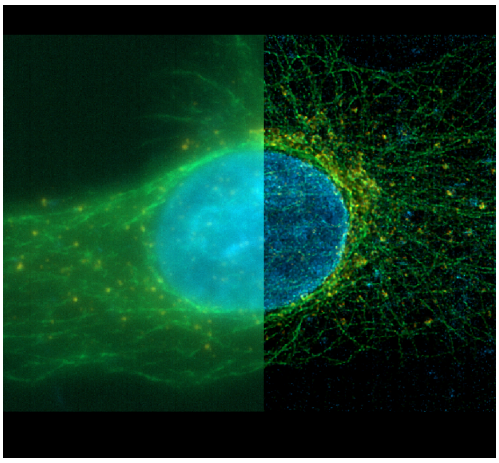
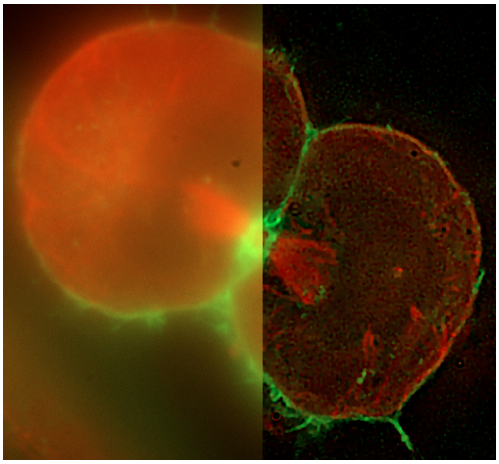
Utrecht University

Scientific Volume Imaging B.V.

Deconvolution of 3D Fluorescence Microscopy Images using Scaled Gradient Methods





Author:

Isabel Droste

Master's Thesis

January 2021

Supervisors:

Utrecht University

Project supervisor:
Dr. Tristan van Leeuwen

Second reader:
Dr. Paul Zegeling

Scientific Volume Imaging

Dr. Frans van der Have

Daniel Sevilla Sánchez, MSc

Dr. Hans van der Voort

Master's Thesis
Deconvolution of 3D Fluorescence Microscopy Images using Scaled Gradient Methods

Isabel Droste, BSc
Student ID: 5492335

Utrecht University
Graduate School of Natural Sciences
Mathematical Sciences

Project supervisor: Dr. Tristan van Leeuwen (UU)
Second reader: Dr. Paul Zegeling (UU)
Daily supervisors: Dr. Frans van der Have; Daniel Sevilla Sánchez, MSc; Dr. Hans van der Voort (Scientific Volume Imaging B.V.)

Front page images: Microscopy images before (left) and after (right) deconvolution with the scaled gradient algorithm.
Left image: Widefield image of a HeLa cell. Data courtesy: Dr. Yury Belyaev, European Molecular Biology Laboratory, Heidelberg, Germany.
Right image: Widefield image of a cell in telophase. Data courtesy: Dr. Ulrike Engel, Nikon Imaging Center, BioQuant Institute, Heidelberg, Germany

Time frame: February 2020 - January 2021
Utrecht, 29 January 2021

Abstract

Fluorescence microscopy images are blurred due to diffraction of light by passage through the optical path of the microscope. The resulting image is the convolution of the original object with a point spread function. We investigate the restoration of 3D fluorescence microscopy images that are affected by convolution and noise. We use a variational approach and investigate both which functional to minimize and how to find the minimizer. Regularization reflects a trade-off between bias and variance that is controlled by the regularization parameter. We apply different methods for finding the optimal value for this parameter and investigate how the optimal value depends on the signal-to-noise ratio of the image. We minimize the functional by using a scaled gradient projection algorithm that aims to improve the convergence rate compared to a standard gradient descent method by multiplication of the search direction by a scaling matrix and choosing an effective step size. The algorithm is applied to real data from confocal and widefield microscopy. The convergence of the algorithm is investigated and different scaling matrices are compared.

Acknowledgements

First of all, I thank Tristan van Leeuwen for being a great teacher and supervisor. Thank you for the interesting course on inverse problems, all your mathematical input into this project and our pleasant weekly meetings. You always came with useful ideas, good advice and were quick to respond to my questions and to give me feedback. I thank Paul Zegeling for being the second reader of this thesis.

I wrote this thesis as part of an internship at Scientific Volume Imaging (SVI). I thank SVI for giving me the opportunity of doing an internship at their company and providing me with a very interesting topic for my Master's thesis. I have learned a lot in the past year. In particular, I want to thank Frans van der Have for our interesting discussions and for always extensively answering my questions and giving me feedback. I thank Daniel Sevilla Sánchez for all his feedback and ideas, for sharing his expertise, and for being a great colleague in general. I thank Hans van der Voort for his guidance during this project and everything that he taught me about deconvolution algorithms. I thank the other developers Kiefer, Michel, Kevin and Steven for their interest in my thesis and for helping me with computer and programming problems. I thank the developers and the sales staff of SVI for making my internship into a pleasant time.

Finally, I would like to thank my family and friends for their love and support. In particular, I thank my mother Marian Droste and my brother Andreas Droste for being a safe and warm home for me that I can always fall back on. I thank Berend Ringeling and Kyra Brakkee for proofreading and for studying together in the mathematics library. Last but not least, I thank Lars van den Berg. Our discussions made me understand the topics more deeply and improved the contents of this thesis. But most importantly, your love and support were so important to me during this project. Thank you for always believing in me.

Contents

Preliminaries

Notation

- An overline, like $\overline{x}$ or $\overline{S}$ can mean two different things, which will become clear from the context. When x is a complex number, then $\overline{x}$ is the complex conjugate of x. When S is an area in $\mathbb{R}^3$, then $\overline{S}$ is used to denote a larger area in which S is contained. Note that with $\overline{S}$ we do not necessary mean the closure of S.
- We use the notation $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ to denote the inner product. When an inner product $\langle \cdot, \cdot \rangle$ is used without a subscript, the standard Hermitian inner product is meant

$$\langle \cdot, \cdot \rangle : \mathbb{C}^N \times \mathbb{C}^N \to \mathbb{C} : (\boldsymbol{x}, \boldsymbol{y}) \to \sum_{i=1}^{N} x_i \overline{y_i}$$

- When a norm $||\boldsymbol{x}||$ without a subscript is used, we mean the Euclidean norm $||\boldsymbol{x}||_2 = \sqrt{\langle x, x \rangle}$.
- The probability of an event A is denoted by $P(A)$. The conditional probability of an event A given the event B is denoted by $P(A|B)$.
- Let X be a random variable. The expected value of X is denoted by $\mathrm{E}(X)$ and the variance of X by $\mathrm{Var}(X)$.

Prerequisites

- Let $\mathcal{U}$ and $\mathcal{V}$ be Hilbert spaces equipped with inner products $\langle \cdot, \cdot \rangle_{\mathcal{U}}$ and $\langle \cdot, \cdot \rangle_{\mathcal{V}}$. Let $A : \mathcal{U} \to \mathcal{V}$ be a bounded, linear operator. The *adjoint operator* A^* is defined as the operator $A^* : \mathcal{V} \to \mathcal{U}$ that satisfies

$$\langle A\boldsymbol{u}, \boldsymbol{v} \rangle_{\mathcal{V}} = \langle \boldsymbol{u}, A^*\boldsymbol{v} \rangle_{\mathcal{U}}$$

  for all $\boldsymbol{u} \in \mathcal{U}$ and $\boldsymbol{v} \in \mathcal{V}$.
- When $\mathcal{U}$ and $\mathcal{V}$ are equal to $\mathbb{C}^N$ and $\mathbb{C}^M$, respectively, with the standard inner product, and $A \in \mathbb{C}^{M \times N}$, then the adjoint operator $A^* \in \mathbb{C}^{N \times M}$ is equal to the conjugate transpose of A, that is

$$(A^*)_{ij} = \overline{A_{ji}}.$$

- Let A be a complex square matrix. Then A is called a *unitary matrix* if its conjugate transpose is equal to its inverse, that is

$$A^*A = AA^* = I$$

- Let A be real square matrix. Then A is called a *positive definite matrix* if and only if

$$\langle \boldsymbol{x}, A\boldsymbol{x} \rangle > 0$$

  for all nonzero $\boldsymbol{x} \in \mathbb{R}^N$.

- An $N \times N$ matrix is called a *circulant matrix* when it is of the form

$$A = \begin{pmatrix} a_0 & a_{N-1} & a_{N-2} & \cdots & a_2 & a_1 \\ a_1 & a_0 & a_{N-1} & \cdots & a_3 & a_2 \\ a_2 & a_1 & a_0 & \cdots & a_4 & a_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{N-1} & a_{N-2} & a_{N-3} & \cdots & a_1 & a_0 \end{pmatrix},$$

  in other words, $A_{mn} = a_{(m-n)}$ for some N-dimensional vector $\boldsymbol{a}$ where $a_i := a_{(i \mod N)}$. Each row is a cyclic permutation of the row above it where each element is shifted one position to the right.

- For $r \in \mathbb{C}$, $r \neq 1$, the sum of the first N terms of the geometric series is given by

$$\sum_{k=0}^{N-1} r^k = \frac{1 - r^N}{1 - r}. \tag{0.1}$$

Chapter 1

Introduction

Microscopy plays an important role in many fields of scientific research such as medicine, environmental science, forensics and material science, because it makes it possible to see objects that are too small to be seen by the naked eye [13, 25]. Fluorescence microscopy is a subdomain of microscopy where the object of interest is labeled with a fluorescent dye. When illuminated with light of a specific wavelength, the excitation wavelength, the specimen will emit light of another, typically larger, emission wavelength. This light is then recorded by a camera or detector. Since the fluorophores can target specific biological molecules, fluorescence microscopy makes it possible to visualize cell structures of interest. In addition to this, the method is sensitive to small amounts of fluorophores and it allows the imaging of live cells or organisms. These advantages make that the technique is widely applied in microbiology [12, 21].

Due to diffraction of light by passage through the optical path of the microscope the image is blurred. This process can be described by a convolution of the object with a point spread function (PSF). Due to the linearity of convolution, the image measured by the microscope is the sum of point spread functions that are scaled by the intensity and translated to be centered at each individual emitting point in the object. In addition to convolution, the image is also distorted by noise. The main source of noise is Poisson noise that results from the discrete nature of photons.

Image restoration is the process of recovering the original object that is degraded by convolution and noise. Restoration is important because it improves the image quality and makes it possible to perform a better visual and quantitative analysis. Scientific Volume Imaging (SVI) develops software for restoration, visualization and analysis of fluorescence microscopy images. Figure 1.1 shows a multichannel widefield microscope image deconvolved with SVI's Huygens software.

An inverse problem is a problem that can be formulated as

$$K(\boldsymbol{w}) = \boldsymbol{f} \tag{1.1}$$

where $\boldsymbol{w} \in U$ are the parameters, $\boldsymbol{f} \in V$ is the measured data, where U, V are Hilbert spaces. The forward operator $K : U \to V$ is a model of some physical process. The goal is to find a solution $\boldsymbol{w}$ for a given measurement $\boldsymbol{f}$. However, since K is only a model for the underlying process and because of measurements errors, it may be that no solution exists that exactly fits the data. Also, it may be that a solution is not unique or that the solution does not depend continuously on the data. In that case, a small measurement error can lead to a large error in the reconstruction.

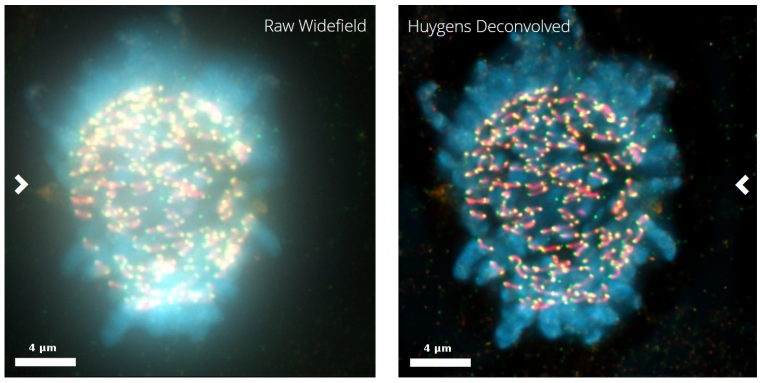


Figure 1.1: A 2D slice of a 3D widefield image of a cell division. Raw image (left) and image deconvolved by SVI's Huygens software (right). The image contains four channels which are shown in the colours of the emission wavelengths. Data courtesy: Dr. Livio Kleij en Martijn Vroomans, Medical Oncology, UMC Utrecht, The Netherlands.

In the inverse problem of deconvolution, the forward operator K models the image formation process in a microscope. Convolution in the spatial domain corresponds to multiplication in the Fourier domain. The Fourier transform of the PSF is called the optical transfer function (OTF). Deconvolution therefore corresponds to division by the OTF in the Fourier domain. However, the OTF usually has finite support and tends to zero for high frequencies. In addition to this, since the noise is described by high frequencies in the Fourier domain, the division by the OTF leads to amplification of noise. Therefore, the solution does not depend continuously on the data, making this problem ill-posed.

The inverse problem can be rewritten as a minimization problem where a function is minimized that measures how well the estimate for the object matches with the measured data. Gradient descent algorithms are well known iterative methods for finding the minimum of a function by taking steps in the direction of the negative gradient. These methods are robust and simple, but their convergence can be slow. The idea of scaled gradient methods is to multiply the descent direction by a scaling matrix to improve the convergence speed.

In this thesis we will explain the most important theory behind inverse problems and regularization, explain how the scaled gradient projection algorithm works and apply it to the deconvolution of three-dimensional fluorescence microscopy images. So far, the scaled gradient algorithm has primarily been tested on synthetic microscopy data that was generated using an ideal PSF and Poisson noise [7, 35, 36]. In reality, the image formation process is more complicated. In [36], the method is tested on real data of confocal and STED microscopy. However, it has not yet been applied to widefield microscopy images, which will be done in this thesis. Deconvolution of widefield images can be regarded as a more difficult problem due to the infinite extent of the point spread function and the boundary problems that arise from this.

The outline of this thesis is as follows. We will start by explaining in more detail the inverse problem we will be dealing with. This will be done in Chapter 2 where we explain more about fluorescence microscopy and define a forward operator K from Equation (1.1) that models image formation process. In Chapter 3 we will treat the theory of solving inverse problems that is relevant for our applications. We will define the pseudo-inverse and spectral regularization. We show how

the problem can be formulated as a minimization problem. In Chapter 4 we explain how the minimization problem can be solved using the scaled gradient projection algorithm. In Chapter 5 we will test our implementation of the scaled gradient algorithm to real microscopy data.

Chapter 2

The inverse problem

In this chapter we explain how the inverse problem of convolution. We start with the basics of fluorescence microscopy and explain how an image of an object is formed in a microscope. We show how the image formation process can be modeled with with a forward operator K such that the inverse problem can be formulated as $K(\boldsymbol{w}) = \boldsymbol{f}$. From the forward model, we derive the likelihood functions in the case of Poisson and Gaussian noise. We explain why the inverse problem of deconvolution is ill-posed and discuss the limitations of the model.

2.1 Fluorescence microscopy

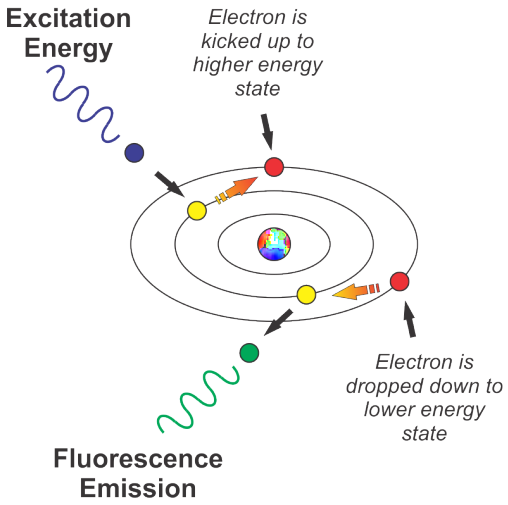


Figure 2.1: The energy levels of an electron. Source: `https://firedivegear.com/the-science-of-fluorescence/`

This section is based on [18, 19, 21, 25]. The principle of fluorescence is illustrated in Figures 2.1 and 2.2. The electrons of an atom can be in different energy levels. The lowest level is called the ground state and is denoted by S_0 as shown in the Jablonski diagram (2.2a). At each energy level, there are multiple vibrational energy levels. A fluorescent molecule that is in the ground state can be excited to a higher vibrational state (S_1^{vib}) by illumination with light. This absorption of photon energy happens most efficiently when the light has the optimal excitation wavelength (2.2b). After this, the electron almost immediately falls back to a lower energy level. First, the electron goes

from the higher vibrational state to the lowest ground level of the excited state S_1. During this, transition energy is released in the form of heat. After that, the electron falls back to the ground state. During this second transition, energy is released in the form of a photon. Because of the loss of energy to the vibrational energy relaxation, the emitted photon has a lower energy than the photon that was absorbed. Therefore, emitted light has a larger wavelength than the excitation light.

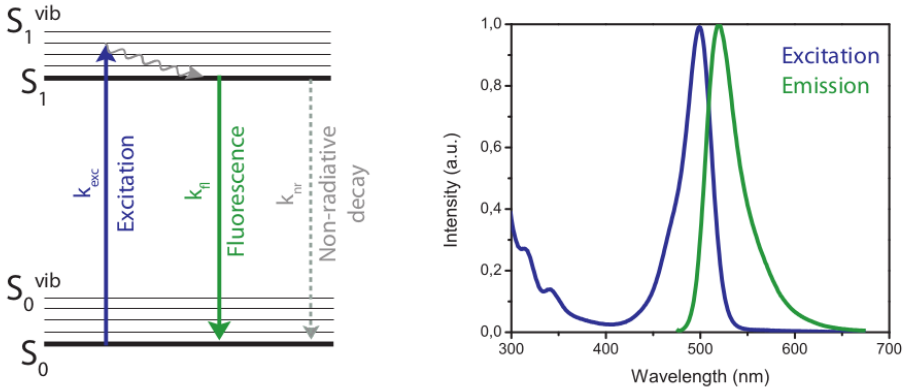


Figure 2.2: Jablonski diagram illustrating the energy state transitions (left). The excitation and emission profile of a commonly used fluorescent dye Alexa 488 (right). Image reproduced from [11].

