We found it more efficient to work with the following simplified optimization problem with only one single regularization parameter. Define the two matrices Q_{VAR} and Q_{meas}:

$$\begin{aligned} Q_{VAR} &:= L\left(A, I_{Ns}, \mathscr{H}, \mathscr{A}_K, \mathbf{X}_1, \mathbf{Y}_1\right), \\ Q_{meas} &:= L\left(D_\alpha^T, D_2, D_\alpha, D_{\mathbf{y}}, \mathbf{X}_2, \mathbf{Y}_2\right). \end{aligned} \tag{3.28}$$

The objective function is simplified to

$$\underset{A,\alpha_k}{\text{minimize}} \quad \lambda||Q_{VAR}||_* + ||Q_{meas}||_*. \tag{3.29}$$

The noise terms $\hat{W}$ and $\hat{V}$ are simply interpreted as the feasibility gap of the Bilinear Matrix Equalitys (BMEs) with the optimal A^* and α_k^*,

$$\begin{aligned} \hat{W}^* &:= \mathscr{A}_K^* - A^*\mathscr{H}^* \\ \hat{V}^* &:= D_{\mathbf{y}} - \left(D_\alpha^*\right)^T D_2 D_\alpha^*. \end{aligned} \tag{3.30}$$

We observe (3.29) minimizes the feasibility gap (interpreted as the norms of $\hat{W}$ and $\hat{V}$), and we therefore drop the two terms in (3.27) that have become redundant.

Since the problem in (3.29) is convex in the parameters A and α_k, it is easy to include several forms of prior information through the use of convex constraints, or regularization of the objective function. Examples are constraints expressing an affine parameter dependence of the matrix A, or the inclusion of an additional term to the objective function such as $\mu\|A\|_F^2$, for some regularization parameter μ, to prevent elements of the matrix A from having large magnitudes.

The optimization in (3.29) can be performed for different choices of the parameters $\mathbf{X}_1, \mathbf{Y}_1, \mathbf{X}_2$ and $\mathbf{Y}_2$. This freedom can be used in an iterative manner, as outlined in Algorithm 3.

Algorithm 3 Sequential Convex Optimization-based Identification (SCOBI)

1: **procedure** SCOBI
2: **while** not converged **do**
3: Solve (3.29) with parameters $\mathbf{X}_1, \mathbf{Y}_1, \mathbf{X}_2, \mathbf{Y}_2$ to obtain optimal $A^*, \mathscr{A}_K^*, D_\alpha^*$ and $D_\mathbf{y}^*$.
4: Set

$$\begin{aligned} \mathbf{X}_1^+ &= -A^*, & \mathbf{Y}_1^+ &= -\mathscr{A}_K^*, \\ \mathbf{X}_2^+ &= -(D_\alpha^*)^T, & \mathbf{Y}_2^+ &= -D_\alpha^*. \end{aligned} \tag{3.31}$$

5: **end while**
6: **end procedure**

3.4. NUMERICAL EXPERIMENTS

3.4.1. EXPERIMENTAL SETTING

To test the performance of Algorithm 1 we generate data for two separate identification experiments as follows.

We assume that the time-varying aberration consists of oblique astigmatism and coma and the diversity of only a defocus. That is, we consider a case with $s = 3$ aberrated Zernike modes, so

$$\overline{\phi}(\alpha_k, \beta_k) = Z_2^{-2}\alpha_k(1) + Z_3^{-1}\alpha_k(2) + Z_3^1\alpha_k(3) + Z_2^0\beta_k. \tag{3.32}$$

The first mode, Z_2^{-2}, is an even mode and the effect is that without an added diversity, $\alpha(1)$ and $-\alpha(1)$ are indistinguishable from a single PSF measurement.

In the first experiment every tenth measurement is taken with a defocus diversity and the remaining 90% of images are taken without diversity ($\beta_k = 0$). That is,

$$\beta_k = \begin{cases} 0.5 & k = 1,\ 11,\ 21,\ \dots \\ 0 & \text{otherwise} \end{cases}. \quad \text{(Experiment 1)} \tag{3.33}$$

The motivation is that the out-of-focus images are used to distinguish between $\alpha_k(1)$ and $-\alpha_k(1)$, and the use of the model-set constraint determines the sign for the remaining in-focus images.

In the second experiment every image is taken out of focus, i.e.

$$\beta_k = 0.5 \quad \forall k. \quad \text{(Experiment 2)} \tag{3.34}$$

For both experiments, the coefficients α evolve according to a VAR(2) model. The state-space formulation of the VAR model has the system matrix

$$A_s = \begin{pmatrix} A_1^{\text{true}} & A_2^{\text{true}} \\ I & 0 \end{pmatrix} \tag{3.35}$$

where the block matrix $\begin{pmatrix} A_1^{\text{true}} & A_2^{\text{true}} \end{pmatrix}$ is random and the poles of A_s have absolute value between 0.75 and 0.9. Thus, the poles are chosen to be relatively 'slow' (towards the

Property	Experiment 1	Experiment 2
w_k	$\mathcal{N}(0, 3 \cdot 10^{-2} I)$	$\mathcal{N}(0, 2 \cdot 10^{-1} I)$
measurement noise in v_k	$\mathcal{N}(0, 1 \cdot 10^{-7} I)$	$\mathcal{N}(0, 1 \cdot 10^{-5} I)$
time series K	100	100
VAR order N	2	2
pixels p^2	25	25
Iterations Alg. 3	150	150
Repetitions	100	100

Table 3.1: Settings for the two numerical experiments

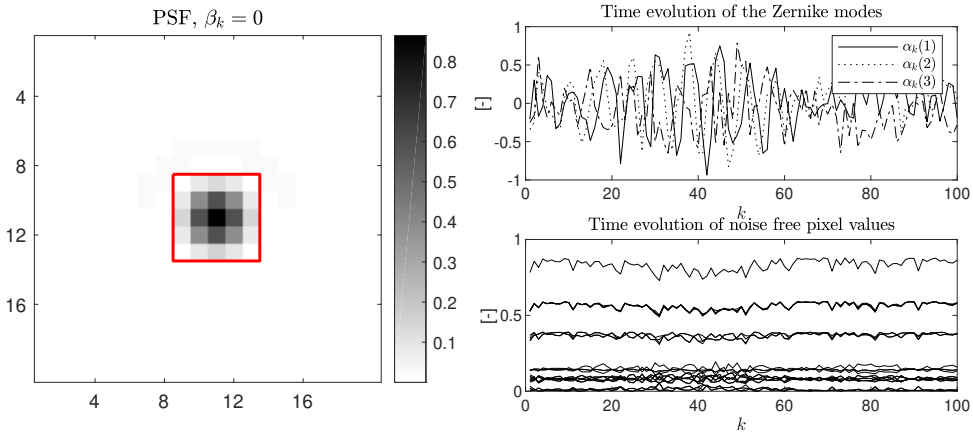


Figure 3.3: On the left, the PSF at time step $k = 100$. Outlined in red, the 25 pixel values used in the identification. On the right, top, the Zernike coefficients for an example dataset for $k = 1, \ldots, 100$. Bottom, the time series for $k = 1, \ldots, 100$ for the corresponding pixel values.

edge of the unit circle). The effect of this choice is that their corresponding dynamics are more clearly present in a relatively short dataset. The decorrelation time of the states of a representative randomly generated system is approximately 40 to 50 samples. These two experiments are repeated 100 times with different system matrices A_s, state and measurement noise sequences for every repetition. Parameter settings are listed in Table 3.1. In both experiments the driving noise w has a noise power that ensures that $\mathcal{O}\left(\|\alpha\|^3\right)$ is small. Both driving noise and measurement noise are Gaussian white noise with the mean and variance as specified in Table 3.1. From the PSF generated according to (3.8) we use a subset of (only) 25 pixels, see Figure 3.3. The number of pixels that are used is limited to reduce the computation time of the optimization, which is roughly 800 seconds on a desktop computer. The initial guess for $\mathbf{X}_1, \mathbf{X}_2, \mathbf{Y}_1, \mathbf{Y}_2$ is drawn randomly from a normal distribution $\mathcal{N}(0, 1 \cdot 10^{-5} I)$. The problem in (3.29) is solved for $\lambda = 0.25, 0.5, \ldots, 0.25 \cdot (z-1), 0.25 \cdot z$ where $\lambda = 0.25 \cdot z$ corresponds to the first solution that worsens Variance Accounted For (VAF) of the states compared to the corresponding solution for $\lambda = 0.25 \cdot (z-1)$. A small regularization is added of the form $0.005 \|A\|_F^2$ to

avoid over-fitting of the model to the estimated state sequence. A side effect is that it reduces the spectral radius [35].

3.4.2. ALTERNATIVE METHODS

For benchmarking purposes, we consider two alternative methods to our proposed method for solving (3.17).

Two-Step Least Squares The first alternative is to estimate the system dynamics in a two-step method based on the linear approximation. It has the following two steps:

$$\begin{aligned} &\text{I)} \quad \hat{\alpha}_k = \arg\min_{\alpha_k} \sum_{i,k} \left\| \mathbf{y}_i(\alpha_k, \beta_k) - D_{0,i}(\beta_k) - D_{1,i}(\beta_k)\alpha_k \right\|_F^2 \\ &\text{II)} \quad \hat{A}_1, \hat{A}_2 = \arg\min_{A_1, A_2} \sum_k \| \hat{\alpha}_k - A_1 \hat{\alpha}_{k-1} - A_2 \hat{\alpha}_{k-2} \|_F^2 . \end{aligned} \tag{3.36}$$

For images taken without diversity, the first problem is ill-conditioned, and this method is not applicable.

Separable non-linear least squares (SNLLS) The second method minimizes a non-linear least squares cost function, that exploits the separability of the optimization problem. For the minimization we use MATLAB's built-in nonlinear least-squares optimizer, where we can make use of an exact or approximate gradient (for settings, see Appendix B.2). That is, the identification problem can also be formulated as

$$\underset{A,\alpha}{\text{minimize}} \quad ||G(\alpha)\bar{A} + h(\alpha)||_2^2 \tag{3.37}$$

where $\bar{A} = \text{vect}(A)$ and G, h are given in Appendix B.1.

The first step is to optimize over $\bar{A}$ and then substitute the optimal solution $\bar{A}(\alpha) = -G^\dagger(\alpha)h(\alpha)$ into (3.37) which yields the problem

$$\begin{aligned} &\underset{A,\alpha}{\text{minimize}} \quad ||G\bar{A} + h||_2^2 \\ =\ &\underset{\alpha}{\text{minimize}} \quad ||(I - GG^\dagger)h||_2^2 \\ =\ &\underset{\alpha}{\text{minimize}} \quad ||P_G^\perp h||_2^2 \end{aligned} \tag{3.38}$$

which may be solved using a nonlinear least square solver. With the residual, $r = G\bar{A}(\alpha) + h = P_G^\perp h$, the solver can either be fed the exact gradient

$$\frac{\partial r}{\partial \alpha} = \frac{\partial P_G^\perp}{\partial \alpha} h + P_G^\perp \frac{\partial h}{\partial \alpha} \tag{3.39}$$

or an approximate gradient

$$\frac{\partial r}{\partial \alpha} \approx P_G^\perp \left(\frac{\partial G}{\partial \alpha} + \frac{\partial h}{\partial \alpha} \right). \tag{3.40}$$

which is considerably faster computationally [36]. The solver was initialized in three different ways: first with the results from the Two-Step Least Squares, and second with the result of our proposed method and third with 100 random initial guesses. This number corresponds to solving the same problem with the proposed method in terms of computational time.

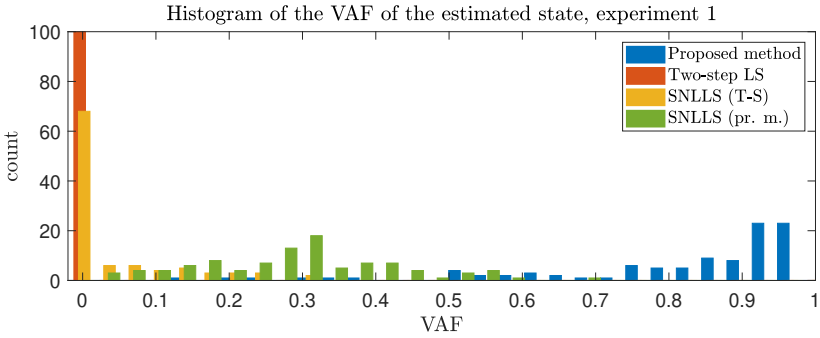


Figure 3.4: The VAF of the estimated state sequences for the first experiment. T-S is for initial guess from Two-Step Least Squares (LS) and pr.m. is for initial guess from the solution of proposed method. Note that from this figure it is apparent that the first least squares problem of the Two-step LS solution (3.36) fails to produce a good estimate of the states.

3.4.3. PERFORMANCE MEASURES

We compare the estimation results in two ways. First, we compare the estimated state sequence (Zernike coefficients). Second, we compare how well the estimated dynamics can be used to predict a state of an independent data set generated under the same circumstances as the data set used for identification. The estimation error for an estimated $\hat{A}_1$ and $\hat{A}_2$ is defined as

$$e_k = \alpha_k - \hat{A}_1 \alpha_{k-1} - \hat{A}_2 \alpha_{k-2}. \tag{3.41}$$

3.4.4. RESULTS AND DISCUSSION

Despite the benefit of several random initial guesses for each experiment, the SNLLS consistently failed to produce good results for any of the experiments. Thus, the results of random initial guesses are omitted from the results. In the experiments it was noted that initialization with the correct solution in the SNLLS algorithm yields the correct solution. However, with small perturbations of this initialization the solution will instead converge to a different local minimum with the SNLLS method. We conclude that the SNLLS method is very sensitive to the initialization. The VAF of the estimated states are displayed in the histogram in figures 3.4 and 3.5. We draw two conclusions from this figure.

First, the proposed method is the only one that can correctly identify the states in Experiment 1 (top histogram). In most instances the estimated states were close to the true states, even though 90% of the images were taken without added phase diversity. The use of the model-set prior information enables this good performance.

Secondly, the proposed method, with its quadratic approximation of the measurements, outperforms the linear model of the measurements (bottom histogram) in Experiment 2.

In Figure 3.6 and 3.7 we compare the average root mean square error of the state estimates for the validation datasets. The SNLLS method produced bad estimates in the terms of RMS, on average about 1000 times worse compared to the proposed method. Thus, it is left out of these figures. We give the Root Mean Square (RMS) of the estima-

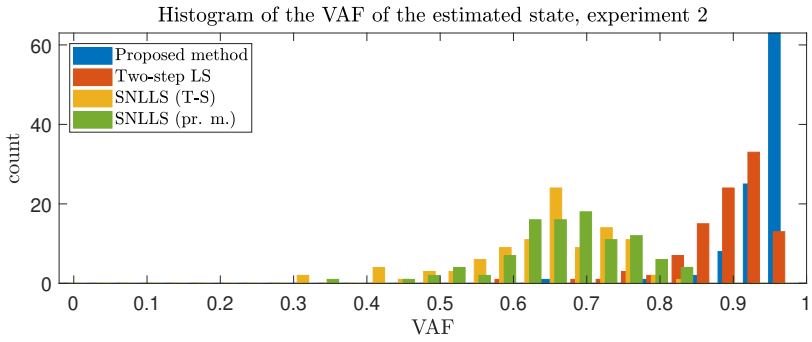


Figure 3.5: The VAF of the estimated state sequences for the second experiment. T-S is for initial guess from Two-Step LS and pr.m. is for initial guess from the solution of proposed method.

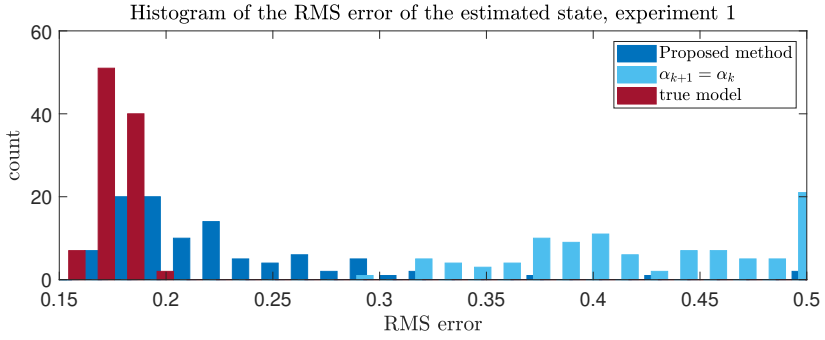


Figure 3.6: Comparison of RMS for the next predicted state by proposed method, states remain constant and true model for the first experiment.

tion error produced by the true model, and the average RMS estimation error produced by the static model $\alpha_{k+1} = \alpha_k$ for comparison. From these figures we conclude that the proposed method can identify the true model with good performance, since its performance is close to that of the true model, and that it significantly improves upon the assumption of a static aberration.

Some of the limitations we found that worsened the estimation results were increasing noise levels, the limitations of the quadratic approximation when the aberration increases in strength. Also, with the relatively short dataset we used, it was more difficult to estimate fast dynamics.

3.5. CONCLUSION AND FUTURE RESEARCH

We presented a method to jointly estimate the temporal dynamics of the phase aberration and the phase aberration itself of an optical system based on measurements of the Point Spread Function of this optical system. The approach is novel firstly in the sense that a model set of temporal dynamics is used as prior information for phase retrieval, and secondly uses a convex heuristic approach with good results to a blind system iden-

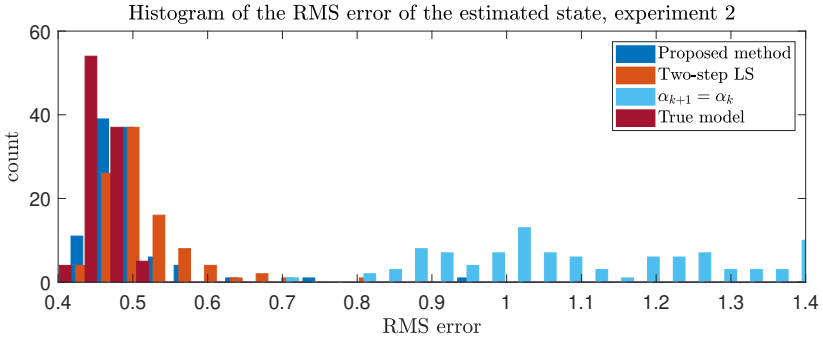


Figure 3.7: Comparison of RMS for the next predicted state by proposed method, states remain constant and true model for the first experiment.

tification problem with a nonlinear output function. Future research lines include modelling spatial dynamics in anisoplanatism instead of temporal dynamics and increasing the accuracy of the (small) phase approximation for larger phase aberrations.

FUNDING INFORMATION

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement No. 339681 and the Swedish Research Council under contract No E05946CI. Both are gratefully acknowledged.

REFERENCES

[1] B. C. Platt and R. Shack, "History and principles of Shack-Hartmann wavefront sensing," Journal of Refractive Surgery, vol. 17, no. 5, pp. S573–S577, 2001.

[2] D. Acton, D. Soltau, and W. Schmidt, "Full-field wavefront measurements with phase diversity.," Astronomy and Astrophysics, vol. 309, pp. 661–672, 1996.

[3] L. Mugnier, J.-F. Sauvage, T. Fusco, A. Cornia, and S. Dandy, "On-line long-exposure phase diversity: a powerful tool for sensing quasi-static aberrations of extreme adaptive optics imaging systems.," Optics Express, vol. 16, no. 22, pp. 18406–18416, 2008.

[4] C. Kulcsár, H.-F. Raynaud, J.-M. Conan, C. Correia, and C. Petit, "Control design and turbulent phase models in adaptive optics: A state-space interpretation," in Adaptive Optics: Methods, Analysis and Applications, p. AOWB1, Optical Society of America, 2009.

[5] B. Le Roux, J.-M. Conan, C. Kulcsár, H.-F. Raynaud, L. M. Mugnier, and T. Fusco, "Optimal control law for classical and multiconjugate adaptive optics," JOSA A, vol. 21, no. 7, pp. 1261–1276, 2004.

[6] K. Hinnen, M. Verhaegen, and N. Doelman, "A data-driven $\mathscr{H}_2$-optimal control approach for adaptive optics," IEEE Transactions on Control Systems Technology, vol. 16, no. 3, pp. 381–395, 2008.

[7] M. Verhaegen and V. Verdult, Filtering and system identification: a least squares approach. Cambridge university press, 2007.

[8] M. Hartung, A. Blanc, T. Fusco, F. Lacombe, L. Mugnier, G. Rousset, and R. Lenzen, "Calibration of NAOS and CONICA static aberrations-experimental results," Astronomy & Astrophysics, vol. 399, no. 1, pp. 385–394, 2003.

[9] A. Blanc, T. Fusco, M. Hartung, L. Mugnier, and G. Rousset, "Calibration of NAOS and CONICA static aberrations-application of the phase diversity technique," Astronomy & Astrophysics, vol. 399, no. 1, pp. 373–383, 2003.

[10] M. A. van Dam, D. Le Mignant, and B. A. Macintosh, "Performance of the Keck Observatory adaptive-optics system," Applied Optics, vol. 43, no. 29, pp. 5458–5467, 2004.

[11] A. Roorda, F. Romero-Borja, W. J. Donnelly III, H. Queener, T. J. Hebert, and M. C. Campbell, "Adaptive optics scanning laser ophthalmoscopy," Optics express, vol. 10, no. 9, pp. 405–412, 2002.

[12] H. Hofer, N. Sredar, H. Queener, C. Li, and J. Porter, "Wavefront sensorless adaptive optics ophthalmoscopy in the human eye," Optics express, vol. 19, no. 15, pp. 14160–14171, 2011.

[13] Y. N. Sulai and A. Dubra, "Non-common path aberration correction in an adaptive optics scanning ophthalmoscope," Biomedical optics express, vol. 5, no. 9, pp. 3059–3073, 2014.

[14] T. Fusco, C. Petit, G. Rousset, J. F. Sauvage, A. Blanc, J. M. Conan, and J. L. Beuzit, "Optimization of the pre-compensation of non-common path aberrations for adaptive optics systems," in Adaptive Optics: Methods, Analysis and Applications, p. AWB2, Optical Society of America, 2005.

[15] J.-F. Sauvage, T. Fusco, G. Rousset, and C. Petit, "Calibration and precompensation of noncommon path aberrations for extreme adaptive optics," JOSA A, vol. 24, no. 8, pp. 2334–2346, 2007.

[16] R. A. Gonsalves, "Phase retrieval and diversity in adaptive optics," Optical Engineering, vol. 21, no. 5, p. 215829, 1982.

[17] J. R. Fienup, "Phase retrieval algorithms: a comparison," Applied optics, vol. 21, no. 15, pp. 2758–2769, 1982.

[18] M. G. Lofdahl, "Multi-frame blind deconvolution with linear equality constraints," in Image reconstruction from incomplete data II, vol. 4792, pp. 146–156, International Society for Optics and Photonics, 2002.

[19] M. Van Noort, L. R. Van Der Voort, and M. G. Löfdahl, "Solar image restoration by use of multi-frame blind de-convolution with multiple objects and phase diversity," Solar Physics, vol. 228, no. 1-2, pp. 191–215, 2005.