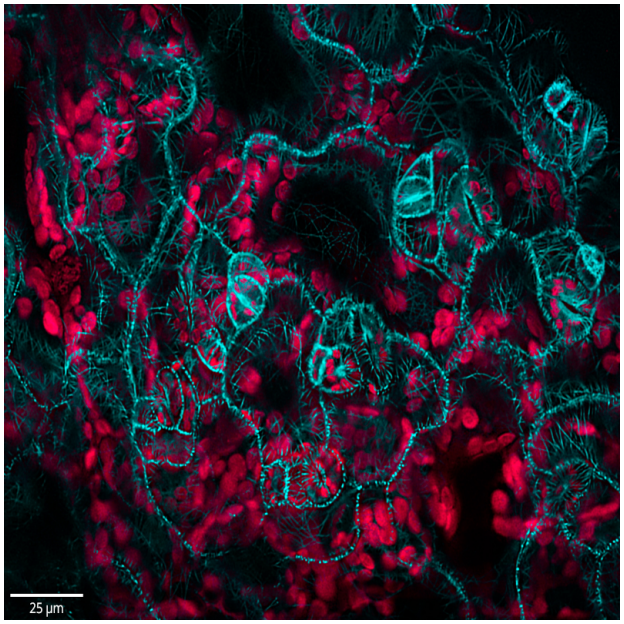


Figure 2.3: Image of a leaf of the plant Arabidopsis thaliana. The image has two channels that are shown in the color corresponding to the emission wavelength. Microtubules are shown in cyan, chloroplasts revealed by their autofluorescence in magenta-red. Data courtesy: Gregory Pozhvanov Ph.D, St. Petersburg State University, Russia.

This principle of fluorescence is used in fluorescence microscopy. Some materials, like chlorophyll and vitamins exhibit auto-fluorescence. Other materials can be made fluorescent by treating it with fluorescent markers. These markers can label specific tissues or structures in a microscopic sample because they bind to certain biochemical targets like DNA, proteins or cellular structures. By using different markers that each have a specific excitation and emission wavelength, different structures can be distinguished. A fluorescence microscope can register light of different emission wavelengths, resulting in an image that consists of multiple channels. Each channel represents the intensities recorded at a specific wavelength. Figure 2.3 shows an example of a microscopic image with two channels that are shown in the color corresponding to their emission wavelength.

The basic principle of a fluorescence microscope is depicted in Figure 2.4. The light from a light source is first filtered by an excitation filter that only lets through light within a specific band of wavelengths. Then, this filtered light reflects off a dichroic mirror. This mirror is designed such that all light above a certain wavelength gets reflected while light below that wavelength passes through. After that, light passes through the objective to the specimen where it excites the fluorophores. The object will then emit light of the emission wavelength. This light travels back through the objective. Because its wavelength is now larger, it passes through the dichroic mirror. After that, the light will be filtered by a emission filter that blocks any light that does not have the emission wavelength. Finally, the light reaches the detector.

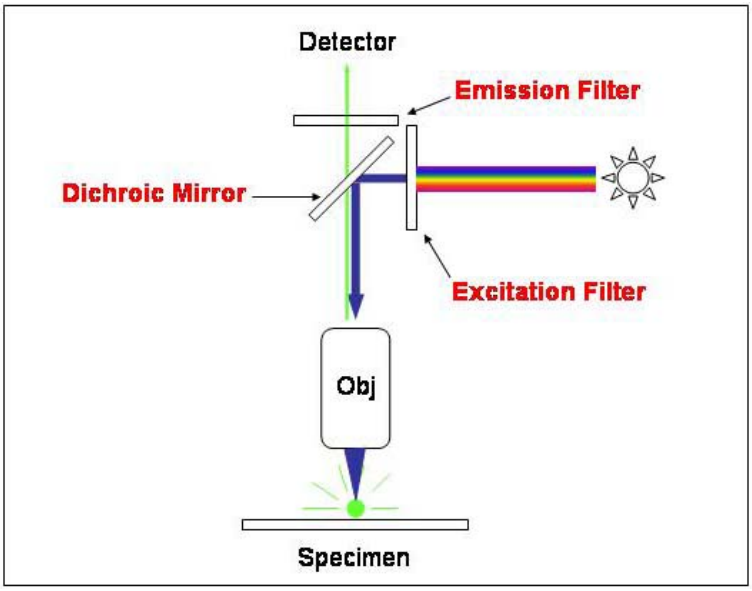


Figure 2.4: A schematic image of a fluorescence microscope. Source: `https://serc.carleton.edu/microbelife/research_methods/microscopy/fluromic.html`

The image that is measured by the detector is a blurred version of the object. To understand this, we look at the image of a point source in the object (Figure 2.5). This point source emits light in all directions that travels as a spherical wave front away from the source. The Huygens-Fresnel principle says that each point on the wave front is excited and then acts as a point source itself that radiates spherical waves. The objective of a microscope contains a number of lenses that focus the light emitted by the object. In Figure 2.5 this is shown in a simplified way with one convergent lens. The lens inverts the shape of the diverging spherical wavefront emitted by the point source into a

converging wave front. This converging wave front is a section of a full spherical wave front. The center of this sphere is the focus point where the light converges. The intensity that is measured is proportional to the squared amplitude of the wave, integrated over a full wavelength. At the focal point, the waves from all the secondary sources arrive in phase. Due to this constructive interference, the largest intensity is measured at this point. Since only part of the full spherical wave front is captured by the lens, not all waves cancel each other out outside of the focal point. At the plane at focal distance from the lens, alternating constructive and destructive interference lead to a pattern of concentric rings that decrease in intensity away from the center. This pattern is called an airy disk (Figure 2.6a). When measuring a 3D image, we need to know the three-dimensional equivalent of an airy disk. This gives rise to the point spread function (PSF), which is the 3D response of an imaging system to a point source (Figure 2.6b). It is common to use a three dimensional coordinate system where the x- and y-axes are parallel to the focal plane. Since the resolution in the z-direction is smaller than in the x- and y-direction, an object is usually recorded such that the interesting features are in the xy-plane. The plane at $z = 0$ is the plane closest to the objective. The larger the z-value, the deeper you travel into the sample.

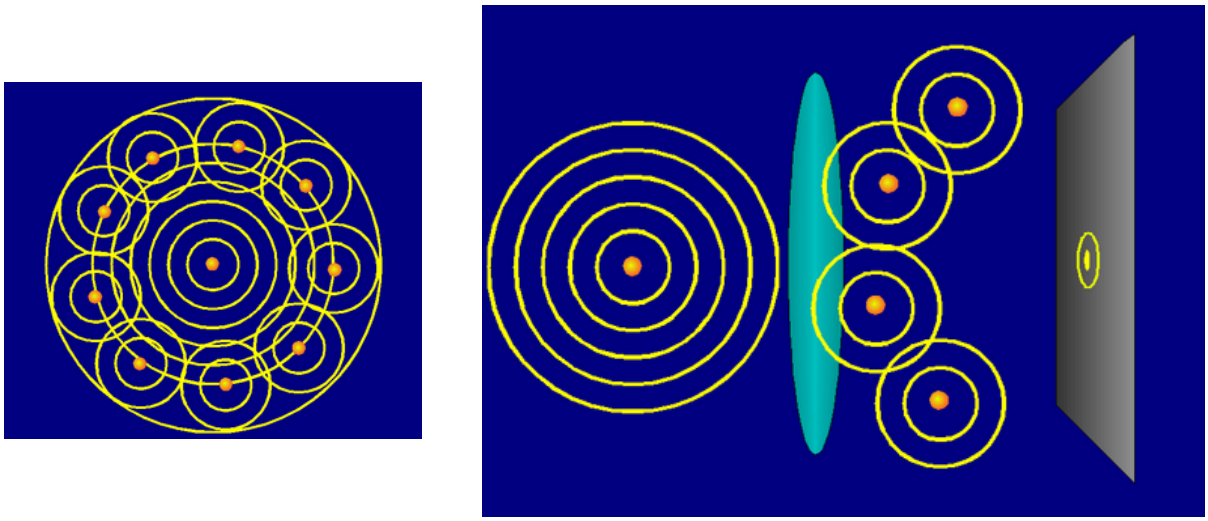
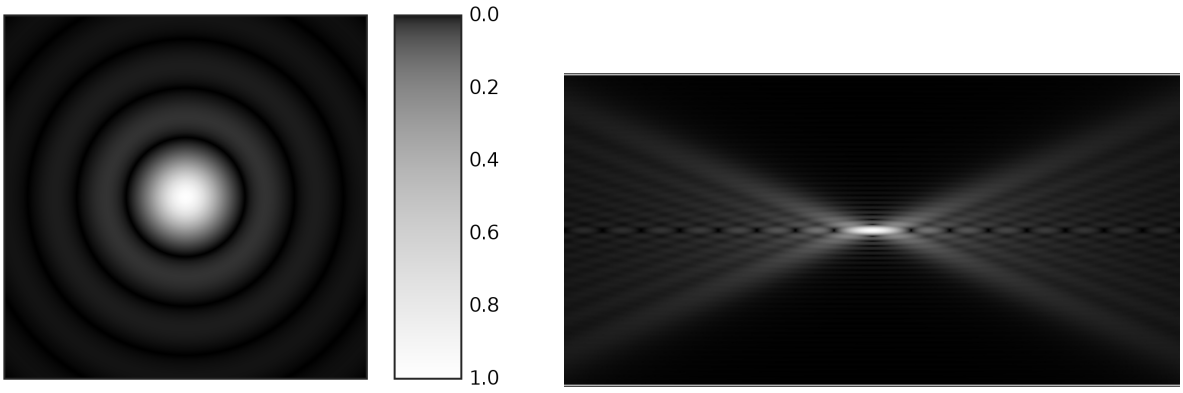


Figure 2.5: The Huygens-Fresnel principle (left). The spherical wave front emitted by the point source is diffracted by the lens leading to waves that converge at the focus point (right). Image reproduced from [18].

The spatial resolution of an optical device is defined as the minimal distance that two points can be apart such that they can still be distinguished as individual points. This depends on which fraction of the full spherical wave front is captured by the lens which is characterized by the numerical aperture (NA) of the lens

$$\mathrm{NA} = n \sin \theta,$$

where θ is maximum angle of a light ray that can still be collected by the objective (Figure 2.7), and n is the refractive index of the medium in which the lens in immersed. It is given by $n = \frac{c}{v}$ where c is the speed of light in vacuum and v is the speed of light in the medium. A larger numerical aperture will lead to a diffraction pattern with a smaller airy disk which improves the resolution of the image. The Rayleigh criterion is a standard to characterize the resolution. It says that two points can be distinguished when their distance is larger than the distance between the central maximum and the first minimum of the airy disk. This distance is called the radius of the airy disk



(a) Airy disk (b) PSF

Figure 2.6: On the left an airy disk is shown in the xy-plane. On the right, a yz-slice of a 3D point spread function is shown with the z-direction on the horizontal axis and the y-direction on the vertical axis.

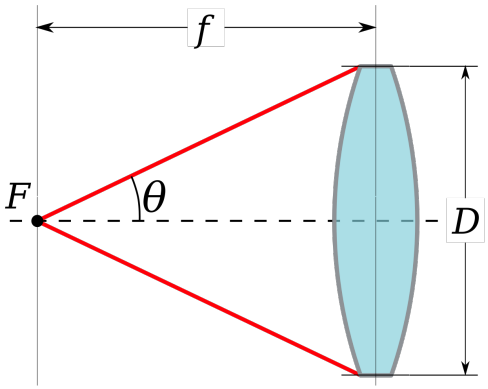


Figure 2.7: The numerical aperture is defined as $\mathrm{NA} = n \sin\theta$ where n is the refractive index.

and is denoted by r. The radius depends on the numerical aperture and the wavelength λ as follows

$$r = 0.61 \frac{\lambda}{\mathrm{NA}}.$$

Since the wavelength of visible light ranges from 380 to 750 nm and the numerical aperture is in the range of 1.0 to 1.5, the resolution of light microscopy images is in the range of a few hundred nanometers. This limitation is called the diffraction limit. However, in recent years, new techniques have been developed that can break this limit by deterministically or stochastically making nearby fluorophores emit light at seperate times. These systems, for example STED, PALM and STORM, are called super-resolution microscopy [6, 17, 30]. In this summary, we will not go into this further.

An optical system is linear when the image of a weighted sum of light sources is equal to the weighted sum of the images of the individual sources. An optical system is said to be translation invariant when a shift of the object in space does not change the form of the output and it only leads to the output to be shifted by the same amount. When an optical system is both linear and translation invariant, it can be characterized by its response to a unit impulse, the PSF. The resulting image of an object is then equal to the sum of all the translated and scaled impulse responses of each point

in the object. The resulting image is the convolution between the object and the PSF, which we will define in section 2.2.

Two important types of fluorescence microscopes are the widefield microscope and the confocal microscope. In a widefield microscope the whole object in illuminated at once. The measurement of the focal plane then contains blurring from planes that are above or below it. In a confocal microscope, the light moves through a pinhole before it reaches the detector. This pinhole is located such that light from the focus point is focused through the pinhole while most of the out-of-focus light is blocked, as illustrated in Figure 2.8. A confocal microscope has a much improved resolution compared to a widefield microscope. Because the only light coming from a single point can be measured at once by a confocal microscope, the object has to be scanned point by point.

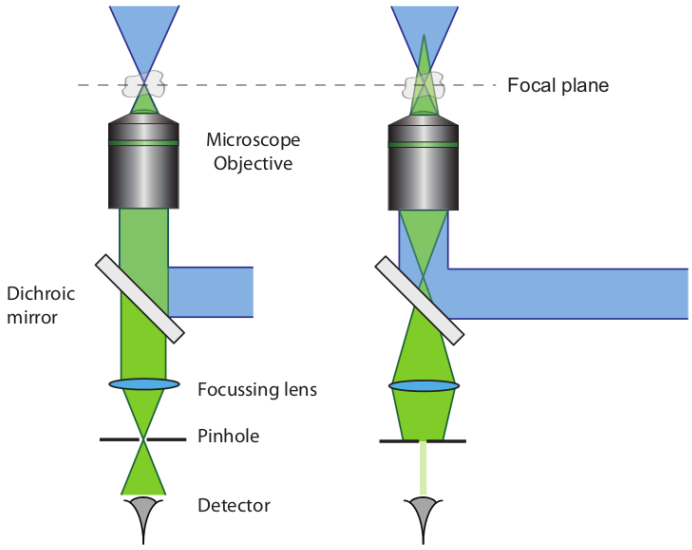
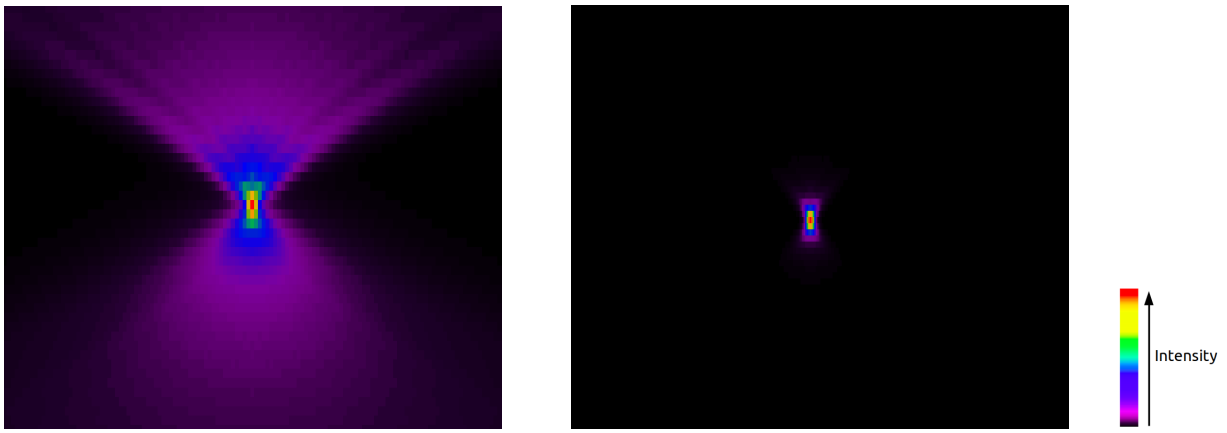


Figure 2.8: Schematic illustration of a confocal microscope. The blue excitation light is focussed at one point in the sample that emits the green emission light. Left: the emission light that comes from the focal point is focussed through the pinhole and reaches the detector. Right: the emission light from an out-of-focus point is mostly blocked by the pinhole. Image reproduced from [11].

Figure 2.9 shows the PSF of a widefield and a confocal microscope. Since a widefield microscope collects light from all planes above and below the focal plane, the PSF extends infinitely. It has the shape of a cone where the intensities are spread out over a larger area away from the center and where each axial plane has the same sum. A confocal PSF is much more compact. It has the shape of an ellipsoid with the longest axis in the axial direction.

In addition to the deterministic blurring by the PSF, the image is also distorted by noise. The most important source of noise comes from the natural randomness of the photon flux. Since photons are discrete and arrive independently of each other, the stream of photons can be modeled with a Poisson process. Since the variance of a Poisson random variable is equal to the expected value, the signal-to-noise ratio (SNR) of an image is equal to the square root of the number of photons. Therefore, the noisiness increases when the number of photons decreases. In a confocal image, part of the light is blocked by the pinhole. In addition to this, scanning of the object takes time, further



(a) Widefield PSF (b) Confocal PSF

Figure 2.9: Sum projections of the PSF onto the axial plane. The z-axis is in the vertical direction.

limiting the number of photons that are measured at each point. Therefore, in general, confocal images are noisier than widefield images.

2.2 The forward model

In the previous section we saw how an image of an object is formed in a microscope. In this section, we will capture the image formation process in the forward model $K(\boldsymbol{w}) = \boldsymbol{f}$.

A 3D image is built up of voxels, the three-dimensional equivalent of pixels. A voxel is a cuboid with dimensions $\Delta x \times \Delta y \times \Delta z$, the sample sizes. An image can contain different channels that correspond to the intensities recorded at specific wavelengths. Each voxel has an intensity value for each channel. We make the assumption that convolution takes place for each channel independently. Therefore, the different channels can be deconvolved separately and it suffices to only consider single-channel images. Let N_x, N_y and N_z be the number of voxels in each direction, which we will call the image dimensions. In typical microscopic images, N_x and N_y are around 500 and N_z is around 50, while much larger images also exist. The z-direction is perpendicular to the focal plane.

An image can be represented as a third-order tensor. Let $W \in \mathbb{R}^{N_x \times N_y \times N_z}$ be the object and $H \in \mathbb{R}^{M_x \times M_y \times M_z}$ the PSF. The PSF is extended periodically in the following way

$$P_{i,j,k} = P_{i \bmod N_x,\, j \bmod N_y,\, k \bmod N_z}$$

Then the discrete convolution of the object W with the point spread function H is an image $F \in \mathbb{R}^{N_x \times N_y \times N_z}$ that is given by

$$F_{i,j,k} = (H * W)_{i,j,k} := \sum_{l=1}^{N_x} \sum_{m=1}^{N_y} \sum_{n=1}^{N_y} H_{i-l,\, j-m,\, k-n} \cdot W_{l,m,n}.$$

We stack the voxels of an image on top of each other to represent an image as a vector in $\mathbb{R}^N$. Let $\boldsymbol{w} \in \mathbb{R}^N$ be the ground truth image and $\boldsymbol{h} \in \mathbb{R}^M$ the PSF that result from stacking the voxels of

W and H, where $N = N_x \cdot N_y \cdot N_z$ and $M = M_x \cdot M_y \cdot M_z$.

The first step in the forward model is the discrete convolution of $\boldsymbol{w}$ with $\boldsymbol{h}$. The convolution can be represented by multiplication with a matrix K that we will call the convolution matrix. Convolution of a one-dimensional signal is illustrated in Figure 2.10 and it can be represented by multiplication with a circulant matrix. In the case of a three-dimensional image, the structure of K becomes more complicated as we will see in Chapter 3.

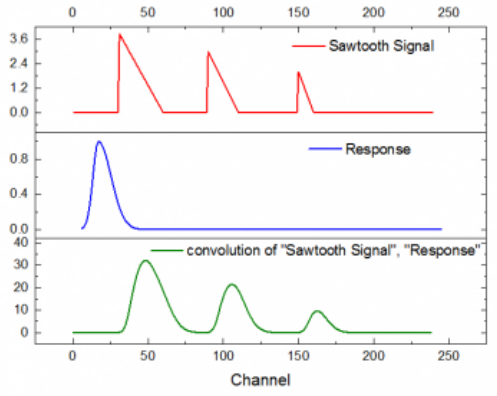


Figure 2.10: The convolution of a one-dimensional signal (red) with the PSF (blue).

The second step is the addition of a background intensity $\boldsymbol{b}$. This background intensity is all intensity that does not come from the object we are interested in. We make the assumption that the background has the same value $b \in \mathbb{R}_{\geq 0}$ everywhere in the image. The convolved image with background is then given by

$$K\boldsymbol{w} + \boldsymbol{b}.$$

The third step is the addition of noise. We model the number of photons that are measured in voxel i in a time interval with a random variable f_i that is Poisson distributed with parameter $\lambda_i > 0$. The value of λ_i is equal to the expected number of photons in that voxel after convolution and addition of background, that is $\lambda_i = (K\boldsymbol{w} + \boldsymbol{b})_i$. The probability of measuring an intensity of k is then

$$P(f_i = k) = \frac{\lambda_i^k \exp(-\lambda_i)}{k!}.$$

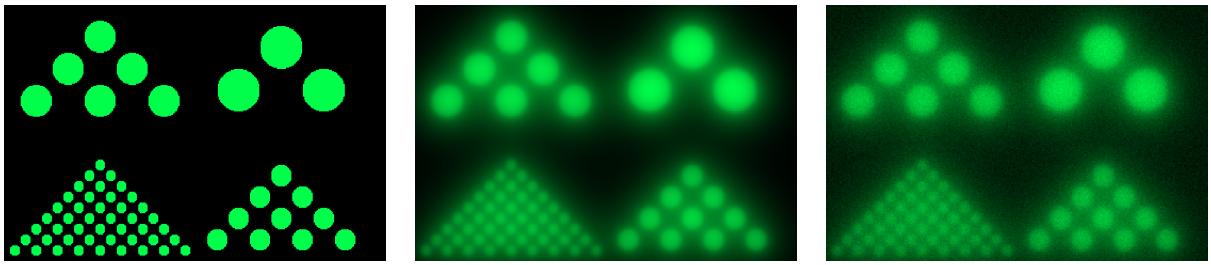
The Poisson distribution has the property that

$$\lambda_i = \mathrm{E}(f_i) = \mathrm{Var}(f_i)$$

To summarize, the total image formation process can be modeled as

$$\boldsymbol{f} = K\boldsymbol{w} + \boldsymbol{b} + \boldsymbol{\delta}$$

where $\boldsymbol{f}$ is the measured image and $\boldsymbol{\delta}$ is the noise. The data f_i is a realization of a Poisson random variable with parameter $\lambda_i = (K\boldsymbol{w} + \boldsymbol{b})_i$. Note that from now on we will use K for the convolution matrix and not for the for the total forward operator. Image 2.11 shows the image formation process of a synthetic image with a widefield PSF.



(a) Ground truth (b) Convolved image (c) Background and noise added

Figure 2.11: Image formation process of a synthetic image. The images are xy-slices.

2.3 Limitations of the model

The relatively simple model that we will use, with a single PSF and Poisson noise, captures the most important aspects of the real processes that take place. However, this model has some limitations, which we will now discuss.

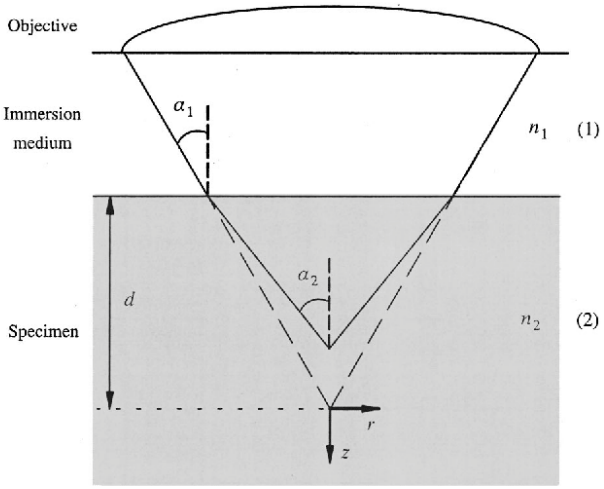


Figure 2.12: A mismatch between the lens immersion medium and the specimen embedding medium leads to refraction of the rays at the boundary of the media. This causes rays that were focused to become out-of-focus which leads to an asymmetric PSF that becomes more elongated with depth. Figure reproduced from [8].

Ideally, the objective converts the wave front into a spherical wave front with the focal point in the center of that sphere. However, when the lens immersion medium does not match the specimen embedding medium, not all light rays have the same focal point. This effect is called spherical aberration. When rays pass from one medium to another, they are bend according to Snell's law

$$\frac{\sin\theta_2}{\sin\theta_1} = \frac{n_1}{n_2},$$

where θ_1 is the angle of incidence, θ_2 is the angle of refraction, n_1 is the refractive index of the lens immersion medium and n_2 is the refractive index of the specimen embedding medium (Figure 2.12). This refraction causes rays with different incident angles to be focused at a different point,

spoiling the focus of the lens. The asymmetry of the PSF increases with the focal depth into the specimen, making the PSF dependent on the z-coordinate. Figure 2.13 shows how the PSF changes with depth when the refractive index of the lens embedding medium is larger than that of the specimen. In that case, the PSF has a tail at the side farthest away from the lens. When it is the other way around, the tail appears at the side closest to the lens. In this case, the angle of the rays increases when moving from the specimen to the lens. Part of the light is then internally reflected back into the specimen, decreasing the effective numerical aperture of the objective. Restoration of images that are affected by spherical aberration requires an algorithm with a depth dependent PSF.

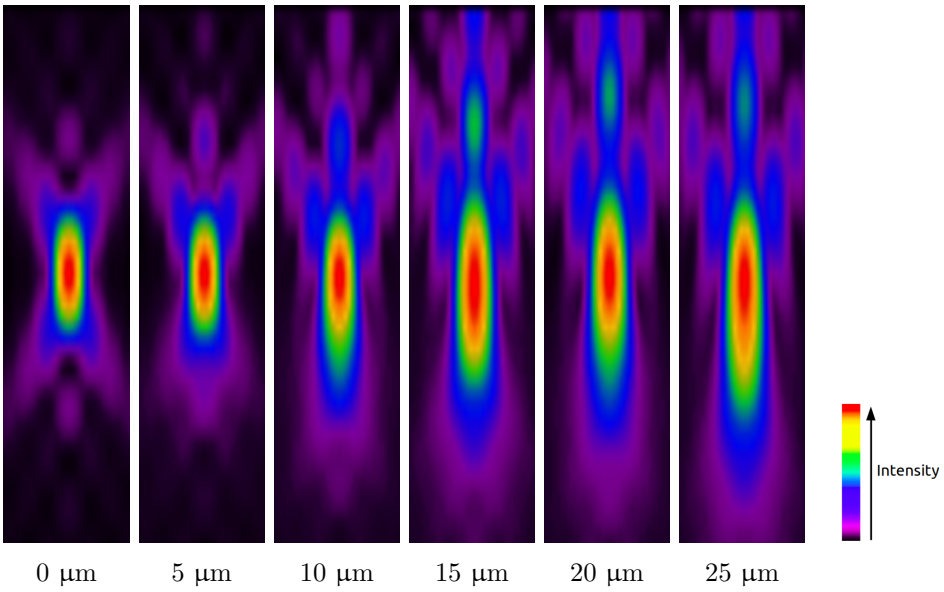


Figure 2.13: The PSF at increasing depth with an oil lens immersion medium ($n_1 = 1.515$) and a watery specimen embedding medium ($n_2 = 1.338$). The lens is at the bottom of the image. The width of each image is 2 µm. Image reproduced from [18].

A fluorescent molecule can only undergo a finite number of absorption-emission cycles. After that, the molecule loses its fluorescence permanently. This process of photobleaching leads to a fading of the emission intensity of the sample over time. This is especially problematic for widefield images because planes above and below the focal plane are also illuminated. This leads to a decrease of intensity in the z-direction of the image, which is problematic when deconvolving such an image.

In the derivation of the likelihood functions, we set the Poisson parameter λ for each voxel equal to the intensity of that voxel after convolution and addition of background. By doing that, we make the assumption that one unit of intensity corresponds to one photon. However, in reality it is more complex than that. Also, the assumption that the noise is purely Poisson noise, is not completely correct. To understand the noise process in more detail, we need to know how the intensities are measured in a microscope. Charge-coupled device (CCD) cameras are often used in fluorescence microscopy. This camera contains a CCD-chip that consists of a two-dimensional array of pixels. The chip makes use of the photoelectric effect to convert electromagnetic radiation into electrons. The quantum efficiency is the rate at which the sensor converts photons to electrons. The total number of electrons is the sum of electrons due to radiation of the object and background radiation. Because the photons arrive randomly, the number of photoconversions can be seen as a Poisson

random process. An amplifier converts the electrons to an electric charge. After that, an AD converter translates the analog signal into a digital signal. The amplifier and the digitization together introduce noise that is normally distributed [5]. Thus, the noise is a combination of Poisson photon noise and Gaussian read-out noise. This leads to a more complex likelihood function as described in [34]. It turns out that this likelihood function is only necessary in very particular circumstances [5].

The number of photons is found by multiplying the intensity by the PPU (photons per unit) of the image. To know the PPU, we need to know the quantum efficiency of the CCD camera. Also, we need to know how the amplifier converts the electrons into a voltage and how this is signal is converted to a digital signal by the AD converter. Generally, the PPU is unknown and therefore, the model that we are fitting to the data is not completely correct. In Chapter 5 we will perform an experiment to measure how much the result of deconvolution is affected when the PPU is far from one.

2.4 The inverse problem and ill-posedness

We showed how the image formation process can be modeled. The inverse problem is now to reconstruct the original object $\boldsymbol{w}$ from the noisy measurement $\boldsymbol{f}$. We want to invert the process of convolution and noise that degraded the original object. To solve the inverse problem, the goal will be to find an image $\tilde{\boldsymbol{w}}$ such that if the image formation process is applied to it, the result matches $\boldsymbol{f}$. In some cases, an inverse problem can be solved in a direct way by applying the inverse operator K^{-1} to $\boldsymbol{f}$. This is possible, for example, when K is a invertible matrix. We will see in this section that convolution has an inverse operator. However, it will be shown that the problem is ill-posed and that the solution of the inverse operator is useless.

An inverse problem $K(\boldsymbol{w}) = \boldsymbol{f}$ is a well-posed problem when

1. A solution exists.
2. The solution is unique.
3. The solution depends continuously on the data. This means that there exists a constant $C \in \mathbb{R}$ such that for all $K(\boldsymbol{w}) = \boldsymbol{f}$ and $K(\boldsymbol{w'}) = \boldsymbol{f'}$ it holds that $|\boldsymbol{w} - \boldsymbol{w'}| \leq C|\boldsymbol{f} - \boldsymbol{f'}|$

A problem is ill-posed when it is not well-posed [16]. To investigate the ill-posedness of deconvolution we will use the Fourier transform $\mathcal{F}$ which decomposes a signal into frequencies. The convolution theorem says that the convolution of two signals H and W in the spatial domain corresponds to an element-wise product in the frequency domain. In other words,

$$\mathcal{F}(\boldsymbol{h} * \boldsymbol{w}) = \mathcal{F}(\boldsymbol{h}) \cdot \mathcal{F}(\boldsymbol{w}). \tag{2.1}$$

The Fourier transform of the PSF is called the optical transfer function (OTF). In Chapter 3 we will formally define the Fourier transform and prove the convolution theorem. It follows from (2.1) that the solution can be found by dividing by the OTF

$$\boldsymbol{w} = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(\boldsymbol{h} * \boldsymbol{w})}{\mathcal{F}(\boldsymbol{h})}\right).$$

Whenever $\mathcal{F}$ does not contain zero values this is well defined and a solution exists. However, this method will lead to heavy amplification of noise in the measurements. Let $\boldsymbol{f} = \boldsymbol{h} * \boldsymbol{w}$ be the image

without noise and $\boldsymbol{f}^{\boldsymbol{\delta}} = \boldsymbol{f} + \boldsymbol{\delta}$ be the noisy data. Let $\boldsymbol{w}^{\boldsymbol{\delta}}$ be the solution that we get from the noisy data

$$\boldsymbol{w}^{\boldsymbol{\delta}} = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(\boldsymbol{f}^{\boldsymbol{\delta}})}{\mathcal{F}(\boldsymbol{h})}\right). \tag{2.2}$$

Then, because both the Fourier transform and the inverse Fourier transform are linear, the error in the solution is

$$||\boldsymbol{w}^{\boldsymbol{\delta}} - \boldsymbol{w}|| = \left|\left|\mathcal{F}^{-1}\left(\frac{\mathcal{F}(\boldsymbol{f})}{\mathcal{F}(\boldsymbol{h})}\right) - \mathcal{F}^{-1}\left(\frac{\mathcal{F}(\boldsymbol{f}^{\boldsymbol{\delta}})}{\mathcal{F}(\boldsymbol{h})}\right)\right|\right| = \left|\left|\mathcal{F}^{-1}\left(\frac{\mathcal{F}(\boldsymbol{\delta})}{\mathcal{F}(\boldsymbol{h})}\right)\right|\right|.$$

For high frequencies, the OTF usually tends to zero while the high frequency noise does not. This leads to amplification of the noise. Therefore, the solution does not depend continuously on the data and so the inverse problem is ill-posed [3]. An example is shown in Figure 2.14 where the noisy measurement of Figure 2.11c is deconvolved using division by the OTF. In this image we can see how the very small values of $\boldsymbol{h}$ lead the to amplification of the high frequencies. The resulting image is then completely dominated by noise.

When a problem is ill-posed, we can instead solve a modified version of the problem $\tilde{K}(\boldsymbol{w}) = \boldsymbol{f}$ where $\tilde{K}$ is a well-posed operator, as we will see in Chapter 3. Another approach is to use variational methods. The original problem is then replaced by a minimization problem in such a way that the solution to the inverse problem is the minimizer of some functional [20]. This functional is often chosen as a likelihood function together with a regularization term. The likelihood function expresses the probability of finding the measured data for a given object and noise model. This function is minimized by the objects that best fits the data. In Section 2.5 we will derive likelihood functions in the case of Poisson and Gaussian noise.

2.5 Maximum likelihood estimation

The image $\boldsymbol{f}$ can be seen as a realization of a random variable $\boldsymbol{F} = (F_1, ..., F_N)$. The maximum likelihood estimator $\hat{\boldsymbol{w}}$ is the object that maximizes the probability of measuring the observed data $\boldsymbol{f}$, that is

$$\hat{\boldsymbol{w}} = \arg\max_{\boldsymbol{w}} \; P(\boldsymbol{F} = \boldsymbol{f} \mid \boldsymbol{w}).$$

Since the logarithmic function is strictly increasing, maximizing the likelihood is equivalent to minimizing the negative log-likelihood

$$y(\boldsymbol{w}) := -\log\left(P(\boldsymbol{F} = \boldsymbol{f} \mid \boldsymbol{w})\right). \tag{2.3}$$

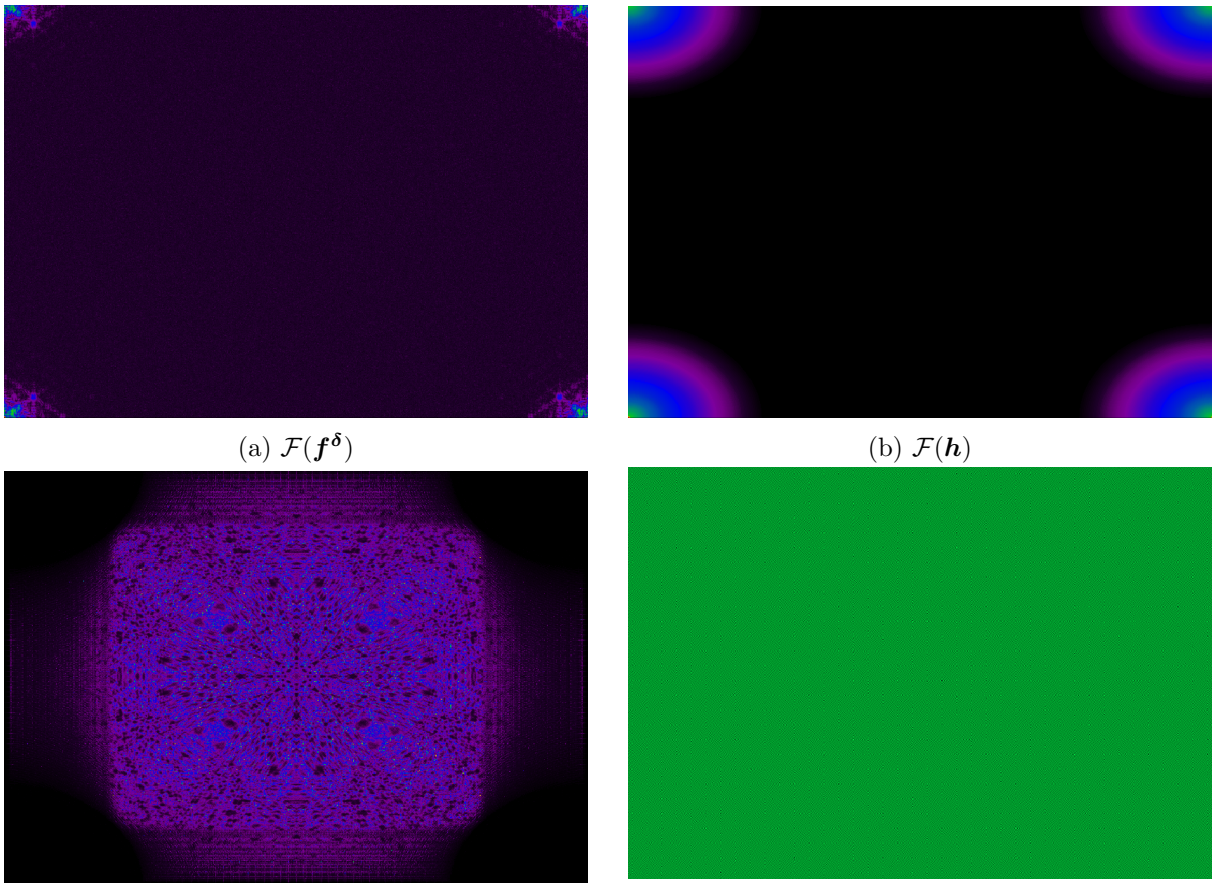
We will now derive a formula for $y(\boldsymbol{w})$ that follows from the forward model of Section 2.2 and the assumption of Poisson noise. We will show that the Poisson distribution can be approximated by the normal distribution and also derive a formula for $y(\boldsymbol{w})$ in with the assumption of Gaussian noise.