[20] L. Vanbeylen, R. Pintelon, and J. Schoukens, "Blind maximum-likelihood identification of Wiener systems," IEEE Transactions on Signal Processing, vol. 57, no. 8, pp. 3017–3029, 2009.

[21] A. Wills, T. B. Schön, L. Ljung, and B. Ninness, "Blind identification of Wiener models," in Proceedings of the 18th IFAC World Congress, Milan, Italy, 2011.

[22] A. Wills, T. B. Schön, L. Ljung, and B. Ninness, "Identification of Hammerstein–Wiener models," Automatica, vol. 49, no. 1, pp. 70–81, 2013.

[23] F. Lindsten, T. B. Schön, and M. I. Jordan, "Bayesian semiparametric Wiener system identification," Automatica, vol. 49, no. 7, pp. 2053–2063, 2013.

[24] F. Lindsten, T. Schön, and M. I. Jordan, A semiparametric Bayesian approach to Wiener system identification. Linköping University Electronic Press, 2011.

[25] R. Marinică, C. S. Smith, and M. Verhaegen, "State feedback control with quadratic output for wavefront correction in adaptive optics," in Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on, pp. 3475–3480, IEEE, 2013.

[26] C. Smith, R. Marinică, A. Den Dekker, M. Verhaegen, V. Korkiakoski, C. Keller, and N. Doelman, "Iterative linear focal-plane wavefront correction," JOSA A, vol. 30, no. 10, pp. 2002–2011, 2013.

[27] G. Monchen, B. Sinquin, and M. Verhaegen, "Recursive kronecker-based vector autoregressive identification for large-scale adaptive optics," IEEE Transactions on Control Systems Technology, no. 99, pp. 1–8, 2018.

[28] B. Sinquin and M. Verhaegen, "Quarks: Identification of large-scale kronecker vector-autoregressive models," IEEE Transactions on Automatic Control, vol. 64, no. 2, pp. 448–463, 2019.

[29] N. Haverbeke, "Efficient numerical methods for moving horizon estimation," Diss., Katholieke Universiteit Leuven, Heverlee, Belgium, 2011.

[30] R. Doelman and M. Verhaegen, "Sequential convex relaxation for convex optimization with bilinear matrix equalities," in Proceedings of the European Control Conference, 2016.

[31] O. Toker and H. Ozbay, "On the NP-hardness of solving bilinear matrix inequalities and simultaneous stabilization with static output feedback," in American Control Conference, Proceedings of the 1995, vol. 4, pp. 2525–2526, IEEE, 1995.

[32] J. Lofberg, "YALMIP: A toolbox for modeling and optimization in MATLAB," in Computer Aided Control Systems Design, 2004 IEEE International Symposium on, pp. 284–289, IEEE, 2004.

[33] M. Grant and S. Boyd, "CVX: MATLAB software for Disciplined Convex Programming, version 2.1." http://cvxr.com/cvx, Mar. 2014.

[34] Y. Hu, D. Zhang, J. Ye, X. Li, and X. He, "Fast and accurate matrix completion via truncated nuclear norm regularization," IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 9, pp. 2117–2130, 2013.

[35] T. Van Gestel, J. A. Suykens, P. Van Dooren, and B. De Moor, "Identification of stable models in subspace identification by using regularization," IEEE Transactions on Automatic control, vol. 46, no. 9, pp. 1416–1420, 2001.

[36] R. Wallin and A. Hansson, "Maximum likelihood estimation of linear SISO models subject to missing output data and missing input data.," International Journal of Control, vol. 87(11), p. 2358, 2014.

4

CONVEX OPTIMIZATION-BASED BLIND DECONVOLUTION FOR IMAGES TAKEN WITH COHERENT ILLUMINATION

A rank-constrained reformulation of the blind deconvolution problem on images taken with coherent illumination is proposed. Since in the reformulation the rank constraint is imposed on a matrix that is affine in the decision variables, we propose a novel convex heuristic for the blind deconvolution problem. The proposed heuristic allows for easy incorporation of prior information on the decision variables and the use of the phase diversity concept. The convex optimization problem can be iteratively re-parameterized to obtain better estimates. The proposed methods are demonstrated on numerically illustrative examples.

Parts of this chapter have been published as a journal publication, “Reinier Doelman and Michel Verhaegen, Convex optimization-based blind deconvolution for images taken with coherent illumination. Journal of the Optical Society of America A, 2019."

4.1. INTRODUCTION

In application areas such as Coherent Diffraction Imaging (CDI) [1], long range horizontal imaging [2], imaging of layered metamaterials [3], or ptychography [4] the image formation process can be described using the expressions for imaging with coherent illumination [5],

$$\mathbf{y} = \left|\mathbf{g}_o \star \mathbf{h}\right|^2, \tag{4.1}$$

where $\mathbf{y}$ denotes the (noiseless) measurement (recorded discretized image), $|\cdot|^2$ denotes the element-wise squared absolute value of the complex-valued argument, in this case the complex field in the imaging plane. $\mathbf{g}_o$ is the object-plane complex amplitude, and $\mathbf{h}$ denotes the amplitude impulse response. $\star$ is the (discrete) convolution operator. The amplitude impulse response is sometimes called the coherent Point Spread Function (coherent PSF). In for example ptychography, the quantity of interests are the Fourier transforms of $\mathbf{g}_o$ or $\mathbf{h}$.

If either $\mathbf{g}_o$ or $\mathbf{h}$ is known, and the other is to be estimated based on the measurements $\mathbf{y}$, then this problem is called a phase retrieval problem. If both quantities are to be estimated based on $\mathbf{y}$, then the problem is called a blind deconvolution problem. The method proposed in this paper can be seen as an extension of a method we proposed in [6] for phase retrieval, where that algorithm is compared to other standard phase retrieval methods.

In this paper we consider the blind deconvolution problem for images taken with coherent illumination, that is

$$\begin{aligned} &\text{find} && \mathbf{g}_o,\ \mathbf{h} \\ &\text{subject to} && \mathbf{y} = \left|\mathbf{g}_o \star \mathbf{h}\right|^2 \\ &&& \mathbf{g}_o \in \mathcal{M}_{\mathbf{g}_o} \\ &&& \mathbf{h} \in \mathcal{M}_{\mathbf{h}} \end{aligned} \tag{4.2}$$

where $\mathcal{M}$ denotes a set of (convex) constraints on the variables that encodes the available prior information.

This blind deconvolution problem is different from what is typically encountered in literature, the blind deconvolution problem for images taken with incoherent illumination [5],

$$\begin{aligned} &\text{find} && \mathbf{f},\ \mathbf{s} \\ &\text{subject to} && \mathbf{y} = \mathbf{f} \star \mathbf{s} \\ &&& \mathbf{f} \in \mathcal{M}_{\mathbf{f}} \\ &&& \mathbf{s} \in \mathcal{M}_{\mathbf{s}} \end{aligned} \tag{4.3}$$

where $\mathbf{f} = \left|\mathbf{g}_o\right|^2$ is the (real and positive valued) intensity of the object in the object plane, and $\mathbf{s} = |\mathbf{h}|^2$ is the intensity impulse response, more often called the Point Spread Function. For the incoherent illumination case, there are several categories of blind deconvolution methods in the literature. Classic iterative projection methods [7–11] use alternating projections of the estimates and their Fourier transform on their respective constraints in the constraint sets. A second group is that of (non-convex,) gradient-based optimization approaches [12], including Bayesian estimation approaches [13–15].

A downside of gradient-based approaches is the initial guess is often crucial for performance. Recently a third group of algorithms is being developed, based on convex optimization of a 'lifted' problem [16–19]. The 'lifting' of the problem hinders the use of phase diversity [20], a powerful type of prior information, which can be described as a linear constraint on **h** for different images with different phase diversities.

For the coherent illumination blind deconvolution problem, we can make the same classifications.

In the first category there is [21, 22], where the extended Ptychographical Iterative Engine (ePIE) is proposed, an iterative transform algorithm for ptychography. Other iterative Fourier transform-based techniques are [23–29].

In the second category, there are the methods proposed in [30, 31] for the estimation of wavefront errors in CDIs and in ptychography [32–34]. [35] compares the performance of several gradient descent schemes showing superior robustness to noise for amplitude based metrics. Refinement of a guessed object and wavefront aberration in a Maximum Likelihood context can be found in [36]. Related to this, gradient-descent schemes are also popular in ptychography for compensating positioning errors [37].

A convex optimization-based approach has to the best of our knowledge not been applied to the coherent blind deconvolution problem in the literature. For example in ptychography, [38] only solves the deconvolution problem, not the blind deconvolution problem, using convex optimization-based heuristic methods.

In this article we propose a blind deconvolution method for the coherent illumination case, based on a rank constrained reformulation of problem (4.2). The reformulation is such, that the use of multiple images and phase diversity is easily incorporated into the reformulation and subsequent optimization problem. To attempt to find a solution with rank constraints satisfied, we propose to use the nuclear norm as a convex heuristic for the rank constraint. An iterative extension of the subsequent convex optimization problem is proposed to possibly improve the convex heuristic approximation. This iterative extension has shown in the validation studies to improve the results. To anticipate the problem that the convex optimization problem results in an unsatisfactory solution, we propose an iterative scheme of convex optimization problems, that produces in our experience iteratively improved results.

The organization of this paper is as follows. In Section 4.2 we formulate the blind deconvolution problem as a problem to estimate a complex-valued object and the affinely parameterized pupil function of the optical system with unknown phase aberration. Section 4.3 explains how to reformulate the blind deconvolution into a rank constraint problem with constraints on matrices affinely parameterized in the object and amplitude impulse response. Section 4.3.4 describes the convex heuristic for the problem and Section 4.3.6 how to incorporate several types of prior information. In Section 4.4 we demonstrate the algorithm on an illustrative numerical example and compare our method to a gradient descent scheme.

4.1.1. NOTATION

The operation $x = \mathrm{vect}(X)$ stacks the columns from left to right of matrix X on top of each other to obtain the vector x. $\otimes$ denotes the Kronecker product. I_n denotes an $n \times n$ identity matrix. $X = \mathrm{d}(x)$ is the diagonal matrix with the values of the vector x on its

diagonal. The Hermitian transpose of X is denoted by X^H. The nuclear norm is denoted as $\|X\|_*$ and the Frobenius norm as $\|X\|_F$.

4.2. PROBLEM DESCRIPTION

The Generalized Pupil Function (GPF) characterizing an optical system [5] is the complex-valued function

$$\mathscr{P}(\rho,\theta) = \mathbf{A}(\rho,\theta)\exp\left(j\phi(\rho,\theta)\right), \tag{4.4}$$

where (ρ,θ) are the radius and angle of the polar coordinates, respectively. $|\rho| \leq 1$ and $\theta \in [0,2\pi)$. $\mathbf{A}(\rho,\theta)$ is the amplitude apodisation function, and $\phi(\rho,\theta) \in \mathbb{R}$ is the phase aberration function of the optical system.

To obtain more measurements of the same object $\mathbf{g}_o$ with different Point Spread Functions (PSFs), a phase diversity $\phi_{\mathbf{d}}$ may be introduced into the system by means of, for example, a deformable mirror. The GPF then becomes

$$\mathscr{P}_{\mathbf{d}}(\rho,\theta) = \mathbf{A}(\rho,\theta)\exp\left(j\phi(\rho,\theta)\right)\exp\left(j\phi_{\mathbf{d}}(\rho,\theta)\right). \tag{4.5}$$

In this paper we consider the problem in modal form: we assume that the GPF can be well-approximated with a weighted sum of basis functions. We use real-valued radial basis functions and complex coefficients to approximate the GPF [39]. Switching from polar coordinates (ρ,θ) to Cartesian coordinates (x,y), the radial basis functions and approximate GPF are given by

$$\begin{aligned} G_i &= \chi(x,y)\exp\left(-\lambda_i\left((x-x_i)^2+(y-y_i)^2\right)\right), \\ \mathscr{P}(x,y) &\approx \tilde{\mathscr{P}}(x,y,\mathbf{v}) = \sum_{i=1}^{s}\mathbf{v}_i G_i(x,y), \end{aligned} \tag{4.6}$$

with (x_i,y_i) being the centers of the radial basis functions, and $\mathbf{v}_i \in \mathbb{C}$. $\chi(x,y)$ is the aperture support function, λ_i is the spread of the radial basis function, and $\mathbf{v} \in \mathbb{C}^s$ is the complex-valued vector of coefficients $\mathbf{v}_i$. Including an introduced diversity $\phi_{\mathbf{d}}(x,y)$, the approximate pupil function reads

$$\tilde{\mathscr{P}}_{\mathbf{d}}(x,y,\mathbf{v}) = \sum_{i=1}^{s}\mathbf{v}_i G_i(x,y)\exp\left(j\phi_{\mathbf{d}}(x,y)\right). \tag{4.7}$$

The amplitude impulse response, also called the coherent Point Spread Function (coherent PSF), $\mathbf{h}_{\mathbf{d}}(u,v)$ is the (2-dimensional) inverse Fourier transform of the GPF:

$$\mathbf{h}_{\mathbf{d}}(u,v) = \sum_{i=1}^{s}\mathbf{v}_i\mathscr{F}^{-1}\left\{G_i(x,y)\exp\left(j\phi_{\mathbf{d}}(x,y)\right)\right\} = \sum_{i=1}^{s}\mathbf{v}_i\mathbf{B}_{\mathbf{d},i}(u,v). \tag{4.8}$$

Here the coordinates (u,v) are the Cartesian coordinates in the image plane of the optical system.

A complex amplitude in the object plane $\mathbf{g}_o$, imaged through this optical system gives, in the case of coherent illumination, the complex amplitude $\mathbf{g}_i$ in the image plane [5]:

$$\mathbf{g}_i(u,v) = \mathbf{g}_o(u,v) \star \mathbf{h}_{\mathbf{d}}(u,v). \tag{4.9}$$

In the noise-free case the intensity of the complex field $\mathbf{g}_i$ is recorded to produce the measurements $\mathbf{y}$:

$$\mathbf{y}(u,v) = \left|\mathbf{g}_i(u,v)\right|^2. \tag{4.10}$$

We now drop the notation for the dependency on coordinates and assume the signals $\mathbf{y}$, $\mathbf{g}_i$, $\mathbf{g}_o$ and $\mathbf{h}_d$ are sampled on square grids of sizes $r \times t$, $r \times t$, $m \times n$ and $p \times q$ respectively, such that we obtain matrices of the corresponding size. Throughout this paper we assume that edges of $\mathbf{g}_o$ and $\mathbf{h}_d$ are zero-padded, which for the discrete two-dimensional convolution results in the relation $r = m + p - 1$, $t = n + q - 1$.

With slight abuse of notation, the blind deconvolution problem (4.2) has now turned into the problem to identify $\mathbf{g}_o$ and $\mathbf{v}$ from measurements $\mathbf{y}$:

$$\begin{aligned} &\text{find} && \mathbf{g}_o,\ \mathbf{v},\ \mathbf{h},\ \mathbf{g}_i \\ &\text{subject to} && \mathbf{y} = \left|\mathbf{g}_i\right|^2 \\ & && \mathbf{g}_i = \mathbf{g}_o \star \mathbf{h} \\ & && \mathbf{h} = \mathbf{B}\mathbf{v} \\ & && \mathbf{g}_o \in \mathcal{M}_{\mathbf{g}_o} \\ & && \mathbf{h} \in \mathcal{M}_{\mathbf{h}} \end{aligned} \tag{4.11}$$

4.3. BLIND DECONVOLUTION AS A RANK-CONSTRAINED FEASIBILITY PROBLEM

The aim of this section is to rewrite (4.11) into a feasibility problem with rank constraints; one rank constraint to replace $\mathbf{y} = \left|\mathbf{g}_i\right|^2$, and one rank constraint to replace $\mathbf{g}_i = \mathbf{g}_o \star \mathbf{h}$.

In the following two subsections we use the following lemma.

Lemma 4.3.1. [40] Define the matrix

$$\begin{aligned} M(C,A,B,Q,X,Y,W_1,W_2) = &\begin{pmatrix} W_1 & 0 \\ 0 & I \end{pmatrix} \times \\ &\begin{pmatrix} C + AQY + XQB + XQY & (A+X)Q \\ Q(B+Y) & Q \end{pmatrix} \begin{pmatrix} W_2 & 0 \\ 0 & I \end{pmatrix}, \end{aligned} \tag{4.12}$$

where I is the identity matrix.

For any X of the same size as A, for any Y of the same size as B, for any invertible matrices W_1, W_2 of a size corresponding to the sizes of matrix C, and for nonzero Q, it holds that the equality

$$\operatorname{rank}(M(C,A,B,Q,X,Y,W_1,W_2)) = \operatorname{rank}(Q)$$

is equivalent to the equality

$$C = AQB. \tag{4.13}$$

Note that variables A and B appear in a product in (4.13), but they do not appear in a product in the matrix M in (4.12).

4.3.1. THE CONVOLUTION CONSTRAINT $\mathbf{g}_i = \mathbf{g}_o \star \mathbf{h}$

The two-dimensional (discrete) convolution of $\mathbf{g}_o$ and $\mathbf{Bv}$ gives the matrix $\mathbf{g}_i$. The elements of the matrix $\mathbf{g}_i$ are given by the summation of products of individual elements of $\mathbf{g}_o$ with individual elements of $\mathbf{Bv}$. Lemma 4.3.2 states how this can be cast into a bilinear matrix equality.

Lemma 4.3.2. The constraint $\mathbf{g}_i = \mathbf{g}_o \star \mathbf{h}$ is equivalent to the bilinear equality

$$\operatorname{vect}(\mathbf{g}_i) = \left(\operatorname{vect}(\mathbf{g}_o)^T \otimes I_{rt}\right) V \operatorname{vect}(\mathbf{Bv}). \tag{4.14}$$

for a matrix of zeros and ones $V \in \mathbb{B}^{mnrt \times pq}$.

Proof. See Appendix C. □

In (4.14) the general bilinear form $C = AQB$ of Lemma 4.3.1 shows through, with

$$\begin{aligned} C &= \operatorname{vect}(\mathbf{g}_i), & A &= \left(\operatorname{vect}(\mathbf{g}_o)^T \otimes I_{rt}\right), \\ Q &= V, & B &= \operatorname{vect}(\mathbf{Bv}). \end{aligned} \tag{4.15}$$

We can therefore replace the constraint $\mathbf{g}_i = \mathbf{g}_o \star \mathbf{h}$ with the rank constraint

$$\operatorname{rank}\left(M\left(\operatorname{vect}(\mathbf{g}_i), \operatorname{vect}(\mathbf{g}_o)^T \otimes I_{rt}, \operatorname{vect}(\mathbf{Bv}), V, X, Y, W_1, W_2\right)\right) = \operatorname{rank}(V). \tag{4.16}$$

The matrices X, Y, W_1 and W_2 are here further specified to

$$\begin{aligned} X &= -\operatorname{vect}(\hat{\mathbf{g}_o})^T \otimes I_{rt}, \\ Y &= -\operatorname{vect}(\mathbf{B}\hat{\mathbf{v}}), \\ W_1 &= I, \\ W_2 &= I, \end{aligned} \tag{4.17}$$

where $\hat{\mathbf{g}_o}$ and $\hat{\mathbf{v}}$ are – for the moment – some guess for $\mathbf{g}_o$ and $\mathbf{v}$ respectively. The expression for the matrix-valued function M we now abbreviate for notational convenience and call this specific abbreviation M_c,

$$\begin{aligned} &M_c(\mathbf{g}_i, \mathbf{g}_o, \mathbf{v}, V, \hat{\mathbf{g}_o}, \hat{\mathbf{v}}) = \\ &M\left(\operatorname{vect}(\mathbf{g}_i), \operatorname{vect}(\mathbf{g}_o)^T \otimes I_{rt}, \operatorname{vect}(\mathbf{Bv}), V, X, Y, W_1, W_2\right), \end{aligned} \tag{4.18}$$

where $M_c \in \mathbb{C}^{(rt+mnrt)\times(1+pq)}$.

4.3.2. THE MEASUREMENT CONSTRAINT $\mathbf{y} = |\mathbf{g}_i|^2$

The treatment of the measurement constraints is similar to [6]. The constraint $\mathbf{y} = |\mathbf{g}_i|^2$ uses the element-wise operator $|\cdot|$. To obtain the relation between $\mathbf{y}$ and $\mathbf{g}_i$ in matrix format, we place the values on a matrix diagonal:

$$\mathbf{y} = |\mathbf{g}_i|^2 \quad \Leftrightarrow \quad \mathrm{d}(\operatorname{vect}(\mathbf{y})) = \mathrm{d}(\operatorname{vect}(\mathbf{g}_i))^H \mathrm{d}(\operatorname{vect}(\mathbf{g}_i)). \tag{4.19}$$

The related rank constraint is

$$\text{rank}\left(M\left(\text{d}\left(\text{vect}\left(\mathbf{y}\right)\right),\text{d}\left(\text{vect}\left(\mathbf{g}_i\right)\right)^H,\text{d}\left(\text{vect}\left(\mathbf{g}_i\right)\right),I_{rt},X,Y,W_1,W_2\right)\right)=rt. \tag{4.20}$$

We further specify here that

$$\begin{aligned} X &= -\text{d}\left(\text{vect}\left(\hat{\mathbf{g}}_i\right)\right)^H, \\ Y &= -\text{d}\left(\text{vect}\left(\hat{\mathbf{g}}_i\right)\right), \\ W_1 &= I, \\ W_2 &= I, \end{aligned} \tag{4.21}$$

where $\hat{\mathbf{g}}_i$ is some guess for $\mathbf{g}_i$. Furthermore, we abbreviate the arguments of M as

$$\begin{aligned} &M_m\left(\mathbf{y},\mathbf{g}_i,\hat{\mathbf{g}}_i\right)= \\ &M\left(\text{d}\left(\text{vect}\left(\mathbf{y}\right)\right),\text{d}\left(\text{vect}\left(\mathbf{g}_i\right)\right)^H,\text{d}\left(\text{vect}\left(\mathbf{g}_i\right)\right),I_{rt},X,Y,W_1,W_2\right), \end{aligned} \tag{4.22}$$

where $M_m \in \mathbb{C}^{2rt\times 2rt}$.

4.3.3. THE RANK-CONSTRAINED BLIND DECONVOLUTION PROBLEM

Using (4.16) and (4.20) the blind deconvolution problem (4.11) can be expressed as

$$\begin{aligned} &\text{find} && \mathbf{g}_o,\mathbf{v},\mathbf{g}_i && \text{(4.23a)} \\ &\text{subject to} && \text{rank}\left(M_m\left(\mathbf{y},\mathbf{g}_i,\hat{\mathbf{g}}_i\right)\right)=rt, && \text{(4.23b)} \\ & && \text{rank}\left(M_c\left(\mathbf{g}_i,\mathbf{g}_o,\mathbf{v},V,\hat{\mathbf{g}}_o,\hat{\mathbf{v}}\right)\right)=\text{rank}\left(V\right), && \text{(4.23c)} \\ & && \mathbf{g}_o\in\mathcal{M}_{\mathbf{g}_o} && \text{(4.23d)} \\ & && \mathbf{h}=\mathbf{B}\mathbf{v}\in\mathcal{M}_{\mathbf{h}} && \text{(4.23e)} \end{aligned}$$

The caveat is that rank-constrained problems are in general Non-deterministic Polynomial-time (NP) hard, that is (informally), in general there do not exist algorithms that can compute a feasible solution, guaranteed, within a time that is bounded by a polynomial in the number of variables. However, we can attempt to compute a solution $\{\mathbf{g}_o{}^*,\mathbf{g}_i{}^*,\mathbf{v}^*\}$ and check whether the matrices $M_m\left(\mathbf{y},\mathbf{g}_i{}^*,\hat{\mathbf{g}}_i\right)$ and $M_c\left(\mathbf{g}_i{}^*,\mathbf{g}_o{}^*,\mathbf{v}^*,V,\hat{\mathbf{g}}_o,\hat{\mathbf{v}}\right)$ have the correct rank.