The forward model is given by

$$\boldsymbol{f} = K\boldsymbol{w} + \boldsymbol{b} + \boldsymbol{\delta}$$

In the case of Poisson noise, the measured intensity f_i is the realization of the Poisson-distributed random variable F_i. For each voxel $i = 1, ..., N$, the probability of measuring f_i photons is given by the probability density function

$$\mathrm{P}\left(F_i = f_i\right) = \frac{\lambda_i^{f_i} \exp(-\lambda_i)}{f_i!} \tag{2.4}$$



(a) $\mathcal{F}(\boldsymbol{f}^{\boldsymbol{\delta}})$ (b) $\mathcal{F}(\boldsymbol{h})$

(c) $\frac{\mathcal{F}(\boldsymbol{f}^{\boldsymbol{\delta}})}{\mathcal{F}(\boldsymbol{h})}$ (d) $\boldsymbol{w}^{\boldsymbol{\delta}} = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(\boldsymbol{f}^{\boldsymbol{\delta}})}{\mathcal{F}(\boldsymbol{h})}\right)$

Figure 2.14: Deconvolution of the noisy image of Figure 2.11c by using equation (2.2). The images are sum projections onto the plane spanned by the first two coordinate axes and the origin lies in the corners of the frequency domain images. Therefore, the corners represent the low frequencies while the high frequencies are depicted in the center of the images. The last image is in the spatial domain.

where $\lambda_i = (K\boldsymbol{w} + \boldsymbol{b})_i$. The measured intensities f_i are independent of each other and therefore the probability of measuring $\boldsymbol{f}$ is the product of the probabilities of measuring the individual intensities.

The function $y(\boldsymbol{w})$ is then given by

$$\begin{aligned} y(\boldsymbol{w}) &= -\log\left(\prod_{i=1}^{N} \mathrm{P}(F_i = f_i \mid \boldsymbol{w})\right) \\ &= -\sum_{i=1}^{N} \log\left(\mathrm{P}(F_i = f_i \mid \boldsymbol{w})\right) \\ &= -\sum_{i=1}^{N} \log\left(\frac{(K\boldsymbol{w}+\boldsymbol{b})_i^{f_i} \exp(-(K\boldsymbol{w}+\boldsymbol{b})_i)}{f_i!}\right) \\ &= \sum_{i=1}^{N} -f_i \log\left((K\boldsymbol{w}+\boldsymbol{b})_i\right) + (K\boldsymbol{w}+\boldsymbol{b})_i + \log(f_i!) \\ &\approx \sum_{i=1}^{N} -f_i \log\left((K\boldsymbol{w}+\boldsymbol{b})_i\right) + (K\boldsymbol{w}+\boldsymbol{b})_i + f_i\log(f_i) - f_i \end{aligned}$$

where we approximated $\log(f_i!)$ using Stirling's approximation $\log(n!) = n\log(n) - n + \mathcal{O}(\log(n))$. The maximum likelihood estimator is then given by

$$\hat{\boldsymbol{w}} = \arg\max_{\boldsymbol{w}} \sum_{i=1}^{N} (K\boldsymbol{w}+\boldsymbol{b})_i - f_i - f_i \log\left(\frac{(K\boldsymbol{w}+\boldsymbol{b})_i}{f_i}\right) \tag{2.5}$$

which is known as the Kullback-Leibler divergence or KL-divergence from $K\boldsymbol{w}+\boldsymbol{b}$ to $\boldsymbol{f}$. The term KL-divergence is often used to denote the slightly different function $\sum_{i=1}^{N} -f_i \log\left(\frac{(K\boldsymbol{w}+\boldsymbol{b})_i}{f_i}\right)$. This function is only applicable when comparing two probability distributions and in that case equal to equation (2.5) since the terms sum up to one. In the remainder of this text, we will always mean equation (2.5) when referring to the KL-divergence.

We will now shown that when λ is large enough, the Poisson(λ) distribution can be approximated by the normal distribution $N(\mu = \lambda, \sigma^2 = \lambda)$. Let $\lambda \in \mathbb{N}$ and let $X_1, X_2, ..., X_\lambda$ be independent Poisson(1) random variables. Then $Y = \sum_{i=1}^{\lambda} X_i$ is Poisson distributed with parameter λ. Then it follows from the central limit theorem that

$$\frac{Y-\lambda}{\sqrt{\lambda}} \xrightarrow{d} N(0,1)$$

as $\lambda \to \infty$. Then it follows that Y approaches $N(\lambda, \lambda)$. So for images with high photon counts, we can assume the noise to be normally distributed. However, for images with a small signal-to-noise ratio, this is not a good approximation.

In the case of additive Gaussian noise, the forward model is given by

$$\boldsymbol{f} = K\boldsymbol{w} + \boldsymbol{b} + \boldsymbol{\eta}$$

where $\eta_i \sim N(0, \sigma^2)$ for some $\sigma \in \mathbb{R}_{\geq 0}$. Note that we assume the same variance for each voxel, in contrast to the Poisson distribution where each voxel has a different variance. The probability of measuring f_i photons is given by

$$\mathrm{P}(F_i = f_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{f_i - (K\boldsymbol{w}+\boldsymbol{b})_i}{\sigma}\right)^2\right)$$

Substituting this into equation (2.3), we get

$$\begin{aligned} y(\boldsymbol{w}) &= -\log\left(\prod_{i=1}^{N} \mathrm{P}(F_i = f_i \mid \boldsymbol{w})\right) \\ &= -\sum_{i=1}^{N} \log\left(\mathrm{P}(F_i = f_i \mid \boldsymbol{w})\right) \\ &= -\sum_{i=1}^{N} \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2}\left(\frac{f_i - (K\boldsymbol{w} + \boldsymbol{b})_i}{\sigma}\right)^2 \end{aligned}$$

Since we are interested in the minimizer of this function, constants can be omitted. We then find that $\hat{\boldsymbol{w}}$ minimizes the residual sum of squares

$$\hat{\boldsymbol{w}} = \arg\max_{\boldsymbol{w}} \sum_{i=1}^{N} (f_i - (K\boldsymbol{w} + \boldsymbol{b})_i)^2 \tag{2.6}$$

Chapter 3

Regularization

In the previous chapter we defined the inverse problem of image reconstruction as finding an object $\boldsymbol{w} \in \mathbb{R}^N$ that fits the model $K\boldsymbol{w} + \boldsymbol{b} = \boldsymbol{f}$ where $\boldsymbol{f} \in \mathbb{R}^M$ is the measurement, $K \in \mathbb{R}^{M\times N}$ is the convolution matrix, and $\boldsymbol{b} \in \mathbb{R}^N$ is the background. By subtracting the background from the original measurement, and regard that vector as the new measurement, it suffices for this chapter to view the problem as only a matrix-vector product. Therefore, we will consider linear inverse problems of the form

$$K\boldsymbol{w} = \boldsymbol{f}, \tag{3.1}$$

where $K \in \mathbb{R}^{M\times N}$, $\boldsymbol{w} \in \mathbb{R}^N$ and $\boldsymbol{f} \in \mathbb{R}^M$. When $M = N$, and K is invertible, the solution is given by $\boldsymbol{w} = K^{-1}\boldsymbol{f}$. When K is not invertible, we can use a generalized inverse called the pseudo-inverse, which we will define in this chapter. We start by defining the singular value decomposition of a matrix. Also, we define the discrete Fourier transform. We show how, in the case that K represents a convolution, the singular value decomposition is given by a discrete Fourier transform. After that, it will be shown how the pseudo-inverse is defined in terms of the singular value decomposition. We will explain how noise in the measurements is amplified by small singular values and how regularization tries to solve this problem. We define Tikhonov regularization as an example of spectral regularization. This chapter is based on [3] and [4].

3.1 Singular value decomposition

Definition 3.1. Let K be an $M \times N$ matrix. The *singular value decomposition (SVD)* of K is a decomposition of the form

$$K = U\Sigma V^*,$$

where $U \in \mathbb{C}^{M\times M}$ and $V \in \mathbb{C}^{N\times N}$ are unitary matrices, and Σ is an $M \times N$ diagonal matrix with real non-negative diagonal elements. The diagonal elements $\sigma_i = \Sigma_{ii}$ are called the *singular values* of K. It is conventional to sort the singular values such that $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_p \geq 0$ where $p = min(M, N)$. The columns $\{\boldsymbol{u_1}, ..., \boldsymbol{u_M}\}$ of U and the columns $\{\boldsymbol{v_1}, ..., \boldsymbol{v_N}\}$ of V are called the *left-singular vectors* and the *right-singular vectors* of K, respectively. □

Each real or complex matrix K has a singular value decomposition. The SVD generalizes the eigendecomposition of a diagonizable square matrix. This can be seen by computing the matrices K^*K and KK^*

$$K^*K = V\Sigma^*U^*\, U\Sigma V^* = V\,(\Sigma^*\Sigma)\,V^*$$
$$KK^* = U\Sigma V^*\, V\Sigma^*U = U\,(\Sigma\Sigma^*)\,U.$$

The right hand sides are now the eigendecompositions and so the singular vectors $\{\boldsymbol{u_1}, ..., \boldsymbol{u_M}\}$ and $\{\boldsymbol{v_1}, ..., \boldsymbol{v_N}\}$ are the eigenvectors of KK^* and K^*K, respectively, and the squared singular values σ_i^2 are the corresponding eigenvalues. The number of nonzero singular values is equal to the rank of K. The sequence $\{(\sigma_j, \boldsymbol{u_j}, \boldsymbol{v_j})\}$ is called the singular system of K.

We now consider K as a linear transformation $K : \mathbb{R}^N \to \mathbb{R}^M$. Applying K to some vector $\boldsymbol{w} \in \mathbb{R}^n$ can now be written as

$$K\boldsymbol{w} = \sum_{j=1}^{k} \sigma_j \langle \boldsymbol{w}, \boldsymbol{v_j} \rangle \boldsymbol{u_j}, \tag{3.2}$$

where k is the rank of K. Since U and V are unitary, $\{\boldsymbol{u_1}, ..., \boldsymbol{u_M}\}$ and $\{\boldsymbol{v_1}, ..., \boldsymbol{v_N}\}$ are orthonormal bases of $\mathbb{C}^m$ and $\mathbb{C}^n$, respectively. The matrix-vector multiplication can now be interpreted as a linear transformation that consists of three steps. The first step is a basis transformation from the standard basis of $\mathbb{R}^n$ to the basis $\{\boldsymbol{v_1}, ..., \boldsymbol{v_N}\}$ where the component of w in the direction of $\boldsymbol{v_j}$ is given by $\langle \boldsymbol{w}, \boldsymbol{v_j} \rangle$. Next, each coordinate j is scaled by the singular value σ_j. The last step is a basis transformation from $\{\boldsymbol{u_1}, ..., \boldsymbol{v_M}\}$ to the standard basis of $\mathbb{C}^m$. The scalar $\sigma_j \langle \boldsymbol{w}, \boldsymbol{v_j} \rangle$ is the component of the solution in the direction of $\boldsymbol{u_j}$ and therefore the desired result is given by equation (3.2).

3.2 The discrete Fourier transform

Definition 3.2. (Discrete Fourier transform)
The *discrete Fourier transform (DFT)* is a linear transformation that maps an N-dimensional complex vector onto another N-dimensional complex vector. Let $\boldsymbol{f} = (f_0, f_1, ..., f_{N-1}) \in \mathbb{C}^N$. Then the discrete Fourier transform $\mathcal{F} : \mathbb{C}^N \to \mathbb{C}^N, \boldsymbol{f} \mapsto \hat{\boldsymbol{f}}$ is defined by

$$\mathcal{F}(\boldsymbol{f})_m = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} f_n \exp\left(-i \frac{2\pi}{N} mn\right) \tag{3.3}$$

for $m = 0, 1, ..., N-1$. The term discrete Fourier transform can refer to both the map $\mathcal{F}$ as the resulting vector $\hat{\boldsymbol{f}} = \left(\hat{f}_0, \hat{f}_1, ..., \hat{f}_{N-1}\right)$. □

The DFT can be seen as a linear combination of vectors of an orthonormal basis. We define

$$\omega = \exp\left(\frac{2\pi i}{N}\right) \tag{3.4}$$

and the *DFT basis vectors* $\boldsymbol{v_0}, \boldsymbol{v_1}, ..., \boldsymbol{v_{N-1}} \in \mathbb{C}^N$ where the n-th element of the m-th vector is given by

$$(\boldsymbol{v_m})_n = \frac{1}{\sqrt{N}} \cdot \omega^{mn}. \tag{3.5}$$

The inner product of two vectors is

$$\langle \boldsymbol{v_m}, \boldsymbol{v_l} \rangle = \frac{1}{N} \sum_{k=0}^{N-1} \exp\left(\frac{2\pi i}{N}(m-l)k\right).$$

When $m = l$, this is equal to 1 and when $m \neq l$, it follows from the summation formula for the first N terms of a geometric series (equation (0.1)), that this is equal to

$$\frac{1}{N}\frac{1-\exp\left(i2\pi(m-l)\right)}{1-\exp\left(i\frac{2\pi}{N}(m-l)\right)} = 0.$$

Therefore, the vectors $\boldsymbol{v_0}, \boldsymbol{v_1}, ..., \boldsymbol{v_{N-1}}$ form an orthonormal basis of $\mathbb{C}^N$.

We can now express the components of the DFT as an inner product of $\boldsymbol{f}$ with a DFT basis vector

$$\hat{f}_m = \langle \boldsymbol{f}, \boldsymbol{v_m} \rangle. \tag{3.6}$$

We define the *DFT matrix* Ω as the matrix that has the DFT basis vectors as columns. That is,

$$\Omega_{m,n} = \frac{1}{\sqrt{N}} \exp\left(\frac{2\pi i}{N} mn\right) \tag{3.7}$$

where we start counting the rows and columns at zero. The matrix then looks like

$$\Omega = \frac{1}{\sqrt{N}} \begin{pmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega & \omega^2 & \omega^3 & \cdots & \omega^{N-1} \\ 1 & \omega^2 & \omega^4 & \omega^6 & \cdots & \omega^{2(N-1)} \\ 1 & \omega^3 & \omega^6 & \omega^9 & \cdots & \omega^{3(N-1)} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{N-1} & \omega^{2(N-1)} & \omega^{3(N-1)} & \cdots & \omega^{(N-1)(N-1)} \end{pmatrix}.$$

Since the DFT basis vectors form an orthonormal basis of $\mathbb{C}^N$, the DFT matrix is unitary. Therefore, its inverse is given by the conjugate transpose Ω^* with

$$\Omega^*_{m,n} = \frac{1}{\sqrt{N}} \exp\left(\frac{-2\pi i}{N} mn\right). \tag{3.8}$$

We can now express the DFT as a matrix-vector multiplication

$$\hat{\boldsymbol{f}} = \Omega^* \boldsymbol{f}.$$

The *inverse discrete Fourier transform (IDFT)* is given by

$$\boldsymbol{f} = \Omega \hat{\boldsymbol{f}} \tag{3.9}$$

which can be expressed in a similar way as equation (3.3)

$$\mathcal{F}^{-1}(\hat{\boldsymbol{f}})_n = \frac{1}{\sqrt{N}} \sum_{m=0}^{N-1} \hat{f}_m \exp\left(i\frac{2\pi}{N} mn\right)$$

or in terms of the DFT basis vectors

$$\boldsymbol{f} = \sum_{m=0}^{N-1} \hat{f}_m \boldsymbol{v_m}. \tag{3.10}$$

Combining equations (3.6) and (3.10) we get

$$\boldsymbol{f} = \sum_{m=0}^{N-1} \langle \boldsymbol{f}, \boldsymbol{v_m} \rangle \boldsymbol{v_m}.$$

In addition to a transformation of vectors, the discrete Fourier transform can also be interpreted as a transformation for periodic sequences. For a given vector $\boldsymbol{f} = (f_0, f_1, ..., f_{N-1})$, we can define the periodic sequence f_n by $f_n := f_{n \bmod N}$ for all $n \in \mathbb{Z}$. Then, the sequence $\hat{f}_m$, with $m \in \mathbb{Z}$, defined by equation (3.3) is also periodic with period N. The DFT basis vectors can be seen as discrete periodic functions of increasing frequencies. The discrete Fourier transform can then be interpreted as decomposing a periodic sequence into a sum of complex exponential functions of different frequencies. The amount that each frequency is present in a signal is determined by taking the inner product of the signal with the corresponding DFT basis vector.

The formulas related to the DFT can differ slightly in different textbooks. Sometimes, the normalization factors of $\frac{1}{\sqrt{N}}$ are chosen in a different way. Instead of a factor of $\frac{1}{\sqrt{N}}$ for both the forward are the inverse transform, there can be a factor of $\frac{1}{N}$ in front of one of them. Also, the sign of the exponent of the basis vectors is sometimes chosen negative. This can lead to small changes in the properties. This is correct as well, as long as the product of the two normalization factors is $\frac{1}{N}$ and the sign of the exponents in the DFT and its inverse are opposite. Our choice of $\frac{1}{\sqrt{N}}$ and the positive sign has the favorable property that the DFT matrix is unitary and that the eigenvectors of a circulant matrix are equal to the DFT basis vectors instead of their complex conjugates.