4.3.4. A CONVEX HEURISTIC FOR BLIND DECONVOLUTION

Even though the reformulated problem with its rank constraints is still non-convex, we propose to use a convex heuristic, the nuclear norm [41], to attempt to minimize the ranks of the matrices involved. The nuclear norm of a matrix is defined as the sum of the singular values of a matrix:

$$\|X\|_*=\sum_i \sigma_i(X), \tag{4.24}$$

where $\sigma_i(X)$ is the i'th largest singular value of X. We can therefore use the nuclear norm as a convex heuristic for the blind deconvolution problem to attempt to find a solution,

but success is not guaranteed. The convex optimization approach for (4.23) is

$$\begin{aligned} \min_{\mathbf{g}_o,\mathbf{v},\mathbf{g}_i} \quad & \mu \left\| M_m\left(\mathbf{y},\mathbf{g}_i,\hat{\mathbf{g}_i}\right)\right\|_* + \left\| M_c(\mathbf{g}_i,\mathbf{g}_o,\mathbf{v},V,\hat{\mathbf{g}_o},\hat{\mathbf{v}})\right\|_*, \\ & \mathbf{g}_o \in \mathcal{M}_{\mathbf{g}_o} \\ & \mathbf{h} = \mathbf{B}\mathbf{v} \in \mathcal{M}_{\mathbf{h}} \end{aligned} \tag{4.25}$$

where the parameter $\mu > 0$ is a tuning parameter that weighs the nuclear norm of the matrix M_m with the nuclear norm of the matrix M_c.

Optimization problem (4.25) is parametrized in (4.17) and (4.21) by $\hat{\mathbf{g}_o}, \hat{\mathbf{v}}$ and $\hat{\mathbf{g}_i}$. The interpretation is that, given some guess $\{\hat{\mathbf{g}_o}, \hat{\mathbf{v}}, \hat{\mathbf{g}_i}\}$, (4.25) produces a new estimate $\{\hat{\mathbf{g}_o}^+, \hat{\mathbf{v}}^+, \hat{\mathbf{g}_i}^+\}$. Motivated by [6, 40], (4.25) can be used in an iterative update scheme, see Algorithm 4.

Algorithm 4 Convex Optimization-based blind deconvolution (COBBD) for images taken with coherent illumination

procedure
  $k = 0$
  while not converged **do**
    Let $\{\hat{\mathbf{g}_i}^{k+1}, \hat{\mathbf{g}_o}^{k+1}, \hat{\mathbf{v}}^{k+1}\}$ be the arguments that minimize (4.25).
    $k = k + 1$
  end while
end procedure

Such an iterative scheme gives rise to three questions. First, do the estimates converge to a fixed point? Second, are the resulting estimates correct solutions to the blind deconvolution problem? Third, if it converges, how fast does it converge? Unfortunately, all three questions are very difficult to answer and we cannot provide a theoretical proof of convergence. We do notice however, that correct solutions of the blind deconvolution problem are fixed points of Algorithm 4. For solutions $\{\mathbf{g}_o{}^*, \mathbf{v}^*, \mathbf{g}_i{}^*\}$ of the blind deconvolution problem, we verify that by substitution

$$\begin{aligned} & \mu \left\| M_m(\mathbf{y},\mathbf{g}_i{}^*,\mathbf{g}_i{}^*)\right\|_* + \left\| M_c(\mathbf{g}_i{}^*,\mathbf{g}_o{}^*,\mathbf{v}^*,V,\mathbf{g}_o{}^*,\mathbf{v}^*)\right\|_* \\ & = \mu \left\| \begin{pmatrix} 0 & 0 \\ 0 & I_{rt} \end{pmatrix} \right\|_* + \left\| \begin{pmatrix} 0 & 0 \\ 0 & V \end{pmatrix} \right\|_* = \mu r t + \|V\|_*, \end{aligned} \tag{4.26}$$

which does not depend on any of the variables. So if $\{\hat{\mathbf{g}_o}, \hat{\mathbf{v}}, \hat{\mathbf{g}_i}\} = \{\mathbf{g}_o{}^*, \mathbf{v}^*, \mathbf{g}_i{}^*\}$ in (4.25), the optimal parameters for (4.25) are $\{\mathbf{g}_o{}^*, \mathbf{v}^*, \mathbf{g}_i{}^*\}$.

The convergence speed properties and success rate of Algorithm 4 depend on the initialization $\{\hat{\mathbf{g}_o}^0, \hat{\mathbf{v}}^0, \hat{\mathbf{g}_i}^0\}$ and tuning of μ and the matrices W_1, W_2 in (4.17) and (4.21). To show the difference that tuning of W_1 and W_2 in (4.17) and (4.21) can make, we solve a small, 1-dimensional blind deconvolution problem with three different sets of tuning parameters.[1] We set $W_1 = W_2 = m_1 I$ in (4.21) , $W_1 = c_1 I$, $W_2 = c_2 I$ in (4.17) and $\mu = 1$. The three sets of parameters (m_1, c_1, c_2) are $(1,1,1)$, $(2,1,4)$ and $(0.6,1,0.6)$. The different convergence speeds can be seen in Figure 4.1. It can be seen that the effect of tuning

[1] See https://bitbucket.org/rdoelman/blinddeconvolution.

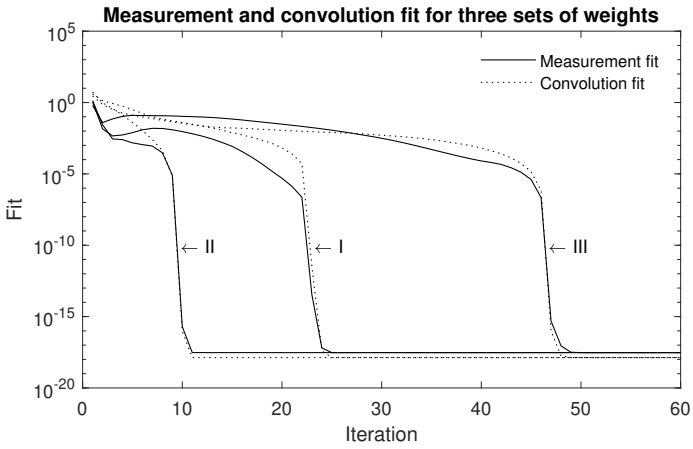


Figure 4.1: The convergence of the constraint violations, $\| y - |\hat{\mathbf{g}}_i|^2 \|_F^2$ (measurement fit) and $\| \hat{\mathbf{g}}_i - \hat{\mathbf{g}}_o \star \mathbf{B}\hat{\mathbf{v}} \|_F^2$ (convolution fit) through updates in Algorithm 4 for three different sets of tuning parameters.

on the convergence speed can be very large. Unfortunately we cannot provide general tuning rules that optimize convergence speed.

4.3.5. COMPUTATIONAL COMPLEXITY OF (4.25)

The computational complexity of the nuclear norm minimization in (4.25) can be estimated as follows. If we assume that minimizing the nuclear norm of a matrix with n variables is of approximately $\mathcal{O}\left(n^6\right)$ when using a standard Semidefite Programming (SDP) solver [42], then solving (4.25) is of complexity $\mathcal{O}\left((rt+mn+pq)^6\right)$ which is very unfavourable for practical applications. An Alternating Direction Method of Multipliers (ADMM), [43, 44] solution for problem (4.25) consists of the singular value decomposition of the matrices $M_c \in \mathbb{C}^{(rt+mnrt)\times(1+pq)}$ and $M_m \in \mathbb{C}^{2rt\times 2rt}$ that are of $\mathcal{O}\left(rt(1+mn)(1+pq)^2+(1+pq)^3\right)$ and $\mathcal{O}\left((rt)^3\right)$ respectively.

Exploiting parallelization opportunities similar to [6], this can be reduced to rt Singular Value Decompositions (SVDs) with complexity $\mathcal{O}\left(\max(mn,pq)^3\right)$ and rt SVDs of matrices of size 2 with complexity $\mathcal{O}(1)$ that can be computed in parallel.

The convergence speed properties and success rate of Algorithm 4 depend on the initialization $\{\hat{\mathbf{g}_o}^0, \hat{\mathbf{v}}^0, \hat{\mathbf{g}_i}^0\}$ and tuning of μ and the matrices W_1, W_2 in (4.17) and (4.21), but we cannot provide successful general tuning rules.

4.3.6. INCLUDING PRIOR INFORMATION AND REGULARIZATION

The optimization in (4.25) is a convex optimization problem in the decision parameters $\mathbf{g}_i, \mathbf{g}_o$ and $\mathbf{v}$. This makes the addition of prior information and regularization very simple, if these can be expressed as convex constraints or convex penalty functions. The convex optimization-based blind deconvolution (for incoherent illumination) techniques such as [17] are based on directly estimating $|\mathbf{g}_i|^2$ and $|\mathbf{h}|^2$, making it difficult to apply constraints on $\mathbf{g}_i$ and $\mathbf{h}$.

We here list some examples of prior information that can be included.

1. The imaged object has a known support (known non-zero-valued pixels). This can be expressed as the constraint $Q\,\mathrm{vect}(\mathbf{g}_o) = 0$ for a selection matrix Q.

2. The imaged object is sparse, in the sense that many pixels of $\mathbf{g}_o$ have value 0. This can be expressed through the addition of a penalty term $\tau \sum_i |\mathbf{g}_{oi}|$ with regularization parameter τ and i denotes the i'th pixel.

3. The extension to the use of multiple images (multi-frame blind deconvolution), taken with different phase diversities, can be done by adding additional terms to the objective function corresponding to the different images, and with addition of the constraints $\mathbf{h}_n = \mathbf{B}_n\mathbf{v} \in \mathscr{M}_{\mathbf{h}_n}$ for the n'th image.

4. In ptychography, overlapping parts of an object positioned in the pupil plane are imaged with the same 'probe' or amplitude transfer function. If we write the Fourier transform as a linear mapping with a matrix $\mathbf{F}$, $\mathrm{vect}(\mathscr{F}\{x\}) = \mathbf{F}\,\mathrm{vect}(x)$, then a shift in the position of the illuminated object can be represented by the constraint $\mathbf{F}_1\,\mathrm{vect}(\mathbf{g}_{o1}) = \mathbf{F}_2\,\mathrm{vect}(\mathbf{g}_{o2})$, where $\mathbf{F}_1$ and $\mathbf{F}_2$ are those parts of the Fourier transform matrices that correspond to the overlapping part of the object. This constraint addresses the problem that a phase aberration of the probe can be attributed to the phase of the object and the other way around. For results on uniqueness and ambiguities, see e.g. [45].

4.4. NUMERICAL EXPERIMENTS

We implemented an ADMM algorithm in MATLAB to compute the updates in Algorithm 4. Although this allows for parallel computations of rtd SVDs with complexity $\mathscr{O}\left(\max(mn, pq)^3\right)$, where d is the number of images taken, and rtd SVDs with complexity $\mathscr{O}(1)$ we computed these in series. Due to the computational complexity, we tested the algorithm for two cases with small dimensions. Furthermore, the ADMM algorithm that iteratively finds the optimal solution to (4.25) is terminated after only 10 iterations.

For comparison, we implemented a gradient descent method comparable to [30, 31, 36], but adapted to our formulation with decision variables defined in the focal plane. An accelerated gradient descent scheme, ADAM [46], is used to speed up the procedure and automatically determine step size. The step size η is tuned once up front to ensure convergence. The settings are: $\beta_1 = 0.8$, $\beta_2 = 0.999$, $\epsilon = 1 \cdot 10^{-8}$, $\eta = 2 \cdot 10^{-4}$.

The experiment models an unknown aberration, consisting of 8 basis functions as in (4.6), that approximate a small defocus, $\phi = 0.2Z_2^0$, where Z_2^0 is the defocus Zernike polynomial. We take three images with phase diversities that are defoci with coefficients $\approx -2, 0$ and 2. Due to the computational complexity the aperture is undersampled when the amplitude impulse response is computed and the resulting matrix is cut to a size of 5×5. The object $\mathbf{g}_o$ is a complex-valued matrix of dimensions 8×8 and the resulting measurements $\mathbf{y}$ are of size 12×12, see Figures 4.2 and 4.3. The value of μ in (4.25) is tuned to 0.55.

Both Algorithm 4 and the gradient descent method are tested on a noiseless case and the same case where measurement noise has been added with a Signal-to-Noise Ratio (SNR) of 20dB. Both algorithms in both cases are initialized with the same initial

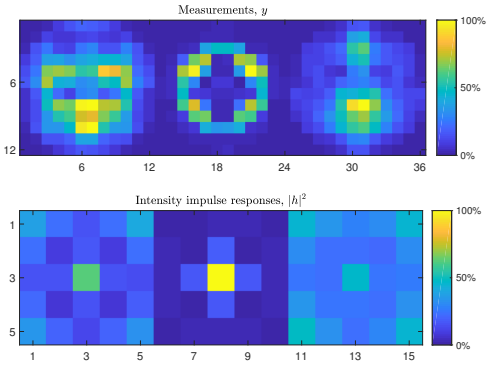


Figure 4.2: Top: the three 12×12 (noiseless) measured intensities $\mathbf{y} = \left|\mathbf{g}_o \star \mathbf{h}\right|^2$. Bottom: the three 5×5 intensity impulse response functions (point spread functions) $\mathbf{s} = |\mathbf{h}|^2$ corresponding to the three different diversities that generate the different $\mathbf{h}$.

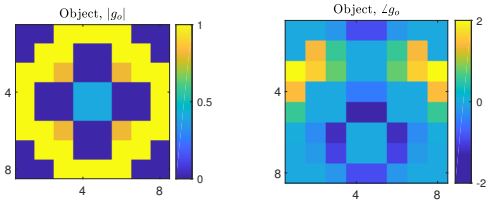


Figure 4.3: Left: the amplitude of the object. Right, the phase in radians of the object.

guess, where the pixels of the initial object estimate are randomly drawn from a Gaussian distribution, and the initial guess for the coefficients are those that best approximate zero aberration. The computation time for the proposed method, implemented without taking advantage of parallel computation of SVDs, is approximately 10 hours for the 15000 iterations as shown here. The computation time consists of roughly 5 hours for computation of SVDs, 40 minutes for solving least squares problems, and the rest being overhead. The gradient descent method is much faster with approximately 18 minutes for 100000 iterations. The resulting norms of the residual between measurements and convolution of the estimated object and amplitude impulse response are plotted in Figure 4.4. As can be seen in this figure, the gradient descent method gets stuck in the noiseless case, whereas the proposed method converges to a feasible solution.

The estimated values $\hat{\mathbf{v}}$ and $\hat{\mathbf{g}}_o$ have an ambiguity, since for a complex scalar c, $c\mathbf{B}\hat{\mathbf{v}} \star \hat{\mathbf{g}}_o = \mathbf{B}\hat{\mathbf{v}} \star c\hat{\mathbf{g}}_o$. We can remove the ambiguity from for example $\hat{\mathbf{v}}$ when reporting the estimation error by computing

$$\min_{c \in \mathbb{C}} \|c\hat{\mathbf{v}} - \mathbf{v}\|_2 . \tag{4.27}$$

After removal of the ambiguity of the estimated values of $\mathbf{g}_o$ and $\mathbf{v}$, we plot in in Figure 4.5 the norms of the residuals between the actual complex amplitude of the object $\mathbf{g}_o$ and coefficients $\mathbf{v}$ and their estimated values. As can be seen in this figure, the proposed method converges in the noiseless case not just to a feasible solutions, but to the correct solution, whereas the gradient descent method stops progressing towards the

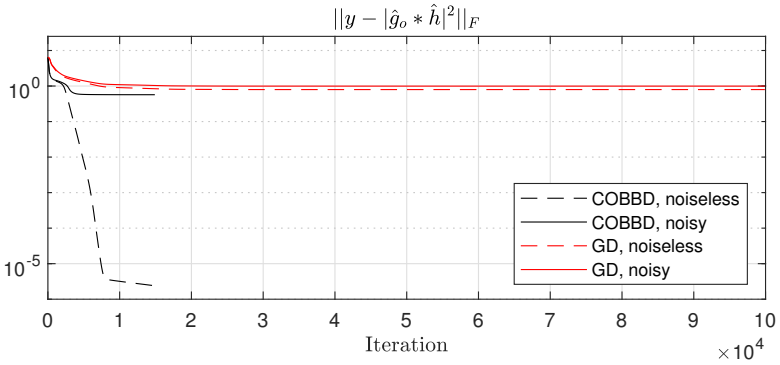


Figure 4.4: The measurement fit generated by the two algorithms. Black: Algorithm 4. Red: gradient descent. Solid lines show the case with noisy measurements (SNR: 20dB), dashed lines show the noiseless case.

solution. In the noisy case Algorithm 4 computes the solution with the best estimated measurements (see Figure 4.4), and with the best estimated object (Figure 4.5, top). The coefficients $\mathbf{v}$ have an estimate that is further from the real coefficients than the estimate of the gradient descent method, but given the measurement fit and fit of $\mathbf{g}_o$, the effect of this error is small. The estimates resulting from Algorithm 4 and from the gradient descent method of the object $\mathbf{g}_o$ are displayed in Figure 4.6. From Figure 4.6 it becomes clear that even though in the noisy case the proposed method does not converge to the exact solution, it converges to a solution that resembles the original object quite well. Inspecting Figure 4.5 shows that the gradient descent method provides estimates of $\mathbf{g}_o$ that are far from it. The resulting estimates of $|\mathbf{h}|$ are shown in Figure 4.7.

4.5. CONCLUSION AND FUTURE RESEARCH

We derived a convex heuristic for the blind deconvolution problem for images taken with coherent illumination that is also able to incorporate the concept of phase diversity. We suggested an update scheme and demonstrated on a numerically illustrative example that it is capable of retrieving the object and PSF from a random initialization, thereby overcoming local minima. At the moment, the method is computationally burdensome, but we expect computational improvements similar to [6] by fully exploiting parallelization opportunities and the structure in the optimization problems. Apart from the nuclear norm heuristic, there are also other methods that attempt to find low rank results, like difference of convex programming (e.g. [47]), or application of the truncated nuclear norm (e.g. [48]), but we leave the evaluation of their performance for future research. Several questions still remain open, concerning optimal tuning rules for the different parameters in the optimization, the conditions on the variables that guarantee uniqueness of the solution, the performance of other non-convex low-rank inducing norms, bounds on convergence speed and the computational speed-up by exploiting parallelization opportunities.

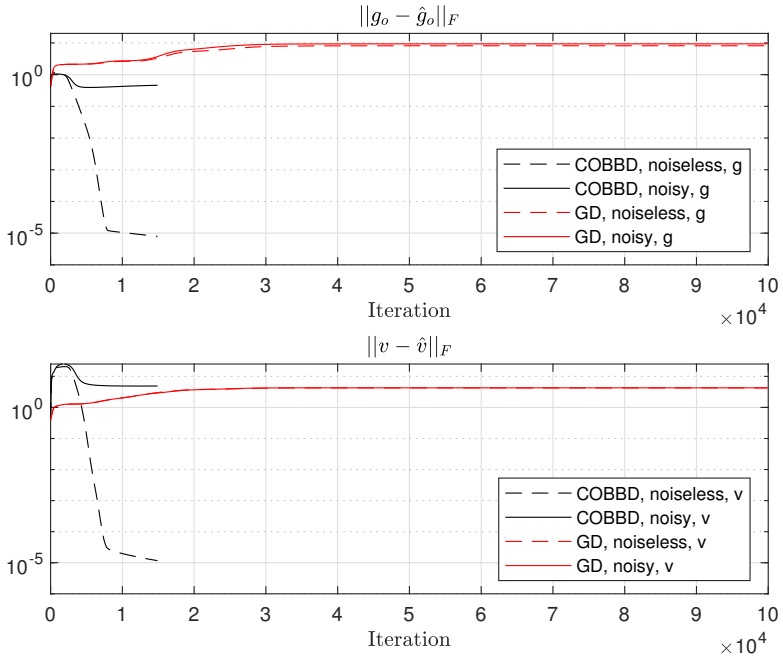


Figure 4.5: The Frobenius norm of the residual between the true variables $\mathbf{g}_o$, the complex-valued object, and $\mathbf{v}$, the Radial Basis Function coefficients, and the (ambiguity removed) estimated variables $\hat{\mathbf{g}_o}$ and $\hat{\mathbf{v}}$. Top figure: residuals for $\mathbf{g}_o$. Bottom figure: residuals for $\mathbf{v}$ Black: Algorithm 4. Red: gradient descent. Solid lines show the case with noisy measurements (SNR: 20dB), dashed lines show the noiseless case.

4.6. FUNDING INFORMATION

The research leading to these results has received funding from the European Research Council (ERC) under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement No. 339681.

REFERENCES

[1] J. Miao, P. Charalambous, J. Kirz, and D. Sayre, "Extending the methodology of X-ray crystallography to allow imaging of micrometre-sized non-crystalline specimens," Nature, vol. 400, no. 6742, p. 342, 1999.

[2] A. MacDonald, Blind deconvolution of anisoplanatic images collected by a partially coherent imaging system. PhD thesis, Air Force Institute of Technology, Wright-Patterson Air Force Base Ohio, 2004.

[3] D. Pastor, T. Stefaniuk, P. Wróbel, C. J. Zapata-Rodríguez, and R. Kotyński, "Determination of the point spread function of layered metamaterials assisted with the blind deconvolution algorithm," Optical and Quantum Electronics, vol. 47, no. 1, pp. 17–26, 2015.

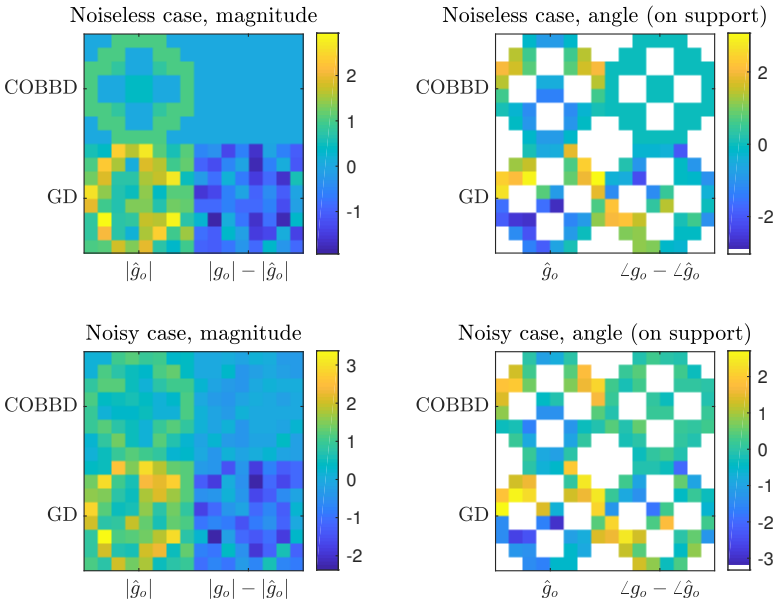


Figure 4.6: The estimates and residuals of $\mathbf{g}_o$ using Algorithm 4 and gradient descent with identical initialization. Top left: the estimated amplitude of $\mathbf{g}_o$ and the residual for the noiseless case; Top right: the estimated angle of $\mathbf{g}_o$ and the residual for the noiseless case; Bottom left: the estimated amplitude of $\mathbf{g}_o$ and the residual for the noisy case; Bottom right: the estimated angle of $\mathbf{g}_o$ and the residual for the noisy case. Since estimates of very small complex values can have a radically different complex angle, the angle is only plotted for the nonzero pixels in the original object of Figure 4.3.

[4] J. M. Rodenburg and H. M. Faulkner, "A phase retrieval algorithm for shifting illumination," Applied physics letters, vol. 85, no. 20, pp. 4795–4797, 2004.

[5] J. Goodman, Introduction to Fourier optics. McGraw-hill, 2008.

[6] R. Doelman, N. H. Thao, and M. Verhaegen, "Solving large-scale general phase retrieval problems via a sequence of convex relaxations," J. Opt. Soc. Am. A, vol. 35, pp. 1410–1419, Aug 2018.

[7] G. Ayers and J. C. Dainty, "Iterative blind deconvolution method and its applications," Optics letters, vol. 13, no. 7, pp. 547–549, 1988.

[8] M. Tofighi, O. Yorulmaz, K. Köse, D. C. Yıldırım, R. Çetin-Atalay, and A. E. Cetin, "Phase and TV based convex sets for blind deconvolution of microscopic images," IEEE Journal of Selected Topics in Signal Processing, vol. 10, no. 1, pp. 81–91, 2016.

[9] D. Wilding, O. Soloviev, P. Pozzi, G. Vdovin, and M. Verhaegen, "Blind multi-frame deconvolution by tangential iterative projections (tip)," Optics Express, vol. 25, no. 26, pp. 32305–32322, 2017.

[10] D. Fish, A. Brinicombe, E. Pike, and J. Walker, "Blind deconvolution by means of the Richardson–Lucy algorithm," JOSA A, vol. 12, no. 1, pp. 58–65, 1995.

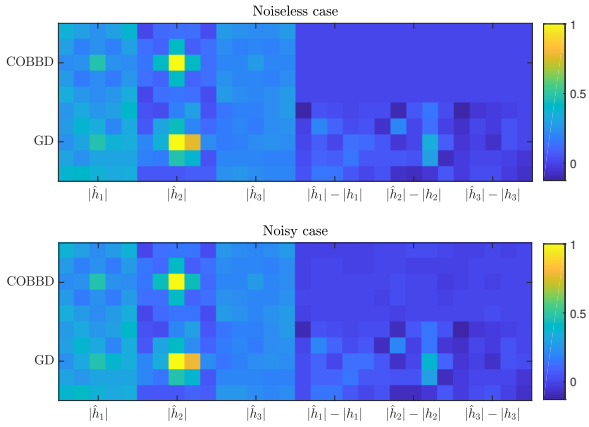


Figure 4.7: The estimates of $|\mathbf{h}|$. The maximum absolute value of $\hat{h}$ is scaled to 1.

[11] P. Sarder and A. Nehorai, "Deconvolution methods for 3-D fluorescence microscopy images," IEEE Signal Processing Magazine, vol. 23, no. 3, pp. 32–45, 2006.

[12] G. Liu, S. Chang, and Y. Ma, "Blind image deblurring using spectral properties of convolution operators," IEEE Transactions on image processing, vol. 23, no. 12, pp. 5047–5056, 2014.

[13] T. J. Schulz, "Multiframe blind deconvolution of astronomical images," JOSA A, vol. 10, no. 5, pp. 1064–1073, 1993.

[14] T. J. Holmes, "Blind deconvolution of quantum-limited incoherent imagery: maximum-likelihood approach," JOSA A, vol. 9, no. 7, pp. 1052–1061, 1992.

[15] R. Mourya, L. Denis, J.-M. Becker, and E. Thiébaut, "A blind deblurring and image decomposition approach for astronomical image restoration," in Signal Processing Conference (EUSIPCO), 2015 23rd European, pp. 1636–1640, IEEE, 2015.

[16] D. Stöger, P. Jung, and F. Krahmer, "Blind deconvolution and compressed sensing," in Compressed Sensing Theory and its Applications to Radar, Sonar and Remote Sensing (CoSeRa), 2016 4th International Workshop on, pp. 24–27, IEEE, 2016.

[17] A. Ahmed, A. Cosse, and L. Demanet, "A convex approach to blind deconvolution with diverse inputs," in Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2015 IEEE 6th International Workshop on, pp. 5–8, IEEE, 2015.

[18] S. Ling and T. Strohmer, "Regularized gradient descent: A nonconvex recipe for fast joint blind deconvolution and demixing," arXiv preprint arXiv:1703.08642, 2017.

[19] P. Jung, F. Krahmer, and D. Stöger, "Blind demixing and deconvolution at near-optimal rate," arXiv preprint arXiv:1704.04178, 2017.

[20] R. A. Gonsalves, "Phase retrieval and diversity in adaptive optics," Optical Engineering, vol. 21, no. 5, p. 215829, 1982.

[21] A. M. Maiden and J. M. Rodenburg, "An improved ptychographical phase retrieval algorithm for diffractive imaging," Ultramicroscopy, vol. 109, no. 10, pp. 1256–1262, 2009.

[22] A. M. Maiden, M. J. Humphry, F. Zhang, and J. M. Rodenburg, "Superresolution imaging via ptychography," JOSA A, vol. 28, no. 4, pp. 604–612, 2011.

[23] R. Horstmeyer, X. Ou, J. Chung, G. Zheng, and C. Yang, "Overlapped Fourier coding for optical aberration removal," Optics express, vol. 22, no. 20, pp. 24062–24080, 2014.

[24] F. Jian and L. Peng, "A general phase retrieval algorithm based on a ptychographical iterative engine for coherent diffractive imaging," Chinese Physics B, vol. 22, no. 1, p. 014204, 2013.

[25] M. Foreman, C. Giusca, P. Török, and R. Leach, "Phase-retrieved pupil function and coherent transfer function in confocal microscopy," Journal of microscopy, vol. 251, no. 1, pp. 99–107, 2013.

[26] X. Ou, G. Zheng, and C. Yang, "Embedded pupil function recovery for Fourier ptychographic microscopy," Optics express, vol. 22, no. 5, pp. 4960–4972, 2014.

[27] P. Thibault, M. Dierolf, O. Bunk, A. Menzel, and F. Pfeiffer, "Probe retrieval in ptychographic coherent diffractive imaging," Ultramicroscopy, vol. 109, no. 4, pp. 338–343, 2009.

[28] R. Hesse, D. R. Luke, S. Sabach, and M. K. Tam, "Proximal heterogeneous block implicit-explicit method and application to blind ptychographic diffraction imaging," SIAM Journal on Imaging Sciences, vol. 8, no. 1, pp. 426–457, 2015.

[29] H. Chang, P. Enfedaque, and S. Marchesini, "Blind ptychographic phase retrieval via convergent alternating direction method of multipliers," arXiv preprint arXiv:1808.05802, 2018.

[30] G. R. Brady, M. Guizar-Sicairos, and J. R. Fienup, "Optical wavefront measurement using phase retrieval with transverse translation diversity," Optics express, vol. 17, no. 2, pp. 624–639, 2009.

[31] M. Guizar-Sicairos and J. R. Fienup, "Measurement of coherent X-ray focused beams by phase retrieval with transverse translation diversity," Optics express, vol. 17, no. 4, pp. 2670–2685, 2009.

[32] Y. S. Nashed, T. Peterka, J. Deng, and C. Jacobsen, "Distributed automatic differentiation for ptychography," Procedia Computer Science, vol. 108, pp. 404–414, 2017.

[33] M. Odstrčil, A. Menzel, and M. Guizar-Sicairos, "Iterative least-squares solver for generalized maximum-likelihood ptychography," Optics express, vol. 26, no. 3, pp. 3108–3123, 2018.

[34] Y. Zhang, W. Jiang, and Q. Dai, "Nonlinear optimization approach for Fourier ptychographic microscopy," Optics Express, vol. 23, no. 26, pp. 33822–33835, 2015.

[35] L.-H. Yeh, J. Dong, J. Zhong, L. Tian, M. Chen, G. Tang, M. Soltanolkotabi, and L. Waller, "Experimental robustness of Fourier ptychography phase retrieval algorithms," Optics express, vol. 23, no. 26, pp. 33214–33240, 2015.

[36] P. Thibault and M. Guizar-Sicairos, "Maximum-likelihood refinement for coherent diffractive imaging," New Journal of Physics, vol. 14, no. 6, p. 063004, 2012.

[37] A. Tripathi, I. McNulty, and O. G. Shpyrko, "Ptychographic overlap constraint errors and the limits of their numerical recovery using conjugate gradient descent methods," Optics Express, vol. 22, no. 2, pp. 1452–1466, 2014.

[38] R. Horstmeyer, R. Y. Chen, X. Ou, B. Ames, J. A. Tropp, and C. Yang, "Solving ptychography with a convex relaxation," New journal of physics, vol. 17, no. 5, p. 053044, 2015.

[39] P. Piscaer, A. Gupta, O. Soloviev, and M. Verhaegen, "Modal-based phase retrieval using Gaussian radial basis functions," JOSA A, 2018.

[40] R. Doelman and M. Verhaegen, "Sequential convex relaxation for convex optimization with bilinear matrix equalities," in European Control Conference (ECC), 2016, pp. 1946–1951, IEEE, 2016.

[41] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," SIAM review, vol. 52, no. 3, pp. 471–501, 2010.

[42] Z. Liu and L. Vandenberghe, "Interior-point method for nuclear norm approximation with application to system identification," SIAM Journal on Matrix Analysis and Applications, vol. 31, no. 3, pp. 1235–1256, 2009.

[43] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al., "Distributed optimization and statistical learning via the alternating direction method of multipliers," Foundations and Trends in Machine learning, vol. 3, no. 1, pp. 1–122, 2011.

[44] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," SIAM Journal on Optimization, vol. 20, no. 4, pp. 1956–1982, 2010.

[45] A. Fannjiang and P. Chen, "Blind ptychography: uniqueness and ambiguities," arXiv preprint arXiv:1806.02674, 2018.

[46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[47] H. Tuy, "DC optimization: theory, methods and algorithms," in Handbook of global optimization, pp. 149–216, Springer, 1995.

[48] Y. Hu, D. Zhang, J. Ye, X. Li, and X. He, "Fast and accurate matrix completion via truncated nuclear norm regularization," IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 9, pp. 2117–2130, 2013.

5

SYSTEMATICALLY STRUCTURED $\mathscr{H}_2$ OPTIMAL CONTROL FOR TRUSS-SUPPORTED SEGMENTED MIRRORS

A systematic distributed optimal control design procedure is proposed for the rejection of wind load induced disturbances on a truss-supported segmented mirror. The distributed nature of the controller is achieved by weighing of the interaction matrices between local (per-segment) controllers in a global $\mathscr{H}_2$ optimization. The procedure allows a tradeoff analysis between the controller implementation complexity versus the improved performance the extra communication brings.

The procedure is demonstrated on a finite-element model of a segmented mirror on a flexible supporting truss to which we apply the combined closed-loop performance and local controller interconnection structure optimization. The resulting set of controllers is compared to a set of baseline controllers including Linear Quadratic Gaussian (LQG) control, Singular Value Decomposition (SVD) control, and a distributed controller where local controllers of neighbouring segments communicate.

The tradeoff analysis for the segmented mirror demonstrates that the communication between the local controllers can be greatly reduced without significantly compromising the rejection of wind-induced wavefront errors.

This chapter is based on Doelman, "Reinier Doelman, Sander Dominicus, Renaud Bastaits and Michel Verhaegen, Systematically structured $\mathscr{H}_2$ optimal control for truss-supported segmented mirrors. IEEE Transactions on Control Systems Technology 99, pp. 1–8. IEEE 2018.", [1]. The segmented mirror model was developed thanks to the support of Fonds National de la Recherche Scientifique, via the grant FRIA FC76554.

5.1. INTRODUCTION

The primary, segmented mirror in large ground-based astronomical telescopes requires active control to compensate for the dynamic disturbance of wind loading on the segments.

As such, the modelling of segmented mirrors, the modelling of dynamical disturbance, design of controllers and the evaluation of the closed loop performance are studied during the design phase of telescopes. See for example for the Keck telescope [2], the European Extremely Large Telescope (E-ELT) [3–5] or for the Thirty Meter Telescope (TMT) [6–9].

The dynamics of the identical segments are not necessarily decoupled. If the segments are mounted on a supporting truss, the backstructure causes interaction between the dynamics of the individual segments. This backstructure interaction is taken into account in the control design [3, 7]. The effect of the backstructure interaction increases with telescope size [10] and depends too on the structural damping in the supporting truss.

From the perspective of the entire closed loop system, not only the backstructure causes spatial dynamics. A second source of the dynamical interaction is the way the position of the mirror segments are measured, which is through edge sensors that measure the relative displacement between neighbouring segments. A third source is spatial correlation in the disturbance model (the force of the wind acting on the segments).

A fourth source are the dynamics of the control loop. For example, [3, 11] feature SVD-based modal controllers. In [12] local (per-segment) controllers are designed that are connected to the controllers of neighbouring segments. In [13] a centralized $\mathscr{H}_2$ optimal controller is designed and a distributed controller based on spatial invariance assumptions.

We observe that, conceptually, the choice of the structure of the controller (centralized, decentralized, hierarchical) is made first, and subsequently the controller is designed to meet a performance criterion. We propose in this article to see the choice of the segmented mirror controller structure and the meeting of a performance criterion as something that should and could be done in a single controller design procedure. In this procedure, there is a multi-criterion optimization that results in a tradeoff curve of controller complexity versus performance.

Even if current design methods can reach the desired performance, the use of optimal control theory and controller structure optimization might open up the possibility of less stringent design criteria for other parts of the system, or might improve the end result of the whole optical system.

Several related approaches for static (state or output) feedback that search for a sparse controller structure regardless of model structure can be found in [14–17].

Instead of static output output feedback, we focus our attention on dynamic output feedback. We analyze the closed loop performance of the system in the case that each segment has a local, low order, dynamic output feedback controller, and and a decision has to be taken on if and how these local controllers should be interconnected.

If the structure of the controllers is fixed up front, existing non-convex (gradient descent) optimization methods could be applied [18, 19] to optimize the controller.

We can apply the relevant theory in [14] on appropriate addition of regularization

on the decision variables to the objective function to induce sparsity in the resulting controller matrices.

We build on the work in [14, 18, 19] to derive analytic expressions to compute the gradients of the multi-criterion performance measure with respect to the controller matrices.

The result of this approach is that the off-line, up front computational cost to design the controller is high, but the on-line computations are light, see also Section 5.6. The gradient descent procedure requires the solution of Lyapunov equations, whose computational complexity grows cubically with the number of states of the closed loop system. The on-line computations are of a computational complexity that grows linearly with the number of interconnections between local controllers, and quadratically with the number of states in a local controller. If the computations are done centrally, the optimization of the interconnection structure leads to scalable on-line computations as well. This is an advantage over an standard optimal LQG controller, whose on-line computational requirements grow with the number of segments and system states.

In this paper we discuss the multi-criterion design procedure in Section 5.2. The controller design procedure is demonstrated on an accurate, but numerically challenging Finite Element-based virtual model of a segmented mirror on a flexible supporting truss. The model is discussed in Section 5.3, the necessary adaptations and transformations for optimal control design in Section 5.4. The complete model includes edge sensors, backstructure interaction and a spatially correlated wind disturbance model. In Section 5.5 we show the results of the multi-criterion tradeoff analysis for this segmented mirror model. The results are discussed in Section 5.6.

5.2. DISTRIBUTED CONTROL APPROACH

As will be discussed in Section 5.4, the combined segmented mirror and wind loading model used for control design can be transformed into the following standard discrete time system description:

$$\begin{pmatrix} x[k+1] \\ z[k] \\ y[k] \end{pmatrix} = \begin{pmatrix} A & B_1 & B \\ C_1 & 0 & 0 \\ C & F & 0 \end{pmatrix} \begin{pmatrix} x[k] \\ w[k] \\ u[k] \end{pmatrix}, \tag{5.1}$$

with $x \in \mathbb{R}^n$, $z \in \mathbb{R}^{r_z}$, $y \in \mathbb{R}^{r_y}$, $w \in \mathbb{R}^{m_w}$, $u \in \mathbb{R}^{m_u}$. For notational simplicity time postscripts will be dropped, i.e. x is short for $x[k]$, z, y , w , u are defined similarly, and x^+ is short for $x[k+1]$. The system matrices have dimensions that can be inferred from the signal dimensions. The vector y contains the edge sensor measurements, the vector u are the segment position actuator inputs. For the segmented mirror system the disturbance vector w is the white Gaussian noise that drives the wind model and the sensor noise. The output vector z describes the mirror shape. The global mirror shape is determined by the top position of each hexagonal segment's 3 position actuators, and the channels of z comprise the deviation of the position of the top of each actuator from the mean of the top positions of all the actuators.

For a system driven by white noise, $w \sim \mathcal{N}(0, I)$, with r_z output channels (indexed

z_j), the squared $\mathscr{H}_2$ norm of the system equals the output variance:

$$\|T_{w\to z}\|_{\mathscr{H}_2}^2 = \sum_{j=1}^{r_z} E\left[z_j^2\right],$$

where $T_{w\to z}$ is the transfer function from the disturbance vector w to performance channel z. If we use the linear approximation of the wavefront error of segmented mirrors [20], then the Root Mean Square (RMS) wavefront error e is related to the $\mathscr{H}_2$ norm by

$$e = \frac{2}{\sqrt{r_z}}\|T_{w\to z}\|_{\mathscr{H}_2},$$

where the factor of 2 comes from the fact that the RMS wavefront error is twice the RMS surface error. Standard Linear Time-Invariant (LTI) optimal control methods can be used to minimize the effects of wind on the error e.

We concentrate on dynamic output feedback controllers without a direct feedthrough term, i.e. controllers of the form:

$$\begin{pmatrix} x_c^+ \\ u \end{pmatrix} = \begin{pmatrix} A_c & B_c \\ C_c & 0 \end{pmatrix}\begin{pmatrix} x_c \\ y \end{pmatrix}, \tag{5.2}$$

where $x_c \in \mathbb{R}^{n_c}$ are the controller states. Closing the loop we obtain the system

$$\begin{pmatrix} \mathbf{x}^+ \\ z \end{pmatrix} = \begin{pmatrix} \mathscr{A} & \mathscr{B} \\ \mathscr{C} & 0 \end{pmatrix}\begin{pmatrix} \mathbf{x} \\ w \end{pmatrix}, \tag{5.3}$$

where

$$\left(\begin{array}{c|c} \mathscr{A} & \mathscr{B} \\ \hline \mathscr{C} & 0 \end{array}\right) = \left(\begin{array}{cc|c} A & BC_c & B_1 \\ B_cC & A_c & B_cF \\ \hline C_1 & 0 & 0 \end{array}\right)$$

and $\mathbf{x} = \begin{pmatrix} x^T & x_c^T \end{pmatrix}^T$.

With a closed loop transfer function $T_{cl,w\to z}$, the optimization problem becomes

$$\min_{A_c,B_c,C_c} J(A_c,B_c,C_c), \tag{5.4}$$

where $J(A_c,B_c,C_c) := \left\|T_{cl,w\to z}\right\|_{\mathscr{H}_2}^2$.

With A_c, B_c and C_c known, the controllability and observability Gramians W_c and W_o are respectively determined by solving the Lyapunov equations

$$\begin{aligned} \mathscr{A}W_c\mathscr{A}^T - W_c + \mathscr{B}\mathscr{B}^T &= 0, \\ \mathscr{A}^T W_o\mathscr{A} - W_o + \mathscr{C}^T\mathscr{C} &= 0. \end{aligned} \tag{5.5}$$

The squared $\mathscr{H}_2$ norm of this system can be computed as follows (see in this context [18, 19]):

$$\left\|T_{cl,w\to z}\right\|_{\mathscr{H}_2}^2 = \operatorname{trace}\left(\mathscr{C}W_c\mathscr{C}^T\right) = \operatorname{trace}\left(\mathscr{B}^T W_o\mathscr{B}\right).$$

Eq. (5.5) shows that this criterion is not convex in the controller parameters, even if the controller parametrization is affine. Transformations exist that render the computation

of an $\mathscr{H}_2$ controller a convex problem [21] through appropriate substitution of products of decision variables, but such a substitution would hamper imposing a desired sparsity structure on the controller matrices.

The squared $\mathscr{H}_2$ norm of the system is however differentiable and analytical expressions can be derived for the gradients of the norm with respect to the controller matrices in (5.2). This also allows us to iteratively update the controller matrices during a gradient descent optimization. The disadvantage is that due to the non-convex nature of the problem it cannot be guaranteed that the global optimum will be found. This consequence can be mitigated by trying multiple different starting points for the optimization.

5.2.1. OPTIMIZATION APPROACH

Using the results in [18, 19] we can derive the following gradients of $J(A_c, B_c, C_c)$ in (5.4) with respect to the controller matrices:

$$\begin{aligned}
&\frac{\partial J(A_c,B_c,C_c)}{\partial A_c} = 2\begin{pmatrix}0 & I\end{pmatrix} W_o \mathscr{A} W_c \begin{pmatrix}0\\ I\end{pmatrix},\\
&\frac{\partial J(A_c,B_c,C_c)}{\partial B_c} =\\
&\quad 2\begin{pmatrix}0 & I\end{pmatrix} W_o \mathscr{A} W_c \begin{pmatrix}C^T\\ 0\end{pmatrix} + 2\begin{pmatrix}0 & I\end{pmatrix} W_o \mathscr{B} F^T,\\
&\frac{\partial J(A_c,B_c,C_c)}{\partial C_c} =\\
&\quad 2\begin{pmatrix}B^T & 0\end{pmatrix} W_o \mathscr{A} W_c \begin{pmatrix}0\\ I\end{pmatrix} + 2E^T \mathscr{C} W_c \begin{pmatrix}0\\ I\end{pmatrix}.
\end{aligned} \tag{5.6}$$

The derivation of these expressions can be found in Appendix D.

The gradients in (5.6) can be used in a gradient descent scheme to find a locally optimal dynamic output feedback controller.

In our implementation we used the accelerated gradient descent method ADAM [22], because the method selects the step sizes automatically, and the method is efficient in both memory usage and the amount of required additional computations. For a gradient descent procedure with n_v variables, the required computations are of $\mathscr{O}(n_v)$ complexity and the additional required memory of $\mathscr{O}(n_v)$ size. The advantage of a gradient-based optimization is that structure can be imposed on the controller system matrices. This will be discussed in the next subsection.