Multidimensional DFT

In Section 2.2 we explained how a 3D image can be represented by a vector in $\mathbb{R}^N$ where the voxels are stack onto each other. However, when computing the Fourier transform of an image it is important to take into account the spatial structure on an image. When we would compute the Fourier transform of the vector representation of an image, we would regard the image as a one-dimensional signal and then decompose that signal into complex exponential functions. However, two intensity values that are far away from each other in the vector representation can actually represent neighboring voxels. Therefore, we want to decompose the image into three-dimensional exponential functions of different frequencies. The theory of discrete Fourier transforms can be extended to multidimensional signals. With our applications in mind we will give the formulas for the DFT of a three-dimensional image. The extensions to other dimensions will then also be clear.

Definition 3.3. Let $F(n_1, n_2, n_3) \in \mathbb{C}^{N_1 \times N_2 \times N_3}$ be a third-order tensor. Then the *three-dimensional discrete Fourier transform* of F is a given by $\hat{F}(m_1, m_2, m_3) \in \mathbb{C}^{N_1 \times N_2 \times N_3}$ which is given by

$$\hat{F}(m_1, m_2, m_3) = \frac{1}{\sqrt{N_1 N_2 N_3}} \sum_{n_1=0}^{N_1-1} \sum_{n_2=0}^{N_2-1} \sum_{n_3=0}^{N_3-1} F(n_1, n_2, n_3) \exp\left(-2\pi i \left(\frac{m_1 n_1}{N_1} + \frac{m_2 n_3}{N_1} + \frac{m_3 n_3}{N_1}\right)\right).$$

and the *inverse three-dimensional discrete Fourier transform* is given by

$$F(n_1, n_2, n_3) = \frac{1}{\sqrt{N_1 N_2 N_3}} \sum_{m_1=0}^{N_1-1} \sum_{m_2=0}^{N_2-1} \sum_{m_3=0}^{N_3-1} F(m_1, m_2, m_3) \exp\left(2\pi i \left(\frac{n_1 m_1}{N_1} + \frac{m_2 n_3}{N_1} + \frac{m_3 n_3}{N_1}\right)\right).$$

□

3.3 Diagonalization of a convolution matrix

In the case that K represents a convolution, the singular value decomposition has a special form. We will treat the case of a one-dimensional signal because this makes the theory easier to understand

and the notation less complex. The extension to 2D-images is described in appendices A and B of [1].

Let $\boldsymbol{w} = (w_0, w_1, ..., w_{N-1}) \in \mathbb{R}^N$ be a signal and $\boldsymbol{h} = (h_0, h_1, ..., h_{n-1}) \in \mathbb{R}^N$ be a point spread function. We extend $\boldsymbol{h}$ periodically by defining $h_i := h_{i \bmod N}$ for all $i \in \mathbb{Z}$. The convolution of $\boldsymbol{w}$ with $\boldsymbol{h}$ is given by

$$(\boldsymbol{h} * \boldsymbol{w})_m = \sum_{n=0}^{N-1} h_{m-n} w_n$$

for $i = 0, 1, ..., N-1$. Convolution of a vector $\boldsymbol{w}$ with a fixed point spread function $\boldsymbol{h}$ defines a linear operator K such that

$$K\boldsymbol{w} = \boldsymbol{h} * \boldsymbol{w}.$$

The elements of K are given by $K_{m,n} = h_{m-n}$ and so K is the circulant matrix

$$K = \begin{pmatrix} h_0 & h_{N-1} & h_{N-2} & \cdots & h_2 & h_1 \\ h_1 & h_0 & h_{N-1} & \cdots & h_3 & h_2 \\ h_2 & h_1 & h_0 & \cdots & h_4 & h_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ h_{N-1} & h_{N-2} & h_{N-3} & \cdots & h_1 & h_0 \end{pmatrix}. \tag{3.11}$$

We will now show the connection between the discrete Fourier transform and the eigenvalues and eigenvectors of a convolution matrix.

Theorem 3.4. *Let K be a circulant matrix that is generated by a vector $\boldsymbol{h} = (h_0, ..., h_{N-1})$, as in equation (3.11). Let $\hat{\boldsymbol{h}} = \left(\hat{h}_0, ..., \hat{h}_{N-1}\right)$ be the DFT of $\boldsymbol{h}$, given by equation (3.3). Let $\boldsymbol{v_0}, ..., \boldsymbol{v_{N-1}}$ be the DFT basisvectors, given by equation (3.5). Then the eigenvectors of K are equal to $\boldsymbol{v_m}$, and the corresponding eigenvalues are equal to $\sqrt{N}\hat{h}_m$ for $m = 0, 1, ..., N-1$.*

Proof. Let $\boldsymbol{v_m}$ be the m-th DFT basis vector. We will show that

$$K\boldsymbol{v_m} = \sqrt{N}\hat{h}_m \boldsymbol{v_m}. \tag{3.12}$$

The k-th element of $K\boldsymbol{v_m}$ is given by

$$(K\boldsymbol{v_m})_k = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} h_{k-n} \exp\left(\frac{2\pi i}{N} mn\right). \tag{3.13}$$

We apply a change of variables by letting $l = k - n$.

$$\begin{aligned}(K\boldsymbol{v_m})_k &= \frac{1}{\sqrt{N}} \sum_{l=k}^{k-N+1} h_l \exp\left(\frac{2\pi i}{N} m(k-l)\right) \\ &= \frac{1}{\sqrt{N}} \left[\sum_{l=k}^{k-N+1} h_l \exp\left(\frac{-2\pi i}{N} ml\right)\right] \left[\exp\left(\frac{2\pi i}{N} mk\right)\right].\end{aligned}$$

Because h_l and $\exp\left(\frac{2\pi i}{N} ml\right)$ are periodic in l with period N this is equal to

$$\begin{aligned}(K\boldsymbol{v_m})_k &= \sqrt{N}\left[\frac{1}{\sqrt{N}} \sum_{l=0}^{N-1} h_l \exp\left(\frac{-2\pi i}{N} ml\right)\right] \left[\frac{1}{\sqrt{N}} \exp\left(\frac{2\pi i}{N} mk\right)\right] \\ &= \sqrt{N}\left(\Omega^* \boldsymbol{h}\right)_m (\boldsymbol{v_m})_k \\ &= \sqrt{N}\hat{h}_m (\boldsymbol{v_m})_k.\end{aligned}$$

□

A consequence of this is that a convolution matrix K can be diagonalized as

$$K = \Omega\Sigma\Omega^* \tag{3.14}$$

where Σ is the diagonal matrix with $\Sigma_{mm} = \sqrt{N}\hat{h}_m$. Since Ω is unitary, this eigendecomposition is equal the singular value decomposition of K in the case that $\hat{\boldsymbol{h}}$ is real.

The convolution theorem tells us that convolution in the spatial domain corresponds to element-wise multiplication in the frequency domain.

Theorem 3.5. *(Convolution theorem)*
Let $\boldsymbol{w} = (w_0, w_1, ..., w_{N-1}) \in \mathbb{R}^N$ be a signal and $\boldsymbol{h} = (h_0, h_1, ..., h_{n-1}) \in \mathbb{R}^N$ be a point spread function. Let $\boldsymbol{f}$ be the convolution of $\boldsymbol{w}$ with $\boldsymbol{h}$. Then $\hat{\boldsymbol{f}}$ is the element-wise product of $\hat{\boldsymbol{w}}$ and $\hat{\boldsymbol{h}}$, up to a constant of $\sqrt{N}$. That is,

$$\hat{f}_m = \sqrt{N}\hat{h}_m\hat{w}_m \tag{3.15}$$

for $m = 0, 1, ..., N-1$.

Proof. Let K be the convolution matrix generated by $\boldsymbol{h}$ with diagonalization $K = \Omega\Sigma\Omega^*$. Then the convolution can be expressed as

$$\begin{aligned}
\boldsymbol{f} &= \boldsymbol{h} * \boldsymbol{w} \\
&= K\boldsymbol{w} \\
&= \Omega\Sigma\Omega^*\boldsymbol{w} \\
&= \sqrt{N}\sum_{k=0}^{N-1} \hat{h}_k\langle \boldsymbol{w}, \boldsymbol{v_k}\rangle \boldsymbol{v_k}.
\end{aligned}$$

From equation (3.6) it follows that this is equal to

$$\boldsymbol{f} = \sqrt{N}\sum_{k=0}^{N-1} \hat{h}_k\hat{w}_k\boldsymbol{v_k}.$$

Equation (3.10) says that $\boldsymbol{f}$ can be expressed as

$$\boldsymbol{f} = \sum_{k=0}^{N-1} \hat{f}_k\boldsymbol{v_k}.$$

Since the DFT basis vectors are linearly independent, it follows that equation (3.15) is true. □

3.4 The pseudo-inverse

We will now use the singular value decomposition to find a solution to the inverse problem (3.1). Let

$$K = U\Sigma V^*$$

be the singular value decomposition of K. Using the singular value decomposition, we can construct a generalization of the inverse of a matrix. This generalized inverse is known as the pseudo-inverse or the Moore-Penrose inverse. Let $V_k \in \mathbb{C}^{n\times k}$ be the matrix that contains the first k right singular vectors as columns, let $U_k \in \mathbb{C}^{m\times k}$ be the matrix that contains the first k left singular vectors

as columns and let $\Sigma_k \in \mathbb{R}^{k\times k}$ be the diagonal matrix with $\sigma_1, ..., \sigma_k$ on the diagonal. Then the pseudo-inverse $K^\dagger$ is defined as

$$K^\dagger = V_k \Sigma_k^{-1} U_k^*.$$

Applying the pseudo-inverse to the measurements gives the solution

$$\widetilde{\boldsymbol{w}} = K^\dagger \boldsymbol{f} = V_k \Sigma_k^{-1} U_k^* \boldsymbol{f} = \sum_{j=1}^{k} \frac{\langle \boldsymbol{f}, \boldsymbol{u_j} \rangle}{\sigma_j} \boldsymbol{v_j}. \tag{3.16}$$

Note that the solution is found by applying the opposite operations as in equation (3.2). The measurements are projected onto the basis $\{\boldsymbol{u_1}, ..., \boldsymbol{u_m}\}$, then divided by the singular values and finally expressed in the standard basis of $\mathbb{C}^m$.

When the columns and/or rows of K are independent, we can give more specific expressions for the pseudo-inverse. We will treat three different cases.

Case 1: $m = n = \text{rank}(K)$.
In this case, the pseudo-inverse corresponds to the normal inverse K^{-1}. In this case the solution exists and is unique.

Case 2: $m > n = \text{rank}(K)$.
In this case there are more measurements than unknowns. When $\boldsymbol{f}$ is not in the range of K, there is no exact solution. It can be shown that the pseudo-inverse solution corresponds to the least-squares solution.

Theorem 3.6. *Let $K \in \mathbb{R}^{m\times n}$ and $\boldsymbol{f} \in \mathbb{R}^m$ with $m > n = \text{rank}(K)$. Let $\widetilde{\boldsymbol{w}}$ be the solution given by equation* (3.16)*. Then $\widetilde{\boldsymbol{w}}$ is equal to the least squares solution, that is*

$$\widetilde{\boldsymbol{w}} = \arg\min_u ||K\boldsymbol{w} - \boldsymbol{f}||^2.$$

Proof. The gradient of $||K\boldsymbol{w} - \boldsymbol{f}||^2$ is given by

$$\nabla ||K\boldsymbol{w} - \boldsymbol{f}||^2(\boldsymbol{w}) = 2K^*K\boldsymbol{w} - 2K^*\boldsymbol{f}.$$

Setting the gradient to zero gives

$$K^*K\boldsymbol{w} = K^*\boldsymbol{f}.$$

Since the $\text{rank}(K) = n$, also $\text{rank}(K^*K) = n$ and so K^*K is invertible. Therefore, the least squares solution is given by

$$\boldsymbol{w_{ls}} = (K^*K)^{-1} K^*\boldsymbol{f}.$$

Filling in the singular value decomposition of K we find

$$\begin{aligned}
\boldsymbol{w_{ls}} &= (V\Sigma^* U^* U \Sigma V^*)^{-1} V\Sigma U^* \boldsymbol{f} \\
&= \left(V\Sigma_n^2 V^*\right)^{-1} V\Sigma U^* \boldsymbol{f} \\
&= V\Sigma_n^{-2} V^* V \Sigma^* U^* \boldsymbol{f} \\
&= V_n \Sigma_n^{-1} U_n^* \boldsymbol{f} \\
&= K^\dagger \boldsymbol{f} = \widetilde{\boldsymbol{w}}.
\end{aligned}$$

□

So in the case that $m > n$ and $\text{rank}(K) = n$, the pseudo-inverse is equal to

$$K^\dagger = (K^*K)^{-1} K^*.$$

Then $K^\dagger K = I_n$ and so $K^\dagger$ is a left-inverse of K.

Case 3: $n > m = \text{rank}(K)$.
In this case, there are more unknowns than measurements. A solution will always exist but it is not unique. It can be shown that the pseudo-inverse solution corresponds to the solution with the smallest norm.

Theorem 3.7. *Let $K \in \mathbb{R}^{m\times n}$ and $f \in \mathbb{R}^m$ with $n > m = rank(K)$. Let $\widetilde{\boldsymbol{w}}$ be the solution given by the pseudo-inverse as in equation (3.16). Let $\mathcal{W} = \{\boldsymbol{w} \in \mathbb{R}^n \mid K\boldsymbol{w} = \boldsymbol{f}\}$. Then $\widetilde{\boldsymbol{w}} \in \mathcal{W}$ and $||\widetilde{\boldsymbol{w}}||^2 < ||\boldsymbol{w}||^2$ for all $\boldsymbol{w} \in \mathcal{W}$.*

For the proof of this theorem we refer to [3]. The smallest solution does not have contributions in the null-space of K. In this case, the pseudo-inverse is equal to

$$K^\dagger = K^* \left(KK^*\right)^{-1}.$$

Then $KK^\dagger = I_m$ and so $K^\dagger$ is a right-inverse of K.

3.5 Spectral regularization

Now suppose that the measured data contains noise. We will see how this affects the solution given by the pseudo-inverse. Let $\boldsymbol{f}^\delta = \boldsymbol{f} + \boldsymbol{\delta}$ be the noisy data with $\boldsymbol{\delta} \in \mathbb{R}^m$. Applying the pseudo-inverse to $\boldsymbol{f}^{\boldsymbol{\delta}}$ gives

$$\begin{aligned}
\widetilde{\boldsymbol{w}}^{\boldsymbol{\delta}} &= K^\dagger \boldsymbol{f}^{\boldsymbol{\delta}} \\
&= \sum_{j=1}^{k} \frac{\langle \boldsymbol{f}^{\boldsymbol{\delta}}, \boldsymbol{u_j}\rangle}{\sigma_j} \boldsymbol{v_j} \\
&= \sum_{j=1}^{k} \frac{\langle \boldsymbol{f}, \boldsymbol{u_j}\rangle}{\sigma_j} \boldsymbol{v_j} + \sum_{j=1}^{k} \frac{\langle \delta, \boldsymbol{u_j}\rangle}{\sigma_j} \boldsymbol{v_j} \\
&= \widetilde{\boldsymbol{w}} + \sum_{j=1}^{k} \frac{\langle \boldsymbol{\delta}, \boldsymbol{u_j}\rangle}{\sigma_j} \boldsymbol{v_j}.
\end{aligned}$$

When there is a fast decay of the singular values and the noise components in the directions of the corresponding basis vectors is nonzero, this leads to amplification of the noise. Spectral regularization tries to solve this problem by applying a function to the singular values that prevents them from becoming too small. We define the family of functions $g_\beta : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ that is parameterized by $\beta \geq 0$ and satisfies

$$\lim_{\beta\to 0} g_\beta(\sigma) = \frac{1}{\sigma}. \tag{3.17}$$

A function g_β defines a regularized pseudo-inverse R_β as follows

$$R_\beta \boldsymbol{f} := \sum_{j=1}^{k} g_\beta(\sigma_j) \langle \boldsymbol{f}, \boldsymbol{u_j}\rangle \boldsymbol{v_j}.$$

The function g_β determines the type of regularization and the values of β determines the amount of regularization that is applied. We want to choose g_β such that the solution of the regularized inverse is close to the true solution while being less sensitive to noise in the data than than the unregularized solution. Also, we want that

$$R_\beta \boldsymbol{f} \to K^\dagger \boldsymbol{f} \quad \text{as} \quad \beta \to 0.$$

An example of such a function is the *trucated singular value decomposition* (TSVD)

$$g_\beta(\sigma) = \begin{cases} \frac{1}{\sigma} & \sigma \geq \beta \\ 0 & \sigma < \beta \end{cases}$$

which sets all singular values below β to zero. This leads to the regularized solution

$$\widetilde{\boldsymbol{w}}_{\boldsymbol{\beta}} = R_\beta \boldsymbol{f} = \sum_{\sigma_j \geq \beta} \frac{\langle \boldsymbol{f}, \boldsymbol{u_j} \rangle}{\sigma_j} \boldsymbol{v_j}.$$

We see that indeed condition (3.17) is satisfied.

3.6 Tikhonov regularization

Tikhonov regularization is a type of spectral regularization where g_β is chosen in the following way

$$g_\beta(\sigma) = \frac{\sigma}{\sigma^2 + \beta}.$$

When σ is large compared to β then $g_\beta(\sigma) \approx \frac{1}{\sigma}$ and so the large singular values are almost unaffected. When σ is small compared to β, then $g_\beta(\sigma) \approx \frac{\sigma}{\beta}$ which prevents division by small values. The condition $g_\beta(\sigma) \to \frac{1}{\sigma}$ as $\beta \to 0$ is satisfied. The resulting Tikhonov regularized solution is given by

$$\widetilde{\boldsymbol{w}}_{\boldsymbol{\beta}} = R_\alpha \boldsymbol{f} = \sum_{j=1}^{k} \frac{\sigma_j}{\sigma_j^2 + \beta} \langle \boldsymbol{f}, \boldsymbol{u_j} \rangle \boldsymbol{v_j}. \tag{3.18}$$

In addition to characterizing Tikhonov regularization as a shift of the sigular values of K, the Tikhonov regularized solution can also be interpreted as the solution of a minimization problem.