5.2.2. DISCOVERING A SPARSELY CONNECTED CONTROLLER IN A USER-MOTIVATED GLOBAL STRUCTURE

From a distributed controller-design point-of-view we can assume that there are N subsystems, with either one or more mirror segments per subsystem, and we would like to assign a local controller to each subsystem. If we assume that each local controller is connected only to the inputs and outputs of its own subsystem and controller states of other local controllers, then we can state that the matrices B_c and C_c have a block-diagonal structure after proper renumbering of system inputs and outputs. Denote this as $B_c \in \mathscr{M}_{B_c}$, $C_c \in \mathscr{M}_{C_c}$.

The matrix A_c can be written in the form

$$A_c = \begin{pmatrix} A_{c,11} & \cdots & A_{c,1N} \\ \vdots & & \vdots \\ A_{c,N1} & \cdots & A_{c,NN} \end{pmatrix}, \tag{5.7}$$

where $A_{c,ii}$, $i = 1,\dots,N$ constitutes the local controller dynamics of subsystem i and $A_{c,ij}$, $i \neq j$ describes together with $A_{c,ji}$ the interaction between local controllers i and j. That is, the local controller update $x_{c,i}^+$ can be described as

$$x_{c,i}^+ = A_{c,ii}x_{c,i} + \sum_{j=1, j\neq i}^{N} A_{c,ij}x_{c,j} + B_{c,i}y_i, \tag{5.8}$$

where $x_c^T = \begin{pmatrix} x_{c,1}^T & \cdots & x_{c,N}^T \end{pmatrix}$ is the controller state, y_i are local measurements and $B_{c,i}$ is the appropriate block on the diagonal of B_c. The number of connections between the states of local controllers is computed by

$$\frac{1}{2}\sum_{i,j,\ i\neq j} \operatorname{card}\left(\left\| \begin{pmatrix} A_{c,ij} \\ A_{c,ji}^T \end{pmatrix} \right\|_F \right), \tag{5.9}$$

where the function $\operatorname{card}(\cdot)$ denotes the cardinality operator:

$$\operatorname{card}(q) = \begin{cases} 1 & q \neq 0, \\ 0 & q = 0. \end{cases}$$

The argument of the cardinality operator in (5.9) accounts for the fact that the local controller state information can flow both ways if there is an interconnection between the states of controller i and j.

To find a distributed controller, one could force the matrix A_c to have blocks equal to zero, indicating that there is no possibility for communication between two local controllers. Two examples are decentralized control and controllers where states of local controllers of neighbouring segments are connected.

Based on [17, 23] we propose to add the block-sparsity promoting term

$$H(A_c) := \frac{1}{2}\sum_{i,j,\ i\neq j} \Gamma_{i,j} \left\| \begin{pmatrix} A_{c,ij} \\ A_{c,ji}^T \end{pmatrix} \right\|_F,$$

as a weighted convex relaxation of (5.9), to the objective function (5.4), in order to trade off performance of the closed loop system with the number of interconnections between local controllers. The block weights $\Gamma_{i,j}$ influence the optimization's preference for having certain blocks put to 0. Using this weighting term, it is possible to systematically weigh the relative ease of implementation of connections between local controllers against an improved performance of the closed-loop system by varying the parameter γ.

The resulting optimization problem is

$$\begin{aligned} \min_{A_c,B_c,C_c} \quad & J(A_c,B_c,C_c) + \gamma H(A_c), \\ \text{s.t.} \quad & B_c \in \mathscr{M}_{B_c}, \\ & C_c \in \mathscr{M}_{C_c}. \end{aligned} \tag{5.10}$$

where γ is a regularization term whereby we can influence the (general) level of block-sparsity in A_c.

The derivative of $H(A_c)$ with respect to the blocks in A_c is

$$\frac{\partial H(A_c)}{\partial A_{c,kl}} = \frac{\Gamma_{k,l} A_{c,kl}}{\left\| \begin{pmatrix} A_{c,kl} \\ A_{c,lk}^T \end{pmatrix} \right\|_F}, \; l \neq k, \left\| \begin{pmatrix} A_{c,kl} \\ A_{c,lk}^T \end{pmatrix} \right\|_F \neq 0$$

From the derivatives with respect to the blocks of A_c the entire gradient of the objective function in (5.10) with respect to A_c can be constructed.

When in addition to the sets $\mathcal{M}_{B_c}, \mathcal{M}_{C_c}$ the set $\mathcal{M}_{A_c}$ is similarly defined as the nonzero block-pattern of the matrix A_c, (5.10) can be optimized with $\gamma = 0$ and the additional constraint $A_c \in \mathcal{M}_{A_c}$:

$$\begin{aligned} \min_{A_c, B_c, C_c} \quad & J(A_c, B_c, C_c), \\ \text{s.t.} \quad & A_c \in \mathcal{M}_{A_c}, \\ & B_c \in \mathcal{M}_{B_c}, \\ & C_c \in \mathcal{M}_{C_c}. \end{aligned} \tag{5.11}$$

The sets $\mathcal{M}_{A_c}, \mathcal{M}_{B_c}, \mathcal{M}_{C_c}$ can be specified by the user. However, the set $\mathcal{M}_{A_c}$ can also be derived from the solution to (5.10) by thresholding the Frobenius norm of the blocks of A_c. The subsequent optimization is called 'polishing' [14].

5.3. A SEGMENTED MIRROR ON A FLEXIBLE SUPPORTING TRUSS

5.3.1. MIRROR MODEL

We use the mirror model as described in [11] and [10] Section 4.3. The mirror model has segments with a diameter of 1.8 m and consists of 2 rings and 18 segments.

The model is created using SAMCEF Finite Element Method (FEM) software and a Craig-Bampton reduction. The FEM model reads:

$$\begin{aligned} &\underbrace{\begin{pmatrix} \hat{M}_{11} & \hat{M}_{12} \\ \hat{M}_{21} & I \end{pmatrix}}_{\hat{M}} \begin{pmatrix} \ddot{x}_1 \\ \ddot{\alpha} \end{pmatrix} + \underbrace{\begin{pmatrix} C_{11} & 0 \\ 0 & 0 \end{pmatrix}}_{\hat{C}} \begin{pmatrix} \dot{x}_1 \\ \dot{\alpha} \end{pmatrix} \\ &+ \underbrace{\begin{pmatrix} \hat{K}_{11} & 0 \\ 0 & \Omega^2 \end{pmatrix}}_{\hat{K}} \begin{pmatrix} x_1 \\ \alpha \end{pmatrix} = \begin{pmatrix} F_1 \\ 0 \end{pmatrix}. \end{aligned} \tag{5.12}$$

The state x_1 contains the bottom and top positions of the actuators, $x_1 = \begin{pmatrix} x_{\text{bottom}}^T & x_{\text{top}}^T \end{pmatrix}$. α is the vector of modal amplitudes of the fixed boundary modes resulting from the Craig-Bampton reduction.

Forces F_1 are external forces on the truss, either through loading of the mirror segments or reaction forces in the truss support. Modal damping is added to the model as in [11] with a damping ratio of 1%, which is a standard value in segmented mirror research [10, 12, 24].

The matrices $\hat{M}_{11}, \hat{M}_{12}, \hat{M}_{21}, \Omega$ and $\hat{K}_{11}$ are obtained from the Craig-Bampton reduction. The diagonal matrix Ω contains the natural frequencies of the fixed boundary modes.

The three actuators that suspend each segment are mounted on top of the supporting truss. The edge sensors are located on the edges of the mirror segments. Up to six sensors are present per segment, depending on the number of neighbouring segments. All sensors measure the relative displacement with respect to the neighbouring segment, in the out-of-plane direction. The six sensors pointing towards the middle, where there is no segment, measure the displacement with respect to the supporting truss.

5.3.2. WIND LOAD DISTURBANCE AND CONTROL OBJECTIVE

The control objective is to minimize the effect of wind loading on the shape of the mirror.

Many different disturbances act on the telescope structure, e.g. wind loading, edge sensor noise and structural vibrations [5, 13, 25, 26]. From these, we incorporate only wind loading and edge sensor noise in the model. Other disturbances can be dealt with by the adaptive optics systems, separate control systems, or are of very little influence on the wavefront.

The wind disturbance considered is the along-wind response of the mirror, modelled with a classic random vibration approach. The turbulent wind force is assumed to follow Davenport's spectrum [27] and act on the model through the force F_1 in (5.12). The reference mean velocity of the wind is $10\,\mathrm{m\,s^{-1}}$ in the direction perpendicular to the mirror with a constant wind profile over the height of the mirror, and cross-correlation of the disturbance on the different segments is assumed to be non-zero and computed according to [10]. The analytically computed cross power spectral density of Davenport's spectrum is approximated by a ninth order band-limited white Gaussian noise driven LTI system using the method in [28].

The edge sensors are assumed to have a noise level of $1\,\mathrm{nm}/\sqrt{\mathrm{Hz}}$.

The Maréchal criterion [29] states that the performance of an optical element is limited by diffraction when the RMS wavefront error is lower than $\sqrt{\lambda_l^2/180} \approx \lambda_l/13.4$, where λ_l is the wavelength of light. Since the smallest λ_l observed by the science instruments in for example E-ELT [30] is $\lambda_l = 370\,\mathrm{nm}$, the objective for the controller design is a closed-loop RMS wavefront error below 27.6 nm. Any performance below this value we consider to be sufficient, though lower values indicate that other system requirements could be set more lenient. We do not consider frequency weighting of errors or the possibility of subsequent error compensation through Adaptive Optics in the performance comparison between different controllers.

5.4. MODEL ADAPTATIONS FOR OPTIMAL CONTROL ENGINEERING

We can write the FEM model into a descriptor (with the subscript d) state-space form:

$$\begin{aligned} E_d \dot{x} &= A_d x + B_d u + B_v v \\ y &= C_d x + e \end{aligned} \tag{5.13}$$

$$E_d = \begin{pmatrix} \hat{M}_{11} & \hat{M}_{12} & 0 & 0 \\ \hat{M}_{21} & I & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \end{pmatrix}, \; B_d = \begin{pmatrix} S_a k_a \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$A_d = \begin{pmatrix} -C_{11} & 0 & -\hat{K}_{11} & 0 \\ 0 & 0 & 0 & -\Omega^2 \\ I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \end{pmatrix}, \; B_v = \begin{pmatrix} I \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$C_d = \begin{pmatrix} 0 & 0 & S_y & 0 \end{pmatrix}.$$

Here $x \in \mathbb{R}^{n_d}$ is the system state of the descriptor system, A_d, $E_d \in \mathbb{R}^{n_d \times n_d}$, $y \in \mathbb{R}^{n_y}$ is the measurement, e is the sensor noise, and $u \in \mathbb{R}^{n_u}$ the input. The matrix $S_a k_a$ in B_d describes the actuator topology and influence of actuator inputs (displacements) u on the system. v is the output of the LTI wind model. The matrix S_y in C_d describes the sensor topology. The descriptor system in (5.13) is from a numerical point of view badly conditioned and not immediately suitable for simulation, optimization and control with standard MATLAB toolboxes. We use the toolbox by Binder et al. [31] to transform system (5.13) in series with the wind model that produces v, into the staircase canonical form, which transforms the matrix E_d into a diagonal matrix. The resulting state-space mirror model can be simulated accurately but still has a badly conditioned E matrix. We make a minimal realization of the system which removes all uncontrollable and unobservable poles. Since for this mirror models at hand this removes the generalized eigenvalues at infinity, E_d is no longer badly conditioned and the model can be rewritten into a standard continuous-time state-space model.

Since the implementation of a controller is expected to be in discrete time, the model is converted to a discrete time-model. All dominant dynamics are contained in a bandwidth of 200 Hz, so the model is discretized with a sampling frequency of 1 kHz.

After this series of transformations, the model is in the form of (5.1) and the techniques outlined in Section 5.2 can be applied. A (numerically stable) square-root covariance filter [32] is used to compute a Kalman gain.

5.5. NUMERICAL RESULTS

The gradient descent procedure described in Section 5.2 was applied to a model of the segmented mirror with 2 rings, for a total number of $N = 18$ segments. This gives $n = 221$ system states, $r_y = 78$ sensors and $r_z = m_u = 54$ actuators. Several baseline controllers are generated for comparison purposes.

1. First of all, a (globally optimal) LQG controller was created.
2. Secondly, an SVD controller similar to the controller in [10] was implemented to compare the performance of the optimal controllers to a controller based on a modal approach.

A range of differently structured dynamic output feedback controllers were obtained.

3. A dynamic output feedback controller with full matrices A_c, B_c and C_c and a reduced number of controller states was designed using the gradient descent pro-

	Controller	n_c	A_c, B_c, C_c	RMS WF error (nm)
1.	LQG	221	f, f, f	6.32
2.	SVD controller	-	-	127.96
3.	Unstruct. red. order	54	f, f, f	6.52
4.	Fully interconnected	54	f, b, b	14.52
5.	Neighbours connected	54	s, b, b	20.26

Table 5.1: Performance of different control strategies of the 2-ring mirror model. The system matrix columns feature an 'f' for ' full matrix', 'b' for 'block-diagonal' or 's' for 'block sparse'.

cedure. Since each segment has 6 degrees of freedom, we choose each local controller to have 3 states, for a total $n_c = 3N = 54$ controller states. We refer to this controller as an unstructured, reduced order controller.

4. All the structured, reduced order ($n_c = 54$) controllers have block diagonal matrices B_c and C_c. The fully interconnected version (all local controllers are connected to all other local controllers) therefore has a full matrix A_c, and block diagonal matrices B_c and C_c.

5. Finally, a structured, reduced order ($n_c = 54$) controller is created where local controllers are connected to those local controllers that are associated with neighbouring segments, which is reflected in the block-structure of A_c. The matrices B_c and C_c are still block diagonal.

Table 5.1 records the RMS wavefront (WF) errors of controllers 1 through 5.

Figure 5.1 displays the performance of the structured, reduced order controllers obtained through optimization of (5.10) for different values of γ, resulting in the points marked 'x'. The values Γ_{ij} we chose to increase proportional to the square root of the Euclidean distance between the centers of segments i and j. A square root was used to not overly penalize the formation of longer distance connections but reflect a preference for shorter connections. The structure of B_c and C_c are fixed to block-diagonal.

For the optimization of (5.10) using gradient descent, different initialization strategies can be applied. First, stabilizing controllers with near-zero values in the system matrices are used. In this way the closed loop system is stable, and a solution to the Lyapunov equations in (5.5) can be found. Secondly, instead of a controller with near-zero matrices, the matrices of baseline controller 4 could have been used. A third option is that the value of γ could have been gradually changed and the optimization started with the controller from the previous optimization. We found that the second option gave the best results.

Once the gradient descent procedure converged, blocks with Frobenius norm below a threshold of 10^{-4} were deemed to not be in the interconnection structure. That is, there is no connection between segment i and j if

$$\left\| \begin{pmatrix} A_{c,kl} \\ A_{c,lk}^T \end{pmatrix} \right\|_F < 10^{-4}. \tag{5.14}$$

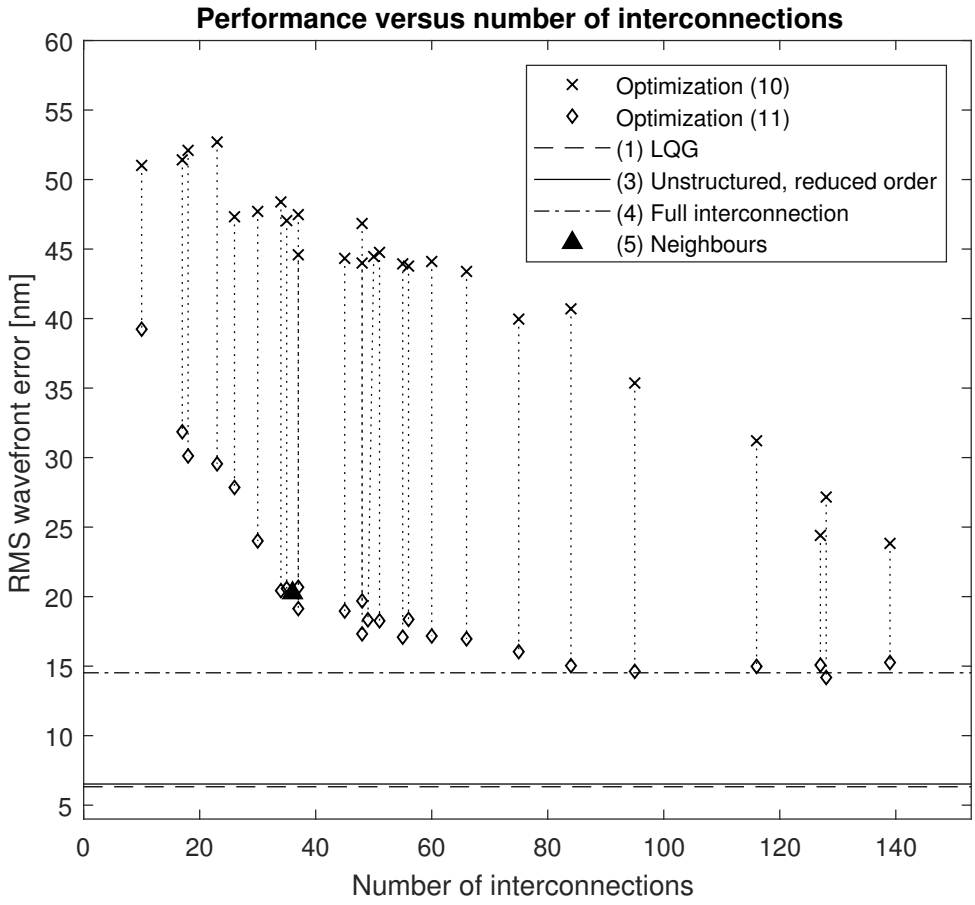


Figure 5.1: The tradeoff between number of interconnections and wavefront error. Controllers marked by 'x' obtained using (5.10). Controllers marked by a diamond are solutions to (5.11). The horizontal axis gives the value of (5.9) (the number of interconnections between local controllers). The vertical axis is the RMS wavefront error as explained in Section 5.3.2. The penalty paid for the use of the L1 norm as stand in for the cardinality operator is indicated by the vertical dotted lines.

The resulting interconnection structure defines $\mathscr{M}_{A_c}$, and (5.11) can be used to compute a locally optimal value for the performance of the structured controller. However, given that proper initialization can make a difference for the result in non-convex optimization, we utilized the availability of the matrices of baseline controller number 4. By setting γ high and Γ_{ij} to zero if segments i and j should be connected according to $\mathscr{M}_{A_c}$, and initializing the gradient descent procedure with the matrices of baseline controller 4, the 'fully interconnected' controller transforms into a controller with the proper structure during the gradient descent procedure. The resulting performance is marked in Figure 5.1 by a diamond marker. Corresponding 'discovery' and 'polishing' controllers are connected by a vertical dotted line. Apart from this initialization, a gradient descent method for problem (5.11) could also have been initialized, like the method for (5.10), with a controller with near-zero matrices, or the controller found in the 'discovery' procedure.

In the same figure, some of the baseline controllers are also indicated by horizontal lines, baseline controller number 5 is indicated with a star. Even though the 'fully interconnected' controller should be plotted by a point in this graph, for comparison with controllers with a sparser interconnection structure, a line is drawn.

In Figure 5.2 one of the discovered interconnection patterns (with 18 connections) is plotted on top of an image of the segmented mirror. A dashed line between two segment centres indicates that the local controllers are connected according to the sparsity structure in A_c.

5.6. DISCUSSION

Immediately apparent from Table 5.1 is the small difference between the global optimum (controller 1, the LQG controller) and the reduced order unstructured controller (controller 3). When controller 3 is compared to controller number 4, where the difference is the structure in B_c and C_c, we see the wavefront error more than doubles.

The points marked with 'x' are the performances of the best controllers found by optimizing (5.10) for different values of γ and different initial guesses for the controller matrices. A tradeoff can clearly be seen between the number of interconnections and performance. The identified interconnection structures for these controllers were used in a structured controller optimization (problem (5.11)), and the resulting performances are indicated with diamond markers. The difference between the tradeoff curves for the discovery procedure (marked with 'x') and the subsequent structured optimization (diamond markers) is relatively large and the resulting curves are not smooth. The large differences in performance clearly show the penalty paid for using the Frobenius-norm as a differentiable substitute for the cardinality operator. What is clear from the 'polished' curve is that fewer than half of the interconnections between mirror segment controllers are not necessary for approximately the same performance as the baseline controller 4. This observation is not clear from the 'discovery'-curve, and neither is it from fixing the controller structure heuristically like in the baseline controllers.

The relatively small degradation in performance for controllers with approximately a third of possible interconnections not only justifies a search for optimal interconnection structures in distributed systems, but is also relevant for situations where the online computation time of the input signal is critical. Even though the block-sparse matrices

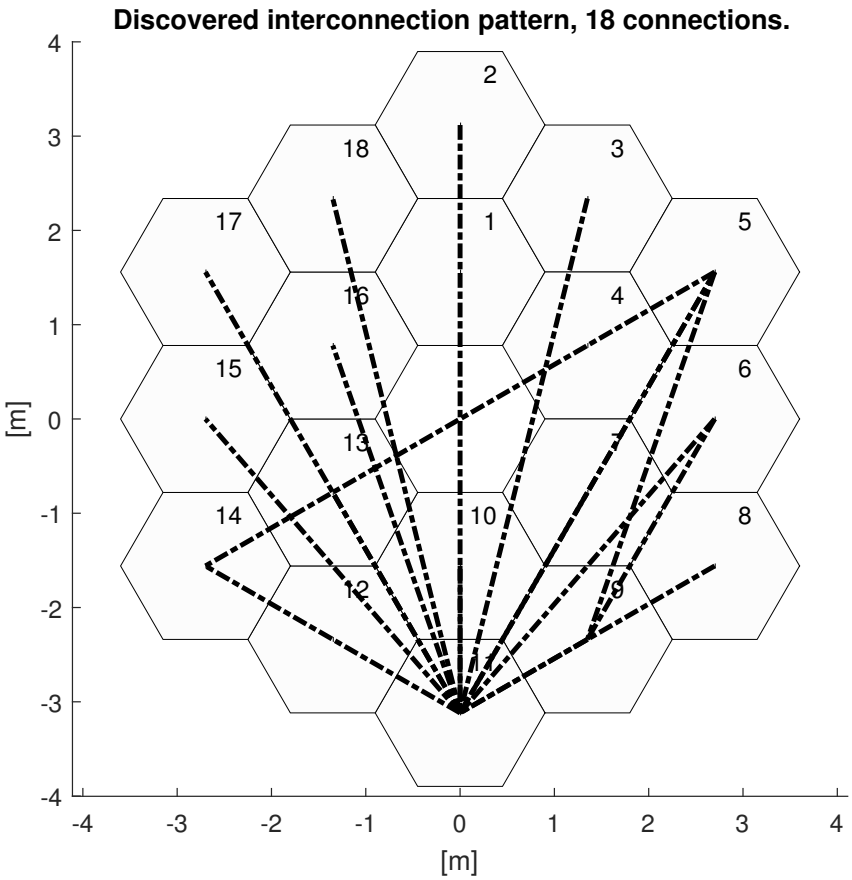


Figure 5.2: An example of an interconnection pattern generated by the discovery procedure.

A_c have an interpretation as being a controller that is distributed, see (5.8), one might choose to do the controller computations in a centralized manner. Let n_{lc} be the dimension of the local controller, such that $n_c = Nn_{lc}$, and let p be the maximum number of blocks on a block-row of A_c,

$$p = \max_i \sum_j^N \text{card}\left(\|A_{c,ij}\|_F\right).$$

The multiplication of $A_c x_c$ has a computational complexity of $\mathscr{O}\left(n_c^2\right) = \mathscr{O}\left(N^2 n_{lc}^2\right)$ for a dense matrix A_c. For a block sparse matrix A_c it is possible to exploit the sparsity and parallellize the computation like in (5.8), and compute the result with computational complexity $\mathscr{O}\left(pn_{lc}^2\right)$. The multiplication $B_c y$ can be computed with computational complexity $\mathscr{O}\left(n_c r_y\right) = \mathscr{O}\left(n_{lc} N r_y\right)$. However, for block diagonal B_c this can be efficiently parallelized and computed in $\mathscr{O}\left(n_{lc} r_{ly}\right)$, where r_{ly} are the maximum number of measurements per segment. A similar argument can be given for the computation of $C_c x_c$. Through efficient use of the sparsity in the controller matrices in a centralized implementation of the controller, the online computation time is not related to the number of subsystems, but to the pre-chosen number of states n_{lc}, and through p to the introduced level of sparsity. We see that for the analyzed segmented mirror there can be a strong improvement in online computation time with only a small loss in performance. By optimizing the interconnection structure, the online computation time can be traded off against system performance.

In Figure 5.2, one of the discovered sparsity patterns is displayed. One thing to notice is the absence of (rotational) symmetry and difference with a controller where neighbouring segments are connected. In a sense such a symmetry was expected, since the segment configuration is rotationally symmetric, all segments are the same, and a flat mean wind velocity profile perpendicular to the mirror plane was assumed for the disturbance. We do see that for this sparsity level, optimization (5.10) resulted in a master-slave type of controller, where the controller of segment 11 connects to nearly all other controllers, and there are only few connections among the other local controllers. Similar master-slave type of controllers can be observed for larger number of interconnections, with a few more local controllers with a 'master' role.

Another interesting property of the controllers is that the computation of the control action does not feature the numerical issues of the original model.

Finally, it is important to note that many of the structured controllers in Figure 5.1 have a remaining wavefront error that is lower than the maximum allowed wavefront error of 27.6 nm, meaning that the tradeoff analysis can and should play a role in system design.

5.7. CONCLUSIONS

We demonstrated how the performance – the residual wavefront error of the closed loop system – of a structured controller, with an interpretation as a distributed controller, can be systematically traded off with the complexity of the distributed controller - interpreted as the number of interconnections between the local controllers. The method was applied to a challenging virtual model of a segmented mirror on a flexible support

truss, and a wind disturbance model with spatial correlation. For this system a range of structured controllers were computed. The results show that compared to the performance of fully interconnected local controllers – controller 4 in Table 5.1 – the amount of interconnections can be greatly reduced without significant loss of performance. Furthermore, for small amounts of interconnections and the particular weighting of interconnections we chose, controllers with a centralized aspect seem to be preferred over controllers where interconnections are distributed in a spatial sense.

For many different amounts of interconnections, the designed controllers have residual wavefront errors below the maximum allowed error and a performance improvement can be found by optimizing the interconnection structure of local controllers with respect to heuristic interconnection structures.

5.8. FUTURE WORK

As future work, we recommend that robustness against modelling uncertainty is investigated. That is, do the resulting interconnection patterns change if model uncertainty is taken into account? Or if the mirror design parameters change? Furthermore, is the centralized character of the optimal interconnection pattern, observed in Figure 5.2, also present in the interconnection pattern of mirrors with more segments? The gap in performance between the controllers with (nearly) full A_c matrices and structured B_c and C_c matrices on the one hand and the unstructured, reduced order controller on the other hand, would motivate the further inclusion of block-sparsity promoting terms in B_c and C_c in the objective function of optimization in the discovery procedure. It also leads to the question whether this inclusion affects the identified interconnection structures to a significant degree.

5.9. FUNDING INFORMATION

The research leading to these results has received funding from the European Research Council (ERC) under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement No. 339681.

REFERENCES

[1] R. Doelman, S. Dominicus, R. Bastaits, and M. Verhaegen, "Systematically structured $\mathscr{H}_2$ optimal control for truss-supported segmented mirrors," IEEE Transactions on Control Systems Technology, no. 99, pp. 1–8, 2018.

[2] J.-N. Aubrun, K. Lorell, T. Mast, and J. Nelson, "Dynamic analysis of the actively controlled segmented mirror of the WM Keck ten-meter telescope," IEEE Control Systems Magazine, vol. 7, no. 6, pp. 3–10, 1987.

[3] B. Sedghi, M. Müller, M. Dimmler, B. Bauvir, T. Erm, H. Bonnet, and M. Cayrel, "Dynamical aspects in control of E-ELT segmented primary mirror (M1)," in Proc. SPIE, vol. 7733, pp. 77332E–77332E, 2010.

[4] B. Sedghi, M. Müller, H. Bonnet, M. Dimmler, and B. Bauvir, "Field stabilization (tip/tilt control) of E-ELT," in Proc. SPIE, vol. 7733, p. 773340, 2010.

[5] M. Dimmler, T. Erm, B. Bauvir, B. Sedghi, H. Bonnet, M. Müller, and A. Wallander, "E-ELT primary mirror control system," in Proc. SPIE Vol, vol. 7012, 2008.

[6] D. G. MacMynowski and T. Andersen, "Wind buffeting of large telescopes," Applied optics, vol. 49, no. 4, pp. 625–636, 2010.

[7] D. G. MacMynowski, P. M. Thompson, J. C. Shelton, L. C. Roberts Jr, M. M. Colavita, and M. J. Sirota, "Control system modeling for the Thirty Meter Telescope primary mirror," Society of Photo-Optical Instrumentation Engineers, 2011.

[8] D. G. MacMartin and K. Vogiatzis, "Unsteady wind loads for TMT: Replacing parametric models with CFD," Society of Photo-Optical Instrumentation Engineers (SPIE), 2014.

[9] D. G. MacMynowski, M. M. Colavita, W. Skidmore, and K. Vogiatzis, "Primary mirror dynamic disturbance models for TMT: Vibration and wind," Society of Photo-optical Instrumentation Engineers (SPIE), 2010.

[10] R. Bastaits, Extremely large segmented mirrors: dynamics, control and scale effects. PhD thesis, 2010.

[11] R. Bastaits, G. Rodrigues, B. Mokrani, and A. Preumont, "Active optics of large segmented mirrors: dynamics and control," Journal of guidance, control, and dynamics, vol. 32, no. 6, p. 1795, 2009.

[12] A. Sarlette and R. J. Sepulchre, "Control limitations from distributed sensing: Theory and extremely large telescope application," Automatica, vol. 50, no. 2, pp. 421–430, 2014.

[13] S. Jiang, P. G. Voulgaris, L. E. Holloway, and L. A. Thompson, "H_2 Control of Large Segmented Telescopes," Journal of Vibration and Control, 2009.

[14] F. Lin, Structure identification and optimal design of large-scale networks of dynamical systems. PhD thesis, University of Minnesota, 2012.

[15] F. Lin, M. Fardad, and M. R. Jovanovic, "Augmented Lagrangian approach to design of structured optimal state feedback gains," IEEE Transactions on Automatic Control, vol. 56, no. 12, pp. 2923–2929, 2011.

[16] F. Lin, M. Fardad, and M. R. Jovanović, "Design of optimal sparse feedback gains via the Alternating Direction Method of Multipliers," IEEE Transactions on Automatic Control, vol. 58, no. 9, pp. 2426–2431, 2013.

[17] M. R. Jovanović and N. K. Dhingra, "Controller architectures: Tradeoffs between performance and structure," European Journal of Control, 2016.

[18] M. Mercadal, H_2, fixed architecture, control design for large scale systems. PhD thesis, Massachusetts Institute of Technology, 1990.

[19] D. Petersson, A nonlinear optimization approach to H_2-optimal modeling and control. PhD thesis, Linköping University Electronic Press, 2013.

[20] G. Chanan, D. G. MacMartin, J. Nelson, and T. Mast, "Control and alignment of segmented-mirror telescopes: matrices, modes, and error propagation," Applied Optics, vol. 43, no. 6, pp. 1223–1232, 2004.

[21] C. Scherer, P. Gahinet, and M. Chilali, "Multiobjective output-feedback control via LMI optimization," IEEE Transactions on automatic control, vol. 42, no. 7, pp. 896–911, 1997.

[22] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[23] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 68, no. 1, pp. 49–67, 2006.

[24] D. G. MacMynowski, C. Blaurock, and G. Z. Angeli, "Initial control results for the thirty meter telescope," in AIAA Guidance, Navigation and Control Conference, 2005.

[25] B. Ulutas, D. Kerley, J. Dunn, A. Suleman, and E. J. Park, "Distributed H_∞ control of dynamically coupled segmented telescope mirrors: Design and simulation," Mechatronics, vol. 22, no. 1, pp. 121–135, 2012.

[26] B. Sedghi, M. Müller, and M. Dimmler, "Analyzing the impact of vibrations on E-ELT primary segmented mirror," in Modeling, Systems Engineering, and Project Management for Astronomy VI, vol. 9911, p. 991111, International Society for Optics and Photonics, 2016.

[27] A. G. Davenport, "The spectrum of horizontal gustiness near the ground in high winds," Quarterly Journal of the Royal Meteorological Society, vol. 87, no. 372, pp. 194–211, 1961.

[28] K. Hinnen, M. Verhaegen, and N. Doelman, "Robust spectral factor approximation of discrete-time frequency domain power spectras," Automatica, vol. 41, no. 10, pp. 1791–1798, 2005.

[29] A. Maréchal, "Étude des effets combinés de la diffraction et des aberrations géométriques sur l'image d'un point lumineux," Rev. Opt, vol. 26, pp. 257–277, 1947.

[30] The E-ELT Project office, E-ELT Construction Proposal. ESO, 2011.

[31] A. Binder, V. Mehrmann, A. Miedlar, and P. Schulze, "A matlab toolbox for the regularization of descriptor systems arising from generalized realization procedures," 2015.

[32] M. Verhaegen and V. Verdult, Filtering and system identification: a least squares approach. Cambridge university press, 2007.

6

CONCLUSIONS AND RECOMMENDATIONS

This thesis discusses a number of estimation and control problems related to optical imaging. Chapter 2 proposes an algorithm to estimate the wavefront aberration based on measurements of the point spread function of the system. Chapter 3 proposes an algorithm to identify the temporal dynamics of the wavefront aberration based on measurements of the points spread function of the system. Chapter 4 proposes a convex optimization-based algorithm to estimate the wavefront aberration based on images taken with coherent illumination. Chapter 5 proposes an algorithm to design distributed locally $\mathscr{H}_2$ optimal controllers to reduce the effect of wind load on wavefront aberrations in telescopes.

In this chapter we summarize the conclusions and propose several lines of research based on the work in this thesis.

6.1. CONCLUSIONS

In this thesis we have proposed a method to solve a range of estimation problems, based on iterative convex optimizations. The method allows for a different approach to a number of problems - not only in the field of optics, but also in the fields of control and identification, as demonstrated in this thesis. From the applications demonstrated here, we conclude that the approach is effective and that there is an opportunity to apply and test these methods to different problems.

Phase retrieval The Convex Optimization-based Phase Retrieval (COPR) algorihm is based on an iteratively applied heuristic convex optimization problem, the nuclear norm. Importantly, the coefficients that are sought-after appear affinely in the optimization problem, enabling the easy incorporation of common types of prior knowledge, such as sparsity. Since the nuclear norm is typically an optimization problem with a high computational complexity, we showed how the Alternating Direction Method of Mul-

tipliers (ADMM) algorithm reduces to updates based on matrix-vector multiplications (solving a least squares problem using a pseudo-inverse) and singular value decompositions of 2-by-2 matrices. We provided some proofs of convergence for specific cases, and demonstrated general applicability on synthetic and experimental phase retrieval problems.

Blind identification of wavefront dynamics A powerful concept in phase retrieval is phase diversity, the availability of several images with a known difference in phase (for example a different defocus). If the phase is time-varying, it is difficult to capture these different images without introducing different optical paths in the system. In Chapter 3 we introduced the use of a model set for the dynamics of the time-varying phase aberration as prior information in the phase retrieval problem. From a system identification perspective, where the dynamic model is of interest and not necessarily the aberration itself, this is the blind identification problem for a Wiener system with squared output measurements. We demonstrated the proposed method on an example and compared its performance with a standard nonlinear least-squared-error minimization method.

Blind deconvolution for coherent imaging Standard phase retrieval algororithms assume that either the image taken with an optical system is that of the point spread function, or that between the imaged object and the point spread function, one of these is known. If this problem is generalized to both object and PSF unknown, we obtain a blind deconvolution problem. Chapter 4 introduced a reformulation of the blind deconvolution problem for images taken with coherent illumination to a rank constrained problem, and proposed the use of a convex heuristic, the nuclear norm, to obtain a solution. We demonstrated how this new method can take into account phase diversity - by which it sets itself apart from other algorithms in its class - and is able to obtain a solution where a standard gradient descent algorithm could not.

Distributed control for telescopes with wind load disturbance Typically the controller for the primary segmented mirror in telescopes is one that is centralized, and in practice based on modal analysis of the dynamics. Given the increasing scale of segemented mirrors in future telescopes, bringing with it an increased amount of communication overhead for the control loop, it is worthwile to consider distributed controllers to address this issue. By their nature a segmented mirror can be seen as a collection of subsystems - the segments -, but not all sources of dynamics are similarly decomposable. For example, the measurements are relative, the wind load is correlated and the supporting structure couples the dynamics of the subsystems. In Chapter 5 we research the design of a distributed, locally $\mathscr{H}_2$ optimal controller, that takes into account explicitly these global dynamics and the amount of interconnections necessary between the different controllers. This gives a trade-off curve between on one hand the optical performance of the telescope and on the other hand the amount of interconnections between the local controllers of the segments.

6.2. RECOMMENDATIONS

Convergence speed: theoretical bounds, tuning rules and low rank inducing norms The sequential application of the nuclear norm on inherently bilinear problems has great practical applicability for a large number of problems in the field of systems and control. Problems are easily reformulated and implementation is facilitated by the use of standard middleware and standard convex optimization software. Feasible solutions to the bilinear problem are not required as starting points and in its most basic form it features a single tuning parameter. Practical use not withstanding, there is not a proof for convergence to a stationary point or to a feasible solution in the most general formulation. Neither general results on convergence speed are obtained and what their relation is to the tuning parameters, even though in practice a clear link can be observed. Finally, the performance related to the use of the nuclear norm can be compared to the performance of different low rank inducing norms.

Parallel processing Several aspects benefit the adoption of a proposed algorithm. Practical applicability, speed and performance on hard optimization problems are some of these. Some of the proposed algorithms in this thesis have the property that their ADMM implementations consist of least squares problems and fully paralellizable singular value decompositions, see also Appendix E. COPRs in Chapter 2 for example requires the computation of a 2-by-2 Singular Value Decomposition (SVD) for every measured pixel, for every ADMM iteration. The blind identification problem likewise has SVDs for every pixel. SVDs of this very small size can be implemented in Graphics Processing Unit (GPU) kernel functions, so that the GPU can compute the ADMM updates in parallel. So far, the SVD is the bottleneck of the computation time of COPRs and an interisting question would be what a potential computational speedup could be accomplished by a GPU. Likewise for the blind deconvolution problem, the SVDs could be distributed across computing cores of a cluster, instead of being computed sequentially on a single computer. Hopefully this would allow for experimental validation with computing times with reasonable limits.

Prior information: anisoplanatism and power series for phase approximations In Chapter 3 the temporal correlation of phase aberrations was taken into account in the estimation. In Chapter 4, we assumed that the image was taken under isoplanatic conditions, i.e. the amplitude transfer function is constant throughout the image. Under anisoplanatic conditions, the amplitude transfer function varies throughout an image. If there is a spatial correlation, it would be interesting to use a spatial dynamics model-set constraints in order to reconstruct the aberrations for different points in an image.

Another crucial assumption in Chapter 4 was the assumption called the 'small phase' approximation. The assumption is based on the first two terms of the Taylor (Maclaurin) series approximating the exponential function. The validity of the method is inherently dependent on the validity of the approximation. Larger phase aberrations require more terms of the power series for the approximation to hold. The insight here that would warrant further investigation is that these truncated power series can be reformulated as bilinear constraints, something that not just holds for scalar functions, but also matrix

functions like the matrix exponential. The question is, would the methods proposed in this thesis also produce good results for these higher order approximations?

Optimal interconnections, robustness and LPV systems Chapter 5 focused on $\mathscr{H}_2$ optimal controllers. The issues of robustness were not addressed. Robustness can be interpreted in two ways. First, the model did not incorporate for example slight differences in masses of the segments, the changing wind conditions, or the presence of other modelling errors. How the model of Chapter 5 should be modified to obtain for example a Linear Paramater-Varying (LPV) system is still to be investigated. But secondly, what effect do these model changes have on the interconnection structure that was deemed to be optimal for a specific model? The $\mathscr{H}_2$ optimal solutions were optimized using gradient descent. This is partly due to the problem size and the numerical stability of controller design software, but also because the $\mathscr{H}_\infty$ performance is not necessarily differentiable with respect to its controller. The methods proposed in this thesis can facilitate the design of structured robust controllers, but how the resulting structure changes under changing parameters is something left to be investigated.

A

APPENDIX FOR CHAPTER 2

A.1. PROOF OF LEMMA 2.5.1

Proof. Let $\mathbf{a}$ satisfy $\mathbf{y} = |U\mathbf{a}|^2$. It suffices to check that $\mathbf{a} \in T(\mathbf{a})$. We first observe that

$$\operatorname{rank}\big(M(U,\mathbf{a},-\mathbf{a},\mathbf{y})\big) = \operatorname{rank}\begin{pmatrix} 0 & 0 \\ 0 & I_{n_y} \end{pmatrix} = n_y.$$

This means that $\mathbf{a}$ is a global minimizer of $\operatorname{rank} M(U,\mathbf{x},-\mathbf{a},\mathbf{y})$ as a function of $\mathbf{x} \in \mathbb{C}^{n_a}$. Since the nuclear norm $\|M(U,\mathbf{x},-\mathbf{a},\mathbf{y})\|_*$ is the convex envelop of the $\operatorname{rank} M(U,\mathbf{x},-\mathbf{a},\mathbf{y})$, they have the same global minimizers. Hence, $\mathbf{a}$ is also a global minimizer of $\|M(U,\mathbf{x},-\mathbf{a},\mathbf{y})\|_*$ as a function of $\mathbf{x}$, that is

$$\mathbf{a} \in \arg\min_{\mathbf{x}\in\mathbb{C}^{n_a}} \|M(U,\mathbf{x},-\mathbf{a},\mathbf{y})\|_*.$$

In other words, $\mathbf{a} \in T(\mathbf{a})$ and the proof is complete. □

A.2. PROOF OF THEOREM 2.5.2

Lemma A.2.1 will serve as the basic step for proving Theorem 2.5.2.

Lemma A.2.1. Let $U = I_{n_a}$ and $\mathbf{a}^* \in \mathbb{C}^{n_a}$ be such that $|U\mathbf{a}^*|^2 = \mathbf{y}$. Then every Picard iteration $\mathbf{a}_{k+1} \in T(\mathbf{a}_k)$ starting sufficiently close to $\mathbf{a}^*$ converges linearly to a point $\tilde{\mathbf{a}} \in \operatorname{Fix} T$ satisfying $|U\tilde{\mathbf{a}}|^2 = \mathbf{y}$.

Proof. Since $U = I_{n_a}$, the nuclear norm of $M(I_{n_a},\mathbf{x},-\mathbf{a},\mathbf{y})$ can be calculated from the nuclear norms of n_a matrices $M(1,x_i,-a_i,y_i) \in \mathbb{C}^{2\times 2}$ $(1 \le i \le n_a)$. Let us do the calculation for an arbitrary $\mathbf{a} \in \mathbb{C}^{n_a}$. We first calculate the nuclear norm of each 2×2 matrix

$$M(1,x_i,-a_i,y_i) = \begin{pmatrix} y_i - 2\mathscr{R}(x_i\overline{a_i}) + |a_i|^2 & x_i - a_i \\ \overline{x_i} - \overline{a_i} & 1 \end{pmatrix}.$$

A

Indeed, we have by direct calculation that

$$\begin{aligned} f_i(x_i) &:= \left\| M(1, x_i, -a_i, y_i) \right\|_*^2 \\ &= \left\| \begin{pmatrix} r & s \\ s & 1 \end{pmatrix} \right\|_*^2 \\ &= r^2 + 2s^2 + 1 + 2|r - s^2|, \end{aligned} \tag{A.1}$$

where

$$r := y_i - 2\mathscr{R}(x_i \overline{a_i}) + |a_i|^2, \quad s := |x_i - a_i|.$$

Let us denote

$$T_i(a_i) := \operatorname*{arg\,min}_{x_i \in \mathbb{C}} f_i(x_i). \tag{A.2}$$

Solving analytically the minimization problem on the right-hand side of (A.2), we obtain the explicit form of T_i as follows

$$T_i(a_i) = \begin{cases} \left\{ z \in \mathbb{C} \mid |z| \le \sqrt{y_i} \right\}, & \text{if} \quad a_i = 0, \\ \left\{ \frac{\sqrt{y_i}}{|a_i|} a_i \right\}, & \text{if} \quad 0 < |a_i| \le \sqrt{\lambda_i}, \\ \left\{ \frac{y_i + |a_i|^2 + 1}{2(|a_i|^2 + 1)} a_i \right\}, & \text{if} \quad |a_i| \ge \sqrt{\lambda_i}, \end{cases} \tag{A.3}$$

where λ_i is the unique real positive root of the real polynomial $g_i(t) := t^3 + 2(1 - y_i)\, t^2 + (y_i^2 - 6y_i + 1)\, t - 4y_i$.

We need to take care of the two possible cases of y_i.

Case 1. $y_i \in (0, 1]$. Then we have $\frac{3}{2}\sqrt{y_i} < \sqrt{\lambda_i} < 2\sqrt{y_i}$ since $g_i\left(\frac{9}{4} y_i\right) < 0$ and $g_i\left(4 y_i\right) > 0$. The following properties of T_i can be verified.

- $\operatorname{Fix} T_i = \left\{ z \in \mathbb{C} \mid |z| = \sqrt{y_i} \right\} \cup \{0\}$, where 0 is an inhomogeneous fixed point of T_i, that is, $T_i(0) \nsubseteq \operatorname{Fix} T_i$.
- The set of homogeneous fixed points of T_i is $S_i := \left\{ z \in \mathbb{C} \mid |z| = \sqrt{y_i} \right\}$.
- T_i is pointwise averaging at every point of S_i on $W_i := \{ z \in \mathbb{C} \mid |z| \ge \sqrt{y_i}/2 \}$ with constant $3/4$.
- The set-valued mapping $\psi_i := T_i - \mathrm{Id}$ is metrically subregular on W_i for 0 with constant $1/2$.
- The technical assumption $\operatorname{dist}(z, S_i) \le \operatorname{dist}(z, \operatorname{Fix} T_i)$ holds for all $z \in W_i$.