Lemma 3.8. *Let $K \in \mathbb{R}^{m \times n}$ with singular system $\{(\sigma_j, \boldsymbol{u_j}, \boldsymbol{v_j})\}$ and let $\boldsymbol{f} \in \mathbb{R}^m$. Let $\beta > 0$ and let $\widetilde{\boldsymbol{w}}_{\boldsymbol{\beta}}$ be the regularized solution as defined by equation* (3.18)*. Then $\widetilde{\boldsymbol{w}}_{\boldsymbol{\beta}}$ is given by*

$$\widetilde{\boldsymbol{w}}_{\boldsymbol{\beta}} = \left(K^* K + \beta I\right)^{-1} K^* \boldsymbol{f}. \tag{3.19}$$

Proof.

$$\begin{aligned}
(K^*K+\beta I)\widetilde{\boldsymbol{w}}_{\boldsymbol{\beta}} &= (K^*K+\beta I)\sum_{j=1}^{k}\frac{\sigma_j}{\sigma_j^2+\beta}\langle \boldsymbol{f},\boldsymbol{u_j}\rangle \boldsymbol{v_j} \\
&= \sum_{j=1}^{k}\frac{\sigma_j}{\sigma_j^2+\beta}\langle \boldsymbol{f},\boldsymbol{u_j}\rangle K^*K\boldsymbol{v_j} + \sum_{j=1}^{k}\frac{\beta\sigma_j}{\sigma_j^2+\beta}\langle \boldsymbol{f},\boldsymbol{u_j}\rangle \boldsymbol{v_j} \\
&= \sum_{j=1}^{k}\frac{\sigma_j}{\sigma_j^2+\beta}\langle \boldsymbol{f},\boldsymbol{u_j}\rangle \sigma_j^2\boldsymbol{v_j} + \sum_{j=1}^{k}\frac{\beta\sigma_j}{\sigma_j^2+\beta}\langle \boldsymbol{f},\boldsymbol{u_j}\rangle \boldsymbol{v_j} \\
&= \sum_{j=1}^{k}\frac{\sigma_j(\sigma_j^2+\beta)}{\sigma_j^2+\beta}\langle \boldsymbol{f},\boldsymbol{u_j}\rangle \boldsymbol{v_j} \\
&= \sum_{j=1}^{k}\sigma_j\langle \boldsymbol{f},\boldsymbol{u_j}\rangle \boldsymbol{v_j} \\
&= K^*\boldsymbol{f}.
\end{aligned}$$

The eigenvalues of $(K^*K+\beta I)$ are equal to $\sigma_j^2+\beta$. Therefore, $(K^*K+\beta I)$ is invertible when $\beta>0$ and so equation (3.19) is satisfied. □

Theorem 3.9. *Let $K\in\mathbb{R}^{m\times n}$ with singular system $\{(\sigma_j,\boldsymbol{u_j},\boldsymbol{v_j})\}$ and let $\boldsymbol{f}\in\mathbb{R}^m$. Let $\beta>0$ and let $\widetilde{\boldsymbol{w}}_{\boldsymbol{\beta}}$ be the regularized solution as defined by equation* (3.18)*. Then $\widetilde{\boldsymbol{w}}_{\boldsymbol{\beta}}$ is the unique solution to the minimization problem*

$$\arg\min_{\boldsymbol{w}} ||K\boldsymbol{w}-\boldsymbol{f}||^2+\beta||\boldsymbol{w}||^2. \tag{3.20}$$

Proof. Define $T_\beta(\boldsymbol{w}) := ||K\boldsymbol{w}-\boldsymbol{f}||^2+\beta||\boldsymbol{w}||^2$. Let $\widetilde{\boldsymbol{w}}_{\boldsymbol{min}}$ be a global minimizer of $T_\beta(\boldsymbol{w})$. We show that $\widetilde{\boldsymbol{w}}_{\boldsymbol{min}}$ is equal to $\widetilde{\boldsymbol{w}}_{\boldsymbol{\beta}}$. It then follows that the minimizer is unique.

For $\boldsymbol{w}=\widetilde{\boldsymbol{w}}_{\boldsymbol{min}}+\tau\boldsymbol{v}$ with $\tau>0$ and an arbitrary direction $\boldsymbol{v}$ we have that

$$\begin{aligned}
0 &\le T_\beta(\boldsymbol{w})-T_\beta(\widetilde{\boldsymbol{w}}_{\boldsymbol{min}}) \\
&= \|K\boldsymbol{w}-\boldsymbol{f}\|^2+\beta\|\boldsymbol{w}\|^2-\|K\widetilde{\boldsymbol{w}}_{\boldsymbol{min}}-\boldsymbol{f}\|^2-\beta\|\widetilde{\boldsymbol{w}}_{\boldsymbol{min}}\|^2 \\
&= \|K\boldsymbol{w}\|^2-2\langle K\boldsymbol{w},\boldsymbol{f}\rangle+\beta\|\boldsymbol{w}\|^2-\|K\widetilde{\boldsymbol{w}}_{\boldsymbol{min}}\|^2+2\langle K\widetilde{\boldsymbol{w}}_{\boldsymbol{min}},\boldsymbol{f}\rangle-\beta\|\widetilde{\boldsymbol{w}}_{\boldsymbol{min}}\|^2 \\
&= \|K\widetilde{\boldsymbol{w}}_{\boldsymbol{min}}+\tau K\boldsymbol{v}\|^2-2\langle K\widetilde{\boldsymbol{w}}_{\boldsymbol{min}}+\tau K\boldsymbol{v},\boldsymbol{f}\rangle+\beta\|\widetilde{\boldsymbol{w}}_{\boldsymbol{min}}+\tau\boldsymbol{v}\|^2 \\
&\quad -\|K\widetilde{\boldsymbol{w}}_{\boldsymbol{min}}\|^2+2\langle K\widetilde{\boldsymbol{w}}_{\boldsymbol{min}},\boldsymbol{f}\rangle-\beta\|\widetilde{\boldsymbol{w}}_{\boldsymbol{min}}\|^2 \\
&= 2\tau\langle K\widetilde{\boldsymbol{w}}_{\boldsymbol{min}},K\boldsymbol{v}\rangle+\tau^2\|K\boldsymbol{v}\|^2-2\tau\langle K\boldsymbol{v},\boldsymbol{f}\rangle+2\beta\tau\langle\widetilde{\boldsymbol{w}}_{\boldsymbol{min}},\boldsymbol{v}\rangle+\beta\tau^2\|\boldsymbol{v}\|^2 \\
&= \tau^2\left(\|K\boldsymbol{v}\|^2+\beta\|\boldsymbol{v}\|^2\right)+2\tau\left(\langle K\widetilde{\boldsymbol{w}}_{\boldsymbol{min}},K\boldsymbol{v}\rangle+\beta\langle\widetilde{\boldsymbol{w}}_{\boldsymbol{min}},\boldsymbol{v}\rangle-\langle K\boldsymbol{v},\boldsymbol{f}\rangle\right) \\
&= \tau^2\left(\|K\boldsymbol{v}\|^2+\beta\|\boldsymbol{v}\|^2\right)+2\tau\left(\langle K^*K\widetilde{\boldsymbol{w}}_{\boldsymbol{min}},\boldsymbol{v}\rangle+\beta\langle\widetilde{\boldsymbol{w}}_{\boldsymbol{min}},\boldsymbol{v}\rangle-\langle K^*\boldsymbol{f},\boldsymbol{v}\rangle\right) \\
&= \tau^2\left(\|K\boldsymbol{v}\|^2+\beta\|\boldsymbol{v}\|^2\right)+2\tau\langle(K^*K+\beta I)\widetilde{\boldsymbol{w}}_{\boldsymbol{min}}-K^*\boldsymbol{f},\boldsymbol{v}\rangle.
\end{aligned}$$

When we divide by τ and take the limit $\tau\downarrow 0$ we obtain

$$\langle(K^*K+\beta I)\widetilde{\boldsymbol{w}}_{\boldsymbol{min}}-K^*\boldsymbol{f},\boldsymbol{v}\rangle\ge 0.$$

Because this holds for all $\boldsymbol{v}$ it follows that

$$(K^*K + \beta I)\widetilde{\boldsymbol{w}}_{\boldsymbol{min}} = K^*\boldsymbol{f}.$$

In Lemma 3.8 we showed that this is also true for $\widetilde{\boldsymbol{w}}_{\boldsymbol{\beta}}$ and so $\widetilde{\boldsymbol{w}}_{\boldsymbol{min}} = \widetilde{\boldsymbol{w}}_{\boldsymbol{\beta}}$. □

Let us now look at what this theory means in the specific case that K represents a convolution. In that case, the singular values σ_j correspond to the DFT of the point spread function. In Chapter 2 we saw that the exact solution of deconvolution is given by

$$\boldsymbol{w}^{\boldsymbol{\delta}} = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(\boldsymbol{f}^{\boldsymbol{\delta}})}{\mathcal{F}(\boldsymbol{h})}\right),$$

where dividing by the OTF leads to a solution that does not depend continuously on the data. This method corresponds to the unregularized inverse operator from Equation (3.16). In the case of Tikhonov regularization, we replace $\frac{1}{\sigma_j}$ by $\frac{\sigma_j}{\sigma_j^2+\beta}$ and in that way prevent division by small values in the OTF.

Chapter 4

Scaled gradient projection methods

In the previous chapters we have seen how the inverse problem of deconvolution can be formulated as a minimization problem of the form

$$\begin{cases} \text{minimize } J(\boldsymbol{w}) \\ \text{subject to } \boldsymbol{w} \in \Psi \end{cases} \tag{4.1}$$

where $J : \mathbb{R}^N \to \mathbb{R}$ is a convex function, $\Psi \in \mathbb{R}^N$ is a closed and convex set, $\boldsymbol{w} \in \mathbb{R}^N$ is the object. We looked at objective functions of the form

$$J(\boldsymbol{w}) = D(\boldsymbol{w}) + \beta R(\boldsymbol{w}) \tag{4.2}$$

that consist of a data fidelity term $D(\boldsymbol{w})$ and a regularization term $R(\boldsymbol{w})$ that are scaled with a regularization parameter β. The set Ψ contains all the allowed solutions. In this chapter, we will show how to find the minimizer of this functional. The methods we discuss are applicable to any problem of the form (4.1), of which our problem is a special case.

4.1 Descent direction

We consider iterative algorithms that start at some initial point $\boldsymbol{w}^{(0)}$ and where the consecutive iterations are given by

$$\boldsymbol{w}^{(k+1)} = \boldsymbol{w}^{(k)} + \lambda_k \boldsymbol{d}^{(k)},$$

where λ_k is the step size and $\boldsymbol{d}^{(k)}$ is the search direction. We want to choose λ_k and $\boldsymbol{d}^{(k)}$ such that $J\left(\boldsymbol{w}^{(k+1)}\right) < J\left(\boldsymbol{w}^{(k)}\right)$ and in that way converge towards the minimum of J.

Definition 4.1. A direction $\boldsymbol{d}^{(k)} \in \mathbb{R}^N$ is called a *descent direction* for the function $J : \mathbb{R}^N \to \mathbb{R}$ at $\boldsymbol{w}^{(k)}$ when

$$\left\langle \nabla J\left(\boldsymbol{w}^{(k)}\right), \boldsymbol{d}^{(k)} \right\rangle < 0. \tag{4.3}$$

□

This definition makes sense because J can be approximated near $\boldsymbol{w}^{(k)}$ by the Taylor series

$$J\left(\boldsymbol{w}^{(k+1)}\right) \approx J\left(\boldsymbol{w}^{(k)}\right) + \left\langle \nabla J\left(\boldsymbol{w}^{(k)}\right), \left(\boldsymbol{w}^{(k+1)} - \boldsymbol{w}^{(k)}\right) \right\rangle$$

and therefore condition (4.3) guarantees that $J\left(\boldsymbol{w}^{(k+1)}\right) < J\left(\boldsymbol{w}^{(k)}\right)$ when the step size λ_k is small enough.

In gradient descent algorithms, the direction is chosen as

$$\boldsymbol{d}^{(k)} = -\nabla J(\boldsymbol{w}^{(k)}).$$

This choice is based on the fact that the gradient is perpendicular to the level set and therefore the function decreases fastest in the direction of the negative gradient (Figure 4.1). Indeed, whenever $\nabla J\left(\boldsymbol{w}^{(k)}\right) \neq 0$, then $\boldsymbol{d}^{(k)}$ is a descent direction because

$$\left\langle \nabla J\left(\boldsymbol{w}^{(k)}\right), \boldsymbol{d}^{(k)} \right\rangle = -\left\| \nabla J\left(\boldsymbol{w}^{(k)}\right) \right\| < 0.$$

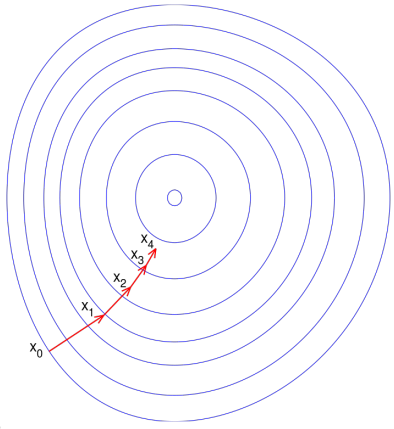


Figure 4.1: Illustration of a gradient descent algorithm on level sets of a function.

4.2 The scaled gradient algorithm

In a scaled gradient algorithm, the idea is to multiply the gradient descent direction of with a scaling matrix D_k in such a way that it improves the convergence. In other words, we want to choose

$$\boldsymbol{d}^{(k)} = -D_k \nabla J\left(\boldsymbol{w}^{(k)}\right).$$

To assure that $\boldsymbol{d}^{(k)}$ is a descent direction, it must hold that

$$0 > \left\langle \nabla J\left(\boldsymbol{w}^{(k)}\right), \boldsymbol{d}^{(k)} \right\rangle = -\left\langle \nabla J\left(\boldsymbol{w}^{(k)}\right), D_k \nabla J\left(\boldsymbol{w}^{(k)}\right) \right\rangle. \tag{4.4}$$

This is satisfied when D_k is a positive definite matrix. In a scaled gradient projection method, in addition to the scaling of the search direction, there is also a projection involved to assure that the solution is in Ψ.

Scaled gradient projection methods applied to image deconvolution are studied in [7, 26–28, 36]. We will now explain the basic algorithm that is summarized in Algorithm 1 and based on [7]. After that, we will discuss different choices that can be made. Each iteration consists of two parts. In

the first part, a search direction $\boldsymbol{d}^{(k)}$ is found. We will prove that the search direction is a descent direction. In the second part, a line search is performed along the search direction to find a point where the objective function decreases sufficiently. We refer to [7] for the full proof that the algorithm converges to a stationary point of J over Ψ.

Algorithm 1: Scaled gradient projection method

Initialization: choose a starting point $\boldsymbol{w}^{(0)} \in \Psi$ and set the parameters $\gamma, \theta \in (0,1)$ and $0 < \alpha_{min} < \alpha_{max}$.

for *k = 0, 1, 2, ...* **do**

STEP 1. Choose the step size $\alpha_k \in [\alpha_{min}, \alpha_{max}]$ and the scaling matrix D_k

STEP 2. Projection: $\boldsymbol{y}^{(k)} = P_{\Psi, D_k^{-1}}\left(\boldsymbol{w}^{(k)} - \alpha_k D_k \nabla J\left(\boldsymbol{w}^{(k)}\right)\right)$.

STEP 3. Search direction: $\boldsymbol{d}^{(k)} = \boldsymbol{y}^{(k)} - \boldsymbol{w}^{(k)}$

STEP 4. Line-search: select a step size λ_k

STEP 5. Set $\boldsymbol{w}^{(k+1)} = \boldsymbol{w}^{(k)} + \lambda_k \boldsymbol{d}^{(k)}$

end

Let $\Psi \in \mathbb{R}^N$ be a closed and convex set and let $D \in \mathbb{R}^{N \times N}$ be a symmetric positive definite matrix. We define the norm $||.||_D$ as

$$||\boldsymbol{w}||_D := \sqrt{\langle \boldsymbol{w}, D\boldsymbol{w} \rangle}. \tag{4.5}$$

The projection operator $P_{\Psi,D}$ is defined as

$$P_{\Psi,D}(\boldsymbol{w}) := \arg\min_{\boldsymbol{v} \in \Psi} ||\boldsymbol{v} - \boldsymbol{w}||_D. \tag{4.6}$$

For later use, we rewrite the projection operator in another form. By using equation (4.5) and the fact the D is symmetric we can write

$$\begin{aligned} ||\boldsymbol{v} - \boldsymbol{w}||_D^2 &= \langle (\boldsymbol{v} - \boldsymbol{w}), D(\boldsymbol{v} - \boldsymbol{w}) \rangle \\ &= \langle \boldsymbol{v}, D\boldsymbol{v} \rangle - \langle \boldsymbol{w}, D\boldsymbol{v} \rangle - \langle \boldsymbol{v}, D\boldsymbol{w} \rangle + \langle \boldsymbol{w}, D\boldsymbol{w} \rangle \\ &= \langle \boldsymbol{v}, D\boldsymbol{v} \rangle - 2 \langle \boldsymbol{v}, D\boldsymbol{w} \rangle + \langle \boldsymbol{w}, D\boldsymbol{w} \rangle . \end{aligned}$$

Therefore, we can write

$$P_{\Psi,D}(\boldsymbol{w}) = \arg\min_{\boldsymbol{v} \in \Psi} \left(\phi(\boldsymbol{v}) := \frac{1}{2} \langle \boldsymbol{v}, D\boldsymbol{v} \rangle - \langle \boldsymbol{v}, D\boldsymbol{w} \rangle \right). \tag{4.7}$$

Now assume that in iteration k we have some symmetric and positive definite scaling matrix $D_k \in \mathbb{R}^{N \times N}$, and we have some step size α_k. In section 4.3 we will explain how to choose D_k and α_k. We define $\boldsymbol{y}^{(k)}$ as

$$\boldsymbol{y}^{(k)} := P_{\Psi, D_k^{-1}}\left(\boldsymbol{w}^{(k)} - \alpha_k D_k \nabla J\left(\boldsymbol{w}^{(k)}\right)\right). \tag{4.8}$$

The search direction is then defined as

$$\boldsymbol{d}^{(k)} := \boldsymbol{y}^{(k)} - \boldsymbol{w}^{(k)} \tag{4.9}$$

as shown in Figure 4.2.