Case 2. $y_i = 0$. Then $\lambda_i = 0$. Note also that $a_i^* = 0$ and the formula (A.3) becomes $T_i(a_i) = \frac{1}{2} a_i$. The following properties of T_i can be verified.

- $\operatorname{Fix} T_i = \{0\}$, where 0 is a homogeneous fixed point of T_i.
- T_i is pointwise averaging at every point of S_i on $\mathbb{C}$ with constant $1/4$.
- The set-valued mapping $\psi_i := T_i - \mathrm{Id}$ is metrically subregular on $\mathbb{C}$ for 0 with constant $1/2$.

- The technical assumption $\operatorname{dist}(z, S_i) \le \operatorname{dist}(z, \operatorname{Fix} T_i)$ holds for all $z \in \mathbb{C}$.

In this case, we denote $S_i := \{0\}$ and $W_i := \mathbb{C}$.

The operator T can be calculated explicitly

$$T(\mathbf{a}) = \arg\min_{\mathbf{x}\in\mathbb{C}^{n_a}} \sum_{i=1}^{n_a} \sqrt{f_i(x_i)}, \quad \forall \mathbf{a} \in \mathbb{C}^{n_a}, \tag{A.4}$$

where the constituent functions $f_i(x_i)$ are given by (A.1).

Minimizing f_i $(i = 1, 2 \ldots, n_a)$ separately yields the explicit form of T as a Cartesian product

$$T(\mathbf{a}) = T_1(a_1) \times T_2(a_2) \cdots \times T_{n_a}(a_{n_a}), \tag{A.5}$$

where the component operators T_i are given by (A.3).

Thanks to the separability structure of T as a Cartesian product at (A.5), the following properties of T in relation to Proposition 2.5.1 can be deduced from the corresponding ones of the component operators T_i.

- $\operatorname{Fix} T = \prod_{i=1}^{n_a} \operatorname{Fix} T_i$ and the set of homogeneous fixed points of T is $S := \prod_{i=1}^{n_a} S_i$. It is clear that $|U\mathbf{a}|^2 = \mathbf{y}$ for $U = I_{n_a}$ and all $\mathbf{a} \in S$.
- T is pointwise averaging at every point of S on $W := \prod_{i=1}^{n_a} W_i$ with constant $\alpha = 3/4$.
- The set-valued mapping $\psi := T - \mathrm{Id}$ is metrically subregular on W for 0 with constant $\kappa = 1/2$.
- The technical assumption (iii) of Proposition 2.5.1 is satisfied on W. That is,

$$\operatorname{dist}(\mathbf{w}, S) \le \operatorname{dist}(\mathbf{w}, \operatorname{Fix} T), \quad \forall \mathbf{w} \in W. \tag{A.6}$$

Now we can apply Proposition 2.5.1 to conclude that every Picard iteration $\mathbf{a}_{k+1} \in T(\mathbf{a}_k)$ starting in W converges linearly to a point in S as claimed. □

Remark A.2.1. Under the assumption that $y_i > 0$ for all $1 \le i \le n_a$, then the linear convergence result established in Lemma A.2.1 can be sharpened to finite convergence.

In order to distinguish the fixed point operator (2.25) corresponding to a general unitary matrix U from the one analyzed in Lemma A.2.1 corresponding to the identity matrix I_{n_a}, in the following proof, we will use the notation $\widehat{T}$ for one specified in Theorem 2.5.2.

Proof. Let T be the fixed point operator (2.25) which corresponds to the identity matrix and has been analyzed in Lemma A.2.1. We start the proof by proving that

$$\widehat{T}(\mathbf{a}) = U^{-1} T(U\mathbf{a}), \quad \forall \mathbf{a} \in \mathbb{C}^{n_a}. \tag{A.7}$$

A

Indeed, let us take an arbitrary $\mathbf{a} \in \mathbb{C}^{n_a}$ and denote $\mathbf{a}' = U\mathbf{a}$. Then we have

$$\begin{aligned}\widehat{T}(\mathbf{a}) &= \arg\min_{\mathbf{x}\in\mathbb{C}^{n_a}} \left\|M(U,\mathbf{x},-\mathbf{a},\mathbf{y})\right\|_* \\ &= \arg\min_{\mathbf{x}\in\mathbb{C}^{n_a}} \left\|M(I_{n_a},U\mathbf{x},-\mathbf{a}',\mathbf{y})\right\|_* \\ &= U^{-1}\left(\arg\min_{\mathbf{x}\in\mathbb{C}^{n_a}} \left\|M(I_{n_a},\mathbf{x},-\mathbf{a}',\mathbf{y})\right\|_*\right) \\ &= U^{-1}\left(T(\mathbf{a}')\right) = U^{-1}\left(T(U\mathbf{a})\right).\end{aligned} \tag{A.8}$$

We have proved (A.7). As a consequence,

$$\begin{aligned}\mathsf{Fix}\ \widehat{T} &= \{\mathbf{a}\in\mathbb{C}^{n_a} \mid \mathbf{a}\in\widehat{T}(\mathbf{a})\} \\ &= \{\mathbf{a}\in\mathbb{C}^{n_a} \mid \mathbf{a}\in U^{-1}T(U\mathbf{a})\} \\ &= \{\mathbf{a}\in\mathbb{C}^{n_a} \mid U\mathbf{a}\in T(U\mathbf{a})\} \\ &= \{\mathbf{a}\in\mathbb{C}^{n_a} \mid U\mathbf{a}\in \mathsf{Fix}\ T\} = U^{-1}\left(\mathsf{Fix}\ T\right).\end{aligned} \tag{A.9}$$

For the sets S and W determined in the proof of Lemma A.2.1, we denote $\widehat{S} := U^{-1}(S)$ and $\widehat{W} := U^{-1}(W)$. Since U is a unitary matrix, the set of homogeneous fixed points of $\widehat{T}$ is $\widehat{S} := U^{-1}(S)$. It also holds by the definition of projection and (A.9) that, for all $\mathbf{w}\in W$,

$$P_{U^{-1}(S)}\left(U^{-1}\mathbf{w}\right) = U^{-1}\left(P_S(\mathbf{w})\right), \tag{A.10}$$

$$\operatorname{dist}\left(U^{-1}\mathbf{w}, U^{-1}(S)\right) = \operatorname{dist}\left(U^{-1}\mathbf{w}, U^{-1}(\mathsf{Fix}\ T)\right). \tag{A.11}$$

By direct calculation one can verify the three assumptions on $\widehat{T}$ imposed in Proposition 2.5.1.

- $\widehat{T}$ is point-wise averaging at every point of $\widehat{S}$ on $\widehat{W}$ with constant $\alpha = 3/4$.
- The set-valued mapping $\widehat{\psi} := \widehat{T} - \mathrm{Id}$ is metrically subregular on $\widehat{W}$ for 0 with constant $\gamma = 1/2$.
- The technical assumption (iii) of Proposition 2.5.1 is satisfied on $\widehat{W}$.

Therefore, we can apply Proposition 2.5.1 to conclude that every Picard iteration $\mathbf{a}_{k+1}\in\widehat{T}(\mathbf{a}_k)$ generated by the COPR algorithm starting in $\widehat{W}$ converges linearly to a point $\tilde{\mathbf{a}}\in\widehat{S}$. Finally, let $\tilde{\mathbf{w}}\in S$ such that $\tilde{\mathbf{a}} = U^{-1}\tilde{\mathbf{w}}$. It holds that $|U\tilde{\mathbf{a}}|^2 = |\tilde{\mathbf{w}}|^2 = \mathbf{y}$ by the structure of S.

The proof is complete. □

B

APPENDIX FOR CHAPTER 3

B.1. THE MATRICES IN EQ. (3.37)

$$G = \begin{pmatrix} I_n \otimes \alpha_{K-1}^T & I_n \otimes \alpha_{K-2}^T & \dots & I_n \otimes \alpha_{K-M}^T \\ I_n \otimes \alpha_{K-2}^T & I_n \otimes \alpha_{K-3}^T & \dots & I_n \otimes \alpha_{K-M-1}^T \\ \vdots & \vdots & \vdots & \vdots \\ I_n \otimes \alpha_{K-N}^T & \dots & \dots & I_n \otimes \alpha_1^T \\ 0 & \dots & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & \dots & 0 \end{pmatrix} \tag{B.1}$$

$$h = \begin{pmatrix} -\alpha_K \\ -\alpha_{K-1} \\ \vdots \\ -\alpha_{N+1} \\ y_{1,K} - D_{0,1}(\beta_k) - D_{1,1}(\beta_k)\alpha_K - \alpha_K^T D_{2,1}(\beta_k)\alpha_K \\ \vdots \\ y_{p^2,K} - D_{0,p^2}(\beta_k) - D_{1,p^2}(\beta_k)\alpha_K - \alpha_K^T D_{2,p^2}(\beta_k)\alpha_K \\ \vdots \\ y_{1,1} - D_{0,1}(\beta_k) - D_{1,1}(\beta_k)\alpha_K - \alpha_K^T D_{2,1}(\beta_k)\alpha_K \\ \vdots \\ y_{p^2,1} - D_{0,p^2}(\beta_k) - D_{1,p^2}(\beta_k)\alpha_1 - \alpha_1^T D_{2,p^2}(\beta_k)\alpha_1 \end{pmatrix}. \tag{B.2}$$

B.2. SETTINGS OF NONLINEAR SOLVER

Apart from `SpecifyObjectiveGradient='true'`, default settings for `lsqnonlin` have been used for all experiments involving SNLLS. For a complete list of the default settings, we refer to MATLAB's official documentation.

C

APPENDIX FOR CHAPTER 4

C.1. PROOF OF LEMMA 4.3.2

The vector $z = \text{vect}\left(\text{vect}\left(\mathbf{g}_o\right)\text{vect}\left(\mathbf{Bv}\right)^T\right)$ lists all possible products between elements of $\mathbf{g}_o$ and elements of $\mathbf{Bv}$. Since the elements of $\mathbf{g}_i$, the result of a discrete convolution, are sums of specific elements of z, it is possible to construct a matrix of zeros and ones $L \in \mathbb{B}^{rt \times mnpq}$, such that

$$\text{vect}\left(\mathbf{g}_i\right) = L\,\text{vect}\left(\text{vect}\left(\mathbf{g}_o\right)\text{vect}\left(\mathbf{Bv}\right)^T\right). \tag{C.1}$$

Application of the identity [1]

$$\text{vect}\left(AXB\right) = \left(B^T \otimes A\right)\text{vect}\left(X\right), \tag{C.2}$$

allows us to rewrite (C.1) into

$$L\left(I_{pq} \otimes \text{vect}\left(\mathbf{g}_o\right)^T\right)\text{vect}\left(\mathbf{Bv}\right). \tag{C.3}$$

Let l_i be the i'th row of L. Then applying (C.2) on the i'th row of (C.3) gives

$$l_i\left(I_{pq} \otimes \text{vect}\left(\mathbf{g}_o\right)^T\right) = \text{vect}\left(\mathbf{g}_o\right)^T L_i, \tag{C.4}$$

where $l_i^T = \text{vect}\left(L_i\right)$ and $L_i \in \mathbb{B}^{mn \times pq}$. Combining the expressions for all rows, the result is

$$\text{vect}\left(\mathbf{g}_i\right) = \left(I_{rt} \otimes \text{vect}\left(\mathbf{g}_o\right)^T\right)\begin{pmatrix} L_1 \\ \vdots \\ L_{rt} \end{pmatrix}\text{vect}\left(\mathbf{Bv}\right). \tag{C.5}$$

Using a row reordering of the matrix with blocks L_i, [1, eq. 2.14], gives us V the expression in (4.14).

REFERENCES

[1] D. A. Turkington, Generalized vectorization, cross-products, and matrix calculus. Cambridge University Press, 2013.

D

Appendix for Chapter 5

D.1. Gradients of the $\mathscr{H}_2$ norm with respect to the controller matrices for the discrete-time case

Using [1] and [2] we can derive that for the discrete time case the gradients of the squared $\mathscr{H}_2$ norm with respect to the closed loop system matrices are:

$$\begin{aligned}
\frac{\partial \operatorname{trace}\left(\mathscr{C} W_c \mathscr{C}^T\right)}{\partial \mathscr{A}} &= 2 W_o \mathscr{A} W_c, \\
\frac{\partial \operatorname{trace}\left(\mathscr{C} W_c \mathscr{C}^T\right)}{\partial \mathscr{B}} &= 2 W_o \mathscr{B}, \\
\frac{\partial \operatorname{trace}\left(\mathscr{B}^T W_o \mathscr{B}\right)}{\partial \mathscr{C}} &= 2 \mathscr{C} W_c.
\end{aligned}$$

Using matrix calculus as described in [3] we have

$$\begin{aligned}
\frac{\partial \operatorname{vect}(\mathscr{B})}{\partial \operatorname{vect}(B_c)} &= \frac{\partial \operatorname{vect}\left(\begin{pmatrix} 0 \\ I \end{pmatrix} B_c F\right)}{\partial \operatorname{vect}(B_c)} \\
&= F \otimes \begin{pmatrix} 0 & I \end{pmatrix}
\end{aligned}$$

and similarly

$$\frac{\partial \operatorname{vect}(\mathscr{A})}{\partial \operatorname{vect}(B_c)} = \begin{pmatrix} C & 0 \end{pmatrix} \otimes \begin{pmatrix} 0 & I \end{pmatrix}.$$

Using the Generalized Chain-Rule (see Thm 5.3 in [3]), we arrive at

$$\begin{aligned}
&\frac{\partial \operatorname{trace}\left(\mathscr{C} W_c \mathscr{C}^T\right)}{\partial B_c} \\
&= 2 \begin{pmatrix} 0 & I \end{pmatrix} W_o \mathscr{A} W_c \begin{pmatrix} C^T \\ 0 \end{pmatrix} + 2 \begin{pmatrix} 0 & I \end{pmatrix} W_o \mathscr{B} F^T.
\end{aligned}$$

The results for the matrices C_c and A_c in (5.6) can be derived along the same lines.

REFERENCES

[1] D. Petersson, A nonlinear optimization approach to H_2-optimal modeling and control. PhD thesis, Linköping University Electronic Press, 2013.

[2] M. Mercadal, H_2, fixed architecture, control design for large scale systems. PhD thesis, Massachusetts Institute of Technology, 1990.

[3] D. A. Turkington, Generalized vectorization, cross-products, and matrix calculus. Cambridge University Press, 2013.

E

CONVEX RELAXATION OF BILINEAR CONSTRAINTS

E.1. INTRODUCTION

The Chapters 2, 3 and 4 have in common that the problems they attempt to solve are particular cases of bilinearly (or just quadratically) constrained optimization problems.[1] All three chapters use the approach as given in [1], an article we submitted for the European Control Conference 2016. In this appendix we discuss the approach in a general context, partly based on [1]. The first three chapters are instances of this this approach, emphasizing its general applicability. We are not aware of any prior similar approaches, other than those outlined in [1], in the field of optimization, identification, control engineering or for any of the applications we developed for optimization in the optical context as in this thesis, or other applications we researched. For this reason we think the approach is novel, even though the method is simple to apply and very versatile, and a summary in the appendix of the different applications and different results throughout the thesis is a useful addition to this thesis.

E.2. EQUIVALENCE OF BILINEAR CONSTRAINTS TO RANK CONSTRAINTS ON MATRICES AFFINE IN THE VARIABLES.

The starting point is the following result on the generalized Schur complement.

Lemma E.2.1 (Carlson [2], generalized Schur complement)**.** Let the matrix X be defined as

$$X = \begin{pmatrix} X_1 & X_2 \\ X_3 & X_4 \end{pmatrix}. \tag{E.1}$$

then

$$\operatorname{rank}(X) = \operatorname{rank}(X_4) + \operatorname{rank}\left(X_1 - X_2 X_4^{+} X_3\right)$$

[1] Although the fourth chapter is also based on a bilinearly constrained optimization problem, the proposed solution in that chapter is different in nature to those for the other chapters.

if and only if

$$X_2\left(I - X_4^+ X_4\right) = 0, \tag{E.2}$$
$$\left(I - X_4 X_4^+\right) X_3 = 0. \tag{E.3}$$

The normal Schur complement requires X_4 to be square and invertible. This result on the generalized Schur complement also applies to non-square matrices. The lemma is essential in proving the equivalence between a bilinear constraint and a rank constraint on a matrix that is affine in the variables.

Theorem E.2.1. Given any matrices $X \in \mathbb{R}^{n_a \times n_b}$, $Y \in \mathbb{R}^{n_c \times n_d}$ and any full rank square matrices $W_1 \in \mathbb{R}^{n_a \times n_a}$, $W_2 \in \mathbb{R}^{n_d \times n_d}$, define the matrix M:

$$M := \begin{pmatrix} W_1 & 0 \\ 0 & I \end{pmatrix} \times \begin{pmatrix} C + XPY + APY + XPB & (A+X)P \\ P(B+Y) & P \end{pmatrix} \times \begin{pmatrix} W_2 & 0 \\ 0 & I \end{pmatrix}.$$

The following two constraints are equivalent:

$$C = APB \iff \operatorname{rank}(M) = \operatorname{rank}(P). \tag{E.4}$$

Proof. To start, notice that constraint $C = APB$ equals a rank constraint on the difference between C and the product APB, i.e.

$$C = APB \iff \operatorname{rank}(C - APB) = 0.$$

Enforcing constraint $\operatorname{rank}(C - APB) = 0$ is difficult for two reasons: it is a rank constraint, and the decision variables do not appear affinely in the constraint. However, using Lemma E.2.1, we can rewrite this constraint.

What we know of matrix M is that the conditions of Lemma E.2.1, (E.2) and (E.3), are fulfilled, since

$$W_1(A+X)P(I - P^+P) = 0, \text{ and}$$
$$(I - PP^+)P(B+Y)W_2 = 0.$$

The generalized Schur complement of P in M is:

$$\begin{aligned} &W_1\left(C + XPY + APY + XPB\right)W_2 \\ &- W_1\left((A+X)P\right)\left(P^+\right)\left(P\left(B+Y\right)\right)W_2 \\ &= W_1\left(C - APB\right)W_2, \end{aligned}$$

so applying Lemma E.2.1 gives us

$$\operatorname{rank}(M) = \operatorname{rank}(P) + \operatorname{rank}\left(W_1\left(C - APB\right)W_2\right).$$

Since W_1, W_2 are square and full rank we have the equivalence

$$\operatorname{rank}(M) = \operatorname{rank}(P) \iff \operatorname{rank}(C - APB) = 0 \iff APB = C.$$

□

As we noted in the proof, enforcing $\text{rank}(C - APB) = 0$ is difficult for two reasons: it is a rank constraint, and $C - APB$ is not affine in A and B. However, the matrix M is affine in all three decision variables. More often in semidefinite programming, we see the normal Schur complement used to render a matrix inequality affine. In this case, its use is slightly different, but not less useful when it comes to rendering expressions affine.

E.3. A CONVEX HEURISTIC FOR SOLVING BILINEAR PROBLEMS

In the previous section the equivalence was shown between a rank constraint on the matrix M and the bilinear constraint. We can use this equivalence to formulate a convex optimization problem as a heuristic for the problem where the bilinear term is causing the non-convexity. We do this by following the following steps, assuming that we have decision variables A and B that appear in the product APB, where P is some non-zero matrix, but not a decision variable.

1. Replace the bilinear term APB with a variable C and add the constraint $C = APB$;
2. Replace the constraint $C = APB$ with the constraint $\text{rank}(M) = \text{rank}(P)$;
3. Drop the rank constraint on the matrix M, and add the convex, low-rank inducing term $\lambda \|M\|_*$ to the objective function, where $\|\cdot\|_*$ denotes the nuclear norm and $\lambda > 0$ is a tuning parameter.

At this point it is important to discuss the parameters W_1, W_2, X and Y. They can be chosen freely (with W_1, W_2 invertible), giving a whole range of convex optimization problems as heuristic approaches to the original problem. Even though the rank constraints are equivalent, changing them influences the convex problem numerically, and (likely) changes the solution that will be obtained. We will demonstrate this with an example.

Example E.3.1 (The constraint $z^2 = 1$)**.** The constraint $z^2 = 1$ is a quadratic constraint[2], and can therefore be reformulated into a rank constraint. Set $W_1 = W_2 = 1$ and $x = X = Y = 0$. We obtain the matrix

$$M(z, x) = M(z, 0) = \begin{pmatrix} 1 & z \\ z & 1 \end{pmatrix}. \tag{E.5}$$

The nuclear norm of this matrix can be found analytically, and attains its minimum on the interval $z \in (-1, 1)$. Typically a numerical solver will return only one optimal value for z, not the interval. In our experience this optimal value that is returned is $z^* = 0$. The use of a different choice of x will lead to different and single-valued optimal solutions, that can also be computed analytically. The optimal solution is

$$z^* = \begin{cases} (-1, 1) & x = 0, \\ -1 & 0 < x < \bar{x}, \\ 1 & -\bar{x} < x < 0, \\ \frac{-x^3 - 2x}{2x^2 + 2} & |x| > \bar{x}, \end{cases} \tag{E.6}$$

[2] This is equivalent to a constraint $s \in \{0, 1\}$ for $z = 2s - 1$.

where the value for $\bar{x}$ is

$$\bar{x} = \frac{1}{3}\left(2 + \sqrt[3]{17+3\sqrt{33}} - \frac{2}{\sqrt[3]{17+3\sqrt{33}}}\right)$$

and is approximately equal to 1.5439. This means that for choices of x close enough (in the intervals $[-\bar{x}, 0)$ and $(0, \bar{x}]$) to a feasible solution, we will obtain a feasible solution by minimizing the convex relaxation. ■

Example E.3.2 (Choice of λ)**.** The convex optimization problem can have an unbounded objective function. For example, if we have the optimization problem

$$\begin{aligned} \min_{z} \quad & z \\ \text{subject to} \quad & z^2 = 1, \end{aligned} \tag{E.7}$$

we obtain the convex problem

$$\min_{z} \quad z + \lambda \|M(z,x)\|_*. \tag{E.8}$$

For $|z| \gg 1$ and $|z| \gg |x|$, the nuclear norm term has the approximate value of $\|M(z,x)\|_* \approx 2\sqrt{1+x^2}\,|z|$. If $\lambda < \left(2\sqrt{1+x^2}\right)^{-1} \leq \frac{1}{2}$, the objective function is unbounded from below. If we consider a choice of x only in the range $-1 < x < 1$, it can be shown that the point where the optimal z changes between $+1$ and -1 lies at $x = -\frac{1}{2\lambda}$. If x is picked randomly in this range, then λ should be tuned as low as possible, i.e. $\lambda = \frac{1}{2}$. ■

Example E.3.3 ([3] Sparse controller design)**.** A slightly more involved problem than the ones above is the following. Consider the continuous time linear, time-invariant (LTI) system

$$\dot{x} = Ax + Bu \tag{E.9}$$

and suppose we are looking for a stabilizing, sparse controller $u = Kx$. This could be formulated [4] as

$$\begin{aligned} \min_{K,P} \quad & \sum_{i,j} \left|K_{[i,j]}\right| \\ \text{subject to} \quad & (A+BK)^T P + P(A+BK) \prec 0 \\ & P \succ 0, \end{aligned} \tag{E.10}$$

where the objective function induces sparsity in the controller and the constraints ensure stability. The bilinear term is the term $E := PBK$, so the reformulated problem is

$$\begin{aligned} \min_{K,P} \quad & \sum_{i,j} \left|K_{[i,j]}\right| \\ \text{subject to} \quad & A^T P + E^T + PA + E \prec 0 \\ & P \succ 0 \\ & \operatorname{rank}(M(E,P,B,K)) = \operatorname{rank}(B), \end{aligned} \tag{E.11}$$

and the resulting heuristic convex optimization problem is

$$\begin{aligned} \min_{K,P} \quad & \sum_{i,j} \left|K_{[i,j]}\right| + \lambda \left\|M(E,P,B,K)\right\|_* \\ \text{subject to} \quad & A^T P + E^T + PA + E \prec 0 \\ & P \succ 0. \end{aligned} \tag{E.12}$$

■

E.4. TWO ITERATIVE USES OF THE RELAXED PROBLEMS

The freedom to parameterize the matrix M also allows for the search for a different, better solution in case the solver returns an infeasible or unsatisfactory solution to the original problem. Our suggestion is to parameterize the new relaxation with the optimal variables from the current optimization problem. This leads to two iterative algorithm variants. Given the convex problem

$$\min_{x,A,B,C} \quad f(x,A,B,C) + \lambda \left\|M(C,A,B,P,X,Y)\right\|_*, \tag{E.13}$$

the first variant uses the updates

$$\begin{aligned} \{x_k, A_k, B_k, C_k\} \in \arg\min_{x,A,B,C} & f(x,A,B,C) + \\ & \lambda \left\|M(C,A,B,P,-A_{k-1},-B_{k-1})\right\|_*, \end{aligned} \tag{E.14}$$

and the second variant uses the updates

$$\begin{aligned} \{A'_k, B'_k\} \in \arg\min_{x,A,B,C} \quad & f(x,A,B,C) + \\ & \lambda \left\|M(C,A,B,P,-A_{k-1},-B_{k-1})\right\|_*, \\ \text{subject to} \quad & (A - A_{k-1})P = 0 \\ \{x_k, A_k, B_k, C_k\} \in \arg\min_{x,A,B,C} \quad & f(x,A,B,C) + \\ & \lambda \left\|M(C,A,B,P,-A'_k,-B'_k)\right\|_*, \\ \text{subject to} \quad & P(B - B'_k) = 0 \end{aligned} \tag{E.15}$$

The first variant is easier to implement. For the second variant we can provide a proof of convergence of the iteration to a fixed point [1]. This fixed point is not necessarily a feasible solution to the bilinearly constrained problem, nor is it necessarily a global optimum of this problem. We will discuss convergence in the next section.

There are a number of advantages to this iterative convex approach.

1. **Ease of implementation.** The resulting problems are convex optimization problems (SDPs) for which solvers exist (for example SeDuMi and SCS) and middleware (YALMIP, CVX or Convex.jl) that allows for easy implementation of the nuclear norm operator.

2. **No feasible starting point required** An alternating minimization approach (fix A, optimize for B, then fix B and optimize for A) typically needs a feasible solution to the bilinear problem. A feasible solution is not always easy to obtain, for example in structured controller design. For the approach we propose, there are no prior conditions on X and Y.

3. **Performance** As shown in [1, 3, 5] the method shows very competitive performance for a range of problems.

4. **Efficiency** The introduction of an extra variable C is relatively efficient, in terms of the number of variables, compared to other approaches that introduce a matrix to replace the product $\text{vect}(A)\,\text{vect}(B)^T$ [1] (called 'lifting').

5. A **and** B **available** The fact that A and B are available in the resulting optimization problem –and are not substituted– enables constraints and objective functions to easily influence their structure and final value.

There are also some downsides.

1. **Convergence** We cannot guarantee convergence (in general) to a feasible solution, nor to a globally optimal solution. In fact, we have not found a general proof of convergence to even a fixed point when using the update rule in (E.14).

2. **Computational complexity** The resulting convex relaxation is an SDP problem, which have relatively high computational complexity of $O(n^6)$ [6].

3. **Numerical accuracy** The numerical solution to the optimization problem typically produces a matrix M for which the $\text{rank}(P)+1$'th singular value has a very small, but non-zero value (for example 10^{-8}). When validating the solution on the original problem, this might result in small but critical violations of constraints.

4. **Tuning** There are five variables that can be tuned: λ, W_1, W_2, X and Y. These variables influence both the success rate and convergence speed in a non-transparent manner.

E.5. A NOTE ON CONVERGENCE OF THE ITERATIVE SOLUTION

There are some results on convergence of the Sequential Convex Relaxation (SCR) algorithms. For the scheme in (E.15), consider the sequence of values

$$g_k = f(x_k, A_k, B_k, C_k) + \|C_k - A_k P B_k\|_*\,,\ k = 1, \ldots, \infty \tag{E.16}$$

This series can be shown to be non-increasing (see [1]). Colloquially, it can be interpreted as either the objective function improves, the constraint satisfaction improves, or both. However, the updates in (E.15) make no sense for a quadratic form, since the variables will be essentially fixed.

The proof is not valid for the update scheme in (E.14). However, in our experience the performance of (E.14) is better, which is why it is the first choice throughout this thesis.

For the case of Example E.3.1, it is easy to show using (E.6) that the updates in (E.14) ensure global convergence of z_k^* to a feasible solution for any starting point x_0. Furthermore, the number of iterations to convergence scales with $\log|x_0|$, since

$$|z_{k+1}^*| < \frac{|z_k^*|}{\bar{x}} \tag{E.17}$$

as long as $|z_k^*| > \bar{x}$,[3] and then it terminates with one more iteration.

The local convergence proof for COPR (Chapter 2) is slightly more elaborate and shows local convergence to a fixed point, under some strong assumption on the underlying problem.

E.6. MULTIPLE BILINEAR CONSTRAINTS AND THE USE OF ADMM

A commonality for the problems in this thesis is that there is not just one bilinear constraint, but that there are many. That is, they are instances of the feasibility problem

$$\begin{aligned} &\text{find} && \theta \in \mathbb{R}^n \\ &\text{subject to} && C_i(\theta) = A_i(\theta) P_i B_i(\theta) \\ &\text{for} && i = 1, \dots, N \end{aligned} \tag{E.18}$$

All the matrices $A_i(\theta)$, $B_i(\theta)$, $C_i(\theta)$ are affinely parameterized in θ, i.e. they can be written as

$$\text{vect}\,(A_i(\theta)) = p_{A_i} + \mathcal{T}_{A_i}\theta \tag{E.19}$$

for a vector p_{A_i}, a matrix $\mathcal{T}_{A_i}$, and similarly for $B_i(\theta)$ and $C_i(\theta)$.

The matrix

$$M_i(\theta, \phi) = M\big(A_i(\theta), B_i(\theta), C_i(\theta), -A_i(\phi), -B_i(\phi)\big) \tag{E.20}$$

is also affine in θ, but not in ϕ. This matrix can be vectorized to obtain

$$\text{vect}\big(M_i(\theta, \phi)\big) = p_{M_i}(\phi) + \mathcal{T}_{M_i}(\phi)\theta. \tag{E.21}$$

Here we list some examples of problems with many bilinear constraints.

Example E.6.1 (Phase retrieval)**.** [5] and Chapter 2. The phase retrieval problem can be described as

$$\begin{aligned} &\text{find} && \mathbf{a} \\ &\text{subject to} && \mathbf{y}_i = |U_i\mathbf{a}|^2 = \mathbf{a}^H U_i^H U_i \mathbf{a} \\ &\text{for} && i = 1, \dots, n_y \end{aligned} \tag{E.22}$$

These constraints are quadratic, a subset of bilinear constrained problems, and there are as many constraint as there are measurements. ■

[3] The gradient of $\frac{\partial z_{k+1}^*}{\partial z_k^*}$ is approximately 0.5 for $\left|z_k^*\right| > \bar{x}$, whereas $\frac{1}{\bar{x}} \approx 0.65$.

Example E.6.2 (Blind deconvolution, Chapter 4)**.** The blind deconvolution problem for the case of coherent illumination is the feasibility problem

$$\begin{aligned} &\text{find} && \mathbf{h},\ \mathbf{g_o},\ \mathbf{g_i} \\ &\text{subject to} && \mathbf{y} = \left|\mathbf{g_i}\right|^2 \Rightarrow \mathrm{d}\left(\mathrm{vect}\left(\mathbf{y}\right)\right) = \mathrm{d}\left(\mathbf{g_i}\right)^H \mathrm{d}\left(\mathbf{g_i}\right) \\ &&& \mathbf{g_i} = \mathbf{h} \star \mathbf{g_o} \Rightarrow \mathrm{vect}\left(\mathbf{g_i}\right) = \left(\mathrm{vect}\left(\mathbf{h}\right)^T \otimes I\right) V \,\mathrm{vect}\left(\mathbf{g_o}\right). \end{aligned} \tag{E.23}$$

These are two bilinear constraints, but a deeper analysis of the matrix V shows this can be further split up into two constraints for every measured pixel, just as in the previous example.

The case of coherent illumination is similar. Let $\mathbf{f} = \left|\mathbf{g_o}\right|^2$ be the intensity of the object distribution and $\mathbf{s}$ the intensity impulse response function. The feasibility problem is

$$\begin{aligned} &\text{find} && \mathbf{h},\ \mathbf{f},\ \mathbf{s} \\ &\text{subject to} && \mathbf{s} = |\mathbf{h}|^2 \Rightarrow \mathrm{d}\left(\mathrm{vect}\left(\mathbf{s}\right)\right) = \mathrm{d}\left(\mathrm{vect}\left(\mathbf{h}\right)\right)^H \mathrm{d}\left(\mathrm{vect}\left(\mathbf{h}\right)\right) \\ &&& \mathbf{y} = \mathbf{s} \star \mathbf{f} \Rightarrow \mathrm{vect}\left(\mathbf{y}\right) = \left(\mathrm{vect}\left(\mathbf{s}\right)^T \otimes I\right) V \,\mathrm{vect}\left(\mathbf{f}\right). \end{aligned} \tag{E.24}$$

E

■

Example E.6.3 (Parameter identification from input and output power spectra [7])**.** Let θ be the unknown parameter vector, $M(\theta)$ a mass matrix, $V(\theta)$ a damping matrix and $K(\theta)$ a stiffness matrix with the spring constants, all affine in θ, u an actuator force and

$$\begin{aligned} M(\theta)\ddot{x} + V(\theta)\dot{x} + K(\theta)x &= Bu \\ y &= Cx \end{aligned} \tag{E.25}$$

the dynamic equations for a system with masses, springs, and dampers and measurements $y = Cx$, where x is the system state. Let the Fourier transforms of $u(t), x(t), y(t)$ be denoted as $U(j\omega), X(j\omega)$ and $Y(j\omega)$ respectively. Denote the transfer function from u to x as $\Gamma(j\omega)$ and from u to y as $H(j\omega)$. For a set of frequency points ω_i, $i = 1, \ldots, N$, we have the N bilinear constraints

$$\begin{pmatrix} M(\theta) & V(\theta) & K(\theta) \\ 0 & 0 & C \end{pmatrix} \begin{pmatrix} -\omega_i^2 I \\ j\omega_i I \\ I \end{pmatrix} \Gamma(\omega_i) = \begin{pmatrix} B \\ H(\omega_i) \end{pmatrix}, \ i = 1, \ldots, N. \tag{E.26}$$

Let the power spectrum of the input be denoted as $\Phi_{uu}(j\omega)$ and that of the output as $\Phi_{yy}(j\omega) = H(j\omega)\Phi_{uu}(j\omega)H(j\omega)^H$, where ^{H} denotes the Hermitian transpose. If both power spectra are known or measured for frequency points ω_i, we have $2N$ bilinear or quadratic constraints in the variables $\theta, H(j\omega_i), \Gamma(j\omega_i)$. ■

Example E.6.4 (Closed loop parameter identification from reference input and output power spectra)**.** Consider the case of example E.6.3, but assume the system is controlled by a known controller with transfer function $\mathscr{K}(j\omega)$. Let the input to the system u be the summation of a reference signal r and the output of the controller, see Figure E.1. The reference signal has a known power spectrum denoted by $\Phi_{rr}(j\omega)$.

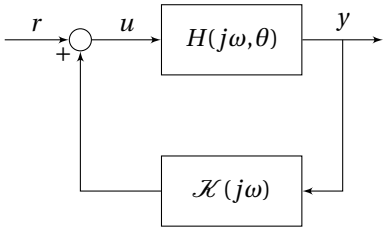


Figure E.1: The identification problem is to estimate θ based on the power spectral densities $\Phi_{rr}(j\omega_i)$ and $\Phi_{yy}(j\omega_i)$, and known controller dynamics $\mathcal{K}(j\omega_i)$ at a set of frequency points ω_i, $i = 1,\dots,N$.

For a set of frequency points ω_i, $i = 1,\dots,N$ we have the constraints

$$\begin{pmatrix} I + \mathcal{K}(j\omega_i)H(j\omega_i) & H(j\omega_i) \end{pmatrix} \begin{pmatrix} \Phi_{yy}(j\omega_i) & 0 \\ 0 & -\Phi_{rr}(j\omega_i) \end{pmatrix} \begin{pmatrix} \big(I + \mathcal{K}(j\omega_i)H(j\omega_i)\big)^H \\ H(j\omega_i)^H \end{pmatrix} = 0. \tag{E.27}$$

Together with (E.26) these give $2N$ bilinear or quadratic constraints. ■

The bilinearly constrained problems above have the convex relaxation

$$\min_{\theta} \quad \sum_{i=1}^{N} \lambda_i \left\| M\big(\big(A_i(\theta), B_i(\theta), C_i(\theta), -A_i(\phi), -B_i(\phi)\big)\right\|_*. \tag{E.28}$$

This optimization problem can be reformulated by introducing the additional variables $\mathbf{X}_i$,

$$\begin{aligned} \min_{\theta,\mathbf{X}} \quad & \sum_{i=1}^{N} \lambda_i \|\mathbf{X}_i\|_* \\ \text{subject to} \quad & \mathbf{X}_i = M_i(\theta,\phi) := M\big(A_i(\theta), B_i(\theta), C_i(\theta), -A_i(\phi), -B_i(\phi)\big) \\ \text{for} \quad & i = 1,\dots,N \end{aligned} \tag{E.29}$$

Following [8], we construct an Alternating Direction Method of Multipliers (ADMM) algorithm [9] to solve this minimization problem. Introducing the dual variables $\mathbf{Y}_i$ and the ADMM parameter ρ we obtain the updates

$$\begin{aligned} \theta^{k+1} &\in \arg\min_{\theta} \frac{\rho}{2} \sum_{i=1}^{N} \left\| \mathbf{X}_i^k - M_i(\theta,\phi) + \frac{1}{\rho}\mathbf{Y}_i^k \right\|_F^2 \\ \mathbf{X}_i^{k+1} &\in \arg\min_{\mathbf{X}_i} \|\mathbf{X}_i\|_* + \frac{\rho}{2\lambda_i} \left\| \mathbf{X}_i - M_i(\theta^{k+1},\phi) + \frac{1}{\rho}\mathbf{Y}_i^k \right\|_F^2 \\ \mathbf{Y}_i^{k+1} &= \mathbf{Y}_i^k + \rho\left(\mathbf{X}_i^{k+1} - M_i(\theta^{k+1},\phi)\right) \end{aligned} \tag{E.30}$$

The update of θ^k is an Ordinary Least Squares (OLS) problem, since M_i is affine in θ. The update of each $\mathbf{X}_i^k$ is completely independent for $i = 1,\dots,N$ and can be computed using singular value soft thresholding. The extension of the use of the nuclear norm to other low-rank inducing norms, such as the truncated nuclear norm, that have similar ADMM implementations, easily follows. The update of each $\mathbf{Y}_i^k$ is also completely independent from the others.

The OLS problem in the update of θ^k can be written in standard form according to

$$\theta^{k+1} \in \arg\min_{\theta} \left\| \underbrace{\begin{pmatrix} \text{vect}\left(\mathbf{X}_1^k + \frac{1}{\rho}\mathbf{Y}_1^k\right) \\ \vdots \\ \text{vect}\left(\mathbf{X}_1^k + \frac{1}{\rho}\mathbf{Y}_N^k\right) \end{pmatrix}}_{b_{\text{ADMM}}^k} - \underbrace{\begin{pmatrix} p_{M_1}(\phi) \\ \vdots \\ p_{M_N}(\phi) \end{pmatrix}}_{b_{\text{SCR}}} - \underbrace{\begin{pmatrix} \mathcal{T}_{M_1} \\ \vdots \\ \mathcal{T}_{M_N} \end{pmatrix}}_{H} \theta \right\|_2^2 \tag{E.31}$$

or simply

$$\theta^{k+1} \in \arg\min_{\theta} \left\| b_{\text{ADMM}}^k - b_{\text{SCR}} - H\theta \right\|_2^2. \tag{E.32}$$

The fact that the matrix H does not change during the ADMM iterations, can be exploited by computing for example its pseudo-inverse H^+ in advance and using the update

$$\theta^{k+1} = H^+ \left(b_{\text{ADMM}}^k - b_{\text{SCR}} \right). \tag{E.33}$$

E

REFERENCES

[1] R. Doelman and M. Verhaegen, "Sequential convex relaxation for convex optimization with bilinear matrix equalities," in Control Conference (ECC), 2016 European, pp. 1946–1951, IEEE, 2016.

[2] D. Carlson, E. Haynsworth, and T. Markham, "A generalization of the Schur complement by means of the moore-penrose inverse," SIAM Journal on Applied Mathematics, vol. 26, no. 1, pp. 169–175, 1974.

[3] R. Doelman and M. Verhaegen, "Sequential convex relaxation for robust static output feedback structured control," IFAC-PapersOnLine, vol. 50, no. 1, pp. 15518–15523, 2017.

[4] C. Scherer and S. Weiland, "Linear matrix inequalities in control," Lecture Notes, Dutch Institute for Systems and Control, Delft, The Netherlands, vol. 3, p. 2, 2000.

[5] R. Doelman, H. T. Nguyen, and M. Verhaegen, "Solving large-scale general phase retrieval problems via a sequence of convex relaxations," arXiv preprint arXiv:1803.02652, 2018.

[6] L. Vandenberghe, V. R. Balakrishnan, R. Wallin, A. Hansson, and T. Roh, "Interior-point algorithms for semidefinite programming problems derived from the KYP lemma," Positive polynomials in control, pp. 579–579, 2005.

[7] W. Krijgsman, "Continuous-time system identification, a bilinear optimization approach," Master's thesis, Delft University of Technology, November 2018.

[8] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," SIAM Journal on Optimization, vol. 20, no. 4, pp. 1956–1982, 2010.

[9] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al., "Distributed optimization and statistical learning via the alternating direction method of multipliers," Foundations and Trends® in Machine learning, vol. 3, no. 1, pp. 1–122, 2011.

List of Acronyms

ADMM	Alternating Direction Method of Multipliers
AO	Adaptive Optics
BME	Bilinear Matrix Equality
CCD	Charge-Coupled Device
CDI	Coherent Diffraction Imaging
CMOS	Complementary Metal Oxide Semiconductor
COBBD	Convex Optimization-based blind deconvolution
COPR	Convex Optimization-based Phase Retrieval
CPRL	Compressive Sensing Phase Retrieval
DM	Deformable Mirror
E-ELT	European Extremely Large Telescope
EM	Expectation Maximization
ENZ	Extended Nijboer-Zernike
ePIE	extended Ptychographical Iterative Engine
ERC	European Research Council
FEM	Finite Element Method
FFT	Fast Fourier Transform
GPF	Generalized Pupil Function
GPU	Graphics Processing Unit
GRBF	Gaussian Radial Basis Function
GS	Gerchberg-Saxton
HIO	Hybrid Input-Output
LPV	Linear Paramater-Varying
LQG	Linear Quadratic Gaussian

LTI Linear Time-Invariant

MAP Maximum a posteriori

ML Maximum Likelihood

NP Non-deterministic Polynomial-time

NCP Non-Common Path

OLS Ordinary Least Squares

OSA The Optical Society

OSS Oversampling Smoothness

OTF Optical Transfer Function

PSF Point Spread Function

RMS Root Mean Square

SCOBI Sequential Convex Optimization-based Identification

SCR Sequential Convex Relaxation

SDP Semidefite Programming

SH Shack-Hartmann

SNLLS Separable non-linear least squares

SNR Signal-to-Noise Ratio

SVD Singular Value Decomposition

TMT Thirty Meter Telescope

VAF Variance Accounted For

VAR Vector Auto-Regressive

VLT Very Large Telescope

CURRICULUM VITÆ

Reinier DOELMAN

EDUCATION

2014–2019	**Doctor of Philosophy** Delft University of Technology *Thesis:* Rank-based optimization techniques for estimation problems in optics. *Promotor:* Prof. dr. ir. M. Verhaegen
2010–2014	**Master of Science in Systems and Control** Delft University of Technology *Thesis:* Optimal Information Architecture for Distributed-Parameter Estimation in Fluid Dynamics. *Specializations:* Model Predictive Control, Pattern Recognition, Machine Learning, (Non-)Convex optimization, Filtering.
2007–2012	**Bachelor of Laws (Rechtsgeleerdheid)** Universiteit Leiden
2009–2010	**ERASMUS exchange programme** Imperial College London Department of Electrical and Electronic Engineering
2007–2010	**Bachelor of Science in Electrical Engineering** Delft University of Technology Graduated Cum Laude
2001–2007	**VWO (Gymnasium)** Stedelijk Gymnasium Leiden

WORK EXPERIENCE

2014–2019	**Teaching assistant, student supervisor** Delft University of Technology Assistant at the course 'Signaalanalyse' (Signal Analysis)
2013–2014	**Engineering intern (graduation project)** TNO
2013 (3 mo.)	**Intern** Ampelmann Operations

LIST OF PUBLICATIONS

JOURNAL PAPERS

Reinier Doelman, Måns Klingspor, Anders Hansson, Johan Löfberg and Michel Verhaegen, Identification of the dynamics of time-varying phase aberrations from time histories of the point-spread function. Journal of the Optical Society of America A, 2019.

Reinier Doelman and Michel Verhaegen, Convex optimization-based blind deconvolution for images taken with coherent illumination. Journal of the Optical Society of America A, 2019.

Reinier Doelman, Nguyen H. Thao and Michel Verhaegen, Solving large-scale general phase retrieval problems via a sequence of convex relaxations. Journal of the Optical Society of America A 35(8), pp. 1410–1419. OSA (2018)

Reinier Doelman, Sander Dominicus, Renaud Bastaits and Michel Verhaegen, Systematically structured $\mathscr{H}_2$ optimal control for truss-supported segmented mirrors. IEEE Transactions on Control Systems Technology 99, pp. 1–8. IEEE 2018.

CONFERENCE PAPERS

Reinier Doelman and Michel Verhaegen, Sequential convex relaxation for robust static output feedback structured control. IFAC-PapersOnLine, 50(1), pp.15518-15523. Elsevier 2017.

Reinier Doelman and Michel Verhaegen, Sequential convex relaxation for convex optimization with bilinear matrix equalities. In European Control Conference (ECC), 2016, pp. 1946-1951. IEEE, 2016.