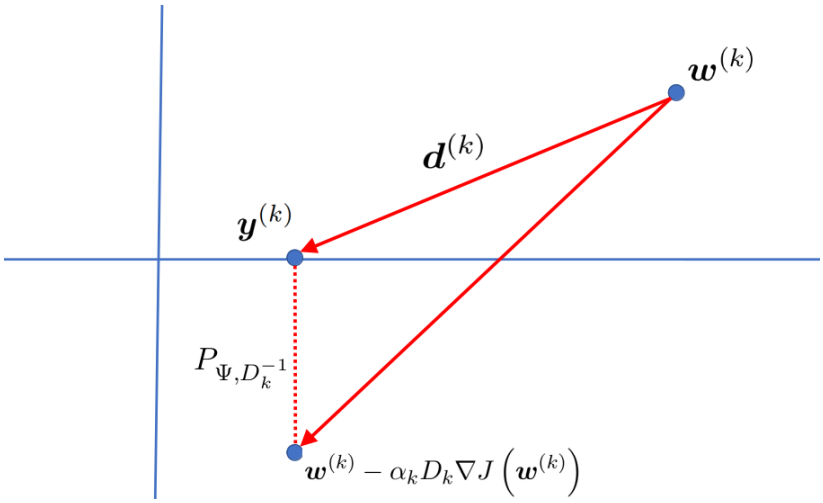


Figure 4.2: Schematic image of the descent direction in two dimensions where Ψ is the first quadrant.

For a constrained minimization problem of the form 4.1, a stationary point of J over Ψ is a point $\boldsymbol{w}^* \in \Psi$ for which

$$\langle \nabla J\left(\boldsymbol{w}^*\right), (\boldsymbol{v} - \boldsymbol{w}^*)\rangle \geq 0 \tag{4.10}$$

for all $\boldsymbol{v} \in \Psi$. This means that when moving a small distance from $\boldsymbol{w}^*$ to any point $\boldsymbol{v}$ in Ψ, the value of J does not decrease, which means that there is no feasible descent direction at $\boldsymbol{w}^*$. We will now show that the search direction is a descent direction, whenever D_k is symmetric and positive definite. The proof is based on [7].

Theorem 4.2. *Let $J : \mathbb{R}^N \to \mathbb{R}$ be a convex funtion, Ψ a closed and convex set and D_k a symmetric positive definite matrix. Define the search direction $\boldsymbol{d}^{(k)}$ as in equations* (4.8) *and* (4.9)*. Then $\boldsymbol{d}^{(k)}$ is a descent direction for J at $\boldsymbol{w}^{(k)}$.*

Proof. We are going to show that $\left\langle \nabla J\left(\boldsymbol{w}^{(k)}\right), \boldsymbol{d}^{(k)}\right\rangle < 0$. We start from equation (4.7) for the projection operator. Since $\phi(\boldsymbol{v})$ is a strictly convex function, $\boldsymbol{v}$ minimizes ϕ over Ψ whenever $\boldsymbol{v}$ is a stationary point of ϕ. The gradient of ϕ is given by

$$\nabla\phi(\boldsymbol{v}) = D(\boldsymbol{v} - \boldsymbol{w}).$$

Therefore, it follows from (4.10) for ϕ at the stationary point $P_{\Psi,D}(\boldsymbol{w})$ that for all $\boldsymbol{v} \in \Psi$

$$\begin{aligned} 0 &\geq \left\langle \nabla\phi\left(P_{\Psi,D}(\boldsymbol{w})\right), P_{\Psi,D}(\boldsymbol{w}) - \boldsymbol{v}\right\rangle \\ &= \left\langle D(P_{\Psi,D}(\boldsymbol{w}) - \boldsymbol{w}), (P_{\Psi,D}(\boldsymbol{w}) - \boldsymbol{v})\right\rangle. \end{aligned} \tag{4.11}$$

Now, by choosing $\boldsymbol{w} = \boldsymbol{w}^{(k)} - \alpha_k D_k \nabla J\left(\boldsymbol{w}^{(k)}\right)$, $D = D_k^{-1}$, and $\boldsymbol{v} = \boldsymbol{w}^{(k)}$ we find

$$\begin{aligned} 0 &\geq \left\langle D_k^{-1}\left(\boldsymbol{y}^{(k)} - \boldsymbol{w}^{(k)} + \alpha_k D_k \nabla J\left(\boldsymbol{w}^{(k)}\right)\right), \left(\boldsymbol{y}^{(k)} - \boldsymbol{w}^{(k)}\right)\right\rangle \\ &= \left\langle D_k^{-1}\left(\boldsymbol{d}^{(k)} + \alpha_k D_k \nabla J\left(\boldsymbol{w}^{(k)}\right)\right), \boldsymbol{d}^{(k)}\right\rangle \\ &= \left\langle D_k^{-1}\boldsymbol{d}^{(k)}, \boldsymbol{d}^{(k)}\right\rangle + \alpha_k \left\langle \nabla J\left(\boldsymbol{w}^{(k)}\right), \boldsymbol{d}^{(k)}\right\rangle. \end{aligned}$$

Then, because D_k^{-1} is symmetric and positive definite, it follows that

$$\left\langle \nabla J\left(\boldsymbol{w}^{(k)}\right), \boldsymbol{d}^{(k)}\right\rangle \leq -\frac{\left\langle \boldsymbol{d}^{(k)}, D_k^{-1}\boldsymbol{d}^{(k)}\right\rangle}{\alpha_k} < 0.$$

□

After finding a descent direction $\boldsymbol{d}^{(k)}$, the second part of the iteration consists of finding a point in this direction that will become the next estimate

$$\boldsymbol{w}^{(k+1)} = \boldsymbol{w}^{(k)} + \lambda_k \boldsymbol{d}^{(k)}.$$

The step size λ_k is selected by some line search method. We want to select choose λ_k in such a way that the decrease in the objective function is large while also taking into account computation time. Ideally, we want that λ_k that satisfies

$$\arg \min_{\lambda_k > 0} \; J\left(\boldsymbol{w}^{(k)} + \lambda_k \boldsymbol{d}^{(k)}\right).$$

However, finding a minimizer is computationally costly because it requires many evaluations of J and possibly ∇J [24].

Note that we refer to both α_k and λ_k as a step size. Figure 4.2 shows the difference between the two. The value α_k is the step size in the direction $-D_k \nabla J\left(\boldsymbol{w}^{(k)}\right)$. After projection, we find a search direction $\boldsymbol{d}^{(k)}$. Then, λ_k refers to the step size in the direction of $\boldsymbol{d}^{(k)}$. For both α_k and λ_k there should be a balance between finding a new descent direction and optimizing the descent in a given direction. When finding a descent direction is costly, it is worth spending some time on finding a good step size in this direction.

So far, we defined the basic scaled gradient algorithm. In the following two sections, we will look at how the scaling matrix, steps size and line search method can be chosen.

4.3 Scaling matrix and step size

In this section we discuss how to choose the scaling matrix D_k and the step size α_k. In literature, the scaling matrix is often chosen as a diagonal matrix

$$D_k = \operatorname{diag}\left(d_1^{(k)}, d_2^{(k)}, ..., d_N^{(k)}\right)$$

to reduce computational costs. One possibility is the quasi-Newton approach to choose for D_k an approximation of the inverse of the Hessian matrix $\nabla^2 J\left(\boldsymbol{w}^{(k)}\right)$. In [7, 36] they choose

$$d_i^{(k)} = \min\left\{L_2, \max\left\{L_1, \left(\frac{\partial^2 J\left(\boldsymbol{w}^{(k)}\right)}{\left(\partial w_i\right)^2}\right)^{-1}\right\}\right\} \tag{4.12}$$

for $i = 1, ..., N$ and where $L_1, L_2 > 0$ are used to bound the eigenvalues of D_k from above and below. This choice for d_k can me motivated by the fact that the second derviative gives information about the curvature. When the error landscape is more curved in some direction, the first order approximation by the gradient is correct over a shorter distance. Therefore, we want to decrease our step in this direction which is achieved by scaling the corresponding component of the descent direction by the inverse of the second derivative. Another approach proposed by [7] is

$$d_i^{(k)} = \min\left\{L_2, \max\left\{L_1, w_i^{(k)}\right\}\right\} \tag{4.13}$$

for $i = 1, ..., N$. This choice is motivated by the fact that, as shown in the Supplementary Information of [36], the iteration step of the Richardson-Lucy algorithm [10, 22, 29, 33] can be written as

$$\boldsymbol{w}^{(k+1)} = \boldsymbol{w}^{(k)} - \text{diag}\left(w_i^{(k)}\right) \nabla J(\boldsymbol{w}^{(k)}),$$

where $J(\boldsymbol{w})$ is the Kullback-Leibler divergence as in (2.5). So with this choice of D_k, the descent direction corresponds to that of the Richardson-Lucy algorithm, corrected with a threshold to make sure that $d_i^{(k)} \in [L_1, L_2]$. Note that this only makes sense if we choose as objective function the Kullback-Leibler divergence.

A third choice for D_k is proposed in [36] where a split-gradient method is used. This method is suitable for the case that the objective function is of the form (4.2) where $D(\boldsymbol{w})$ is the Kullback-Leibler divergence as in equation (2.5) and $R(\boldsymbol{w})$ is some regularization function. We use two non-negative vectors $U(\boldsymbol{w})$ and $V(\boldsymbol{w})$ such that

$$\nabla R(\boldsymbol{w}) = V(\boldsymbol{w}) - U(\boldsymbol{w}).$$

The diagonal components of D_k are now defined as

$$d_i^{(k)} = \min\left\{L_1, \max\left\{L_2, \frac{w_i^{(k)}}{1 + \beta\left(V\left(\boldsymbol{w}^{(k)}\right)\right)_i}\right\}\right\}. \tag{4.14}$$

This choice for the scaling matrix can be motivated as follows. Compared to the previous method, the value of $w_i^{(k)}$ is divided by $1 + \beta\left(V\left(\boldsymbol{w}^{(k)}\right)\right)_i$. For a given voxel i, the value of $V(\boldsymbol{w})_i$ is positive whenever the derivative of the regularization with respect to this voxel is positive. This means that increasing the intensity of this voxel would lead to an increase of the regularization. By dividing by $1 + \beta\left(V\left(\boldsymbol{w}^{(k)}\right)\right)_i$, the step in the direction of this voxel is decreased. Therefore, this matrix can be interpreted as a matrix that speeds up the convergence toward a more regularized solution.

The Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm [9, 14, 15, 32] is Quasi-Newton method that can be seen as a scaled gradient method where the scaling matrix is an approximation of the inverse of the Hessian. Unlike before, this scaling matrix is not a diagonal matrix. The scaling matrix is updated at the end of each iteration step as follows

$$D_{k+1} = \left(I - \frac{\boldsymbol{s}_k \boldsymbol{z}_k^T}{\langle \boldsymbol{z}_k, \boldsymbol{s}_k \rangle}\right) D_k \left(I - \frac{\boldsymbol{z}_k \boldsymbol{s}_k^T}{\langle \boldsymbol{z}_k, \boldsymbol{s}_k \rangle}\right) + \frac{\boldsymbol{s}_k \boldsymbol{s}_k^T}{\langle \boldsymbol{z}_k, \boldsymbol{s}_k \rangle},$$

where $\boldsymbol{s}_k = -\alpha_k D_k \nabla J\left(\boldsymbol{w}^{(k)}\right)$ and $\boldsymbol{z}_k = \nabla J\left(\boldsymbol{w}^{(k+1)}\right) - \nabla J\left(\boldsymbol{w}^{(k)}\right)$. Here we have to multiply $N \times N$ matrices which is costly. However, limited memory versions of the BFGS method are described by [23] and [31].

In the specific case that D is a diagonal matrix with non-negative elements d_i for $i = 1, ..., N$ and Ψ is the set of non-negative solutions it follows from equation (4.11) that

$$\begin{aligned} 0 &\geq \sum_{i=1}^{N} \left(P_{\Psi,D}(\boldsymbol{w})_i - w_i\right) d_i \left(P_{\Psi,D}(\boldsymbol{w})_i - v_i\right) \\ &= \sum_{i=1}^{N} d_i \left(P_{\Psi,D}(\boldsymbol{w})_i^2 - v_i P_{\Psi,D}(\boldsymbol{w})_i - w_i P_{\Psi,D}(\boldsymbol{w})_i + w_i v_i\right) \end{aligned}$$

for all $v \in \Psi$. This inequality is satisfied when

$$P_{\Psi,D}(\boldsymbol{w})_i = \begin{cases} w_i & \text{if} \quad w_i \geq 0 \\ 0 & \text{if} \quad w_i < 0 \end{cases}$$

and therefore the projection reduces to setting the negative elements to zero.

We will now discuss a method for selecting the step size α_k. We want to choose the step size such that the decrease of the functional in the direction $-D_k \nabla J\left(\boldsymbol{w}^{(k)}\right)$ is maximized. The Barzilai-Borwein rules [2] are rules for choosing the steplength in a gradient descent algorithm that are based on a quasi-Newton approach. We want to find a step size α such that αI approximates the inverse Hessian. The reasoning behind this can be understood by using the secant method. This is a finite difference approximation to Newton's method where the root of a function $f : \mathbb{R} \to \mathbb{R}$ is found by using the iteration

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f\left(x_k\right) - f\left(x_{k-1}\right)} f\left(x_k\right). \tag{4.15}$$

which is visualized in Figure 4.3.

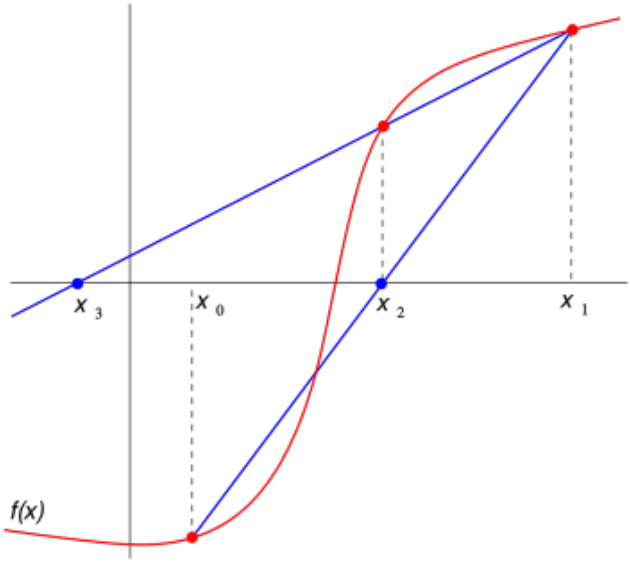


Figure 4.3: Finding the root of a function $f(x)$ using the secant method. We start with x_0 and x_1 and then find x_2 and x_3 using Equation (4.15).

The standard gradient descent method uses the iteration

$$\boldsymbol{w}^{k+1} = \boldsymbol{w}^k - \alpha_k \nabla J(\boldsymbol{w}^k).$$

The idea is to choose α_k such that the gradient of J is zero at $\boldsymbol{w}^{(k+1)}$. If the gradient would be larger than zero, we could take a larger step in that direction. Translating the secant method to multiple dimensions and taking ∇J for f we want choose α_k such that

$$\boldsymbol{w}^{(k)} - \boldsymbol{w}^{(k-1)} = \alpha_k \left(\nabla J\left(\boldsymbol{w}^{(k)}\right) - \nabla J\left(\boldsymbol{w}^{(k-1)}\right)\right).$$

Define $\boldsymbol{s}_k = \boldsymbol{w}^{(k)} - \boldsymbol{w}^{(k-1)}$ and $z_k = \nabla J\left(\boldsymbol{w}^{(k)}\right) - \nabla J\left(\boldsymbol{w}^{(k-1)}\right)$. Then α_k is chosen as

$$\arg\min_{\alpha} \|\alpha \boldsymbol{s}_k - \boldsymbol{z}_k\|^2 \qquad \text{or} \qquad \arg\min_{\alpha} \|\boldsymbol{s}_k - \alpha \boldsymbol{z}_k\|^2.$$

Taking the derivative, setting it to zero gives and solving for α gives

$$\alpha_k^{(1)} = \frac{\langle \boldsymbol{s}_k, \boldsymbol{s}_k \rangle}{\langle \boldsymbol{s}_k, \boldsymbol{z}_k \rangle} \qquad \text{and} \qquad \alpha_k^{(2)} = \frac{\langle \boldsymbol{s}_k, \boldsymbol{z}_k \rangle}{\langle \boldsymbol{z}_k, \boldsymbol{z}_k \rangle}. \tag{4.16}$$

These are known as the Barzilai-Borwein (BB) rules. In [7, 36] the BB rules are adapted to the case of a scaled gradient descent direction such that $\alpha_k D_k$ approximates $\left(\nabla^2 J\left(\boldsymbol{w}^{(k)}\right)\right)^{-1}$. For that reason α_k is chosen such that it satisfies

$$\arg\min_{\alpha} \| \left(\alpha_k D_k\right)^{-1} \boldsymbol{s}_k - \boldsymbol{z}_k\|^2 \qquad \text{or} \qquad \arg\min_{\alpha} \|\boldsymbol{s}_k - \alpha_k D_k \boldsymbol{z}_k\|^2$$

from which it follows in an equivalent way that

$$\alpha_k^{(1)} = \frac{\left\langle \boldsymbol{s}_k, D_k^{-1} D_k^{-1} \boldsymbol{s}_k \right\rangle}{\left\langle \boldsymbol{s}_k, D_k^{-1} \boldsymbol{z}_k \right\rangle} \qquad \text{and} \qquad \alpha_k^{(2)} = \frac{\langle \boldsymbol{s}_k, D_k \boldsymbol{z}_k \rangle}{\langle \boldsymbol{z}_k, D_k D_k \boldsymbol{z}_k \rangle}. \tag{4.17}$$

When $D_k = I$, the step sizes reduce to (4.16) and therefore this approach generalizes the standard BB-rules.

4.4 Line search

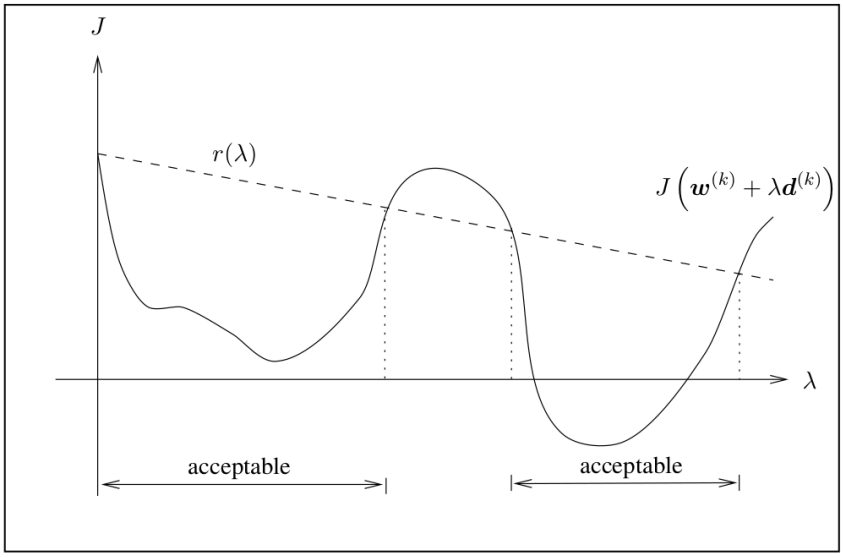


Figure 4.4: The sufficient decrease condition. The function $J\left(\boldsymbol{w}^{(k)} + \lambda \boldsymbol{d}^{(k)}\right)$ is plotted as a function of λ. The dashed line $r(\lambda)$ represents the right hand side of equation (4.18) that is a linear function of λ. Figure adapted from [24] p.33.

In this section we discuss a line search method. When using a step size α_k that is determined with a Quasi-Newton method as just discussed, we already have a good step length to start from. In that

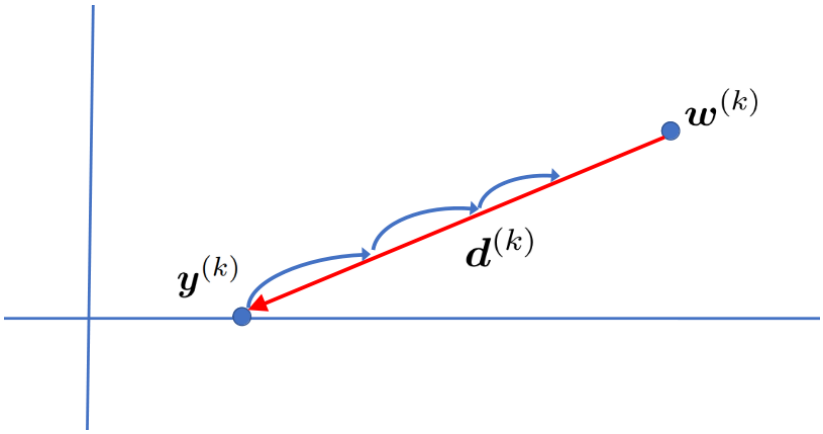


Figure 4.5: Line search.

case it suffices to use a simple backtracking method to determine λ_k. In the backtracking method, a value $\theta \in (0,1)$ is chosen which is the factor by which the step size will be reduced each time. Also, a value $\gamma \in (0,1)$ is selected that determines the minimal decrease in the objective function that has to be achieved. We say that the step size λ_k gives sufficent decrease when

$$J\left(\boldsymbol{w}^{(k)} + \lambda_k \boldsymbol{d}^{(k)}\right) \leq J\left(\boldsymbol{w}^{(k)}\right) + \gamma\lambda_k \left\langle \nabla J\left(\boldsymbol{w}^{(k)}\right), \boldsymbol{d}^{(k)} \right\rangle. \tag{4.18}$$

That is, $J\left(\boldsymbol{w}^{(k)} + \lambda_k \boldsymbol{d}^{(k)}\right)$ is as small as it is expected to become based on a fraction γ of the gradient at $\boldsymbol{w}^{(k)}$ and a step length λ_k in the direction $\boldsymbol{d}^{(k)}$. This is shown in Figure 4.4 where $r(\lambda)$ is the right hand side of (4.18). We start with $\lambda_k = 1$ and stop as soon as (4.18) is satisfied. If not, λ_k is decreased by a factor θ, as depicted in Figure 4.5. Since $\gamma \in (0,1)$, the sufficient decrease condition is guaranteed to be true when λ_k is small enough.

Chapter 5

Numerical experiments

This chapter is not accessible.

Chapter 6

Conclusion and future work

This chapter is not accessible.

Appendix A

Derivatives

This appendix contains the gradient of the different components of the objective function that are used. Also, we derive formulas for the diagonal of the Hessian that is used for one of the scaling matrices.
The objective function has the form

$$J(\boldsymbol{w}) = D(\boldsymbol{w}, \boldsymbol{f}) + \beta \cdot R(\boldsymbol{w}).$$

In order to express the objective functions correctly we define the mask M_S as the square, diagonal matrix that is 1 for the indices that belong to S and 0 for the indices outside of S. Multiplication of M_S with a vector $\boldsymbol{v}$ will set all elements of $\boldsymbol{v}$ outside S to zero.

Calculating $K^T\boldsymbol{w}$
In the calculation of the derivatives we will encounter multiplication of a vector with the transpose of the blurring matrix. To calculate $K^T\boldsymbol{w}$ we first realize that K is structured as in equation 3.11 where the PSF appears in the reverse order on the last row and each row is shifted towards the right by one position compared to the row above it. Because $K_{m,n} = h_{m-n}$ and the periodicity of $\boldsymbol{h}$, the last row of K^T is given by

$$\left(K^T\right)_{N,n} = K_{n,N} = h_{n-N} = h_{N+n-N} = h_{N-(N-n)} = K_{N,N-n}.$$

Therefore, $K^T\boldsymbol{w}$ represents convolution of $\boldsymbol{w}$ with the reversed PSF. That is, when h_m is the m-th elements of the PSF $\boldsymbol{h}$, then $K^T\boldsymbol{w}$ represents a convolution with the PSF $\boldsymbol{h'}$ for which $h'_m = h_{N-m}$. To calculate this convolution using Fourier transforms, we use the property that reversal of space corresponds to reversing the frequency

$$\mathcal{F}\left(\{h_{N-n}\}\right)_m = \mathcal{F}\left(\{h_n\}\right)_{N-m}.$$

In addition to this, when $\boldsymbol{h}$ is real, then the reversal of the frequencies corresponds to complex conjugation

$$\mathcal{F}\left(h_n\right)_{N-m} = \overline{\mathcal{F}\left(h_n\right)_m}.$$

Then it follows that

$$h'_m = \overline{\mathcal{F}\left(h_n\right)_m},$$

and so transposition of the blurring matrix corresponds to complex conjugation of the OTF.

Residual sum of squares
In this case, the data fidelity function is given by

$$D(\boldsymbol{w}, \boldsymbol{f}) = ||M_S(K\boldsymbol{w} + \boldsymbol{b} - \boldsymbol{f})||^2.$$

The gradient is given by

$$\nabla_{\boldsymbol{w}} D(\boldsymbol{w}, \boldsymbol{f}) = 2K^T M_S(K\boldsymbol{w} + \boldsymbol{b} - \boldsymbol{f})$$

and the Hessian is then equal to

$$2K^T M_S K.$$

The diagonal elements are given by

$$\left(2K^T M_S K\right)_{ii} = \sum_{l=1}^{N} \sum_{m=1}^{N} K_{mi}(M_S)_{ml} K_{li}.$$

Since M_S is zero when $m \neq l$ we can leave out of one of the summations

$$\left(2K^T M_S K\right)_{ii} = \sum_{m=1}^{N} (K_{mi})^2 (M_S)_{mm}, \tag{A.1}$$

and so the i-th element of the diagonal of the Hessian is given by the sum of the squares of the elements of the i-th column of the blurring matrix where we only count the indices for which the mask is nonzero.

KL-divergence
In this case, the data fidelity is given by

$$D(\boldsymbol{w}, \boldsymbol{f}) = \sum_{i=1}^{N} (M_S \left(K\boldsymbol{w} + \boldsymbol{b} - \boldsymbol{f}\right))_i - M_S \left(\log\left(\frac{K\boldsymbol{w} + \boldsymbol{b}}{\boldsymbol{f}}\right)\right)_i$$

where the division inside the logarithm is element-wise. The gradient is given by

$$\nabla_{\boldsymbol{w}} D(\boldsymbol{w}, \boldsymbol{f}) = K^T M_S \left(1 - \frac{\boldsymbol{f}}{K\boldsymbol{w} + \boldsymbol{b}}\right).$$

The ii-th element of the Hessian is given by

$$\begin{aligned}
\frac{\partial^2}{\partial w_i^2} D(\boldsymbol{w}, \boldsymbol{f}) &= \frac{\partial}{\partial w_i} \left(\nabla_{\boldsymbol{w}} D(\boldsymbol{w}, \boldsymbol{f})\right)_i \\
&= \left(\frac{\partial}{\partial w_i} K^T M_S \left(1 - \frac{\boldsymbol{f}}{K\boldsymbol{w} + \boldsymbol{b}}\right)\right)_i \\
&= \left(-K^T M_S \frac{\partial}{\partial w_i} \left(\frac{\boldsymbol{f}}{K\boldsymbol{w} + \boldsymbol{b}}\right)\right)_i \\
&= \left(-K^T M_S \frac{\partial}{\partial w_i} \left(\frac{f_1}{K_{11}w_1 + ... + K_{1N}w_N + b_1}, \ ... \ , \ \frac{f_N}{K_{N1}w_1 + ... + K_{NN}w_N + b_n}\right)^T\right)_i \\
&= \left(-K^T M_S \left(\frac{-K_{1i}f_1}{(K\boldsymbol{w} + \boldsymbol{b})_1^2}, \ ... \ , \ \frac{-K_{Ni}f_N}{(K\boldsymbol{w} + \boldsymbol{b})_N^2}\right)^T\right)_i \\
&= \sum_{j=1}^{N} \left((K^T M_S)_{ij} \left(\frac{K_{ji}f_j}{(K\boldsymbol{w} + \boldsymbol{b})_j^2}\right)\right) \\
&= \sum_{j=1}^{N} \left(\sum_{l=1}^{N} K_{li}(M_S)_{lj} \left(\frac{K_{ji}f_j}{(K\boldsymbol{w} + \boldsymbol{b})_j^2}\right)\right) \\
&= \sum_{j=1}^{N} \left(K_{ji}(M_S)_{jj} \left(\frac{K_{ji}f_j}{(K\boldsymbol{w} + \boldsymbol{b})_j^2}\right)\right) \\
&= \sum_{j=1}^{N} \left((K_{ji})^2(M_S)_{jj} \left(\frac{f_j}{(K\boldsymbol{w} + \boldsymbol{b})_j^2}\right)\right).
\end{aligned}$$

Therefore, the diagonal of the Hessian is calculated by applying the mask to the image $\frac{\boldsymbol{f}}{(K\boldsymbol{w}+\boldsymbol{b})^2}$ and then convolving it with the transpose of the element-wise squared PSF.

Tikhonov regularization
In this case, the regularization is given by

$$R(\boldsymbol{w}) = \frac{1}{2}||M_S\boldsymbol{w}||^2.$$

Then, the gradient is given by

$$\begin{aligned}
\nabla R(\boldsymbol{w}) &= M_S^T M_S \boldsymbol{w} \\
&= M_S \boldsymbol{w}.
\end{aligned}$$

The Hessian is given by M_S and therefore the diagonal of the image is 1 inside S and 0 outside of S.

Bibliography

[1] Mongi A. Abidi, Andrei V. Gribok, and Joonki Paik. *Optimization Techniques in Computer Vision*. Springer, 2016. ISBN: 978-3-319-46363-6. DOI: 10.1007/978-3-319-46364-3.

[2] J. Barzilai and J. M. Borwein. "Two-Point Step Size Gradient Methods". In: *IMA Journal of Numerical Analysis* 8.1 (Jan. 1988), pp. 141–148. ISSN: 0272-4979. DOI: 10.1093/imanum/8.1.141.

[3] Martin Benning and Matthias J. Ehrhardt. *Inverse problems in Imaging*. Lecture Notes. Nov. 2016.

[4] M. Bertero and P. Boccacci. *Introduction to Inverse Problems in Imaging*. IOP Publishing Ltd, 1998. ISBN: 0-7503-0435-9.

[5] M Bertero et al. "A discrepancy principle for Poisson data". In: *Inverse Problems* 26.10 (Aug. 2010), p. 105004. DOI: 10.1088/0266-5611/26/10/105004.

[6] Eric Betzig et al. "Imaging Intracellular Fluorescent Proteins at Nanometer Resolution". In: *Science* 313.5793 (2006), pp. 1642–1645. ISSN: 0036-8075. DOI: 10.1126/science.1127344.

[7] S. Bonettini, R. Zanella, and L. Zanni. "A Scaled Gradient Projection Method for Constrained Image Deblurring". In: *Inverse Problems* 25.1 (2008), pp. 1–28.

[8] M. Booth and T. Wilson. "Refractive-index-mismatch induced aberrations in single-photon and two-photon microscopy and the use of aberration correction." In: *Journal of biomedical optics* 6.3 (July 2001), pp. 266–272.

[9] C.G. Broyden. "The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations". In: *IMA Journal of Applied Mathematics* 6.1 (Mar. 1970), pp. 76–90. ISSN: 0272-4960. DOI: 10.1093/imamat/6.1.76.

[10] A.P. Dempster, N.M. Laird, and D.B. Rubin. "Maximum Likelihood from Incomplete Data via the EM Algorithm". In: *Journal of the Royal Statistical Society* 39.1 (1977), pp. 1–38.

[11] Remko Dijksta. "Design and realization of a CW-STED super-resolution microscope setup". Master's thesis. University of Twente, Oct. 2012.

[12] J. Enderlein. "4.09 - Advanced Fluorescence Microscopy". In: *Comprehensive Biomedical Physics*. Ed. by Anders Brahme. Oxford: Elsevier, 2014, pp. 111–151. ISBN: 978-0-444-53633-4. DOI: https://doi.org/10.1016/B978-0-444-53632-7.00409-3.

[13] Richard J. Evans. "Microscopy". In: *Encyclopedia of Physical Science and Technology*. Ed. by Robert A. Meyers. Third Edition. New York: Academic Press, 2003, pp. 765–775. ISBN: 978-0-12-227410-7. DOI: https://doi.org/10.1016/B0-12-227410-5/00444-0.

[14] R. Fletcher. "A new approach to variable metric algorithms". In: *The Computer Journal* 13.3 (Jan. 1970), pp. 317–322. ISSN: 0010-4620. DOI: 10.1093/comjnl/13.3.317.

[15] D. Goldfarb. "A Family of Variable-Metric Methods Derived by Variational Means". In: *Mathematics of Computation* 24 (1970), pp. 23–26. DOI: https://doi.org/10.1090/S0025-5718-1970-0258249-6.

[16] Jacques Hadamard. "Sur les problémes aux dérivées partielles et leur signification physique". In: *Princeton University Bulletin* 13 (1902), pp. 49–52.

[17] Stefan W. Hell and Jan Wichmann. "Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy". In: *Opt. Lett.* 19.11 (June 1994), pp. 780–782. DOI: 10.1364/OL.19.000780.

[18] *Huygens Imaging Academy*. Scientific Volume Imaging. URL: https://svi.nl/Huygens-Imaging-Academy (visited on 12/20/2020).

[19] Geert M.P. van Kempen. "Image Restoration in Fluorescence Microscopy". PhD thesis. Delft University of Technology, 1999.

[20] Tristan van Leeuwen and Christoph Brune. *10 Lectures on Inverse Problems and Imaging*. Nov. 5, 2020. URL: https://tristanvanleeuwen.github.io/IP_and_Im_Lectures/intro.html (visited on 12/27/2020).

[21] Jeff W Lichtman and José-Angel Conchello. "Fluorescence microscopy". In: *Nature Methods* 2.12 (Dec. 2005), pp. 910–919. DOI: 10.1038/nmeth817.

[22] L. B. Lucy. "An Iterative Technique for the Rectification of Observed Distributions". In: *The astronomical journal* 79.6 (1974), pp. 745–754.

[23] D.C. Lui and J. Nocedal. "On the limited memory BFGS method for large scale optimization". In: *Mathematical Programming* 45 (Aug. 1989), pp. 503–528. DOI: 10.1007/BF01589116.

[24] Jorge Nocedal and Stephen Wright. *Numerical Optimization*. 2nd ed. Springer Series in Operations Research and Financial Engineering. location: Springer, 2006. 664 pp. ISBN: 9780387303031.

[25] *Optical Microscopy Primer. Introduction to Optical Microscopy, Digital Imaging, and Photomicrography*. URL: https://micro.magnet.fsu.edu/primer/index.html (visited on 12/20/2020).

[26] F. Porta, M. Prato, and L. Zanni. "A New Steplength Selection for Scaled Gradient Methods with Application to Image Deblurring". In: *Journal of Scientific Computing* 65.1 (2015), pp. 895–919. DOI: 10.1007/s10915-015-9991-9.

[27] F. Porta et al. "Limited-memory scaled gradient projection methods for real-time image deconvolution in microscopy". In: *Communications in Nonlinear Science and Numerical Simulation* 21.1 (2015). Numerical Computations: Theory and Algorithms (NUMTA 2013), International Conference and Summer School, pp. 112–127. ISSN: 1007-5704. DOI: 10.1016/j.cnsns.2014.08.035.

[28] Marco Prato et al. "The scaled gradient projection method: an application to nonconvex optimization". In: *PIERS 2015 Progress in Electromagnetics Research Symposium*. The Electromagnetics Academy. 2015, pp. 2332–2336.

[29] W. H. Richardson. "Bayesian-Based Iterative Method of Image Restoration". In: *Journal of the optical society of America* 62.1 (1972), pp. 55–59.

[30] Michael J. Rust, Mark Bates, and Xiaowei Zhuang. "Stochastic optical reconstruction microscopy (STORM) provides sub-diffraction-limit image resolution". In: *Nature methods* 3.10 (2006), pp. 793–795. DOI: 10.1038/nmeth929.

[31] Mark Schmidt et al. "Optimizing Costly Functions with Simple Constraints: A Limited-Memory Projected Quasi-Newton Algorithm". In: *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*. Vol. 5. Proceedings of Machine Learning Research. Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA: PMLR, Apr. 2009, pp. 456–463.

[32] D. F. Shanno. "Conditioning of quasi-Newton methods for function minimization". In: *Mathematics of Computation* 24.111 (July 1970), pp. 647–656. DOI: https://doi.org/10.1090/S0025-5718-1970-0274029-X.

[33] L.A. Shepp and Y. Verdi. "Maximum Likelihood Reconstruction Emission Tomography". In: *IEEE Transactions on Medical Imaging* 1 (1982), pp. 113–121.

[34] Donald L. Snyder, Abed M. Hammoud, and Richard L. White. "Image recovery from data acquired with a charge-coupled-device camera". In: *J. Opt. Soc. Am. A* 10.5 (May 1993), pp. 1014–1023. DOI: 10.1364/JOSAA.10.001014.

[35] R. Zanella et al. "Towards Real-time Image Deconvolution: Application to Confocal and STED Microscopy". In: *Scientific Reports* 3.2532 (Aug. 2013). DOI: 10.1038/srep02523.

[36] R Zanella et al. "Corrigendum: Efficient gradient projection methods for edge-preserving removal of Poisson noise". In: *Inverse Problems* 29.11 (Sept. 2013), p. 119501. DOI: 10.1088/0266-5611/29/11/119501.